Information Geometric Aspects of Learning Systems

Dissertation (cumulative) submitted
to the Doctoral Degree Committee of
Hamburg University of Technology
in pursuit of the academic degree of

Doktor der Naturwissenschaften (Dr. rer. nat.)

written by

Jesse van Oostrum

from

Hilversum, Netherlands

2026

Supervisor: Prof. Dr. Nihat Ay

Contents

I Introduction **1**

1 Introduction **3**
1.1 Setting the stage . . . 3
1.2 Generative models and the two problems of perception . . . 4
1.3 Recognition models . . . 5
1.4 Learning: the natural gradient . . . 6
1.5 Active inference and the free energy principle . . . 7
1.6 Information geometry as a unifying language . . . 8
1.7 Overview of the papers . . . 9
Prelude to Chapter 2 . . . 11

2 Introduction to Information Geometry **15**
2.1 Information geometry setup . . . 16
2.2 Cramér–Rao Lower Bound . . . 20
2.3 Sufficient statistics and Chentsov's theorem . . . 31
2.4 Appendix . . . 41

II Papers **45**

3 Bures–Wasserstein Geometry for Positive-Definite Hermitian Matrices and Their Trace-One Subset **47**
3.1 Introduction . . . 48
3.2 Preliminaries . . . 49
3.2.1 Differential geometry . . . 49
3.2.2 Representations of the tangent space . . . 50
3.2.3 Matrix identities . . . 51
3.3 Classical information geometry . . . 51
3.3.1 Hellinger distance, Fisher metric and Fisher distance . . . 52
3.4 Quantum information geometry . . . 52
3.5 Bures-Wasserstein Geometry . . . 54
3.5.1 Geometry on the space $\mathbb{P}(n)$ . . . 56
3.5.2 Geometry on the space $\mathcal{D}(n)$ . . . 59
3.6 BW geometry and quantum information . . . 61

4 Invariance Properties of the Natural Gradient in Overparametrised Systems **67**
4.1 Introduction . . . 67

4.2 The natural gradient . . . 68
4.3 Invariance properties of the natural parameter gradient . . . 71
4.3.1 Parametrisation dependence and reparametrisation invariance on the model . . . 71
4.3.2 Reparametrisation (in)variance on the parameter space . . . 74
4.4 Practical considerations for the natural gradient method . . . 76
4.5 Conclusion . . . 78
4.A Appendix . . . 78
4.A.1 On the limitation of Definition 4.4 . . . 78
4.A.2 Example calculation of parametrisation dependence on the model 79
4.A.3 Example calculation of reparametrisation (in)variance on the parameter space . . . 80

5 Inversion of Bayesian Networks 83
5.1 Introduction . . . 83
5.1.1 Example . . . 85
5.2 Notation . . . 85
5.2.1 Graph theory . . . 85
5.2.2 Probability on graphs . . . 87
5.3 Problem statement . . . 88
5.4 Preliminaries . . . 89
5.5 Results . . . 90
5.5.1 Equivalence of two goals . . . 90
5.5.2 Conditions in terms of d-separation . . . 91
5.5.3 Conditions in terms of perfectness . . . 91
5.5.4 Conditions in terms of single edge operations . . . 97
5.5.5 Restricting the set of possible kernel functions . . . 98

6 On the Natural Gradient of the Evidence Lower Bound 103
6.1 Introduction . . . 103
6.2 Learning a Target Distribution and the Evidence Lower Bound . . . 106
6.3 Information-Geometric Preliminaries . . . 108
6.4 The Extended Problem with Hidden Units . . . 113
6.5 The Natural Gradient of the Evidence Lower Bound . . . 117
6.6 Simplification of the Learning in the Extended Space . . . 124
6.A Examples of Cylindrical and Non-Cylindrical Models . . . 126
6.B The Natural Versus the Euclidean Gradient of the Variational Gap . . . 132
6.C Proof of Proposition 6.2 . . . 135

7 A Concise Mathematical Description of Active Inference in Discrete Time 139
7.1 Set-up and notation . . . 140
7.2 Inference . . . 141
7.2.1 Action selection according to active inference . . . 141
7.2.2 Expected free energy . . . 141
7.2.3 State inference . . . 142
7.3 Learning . . . 144
7.4 Example: T-maze example . . . 145
7.A Variational free energy minimization . . . 149

7.B Learning Preliminaries 153
7.C Further details 157

Part I

Introduction

Chapter 1

Introduction

1.1 Setting the stage

Please close your eyes for a few seconds... Did you notice that while your eyes were closed you still knew what objects were around you, where the light in the room was coming from, what the world would look like when you opened your eyes again? There is, in effect, a simplified copy of the outside world in your brain that gets updated every time you receive new sensory information. You do not directly experience the outside world, but rather it is this copy (or model) that determines your experience. How the brain creates this model has been a guiding question throughout my PhD. It is a question that sits at the intersection of neuroscience, philosophy of mind, and mathematics. The work of Metzinger (2009), and in particular his account of self-models and the constructed nature of subjective experience, was one early influence on how I came to think about perception. The brain, on this view, is not a mirror but a modeling device, continuously generating hypotheses about the causes of its own sensory inputs (Clark, 2013).

The mathematical ideas developed to formalize this picture have, in recent years, become central to artificial intelligence. Generative models — probabilistic models that learn to synthesize realistic images, text, audio, and video — are now among the most widely deployed and studied architectures in machine learning (Goodfellow et al., 2016; LeCun et al., 2015). Large language models (Brown et al., 2020), diffusion models (Ho et al., 2020), and variational autoencoders (Kingma and Welling, 2014) are all instances, at different levels of abstraction, of systems that learn an internal model of a data-generating process. The mathematical objects studied in this dissertation — the evidence lower bound, the natural gradient, variational inference, probabilistic graphical models — are precisely the theoretical foundations on which many of these architectures rest. Understanding them rigorously is therefore not only a question for theoretical neuroscience or philosophy of mind, but one with direct relevance to how modern AI systems are designed, trained, and understood.

I would like to state that the five papers comprising this dissertation do not directly answer questions about the brain or consciousness. They are mathematical in character, and their contributions belong to the fields of information geometry, probabilistic graphical models, and variational inference. What the brain perspective provides is not a topic but a vantage point — a set of questions that give the mathematical objects in

these papers their meaning and motivation. A reader looking for neuroscientific claims will not find them here. A reader asking why certain mathematical structures matter for understanding intelligent behavior in the world will, we hope, find the following introduction useful.

The conceptual setting for this dissertation can be described compactly through the idea of the *sensorimotor loop*: an agent, embedded in an environment, receives sensory inputs and produces actions that in turn change the state of the environment, which produces new sensory inputs, and so on. This loop provides the minimal structure within which questions of perception, learning, and action can be posed precisely. The papers in this dissertation are concerned with different aspects of what happens inside the agent — how it represents the world, how it learns from experience, and how it selects actions.

1.2 Generative models and the two problems of perception

The dominant mathematical framework for formalizing the above picture is that of *generative models*. Rather than attempting to learn a direct mapping from observations to interpretations, a generative model works in the opposite direction: it specifies how latent or internal states of the agent give rise to observable sensory signals. Formally, a generative model is a parametrized family of joint probability distributions over latent variables z and observed variables x, written $p_\theta(x, z)$. The model encodes the agent's beliefs about the causal structure of the world: what kinds of internal states exist, and what observations each internal state is likely to produce (Bishop, 2006; Murphy, 2012).

This framework is motivated by a simple intuition that goes back at least to Helmholtz, who argued that perception is a form of unconscious inference — the brain actively infers the most probable cause of its sensory signals rather than passively registering them (Helmholtz, 1867). The same idea is expressed memorably by Richard Feynman — "What I cannot create, I do not understand" — and echoed in Hinton et al. (1995)'s remark that to recognize shapes, one should first learn to generate images. A model that can generate plausible sensory data has, in some sense, understood the structure that underlies it.

Two distinct problems arise within this framework, and they play different roles throughout this dissertation. The first is *learning*: given a collection of observed data, how should the parameters θ of the generative model be adjusted so that the model accurately reflects the statistical structure of the world? This is an optimization problem, and the choice of optimization algorithm matters greatly. A central theme in several of the papers here is the natural gradient — a principled modification of ordinary gradient descent that accounts for the intrinsic geometry of the space of probability distributions, rather than treating parameter space as if it were Euclidean (Amari, 1998). In practice, computing the true learning objective requires marginalizing over the latent variables, which is generally intractable. One therefore optimizes a tractable surrogate — the evidence lower bound (ELBO) — instead (Jordan et al., 1999; Wainwright and Jordan, 2008). The ELBO is the central training objective of the *variational autoencoder* (VAE) (Kingma and Welling, 2014; Rezende et al., 2014), a generative ar-

chitecture that has had enormous influence on modern deep learning and that underlies many contemporary generative AI systems. The paper *On the Natural Gradient of the Evidence Lower Bound* (Ay et al., 2025) studies the natural gradient of precisely this objective, and investigates under what geometric conditions optimizing the ELBO is equivalent to optimizing the true learning objective.

The second is *inference*: given a trained generative model and a new observation x, how should the agent determine which internal state z most plausibly gave rise to it? This requires computing or approximating the posterior distribution $p_\theta(z \mid x)$ — the distribution over latent states conditioned on the observation. Inference is generally intractable for models of real-world complexity (Koller and Friedman, 2009), and a substantial body of work in machine learning is devoted to principled approximations (Blei et al., 2017). Two broad strategies exist: non-amortized inference, in which the posterior is approximated fresh for each new observation, and amortized inference, in which a separate recognition model $q_\phi(z \mid x)$ is trained to produce posterior approximations directly (Gershman and Goodman, 2014). The latter approach is the basis of variational autoencoders and related architectures (Kingma and Welling, 2014).

1.3 Recognition models and inversion of Bayesian networks

Amortized inference introduces a new question that is both practically and theoretically interesting: given a generative model with a particular graphical structure — a Bayesian network specifying which variables influence which — what structure should the recognition model have in order to be capable of exactly representing the posterior distribution? This is not merely a question about approximation quality. It is a structural question: are there configurations of the recognition network that make exact posterior inference impossible in principle, regardless of the parameter values chosen? Conversely, what is the minimal architecture that guarantees exact representability?

The paper *Inversion of Bayesian Networks* (Oostrum et al., 2024) addresses precisely this question. The generative model is a directed acyclic graph (DAG) specifying a Bayesian network, and the recognition model is another DAG whose structure must be chosen such that the family of distributions it can represent contains all posterior distributions of the generative model. The paper provides necessary and sufficient conditions — expressed in terms of the graphical structure of the two networks — for this containment to hold. In the special case of a single-root generative network, the conditions coincide; for the general case of multiple roots, a gap remains between necessity and sufficiency, and characterizing this gap is identified as a direction for further work.

The problem can be understood intuitively as follows. The generative model encodes a particular set of conditional independence assumptions about the latent variables, as expressed through the d-separation structure of its graph. The posterior distribution, in general, violates some of these independence assumptions — conditioning on observations can introduce dependencies between previously independent latent variables, a phenomenon sometimes called *explaining away* (Pearl, 1988). For the recognition model to exactly capture the posterior, its graph must be rich enough to represent these induced dependencies. The paper makes this intuition precise and provides an

explicit procedure for constructing a valid recognition network from a given generative network.

1.4 Learning: the natural gradient

Once a generative model has been specified, the agent must fit its parameters to observed data. The standard approach is gradient descent: compute the gradient of a loss function with respect to the parameters and take a step in the direction that improves it. This is a sensible procedure, but it carries an implicit assumption — that parameter space is Euclidean, meaning that all directions and all step sizes are treated equally regardless of what they correspond to on the model. This assumption is rarely justified. Two parameter vectors that are close in Euclidean distance may correspond to probability distributions that are very different, and vice versa.

A more principled approach is to equip the model — the space of probability distributions — with a geometry that reflects the statistical structure of the problem, and then to follow the gradient with respect to that geometry. The resulting update direction is called the *natural gradient*. In information geometry, the canonical choice of geometry on a space of probability distributions is the one induced by the Fisher metric, which is studied in detail in Chapter 2. The natural gradient method was first proposed by Amari (1998) and has since been shown to be effective across a wide range of applications, including neural network training (Martens and Grosse, 2015) and reinforcement learning (Kakade, 2002). A key property that motivates its use is *parametrization invariance*: because the natural gradient lives on the model rather than on parameter space, it does not depend on the particular parametrization chosen. Two researchers who parametrize the same model differently should, in principle, follow the same update direction on the model.

In practice, however, the situation is more subtle. The natural gradient on the model must be pulled back to parameter space in order to actually update the parameters, and this step is where invariance can break down. In particular, modern machine learning models are frequently overparametrized — the map from parameter space to the model is not injective, meaning that different parameter vectors correspond to the same point on the model. In this setting, the Gram matrix associated with the Fisher metric is degenerate, and one must resort to a generalized inverse to compute the natural parameter gradient. The paper *Invariance Properties of the Natural Gradient in Overparametrized Systems* (Oostrum et al., 2023) studies precisely this situation. It identifies two distinct notions of invariance and determines which of them survive in the overparametrized regime, clarifying when the natural parameter gradient can be trusted to behave as the natural gradient it is intended to approximate.

A further subtlety arises when the natural gradient is applied to the ELBO, as introduced in § 1.2. The ELBO is defined on the full joint model over both visible and hidden variables, whereas the true learning objective — minimizing the KL divergence between the data distribution and the model — concerns only the marginal over visible variables. These two models carry different geometries, and the natural gradient computed on the full model need not agree with the natural gradient of the true objective on the marginal model. The paper *On the Natural Gradient of the Evidence Lower Bound* studies exactly this relationship. It shows that the relevant question is not whether the

variational gap is large in magnitude, but whether its natural gradient vanishes — in which case the gap has no effect on the learning trajectory. The geometric condition that ensures agreement between the two natural gradients is the cylindricity of the model, connecting this paper directly to the discussion of Riemannian submersions in § 1.6.

1.5 Active inference and the free energy principle

The framework of generative models and variational inference describes how an agent can learn a model of the world and use it to interpret sensory input. A natural extension of this framework asks a further question: how should the agent *act*? Active inference, developed by Friston (2010) and collaborators, provides a framework for treating perception, learning, and action within a single objective by proposing that agents minimize a quantity called the *variational free energy* (Friston et al., 2006; Friston et al., 2017).

The free energy is equal to the negative of the ELBO introduced in the previous section. Minimizing it with respect to the internal states of the agent corresponds to inference — the agent updates its beliefs to better explain its sensory input. Minimizing it with respect to the parameters of the generative model corresponds to learning — the agent adjusts its internal model to better reflect the structure of the world. Minimizing it with respect to actions corresponds to a form of planning — the agent selects actions that are expected to lead to observations consistent with its preferences. In this way, active inference subsumes learning, inference, and action selection within a single variational objective, and connects naturally to the sensorimotor loop described in § 1.1.

The free energy and the quantities that arise from minimizing it are deeply information-theoretic in nature. The KL divergence, a central object in information geometry that measures the dissimilarity between two probability distributions, appears explicitly in the decomposition of the free energy. Other quantities arising in active inference, such as ambiguity and risk, can similarly be interpreted in information-geometric terms. This makes active inference a natural application domain for the tools developed throughout this dissertation.

Active inference has attracted substantial and growing interest across multiple communities. As a theory of brain function it has been applied to a wide range of behavioral and neuroscientific phenomena, including attention (Feldman and Friston, 2010), action and motor control (Friston et al., 2010), psychiatric conditions (Adams et al., 2013), and perceptual learning (Friston and Kiebel, 2009). Beyond neuroscience, the framework has been taken up in robotics and artificial agents (Lanillos et al., 2021), making it one of the more ambitious attempts to unify biological and artificial intelligence under a single normative principle. The number of publications on active inference has grown rapidly over the past decade, and dedicated software toolboxes, tutorials, and a textbook (Parr et al., 2022) now exist to support the community. The mathematical formulation of the theory, however, has often been presented in a way that makes it difficult to follow the derivations closely. The paper *A Concise Mathematical Description of Active Inference in Discrete Time* (Oostrum et al., 2025) addresses this directly. It presents the theory in discrete time using precise and standard mathematical notation, with the goal of making the derivations accessible to readers with a

mathematical background. The paper covers the inference, learning, and action selection components of active inference and includes a detailed worked example, as well as a Python implementation compatible with standard active inference environments.

1.6 Information geometry as a unifying language

The concepts introduced in the preceding sections — the Fisher metric, the natural gradient, the KL divergence, the ELBO — do not belong to separate mathematical traditions. They are all instances of a single underlying framework: *information geometry*, the study of geometric structures on spaces of probability distributions.

The origins of the field are usually traced to a 1945 paper by C. R. Rao (Rao, 1945), who observed that the Fisher information matrix defines a Riemannian metric on parametric families of probability distributions. This observation went largely unnoticed for several decades, until Amari (1998) and Chentsov (1982) developed it into a systematic theory in the 1970s and 80s. Chentsov showed that the Fisher metric is, in a precise sense, the unique Riemannian metric on the probability simplex that is compatible with the natural symmetries of statistics. Amari (1998) introduced the dual geometric structure — a pair of conjugate connections that give information geometry much of its richness — and connected the framework to statistical estimation, neural networks, and learning theory. A comprehensive modern treatment can be found in the monograph by Ay et al. (2017).

Several of the papers in this dissertation are directly concerned with information geometry. The paper on Bures–Wasserstein geometry (Oostrum, 2022) studies the geometric properties of a distance measure on the space of positive-definite Hermitian matrices, establishing its Riemannian structure via a Riemannian submersion and drawing an explicit analogy with the classical Fisher geometry on probability distributions. The papers on the natural gradient are concerned with the geometry induced by the Fisher metric on spaces of probability distributions, and in particular with how this geometry interacts with parametrization. The paper on the natural gradient of the ELBO applies these ideas to a concrete optimization problem arising in variational inference. The paper on active inference uses information-geometric quantities — KL divergence, free energy — as its primary building blocks, even if it does not develop the geometric theory explicitly.

A geometric construction that plays a recurring role across several of these papers is the *Riemannian submersion*. Intuitively, a Riemannian submersion is a map between two Riemannian manifolds that preserves the lengths of horizontal curves — those that move perpendicularly to the fibers of the map. This makes it a useful tool for relating the geometry of a larger or more tractable space to that of a quotient or reduced space, carrying over geodesics and distance functions in a controlled way. In the Bures–Wasserstein paper, a Riemannian submersion is used to derive the geometry of the space of positive-definite matrices from that of the general linear group. In the natural gradient of the ELBO paper, the same defining condition of a Riemannian submersion reappears in a different guise, through the notion of a *cylindrical model*. A statistical model is called cylindrical if, at every non-singular point, its tangent space decomposes as a direct sum of its intersections with the horizontal and vertical subspaces of the ambient space. This is precisely the local condition that a Riemannian submersion

requires to hold globally, meaning that cylindrical models can be thought of as the statistical analog of a Riemannian submersion, relaxed to hold only at non-singular points. The significance of this condition is that it guarantees the invariance of the natural gradient under the marginalization map — that is, gradient descent on the full model and gradient descent on the marginal model yield the same trajectories. Chapter 2 develops the necessary geometric background, including a careful treatment of Riemannian submersions and their role in information geometry.

1.7 Overview of the papers

The five papers that form the core of this dissertation span two interconnected themes: the geometry of spaces of probability distributions, and the mathematical foundations of learning and inference in generative models. The following paragraphs describe each paper in turn, indicating where it fits within the narrative developed above and what its main contributions are.

Jesse van Oostrum. "Bures–Wasserstein geometry for positive-definite Hermitian matrices and their trace-one subset." ***Information Geometry.*** **Springer, 2022.**
This paper was the first research work of the dissertation and, in many ways, the author's introduction to information geometry. It studies the geometric properties of the Bures–Wasserstein (BW) distance on the space of positive-definite Hermitian matrices — a distance measure that appears independently in quantum information theory (as the Bures distance between quantum states) and in optimal transport (as the Wasserstein distance between Gaussian distributions). The central argument follows the classical derivation of Rao: just as the Fisher metric on probability measures can be obtained by pulling back the Euclidean metric via a square-root map, the BW metric on positive-definite matrices is derived by viewing the space as a quotient of the general linear group under a Riemannian submersion. The paper then restricts this geometry to the subset of trace-one matrices — the quantum analog of the probability simplex — and derives the corresponding geodesics and Riemannian distance. The close structural parallel between the classical and quantum cases is a central theme.

Jesse van Oostrum, Johannes Müller, Nihat Ay. "Invariance properties of the natural gradient in overparametrised systems." ***Information Geometry.*** **Springer, 2023.**
This paper studies the natural gradient method in the setting of overparametrized models — the regime relevant to modern deep learning, where the map from parameter space to the space of distributions is not injective. The natural gradient is defined on the model, independently of any parametrization, and represents the direction of steepest ascent of an objective function with respect to the Fisher metric. In practice, it must be approximated by a vector on parameter space, obtained by multiplying the ordinary gradient by the inverse of the Gram matrix of the metric. In the overparametrized case, the Gram matrix is degenerate and a generalized inverse must be used, which raises the question of whether the resulting natural parameter gradient retains the invariance properties of the natural gradient it approximates. The paper introduces two distinct notions of invariance and determines which of them hold in the

overparametrized setting, providing a theoretical foundation for understanding when and why the natural gradient method can be expected to behave well in practice.

Jesse van Oostrum, Peter van Hintum, Nihat Ay. "Inversion of Bayesian networks." ***International Journal of Approximate Reasoning.*** **Elsevier, 2024.**
When an agent has learned a generative model of the world, it faces the problem of inverting it: given a sensory observation, which internal state most plausibly caused it? This paper addresses the structural question at the heart of amortized inference: given a generative Bayesian network, what architecture must a recognition network have in order to be capable of exactly representing the posterior distribution? The paper provides necessary and sufficient conditions for this, expressed in terms of the d-separation structure of the two directed acyclic graphs. When the generative network has a single root node, the conditions coincide and a complete characterization is obtained. For networks with multiple roots, the paper provides a sufficient condition and a weaker necessary condition, with the gap between them identified as a direction for further work. The paper also develops an explicit construction procedure — the inversion procedure — that, given a generative network satisfying a certain condition, produces a recognition network capable of exact posterior inference.

Nihat Ay, Jesse van Oostrum, Adwait Datar. "On the natural gradient of the evidence lower bound." ***Journal of Machine Learning Research.*** **2025.**

This paper sits at the intersection of the two main themes of the dissertation: it applies the natural gradient — a tool from information geometry — to the ELBO, the central optimization objective in variational inference and the variational autoencoder framework. The starting point is a fundamental tension in variational learning. The true learning objective is to minimize the KL divergence between the target distribution and the marginal of the generative model over the visible variables. This marginal is typically intractable, which is why the ELBO — a lower bound on the log-evidence — is used as a surrogate instead. The difference between the two, known as the *variational gap*, can in principle be large, raising the concern that optimizing the ELBO may drive learning in a different direction than optimizing the true objective. The paper argues that this concern should be assessed not by looking at the magnitude of the variational gap itself, but by examining the geometry of its gradient. If the natural gradient of the variational gap vanishes, then the gap is irrelevant to the learning trajectory, regardless of its size. The geometric condition that ensures agreement between the two natural gradients is the cylindricity of the model, connecting this paper directly to the discussion in § 1.6.

Jesse van Oostrum, Carlotta Langer, Nihat Ay. "A concise mathematical description of active inference in discrete time." ***Journal of Mathematical Psychology.*** **Elsevier, 2025.**
This paper presents the active inference framework — a theory of perception, learning, and action rooted in the free energy principle — in a mathematically precise and self-contained form. The existing active inference literature contains a number of derivations that are difficult to follow due to non-standard notation and implicit assumptions. The paper addresses this by reformulating the theory using standard

mathematical notation, with all dependencies made explicit. It covers the inference and learning mechanisms, includes a detailed worked example of the action selection procedure, and provides a Python implementation compatible with standard active inference environments. The paper can be read as both a contribution to the clarification of the active inference literature and as a demonstration of how the mathematical tools developed elsewhere in this dissertation — variational objectives, information-theoretic quantities, probabilistic graphical models — come together in a unified framework for intelligent behavior.

Prelude to Part I, Chapter 2

The following chapter provides a self-contained introduction to information geometry, aimed at readers with a background in Riemannian geometry. It was written with a specific goal in mind: to introduce as little machinery as possible while still arriving at results that demonstrate the genuine power of the framework. Central to information geometry is the idea that a parametric family of probability distributions can be equipped with a Riemannian metric — the *Fisher metric* — turning it into a geometric object amenable to the tools of differential geometry. This is not an arbitrary choice: the chapter motivates the Fisher metric through two complementary results, each of which singles it out as the canonical geometric structure on a statistical model. The first is the Cramér–Rao lower bound, which shows that the Fisher metric governs the fundamental limits of statistical estimation. The second is Chentsov's theorem, discussed below, which characterizes the Fisher metric as the unique Riemannian metric invariant under sufficient statistics.

Much of the existing literature on information geometry — including the foundational texts of Amari and Nagaoka (2000) and of Ay et al. (2017) — introduces a substantial amount of structure before any application becomes visible. This is appropriate for a comprehensive treatment, but can make it difficult for a newcomer to understand why the framework is worth the effort. The present chapter takes a different approach: it develops the Fisher metric and its properties from first principles, works through concrete examples at each step, and culminates in a proof of the Cramér–Rao lower bound — one of the most elegant results in mathematical statistics, and in the author's view the clearest illustration of what information geometry is for. The bound quantifies the minimum variance any unbiased estimator can achieve, and as such provides a natural standard against which learning algorithms can be assessed. The natural gradient method, which features in several of the papers here, was shown by Amari (1998) to be Fisher efficient — meaning its asymptotic variance achieves precisely this bound. The Cramér–Rao bound thus serves a dual purpose in this chapter: it illustrates, concretely, what the geometric machinery is good for, and it connects the introductory material to one of the central tools of the dissertation.

Its proof, in the information-geometric formulation given here, amounts to an application of the Cauchy–Schwarz inequality on the tangent space of the statistical manifold — a fact that is almost obvious once the geometric picture is in place, but far from obvious without it. The proof can also be found in the monograph of Amari and Nagaoka (2000) and in the book of Ay et al. (2017), but in the former the notation is scattered and the argument difficult to follow in one place, while in the latter the result is established in considerable generality, which comes at the cost of accessibility. The

treatment here aims to occupy a middle ground: complete and rigorous, but focused on the finite discrete case where the essential ideas are already present.

The chapter also develops a second line of justification for the Fisher metric, complementary to the Cramér–Rao bound, through the study of *sufficient statistics* and *Chentsov's theorem* (Chentsov, 1982; Ay et al., 2017). A statistic — a function from the sample space to some other space — is called *sufficient* for a parametric model if it retains all information about the parameter that is present in the data: once the value of the statistic is known, the raw observations provide no further evidence about the parameter. Chentsov's theorem states that the Fisher metric is, up to a constant, the *unique* Riemannian metric on the probability simplex that is invariant under every sufficient statistic. This gives the Fisher metric a canonical status that goes beyond its role in estimation theory: it is the only metric compatible with the natural notion of lossless data reduction in statistics.

The chapter closes with a reformulation of Chentsov's theorem in the language of Riemannian submersions. As discussed in § 1.6, Riemannian submersions play a recurring role in this thesis. The reformulation establishes that the invariance condition in Chentsov's theorem is equivalent to asking that every statistic induces a Riemannian submersion from the source probability simplex to the target one. This restatement places Chentsov's theorem within the same geometric framework that appears in the Bures–Wasserstein paper and in the paper on the natural gradient of the ELBO, and makes the connection between statistical sufficiency and the broader submersion geometry of the dissertation explicit.

References

Adams, Rick A., Stewart Shipp, and Karl J. Friston (2013). "Predictions Not Commands: Active Inference in the Motor System". In: *Brain Structure and Function* 218.3, pp. 611–643.

Amari, Shun-ichi (1998). "Natural Gradient Works Efficiently in Learning". In: *Neural Computation* 10.2, pp. 251–276.

Amari, Shun-ichi and Hiroshi Nagaoka (2000). *Methods of Information Geometry*. Vol. 191. Translations of Mathematical Monographs. Providence, RI: American Mathematical Society.

Ay, Nihat, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer (2017). *Information Geometry*. Vol. 64. Ergebnisse der Mathematik und ihrer Grenzgebiete. Cham: Springer.

Ay, Nihat, Jesse van Oostrum, and Adwait Datar (2025). "On the Natural Gradient of the Evidence Lower Bound". In: *Journal of Machine Learning Research*.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe (2017). "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518, pp. 859–877.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems* 33, pp. 1877–1901.

Chentsov, Nikolai Nikolaevich (1982). *Statistical Decision Rules and Optimal Inference*. Vol. 53. Translations of Mathematical Monographs. Providence, RI: American Mathematical Society.

Clark, Andy (2013). "Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science". In: *Behavioral and Brain Sciences* 36.3, pp. 181–204.

Feldman, Harriet and Karl Friston (2010). "Attention, Uncertainty, and Free-Energy". In: *Frontiers in Human Neuroscience* 4, p. 215.

Friston, Karl (2010). "The Free-Energy Principle: A Unified Brain Theory?" In: *Nature Reviews Neuroscience* 11.2, pp. 127–138.

Friston, Karl, Jean Daunizeau, James Kilner, and Stefan J. Kiebel (2010). "Action and Behavior: A Free-Energy Formulation". In: *Biological Cybernetics* 102.3, pp. 227–260.

Friston, Karl, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo (2017). "Active Inference: A Process Theory". In: *Neural Computation* 29.1, pp. 1–49.

Friston, Karl and Stefan Kiebel (2009). "Predictive Coding under the Free-Energy Principle". In: *Philosophical Transactions of the Royal Society B* 364.1521, pp. 1211–1221.

Friston, Karl, James Kilner, and Lee Harrison (2006). "A Free Energy Principle for the Brain". In: *Journal of Physiology-Paris* 100.1–3, pp. 70–87.

Gershman, Samuel J. and Noah D. Goodman (2014). "Amortized Inference in Probabilistic Reasoning". In: *Proceedings of the Annual Conference of the Cognitive Science Society* 36.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Helmholtz, Hermann von (1867). *Handbuch der Physiologischen Optik*. Leipzig: Voss.

Hinton, Geoffrey E., Peter Dayan, Brendan J. Frey, and Radford M. Neal (1995). "The "Wake-Sleep" Algorithm for Unsupervised Neural Networks". In: *Science* 268.5214, pp. 1158–1161.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems* 33, pp. 6840–6851.

Jordan, Michael I., Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul (1999). "An Introduction to Variational Methods for Graphical Models". In: *Machine Learning* 37.2, pp. 183–233.

Kakade, Sham (2002). "A Natural Policy Gradient". In: *Advances in Neural Information Processing Systems* 14.

Kingma, Diederik P. and Max Welling (2014). "Auto-Encoding Variational Bayes". In: *International Conference on Learning Representations*.

Koller, Daphne and Nir Friedman (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press.

Lanillos, Pablo, Cristian Meo, Corrado Pezzato, Ajith Anil Meera, Mohamed Baioumy, Wataru Ohata, Alexander Tschantz, Beren Millidge, Martijn Wisse, Christopher L Buckley, et al. (2021). "Active inference in robotics and artificial agents: Survey and challenges". In: *arXiv preprint arXiv:2112.01871*.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep Learning". In: *Nature* 521.7553, pp. 436–444.

Martens, James and Roger Grosse (2015). "Optimizing Neural Networks with Kronecker-Factored Approximate Curvature". In: *Proceedings of the International Conference on Machine Learning*, pp. 2408–2417.

Metzinger, Thomas (2009). *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. New York: Basic Books.

Murphy, Kevin P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.

Oostrum, Jesse van (2022). "Bures–Wasserstein Geometry for Positive-Definite Hermitian Matrices and Their Trace-One Subset". In: *Information Geometry* 5.2, pp. 405–425.

Oostrum, Jesse van, Peter van Hintum, and Nihat Ay (2024). "Inversion of Bayesian Networks". In: *International Journal of Approximate Reasoning* 164.

Oostrum, Jesse van, Carlotta Langer, and Nihat Ay (2025). "A Concise Mathematical Description of Active Inference in Discrete Time". In: *Journal of Mathematical Psychology*.

Oostrum, Jesse van, Johannes Müller, and Nihat Ay (2023). "Invariance Properties of the Natural Gradient in Overparametrised Systems". In: *Information Geometry* 6.1, pp. 51–67.

Parr, Thomas, Giovanni Pezzulo, and Karl J. Friston (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. Cambridge, MA: MIT Press.

Pearl, Judea (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA: Morgan Kaufmann.

Rao, C. Radhakrishna (1945). "Information and the Accuracy Attainable in the Estimation of Statistical Parameters". In: *Bulletin of the Calcutta Mathematical Society* 37, pp. 81–91.

Rezende, Danilo Jimenez, Shakir Mohamed, and Daan Wierstra (2014). "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". In: *Proceedings of the International Conference on Machine Learning*, pp. 1278–1286.

Wainwright, Martin J. and Michael I. Jordan (2008). *Graphical Models, Exponential Families, and Variational Inference*. Vol. 1. Foundations and Trends in Machine Learning 1–2. Hanover, MA: Now Publishers.

Chapter 2

Introduction to Information Geometry

This chapter gives a self-contained introduction to information geometry and its application to classical problems in statistics. The central idea, due to Rao (1945) and developed extensively by Chentsov (1982), Amari and Nagaoka (2000), and others, is that a family of probability distributions can be viewed as a Riemannian manifold, equipped with a canonical metric—the *Fisher metric*—that organizes familiar statistical concepts in geometric terms. The chapter has a deliberate two-fold purpose. On the one hand, the Fisher metric is motivated through two complementary classical results: the Cramér–Rao lower bound, which casts it as the geometric object governing the precision of parameter estimation, and Chentsov's theorem, which characterizes it as the unique Riemannian metric invariant under sufficient statistics. On the other hand, information geometry is itself put to work, supplying a transparent geometric proof of the Cramér–Rao bound. In line with the broader theme of this thesis, particular attention is paid to the role of *Riemannian submersions*, in terms of which we also provide an alternative formulation of Chentsov's theorem. The reader is expected to be familiar with basic concepts of Riemannian geometry, which can be found in Lee (2018).

Section 2.1 develops the geometric framework on the space of strictly positive measures over a finite state space, introducing the Fisher metric and the geometry of parametrized families. Section 2.2 applies the framework to the problem of parameter estimation. The main result is a geometric proof of the Cramér–Rao lower bound (Cramér, 1946; Rao, 1945), which states that the variance of any unbiased estimator is bounded below by the inverse of the Fisher information. In the geometric picture, this bound reduces to a simple inequality on the tangent space of the statistical manifold, making both its content and its proof considerably more transparent than in the classical treatment.

Section 2.3 studies maps between sample spaces, called *statistics*, which naturally induce maps between spaces of probability distributions. Of particular interest are *sufficient* statistics: those that retain all information about the parameter that is present in the original data. We show that sufficiency has a clean characterization in terms of the orthogonal decomposition of tangent vectors, and that the induced maps are Riemannian submersions with respect to the Fisher metric. This sets the stage for

Chentsov's theorem (Chentsov, 1982; Ay et al., 2017), which singles out the Fisher metric (up to a positive constant) as the unique Riemannian metric invariant under sufficient statistics. We close the chapter with a reformulation of the theorem entirely in the language of Riemannian submersions.

2.1 Information geometry setup

We consider a finite and non-empty state space $\mathcal{X}$ with typical elements x. We denote the space of strictly positive measures on $\mathcal{X}$ by $\mathcal{M}$ with element μ and the subspace of probability measures by $\mathcal{P}$ with elements p. We can use the Dirac delta measures δ^x to write each measure $\mu \in \mathcal{M}$ as follows:

$$\mu = \sum_{x \in \mathcal{X}} \mu(x) \delta^x,$$

where

$$\delta^x(x') = \begin{cases} 1 & \text{if } x = x', \\ 0 & \text{otherwise.} \end{cases}$$

We then have

$$\mathcal{M} = \{\mu : \mu(x) > 0, \forall x \in \mathcal{X}\},$$
$$\mathcal{P} = \{p \in \mathcal{M} : \sum_x p(x) = 1\}.$$

As an example we can consider the case $\mathcal{X} = \{1, 2\}$. The spaces $\mathcal{M}, \mathcal{P}$ and the measures δ^1, δ^2 are depicted in Figure 2.1.

The tangent space $T_\mu \mathcal{M}$ is spanned by the vectors $\{\partial_x|_\mu\}_{x \in \mathcal{X}}$, defined as follows:

$$\partial_x\big|_\mu V = \partial_x V(\mu) = \lim_{t \downarrow 0} \frac{V(\mu + t\delta^x) - V(\mu)}{t}, \tag{2.1}$$

where $V \colon \mathcal{M} \to \mathbb{R}$ is a smooth function. Equivalently, one can define the curve

$$\gamma^x \colon [0, \epsilon) \to \mathcal{M},$$
$$\gamma^x(t) = \mu + t\delta^x,$$

and let

$$\partial_x\big|_\mu = d\gamma_0^x \left(\left. \frac{d}{dt} \right|_{t=0} \right).$$

We will use $A = \sum_x A(x) \partial_x\big|_\mu$ to denote a generic element of $T_\mu \mathcal{M}$. Now let

$$\gamma^A \colon [0, \epsilon) \to \mathcal{M},$$
$$\gamma^A(t) = \mu + t \sum_x A(x) \delta^x.$$

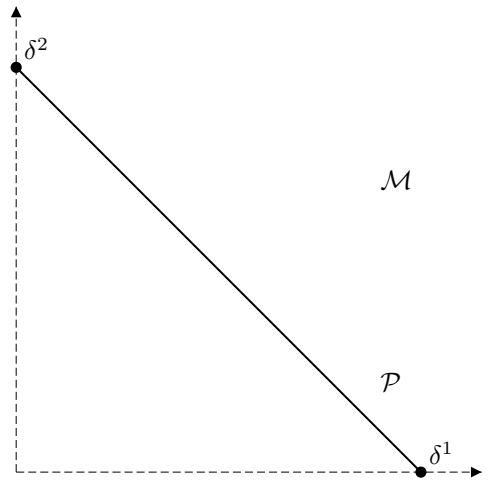


Figure 2.1: The spaces $\mathcal{M}$ and $\mathcal{P}$

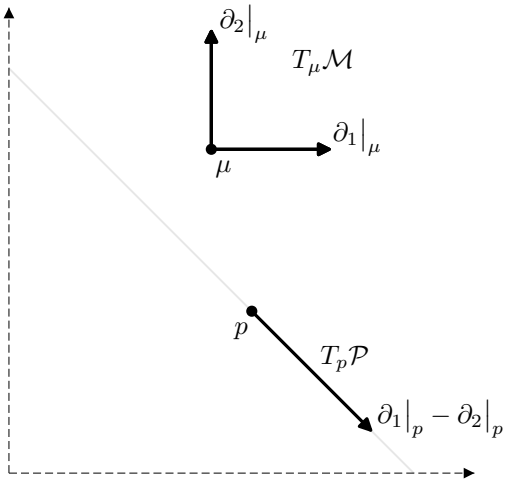


Figure 2.2: The tangent spaces $T_\mu\mathcal{M}$ and $T_p\mathcal{P}$

Note that $d\gamma_0^A\left(\frac{d}{dt}\big|_{t=0}\right) = A$. Furthermore, when γ^A is in $\mathcal{P}$, then $\sum_x \gamma^A(t)(x) = 1$ and therefore the tangent space $T_p\mathcal{P}$ is then given by

$$T_p\mathcal{P} = \left\{ A \in T_p\mathcal{M} : \sum_x A(x) = 0 \right\}.$$

These spaces are depicted in Figure 2.2.

Definition 2.1. For two tangent vectors $A, B \in T_\mu\mathcal{M}$ the *Fisher metric* is given by:

$$\mathfrak{g}_\mu(A, B) = \sum_x \frac{1}{\mu(x)} A(x)B(x). \tag{2.2}$$

We will often suppress the dependence on the measure μ in the notation.

We denote the length of a tangent vector A induced by the Fisher metric

$$\|A\|_\mu = \mathfrak{g}_\mu(A, A)^{1/2}.$$

Remark 2.1. The Fisher metric induces a Riemannian distance on the space of probability measures, which admits a natural statistical interpretation: it measures how easily two distributions can be distinguished on the basis of samples. Intuitively, two distributions that are far apart in Fisher distance are easy to tell apart, whereas distributions that are close together are hard to discriminate. We will make this precise after stating the Cramér–Rao lower bound (Theorem 2.3).
The Fisher metric also has a separate, purely geometric justification: after introducing sufficient statistics, we will state Chentsov's theorem, which characterizes the Fisher metric as the unique Riemannian metric (up to a constant factor) that is invariant under the action of sufficient statistics.

Using (2.2), the (Fisher) gradient of a smooth function $V\colon \mathcal{M} \to \mathbb{R}$ is given by

$$\operatorname{grad}_\mu V = \sum_x \mu(x)\partial_x V(\mu)\partial_x\big|_\mu.$$

For a tangent vector $A \in T_p\mathcal{M}$, $p \in \mathcal{P}$, we have the orthogonal decomposition

$$A^\top = \sum_x \left(A(x) - p(x)\sum_{\tilde{x}} A(\tilde{x})\right)\partial_x\big|_p \in T_p\mathcal{P},$$
$$A^\perp = \sum_x \left(p(x)\sum_{\tilde{x}} A(\tilde{x})\right)\partial_x\big|_p \in T_p\mathcal{P}^\perp.$$

The gradient of the restricted function $V\big|_{\mathcal{P}}$ is given by the projection of $\operatorname{grad}_p V$ onto $T_p\mathcal{P}$ (Lemma 2.3 in the appendix). Therefore for any function $V\colon \mathcal{P} \to \mathbb{R}$ with smooth extension $\tilde{V}$ to $\mathcal{M}$ we have

$$\operatorname{grad}_p V = \sum_x p(x)\left(\partial_x\tilde{V}(p) - \sum_{\tilde{x}} p(\tilde{x})\partial_{\tilde{x}}\tilde{V}(p)\right)\partial_x\big|_p. \tag{2.3}$$

Parametrizations Often we consider a parametrization of a subset of $\mathcal{P}$, which is a map

$$\phi\colon \mathbb{R}^d \supset \Theta \ni \theta \mapsto p_\theta \in \mathcal{P},$$

that is a diffeomorphism onto its image $\mathcal{S} = \phi(\Theta)$. The image space $\mathcal{S}$ is sometimes referred to as the *model*.

For example, we have

$$\begin{aligned}&\phi\colon (0,1) \to \mathcal{P},\\ &\phi(\theta) = p_\theta = (1-\theta)\delta^0 + \theta\delta^1,\end{aligned} \tag{2.4}$$

which we will call the *Bernoulli parametrization*.

We write $\partial_i(\theta) \in T_{p_\theta}\mathcal{P}$ for the pushforward of the tangent vector $\frac{\partial}{\partial\theta^i}\big|_\theta \in T_\theta\Theta$, that is[1] (see Figure 2.3)

$$\begin{aligned}\partial_i(\theta) &= d\phi_\theta\left(\frac{\partial}{\partial\theta^i}\bigg|_\theta\right)\\ &= \sum_x \frac{\partial p_\theta(x)}{\partial\theta^i}\partial_x\big|_{p_\theta}.\end{aligned}$$

[1]Note that the tangent vector ∂_x defined in (2.1) can also be interpreted as the pushforward of a parameter tangent vector by letting $\mathcal{X} = \{1, ..., d\}$ and considering the parametrization

$$\phi\colon \mathbb{R}^d_+ \ni \theta \mapsto p_\theta = \sum_i \theta^i\delta^i.$$

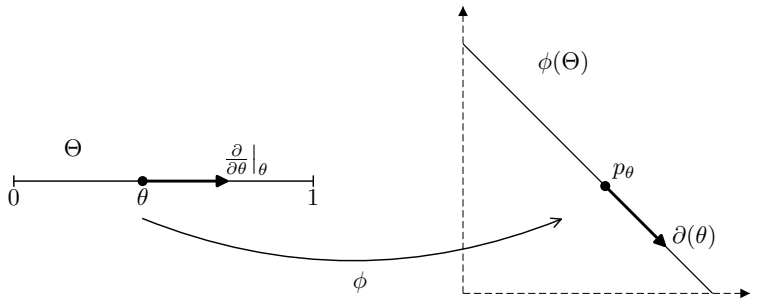


Figure 2.3: The parametrization ϕ and tangent vectors $\frac{\partial}{\partial\theta^i}\big|_\theta$ and $\partial(\theta)$

We can now use ϕ to pull back the Fisher metric defined in (2.2) to a metric g on Θ. For $v = \sum_i v^i \frac{\partial}{\partial\theta^i}\big|_\theta, w = \sum_i w^i \frac{\partial}{\partial\theta^i}\big|_\theta \in T_\theta\Theta$, we have

$$\begin{aligned} g_\theta(v,w) &= \mathfrak{g}(d\phi_\theta(v), d\phi_\theta(w)) \\ &= \mathfrak{g}\left(d\phi_\theta\left(\sum_i v^i \left.\frac{\partial}{\partial\theta^i}\right|_\theta\right), d\phi_\theta\left(\sum_i w^i \left.\frac{\partial}{\partial\theta^i}\right|_\theta\right)\right) \\ &= \sum_{i,j} v^i w^j \mathfrak{g}\left(\partial_i(\theta), \partial_j(\theta)\right) \\ &= \sum_{i,j} v^i w^j G_{ij}(\theta), \end{aligned}$$

where

$$\begin{aligned} G_{ij}(\theta) &= \mathfrak{g}\left(\partial_i(\theta), \partial_j(\theta)\right) \\ &= \mathfrak{g}\left(\sum_x \frac{\partial p_\theta(x)}{\partial\theta^i}\partial_x\big|_p, \sum_x \frac{\partial p_\theta(x)}{\partial\theta^j}\partial_x\big|_p\right) \\ &= \sum_x \frac{1}{p_\theta(x)}\frac{\partial p_\theta(x)}{\partial\theta^i}\frac{\partial p_\theta(x)}{\partial\theta^j} \\ &= \sum_x p_\theta(x)\frac{\partial \log p_\theta(x)}{\partial\theta^i}\frac{\partial \log p_\theta(x)}{\partial\theta^j}. \end{aligned} \tag{2.5}$$

The matrix $G(\theta)$ is referred to as the *Fisher information matrix* (FIM). Another useful form (see Lemma 2.4 in the appendix) of the FIM, which is in general only valid on $\mathcal{P}$ not on $\mathcal{M}$, is given by

$$G_{ij}(\theta) = -\sum_x p_\theta(x)\frac{\partial^2 \log p_\theta(x)}{\partial\theta^i\partial\theta^j}. \tag{2.6}$$

We can work out the FIM for the Bernoulli parametrization given in (2.4). We get

$$G(\theta) = \frac{1}{\theta(1-\theta)}.$$

In Figure 2.4 we plot this function. For one-dimensional parametrizations, the distance

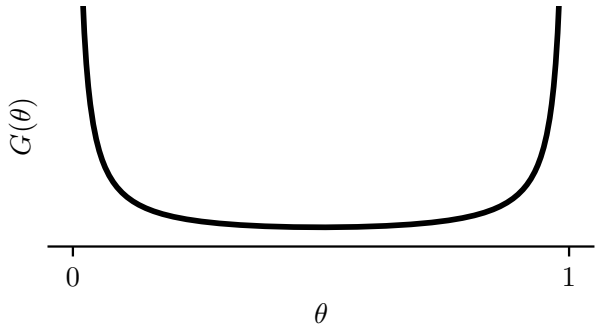


Figure 2.4: Fisher information matrix for the Bernoulli parametrization

induced by the Fisher metric between two distributions p_{θ_a} and p_{θ_b}, $\theta_a \leq \theta_b$ is given by

$$d^F(p_{\theta_a}, p_{\theta_b}) = \int_{\theta_a}^{\theta_b} \sqrt{G(\theta)} d\theta.$$

We can conclude that for the Bernoulli parametrization the distance between p_θ and $p_{\theta+\epsilon}$ is large when θ is close to 0 or 1 and small when θ is close to $1/2$. In the next section we will see why this is a desirable feature of the Fisher distance from the perspective of statistical estimation.

2.2 Cramér–Rao Lower Bound

Let $X \colon \mathcal{X} \to \mathbb{R}$ be a random variable.[2] We denote the expectation of X as a function on $\mathcal{P}$ by

$$\begin{aligned} \mathbb{E}[X] \colon \mathcal{P} &\to \mathbb{R} \\ p &\mapsto \mathbb{E}_p[X] = \sum_{x \in \mathcal{X}} p(x) X(x). \end{aligned}$$

Note that this function can be extended to $\mathcal{M}$ as follows

$$\begin{aligned} \tilde{\mathbb{E}}[X] \colon \mathcal{M} &\to \mathbb{R} \\ \mu &\mapsto \tilde{\mathbb{E}}_\mu[X] = \sum_{x \in \mathcal{X}} \mu(x) X(x). \end{aligned}$$

We also note that

$$\partial_x \tilde{\mathbb{E}}[X](\mu) = X(x). \tag{2.7}$$

Combining equations (2.3) and (2.7), we know the gradient of $\mathbb{E}[X]$ is therefore given by

$$\operatorname{grad}_p \mathbb{E}[X] = \sum_x p(x) \left(X(x) - \mathbb{E}_p[X]\right) \partial_x \big|_p. \tag{2.8}$$

[2]Note that x is the argument of X and not a possible outcome of X, as you might see in other probability theory settings.

We denote the variance of X at p by $\mathbb{V}_p[X]$, i.e.

$$\begin{aligned}\mathbb{V}[X]\colon \mathcal{P} &\to \mathbb{R}_{\geq 0}\\ p &\mapsto \mathbb{V}_p[X] = \sum_{x\in\mathcal{X}} p(x)\left(X(x) - \mathbb{E}_p[X]\right)^2.\end{aligned}$$

Theorem 2.1. *The variance of X is given by the square (Fisher) length of the gradient of the expected value of X, that is*

$$\mathbb{V}_p[X] = \|\operatorname{grad}_p \mathbb{E}[X]\|_p^2.$$

Proof.

$$\begin{aligned}\|\operatorname{grad}_p \mathbb{E}[X]\|_p^2 &= \mathfrak{g}(\operatorname{grad}_p \mathbb{E}[X], \operatorname{grad}_p \mathbb{E}[X])\\ &= \sum_x \frac{1}{p(x)} p(x)^2 (X(x) - \mathbb{E}_p[X])^2\\ &= \mathbb{V}_p[X].\end{aligned}$$

□

Remark 2.2. This is an interesting insight. The length of the gradient tells the maximum rate of change of a function. In this case the function is the expected value of a random variable, and the length of the gradient gives the maximum rate of change of the expectation when moving the base measure of the expectation. Intuitively, one can see that when the values $X(x), x \in \mathcal{X}$ are relatively similar, changing the base measure will not have a large effect on the expectation. On the other hand, when the values $X(x), x \in \mathcal{X}$ are very different, then changing the base measure can have a large effect on the expectation. This theorem tells us that if we use the Fisher metric as the metric on the space of probability distributions, then the length of the gradient of the expectation is exactly equal to the standard deviation of the random variable.

Now consider the parametrization

$$\phi\colon \mathbb{R} \supset \Theta \ni \theta \mapsto p_\theta(\cdot) \in \mathcal{P}$$

and examine the gradient on $\mathcal{S} = \phi(\Theta)$ of $\mathbb{E}[X]$, denoted $\operatorname{grad}^{\mathcal{S}} \mathbb{E}[X]$. Since this is an orthogonal projection of $\operatorname{grad} \mathbb{E}[X]$ we have the following inequality:

$$\|\operatorname{grad}_p^{\mathcal{S}} \mathbb{E}[X]\|_p^2 \leq \|\operatorname{grad}_p \mathbb{E}[X]\|_p^2 = \mathbb{V}_p[X]. \tag{2.9}$$

Example. Let $\mathcal{X} = \{1, 2, 3\}$. We let the random variable X be defined as follows:

$$X(1) = 0.6, \quad X(2) = 0.9, \quad X(3) = 1.6.$$

Figure 2.5 visualizes the probability simplex $\mathcal{P}$ on $\mathcal{X}$, and the colors represent the value of the function $\mathbb{E}[X]\colon \mathcal{P} \to \mathbb{R}$ on each point of the simplex.

At $p = \frac{3}{8}\delta^1 + \frac{3}{8}\delta^2 + \frac{1}{4}\delta^3$, the gradient of $\mathbb{E}[X]$ is given by equation (2.8):

$$\begin{aligned}\operatorname{grad}_p \mathbb{E}[X] &= \sum_x p(x)\left(X(x) - \mathbb{E}_p[X]\right)\partial_x\big|_p\\ &\approx -0.14\partial_1|_p - 0.02\partial_2|_p + 0.16\partial_3|_p.\end{aligned}$$

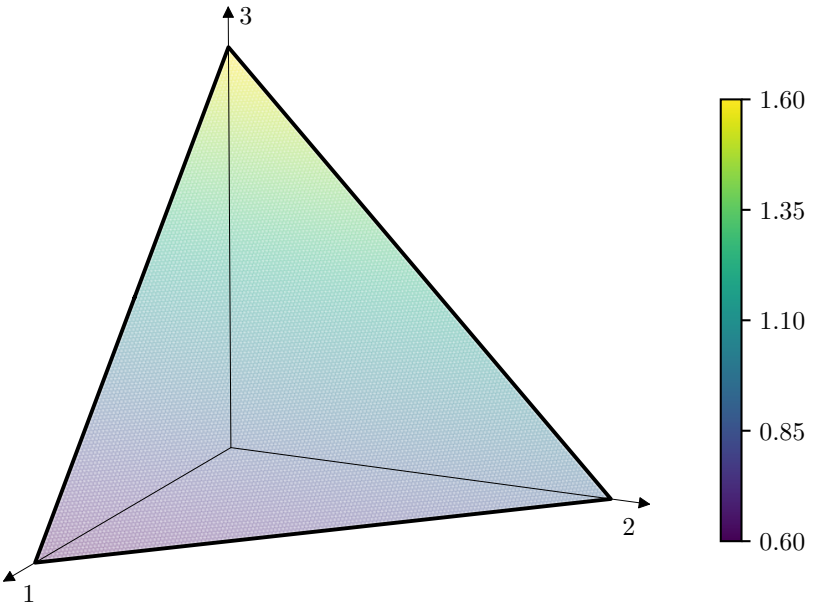


Figure 2.5: The probability simplex $\mathcal{P}$ on $\mathcal{X} = \{1, 2, 3\}$ and a heatmap of the function $\mathbb{E}[X] \colon \mathcal{P} \to \mathbb{R}$

We know from Theorem 2.1 that the (Fisher) length of this gradient vector is equal to the standard deviation of X at p. Both the gradient and its length are depicted in Figure 2.6.

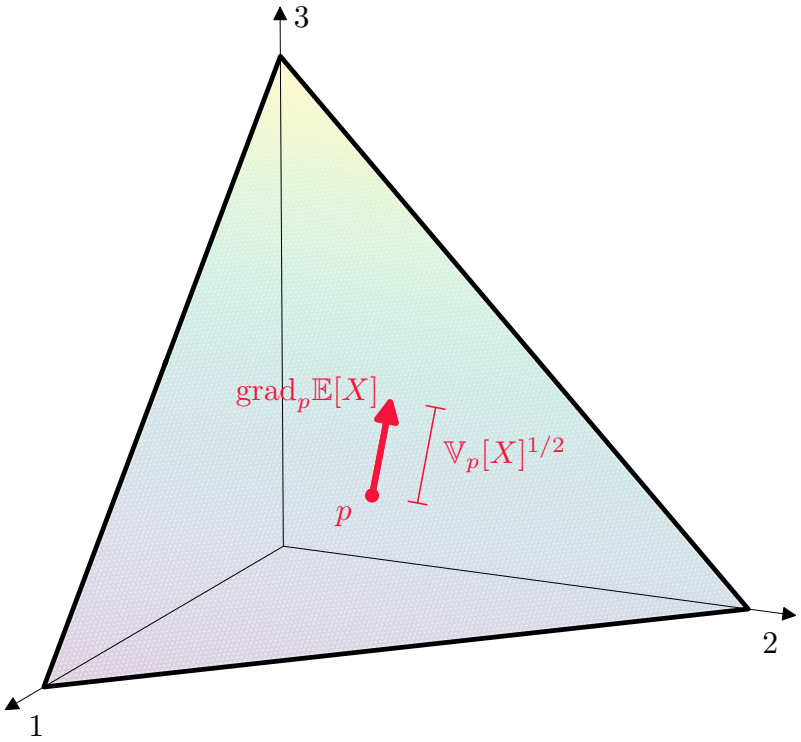


Figure 2.6: The gradient of $\mathbb{E}[X]$ at p and its length, equal to the square root of the variance of X at p

Now we consider a parametrization ϕ of a 1-dimensional model $\mathcal{S} = \phi(\Theta)$ on $\mathcal{P}$ given

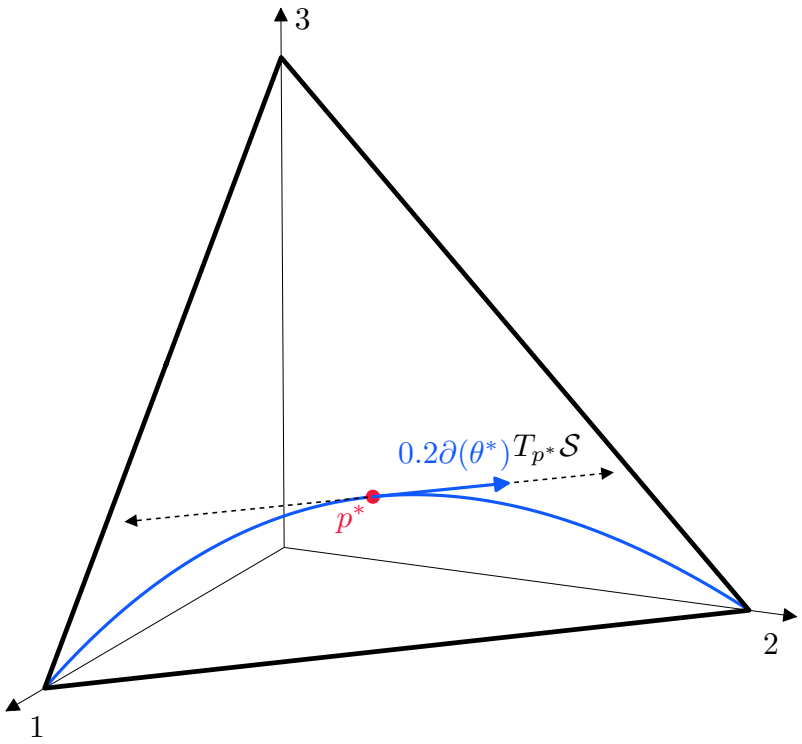


Figure 2.7: The probability simplex $\mathcal{P}$ over $\{1, 2, 3\}$. The red dot represents the point p_θ, the blue line the image space $\mathcal{S} = \phi(\Theta)$ of the parametrization, the dotted lines the tangent space $T_{p_\theta}\mathcal{S}$ and the blue arrow the (scaled) parameter tangent vector $\partial(\theta)$.

by

$$p_\theta(1) = (1 - \theta) - 0.5p_\theta(3), \tag{2.10}$$
$$p_\theta(2) = \theta - 0.5p_\theta(3), \tag{2.11}$$
$$p_\theta(3) = -\theta^2 + \theta, \tag{2.12}$$
$$\Theta = (0, 1).$$

The space $\mathcal{S}$ is depicted by the blue line in Figure 2.7.

We now fix $\theta^* = 0.5$ and let $p^* = \phi(\theta^*)$. We get

$$p^*(1) = \frac{3}{8}, \quad p^*(2) = \frac{3}{8}, \quad p^*(3) = \frac{1}{4}.$$

The tangent space of $\mathcal{S}$ is given by

$$T_{p^*}\mathcal{S} = \operatorname{span}\{\partial(\theta^*)\},$$
$$\partial(\theta^*) = \partial_2|_{p^*} - \partial_1|_{p^*}.$$

The tangent space and tangent vector are depicted in Figure 2.7.

The projection of the gradient is given by

$$\operatorname{grad}^{\mathcal{S}}_{p^*} \mathbb{E}[X] = \frac{\mathfrak{g}(\operatorname{grad}_{p^*} \mathbb{E}[X], \partial(\theta^*))}{\|\partial(\theta^*)\|^2}\partial(\theta^*).$$

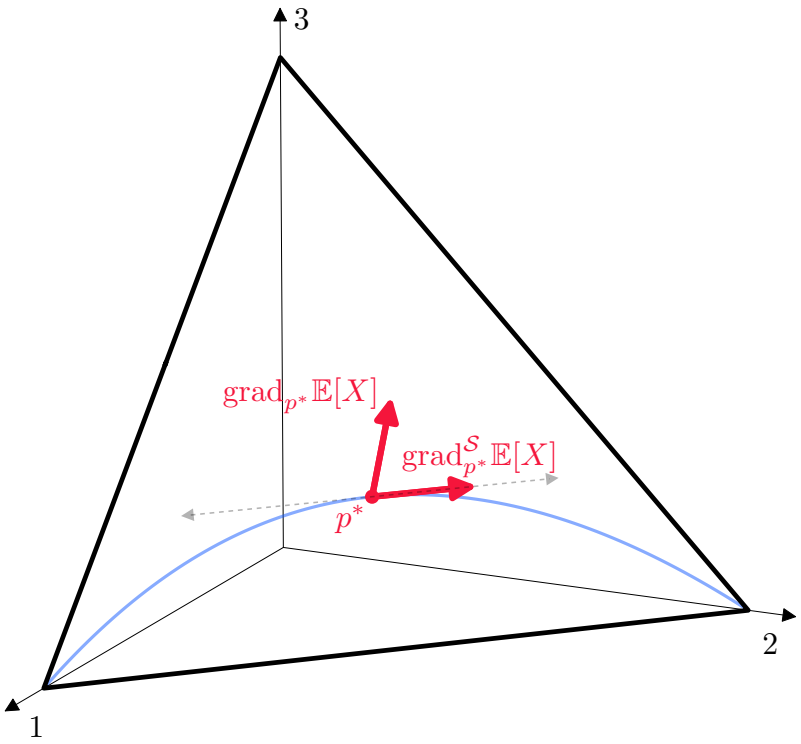


Figure 2.8: The gradient of $\mathbb{E}[X]$ at p^* and its projection onto $T_{p^*}\mathcal{S}$

Using the defining property of the gradient, $\mathfrak{g}(\operatorname{grad}_p V, A) = A(V)$, we get

$$\begin{aligned}\mathfrak{g}(\operatorname{grad}_{p^*} \mathbb{E}[X], \partial(\theta^*)) &= \partial(\theta^*)(\mathbb{E}[X]) \\ &= X(2) - X(1) \\ &= 0.3\end{aligned}$$

and therefore

$$\operatorname{grad}^{\mathcal{S}}_{p^*} \mathbb{E}[X] = 0.15(\partial_2|_{p^*} - \partial_1|_{p^*}).$$

It is clear from the picture that the Fisher projection is unequal to the Euclidean one. We can verify equation (2.9) in this case. We have

$$\begin{aligned}\| \operatorname{grad}^{\mathcal{S}}_{p^*} \mathbb{E}[X]\|^2_{p^*} &= \mathfrak{g}(\operatorname{grad}^{\mathcal{S}}_{p^*} \mathbb{E}[X], \operatorname{grad}^{\mathcal{S}}_{p^*} \mathbb{E}[X]) \\ &= \frac{1}{3/8}(-0.15)^2 + \frac{1}{3/8}0.15^2 \\ &= 0.12,\end{aligned}$$

whereas the variance of X at p^* is approximately 0.15. //

The previous theorem applies to any random variable X. We now specialize to a particular kind: those used to estimate the parameter of a parametrized family. Suppose we are given a parametrization $\phi\colon \Theta \to \mathcal{P}$ and observe a single sample $x \in \mathcal{X}$ drawn from p_θ for some unknown θ. An *estimator* is a function

$$\hat{\theta}\colon \mathcal{X} \to \Theta$$

that maps the observation to a guess for the parameter. Since x is random, $\hat{\theta}$ is itself a random variable taking values in Θ, and we can speak of its expectation $\mathbb{E}_{p_\theta}[\hat{\theta}]$ and variance $\mathbb{V}_{p_\theta}[\hat{\theta}]$ under the true distribution.

A minimal requirement to ask of an estimator is that it be correct on average: $\mathbb{E}_{p_\theta}[\hat{\theta}] = \theta$ for all $\theta \in \Theta$. An estimator with this property is called *unbiased*. Among unbiased estimators, the natural measure of quality is the variance, which controls how tightly the guesses concentrate around θ. The Cramér–Rao lower bound (Cramér, 1946; Rao, 1945), which we now prove, tells us how small this variance can be.

Theorem 2.2 (Cramér–Rao Lower Bound)**.** *The variance of an unbiased estimator is lower bounded by the inverse of the Fisher information matrix. That is,*

$$\mathbb{V}_{p_\theta}[\hat{\theta}] \geq G^{-1}(\theta).$$

Proof. Note that an estimator is simply a random variable. We can therefore use Theorem 2.1 and equation (2.9) to get

$$\begin{aligned}\mathbb{V}_{p_\theta}[\hat{\theta}] &= \| \operatorname{grad}_{p_\theta} \mathbb{E}[\hat{\theta}] \|^2_{p_\theta} \\ &\geq \| \operatorname{grad}^{\mathcal{S}}_{p_\theta} \mathbb{E}[\hat{\theta}] \|^2_{p_\theta}.\end{aligned}$$

Since $\hat{\theta}$ is unbiased we have

$$\mathbb{E}_{p_\theta}[\hat{\theta}] = \theta,$$

that is, evaluating $\mathbb{E}[\hat{\theta}]$ at p_θ returns the parameter that produced p_θ. Restricted to the model, the expectation therefore coincides with the inverse parametrization,

$$\mathbb{E}[\hat{\theta}]\Big|_{\mathcal{S}} = \phi^{-1} \colon \mathcal{S} \to \Theta \subset \mathbb{R},$$

which is exactly the local coordinate function on $\mathcal{S}$ induced by ϕ. Applying Lemma 2.5 from the appendix to this coordinate function gives

$$\| \operatorname{grad}^{\mathcal{S}}_{p_\theta} \mathbb{E}[\hat{\theta}] \|^2_{p_\theta} = G^{-1}(\theta).$$

This gives the result. □

Usually a parameter is estimated after observing multiple samples, say N. In this case an estimator is defined as follows:

$$\hat{\theta}^N \colon \mathcal{X}^N \to \Theta.$$

We write $x = (x^1, \ldots, x^N) \in \mathcal{X}^N$ and let $\mathcal{P}^N$ be the space of all strictly positive probability measures on $\mathcal{X}^N$. We model the N samples as drawn i.i.d. from p_θ, so that the joint distribution on $\mathcal{X}^N$ is the product $p_\theta(x^1) \cdots p_\theta(x^N)$. The corresponding parametrization into this larger space is given by

$$\begin{aligned}\phi^N \colon \Theta &\to \mathcal{P}^N, \\ \theta &\mapsto \phi^N(\theta)(x) = p^N_\theta(x), \\ p^N_\theta(x) &= \prod_{n=1}^{N} p_\theta(x^n),\end{aligned}$$

with $\mathcal{S}^N = \phi^N(\Theta)$. Finally, we let $G^N(\theta)$ be the Fisher information matrix for this parametrization.

Lemma 2.1. *The FIM G^N on $\mathcal{S}^N$ and the FIM G on $\mathcal{S}$ have the following relationship:*

$$G^N(\theta) = NG(\theta).$$

Proof. Using (2.6) we get

$$\begin{aligned} G^N(\theta) &= -\sum_x p_\theta^N(x) \frac{\partial^2 \log p_\theta^N(x)}{\partial \theta^2} \\ &= -\sum_x p_\theta^N(x) \sum_{n=1}^N \frac{\partial^2 \log p_\theta(x^n)}{\partial \theta^2} \\ &= -\sum_{n=1}^N \sum_x p_\theta^N(x) \frac{\partial^2 \log p_\theta(x^n)}{\partial \theta^2} \\ &= -\sum_{n=1}^N \sum_{x^n} p_\theta(x^n) \frac{\partial^2 \log p_\theta(x^n)}{\partial \theta^2} \\ &= NG(\theta). \end{aligned}$$

□

Now we are in position to generalize the Cramér–Rao Lower Bound to multiple samples.

Theorem 2.3 (Cramér–Rao Lower Bound)**.** *Let $\phi : \Theta \to \mathcal{P}$ be a parametrization, $G(\theta)$ the associated FIM and $\hat{\theta}^N : \mathcal{X}^N \to \Theta$ an unbiased estimator. Then we have*

$$\mathbb{V}_{p_\theta^N}[\hat{\theta}^N] \geq \frac{1}{N} G^{-1}(\theta).$$

Proof. This follows directly from Theorem 2.2 and Lemma 2.1. □

The Cramér–Rao Lower Bound tells us the minimum amount of variance any estimator will have for a given parametrization. When the variance of an estimator is equal to the lower bound for all $\theta \in \Theta$, we say this estimator is *efficient*. The following corollary gives a geometric condition for when an estimator is efficient.

Corollary 2.1. *An estimator $\hat{\theta}$ is efficient if and only if for all $\theta \in \Theta$*

$$\operatorname{grad}_{p_\theta^N} \mathbb{E}[\hat{\theta}] \in T_{p_\theta^N} \mathcal{S}^N.$$

Proof. This follows directly from the fact that we have equality in (2.9) if and only if $\operatorname{grad}_{p_\theta} \mathbb{E}[X] \in T_{p_\theta} \mathcal{S}$. □

Example. Let us consider the case that $X = \{0, 1\}$, $\Theta = (0, 1)$, ϕ the Bernoulli parametrization, and $N = 2$. In this case ϕ^N is given by

$$\phi^N(\theta) = p_\theta^N = (1-\theta)^2 \delta^{0,0} + (1-\theta)\theta \delta^{0,1} + \theta(1-\theta)\delta^{1,0} + \theta^2 \delta^{1,1}.$$

The space $\mathcal{S}^N = \phi^N(\Theta)$ is the 1-dimensional line depicted in Figure 2.9. Note that the parameter tangent vector is given by

$$\begin{aligned} \partial(\theta) &= d\phi_\theta \left. \frac{\partial}{\partial \theta^i} \right|_\theta \\ &= -2(1-\theta)\partial_{0,0} + (1-2\theta)\partial_{0,1} + (1-2\theta)\partial_{1,0} + 2\theta \partial_{1,1}. \end{aligned} \tag{2.13}$$

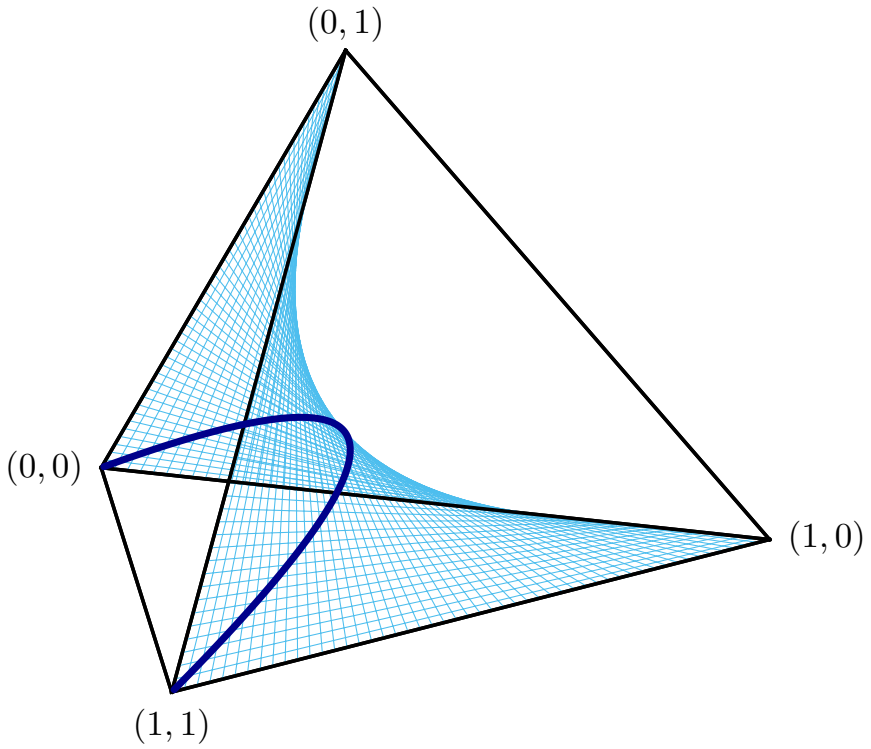


Figure 2.9: This tetrahedron visualizes the probability simplex for a sample space consisting of two binary variables, with each vertex corresponding to a point mass (delta distribution) on one of the four joint outcomes. The light-blue surface depicts the independence model, consisting of all distributions of the form $p(x^1)q(x^2)$. The dark-blue line shows the subset where the two marginals coincide, i.e., all distributions of the form $p(x^1)p(x^2)$.

We consider the following two estimators:

$$\hat{\theta}_a(x) = \frac{x^1 + x^2}{2},$$
$$\hat{\theta}_b(x) = x^1.$$

In order to illustrate the Cramér–Rao Lower Bound, we will show that the estimator $\hat{\theta}_a$ is efficient and $\hat{\theta}_b$ is not, by examining the gradients of the expectations of both estimators.

From equation (2.8) we get

$$\operatorname{grad}_{p_\theta^N} \mathbb{E}[\hat{\theta}] = \sum_{x \in \mathcal{X}^N} p_\theta^N(x) \left(\hat{\theta}(x) - \mathbb{E}_{p_\theta^N}[\hat{\theta}] \right) \partial_x.$$

Plugging in $\hat{\theta}_a$ and $\hat{\theta}_b$ and using that they are both unbiased ($\mathbb{E}_{p_\theta^N}[\hat{\theta}] = \theta$) gives

$$\begin{aligned}
\operatorname{grad}_{p_\theta^N} \mathbb{E}[\hat{\theta}_a] &= \sum_{x \in \mathcal{X}^N} p_\theta^N(x) \left(\hat{\theta}_a(x) - \theta \right) \partial_x. \\
&= \begin{aligned}[t] &(1-\theta)^2(-\theta)\partial_{0,0} + (1-\theta)\theta(1/2-\theta)\partial_{0,1} + \\ &\quad \theta(1-\theta)(1/2-\theta)\partial_{1,0} + \theta^2(1-\theta)\partial_{1,1} \end{aligned} \\
&= \begin{aligned}[t] &(1-\theta)\theta\Big(-(1-\theta)\partial_{0,0} + (1/2-\theta)\partial_{0,1} + \\ &\quad (1/2-\theta)\partial_{1,0} + \theta\partial_{1,1} \Big), \end{aligned}
\end{aligned}$$

$$\begin{aligned}
\operatorname{grad}_{p_\theta^N} \mathbb{E}[\hat{\theta}_b] &= \sum_{x \in \mathcal{X}^N} p_\theta^N(x) \left(\hat{\theta}_b(x) - \theta \right) \partial_x. \\
&= \begin{aligned}[t] &(1-\theta)^2(-\theta)\partial_{0,0} + (1-\theta)\theta(-\theta)\partial_{0,1} + \\ &\quad \theta(1-\theta)(1-\theta)\partial_{1,0} + \theta^2(1-\theta)\partial_{1,1} \end{aligned} \\
&= \begin{aligned}[t] &(1-\theta)\theta\Big(-(1-\theta)\partial_{0,0} + (-\theta)\partial_{0,1} + \\ &\quad (1-\theta)\partial_{1,0} + \theta\partial_{1,1} \Big). \end{aligned}
\end{aligned}$$

From (2.13) it follows that $\operatorname{grad}_{p_\theta^N} \mathbb{E}[\hat{\theta}_a] \in T_{p_\theta^N} \mathcal{S}^N$, and hence $\hat{\theta}_a$ is efficient. The vector $\operatorname{grad}_{p_\theta^N} \mathbb{E}[\hat{\theta}_b]$ on the other hand is not tangent to $\mathcal{S}^N$, so $\hat{\theta}_b$ has strictly higher variance.

Both gradients are plotted in Figure 2.10 for $\theta^* = 0.76$, $p^* = \phi^N(\theta^*)$. Note that $\operatorname{grad}_{p^*} \mathbb{E}[\hat{\theta}_b]$ is pointing more in the direction of $\delta^{1,0}$ which is expected since $\hat{\theta}_b$ only depends on x^1. ∥

Multidimensional parameter space Now we extend the above discussion to multidimensional parameter spaces. The covariance and Fisher information are now matrices, and for any symmetric matrices A, B the inequality $A \geq B$ means $A - B$ is positive semidefinite.

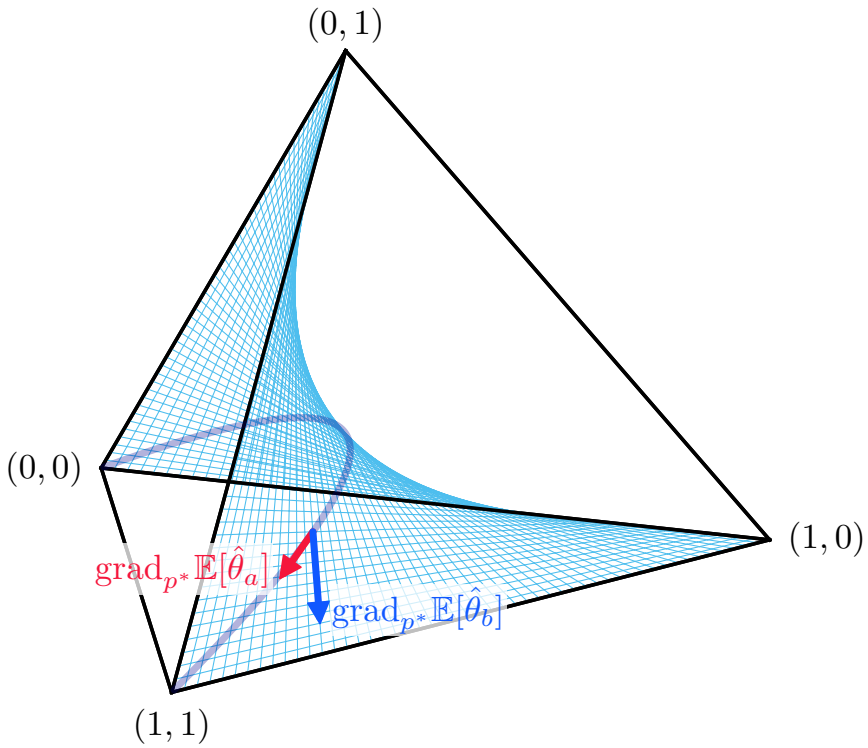


Figure 2.10: Gradient vectors for the expectation of the estimators $\hat{\theta}_a$ and $\hat{\theta}_b$

We start by considering a general random variable

$$X : \mathcal{X} \to \mathbb{R}^d.$$

We write the covariance matrix of X as follows:

$$[\mathbb{V}_p[X]]_{kl} = \mathrm{Cov}(X_k, X_l).$$

Now we consider the parametrization

$$\phi \colon \Theta \to \mathcal{P}$$

and examine the gradients of $\mathbb{E}[X]$. For $p \in \mathcal{S}$ we define

$$\left[\tilde{H}_p^X\right]^{kl} = \mathfrak{g}\left(\mathrm{grad}_p \mathbb{E}[X_k], \mathrm{grad}_p \mathbb{E}[X_l]\right),$$
$$\left[H_p^X\right]^{kl} = \mathfrak{g}\left(\mathrm{grad}_p^{\mathcal{S}} \mathbb{E}[X_k], \mathrm{grad}_p^{\mathcal{S}} \mathbb{E}[X_l]\right).$$

Using (2.3) and Lemma 2.6 we have

$$\mathbb{V}_p[X] = \tilde{H}_p^X \geq H_p^X. \tag{2.14}$$

Now we let $\hat{\theta} \colon \mathcal{X} \to \Theta \subset \mathbb{R}^d$ be an unbiased estimator. Specializing (2.14) to $X = \hat{\theta}$, the unbiasedness condition $\mathbb{E}_{p_\theta}[\hat{\theta}_i] = \theta^i$ implies that, restricted to $\mathcal{S}$, each component $\mathbb{E}[\hat{\theta}_i]|_{\mathcal{S}} = \phi_i^{-1}$ is exactly the ith local coordinate function on $\mathcal{S}$ induced by ϕ. Applying Lemma 2.5 to these coordinate functions gives

$$\left[H_{p_\theta}^{\hat{\theta}}\right]^{ij} = \mathfrak{g}(\mathrm{grad}_{p_\theta}^{\mathcal{S}} \mathbb{E}[\hat{\theta}_i], \mathrm{grad}_{p_\theta}^{\mathcal{S}} \mathbb{E}[\hat{\theta}_j]) = \left[G^{-1}(\theta)\right]^{ij}.$$

Combining this with (2.14) gives

$$\mathbb{V}_{p_\theta}[\hat{\theta}] \geq G^{-1}(\theta). \tag{2.15}$$

We can also extend this result to multiple samples. We first generalize Lemma 2.1 to multidimensional parameter space.

Lemma 2.2. *The FIM G^N on $\mathcal{S}^N$ and the FIM G on $\mathcal{S}$ have the following relationship:*

$$G^N(\theta) = NG(\theta).$$

Proof.

$$\begin{aligned} G^N_{ij}(\theta) &= -\sum_x p^N_\theta(x) \frac{\partial^2 \log p^N_\theta(x)}{\partial\theta^i \partial\theta^j} \\ &= -\sum_{n=1}^{N} \sum_{x^n} p_\theta(x^n) \frac{\partial^2 \log p_\theta(x^n)}{\partial\theta^i \partial\theta^j} \\ &= NG_{ij}(\theta). \end{aligned}$$

□

Now we can formulate the multidimensional and multi-sample version of the Cramér–Rao Lower Bound.

Theorem 2.4 (Cramér–Rao Lower Bound)**.** *Let $\phi : \Theta \to \mathcal{P}$ be a parametrization, $G(\theta)$ the associated FIM and $\hat{\theta}^N : \mathcal{X}^N \to \Theta$ an unbiased estimator. Then we have*

$$\mathbb{V}_{p^N_\theta}[\hat{\theta}^N] \geq \frac{1}{N} G^{-1}(\theta).$$

Proof. This follows directly from (2.15) and Lemma 2.2. □

Remark 2.3. We can now make the intuition from Remark 2.1 more precise. For an efficient estimator $\hat{\theta}^N$ we have

$$\mathbb{V}_{p^N_\theta}[\hat{\theta}^N] = \frac{1}{N} G^{-1}(\theta),$$

so the Fisher information is (up to the factor $1/N$) exactly the inverse covariance of the estimator. This allows us to interpret the Fisher metric as a direction-wise measure of certainty about θ. For $v = \sum_i v^i \partial_i(\theta) \in T_{p_\theta}\mathcal{S}$ a large value of $\|v\|_{p_\theta}$ means a small estimator variance in the direction v, and hence a high certainty about the parameter along that direction.

This in turn controls how easily nearby distributions can be distinguished. The Fisher length $\|\epsilon v\|_{p_\theta}$ of an infinitesimal shift ϵv tells us how well p_θ and $p_{\theta+\epsilon v}$ can be told apart using an efficient estimator. The Fisher metric therefore gives a local, direction-wise measure of statistical distinguishability — exactly the intuition anticipated in Remark 2.1.

2.3 Sufficient statistics and Chentsov's theorem

A *statistic* is simply a function from the sample space $\mathcal{X}$ to some other space $\mathcal{X}'$; it can be thought of as a transformation or summary of the data. The estimators introduced in Section 2.2 are themselves statistics, the codomain being the parameter space Θ. In this section we study how statistics interact with the geometry of $\mathcal{P}$, and identify the *sufficient* statistics as those that preserve all information about the parameter.

We let

$$f\colon \mathcal{X} \to \mathcal{X}'$$

be a statistic that we assume to be surjective. We use y to denote a typical element of $\mathcal{X}'$ and write $\mathcal{P}'$ for the space of probability measures over $\mathcal{X}'$ and $\mathfrak{g}'$ for the Fisher metric on this space. Furthermore, we write $\mathcal{X}_y = f^{-1}(y)$ for the fibers of f. We can use f to define the following map:

$$\begin{aligned} f_*\colon \mathcal{P} &\to \mathcal{P}' \\ \sum_x p(x)\delta^x &\mapsto \sum_y \left(\sum_{x \in \mathcal{X}_y} p(x) \right) \delta^y. \end{aligned}$$

We can write $p \in \mathcal{P}$ as follows:

$$\begin{aligned} p(x) &= p(x|f(x))p(f(x)), \\ p(y) &= \sum_{\tilde{x} \in \mathcal{X}_y} p(\tilde{x}) = f_*(p)(y), \\ p(x|y) &= \begin{cases} \frac{p(x)}{p(y)} & x \in \mathcal{X}_y, \\ 0 & x \notin \mathcal{X}_y. \end{cases} \end{aligned}$$

For $A \in T_p\mathcal{P}$ we introduce similar notation, namely

$$\begin{aligned} A(y) &= \sum_{x \in \mathcal{X}_y} A(x), \\ A(x|y) &= \begin{cases} \frac{A(x)}{p(y)} & x \in \mathcal{X}_y, \\ 0 & x \notin \mathcal{X}_y. \end{cases} \end{aligned}$$

Note that for $A = \sum_x A(x)\partial_x \in T_p\mathcal{P}$ we have

$$df_*(A) = \sum_y A(y)\partial_y\big|_p.$$

We can consider the following subspaces of $T_p\mathcal{P}$:

$$\begin{aligned} \mathcal{V}_p &= \{A \in T_p\mathcal{P} : (df_*)_p A = 0\}, \\ \mathcal{H}_p &= \{B \in T_p\mathcal{P} : \mathfrak{g}(A, B) = 0, A \in \mathcal{V}_p\}, \end{aligned}$$

which are called the *vertical* and *horizontal* space respectively.

Example. We can consider the spaces $\mathcal{X} = \{1, 2, 3\}$ and $\mathcal{X}' = \{a, b\}$ and the following statistic:

$$f(x) = \begin{cases} a & x \in \{1, 3\}, \\ b & x = 2. \end{cases} \tag{2.16}$$

We compute the vertical space for the statistic. It is easy to see that this is given by

$$\mathcal{V}_p = \{A \in T_p\mathcal{P} : A(1) = -A(3), A(2) = 0\}.$$

From the definition of the horizontal space we know that $B(x)/p(x)$ needs to be constant on the fibers of f. In this case this translates to the requirement $\frac{B(1)}{p(1)} = \frac{B(3)}{p(3)}$. The horizontal space is therefore given by

$$\begin{aligned} \mathcal{H}_p &= \operatorname{span}\{B\}, \\ B &= p(1)\partial_1|_p - (p(1) + p(3))\partial_2|_p + p(3)\partial_3|_p. \end{aligned}$$

//

The following proposition gives a useful expression for vertical and horizontal vectors.

Proposition 2.1. *We have that $A \in \mathcal{V}_p$ if and only if $\sum_{x \in \mathcal{X}_y} A(x|y) = 0$ for all $y \in \mathcal{X}'$. This gives the following expression for $A \in \mathcal{V}_p$:*

$$A(x) = p(y)A(x|y),$$

with $\sum_{x \in \mathcal{X}_y} A(x|y) = 0$.

We have that $B \in \mathcal{H}_p$ if and only if for $x \in \mathcal{X}_y$, it can be written as follows:

$$B(x) = p(x|y)B(y),$$

with $\sum_{y \in \mathcal{X}'} B(y) = 0$.

Proof. The facts about A follow straight from the definitions. To see that B is true, we use that B is horizontal if and only if the term $B(x)/p(x)$ is constant on the fibers $\mathcal{X}_y$. This immediately proves the reverse implication. For the forward implication we have

$$\begin{aligned} B(y) &= \sum_{x \in \mathcal{X}_y} B(x) \\ &= \sum_{x \in \mathcal{X}_y} p(x|y) \frac{B(x)}{p(x|y)} \\ &= \frac{B(\tilde{x})}{p(\tilde{x}|y)} \sum_{x \in \mathcal{X}_y} p(x|y) \\ &= \frac{B(\tilde{x})}{p(\tilde{x}|y)} \end{aligned}$$

for any $\tilde{x} \in \mathcal{X}_y$. □

The following proposition shows that when moving in the vertical direction, only the conditional component $p(x|y)$ of the distribution p changes, and when moving along the horizontal direction, only the marginal component $p(y)$.

Proposition 2.2. *Let $A \in \mathcal{V}_p$ and p_t^A such that*

$$p_t^A(x) = p(x) + tA(x)$$

then,

$$\left.\frac{d}{dt}\right|_{t=0} p_t^A(y) = 0.$$

For $B \in \mathcal{H}_p$ and p_t^B such that

$$p_t^B(x) = p(x) + tB(x)$$

we have that

$$\left.\frac{d}{dt}\right|_{t=0} p_t^B(x|y) = 0.$$

Proof. Note that

$$\begin{aligned} p_t^A(y) &= p(y) + t\sum_{\tilde{x}\in\mathcal{X}_y} A(\tilde{x}) \\ &= p(y) \end{aligned}$$

from the definition of the vertical space. From Proposition 2.1 we know that we can write $B(x) = p(x|y)B(y)$, and therefore we have

$$\begin{aligned} p_t^B(x) &= p(x|y)p(y) + tp(x|y)B(y) \\ &= p(x|y)\left(p(y) + tB(y)\right). \end{aligned}$$

Now we can compute

$$\begin{aligned} p_t^B(x|y) &= \frac{p_t^B(x)}{\sum_{\tilde{x}\in\mathcal{X}_y} p_t^B(\tilde{x})} \\ &= p(x|y). \end{aligned}$$

□

We can project any tangent vector onto the vertical and horizontal space and get a unique decomposition. The following proposition tells us how to compute these projections.

Proposition 2.3. *Let $A \in T_p\mathcal{P}$ and $y = f(x)$. The projections on the horizontal and vertical space are given by*

$$\begin{aligned} A^H(x) &= p(x|y)A(y), \\ A^V(x) &= A(x) - p(x|y)A(y). \end{aligned}$$

Proof. The fact that A^H is horizontal follows directly from Proposition 2.1.

For A^V to be vertical we need that $\sum_{x\in\mathcal{X}_y} A^V(x) = 0$ for all $y \in \mathcal{X}'$. This follows immediately from the fact that $\sum_{x\in\mathcal{X}_y} p(x|y) = 1$ for all $y \in \mathcal{X}'$.

Lastly it is also easy to see that $A = A^H + A^V$. □

The next proposition shows that we can also obtain the decomposition by applying the chain rule.

Proposition 2.4. *Let $p \in \mathcal{P}, A \in T_p\mathcal{P}$ and p_t a curve given by*

$$p_t(x) = p(x) + tA(x)$$

and $x \in \mathcal{X}_y$. Then we have

$$A(x) = \left.\frac{d}{dt}\right|_{t=0} p_t(x) = A^H(x) + A^V(x), \tag{2.17}$$

where

$$A^H(x) = p(x|y) \left.\frac{d}{dt}\right|_{t=0} p_t(y),$$
$$A^V(x) = p(y) \left.\frac{d}{dt}\right|_{t=0} p_t(x|y).$$

Proof. Equation (2.17) follows directly from the chain rule for $p_t(x) = p_t(y)p_t(x|y)$. Horizontality and verticality follow directly from Proposition 2.1. □

The following proposition gives a final insightful form of the vertical component of a tangent vector.

Proposition 2.5. *Let $A \in T_p\mathcal{P}$ such that $\tilde{p} = p + \sum_x A(x)\delta^x \in \mathcal{P}$. For $x \in \mathcal{X}_y$, the vertical component can be written as follows:*

$$A^V(x) = \tilde{p}(y)(\tilde{p}(x|y) - p(x|y)),$$

where $\tilde{p} = p + \sum_x A(x)\delta^x$.

Proof.

$$\begin{aligned} A^V(x) &= \tilde{p}(y)(\tilde{p}(x|y) - p(x|y)) \\ &= p(x) + A(x) - p(x) - A(y)p(x|y) \\ &= A(x) - A(y)p(x|y). \end{aligned}$$

□

Example. We can consider the case that $f : \mathcal{U} \times \mathcal{V} \to \mathcal{U}$ is the marginalization map. We get

$$f_*(p)(u) = \sum_v p(u, v),$$
$$df_*(A) = \sum_u \left(\sum_v A(u, v) \right) \partial_u.$$

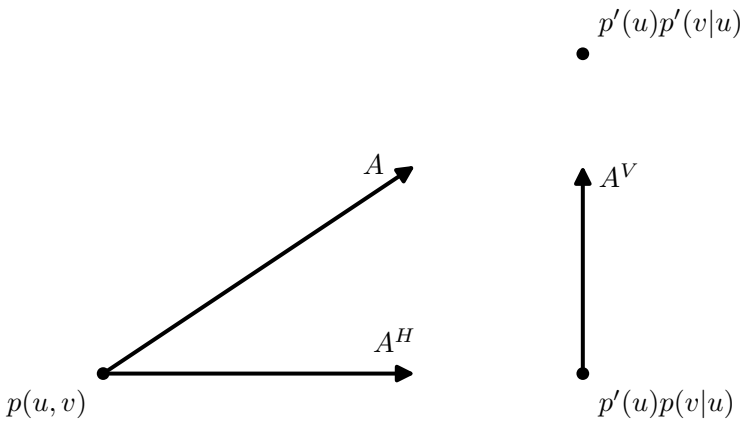


Figure 2.11: The orthogonal decomposition of A into horizontal and vertical components

If we decompose $p(u,v) = p(v|u)p(u)$, the vertical vectors only change the $p(v|u)$ component, whereas the horizontal vectors only change the marginal $p(u)$.

Now let $A \in T_p\mathcal{P}$ be such that $\tilde{p} = p + \sum_{u,v} A(u,v) \in \mathcal{P}$. From Proposition 2.3 and Proposition 2.5 we know that the orthogonal decomposition of $A \in T_p\mathcal{P}$ into horizontal and vertical components is given by

$$A^H(u,v) = p(v|u) \sum_v A(u,v),$$
$$A^V(u,v) = p'(u)\left(p'(v|u) - p(v|u)\right),$$

where $p'(u,v) = p(u,v) + A(u,v)$. See Figure 2.11 for an illustration of this. ∥

Now we have built up some intuition for horizontal and vertical vectors, we can define the concept of a *Riemannian submersion*.

Definition 2.2. Let M, g and N, h be two Riemannian manifolds. A *Riemannian submersion* $\pi\colon M \to N$ is smooth submersion that satisfies the following condition: for horizontal vectors $A, B \in \mathcal{H}_p^\pi$ we have

$$g(A,B) = h(d\pi_p A, d\pi_p B).$$

Intuitively, a Riemannian submersion is a projection $\pi\colon M \to N$ that collapses each fiber $\pi^{-1}(q)$ to a single point while preserving the geometry transverse to the fibers. The vertical directions—those tangent to the fibers—lie in the kernel of $d\pi_p$ and are therefore invisible downstairs: moving along a fiber does not change where we land in N. The horizontal directions are the orthogonal complement, and the defining condition demands that $d\pi_p$ restricts to an isometry $\mathcal{H}_p^\pi \to T_{\pi(p)}N$. In other words, N inherits its geometry from horizontal motion in M: lengths and angles measured in N are exactly those of the horizontal lifts in M.

In our setting this picture becomes concrete. Each fiber of the marginalization map f_* consists of joint distributions sharing a common marginal, vertical motion alters only the conditional $p(v|u)$—precisely the information discarded by marginalizing—and horizontal motion varies the marginal in a way that is faithfully recorded by f_*. The next proposition confirms that this recording is in fact an isometry.

Proposition 2.6. *The map f_* is a Riemannian submersion.*

Proof. For $A, B \in \mathcal{H}_p$ with $p' = f_*(p)$ we need to show

$$\mathfrak{g}_p(A, B) = \mathfrak{g}'_{p'}(df_*A, df_*B).$$

Using Proposition 2.1 we get

$$\begin{aligned}
\mathfrak{g}_p(A, B) &= \sum_x \frac{1}{p(x)} A(x)B(x) \\
&= \sum_y \sum_{x \in \mathcal{X}_y} \frac{1}{p(y)p(x|y)} p(x|y)A(y)p(x|y)B(y) \\
&= \sum_y \frac{1}{p(y)} A(y)B(y) \\
&= \mathfrak{g}'_{p'}(df_*A, df_*B).
\end{aligned}$$

□

Now we let $\phi\colon \Theta \to \mathcal{P}$ be a parametrization with $\mathcal{S} = \phi(\Theta)$.

Proposition 2.7. *The pushforward parameter tangent vector $\partial_i(\theta)$ can be decomposed into a horizontal and vertical component as follows:*

$$\begin{aligned}
\partial_i^H(\theta)(x) &= p_\theta(x|y) \frac{\partial}{\partial \theta^i} p_\theta(y), \\
\partial_i^V(\theta)(x) &= p_\theta(y) \frac{\partial}{\partial \theta^i} p_\theta(x|y),
\end{aligned}$$

where $y = f(x)$.

Proof. This follows directly from Proposition 2.4. □

Definition 2.3. Let us fix a parametrization $\phi\colon \theta \mapsto p_\theta$. A statistic f is called *sufficient* (for ϕ) if the conditional distribution $p_\theta(x|y)$ is independent of θ.

Theorem 2.5. *A statistic f is sufficient if and only if $T_{p_\theta}\mathcal{S} \subset \mathcal{H}_{p_\theta}$ for all $\theta \in \Theta$.*

Proof. From the definition we know that f is sufficient if and only if $p_\theta(x|y)$ is independent of θ. This means that $\frac{\partial}{\partial \theta^i} p_\theta(x|y)$ has to be zero for all i. This is true if and only if $\partial_i^V(\theta)$ is zero and thus $\partial_i(\theta) \in \mathcal{H}_{p_\theta}$ for all i. □

The Fisher matrix on $(f_* \circ \phi)(\Theta) = \mathcal{S}'$, denoted G', is given by:

$$G'_{ij}(\theta) = \mathfrak{g}'(df_*(\partial_i(\theta)), df_*(\partial_j(\theta))).$$

Theorem 2.6. *Let f be a statistic. In general we have*

$$G(\theta) \geq G'(\theta).$$

The statistic f is sufficient if and only if

$$G(\theta) = G'(\theta).$$

Proof. Note that we have

$$\mathfrak{g}(A, B) = \mathfrak{g}(A^H, B^H) + \mathfrak{g}(A^V, B^V).$$

The Fisher matrix G can be split up into

$$\begin{aligned} G(\theta) &= G^H(\theta) + G^V(\theta), \\ G^H_{ij}(\theta) &= \mathfrak{g}\left(\partial_i^H(\theta), \partial_j^H(\theta)\right), \\ G^V_{ij}(\theta) &= \mathfrak{g}\left(\partial_i^V(\theta), \partial_j^V(\theta)\right). \end{aligned} \tag{2.18}$$

Since we know that $df_*(\partial_i(\theta)) = df_*(\partial_i^H(\theta))$ and f_* is a Riemannian submersion we have

$$G^H(\theta) = G'(\theta). \tag{2.19}$$

Since $G^V(\theta)$ is positive semidefinite, this implies $G(\theta) \geq G'(\theta)$.

From Theorem 2.5 we know that $G^V(\theta)$ is zero if and only if f is sufficient. This proves the second statement. □

Definition 2.4. The *information loss* of the statistic f at a point $\theta \in \Theta$ is given by

$$\Delta G(\theta) = G(\theta) - G'(\theta).$$

Corollary 2.2. *The information loss is given by*

$$(\Delta G)_{ij}(\theta) = \mathfrak{g}\left(\partial_i^V(\theta), \partial_j^V(\theta)\right).$$

Proof. This follows directly from equations (2.18) and (2.19) in the proof of Theorem 2.6. □

Example. We consider the statistic from equation (2.16). We first consider the following parametrization:

$$p_\theta(x) = \begin{cases} 0.6(1-\theta) & \text{if } x = 1, \\ \theta & \text{if } x = 2, \\ 0.4(1-\theta) & \text{if } x = 3, \end{cases} \tag{2.20}$$

which is depicted in Figure 2.12. Note that the pushforward of the parameter tangent vector is given by

$$\partial(\theta)(x) = \begin{cases} -0.6 & \text{if } x = 1, \\ 1 & \text{if } x = 2, \\ -0.4 & \text{if } x = 3. \end{cases}$$

Therefore, we have $\frac{\partial(\theta)(1)}{p_\theta(1)} = \frac{\partial(\theta)(3)}{p_\theta(3)}$ and thus $\partial(\theta) \in \mathcal{H}_{p_\theta}$. From Theorem 2.5 we conclude that the statistic is sufficient for parametrization (2.20).

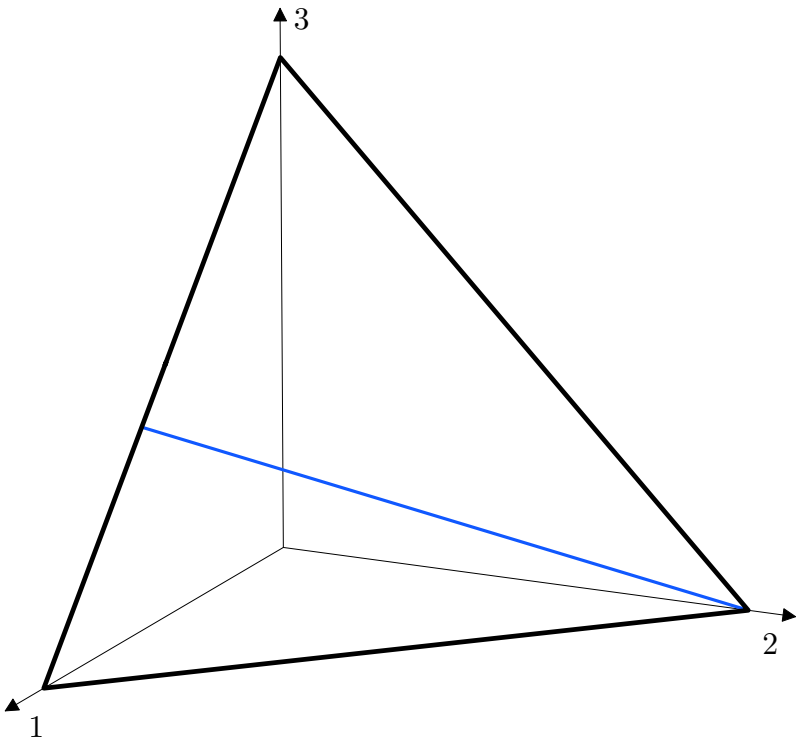


Figure 2.12: Parametrization from equation (2.20)

Now consider the parametrization from (2.10) – (2.12), depicted in Figure 2.13. We saw in the previous section that at $p^* = \phi(\theta^*), \theta^* = 0.5$, the pushforward of the parameter tangent vector is given by

$$\partial(\theta^*) = \partial_2|_{p^*} - \partial_1|_{p^*}.$$

The horizontal space at p^* is spanned by the vector

$$B = \frac{3}{8}\partial_1|_{p^*} - \frac{6}{8}\partial_2|_{p^*} + \frac{3}{8}\partial_3|_{p^*}.$$

Since these two vectors are not parallel, we can conclude that the statistic f is not sufficient for this parametrization. ∥

We can now state Chentsov's theorem (Chentsov, 1982; Ay et al., 2017), which provides a second justification for the Fisher metric.

Theorem 2.7 (Chentsov's theorem)**.** *Up to a positive constant factor, the Fisher metric is the unique Riemannian metric that is invariant under sufficient statistics. More precisely, suppose that for every finite state space $\mathcal{X}$ we are given a Riemannian metric $\mathfrak{h}^{\mathcal{X}}$ on $\mathcal{P}^{\mathcal{X}}$, and these metrics are compatible with sufficient statistics in the following sense: for every statistical model $\mathcal{S} \subset \mathcal{P}^{\mathcal{X}}$ and every statistic $f\colon \mathcal{X} \to \mathcal{X}'$ that is sufficient for $\mathcal{S}$,*

$$\mathfrak{h}^{\mathcal{X}}_p(A, B) = \mathfrak{h}^{\mathcal{X}'}_{f_*(p)}(df_*(A), df_*(B)) \quad \text{for all } p \in \mathcal{S},\ A, B \in T_p\mathcal{S}.$$

Then there exists a constant $c > 0$ such that for all $\mathcal{X}$, $\mathfrak{h}^{\mathcal{X}}$ coincides with $c \cdot \mathfrak{g}$ on $\mathcal{P}^{\mathcal{X}}$.

Chentsov's theorem and the Cramér–Rao lower bound give two complementary justifications for the Fisher metric as the canonical geometric structure on a statistical model.

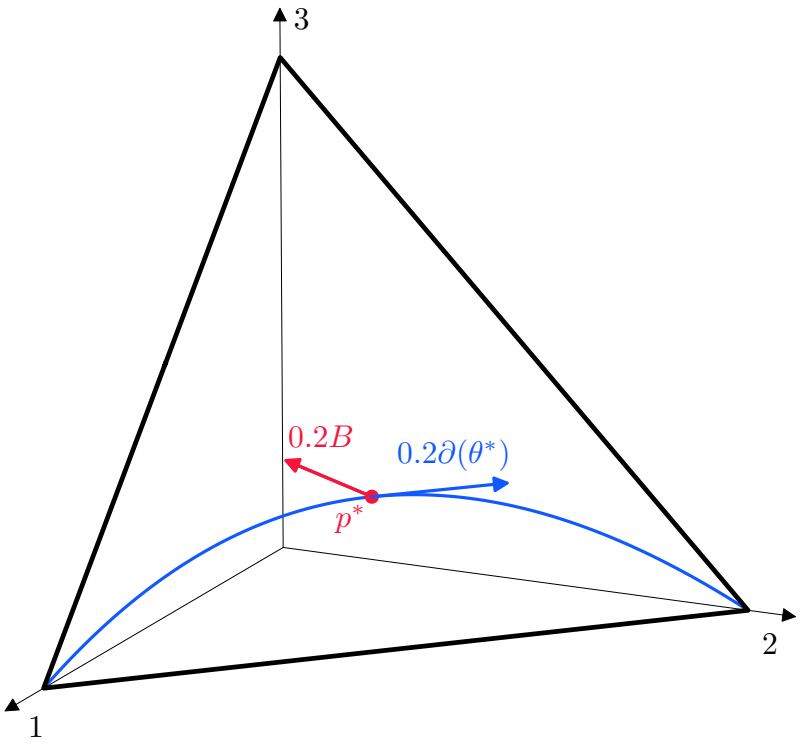


Figure 2.13: Parametrization from equations (2.10) – (2.12) and the (scaled) pushforward of the parameter tangent vector and the vector spanning the horizontal space at p^*

On the one hand, it is the unique metric preserved by every sufficient statistic — that is, by any data reduction that retains all information about the parameter. A sufficient statistic can be regarded as a lossless compression of the data, and the theorem says that the Fisher metric is exactly the structure that such compressions preserve. On the other hand, via the Cramér–Rao lower bound, the Fisher metric governs the fundamental limits of statistical estimation: it measures how precisely the parameter can be recovered from observations, and thereby how easily nearby distributions can be distinguished.

We close the chapter with a reformulation of Chentsov's theorem in the language of Riemannian submersions. Rather than singling out, for each model, a class of statistics under which the metric must be invariant, the next theorem asks that *every* statistic be a Riemannian submersion of the corresponding probability simplices — and concludes that the Fisher metric is the only metric for which this can hold.

Theorem 2.8. *Suppose that for every finite state space $\mathcal{X}$ we are given a Riemannian metric $\mathfrak{h}^{\mathcal{X}}$ on $\mathcal{P}^{\mathcal{X}}$ such that every statistic f is a Riemannian submersion. Then there exists a constant $c > 0$ such that for all $\mathcal{X}$, $\mathfrak{h}^{\mathcal{X}}$ coincides with $c \cdot \mathfrak{g}$ on $\mathcal{P}^{\mathcal{X}}$.*

Proof. We derive the theorem from a related formulation of Chentsov's theorem in (Ay et al., 2017), which characterizes the Fisher metric through invariance under *congruent Markov kernels* rather than statistics.

A *Markov kernel* from $\mathcal{X}'$ to $\mathcal{X}$ is a map

$$\begin{aligned} K\colon \mathcal{X}' &\to \mathcal{P}^{\mathcal{X}}, \\ y &\mapsto K(y) = \sum_{x\in\mathcal{X}} k(x|y)\,\delta^x, \end{aligned}$$

which assigns to each $y \in \mathcal{X}'$ a probability distribution $K(y)$ on $\mathcal{X}$. Each Markov kernel induces a map between probability distributions

$$\begin{aligned} K_*\colon \mathcal{P}^{\mathcal{X}'} &\to \mathcal{P}^{\mathcal{X}}, \\ p' &\mapsto \sum_{y\in\mathcal{X}'} p'(y)\,K(y) = \sum_{x\in\mathcal{X}} \left(\sum_{y\in\mathcal{X}'} p'(y)\,k(x|y)\right)\delta^x. \end{aligned}$$

A Markov kernel $K\colon \mathcal{X}' \to \mathcal{P}^{\mathcal{X}}$ with $|\mathcal{X}'| \leq |\mathcal{X}|$ is called *congruent* if there exists a partition $\{A_y\}_{y\in\mathcal{X}'}$ of $\mathcal{X}$ such that

$$k(x|y) > 0 \quad\Longleftrightarrow\quad x \in A_y. \tag{2.21}$$

The version of Chentsov's theorem from (Ay et al., 2017) states that if for every finite state space $\mathcal{X}$ we are given a Riemannian metric $\mathfrak{h}^{\mathcal{X}}$ on $\mathcal{P}^{\mathcal{X}}$, and these metrics are compatible with every congruent Markov kernel $K\colon \mathcal{X}' \to \mathcal{P}^{\mathcal{X}}$ in the sense that

$$\mathfrak{h}^{\mathcal{X}'}_{p'}(A', B') = \mathfrak{h}^{\mathcal{X}}_{K_*(p')}\big(dK_*(A'),\, dK_*(B')\big) \tag{2.22}$$

for all $p' \in \mathcal{P}^{\mathcal{X}'}$ and $A', B' \in T_{p'}\mathcal{P}^{\mathcal{X}'}$, then there exists a constant $c > 0$ such that $\mathfrak{h}^{\mathcal{X}} = c \cdot \mathfrak{g}$ on $\mathcal{P}^{\mathcal{X}}$ for every $\mathcal{X}$.

It therefore suffices to show that the assumption of the theorem—that every statistic is a Riemannian submersion with respect to the metrics $\mathfrak{h}^{\mathcal{X}}$—implies the invariance (2.22) for every congruent Markov kernel.

Let $K\colon \mathcal{X}' \to \mathcal{P}^{\mathcal{X}}$ be a congruent Markov kernel with associated partition $\{A_y\}_{y\in\mathcal{X}'}$ of $\mathcal{X}$. Since the partition sets A_y are non-empty, we may define a surjective statistic

$$f\colon \mathcal{X} \to \mathcal{X}', \qquad f(x) = y \;\text{ if and only if }\; x \in A_y,$$

so that the fibers of f are exactly the partition sets, $f^{-1}(y) = A_y$.

Step 1: K_* is a section of f_*. Let $p' \in \mathcal{P}^{\mathcal{X}'}$ and write $p = K_*(p')$. The congruence condition (2.21) implies that $k(x|y)$ is non-zero only for $y = f(x)$, so

$$p(x) = \sum_{y\in\mathcal{X}'} p'(y)\,k(x|y) = p'(f(x))\,k(x|f(x)). \tag{2.23}$$

Summing over the fiber $f^{-1}(y) = A_y$ and using $\sum_{x\in A_y} k(x|y) = 1$, we obtain

$$f_*(p)(y) = \sum_{x\in A_y} p(x) = p'(y) \sum_{x\in A_y} k(x|y) = p'(y),$$

so $f_* \circ K_* = \mathrm{id}_{\mathcal{P}^{\mathcal{X}'}}$. Differentiating yields

$$df_* \circ dK_* = \mathrm{id}_{T\mathcal{P}^{\mathcal{X}'}}\,. \tag{2.24}$$

Step 2: dK_* takes values in the horizontal space. Let $A' \in T_{p'}\mathcal{P}^{\mathcal{X}'}$ and consider the curve $p_t^{A'} = p' + tA'$. From (2.23) we have

$$K_*(p_t^{A'})(x) = \big(p'(f(x)) + t\,A'(f(x))\big)\,k(x|f(x)),$$

and therefore

$$dK_*(A')(x) = A'(f(x))\,k(x|f(x)) = p(x|f(x))\,A'(f(x)),$$

where in the last step we used (2.23) to identify $k(x|f(x)) = p(x)/p'(f(x)) = p(x|f(x))$. By Proposition 2.1, this is exactly the form of a horizontal vector at p with respect to f, so $dK_*(A') \in \mathcal{H}_p$.

Step 3: Invariance from the Riemannian submersion property. Let $A', B' \in T_{p'}\mathcal{P}^{\mathcal{X}'}$. By Step 2, $dK_*(A'), dK_*(B') \in \mathcal{H}_p$, so the assumption that f_* is a Riemannian submersion with respect to $\mathfrak{h}^{\mathcal{X}}$ and $\mathfrak{h}^{\mathcal{X}'}$ gives

$$\begin{aligned}\mathfrak{h}^{\mathcal{X}}_p\big(dK_*(A'),\, dK_*(B')\big) &= \mathfrak{h}^{\mathcal{X}'}_{f_*(p)}\big(df_*(dK_*(A')),\, df_*(dK_*(B'))\big)\\ &= \mathfrak{h}^{\mathcal{X}'}_{p'}(A', B'),\end{aligned}$$

where in the second line we used $f_*(p) = p'$ together with (2.24). This is precisely (2.22), and the theorem follows. □

The submersion formulation of Chentsov's theorem is appealing for two reasons. First, it is intrinsically Riemannian: the invariance condition is rephrased as a property of the metric itself, without reference to any particular statistical model. Second, it makes the symmetry between sample-space reductions and the geometry of probability simplices fully explicit — every statistic is, geometrically, a Riemannian submersion, and the Fisher metric is exactly the structure that makes this so. Riemannian submersions will reappear several times in this thesis as a tool for relating geometries on related spaces; the present formulation of Chentsov's theorem can be read as a statement that, in the statistical setting, this geometric notion already captures the right notion of sufficiency.

2.4 Appendix

Lemma 2.3. *Let M, g be a Riemannian manifold and $S \subset M$ a submanifold. Let $f\colon M \to \mathbb{R}$ be a function on M. Let us denote the gradient of f on M at $p \in M$ by $\operatorname{grad}_p f$, and the gradient of f on S at $p \in S$ by $\operatorname{grad}^S_p f$, and finally for a tangent vector $v \in T_pM$, $p \in S$ the orthogonal projection of v on T_pS by Π^S. Then we have:*

$$\operatorname{grad}^S_p f = \Pi^S(\operatorname{grad}_p f).$$

Proof. We know that $\operatorname{grad}_p f$ is characterized by the fact that for all $v \in T_pM$

$$g(\operatorname{grad}_p f, v) = df_p(v).$$

We therefore need to show that for all $v \in T_pS$ we have

$$g(\Pi^S(\operatorname{grad}_p f), v) = df_p(v).$$

Since Π^S is an orthogonal projection we know that

$$g(\Pi^S(\operatorname{grad}_p f), v) = g(\operatorname{grad}_p f, v)$$

for all $v \in T_pS$. Therefore the result holds. □

Lemma 2.4. *The (i,j)th element of the Fisher information matrix is given by*

$$G_{ij}(\theta) = -\sum_k (p_\theta)_k \frac{\partial^2 \log(p_\theta)_k}{\partial\theta^i \partial\theta^j}.$$

Proof. First note that

$$\begin{aligned}\sum_k (p_\theta)_k \frac{\partial \log(p_\theta)_k}{\partial\theta^i} &= \sum_k \frac{\partial (p_\theta)_k}{\partial\theta^i} \\ &= \frac{\partial}{\partial\theta^i}\sum_k (p_\theta)_k \\ &= \frac{\partial}{\partial\theta^i} 1 \\ &= 0 \end{aligned} \tag{2.25}$$

Now we compute

$$\begin{aligned}&\frac{\partial}{\partial\theta^j}\sum_k (p_\theta)_k \frac{\partial \log(p_\theta)_k}{\partial\theta^i} \\ &\quad= \sum_k \frac{\partial (p_\theta)_k}{\partial\theta^j}\frac{\partial \log(p_\theta)_k}{\partial\theta^i} + (p_\theta)_k \frac{\partial^2 \log(p_\theta)_k}{\partial\theta^i\partial\theta^j} \\ &\quad= \sum_k (p_\theta) \frac{\partial \log(p_\theta)_k}{\partial\theta^j}\frac{\partial \log(p_\theta)_k}{\partial\theta^i} + \sum_k (p_\theta)_k \frac{\partial^2 \log(p_\theta)_k}{\partial\theta^i\partial\theta^j}\end{aligned}$$

Now using both (2.5) and (2.25) gives the result. □

Lemma 2.5. *Let M, g be a Riemannian manifold $p \in U \subset M$ and $\{x^i\}_{i\in\{1,\dots,n\}}$ local coordinates. Then we have $g(\operatorname{grad}_p x^i, \operatorname{grad}_p x^j) = \left[G_p^{-1}\right]^{ij}$ where $[G_p]_{ij} = g\left(\frac{\partial}{\partial x^i}\big|_p, \frac{\partial}{\partial x^j}\big|_p\right)$.*

Proof. For any smooth function $f : M \to \mathbb{R}$ we know

$$\operatorname{grad}_p f = \sum_{i,j} \left[G_p^{-1}\right]^{ij} \frac{\partial f}{\partial x^j} \left.\frac{\partial}{\partial x^i}\right|_p.$$

Therefore we have

$$\begin{aligned} g(\operatorname{grad}_p x^i, \operatorname{grad}_p x^j) &= g\left(\sum_k \left[G_p^{-1}\right]^{ki} \left.\frac{\partial}{\partial x^k}\right|_p, \sum_l \left[G_p^{-1}\right]^{lj} \left.\frac{\partial}{\partial x^l}\right|_p\right) \\ &= \sum_{k,l} \left[G_p^{-1}\right]^{ki} \left[G_p^{-1}\right]^{lj} [G_p]_{kl} \\ &= \left[G_p^{-1}\right]^{ij}. \end{aligned}$$

□

Lemma 2.6. *Let V be an inner product space, and let $W \subseteq V$ be a subspace. For $v^1, \dots, v^d \in V$, define the Gram matrices*

$$\begin{aligned}\tilde{H}^{ij} &= \langle v^i, v^j\rangle, \\ H^{ij} &= \langle \Pi(v^i), \Pi(v^j)\rangle.\end{aligned}$$

where $\Pi : V \to W$ denotes the orthogonal projection onto W. Then the matrix $\tilde{H} - H$ is positive semidefinite.

Proof. Let $c = (c_1, ..., c_d) \in \mathbb{R}^d$. We need to show that

$$c^\intercal(\tilde{H} - H)c \geq 0.$$

Note that

$$c^\intercal \tilde{H} c = \langle u, u \rangle,$$
$$c^\intercal H c = \langle \Pi(u), \Pi(u) \rangle,$$

where $u = \sum_{i=1}^{d} c_i v^i \in V$. Since Π is a projection we know $\|u\| \geq \|\Pi(u)\|$ which gives the result. □

References

Amari, Shun-ichi and Hiroshi Nagaoka (2000). *Methods of Information Geometry*. Vol. 191. Translations of Mathematical Monographs. Providence, RI: American Mathematical Society.

Ay, Nihat, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer (2017). *Information Geometry*. Vol. 64. Ergebnisse der Mathematik und ihrer Grenzgebiete. Cham: Springer.

Chentsov, Nikolai Nikolaevich (1982). *Statistical Decision Rules and Optimal Inference*. Vol. 53. Translations of Mathematical Monographs. Providence, RI: American Mathematical Society.

Cramér, Harald (1946). *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.

Lee, John M (2018). *Introduction to Riemannian manifolds*. Vol. 2. Springer.

Rao, C. Radhakrishna (1945). "Information and the Accuracy Attainable in the Estimation of Statistical Parameters". In: *Bulletin of the Calcutta Mathematical Society* 37, pp. 81–91.

Part II

Papers

Chapter 3

Bures–Wasserstein Geometry for Positive-Definite Hermitian Matrices and Their Trace-One Subset

Jesse van Oostrum

Information Geometry, Springer, 2022

Abstract. In his classical argument, Rao derives the Riemannian distance corresponding to the Fisher metric using a mapping between the space of positive measures and Euclidean space. He obtains the Hellinger distance on the full space of measures and the Fisher distance on the subset of probability measures. In order to highlight the interplay between Fisher theory and quantum information theory, we extend this construction to the space of positive-definite Hermitian matrices using Riemannian submersions and quotient manifolds. The analog of the Hellinger distance turns out to be the Bures-Wasserstein (BW) distance, a distance measure appearing in optimal transport, quantum information, and optimisation theory. First we present an existing derivation of the Riemannian metric and geodesics associated with this distance. Subsequently, we present a novel derivation of the Riemannian distance and geodesics for this metric on the subset of trace-one matrices, analogous to the Fisher distance for probability measures.

Keywords: Information geometry • positive-definite matrices • Bures distance • Wasserstein metric • Optimal transport • Quantum information

[0]The author wishes to thank Giovanni Pistone for supervising this research project and acknowledges the support of the Collegio Carlo Alberto and the Prins Bernhard Cultuurfonds. This work was presented at the Information Geometry seminar in Turin, December 2019.

3.1 Introduction

In this paper we investigate the geometrical properties of the Bures-Wasserstein (BW) distance on the space of positive-definite Hermitian matrices, $\mathbb{P}(n)$. For $\Sigma, T \in \mathbb{P}(n)$, this distance function is given by:

$$d^{BW}_{\mathbb{P}(n)}(\Sigma, T) = \left[\operatorname{tr}(\Sigma) + \operatorname{tr}(T) - 2\operatorname{tr}\left(\left(\Sigma^{1/2} T \Sigma^{1/2}\right)^{1/2}\right)\right]^{1/2}. \tag{3.1}$$

This function appears in optimal transport as a distance measure on the space of mean-zero Gaussian densities, where it is called the Wasserstein distance. In quantum information theory, this is a distance measure between quantum states or density matrices, called the Bures distance.

The development of this subject started when Rao realised that the Fisher information defines a Riemannian metric on the space of probability measures (Rao, 1992). He obtained this metric by finding an isometry between the positive orthant of the unit sphere, equipped with the Euclidean metric, and the probability simplex. Later, the study of the geometrical properties of the probability simplex was extended to the quantum realm with notable contributions of Nagaoka, Petz and Hasegawa. An overview of this field can be found in Hayashi (2017) and Bengtsson and Życzkowski (2017). The distance measure defined in (3.1) was introduced by Helstrom (1967) and Bures (1969) as a measure of similarity between quantum states. Uhlmann (1993) derives the geometrical properties of this distance measure.

In the context of optimal transport, this distance measure was derived to be the L^2-Wasserstein distance on the space of covariance matrices for mean zero Gaussian distributions (Olkin and Pukelsheim, 1982). Its geometrical properties were first studied in this context by Takatsu (2011). It is interesting to note that the obtained results are equivalent to but independent of the work of Uhlmann fifteen years prior. See Section 3.6 for more on this. The Levi-Civita connection and exponential map are computed in Malagò et al. (2018). A further question at the end of Malagò et al. (2018) on the restriction to trace-one matrices formed the start of the research for this paper. Bhatia et al. (2019) discusses both the quantum and the optimal transport interpretation of the distance and introduces the name Bures-Wasserstein distance. Furthermore, the approach of Takatsu is refined using facts on Riemannian submersions and quotient manifolds. The curvature and parallel transport for the BW metric are examined in Massart et al. (2019) and Thanwerdas and Pennec (2021). More recently, the geometrical structure discussed in this paper is of interest in the field of optimisation. It turns out that for this choice of geometry the exponential and logarithmic map are cheap to evaluate, which makes it particularly suitable for numerical computations (Massart and Absil, 2020). In the present paper, we adapt the quotient manifold construction from Bhatia to the subspace of trace-one positive-definite Hermitian matrices, which is of particular importance in quantum information. We recover the Riemannian distance for the BW metric on this space, which is called the *Bures-Wasserstein angle*, and its the geodesics. This distance measure and geodesics were first derived by Uhlmann in the context of quantum information (Uhlmann, 1993).

The main goal of this paper is to highlight the interplay between Fisher theory and quantum information theory. This is done by showing the parallel between the classical derivation of the Hellinger and Fisher distance due to Rao, and the derivation of the BW distance and angle. For this, we present a novel rigorous derivation of the BW angle that emphasises the geometrical aspects of the BW metric and does not utilise concepts from quantum information. This derivation is an extension of recent derivations for general positive-definite matrices, as presented in (Bhatia et al., 2019; Massart and Absil, 2020). We then show how our derivation can be translated into the language of quantum information. A particular source of confusion is the use of different representations of the tangent space, as is also noted in (Thanwerdas and Pennec, 2021). We present clear and general definitions of the the (m)- and (e)-representation used in quantum information in the language of differential geometry and show how different quantum information metrics can be denoted in each representation.

The next section of the paper discusses preliminary facts on differential geometry and matrix identities and gives general definitions for the (m)- and (e)-representation, which coincide with existing definitions from both classical and quantum information geometry. Section 3.3 illustrates the use of these representations in the context of classical information geometry and describes the derivation of the Hellinger and Fisher distance due to Rao. In Section 3.4 the representations of the tangent space are worked out explicitly for the SLD Fisher metric and the Bogoliubov metric used in quantum information. The main derivations of the paper can be found in Section 3.5, where the geometrical structure of the BW distance is investigated first on $\mathbb{P}(n)$ and subsequently on the the trace-one subset, $\mathcal{D}(n)$. The last section of the paper compares the geometrical structure obtained in the foregoing to similar results in the field. An overview of the notation used in the paper can be found on page 64.

3.2 Preliminaries

3.2.1 Differential geometry

Let $\mathcal{M}, \mathcal{N}$ be smooth manifolds and $F : \mathcal{M} \to \mathcal{N}$ a smooth map. We will denote the differential of F at $p \in \mathcal{M}$ by: $dF_p : T_p\mathcal{M} \to T_{F(p)}\mathcal{N}$. For g a Riemannian metric on $\mathcal{M}$, the length of a tangent vector $v \in T_p\mathcal{M}$ is given by: $||v||_g = (g(v,v))^{1/2}$, and the length of a curve $\gamma : [a,b] \to \mathcal{M}$ is given by:

$$L(\gamma) = \int_a^b ||\gamma'(t)||_g dt$$

where $\gamma'(t_0) = d\gamma_{t_0}(\frac{d}{dt}|_{t_0})$. The Riemannian distance between $p, q \in \mathcal{M}$ is defined to be:

$$d_{\mathcal{M}}(p,q) = \inf\{L(\gamma) \mid \gamma : [0,1] \to \mathcal{M}, \gamma(0) = p, \gamma(1) = q\}. \tag{3.2}$$

See chapter 2 of Lee (2018) for details.

Definition 3.1. Let V be a real or complex vector space. An *affine subspace* of V is a subset $A \subset V$ together with a vector subspace $\widetilde{V} \subset V$ such that:

- $\forall a, b \in A,\ \exists v \in \widetilde{V}$ such that $a + v = b$

- $\forall v \in \widetilde{V}$ and $a \in A$ we have: $a + v \in A$

Now let $\mathcal{M}$ be an open convex subset of an affine subspace A of V with associated vector space $\widetilde{V}$ and $p \in \mathcal{M}$ fixed. We have that the following map is a vector space isomorphism (Lee, 2012):

$$\begin{aligned} \mathrm{id}_1 : \widetilde{V} &\to T_p\mathcal{M} \\ \tilde{v} &\mapsto v \\ \mathrm{id}_1(\tilde{v})(f) &\coloneqq \left.\frac{d}{dt}\right|_{t=0} f(p + t\tilde{v}), \end{aligned}$$

where f is any smooth function on $\mathcal{M}$. We can therefore identify every tangent vector in $T_p\mathcal{M}$ with an element of $\widetilde{V}$ through id_1. Given a basis $(e_1, ..., e_m)$ for V, we define the Euclidean inner product on V to be such that $\langle e_i, e_j\rangle^{\mathrm{Eucl}} \coloneqq \delta_{ij}$. The *Euclidean metric* $\bar{\mathfrak{g}}$ on $\mathcal{M}$ is defined such that for $\tilde{v}, \tilde{w} \in \widetilde{V}$, we have $\bar{\mathfrak{g}}_p(\mathrm{id}_1(\tilde{v}), \mathrm{id}_1(\tilde{w})) = \mathrm{Re}\left(\langle \tilde{v}, \tilde{w}\rangle^{\mathrm{Eucl}}\right)$. The Riemannian distance associated with $\bar{\mathfrak{g}}$ is denoted $\bar{d}$.

3.2.2 Representations of the tangent space

Let $\mathbb{K}$ be either $\mathbb{R}$ or $\mathbb{C}$ and $(e_1, ..., e_m)$ a fixed basis for V. This basis induces a coordinate map $k : V \to \mathbb{K}^m$ such that for $k(v) = x$ we have $v = \sum_i x_i e_i$. We define the (m)-*representation* of an element in $T_p\mathcal{M}$ to be the coordinate representation of its id_1-associated element in $\widetilde{V}$. Or in symbols,

$$(m) : T_p\mathcal{M} \xrightarrow{\mathrm{id}_1} \widetilde{V} \xrightarrow{k} \mathbb{K}^m.$$

Using the Riesz representation theorem we let id_2 be the identification between $T_p\mathcal{M}$ with its dual $T_p^*\mathcal{M}$ through the Euclidean metric such that: $v \leftrightarrow v'(\cdot) = \bar{\mathfrak{g}}_p(v, \cdot)$. We define the (m^*)-representation of an element of $T_p^*\mathcal{M}$ to be the (m)-representation of its id_2-associated element in $T_p\mathcal{M}$. In symbols,

$$(m^*) : T_p^*\mathcal{M} \xrightarrow{\mathrm{id}_2} T_p\mathcal{M} \xrightarrow{(m)} \mathbb{K}^m.$$

A general Riemannian metric $\mathfrak{g}$ on $\mathcal{M}$ gives a final identification, $\mathrm{id}_3(\mathfrak{g})$, between $T_p\mathcal{M}$ and $T_p^*\mathcal{M}$: $v \leftrightarrow v'(\cdot) = \mathfrak{g}_p(v, \cdot)$. Given a metric $\mathfrak{g}$, we define the (e)-*representation* of an element of $T_p\mathcal{M}$ as the (m^*)-representation of its $\mathrm{id}_3(\mathfrak{g})$-associated element in $T_p^*\mathcal{M}$. In symbols,

$$(e) : T_p\mathcal{M} \xrightarrow{\mathrm{id}_3(\mathfrak{g})} T_p^*\mathcal{M} \xrightarrow{(m^*)} \mathbb{K}^m.$$

Note that the definition of the (m)- and (e)-representation implies:

$$\mathfrak{g}_p(v, w) = \mathrm{Re}\left(\langle v^{(e)}, w^{(m)}\rangle\right),$$

with on the right the standard inner product on $\mathbb{K}^m$.

Remark 3.1. The definitions above are inspired by Chapter 2 of Ay et al. (2017). We will see in Sections 3.3 and 3.4 that they correspond to the standard definitions in classical and quantum information theory given in e.g. (Amari and Nagaoka, 2007) and (Hayashi, 2017).

3.2.3 Matrix identities

Let $\mathrm{GL}(n)$ be the space of invertible complex matrices and $\mathrm{U}(n)$ the unitary matrices. Every $M \in \mathrm{GL}(n)$ can be written as $M = U\Sigma$ where $\Sigma = (M^*M)^{1/2} \in \mathbb{P}(n)$ and $U = M\Sigma^{-1} \in \mathrm{U}(n)$. This is called the *polar decomposition* of M and U is called the *unitary polar factor*. In the following result, related to Uhlmann's theorem and the Procrustes problem, the unitary polar factor shows up in a maximisation problem.

Theorem 3.1. *Let $\Sigma, T \in \mathbb{P}(n)$ and consider the following maximisation problem:*

$$\sup_{V \in \mathrm{U}(n)} \mathrm{Re}\,\mathrm{tr}\,(\Sigma V T)\,.$$

Then the supremum is attained for $V = U^*$, with U the unitary polar factor of $T\Sigma$.

Proof. See the proof of Theorem 1 in Bhatia et al. (2019), or Theorem 9.2 in Bengtsson and Życzkowski (2017). □ □

Definition 3.2. The solution for X to the *Lyapunov equation*: $\Sigma X + X\Sigma = H$ with $\Sigma \in \mathbb{P}(n), H \in \mathbb{H}(n)$ will be denoted $\mathcal{L}_\Sigma(H)$. It turns out that this solution exists and is unique (Bhatia, 2009).

3.3 Classical information geometry

In this section we will study $\mathcal{M}_+(\Omega)$ and $\mathcal{P}_+(\Omega)$, the space of stricly positive (resp. probability) measures on $\Omega = \{\omega_1, ..., \omega_n\}$. These spaces are open convex subsets of affine subspaces of the vector space of signed measures $\mathcal{S}(\Omega)$ and therefore we can use the definitions from the previous section. The canonical basis of $\mathcal{S}(\Omega)$ is given by the set of Dirac delta measures $(\delta_1, ..., \delta_n)$ such that $\delta_i(\omega_j) = \delta_{ij}$. This basis gives us the (m)-representation as described in the preliminaries. Now we let the metric on $\mathcal{M}_+(\Omega)$ and $\mathcal{P}_+(\Omega)$ be the Fisher information metric, given by:

$$\mathfrak{g}^F_\mu(a, b) = \sum_{i=1}^n \frac{a_i^{(m)} b_i^{(m)}}{\mu(\omega_i)},$$

with a, b in the tangent space of $\mathcal{M}_+(\Omega)$ respectively $\mathcal{P}_+(\Omega)$ at μ. The (e)-representation is given by:

$$a_i^{(e)} = \frac{a_i^{(m)}}{\mu(\omega_i)}.$$

Another way of obtaining the (e)-representation for this case is by applying the (m)-representation to the pushforward of a under the logarithm map. If $\log : \mathcal{M}_+(\Omega) \ni \mu \mapsto \log(\mu) \in \mathcal{S}(\Omega)$ such that $\log(\mu)(\omega) = \log(\mu(\omega))\ \forall \omega \in \Omega$, then:

$$a^{(e)} = \big(d\log_\mu(a)\big)^{(m)}\,.$$

We will see however that this expression of the (e)-representation is not general enough for the quantum case.

3.3.1 Hellinger distance, Fisher metric and Fisher distance

In this section we derive the Riemannian metric corresponding to the Hellinger distance on $\mathcal{M}_+(\Omega)$. Then we find the Riemannian distance corresponding to the restriction of this metric to $\mathcal{P}_+(\Omega)$. These will turn out to be the Fisher metric and Fisher distance respectively. This derivation was first due to Rao (Rao, 1992) and can be seen as a special case of the derivation in Section 3.5.

The Hellinger distance is given by:

$$d^H(\mu,\nu) = \sqrt{\sum_{i=1}^{n} \left(\mu(\omega_i)^{1/2} - \nu(\omega_i)^{1/2}\right)^2}.$$

It can be obtained as the pushforward of the Euclidean distance, $\bar{d}$, under the square map:

$$(\mathcal{M}_+(\Omega), \bar{d}) \ni \mu \mapsto \mu^2 \in (\mathcal{M}_+(\Omega), d^H), \tag{3.3}$$

where $\mu^2(\omega) \coloneqq \mu(\omega)^2 \quad \forall \omega \in \Omega$. Extending the geometrical structure on the left from a distance function to the corresponding Riemannian Euclidean metric $\bar{\mathfrak{g}}$, we have the following isometry[1]:

$$(\mathcal{M}_+(\Omega), \bar{\mathfrak{g}}) \ni \mu \mapsto \mu^2 \in \left(\mathcal{M}_+(\Omega), \mathfrak{g}^F\right). \tag{3.4}$$

From (3.3) and (3.4) it follows that the Hellinger distance is the geodesic distance for the Fisher metric in $\mathcal{M}_+(\Omega)$.

We now aim to find the Riemannian distance for the Fisher metric restricted to $\mathcal{P}_+(\Omega)$. It turns out that for this subset, the geodesic distance is no longer the Hellinger distance. In order to find the right geodesic distance, we can use the fact that the following restriction of (3.4) remains an isometry:

$$\left(\mathbb{S}_{\mathcal{M}_+(\Omega)}, \bar{\mathfrak{g}}\right) \ni \mu \mapsto \mu^2 \in \left(\mathcal{P}_+(\Omega), \mathfrak{g}^F\right).$$

where $\mathbb{S}_{\mathcal{M}_+(\Omega)} \coloneqq \{\mu \in \mathcal{M}_+(\Omega) : \sum_i \mu(\omega_i)^2 = 1\}$, the unit sphere in $(\mathcal{M}_+(\Omega), \bar{\mathfrak{g}})$ (see Figure 3.1). We know that on this space, the geodesics are given by great circles and therefore we can also compute the Riemannian distance. Using the fact that this distance is carried over by the isometry, we obtain the Riemannian distance for the space of probability measures with the Fisher metric, called the Fisher distance. This is given by:

$$d^F(p,q) = \arccos\left(\sum_{i=1}^{n} \left(p(\omega_i)q(\omega_i)\right)^{1/2}\right).$$

3.4 Quantum information geometry

Let $\mathbb{H}(n)$ be the set of Hermitian matrices and $\mathcal{D}(n)$ be the subset of positive-definite Hermitian matrices with trace one. Within the context of Section 3.2, we have $\mathcal{M} =$

[1]The isometries in this section are defined up to a constant factor.

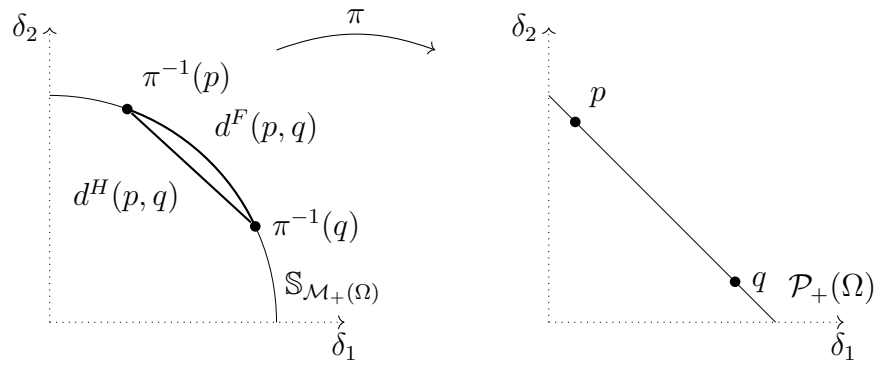


Figure 3.1: A mapping (π) from the sphere ($\mathbb{S}_{\mathcal{M}_+}(\Omega)$) to the simplex ($\mathcal{P}_+(\Omega)$) and an illustration of the relation between the Hellinger (d^H) and Fisher (d^F) distance

$\mathcal{D}(n)$, $V = \mathbb{C}^{n\times n}$ and $\widetilde{V} = \{H \in \mathbb{H}(n) : \mathrm{tr}(H) = 0\}$. The basis vectors for $\mathbb{C}^{n\times n}$ are simply given by $(A_{11}, A_{12}, ..., A_{nn})$, where the ij-th entry of A_{ij} is one and the rest zero. From this, we get the (m)-representation for $T_\rho\mathcal{D}(n)$. For the submanifold of diagonal matrices (probability measures) Chentsov showed that the Fisher metric is the unique metric satisfying certain (statistically) natural conditions on the metric (Morozova and Chentsov, 1989). Petz proved that for $\mathcal{D}(n)$, this uniqueness no longer exists (Petz, 1996). One of the suggested generelisations is the symmetrised logarithmic derivative (SLD) Fisher metric, see e.g. (Amari and Nagaoka, 2007; Hayashi, 2017). For $H, K \in T_\rho\mathcal{D}(n)$ this Hermitian metric is given explicitly by:

$$\mathfrak{g}^{SLD}_\rho(H,K) = 2\,\mathrm{tr}\left(\mathcal{L}_\rho\left(H^{(m)}\right)K^{(m)}\right).$$

We will derive in Section 3.5 that the Riemannian metric corresponding to the BW distance on $\mathbb{P}(n)$ is given by:

$$\mathfrak{g}^{BW}_\Sigma(H,K) = \frac{1}{2}\,\mathrm{Re}\,\mathrm{tr}\left(\mathcal{L}_\Sigma\left(H^{(m)}\right)K^{(m)}\right).$$

Furthermore we will prove that the Riemannian distance on $\mathcal{D}(n)$ for this metric is given by:

$$d^{BW}_{\mathcal{D}(n)}(\rho_1,\rho_2) = \arccos\left(\mathrm{Re}\,\mathrm{tr}\left(\left(\rho_2^{1/2}\rho_1\rho_2^{1/2}\right)^{1/2}\right)\right).$$

Because the real parts of $\mathfrak{g}^{SLD}$ and $\mathfrak{g}^{BW}$ are equal on $\mathcal{D}(n)$, we can conclude that $d^{BW}_{\mathcal{D}(n)}$ is the distance function for $\mathfrak{g}^{SLD}$ (up to a constant factor).

(e)-representations in quantum information geometry

The SLD Fisher metric for $H, K \in T_\rho\mathcal{D}(n)$ is given up to a constant by:

$$\mathfrak{g}^{SLD}_\rho(H,K) = \mathrm{tr}\left(\mathcal{L}_\rho\left(H^{(m)}\right)K^{(m)}\right).$$

From the preliminaries it follows that for this choice of metric we have the following relation between the (m)- and (e)-representation:

$$H^{(e)} = \mathcal{L}_\rho\left(H^{(m)}\right),$$
$$H^{(m)} = H^{(e)}\rho + \rho H^{(e)}.$$

Expressing the SLD Fisher metric in terms of the (e)-representation therefore gives the potentially more familiar expression:

$$\mathfrak{g}_\rho^{SLD}(H,K) = \operatorname{tr}\left(H^{(e)}\left(K^{(e)}\rho + \rho K^{(e)}\right)\right).$$

Another common metric is the Bogoliubov metric. In the (m)-representation this is given by:

$$\mathfrak{g}_\rho^{Bo}(H,K) = \operatorname{tr}\left((d\log_\rho(H))^{(m)}K^{(m)}\right),$$

where log is the inverse of the matrix exponential $\exp(\Sigma) = \sum_{k=0}^{\infty}\frac{1}{k}\Sigma^k$. The relation between the (m)- and (e)-representation is given by:

$$H^{(e)} = (d\log_\rho(H))^{(m)},$$
$$H^{(m)} = \int_0^1 \rho^\lambda H^{(e)}\rho^{1-\lambda}d\lambda.$$

The Bogoliubov metric in (e)-representation is therefore given by:

$$\mathfrak{g}_\rho^{Bo}(H,K) = \operatorname{tr}\left(H^{(e)}\int_0^1 \rho^\lambda K^{(e)}\rho^{1-\lambda}d\lambda\right).$$

Remark 3.2. In the rest of the paper we will exclusively and implicitely use the (m)-representation for the elements of the tangent bundle of $\mathbb{P}(n)$ and $\mathcal{D}(n)$.

3.5 Bures-Wasserstein Geometry

In this section, we explore the geometry induced by the Bures-Wasserstein distance. We start by recovering the metric and geodesics corresponding to the BW distance on $\mathbb{P}(n)$. For this, we follow the exposition in Bhatia et al. (2019) and provide extra justification for some of the results.[2] Subsequently, we restrict the obtained metric to $\mathcal{D}(n)$ and derive the corresponding distance function and geodesics. The flow of the argument is analogous to Section 3.3.1, where we start from the Hellinger distance on $\mathcal{M}_+(\Omega)$, derive the Fisher metric, and subsequently find the Riemannian distance and geodesics for this metric restricted to the submanifold $\mathcal{P}_+(\Omega)$. We start by discussing some general results from Riemannian geometry.

Let $(\mathcal{M},g)$ and $(\mathcal{N},h)$ be Riemannian manifolds and $\pi:(\mathcal{M},g)\to(\mathcal{N},h)$ a smooth submersion. We can make the following orthogonal decomposition of the tangent space at $p\in\mathcal{M}$:

$$T_p\mathcal{M} = \mathcal{V}(\pi,p)\oplus\mathcal{H}(\pi,p,g),$$

where $\mathcal{V}(\pi,p)$ is the kernel of $d\pi_p$ and $\mathcal{H}(\pi,p,g)$ is its orthogonal complement with respect to the metric at p. We will refer to these subspaces and their elements as *vertical* and *horizontal* respectively. A curve γ in $\mathcal{M}$ is said to be horizontal if $\gamma'(t)$ is horizontal for all t. We say that a submersion π is *Riemannian* if for all $p\in\mathcal{M}$ and $v,w\in\mathcal{H}(\pi,p,g)$ the following holds:

$$g_p(v,w) = h_{\pi(p)}(d\pi_p v, d\pi_p w). \tag{3.5}$$

[2]This derivation is presented here because it lays the foundation for the results derived in the second part of the section and it allows for a better comparison with the classical argument of Rao.

That is, the map $d\pi_p|_{\mathcal{H}(\pi,p,g)}$ is an isometric isomorphism.

Theorem 3.2. *Let $\pi : (\mathcal{M}, g) \to (\mathcal{N}, h)$ be a Riemannian submersion. For every horizontal curve γ in $\mathcal{M}$ we have:*

- $L(\gamma) = L(\pi \circ \gamma)$.

For every geodesic γ in $\mathcal{M}$ such that $\gamma'(0)$ is horizontal we have:

- *$\gamma'(t)$ is horizontal for all t,*
- *$\pi \circ \gamma$ is a geodesic in $\mathcal{N}$.*

Proof. Lemma 26.11 in Michor (2008) and Proposition 2.109 in Gallot et al. (1990). □ □

Theorem 3.3. *Suppose $(\mathcal{M}, g)$ is a Riemannian manifold and G a compact Lie group of isometries of $(\mathcal{M}, g)$ acting freely and smoothly on $\mathcal{M}$. Then the orbit space $\mathcal{N} = \mathcal{M}/G$ has a unique smooth manifold structure and there exists a unique Riemannian metric h such that the quotient map $\pi : (\mathcal{M}, g) \to (\mathcal{N}, h)$ is a Riemannian submersion. Furthermore the Riemannian distance $d_{\mathcal{N}}$ is given by quotient distance[3]:*

$$d'_{\mathcal{N}}(p, q) \coloneqq \inf_{\substack{\tilde{p}\in\pi^{-1}(p)\\ \tilde{q}\in\pi^{-1}(q)}} d_{\mathcal{M}}(\tilde{p}, \tilde{q}). \tag{3.6}$$

Proof. The first part of the theorem is a standard result and can be found in e.g. Corollary 2.29 of Lee (2018). We show the second part explicitly. Recall from equation (3.2) in the preliminaries that the Riemannian distance function on $\mathcal{N}$ is given by:

$$d_{\mathcal{N}}(p, q) = \inf\{L(\gamma) \mid \gamma : [0, 1] \to \mathcal{N}, \gamma(0) = p, \gamma(1) = q\}.$$

From Michor (2008) we know that for the type of Riemannian submersion described in the theorem, we can find for every curve γ on the RHS, a horizontal curve $\tilde{\gamma}$ in $\mathcal{M}$ such that $\pi \circ \tilde{\gamma} = \gamma$. So in particular we have $\tilde{\gamma}(0) \in \pi^{-1}(p)$ and $\tilde{\gamma}(1) \in \pi^{-1}(p)$. From Theorem 3.2 it follows that $L(\tilde{\gamma}) = L(\gamma)$. Therefore we have $d_{\mathcal{N}}(p, q) \geq d'_{\mathcal{N}}(p, q)$.

For the reverse inequality, we note that for every curve $\tilde{\gamma}$ in $\mathcal{M}$ we have:

$$\begin{aligned} L(\tilde{\gamma}) &= \int_0^1 ||\tilde{\gamma}'(t)||_g \, dt \\ &\geq \int_0^1 ||d\pi_{\tilde{\gamma}(t)}(\tilde{\gamma}'(t))||_h \, dt \\ &= \int_0^1 ||(\pi \circ \tilde{\gamma})'(t))||_h \, dt \\ &= L(\pi \circ \tilde{\gamma}), \end{aligned}$$

where in the second line we use that π is a Riemannian submersion. From this it follows immediately that $d_{\mathcal{N}}(p, q) \leq d'_{\mathcal{N}}(p, q)$. □ □

[3]The distance of a quotient metric space is formally defined as: $\inf\{d(p_1, q_1) + ... + d(p_n, q_n)\}$ with the infimum over all finite sequences $(p_1, ..., p_n), (q_1, ..., q_n)$ such that $p_1 \in \pi^{-1}(p), q_n \in \pi^{-1}(q)$. Proposition 4.1 in Lang (2006) shows that in our setting this distance reduces to (3.6).

3.5.1 Geometry on the space $\mathbb{P}(n)$

Riemannian metric and distance function

In this section we will prove the following theorem.

Theorem 3.4. *The Bures-Wasserstein distance on $\mathbb{P}(n)$ given in (3.1) is a Riemannian distance. The corresponding metric is given by:*

$$\mathfrak{g}^{BW}_\Sigma(H,K) = \operatorname{Re}\operatorname{tr}(\mathcal{L}_\Sigma(H)\Sigma\mathcal{L}_\Sigma(K)) = \frac{1}{2}\operatorname{Re}\operatorname{tr}(\mathcal{L}_\Sigma(H)K), \tag{3.7}$$

with $H, K \in T_\Sigma\mathbb{P}(n) \cong \mathbb{H}(n)$.

In order to prove Theorem 3.4 we need some preliminary results. The final proof can be found on page 58.

From Theorem 3.3 we know that there exists a metric $\tilde{h}$ such that the quotient map $(\mathrm{GL}(n), \bar{\mathfrak{g}}) \to (\mathrm{GL}(n)/\mathrm{U}(n), \tilde{h})$ is a Riemannian submersion. We make the following identification: $\mathrm{GL}(n)/\mathrm{U}(n) \ni M \cdot \mathrm{U}(n) \leftrightarrow MM^* \in \mathbb{P}(n)$. This gives us the following map:

$$\begin{aligned} \pi : (\mathrm{GL}(n), \bar{\mathfrak{g}}) &\to (\mathbb{P}(n), h) \\ M &\mapsto MM^*. \end{aligned} \tag{3.8}$$

We will derive the prelimary results in the following order. First, we find the horizontal and vertical subspaces for π (Proposition 3.1), which we use to show that h is given by (3.7) (Proposition 3.2). Next, we show that the BW distance on $\mathbb{P}(n)$ is equal to the quotient distance given in (3.6) for π (Proposition 3.3). Then we can use Theorem 3.3 to conclude that the BW distance is actually the Riemannian distance for $\mathfrak{g}^{BW}$.

Proposition 3.1. *Let $\mathbb{H}(n)$ and $\mathbb{H}^\perp(n)$ be the set of Hermitian and skew-Hermitian matrices, respectively. The vertical and horizontal space of π at $M \in (\mathrm{GL}(n), \bar{\mathfrak{g}})$ are given by:*

$$\mathcal{V}(\pi, M) = \left\{K\left(M^{-1}\right)^* : K \in \mathbb{H}^\perp(n)\right\} \tag{3.9}$$

$$\mathcal{H}(\pi, M, \bar{\mathfrak{g}}) = \left\{HM : H \in \mathbb{H}(n)\right\}. \tag{3.10}$$

Proof. We have:

$$d\pi_M(A) = AM^* + MA^*.$$

Therefore $d\pi_M(A) = 0 \iff A \in \mathcal{V}(\pi, M)$. Furthermore, we have that $\bar{\mathfrak{g}}_M\left(K\left(M^{-1}\right)^*, A\right) = \operatorname{Re}\operatorname{tr}\left(K\left(M^{-1}\right)^* A^*\right) = 0\ \forall K \in \mathbb{H}^\perp(n) \iff A \in \mathcal{H}(\pi, M, \bar{\mathfrak{g}})$. □ □

Proposition 3.2. *The metric h on $\mathbb{P}(n)$ is given by (3.7).*

Proof. Because π is a Riemannian submersion, we know from (3.5) that for $A, B \in \mathcal{H}(\pi, M, \bar{\mathfrak{g}})$, h needs to satisfy:

$$\bar{\mathfrak{g}}_M(A,B) = h_{MM^*}(d\pi_M A, d\pi_M B).$$

Working this out gives:

$$\operatorname{Re}\operatorname{tr}(AB^*) = h_{MM^*}(MA^* + AM^*, MB^* + BM^*). \tag{3.11}$$

Now we plug in $A = \tilde{H}M, B = \tilde{K}M$ for $\tilde{H}, \tilde{K} \in \mathbb{H}(n)$. Then (3.11) becomes:

$$\operatorname{Re}\operatorname{tr}(\tilde{H}MM^*\tilde{K}) = h_{MM^*}(MM^*\tilde{H} + \tilde{H}MM^*, MM^*\tilde{K} + \tilde{K}MM^*).$$

If we set $M = \Sigma^{1/2}$ and $\tilde{H} = \mathcal{L}_\Sigma(H), \tilde{K} = \mathcal{L}_\Sigma(K)$, we get for general $\Sigma \in \mathbb{P}(n)$ and $H, K \in \mathbb{H}(n)$:

$$h_\Sigma(H, K) = \operatorname{Re}\operatorname{tr}(\mathcal{L}_\Sigma(H)\Sigma\mathcal{L}_\Sigma(K)).$$

Using the properties of the trace we have:

$$\operatorname{Re}\operatorname{tr}(\mathcal{L}_\Sigma(H)\Sigma\mathcal{L}_\Sigma(K)) = \operatorname{Re}\operatorname{tr}(\mathcal{L}_\Sigma(K)\Sigma\mathcal{L}_\Sigma(H)) = \operatorname{Re}\operatorname{tr}(\mathcal{L}_\Sigma(H)\mathcal{L}_\Sigma(K)\Sigma).$$

Adding the first and last expression gives:

$$2h_\Sigma(H, K) = \operatorname{tr}\left[\mathcal{L}_\Sigma(H)\Big(\Sigma\mathcal{L}_\Sigma(K) + \mathcal{L}_\Sigma(K)\Sigma\Big)\right] = \operatorname{Re}\operatorname{tr}(\mathcal{L}_\Sigma(H)K).$$

Dividing both sides by two gives the final result. □ □

In order to show that the BW distance on $\mathbb{P}(n)$ is equal to the quotient distance of π, we first have to investigate the distance on $(\mathrm{GL}(n), \bar{\mathfrak{g}})$. We know that on $(\mathbb{C}^{n\times n}, \bar{\mathfrak{g}})$ the distance is given by: $\bar{d}_{\mathbb{C}^{n\times n}}(A, B) = ||A - B||_2 = [\operatorname{tr}((A - B)(A - B)^*)]^{1/2}$. Because $\mathrm{GL}(n) \subset \mathbb{C}^{n\times n}$ we have $\bar{d}_{\mathrm{GL}(n)} \geq \bar{d}_{\mathbb{C}^{n\times n}}$. However we can show, using the following lemmata, that for some choices of A and B the curve $\gamma(t) = (1 - t)A + tB$ stays in $\mathrm{GL}(n)$ and thus the two distances are equal.

Lemma 3.1. *For $\Sigma, T \in \mathbb{P}(n)$ and $U = T\Sigma(\Sigma T^2\Sigma)^{-1/2}$, the unitary polar factor of $T\Sigma$, we have that $TU\Sigma^{-1} \in \mathbb{P}(n)$.*

Proof. Substituting U gives:

$$\begin{aligned} TU\Sigma^{-1} &= T^2\Sigma(\Sigma T^2\Sigma)^{-1/2}\Sigma^{-1} \\ &= \Sigma^{-1}\Sigma T^2\Sigma(\Sigma T^2\Sigma)^{-1/2}\Sigma^{-1} \\ &= \Sigma^{-1}(\Sigma T^2\Sigma)^{1/2}\Sigma^{-1}. \end{aligned}$$

It is easy to see that this is Hermitian. Positive definiteness follows from the fact that $(T\Sigma^2T)^{-1/2}$ is positive-definite. □ □

Lemma 3.2. *Let $\Sigma, T \in \mathbb{P}(n)$ and U, V such that $UV^* = T\Sigma(\Sigma T^2\Sigma)^{-1/2}$, the unitary polar factor of $T\Sigma$. Then, $\gamma(t) = (1 - t)\Sigma V + tTU$ is in $\mathrm{GL}(n)$ for $t \in [0, 1]$.*

Proof. We can write:

$$\gamma(t) = \left((1 - t)I + tTUV^*\Sigma^{-1}\right)\Sigma V.$$

By the previous lemma know that $TUV^*\Sigma^{-1}$ is positive-definite. Therefore we have that $((1 - t)I + tTUV^*\Sigma^{-1})$ is positive-definite for $t \in [0, 1]$ and thus in $\mathrm{GL}(n)$. Since $\mathrm{GL}(n)$ is closed under multiplication we have that $\gamma(t) \in \mathrm{GL}(n)$. □ □

Now we are in position to study the quotient distance (3.6) for π. We have that $(\mathcal{M}, g) = (\mathrm{GL}(n), \bar{\mathfrak{g}})$ and $\mathcal{N} = \mathbb{P}(n)$. Plugging this in gives the following distance function:

$$\begin{aligned} d'_{\mathbb{P}(n)}(\Sigma_1, \Sigma_2) &= \inf\{\bar{d}_{\mathrm{GL}(n)}(M_1, M_2) : M_i \in \pi^{-1}(\Sigma_i)\} \\ &= \inf_{U,V\in\mathrm{U}(n)} \bar{d}_{\mathrm{GL}(n)}(\Sigma_1^{1/2}V, \Sigma_2^{1/2}U). \end{aligned} \tag{3.12}$$

Proposition 3.3. *The BW distance on $\mathbb{P}(n)$ is equal to the quotient distance $d'_{\mathbb{P}(n)}$. That is,*

$$\left[tr(\Sigma_1) + \operatorname{tr}(\Sigma_2) - 2\operatorname{tr}\left(\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right)\right]^{1/2} = \inf_{U,V\in\mathrm{U}(n)} \bar{d}_{\mathrm{GL}(n)}(\Sigma_1^{1/2}V, \Sigma_2^{1/2}U).$$

Moreover, the infimum on the right is attained when UV^* is the unitary polar factor of $\Sigma_2^{1/2}\Sigma_1^{1/2}$, given by $\Sigma_2^{1/2}\Sigma_1^{1/2}\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{-1/2}$.

Proof. From the discussion above we know:

$$\begin{aligned} d'^2_{\mathbb{P}(n)}(\Sigma_1,\Sigma_2) &= \inf_{U,V\in\mathrm{U}(n)} \bar{d}^2_{\mathrm{GL}(n)}(\Sigma_1^{1/2}V, \Sigma_2^{1/2}U) \\ &\geq \inf_{U,V\in\mathrm{U}(n)} \bar{d}^2_{\mathbb{C}^{n\times n}}(\Sigma_1^{1/2}V, \Sigma_2^{1/2}U) \qquad (3.13) \\ &= \inf_{U,V\in\mathrm{U}(n)} ||\Sigma_1^{1/2}V - \Sigma_2^{1/2}U||_2^2. \\ &= \inf_{U,V\in\mathrm{U}(n)} \operatorname{tr}\left(\left(\Sigma_1^{1/2}V - \Sigma_2^{1/2}U\right)\left(\Sigma_1^{1/2}V - \Sigma_2^{1/2}U\right)^*\right) \\ &= \operatorname{tr}(\Sigma_1) + \operatorname{tr}(\Sigma_2) - 2\sup_{U,V\in\mathrm{U}(n)} \operatorname{Re}\operatorname{tr}\left(\Sigma_1^{1/2}VU^*\Sigma_2^{1/2}\right). \end{aligned}$$

We saw in Theorem 3.1 of the preliminaries that the supremum on the right is obtained for UV^* as in the proposition. Moreover, by Lemma 3.2 we have that for this choice $(1-t)\Sigma_1^{1/2}V + t\Sigma_2^{1/2}U$ stays in $\mathrm{GL}(n)$ and thus we have $\bar{d}_{\mathrm{GL}(n)}(\Sigma_1^{1/2}V, \Sigma_2^{1/2}U) = \bar{d}_{\mathbb{C}^{n\times n}}(\Sigma_1^{1/2}V, \Sigma_2^{1/2}U)$. Therefore we get equality in (3.13) and conclude:

$$\begin{aligned} d'_{\mathbb{P}(n)}(\Sigma_1,\Sigma_2) &= \left[\operatorname{tr}(\Sigma_1) + \operatorname{tr}(\Sigma_2) - 2\operatorname{tr}\left(\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right)\right]^{1/2} \\ &= d^{BW}_{\mathbb{P}(n)}(\Sigma_1,\Sigma_2). \end{aligned}$$

□ □

Proof. (of Theorem 3.4) By Theorem 3.3 we know that there exists a unique metric h such that π as defined in equation (3.8) is a Riemannian submersion. In Proposition 3.2 we saw that this metric is given by $\mathfrak{g}^{BW}$, as defined in (3.7) in the statement of the theorem. From Theorem 3.3 we know that the Riemannian distance for this metric is given by the quotient distance for π, (3.12). In Proposition 3.3 we saw that this distance is equal to the BW-distance on $\mathbb{P}(n)$. This is what we set out to proof. □ □

Geodesics

In order to find a geodesic between Σ_1 and Σ_2 in $\mathbb{P}(n)$, according to Theorem 3.2, we need to find a geodesic γ in $\mathrm{GL}(n)$ between points in $\pi^{-1}(\Sigma_1)$ and $\pi^{-1}(\Sigma_2)$ such that $\gamma'(0)$ is horizontal.

Theorem 3.5. *A geodesic between Σ_1 and Σ_2 in $\left(\mathbb{P}(n), \mathfrak{g}^{BW}\right)$ is given by $\pi\circ\gamma$, where*

$$\gamma(t) = (1-t)\Sigma_1^{1/2}V + t\Sigma_2^{1/2}U$$

and UV^* is again the unitary polar factor of $\Sigma_2^{1/2}\Sigma_1^{1/2}$.

Proof. It is clear that γ is a geodesic in $(\mathbb{C}^{n\times n}, \bar{\mathfrak{g}})$. We saw in Lemma 3.2 that γ stays in $\mathrm{GL}(n)$. It remains to show $\gamma'(0) \in \mathcal{H}(\pi, \Sigma_1^{1/2}, \bar{\mathfrak{g}})$. We have:

$$\begin{aligned}\gamma'(0) &= \Sigma_2^{1/2}U - \Sigma_1^{1/2}V \\ &= \left(\Sigma_2^{1/2}UV^*\Sigma_1^{-1/2} - I\right)\Sigma_1^{1/2}V.\end{aligned}$$

By Lemma 3.1 we have that $\Sigma_2^{1/2}UV^*\Sigma_1^{-1/2}$ is Hermitian and thus the same holds for $\Sigma_2^{1/2}UV^*\Sigma_1^{-1/2} - I$. Using Proposition 3.1 we conclude $\gamma'(0)$ is horizontal. The statement of the theorem now follows from Theorem 3.2. □ □

3.5.2 Geometry on the space $\mathcal{D}(n)$

Inner product and distance function

Let us denote the unit sphere in $(\mathbb{C}^{n\times n}, \bar{\mathfrak{g}})$ by:

$$\mathbb{S}_{\mathbb{C}^{n\times n}} \coloneqq \{A \in \mathbb{C}^{n\times n} : \mathrm{tr}(AA^*) = 1\}$$

and $\mathbb{S}_{\mathrm{GL}(n)} \coloneqq \mathbb{S}_{\mathbb{C}^{n\times n}} \cap \mathrm{GL}(n)$, its restriction to $\mathrm{GL}(n)$. Note that:

$$\pi^{-1}(\mathcal{D}(n)) = \mathbb{S}_{\mathrm{GL}(n)}.$$

We now apply Theorem 3.3 to this submanifold of $\mathrm{GL}(n)$. We choose the metric g to be the restriction of $\bar{\mathfrak{g}}$ to $\mathbb{S}_{\mathrm{GL}(n)}$ and the Lie group G again $\mathrm{U}(n)$. Since both the metric and the quotient map are just restrictions of the ones in Theorem 3.4, the resulting metric on $\mathbb{S}_{\mathrm{GL}(n)}/\mathrm{U}(n) \cong \mathcal{D}(n)$ will also be the restriction of $\mathfrak{g}^{BW}$. From Theorem 3.3 we therefore know that the Riemannian distance function on $\mathcal{D}(n)$ corresponding to this restricted metric is given by the quotient distance defined in (3.6). Just as in the classical case (Section 3.3) where the Fisher distance on $\mathcal{P}_+(\Omega)$ is different from the Hellinger distance, it will turn out that the Riemannian distance on $\mathcal{D}(n)$ is different from the BW distance on $\mathbb{P}(n)$. In order to compute the distance on $\mathcal{D}(n)$, we first investigate the geometry on $\mathbb{S}_{\mathrm{GL}(n)}$.

Geodesics on a Euclidean sphere are obtained by intersecting the sphere with (hyper)planes through the origin. If M and N are two non-antipodal point on $\mathbb{S}_{\mathbb{C}^{n\times n}}$ we can obtain the unnormalised geodesic by projecting the geodesic in $\mathbb{C}^{n\times n}$ onto $\mathbb{S}_{\mathbb{C}^{n\times n}}$. More specifically, if $\gamma(t) = (1-t)M + (t)N$ is the geodesic in $\mathbb{C}^{n\times n}$, then

$$\tilde{\gamma}(t) = \frac{\gamma(t)}{||\gamma(t)||_2}$$

is the unnormalised geodesic in $\mathbb{S}_{\mathbb{C}^{n\times n}}$. Moreover, we have that $\gamma(t) \in \mathrm{GL}(n) \implies \tilde{\gamma}(t) \in \mathrm{GL}(n)$ since they are scalar multiples of each other. The Euclidean distance on $\mathbb{S}_{\mathbb{C}^{n\times n}}$ is given by:

$$\bar{d}_{\mathbb{S}_{\mathbb{C}^{n\times n}}}(M, N) = \arccos(\mathrm{Re}\,\mathrm{tr}(MN^*)).$$

Just as before, we have that in general $\bar{d}_{\mathbb{S}_{\mathbb{C}^{n\times n}}} \leq \bar{d}_{\mathbb{S}_{\mathrm{GL}(n)}}$, but when $\tilde{\gamma}$ stays in $\mathrm{GL}(n)$, we have that the two distances are equal. We are now in position to deduce the Riemannian distance function on $\mathcal{D}(n)$.

Theorem 3.6. *On $\mathcal{D}(n)$, the Riemannian distance for the Bures-Wasserstein metric is given by:*

$$d^{BW}_{\mathcal{D}(n)}(\rho_1,\rho_2) = \arccos\left(\operatorname{tr}\left(\left(\rho_2^{1/2}\rho_1\rho_2^{1/2}\right)^{1/2}\right)\right).$$

We will refer to this as the Bures-Wasserstein angle

Proof. From the definition of the quotient distance we have:

$$\begin{aligned} d^{BW}_{\mathcal{D}(n)}(\rho_1,\rho_2) &= \inf_{U,V\in \mathrm{U}(n)} \bar{d}_{\mathbb{S}_{\mathrm{GL}(n)}}\left(\rho_1^{1/2}V,\rho_2^{1/2}U\right) \\ &\geq \inf_{U,V\in \mathrm{U}(n)} \bar{d}_{\mathbb{S}_{\mathbb{C}^{n\times n}}}\left(\rho_1^{1/2}V,\rho_2^{1/2}U\right) \qquad (3.14) \\ &= \inf_{U,V\in \mathrm{U}(n)} \arccos\left(\operatorname{Re}\operatorname{tr}\left(\rho_1^{1/2}VU^*\rho_2^{1/2}\right)\right). \end{aligned}$$

As before, the infimum is attained when UV^* is the unitary polar factor of $\rho_2^{1/2}\rho_1^{1/2}$. From Lemma 3.2 and the above discussion we know that for this choice of U and V we have $(1-t)\rho_1^{1/2}V + t\rho_2^{1/2}U \in \mathrm{GL}(n)$. Therefore we have equality in (3.14) and conclude the proof. □ □

Geodesics

Analogous to Theorem 3.5, we get the following result for the geodesics in $\mathcal{D}(n)$:

Theorem 3.7. *An unnormalised geodesic between ρ_1 and ρ_2 in $(\mathcal{D}(n),\mathfrak{g}^{BW})$ is given by $\pi\circ\tilde{\gamma}$, where*

$$\tilde{\gamma}(t) = \frac{\gamma(t)}{||\gamma(t)||_2},$$

with

$$\gamma(t) = (1-t)\rho_1^{1/2}V + t\rho_2^{1/2}U$$

and UV^* is the unitary polar factor of $\rho_2^{1/2}\rho_1^{1/2}$.

Proof. Start by noting that the vertical and horizontal spaces are simply the restrictions of (3.9) and (3.10) to $T_{\rho^{1/2}}\mathbb{S}_{\mathrm{GL}(n)}$. We saw above that $\tilde{\gamma}(t)\in\mathrm{GL}(n)$ and therefore by Theorem 3.2 it is enough to show $\tilde{\gamma}'(0)$ is horizontal. The direction of $\tilde{\gamma}'(0)$ can be obtained by projecting $\gamma'(0)\in T_{\rho_1^{1/2}V}\mathrm{GL}(n)$ onto the subspace $T_{\rho_1^{1/2}V}\mathbb{S}_{\mathrm{GL}(n)}$. Since $\operatorname{span}\{\rho_1^{1/2}V\}$ is the orthogonal complement of $T_{\rho_1^{1/2}V}\mathbb{S}_{\mathrm{GL}(n)}$ within $T_{\rho_1^{1/2}V}\mathrm{GL}(n)$, this projection is given by:

$$\tilde{\gamma}'(0) \propto \gamma'(0) - \operatorname{Re}\operatorname{tr}\left(\rho_1^{1/2}V\gamma'(0)^*\right)\rho_1^{1/2}V.$$

From the proof of Theorem 3.5 it follows that $\gamma'(0)$ is horizontal in $T_{\rho_1^{1/2}V}\mathrm{GL}(n)$. Since $\rho_1^{1/2}V$, viewed as a tangent vector, is also horizontal in $T_{\rho_1^{1/2}V}\mathrm{GL}(n)$, it follows that $\tilde{\gamma}'(0)$ is horizontal in $T_{\rho_1^{1/2}V}\mathbb{S}_{\mathrm{GL}(n)}$. □ □

3.6 BW geometry and quantum information

In quantum information theory, a quantum state is represented by a trace-one positive semi-definite matrix ρ called a density operator. We will denote the set of these matrices by $\mathcal{D}_0(n)$. From the spectral theorem it follows that any $\rho \in \mathcal{D}_0(n)$ can be written as $\rho = UDU^*$ with $U \in \mathrm{U}(n)$ and D a real non-negative diagonal matrix with trace one. Since this latter matrix can interpreted as a probability distribution on $\{1, 2, ..., n\}$, $\mathcal{D}_0(n)$ can be viewed as a generelisation of the space of probability distributions. See Wilde (2013) for a more detailed account of this statement. The ρ's for which the diagonal of D is a Dirac delta function are called pure states and their set is denoted $\mathcal{D}_{\mathbf{p}}(n)$.

Fubini-Study metric

Note that a pure state $\rho \in \mathcal{D}_{\mathbf{p}}(n)$ can also be written as follows: $\rho = \phi\phi^T$ where $\phi \in \mathbb{S}_{\mathbb{C}^n} \coloneqq \{\psi \in \mathbb{C}^n : ||\psi||_2 = 1\}$. The set of pure states can be identified with the complex projective space. This space is obtained by identifying two elements in $\mathbb{S}_{\mathbb{C}^n}$ that differ a complex factor, i.e. the quotient space $\mathbb{S}_{\mathbb{C}^n}/U(1)$. The identification between the pure states and $\mathbb{S}_{\mathbb{C}^n}/U(1)$ is given explicitly as follows: $\mathcal{D}_{\mathbf{p}}(n) \ni \rho = \phi\phi^T \leftrightarrow [\phi] \in \mathbb{S}_{\mathbb{C}^n}/U(1)$. Combining the above, we get the following quotient map:

$$\mathbb{S}_{\mathbb{C}^n} \ni \phi \mapsto \phi\phi^T \in \mathcal{D}_{\mathbf{p}}(n).$$

If we equip the unit sphere with the Euclidean metric, the resulting quotient metric on the set of pure states will be the Fubini-Study metric with corresponding distance measure $d^{FS}([\phi], [\psi]) = |\langle\phi, \psi\rangle|$. See Bengtsson and Życzkowski (2017) for a more comprehensive description.

From Section 3.5 we know that the BW metric is obtained in a similar way. If we let $\mathbb{S}_{GL} = \{A \in \mathrm{GL}(n) : \mathrm{tr}(AA^*) = 1\}$, the unit sphere in $\mathrm{GL}(n)$, then $\mathcal{D}(n)$ can be identified with $\mathbb{S}_{GL}/\mathrm{U}(n)$ such that $\rho \leftrightarrow [\rho^{1/2}]$. We have shown that when we equip $\mathbb{S}_{GL}$ with the Euclidean metric, the resulting quotient metric on $\mathcal{D}(n)$ is the BW metric. This shows that the BW metric and angle can be viewed as generalizations of the Fubini-Study metric and distance for mixed states.

Uhlmann and Takatsu

In 1992 the German theoretical physicist Armin Uhlmann gave a lecture at the Symposium of mathematical physics in Toruń, Poland titled *"Density operators as an arena for geometry"* (Uhlmann, 1993). In this talk he considers density operators as reductions of elements of a larger Hilbert space $\mathcal{H}^{\mathrm{ext}}$ called the purification space. If the reduction of a vector $M \in \mathcal{H}^{\mathrm{ext}}$ is equal to a density operator ρ we call this vector a purification of ρ. This modus operandi is often referred to as: *"Going to the church of the larger Hilbert space"* (Datta, 2018). $\mathcal{H}^{\mathrm{ext}}$ is given by the space of n-dimensional operators with the Hilbert-Schmidt (Euclidean) inner product and the reduction map is given by: $M \mapsto MM^* = \rho$. Note that the fiber of ρ is given by: $\{\rho^{1/2}U : U \in \mathrm{U}(n)\}$. The fact that multiple vectors in $\mathcal{H}^{\mathrm{ext}}$ correspond to the same density operators is referred to as *gauge freedom*. Minimizing the distance between purifications of ρ_1 and ρ_2 using this freedom gives the BW distance and computing the pushforward of the

Hilbert-Schmidt inner product by the reduction map gives the BW metric.

As one can see, the approach taken by Uhlmann is similar as the argument in Section 3.5. The purification space takes the role of $\mathrm{GL}(n)$, the reduction map is the quotient map π and the gauge freedom is justified by Theorem 3.3. In 2008 an argument in a similar language to ours was given independently by Takatsu, this time for the Wasserstein distance between two mean-zero Gaussian distributions (Takatsu, 2008). Since this distance is identical to the Bures distance, results from both fields can be carried over.

Wigner-Yanase information

In Section 3.3 it was described that the Fisher metric is the pushforward metric of the Euclidean metric under the square map. That is, the following map is an isometry:

$$(\mathcal{M}_+(\Omega), \bar{\mathfrak{g}}) \ni \mu \mapsto \mu^2 \in \left(\mathcal{M}_+(\Omega), \mathfrak{g}^F\right).$$

If one would replace $\mathcal{M}_+(\Omega)$ by $\mathbb{P}(n)$, the pushforward of $\bar{\mathfrak{g}}$ will be the Wigner-Yanase metric as described in Gibilisco and Isola (2003).

We will now describe what the metric on the left should be so that its pushforward becomes the BW metric. We write $(\cdot)_{\mathcal{H}(M)} : T_M\mathrm{GL}(n) \to \mathcal{H}(\pi, M, \bar{\mathfrak{g}})$ for the orthogonal projection on the horizontal subspace of $T_M\mathrm{GL}(n)$ where π is still as defined in (3.8).

Lemma 3.3. *For $M \in \mathrm{GL}(n)$ and $A \in \mathcal{H}(\pi, M, \bar{\mathfrak{g}})$, we have:*

$$\mathcal{L}_{MM^*}(d\pi_M A) = AM^{-1}.$$

Proof. Since $A \in \mathcal{H}(\pi, M, \bar{\mathfrak{g}})$ we have that $AM^{-1} \in \mathbb{H}(n)$. We check:

$$\begin{aligned} MM^*\left(AM^{-1}\right) + \left(AM^{-1}\right)MM^* &= MM^*\left(AM^{-1}\right)^* + \left(AM^{-1}\right)MM^* \\ &= MA^* + AM^* \\ &= d\pi_M A. \end{aligned}$$

□ □

We define the following metric on $\mathbb{P}(n)$ for $H, K \in T_\Sigma\mathbb{P}(n)$:

$$\mathfrak{g}^{\mathcal{H}}_\Sigma(H, K) \coloneqq \mathrm{Re}\,\mathrm{tr}\left(H_{\mathcal{H}(\Sigma)}\left(K_{\mathcal{H}(\Sigma)}\right)^*\right).$$

Note that the projection $(\cdot)_{\mathcal{H}(\Sigma)}$ happens in the ambient space $T_\Sigma\mathrm{GL}(n)$ and therefore in general $H_{\mathcal{H}(\Sigma)}, K_{\mathcal{H}(\Sigma)} \notin T_\Sigma\mathbb{P}(n)$.

Theorem 3.8. *The pushforward metric of $\mathfrak{g}^{\mathcal{H}}$ under the square map is equal to the BW metric $\mathfrak{g}^{BW}$. That is, the following map is an isometry:*

$$\begin{aligned} \pi : \left(\mathbb{P}(n), \mathfrak{g}^{\mathcal{H}}\right) &\to \left(\mathbb{P}(n), \mathfrak{g}^{BW}\right) \\ \Sigma &\mapsto \Sigma^2. \end{aligned}$$

Proof. We have:

$$\begin{aligned}
\mathfrak{g}^{BW}_{\Sigma^2}(d\pi_\Sigma H, d\pi_\Sigma K) &= \mathfrak{g}^{BW}_{\Sigma^2}(d\pi_\Sigma H_{\mathcal{H}(\Sigma)}, d\pi_\Sigma K_{\mathcal{H}(\Sigma)}) \\
&= \frac{1}{2}\operatorname{Re}\operatorname{tr}\left(\mathcal{L}_{\Sigma^2}(d\pi_\Sigma H_{\mathcal{H}(\Sigma)}) d\pi_\Sigma K_{\mathcal{H}(\Sigma)}\right) \\
&= \frac{1}{2}\operatorname{Re}\operatorname{tr}\left(H_{\mathcal{H}(\Sigma)}\Sigma^{-1}\left(\Sigma(K_{\mathcal{H}(\Sigma)})^* + K_{\mathcal{H}(\Sigma)}\Sigma\right)\right) \\
&= \operatorname{Re}\operatorname{tr}\left(H_{\mathcal{H}(\Sigma)}(K_{\mathcal{H}(\Sigma)})^*\right) \\
&= \mathfrak{g}^{\mathcal{H}}_{\Sigma}(H, K).
\end{aligned}$$

□ □

Conclusion

In this paper we have highlighted the interplay between Fisher theory and quantum information theory. This has been done by presenting an existing derivation of the Riemannian metric corresponding to the BW distance on $\mathbb{P}(n)$ and subsequently adapting this argument so that the Riemannian distance and geodesics could be recovered for this metric on the subset $\mathcal{D}(n)$. In the last part we have compared the geometrical structure to similar structures within quantum information.

Further questions

The geometrical structure derived in this paper lives on the space of positive-definite matrices. In quantum information, the larger set of positive semi-definite matrices is studied. It would be interesting to investigate whether the argument to obtain this geometrical structure can be generalised to this set.

The Talagrand inequality in its original form (Talagrand, 1996) gives a bound on the 2-Wasserstein distance between two Gaussian distributions in terms of their relative entropy:

$$W^2(\mu, \nu) \leq D(\mu||\nu).$$

The Wasserstein distance can be written in terms of the covariance matrices of its arguments. The current exposition shows that written in this form, this distance measure appears in quantum information. A related distance measure on the space of covariance matrices is the quantum relative entropy. A further study of the relation between these two distances could result in a distance measure between two (classical) distributions in terms of the quantum relative entropy of their covariance matrices, similar to (Georgiou, 2003), and could potentially lead to a Talagrand-type inequality.

List of symbols

Ω :	Sample space
	$\{\omega_1, ..., \omega_n\}$
$\mathcal{S}(\Omega)$:	Space of signed measures
	$\{\mu : \Omega \to \mathbb{R}\}$
$\mathcal{M}_+(\Omega)$:	Space of strictly positive measures
	$\{\mu \in \mathcal{S}(\Omega) : \mu(\{\omega_i\}) > 0,\ \forall i \in \{1, ..., n\}\}$
$\mathcal{P}_+(\Omega)$	Space of strictly positive probability measures
	$\{p \in \mathcal{M}_+(\Omega) : \sum_i p(\omega_i) = 1\}$
$\mathrm{GL}(n)$:	General Linear group
	$\{M \in \mathbb{C}^{n\times n} : \det(M) \neq 0\}$
$\mathrm{U}(n)$:	Unitary matrices
	$\{U \in \mathrm{GL}(n) : U^*U = I\}$
$\mathbb{H}(n)$:	Hermitian matrices
	$\{H \in \mathbb{C}^{n\times n} : H = H^*\}$
$\mathbb{H}^{\perp}(n)$:	Skew-Hermitian matrices
	$\{K \in \mathbb{C}^{n\times n} : K = -K^*\}$
$\mathbb{P}(n)$:	Positive-definite Hermitian matrices
	$\{\Sigma \in \mathbb{H}(n) : \phi^*\Sigma\phi > 0, \forall \phi \in \mathbb{C}^n\}$
$\mathcal{D}(n)$:	Trace-one positive-definite Hermitian matrices
	$\{\rho \in \mathbb{P}(n) : \mathrm{tr}(\rho) = 1\}$
$\mathbb{S}_{\mathbb{C}^{n\times n}}$:	Unit sphere in matrix space
	$\{A \in \mathbb{C}^{n\times n} : \mathrm{tr}(AA^*) = 1\}$
$\mathbb{S}_{\mathrm{GL}(n)}$:	Unit sphere in invertible matrix space
	$\mathbb{S}_{\mathbb{C}^{n\times n}} \cap \mathrm{GL}(n)$
$\mathcal{L}_\Sigma(H)$:	Lyapunov operator
	X such that $X\Sigma + \Sigma X = H$
$d^{BW}_{\mathbb{P}(n)}(\Sigma, T)$:	Bures-Wasserstein distance
	$\left[\mathrm{tr}(\Sigma) + \mathrm{tr}(T) - 2\,\mathrm{tr}\left(\left(\Sigma^{1/2}T\Sigma^{1/2}\right)^{1/2}\right)\right]^{1/2}$
$d^{BW}_{\mathcal{D}(n)}(\rho_1, \rho_2)$	Bures-Wasserstein angle
	$\arccos\left(\mathrm{tr}\left(\left(\rho_2^{1/2}\rho_1\rho_2^{1/2}\right)^{1/2}\right)\right)$
$\mathfrak{g}^{BW}_\Sigma(H, K)$:	Bures-Wasserstein metric
	$\frac{1}{2}\,\mathrm{Re}\,\mathrm{tr}\left(\mathcal{L}_\Sigma\left(H^{(m)}\right)K^{(m)}\right)$
$L(\gamma)$:	Length of a curve
	$\int_a^b ||\gamma'(t)||_g dt$
$d_{\mathcal{M}}(p, q)$:	Riemannian distance on a manifold $\mathcal{M}$
	$\inf\{L(\gamma) \mid \gamma : [0, 1] \to \mathcal{M}, \gamma(0) = p, \gamma(1) = q\}$
$\bar{\mathfrak{g}}$:	Euclidean metric
$\bar{d}$:	Euclidean distance

Data Availability

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

References

Amari, Shun-ichi and Hiroshi Nagaoka (2007). *Methods of information geometry*. Vol. 191. American Mathematical Soc.

Ay, Nihat, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer (2017). *Information geometry*. Vol. 64. Springer.

Bengtsson, Ingemar and Karol Życzkowski (2017). *Geometry of quantum states: an introduction to quantum entanglement*. Cambridge university press.

Bhatia, Rajendra (2009). *Positive definite matrices*. Vol. 24. Princeton Series in Applied Mathematics. Princeton University Press.

Bhatia, Rajendra, Tanvi Jain, and Yongdo Lim (2019). "On the Bures–Wasserstein distance between positive definite matrices". In: *Expositiones Mathematicae* 37.2, pp. 165–191.

Bures, Donald (1969). "An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w*-algebras". In: *Transactions of the American Mathematical Society* 135, pp. 199–212.

Datta, Nilanjana (Lent 2018). *Lecture notes on Quantum Information Theory*.

Gallot, Sylvestre, Dominique Hulin, and Jacques Lafontaine (1990). *Riemannian geometry*. Vol. 2. Springer.

Georgiou, Tryphon T (2003). "Toeplitz covariance matrices and the Von Neumann relative entropy". In: *Control and Modeling of Complex Systems*. Ed. by Yutaka Yamamoto Koichi Hashimoto Yasuaki Oishi. Springer, pp. 23–29.

Gibilisco, Paolo and Tommaso Isola (2003). "Wigner-Yanase information on quantum state space: the geometric approach". In: *Journal of Mathematical Physics* 44.9, pp. 3752–3762.

Hayashi, Masahito (2017). *Quantum Information Theory*. Springer.

Helstrom, Carl W (1967). "Minimum mean-squared error of estimates in quantum statistics". In: *Physics letters A* 25.2, pp. 101–102.

Lang, Urs (2006). *Length spaces*. www.math.ethz.ch/lang/mg.pdf. Date of Access: 10-10-2021.

Lee, John M (2012). *Introduction to Smooth Manifolds*. Vol. 218. Springer.

– (2018). *Introduction to Riemannian Manifolds*. Vol. 176. Springer.

Malagò, Luigi, Luigi Montrucchio, and Giovanni Pistone (Dec. 2018). "Wasserstein Riemannian geometry of Gaussian densities". In: *Information Geometry* 1.2, pp. 137–179. ISSN: 2511-249X. DOI: 10.1007/s41884-018-0014-4. URL: https://doi.org/10.1007/s41884-018-0014-4.

Massart, Estelle and Pierre-Antoine Absil (2020). "Quotient Geometry with Simple Geodesics for the Manifold of Fixed-Rank Positive-Semidefinite Matrices". In: *SIAM Journal on Matrix Analysis and Applications* 41.1, pp. 171–198. DOI: 10.1137/18M1231389.

Massart, Estelle, Julien M Hendrickx, and P-A Absil (2019). "Curvature of the manifold of fixed-rank positive-semidefinite matrices endowed with the Bures–Wasserstein

metric". In: *International Conference on Geometric Science of Information*. Springer, pp. 739–748.

Michor, Peter W (2008). *Topics in differential geometry*. Vol. 93. American Mathematical Soc.

Morozova, Elena Aleksandrovna and Nikolai Nikolaevich Chentsov (1989). "Markov invariant geometry on state manifolds". In: *Itogi Nauki i Tekhniki. Seriya" Sovremennye Problemy Matematiki. Noveishie Dostizheniya"* 36, pp. 69–102.

Olkin, Ingram and Friedrich Pukelsheim (1982). "The distance between two random vectors with given dispersion matrices". In: *Linear Algebra and its Applications* 48, pp. 257–263.

Petz, Dénes (1996). "Monotone metrics on matrix spaces". In: *Linear algebra and its applications* 244, pp. 81–96.

Rao, C Radhakrishna (1992). "Information and the accuracy attainable in the estimation of statistical parameters". In: *Breakthroughs in statistics*. Ed. by Norman L. Johnson Samuel Kotz. Springer, pp. 235–247.

Takatsu, Asuka (2008). "On Wasserstein geometry of the space of Gaussian measures". In: *arXiv preprint arXiv:0801.2250*.

– (2011). "Wasserstein geometry of Gaussian measures". In: *Osaka J. Math.* 48.4, pp. 1005–1026. ISSN: 0030-6126. URL: `http://projecteuclid.org/euclid.ojm/1326291215`.

Talagrand, Michel (1996). "Transportation cost for Gaussian and other product measures". In: *Geometric & Functional Analysis GAFA* 6.3, pp. 587–600.

Thanwerdas, Yann and Xavier Pennec (2021). "$O(n)$-invariant Riemannian metrics on SPD matrices". In: *arXiv preprint arXiv:2109.05768*.

Uhlmann, Armin (1993). "Density operators as an arena for differential geometry". In: *Rep. Math. Phys* 33, pp. 253–263.

Wilde, Mark M (2013). *Quantum information theory*. Cambridge University Press.

Chapter 4

Invariance Properties of the Natural Gradient in Overparametrised Systems

Jesse van Oostrum, Johannes Müller, Nihat Ay

Information Geometry, Springer, 2023

Abstract. The natural gradient field is a vector field that lives on a model equipped with a distinguished Riemannian metric, e.g. the Fisher-Rao metric, and represents the direction of steepest ascent of an objective function on the model with respect to this metric. In practice, one tries to obtain the corresponding direction on the parameter space by multiplying the ordinary gradient by the inverse of the Gram matrix associated with the metric. We refer to this vector on the parameter space as the natural parameter gradient. In this paper we study when the pushforward of the natural parameter gradient is equal to the natural gradient. Furthermore we investigate the invariance properties of the natural parameter gradient. Both questions are addressed in an overparametrised setting.

4.1 Introduction

Within the field of deep learning, gradient methods have become ubiquitous tools for parameter optimisation. Standard gradient optimisation procedures use the vector of coordinate derivatives of the objective function as the update direction of the parameters. This is implicitly assuming a Euclidean geometry on the space of parameters. It can be argued that this is not always the most natural choice of geometry. Instead one can choose a more natural geometry for the problem at hand and then determine the Riemannian gradient of the objective function for this natural geometry, resulting in the so called *natural gradient*. The *natural gradient method* is the optimisation algorithm that performs discrete parameter updates in the direction of the natural gradient. This method was first proposed by Amari Amari, 1998 using the geometry induced by the Fisher-Rao metric. It is an active field of study within information

geometry Ay, 2020; Lin et al., 2021; Grosse and Martens, 2016 and has been shown extremely effective in many applications Ay et al., 2013; Van Hasselt, 2012; Várady et al., 2020. More recently, also other geometries on the model have been studied, such as the Wasserstein geometry Malagò et al., 2018; Li and Montúfar, 2018. The natural gradient is defined independently of a specific parametrisation. Although it is an open problem, there is work supporting the idea that the efficiency of learning of the method is due to this invariance Zhang et al., 2019.

In practice, the update direction of the parameters is given by the ordinary gradient multiplied by the inverse of the Gram matrix associated with the metric on the model. We will refer to this vector on the parameter space as the *natural parameter gradient*. In order to determine whether this direction is desired we have to map this vector to the model, since it is the location on the model, not on the parameter space, that determines the performance of the model. In non-overparametrised systems it can be shown that in a non-singular point on the model, the pushforward of the natural parameter gradient is equal to the natural gradient. Furthermore the natural parameter gradient can be called parametrisation invariant in this case Martens, 2020. In many practical applications of machine learning, and in particular deep learning, one deals however with overparametrised models, in which different directions on the parameter space correspond to a single direction on the model. In this case, the Gram matrix is degenerate and we use a generalised inverse to calculate the natural parameter gradient. In this paper, we will investigate whether the pushforward of the natural parameter gradient remains equal to the natural gradient in the overparametrised setting.

The Moore-Penrose (MP) inverse is the canonical choice of generalised inverse for the natural parameter gradient Bernacchia et al., 2018. The definition of the MP inverse is based on the Euclidean inner product defined on the parameter space. Using the MP inverse is therefore thought to affect the parametrisation invariance of the natural parameter gradient Ollivier, 2015, and thus potentially the performance of the natural gradient method. In this paper we propose two different notions of invariance. The first evaluates the invariance of the natural parameter gradient by examining the behaviour of its pushforward on the model. The second looks at the behaviour on the parameter space itself. Since the location and direction on the model is what matters, we argue that the former is of greater importance.

4.2 The natural gradient

Let $(\mathcal{Z}, g)$ be a Riemannian manifold, Ξ be a parameter space that we assume to be an open subset of $\mathbb{R}^d$, $\phi\colon \Xi \to \mathcal{Z}$ a smooth map (taking the role of a parametrisation[1]), $\mathcal{M} \coloneqq \phi(\Xi) \subset \mathcal{Z}$ a model[2], and $\mathcal{L}\colon \mathcal{Z} \to \mathbb{R}$ a smooth (objective) function (see Figure 4.1). We call $p \in \mathcal{M}$ *non-singular* if $\mathcal{M}$ is locally an embedded submanifold of $\mathcal{Z}$ around p and we denote the set of non-singular points with $\mathrm{Smooth}(\mathcal{M})$. A point p is

[1]Within the context of differential geometry, a parametrisation usually means a local diffeomorphism onto its image. This is also referred to as a coordinate system. In our context, we use a more general definition of parametrisation, where $\phi\colon \Xi \to \mathcal{Z}$ no longer needs to be a diffeomorphism onto its image but only smooth.

[2]Note that in the literature this is also called a parametrised model, statistical manifold, or in the context of machine learning, a neuromanifold. We choose the word model, however, to emphasize that we do not assume a manifold structure on $\mathcal{M}$.

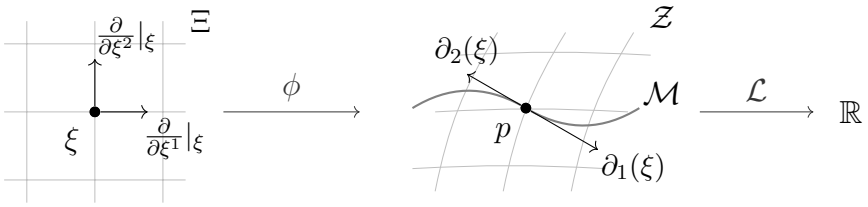


Figure 4.1: Parametrisation and objective function

called *singular* if it is not non-singular. The Riemannian gradient of $\mathcal{L}$ on $\mathcal{Z}$ is defined implicitly as follows:

$$g_p\left(\operatorname{grad}_p^{\mathcal{Z}} \mathcal{L}, \cdot\right) = d\mathcal{L}_p(\cdot).$$

By the Riesz representation theorem, this defines the gradient uniquely.

Definition 4.1 (Natural gradient)**.** For $p \in \text{Smooth}(\mathcal{M})$ the Riemannian gradient of $\mathcal{L}|_{\mathcal{M}}$ on the model $\mathcal{M}$ is called the *natural gradient* and is denoted $\operatorname{grad}_p^{\mathcal{M}} \mathcal{L}$.

It is easy to show that:

$$\operatorname{grad}_p^{\mathcal{M}} \mathcal{L} = \Pi_p(\operatorname{grad}_p^{\mathcal{Z}} \mathcal{L}),$$

where Π_p is the projection onto $T_p\mathcal{M}$. We define the pushforward of the tangent vector on the parameter space through the parametrisation as $\partial_i(\xi) \coloneqq d\phi_\xi \left(\frac{\partial}{\partial \xi^i}\Big|_\xi \right)$, and the Gram matrix $G(\xi)$ by $G_{ij}(\xi) \coloneqq g_{\phi(\xi)}\left(\partial_i(\xi), \partial_j(\xi)\right)$. We denote the vector of coordinate derivatives with $\nabla_\xi \mathcal{L} \coloneqq (\partial_1(\xi)\mathcal{L}, ..., \partial_d(\xi)\mathcal{L}) = \left(\frac{\partial \mathcal{L} \circ \phi}{\partial \xi^1}(\xi), ..., \frac{\partial \mathcal{L} \circ \phi}{\partial \xi^d}(\xi) \right) \in \mathbb{R}^d$. Let ξ be such that $\phi(\xi) \in \text{Smooth}(\mathcal{M})$. We say that a parametrisation is *proper* in ξ when: $\operatorname{span}\left(\{\partial_1(\xi), ..., \partial_d(\xi)\}\right) = T_{\phi(\xi)}\mathcal{M}$. Furthermore, following the Einstein summation convention, we write $a^i b_i$ for the sum $\sum_i a^i b_i$.

Definition 4.2 (Generalised inverse)**.** A *generalised inverse* of an $n \times m$ matrix A, denoted A^+, is an $m \times n$ matrix satisfying the following property:

$$AA^+A = A.$$

Note that this definition implies that for $w \in \mathbb{R}^n$ in the image of A, i.e. $w = Av$ for some $v \in \mathbb{R}^m$, we have:

$$AA^+w = AA^+Av = Av = w.$$

This shows that AA^+ is the identity operator on the image of A.

Definition 4.3 (Natural parameter gradient)**.** We define the *natural parameter gradient* to be the following vector on the parameter space:

$$\widetilde{\nabla}_\xi \mathcal{L} \coloneqq \left(G^+(\xi)\nabla_\xi \mathcal{L}\right)^i \left.\frac{\partial}{\partial \xi^i}\right|_\xi. \tag{4.1}$$

The pushforward of this vector, given by:

$$d\phi_\xi \widetilde{\nabla}_\xi \mathcal{L} = \left(G^+(\xi)\nabla_\xi \mathcal{L}\right)^i \partial_i,$$

is called the *natural parameter gradient on* $\mathcal{M}$.

Often, the natural parameter gradient is denoted with matrix notation as follows:

$$\widetilde{\nabla}_\xi \mathcal{L} := G^+(\xi)\nabla_\xi \mathcal{L},$$

where an identification between the canonical basis of $\mathbb{R}^d$ and the vectors $\frac{\partial}{\partial \xi^i}\Big|_\xi$ is made implicitly.

We are now in the position to state the main result of the paper:

Theorem 4.1. *Let $\xi \in \Xi$ and $p = \phi(\xi) \in \mathcal{M}$. We have:*

$$d\phi_\xi \widetilde{\nabla}_\xi \mathcal{L} = \Pi_\xi \left(\operatorname{grad}_p^{\mathcal{Z}} \mathcal{L}\right), \tag{4.2}$$

where Π_ξ is the projection onto $\operatorname{span}\{\partial_i(\xi)\}_i$. In particular, when $\phi(\xi)$ is non-singular and $\operatorname{span}\{\partial_i(\xi)\}_i = T_p\mathcal{M}$ we have:

$$d\phi_\xi \widetilde{\nabla}_\xi \mathcal{L} = \operatorname{grad}_p^{\mathcal{M}} \mathcal{L}. \tag{4.3}$$

This theorem implies that under certain conditions the pushforward of the natural parameter gradient is equal to the natural gradient. Furthermore we see that in general the natural parameter gradient on $\mathcal{M}$ is dependent on the choice of parametrisation through Π_ξ, but becomes invariant when the coordinate vectors span the full tangent space of $\mathcal{M}$. In the next section we will study in more detail the invariance properties of the natural parameter gradient.

The proof of Theorem 4.1 will be based on the following result from linear algebra:

Lemma 4.1. *Let $(V, \langle\cdot,\cdot\rangle)$ be a finite-dimensional inner product space and V^* its dual space. Let $\{e_i\}_{i\in\{1,\dots,d\}} \subset V$ (not necessarily linearly independent), G the matrix defined by $G_{ij} = \langle e_i, e_j\rangle$, $\omega \in V^*$, v such that $\langle v, \cdot\rangle = \omega(\cdot)$ and Π the projection on the space $\operatorname{span}\{e_i\}_i$. Then,*

$$\Pi(v) = \left(G^+\right)^{ij} \omega(e_j) e_i. \tag{4.4}$$

Proof. Start by noting $\Pi(v)$ is uniquely defined by the fact that $\langle \Pi(v), w\rangle = \omega(w)$, for $w \in \operatorname{span}\{e_i\}_i$ and $\langle \Pi(v), w\rangle = 0$ for $w \in (\operatorname{span}\{e_i\}_i)^\perp$. Since the RHS of (4.4) lies in the span of $\{e_i\}_i$, it remains to show that for an arbitrary vector $w = w^i e_i \in \operatorname{span}\{e_i\}_i$ we have:

$$\langle \left(G^+\right)^{ij} \omega(e_j) e_i, w^k e_k\rangle = \omega(w).$$

Working out the LHS gives:

$$\begin{aligned}
\langle \left(G^+\right)^{ij} \omega(e_j) e_i, w^k e_k\rangle &= G_{ik} \left(G^+\right)^{ij} \omega(e_j) w^k \\
&= w^k G_{ki} \left(G^+\right)^{ij} G_{jl} v^l \\
&= w^k G_{kl} v^l \\
&= w^k \omega(e_k) \\
&= \omega(w),
\end{aligned}$$

where we use the fact that: $\omega(e_i) = G_{ij} v^j$ and the symmetry of G in the second equality.

□

Proof of Theorem 4.1. We now let $T_p\mathcal{M}$ take the role of V, $d\mathcal{L}_p$ the role of ω, $\partial_i(\xi)$ the role of e_i, and $\text{grad}_p \mathcal{L}$ the role of v. Equation (4.2) now follows immediately. When the tangent vectors $\{\partial_i(\xi)\}_i$ span the whole tangent space of $\mathcal{M}$ at p, Π_ξ becomes the identity on $T_p\mathcal{M}$. This gives Equation (4.3). □

4.3 Invariance properties of the natural parameter gradient

In this section we study the invariance properties of the natural parameter gradient by using an alternative parametrisation of $\mathcal{M}$ given by:

$$\psi \colon \Theta \ni \theta \mapsto \psi(\theta) \in \mathcal{Z}.$$

Note that $G^+(\xi), \nabla_\xi \mathcal{L}$ and $\partial_i(\xi)$ in the definition of $d\phi_\xi \widetilde{\nabla}_\xi \mathcal{L}$ all implicitly depend on the parametrisation ϕ. For an alternative parametrisation ψ we will therefore write: $\partial_i(\theta) \coloneqq d\psi_\theta \left(\frac{\partial}{\partial\theta}|_\theta\right)$, $G_{ij}(\theta) \coloneqq g_{\psi(\theta)}\left(\partial_i(\theta), \partial_j(\theta)\right)$, and $\nabla_\theta \mathcal{L} \coloneqq (\partial_1(\theta)\mathcal{L}, ..., \partial_d(\theta)\mathcal{L})$.

The invariance properties can be studied from the perspective of the model and from the perspective of the parameter space itself. Since the former is of more importance, we will start with this one.

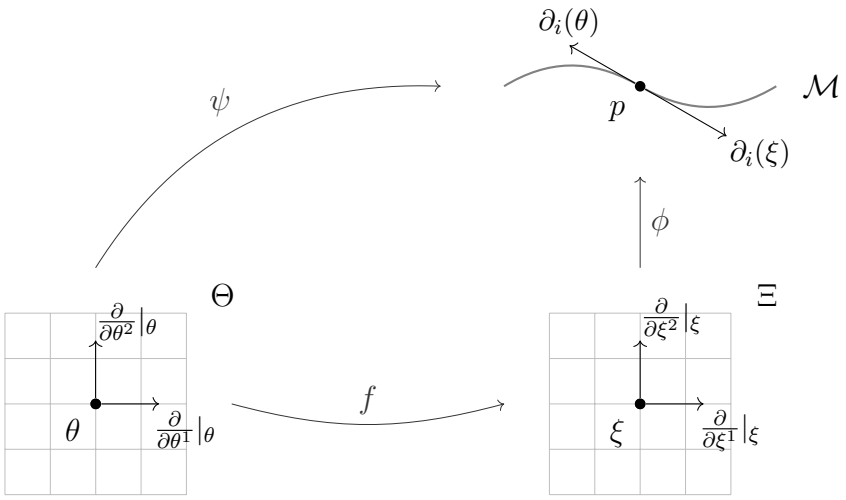


Figure 4.2: Two parametrisations of $\mathcal{M}$

4.3.1 Parametrisation dependence and reparametrisation invariance on the model

A parametrisation can be used to represent tangent vectors on the model space by elements of $\mathbb{R}^d$. A representation (of vectors on $\mathcal{M}$) can be interpreted as the map $\mathcal{O} \colon (\phi, \xi) \mapsto \mathcal{O}(\phi, \xi) \in T_\xi \Xi$ ($\cong \mathbb{R}^d$) that takes a parametrisation-coordinate pair and assigns a tangent vector on the parameter space to it. The natural parameter gradient defined by $\widetilde{\nabla}_\xi \mathcal{L} = \left(G^+(\xi)\nabla_\xi \mathcal{L}\right)^i \frac{\partial}{\partial\xi^i}\Big|_\xi$ in Equation (4.1) is an example of a representation, where the dependence on ϕ on the RHS is implicit. Naively, one could define invariance of a representation in the following way:

Definition 4.4 (Parametrisation independence)**.** Let $\mathcal{M}$ be a model. A representation $\mathcal{O}(\cdot,\cdot)$ is called *parametrisation independent* if for any pair ϕ, ψ of parametrisations of $\mathcal{M}$, and coordinates ξ, θ such that $\psi(\theta) = \phi(\xi)$, the following holds:

$$d\psi_\theta \mathcal{O}(\psi, \theta) = d\phi_\xi \mathcal{O}(\phi, \xi). \tag{4.5}$$

It turns out that this is not a very useful definition. As we will see, no non-trivial representation can be parametrisation independent in the sense of this definition. We will illustrate this in Example 1 and 2 below for the natural parameter gradient on specific models. A formal proof can be found in the Appendix 4.A.1.
In order to overcome the limitation of Definition 4, we propose the following more suitable definition of invariance of a representation:

Definition 4.5 (Reparametrisation invariance)**.** Let $\mathcal{M}$ be a model. A representation $\mathcal{O}(\cdot,\cdot)$ is called *reparametrisation invariant* if for any pair ϕ, ψ of parametrisations of $\mathcal{M}$, such that $\psi = \phi \circ f$ for a diffeomorphism $f\colon \Theta \to \Xi$, and coordinates ξ, θ such that $\theta = f^{-1}(\xi)$, the equality (4.5) holds.

Due to the extra requirement of the existence of the reparametrisation function f in Definition 4.5, we get the following central result of this paper:

Theorem 4.2. *The natural parameter gradient is reparametrisation invariant.*

Proof. By Definition 4.5, we need to show that for $\psi = \phi \circ f$ and $\theta = f^{-1}(\xi)$ we have:

$$d\psi_\theta \widetilde{\nabla}_\theta \mathcal{L} = d\phi_\xi \widetilde{\nabla}_\xi \mathcal{L}.$$

Since the differential df_θ is surjective, we have that $\operatorname{span}\{\partial_i(\xi)\}_i = \operatorname{span}\{\partial_j(\theta)\}_j$. Therefore, by using equation Equation (4.2) of Theorem 4.1, we get:

$$d\psi_\theta \widetilde{\nabla}_\theta \mathcal{L} = \Pi_\theta \left(\operatorname{grad}^{\mathcal{Z}}_{\psi(\theta)} \mathcal{L}\right) = \Pi_\xi \left(\operatorname{grad}^{\mathcal{Z}}_{\phi(\xi)} \mathcal{L}\right) = d\phi_\xi \widetilde{\nabla}_\xi \mathcal{L},$$

which is what we wanted to show. □

Remark 4.1. Note that under the extra assumptions that $\mathcal{M}$ is a smooth manifold and all parametrisations are required to be diffeomorphisms Definitions 4.4 and 4.5 are equivalent, since the composition $f = \psi \circ \phi^{-1}$ is a diffeomorphism. These assumptions are often implicitly made when referring to the invariance of the natural gradient. However, as we will see below, this is no longer the case in our more general setting.

Example 1

This example will be of a graphical nature. Consider the parametrisation ϕ that is the composition of the 2 maps in Figure 4.3. Now let $\psi\colon \Xi \to \mathcal{M}$ be this parametrisation but with a 90 degree rotation around $\phi(\xi)$ applied before projecting down to $\mathcal{M}$.

Note that the spans of the parameter vectors have trivial intersection as depicted in Figure 4.4a. This immediately implies that for any representation we have: $d\psi_\theta \mathcal{O}(\psi, \theta) \neq d\phi_\xi \mathcal{O}(\phi, \xi)$ except when both sides are equal to zero. In particular, we can let the natural gradient be as in Figure 4.4b. We know from Theorem 4.1 that the natural parameter gradients on $\mathcal{M}$ will be the projections of the natural gradient onto the

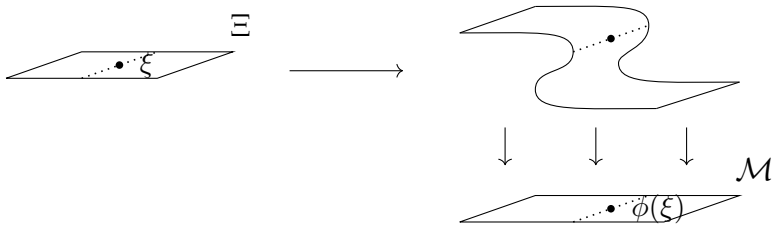
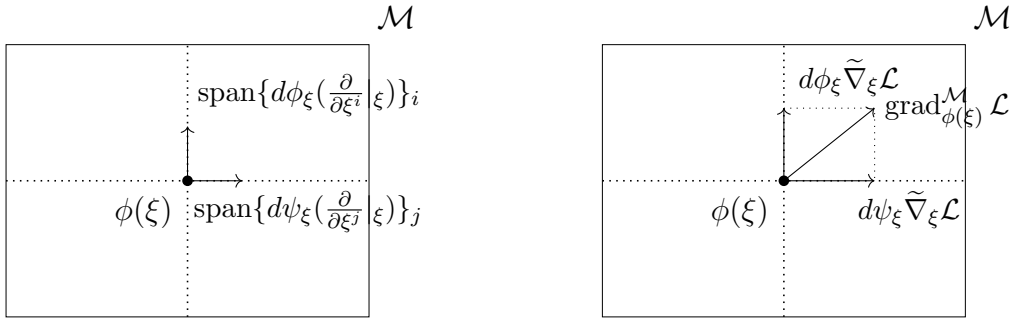


Figure 4.3: Parametrisation with a non-surjective span of the parameter vectors



(a) Spans of both sets of parameter vectors

(b) Projected gradient vectors

Figure 4.4: Spans and gradient vectors on the model of both parametrisations

respective spans of the parameter vectors as depicted in the figure. Note that the projection should be orthogonal with respect to the inner product g_p, which we have chosen here to be Euclidean for ease of illustration.

This example shows that for non-singular points, we can construct two parametrisations that give different natural parameter gradients on the same point of the model. Note that this is not in violation of Theorem 4.2 since there does not exist a diffeomorphism f such that $\psi = \phi \circ f$.

Example 2

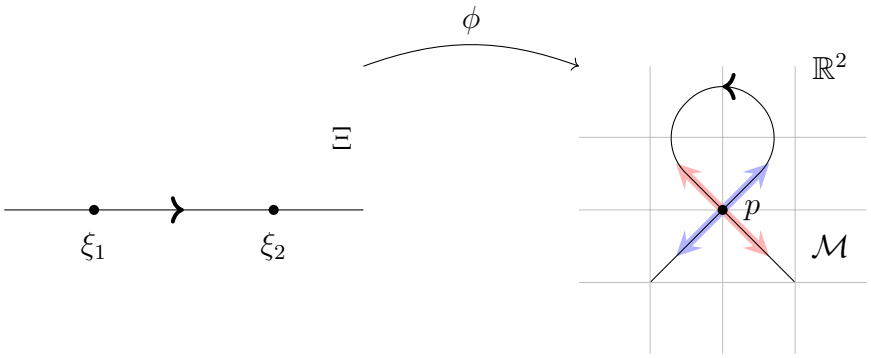


Figure 4.5: Parametrisation that contains a singular point

Let us consider the case in which ϕ is a smooth map from an interval on the real line to $\mathbb{R}^2$ as depicted in Figure 4.5. We have that ξ_1 and ξ_2 are both mapped to the same point p in $\mathbb{R}^2$. Note that $\mathcal{M}$ is in this case not a locally embedded submanifold around p and thus p is a singular point. Note that $G(\xi_1)$ is a real number different from zero

and therefore non-degenerate. Calculating the natural parameter gradient on $\mathcal{M}$ for $\xi = \xi_1$ gives:

$$\begin{aligned} d\phi_{\xi_1} \widetilde{\nabla}_{\xi_1} \mathcal{L} &= G^+(\xi_1) \nabla_\xi \mathcal{L} \; \partial(\xi_1) \\ &= G^{-1}(\xi_1) \frac{\partial \mathcal{L} \circ \phi}{\partial \xi}(\xi_1) \partial(\xi_1). \end{aligned} \tag{4.6}$$

Since $G^{-1}(\xi_1)\frac{\partial \mathcal{L} \circ \phi}{\partial \xi}(\xi_1)$ is a scalar, the resulting vector will lie in the span of $\partial(\xi_1)$ illustrated by the blue arrows in the figure.

Now let $f \colon \Theta \to \Xi$ be a diffeomorphism such that $f(\theta_1) = \xi_2$. An alternative parametrisation of $\mathcal{M}$ is given by:

$$\psi = \phi \circ f.$$

Calculating the natural parameter gradient at θ_1 for this parametrisation gives:

$$d\psi_{\theta_1} \widetilde{\nabla}_{\theta_1} \mathcal{L} = G^{-1}(\theta_1) \frac{\partial \mathcal{L} \circ \psi}{\partial \theta}(\theta_1) \partial(\theta_1).$$

Note that this vector is in the span of $\partial(\theta_1)$ denoted by the red arrows in the figure and therefore in general different from (4.6). This shows that when $\text{span}\{\partial_i(\xi)\}_i \neq \text{span}\{\partial_j(\theta)\}_j$ the outcome of $\left(G^+(\xi)\nabla_\xi \mathcal{L}\right)^i \partial_i(\xi)$ can be dependent on the choice of parametrisation and therefore the natural parameter gradient is not parametrisation independent. Note however that this result is not in contradiction with Theorem 4.2 since we do not have $\theta_1 = f^{-1}(\xi_1)$. See Appendix 4.A.2 for a worked-out example of this.

4.3.2 Reparametrisation (in)variance on the parameter space

In the previous section we have looked at the invariance properties of the natural parameter gradient from the perspective of the model. One can also study the invariance properties from the perspective of the parameter space as is done for example in Section 12 of Martens, 2020. Translating the definition of invariance given there to our notation gives the following:

Definition 4.6 (Reparametrisation invariance on the parameter space)**.** A representation $\mathcal{O}(\cdot, \cdot)$ is called *reparametrisation invariant on the parameter space* if for any pair of parametrisations ϕ, ψ such that $\psi = \phi \circ f$ for a diffeomorphism $f \colon \Theta \to \Xi$, and coordinates ξ, θ such that $\theta = f^{-1}(\xi)$, we have:

$$df_\theta \mathcal{O}(\psi, \theta) = \mathcal{O}(\phi, \xi). \tag{4.7}$$

Note that reparametrisation invariance on the parameter space implies reparametrisation invariance on the model as defined in Definition 4.5. Furthermore, it can be shown that when $\mathcal{M}$ is a smooth manifold and all parametrisations are required to be diffeomorphisms, like in Remark 4.1, this definition is equivalent to Definitions 4.4 and 4.5. In that case, the natural parameter gradient satisfies Equation (4.7). As we will see below, this is not true for general ϕ. We would like to argue, however, that this is not a suitable definition of invariance, since multiple vectors on the parameter space can

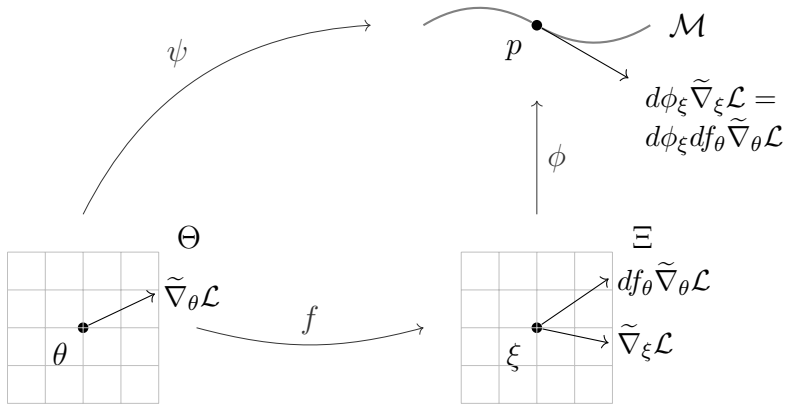


Figure 4.6: Two parametrisations of $\mathcal{M}$ with different gradient vectors on the parameter space

be mapped to the same vector on the model. Therefore inequality on the parameter space does not have to imply inequality on the model.

We will now make the above explicit. Let us choose the MP inverse as generalised inverse for $\widetilde{\nabla}_\xi \mathcal{L}$ and consider an alternative parametrisation $\psi = \phi \circ f$ for a diffeomorphism $f \colon \Theta \to \Xi$ (see Figure 4.6). We denote the matrix of partial derivatives of f at θ with $F_i^j(\theta) = \frac{\partial f^j}{\partial \theta^i}(\theta)$. For $\xi = f(\theta)$ we get the following relations:

$$\begin{aligned}
\partial_i(\theta) &= F_i^j(\theta)\, \partial_j(\xi) \\
\nabla_\theta \mathcal{L} &= F(\theta)\, \nabla_\xi \mathcal{L} \\
G(\theta) &= F(\theta)\, G(\xi)\, F^T(\theta).
\end{aligned}$$

We map $\widetilde{\nabla}_\theta \mathcal{L}$ to $T_\xi \Xi$ through df_θ and get:

$$\begin{aligned}
df_\theta \widetilde{\nabla}_\theta \mathcal{L} &= df_\theta \left(\left(G^+(\theta) \nabla_\theta \mathcal{L} \right)^i \left. \frac{\partial}{\partial \theta^i} \right|_\theta \right) \\
&= \left(\left(F(\theta) G(\xi) F^T(\theta) \right)^+ F(\theta)\, \nabla_\xi \mathcal{L} \right)^i F_i^j(\theta) \left. \frac{\partial}{\partial \xi^j} \right|_\xi \\
&= \left(F^T(\theta) \left(F(\theta) G(\xi) F^T(\theta) \right)^+ F(\theta)\, \nabla_\xi \mathcal{L} \right)^j \left. \frac{\partial}{\partial \xi^j} \right|_\xi . \qquad (4.8)
\end{aligned}$$

We will write y_Ξ, y_Θ for the coefficients of $\widetilde{\nabla}_\xi \mathcal{L}$ and $df_\theta \widetilde{\nabla}_\theta \mathcal{L}$ respectively. From Theorem 4.1 and the fact that $F(\theta)$ is of full rank we know that $F(\theta)\, \nabla_\xi \mathcal{L}$ lies in the image of $F(\theta) G(\xi) F^T(\theta)$. Therefore, by the definition of the MP inverse, we have that $\left(F(\theta) G(\xi) F^T(\theta) \right)^+ F(\theta)\, \nabla_\xi \mathcal{L} = \arg\min_x \{ ||x|| : F(\theta) G(\xi) F^T(\theta) x = F(\theta)\, \nabla_\xi \mathcal{L} \}$, where $|| \cdot ||$ is the Euclidean norm on $\mathbb{R}^d$. The coefficients in (4.8) become:

$$\begin{aligned}
y_\Theta &= F^T(\theta) \left(F(\theta) G(\xi) F^T(\theta) \right)^+ F(\theta)\, \nabla_\xi \mathcal{L} \\
&= \arg\min_y \left\{ || \left(F^T(\theta) \right)^{-1} y || : G(\xi) y = \nabla_\xi \mathcal{L} \right\}, \qquad (4.9)
\end{aligned}$$

where we substitute $y = F^T(\theta) x$ in the last line.

Remark 4.2. Note that $|| \left(F^T \right)^{-1} (\cdot) ||$ is the pushforward of the norm on Θ through f. This shows nicely the equivalence of the gradient for on the one hand constructing a

different parametrisation (ψ), and on the other hand defining a different inner product $\left(|| \left(F^T\right)^{-1} (\cdot)||\right)$ for the existing parametrisation (ϕ).

Comparing the result to the natural parameter gradient on Ξ gives:

$$\begin{aligned}\widetilde{\nabla}_\xi \mathcal{L} &= \left(G^+(\xi)\nabla_\xi \mathcal{L}\right)^i \left.\frac{\partial}{\partial \xi^i}\right|_\xi \\ &= (y_\Xi)^i \left.\frac{\partial}{\partial \xi^i}\right|_\xi \\ y_\Xi &= \arg\min_y \{||y|| : G(\xi)y = \nabla_\xi \mathcal{L}\}. \end{aligned} \tag{4.10}$$

Because the norms in (4.9) and (4.10) are different, generally $y_\Theta \neq y_\Xi$. However, both satisfy $G(\xi)y = \nabla_\xi \mathcal{L}$ and therefore $G(\xi)(y_\Theta - y_\Xi) = 0$. This implies:

$$\begin{aligned} d\phi_\xi \left(df_\theta \widetilde{\nabla}_\theta \mathcal{L} - \widetilde{\nabla}_\xi \mathcal{L}\right) &= (y_\Theta - y_\Xi)^i \partial_i(\xi) \\ &= 0, \end{aligned} \tag{4.11}$$

where the last equality can be verified by taking the norm on the RHS of (4.11) using that it is non-degenerate, like so:

$$|| (y_\Theta - y_\Xi)^i \partial_i(\xi)||_g^2 = (y_\Theta - y_\Xi)^T G(\xi) (y_\Theta - y_\Xi) = 0.$$

This shows that for overparametrised systems the natural parameter gradient is not reparametrisation invariant on the parameter space. However, as implied by Theorem 4.1, the dependency on the parametrisation disappears when the gradient is mapped to the model. See Appendix 4.A.3 for a worked-out example of the above discussion.

4.4 Practical considerations for the natural gradient method

The natural gradient method is performed by updating the current parameter vector $\xi \in \Xi$ in the direction of the vector $G^+(\xi)\nabla_\xi \mathcal{L} \in \mathbb{R}^d$. In case of a constrained parameter space, i.e. Ξ is not the full space $\mathbb{R}^d$, such as the space of covariance matrices, one runs the risk of stepping outside the parameter space, see also Arjovsky et al., 2017. This is called a constraint violation. One can use backprojection Kushner and Clark, 1978, addition of a penalty, and weight clipping Gulrajani et al., 2017 to avoid these violations. Note that for a variety of neural network applications, including many supervised learning tasks, the parameter space is unconstrained. For these models however, the generalised inverse is often hard to compute due to the high number of parameters. In this context, the Woodbury matrix identity with damping is often used instead Singh and Alistarh, 2020. Investigating these topics further falls outside the scope of this paper.

Reparametrisation (in)variance of the natural gradient method trajectory

We have shown in Theorem 4.2 that from the perspective of the model, the natural parameter gradient is reparametrisation invariant. That is, for two parametrisations

ϕ, ψ for which $\psi = \phi \circ f$ for a diffeomorphism f and $\theta = f^{-1}(\xi)$, we have that,

$$d\phi_\xi \widetilde{\nabla}_\xi \mathcal{L} = \left(G^+(\xi)\nabla_\xi \mathcal{L}\right)^i \partial_i = \left(G^+(\theta)\nabla_\theta \mathcal{L}\right)^j \partial_j = d\psi_\theta \widetilde{\nabla}_\theta \mathcal{L}. \tag{4.12}$$

The natural gradient method is performed by updating the current parameter vector $\xi \in \mathbb{R}^d$ in the direction of the vector $G^+(\xi)\nabla_\xi \mathcal{L} \in \mathbb{R}^d$. Equation (4.12) implies that if we would update the parameters for both parametrisations an infinitesimal amount, this would give us the same result on the model. We would like to emphasise however that updating the parameters by a finite amount will in general result in different locations on the model. Therefore the natural gradient method trajectory is dependent on the choice of parametrisation. This is however not an issue specific to overparametrised models but with the natural gradient method in general. See Section 12 of Martens, 2020 for exact bounds on the invariance.

Occurrence of non-proper points

We saw in the proof of Theorem 4.2 that when $\text{span}\{\partial_i(\xi)\}_i = \text{span}\{\partial_j(\theta)\}_j$ for two parametrisations ϕ and ψ with $\phi(\xi) = \psi(\theta)$, we have $d\phi_\xi \widetilde{\nabla}_\xi \mathcal{L} = d\psi_\theta \widetilde{\nabla}_\theta \mathcal{L}$. For $\phi(\xi) \in \text{Smooth}(\mathcal{M})$ note that this equality holds in particular when $\text{span}\{\partial_i(\xi)\}_i = \text{span}\{\partial_j(\theta)\}_j = T_p\mathcal{M}$, i.e. ϕ is proper in ξ. Therefore we will now study when this is the case. We start by recalling some basic facts from smooth manifold theory: Let M, N be smooth manifolds and $F : M \to N$ a smooth map. We call a point $p \in M$ a *regular point* if $dF_p : T_pM \to T_{F(p)}N$ is surjective and a *critical point* otherwise. A point $q \in N$ is called a *regular value* if all the elements in $F^{-1}(q)$ are regular points, and a *critical value* otherwise. If M is n-dimensional, we say that a subset $S \subset M$ has measure zero in M, if for every smooth chart (U, ψ) for M, the subset $\psi(S \cap U) \subset \mathbb{R}^n$ has n-dimensional measure zero. That is: $\forall \delta > 0$, there exists a countable cover of $\psi(S \cap U)$ consisting of open rectangles, the sum of whose volumes is less than δ. We have the following result based on Sard's theorem:

Proposition 4.1. *If* $\text{Smooth}(\mathcal{M})$ *is a manifold, then the image of the set of points for which* ϕ *is not proper has measure zero in* $\text{Smooth}(\mathcal{M})$.

Proof. From the definition of $\text{Smooth}(\mathcal{M})$ we know that for every $p \in \text{Smooth}(\mathcal{M})$ there exists a U_p open in $\mathcal{Z}$ such that $U_p \cap \mathcal{M}$ is an embedded submanifold of $\mathcal{Z}$. Let $U \coloneqq \bigcup_{p \in \text{Smooth}(\mathcal{M})} U_p$. Note that: $U \cap \mathcal{M} = \text{Smooth}(\mathcal{M})$ and therefore $\phi^{-1}(U) = \phi^{-1}(\text{Smooth}(\mathcal{M}))$. Since U is open in $\mathcal{Z}$, $\phi^{-1}(\text{Smooth}(\mathcal{M}))$ is an open subset of Ξ and thus an embedded submanifold. Therefore we can consider the map:

$$\phi|_{\phi^{-1}(\text{Smooth}(\mathcal{M}))} : \phi^{-1}(\text{Smooth}(\mathcal{M})) \to \text{Smooth}(\mathcal{M})$$

and note that the image of the set of points for which ϕ is not proper is equal to the set of critical values of $\phi|_{\phi^{-1}(\text{Smooth}(\mathcal{M}))}$ in $\text{Smooth}(\mathcal{M})$. A simple application of Sard's theorem gives the result. □

This proposition implies that when $\text{Smooth}(\mathcal{M})$ is a manifold, the set of points for which the pushforward of the natural parameter gradient is unequal to the natural gradient has measure zero in $\text{Smooth}(\mathcal{M})$.

4.5 Conclusion

In this paper we have studied the natural parameter gradient, which was defined as the update direction of the natural gradient method, and its pushforward to the model in an overparametrised setting. We have seen that the latter is equal to the natural gradient under certain conditions. Furthermore we have proposed different notions of invariance and studied whether the natural parameter gradient satisfies these. From the perspective of the model, we have seen that the natural parameter gradient is reparametrisation invariant but that it is not parametrisation independent. Additionally, we saw that the natural parameter gradient is not reparametrisation invariant on the parameter space. We have argued, however, that this notion is less suitable in an overparametrised setting since multiple vectors on the parameter space can correspond to the same vector on the model. Finally we have given some practical considerations for the natural gradient method.

Acknowledgements

JvO and NA acknowledge the support of the Deutsche Forschungsgemeinschaft Priority Programme "The Active Self" (SPP 2134). JM acknowledges support by the ERC under the European Union's Horizon 2020 research and innovation programme (grant agreement no 757983), by the International Max Planck Research School for Mathematics in the Sciences and the Evangelisches Studienwerk Villigst e.V..

4.A Appendix

4.A.1 On the limitation of Definition 4.4

Proposition 4.2. *No non-trivial representation of a vector can be parametrisation independent in the sense of Definition 4.4. More precisely, any representation satisfying Definition 4.4 has the property that for all parametrisation-coordinate pairs ϕ, ξ the following holds:*

$$d\phi_\xi \mathcal{O}(\phi, \xi) = 0.$$

Proof. Let $\mathcal{M}$ be a model and assume that $\mathcal{O}$ is a representation satisfying Definition 4.4. Let ϕ be a parametrisation of $\mathcal{M}$ and $\xi^* \in \Xi$ be a fixed (arbitrary) element on the domain of ϕ. Now consider the following function:

$$f \colon \mathbb{R}^d \to \mathbb{R}^d$$
$$\theta = \left(\theta^1, .., \theta^d\right) \mapsto \left(\left(\theta^1\right)^3, ..., \left(\theta^d\right)^3\right) + \xi^*.$$

We define $\Theta = f^{-1}(\Xi)$ and $\psi = \phi \circ f|_\Theta$. First note that, since f is continuous, Θ is an open set. Secondly, since f is surjective, we have $\psi(\Theta) = \mathcal{M}$. Therefore ψ is a parametrisation of $\mathcal{M}$. It is easy to see that the differential of f at $\theta = 0$, df_0, is equal to zero and therefore by the chain rule we have $d\psi_0 = d(\phi \circ f|_\Theta)_0 = d\phi_{f(0)} \circ df_0 = 0$. Furthermore we have: $\psi(0) = \phi(\xi^*)$. Therefore in order for $\mathcal{O}$ to satisfy equation (4.5) we need that:

$$d\phi^*_\xi \mathcal{O}(\phi, \xi^*) = d\psi_0 \mathcal{O}(\psi, 0) = 0.$$

Since ξ^* was chosen arbitrarily, this implies that $\mathcal{O}$ is a trivial representation. □

Remark 4.3. The function f used in the proof above is actually a homeomorphism since it is a continuous bijection and its inverse, given by $f^{-1}(\xi) = \left(\sqrt[3]{\xi^1 - (\xi^*)^1}, ..., \sqrt[3]{\xi^d - (\xi^*)^d}\right)$, is also continuous. The inverse is however not differentiable and therefore f is not a diffeomorphism, which is required for Definition 4.5.

4.A.2 Example calculation of parametrisation dependence on the model

We illustrate Example 2 in Section 4.3.1 with a specific calculation. Let us consider the following parametrisation:

$$\begin{aligned}\phi\colon (-1/8\pi, 5/8\pi) &\to (\mathbb{R}^2, \bar{g})\\ \xi &\mapsto (x, y) = (\sin(2\xi), \sin(\xi)).\end{aligned}$$

This gives $\xi_1 = 0, \xi_2 = \frac{1}{2}\pi$ in the above discussion. We get the following calculation for $d\phi_0 \widetilde{\nabla}_0 \mathcal{L}$:

$$\begin{aligned}\partial(\xi) &= d\phi_\xi \left(\frac{\partial}{\partial \xi}\bigg|_\xi \right)\\ &= 2\cos(2\xi)\, \frac{\partial}{\partial x}\bigg|_{\phi(\xi)} + \cos(\xi)\, \frac{\partial}{\partial y}\bigg|_{\phi(\xi)}\\ \partial(\xi_1) &= 2\, \frac{\partial}{\partial x}\bigg|_{(0,0)} + \frac{\partial}{\partial y}\bigg|_{(0,0)}\\ G(\xi_1) &= 2^2 + 1^1 = 5\\ d\phi_{\xi_1} \widetilde{\nabla}_{\xi_1} \mathcal{L} &= \frac{1}{5}\, \frac{\partial}{\partial \xi}\bigg|_{\xi=0} \left(\mathcal{L}(\sin(2\xi), \sin(\xi)) \right) \left(2\, \frac{\partial}{\partial x}\bigg|_{(0,0)} + \frac{\partial}{\partial y}\bigg|_{(0,0)} \right). \qquad (4.13)\end{aligned}$$

Now let:

$$\begin{aligned}f\colon (-1/8\pi, 5/8\pi) &\to (-1/8\pi, 5/8\pi)\\ \theta &\mapsto -(\theta - \frac{1}{4}\pi).\end{aligned}$$

This implies that $\theta_1 = f^{-1}(\xi_2) = 0$. We define the alternative parametrisation $\psi = \phi \circ f$. Note that we have $\psi(\theta) = (-\sin(2\theta), \sin(\theta))$ and thus $\psi(\theta_1) = \phi(\xi_1) = (0, 0)$. A similar calculation as before gives:

$$\begin{aligned}\partial(\theta_1) &= -2\, \frac{\partial}{\partial x}\bigg|_{(0,0)} + \frac{\partial}{\partial y}\bigg|_{(0,0)}\\ d\psi_{\theta_1} \widetilde{\nabla}_{\theta_1} \mathcal{L} &= \frac{1}{5}\, \frac{\partial}{\partial \theta}\bigg|_{\theta=0} \left(\mathcal{L}(-\sin(2\theta), \sin(\theta)) \right) \left(-2\, \frac{\partial}{\partial x}\bigg|_{(0,0)} + \frac{\partial}{\partial y}\bigg|_{(0,0)} \right). \qquad (4.14)\end{aligned}$$

Note that because $\partial(\xi_1) \neq \partial(\theta_1)$, (4.13) and (4.14) are not equal to each other. We can therefore conclude that the natural parameter gradient is not parametrisation independent.

4.A.3 Example calculation of reparametrisation (in)variance on the parameter space

We illustrate the discussion in Section 4.3.2 with a specific calculation. Let us consider the following setting:

$$\begin{aligned} \mathcal{Z} = \Xi = \Theta &= \mathbb{R}^2 \\ \phi(\xi_1, \xi_2) &= (\xi_1 + \xi_2, 0) \\ f(\theta_1, \theta_2) &= (2\theta_1, \theta_2) \\ \mathcal{L}(x, y) &= x^2 \end{aligned}$$

Plugging this into the expressions derived above gives:

$$\begin{aligned} \psi(\theta_1, \theta_2) &= (2\theta_1 + \theta_2, 0) \\ \partial_1(\xi) = \partial_2(\xi) &= \left.\frac{\partial}{\partial x}\right|_{\phi(\xi)} \qquad & F(\theta) &= \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \\ \partial_1(\theta) &= 2 \left.\frac{\partial}{\partial x}\right|_{\psi(\theta)} & G(\xi) &= \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \\ \partial_2(\theta) &= \left.\frac{\partial}{\partial x}\right|_{\psi(\theta)} & G(\theta) &= \begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix} \\ \nabla_\xi \mathcal{L} &= (2(\xi_1 + \xi_2), 2(\xi_1 + \xi_2)) \\ \nabla_\theta \mathcal{L} &= (4(2\theta_1 + \theta_2), 2(2\theta_1 + \theta_2)). \end{aligned}$$

Now we fix $\theta = (1, 1)$ and $\xi = f(\theta) = (2, 1)$. We start by computing y_Ξ. From the above we know that:

$$y_\Xi = \arg\min_y \{||y|| : G(\xi)y = \nabla_\xi \mathcal{L}\}.$$

It can be easily verified that this gives $y_\Xi = (3, 3)$. For y_Θ we get:

$$y_\Theta = \arg\min_y \{|| \left(F^T(\theta)\right)^{-1} y|| : G(\xi)y = \nabla_\xi \mathcal{L}\},$$

which gives: $y_\Theta = (4\frac{4}{5}, 1\frac{1}{5})$. Evidently we have $y_\Xi \neq y_\Theta$. Note however that when we map the difference of the two gradient vectors from $T_{(2,1)}\Xi$ to $T_{(3,0)}\mathcal{M}$ through $d\phi_{(2,1)}$ we get:

$$\begin{aligned} d\phi_\xi \left(df_\theta \widetilde{\nabla}_\theta \mathcal{L} - \widetilde{\nabla}_\xi \mathcal{L} \right) &= (y_\Theta - y_\Xi)^i \, \partial_i(\xi) \\ &= (4\frac{4}{5} - 3)\frac{\partial}{\partial x}|_{(3,0)} + (1\frac{1}{5} - 3)\frac{\partial}{\partial x}|_{(3,0)} \\ &= 0. \end{aligned}$$

This shows that the natural parameter gradient is in general not reparametrisation invariant on the parameter space, but that the dependency on the parametrisation disappears when the gradient is mapped to the model.

References

Amari, Shun-ichi (1998). "Natural Gradient Works Efficiently in Learning". In: *Neural Computation* 10.2, pp. 251–276.

Arjovsky, Martin, Soumith Chintala, and Léon Bottou (2017). "Wasserstein generative adversarial networks". In: *International conference on machine learning*. PMLR, pp. 214–223.

Ay, Nihat (2020). "On the locality of the natural gradient for learning in deep Bayesian networks". In: *Information Geometry*, pp. 1–49.

Ay, Nihat, Guido Montúfar, and Johannes Rauh (2013). "Selection criteria for neuromanifolds of stochastic dynamics". In: *Advances in Cognitive Neurodynamics*. Springer, pp. 147–154.

Bernacchia, Alberto, Mate Lengyel, and Guillaume Hennequin (2018). "Exact natural gradient in deep linear networks and its application to the nonlinear case". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc.

Grosse, Roger and James Martens (2016). "A Kronecker-factored approximate Fisher matrix for convolution layers". In: *International Conference on Machine Learning*. PMLR, pp. 573–582.

Gulrajani, Ishaan, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville (2017). "Improved training of Wasserstein GANs". In: *Advances in neural information processing systems* 30.

Kushner, Harold J. and Dean S. Clark (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer New York.

Li, Wuchen and Guido Montúfar (2018). "Natural gradient via optimal transport". In: *Information Geometry* 1.2, pp. 181–214.

Lin, Wu, Frank Nielsen, Mohammad Emtiyaz Khan, and Mark Schmidt (2021). "Tractable structured natural gradient descent using local parameterizations". In: *International Conference on Machine Learning*. PMLR, pp. 6680–6691.

Malagò, Luigi, Luigi Montrucchio, and Giovanni Pistone (2018). "Wasserstein Riemannian geometry of Gaussian densities". In: *Information Geometry* 1.2, pp. 137–179.

Martens, James (2020). "New Insights and Perspectives on the Natural Gradient Method". In: *Journal of Machine Learning Research* 21, pp. 1–76.

Ollivier, Yann (2015). "Riemannian metrics for neural networks I: feedforward networks". In: *Information and Inference: A Journal of the IMA* 4.2, pp. 108–153.

Singh, Sidak Pal and Dan Alistarh (2020). *WoodFisher: Efficient Second-Order Approximation for Neural Network Compression*. DOI: 10.48550/ARXIV.2004.14340. URL: https://arxiv.org/abs/2004.14340.

Van Hasselt, Hado (2012). "Reinforcement learning in continuous state and action spaces". In: *Reinforcement learning*. Springer, pp. 207–251.

Várady, Csongor, Riccardo Volpi, Luigi Malagò, and Nihat Ay (Dec. 2020). "Natural wake-sleep algorithm". In: *NeurIPS. Deep Learning through Information Geometry Worksop*.

Zhang, Guodong, James Martens, and Roger Grosse (2019). *Fast Convergence of Natural Gradient Descent for Overparameterized Neural Networks*. DOI: 10.48550/ARXIV.1905.10961. URL: https://arxiv.org/abs/1905.10961.

Chapter 5

Inversion of Bayesian Networks

Jesse van Oostrum, Peter van Hintum, Nihat Ay

International Journal of Approximate Reasoning, Elsevier, 2024

Abstract. Variational autoencoders and Helmholtz machines use a recognition network (encoder) to approximate the posterior distribution of a generative model (decoder). In this paper we establish some necessary and some sufficient properties of a recognition network so that it can model the true posterior distribution exactly. These results are derived in the general context of probabilistic graphical modelling / Bayesian networks, for which the network represents a set of conditional independence statements. We derive both global conditions, in terms of d-separation, and local conditions for the recognition network to have the desired qualities. It turns out that for the local conditions the *perfectness* property (for every node, all parents are joined) plays an important role.

5.1 Introduction

A generative model is a set of probability distributions that models the distribution of observed and latent variables. Generative models are used in many machine learning applications. One is often interested in inferring the latent variable on the basis of an observation, i.e. obtaining the posterior distribution. For complex generative models it is often hard to calculate the posterior distribution analytically. The field of variational Bayesian inference (Wainwright, Jordan, et al., 2008) studies different ways of approximating the true posterior. One approach within this field is called *amortised inference* (Gershman and Goodman, 2014). This approach distinguishes itself through using one set of parameters for recognition that is optimised over multiple data points. This can be contrasted with "memoryless" inference algorithms, such as the message passing algorithm (Pearl, 1982; Cowell et al., 1999), which finds a separate set of parameters for every data point. Both the variational autoencoder (VAE) (Kingma and Welling, 2013) and the Helmholtz machine (Dayan et al., 1995) are examples of amortised inference. In their most general form, these consist of a Bayesian network that is used to model the generative distribution. A second network, called the recognition model, is used to model the posterior distribution. Both these networks have the same set of nodes,

namely the union of the observed and latent variables. In the generative network the arrows point from the latent to the observed nodes but in the recognition network it is the other way around. The recognition network is therefore in some sense an inversion of the generative network. In many applications, one simply reverses the direction of the edges of the generative network to obtain the recognition network. However, as the simple example in Figure 5.1 shows, this does not guarantee that the recognition model is actually able to model the true posterior distribution of the generative model. In this paper, we establish some necessary and some sufficient properties of the recognition network such that we do have this guarantee. We first discuss these properties in terms of d-separation, subsequently in terms of perfectness, and finally in terms of single edge operations using the Meek conjecture (Meek, 1997).

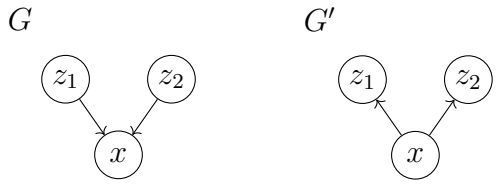


Figure 5.1: Pair of DAGs G (left) and G' (right) where G' is obtained by reversing the direction of the edges in G. The variables z_1, z_2 represent the latent variables and x the observed variable. The distribution p such that z_1, z_2 are independent Bernoulli(0.5) and $x = z_1 + z_2 \mod 2$ can be modelled by G, but the conditional distribution $p_{z_1,z_2|x}$ cannot be modelled by G'.

In practice, one often puts further restrictions on the probability distributions the networks can model. For example, the distribution of an individual node can be required to be Gaussian. The mean (and variance) are in this case a function of the values of the parent nodes. We discuss the case of a restricted set of probability distributions in the last part of the results section.

The question of finding a sparse G' that can approximate the posterior distribution of the generative model well is also studied from a more practical perspective, using methods from machine learning. One can use a sparsity prior when determining the recognition model, to encourage that only the edges really necessary for modelling the posterior are added. Löwe et al. (2022), Louizos et al. (2017), and Molchanov et al. (2019) present several approaches.

Markov equivalence is a property of a pair of Bayesian networks that indicates that they encode the same set of conditional independence statements (Verma and Pearl, 1990; Flesch and Lucas, 2007). A generalisation of this, which we will call *Markov inclusion*, is when the set of conditional independence statements encoded in one graph is a subset of the conditional independence statements encoded in the other graph (Castelo and Kočka, 2003). We will see in Proposition 5.1 that the results in this paper can also be viewed as describing under which conditions one Bayesian network is Markov inclusive of another.

Webb et al. (2018) deal with a closely related problem. They present an algorithm for inverting the generative network that gives a recognition network with the desired properties. While the present article also deals with the algorithmic aspects, it puts more emphasis on the conditions for the recognition network to have the desired properties. The authors were unaware of the publication (Webb et al., 2018) prior to the acceptance of the present paper.

5.1.1 Example

Before giving a formal definition of the problem, we illustrate the topic of this paper by an intuitive example that provides context for the rest of the paper. Consider the generative model for diseases and their symptoms in Figure 5.2. Our goal is to find a

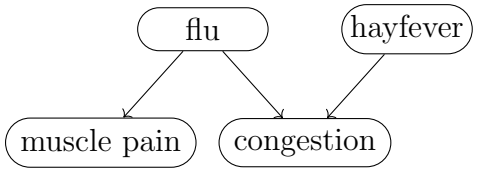


Figure 5.2: Example from: Koller and Friedman, 2009

model to perform inference on this generative model, i.e. to find the posterior distribution $P(\text{diseases} \mid \text{symptoms})$. Naively, one could reverse the edges of the generative model to obtain the recognition model in Figure 5.3.

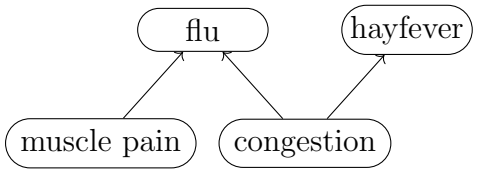


Figure 5.3: Recognition model obtained from reversing the edges of the generative model

It is clear that when someone is congested, whether or not they have muscle pain affects the likelihood of that person having hayfever. If someone is congested and also has muscle pain, the congestion is more likely to be caused by the flu. On the other hand, the absence of muscle pain would make it more likely for the congestion to be caused by hayfever. This dependence is however not captured in the graph in Figure 5.3, because no information can flow from muscle pain to hayfever. By adding an edge between muscle pain and hayfever, or between flu and hayfever, this dependence can be captured. (Figure 5.4).

5.2 Notation

5.2.1 Graph theory

For a comprehensive overview of the theory and terminology of probabilistic graphical models, we refer to (Lauritzen, 1996; Cowell et al., 1999; Studeny, 2005). We define

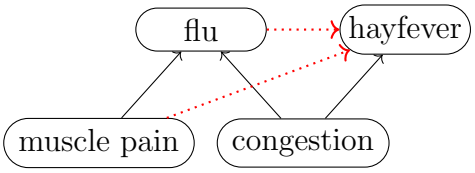


Figure 5.4: Recognition model with optional arrows in red to capture the dependence between muscle pain and hayfever

a *graph* G to be a pair $G = (N, E)$ where N is a non-empty finite set of *vertices* or *nodes* and $E \subset N \times N$ such that $(s, s) \notin E$ for all $s \in N$. A graph $H = (A, \tilde{E})$ is called a *subgraph* of G if $A \subset N$ and $\tilde{E} \subset E$ and we write $H \subset G$. For a subset $A \subset N$, the *vertex-induced subgraph* of G is denoted $G[A]$ and is given by $(A, E[A])$, with $E[A] = \{(s, t) \in E : s, t \in A\}$. When both (s, t) and (t, s) are in E, we say that there is an *undirected edge* between s and t. When $(s, t) \in E$ and $(t, s) \notin E$ we say that there is a *directed edge* going from s to t and write $s \to t$. We say that a graph is *directed* if all edges are directed.

In the following we let $G = (N, E)$ be a directed graph. We say that two vertices $s, t \in N$ are *joined* if there is an edge between the two. A set of vertices is called *complete* if all pairs of its elements are joined. A *path* in G from s to t is a sequence of distinct nodes $s = u_0, ..., u_n = t$ such that $(u_i, u_{i+1}) \in E$ for all $i \in \{0, ..., n-1\}$. A *cycle* is a path of length $n > 1$ with the modification that the end points are identical. We say that a graph is *acyclic* if it does not possess any cycles. A directed graph which is acyclic is called a *directed acyclic graph,* or DAG.

In the following we let $G = (N, E)$ be a DAG. A *trail* from s to t is a sequence of distinct nodes $s = u_0, ..., u_n = t$ such that $(u_i, u_{i+1}) \in E$ or $(u_{i+1}, u_i) \in E$ or both, for all $i \in \{0, ..., n-1\}$. Note that movement along a trail could go against the direction of the arrows, in contrast to the case of a path. A *loop* is a trail of length $n > 1$ with the modification that the end points are identical. If $(s, t) \in E$ we call s a *parent* of t and t a *child* of s. The set of parents of a node t is denoted $\mathrm{pa}_G(t)$ and the set of children of a node s is denoted $\mathrm{ch}_G(s)$. A node $t \neq s$ is called a *descendant* of s if there exists a path from s to t. The set of descendants of a node s is denoted $\mathrm{des}_G(s)$. The set of *non-descendants* of s is given by $N \setminus (\{s\} \cup \mathrm{des}_G(s))$ and is denoted by $\mathrm{nd}_G(s)$. G is called *perfect* if for all s, the set $\mathrm{pa}_G(s)$ is complete. We let $\mathrm{Leaves}(G) = \{s \in N : \mathrm{ch}_G(s) = \emptyset\}$ be the set of nodes without children, and $\mathrm{Roots}(G) = \{s \in N : \mathrm{pa}_G(s) = \emptyset\}$ be the set of nodes without parents (see Figure 5.5). For $e = (s, t) \in E$, let $e^* = (t, s)$, $E^* = \{e^* : e \in E\}$, $G^* = (N, E^*)$ the graph G with its edges reversed, $E^\sim = E \cup E^*, G^\sim = (N, E^\sim)$ the *skeleton* (i.e. undirected version) of G, and $E^\sim_{\mathrm{pa}(G)} = \{(t_1, t_2) : \exists s \in N, t_1, t_2 \in \mathrm{pa}_G(s)\}, G^{\mathrm{M}} = (N, E^\sim \cup E^\sim_{\mathrm{pa}(G)})$ the *moral graph* of G, which is the skeleton of G, with extra (undirected) edges between all parents of every vertex in G.

For a trail $\gamma = (u_0, .., u_n)$, we let $\gamma^* = (u_n, ..., u_0)$ be its reversed version, and $(\gamma, s) = (u_0, .., u_n, s)$ for $s \notin \gamma$ such that $(u_n, s) \in E$ or $(s, u_n) \in E$ be its prolongation. Now let $\gamma_1 = (u_0, .., u_n)$ and $\gamma_2 = (v_0, .., v_m)$ be trails such that $u_n = v_0$. We define $\gamma = \gamma_1; \gamma_2$

to be the *concatenation* of γ_1 and γ_2 such that $\gamma = (u_0, ..., u_i, s, v_j, ..., v_m)$, with s the first node in γ_1 which belongs to $\gamma_1 \cap \gamma_2$. Let $\gamma = (u_0, .., u_n)$ be a trail and u_i a node on this trail that is not one of the endpoints, i.e. $0 < i < n$. u_i is called a *v-structure* if $u_{i-1} \to u_i \leftarrow u_{i+1}$. γ is said to be *blocked* by $S \subset N$ if γ contains a vertex u such that either: (a) $u \in S$ and u is not a v-structure; (b) u and $\mathrm{des}_G(u)$ are not in S and u is a v-structure. Note that when one of the endpoints of the trail is in S, the trail is definitely blocked, since endpoints can never be v-structures. Let $A, B, S \subset N$ (not necessarily disjoint). A, B are said to be *d-separated* by S if all trails from A to B are blocked by S and we write $A \perp_G B \mid S$. A *topological ordering of* G is an injective map $\mathcal{O} : N \to \{1, ..., |N|\}$ that assigns to every node a number such that, if two nodes are joined, the edge points from the lower to the higher numbered node, i.e. $(s, t) \in E$ implies $\mathcal{O}(s) < \mathcal{O}(t)$. When $s, t \in N$ are such that $\mathcal{O}(s) < \mathcal{O}(t)$ we say s is *older* than t, and t is *younger* than s. Given a topological ordering $\mathcal{O}$, the set of *predecessors* of a node s, denoted $\mathrm{pr}^{\mathcal{O}}(s)$, is the set of all nodes with a lower topological number[1], i.e. $\mathrm{pr}^{\mathcal{O}}(s) = \{t \in N : \mathcal{O}(t) < \mathcal{O}(s)\}$. Note that, although $\mathrm{des}_G(s) \cap \mathrm{pr}^{\mathcal{O}}(s) = \emptyset$, the set $\mathrm{pr}^{\mathcal{O}}(s)$ in general depends on the choice of topological ordering (see Figure 5.6).

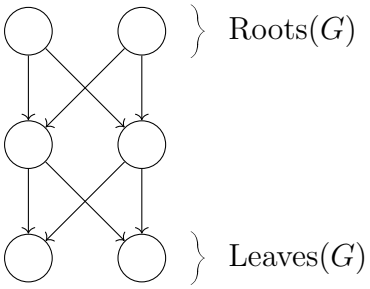


Figure 5.5: Different subsets of N for a graph G

5.2.2 Probability on graphs

Consider a DAG $G = (N, E)$. To every node $s \in N$ we associate a measurable space $(\mathsf{X}_s, \mathcal{X}_s)$. The state spaces are either real finite-dimensional vector spaces or finite sets and to each measurable space we associate a (σ-finite) base measure μ_s which is typically the Lebesgue measure or counting measure respectively. For a non-empty[2] subset $A \subset N$ we let $\mathsf{X}_A = \times_{s \in A} \mathsf{X}_s$ be the Cartesian product of the individual X_s and $\mathcal{X}_A = \otimes_{s \in A} \mathcal{X}_s$ be the product σ-algebra. We write $(\mathsf{X}, \mathcal{X}) = (\mathsf{X}_N, \mathcal{X}_N)$ and assign to this space the base measure $\mu = \otimes_{s \in N} \mu_s$, which is the product measure. In this paper, we consider probability distributions P over the space $(\mathsf{X}, \mathcal{X})$. For every $s \in N$ we let $X_s \colon \mathsf{X} \to \mathsf{X}_s$ be the random variable projecting onto the individual spaces, and similarly $X_A = (X_s)_{s \in A}$ and $X = X_N$. A typical element of X_s is denoted x_s with $x_A = (x_s)_{s \in A}$ and $x = (x_s)_{s \in N}$. We write P_A for the distribution of X_A on $(\mathsf{X}_A, \mathcal{X}_A)$, i.e. for $\mathsf{A} \in \mathcal{X}_A$,

$$P_A(\mathsf{A}) \coloneqq P\left((X_A)^{-1}(\mathsf{A})\right) = P\left(\mathsf{A} \times \mathsf{X}_{N \setminus A}\right).$$

[1]Note that some authors define the set of predecessors to be the set of nodes with a lower topological number in all possible topological orderings.

[2]When $A = \emptyset$ we let $\mathsf{X}_\emptyset = \{()\}$ be the space containing only the empty sequence and $\mathcal{X} = \{\emptyset, \mathsf{X}_\emptyset\}$ be the trivial σ-algebra. $X_\emptyset$ is simply given by the constant map $X_\emptyset \colon \mathcal{X} \ni x \mapsto () \in \mathsf{X}_\emptyset$. See Studeny, 2005 Section 2.1 for details.

For $A, C \subset N$, we say that a map $K\colon \mathcal{X}_A \times \mathsf{X}_C \to [0,1]$ is a *Markov kernel* if

$$\begin{aligned} \mathsf{A} &\mapsto K(\mathsf{A}, x_C) && \text{is a probability measure on } \mathcal{X}_A \text{ for all } x_C \in \mathsf{X}_C, \\ x_C &\mapsto K(\mathsf{A}, x_C) && \text{is } \mathcal{X}_C\text{-measurable for all } \mathsf{A} \in \mathcal{X}_A. \end{aligned}$$

Furthermore, we say that K is a *(regular) version of the conditional probability of A given C* if it is Markov kernel and for all $\mathsf{A} \in \mathcal{X}_A, \mathsf{C} \in \mathcal{X}_C$

$$P_{A\cup C}(\mathsf{A} \times \mathsf{C}) = \int_{\mathsf{C}} K(\mathsf{A}, x_C) dP_C(x_C)$$

holds. It can be shown that in our setting, one can always find such a Markov kernel that is unique P_C-a.e. (Dudley, 2018). We therefore also denote such a Markov kernel by $P_{A|C}$. For $A, B, C \subseteq N$ we say that A is *conditionally independent* of B given C and write $A \perp\!\!\!\perp B \mid C$ if for every $\mathsf{A} \in \mathcal{X}_A$ and $\mathsf{B} \in \mathcal{X}_B$

$$P_{A\cup B|C}(\mathsf{A} \times \mathsf{B} \mid x) = P_{A|C}(\mathsf{A} \mid x) \cdot P_{B|C}(\mathsf{B} \mid x) \quad P_C\text{-a.e..}$$

For $s \in N$, a *kernel function* will be a map $k^s(\cdot|\cdot)\colon \mathsf{X}_s \times \mathsf{X}_{\mathrm{pa}_G(s)} \to \mathbb{R}_{\geq 0}$ such that for all $x_{\mathrm{pa}_G(s)} \in \mathsf{X}_{\mathrm{pa}_G(s)}$

$$\int k^s\left(x_s | x_{\mathrm{pa}_G(s)}\right) d\mu_s\left(x_s\right) = 1.$$

A probability distribution P is said to *factorise* over a DAG G if it has a density p w.r.t. the product measure μ and there exist kernel functions $(k^s)_{s\in N}$ such that

$$p(x) = \prod_{s\in N} k^s\left(x_s | x_{\mathrm{pa}_G(s)}\right).$$

We denote the set of probability distributions on X that factorise over G by $\mathcal{P}^G$. Now let $A \subset N$ be such that its parents are themselves in A. A Markov kernel $K\colon \mathcal{X}_{N\setminus A} \times \mathsf{X}_A \to [0,1]$ is said to *factorise* over a DAG G if there exist kernel functions $(k^s)_{s\in N\setminus A}$ such that for every $x_A \in \mathsf{X}_A$, $K\left((\cdot), x_A\right)$ has a density $p\left((\cdot)|x_A\right)$, such that

$$p\left(x_{N\setminus A} | x_A\right) = \prod_{s\in N\setminus A} k^s\left(x_s | x_{\mathrm{pa}_G(s)}\right).$$

We denote the set of such Markov kernels by $\mathcal{K}^G$.

5.3 Problem statement

Goal I Given a DAG $G = (N, E)$, find a DAG $G' = (N, E')$ such that $\mathrm{Roots}(G') \supset \mathrm{Leaves}(G)$ and for every $P \in \mathcal{P}^G$, there exists a $K \in \mathcal{K}^{G'}$ that is a version of the conditional distribution $P_{N\setminus \mathrm{Leaves}(G)|\,\mathrm{Leaves}(G)}$.

It turns out (Proposition 5.1) that this goal is equivalent (up to edges in G' between nodes in $\mathrm{Leaves}(G)$) to the following goal:

Goal II Given a DAG $G = (N, E)$, find a DAG $G' = (N, E')$ such that $\mathcal{P}^{G'} \supset \mathcal{P}^G$ and the parents in G' of the nodes in $\mathrm{Leaves}(G)$ are themselves in $\mathrm{Leaves}(G)$.

Note that even though the parents in G' of the nodes in Leaves(G) are themselves in Leaves(G), it is not necessarily the case that Leaves(G) are oldest in every topological ordering of G'. See Figure 5.6 for an example. However, there exists a topological ordering of G' for which the nodes in Leaves(G) precede the other nodes if and only if the parents in G' of the nodes in Leaves(G) are themselves in Leaves(G).

In the remainder of the paper, we will focus on Goal II. Moreover, we sometimes impose the following extra condition:

$$G' \supset G^*.$$

It can be argued that this is a natural condition since it enforces that the hierarchical structure of the generative model G is preserved when finding a suitable G'. Furthermore, note that G^* satisfies the requirement that the parents in G^* of Leaves(G) are themselves in Leaves(G).

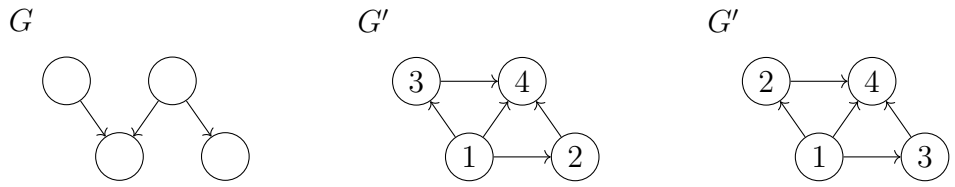


Figure 5.6: Pair of DAGs G, G' that satisfy the second requirement of Goal II, but for which there exists a topological ordering of G' (the one on the right) that does not reflect this.

5.4 Preliminaries

Lemma 5.1. *Let $G = (N, E_G), H = (N, E_H)$ be DAGs such that $E_G \subset E_H$. Then $\mathcal{P}^G \subset \mathcal{P}^H$.*

Proof. Since $\mathrm{pa}_G(s) \subset \mathrm{pa}_H(s)$ for every node s, a density that can be written as $\prod_s k^s\left(x_s|x_{\mathrm{pa}_G(s)}\right)$ can also be written as $\prod_s k^s\left(x_s|x_{\mathrm{pa}_H(s)}\right)$. □

Lemma 5.2. *(Theorem 5.14 in Cowell et al. (1999)) Let G be a DAG with a topological ordering $\mathcal{O}$. For a probability distribution P on X, the following conditions are equivalent:*

(i) $P \in \mathcal{P}^G$,

(ii) for all sets $A, B, S \subset N$ *such that* $A \perp_G B \mid S$ *we have* $A \perp\!\!\!\perp B \mid S$ *w.r.t.* P,

(iii) for all s *we have* $s \perp\!\!\!\perp \mathrm{nd}_G(s) \mid \mathrm{pa}_G(s)$ *w.r.t.* P,

(iv) for all s *we have* $s \perp\!\!\!\perp \mathrm{pr}^{\mathcal{O}}(s) \mid \mathrm{pa}_G(s)$ *w.r.t.* P.

Corollary 5.1. *Let $\mathcal{O}, \tilde{\mathcal{O}}$ be two topological orderings of G. If P satisfies property (iv) of Lemma 5.2 w.r.t. $\mathcal{O}$, then the same is true for $\tilde{\mathcal{O}}$.*

Proof. Note that $(i) - (iii)$ of Lemma 5.2 are independent of the topological ordering. Therefore we have the following implications: for all s we have $s \perp\!\!\!\perp \mathrm{pr}^{\mathcal{O}}(s) \mid \mathrm{pa}_G(s)$

w.r.t. $P \implies P \in \mathcal{P}^G$ (with topological ordering $\mathcal{O}$) $\implies P \in \mathcal{P}^G$ (with topological ordering $\tilde{\mathcal{O}}$) $\implies$ for all s we have $s \perp\!\!\!\perp \mathrm{pr}^{\tilde{\mathcal{O}}}(s) \mid \mathrm{pa}_G(s)$ w.r.t. P. □

5.5 Results

5.5.1 Equivalence of two goals

Proposition 5.1. *Let G be a DAG. The following statements hold:*

(a) Let $G' = (N, E')$ be a DAG that satisfies Goal I and let $\mathcal{O}'$ be a topological ordering of G'. Then $\tilde{G}' = (N, E' \cup E_{\text{Leaves}})$, with $E_{\text{Leaves}} = \{(s,t) : s,t \in \text{Leaves}(G), \mathcal{O}'(s) < \mathcal{O}'(t)\}$ satisfies Goal II.

(b) Let $G' = (N, E')$ be a DAG that satisfies Goal II. Then $\tilde{G}' = (N, E' \setminus E_{\text{Leaves}})$, with $E_{\text{Leaves}} = \{(s,t) \in E' : s,t \in \text{Leaves}(G)\}$ satisfies Goal I.

The proof of this proposition is based on the following lemma, in which G plays the role of G' in the proposition.

Lemma 5.3. *Let P be a distribution on X that has a density p w.r.t. μ.*

(a) Let $G = (N, E)$ be a DAG with topological ordering $\mathcal{O}$, $A \subset \text{Roots}(G)$, and $H = (N, E \cup E_A)$, with $E_A = \{(s,t) : s,t \in A, \mathcal{O}(s) < \mathcal{O}(t)\}$. If there exists a Markov kernel $K \in \mathcal{K}^G$ that is a version of the conditional distribution $P_{N\setminus A|A}$, then $P \in \mathcal{P}^H$.

(b) Let $G = (N, E)$ be a DAG, $A \subset N$ such that the parents of nodes in A are themselves in A, and $H = (N, E \setminus E_A)$, $E_A = \{(s,t) \in E : s,t \in A\}$. If $P \in \mathcal{P}^G$, then there exists a Markov kernel $K \in \mathcal{K}^H$ that is a version of the conditional distribution $P_{N\setminus A|A}$.

Proof. (a) Suppose K is a version of $P_{N\setminus A|A}$ and $K \in \mathcal{K}^G$. We need to show $P \in \mathcal{P}^H$. We can write p as follows:

$$p(x) = p\left(x_{N\setminus A}|x_A\right) p\left(x_A\right),$$

where $p\left(x_{N\setminus A}|x_A\right)$ is the density corresponding to K (Dudley, 2018). From the fact that $K \in \mathcal{K}^G$ we know

$$p\left(x_{N\setminus A}|x_A\right) = \prod_{s\in N\setminus A} k^s\left(x_s|x_{\mathrm{pa}_G(s)}\right).$$

Since all the nodes in A are joined in H we have

$$p\left(x_A\right) = \prod_{s\in A} k^s\left(x_s|x_{\mathrm{pa}_H(s)}\right).$$

Combining the above gives

$$\begin{aligned} p(x) &= \prod_{s\in N\setminus A} k^s\left(x_s|x_{\mathrm{pa}_G(s)}\right) \prod_{s\in A} k^s\left(x_s|x_{\mathrm{pa}_H(s)}\right) \\ &= \prod_{s\in N} k^s\left(x_s|x_{\mathrm{pa}_H(s)}\right), \end{aligned}$$

and therefore $P \in \mathcal{P}^H$.

(b) Now suppose $P \in \mathcal{P}^G$ and $x \in \mathsf{X}$ such that $p(x_A) > 0$. We can write

$$\begin{aligned} p(x) &= \prod_{s \in N} k^s \left(x_s | x_{\mathrm{pa}_G(s)}\right) \\ &= \prod_{s \in N \setminus A} k^s \left(x_s | x_{\mathrm{pa}_G(s)}\right) \prod_{s \in A} k^s \left(x_s | x_{\mathrm{pa}_G(s)}\right) \\ &= \prod_{s \in N \setminus A} k^s \left(x_s | x_{\mathrm{pa}_H(s)}\right) \prod_{s \in A} k^s \left(x_s | x_{\mathrm{pa}_G(s)}\right), \end{aligned}$$

where we can switch from pa_G to pa_H in the third equality because there are only edges removed between nodes in A to obtain H. From the definition of A it follows that $\mathrm{pa}_G(s) \subset A$ for all $s \in A$, hence, $\prod_{s \in A} k^s \left(x_s | x_{\mathrm{pa}_G(s)}\right) = p(x_A)$. Dividing by $p(x_A)$ on both sides gives:

$$p\left(x_{N \setminus A} | x_A\right) = \prod_{s \in N \setminus A} k^s \left(x_s | x_{\mathrm{pa}_H(s)}\right). \tag{5.1}$$

We know that there exists a Markov kernel K that is a version of the conditional distribution of $N \setminus A$ given A and that this kernel has density $p\left(x_{N \setminus A} | x_A\right)$ (Dudley, 2018). Equation (5.1) shows that the density factorises and therefore $K \in \mathcal{K}^H$. □

5.5.2 Conditions in terms of d-separation

Necessary and sufficient conditions for our goal can be deduced from the following theorem:

Theorem 5.1. *Let $G = (N, E), G' = (N, E')$ be DAGs, with $\mathcal{O}'$ a topological ordering for G'. The following statements are equivalent:*

(i) $\mathcal{P}^{G'} \supset \mathcal{P}^G$,

(ii) For all sets $A, B, S \subset N$ such that $A \perp_{G'} B \mid S$, we have $A \perp_G B \mid S$,

(iii) For all $s \in N$, we have $s \perp_G \mathrm{nd}_{G'}(s) \mid \mathrm{pa}_{G'}(s)$,

(iv) For all $s \in N$, we have $s \perp_G \mathrm{pr}^{\mathcal{O}'}(s) \mid \mathrm{pa}_{G'}(s)$.

Proof. $(i) \implies (ii)$ (by contradiction) Suppose there exist A, B, S such that $A \perp_{G'} B \mid S$, but $A \not\perp_G B \mid S$. From equation (5.7) in Cowell et al. (1999) we know that there exists a $P \in \mathcal{P}^G$ for which $A \not\perp\!\!\!\perp B \mid S$. This violates (ii) of Lemma 5.2 and therefore $P \notin \mathcal{P}^{G'}$.
$(ii) \implies (i)$ Let $P \in \mathcal{P}^G$. We need to show $P \in \mathcal{P}^{G'}$. From $(i) \implies (ii)$ in Lemma 5.2 we know that $A \perp_G B \mid S \implies A \perp\!\!\!\perp B \mid S$ for P. Combining this with the assumption $A \perp_{G'} B \mid S \implies A \perp_G B \mid S$ gives $A \perp_{G'} B \mid S \implies A \perp\!\!\!\perp B \mid S$. This means that P satisfies (ii) of Lemma 5.2 w.r.t. G' and therefore $P \in \mathcal{P}^{G'}$.
$(i) \iff (iii)$ and $(i) \iff (iv)$ can be shown in a similar way. □

5.5.3 Conditions in terms of perfectness

A sufficient condition for our goal can be deduced from the following theorem:

Theorem 5.2. *Let $G = (N, E), G' = (N, E')$ be two DAGs. If G' contains a subgraph H such that H is perfect and its undirected version $H^{\sim}$ contains the moral graph G^{M}, then $\mathcal{P}^{G'} \supset \mathcal{P}^{G}$.*

Proof. Let $P \in \mathcal{P}^{G}$. By Lemma 5.9 from Cowell et al. (1999) we know that P factorises undirectedly[3] over the undirected graph G^{M} and thus over any undirected graph $L = (N, E_L)$ containing G^{M}. From Proposition 5.15 in Cowell et al. (1999) we know that P factorises (directedly) over any perfect directed graph H such that $H^{\sim} = L$. Thus, $P \in \mathcal{P}^{H}$, and by Lemma 5.1, $P \in \mathcal{P}^{G'}$. Therefore when $H^{\sim} \supset G^{\mathrm{M}}$ we have $\mathcal{P}^{G'} \supset \mathcal{P}^{H} \supset \mathcal{P}^{G}$. □

From this theorem we can conclude that if we reverse all the edges of G, fix a topological ordering such that Leaves(G) are oldest, and then add edges consonant with this topological ordering until both G' is perfect and $G'^{\sim} \supset G^{\mathrm{M}}$, we obtain an inverse of G that satisfies our goal. The example in Figure 5.7 shows however that the condition that G' needs to contain a perfect subgraph H such that $H^{\sim} \supset G^{\mathrm{M}}$ is not a necessary condition.

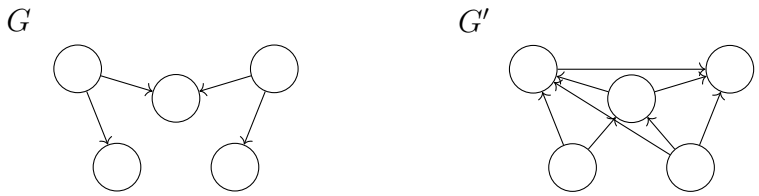


Figure 5.7: Pair of DAGs G, G' that satisfies Goal II but G' does not satisfy the condition in Theorem 5.2. One can check that $\mathcal{P}^{G'} \supset \mathcal{P}^{G}$ by verifying that Condition (iv) of Theorem 5.1 is satisfied for all nodes.

We do have the following necessary conditions on the graph G' to satisfy our goal:

Theorem 5.3. *Let G, G' be DAGs. If G' is such that $\mathcal{P}^{G'} \supset \mathcal{P}^{G}$ and $G' \supset G^{*}$, then $G'^{\sim} \supset G^{\mathrm{M}}$ and for every s in N, the vertex-induced subgraph $G'[\{s\} \cup \mathrm{des}_G(s)]$ contains a perfect subgraph H_s, such that $H_s^{\sim} \supset G^{\mathrm{M}}[\{s\} \cup \mathrm{des}_G(s)]$.*

This theorem is based on the following two propositions. We will first prove both propositions and then show how Theorem 5.3 can be obtained from it.

Proposition 5.2. *Let $\mathcal{P}^{G'} \supset \mathcal{P}^{G}$ and $G' \supset G^{*}$. Then $G'^{\sim} \supset G^{\mathrm{M}}$.*

Proof. Let $t_1, t_2 \in \mathrm{pa}_G(s)$ such that t_1 and t_2 are not joined in G and assume WLOG $\mathcal{O}'(t_2) < \mathcal{O}'(t_1)$. Assume for a contradiction that $t_2 \notin \mathrm{pa}_{G'}(t_1)$. By Condition (iv) of Theorem 5.1 we must have that $t_1 \perp_G t_2 | \mathrm{pa}_{G'}(t_1)$. However, since $G' \supset G^{*}$, we know $s \in \mathrm{pa}_{G'}(t_1)$. Furthermore, s is a v-structure on the trail t_1, s, t_2 in G. Therefore this trail is unblocked by $\mathrm{pa}_{G'}(t_1)$ and we arrive at a contradiction. Thus, necessarily $t_2 \to t_1$ in G'. □

Proposition 5.3. *If $\mathcal{P}^{G'} \supset \mathcal{P}^{G}$, $G' \supset G^{*}$, and $|\mathrm{Roots}(G)| = 1$, then G' contains a perfect subgraph H such that $H^{\sim} \supset G^{\mathrm{M}}$.*

[3]For a definition of this type of factorisation see Section 5.2 of Cowell et al. (1999)

Proof. Our approach will be to construct a subgraph H of G', by starting from G^* and iteratively adding edges. We show that every edge in H is also in G' and that the end result is both perfect and such that $H^{\sim} \supset G^{\mathrm{M}}$.

Construction of H

Let $\mathcal{O}'$ be a topological ordering of G'. We use an iterative process $H_i = (N, E_i)$ such that $H_{|N|} = H$. We start from $H_{-1} = G^*$. Then we join all parents of the same node in G according to $\mathcal{O}'$ to obtain H_0. In every subsequent step we join the parents in H_i of the $(i+1)$th youngest node according to $\mathcal{O}'$. Formally this procedure can be defined as follows:

$$\begin{aligned} E_{-1} &= E^* \\ E_0 &= E_{-1} \cup E_{\mathrm{pa}_G(G)} \\ E_{i+1} &= E_i \cup E_{\mathrm{pa}_{H_i}(r_i)}, \quad \text{for } 0 \leq i \leq |N| - 1, \end{aligned} \tag{5.2}$$

with,

$$\begin{aligned} E_{\mathrm{pa}_G(G)} &= \{(t_1, t_2) : \exists s \in N \text{ such that } t_1, t_2 \in \mathrm{pa}_G(s) \text{ and } \mathcal{O}'(t_1) < \mathcal{O}'(t_2)\} \\ E_{\mathrm{pa}_{H_i}(r_i)} &= \{(t_1, t_2) : t_1, t_2 \in \mathrm{pa}_{H_i}(r_i), \mathcal{O}'(t_1) < \mathcal{O}'(t_2)\} \\ r_i &\in N \text{ such that } \mathcal{O}'(r_i) = |N| - i. \end{aligned} \tag{5.3}$$

See Figure 5.8 for an example.

H_{-1} is a subgraph of G'

This follows directly from the fact that $G' \supset G^*$.

H_0 is a subgraph of G'

This follows from Proposition 5.2 and the fact that both H_0 and G' are consonant with $\mathcal{O}'$.

H_{i+1} is a subgraph of G' for $0 \leq i \leq |N| - 1$

Lemma 5.4. *Let $\mathcal{P}^{G'} \supset \mathcal{P}^G$, $G' \supset G^*$, $|\operatorname{Roots}(G)| = 1$ and $\mathcal{O}'$ a topological ordering for G'. Now let $r, u, v \in N$ such that $\mathcal{O}'(v) < \mathcal{O}'(u) < \mathcal{O}'(r)$ and $r \to v$ in G. Then, $v \to u$ in G'.*

Proof. Let r_0 be the youngest node in $\mathcal{O}'$ and γ_1 be a path from r_0 to u in G. Since $\mathcal{O}'(u) > \mathcal{O}'(v)$ we know $v \notin \gamma_1$. Now let γ_2 be a path in G from r_0 to r. Since $\mathcal{O}'(u) < \mathcal{O}'(r)$ we know $u \notin \gamma_2$. Now let $\gamma = \gamma_2^*; (\gamma_1, v)$ be a trail in G from u to v. Since there are no v-structures on this trail and all nodes except v are younger in $\mathcal{O}'$ than u it follows from property (iv) of Theorem 5.1 that v must be a parent of u in G'. □

Suppose $(t_2, t_1) \in E_{\mathrm{pa}_{H_i}(r_i)}$, i.e. $t_2, t_1 \in \mathrm{pa}_{H_i}(r_i)$ and $\mathcal{O}'(t_2) < \mathcal{O}'(t_1)$. We need to show that $t_2 \to t_1$ in G'. We distinguish two cases:

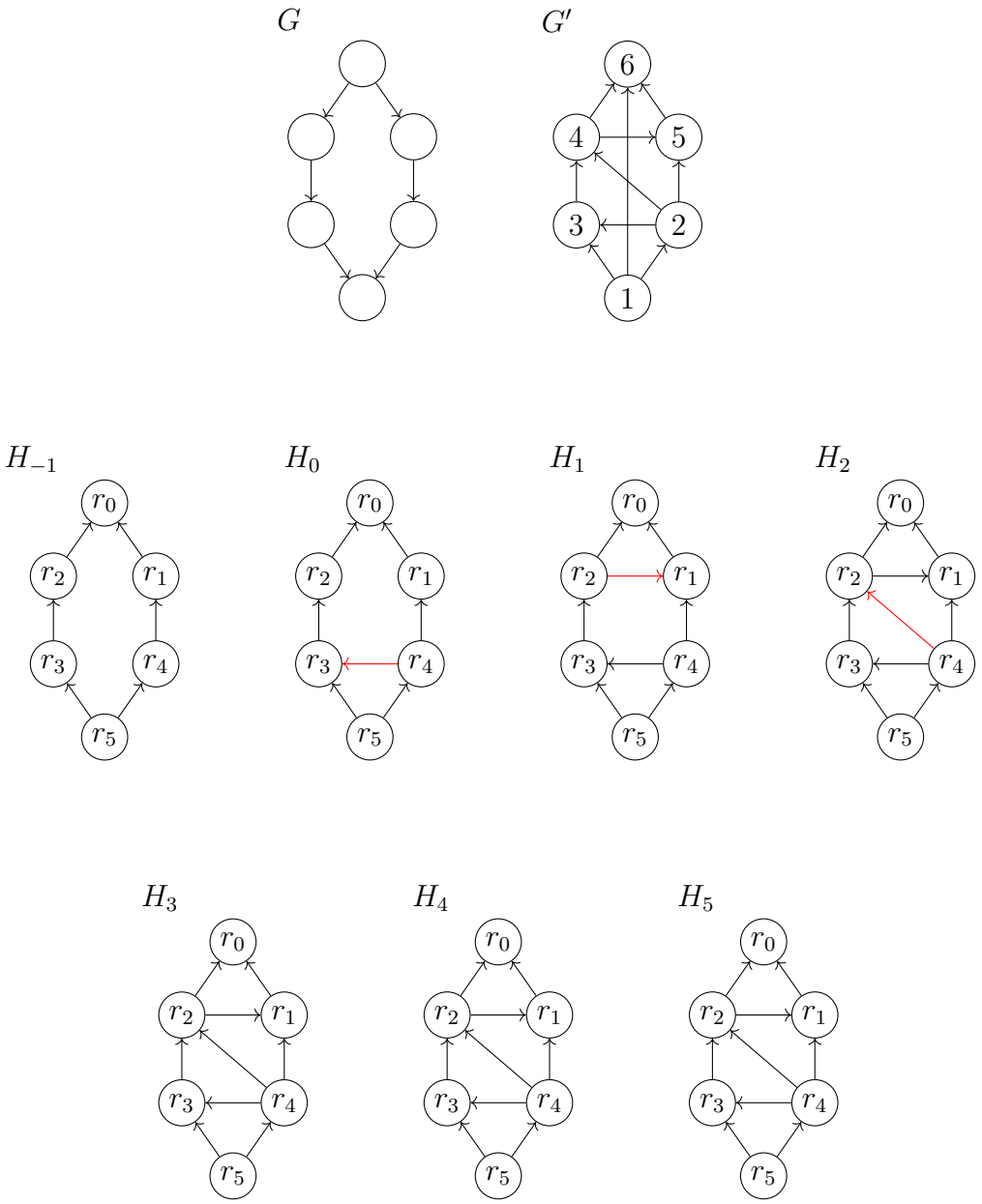


Figure 5.8: (**1st row**) Example pair of DAGs G, G' satisfying the requirements of Proposition 5.3. The fact that $\mathcal{P}^{G'} \supset \mathcal{P}^{G}$ can be checked by verifying that Condition (iv) of Theorem 5.1 is satisfied for all nodes. (**2nd and 3rd row**) Example course of the inversion procedure (5.2)–(5.3) for the pair G, G' depicted in top row. H_{-1} is the version with the edges of G reversed. In H_0 edges $E_{\mathrm{pa}_G(G)}$ for joining the parents in G are added. For $H_{i+1}, i \geq 0$ the parents in H_i of r_i are joined, by adding $E_{\mathrm{pa}_{H_i}(r_i)}$ to the edge set. Note that because the parent set of r_i in H_i for $i \geq 2$ is already complete, no new edges are added in H_3, H_4 and H_5.

Case 1: There exists a $0 \leq j \leq i$ *such that* $r_j \to t_2$ *in* G.
This follows straight from Lemma 5.4 by setting $r = r_j, u = t_1, v = t_2$.

Case 2: For all $0 \leq j \leq i$ *we have* $r_j \not\to t_2$ *in* G.
See Figure 5.9 for an illustration. Let $k = \min\{0 \leq j \leq i : (t_2, r_j) \in E_j\}$. Note that $(r_k, t_2) \in E_0$ follows from the definition of k by contradiction. The definition of H_0 implies the existence of s with $t_2 \to s \leftarrow r_k$ in G. The application of Lemma 5.4 with $r := r_k, u = t_1$ and $v = s$ gives $s \in \mathrm{pa}_{G'}(t_1)$. Assume $t_2 \not\to t_1$ in G' for a contradiction.
We let γ_1 be a path in G from r_0 to t_1, γ_2 a path in G from r_0 to r_k in G, and $\gamma = \gamma_2^*; \gamma_1$ the concatenation of the reversed γ_2 and γ_1. Note that the trail $((\gamma, s), t_2)$ is not blocked by $\mathrm{pa}_{G'}(t_1)$, since for the only v-structure (r_k, s, t_2) we have $s \in \mathrm{pa}_{G'}(t_1)$, and all other nodes on the path, except for t_2, are younger than t_1 in $\mathcal{O}'$. However, by Condition (iv) of Theorem 5.1 we know that this trail must be blocked and we arrive at a contradiction. Therefore we conclude $t_2 \to t_1$ in G'.

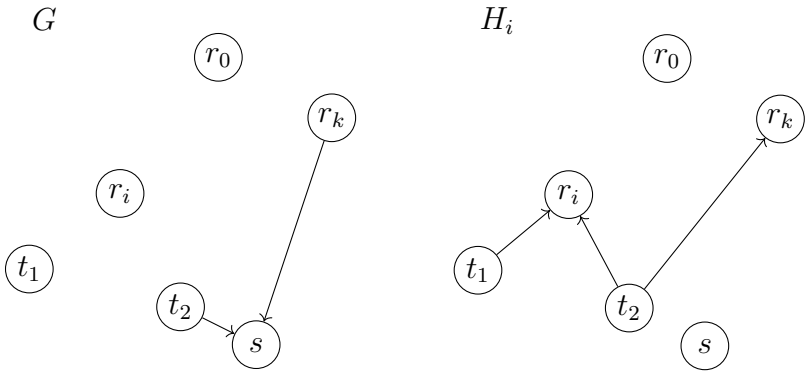


Figure 5.9: Illustration of Case 2 in the proof of Proposition 5.3

H is perfect

By adding the set $E_{\mathrm{pa}_{H_i}(r_i)}$ to the edge set, we join all the parents of r_i. After this step, no new parents of r_i are created. We perform this step for every $i \in \{0, ..., |N| - 1\}$. Since $\{r_i : i = 0, \ldots, |N| - 1\} = N$, the end result is perfect.

$H^{\sim}$ contains G^{M}

This is ensured by adding the set $E_{\mathrm{pa}_G(G)}$.

□

Remark 5.1. Note that from the proof of Proposition 5.3 it follows that the steps in the procedure described in (5.2)–(5.3) are actually necessary for obtaining a graph L that satisfies the following conditions:

(i) $\mathcal{P}^L \supset \mathcal{P}^G$,

(ii) $L \supset G^*$,

(iii) $\mathcal{O}'$ is a topological ordering for L.

This means that the end result H of the procedure will be a subset of any other graph L satisfying conditions (i)-(iii). By Theorem 5.2, H is also sufficient for satisfying these

conditions. Furthermore, any other minimal inversion of G can be obtained by using the same procedure (5.2)–(5.3) but instead of $\mathcal{O}'$ fixing a different topological ordering that is compatible with G^*.

Remark 5.2. Since any perfect graph with a single leave has a unique topological ordering[4], it follows from the proposition that any DAG G' for which there exists a DAG G such that $\mathcal{P}^{G'} \supset \mathcal{P}^G$, $G' \supset G^*$ and $\text{Roots}(G) = 1$, has a unique topological ordering as well.

Lemma 5.5. *Let G, H be DAGs and $A \subset N$. If G, H are such that $\mathcal{P}^G \subset \mathcal{P}^H$, then the same holds for the vertex-induced subgraph of both graphs: $\mathcal{P}^{G[A]} \subset \mathcal{P}^{H[A]}$.*

Proof. Let $\mathcal{O}_H$ be a topological ordering for H and $\mathcal{O}_{H[A]}$ a topological ordering for $H[A]$ that preserves the order of $\mathcal{O}_H$ and assume that G, H are such that $\mathcal{P}^G \subset \mathcal{P}^H$. By Condition (iv) of Theorem 5.1 we need to show that for all $s \in A$ we have $s \perp_{G[A]} \text{pr}^{\mathcal{O}_{H[A]}}(s) | \, \text{pa}_{H[A]}(s)$. Therefore let $s \in A$ and $t \in \text{pr}^{\mathcal{O}_{H[A]}}(s)$ for which there exists a trail γ in $G[A]$ from s to t. We need to show that this trail is blocked by $\text{pa}_{H[A]}(s)$. Since Condition (iv) of Theorem 5.1 holds for the original graphs G and H, we know that this trail must be blocked by $\text{pa}_H(s)$ in G. This implies that at least one of the following conditions must hold:

(i) There is a v-structure v on the trail such that neither v nor its descendants in G are in $\text{pa}_H(s)$.

(ii) There is a node u on the trail that is not a v-structure and is in $\text{pa}_H(s)$.

Since $\text{pa}_{H[A]}(s) \subset \text{pa}_H(s)$ and $\text{des}_{G[A]}(v) \subset \text{des}_G(v)$, the first condition remains valid. In case of the second condition, since all nodes on the trail are in A, u must also be in $\text{pa}_{H[A]}(s)$. Therefore, if the trail is blocked by $\text{pa}_H(s)$ it is also blocked by $\text{pa}_{H[A]}(s)$. □

Proof of Theorem 5.3. The first part of the theorem follows directly from Proposition 5.2. Note that by Lemma 5.5, $\mathcal{P}^{G'} \supset \mathcal{P}^G$ implies $\mathcal{P}^{G'[\{s\} \cup \text{des}_G(s)]} \supset \mathcal{P}^{G[\{s\} \cup \text{des}_G(s)]}$, for any $s \in N$. Since s is the unique root for $G[\{s\} \cup \text{des}_G(s)]$, we know from Proposition 5.3 that this implies that $G'[\{s\} \cup \text{des}_G(s)]$ contains a perfect subgraph H_s, such that $H_s \supset G^{\text{M}}[\{s\} \cup \text{des}_G(s)]$. □

In practice, the inverse G' is often obtained by simply inverting the edges in G. In this case we have the following necessary and sufficient condition to satisfy our goal.

Theorem 5.4. *Let G, G' be DAGs. If $G' = G^*$, then: $\mathcal{P}^{G'} \supset \mathcal{P}^G \iff \text{pa}_G(s), \text{ch}_G(s)$ are complete for all $s \in N$.*

Proof. ($\Longleftarrow$) If $\text{pa}_G(s), \text{ch}_G(s)$ are complete for all $s \in N$ and $G' = G^*$ this implies that $G'^{\sim} \supset G^{\text{M}}$ and G' is perfect. The result now follows from Theorem 5.2.
($\Longrightarrow$) The fact that the parents much be complete follows from Proposition 5.2. Now assume by contradiction that there exists an $s \in N$ such that $u_1, u_2 \in \text{ch}_G(s)$ are not joined. Consider the distribution $P \in \mathcal{P}^G$ such that X_{u_1} and X_{u_2} are equal to X_s

[4]This follows from the fact that every node has a unique youngest parent, and the single leave is the unique youngest node of the graph.

and all other nodes (including X_s itself) are independently distributed. Since P must factorise over G' we can write its density p as

$$p(x) = \prod_{s \in N} k^s \left(x_s | x_{\mathrm{pa}_{G'}(s)} \right).$$

Since u_1 and u_2 are both independent of their parents in G', their kernels will be simply of the form $k^{u_1}(x_{u_1}), k^{u_2}(x_{u_2})$ respectively. This however implies that u_1 and u_2 are themselves independent which contradicts the construction of P. □

Corollary 5.2. *Let G be a DAG. $\mathcal{P}^G = \mathcal{P}^{G^*}$ if and only if $\mathrm{pa}_G(s)$ and $\mathrm{ch}_G(s)$ are complete for all $s \in N$.*

Proof. This follows from twice applying Theorem 5.4 for $G = G$ and $G = G^*$ respectively and noting that completeness of $\mathrm{pa}_G(s)$ and $\mathrm{ch}_G(s)$ implies completeness of $\mathrm{pa}_{G^*}(s)$ and $\mathrm{ch}_{G^*}(s)$. □

5.5.4 Conditions in terms of single edge operations

This paper discusses results related to the inversion of Bayesian networks in the sense of Goal II. In the proof of Proposition 5.3, we suggested an algorithm for inverting G, that starts by reversing all edges of G at once and then add edges where necessary. In this section we are looking at obtaining an inverse of G by reversing the edges one by one, and potentially adding edges where necessary, based on a classical result often called Meek conjecture[5] stated below. An edge (s, t) is called *covered* when $\mathrm{pa}_G(t) = \mathrm{pa}_G(s) \cup \{s\}$. Note that reversing a covered edge preserves acyclicity of the graph.

Meek (1997) states the following conjecture:

Theorem 5.5 (Meek conjecture)**.** *Let $G = (N, E)$ and $G' = (N, E')$ be DAGs. $\mathcal{P}^{G'} \supset \mathcal{P}^G$ if and only if there exists a sequence of DAGs $H_1, ..., H_n$ such that $H_1 = G'$ and $H_n = G$ and H_{i+1} is obtained from H_i by one of the following operations:*

(a) covered edge reversal,

(b) edge removal.

This result suggests the outline of an algorithm for the inversion of a Bayesian network G. This algorithm starts with $G = L_1$. L_{i+1} is obtained from L_i by performing the inverse of operation (a) or (b), i.e. going from L_{i+1} to L_i can be done by applying operation (a) or (b). Since a covered edge remains covered after reversal, this operation is equal to its inverse. The inverse of removing an edge is simply adding an edge such that acyclicity is maintained. The goal is to reverse all edges present in G without adding parents of Leaves(G) that are not themselves in Leaves(G).

[5] It should be maybe more appropriately called Meek-Chickering Theorem, since it was proved in Chickering (2002).

5.5.5 Restricting the set of possible kernel functions

The results derived in the above discuss the question what conditions G' must satisfy such that for every $P \in \mathcal{P}^G$, $\mathcal{K}^{G'}$ contains a version of the conditional distribution $P_{N\setminus \text{Leaves}(G)|\,\text{Leaves}(G)}$. Here it is implied that we allow for all possible kernel functions k^s in the definitions of $\mathcal{P}^G$ and $\mathcal{K}^{G'}$. In practice, however, restrictions are often put on the space of possible kernel functions. A common choice (Kingma and Welling, 2013) is to allow for only Gaussian kernel functions, of the form

$$k^s\left(x_s|x_{\text{pa}(s)};\theta_s\right) \propto \exp\left(-\frac{1}{2}\left(\mu_{\theta_s}\left(x_{\text{pa}(s)}\right) - x_s\right)^2\right)$$
$$\mu_{\theta_s}\left(x_{\text{pa}(s)}\right) = f\left(\sum_{t\in\text{pa}(s)} \theta_{(s;t)}x_t\right),$$

with f some fixed possibly nonlinear function, and parametrised by the vectors $\theta_s = \{\theta_{(s,t)} \in \mathbb{R} : t \in \text{pa}(s)\}$. We will now investigate which results remain valid for the restricted case. Given a subset R of kernel functions, we will denote the restricted spaces of probability distributions and Markov kernels factorising over G by $\mathcal{P}^G_R, \mathcal{K}^G_R$ respectively. Before we dive into the results for general restrictions, we start by examining the case where R is the set of Gaussian kernel functions defined above. Consider the pair of graphs G, G' in Figure 5.10. It is clear that this pair of graphs satisfies our

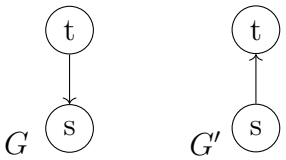


Figure 5.10: Pair of graphs G, G'

original Goal I. However, when we restrict to the set Gaussian kernel functions, we are no longer able to model the posterior distribution exactly, as we will show now. Consider the distribution in $\mathcal{P}^G_R$ given by

$$X_t \sim \mathcal{N}(0,1)$$
$$X_s \sim \mathcal{N}\left(f\left(X_t\right),1\right).$$

If the distribution $P_{t|s}$ would be in $\mathcal{K}^{G'}_R$, we would need that the joint density of X_t, X_s satisfies the following proportionality as a function of x_t

$$p(x_t, x_s) \propto \exp(-ax_t^2 + bx_t),$$

where only b may depend on x_s. Working out the actual joint density gives

$$p(x_t, x_s) \propto \exp(-\frac{1}{2}(x_t^2 + f(x_t)^2 + x_s f(x_t) + x_s^2)).$$

We can conclude that we only have that $P_{t|s} \in \mathcal{K}^{G'}_R$ if f is a linear function.[6] From this example, we can conclude that the conditions that were sufficient for the unrestricted

[6]Note that this remains true even when we loosen the restriction and allow $\mathcal{K}^{G'}_{R_{\tilde{f}}}$ to use a different function $\tilde{f}$, e.g. $\tilde{f} = f^{-1}$.

case, are in general not sufficient in the restricted case.

Now we look at the validity of the presented results under some general restrictions laid on the class of all kernel functions, i.e. R can be any subclass of kernel functions. We start with the equivalence of the two goals, Proposition 5.1. Recall that the proposition shows that finding a G' such that $\mathcal{P}^{G'} \supset \mathcal{P}^{G}$ and all parents in G' of nodes in Leaves(G) are themselves in Leaves(G), is both a necessary (statement (a)) and sufficient condition (statement (b)) to satisfy Goal I. It is easy to see that statement (b) is still valid for the restricted case, i.e.it is still a sufficient condition. However for (a) we used that when all the nodes in Roots(G) are connected, any density function can be written as $p\left(x_{\text{Roots}(G)}\right) = \prod_{s\in\text{Roots}(G)} k^s\left(x_s|x_{\text{pa}_G(s)}\right)$. This is no longer the case when we restrict the space of possible kernel functions. We have that the condition is only necessary if for every $P \in \mathcal{P}_R^G$, the marginal distribution $P_{\text{Leaves}(G)}$ factorises over a complete DAG of the leaves of G. A slightly weaker necessary condition for Goal I still holds in general, namely that for every subset $S \subset N\backslash\text{Leaves}(G)$ we need that $\mathcal{P}_R^{G'[S]} \supset \mathcal{P}_R^{G[S]}$.

For Theorem 5.1, note that conditions $(ii) - (iv)$ only relate to the graph structures of G and G'. Therefore these conditions will still be equivalent for the restricted case. The implication $(ii) \implies (i)$ does not hold in general, which was exemplified by the Gaussian kernel functions above. The implication $(i) \implies (ii)$, on the other hand, does still hold, under the extra assumption that the restriction R is such that for any graph G, for all $A, B, S \subset N$ such that $A \not\perp_G B \mid S$, there is a $P \in \mathcal{P}_R^G$ for which $A \not\perp\!\!\!\perp B \mid S$. We will sketch how this assumption is satisfied for the Gaussian kernel functions described above. Let $A, B, S \subset N$ such that $A \not\perp_G B \mid S$. This implies that there is a trail $\gamma = (u_0, ..., u_n)$ in G from $u_0 = a \in A$ to $u_n = b \in B$ that is unblocked by S. We assume that the descendants of the v-structures on this path do not intersect the trail. If this is the case we replace the part of the original trail between the v-structure and the intersecting node by the trail going via the descendants of the former v-structure and keep doing this until all these type of v-structures are removed. Now let $\theta_{(s,t)} = 1$ for all $s = u_i, t = u_j$ with $u_i, u_j \in \gamma, u_j \in \text{pa}_G(u_i), |j - i| = 1$ and all s, t subsequent descendants of a v-structure on the trail and zero otherwise. It can be shown[7] that for this distribution $a \not\perp\!\!\!\perp b \mid S$ and therefore $A \not\perp\!\!\!\perp B \mid S$. With this extra assumption we will now show $(i) \implies (ii)$. Suppose $A, B, S \subset N$ such that $A \perp_{G'} B \mid S$. This implies that for all $P \in \mathcal{P}_R^{G'}$, we have $A \perp\!\!\!\perp B \mid S$. Now suppose by contradiction that $A \not\perp_G B \mid S$. By the assumption, there must be a $P \in \mathcal{P}_R^G$ for which $A \not\perp\!\!\!\perp B \mid S$, which would contradict (a). Therefore $A \perp_G B \mid S$ which shows $(i) \implies (ii)$.

Theorem 5.2 is only a sufficient condition which is, by the Gaussian kernel function example, not sufficient any more in the restricted case. Theorem 5.3 on the other hand is only a necessary condition. The proof of this theorem only uses the necessity of the conditions in Theorem 5.1 which we showed above are still valid in the restricted case. We conclude that therefore Theorem 3 also still holds in the restricted case.

[7] As an example, one can consider a DAG consisting of a single trail $(a, ..., b)$ of binary nodes for which the nodes without parents are sampled i.i.d. (Bernoulli(.5)) and the rest of the nodes are the sum of their parents modulo 2. If we condition on the nodes that are v-structures (i.e. have parents on both sides) it is easy to see that a and b are not independent.

To conclude this section we summarise the results for the restricted case. We saw that we only have a slightly weaker necessary condition for Goal I, namely that for every subset $S \subset N \setminus \text{Leaves}(G)$ we need that $\mathcal{P}_R^{G'[S]} \supset \mathcal{P}_R^{G[S]}$. Necessary conditions for this latter condition are then provided by Theorem 5.1 and 5.3, which are still valid for the restricted case.

Conclusion

In this paper, we introduce some necessary and some sufficient conditions for the recognition network to be able to model the exact posterior distribution of a generative Bayesian network. In case that the generative network has a single root, the necessary and sufficient conditions coincide. However, for multiple roots there is still a gap between both conditions.

Further study directions

A further direction of study could be to find a single necessary and sufficient condition for the general case.
Another interesting question is the following: "What is the smallest number of edges in an inversion G' of G?". Using the results on single edge operations, one could try to find an algorithm that finds an optimal inversion of G.
It is generally believed that the recognition network needs many edges to make exact modelling of the posterior distribution possible (Welling, personal communication, 2022). Therefore, the number of edges in the recognition network will be reduced to make it computationally efficient. In practice, this approximation does not seem to affect the quality of the inference. This phenomenon remains poorly understood, but very relevant to the computational side of machine learning.

Acknowledgements

The authors would like to thank the reviewers for helpful comments. JvO would like to thank Milan Studený and Martijn Oei for helpful discussions and comments. JvO and NA acknowledge the support of the Deutsche Forschungsgemeinschaft Priority Programme "The Active Self" (SPP 2134). Lastly, JvO and PvH would like to thank Floris Triest for providing a conducive working environment.

References

Castelo, Robert and Tomás Kočka (2003). "On inclusion-driven learning of Bayesian networks". In: *Journal of Machine Learning Research* 4.Sep, pp. 527–574.

Chickering, David Maxwell (2002). "Optimal structure identification with greedy search". In: *Journal of machine learning research* 3.Nov, pp. 507–554.

Cowell, Robert G, Philip Dawid, Steffen L Lauritzen, and David J Spiegelhalter (1999). *Probabilistic Networks and Expert Systems*. Springer-Verlag New York.

Dayan, Peter, Geoffrey E Hinton, Radford M Neal, and Richard S Zemel (1995). "The Helmholtz Machine". In: *Neural computation* 7.5, pp. 889–904.

Dudley, Richard M (2018). *Real analysis and probability*. CRC Press.

Flesch, Ildikó and Peter JF Lucas (2007). "Markov equivalence in Bayesian networks". In: *Advances in probabilistic graphical models*. Springer, pp. 3–38.

Gershman, Samuel and Noah Goodman (2014). "Amortized inference in probabilistic reasoning". In: *Proceedings of the annual meeting of the cognitive science society*. Vol. 36.

Kingma, Diederik P and Max Welling (2013). "Auto-Encoding Variational Bayes". In: *arXiv preprint arXiv:1312.6114*.

Koller, Daphne and Nir Friedman (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Lauritzen, Steffen L (1996). *Graphical models*. Vol. 17. Clarendon Press.

Louizos, Christos, Max Welling, and Diederik P Kingma (2017). "Learning sparse neural networks through L_0 regularization". In: *arXiv preprint arXiv:1712.01312*.

Löwe, Sindy, David Madras, Richard Zemel, and Max Welling (2022). "Amortized causal discovery: Learning to infer causal graphs from time-series data". In: *Conference on Causal Learning and Reasoning*. PMLR, pp. 509–525.

Meek, Christopher (1997). "Graphical Models: Selecting causal and statistical models". PhD thesis. PhD thesis, Carnegie Mellon University.

Molchanov, Dmitry, Valery Kharitonov, Artem Sobolev, and Dmitry Vetrov (2019). "Doubly semi-implicit variational inference". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2593–2602.

Pearl, Judea (1982). "Reverend Bayes on inference engines: A distributed hierarchical approach". In: *Proceedings of the Second National Conference on Artificial Intelligence*, pp. 133–136.

Studeny, Milan (2005). *Probabilistic Conditional Independence Structures*. Information Science and Statistics. Springer London.

Verma, Thomas and Judea Pearl (1990). "Equivalence and synthesis of causal models". In: *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 255–270.

Wainwright, Martin J, Michael I Jordan, et al. (2008). "Graphical models, exponential families, and variational inference". In: *Foundations and Trends® in Machine Learning* 1.1–2, pp. 1–305.

Webb, Stefan, Adam Golinski, Rob Zinkov, Tom Rainforth, Yee Whye Teh, Frank Wood, et al. (2018). "Faithful inversion of generative models for effective amortized inference". In: *Advances in Neural Information Processing Systems* 31.

Chapter 6

On the Natural Gradient of the Evidence Lower Bound

Nihat Ay, Jesse van Oostrum, Adwait Datar

Journal of Machine Learning Research, 2025

Abstract. This article studies the Fisher-Rao gradient, also referred to as the natural gradient, of the evidence lower bound (ELBO) which plays a central role in generative machine learning. It reveals that the gap between the evidence and its lower bound, the ELBO, has essentially a vanishing natural gradient within unconstrained optimization. As a result, maximization of the ELBO is equivalent to minimization of the Kullback-Leibler divergence from a target distribution, the primary objective function of learning. Building on this insight, we derive a condition under which this equivalence persists even when optimization is constrained to a model. This condition yields a geometric characterization, which we formalize through the notion of a *cylindrical model*.

Keywords: Evidence lower bound, variational gap, natural gradient, information geometry, variational inference

6.1 Introduction

Generating samples from a complex target probability distribution represents the key challenge of generative machine learning. Typical examples of such a distribution are given in terms of natural images or token sequences in large language models. A primary objective function for training a generative network is based on the log-likelihood of samples, referred to as the *evidence*. In order to train a generative network, a corresponding recognition network has to be trained, with which the evidence is replaced as an objective function by the *evidence lower bound* (*ELBO*). This bound has its roots in variational methods, originally developed by Feynman (see, for example, Chapter 3, Section 3.4 of Feynman, 1972) in the context of statistical physics, where it was employed to approximate free energy. These variational methods have since been adapted for statistical inference and machine learning, proving especially useful

in the formulation of the Helmholtz Machine (Dayan et al., 1995; Ikeda et al., 1998) and other early applications (Hinton and Van Camp, 1993; Hinton and Zemel, 1993; MacKay, 1995). In more recent applications, the ELBO has become a core objective for training deep generative models, such as the Variational Autoencoder (VAE) (Kingma and Welling, 2013) and other deep generative models (Rezende et al., 2014). Beyond machine learning, the ELBO also holds a central role in cognitive science and neuroscience, underpinning the Free Energy Principle (Friston, 2005). More recently, a generalization of the ELBO called the *generalized evidence lower bound* (GLBO) has been proposed for model selection to ensure better generalization (Chen et al., 2018). A closely related idea of using a so-called Stein gradient instead of the usual gradient of the lower bound has been pursued in (Pu et al., 2017). As a generalization of the Kullback-Leibler divergence, Rényi's α-divergences (related to but different from the α-divergence in information geometry) have been studied in (Li and Turner, 2016) where a smooth interpolation between the evidence lower bound and the log (marginal) likelihood is formalized via the parameter α thereby unifying a number previously existing approaches.

Intuitively, one should expect that the gap between the evidence and its lower bound, the so-called *variational gap*, crucially affects the quality of learning. Various components of the variational gap and their influence on the learning have been studied, including the *approximation gap*, the *amortisation gap*, and the *conditioning gap* (Bayer et al., 2021). While tightening the bound appears to be beneficial at first sight, it has also been observed that a tighter bound does not necessarily imply an improvement and can even compromise the objective of learning (Rainforth et al., 2018). Aiming at an explanation of this phenomenon, our article is based on the simple idea that learning in terms of gradient methods is not so much dependent on the variational gap itself but on its gradient. While the variational gap can be rather large, its gradient might vanish (in a particular sense that we are going to specify) and therefore has no effect on the learning. We will pursue this idea with the help of information geometry (Amari and Nagaoka, 2000; Amari, 2016; Ay et al., 2017), a framework that is particularly appropriate for analyzing the evidence lower bound and the variational gap. In particular, we will study the natural gradient (Amari, 1998) of both quantities, that is the gradient with respect to the Fisher-Rao metric which we will introduce below. Our analysis will reveal a geometric criterion for the variational gap to have no effect on the learning. This core result depends crucially on the information-geometric structures and does not hold for the standard Euclidean geometry that underlies most existing gradient-based algorithms. In what follows, we briefly outline the framework of information geometry.

Originating from statistics, information geometry provides efficient methods for the field of machine learning which are based on duality concepts from differential geometry (Amari and Nagaoka, 2000; Amari, 2016; Ay et al., 2017). Most prominently, it suggests as a fundamental structure a Riemannian manifold $(\mathcal{P}, g)$, equipped with a pair (∇, ∇^*) of affine connections that are dual with respect to the Riemannian metric g. A particularly important situation is given when the two connections are flat, which implies the existence of a pair of dual affine coordinate systems and a corresponding canonical divergence $D : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$. In this case, the geometry is comparable with the Euclidean geometry of $\mathbb{R}^d$, with D being proportional to the standard squared distance function in $\mathbb{R}^d$. These structures can lead to highly efficient learning algorithms

when consistently used together. To be more precise, the distinguished canonical divergence D offers a natural way to define an objective or risk function $\mathcal{L} : \mathcal{P} \to \mathbb{R}$ for learning. When optimizing this divergence in terms of the gradient descent method, the Riemannian metric g should be applied to define the natural gradient $\text{grad}_p\mathcal{L}$ in $p \in \mathcal{P}$ via the equation

$$d\mathcal{L}_p(A) \;=\; g_p(\text{grad}_p\mathcal{L}, A) \tag{6.1}$$

for all tangent vectors A in the tangent space $T_p\mathcal{P}$. This leads to the *natural gradient method* which plays a crucial role in the theory of neural networks and machine learning (Amari, 1998; Ollivier, 2015; Martens, 2020). With these choices, the learning trajectories are then simply straight lines in the above-mentioned dual affine coordinate systems. Loosely speaking, the learning converges to a solution in the most direct way (Fujiwara and Amari, 1995; Datar and Ay, 2025). This demonstrates the simplicity and efficiency of learning as a result of a consistent combination of the underlying geometric structures.

Despite the great advantages of the outlined information-geometric approach to learning, it is a highly non-trivial task to actually utilize and implement this approach within the setting of machine learning. In what follows, we highlight two complications that are particularly relevant for this article.

1. The manifold $\mathcal{P}$ of the above paragraph plays the role of a high-dimensional ambient space, equipped with a dually flat structure g, ∇, and ∇^*. Thus, it comes with a canonical divergence for learning, as outlined above. The learning, however, is typically restricted to a lower-dimensional model $\mathcal{M} \subseteq \mathcal{P}$. The restriction of the convenient geometric structures on $\mathcal{P}$ to the model $\mathcal{M}$ is typically much more complex. Only in exceptional cases, this restriction preserves the simplicity of the geometry of $\mathcal{P}$.

2. In addition to that, we face another potential source of complication. Typically, the expressive power of a learning system has to be increased in terms of a set of latent or hidden units denoted by H. In this case, the primary model for learning is associated with the observed or visible units denoted by V. It is obtained as the image $\mathcal{M}_V$ of a model $\mathcal{M}$ under the marginalization map. Even if $\mathcal{M}$ inherits geometric properties from its ambient space $\mathcal{P}$ that are advantageous for learning, these properties need not be preserved under this marginalization.

To summarize, we face two sources of complexity when designing information-geometric learning algorithms, the restriction of natural structures from the ambient space $\mathcal{P}$ to the model $\mathcal{M}$, and the marginalization which maps $\mathcal{M}$ to the model $\mathcal{M}_V$. In this article, we aim to disentangle the individual complexities resulting from these two operations by studying the optimization processes first on $\mathcal{P}$ and then extend the analysis to the constrained setting $\mathcal{M}$. We follow this reasoning in order to discuss the evidence lower bound and the variational gap from an information-geometric perspective. We relate the maximization of the evidence to the maximization of its lower bound in view of information geometry and highlight the simplicity and consistency of both optimization problems when considered in the full ambient space, without restricting it to a model $\mathcal{M}$. We show that in this case the evidence lower bound leads to the same natural gradient field as the original objective function, the evidence, which we find remarkable. This equivalence is not necessarily preserved when restricting the

optimization to a model $\mathcal{M}$. We provide a sufficient condition for this to hold, which requires the notion of a cylindrical model.

In this article, we follow two story lines, one referring to the evidence and its lower bound and one referring to corresponding Kullback-Leibler divergences. We use the former story line to formulate the main problem and to convey the key findings without assuming a background in information geometry. The latter story line is more convenient for our information-geometric studies. Section 6.2 introduces the primary objective of learning, minimizing the Kullback-Leibler divergence from a target distribution on states of the visible units, and briefly outlines its relation to the evidence and its lower bound. This section is generally accessible, without a background in information geometry. In Section 6.3, we are then going to review basic information-geometric structures, thereby introducing the notation used in this article. This section also includes results from the previous work (Ay, 2020) on which this article is based. Section 6.4 deals with the analysis of the optimization problem for the extended system, including visible and hidden units, and relates it to the primary optimization problem defined for its visible part. Section 6.5 relates these results to the evidence and its lower bound, thereby making statements on their respective natural gradients. Section 6.6 concludes with a result that is particularly helpful when dealing specifically with Bayesian graphical models.

6.2 Learning a Target Distribution and the Evidence Lower Bound

Throughout this article, we consider a system consisting of visible units V and hidden units H taking values in state sets X_V and X_H, respectively. For simplicity, we assume X_V and X_H to be finite. The set of all strictly positive probability distributions on joint states (x_V, x_H) is denoted by $\mathcal{P}_{V,H}$, which we also abbreviate as $\mathcal{P}$. In order to study learning in terms of the natural gradient method, we consider a model $\mathcal{M}$ consisting of probability distributions $p_\theta(x_V, x_H) = p(x_V, x_H; \theta)$ which are parametrized by a parameter vector θ in $\mathbb{R}^d$. Typically, the parameter set is an open subset Θ of $\mathbb{R}^d$, and we obtain $\mathcal{M}$ as the image of the parametrization

$$\varphi : \Theta \to \mathcal{M} \subseteq \mathcal{P}, \qquad \theta \mapsto p_\theta. \tag{6.2}$$

The model $\mathcal{M}$ is referred to as a *generative model*. The objective of learning is to generate a probability distribution on visible states x_V that is close to some target distribution. Here, we interpret the hidden units merely as auxiliary units to increase the expressive power. The learning objective should therefore only refer to the visible units. To be more precise, we denote by $\mathcal{P}_V$ the set of strictly positive probability distributions on states x_V and consider the natural marginalization map

$$\pi_V : \mathcal{P} \to \mathcal{P}_V,$$

which assigns to a joint probability distribution $p(x_V, x_H)$ the marginal distribution

$$p(x_V) := \sum_{x_H} p(x_V, x_H). \tag{6.3}$$

The image of the model $\mathcal{M}$, that is $\pi_V(\mathcal{M})$, is denoted by $\mathcal{M}_V$. It consists of all probability distributions that can be generated by the learning system. With the parametrization (6.2), we can parametrize $\mathcal{M}_V$ in terms of

$$\pi_V \circ \varphi : \Theta \to \mathcal{M}_V \subseteq \mathcal{P}_V, \qquad \theta \mapsto \pi_V(p_\theta).$$

In this article, we will mostly omit the parameter and simply write $p \in \mathcal{M}$ and $p \in \mathcal{M}_V$, respectively.

Now consider a target distribution $p^* \in \mathcal{P}_V$. The objective of learning is to find $p \in \mathcal{M}_V$ that is close to p^*. To achieve that, we minimize the KL-divergence of p^* from a distribution $p \in \mathcal{M}_V$, that is,

$$D(p^*\|p) := \sum_{x_V} p^*(x_V) \ln \frac{p^*(x_V)}{p(x_V)}. \tag{6.4}$$

Throughout this article, we refer to this function as a primary objective function defined on $\mathcal{M}_V$ and therefore only involving visible units. This will be compared with corresponding lifted objective functions defined on $\mathcal{M}$ which involve the visible as well as the hidden units. Observe that the minimization of $D(p^*\|\cdot)$ is equivalent to the minimization of the *cross entropy*

$$-\sum_{x_V} p^*(x_V) \ln p(x_V)$$

because these two functions differ only by a constant, the *entropy* of p^*, which is given by

$$-\sum_{x_V} p^*(x_V) \ln p^*(x_V).$$

The cross entropy is nothing but the mean value of the *surprise*, $-\ln p(x_V)$. Alternatively, we can change the sign of the surprise and consider the *evidence*, $\ln p(x_V)$, leading to the mean value

$$\text{EVIDENCE}(p) := \sum_{x_V} p^*(x_V) \ln p(x_V), \tag{6.5}$$

which we also refer to as the *evidence* without explicitly highlighting the fact that it is an integrated quantity. Minimizing the KL-divergence (6.4) is then equivalent to maximizing the evidence (6.5). In order to be tractable, we bound the evidence from below by considering the set H of hidden units. For any conditional probability measure $q(x_H|x_V)$ and $p \in \mathcal{M}$, we then have

$$\begin{aligned}
\text{EVIDENCE}(\pi_V(p)) &= -\sum_{x_V,x_H} p^*(x_V) q(x_H|x_V) \ln \frac{q(x_H|x_V)}{p(x_V,x_H)} \\
&\quad + \sum_{x_V} p^*(x_V) \sum_{x_H} q(x_H|x_V) \ln \frac{q(x_H|x_V)}{p(x_H|x_V)} \qquad (6.6) \\
&\geq -\sum_{x_V,x_H} p^*(x_V) q(x_H|x_V) \ln \frac{q(x_H|x_V)}{p(x_V,x_H)} \qquad (6.7) \\
&=: \text{ELBO}(q,p). \qquad (6.8)
\end{aligned}$$

The inequality (6.7) follows from the non-negativity of the KL-divergences between the conditional probability distributions $q(\cdot|x_V)$ and $p(\cdot|x_V)$ in (6.6). The bound (6.8) is referred to as the *evidence lower bound*. It coincides with the negative of the *variational free energy*. The importance of this quantity has been highlighted in the introduction. The evidence lower bound gives rise to the function

$$\mathrm{ELBO}(q,\cdot) : \mathcal{M} \to \mathbb{R}, \qquad p \mapsto \mathrm{ELBO}(q,p).$$

Replacing the evidence by the evidence lower bound implies a number of simplifications of the optimization in terms of gradient methods. One instance of these simplifications will be outlined in some more detail in Section 6.6. But how much do we alter the original optimization problem by this replacement? To get a first intuition, observe that the gap between the evidence and its lower bound is given by the mean value (6.6) of KL-divergences,

$$\mathrm{GAP}(q,p) \;:=\; \sum_{x_V,x_H} p^*(x_V)q(x_H|x_V) \ln \frac{q(x_H|x_V)}{p(x_H|x_V)}. \tag{6.9}$$

In summary, we have the following relationship between the introduced quantities:

$$\mathrm{EVIDENCE}(\pi_V(p)) \;=\; \mathrm{ELBO}(q,p) + \mathrm{GAP}(q,p).$$

In Sections 6.4 and 6.5, we shall provide arguments supporting the hypothesis that the gap does not play a major role in learning. The main target of this article is to compare the natural gradient of $\mathrm{ELBO}(q,\cdot)$ on $\mathcal{M}$ with the natural gradient of EVIDENCE or, equivalently, the objective function $D(p^*\|\cdot)$ on $\mathcal{M}_V$. In order to imply the same learning process based on the natural gradient method, the respective gradients should be consistent in a sense that we are going to specify. To reveal a condition for such a consistency, we are going to interpret the derivations of this section in a more geometric way. Before coming to this, we first review some information-geometric preliminaries.

6.3 Information-Geometric Preliminaries

The set $\mathcal{P}$ of strictly positive probability distributions on some finite set X of states x represents the most basic example of a model within information geometry. We write a point $p \in \mathcal{P}$ as

$$p \;=\; \sum_x p(x)\,\delta^x, \tag{6.10}$$

where δ^x denotes the Dirac measure concentrated in x. The tangent space of $\mathcal{P}$ in p is given by

$$T_p\mathcal{P} \;=\; \left\{ A = \sum_x A(x)\,\delta^x \;:\; \sum_x A(x) = 0 \right\}.$$

For two vectors $A, B \in T_p\mathcal{P}$, we have the *Fisher-Rao metric*

$$g_p^{\mathrm{FR}}(A,B) \;=\; \sum_x \frac{1}{p(x)} A(x)B(x), \tag{6.11}$$

which is a Riemannian metric on $\mathcal{P}$. (Throughout this article, we also write $\langle A, B\rangle$ if there is no ambiguity regarding the Riemannian metric and the base point.) Furthermore, we consider the *Kullback-Leibler divergence* (*KL-divergence*) which is defined on $\mathcal{P} \times \mathcal{P}$ by

$$D(q\|p) \;=\; \sum_x q(x) \ln \frac{q(x)}{p(x)}. \tag{6.12}$$

Note that we already used the KL-divergence to define the primary objective function (6.4). We can express the Fisher-Rao gradients of the KL-divergence in both arguments:

$$\begin{aligned} \operatorname{grad}_p D(q\|\cdot) &= \sum_x (p(x) - q(x))\, \delta^x \\ &= p - q \;\in\; T_p\mathcal{P}, \\ \operatorname{grad}_q D(\cdot\|p) &= \sum_x q(x) \left(\ln \frac{q(x)}{p(x)} - \sum_{x'} q(x') \left(\ln \frac{q(x')}{p(x')} \right) \right) \delta^x \\ &= q \left(\ln \frac{q}{p} - \mathbb{E}_q \left(\ln \frac{q}{p} \right) \right) \;\in\; T_q\mathcal{P}. \end{aligned} \tag{6.13}$$

These gradients satisfy the defining condition (6.1), where $\mathcal{L}$ is the KL-divergence (6.12) in the first and the second argument, respectively, and $g = g^{\mathrm{FR}}$ as defined by (6.11). For more details, see (Ay and Amari, 2015; Ay et al., 2017).

Now we consider the marginalization map $\pi_V : \mathcal{P} \to \mathcal{P}_V$, defined in terms of (6.3). In order to relate tangent vectors in $T_p\mathcal{P}$ to tangent vectors in $T_{\pi_V(p)}\mathcal{P}_V$, we consider the differential

$$d\pi_V : T_p\mathcal{P} \to T_{\pi_V(p)}\mathcal{P}_V,$$

given by

$$d\pi_V(A)(x_V) \;=\; \sum_{x_H} A(x_V, x_H). \tag{6.14}$$

Furthermore, we introduce the following orthogonal spaces:

$$\mathcal{V}_p := \ker d\pi_V, \qquad \mathcal{H}_p := \mathcal{V}_p{}^{\perp},$$

where the orthogonal complement in the definition of $\mathcal{H}_p$ is meant to be with respect to the Fisher-Rao metric in $p \in \mathcal{P}$. We refer to $\mathcal{V}_p$ as the *vertical space* and to $\mathcal{H}_p$ as the *horizontal space* in p, which is in line with the differential-geometric terminology. This should not be confused with the symbols V and H for the visible and hidden units, respectively. In fact, by an unfortunate coincidence, the latter meaning of the symbols might even suggest the opposite naming. More precisely, the tangent space of $\mathcal{P}_V$ can be identified with the horizontal space $\mathcal{H}_p$ and not, as the symbol V in $\mathcal{P}_V$ might suggest, with the vertical space. Clearly, we have the orthogonal decomposition

$$T_p\mathcal{P} \;=\; \mathcal{H}_p \oplus \mathcal{V}_p\,.$$

Every vector A in $T_p\mathcal{P}$ has a unique representation as

$$A \;=\; A^{\mathcal{H}} + A^{\mathcal{V}},$$

where $A^{\mathcal{H}} \in \mathcal{H}_p$ and $A^{\mathcal{V}} \in \mathcal{V}_p$.

We now consider a model $\mathcal{M}$ in $\mathcal{P}$ and its π_V-image $\mathcal{M}_V$ and thereby restrict attention to non-singular points. A point $p \in \mathcal{M}$ is *admissible* if p and $\pi_V(p)$ are non-singular points of $\mathcal{M}$ and $\mathcal{M}_V$, respectively, and $d\pi_V(T_p\mathcal{M}) = T_{\pi_V(p)}\mathcal{M}_V$. Admissible points allow us to locally define the geometric structures that are relevant from the perspective of information geometry. In particular, the model $\mathcal{M}$ carries the induced geometry of $\mathcal{P}$ in an admissible point p, and $\mathcal{M}_V$ carries the corresponding induced geometry of $\mathcal{P}_V$ in $\pi_V(p)$. This will allow us to consider the gradient on $\mathcal{M}$, denoted by $\text{grad}^{\mathcal{M}}$, and the gradient on $\mathcal{M}_V$, denoted by $\text{grad}^{\mathcal{M}_V}$.

The objective of learning can be formulated as the optimization of a differentiable function $\mathcal{L} : \mathcal{P}_V \to \mathbb{R}$ on $\mathcal{M}_V$ which plays the role of a primary objective function. Examples are given by the KL-divergence (6.4), which we should minimize, and the mean evidence (6.5), which we should maximize. In what follows, we will mainly refer to the case of minimizing $\mathcal{L}$ on $\mathcal{M}_V$ by means of the gradient descent method. Alternatively, one could also minimize the corresponding lifted function $\mathcal{L} \circ \pi_V$ defined on $\mathcal{M}$. More precisely, consider a curve γ in $\mathcal{M}$ that solves the differential equation

$$\dot{\gamma}(t) \;=\; -\text{grad}^{\mathcal{M}}_{\gamma(t)}(\mathcal{L} \circ \pi_V), \qquad \gamma(0) = p,$$

where we assume that all points $\gamma(t)$ are admissible and $\dot{\gamma}(0) \neq 0$. (Throughout this article, we assume the existence and uniqueness of maximal solutions of differential equations without explicitly stating the conditions for this to hold.) Furthermore, let $\sigma := \pi_V \circ \gamma$ be the projected curve in $\mathcal{M}_V$. The change of $\mathcal{L}$ along σ is then given by:

$$\frac{d}{dt}\,\mathcal{L}(\sigma(t)) \;=\; \frac{d}{dt}\,\mathcal{L}(\pi_V(\gamma(t))) \;=\; \frac{d}{dt}\,(\mathcal{L} \circ \pi_V)(\gamma(t)) \;<\; 0.$$

This shows that the minimization of the lifted function $\mathcal{L} \circ \pi_V$ on $\mathcal{M}$ provides a useful strategy for minimizing the primary objective function $\mathcal{L}$ on $\mathcal{M}_V$. However, even though $\mathcal{L}$ is decreasing along σ, it will typically not be following minus the gradient of $\mathcal{L}$ on $\mathcal{M}_V$. From the chain rule we have in general that

$$\begin{aligned} \dot{\sigma}(t) &= d\pi_V\left(\dot{\gamma}(t)\right) \\ &= -d\pi_V\left(\text{grad}^{\mathcal{M}}_{\gamma(t)}(\mathcal{L} \circ \pi_V)\right). \end{aligned} \tag{6.15}$$

Following the Fisher-Rao gradient on $\mathcal{M}_V$, however, would require

$$\dot{\sigma}(t) \;=\; -\text{grad}^{\mathcal{M}_V}_{\sigma(t)}\mathcal{L}. \tag{6.16}$$

The vector fields defined by the respective RHS of (6.15) and (6.16) are typically different, but they point, at least, in a similar direction, which is shown in the following proposition.

Proposition 6.1. *Let $\mathcal{M}$ be a model in $\mathcal{P}$, let $\mathcal{L} : \mathcal{P}_V \to \mathbb{R}$ be a differentiable objective function, and let $p \in \mathcal{M}$ be an admissible point. Then,*

$$\text{grad}^{\mathcal{M}}_p\left(\mathcal{L} \circ \pi_V\right) \;=\; 0 \quad \Leftrightarrow \quad \text{grad}^{\mathcal{M}_V}_{\pi_V(p)}\mathcal{L} \;=\; 0.$$

Furthermore, if one of the two gradients does not vanish, we have

$$\left\langle d\pi_V\left(\text{grad}^{\mathcal{M}}_p\left(\mathcal{L} \circ \pi_V\right)\right), \text{grad}^{\mathcal{M}_V}_{\pi_V(p)}\mathcal{L}\right\rangle \;>\; 0.$$

Proof. For an arbitrary $A \in T_p\mathcal{M}$, we have

$$\begin{aligned}\left\langle \operatorname{grad}_p^{\mathcal{M}}(\mathcal{L}\circ\pi_V), A\right\rangle &= d(\mathcal{L}\circ\pi_V)_p(A)\\ &= \left(d\mathcal{L}_{\pi_V(p)}\circ d\pi_V\right)(A)\\ &= d\mathcal{L}_{\pi_V(p)}(d\pi_V(A))\\ &= \left\langle \operatorname{grad}^{\mathcal{M}_V}_{\pi_V(p)}\mathcal{L}, d\pi_V(A)\right\rangle.\end{aligned}$$

This implies that $\operatorname{grad}_p^{\mathcal{M}}(\mathcal{L}\circ\pi_V)$ vanishes if and only if $\operatorname{grad}^{\mathcal{M}_V}_{\pi_V(p)}\mathcal{L}$ vanishes (note that the $d\pi_V(A)$, $A \in T_p\mathcal{M}$, span the tangent space $T_{\pi_V(p)}\mathcal{M}_V$ because p is assumed to be admissible). Furthermore, for the special case $A = \operatorname{grad}_p^{\mathcal{M}}(\mathcal{L}\circ\pi_V) \neq 0$, we obtain

$$\begin{aligned}\left\langle d\pi_V\left(\operatorname{grad}_p^{\mathcal{M}}(\mathcal{L}\circ\pi_V)\right), \operatorname{grad}^{\mathcal{M}_V}_{\pi_V(p)}\mathcal{L}\right\rangle &= \left\langle \operatorname{grad}_p^{\mathcal{M}}(\mathcal{L}\circ\pi_V), \operatorname{grad}_p^{\mathcal{M}}(\mathcal{L}\circ\pi_V)\right\rangle\\ &> 0.\end{aligned}$$

□

We now ask the question under which conditions the two gradient fields of Proposition 6.1 are not only pointing in a similar direction but are actually equal. The following definition specifies the models $\mathcal{M}$ for which this is satisfied for any objective function $\mathcal{L} : \mathcal{M} \to \mathbb{R}$, as stated in Theorem 6.1. For such models, a projected solution curve σ in $\mathcal{M}_V$ satisfies equation (6.16).

Definition 6.1 (Definition 1 of Ay, 2020)**.** We call a model $\mathcal{M} \subseteq \mathcal{P}$ *cylindrical* in a non-singular point $p \in \mathcal{M}$, if

$$T_p\mathcal{M} = (T_p\mathcal{M} \cap \mathcal{H}_p) \oplus (T_p\mathcal{M} \cap \mathcal{V}_p).$$

If the model is cylindrical in all non-singular points $p \in \mathcal{M}$ then we call it *(pointwise) cylindrical*.

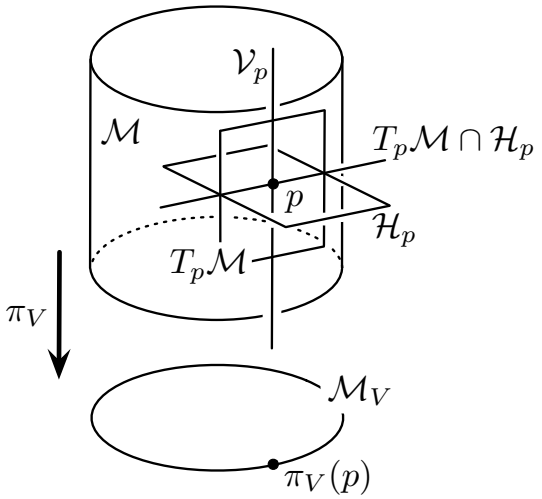


Figure 6.1: Illustration of a cylindrical model $\mathcal{M}$ in terms of a cylinder, the Cartesian product of a circle with a finite interval. The tangent space $T_p\mathcal{M}$ equals the sum of its intersections with $\mathcal{H}_p$ and $\mathcal{V}_p$.

See Figure 6.1 for an illustration of a cylindrical model and Appendix 6.A for examples of cylindrical and non-cylindrical models. It has been shown in Ay, 2020, Theorem 3

that a model $\mathcal{M}$ is cylindrical if and only if for the restriction $\pi_V|_{\mathcal{M}} : \mathcal{M} \to \mathcal{M}_V$ the following holds: Given $A, B \in (\ker d\pi_V|_{\mathcal{M}})^\perp$, we have

$$g_p^{\mathrm{FR}}(A, B) \;=\; g_{\pi_V(p)}^{\mathrm{FR}}\left(d\pi_V(A), d\pi_V(B)\right), \tag{6.17}$$

whenever p is *admissible*. The equality (6.17) is central in the definition of a Riemannian submersion. The property of $\mathcal{M}$ being cylindrical ensures the invariance of the natural gradient, as stated in the following theorem. (This is different from the often stated invariance of the natural gradient under coordinate transformations, elaborated on in (Oostrum et al., 2023).)

Theorem 6.1 (Theorem 5 of Ay, 2020)**.** *Let $\mathcal{M}$ be a cylindrical model, let $\mathcal{L} : \mathcal{M}_V \to \mathbb{R}$ be a differentiable objective function, and let $p \in \mathcal{M}$ be an admissible point. Then,*

$$d\pi_V\left(\mathrm{grad}_p^{\mathcal{M}}(\mathcal{L} \circ \pi_V)\right) \;=\; \mathrm{grad}_{\pi_V(p)}^{\mathcal{M}_V}\mathcal{L}. \tag{6.18}$$

Note that the gradient on the LHS of (6.18) refers to the Fisher-Rao metric on $\mathcal{M} \subseteq \mathcal{P} = \mathcal{P}_{V,H}$, whereas the RHS refers to the Fisher-Rao metric on $\mathcal{M}_V \subseteq \mathcal{P}_V$. The invariance of the gradient as formulated in Theorem 6.1 is quite special and holds only for the Fisher-Rao metric and cylindrical models (see (Ay, 2020) for further details). Our main example of a cylindrical model will be the full model $\mathcal{M} = \mathcal{P}$. More precisely, all points p are non-singular and we obviously have $T_p\mathcal{P} = \mathcal{H}_p \oplus \mathcal{V}_p = (T_p\mathcal{P} \cap \mathcal{H}_p) \oplus (T_p\mathcal{P} \cap \mathcal{V}_p)$. This example will provide the setting in which information-geometric quantities are studied in the absence of constraints through a lower-dimensional model. Clearly, when dealing with learning systems, we typically do have constraints. By relating this typical situation to the situation without constraints we are able to reveal the geometric effect of these constraints.

We conclude this section with a simple statement about the orthogonal projection onto the tangent space of a cylindrical model.

Lemma 6.1. *Let $\mathcal{M}$ be a cylindrical model in $\mathcal{P}$, let p be a non-singular point of $\mathcal{M}$, and let Π_p denote the orthogonal projection of $T_p\mathcal{P}$ onto $T_p\mathcal{M}$. Then,*

$$\Pi_p(\mathcal{H}_p) \;\subseteq\; \mathcal{H}_p, \qquad \Pi_p(\mathcal{V}_p) \;\subseteq\; \mathcal{V}_p$$

Proof. Definition 6.1 implies the following orthogonal decomposition:

$$\begin{aligned} T_p\mathcal{P} &= T_p\mathcal{M} \oplus (T_p\mathcal{M})^\perp \\ &= (T_p\mathcal{M} \cap \mathcal{H}_p) \oplus (T_p\mathcal{M} \cap \mathcal{V}_p) \oplus (T_p\mathcal{M})^\perp . \end{aligned}$$

Thus, every vector $X \in T_p\mathcal{P}$ has a unique orthogonal decomposition as $X = A+B+C$, where $A \in (T_p\mathcal{M} \cap \mathcal{H}_p)$, $B \in (T_p\mathcal{M} \cap \mathcal{V}_p)$, and $C \in (T_p\mathcal{M})^\perp$. With this decomposition, we have $\Pi_p(A) = A$, $\Pi_p(B) = B$, and $\Pi_p(C) = 0$. Now, if $X \in \mathcal{H}_p$ then its B component vanishes, so that $\Pi_p(X) = \Pi_p(A + C) = \Pi_p(A) + \Pi_p(C) = A \in \mathcal{H}_p$. If, on the other hand, $X \in \mathcal{V}_p$ then its A component vanishes, so that $\Pi_p(X) = \Pi_p(B + C) = \Pi_p(B) + \Pi_p(C) = B \in \mathcal{V}_p$. □

6.4 The Extended Problem with Hidden Units

In this section, we are going to relate the minimization of the KL-divergence (6.4), $\mathcal{L} := D(p^*\|\cdot)$, on $\mathcal{M}_V$ to the minimization of the lifted function $\mathcal{L} \circ \pi_V$ on $\mathcal{M}$. In general, it is difficult to minimize $\mathcal{L}$. In particular, we face here various challenges when trying to apply the natural gradient descent method. On the one hand, $\mathcal{M}_V$ will typically have singularities so that gradients cannot be evaluated in these points. On the other hand, even for non-singular points the Fisher-Rao metric will be difficult to evaluate if we do not assume $\mathcal{M}_V$ to have a particularly simple structure. To be more concrete, we first evaluate the gradient of $D(p^*\|\cdot)$, considered as a function on $\mathcal{P}_V$ (see equation (6.13)):

$$\operatorname{grad}_p^{\mathcal{P}_V} D(p^*\|\cdot) \;=\; p - p^* \;\in\; T_p\mathcal{P}_V. \tag{6.19}$$

For the gradient on the model $\mathcal{M}_V$, we then have to project the gradient (6.19) in p onto the tangent space $T_p\mathcal{M}_V$, thereby assuming that p is a non-singular point of $\mathcal{M}_V$. This leads to

$$\operatorname{grad}_p^{\mathcal{M}_V} D(p^*\|\cdot) \;=\; \Pi_p(p - p^*) \;\in\; T_p\mathcal{M}_V, \tag{6.20}$$

where Π_p denotes the orthogonal projection onto the tangent space $T_p\mathcal{M}_V$. Note that the projected vector $\Pi_p(p - p^*)$ does not have to be particularly simple, even though the difference vector $p - p^*$, the gradient in the ambient space, is simple.

We are now going to modify the problem of minimizing the KL-divergence (6.4) in several simplifying steps, thereby tracing the geometric implication of each individual step. The overall aim of this modification is to relate the minimization of (6.4), or equivalently the maximization of the evidence, to the corresponding maximization of the evidence lower bound, which will be finally addressed in Section 6.5.

It is well-known that the minimization of the KL-divergence (6.4) can be simplified by extending the problem to the space of probability distributions on joint states (x_V, x_H) that is $\mathcal{P}_{V,H}$ (see Amari, 2016, Chapter 8). For that, we consider the so-called *data manifold*

$$\mathcal{Q} := \{q \in \mathcal{P}_{V,H} \;:\; \pi_V(q) = p^*\}.$$

Note that the symbol q here denotes a joint probability distribution whereas previously we have used the same symbol for the conditional probability distribution. The relation is given by $q(x_V, x_H) = p^*(x_V)q(x_H|x_V)$. Thus, even though it is not visible at first sight, the data manifold $\mathcal{Q}$ incorporates the data distribution p^*. With the monotonicity of the KL-divergence, we obtain for any $p \in \mathcal{M}$ and $q \in \mathcal{Q}$

$$\begin{aligned}(\mathcal{L} \circ \pi_V)(p) &= D(p^*\|\pi_V(p)) \\ &= D(\pi_V(q)\|\pi_V(p)) \\ &\leq D(q\|p),\end{aligned}$$

where equality holds for $q = \pi_{\mathcal{Q}}(p)$ defined by

$$\pi_{\mathcal{Q}}(p)(x_V, x_H) \;=\; p^*(x_V)p(x_H|x_V). \tag{6.21}$$

Thus, we have

$$\begin{aligned}(\mathcal{L} \circ \pi_V)(p) &= D(\pi_{\mathcal{Q}}(p)\|p) \\ &= \inf_{q\in\mathcal{Q}} D(q\|p) \\ &=: D(\mathcal{Q}\|p).\end{aligned}$$

Clearly, a point $\hat{p}$ minimizes $\mathcal{L} \circ \pi_V = D(\mathcal{Q}\|\cdot)$ in $\mathcal{M}$ if and only if $\pi_V(\hat{p})$ minimises $\mathcal{L} = D(p^*\|\cdot)$ in $\mathcal{M}_V$. However, there are important differences between the corresponding optimizations in terms of the natural gradient method. On the one hand, $\mathcal{M}$ typically comes with a geometric structure that simplifies the optimization of $\mathcal{L} \circ \pi_V$. On the other hand, for the optimization of $\mathcal{L}$ it is natural to use the Fisher-Rao metric on $\mathcal{M}_V$, whereas $\mathcal{L} \circ \pi_V$ is defined on $\mathcal{M}$ and should be optimized with respect to the corresponding Fisher-Rao gradient on $\mathcal{M}$. In general, the two ways to optimize basically the same function will not be equivalent. However, according to Theorem 6.1, they will be equivalent whenever the model $\mathcal{M}$ is cylindrical.

Theorem 6.2. **(a)** *Consider first the function $D(\mathcal{Q}\|\cdot)$ on $\mathcal{P}$. Then*

$$\operatorname{grad}_p^{\mathcal{P}} D(\mathcal{Q}\|\cdot) \;=\; p - \pi_{\mathcal{Q}}(p), \tag{6.22}$$

where $\pi_{\mathcal{Q}}(p)$ is defined by (6.21)*. In order to obtain the gradient of $D(\mathcal{Q}\|\cdot)$ in a non-singular point $p \in \mathcal{M}$, we have to project* (6.22) *onto $T_p\mathcal{M}$, that is*

$$\operatorname{grad}_p^{\mathcal{M}} D(\mathcal{Q}\|\cdot) \;=\; \Pi_p(p - \pi_{\mathcal{Q}}(p)), \tag{6.23}$$

where Π_p denotes the orthogonal projection $T_p\mathcal{P} \to T_p\mathcal{M}$ with respect to the Fisher-Rao metric on $\mathcal{P}$.

(b) *If $\mathcal{M}$ is cylindrical and $p \in \mathcal{M}$ admissible then*

$$d\pi_V \left(\operatorname{grad}_p^{\mathcal{M}} D(\mathcal{Q}\|\cdot)\right) \;=\; \operatorname{grad}_{\pi_V(p)}^{\mathcal{M}_V} D(p^*\|\cdot). \tag{6.24}$$

In particular, the equality (6.24) *holds in all points of the maximal model $\mathcal{M} = \mathcal{P}$ where $\mathcal{M}_V = \mathcal{P}_V$.*

Proof. We know that $D(\mathcal{Q}\|\cdot) = D(p^*\|\pi_V(\cdot))$, which is a function of p or, equivalently, a function of its coordinates $p(x_V, x_H)$ with respect to the basis vectors δ^{x_V,x_H} (see equation (6.10)). We evaluate the partial derivatives with respect to these coordinates,

$$\frac{\partial}{\partial p(x_V, x_H)} D(p^*\|\pi_V(\cdot)) \;=\; -\frac{p^*(x_V)}{p(x_V)},$$

and obtain for the (x_V, x_H)-component of the natural gradient (see (Ay et al., 2017), Proposition 2.2)

$$\begin{aligned}
&\left(\operatorname{grad}_p^{\mathcal{P}} D(\mathcal{Q}\|\cdot)\right)(x_V, x_H) \\
&= p(x_V, x_H) \left(-\frac{p^*(x_V)}{p(x_V)} + \sum_{x'_V, x'_H} p(x'_V, x'_H) \frac{p^*(x_V)}{p(x_V)} \right) \\
&= p(x_V, x_H) \left(-\frac{p^*(x_V)}{p(x_V)} + 1 \right) \\
&= p(x_V, x_H) - p(x_H|x_V) p^*(x_V) \\
&= p(x_V, x_H) - \pi_{\mathcal{Q}}(p)(x_V, x_H).
\end{aligned}$$

This proves equation (6.22), and equation (6.23) follows immediately from that. Finally, the invariance (6.24) is a direct consequence of Theorem 6.1. □

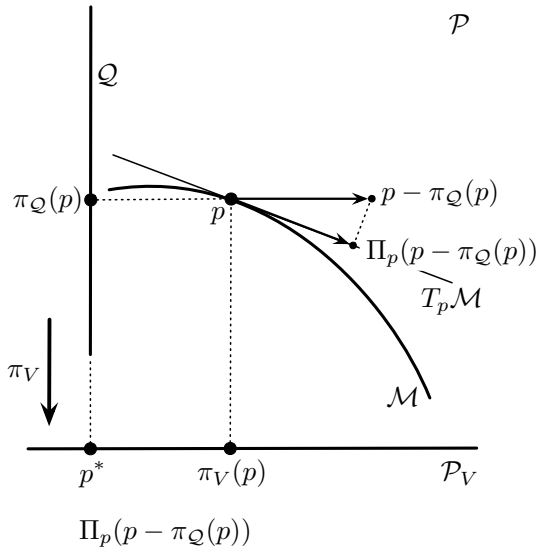


Figure 6.2: Illustration of the gradients considered in Theorem 6.2.

The gradients considered in Theorem 6.2 are graphically illustrated in Figure 6.2. Theorem 6.2 reveals a number of insights concerning the complexity and the invariance of the natural gradients which we are now going to elaborate on. First of all, it highlights the simplicity of the natural gradient of $D(\mathcal{Q}\|\cdot)$ in $p \in \mathcal{P}$. It is nothing but the difference vector between p and its projection $\pi_{\mathcal{Q}}(p)$. Thus, any complexity of the natural gradient of $D(\mathcal{Q}\|\cdot)$ on a model $\mathcal{M}$ arises from the projection of that difference vector onto the tangent space $T_p\mathcal{M}$ and therefore depends very much on the structure of $\mathcal{M}$. For a Bayesian graphical model, $T_p\mathcal{M}$ decomposes in a convenient way so that some of the original simplicity is preserved after projection. A corresponding more precise statement will be formulated at the end of this article, in Proposition 6.2. Furthermore, the gradient (6.23) of the function $D(\mathcal{Q}\|\cdot)$, defined on $\mathcal{M}$, can now be compared with the gradient (6.20) of the original function $D(p^*\|\cdot)$ which is defined on $\mathcal{M}_V$. According to the invariance (6.24), these two gradients are equivalent, if $\mathcal{M}$ is cylindrical, which implies that gradient descent learning in $\mathcal{M}$ yields exactly the same trajectories as the gradient descent learning in $\mathcal{M}_V$. This is a consequence of the corresponding invariance of the Fisher-Rao metric as formulated by Chentsov and not at all given for other choices of Riemannian metrics (Chentsov, 1982). While the requirement for a model to be cylindrical is quite restrictive, it holds for the full model $\mathcal{M} = \mathcal{P}$. This brings us to the last insight of Theorem 6.2. If we do not restrict the optimization to a lower-dimensional model $\mathcal{M}$ then all information-geometric structures are consistent in the sense that the optimization in the extended system, with hidden units, is equivalent to the original optimization with only visible units. Again, any deviation from the invariance (6.24) arises from the restriction of the optimization to $\mathcal{M}$.

We can illustrate the invariance (6.24) and a possible deviation from it using two example models, denoted by $\mathcal{M}^{(a)}$ and $\mathcal{M}^{(b)}$, which we introduce in what follows. (The code for reproducing the data and figures in this paper is made available at Datar et al., 2024.) Consider three binary random variables X_s, X_{t_1}, X_{t_2}. The manifold $\mathcal{P}$ consists of all joint probability distributions $p(x_s, x_{t_1}, x_{t_2})$. With $H = \{s\}$ and $V = \{t_1, t_2\}$, let

π_V be the marginalization map over X_s, i.e.

$$\begin{aligned} \pi_V \colon \mathcal{P} &\to \mathcal{P}_V \\ p(x_s, x_{t_1}, x_{t_2}) &\mapsto p(x_{t_1}, x_{t_2}) = \sum_{x_s} p(x_s, x_{t_1}, x_{t_2}). \end{aligned}$$

For a general model $\mathcal{M} \subseteq \mathcal{P}$ we study the following curves. We let $\sigma_0 \in \mathcal{M}_V$ be the integral curve of the gradient of the primary objective function $D(p^*\|\cdot)$, i.e. solving the differential equation

$$\dot{\sigma}_0(t) = -\mathrm{grad}^{\mathcal{M}_V}_{\sigma_0(t)} D(p^*\|\cdot), \quad \sigma_0(0) = \pi_V(p). \tag{6.25}$$

Furthermore, we let the curve γ_1 in $\mathcal{M}$ be the solution of the differential equation

$$\dot{\gamma}_1(t) = -\mathrm{grad}^{\mathcal{M}}_{\gamma_1(t)} D(\mathcal{Q}\|\cdot), \quad \gamma_1(0) = p, \tag{6.26}$$

and $\sigma_1 = \pi_V \circ \gamma_1$ in $\mathcal{M}_V$ be the projection of γ_1.

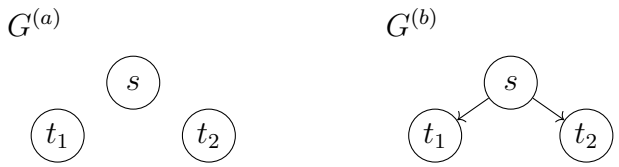


Figure 6.3: Graphical representations of the models $\mathcal{M}^{(a)}$ and $\mathcal{M}^{(b)}$.

We are now going to define two models in $\mathcal{P}$, $\mathcal{M}^{(a)}$ and $\mathcal{M}^{(b)}$, given by the corresponding graphs $G^{(a)}$ and $G^{(b)}$ in Figure 6.3. We begin with the model $\mathcal{M}^{(a)}$, which we define as the set of probability distributions for which X_s, X_{t_1}, X_{t_2} are independent, that is,

$$\mathcal{M}^{(a)} = \{p \in \mathcal{P} : p(x_s, x_{t_1}, x_{t_2}) = p(x_s)p(x_{t_1})p(x_{t_2})\}. \tag{6.27}$$

Graphically, these are all the distributions factorizing over the graph $G^{(a)}$ in Figure 6.3. It can be shown that this model is cylindrical.[1] Owing to Theorem 6.2, we know that the curves σ_0 and σ_1 are identical. This is shown in Figure 6.5, where the model $\mathcal{M}^{(a)}_V$ is plotted by the blue grid and the indistinguishable curves σ_0 and σ_1 are denoted by the solid black line. Note that the black line also represents the gradient curve coming from the evidence lower bound which is going to be discussed in the next section.

Similarly, we now define

$$\mathcal{M}^{(b)} = \{p \in \mathcal{P} : p(x_s, x_{t_1}, x_{t_2}) = p(x_s)p(x_{t_1}|x_s)p(x_{t_2}|x_s)\}. \tag{6.28}$$

It consists of those distributions that factorize over the graph $G^{(b)}$ in Figure 6.3. In Example 3 of Appendix 6.A we show that this model is not cylindrical. The model $\mathcal{M}^{(b)}_V$ is in this case the full simplex $\mathcal{P}_V$. Figure 6.6 (top) shows the trajectories of the curves σ_0 and σ_1 defined by (6.25) and (6.26), respectively, using dashed blue and solid green lines. (The solid red lines are related to the ELBO objective function elaborated on in the next section.) Figure 6.6 (bottom-left) shows the same trajectories in coordinates as functions of time and Figure 6.6 (bottom-right) shows the KL-divergence evaluated

[1] See Example 1 in Appendix 6.A for a two-node example of this.

on these trajectories as a function of time. Note that now the trajectories of σ_0 and σ_1 do not coincide, deviating from the situation of Figure 6.5, due to $\mathcal{M}^{(b)}$ not being cylindrical. In spite of this, the trajectories converge to the target distribution as evident from the top two rows. Furthermore, the evaluations of the KL-divergence along these different trajectories is almost indistinguishable. Since this decay of KL-divergence corresponds directly to the speed of learning, understanding the effect of a model being cylindrical on the speed of learning is an important question for future research. Finally, observe that since $\mathcal{M}_V^{(b)} = \mathcal{P}_V$, the trajectories of the integral curves σ_0 (dashed blue) of the gradient of $D(p^*\|\cdot)$ are straight lines. This is a consequence of (6.13) and has more general implications on the learning (Datar and Ay, 2025).

In order to measure the deviation from the invariance (6.24) we evaluate the cosine similarity of the involved vector fields. More precisely, we compute the cosine similarity between the vectors $\operatorname{grad}^{\mathcal{M}_V}_{\pi_V(p)} D(p^*\|\cdot)$ and $d\pi_V\left(\operatorname{grad}^{\mathcal{M}}_p D(\mathcal{Q}\|\cdot)\right)$ in $T_p\mathcal{M}_V^{(b)}$, thereby assuming them to be non-zero. In general, the cosine similarity between two non-zero vectors A and B in an inner product space is defined as

$$\cos(\alpha) \;=\; \frac{\langle A, B\rangle}{\|A\|\|B\|}, \tag{6.29}$$

where α is the angle between A and B. The cosine similarity reaches its maximal value 1 when A and B point in the same direction and its minimal value -1 when A and B point in opposite directions. Clearly, in our setting the inner product and the norm are given by the Fisher-Rao metric in $T_p\mathcal{M}_V^{(b)}$. For a fixed p^*, the cosine similarity depends on the base point p which is sampled from $\mathcal{M}^{(b)}$ according to Jeffrey's prior. The results are plotted in the histograms in Figure 6.7 (left). As one can see, most points are close to 1. In fact, more than 83% of the samples are above 0.7. Furthermore, all of the values are larger than zero, which means that both vector fields qualify for optimizing the primary objective function $D(p^*\|\cdot)$. This is not a coincidence and follows directly from Proposition 6.1.

6.5 The Natural Gradient of the Evidence Lower Bound

In Theorem 6.2, we have related the gradient of the primary objective functions $D(p^*\|\cdot)$ to the gradient of $D(\mathcal{Q}\|\cdot)$. We are now going to replace the entire set $\mathcal{Q}$ in $D(\mathcal{Q}\|\cdot)$ by a single point $q \in \mathcal{Q}$, leading to the upper bound $D(q\|\cdot)$. In Theorem 6.3 below, we will then relate the gradient of the primary objective function $D(p^*\|\cdot)$ to the gradient of $D(q\|\cdot)$. This will finally allow us to study the natural gradient of the evidence lower bound in relation to the natural gradient of the evidence.

For any $q \in \mathcal{Q}$ and $p \in \mathcal{M}$, we apply the Pythagorian relation and obtain

$$\begin{aligned} D(\mathcal{Q}\|p) &\leq D(\mathcal{Q}\|p) + D(q\|\pi_{\mathcal{Q}}(p)) && (6.30)\\ &= D(\pi_{\mathcal{Q}}(p)\|p) + D(q\|\pi_{\mathcal{Q}}(p)) \\ &= D(q\|p) && (6.31)\end{aligned}$$

It can be easily verified that the additional term in (6.30), $D(q\|\pi_{\mathcal{Q}}(p))$, coincides with $\mathrm{GAP}(q,p)$ as defined by (6.9). Now, instead of taking the gradient of $D(\mathcal{Q}\|\cdot)$ we take

the gradient of the upper bound (6.31), with a fixed q, and analyze the effect of this replacement. For that, let us first rewrite this upper bound as follows:

$$D(q\|p) \;=\; D(\mathcal{Q}\|p) + \mathrm{GAP}(q,p). \tag{6.32}$$

For the gradient of $D(q\|\cdot)$ in $\mathcal{P}$, we obtain

$$\begin{aligned}\mathrm{grad}^{\mathcal{P}}_p D(q\|\cdot) &= \mathrm{grad}^{\mathcal{P}}_p D(\mathcal{Q}\|\cdot) + \mathrm{grad}^{\mathcal{P}}_p \mathrm{GAP}(q,\cdot)\\ &= (p-\pi_{\mathcal{Q}}(p)) + (\pi_{\mathcal{Q}}(p)-q). \qquad \text{(by (6.22) and (6.13))}\end{aligned} \tag{6.33}$$

It is easy to see that the first difference vector in (6.33), $p-\pi_{\mathcal{Q}}(p)$, is an element of the horizontal space $\mathcal{H}_p$, whereas the second one, $\pi_{\mathcal{Q}}(p)-q$, is an element of the vertical space $\mathcal{V}_p$. Therefore, the $d\pi_V$-image of the latter difference vector vanishes:

$$\begin{aligned} d\pi_V(\pi_{\mathcal{Q}}(p)-q) \;&=\; \sum_{x_H} (p^*(x_V)p(x_H|x_V) - q(x_V,x_H)) \qquad \text{(by (6.14))}\\ &=\; p^*(x_V)\sum_{x_H}(p(x_H|x_V)-q(x_H|x_V))\\ &=\; p^*(x_V)(1-1) \;=\; 0,\end{aligned}$$

or equivalently,

$$d\pi_V\left(\mathrm{grad}^{\mathcal{P}}_p \mathrm{GAP}(q,\cdot)\right) \;=\; 0.$$

In summary, we obtain

$$\begin{aligned} d\pi_V\left(\mathrm{grad}^{\mathcal{P}}_p D(q\|\cdot)\right) \;&=\; d\pi_V\left(p-\pi_{\mathcal{Q}}(p)\right)\\ &=\; \pi_V(p) - p^*\\ &=\; \mathrm{grad}^{\mathcal{P}_V}_{\pi_V(p)} D(p^*\|\cdot).\end{aligned}$$

This shows that the Fisher-Rao gradient of the primary objective function $D(p^*\|\cdot)$ on $\mathcal{P}_V$ is not affected at all by the extension of the problem to the set $\mathcal{P}=\mathcal{P}_{V,H}$. When we replace $\mathcal{P}$ by a more general model $\mathcal{M}$ this invariance only holds, if $\mathcal{M}$ is cylindrical.

Theorem 6.3. *Let $\mathcal{M}$ be a cylindrical model in $\mathcal{P}$, let $p\in\mathcal{M}$ be admissible, and let $q\in\mathcal{Q}$. Then*

$$d\pi_V\left(\mathrm{grad}^{\mathcal{M}}_p \mathrm{GAP}(q,\cdot)\right) \;=\; 0, \tag{6.34}$$

and therefore

$$d\pi_V\left(\mathrm{grad}^{\mathcal{M}}_p D(q\|\cdot)\right) \;=\; \mathrm{grad}^{\mathcal{M}_V}_{\pi_V(p)} D(p^*\|\cdot). \tag{6.35}$$

Proof. We begin with the gradient of $\mathrm{GAP}(q,\cdot)$:

$$\mathrm{grad}^{\mathcal{M}}_p \mathrm{GAP}(q,\cdot) \;=\; \Pi_p\left(\mathrm{grad}^{\mathcal{P}}_p \mathrm{GAP}(q,\cdot)\right) \;=\; \Pi_p\left(\pi_{\mathcal{Q}}(p)-q\right) \;=\; \Pi_p\left((p-q)^{\mathcal{V}}\right).$$

We know, by definition, that $(p-q)^{\mathcal{V}}$ is contained in $\mathcal{V}_p$. According to Lemma 6.1, the vector $(p-q)^{\mathcal{V}}$ remains in $\mathcal{V}_p$ after projecting it onto the tangent space $T_p\mathcal{M}$, that is,

$$\Pi_p\left((p-q)^{\mathcal{V}}\right) \;\in\; \mathcal{V}_p.$$

Therefore, this vector is mapped via $d\pi_V$ to 0, which proves (6.34). With this, we have

$$\begin{aligned} d\pi_V\left(\mathrm{grad}^{\mathcal{M}}_p D(q\|\cdot)\right) \;&=\; d\pi_V\left(\mathrm{grad}^{\mathcal{M}}_p D(\mathcal{Q}\|\cdot)\right) + d\pi_V\left(\mathrm{grad}^{\mathcal{M}}_p \mathrm{GAP}(q,\cdot)\right)\\ &=\; d\pi_V\left(\mathrm{grad}^{\mathcal{M}}_p D(\mathcal{Q}\|\cdot)\right),\end{aligned}$$

and with (6.24) of Theorem 6.2 we finally obtain (6.35). □

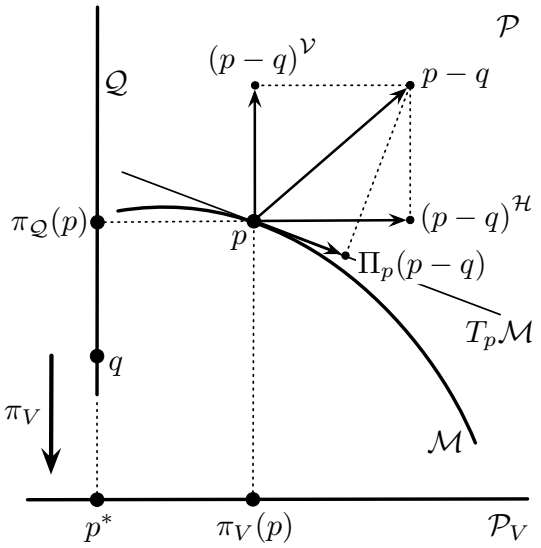


Figure 6.4: Illustration of gradients considered in Theorem 6.3.

The gradients considered in Theorem 6.3 are graphically illustrated in Figure 6.4. Note that while the invariance (6.35) appears to be very similar to the invariance (6.24), it is in fact quite different. The main difference is that the objective function on $\mathcal{M}$, the function $D(q\|\cdot)$, is not "just" the pull-back of an objective function on $\mathcal{M}_V$. It consists of a pull-back component, the first term on the RHS of (6.32), and the function $\mathrm{GAP}(q,\cdot)$, the second term on the RHS of (6.32), which varies only in vertical direction so that the $d\pi_V$-image of its gradient vanishes. Note that this result, expressed by (6.34), depends crucially on the information-geometric structures and does not hold for the standard Euclidean geometry defined in terms of a coordinate system. We give an example in Appendix 6.B.

We illustrate the invariance (6.35) and a possible deviation from it by revisiting the example models $\mathcal{M}^{(a)}$ and $\mathcal{M}^{(b)}$ from the previous section depicted in Figure 6.3. Now, for a general model $\mathcal{M}$ we let the curve γ_2 in $\mathcal{M}$ be the solution to the differential equation

$$\dot{\gamma}_2(t) \;=\; -\mathrm{grad}^{\mathcal{M}}_{\gamma_2(t)}\, D(q\|\cdot), \quad \gamma_2(0) \;=\; p,$$

and $\sigma_2 = \pi_V \circ \gamma_2$ the projection of γ_2. Since $\mathcal{M}^{(a)}$ is cylindrical, we can use Theorems 6.2 and 6.3 to conclude that the curves σ_0, σ_1 and σ_2 are identical. This is depicted in Figure 6.5 where the blue grid represents the cylindrical model $\mathcal{M}^{(a)}_V$, as a subset of $\mathcal{P}_V$, and the solid black line represents the indistinguishable curves σ_0, σ_1 and σ_2.

For the non-cylindrical model $\mathcal{M}^{(b)}$ we plot the trajectory of σ_2 in Figure 6.6 (top) in solid red and compare it with the trajectory of σ_0 (shown in dashed blue) and σ_1 (shown in solid green). Figure 6.6 (bottom-left) shows the trajectories in coordinates as functions of time and Figure 6.6 (bottom-right) shows the KL-divergence evaluated on these trajectories as a function of time. Note again that the trajectory σ_2 is distinct from σ_0 and σ_1. However, σ_2 also converges to the same target distribution as that of σ_0 or σ_1 and the KL-divergence evaluation is barely distinguishable. In line with the discussion on cosine similarity from the last section, we evaluate the cosine similarity between the vectors $\mathrm{grad}^{\mathcal{M}_V}_{\pi_V(p)} D(p^*\|\cdot)$ and $d\pi_V\left(\mathrm{grad}^{\mathcal{M}}_p D(q\|\cdot)\right)$ defined in equation (6.29). We sample p in $\mathcal{M}^{(b)}$ according to Jeffrey's prior and plot the re-

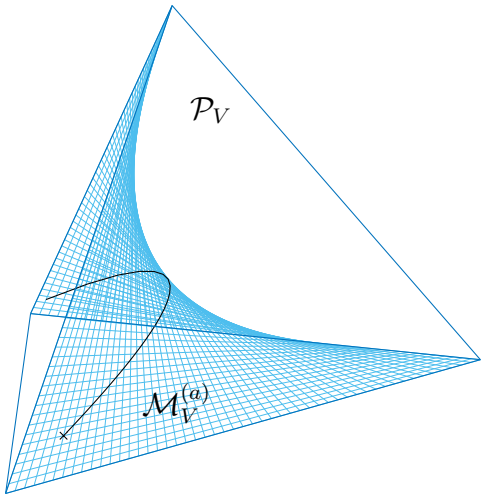


Figure 6.5: The blue grid represents the set of independent probability distributions over two random variables corresponding to the cylindrical model $\mathcal{M}_V^{(a)}$, as a subset of $\mathcal{P}_V$. The cylindrical model $\mathcal{M}^{(a)}$ is defined in (6.27) and depicted in Figure 6.3. The solid black line shows the overlapping trajectories σ_0, σ_1 and σ_2, where σ_0 satisfies $\dot{\sigma}_0(t) = -\text{grad}^{\mathcal{M}_V^{(a)}}_{\sigma_0(t)} D(p^*\|\cdot)$, $\sigma_1 = \pi_V \circ \gamma_1$ is the projection of the negative gradient curve γ_1 satisfying $\dot{\gamma}_1(t) = -\text{grad}^{\mathcal{M}^{(a)}}_{\gamma_1(t)} D(\mathcal{Q}\|\cdot)$ and $\sigma_2 = \pi_V \circ \gamma_2$ is the projection of the negative gradient curve γ_2 satisfying $\dot{\gamma}_2(t) = -\text{grad}^{\mathcal{M}^{(a)}}_{\gamma_2(t)} D(q\|\cdot)$. Theorem 6.2 and 6.3 imply that σ_0, σ_1 and σ_2 are identical.

sults in the histograms in Figure 6.7 (right). As one can see, the points are still close to 1 but on average lower than the cosine similarities between $\text{grad}^{\mathcal{M}_V}_{\pi_V(p)} D(p^*\|\cdot)$ and $d\pi_V\left(\text{grad}^{\mathcal{M}}_p D(\mathcal{Q}\|\cdot)\right)$. Moreover, note that in this case not all points lie above zero anymore. Investigating conditions under which this cosine similarity is negative is an interesting question for future research. According to equation (6.35), minimizing the primary objective function $D(p^*\|\cdot)$ on $\mathcal{M}_V$ is equivalent to minimizing the function $D(q\|\cdot)$ on $\mathcal{M}$ whenever $\mathcal{M}$ is cylindrical. We now show that this is, at the same time, equivalent to maximizing the evidence lower bound. For any $q \in \mathcal{Q}$ and $p \in \mathcal{M}$, we have

$$\begin{aligned} D(q\|p) &= \sum_{x_V, x_H} p^*(x_V)\, q(x_H|x_V) \ln \frac{p^*(x_V)\, q(x_H|x_V)}{p(x_V, x_H)} \\ &= \underbrace{\sum_{x_V} p^*(x_V) \ln p^*(x_V)}_{\leq 0} + \sum_{x_V, x_H} p^*(x_V)\, q(x_H|x_V) \ln \frac{q(x_H|x_V)}{p(x_V, x_H)} \\ &= -\text{ELBO}(q, p) + \text{const.}, \end{aligned}$$

where $\text{ELBO}(q, p)$ is defined by (6.8). Thus, the gradient of the function $D(q\|\cdot)$ will be the same as the gradient of $-\text{ELBO}(q, \cdot)$, because the two functions differ only by a constant. More precisely, we have for all non-singular points p of $\mathcal{M}$

$$\text{grad}^{\mathcal{M}}_p \text{ELBO}(q, \cdot) = -\text{grad}^{\mathcal{M}}_p D(q\|\cdot).$$

As we know that the minimization of $D(p^*\|\cdot)$ is equivalent to the maximization of the evidence, we have the following immediate consequence of Theorem 6.3.

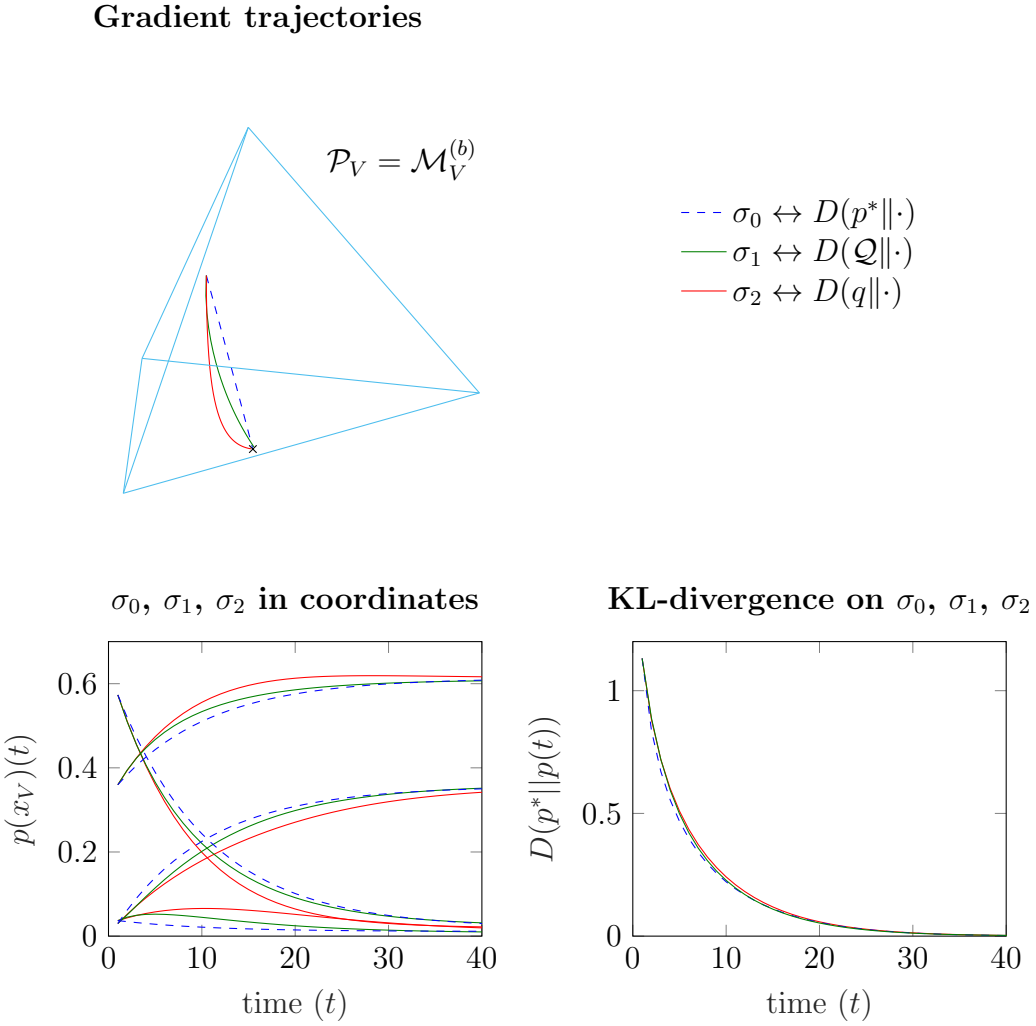


Figure 6.6: These figures study different gradient trajectories on the model $\mathcal{M} = \mathcal{M}^{(b)}$ defined in (6.28) and discussed on pages 116 and 119. The top figure shows the curve σ_0 satisfying $\dot{\sigma}_0(t) = -\text{grad}^{\mathcal{M}_V}_{\sigma_0(t)} D(p^*\|\cdot)$ (dashed-blue), the curve $\sigma_1 = \pi_V \circ \gamma_1$, with γ_1 satisfying $\dot{\gamma}_1(t) = -\text{grad}^{\mathcal{M}}_{\gamma_1(t)} D(\mathcal{Q}\|\cdot)$ (solid green) and the curve $\sigma_2 = \pi_V \circ \gamma_2$, with γ_2 satisfying $\dot{\gamma}_2(t) = -\text{grad}^{\mathcal{M}}_{\gamma_2(t)} D(q\|\cdot)$ (solid red) for a fixed target distribution (cross). The bottom-left figure shows the trajectories σ_0, σ_1 and σ_2 in coordinates as functions of time. The bottom-right figure shows the KL-divergence evaluated on the trajectories σ_0, σ_1 and σ_2.

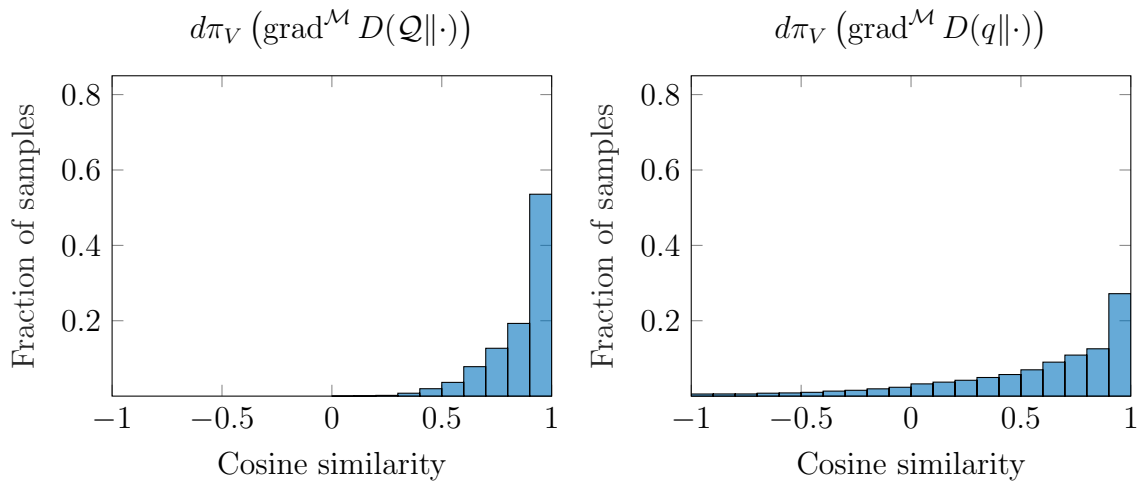


Figure 6.7: Histograms showing the cosine similarity between $d\pi_V\left(\operatorname{grad}^{\mathcal{M}} D(\mathcal{Q}\|\cdot)\right)$ and $\operatorname{grad}^{\mathcal{M}_V} D(p^*\|\cdot)$ (left) and the cosine similarity between $d\pi_V\left(\operatorname{grad}^{\mathcal{M}} D(q\|\cdot)\right)$ and $\operatorname{grad}^{\mathcal{M}_V} D(p^*\|\cdot)$ (right), both with respect to the Fisher-Rao metric.

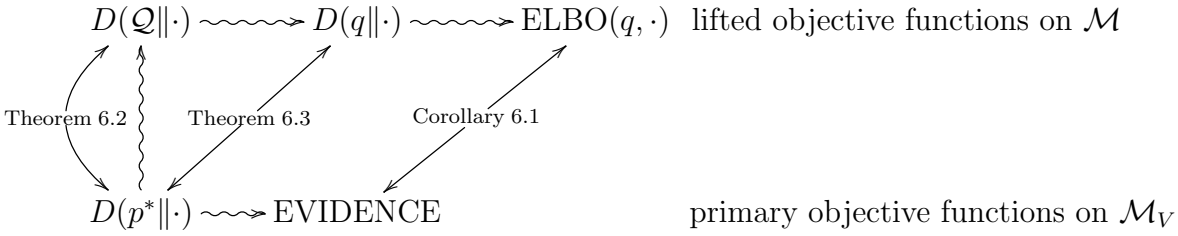
Corollary 6.1. *Let $\mathcal{M}$ be a cylindrical model in $\mathcal{P}$, let $p \in \mathcal{M}$ be admissible, and let $q \in \mathcal{Q}$. Then*

$$d\pi_V\left(\operatorname{grad}_p^{\mathcal{M}} \mathrm{ELBO}(q,\cdot)\right) \;=\; \operatorname{grad}_{\pi_V(p)}^{\mathcal{M}_V} \mathrm{EVIDENCE}. \tag{6.36}$$

In particular, the invariance (6.36) *holds in all points of the maximal model $\mathcal{M} = \mathcal{P}$ where $\mathcal{M}_V = \mathcal{P}_V$.*

The central insight that underlies this result is the following. Under a certain condition, even though the evidence lower bound "lives" in an extended space and provides a bound for the evidence, it is equivalent to it in terms of the natural gradient. The gap does not play any role here. The condition is met, in particular, if we evaluate the gradients on the corresponding maximal models $\mathcal{P}$ and $\mathcal{P}_V$. If we replace these maximal models by $\mathcal{M}$ and $\mathcal{M}_V$, respectively, then we have to impose a quite strong assumption on $\mathcal{M}$ for the equivalence to hold. Therefore, our result has a conceptual rather than a direct methodological value. It states that in the absence of constraints by a model, the evidence lower bound does not alter the original optimization at all. This is remarkable and demonstrates the consistency of the information-geometric structures, which involve the Fisher-Rao metric and the KL-divergence on $\mathcal{P}$ and $\mathcal{P}_V$. Any deviation from the invariance is caused by the restriction of the optimization to a model.

We now summarize the path that we pursued in this article by means of the following diagram:



The overall intention was to relate the maximization of the evidence to the maximization of the evidence lower bound. This has been achieved with Corollary 6.1. To get

there, we have translated the problem to the information-geometric setting, where the primary objective function is given by the KL-divergence $D(p^*\|\cdot)$ defined on $\mathcal{M}_V$. This has then been modified in two steps. In the first step, we replaced the primary objective function by $D(\mathcal{Q}\|\cdot)$ defined on $\mathcal{M}$. The interplay between $D(p^*\|\cdot)$ and $D(\mathcal{Q}\|\cdot)$ was subject of Theorem 6.2. In the second step, we then replaced $D(\mathcal{Q}\|\cdot)$ by $D(q\|\cdot)$ which is also defined on $\mathcal{M}$. The interplay of $D(q\|\cdot)$ and the primary objective function was the subject of Theorem 6.3. In a somewhat parallel story line, we can translate this theorem to a statement about the evidence and its lower bound. This is the subject of Corollary 6.1. Note that the main results of this article are crucially dependent on the information-geometric structures, suggesting that the variational gap does not alter the original objective of learning too much if the corresponding algorithms are based on the natural gradient. The standard Euclidean gradient, on the other hand, depends on the parametrization of the given model and therefore does not yield such parametrization-independent results. An example demonstrating this is presented in the Appendix 6.B. However, empirical case studies in higher dimensions are required for more conclusive statements.

We conclude this section with Remarks 6.1 and 6.2 which outline further research directions as possible continuations of the present work.

Remark 6.1. Theorem 6.3 states that the gap $D(q\|\pi_{\mathcal{Q}}(p))$ has no effect on the learning of the primary objective function $D(p^*\|\cdot)$, if the model is cylindrical. It is remarkable that this statement is independent of the choice of q so that no adjustment of q is required. However, if the model is not cylindrical, then the gap *will* have an effect on the learning. In that case, one could try to adjust q in such a way that the effect is minimal. A natural way to do so is by moving q towards the projection $\pi_{\mathcal{Q}}(p)$. Ideally, the gap will then vanish and the objective function $D(q\|p)$ reduces to $D(\mathcal{Q}\|p)$. However, in a typical learning scenario, q is constrained to a so-called *recognition model* $\mathcal{Q}'$ which is much smaller dimensional than the data manifold $\mathcal{Q}$ defined by (6.4). Denoting a minimizer of $D(q\|p)$ with respect to $q \in \mathcal{Q}'$ by $\pi_{\mathcal{Q}'}(p)$, we consider the *residual gap*

$$D(\pi_{\mathcal{Q}'}(p)\|\pi_{\mathcal{Q}}(p)).$$

This gap vanishes for all $p \in \mathcal{M}$, if $\mathcal{Q}'$ is sufficient in the sense that it already contains all projections of points $p \in \mathcal{M}$ onto the maximally possible recognition model, the data manifold $\mathcal{Q}$, that is

$$\mathcal{Q}' \supseteq \{\pi_{\mathcal{Q}}(p) : p \in \mathcal{M}\}.$$

In principle, such a recognition model $\mathcal{Q}'$ has the dimensionality of $\mathcal{M}$. However, representing it in terms of a graphical model typically leads to a blowup of dimensionality (Oostrum et al., 2024; Webb et al., 2018), which forces us to consider smaller recognition models $\mathcal{Q}'$ for learning with a non-vanishing and even large residual gap. The present work suggests, on the other hand, that even in this case, the effect on the learning of the primary objective function $D(p^*\|\cdot)$ can be rather small, if the model is close to being cylindrical. That opens up a way to define concise recognition models $\mathcal{Q}'$ with limited perturbation of the primary optimization problem.

Remark 6.2. In this remark, we outline a way to extend the analysis of the present article to the general situation where the model $\mathcal{M}$ is not assumed to be cylindrical. (Note that all our results for non-cylindrical models refer to a particular example and

are numerical in nature.) For that, we require the notion of a *cylindrical extension* $\widetilde{\mathcal{M}}$ of $\mathcal{M}$. This is a submodel of $\mathcal{P}$ that satisfies the following conditions:

$$\text{(a)} \quad \mathcal{M} \subseteq \widetilde{\mathcal{M}}, \qquad \text{(b)} \quad \pi_V(\mathcal{M}) = \pi_V(\widetilde{\mathcal{M}}), \qquad \text{(c)} \quad \widetilde{\mathcal{M}} \text{ is cylindrical.}$$

It is easy to show that any model $\mathcal{M}$ in $\mathcal{P}$ has a cylindrical extension. For instance, we can simply consider the set

$$\widetilde{\mathcal{M}} \;=\; \{p \in \mathcal{P} \;:\; \pi_V(p) \in \mathcal{M}_V\}\,,$$

which is the maximal cylindrical extension of $\mathcal{M}$ with respect to set inclusion. We can now apply Theorem 6.3 to a cylindrical extension and obtain

$$\begin{aligned} d\pi_V\left(\operatorname{grad}_p^{\widetilde{\mathcal{M}}} D(q\|\cdot)\right) &= d\pi_V\left(\operatorname{grad}_p^{\mathcal{M}} D(q\|\cdot)\right) + d\pi_V\left(\operatorname{grad}_p^{\perp} D(q\|\cdot)\right) \\ &= \operatorname{grad}_{\pi_V(p)}^{\mathcal{M}_V} D(p^*\|\cdot), \end{aligned}$$

where $\operatorname{grad}_p^{\perp} D(q\|\cdot)$ denotes the projection of $\operatorname{grad}_p^{\widetilde{\mathcal{M}}} D(q\|\cdot)$ onto the orthogonal complement of $T_p\mathcal{M}$ in $T_p\widetilde{\mathcal{M}}$. This finally gives us the following generalization of (6.35):

$$d\pi_V\left(\operatorname{grad}_p^{\mathcal{M}} D(q\|\cdot)\right) \;=\; \operatorname{grad}_{\pi_V(p)}^{\mathcal{M}_V} D(p^*\|\cdot) - d\pi_V\left(\operatorname{grad}_p^{\perp} D(q\|\cdot)\right). \tag{6.37}$$

Equation (6.37) suggests a way to establish a relation between the natural gradient of $D(q\|\cdot)$ and the natural gradient of the primary objective function $D(p^*\|\cdot)$ in the general case, with no restriction to cylindrical models.

6.6 Simplification of the Learning in the Extended Space

In this article, we have studied the optimization of a primary objective function defined on a model $\mathcal{M}_V$, where V denotes the set of visible units. We have compared this optimization with a corresponding optimization on an extended model $\mathcal{M}$ which incorporates hidden units. More precisely, the former objective function on $\mathcal{M}_V$ is the mean evidence, whereas the latter is given by the evidence lower bound defined on $\mathcal{M}$. We have stated that the replacement of the primary objective function by a lower bound can greatly simplify the optimization process. In this section, we are now going to provide an instance of this simplification in the context of Bayesian graphical models. Such a model is defined in terms of a directed acyclic graph $G = (N, E)$ with node set $N = V \cup H$. The points of the corresponding Bayesian graphical model, which we denote by $\mathcal{P}^G$, are those probability distributions p in $\mathcal{P}$ that factorize according to G, that is

$$p(x) \;=\; \prod_{s \in N} p(x_s | x_{\mathrm{pa}(s)}).$$

Here, $\mathrm{pa}(s)$ denotes the parents of unit s, those units $r \in N$ for which $(r, s) \in E$. Typically, each conditional probability distribution $p(x_s|x_{\mathrm{pa}(s)})$, which we interpret as the local generative mechanism of unit s, is parametrized in terms of a local parameter vector $\theta_s = (\theta_{s,1}, \ldots, \theta_{s,d_s})$, indicated by $p(x_s|x_{\mathrm{pa}(s)}; \theta_s)$. Concatenating all the parameter vectors θ_s to one vector $\theta = (\theta_s)_{s \in N}$ of size $d = \sum_{s \in N} d_s$, the overall probability distribution is parametrized as

$$p_\theta(x) \;:=\; p(x; \theta) \;:=\; \prod_{s \in N} p(x_s | x_{\mathrm{pa}(s)}; \theta_s). \tag{6.38}$$

The model $\mathcal{M}$, given by the collection $p(\cdot;\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^d$, is a submodel of $\mathcal{P}^G$. Whenever referring to a submodel of a Bayesian graphical model in the following, we mean this kind of a submodel without explicitly mentioning it. The product structure (6.38) implies a number of simplifications which have been discussed in (Ay, 2020). In this article, we add a somewhat simple but illuminating point to this discussion.

In order to compute the gradient of an objective function in a non-singular point p of $\mathcal{M}$, we need to project the corresponding gradient in $\mathcal{P}$ onto the tangent space $T_p\mathcal{M}$ of $\mathcal{M}$ in p, in terms of the orthogonal projection Π_p. We have encountered this method of determining the gradient several times in this article. The projection Π_p will be particularly simple, if the tangent space decomposes into orthogonal lower-dimensional spaces. We are now going to highlight this structure for any submodel $\mathcal{M}$ of a Bayesian graphical model. For $\theta \in \Theta$, the tangent space of $\mathcal{M}$ in p_θ, denoted by $T_{p_\theta}\mathcal{M}$, is typically expressed in terms of the derivatives

$$\partial_{s,k}(\theta) \;=\; \sum_x \partial_{s,k}(x;\theta)\,\delta^x,$$

with

$$\begin{aligned}\partial_{s,k}(x;\theta) &:= \frac{\partial}{\partial\theta_{s,k}}\,p(x;\theta)\\ &= p(x;\theta)\,\frac{\partial}{\partial\theta_{s,k}}\,\ln p(x;\theta)\\ &= p(x;\theta)\,\frac{\partial}{\partial\theta_{s,k}}\,\ln p(x_s|x_{\mathrm{pa}(s)};\theta_s). \end{aligned} \tag{6.39}$$

The vectors $\partial_{s,k}(\theta)$, $s \in N$, $k = 1,\dots,d_s$, are clearly contained in $T_{p_\theta}\mathcal{M}$ but in general they need not to span $T_{p_\theta}\mathcal{M}$. To express all elements of the tangent space, we assume that the parametrization $\theta \mapsto p_\theta = p(\cdot;\theta)$ is *proper* in the sense that the vectors $\partial_{s,k}(\theta)$, $s \in N$, $k = 1,\dots,d_s$, span $T_\theta\mathcal{M}$. On the other hand, assuming a proper parametrization, we cannot expect these vectors to form a basis of $T_{p_\theta}\mathcal{M}$ as they do not have to be linearly inedependent. For a submodel $\mathcal{M}$ of a Bayesian graphical model, however, these vectors give rise to a natural orthogonal decomposition, which simplifies the projection onto the tangent space $T_{p_\theta}\mathcal{M}$.

Proposition 6.2. *Let $\mathcal{M}$ be a submodel of a Bayesian graphical model, parametrized by θ (not necessarily by a proper parametrization). Then, for $s \neq t$, $1 \leq k \leq d_s$, $1 \leq l \leq d_t$, we have*

$$g^{\mathrm{FR}}_\theta\left(\partial_{s,k}(\theta), \partial_{t,l}(\theta)\right) \;=\; 0.$$

Assuming that the parametrization $\theta \mapsto p_\theta$ is proper, we obtain an orthogonal decomposition of the tangent space $T_{p_\theta}\mathcal{M}$ into the subspaces

$$T^{(s)}_\theta\mathcal{M} \;:=\; \mathrm{span}\{\partial_{s,k}(\theta) \;:\; k = 1,\dots,d_s\}, \quad s \in N.$$

See Appendix 6.C for a proof.

Acknowledgements. NA and JvO acknowledge the support of the Deutsche Forschungsgemeinschaft Priority Programme "The Active Self" (SPP 2134).

6.A Examples of Cylindrical and Non-Cylindrical Models

Example 1: The Independence Model

Let us consider the setting in which $\mathcal{P}$ is the set of distributions over two binary nodes s and t. The state space is given by

$$\begin{aligned} \mathsf{X} &= \mathsf{X}_s \times \mathsf{X}_t, \\ \mathsf{X}_s &= \mathsf{X}_t = \{0,1\}, \end{aligned}$$

and we let $X_r : \mathsf{X} \to \mathsf{X}_r, r \in \{s,t\}$ be the projections. The marginalization map is given by

$$\begin{aligned} &\pi_V \colon \mathcal{P} \to \mathcal{P}_V, \\ &p(x_s,x_t) \mapsto p(x_t) = \sum_{x_s} p(x_s,x_t), \end{aligned} \tag{6.40}$$

and its differential

$$\begin{aligned} d\pi_V \colon T_p\mathcal{P} &\to T_{\pi_V(p)}\mathcal{P}_V, \\ A = \sum_{x_s,x_t} A(x_s,x_t)\delta^{(x_s,x_t)} &\mapsto \sum_{x_t}\left(\sum_{x_s} A(x_s,x_t)\right)\delta^{x_t}. \end{aligned}$$

The vertical and horizontal spaces are given by

$$\begin{aligned} \mathcal{V}_p &= \ker d\pi_V \\ &= \left\{ A \in T_p\mathcal{P} : \sum_{x_s} A(x_s,x_t) = 0, x_t \in \mathsf{X}_t \right\} \\ &= \operatorname{span}\left\{ \delta^{(0,0)} - \delta^{(1,0)}, \delta^{(0,1)} - \delta^{(1,1)} \right\}, \\ \mathcal{H}_p &= (\ker d\pi_V)^{\perp} \\ &= \left\{ A \in T_p\mathcal{P} : \frac{A(0,x_t)}{p(0,x_t)} - \frac{A(1,x_t)}{p(1,x_t)} = 0, x_t \in \mathsf{X}_t \right\} \\ &= \operatorname{span}\left\{ p(0,0)\delta^{(0,0)} + p(1,0)\delta^{(1,0)}, p(0,1)\delta^{(0,1)} + p(1,1)\delta^{(1,1)} \right\}. \end{aligned} \tag{6.41}$$

Now we let the model be the independence model, given by

$$\mathcal{M} = \{ p \in \mathcal{P} : p(x_s,x_t) = p(x_s)p(x_t) \}.$$

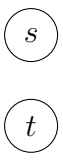


Figure 6.8: Graph G

This model factorizes over the graph depicted in Figure 6.8 and can be parameterized as follows:

$$\begin{aligned} p(X_s = 1; \theta) &= \theta_s, \\ p(X_t = 1; \theta) &= \theta_t. \end{aligned}$$

This parametrization gives

$$p_\theta = (1-\theta_s)(1-\theta_t)\delta^{(0,0)} + (1-\theta_s)\theta_t\delta^{(0,1)} + \theta_s(1-\theta_t)\delta^{(1,0)} + \theta_s\theta_t\delta^{(1,1)}.$$

The tangent space $T_{p_\theta}\mathcal{M}$ is spanned by the parameter tangent vectors given by

$$\begin{aligned} \partial_s(\theta) &= -(1-\theta_t)\delta^{(0,0)} - \theta_t\delta^{(0,1)} + (1-\theta_t)\delta^{(1,0)} + \theta_t\delta^{(1,1)}, \\ \partial_t(\theta) &= -(1-\theta_s)\delta^{(0,0)} + (1-\theta_s)\delta^{(0,1)} - \theta_s\delta^{(1,0)} + \theta_s\delta^{(1,1)}. \end{aligned}$$

Note that

$$\begin{aligned} \partial_s(\theta) &= -(1-\theta_t)\left(\delta^{(0,0)} - \delta^{(1,0)}\right) - \theta_t\left(\delta^{(0,1)} - \delta^{(1,1)}\right) \in \mathcal{V}_{p_\theta}, \\ \partial_t(\theta) &= -\frac{1}{(1-\theta_t)}\left(p_\theta(0,0)\delta^{(0,0)} + p_\theta(1,0)\delta^{(1,0)}\right) + \frac{1}{\theta_t}\left(p_\theta(0,1)\delta^{(0,1)} + p_\theta(1,1)\delta^{(1,1)}\right) \in \mathcal{H}_{p_\theta}. \end{aligned}$$

Therefore, this model is cylindrical.

Example 2: Non-Cylindrical Two-Node Model

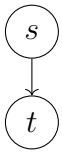


Figure 6.9: Graph G

For the same $\mathcal{P}$ and the same π_V as in Example 1 (see equation (6.40)), let us fix a distribution $\bar{p}(x_s)$ and consider the following model:

$$\mathcal{M} = \{p \in \mathcal{P} : p(x_s, x_t) = \bar{p}(x_s)p(x_t|x_s)\},$$

which factorizes over the graph from Figure 6.9. This model can be parametrized by

$$\begin{aligned} p(X_t = 1|X_s = 0; \theta) &= \theta_{t,1}, \\ p(X_t = 1|X_s = 1; \theta) &= \theta_{t,2}. \end{aligned}$$

This parametrization gives

$$p_\theta = (1-\theta_{t,1})\bar{p}(0)\delta^{(0,0)} + \theta_{t,1}\bar{p}(0)\delta^{(0,1)} + (1-\theta_{t,2})\bar{p}(1)\delta^{(1,0)} + \theta_{t,2}\bar{p}(1)\delta^{(1,1)}.$$

The parameter tangent vectors of $T_{p_\theta}\mathcal{M}$ are given by

$$\begin{aligned} \partial_1(\theta) &= -\bar{p}(0)\delta^{(0,0)} + \bar{p}(0)\delta^{(0,1)}, \\ \partial_2(\theta) &= -\bar{p}(1)\delta^{(1,0)} + \bar{p}(1)\delta^{(1,1)}. \end{aligned}$$

For the intersection of $T_{p_\theta}\mathcal{M}$ with $\mathcal{V}_{p_\theta}$ we have

$$T_{p_\theta}\mathcal{M} \cap \mathcal{V}_{p_\theta} \;=\; \operatorname{span}\left\{ \frac{1}{\bar{p}(0)}\partial_1(\theta) - \frac{1}{\bar{p}(1)}\partial_2(\theta) \right\}.$$

Note that this space is only one-dimensional. In order for $\mathcal{M}$ to be cylindrical, we would therefore need that the intersection of $T_{p_\theta}\mathcal{M}$ with $\mathcal{H}_{p_\theta}$ is non-trivial. Let us assume by contradiction that there exists α, β such that $\alpha\partial_1(\theta) + \beta\partial_2(\theta) \in \mathcal{H}_{p_\theta}$. WLOG assume $\alpha = 1$. Using the definition of $\mathcal{H}_{p_\theta}$ from equation (6.41), we get the following conditions:

$$\frac{-\bar{p}(0)}{p_\theta(0,0)} \;=\; \beta\frac{-\bar{p}(1)}{p_\theta(1,0)}$$

and,

$$\frac{\bar{p}(0)}{p_\theta(0,1)} \;=\; \beta\frac{\bar{p}(1)}{p_\theta(1,1)}.$$

Working out these conditions gives $\beta = \frac{1-\theta_{t,2}}{1-\theta_{t,1}}$ and $\beta = \frac{\theta_{t,2}}{\theta_{t,1}}$ respectively. Therefore, we conclude that this model is only cylindrical in the points where $\theta_{t,1} = \theta_{t,2}$ which are exactly the points for which s and t are independent, and is in general not cylindrical.

Example 3: Non-Cylindrical Three-Node Model

Now, let $\mathcal{P}$ be the space of probability measures over the sample space X given by

$$\begin{aligned} \mathsf{X} \;&=\; \mathsf{X}_s \times \mathsf{X}_{t_1} \times \mathsf{X}_{t_2}, \\ \mathsf{X}_s \;&=\; \mathsf{X}_{t_1} \;=\; \mathsf{X}_{t_2} \;=\; \{0,1\}. \end{aligned}$$

We let $X_r : \mathsf{X} \to \mathsf{X}_r, r \in \{s, t_1, t_2\}$ be the projections. The marginalization map is given by

$$\begin{aligned} \pi_V \colon \mathcal{P} \;&\to\; \mathcal{P}_V, \\ p(x_s, x_{t_1}, x_{t_2}) \;&\mapsto\; p(x_{t_1}, x_{t_2}) = \sum_{x_s} p(x_s, x_{t_1}, x_{t_2}), \end{aligned}$$

and its differential

$$d\pi_V \colon T_p\mathcal{P} \to T_{\pi_V(p)}\mathcal{P}_V,$$

$$A = \sum_{x_s, x_{t_1}, x_{t_2}} A(x_s, x_{t_1}, x_{t_2})\delta^{(x_s, x_{t_1}, x_{t_2})} \mapsto \sum_{x_{t_1}, x_{t_2}} \left(\sum_{x_s} A(x_s, x_{t_1}, x_{t_2}) \right) \delta^{(x_{t_1}, x_{t_2})}. \tag{6.42}$$

The vertical and horizontal spaces are given by

$$\begin{aligned} \mathcal{V}_p \;&=\; \ker d\pi_V = \left\{ A \in T_p\mathcal{P} : \sum_{x_s} A(x_s, x_{t_1}, x_{t_2}) = 0 \right\}, \\ \mathcal{H}_p \;&=\; (\ker d\pi_V)^\perp = \left\{ A \in T_p\mathcal{P} : \sum_{x_s} \frac{A(x_s, x_{t_1}, x_{t_2})}{p(x_s, x_{t_1}, x_{t_2})}(-1)^{x_s} = 0 \right\}. \end{aligned}$$

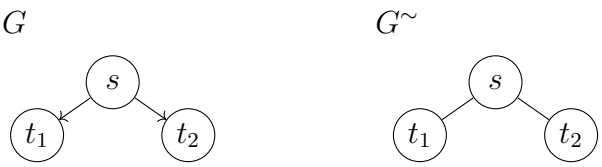


Figure 6.10: (left) Directed graph G; (right) Undirected graph $G^{\sim}$.

Now we consider the model given by

$$\mathcal{M} \;=\; \{p \in \mathcal{P} : p(x_s, x_{t_1}, x_{t_2}) = p(x_s)p(x_{t_1}|x_s)p(x_{t_2}|x_s)\}.$$

Note that this model is both equal to the Bayesian graphical model of distributions that factorise over the graph G, and equal to the distributions corresponding to the Boltzmann machine with the undirected graph $G^{\sim}$, both in Figure 6.10. The model can be parameterized as follows:

$$\begin{aligned} p(X_s = 1;\theta) &= \theta_s, \\ p(X_{t_1} = 1|X_s = 0;\theta) &= \theta_{t_1,1}, \\ p(X_{t_1} = 1|X_s = 1;\theta) &= \theta_{t_1,2}, \\ p(X_{t_2} = 1|X_s = 0;\theta) &= \theta_{t_2,1}, \\ p(X_{t_2} = 1|X_s = 1;\theta) &= \theta_{t_2,2}. \end{aligned}$$

As in the previous examples, this parametrization gives

$$\begin{aligned} p_\theta \;=\; & (1-\theta_s)(1-\theta_{t_1,1})(1-\theta_{t_2,1})\delta^{(0,0,0)} + (1-\theta_s)(1-\theta_{t_1,1})\theta_{t_2,1}\delta^{(0,0,1)} \\ & + (1-\theta_s)\theta_{t_1,1}(1-\theta_{t_2,1})\delta^{(0,1,0)} + (1-\theta_s)\theta_{t_1,1}\theta_{t_2,1}\delta^{(0,1,1)} \\ & + \theta_s(1-\theta_{t_1,2})(1-\theta_{t_2,2})\delta^{(1,0,0)} + \theta_s(1-\theta_{t_1,2})\theta_{t_2,2}\delta^{(1,0,1)} \\ & + \theta_s\theta_{t_1,2}(1-\theta_{t_2,2})\delta^{(1,1,0)} + \theta_s\theta_{t_1,2}\theta_{t_2,2}\delta^{(1,1,1)}. \end{aligned}$$

To simplify the ensuing long expressions, we next identify the space of signed measures on X with $\mathbb{R}^8$, where we use the following enumeration of the sample space X:

$$((0,0,0),(0,0,1),(0,1,0),(0,1,1),(1,0,0),(1,0,1),(1,1,0),(1,1,1)).$$

This gives for example

$$\delta^{(0,1,0)} \;=\; \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad \text{and} \qquad p_\theta \;=\; \begin{bmatrix} (1-\theta_s)(1-\theta_{t_1,1})(1-\theta_{t_2,1}) \\ (1-\theta_s)(1-\theta_{t_1,1})\theta_{t_2,1} \\ (1-\theta_s)\theta_{t_1,1}(1-\theta_{t_2,1}) \\ (1-\theta_s)\theta_{t_1,1}\theta_{t_2,1} \\ \theta_s(1-\theta_{t_1,2})(1-\theta_{t_2,2}) \\ \theta_s(1-\theta_{t_1,2})\theta_{t_2,2} \\ \theta_s\theta_{t_1,2}(1-\theta_{t_2,2}) \\ \theta_s\theta_{t_1,2}\theta_{t_2,2} \end{bmatrix}.$$

The parameter tangent vectors of $T_p\mathcal{M}$ can similarly be identified as

$$\partial_s(\theta) = \begin{bmatrix} -(1-\theta_{t_1,1})(1-\theta_{t_2,1}) \\ -(1-\theta_{t_1,1})\theta_{t_2,1} \\ -\theta_{t_1,1}(1-\theta_{t_2,1}) \\ -\theta_{t_1,1}\theta_{t_2,1} \\ (1-\theta_{t_1,2})(1-\theta_{t_2,2}) \\ (1-\theta_{t_1,2})\theta_{t_2,2} \\ \theta_{t_1,2}(1-\theta_{t_2,2}) \\ \theta_{t_1,2}\theta_{t_2,2} \end{bmatrix},$$

$$\partial_{t_1,1}(\theta) = \begin{bmatrix} -(1-\theta_s)(1-\theta_{t_2,1}) \\ -(1-\theta_s)\theta_{t_2,1} \\ (1-\theta_s)(1-\theta_{t_2,1}) \\ (1-\theta_s)\theta_{t_2,1} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \partial_{t_1,2}(\theta) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -\theta_s(1-\theta_{t_2,2}) \\ -\theta_s\theta_{t_2,2} \\ \theta_s(1-\theta_{t_2,2}) \\ \theta_s\theta_{t_2,2} \end{bmatrix},$$

$$\partial_{t_2,1}(\theta) = \begin{bmatrix} -(1-\theta_s)(1-\theta_{t_1,1}) \\ (1-\theta_s)(1-\theta_{t_1,1}) \\ -(1-\theta_s)\theta_{t_1,1} \\ (1-\theta_s)\theta_{t_1,1} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \partial_{t_2,2}(\theta) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -\theta_s(1-\theta_{t_1,2}) \\ \theta_s(1-\theta_{t_1,2}) \\ -\theta_s\theta_{t_1,2} \\ \theta_s\theta_{t_1,2} \end{bmatrix}.$$

We let $B_{p_\theta} \in \mathbb{R}^{8\times 5}$ be the matrix with these parameter vectors as columns, i.e.,

$$B_{p_\theta} = \begin{bmatrix} | & | & | & | & | \\ \partial_s(\theta) & \partial_{t_1,1}(\theta) & \partial_{t_1,2}(\theta) & \partial_{t_2,1}(\theta) & \partial_{t_2,2}(\theta) \\ | & | & | & | & | \end{bmatrix}.$$

In the same spirit as above, the space of signed measures on $\mathsf{X}_{t_1} \times \mathsf{X}_{t_2}$ can be identified with $\mathbb{R}^4$, where we use enumerate the sample space $\mathsf{X}_{t_1} \times \mathsf{X}_{t_2}$ as $((0,0),(0,1),(1,0),(1,1))$. For example, this gives

$$\delta^{(1,0)} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}.$$

With the identification of $p \in \mathcal{P}$ with vectors in $\mathbb{R}^8$ and the identification of $p_V \in \mathcal{P}_V$ with vectors in $\mathbb{R}^4$, we can identify the map $d\pi_p$ defined in (6.42) with a matrix $J \in \mathbb{R}^{4\times 8}$ given by

$$J = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Note that

$$\dim\left(T_{p_\theta}\mathcal{M} \cap \mathcal{V}_{p_\theta}\right) \;=\; \dim\left(\ker\, JB_{p_\theta}\right).$$

Similarly, one can derive

$$\dim\left(T_{p_\theta}\mathcal{M} \cap \mathcal{H}_{p_\theta}\right) \;=\; \dim\left(\ker\, \tilde{J}G_{p_\theta}B_{p_\theta}\right),$$

where

$$\tilde{J} \;=\; \begin{bmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{bmatrix}$$

and G_{p_θ} is the matrix representative of the Fisher-Rao metric at p_θ, given by

$$G_{p_\theta} \;=\; \begin{bmatrix} 1/p_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/p_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/p_3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/p_4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/p_5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/p_6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/p_7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/p_8 \end{bmatrix},$$

with $p_i = p_\theta(x_i)$, where x_i is the ith element of the sample space X.

In order to show that this model is not cylindrical, we only have to show this for one specific point. We choose the point $\theta_s = \theta_{t_1,1} = \theta_{t_2,1} = 1/2, \theta_{t_1,2} = \theta_{t_2,2} = 1/3$.

For this choice of θ, B_{p_θ} becomes

$$B_{p_\theta} \;=\; \begin{bmatrix} -1/4 & -1/4 & 0 & -1/4 & 0 \\ -1/4 & -1/4 & 0 & 1/4 & 0 \\ -1/4 & 1/4 & 0 & -1/4 & 0 \\ -1/4 & 1/4 & 0 & 1/4 & 0 \\ 4/9 & 0 & -1/3 & 0 & -1/3 \\ 2/9 & 0 & -1/6 & 0 & 1/3 \\ 2/9 & 0 & 1/3 & 0 & -1/6 \\ 1/9 & 0 & 1/6 & 0 & 1/6 \end{bmatrix}.$$

It can be verified that the space $\ker\, JB_{p_\theta}$ is spanned by the following vectors:

$$\begin{bmatrix} 3 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix},$$

and is therefore two-dimensional.

Similarly, the space $\ker \tilde{J} G_{p_\theta} B_{p_\theta}$ is spanned by

$$\begin{bmatrix} -3/16 \\ 9/8 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -3/16 \\ 0 \\ 0 \\ 9/8 \\ 1 \end{bmatrix},$$

and is therefore also two-dimensional. This means that $(T_{p_\theta}\mathcal{M} \cap \mathcal{V}_{p_\theta}) \oplus (T_{p_\theta}\mathcal{M} \cap \mathcal{H}_{p_\theta})$ is four-dimensional and therefore unequal to $T_{p_\theta}\mathcal{M}$ which is five-dimensional. We therefore conclude that $\mathcal{M}$ is not cylindrical.

6.B The Natural Versus the Euclidean Gradient of the Variational Gap

Let us again consider the setting in which $\mathcal{P}$ is the set of distributions over two binary nodes s and t, where t is the visible and s is the hidden node. The state space is given by

$$\begin{aligned} \mathsf{X} &= \mathsf{X}_s \times \mathsf{X}_t, \\ \mathsf{X}_s &= \mathsf{X}_t = \{0, 1\}. \end{aligned}$$

Thus, we have the four states $(0,0)$, $(1,0)$, $(0,1)$, and $(1,1)$ and the corresponding Dirac measures $\delta^{(0,0)}$, $\delta^{(1,0)}$, $\delta^{(0,1)}$, and $\delta^{(1,1)}$. In what follows, we parametrize $\mathcal{P}$ in terms of

$$\varphi : \mathbb{R}^3 \ni \theta = (\theta_1, \theta_2, \theta_3) \mapsto \theta_1 \delta^{(0,0)} + \theta_2 \delta^{(1,0)} + \theta_3 \delta^{(0,1)} + (1 - \theta_1 - \theta_2 - \theta_3) \delta^{(1,1)} \in \mathcal{P},$$

where we assume $\theta_1, \theta_2, \theta_3 > 0$ and $\theta_1 + \theta_2 + \theta_3 < 1$. To simplify the notation, we abbreviate $1 - (\theta_1 + \theta_2 + \theta_3)$ by θ_4. The tangent space of $\mathcal{P}$ in $\varphi(\theta)$ is spanned by the basis

$$\begin{aligned} \partial_1(\theta) &= \frac{\partial \varphi}{\partial \theta_1}(\theta) = \delta^{(0,0)} - \delta^{(1,1)}, \\ \partial_2(\theta) &= \frac{\partial \varphi}{\partial \theta_2}(\theta) = \delta^{(1,0)} - \delta^{(1,1)}, \\ \partial_3(\theta) &= \frac{\partial \varphi}{\partial \theta_3}(\theta) = \delta^{(0,1)} - \delta^{(1,1)}. \end{aligned}$$

Application of the differential (6.14) of the marginalization map $\pi_V : \mathcal{P} \mapsto \mathcal{P}_V = \mathcal{P}_{\{t\}}$ gives us

$$\begin{aligned} d\pi_V(\partial_1(\theta)) &= \delta^0 - \delta^1, \\ d\pi_V(\partial_2(\theta)) &= 0, \\ d\pi_V(\partial_3(\theta)) &= \delta^0 - \delta^1. \end{aligned}$$

Here, δ^0 and δ^1 denote the Dirac measures of the states 0 and 1 of the visible node t. The Fisher information matrix $G(\theta)$ with components $g^{\mathrm{FR}}(\partial_i(\theta), \partial_j(\theta))$ is given as

$$G(\theta) \;=\; \frac{1}{\theta_4}\begin{bmatrix} \frac{\theta_4}{\theta_1}+1 & 1 & 1 \\ 1 & \frac{\theta_4}{\theta_2}+1 & 1 \\ 1 & 1 & \frac{\theta_4}{\theta_3}+1 \end{bmatrix},$$

with inverse

$$G^{-1}(\theta) \;=\; \begin{bmatrix} \theta_1(1-\theta_1) & -\theta_1\theta_2 & -\theta_1\theta_3 \\ -\theta_2\theta_1 & \theta_2(1-\theta_2) & -\theta_2\theta_3 \\ -\theta_3\theta_1 & -\theta_3\theta_2 & \theta_3(1-\theta_3) \end{bmatrix}. \tag{6.43}$$

Given a differentiable function $\mathcal{L} : \mathcal{P} \to \mathbb{R}$, we set

$$\nabla\mathcal{L}(\theta) :=\; \begin{bmatrix} \dfrac{\partial \mathcal{L}\circ\varphi}{\partial\theta_1}(\theta) \\ \dfrac{\partial \mathcal{L}\circ\varphi}{\partial\theta_2}(\theta) \\ \dfrac{\partial \mathcal{L}\circ\varphi}{\partial\theta_3}(\theta) \end{bmatrix}$$

and

$$\widetilde{\nabla}\mathcal{L}(\theta) :=\; G^{-1}(\theta)\,\nabla\mathcal{L}(\theta).$$

The Euclidean gradient with respect to the standard inner product in $\mathbb{R}^3$ is given by

$$\sum_{i=1}^{3} [\nabla\mathcal{L}(\theta)]_i \;\partial_i(\theta),$$

whereas the natural gradient involves the Fisher information matrix:

$$\sum_{i=1}^{3} \left[\widetilde{\nabla}\mathcal{L}(\theta)\right]_i \;\partial_i(\theta).$$

Learning based on the Euclidean gradient ascent method follows the update rule

$$\theta_{m+1} \;=\; \theta_m + \varepsilon\cdot\nabla\mathcal{L}(\theta_m), \qquad m = 0, 1, 2, \ldots, \tag{6.44}$$

whereas the natural gradient method suggests

$$\theta_{m+1} \;=\; \theta_m + \varepsilon\cdot\widetilde{\nabla}\mathcal{L}(\theta_m), \qquad m = 0, 1, 2, \ldots. \tag{6.45}$$

One could apply these iteration rules, for instance, to maximize the evidence and its lower bound, respectively. This article suggests that the replacement of the evidence by its lower bound will have a less "visible" effect if we use the natural gradient iteration rule (6.45) in comparison with the Euclidean iteration rule (6.44). This can be formally studied by mapping the gradient of the variational gap via the differential $d\pi_V$. In what

follows, we evaluate the Euclidean as well as the natural gradient of $\mathrm{GAP}_q := \mathrm{GAP}(q,\cdot)$, defined by (6.9). After some straightforward calculations, we obtain

$$\begin{aligned}
\left[\nabla \mathrm{GAP}_q(\theta)\right]_1 &= \frac{p^*(0)}{\theta_1+\theta_3} - \frac{p^*(1)}{\theta_2+\theta_4} - \frac{p^*(0)q(0|0)}{\theta_1} + \frac{p^*(1)q(1|1)}{\theta_4},\\
\left[\nabla \mathrm{GAP}_q(\theta)\right]_2 &= -\frac{p^*(1)q(0|1)}{\theta_2} + \frac{p^*(1)q(1|1)}{\theta_4},\\
\left[\nabla \mathrm{GAP}_q(\theta)\right]_3 &= \frac{p^*(0)}{\theta_1+\theta_3} - \frac{p^*(1)}{\theta_2+\theta_4} - \frac{p^*(0)q(1|0)}{\theta_3} + \frac{p^*(1)q(1|1)}{\theta_4}.
\end{aligned}$$

With the inverse of the Fisher information matrix, (6.43), this yields

$$\begin{aligned}
\left[\widetilde{\nabla} \mathrm{GAP}_q(\theta)\right]_1 &= \left(\frac{p^*(0)}{\theta_1+\theta_3} - \frac{p^*(1)}{\theta_2+\theta_4}\right)\theta_1(\theta_2+\theta_4) - p^*(0)q(0|0) + \theta_1,\\
\left[\widetilde{\nabla} \mathrm{GAP}_q(\theta)\right]_2 &= -\left(\frac{p^*(0)}{\theta_1+\theta_3} - \frac{p^*(1)}{\theta_2+\theta_4}\right)\theta_2(\theta_1+\theta_3) - p^*(1)q(0|1) + \theta_2,\\
\left[\widetilde{\nabla} \mathrm{GAP}_q(\theta)\right]_3 &= \left(\frac{p^*(0)}{\theta_1+\theta_3} - \frac{p^*(1)}{\theta_2+\theta_4}\right)\theta_3(\theta_2+\theta_4) - p^*(0)q(1|0) + \theta_3.
\end{aligned}$$

Mapping the Euclidean gradient with $d\pi_V$ yields

$$\begin{aligned}
d\pi_V\left(\sum_{i=1}^3 \left[\nabla \mathrm{GAP}_q(\theta)\right]_i \partial_i(\theta)\right) &= \sum_{i=1}^3 \left[\nabla \mathrm{GAP}_q(\theta)\right]_i d\pi_V(\partial_i(\theta))\\
&= \left(\left[\nabla \mathrm{GAP}_q(\theta)\right]_1 + \left[\nabla \mathrm{GAP}_q(\theta)\right]_3\right)(\delta^0 - \delta^1).
\end{aligned}$$

The same formula holds for the natural gradient where we simply replace ∇ by $\widetilde{\nabla}$. Thus, in both cases we can analyze whether the image of the gradient under $d\pi_V$ vanishes by simply adding the respective first and third components. Let us begin with the natural gradient:

$$\begin{aligned}
&\left[\widetilde{\nabla} \mathrm{GAP}_q(\theta)\right]_1 + \left[\widetilde{\nabla} \mathrm{GAP}_q(\theta)\right]_3\\
&= \left(\frac{p^*(0)}{\theta_1+\theta_3} - \frac{p^*(1)}{\theta_2+\theta_4}\right)\theta_1(\theta_2+\theta_4) - p^*(0)q(0|0) + \theta_1\\
&\quad + \left(\frac{p^*(0)}{\theta_1+\theta_3} - \frac{p^*(1)}{\theta_2+\theta_4}\right)\theta_3(\theta_2+\theta_4) - p^*(0)q(1|0) + \theta_3\\
&= \left(\frac{p^*(0)}{\theta_1+\theta_3} - \frac{p^*(1)}{\theta_2+\theta_4}\right)(\theta_1+\theta_3)(\theta_2+\theta_4) \underbrace{-p^*(0)q(0|0) - p^*(0)q(1|0)}_{-p^*(0)} + \theta_1 + \theta_3\\
&= p^*(0)(\theta_2+\theta_4) - p^*(1)(\theta_1+\theta_3) - p^*(0) + \theta_1 + \theta_3\\
&= p^*(0) \underbrace{-p^*(0)(\theta_1+\theta_3) - p^*(1)(\theta_1+\theta_3)}_{-(\theta_1+\theta_3)} - p^*(0) + \theta_1 + \theta_3\\
&= 0.
\end{aligned}$$

This exemplifies our core result (6.34) in terms of local coordinates. The same calculation for the Euclidean gradient does not lead to this result. Thus, generically we have

$$\left[\nabla \mathrm{GAP}_q(\theta)\right]_1 + \left[\nabla \mathrm{GAP}_q(\theta)\right]_3 \neq 0.$$

6.C Proof of Proposition 6.2

Proof of Proposition 6.2. Without loss of generality, we identify the unit set N with the set $\{1, \dots, n\}$, $n = |N|$, in a way that is consistent with the graph $G = (N, E)$. That means, whenever $i \in \mathrm{pa}(s)$ we have $i < s$. Note that such an identification is always possible for a directed acyclic graph. Furthermore, we assume $s < t$. Then:

$$
\begin{aligned}
& g_\theta^{\mathrm{FR}}\left(\partial_{s,k}(\theta), \partial_{t,l}(\theta)\right) \\
&= \sum_x \frac{1}{p(x;\theta)} \partial_{s,k}(x;\theta)\, \partial_{t,l}(x;\theta) \qquad \text{(by (6.11))} \\
&= \sum_x \frac{1}{p(x;\theta)} \qquad \text{(by (6.39))} \\
&\quad \times \left(p(x;\theta) \frac{\partial}{\partial \theta_{s,k}} \ln p(x_s | x_{\mathrm{pa}(s)}; \theta_s) \right) \left(p(x;\theta) \frac{\partial}{\partial \theta_{t,l}} \ln p(x_t | x_{\mathrm{pa}(t)}; \theta_t) \right) \\
&= \sum_x p(x;\theta) \frac{\partial}{\partial \theta_{s,k}} \ln p(x_s | x_{\mathrm{pa}(s)}; \theta_s) \frac{\partial}{\partial \theta_{t,l}} \ln p(x_t | x_{\mathrm{pa}(t)}; \theta_t)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_x \frac{\partial}{\partial \theta_{s,k}} \ln p(x_s | x_{\mathrm{pa}(s)}; \theta_s) \prod_{i=1}^{t} p(x_i | x_{\mathrm{pa}(i)}; \theta_i) \frac{\partial}{\partial \theta_{t,l}} \ln p(x_t | x_{\mathrm{pa}(t)}; \theta_t) \\
&\quad \times \underbrace{\prod_{i=t+1}^{n} p(x_i | x_{\mathrm{pa}(i)}; \theta_i)}_{=1} \\
&= \sum_{x_1, \dots, x_t} \frac{\partial}{\partial \theta_{s,k}} \ln p(x_s | x_{\mathrm{pa}(s)}; \theta_s) \prod_{i=1}^{t} p(x_i | x_{\mathrm{pa}(i)}; \theta_i) \frac{\partial}{\partial \theta_{t,l}} \ln p(x_t | x_{\mathrm{pa}(t)}; \theta_t) \\
&= \sum_{x_1, \dots, x_{t-1}} \frac{\partial}{\partial \theta_{s,k}} \ln p(x_s | x_{\mathrm{pa}(s)}; \theta_s) \sum_{x_t} \prod_{i=1}^{t} p(x_i | x_{\mathrm{pa}(i)}; \theta_i) \\
&\quad \times \frac{\partial}{\partial \theta_{t,l}} \ln p(x_t | x_{\mathrm{pa}(t)}; \theta_t) \\
&= \sum_{x_1, \dots, x_{t-1}} \frac{\partial}{\partial \theta_{s,k}} \ln p(x_s | x_{\mathrm{pa}(s)}; \theta_s) \prod_{i=1}^{t-1} p(x_i | x_{\mathrm{pa}(i)}; \theta_i) \sum_{x_t} p(x_t | x_{\mathrm{pa}(t)}; \theta_t) \\
&\quad \times \frac{\partial}{\partial \theta_{t,l}} \ln p(x_t | x_{\mathrm{pa}(t)}; \theta_t) \\
&= \sum_{x_1, \dots, x_{t-1}} \frac{\partial}{\partial \theta_{s,k}} \ln p(x_s | x_{\mathrm{pa}(s)}; \theta_s) \prod_{i=1}^{t-1} p(x_i | x_{\mathrm{pa}(i)}; \theta_i) \frac{\partial}{\partial \theta_{t,l}} \sum_{x_t} p(x_t | x_{\mathrm{pa}(t)}; \theta_t) \\
&= \sum_{x_1, \dots, x_{t-1}} \frac{\partial}{\partial \theta_{s,k}} \ln p(x_s | x_{\mathrm{pa}(s)}; \theta_s) \prod_{i=1}^{t-1} p(x_i | x_{\mathrm{pa}(i)}; \theta) \frac{\partial}{\partial \theta_{t,l}} 1 \\
&= 0.
\end{aligned}
$$

□

References

Amari, Shun-ichi (1998). "Natural Gradient Works Efficiently in Learning". In: *Neural Computation* 10.2, pp. 251–276.

– (2016). *Information geometry and its applications*. Vol. 194. Springer.

Amari, Shun-ichi and Hiroshi Nagaoka (2000). *Methods of information geometry*. Vol. 191. American Mathematical Soc.

Ay, Nihat (2020). "On the locality of the natural gradient for learning in deep Bayesian networks". In: *Information Geometry*, pp. 1–49.

Ay, Nihat and Shun-ichi Amari (2015). "A Novel Approach to Canonical Divergences within Information Geometry". In: *Entropy* 17.12, pp. 8111–8129. ISSN: 1099-4300. DOI: 10.3390/e17127866. URL: https://www.mdpi.com/1099-4300/17/12/7866.

Ay, Nihat, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer (2017). *Information geometry*. Vol. 64. Springer.

Bayer, Justin, Maximilian Soelch, Atanas Mirchev, Baris Kayalibay, and Patrick van der Smagt (2021). "Mind the Gap when Conditioning Amortised Inference in Sequential Latent-Variable Models". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=a2gqxKDvYys.

Chen, Liqun, Chenyang Tao, Ruiyi Zhang, Ricardo Henao, and Lawrence Carin Duke (2018). "Variational inference and model selection with generalized evidence bounds". In: *International conference on machine learning*. PMLR, pp. 893–902.

Chentsov, Nikolai Nikolaevich (1982). "Statiscal decision rules and optimal inference". In: *Monog* 53.

Datar, Adwait and Nihat Ay (2025). "Convergence Properties of Natural Gradient Descent for Minimizing KL Divergence". In: *Transactions on Machine Learning Research*. ISSN: 2835-8856. URL: https://openreview.net/forum?id=h6hjjAF5Bj.

Datar, Adwait, Jesse van Oostrum, and Nihat Ay (2024). *Code for paper: On the Natural Gradient of the Evidence Lower Bound*. https://github.com/addat10/Nat-Gradient-ELBO.git.

Dayan, Peter, Geoffrey E Hinton, Radford M Neal, and Richard S Zemel (1995). "The Helmholtz Machine". In: *Neural computation* 7.5, pp. 889–904.

Feynman, Richard P (1972). *Statistical mechanics: a set of lectures*. W.A. Benjamin.

Friston, Karl (2005). "A theory of cortical responses". In: *Philosophical transactions of the Royal Society B: Biological sciences* 360.1456, pp. 815–836.

Fujiwara, Akio and Shun-ichi Amari (1995). "Gradient systems in view of information geometry". In: *Physica D: Nonlinear Phenomena* 80.3, pp. 317–327.

Hinton, Geoffrey E and Drew Van Camp (1993). "Keeping the neural networks simple by minimizing the description length of the weights". In: *Proceedings of the sixth annual conference on Computational learning theory*, pp. 5–13.

Hinton, Geoffrey E and Richard Zemel (1993). "Autoencoders, minimum description length and Helmholtz free energy". In: *Advances in neural information processing systems* 6.

Ikeda, Shiro, Shun-ichi Amari, and Hiroyuki Nakahara (1998). "Convergence of the wake-sleep algorithm". In: *Advances in neural information processing systems* 11.

Kingma, Diederik P and Max Welling (2013). "Auto-Encoding Variational Bayes". In: *arXiv preprint arXiv:1312.6114*.

Li, Yingzhen and Richard E Turner (2016). "Rényi divergence variational inference". In: *Advances in neural information processing systems* 29.

MacKay, David JC (1995). "Developments in probabilistic modelling with neural networks—ensemble learning". In: *Neural Networks: Artificial Intelligence and Industrial Applications: Proceedings of the Third Annual SNN Symposium on Neural Networks, Nijmegen, The Netherlands, 14–15 September 1995*. Springer, pp. 191–198.

Martens, James (2020). "New Insights and Perspectives on the Natural Gradient Method". In: *Journal of Machine Learning Research* 21, pp. 1–76.

Ollivier, Yann (2015). "Riemannian metrics for neural networks I: feedforward networks". In: *Information and Inference: A Journal of the IMA* 4.2, pp. 108–153.

Oostrum, Jesse van, Peter van Hintum, and Nihat Ay (Feb. 2024). "Inversion of Bayesian networks". In: *Int. J. Approx. Reasoning* 164.C. ISSN: 0888-613X. DOI: 10.1016/j.ijar.2023.109042. URL: https://doi.org/10.1016/j.ijar.2023.109042.

Oostrum, Jesse van, Johannes Müller, and Nihat Ay (2023). "Invariance properties of the natural gradient in overparametrised systems". In: *Information geometry* 6.1, pp. 51–67.

Pu, Yuchen, Zhe Gan, Ricardo Henao, Chunyuan Li, Shaobo Han, and Lawrence Carin (2017). "VAE learning via Stein variational gradient descent". In: *Advances in Neural Information Processing Systems* 30.

Rainforth, Tom, Adam R. Kosiorek, Tuan Anh Le, Chris J. Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh (2018). "Tighter Variational Bounds are Not Necessarily Better". In: *International Conference on Machine Learning*. URL: https://api.semanticscholar.org/CorpusID:3281926.

Rezende, Danilo Jimenez, Shakir Mohamed, and Daan Wierstra (2014). "Stochastic backpropagation and approximate inference in deep generative models". In: *International conference on machine learning*. PMLR, pp. 1278–1286.

Webb, Stefan, Adam Golinski, Rob Zinkov, Siddharth N, Tom Rainforth, Yee Whye Teh, and Frank Wood (2018). "Faithful Inversion of Generative Models for Effective Amortized Inference". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/894b77f805bd94d292574c38c5d628d5-Paper.pdf.

Chapter 7

A Concise Mathematical Description of Active Inference in Discrete Time

Jesse van Oostrum, Carlotta Langer, Nihat Ay

Journal of Mathematical Psychology, Elsevier, 2025

Abstract. In this paper we present a concise mathematical description of active inference in discrete time. The main part of the paper serves as a basic introduction to the topic, including a detailed example of the action selection mechanism. The appendix discusses the more subtle mathematical details, targeting readers who have already studied the active inference literature but struggle to make sense of the mathematical details and derivations. Throughout, we emphasize precise and standard mathematical notation, ensuring consistency with existing texts and linking all equations to widely used references on active inference. Additionally, we provide Python code that implements the action selection and learning mechanisms described in this paper and is compatible with `pymdp` environments.

Introduction

Active inference is a theory that describes the action selection and learning mechanisms of an agent in an environment. We aim to present a concise mathematical description of the theory so that readers interested in the mathematical details can quickly find what they are looking for. We have paid special attention to choosing notation that is more in line with standard mathematical texts and is also descriptive, in the sense that dependencies are made explicit. Hence, the focus of this paper lies on the mathematical details and derivations rather than verbal motivations and justifications.

The paper consists of a main text and an appendix. The main text provides a clear introduction to active inference in discrete time, accessible for people new to the topic. It is divided into two parts: inference, which assumes a given generative model, and learning, which explains how the agent acquires this model. The main text concludes

with a worked example of action selection. The appendix delves into finer details and derivations, catering to readers familiar with active inference who seek clarity on the mathematical aspects.

To complement our theoretical exposition, we provide a Python implementation[1] of the action selection and learning mechanisms described in this paper, which is compatible with `pymdp` environments. This code is more minimalistic, which makes it easier to understand than other implementations such as `SPM` and `pymdp`.

7.1 Set-up and notation

In this paper we consider an active inference agent acting in a discrete-time setting with a finite-time horizon. This means that we consider a sequence of T time steps and at every time step τ the agent receives an observation o_τ, and performs an action a_τ. We use τ for arbitrary time steps and the letter t to denote the current time step. We use the subscript $_{\tau:\tau'}$ to denote a sequence of variables, e.g. $o_{\tau:\tau'} = (o_\tau, \dots, o_{\tau'})$. A sequence of (future) actions is called a *policy*[2] and is denoted by $\pi_t = a_{t:T}$, with $\pi = \pi_1$. We write $a_{1:t-1}$ for actions that were performed in the past and π_t for future actions that still need to be selected.

The agent models the dynamics of the environment using an internal generative model. This model uses a variable s_τ, called an internal state, to represent the state of the environment[3] at time step τ. The model is given by the following probability distribution:

$$p(o_{1:T}, s_{1:T} | a_{1:T-1}, \theta).$$

This probability distribution factorizes according to the graph in Figure 7.1. In the

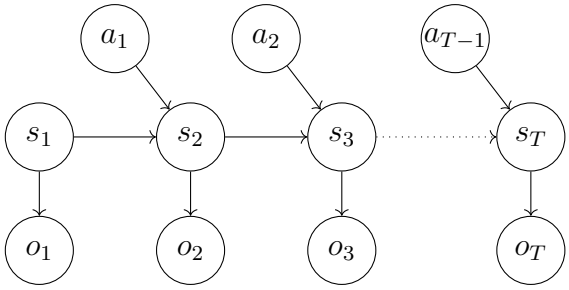


Figure 7.1: Graphical representation of the generative model

first part of this paper we assume that this generative model is given and need not be learned, and we therefore suppress the dependence on the parameter θ. In the second part we discuss how the model is learned.

Suppose the agent is at time step t. It will have received observations $o_{1:t}$ and performed actions $a_{1:t-1}$. We use $q_t(s_{\tau:\tau'})$ to denote the (approximate) posterior distribution of the generative model, $p(s_{\tau:\tau'} | o_{1:t}, a_{1:t-1})$, and also refer to this as the belief of the agent about the variable $s_{\tau:\tau'}$.

[1] `https://github.com/jessevoostrum/active-inference`

[2] Note that in a reinforcement learning context the term "policy" has a different meaning.

[3] Note that the number of possible internal states is usually much smaller than the actual number of states the environment can be in.

7.2 Inference

7.2.1 Action selection according to active inference

Let an agent be at time step t, having received observations $o_{1:t}$ and performed actions $a_{1:t-1}$. According to active inference an agent selects its next action by sampling a policy π_t from the following distribution (equation (10) in Da Costa et al., 2020):

$$\sigma(-G(\pi_t|o_{1:t}, a_{1:t-1})) \tag{7.1}$$

and selecting the action a_t corresponding to that policy.[4] The function σ denotes the softmax function defined in (7.48) and G is the expected free energy function given by

$$\begin{aligned} G(\pi_t|o_{1:t}, a_{1:t-1}) = -\Bigg(&\mathbb{E}_{q_t(o_{t+1:T}|\pi_t)}\Big[D_{\mathrm{KL}}\Big(q_t(s_{t+1:T}|o_{t+1:T}, \pi_t) \parallel q_t(s_{t+1:T}|\pi_t)\Big)\Big] \\ &+ \mathbb{E}_{q_t(o_{t+1:T}|\pi_t)}\Big[\ln p_C(o_{t+1:T})\Big]\Bigg). \end{aligned} \tag{7.2}$$

Note that according to (7.1) the agent is more likely to sample policies π_t that have a low expected free energy $G(\pi_t, o_{1:t}, a_{1:t-1})$. In Section 7.2.2 we discuss equivalent formulations and different interpretations of the expected free energy. The distributions $q_t(o_{t+1:T}|\pi_t)$, $q_t(s_{t+1:T}|o_{t+1:T}, \pi_t)$, $q_t(s_{t+1:T}|\pi_t)$, needed for the calculation of G, are (approximate) posterior distributions of the generative model of the agent after having observed $o_{1:t}$ and performed $a_{1:t-1}$. In Section 7.2.3 we describe how the agent infers these posterior distributions. The distribution p_C is a preference distribution over observations that we assume is given to the agent. This distribution is distinct from the generative model p.

Remark 7.1. Note that in certain descriptions of active inference in discrete time also a variational free energy term F appears in the distribution in equation (7.1) (e.g. equation (B.9) in Parr et al., 2022). This term is only relevant in specific cases that we will discuss in Remark 7.2 in Appendix 7.A. Furthermore a habit term E is sometimes included that is also considered in Appendix 7.A, but discarded here for simplicity.

7.2.2 Expected free energy

Recall that equation (7.2) gives the following expression for the expected free energy function:

$$\begin{aligned} G(\pi_t|o_{1:t}, a_{1:t-1}) = -\Bigg(&\mathbb{E}_{q_t(o_{t+1:T}|\pi_t)}\Big[D_{\mathrm{KL}}\Big(q_t(s_{t+1:T}|o_{t+1:T}, \pi_t) \parallel q_t(s_{t+1:T}|\pi_t)\Big)\Big] \\ &+ \mathbb{E}_{q_t(o_{t+1:T}|\pi_t)}\Big[\ln p_C(o_{t+1:T})\Big]\Bigg). \end{aligned}$$

The first term between the brackets on the RHS is called *epistemic value* or *information gain*. It measures the average change in belief about the future states $s_{t+1:T}$ due to receiving future observations $o_{t+1:T}$. The second term, known as *utility*, quantifies the similarity between the expected future observation distribution and the preferred

[4] Note that at every time step a new policy is sampled, only the action a_t for the corresponding time step t is executed, and the rest of the actions in π_t are discarded.

observation distribution. As previously mentioned, the agent is more likely to sample policies with low expected free energy, which correspond to high information gain and utility.

An equivalent formulation of expected free energy is given by

$$\begin{aligned} G(\pi_t|o_{1:t}, a_{1:t-1}) = \mathbb{E}_{q_t(s_{t+1:T}|\pi_t)}\Big[\mathrm{H}\big[p(o_{t+1:T}|s_{t+1:T})\big]\Big] \\ + D_{\mathrm{KL}}\Big(q_t(o_{t+1:T}|\pi_t) \parallel p_C(o_{t+1:T})\Big). \end{aligned} \tag{7.3}$$

The first term on the RHS is referred to as *ambiguity*. It measures the average uncertainty an agent has about its future observations given knowledge of its future states. The second term is called *expected complexity* or *risk*. It represents the divergence between expected and preferred future observations. The agent favors policies with low ambiguity and risk.

In Appendix 7.C we show that both expressions of the expected free energy are equal.

In practice, often the following mean field approximations are made:

$$\begin{aligned} q_t(s_{t+1:T}) &= \prod_{\tau=t+1}^{T} q_t(s_\tau), \\ p_C(o_{t+1:T}) &= \prod_{\tau=t+1}^{T} p_C(o_\tau). \end{aligned}$$

Equation (7.2) and (7.3) can then be written as follows:

$$G(\pi_t|o_{1:t}, a_{1:t-1}) = \sum_{\tau=t+1}^{T} G_\tau(\pi_t, o_{1:t}, a_{1:t-1}), \tag{7.4}$$

with

$$\begin{aligned} G_\tau(\pi_t|o_{1:t}, a_{1:t-1}) &= -\Bigg(\mathbb{E}_{q_t(o_\tau|\pi_t)}\Big[D_{\mathrm{KL}}\Big(q_t(s_\tau|o_\tau, \pi_t) \parallel q_t(s_\tau|\pi_t)\Big)\Big] \\ &\qquad + \mathbb{E}_{q_t(o_\tau|\pi_t)}\Big[\ln p_C(o_\tau)\Big]\Bigg) \\ &= \mathbb{E}_{q_t(s_\tau|\pi_t)}\Big[\mathrm{H}\big[p(o_\tau|s_\tau)\big]\Big] + D_{\mathrm{KL}}\Big(q_t(o_\tau|\pi_t) \parallel p_C(o_\tau)\Big). \end{aligned} \tag{7.5}$$

It is outside the scope of this paper to further derive or motivate the expected free energy. We refer the reader to Appendix B.2.5 in Parr et al., 2022 and Da Costa et al., 2024; Wei, 2024; Millidge et al., 2021 for more details.

7.2.3 State inference

In this section we describe the simplest form of state inference, which is obtained by applying Bayes' rule. State inference methods as described in e.g. Parr et al., 2022; Heins et al., 2022; Smith et al., 2022 can be thought of as computationally efficient approximations of what is described here. See Appendix 7.A for more details on these methods.

Above we defined q_t to be the (approximate) posterior of the generative model given $o_{1:t}$, $a_{1:t-1}$. In this section we make the conditioning variables explicit and write $q(\,\cdot\,|o_{1:t}, a_{1:t-1})$ instead.

Current and future state inference

We start by studying the generative model that is assumed to be given to the agent. The generative model can be decomposed as follows: (see Figure 7.1)

$$\begin{aligned} p(o_{1:T}, s_{1:T}|a_{1:T-1}) &= p(s_{1:T}|a_{1:T-1})p(o_{1:T}|s_{1:T}) \\ p(s_{1:T}|a_{1:T-1}) &= p(s_1)\prod_{\tau=2}^{T} p(s_\tau|s_{\tau-1}, a_{\tau-1}) \\ p(o_{1:T}|s_{1:T}) &= \prod_{\tau=1}^{T} p(o_\tau|s_\tau). \end{aligned}$$

At every time step the agent updates its belief about the current state it is in. Before having performed any observations, its belief about the current state is equal to the prior belief $p(s_1)$. After receiving o_1 it will update its belief using Bayes' rule:

$$q(s_1|o_1) \propto p(o_1|s_1)p(s_1).$$

Subsequently, it will perform an action a_1 (selected as described in Section 7.2.1) and receive a next observation o_2. The belief about s_2 is given by

$$\begin{aligned} q(s_2|o_{1:2}, a_1) &\propto p(o_2|s_2, o_1, a_1)p(s_2|o_1, a_1) \\ &= p(o_2|s_2)\sum_{s_1} p(s_2|s_1, o_1, a_1)p(s_1|a_1, o_1) \\ &= p(o_2|s_2)\sum_{s_1} p(s_2|s_1, a_1)q(s_1|o_1). \end{aligned}$$

For a general t the belief about s_t is updated as follows:

$$q(s_t|o_{1:t}, a_{1:t-1}) \propto p(o_t|s_t)\sum_{s_{t-1}} p(s_t|s_{t-1}, a_{t-1})q(s_{t-1}|o_{1:t-1}, a_{1:t-2}). \tag{7.6}$$

For future time point $\tau > t$, the belief about the state s_τ is given by

$$q(s_\tau|o_{1:t}, a_{1:\tau-1}) = \sum_{s_{\tau-1}} p(s_\tau|s_{\tau-1}, a_{\tau-1})q(s_{\tau-1}|o_{1:t}, a_{1:\tau-2}). \tag{7.7}$$

The belief about a future state given a future observation is calculated as follows:

$$q(s_\tau|o_\tau, o_{1:t}, a_{1:\tau-1}) \propto p(o_\tau|s_\tau)q(s_\tau|o_{1:t}, a_{1:\tau-1}). \tag{7.8}$$

Future observation inference

In order to compute G in (7.2), we need a posterior distribution $q(o_\tau|o_{1:t}, a_{1:\tau-1})$ over future observations o_τ. This can be computed as follows:

$$q(o_\tau|o_{1:t}, a_{1:\tau-1}) = \sum_{s_\tau} p(o_\tau|s_\tau)q(s_\tau|o_{1:t}, a_{1:\tau-1}). \tag{7.9}$$

Past, current, and future state inference

In order to perform inference over states in the past, present and future (which is needed for the learning of the generative model and for the computation of the variational free energy over states), the agent can use the following formula:

$$\begin{aligned} q(s_{1:T}|o_{1:t}, a_{1:t-1}, \pi_t) &\propto p(o_{1:t}|s_{1:T}, a_{1:t-1}, \pi_t)p(s_{1:T}|a_{1:t-1}, \pi_t) \\ &= p(o_{1:t}|s_{1:t})p(s_{1:T}|a_{1:t-1}, \pi_t) \\ &= \prod_{\tau=1}^{t} p(o_\tau|s_\tau)\ p(s_1) \prod_{\tau=2}^{T} p(s_\tau|s_{\tau-1}, a_{\tau-1}), \end{aligned} \tag{7.10}$$

where we use $a_{\tau-1}$ in the last term for elements of both $a_{1:t-1}$ and π_t.

7.3 Learning

In the above section, we have assumed that the agent has access to a generative model $p(o_{1:T}, s_{1:T}|a_{1:T-1}, \theta)$. In this section we discuss how the parameter θ of this model is learned. The generative model consists of three (conditional) categorical distributions that are parametrized by $\theta = (\theta^D, \theta^A, \theta^B)$ in the following way:

$$p(s_1^{(j)}|\theta^D) = \theta_j^D, \tag{7.11}$$

$$p(o_\tau^{(i)}|s_\tau^{(j)}, \theta^A) = \theta_{ij}^A, \tag{7.12}$$

$$p(s_\tau^{(j)}|s_{\tau-1}^{(k)}, a_{\tau-1}^{(l)}, \theta^B) = \theta_{jkl}^B, \tag{7.13}$$

where we have enumerated the elements of the observation, action and state space with the bracketed superscript $^{(\cdot)}$. In order to learn the parameters of the generative model, we adopt a Bayesian belief updating scheme with a Dirichlet prior. (See Appendix 7.B for details on this.) More specifically, the prior over θ is parametrized by $\alpha = (\alpha^D, \alpha^A, \alpha^B)$ and is given by

$$p(\theta|\alpha) = p(\theta^D|\alpha^D) \prod_j p(\theta_{\bullet j}^A|\alpha^A) \prod_{k,l} p(\theta_{\bullet kl}^B|\alpha^B) \tag{7.14}$$

$$p(\theta^D|\alpha^D) \propto \prod_j \left(\theta_j^D\right)^{\alpha_j^D - 1}, \tag{7.15}$$

$$p(\theta_{\bullet j}^A|\alpha^A) \propto \prod_i \left(\theta_{ij}^A\right)^{\alpha_{ij}^A - 1}, \tag{7.16}$$

$$p(\theta_{\bullet kl}^B|\alpha^B) \propto \prod_j \left(\theta_{jkl}^B\right)^{\alpha_{jkl}^B - 1}, \tag{7.17}$$

where $\theta_{\bullet j}$ denotes the vector $(\theta_{1j}, \ldots, \theta_{nj})$.

Now after performing actions $a_{1:T-1}$ and receiving observations $o_{1:T}$ we want to update our belief about θ according to Bayes' rule. The true posteriors over these parameters are not Dirichlet distributions. (See Appendix 7.B for details.) The active inference literature suggests to set the approximate posterior distribution to be a Dirichlet distribution and update the hyperparameter α in the following way (equation (B.12) in

Parr et al., 2022 and equation (21), (A.6) and (A.7) in Da Costa et al., 2020):

$$\alpha_j^{D\prime} = \alpha_j^D + q_T\left(s_1^{(j)}\right), \tag{7.18}$$

$$\alpha_{ij}^{A\prime} = \alpha_{ij}^A + \sum_{\tau=1}^{T} \mathbb{1}_{o^{(i)}}(o_\tau) q_T\left(s_\tau^{(j)}\right), \tag{7.19}$$

$$\alpha_{jkl}^{B\ \prime} = \alpha_{jkl}^B + \sum_{\tau=2}^{T} q_T\left(s_\tau^{(j)}\right) q_T\left(s_{\tau-1}^{(k)}\right) \mathbb{1}_{a^{(l)}}(a_{\tau-1}). \tag{7.20}$$

The distributions $q_T(s_\tau), \tau \in \{1, \ldots, T\}$ are approximate posteriors obtained[5] using the current version of the generative model (before θ has been updated). In Appendix 7.B we elaborate on the origin of this learning rule.

Note the similarity with the standard update rule for Dirichlet priors given in (7.31). In the standard update rule the element α_{i^*} of the hyperparameter corresponding to the observation $x^{(i^*)}$ is incremented by 1, which makes this observation more likely in the updated distribution. In the updates (7.18)–(7.20) the hyperparameters are incremented by the amount of posterior belief in that state or state transition, e.g. α_j^D is incremented by $q\left(s_1^{(j)} | o_{1:T}, a_{1:T-1}\right)$.

In order to go from a Dirichlet distribution $p(\theta|\alpha)$ to an actual value of the parameter that can be used for the generative model, the mean of the distribution can be used, which is given by

$$\begin{aligned} \hat{\theta}_i &= \mathbb{E}_{p(\theta|\alpha)}[\theta_i] \\ &= \frac{\alpha_i}{\sum_j \alpha_j}. \end{aligned} \tag{7.21}$$

For example, after the learning step, the new distribution over s_1 is given by

$$p(s_1^{(j)} | \hat{\theta}^D) = \frac{\alpha_j^{D\prime}}{\sum_j \alpha_k^{D\prime}}.$$

This concludes the discussion of learning in the context of the active inference framework.

7.4 Example: T-maze example

In this section we discuss the action selection mechanism of an active inference agent in the T-maze environment depicted in Figure 7.2. (See also Section 7.3 of Parr et al., 2022 and Heins, n.d.) This illustrates the theory of action selection and state inference that is presented in Section 7.2.

[5] See Section 7.2.3.

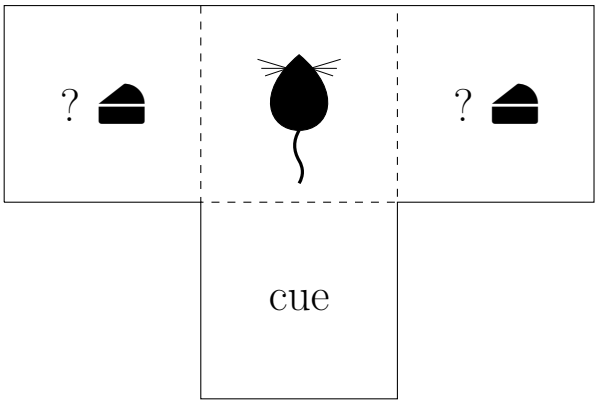


Figure 7.2: T-maze environment

Description of the (internal) generative model of the agent

State, observation and action spaces

The internal state space of the agent[6] $\mathcal{S}$ has two dimensions,[7], *Location* and *Reward condition*, and can be described as follows:

$$\begin{aligned}\mathcal{S} &= \mathcal{S}^L \times \mathcal{S}^R,\\ \mathcal{S}^L &= \{\texttt{center}, \texttt{right arm}, \texttt{left arm}, \texttt{cue location}\},\\ \mathcal{S}^R &= \{\texttt{reward on right}, \texttt{reward on left}\}.\end{aligned}$$

A typical element of the state space is written as $s = (s^L, s^R)$.

The observation space $\mathcal{O}$ has three dimensions[8,9], *Location*, *Reward*, and *Cue*, and can be described as follows:

$$\begin{aligned}\mathcal{O} &= \mathcal{O}^L \times \mathcal{O}^R \times \mathcal{O}^C,\\ \mathcal{O}^L &= \{\texttt{center}, \texttt{right arm}, \texttt{left arm}, \texttt{cue location}\},\\ \mathcal{O}^R &= \{\texttt{no reward}, \texttt{reward}, \texttt{loss}\},\\ \mathcal{O}^C &= \{\texttt{cue right}, \texttt{cue left}\}.\end{aligned}$$

A typical element of the observation space is written as $o = (o^L, o^R, o^C)$.

The space of actions $\mathcal{A}$ is described by

$$\begin{aligned}\mathcal{A} = \{&\texttt{move to center}, \texttt{move to right arm},\\ &\texttt{move to left arm}, \texttt{move to cue location}\}.\end{aligned}$$

Note that these actions are always available, independent of the current location of the agent. A typical element of the action space is written as a.

[6]In this example the state of the environment (generative process) is the same as the state space of the internal world model of the agent (generative model). Note that this is in general not the case. In a more realistic setting the state space of the environment will be much more complex than the internal state space.

[7]The dimensions of the state space are sometimes referred to as state factors.

[8]The dimensions of the observation space are sometimes referred to as observation modalities.

[9]Note that we follow here the description from Heins, n.d. In Parr et al., 2022 the cue observation is absorbed into the location observation.

In the following we use `[dir]` as a placeholder for `left` and `right`, and `[loc]` as a placeholder for the four locations `center`, `right arm`, `left arm` and `cue location`.

Observation kernel $p(o|s)$

We now specify the observation kernel $p(o|s)$ of the generative model of the agent. First note that the observation dimensions are independent, that is

$$p(o|s) = p(o^L|s)p(o^R|s).$$

The beliefs about the location observation given the state are modelled as follows:

$$p(o^L|s) = \begin{cases} 1 & \text{if } o^L = s^L \\ 0 & \text{otherwise} \end{cases},$$

which implies that the location can be unambiguously inferred from the observation.

The beliefs about the reward observation given the state are modelled as follows:

$$\begin{aligned} p(o^R = \texttt{no reward}|s^L \in \{\texttt{center}, \texttt{cue location}\}, s^R) &= 1, \\ p(o^R = \texttt{no reward}|s^L \notin \{\texttt{center}, \texttt{cue location}\}, s^R) &= 0, \\ p(o^R = \texttt{reward}|s^L = \texttt{[dir] arm}, s^R = \texttt{reward on [dir]}) &= 0.98, \\ p(o^R = \texttt{loss}|s^L = \texttt{right arm}, s^R = \texttt{reward on left}) &= 0.98, \\ p(o^R = \texttt{loss}|s^L = \texttt{left arm}, s^R = \texttt{reward on right}) &= 0.98. \end{aligned}$$

This implies that the agent observes `no reward` when it is in location `center` or `cue location`, it observes `reward` when it is in the same arm as specified by the reward condition with high probability, and it observes `loss` when it is in the opposite arm of the reward condition with high probability.

The beliefs about the cue observation given the state are modelled as follows:

$$\begin{aligned} &p(o^C = \texttt{cue [dir]}|s^L = \texttt{cue location}, s^R = \texttt{reward on [dir]}) = 1, \\ &p(o^C = \texttt{cue [dir]}|s^L \in \mathcal{S}^L \setminus \{\texttt{cue location}\}, s^R) = 0.5. \end{aligned}$$

This implies that the cue observation is completely informative about the reward condition when the agent is at the cue location, and otherwise independent of the actual reward condition.

Transition dynamics kernel $p(s_{\tau+1}|s_\tau, a_\tau)$

We continue by describing the transition dynamics kernel $p(s_{\tau+1}|s_\tau, a_\tau)$.

$$\begin{aligned} &p(s^L_{\tau+1} = \texttt{[loc]}|s^L_\tau \in \{\texttt{center}, \texttt{cue location}\}, s^R_\tau, a_\tau = \texttt{go to [loc]}) = 1, \\ &p(s^L_{\tau+1} = \texttt{[dir] arm}|s^L_\tau = \texttt{[dir] arm}, s^R_\tau, a_\tau) = 1. \end{aligned}$$

This implies that if the agent is in `center` or `cue location` it will be in the location specified by the action in the next time step. If it is in one of the arms however, it will stay there, independent of the choice of action a_τ.

$$p(s^R_{\tau+1} = \texttt{reward on [dir]}|s^R_\tau = \texttt{reward on [dir]}, s^L_\tau, a_\tau) = 1.$$

This implies that the reward condition stays constant throughout the trajectory.

Preference distribution p_C and prior over states p_D

The unnormalized preference distribution p_C can be chosen to favor observations with `reward` and discourage observations with `loss` as follows:

$$p_C((\texttt{[loc]}, \texttt{no reward}, \texttt{cue [dir]})) = 2,$$
$$p_C((\texttt{[loc]}, \texttt{reward}, \texttt{cue [dir]})) = 3,$$
$$p_C((\texttt{[loc]}, \texttt{loss}, \texttt{cue [dir]})) = 1.$$

Finally we let the prior belief over states p_D be uniform.

Action selection procedure

We will now simulate the trajectory of an agent acting according to active inference. We set the time horizon to $T = 3$, which implies that the policies will have length 2.

Time step 1

The agent starts by receiving an observation $o_1 = (\texttt{center}, \texttt{no reward}, \texttt{cue right})$. It now updates its beliefs about the current state such that

$$q_t(s_1^L = \texttt{center}, s_1^R = \texttt{reward on [dir]}) = 0.5.$$

Subsequently it computes its beliefs about future states and observations given a policy using equation (7.7), (7.8), (7.9). For example for $\pi_1^* = (\texttt{move to cue location}, \texttt{move to left arm})$ we have

$$q_1(s_2^L = \texttt{cue location}, s_2^R = \texttt{reward on [dir]} \mid \pi_1 = \pi_1^*) = 0.5, \tag{7.22}$$
$$q_1(s_3^L = \texttt{left arm}, s_3^R = \texttt{reward on [dir]} \mid \pi_1 = \pi_1^*) = 0.5,$$

and

$$q_1(o_2^L = \texttt{cue location}, o_2^R = \texttt{no reward}, o_2^C = \texttt{cue [dir]} \mid \pi_1 = \pi_1^*) = 0.5,$$
$$q_1(o_3^L = \texttt{left arm}, o_3^R = \texttt{reward}, o_3^C = \texttt{cue [dir]} \mid \pi_1 = \pi_1^*) = 0.25,$$
$$q_1(o_3^L = \texttt{left arm}, o_3^R = \texttt{loss}, o_3^C = \texttt{cue [dir]} \mid \pi_1 = \pi_1^*) = 0.25,$$

and for example for $o_2^* = (\texttt{cue location}, \texttt{no reward}, \texttt{cue left})$ we have

$$q_1(s_2^L = \texttt{cue location}, s_2^R = \texttt{reward on left} \mid o_2 = o_2^*, \pi_1 = \pi_1^*) = 1. \tag{7.23}$$

Note here the reduction of uncertainty about s_2 due to the observation o_2^*, represented by the epistemic value defined in Section 7.2.2 given by the KL divergence between the distributions (7.23) and (7.22).

The agent now computes G and plugs this into equation (7.1) and gets the following posterior distribution over policies:

$a_1 \downarrow$ \ $a_2 \rightarrow$	`center`	`right arm`	`left arm`	`cue location`
`center`	0.022	0.041	0.041	0.046
`right arm`	0.041	0.075	0.075	0.083
`left arm`	0.041	0.075	0.075	0.083
`cue location`	0.046	0.083	0.083	0.091

and in this scenario it samples a policy with as first action `move to cue location` with highest probability.

Time step 2

After having performed action $a_1^* =$ `move to cue location`, the next observation it receives is $o_2^* =$ (`cue location`, `no reward`, `cue right`). Its belief about the current state is now given by

$$q_2(s_2^L = \texttt{cue location}, s_2^R = \texttt{reward on right}) = 1.$$

For instance, when $\pi_2^* =$ (`move to left arm`), the beliefs about future states are

$$q_2(s_3^L = \texttt{left arm}, s_3^R = \texttt{reward on right} \mid \pi_2 = \pi_2^*) = 1,$$

and the observation beliefs

$$q_2(o_3^L = \texttt{left arm}, o_3^R = \texttt{reward}, o_3^C = \texttt{cue [dir]} \mid \pi_2 = \pi_2^*) = 0.01,$$
$$q_2(o_3^L = \texttt{left arm}, o_3^R = \texttt{loss}, o_3^C = \texttt{cue [dir]} \mid \pi_2 = \pi_2^*) = 0.49.$$

Since all uncertainty has already been taken away by the last observation, conditioning on $o_3^* =$ (`left arm`, `reward`, `cue left`) will make no difference to the belief about the state s_3, i.e.

$$q_2(s_3^L = \texttt{left arm}, s_3^R = \texttt{reward on right} \mid o_3 = o_3^*, \pi_2 = \pi_2^*) = q_2(s_3^L = \texttt{left arm}, s_3^R = \texttt{reward on right} \mid \pi_2 = \pi_2^*) = 1,$$

which will cause the epistemic value term in G to be zero.

The agent now calculates G again and obtains the following distribution over policies:

a_2	`center`	`right arm`	`left arm`	`cue location`
	0.20	0.52	0.08	0.20

and will then sample the policy (`move to right arm`) with highest probability.

Acknowledgements

The authors would like to thank Thomas Parr, Conor Heins, Ryan Smith, Beren Millidge, Pablo Lanillos, Sean Tull, Stephen Mann, Pradeep Kumar Banerjee, Frank Röder and Lance Da Costa for helpful discussions and comments and acknowledge the support of the Deutsche Forschungsgemeinschaft Priority Programme "The Active Self" (SPP 2134).

7.A Variational free energy minimization

Preliminaries

In this section we use x for a general observed variable and z for a general latent variable.

Let $p(x, z) = p(x|z)p(z)$ be a generative model. In order to perform inference over the latent variables after making an observation x, one has to compute the posterior $p(z|x)$. This posterior is often hard to compute directly. One can instead approximate this posterior by finding the distribution $q_x(z)$ in a family of distributions $\mathcal{Q}$ that minimizes the following function:

$$F(\tilde{q}|x) = \sum_z \tilde{q}(z) \left(\ln \tilde{q}(z) - \ln p(x, z)\right),$$

called the variational free energy. Note that we distinguish notationally $\tilde{q}$, which is a generic element of $\mathcal{Q}$ and a variable in F, and q_x, which is the minimizer of F for a fixed x, i.e.

$$q_x = \arg\min_{\tilde{q} \in \mathcal{Q}} F(\tilde{q}|x).$$

Note that when $\mathcal{Q}$ is large enough, for example when z is discrete and $\mathcal{Q}$ is the set of all probability distributions over z, then the minimizer of the free energy is equal to the exact posterior distribution and we have

$$q_x(z) = p(z|x).$$

We can replace $\ln p(x, z)$ in F by a general function f, i.e.

$$F_f(\tilde{q}|x) = \sum_z \tilde{q}(z) \left(\ln \tilde{q}(z) - f(x, z)\right).$$

The minimizer can be found by substituting $g(x, z) = e^{f(x,z)}$ as follows:

$$\begin{aligned} F_f(\tilde{q}, x) &= \sum_z \tilde{q}(z) \left(\ln \tilde{q}(z) - \ln g(x, z)\right) \\ &= \left(\sum_z \tilde{q}(z) \left(\ln \tilde{q}(z) - \ln \frac{g(x, z)}{\sum_{z'} g(x, z')}\right)\right) + \ln \sum_{z'} g(x, z'). \end{aligned}$$

If $\mathcal{Q}$ is again large enough, the minimizer is given by:

$$\begin{aligned} q_x(z) &= \frac{g(x, z)}{\sum_z g(x, z)} \\ &= \frac{e^{f(x,z)}}{\sum_z e^{f(x,z)}} \\ &= \sigma(f(x, z)), \end{aligned} \tag{7.24}$$

where σ is the softmax function defined in (7.48) and in the case that $f(x, z) = \ln p(x, z)$ we have

$$\begin{aligned} q_x(z) &= \sigma(\ln p(x, z)) \\ &= p(z|x). \end{aligned}$$

The variational free energy can be written as follows:

$$\begin{aligned} F(\tilde{q}|x) &= D_{\mathrm{KL}}(\tilde{q}(z) \parallel p(z|x)) - \ln p(x) \\ &\geq -\ln p(x), \end{aligned}$$

which shows that the negative of the variational free energy is a lower bound on the evidence (ELBO). By minimizing F w.r.t. $\tilde{q}$ we get the following approximate equality (equation (B.2) in Parr et al., 2022):

$$F(q_x|x) \approx -\ln p(x). \tag{7.25}$$

Variational free energy minimization in active inference

Active inference adopts the perspective that perception, action selection and learning can be interpreted as minimizing one single variational free energy function F. In its complete form it can be written as follows:

$$F(\tilde{q}|o_{1:t}, a_{1:t-1}) = \mathbb{E}_{\tilde{q}(s_{1:T},\theta,\pi_t)}\left[\ln \tilde{q}(s_{1:T}, \theta, \pi_t) - \ln p(s_{1:T}, o_{1:t}, \theta, \pi_t|a_{1:t-1})\right].$$

The distributions in the family $\mathcal{Q}$ are assumed to factorize as follows:

$$q(s_{1:T}, \theta, \pi) = q(\theta^D)q(\theta^A)q(\theta^B)q(\pi)\prod_{\tau=1}^{T} q(s_\tau),$$

which is sometimes referred to as the mean-field approximation. Now F can be written as follows:

$$F(\tilde{q}|o_{1:t}, a_{1:t-1}) = \mathbb{E}_{\tilde{q}(s_{1:T},\theta,\pi_t)}\Bigg[\ln \tilde{q}(\pi_t) + \ln \tilde{q}(\theta) + \sum_{\tau=1}^{T} \ln \tilde{q}(s_\tau) - \ln p(\pi_t) \\ - \ln p(\theta) - \ln p(s_1|\theta^D) - \sum_{\tau=1}^{t} \ln p(o_\tau|s_\tau, \theta^A) - \sum_{\tau=2}^{T} \ln p(s_\tau|s_{\tau-1}, a_{\tau-1}, \theta^B)\Bigg], \quad (7.26)$$

where we use $a_{\tau-1}$ in the last term for elements of both $a_{1:t-1}$ and π_t.

Perception

For studying perception, equation (7.26) can be rewritten as follows:

$$F(\tilde{q}|o_{1:t}, a_{1:t-1}) = \mathbb{E}_{\tilde{q}(\pi_t)}\left[F_{\pi_t}(\tilde{q}|o_{1:t}, a_{1:t-1})\right] + C_{\setminus s}$$

where $C_{\setminus s}$ is independent of $\tilde{q}(s_{1:T})$ and

$$F_{\pi_t}(\tilde{q}|o_{1:t}, a_{1:t-1}) = \\ \mathbb{E}_{\tilde{q}(s_{1:T})}\left[\ln \tilde{q}(s_{1:T}) - \mathbb{E}_{\tilde{q}(\theta)}[\ln p(s_{1:T}, o_{1:t}|\theta, a_{1:t-1}, \pi_t)]\right]. \quad (7.27)$$

Perception according to active inference is minimizing F_{π_t} w.r.t. $\tilde{q}(s_{1:T})$. The minimizer is written $q_t(s_{1:T}|\pi_t) = q(s_{1:T}|\pi_t, o_{1:t}, a_{1:t-1})$.

We can use the factorizing properties to rewrite (7.27) as follows:

$$F_{\pi_t}(\tilde{q}|o_{1:t}, a_{1:t-1}) = \sum_{\tau=1}^{T} \mathbb{E}_{\tilde{q}(s_\tau)}[\ln \tilde{q}(s_\tau)] - \mathbb{E}_{\tilde{q}(s_1,\theta^D)}[\ln p(s_1|\theta^D)] \\ - \sum_{\tau=1}^{t} \mathbb{E}_{\tilde{q}(s_\tau,\theta^A)} p(o_\tau|s_\tau, \theta^A) - \sum_{\tau=2}^{T} \mathbb{E}_{\tilde{q}(s_\tau,s_{\tau-1},\theta^B)} p(s_\tau|s_{\tau-1}, a_\tau, \theta^B),$$

which is equivalent to equation (6) in Da Costa et al., 2020.[10] One can also get rid of

[10]Note that in Da Costa et al., 2020 only the parameter θ^A is treated as a variable and θ^D and θ^B are considered fixed.

the expectations over θ by replacing them by an estimator $\hat{\theta}$. Then we get

$$F_{\pi_t}(\tilde{q}|o_{1:t}, a_{1:t-1}) = \sum_{\tau=1}^{T} \mathbb{E}_{\tilde{q}(s_\tau)}[\ln \tilde{q}(s_\tau)] - \mathbb{E}_{\tilde{q}(s_1)}[\ln p(s_1|\hat{\theta}^D)] \\ - \sum_{\tau=1}^{t} \mathbb{E}_{\tilde{q}(s_\tau)} p(o_\tau|s_\tau, \hat{\theta}^A) - \sum_{\tau=2}^{T} \mathbb{E}_{\tilde{q}(s_\tau, s_{\tau-1})} p(s_\tau|s_{\tau-1}, a_\tau, \hat{\theta}^B),$$

which is equivalent to (B.4) in Parr et al., 2022.

Learning

Updating the θ parameter (learning) happens at the end of an episode ($t = T$). The agent has observed $o_{1:T}$ and performed $a_{1:T-1}$. The variational free energy from equation (7.26) can be written as follows:

$$F(\tilde{q}|o_{1:T}, a_{1:T-1}) = \\ \mathbb{E}_{\tilde{q}(\theta)} \left[\ln \tilde{q}(\theta) - \mathbb{E}_{q_T(s_{1:T})}[\ln p(o_{1:T}, s_{1:T}, \theta|a_{1:T-1})]\right] + C_{\setminus\theta}, \quad (7.28)$$

where $C_{\setminus\theta}$ is independent of $\tilde{q}(\theta)$ and we have fixed $q_T(s_{1:T})$ to be the approximate posterior over states, inferred using the current (not-updated) belief $p(\theta)$. If the current belief $p(\theta)$ is a Dirichlet distribution with hyperparameter α, then the minimizer $q_T(\theta)$ of F will also be a Dirichlet distribution with hyperparameter α' as given in (7.18)–(7.20). In Appendix 7.B we derive this, and relate it to a well known variational inference algorithm called coordinate ascent variational inference (CAVI).

Action selection

Finally we can also view action selection as the minimization of the variational free energy function (7.26). We can rewrite this function as follows:

$$\begin{aligned} F(\tilde{q}|o_{1:t}) &= \mathbb{E}_{\tilde{q}(\pi)} \left[\ln \tilde{q}(\pi) - \ln p(\pi) + \mathbb{E}_{\tilde{q}(s_{1:T})} \left[\ln \tilde{q}(s_{1:T}) - \ln p(o_{1:t}, s_{1:T}|\pi)\right]\right] + C_{\setminus\pi} \\ &= \mathbb{E}_{\tilde{q}(\pi)} \left[\ln \tilde{q}(\pi) - \ln p(\pi) + F_\pi(\tilde{q}(s_{1:T})|o_{1:t})\right] + C_{\setminus\pi}, \end{aligned}$$

where F_π is defined in (7.27), $C_{\setminus\pi}$ is independent of $\tilde{q}(\pi)$, and we replaced the expectation over $\tilde{q}(\theta)$ by an estimator $\hat{\theta}$ and suppress the dependence of the generative model p on $\hat{\theta}$ in the notation. (This is equivalent to the second line in equation (B.7) in Parr et al., 2022.) What is important to note here, is that the agent is trying to infer an action sequence (policy) π of both future and past actions. We have therefore dropped the dependence on $a_{1:t-1}$ in both F and F_π, and instead π is a sequence of action starting at $\tau = 1$ instead of $\tau = t$. In Section 7.2, we always fixed the past actions to the actions that were actually performed, which is no longer the case here.

We minimize F w.r.t. $\tilde{q}(\pi)$ and $\tilde{q}(s_{1:T})$ and using (7.24) we get for the minimizers respectively

$$q_t(\pi) = \sigma\left(-\ln p(\pi) + F_\pi(q_t(s_{1:T})|o_{1:t})\right), \quad (7.29)$$

and $q_t(s_{1:T})$ is the minimizer of F_π. Now we can use

$$p(\pi) = \sigma(\ln E(\pi) - G(\pi_t|o_{1:t}, a_{1:t-1})),$$

which corresponds to the last line equation (B.7) in Parr et al., 2022.[11] The term $E(\pi)$ is a habit term, signifying what policies the agent is usually exercising. Plugging this back into (7.29) gives

$$q_t(\pi) = \sigma\left(-\ln E(\pi) + G(\pi_t|o_{1:t}, a_{1:t-1}) + F_\pi(q_t|o_{1:t})\right), \tag{7.30}$$

which corresponds to equation (B.9) in Parr et al., 2022.

Remark 7.2. We now try to interpret this derivation conceptually. Note that due to (7.25) we have the following approximate equality:

$$\begin{aligned} F(\tilde{q}|o_{1:t}) &= \mathbb{E}_{\tilde{q}(\pi)}\left[\ln \tilde{q}(\pi) - \ln p(\pi) + F_\pi(\tilde{q}(s_{1:T}), o_{1:t})\right] + C_{\setminus\pi} \\ &\approx \mathbb{E}_{\tilde{q}(\pi)}\left[\ln \tilde{q}(\pi) - \ln p(\pi) - \ln p(o_{1:t}|\pi)\right] + C_{\setminus\pi} \\ &= \mathbb{E}_{\tilde{q}(\pi)}\left[\ln \tilde{q}(\pi) - \ln p(o_{1:t}, \pi)\right] + C_{\setminus\pi}, \end{aligned}$$

which implies that the minimizer $q_t(\pi)$ is approximately equal to the posterior $p(\pi|o_{1:t})$. In other words, this says that we select the policy that is most probable given the past observations. That is, the agent forgets which past actions it has performed, and tries to infer these based on the past observations. Then it tries to find the most likely sequence of future actions to go with this sequence of past actions. Selecting future actions in this way however only makes sense when certain past action sequences make certain future action sequences more likely. For example, let our action space consist of two actions $\{\texttt{left}, \texttt{right}\}$ and policies consist of sequences of actions of length two. Now suppose that the prior over policies dictates that the agent almost certainly performs the policies $(\texttt{left}, \texttt{left})$ or $(\texttt{right}, \texttt{right})$. This implies that having inferred the first action gives the agent more information about the most likely next action. However, if both next actions are equally likely given a first action, according to the prior, then the likelihood term $p(o_{1:t}|\pi)$ does not have any information about the next action. Note that in the calculation of G the past actions are fixed to the actions that have been performed. Therefore G will not make certain future action sequences more likely based on possible past action sequences. Therefore, the term F_π in (7.30) only becomes relevant when the habit term E makes certain future action sequences more likely based on past action sequences.

7.B Learning Preliminaries

Bayesian belief updating

The learning process of an active inference agent is formulated as Bayesian belief updating over the parameters. In general, Bayesian belief updating can be described as follows. Let θ be the parameter of a model p_θ we want to learn and x the output of this model. We start with a prior belief $p(\theta|\alpha)$ which is parametrized by the hyperparameter α. Now our posterior belief about θ is given by the distribution $p(\theta|x, \alpha)$ which is obtained by Bayes' rule. In some special cases[12], the posterior distribution belongs to the same parametrized family as the prior, such that $p(\theta|x, \alpha) = p(\theta|\alpha')$. Then the learning can be summarized by the update from α to α'.

[11] It can be argued that calling this a prior is incorrect, since it actually depends on the observations $o_{1:t}$.

[12] For details, see the theory of conjugate priors.

Categorical model without latent variables

Now we let the model be a categorical distribution over elements $\{x^{(1)}, \dots, x^{(n)}\}$ parametrized by $\theta = (\theta_1, \dots, \theta_n)$. That is,

$$p(x^{(i)}|\theta) = \theta_i, \quad \forall i \in \{1, \dots, n\}.$$

The prior over the parameter θ is given by the Dirichlet distribution parametrized by the hyperparameter $\alpha = (\alpha_1, \dots, \alpha_n)$. That is,

$$p(\theta|\alpha) \propto \prod_i \theta_i^{\alpha_i - 1}.$$

After observing x^*, the posterior is given by

$$\begin{aligned} p(\theta|x^*, \alpha) &\propto p(x^*|\theta)p(\theta|\alpha) \\ &= \prod_i \theta_i^{\alpha_i - 1 + \mathbb{1}_{x^{(i)}}(x^*)}. \end{aligned}$$

Note that this is again a Dirichlet distribution with hyperparameter α' such that

$$\alpha'_i = \alpha_i + \mathbb{1}_{x^{(i)}}(x^*), \quad \forall i \in \{1, \dots, n\}. \tag{7.31}$$

That is, the hyperparameter corresponding to the observation x^* is increased by one. This will make this observation more likely in the updated distribution.

Categorical model with latent variables

Exact posterior

Now we let the model be a joint distribution over the product space of observations x and latent states z, given by $\{x^{(1)}, \dots, x^{(n)}\} \times \{z^{(1)}, \dots, z^{(m)}\}$. The model is parametrized by $\theta = (\theta^D, \theta^A)$ as follows:

$$p(z^{(j)}|\theta^D) = \theta^D_j, \tag{7.32}$$

$$p(x^{(i)}|z^{(j)}, \theta^A) = \theta^A_{ij}. \tag{7.33}$$

The prior over the θ is defined as follows:

$$p(\theta|\alpha) = p(\theta^D|\alpha^D) \prod_j p(\theta^A_{\bullet j}|\alpha^A) \tag{7.34}$$

$$p(\theta^D|\alpha^D) \propto \prod_j \left(\theta^D_j\right)^{\alpha^D_j - 1}, \tag{7.35}$$

$$p(\theta^A_{\bullet j}|\alpha^A) \propto \prod_i \left(\theta^A_{ij}\right)^{\alpha^A_{ij} - 1}, \tag{7.36}$$

where $\theta_{\bullet j}$ denotes the vector $(\theta_{1j}, \ldots, \theta_{nj})$. After observing $x^* = x^{(i^*)}$, the exact posterior is given by

$$\begin{aligned} p(\theta|x^*, \alpha) &\propto p(x^*|\theta)p(\theta|\alpha) \\ &= \sum_j p(x^*|z^{(j)}, \theta^A)p(z^{(j)}|\theta^D)p(\theta^A|\alpha^A)p(\theta^D|\alpha^D) \\ &\propto \sum_j \theta^A_{i^*j}\theta^D_j \left(\prod_{j'}\prod_{i'} \left(\theta^A_{i'j'}\right)^{\alpha^A_{i'j'}-1}\right)\left(\prod_{j''}\left(\theta^D_{j''}\right)^{\alpha^D_{j''}-1}\right) \\ &= \sum_j \left(\prod_{j'}\prod_{i'} \left(\theta^A_{i'j'}\right)^{\alpha^A_{i'j'}-1+\mathbb{1}_{i^*}(i')\mathbb{1}_j(j')}\right)\left(\prod_{j''}\left(\theta^D_{j''}\right)^{\alpha^D_{j''}-1+\mathbb{1}_j(j'')}\right). \end{aligned}$$

Note that this is no longer a Dirichlet distribution. Below we discuss how the Dirichlet distribution shows up in an algorithm for approximating the true posterior.

Mean-field approximation

We can also approximate the posterior over θ by minimizing the following variational free energy function:

$$F(\tilde{q}|x^*) = \mathbb{E}_{\tilde{q}(z,\theta)}\left[\ln \tilde{q}(z, \theta) - \ln p(x^*, z, \theta|\alpha)\right], \tag{7.37}$$

where we assume the approximate posterior over both θ and z factorizes as follows:

$$q(z, \theta) = q(z)q(\theta).$$

The coordinate ascent variational inference (CAVI) algorithm Blei et al., 2017; Bishop and Nasrabadi, 2006 updates the distributions $q(z)$ and $q(\theta)$ iteratively, each time holding one distribution fixed while updating the other. More specifically, we can initialize $q(\theta)$ with our prior belief $p(\theta|\alpha)$ and minimize F w.r.t. $\tilde{q}(z)$. The variational free energy now becomes

$$F(\tilde{q}|x^*, q(\theta)) = \mathbb{E}_{\tilde{q}(z)}\left[\ln \tilde{q}(z) - \mathbb{E}_{q(\theta)}[\ln p(z|x^*, \theta)]\right] + C_{\backslash z}, \tag{7.38}$$

where $C_{\backslash z}$ is independent of $\tilde{q}(z)$. The minimizer $q(z)$ is proportional to

$$q(z) \propto \exp\left(\mathbb{E}_{q(\theta)}[\ln p(z|x^*, \theta)]\right). \tag{7.39}$$

(See equation (7.24).) We then fix this $q(z)$ and optimize F w.r.t. $\tilde{q}(\theta)$ and get

$$F(\tilde{q}|x^*, q(z)) = \mathbb{E}_{\tilde{q}(\theta)}\left[\ln \tilde{q}(\theta) - \mathbb{E}_{q(z)}[\ln p(x^*, z, \theta|\alpha)]\right] + C_{\backslash \theta}, \tag{7.40}$$

where $C_{\backslash \theta}$ is independent of $\tilde{q}(\theta)$. The minimizer $q(\theta)$ is proportional to

$$q(\theta) \propto \exp\left(\mathbb{E}_{q(z)}[\ln p(x^*, z, \theta|\alpha)]\right). \tag{7.41}$$

These two steps are performed iteratively until the beliefs about θ and z have converged.

Note that when $p(x, z, \theta|\alpha)$ is a categorical model with Dirichlet priors, as defined in (7.32)–(7.36), we can rewrite equation (7.41) as follows:

$$\begin{aligned}
q(\theta) &\propto \exp\left(\mathbb{E}_{q(z)}[\ln p(x^*, z, \theta|\alpha)]\right) \\
&= \exp\left(\mathbb{E}_{q(z)}\left[\ln p(x^*|z, \theta^A) + \ln p(z|\theta^D) + \ln p(\theta|\alpha)\right]\right) \\
&= \exp\left(\sum_j q(z^{(j)})\left[\ln p(x^*|z^{(j)}, \theta^A) + \ln p(z^{(j)}|\theta^D)\right]\right) p(\theta|\alpha) \\
&= \prod_j \left(\theta^A_{i^*j}\right)^{q(z^{(j)})} \left(\theta^D_j\right)^{q(z^{(j)})} \left(\theta^D_j\right)^{\alpha^D_j - 1} \prod_i \left(\theta^A_{ij}\right)^{\alpha^A_{ij}-1} \\
&= \prod_j \left(\theta^D_j\right)^{\alpha^D_j + q(z^{(j)}) - 1} \prod_i \left(\theta^A_{ij}\right)^{\alpha^A_{ij} + \mathbb{1}_{i^*}(i) q(z^{(j)}) - 1}
\end{aligned}$$

Note that this is again a Dirichlet distribution with updated parameter $\alpha' = (\alpha^{D\prime}, \alpha^{A\prime})$ given by

$$\alpha^{D\prime}_j = \alpha^D_j + q(z^{(j)}), \tag{7.42}$$

$$\alpha^{A\prime}_{ij} = \alpha^A_{ij} + \mathbb{1}_{i^*}(i) q(z^{(j)}). \tag{7.43}$$

Update (7.42) makes latent states with high $q(z)$ more likely in the updated distribution. Update (7.43) makes sure that latent states are more likely to generate the observation $x^{(i^*)}$, especially those with high $q(z)$. Note the similarity with the update rule (7.31) for categorical models without latent variables.

POMDP model

Now we let p be the generative model from an active inference agent as described in (7.11)–(7.17). Similar to (7.37) the variational free energy now becomes

$$F(\tilde{q}|o_{1:T}, a_{1:T-1}) = \mathbb{E}_{\tilde{q}(s_{1:T},\theta)}\left[\ln \tilde{q}(s_{1:T}, \theta) - \ln p(o_{1:T}, s_{1:T}, \theta|a_{1:T-1}, \alpha)\right]. \tag{7.44}$$

We use the mean-field approximation $q(s_{1:T}, \theta) = q(s_{1:T})q(\theta)$. Equivalent to (7.38)–(7.39) we can start by minimizing F w.r.t. $\tilde{q}(s_{1:T})$ and get a minimizer $q_T(s_{1:T})$ using the current belief about θ. Then, similar to (7.40)–(7.41) we can use this minimizer to write F as follows:

$$F(\tilde{q}|o_{1:T}, a_{1:T-1}, q_T(s_{1:T})) = \mathbb{E}_{\tilde{q}(\theta)}\big[\ln \tilde{q}(\theta) - \mathbb{E}_{q_T(s_{1:T})}[\ln p(o_{1:T}, s_{1:T}, \theta|a_{1:T-1}, \alpha)]\big] + C_{\setminus\theta}, \tag{7.45}$$

and work out the minimizer. This gives

$$\begin{aligned}
q(\theta) &\propto \exp\left(\mathbb{E}_{q_T(s_{1:T})}[\ln p(o_{1:T}, s_{1:T}, \theta | a_{1:T-1}, \alpha)]\right) \\
&= \exp\left(\mathbb{E}_{q_T(s_{1:T})}\left[\sum_{\tau=1}^{T} \ln p(o_\tau|s_\tau, \theta^A) + \ln p(s_1|\theta^D) \right.\right. \\
&\qquad \left.\left. + \sum_{\tau=2}^{T} \ln p(s_\tau|s_{\tau-1}, a_{\tau-1}, \theta^B)\right]\right) p(\theta|\alpha) \\
&= \exp\left(\sum_{s_{1:T}} q_T(s_{1:T})\left[\sum_{\tau=1}^{T} \ln p(o_\tau|s_\tau, \theta^A) + \ln p(s_1|\theta^D) \right.\right. \\
&\qquad \left.\left. + \sum_{\tau=2}^{T} \ln p(s_\tau|s_{\tau-1}, a_{\tau-1}, \theta^B)\right]\right) p(\theta|\alpha) \\
&= \prod_j \left(\theta^D_j\right)^{q_T(s_1^{(j)})} \prod_i \left(\theta^A_{ij}\right)^{\sum_{\tau=1}^{T} \mathbb{1}_{o^{(i)}}(o_\tau) q_T(s_\tau^{(j)})} \\
&\quad \prod_k \left(\theta^B_{jkl}\right)^{\sum_{\tau=2}^{T} q_T(s_\tau^{(j)}) q_T(s_{\tau-1}^{(k)}) \mathbb{1}_{a^{(l)}}(a_{\tau-1})} p(\theta|\alpha) \\
&= \prod_j \left(\theta^D_j\right)^{\alpha^D_j + q_T(s_1^{(j)}) - 1} \prod_i \left(\theta^A_{ij}\right)^{\alpha^A_{ij} + \sum_{\tau=1}^{T} \mathbb{1}_{o^{(i)}}(o_\tau) q_T(s_\tau^{(j)}) - 1} \\
&\qquad \prod_k \left(\theta^B_{jkl}\right)^{\alpha^B_{jkl} + \sum_{\tau=2}^{T} q_T(s_\tau^{(j)}) q_T(s_{\tau-1}^{(k)}) \mathbb{1}_{a^{(l)}}(a_{\tau-1}) - 1}.
\end{aligned}$$

This gives the update rules for α given in (7.18)–(7.20).

Remark 7.3. Note that these update rules are actually just the first iteration of the CAVI algorithm described above. It will therefore in general not minimize the variational free energy in equation (7.44). Instead it minimizes the quantity given in equation (7.28) and (7.45) where $q_T(s_{1:T})$ is fixed. Note however that if one would assume $q_T(s_{1:T})$ to be given, the following variational free energy would be the natural choice to minimize:

$$\mathbb{E}_{\tilde{q}(\theta)}\left[\ln \tilde{q}(\theta) - \ln\left(\mathbb{E}_{q_T(s_{1:T})}[p(o_{1:T}, \theta | s_{1:T}, a_{1:T-1}, \alpha)]\right)\right],$$

since this has as minimizer the exact posterior distribution.

7.C Further details

Equivalent formulations of expected free energy

Recall that the expected free energy in equation (7.2) is given by

$$\begin{aligned}
G(\pi_t | o_{1:t}, a_{1:t-1}) = -\Bigg(&\mathbb{E}_{q_t(o_{t+1:T}|\pi_t)}\left[D_{\mathrm{KL}}\Big(q_t(s_{t+1:T}|o_{t+1:T}, \pi_t) \parallel q_t(s_{t+1:T}|\pi_t)\Big)\right] \\
&+ \mathbb{E}_{q_t(o_{t+1:T}|\pi_t)}\Big[\ln p_C(o_{t+1:T})\Big]\Bigg).
\end{aligned}$$

We can derive the equivalent formulation from equation (7.3) as follows. We first expand the KL divergence term to get

$$G(\pi_t|o_{1:t}, a_{1:t-1}) = \mathbb{E}_{q_t(s_{t+1:T}, o_{t+1:T}|\pi_t)}\Big[\ln q_t(s_{t+1:T}|\pi_t) - \ln q_t(s_{t+1:T}|o_{t+1:T}, \pi_t) - \ln p_C(o_{t+1:T})\Big]. \tag{7.46}$$

Using Bayes' rule we can rewrite

$$\begin{aligned}-\ln q_t(s_{t+1:T}|o_{t+1:T}, \pi_t) = &-\ln q_t(o_{t+1:T}|s_{t+1:T}, \pi_t) - \ln q_t(s_{t+1:T}|\pi_t) \\ &+ \ln q_t(o_{t+1:T}|\pi_t).\end{aligned}$$

Plugging this into (7.46) and using that $q_t(o_{t+1:T}|s_{t+1:T}, \pi_t) = p(o_{t+1:T}|s_{t+1:T})$ we get

$$\begin{aligned}G(\pi_t|o_{1:t}, a_{1:t-1}) &= \mathbb{E}_{q_t(s_{t+1:T}, o_{t+1:T}|\pi_t)}\Big[\ln p(o_{t+1:T}|s_{t+1:T}) + \ln q_t(o_{t+1:T}|\pi_t) \\ &\qquad - \ln p_C(o_{t+1:T})\Big] \\ &= \mathbb{E}_{q_t(s_{t+1:T}|\pi_t)}\Big[\mathrm{H}\big[p(o_{t+1:T}|s_{t+1:T})\big]\Big] \\ &\quad + D_{\mathrm{KL}}\Big(q_t(o_{t+1:T}|\pi_t) \parallel p_C(o_{t+1:T})\Big),\end{aligned}$$

where the last line is equal to equation (7.3).

Independence between state factors and observation modalities

In order to make the computation of state inference more efficient, the agent can use independencies between different state factors and observation modalities (different dimensions of state and observation space). More specifically, we assume that given a state, the different observation modalities are independent, which translates to:

$$p(o_\tau|s_\tau) = \prod_m p(o_\tau^m|s_\tau).$$

We use superscript m and f to denote a specific observation modalities and state factors respectively. Furthermore, we assume a certain state factor to be independent of all other state factors in the same and previous time step, given the same state factor in the previous time step and the last action, i.e.:

$$p(s_\tau|s_{\tau-1}, a_{\tau-1}) = \prod_f p(s_\tau^f|s_{\tau-1}^f, a_{\tau-1}).$$

For a fixed state factor f equation (7.6) and (7.7) now become

$$q_t(s_t^f) \propto p(o_t|s_t^f) \sum_{s_{t-1}^f} p(s_t^f|s_{t-1}^f, a_{t-1}) q_{t-1}(s_{t-1}^f) \tag{7.47}$$

$$q_t(s_\tau^f|a_{1:\tau-1}) = \sum_{s_{\tau-1}^f} p(s_\tau^f|s_{\tau-1}^f, a_{\tau-1}) q_t(s_{\tau-1}^f|a_{1:\tau-2}),$$

and equation (7.10) becomes

$$q_t(s_{1:T}|\pi_t) \propto \prod_{\tau=1}^{t}\prod_m p(o_\tau^m|s_\tau)\ p(s_1) \prod_{\tau=2}^{T}\prod_f p(s_\tau^f|s_{\tau-1}^f, a_{\tau-1}).$$

Fixed point iteration

Equation (7.47) involves the distribution $p(o_t|s_t^f)$. We can however not access this directly. To find an approximate solution, we can use fixed point iteration as follows:

$$q_t^{(i+1)}(s_t^f) \propto \sum_{s_t^{\setminus f}} q_t^{(i)}(s_t^{\setminus f}) p(o_t|s_t) \sum_{s_{t-1}^f} p(s_t^f|s_{t-1}^f, a_{t-1}) q_{t-1}(s_{t-1}^f),$$

where $\setminus f$ denotes the set of all state factors apart from f.

Softmax function

Definition 7.1. Let $\mathcal{S} = \{x^{(1)}, \dots, x^{(n)}\}$ be a finite set and $\mu : \mathcal{S} \to \mathbb{R}$ a function. The *softmax function* σ is given by

$$\sigma(\mu(x^{(i)})) = \frac{e^{\mu(x^{(i)})}}{\sum_j e^{\mu(x^{(j)})}}. \tag{7.48}$$

Note that $\sigma(\mu(x^{(i)})) > 1$ and $\sum_{\mathcal{S}} \sigma(\mu(x^{(i)})) = 1$. Therefore the softmax function can be used to turn μ into a probability distribution.

References

Bishop, Christopher M and Nasser M Nasrabadi (2006). *Pattern recognition and machine learning*. Vol. 4. 4. Springer.

Blei, David M, Alp Kucukelbir, and Jon D McAuliffe (2017). "Variational inference: A review for statisticians". In: *Journal of the American statistical Association* 112.518, pp. 859–877.

Da Costa, Lancelot, Thomas Parr, Noor Sajid, Sebastijan Veselic, Victorita Neacsu, and Karl Friston (2020). "Active inference on discrete state-spaces: A synthesis". In: *Journal of Mathematical Psychology* 99, p. 102447. ISSN: 0022-2496. DOI: `https://doi.org/10.1016/j.jmp.2020.102447`.

Da Costa, Lancelot, Samuel Tenka, Dominic Zhao, and Noor Sajid (2024). "Active inference as a model of agency". In: *arXiv preprint arXiv:2401.12917*.

Heins, Conor (n.d.). *Active Inference Demo: T-Maze Environment*. `https://pymdp-rtd.readthedocs.io/en/latest/notebooks/tmaze_demo.html`.

Heins, Conor, Beren Millidge, Daphne Demekas, Brennan Klein, Karl Friston, Iain Couzin, and Alexander Tschantz (2022). "pymdp: A Python library for active inference in discrete state spaces". In: *The Journal of Open Source Software* 7.73, p. 4098.

Millidge, Beren, Alexander Tschantz, and Christopher L Buckley (2021). "Whence the expected free energy?" In: *Neural Computation* 33.2, pp. 447–482.

Parr, Thomas, Giovanni Pezzulo, and Karl J Friston (2022). *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press.

Smith, Ryan, Karl J Friston, and Christopher J Whyte (2022). "A step-by-step tutorial on active inference and its application to empirical data". In: *Journal of mathematical psychology* 107, p. 102632.

Wei, Ran (2024). "Value of Information and Reward Specification in Active Inference and POMDPs". In: *arXiv preprint arXiv:2408.06542*.

Curriculum Vitae

Last name	van Oostrum
First name	Jesse
Nationality	Dutch
Date of birth	04-10-1995
Place of birth	the Netherlands
Email	jesse.van@tuhh.de
Address	Kasernenstraße 12, 21073 Hamburg, Germany

09.1999 – 07.2007	Primary school, Montessori School Zuid Hilversum, Netherlands
09.2007 – 07.2013	High school, Gemeentelijk Gymnasium Hilversum, Netherlands
09.2013 – 07.2016	Liberal Arts and Sciences (Mathematics, Physics & Economics) University College Utrecht | GPA: 3.82/4.0, Magna cum Laude
09.2016 – 07.2017	Pre-Master Mathematics, Utrecht University | GPA: 8.5/10
09.2017 – 06.2018	Master of Advanced Studies in Mathematical Statistics University of Cambridge | Graduated with Honours
07.2018 – 09.2018	Computational Genomics Researcher Wellcome Sanger Institute, Hinxton, UK
10.2018 – 01.2019	Self study differential geometry
02.2019 – 08.2019	Information Geometry Researcher Collegio Carlo Alberto, Turin, Italy
10.2019 – 02.2020	Machine Learning Intern, Geronimo.AI, Delft, the Netherlands
03.2020 – 07.2020	Digital Strategy Consultant Intern, BLOOM, Amsterdam, the Netherlands
10.2020 – present	PhD Candidate, Hamburg University of Technology & Max Planck Institute for Mathematics in the Sciences, Leipzig