Journal of Machine Learning Research 26 (2025) 1-37 Submitted 4/24; Revised 6/25; Published 9/25

On the Natural Gradient of the Evidence Lower Bound

Nihat Ay NIHAT.AY@TUHH.DE
Institute for Data Science Foundations
Hamburg University of Technology
21073 Hamburg, Germany

Santa Fe Institute
Santa Fe, NM 87501, USA

Leipzig University
04109 Leipzig, Germany

Jesse van Oostrum JESSE.VAN@TUHH.DE
Institute for Data Science Foundations
Hamburg University of Technology
21073 Hamburg, Germany

Adwait Datar ADWAIT.DATAR@TUHH.DE
Institute for Data Science Foundations
Hamburg University of Technology
21073 Hamburg, Germany

Editor: Dan Alistarh

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

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 $\mathrm{grad}_p \mathcal{L}$ in $p \in \mathcal{P}$ via the equation

$$d\mathcal{L}_p(A) \;=\; g_p(\mathrm{grad}_p \mathcal{L}, A) \tag{1}$$

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 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 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 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 5 relates these results to the evidence and its lower bound, thereby making statements on their respective natural gradients. Section 6 concludes with a result that is particularly helpful when dealing specifically with Bayesian graphical models.

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

$$\varphi : \Theta \to \mathcal{M} \subseteq \mathcal{P}, \qquad \theta \mapsto p_\theta. \tag{2}$$

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

$$p(x_V) := \sum_{x_H} p(x_V, x_H). \tag{3}$$

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 (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,

$$D(p^* \| p) \ := \ \sum_{x_V} p^*(x_V) \ln \frac{p^*(x_V)}{p(x_V)}. \tag{4}$$

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

$$\text{EVIDENCE}(p) \ := \ \sum_{x_V} p^*(x_V) \ln p(x_V), \tag{5}$$

which we also refer to as the *evidence* without explicitly highlighting the fact that it is an integrated quantity. Minimizing the KL-divergence (4) is then equivalent to maximizing the evidence (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

$$\begin{aligned}
\mathrm{EVIDENCE}(\pi_V(p)) \;&=\; -\sum_{x_V,x_H} p^*(x_V)q(x_H|x_V)\ln\frac{q(x_H|x_V)}{p(x_V,x_H)} \\
&\quad +\sum_{x_V} p^*(x_V)\sum_{x_H} q(x_H|x_V)\ln\frac{q(x_H|x_V)}{p(x_H|x_V)} \qquad (6)\\
&\geq\; -\sum_{x_V,x_H} p^*(x_V)q(x_H|x_V)\ln\frac{q(x_H|x_V)}{p(x_V,x_H)} \qquad (7)\\
&=:\; \mathrm{ELBO}(q,p). \qquad (8)
\end{aligned}$$

The inequality (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). The bound (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. 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) of KL-divergences,

$$\mathrm{GAP}(q,p) \;:=\; \sum_{x_V,x_H} p^*(x_V)q(x_H|x_V)\ln\frac{q(x_H|x_V)}{p(x_H|x_V)}. \qquad (9)$$

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 4 and 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.

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

$$p \;=\; \sum_x p(x)\,\delta^x, \qquad (10)$$

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*

$$g_p^{\mathrm{FR}}(A,B) \;=\; \sum_x \frac{1}{p(x)} A(x)B(x), \tag{11}$$

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

$$D(q\|p) \;=\; \sum_x q(x) \ln \frac{q(x)}{p(x)}. \tag{12}$$

Note that we already used the KL-divergence to define the primary objective function (4). We can express the Fisher-Rao gradients of the KL-divergence in both arguments:

$$\begin{aligned}
\mathrm{grad}_p D(q\|\cdot) \;&=\; \sum_x (p(x) - q(x))\,\delta^x \\
&=\; p - q \;\in\; T_p\mathcal{P}, \\
\mathrm{grad}_q D(\cdot\|p) \;&=\; \sum_x q(x) \left(\ln \frac{q(x)}{p(x)} - \sum_{x'} q(x') \left(\ln \frac{q(x')}{p(x')} \right) \right) \delta^x \\
&=\; q \left(\ln \frac{q}{p} - \mathbb{E}_q \left(\ln \frac{q}{p} \right) \right) \;\in\; T_q\mathcal{P}.
\end{aligned} \tag{13}$$

These gradients satisfy the defining condition (1), where $\mathcal{L}$ is the KL-divergence (12) in the first and the second argument, respectively, and $g = g^{\mathrm{FR}}$ as defined by (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 (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

$$d\pi_V(A)(x_V) \;=\; \sum_{x_H} A(x_V, x_H). \tag{14}$$

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 $\mathrm{grad}^{\mathcal{M}}$, and the gradient on $\mathcal{M}_V$, denoted by $\mathrm{grad}^{\mathcal{M}_V}$.

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 (4), which we should minimize, and the mean evidence (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

$$\dot\gamma(t) \;=\; -\mathrm{grad}^{\mathcal{M}}_{\gamma(t)}(\mathcal{L} \circ \pi_V), \qquad \gamma(0) = p,$$

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

$$\begin{aligned} \dot\sigma(t) &= d\pi_V\left(\dot\gamma(t)\right) \\ &= -d\pi_V\left(\mathrm{grad}^{\mathcal{M}}_{\gamma(t)}(\mathcal{L} \circ \pi_V)\right). \end{aligned} \tag{15}$$

Following the Fisher-Rao gradient on $\mathcal{M}_V$, however, would require

$$\dot{\sigma}(t) \; = \; -\mathrm{grad}^{\mathcal{M}_V}_{\sigma(t)} \mathcal{L}. \tag{16}$$

The vector fields defined by the respective RHS of (15) and (16) are typically different, but they point, at least, in a similar direction, which is shown in the following proposition.

Proposition 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,*

$$\mathrm{grad}^{\mathcal{M}}_p \left(\mathcal{L} \circ \pi_V\right) \; = \; 0 \quad \Leftrightarrow \quad \mathrm{grad}^{\mathcal{M}_V}_{\pi_V(p)} \mathcal{L} \; = \; 0.$$

Furthermore, if one of the two gradients does not vanish, we have

$$\left\langle d\pi_V \left(\mathrm{grad}^{\mathcal{M}}_p \left(\mathcal{L} \circ \pi_V\right)\right), \mathrm{grad}^{\mathcal{M}_V}_{\pi_V(p)} \mathcal{L} \right\rangle \; > \; 0.$$

Proof For an arbitrary $A \in T_p\mathcal{M}$, we have

$$\begin{aligned} \left\langle \mathrm{grad}^{\mathcal{M}}_p \left(\mathcal{L} \circ \pi_V\right), A \right\rangle \; &= \; d\left(\mathcal{L} \circ \pi_V\right)_p (A) \\ &= \; \left(d\mathcal{L}_{\pi_V(p)} \circ d\pi_V\right)(A) \\ &= \; d\mathcal{L}_{\pi_V(p)} \left(d\pi_V(A)\right) \\ &= \; \left\langle \mathrm{grad}^{\mathcal{M}_V}_{\pi_V(p)} \mathcal{L}, d\pi_V(A) \right\rangle. \end{aligned}$$

This implies that $\mathrm{grad}^{\mathcal{M}}_p \left(\mathcal{L} \circ \pi_V\right)$ vanishes if and only if $\mathrm{grad}^{\mathcal{M}_V}_{\pi_V(p)} \mathcal{L}$ 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 $A = \mathrm{grad}^{\mathcal{M}}_p \left(\mathcal{L} \circ \pi_V\right) \neq 0$, we obtain

$$\begin{aligned} \left\langle d\pi_V \left(\mathrm{grad}^{\mathcal{M}}_p \left(\mathcal{L} \circ \pi_V\right)\right), \mathrm{grad}^{\mathcal{M}_V}_{\pi_V(p)} \mathcal{L} \right\rangle \; &= \; \left\langle \mathrm{grad}^{\mathcal{M}}_p \left(\mathcal{L} \circ \pi_V\right), \mathrm{grad}^{\mathcal{M}}_p \left(\mathcal{L} \circ \pi_V\right) \right\rangle \\ &> \; 0. \end{aligned}$$

■

We now ask the question under which conditions the two gradient fields of Proposition 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 3. For such models, a projected solution curve σ in $\mathcal{M}_V$ satisfies equation (16).

Definition 2 (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*

$$T_p\mathcal{M} \; = \; \left(T_p\mathcal{M} \cap \mathcal{H}_p\right) \oplus \left(T_p\mathcal{M} \cap \mathcal{V}_p\right).$$

If the model is cylindrical in all non-singular points $p \in \mathcal{M}$ then we call it (pointwise) cylindrical.

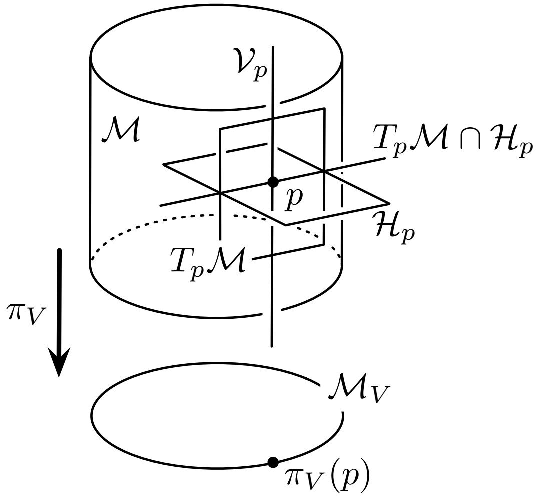

Figure 1: Illustration of a cylindrical model $\mathcal{M}$ in terms of a cylinder, the Cartesian product of a circle with a finite interval. The tangent space $T_p\mathcal{M}$ equals the sum of its intersections with $\mathcal{H}_p$ and $\mathcal{V}_p$.

See Figure 1 for an illustration of a cylindrical model and Appendix 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

$$g_p^{\mathrm{FR}}(A,B) \;=\; g_{\pi_V(p)}^{\mathrm{FR}}\left(d\pi_V(A), d\pi_V(B)\right), \tag{17}$$

whenever p is *admissible*. The equality (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 (van Oostrum et al., 2023).)

Theorem 3 (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,*

$$d\pi_V\left(\mathrm{grad}_p^{\mathcal{M}}(\mathcal{L} \circ \pi_V)\right) \;=\; \mathrm{grad}_{\pi_V(p)}^{\mathcal{M}_V}\mathcal{L}. \tag{18}$$

Note that the gradient on the LHS of (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 3 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 4 *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 2 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 $\Pi_p(X) = \Pi_p(B + C) = \Pi_p(B) + \Pi_p(C) = B \in \mathcal{V}_p$. ■

4. The Extended Problem with Hidden Units

In this section, we are going to relate the minimization of the KL-divergence (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 (13)):

$$\mathrm{grad}_p^{\mathcal{P}_V} D(p^*\|\cdot) \;=\; p - p^* \;\in\; T_p\mathcal{P}_V. \tag{19}$$

For the gradient on the model $\mathcal{M}_V$, we then have to project the gradient (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

$$\mathrm{grad}_p^{\mathcal{M}_V} D(p^*\|\cdot) \;=\; \Pi_p(p - p^*) \;\in\; T_p\mathcal{M}_V, \tag{20}$$

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 (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 (4), or equivalently the maximization of the evidence, to the corresponding maximization of the evidence lower bound, which will be finally addressed in Section 5.

It is well-known that the minimization of the KL-divergence (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

$$\pi_{\mathcal{Q}}(p)(x_V, x_H) \;=\; p^*(x_V)p(x_H|x_V). \tag{21}$$

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 3, they will be equivalent whenever the model $\mathcal{M}$ is cylindrical.

Theorem 5 **(a)** *Consider first the function $D(\mathcal{Q} \| \cdot)$ on $\mathcal{P}$. Then*

$$\mathrm{grad}_p^{\mathcal{P}} D(\mathcal{Q} \| \cdot) \;=\; p - \pi_{\mathcal{Q}}(p), \tag{22}$$

where $\pi_{\mathcal{Q}}(p)$ is defined by (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 (22) onto $T_p\mathcal{M}$, that is

$$\mathrm{grad}_p^{\mathcal{M}} D(\mathcal{Q}\|\cdot) \;=\; \Pi_p(p - \pi_{\mathcal{Q}}(p)), \tag{23}$$

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*

$$d\pi_V\left(\mathrm{grad}_p^{\mathcal{M}} D(\mathcal{Q}\|\cdot)\right) \;=\; \mathrm{grad}_{\pi_V(p)}^{\mathcal{M}_V} D(p^*\|\cdot). \tag{24}$$

In particular, the equality (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 (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)

$$\begin{aligned}
&\left(\mathrm{grad}_p^{\mathcal{P}} D(\mathcal{Q}\|\cdot)\right)(x_V, x_H)\\
&\;=\; p(x_V, x_H)\left(-\frac{p^*(x_V)}{p(x_V)} + \sum_{x'_V, x'_H} p(x'_V, x'_H)\frac{p^*(x_V)}{p(x_V)}\right)\\
&\;=\; p(x_V, x_H)\left(-\frac{p^*(x_V)}{p(x_V)} + 1\right)\\
&\;=\; p(x_V, x_H) - p(x_H|x_V)p^*(x_V)\\
&\;=\; p(x_V, x_H) - \pi_{\mathcal{Q}}(p)(x_V, x_H).
\end{aligned}$$

This proves equation (22), and equation (23) follows immediately from that. Finally, the invariance (24) is a direct consequence of Theorem 3. ■

The gradients considered in Theorem 5 are graphically illustrated in Figure 2. Theorem 5 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 10. Furthermore, the gradient (23) of the function $D(\mathcal{Q}\|\cdot)$, defined

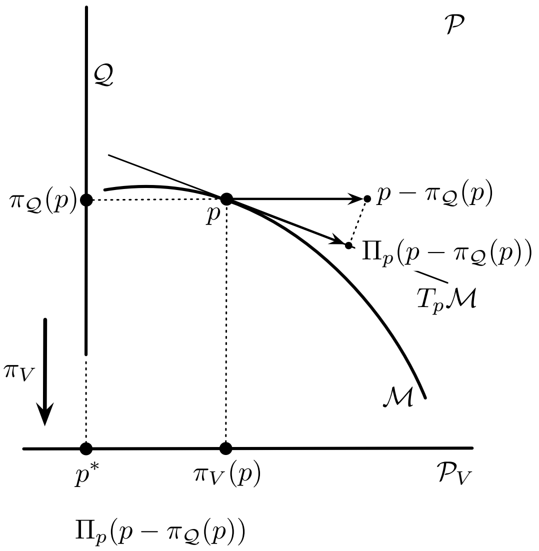

Figure 2: Illustration of the gradients considered in Theorem 5.

on $\mathcal{M}$, can now be compared with the gradient (20) of the original function $D(p^* \| \cdot)$ which is defined on $\mathcal{M}_V$. According to the invariance (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 5. 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 (24) arises from the restriction of the optimization to $\mathcal{M}$.

We can illustrate the invariance (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.

$$\begin{array}{rcl} \pi_V : \mathcal{P} & \to & \mathcal{P}_V \\ p(x_s, x_{t_1}, x_{t_2}) & \mapsto & p(x_{t_1}, x_{t_2}) = \displaystyle\sum_{x_s} p(x_s, x_{t_1}, x_{t_2}). \end{array}$$

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

$$\dot{\sigma}_0(t) \;=\; -\mathrm{grad}^{\mathcal{M}_V}_{\sigma_0(t)} D(p^* \| \cdot), \quad \sigma_0(0) \;=\; \pi_V(p). \tag{25}$$

Furthermore, we let the curve γ_1 in $\mathcal{M}$ be the solution of the differential equation

$$\dot{\gamma}_1(t) \;=\; -\mathrm{grad}^{\mathcal{M}}_{\gamma_1(t)} D(\mathcal{Q}\|\cdot), \quad \gamma_1(0) \;=\; p, \tag{26}$$

and $\sigma_1 = \pi_V \circ \gamma_1$ in $\mathcal{M}_V$ be the projection of γ_1.

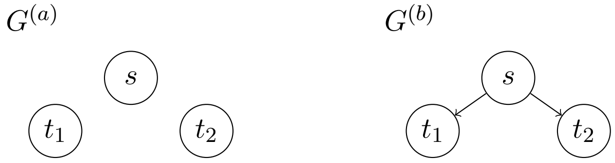

Figure 3: Graphical representations of the models $\mathcal{M}^{(a)}$ and $\mathcal{M}^{(b)}$.

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 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,

$$\mathcal{M}^{(a)} \;=\; \{p \in \mathcal{P} \,:\, p(x_s, x_{t_1}, x_{t_2}) \;=\; p(x_s)p(x_{t_1})p(x_{t_2})\}. \tag{27}$$

Graphically, these are all the distributions factorizing over the graph $G^{(a)}$ in Figure 3. It can be shown that this model is cylindrical.[1] Owing to Theorem 5, we know that the curves σ_0 and σ_1 are identical. This is shown in Figure 5, where the model $\mathcal{M}_V^{(a)}$ 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

$$\mathcal{M}^{(b)} \;=\; \{p \in \mathcal{P} \,:\, p(x_s, x_{t_1}, x_{t_2}) \;=\; p(x_s)p(x_{t_1}|x_s)p(x_{t_2}|x_s)\}. \tag{28}$$

It consists of those distributions that factorize over the graph $G^{(b)}$ in Figure 3. In Example 3 of Appendix A we show that this model is not cylindrical. The model $\mathcal{M}_V^{(b)}$ is in this case the full simplex $\mathcal{P}_V$. Figure 6 (top) shows the trajectories of the curves σ_0 and σ_1 defined by (25) and (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 (bottom-left) shows the same trajectories in coordinates as functions of time and Figure 6 (bottom-right) shows the KL-divergence evaluated 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 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

[1] See Example 1 in Appendix A for a two-node example of this.

curves σ_0 (dashed blue) of the gradient of $D(p^*\|\cdot)$ are straight lines. This is a consequence of (13) and has more general implications on the learning (Datar and Ay, 2025).

In order to measure the deviation from the invariance (24) we evaluate the cosine similarity of the involved vector fields. More precisely, we compute the cosine similarity between the vectors $\mathrm{grad}^{\mathcal{M}_V}_{\pi_V(p)} D(p^*\|\cdot)$ and $d\pi_V\left(\mathrm{grad}^{\mathcal{M}}_p D(\mathcal{Q}\|\cdot)\right)$ in $T_p\mathcal{M}^{(b)}_V$, 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

$$\cos(\alpha) \;=\; \frac{\langle A, B\rangle}{\|A\|\|B\|}, \tag{29}$$

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 $T_p\mathcal{M}^{(b)}_V$. 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 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 1.

5. The Natural Gradient of the Evidence Lower Bound

In Theorem 5, 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 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

$$\begin{aligned} D(\mathcal{Q}\|p) \;&\le\; D(\mathcal{Q}\|p) + D(q\|\pi_{\mathcal{Q}}(p)) && (30)\\ &=\; D(\pi_{\mathcal{Q}}(p)\|p) + D(q\|\pi_{\mathcal{Q}}(p)). \\ &=\; D(q\|p) && (31)\end{aligned}$$

It can be easily verified that the additional term in (30), $D(q\|\pi_{\mathcal{Q}}(p))$, coincides with $\mathrm{GAP}(q,p)$ as defined by (9). Now, instead of taking the gradient of $D(\mathcal{Q}\|\cdot)$ we take the gradient of the upper bound (31), with a fixed q, and analyze the effect of this replacement. For that, let us first rewrite this upper bound as follows:

$$D(q\|p) \;=\; D(\mathcal{Q}\|p) + \mathrm{GAP}(q,p). \tag{32}$$

For the gradient of $D(q\|\cdot)$ in $\mathcal{P}$, we obtain

$$\begin{aligned} \mathrm{grad}^{\mathcal{P}}_p D(q\|\cdot) \;&=\; \mathrm{grad}^{\mathcal{P}}_p\, D(\mathcal{Q}\|\cdot) + \mathrm{grad}^{\mathcal{P}}_p\, \mathrm{GAP}(q,\cdot) \\ &=\; (p - \pi_{\mathcal{Q}}(p)) + (\pi_{\mathcal{Q}}(p) - q). \qquad \text{(by (22) and (13))} && (33)\end{aligned}$$

It is easy to see that the first difference vector in (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:

$$\begin{aligned} d\pi_V(\pi_{\mathcal{Q}}(p) - q) &= \sum_{x_H} (p^*(x_V) p(x_H|x_V) - q(x_V, x_H)) \qquad \text{(by (14))} \\ &= p^*(x_V) \sum_{x_H} (p(x_H|x_V) - q(x_H|x_V)) \\ &= p^*(x_V)(1-1) \;=\; 0, \end{aligned}$$

or equivalently,

$$d\pi_V\left(\mathrm{grad}_p^{\mathcal{P}}\, \mathrm{GAP}(q,\cdot)\right) \;=\; 0.$$

In summary, we obtain

$$\begin{aligned} d\pi_V\left(\mathrm{grad}_p^{\mathcal{P}} D(q\|\cdot)\right) &= d\pi_V\left(p - \pi_{\mathcal{Q}}(p)\right) \\ &= \pi_V(p) - p^* \\ &= \mathrm{grad}_{\pi_V(p)}^{\mathcal{P}_V} D(p^*\|\cdot). \end{aligned}$$

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 *Let $\mathcal{M}$ be a cylindrical model in $\mathcal{P}$, let $p \in \mathcal{M}$ be admissible, and let $q \in \mathcal{Q}$. Then*

$$d\pi_V\left(\mathrm{grad}_p^{\mathcal{M}}\, \mathrm{GAP}(q,\cdot)\right) \;=\; 0, \tag{34}$$

and therefore

$$d\pi_V\left(\mathrm{grad}_p^{\mathcal{M}} D(q\|\cdot)\right) \;=\; \mathrm{grad}_{\pi_V(p)}^{\mathcal{M}_V} D(p^*\|\cdot). \tag{35}$$

Proof We begin with the gradient of $\mathrm{GAP}(q,\cdot)$:

$$\mathrm{grad}_p^{\mathcal{M}}\, \mathrm{GAP}(q,\cdot) \;=\; \Pi_p\left(\mathrm{grad}_p^{\mathcal{P}}\, \mathrm{GAP}(q,\cdot)\right) \;=\; \Pi_p\left(\pi_{\mathcal{Q}}(p) - q\right) \;=\; \Pi_p\left((p-q)^{\mathcal{V}}\right).$$

We know, by definition, that $(p-q)^{\mathcal{V}}$ is contained in $\mathcal{V}_p$. According to Lemma 4, 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 (34). With this, we have

$$\begin{aligned} d\pi_V\left(\mathrm{grad}_p^{\mathcal{M}} D(q\|\cdot)\right) &= d\pi_V\left(\mathrm{grad}_p^{\mathcal{M}} D(\mathcal{Q}\|\cdot)\right) + d\pi_V\left(\mathrm{grad}_p^{\mathcal{M}}\, \mathrm{GAP}(q,\cdot)\right) \\ &= d\pi_V\left(\mathrm{grad}_p^{\mathcal{M}} D(\mathcal{Q}\|\cdot)\right), \end{aligned}$$

and with (24) of Theorem 5 we finally obtain (35). ■

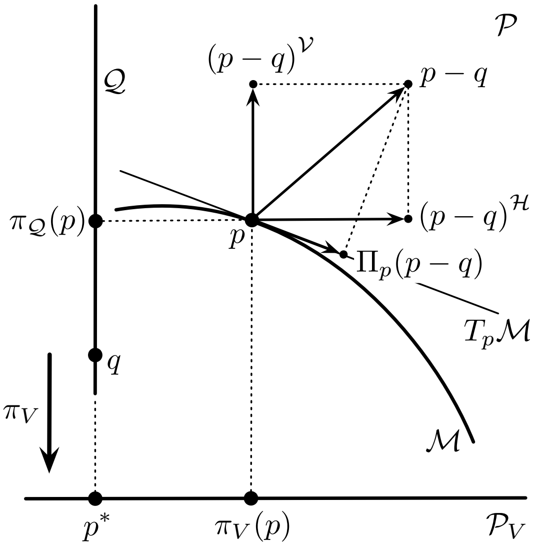

Figure 4: Illustration of gradients considered in Theorem 6.

The gradients considered in Theorem 6 are graphically illustrated in Figure 4. Note that while the invariance (35) appears to be very similar to the invariance (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 (32), and the function $\mathrm{GAP}(q,\cdot)$, the second term on the RHS of (32), which varies only in vertical direction so that the $d\pi_V$-image of its gradient vanishes. Note that this result, expressed by (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 B.

We illustrate the invariance (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 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 5 and 6 to conclude that the curves σ_0, σ_1 and σ_2 are identical. This is depicted in Figure 5 where the blue grid represents the cylindrical model $\mathcal{M}_V^{(a)}$, 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 (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 (bottom-left) shows the trajectories in coordinates as functions of time and Figure 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

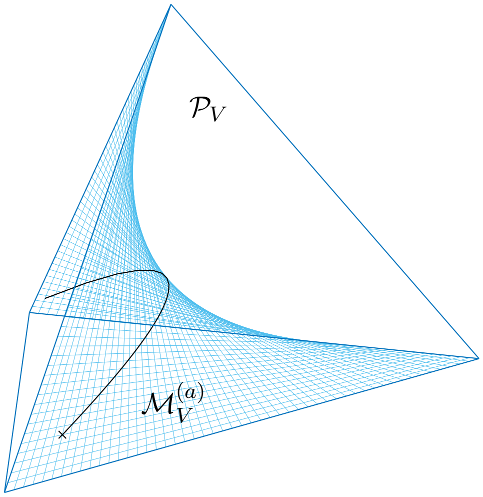

Figure 5: The blue grid represents the set of independent probability distributions over two random variables corresponding to the cylindrical model $\mathcal{M}_V^{(a)}$, as a subset of $\mathcal{P}_V$. The cylindrical model $\mathcal{M}^{(a)}$ is defined in (27) and depicted in Figure 3. The solid black line shows the overlapping trajectories σ_0, σ_1 and σ_2, where σ_0 satisfies $\dot{\sigma}_0(t) = -\mathrm{grad}^{\mathcal{M}_V^{(a)}}_{\sigma_0(t)} D(p^*\|\cdot)$, $\sigma_1 = \pi_V \circ \gamma_1$ is the projection of the negative gradient curve γ_1 satisfying $\dot{\gamma}_1(t) = -\mathrm{grad}^{\mathcal{M}^{(a)}}_{\gamma_1(t)} D(\mathcal{Q}\|\cdot)$ and $\sigma_2 = \pi_V \circ \gamma_2$ is the projection of the negative gradient curve γ_2 satisfying $\dot{\gamma}_2(t) = -\mathrm{grad}^{\mathcal{M}^{(a)}}_{\gamma_2(t)} D(q\|\cdot)$. Theorem 5 and 6 imply that σ_0, σ_1 and σ_2 are identical.

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 (29). We sample p in $\mathcal{M}^{(b)}$ according to Jeffrey's prior and plot the results in the histograms in Figure 7 (right). As one can see, the points are still close to 1 but on average lower than the cosine similarities between $\mathrm{grad}^{\mathcal{M}_V}_{\pi_V(p)} D(p^*\|\cdot)$ and $d\pi_V\left(\mathrm{grad}^{\mathcal{M}}_p D(\mathcal{Q}\|\cdot)\right)$. 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 (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

$$\begin{aligned} D(q\|p) &= \sum_{x_V, x_H} p^*(x_V)\, q(x_H|x_V) \ln \frac{p^*(x_V)\, q(x_H|x_V)}{p(x_V, x_H)} \\ &= \underbrace{\sum_{x_V} p^*(x_V) \ln p^*(x_V)}_{\leq 0} + \sum_{x_V, x_H} p^*(x_V)\, q(x_H|x_V) \ln \frac{q(x_H|x_V)}{p(x_V, x_H)} \\ &= -\mathrm{ELBO}(q, p) + \text{const.}, \end{aligned}$$

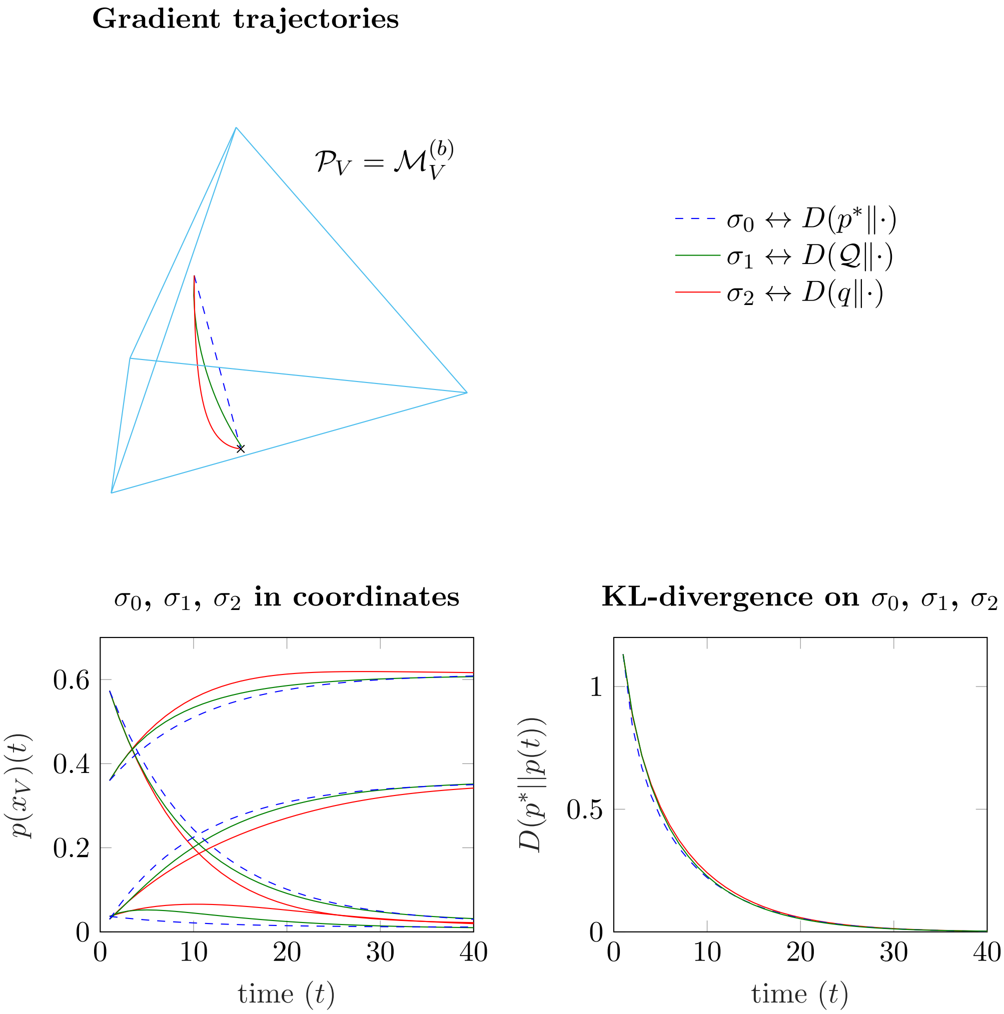

Figure 6: These figures study different gradient trajectories on the model $\mathcal{M} = \mathcal{M}^{(b)}$ defined in (28) and discussed on pages 15 and 18. The top figure shows the curve σ_0 satisfying $\dot{\sigma}_0(t) = -\mathrm{grad}^{\mathcal{M}_V}_{\sigma_0(t)} D(p^* \| \cdot)$ (dashed-blue), the curve $\sigma_1 = \pi_V \circ \gamma_1$, with γ_1 satisfying $\dot{\gamma}_1(t) = -\mathrm{grad}^{\mathcal{M}}_{\gamma_1(t)} D(\mathcal{Q} \| \cdot)$ (solid green) and the curve $\sigma_2 = \pi_V \circ \gamma_2$, with γ_2 satisfying $\dot{\gamma}_2(t) = -\mathrm{grad}^{\mathcal{M}}_{\gamma_2(t)} D(q \| \cdot)$ (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.

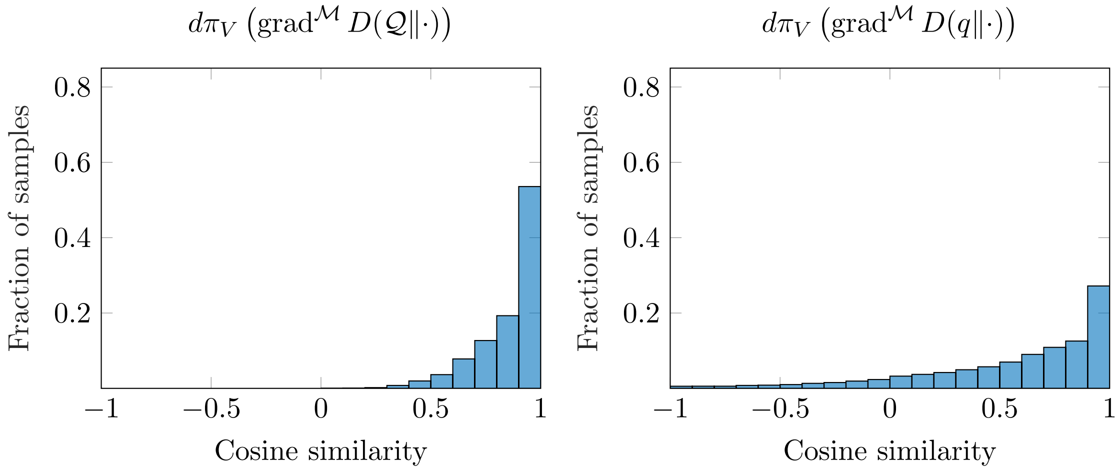

Figure 7: Histograms showing the cosine similarity between $d\pi_V \left(\mathrm{grad}^{\mathcal{M}} D(\mathcal{Q}\|\cdot)\right)$ and $\mathrm{grad}^{\mathcal{M}_V} D(p^*\|\cdot)$ (left) and the cosine similarity between $d\pi_V \left(\mathrm{grad}^{\mathcal{M}} D(q\|\cdot)\right)$ and $\mathrm{grad}^{\mathcal{M}_V} D(p^*\|\cdot)$ (right), both with respect to the Fisher-Rao metric.

where $\mathrm{ELBO}(q, p)$ is defined by (8). Thus, the gradient of the function $D(q\|\cdot)$ will be the same as the gradient of $-\mathrm{ELBO}(q, \cdot)$, because the two functions differ only by a constant. More precisely, we have for all non-singular points p of $\mathcal{M}$

$$\mathrm{grad}_p^{\mathcal{M}} \mathrm{ELBO}(q, \cdot) \;=\; -\mathrm{grad}_p^{\mathcal{M}} D(q\|\cdot).$$

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.

Corollary 7 *Let $\mathcal{M}$ be a cylindrical model in $\mathcal{P}$, let $p \in \mathcal{M}$ be admissible, and let $q \in \mathcal{Q}$. Then*

$$d\pi_V \left(\mathrm{grad}_p^{\mathcal{M}} \mathrm{ELBO}(q, \cdot)\right) \;=\; \mathrm{grad}_{\pi_V(p)}^{\mathcal{M}_V} \mathrm{EVIDENCE}. \tag{36}$$

In particular, the invariance (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:

$$\begin{tikzcd}
D(\mathcal{Q}\|\cdot) \arrow[r, squiggly] \arrow[d, bend right, leftrightarrow, "\text{Theorem 5}"'] & D(q\|\cdot) \arrow[r, squiggly] \arrow[dl, leftrightarrow, "\text{Theorem 6}"'] & \mathrm{ELBO}(q,\cdot) \arrow[dl, leftrightarrow, "\text{Corollary 7}"'] & \text{lifted objective functions on } \mathcal{M} \\
D(p^*\|\cdot) \arrow[r, squiggly] \arrow[u, squiggly] & \mathrm{EVIDENCE} & & \text{primary objective functions on } \mathcal{M}_V
\end{tikzcd}$$

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 7. 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 5. 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. 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 7. 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 B. However, empirical case studies in higher dimensions are required for more conclusive statements.

We conclude this section with Remarks 8 and 9 which outline further research directions as possible continuations of the present work.

Remark 8 *Theorem 6 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 (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 (van 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 9 *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:*

$$\text{(a)}\quad \mathcal{M} \subseteq \widetilde{\mathcal{M}}, \qquad \text{(b)}\quad \pi_V(\mathcal{M}) = \pi_V(\widetilde{\mathcal{M}}), \qquad \text{(c)}\quad \widetilde{\mathcal{M}} \textit{ is cylindrical.}$$

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 to a cylindrical extension and obtain

$$\begin{aligned} d\pi_V\left(\mathrm{grad}_p^{\widetilde{\mathcal{M}}} D(q\|\cdot)\right) &= d\pi_V\left(\mathrm{grad}_p^{\mathcal{M}} D(q\|\cdot)\right) + d\pi_V\left(\mathrm{grad}_p^{\perp} D(q\|\cdot)\right) \\ &= \mathrm{grad}_{\pi_V(p)}^{\mathcal{M}_V} D(p^*\|\cdot), \end{aligned}$$

where $\mathrm{grad}_p^{\perp} D(q\|\cdot)$ denotes the projection of $\mathrm{grad}_p^{\widetilde{\mathcal{M}}} D(q\|\cdot)$ onto the orthogonal complement of $T_p\mathcal{M}$ in $T_p\widetilde{\mathcal{M}}$. This finally gives us the following generalization of (35):

$$d\pi_V\left(\mathrm{grad}_p^{\mathcal{M}} D(q\|\cdot)\right) \;=\; \mathrm{grad}_{\pi_V(p)}^{\mathcal{M}_V} D(p^*\|\cdot) - d\pi_V\left(\mathrm{grad}_p^{\perp} D(q\|\cdot)\right). \tag{37}$$

Equation (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. 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

$$p_\theta(x) \;:=\; p(x;\theta) \;:=\; \prod_{s \in N} p(x_s|x_{\mathrm{pa}(s)}; \theta_s). \tag{38}$$

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 (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

$$\begin{aligned} \partial_{s,k}(x;\theta) \;&:=\; \frac{\partial}{\partial \theta_{s,k}}\, p(x;\theta) \\ &=\; p(x;\theta)\, \frac{\partial}{\partial \theta_{s,k}}\, \ln p(x;\theta) \\ &=\; p(x;\theta)\, \frac{\partial}{\partial \theta_{s,k}}\, \ln p(x_s|x_{\mathrm{pa}(s)};\theta_s). \end{aligned} \tag{39}$$

The vectors $\partial_{s,k}(\theta)$, $s \in N$, $k = 1, \ldots, d_s$, 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 10 *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*

$$g_\theta^{\mathrm{FR}}\left(\partial_{s,k}(\theta), \partial_{t,l}(\theta)\right) \;=\; 0.$$

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

$$T_\theta^{(s)}\mathcal{M} \;:=\; \mathrm{span}\{\partial_{s,k}(\theta) \;:\; k = 1, \dots, d_s\}, \quad s \in N.$$

See Appendix C for a proof.

Acknowledgments

NA and JvO acknowledge the support of the Deutsche Forschungsgemeinschaft Priority Programme “The Active Self” (SPP 2134).

Appendix 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

$$\begin{aligned} \pi_V \colon \mathcal{P} &\to \mathcal{P}_V, \\ p(x_s, x_t) &\mapsto p(x_t) \;=\; \sum_{x_s} p(x_s, x_t), \end{aligned} \tag{40}$$

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

$$\begin{aligned}
\mathcal{V}_p &= \ker d\pi_V \\
&= \left\{ A \in T_p\mathcal{P} : \sum_{x_s} A(x_s, x_t) \ = \ 0, x_t \in \mathsf{X}_t \right\} \\
&= \operatorname{span}\left\{ \delta^{(0,0)} - \delta^{(1,0)}, \delta^{(0,1)} - \delta^{(1,1)} \right\}, \\
\mathcal{H}_p &= (\ker d\pi_V)^{\perp} \\
&= \left\{ A \in T_p\mathcal{P} : \frac{A(0,x_t)}{p(0,x_t)} - \frac{A(1,x_t)}{p(1,x_t)} \ = \ 0, x_t \in \mathsf{X}_t \right\} \qquad (41) \\
&= \operatorname{span}\left\{ p(0,0)\delta^{(0,0)} + p(1,0)\delta^{(1,0)}, p(0,1)\delta^{(0,1)} + p(1,1)\delta^{(1,1)} \right\}.
\end{aligned}$$

Now we let the model be the independence model, given by

$$\mathcal{M} \ = \ \{p \in \mathcal{P} : p(x_s, x_t) \ = \ p(x_s)p(x_t)\}.$$

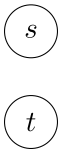

Figure 8: Graph G

This model factorizes over the graph depicted in Figure 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.

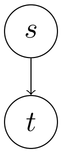

Figure 9: Graph G

Example 2: Non-Cylindrical Two-Node Model

For the same $\mathcal{P}$ and the same π_V as in Example 1 (see equation (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 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 (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.

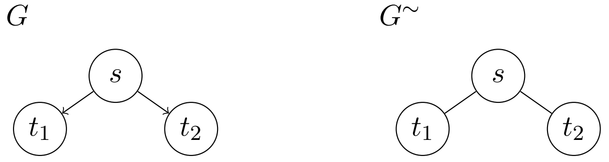

Figure 10: (left) Directed graph G; (right) Undirected graph $G^{\sim}$.

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

$$\begin{aligned} \pi_V : \mathcal{P} &\to \mathcal{P}_V, \\ p(x_s, x_{t_1}, x_{t_2}) &\mapsto p(x_{t_1}, x_{t_2}) = \sum_{x_s} p(x_s, x_{t_1}, x_{t_2}), \end{aligned}$$

and its differential

$$\begin{aligned} d\pi_V : T_p\mathcal{P} &\to T_{\pi_V(p)}\mathcal{P}_V, \\ A = \sum_{x_s, x_{t_1}, x_{t_2}} A(x_s, x_{t_1}, x_{t_2}) \delta^{(x_s, x_{t_1}, x_{t_2})} &\mapsto \sum_{x_{t_1}, x_{t_2}} \left(\sum_{x_s} A(x_s, x_{t_1}, x_{t_2}) \right) \delta^{(x_{t_1}, x_{t_2})}. \end{aligned} \tag{42}$$

The vertical and horizontal spaces are given by

$$\begin{aligned} \mathcal{V}_p &= \ker d\pi_V = \left\{ A \in T_p\mathcal{P} : \sum_{x_s} A(x_s, x_{t_1}, x_{t_2}) = 0 \right\}, \\ \mathcal{H}_p &= (\ker d\pi_V)^{\perp} = \left\{ A \in T_p\mathcal{P} : \sum_{x_s} \frac{A(x_s, x_{t_1}, x_{t_2})}{p(x_s, x_{t_1}, x_{t_2})} (-1)^{x_s} = 0 \right\}. \end{aligned}$$

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 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

$$\begin{aligned} \partial_s(\theta) \;&=\; \begin{bmatrix} -(1-\theta_{t_1,1})(1-\theta_{t_2,1}) \\ -(1-\theta_{t_1,1})\theta_{t_2,1} \\ -\theta_{t_1,1}(1-\theta_{t_2,1}) \\ -\theta_{t_1,1}\theta_{t_2,1} \\ (1-\theta_{t_1,2})(1-\theta_{t_2,2}) \\ (1-\theta_{t_1,2})\theta_{t_2,2} \\ \theta_{t_1,2}(1-\theta_{t_2,2}) \\ \theta_{t_1,2}\theta_{t_2,2} \end{bmatrix}, \\ \partial_{t_1,1}(\theta) \;&=\; \begin{bmatrix} -(1-\theta_s)(1-\theta_{t_2,1}) \\ -(1-\theta_s)\theta_{t_2,1} \\ (1-\theta_s)(1-\theta_{t_2,1}) \\ (1-\theta_s)\theta_{t_2,1} \\ 0\\0\\0\\0 \end{bmatrix}, \quad \partial_{t_1,2}(\theta) \;=\; \begin{bmatrix} 0\\0\\0\\0 \\ -\theta_s(1-\theta_{t_2,2}) \\ -\theta_s\theta_{t_2,2} \\ \theta_s(1-\theta_{t_2,2}) \\ \theta_s\theta_{t_2,2} \end{bmatrix}, \end{aligned}$$

$$\partial_{t_2,1}(\theta) \;=\; \begin{bmatrix} -(1-\theta_s)(1-\theta_{t_1,1}) \\ (1-\theta_s)(1-\theta_{t_1,1}) \\ -(1-\theta_s)\theta_{t_1,1} \\ (1-\theta_s)\theta_{t_1,1} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \partial_{t_2,2}(\theta) \;=\; \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -\theta_s(1-\theta_{t_1,2}) \\ \theta_s(1-\theta_{t_1,2}) \\ -\theta_s\theta_{t_1,2} \\ \theta_s\theta_{t_1,2} \end{bmatrix}.$$

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 (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_θ} 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.

Appendix 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 (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

$$G^{-1}(\theta) = \begin{bmatrix} \theta_1(1-\theta_1) & -\theta_1\theta_2 & -\theta_1\theta_3 \\ -\theta_2\theta_1 & \theta_2(1-\theta_2) & -\theta_2\theta_3 \\ -\theta_3\theta_1 & -\theta_3\theta_2 & \theta_3(1-\theta_3) \end{bmatrix}. \tag{43}$$

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

$$\sum_{i=1}^{3} \left[\nabla\mathcal{L}(\theta)\right]_i\, \partial_i(\theta),$$

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

$$\theta_{m+1} \;=\; \theta_m + \varepsilon \cdot \nabla\mathcal{L}(\theta_m), \qquad m = 0, 1, 2, \ldots, \tag{44}$$

whereas the natural gradient method suggests

$$\theta_{m+1} \;=\; \theta_m + \varepsilon \cdot \widetilde{\nabla}\mathcal{L}(\theta_m), \qquad m = 0, 1, 2, \ldots. \tag{45}$$

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 (45) in comparison with the Euclidean iteration rule (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 (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, (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

$$\begin{aligned} d\pi_V \left(\sum_{i=1}^{3} [\nabla \operatorname{GAP}_q(\theta)]_i \, \partial_i(\theta) \right) &= \sum_{i=1}^{3} [\nabla \operatorname{GAP}_q(\theta)]_i \, d\pi_V(\partial_i(\theta)) \\ &= \left([\nabla \operatorname{GAP}_q(\theta)]_1 + [\nabla \operatorname{GAP}_q(\theta)]_3 \right) (\delta^0 - \delta^1). \end{aligned}$$

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:

$$\begin{aligned} & \left[\widetilde{\nabla} \operatorname{GAP}_q(\theta) \right]_1 + \left[\widetilde{\nabla} \operatorname{GAP}_q(\theta) \right]_3 \\ &= \left(\frac{p^*(0)}{\theta_1 + \theta_3} - \frac{p^*(1)}{\theta_2 + \theta_4} \right) \theta_1(\theta_2 + \theta_4) - p^*(0) q(0|0) + \theta_1 \\ &\quad + \left(\frac{p^*(0)}{\theta_1 + \theta_3} - \frac{p^*(1)}{\theta_2 + \theta_4} \right) \theta_3(\theta_2 + \theta_4) - p^*(0) q(1|0) + \theta_3 \\ &= \left(\frac{p^*(0)}{\theta_1 + \theta_3} - \frac{p^*(1)}{\theta_2 + \theta_4} \right) (\theta_1 + \theta_3)(\theta_2 + \theta_4) \underbrace{- p^*(0) q(0|0) - p^*(0) q(1|0)}_{-p^*(0)} + \theta_1 + \theta_3 \\ &= p^*(0)(\theta_2 + \theta_4) - p^*(1)(\theta_1 + \theta_3) - p^*(0) + \theta_1 + \theta_3 \\ &= p^*(0) \underbrace{- p^*(0)(\theta_1 + \theta_3) - p^*(1)(\theta_1 + \theta_3)}_{-(\theta_1 + \theta_3)} - p^*(0) + \theta_1 + \theta_3 \\ &= 0. \end{aligned}$$

This exemplifies our core result (34) in terms of local coordinates. The same calculation for the Euclidean gradient does not lead to this result. Thus, generically we have

$$[\nabla \operatorname{GAP}_q(\theta)]_1 + [\nabla \operatorname{GAP}_q(\theta)]_3 \neq 0.$$

Appendix C. Proof of Proposition 10

Proof [Proof of Proposition 10] Without loss of generality, we identify the unit set N with the set $\{1, \ldots, n\}$, $n = |N|$, 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:

$$\begin{aligned} & g_\theta^{\mathrm{FR}} \left(\partial_{s,k}(\theta), \partial_{t,l}(\theta) \right) \\ &= \sum_x \frac{1}{p(x;\theta)} \partial_{s,k}(x;\theta) \, \partial_{t,l}(x;\theta) \qquad \text{(by (11))} \\ &= \sum_x \frac{1}{p(x;\theta)} \qquad \text{(by (39))} \\ &\quad \times \left(p(x;\theta) \frac{\partial}{\partial \theta_{s,k}} \ln p(x_s | x_{\mathrm{pa}(s)}; \theta_s) \right) \left(p(x;\theta) \frac{\partial}{\partial \theta_{t,l}} \ln p(x_t | x_{\mathrm{pa}(t)}; \theta_t) \right) \\ &= \sum_x p(x;\theta) \, \frac{\partial}{\partial \theta_{s,k}} \ln p(x_s | x_{\mathrm{pa}(s)}; \theta_s) \frac{\partial}{\partial \theta_{t,l}} \ln p(x_t | x_{\mathrm{pa}(t)}; \theta_t) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{x} \frac{\partial}{\partial \theta_{s,k}} \ln p(x_s|x_{\mathrm{pa}(s)};\theta_s) \prod_{i=1}^{t} p(x_i|x_{\mathrm{pa}(i)};\theta_i) \frac{\partial}{\partial \theta_{t,l}} \ln p(x_t|x_{\mathrm{pa}(t)};\theta_t) \\
&\quad \times \underbrace{\prod_{i=t+1}^{n} p(x_i|x_{\mathrm{pa}(i)};\theta_i)}_{=1} \\
&= \sum_{x_1,\ldots,x_t} \frac{\partial}{\partial \theta_{s,k}} \ln p(x_s|x_{\mathrm{pa}(s)};\theta_s) \prod_{i=1}^{t} p(x_i|x_{\mathrm{pa}(i)};\theta_i) \frac{\partial}{\partial \theta_{t,l}} \ln p(x_t|x_{\mathrm{pa}(t)};\theta_t) \\
&= \sum_{x_1,\ldots,x_{t-1}} \frac{\partial}{\partial \theta_{s,k}} \ln p(x_s|x_{\mathrm{pa}(s)};\theta_s) \sum_{x_t} \prod_{i=1}^{t} p(x_i|x_{\mathrm{pa}(i)};\theta_i) \\
&\quad \times \frac{\partial}{\partial \theta_{t,l}} \ln p(x_t|x_{\mathrm{pa}(t)};\theta_t) \\
&= \sum_{x_1,\ldots,x_{t-1}} \frac{\partial}{\partial \theta_{s,k}} \ln p(x_s|x_{\mathrm{pa}(s)};\theta_s) \prod_{i=1}^{t-1} p(x_i|x_{\mathrm{pa}(i)};\theta_i) \sum_{x_t} p(x_t|x_{\mathrm{pa}(t)};\theta_t) \\
&\quad \times \frac{\partial}{\partial \theta_{t,l}} \ln p(x_t|x_{\mathrm{pa}(t)};\theta_t) \\
&= \sum_{x_1,\ldots,x_{t-1}} \frac{\partial}{\partial \theta_{s,k}} \ln p(x_s|x_{\mathrm{pa}(s)};\theta_s) \prod_{i=1}^{t-1} p(x_i|x_{\mathrm{pa}(i)};\theta_i) \frac{\partial}{\partial \theta_{t,l}} \sum_{x_t} p(x_t|x_{\mathrm{pa}(t)};\theta_t) \\
&= \sum_{x_1,\ldots,x_{t-1}} \frac{\partial}{\partial \theta_{s,k}} \ln p(x_s|x_{\mathrm{pa}(s)};\theta_s) \prod_{i=1}^{t-1} p(x_i|x_{\mathrm{pa}(i)};\theta) \frac{\partial}{\partial \theta_{t,l}} 1 \\
&= 0.
\end{aligned}
$$

■

References

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2): 251–276, 1998.

Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.

Nihat Ay. On the locality of the natural gradient for learning in deep Bayesian networks. *Information Geometry*, pages 1–49, 2020.

Nihat Ay and Shun-ichi Amari. A novel approach to canonical divergences within information geometry. *Entropy*, 17(12):8111–8129, 2015. ISSN 1099-4300. doi: 10.3390/e17127866. URL https://www.mdpi.com/1099-4300/17/12/7866.

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

Justin Bayer, Maximilian Soelch, Atanas Mirchev, Baris Kayalibay, and Patrick van der Smagt. Mind the gap when conditioning amortised inference in sequential latent-variable models. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=a2gqxKDvYys`.

Liqun Chen, Chenyang Tao, Ruiyi Zhang, Ricardo Henao, and Lawrence Carin Duke. Variational inference and model selection with generalized evidence bounds. In *International conference on machine learning*, pages 893–902. PMLR, 2018.

Nikolai Nikolaevich Chentsov. Statiscal decision rules and optimal inference. *Monog*, 53, 1982.

Adwait Datar and Nihat Ay. Convergence properties of natural gradient descent for minimizing KL divergence. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL `https://openreview.net/forum?id=h6hjjAF5Bj`.

Adwait Datar, Jesse van Oostrum, and Nihat Ay. Code for paper: On the natural gradient of the evidence lower bound. `https://github.com/addat10/Nat-Gradient-ELBO.git`, 2024.

Peter Dayan, Geoffrey E Hinton, Radford M Neal, and Richard S Zemel. The Helmholtz machine. *Neural computation*, 7(5):889–904, 1995.

Richard P Feynman. *Statistical mechanics: a set of lectures.* W.A. Benjamin, 1972.

Karl Friston. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836, 2005.

Akio Fujiwara and Shun-ichi Amari. Gradient systems in view of information geometry. *Physica D: Nonlinear Phenomena*, 80(3):317–327, 1995.

Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13, 1993.

Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and Helmholtz free energy. *Advances in neural information processing systems*, 6, 1993.

Shiro Ikeda, Shun-ichi Amari, and Hiroyuki Nakahara. Convergence of the wake-sleep algorithm. *Advances in neural information processing systems*, 11, 1998.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Yingzhen Li and Richard E Turner. Rényi divergence variational inference. *Advances in neural information processing systems*, 29, 2016.

David JC MacKay. 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*, pages 191–198. Springer, 1995.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21:1–76, 2020.

Yann Ollivier. Riemannian metrics for neural networks I: Feedforward networks. *Information and Inference: A Journal of the IMA*, 4(2):108–153, 2015.

Yuchen Pu, Zhe Gan, Ricardo Henao, Chunyuan Li, Shaobo Han, and Lawrence Carin. VAE learning via Stein variational gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

Tom Rainforth, Adam R. Kosiorek, Tuan Anh Le, Chris J. Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. In *International Conference on Machine Learning*, 2018. URL `https://api.semanticscholar.org/CorpusID:3281926`.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

Jesse van Oostrum, Johannes Müller, and Nihat Ay. Invariance properties of the natural gradient in overparametrised systems. *Information geometry*, 6(1):51–67, 2023.

Jesse van Oostrum, Peter van Hintum, and Nihat Ay. Inversion of Bayesian networks. *Int. J. Approx. Reasoning*, 164(C), feb 2024. ISSN 0888-613X. doi: 10.1016/j.ijar.2023.109042. URL `https://doi.org/10.1016/j.ijar.2023.109042`.

Stefan Webb, Adam Golinski, Rob Zinkov, Siddharth N, Tom Rainforth, Yee Whye Teh, and Frank Wood. Faithful inversion of generative models for effective amortized inference. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper_files/paper/2018/file/894b77f805bd94d292574c38c5d628d5-Paper.pdf`.