On the Fisher-Rao Gradient of the Evidence Lower Bound

Nihat Ay [1 2 3] **Jesse van Oostrum** [1]

Abstract

This article studies the Fisher-Rao gradient, also referred to as the natural gradient, of the evidence lower bound, the ELBO, which plays a crucial role within the theory of the Variational Autonecoder, the Helmholtz Machine and the Free Energy Principle. The natural gradient of the ELBO is related to the natural gradient of the Kullback-Leibler divergence from a target distribution, the prime objective function of learning. Based on invariance properties of gradients within information geometry, conditions on the underlying model are provided that ensure the equivalence of minimising the prime objective function and the maximisation of the ELBO.

1. Introduction

Originating from statistics, information geometry provides efficient methods in machine learning that are based on duality concepts from differential geometry (Amari & Nagaoka, 2000; Amari, 2016; Ay et al., 2017). Most prominently, it suggests as a fundamental structure a Riemannian manifold $(\mathcal{M}, g)$, 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 $D : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+$. These structures can lead to highly efficient learning algorithms when used together. On the one hand, the distinguished canonical divergence D offers a natural way to define an objective or risk function for learning. When optimising this divergence in terms of the gradient descent method, the Riemannian metric g should be applied, leading 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 affine coordinate systems. Loosely speaking, the learning converges to a solution in the most direct way (Fujiwara & Amari, 1995). This demonstrates the simplicity and efficiency of learning as a result of a consistent combination of the underlying structures.

Despite the great advantages of the outlined information-geometric approach to learning, it is a highly non-trivial task to actually utilise and implement this approach within the setting of machine learning. This is partly due to the fact that the outlined dually flat structure, consiting of g, ∇, and ∇^*, does not come with $\mathcal{M}$ itself but with a typically high-dimensional ambient space $\mathcal{P}$ of $\mathcal{M}$. When induced to $\mathcal{M}$, the resulting structure $g_{\mathcal{M}}$, $\nabla_{\mathcal{M}}$, and $\nabla^*_{\mathcal{M}}$ may loose much of its simplicity without further assumptions. Assuming $\mathcal{M}$ to be autoparallel with respect to ∇ or ∇^* is sufficient for a dually flat induced geometry (Theorem 3.5 of (Amari & Nagaoka, 2000)). Such an example is given by a Boltzmann machine without hidden units (Amari et al., 1992). In that case, the existence and uniqueness of projections based on the canonical divergence are guaranteed. Furthermore, learning according to the natural gradient method is consistent in the sense that it follows straight lines defined in terms of the induced dually flat structure. However, typically the expressive power of a learning system has to be increased in terms of latent or hidden units denoted by H. In this case, the prime model for learning is associated with the observed or visible units denoted by V. It is obtained as the image of $\mathcal{M}$ under the marginalisation map from the full system to its visible part and will therefore be denoted by $\mathcal{M}_V$. Even if $\mathcal{M}$ inherits properties from its ambient space that are advantageous for learning, these properties need not be preserved under this marginalisation. Thus, we are faced with 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 marginalisation which maps $\mathcal{M}$ to the model $\mathcal{M}_V$. In order to disentangle the complexity as a consequence of these two operations from the complexity based on the information-geometric structures on the respective ambient spaces, it is important to study learning processes in the absence of any constraints through $\mathcal{M}$.

In this article, we follow the above reasoning in order to discuss the evidence lower bound, referred to as the ELBO,

[1] Institute for Data Science Foundation, Hamburg University of Technology, Hamburg, Germany [2] Santa Fe Institute, Santa Fe, USA [3] Leipzig University, Leipzig, Germany. Correspondence to: Nihat Ay <nihat.ay@tuhh.de>.

ICML 2023 Workshop on Duality for Modern Machine Learning, Honolulu, Hawaii, USA.

from an information-geometric perspective. This bound plays a fundamental role in the theory of the Variational Autoencoder (VAE) (Kingma & Welling, 2013), the Helmholtz machine (Dayan et al., 1995; Ikeda et al., 1998), and the Free Energy Principle (Friston, 2005). While the prime objective of learning in this context is the minimisation of the Kullback-Leibler divergence from a target distribution on states of visible units, the derivation of the ELBO leads to a different objective and aims at maximising that bound. We relate the two optimisation problems to each other by studying them in view of information geometry. We highlight the simplicity and consistency of both problems when considered in the full ambient space, without restricting it to a model $\mathcal{M}$. It is remarkable that in this case, the ELBO leads to the same gradient field as the original objective function, the Kullback-Leibler divergence from a target distribution on states of the visible units. This equivalence is not necessarily preserved when restricting the optimisation to a model $\mathcal{M}$. We provide a sufficient condition for this to hold, which requires the notion of a cylindrical model.

In Section 2 we are going to review basic information-geometric structures, thereby introducing the notation used in this article. That section also includes results from the previous work (Ay, 2020) on which this article is based. Section 3 introduces the prime objective of learning, minimising the Kullback-Leibler divergence from a target distribution on states of the visible units and briefly outlines its relation to the ELBO. Section 4 deals with the analysis of the optimisation problem for the extended full system and relates it to the prime optimisation problem defined for its visible part. This section contains the main results of this article. Section 5 relates these results to the ELBO, thereby making statements on its natural gradient. Section 6 concludes with a result that is particularly helpful when dealing specifically with Bayesian graphical models.

2. Basic information-geometric structures

The set of strictly positive probability distributions $\mathcal{P}$ on some finite set of states x represents a basic example within information geometry. It carries a natural dually flat structure, given by the Fisher-Rao metric g^{FR}, the mixture connection $\nabla^{(m)}$ and the exponential connection $\nabla^{(e)}$. We write a point $p \in \mathcal{P}$ as

$$p = \sum_x p(x)\, \delta^x, \qquad (1)$$

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

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

In this article, the dual connections will only implicitly play a role, through their relation to the Kullback-Leibler divergence (KL-divergence), which is defined on $\mathcal{P} \times \mathcal{P}$ by

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

We can express the Fisher-Rao gradients of the KL-divergence in both arguments and obtain

$$\begin{aligned} \mathrm{grad}_p D(q\|\cdot) \;&=\; \sum_x (p(x) - q(x))\, \delta^x \\ &=\; p - q \;\in\; T_p\mathcal{P}, \end{aligned}$$

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

These gradients coincide with the inverse of the respective exponential maps of the connections $\nabla^{(m)}$ and $\nabla^{(e)}$ (Ay & Amari, 2015; Ay et al., 2017).

We are now going to use these structures in a natural setting of a learning system. We consider a system consisting of visible and hidden units V and H, respectively. The set of all strictly positive probability distributions on joint states (x_V, x_H) is denoted by $\mathcal{P}_{V,H}$, which we typically abbreviate as $\mathcal{P}$. Learning takes place in a model $\mathcal{M} \subseteq \mathcal{P}$ with probability distributions $p(x_V, x_H; \theta)$ parametrised in terms of a parameter vector $\theta \in \mathbb{R}^d$. (Typically, the parameter set is an open subset Θ of $\mathbb{R}^d$.) In this article, we will mostly omit the parameter and simply write $p \in \mathcal{M}$. The model $\mathcal{M}$ carries the induced geometry of $\mathcal{P}$. As mentioned in the introduction, one complication of learning emerges from the fact that hidden units are often considered to be auxiliary units which increase the expressive power of the network. The actual objective of learning refers only to the visible units. Therefore, we have to consider the restricted model, defined for states x_V of the visible units only. We denote by $\mathcal{P}_V$ the set of strictly positive probability distributions on states x_V and consider the natural marginalisation map

$$\pi_V : \mathcal{P} \to \mathcal{P}_V,$$

which assigns to a joint probability distribution $p(x_V, x_H)$ the marginal distribution

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

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

$$d\pi_V : T_p\mathcal{P} \to T_{\pi_V(p)}\mathcal{P}_V,$$

defined by

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

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 $\mathcal{H}_p$ refers to the Fisher-Rao metric in $p \in \mathcal{P}$. With this orthogonal decomposition, we can write

$$T_p\mathcal{P} \;=\; \mathcal{H}_p \oplus \mathcal{V}_p\,.$$

Every vector A in $T_p\mathcal{P}$ has a unique decomposition as

$$A \;=\; A^{\mathcal{H}} + A^{\mathcal{V}},$$

where $A^{\mathcal{H}} \in \mathcal{H}_p$ and $A^{\mathcal{V}} \in \mathcal{V}_p$.

The image of the model $\mathcal{M} \subseteq \mathcal{P}$, that is $\pi_V(\mathcal{M})$, is denoted by $\mathcal{M}_V$. It carries the Fisher-Rao metric induced from $\mathcal{P}_V$.

Definition 2.1 (Definition 1 of (Ay, 2020))**.** We call a model $\mathcal{M} \subseteq \mathcal{P}$ *cylindrical* in a non-singular point $p \in \mathcal{M}$, if

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

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

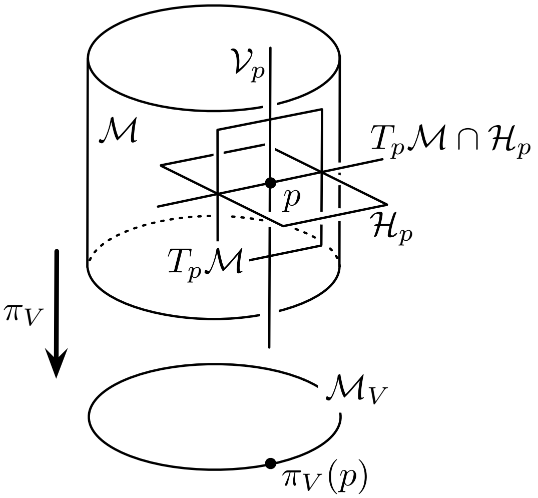

Figure 1. Illustration of a cylindrical model $\mathcal{M}$ in terms of a cylinder, the Cartesian product of a circle with a finite interval. The tangent space $T_p\mathcal{M}$ equals the sum of its intersections with $\mathcal{H}_p$ and $\mathcal{V}_p$.

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

$$g_p^{\mathrm{FR}}(A,B) \;=\; g_{\pi_V(p)}^{\mathrm{FR}}\left(d\pi_V(A), d\pi_V(B)\right), \qquad (2)$$

whenever p and $\pi_V(p)$ are non-singular points of $\mathcal{M}$ and $\mathcal{M}_V$, respectively, and $d\pi_V(T_p\mathcal{M}) = T_{\pi_V(p)}\mathcal{M}_V$. The equality (2) 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.

Theorem 2.2 (Theorem 5 of (Ay, 2020))**.** *Let $\mathcal{M}$ be a cylindrical model, and let $\mathcal{L} : \mathcal{M}_V \to \mathbb{R}$ be a differentiable objective function. Then*

$$d\pi_V\left(\mathrm{grad}_p^{\mathcal{M}}(\mathcal{L} \circ \pi_V)\right) \;=\; \mathrm{grad}_{\pi_V(p)}^{\mathcal{M}_V}\mathcal{L}. \qquad (3)$$

Here, we assume that $\mathcal{M}$ and $\mathcal{M}_V$ are non-singular in p and $\pi_V(p)$, respectively, and that $d\pi_V(T_p\mathcal{M}) = T_{\pi_V(p)}\mathcal{M}_V$.

Note that the gradient on the LHS of (3) 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 2.2 is quite restrictive and basically holds only for the Fisher-Rao metric and cylindrical models.

Our main example of a cylindrical model will be the full model $\mathcal{M} = \mathcal{P}$. Here, all points p are non-singular and $T_p\mathcal{P} = \mathcal{H}_p \oplus \mathcal{V}_p$. 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 with no constraints, however, we are able to reveal the geometric effect of these constraints. In addition to this reasoning, we also aim at studying models that are cylindrical and not just embedded in a cylindrical ambient space. The most natural choice here is a model of maximal dimension, for instance an open subset of $\mathcal{P}$. When representing such a model in terms of a neural network, one often refers to it as an overparametrised model. For lower-dimensional cylindrical models one would have to develop design principles. Initial work in that direction is provided in (Ay, 2020).

3. Learning a target distribution for visible units

The objective of learning is to find a parameter vector θ such that $p(x_V;\theta)$ mimics a target distribution $p^* \in \mathcal{P}_V$. Interpreting $\mathcal{M}_V$ as a model defined by a generative network, we can say that the network learns to generate patterns x_V that are distributed according to p^*. To achieve that, we minimise the KL-divergence of a distribution $p \in \mathcal{M}_V$ from a target distribution p^*, that is, we minimise the following objective function:

$$\begin{aligned} \mathcal{L}(p) &= D(p^* \| p) \\ &= \sum_{x_V} p^*(x_V) \ln \frac{p^*(x_V)}{p(x_V)}. \end{aligned} \tag{4}$$

In general, this is a difficult problem. On the one hand, $\mathcal{M}_V$ can be complicated with singularities. On the other hand, the Fisher-Rao metric is difficult to evaluate if $\mathcal{M}_V$ does not have a particularly nice structure. To be more concrete, we first evaluate the gradient of $D(p^* \| \cdot)$, considered as a function on $\mathcal{P}_V$:

$$\mathrm{grad}_p^{\mathcal{P}_V} D(p^* \| \cdot) \; = \; p - p^* \; \in \; T_p \mathcal{P}_V. \tag{5}$$

For the gradient on the model $\mathcal{M}_V$, we have to project the gradient (5) 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

$$\mathrm{grad}_p^{\mathcal{M}_V} D(p^* \| \cdot) \; = \; \Pi_p(p - p^*) \; \in \; T_p \mathcal{M}_V, \tag{6}$$

where Π_p denotes the orthogonal projection onto the tangent space $T_p \mathcal{M}_p$. Note that this projection does not have to be particularly simple, even though the difference vector $p - p^*$, the gradient in the ambient space, is simple.

We will modify the problem of minimising the KL-divergence in several simplifying steps thereby tracing the geometric implication of each individual step. In the first step, we observe that the minimisation of the KL-divergence (4) with respect to p is equivalent to minimising the cross entropy

$$-\sum_{x_V} p^*(x_V) \ln p(x_V) \tag{7}$$

because these two functions differ only by a constant, the negative entropy of p^*. Here, both distributions are defined for states x_V of the visible units V. In order to be tractable, one upper bounds the cross entropy (7) using the evidence lower bound which involves the extension to the set H of hidden units:

$$\ln p(x_V) \; \geq \; -\sum_{x_H} q(x_H | x_V) \ln \frac{q(x_H | x_V)}{p(x_V, x_H)}. \tag{8}$$

This leads to

$$\begin{aligned} & -\sum_{x_V} p^*(x_V) \ln p(x_V) \\ & \leq \; \sum_{x_V} p^*(x_V) \sum_{x_H} q(x_H | x_V) \ln \frac{q(x_H | x_V)}{p(x_V, x_H)}. \end{aligned} \tag{9}$$

In this article, we compare the natural gradient of the bound (9) in the extended system with the natural gradient of the original objective function $D(p^* \| p)$ defined on its visible part. In order to imply the same learning processes based on the gradient descent method, the respective gradients should coincide. We provide a criterion for this to be the case.

4. The extended problem with hidden units

It is well-known that the minimisation 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 $\mathcal{P}_{V,H}$ (Amari, 2016). For that, we consider the so-called *data manifold*

$$\mathcal{Q} \; := \; \{ q \in \mathcal{P}_{V,H} \; : \; \pi_V(q) = p^* \} \, . \tag{10}$$

With the monotonicity of the KL-divergence, we obtain for any $p \in \mathcal{M}$ and $q \in \mathcal{Q}$

$$\begin{aligned} (\mathcal{L} \circ \pi_V)(p) &= D(p^* \| \pi_V(p)) \\ &\leq D(q \| p), \end{aligned} \tag{11}$$

where equality holds for $q = \pi_{\mathcal{Q}}(p)$ defined by

$$\pi_{\mathcal{Q}}(p)(x_V, x_H) \; = \; p^*(x_V) p(x_H | x_V). \tag{12}$$

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 $\hat{p}$ minimises $\mathcal{L} \circ \pi_V = D(\mathcal{Q} \| \cdot)$ in $\mathcal{M}$ if and only if $\pi_V(\hat{p})$ minimises $\mathcal{L} = D(p^* \| \cdot)$ in $\mathcal{M}_V$. However, there is one important difference between the corresponding optimisations in terms of the natural gradient method. For the optimisation 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 optimised with respect to the corresponding Fisher-Rao gradient on $\mathcal{M}$. In general, the two ways to optimise basically the same function will not be equivalent. However, according to Theorem 2.2, they will be equivalent whenever the model $\mathcal{M}$ is cylindrical.

Theorem 4.1. **(a)** *Consider first the function $D(\mathcal{Q} \| \cdot)$ on $\mathcal{P}$. Then*

$$\mathrm{grad}_p^{\mathcal{P}} D(\mathcal{Q} \| \cdot) \; = \; p - \pi_{\mathcal{Q}}(p) \tag{13}$$

where $\pi_{\mathcal{Q}}(p)$ is defined by (12). In order to obtain the gradient of $D(\mathcal{Q} \| \cdot)$ in a non-singular point $p \in \mathcal{M}$, we have to project (13) onto $T_p \mathcal{M}$, that is

$$\mathrm{grad}_p^{\mathcal{M}} D(\mathcal{Q} \| \cdot) \; = \; \Pi_p(p - \pi_{\mathcal{Q}}(p)), \tag{14}$$

where Π_p denotes the orthogonal projection $T_p \mathcal{P} \to T_p \mathcal{M}$ with repect to the Fisher-Rao metric on $\mathcal{P}$.
(b) *If $\mathcal{M}$ is cylindrical then*

$$d\pi_V \left(\mathrm{grad}_p^{\mathcal{M}} D(\mathcal{Q} \| \cdot) \right) \; = \; \mathrm{grad}_{\pi_V(p)}^{\mathcal{M}_V} D(p^* \| \cdot) \tag{15}$$

Here, we assume that $\mathcal{M}$ and $\mathcal{M}_V$ are non-singular in p and $\pi_V(p)$, respectively, and that $d\pi_V(T_p \mathcal{M}) = T_{\pi_V(p)} \mathcal{M}_V$. In particular, this holds for $\mathcal{M} = \mathcal{P}$ and $\mathcal{M}_V = \mathcal{P}_V$.

Proof. We know that $D(\mathcal{Q}\|\cdot) = D(p^*\|\pi_V(\cdot))$. With the partial derivatives

$$\frac{\partial}{\partial p(x_V, x_H)} D(p^*\|\pi_V(\cdot)) \; = \; -\frac{p^*(x_V)}{p(x_V)},$$

this implies 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 (13), and equation (14) follows immediately from that. Finally, the invariance (15) is a direct consequence of Theorem 2.2. □

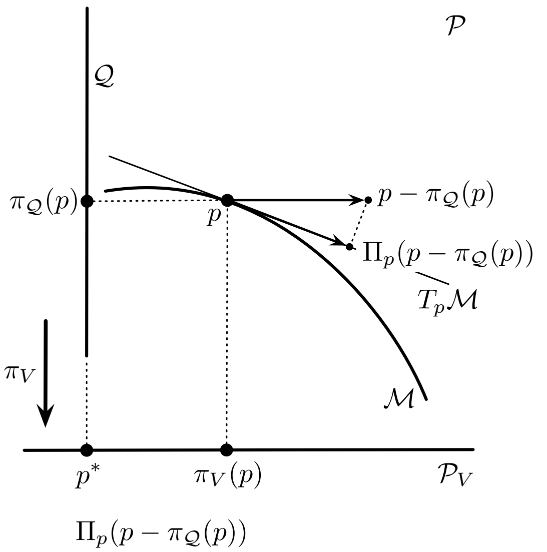

Figure 2. Illustration of gradients considered in Theorem 4.1.

Theorem 4.1 implies a number of conceptual insights which we are now going to elaborate on. First of all, it highlights the simplicity of the natural gradient of $D(\mathcal{Q}\|\cdot)$ in $p \in \mathcal{P}$. It is nothing but the difference vector between p and its projection $\pi_{\mathcal{Q}}(p)$. Thus, any complexity of the natural gradient of $D(\mathcal{Q}\|\cdot)$ on a model $\mathcal{M}$ arises from the projection of that difference vector onto the tangent space $T_p\mathcal{M}$ and therefore depends very much on the structure of $\mathcal{M}$. For a Bayesian graphical model, $T_p\mathcal{M}$ decomposes in a convenient way so that some of the original simplicity is preserved after projection. A corresponding more precise statement will be formulated at the end of this article, in Proposition 6.1. Furthermore, the gradient (14) of the function $D(\mathcal{Q}\|\cdot)$, defined on $\mathcal{M}$, can now be compared with the gradient (6) of the original function $D(p^*\|\cdot)$ which is defined on $\mathcal{M}_V$. According to the invariance (15), these two gradients are equivalent, if $\mathcal{M}$ is cylindrical, which implies that gradient descent learning in $\mathcal{M}$ has 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 4.1. If we do not restrict the optimisation to a lower-dimensional model $\mathcal{M}$ then all information-geometric structures are consistent in the sense that the optimisation in the extended system, with hidden units, is equivalent to the original optimisation with only visible units. Again, any deviation from the invariance (15) arises from the restriction of the optimisation to $\mathcal{M}$.

We are now going to extend Theorem 4.1 to a more general setting. This will allows us to study the natural gradient of the evidence lower bound and thereby show that it is "exact" in the sense that without restricting the optimisation to a lower-dimensional model, its optimisation is equivalent to the optimisation of (4) in $\mathcal{P}_V$. Our extension involves adding a simplifying term to the objective function $D(\mathcal{Q}\|\cdot)$ that ideally leaves the natural gradient invariant. More precisely, for any $q \in \mathcal{Q}$ and $p \in \mathcal{M}$, we have

$$\begin{aligned}
D(\mathcal{Q}\|p) \;&=\; D(\pi_{\mathcal{Q}}(p)\|p) \\
&\leq\; D(q\|\pi_{\mathcal{Q}}(p)) + D(\pi_{\mathcal{Q}}(p)\|p) \qquad (16)\\
&=\; D(q\|p). \qquad (17)
\end{aligned}$$

Instead of taking the gradient of $D(\mathcal{Q}\|\cdot)$ we take the gradient of the upper bound (17), with fixed q, and analyse the effect of this replacement. We have

$$\begin{aligned}
&\mathrm{grad}_p^{\mathcal{P}} D(q\|\cdot) \\
&= \; p - q \\
&= \; (p - \pi_{\mathcal{Q}}(p)) + (\pi_{\mathcal{Q}}(p) - q) \qquad (18)\\
&= \; (p - q)^{\mathcal{H}} + (p - q)^{\mathcal{V}}. \qquad (19)
\end{aligned}$$

Thus, by adding a term we have created a second difference vector as part of the gradient, namely $\pi_{\mathcal{Q}}(p) - q$. When mapping it down, however, this difference vector vanishes, and 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 original function on $\mathcal{P}_V$, the set of distributions on the visible units, is not affected at all by the simplifying extension of the problem to the set $\mathcal{P} = \mathcal{P}_{V,H}$ of distributions on the visible and hidden units together. If we replace $\mathcal{P}$ by a more general model $\mathcal{M}$, then this invariance only holds, if $\mathcal{M}$ is cylindrical.

Theorem 4.2. *Let $\mathcal{M}$ be a cylindrical model in $\mathcal{P}$, and let $q \in \mathcal{Q}$. Then*

$$d\pi_V \left(\mathrm{grad}_p^{\mathcal{M}} D(q\|\cdot)\right) \;=\; \mathrm{grad}_{\pi_V(p)}^{\mathcal{M}_V} D(p^*\|\cdot). \tag{20}$$

Here, we assume that $\mathcal{M}$ and $\mathcal{M}_V$ are non-singular in p and $\pi_V(p)$, respectively, and that $d\pi_V(T_p\mathcal{M}) = T_{\pi_V(p)}\mathcal{M}_V$.

Proof. With (25), it follows

$$\begin{aligned} &\mathrm{grad}_p^{\mathcal{M}} D(q\|\cdot) \\ &\quad= \Pi_p\left(\mathrm{grad}_p^{\mathcal{P}} D(q\|\cdot)\right) \\ &\quad= \Pi_p\,(p-q) \\ &\quad= \Pi_p\left((p-q)^{\mathcal{H}} + (p-q)^{\mathcal{V}}\right) \\ &\quad= \Pi_p\left((p-q)^{\mathcal{H}}\right) + \Pi_p\left((p-q)^{\mathcal{V}}\right) \end{aligned} \tag{21}$$

Let us first consider the second component. We know, by definition, that $(p-q)^{\mathcal{V}}$ is contained in $\mathcal{V}_p$. Given that $\mathcal{M}$ is cylindrical in p, 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 and we only have to consider the first term in (21). With

$$(p-q)^{\mathcal{H}} \;=\; p - \pi_{\mathcal{Q}}(p),$$

we have

$$\begin{aligned} &\Pi_p\left((p-q)^{\mathcal{H}}\right) \\ &\quad= \Pi_p\,(p - \pi_{\mathcal{Q}}(p)) \\ &\quad= \Pi_p\left(\mathrm{grad}_p^{\mathcal{P}} D(\mathcal{Q}\|\cdot)\right) \qquad \text{(Theorem 4.1 (a))} \\ &\quad= \Pi_p\left(\mathrm{grad}_p^{\mathcal{P}} D(p^*\|\pi_V(\cdot))\right) \\ &\quad= \mathrm{grad}_p^{\mathcal{M}} D(p^*\|\pi_V(\cdot)) \\ &\quad= \mathrm{grad}_{\pi_V(p)}^{\mathcal{M}_V} D(p^*\|\cdot) \qquad \text{(Theorem 4.1 (b)).} \end{aligned}$$

□

Note that while the invariance (20) appears very similar to the invariance (15), 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 plus a simplifying term that varies only in vertical direction so that the $d\pi$ image of its gradient vanishes.

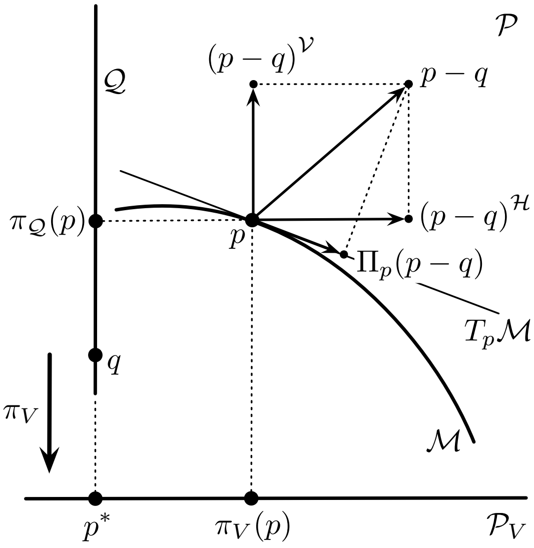

Figure 3. Illustration of gradients considered in Theorem 4.2.

5. The natural gradient of the evidence lower bound

In the previous section, we observed that the extension of the original problem of minimising the divergence from a target distribution p^* to the simpler setting that involves hidden units does not change the gradient for cylindrical models. We now show that this is, at the same time, equivalent to maximising the evidence lower bound. For any $q \in \mathcal{Q}$ and $p \in \mathcal{M}$, we have

$$\begin{aligned} &D(q\|p) \qquad\qquad (22) \\ &\quad= \sum_{x_V, x_H} p^*(x_V)\, q(x_H|x_V) \ln \frac{p^*(x_V)\, q(x_H|x_V)}{p(x_V, x_H)} \\ &\quad= \underbrace{\sum_{x_V} p^*(x_V) \ln p^*(x_V) +}_{\leq 0} \\ &\qquad\quad \sum_{x_V, x_H} p^*(x_V)\, q(x_H|x_V) \ln \frac{q(x_H|x_V)}{p(x_V, x_H)} \\ &\quad\leq \sum_{x_V, x_H} p^*(x_V)\, q(x_H|x_V) \ln \frac{q(x_H|x_V)}{p(x_V, x_H)} \qquad (23) \end{aligned}$$

This derivation is valid for any distribution p^*. In particular, we can choose p^* to be concentrated in one configuration x_V and obtain

$$\ln p(x_V) \;\geq\; -\sum_{x_H} q(x_H|x_V) \ln \frac{q(x_H|x_V)}{p(x_V, x_H)}. \tag{24}$$

The LHS of the inequality (24) is referred to as the *evidence* for x_V, while the bound on its RHS is the ELBO and is equal to the negative of the *variational free energy* for x_V. These quantities play a crucial role in the theory of the Variational Autoencoder (VAE) (Kingma & Welling, 2013), the Helmholtz machine (Dayan et al., 1995; Ikeda et al., 1998) and the Free Energy Principle (Friston, 2005). In these terms, the upper bound in (23) is the negative of the expectation value of the evidence lower bound with respect to p^*. We introduce this quantity as a function on $\mathcal{M}$ (with fixed $q \in \mathcal{Q}$),

$$\mathrm{ELBO}(q, \cdot) : \mathcal{M} \to \mathbb{R},$$

with

$$\mathrm{ELBO}(q,p) \;:=\; -\sum_{x_V, x_H} q(x_V, x_H) \ln \frac{q(x_H|x_V)}{p(x_V, x_H)}$$

Note that the gradient of the function $D(q\|\cdot)$ with respect to p will be the same as the gradient of its upper bound (23), because the two functions differ only by a constant, the entropy of p^*. More formally, we have

$$\mathrm{grad}^{\mathcal{M}}_p \, \mathrm{ELBO}(q,\cdot) \;=\; -\mathrm{grad}^{\mathcal{M}}_p D(q\|\cdot). \tag{25}$$

This implies that the inequality (23) does not play a role in optimisation in terms of gradient methods. We obtain the following immediate consequence of Theorem 4.2.

Corollary 5.1. *Let $\mathcal{M}$ be a cylindrical model in $\mathcal{P}$, and let $q \in \mathcal{Q}$. Then*

$$\begin{aligned} d\pi_V \left(\mathrm{grad}^{\mathcal{M}}_p \, \mathrm{ELBO}(q,\cdot)\right) & \\ = \; -\mathrm{grad}^{\mathcal{M}_V}_{\pi_V(p)} D(p^*\|\cdot). & \quad (26) \end{aligned}$$

Here, we assume that $\mathcal{M}$ and $\mathcal{M}_V$ are non-singular in p and $\pi_V(p)$, respectively, and that $d\pi_V(T_p\mathcal{M}) = T_{\pi_V(p)}\mathcal{M}_V$. In particular, all conditions are satisfied for $\mathcal{M} = \mathcal{P} = \mathcal{P}_{V,H}$ and $\mathcal{M}_V = \mathcal{P}_V$ so that (26) holds in this case.

This proves that, even though the ELBO "lives" in an extended space and provides a bound for our objective function, the KL-divergence on the visible units, it is equivalent to that objective function in terms of the gradient. However, this statement only holds if we evaluate the gradients on the corresponding maximal models $\mathcal{P} = \mathcal{P}_{V,H}$ 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 terms of the objective function, the ELBO does not alter the original optimisation at all. This is remarkable and demonstrates the consistency of the information-geometric structures, which involves the Fisher-Rao metric and the KL-divergence on $\mathcal{P}$ and $\mathcal{P}_V$. However, this invariance of the ELBO is not necessarily preserved when the optimisation is restricted to a model $\mathcal{M}$ and its image $\mathcal{M}_V$, respectively. Therefore, any deviation from the invariance of the ELBO is caused by the restriction of the optimisation to a model. This suggests to study structures of models that preserve natural properties for learning. The following section outlines one instance of such a model.

6. Projecting onto the tangent space of a Bayesian graphical model

In many cases the model $\mathcal{M}$ is given by a Bayesian graphical model $\mathcal{P}^G$, which is a collection of distributions p that factorise over the graph $G = (N, E)$, i.e.

$$p(x) = \prod_{s \in N} p(x_s | x_{\mathrm{pa}(s)}). \tag{27}$$

We can use $\theta_{s,k}$ to parametrise the model, such that

$$p(x) = \prod_{s \in N} p(x_s | x_{\mathrm{pa}(s)}; \theta_s). \tag{28}$$

In order to compute the gradient (14), we need to project onto the tangent space of the model, in terms of Π_p. We have the following result for the projection Π_p on the tangent space of Bayesian graphical models.

Proposition 6.1. *Let $\mathcal{M} = \mathcal{P}^G$ be a graphical model parametrised by θ and $\partial_{s,k} = \frac{\partial}{\partial \theta_{s,k}}$ a basis for $T_p\mathcal{M}$. Then, for $s \neq t$ we have*

$$\langle \partial_{s,k}, \partial_{t,l} \rangle = 0. \tag{29}$$

This implies that $\{\partial_{s,k}\}_{s,k}$ form an orthogonal basis for $T_p\mathcal{M}$. See the appendix for a proof.

References

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

Amari, S.-i. *Information geometry and its applications*, volume 194. Springer, 2016.

Amari, S.-i. and Nagaoka, H. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.

Amari, S.-i., Kurata, K., and Nagaoka, H. Information geometry of boltzmann machines. *IEEE Transactions on neural networks*, 3(2):260–271, 1992.

Ay, N. On the locality of the natural gradient for learning in deep bayesian networks. *Information Geometry*, pp. 1–49, 2020.

Ay, N. and Amari, S.-i. 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.

Ay, N., Jost, J., Vân Lê, H., and Schwachhöfer, L. *Information geometry*, volume 64. Springer, 2017.

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

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. The helmholtz machine. *Neural computation*, 7(5):889–904, 1995.

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

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

Ikeda, S., Amari, S.-i., and Nakahara, H. Convergence of the wake-sleep algorithm. *Advances in neural information processing systems*, 11, 1998.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

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

A. Appendix

Proof of Proposition 6.1. In the proof we suppress the second index of the parameter. Assume $s < t$

$$
\begin{aligned}
\langle \partial_s, \partial_t \rangle &= \sum_x \frac{1}{p(x;\theta)} \partial_s(x) \partial_t(x) \\
&= \sum_x \frac{1}{p(x;\theta)} \left(p(x;\theta) \frac{\partial}{\partial \theta_s} \ln p(x_s | x_{\mathrm{pa}(s)}; \theta_s) \right) \left(p(x;\theta) \frac{\partial}{\partial \theta_t} \ln p(x_t | x_{\mathrm{pa}(t)}; \theta_t) \right) \\
&= \sum_x \frac{\partial}{\partial \theta_s} \ln p(x_s | x_{\mathrm{pa}(s)}; \theta_s) p(x;\theta) \frac{\partial}{\partial \theta_t} \ln p(x_t | x_{\mathrm{pa}(t)}; \theta_t) \\
&= \sum_x \frac{\partial}{\partial \theta_s} \ln p(x_s | x_{\mathrm{pa}(s)}; \theta_s) \prod_{i=1}^{l} p(x_i | x_{\mathrm{pa}(i)}; \theta) \frac{\partial}{\partial \theta_t} \ln p(x_t | x_{\mathrm{pa}(t)}; \theta_t) \prod_{i=t+1}^{n} p(x_i | x_{\mathrm{pa}(i)}; \theta) \\
&= \sum_{x_1, \ldots, x_t} \frac{\partial}{\partial \theta_s} \ln p(x_s | x_{\mathrm{pa}(s)}; \theta_s) \prod_{i=1}^{t} p(x_i | x_{\mathrm{pa}(i)}; \theta) \frac{\partial}{\partial \theta_t} \ln p(x_t | x_{\mathrm{pa}(t)}; \theta_t) \\
&= \sum_{x_1, \ldots, x_{t-1}} \frac{\partial}{\partial \theta_s} \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) \frac{\partial}{\partial \theta_t} \ln p(x_t | x_{\mathrm{pa}(t)}; \theta_t) \\
&= \sum_{x_1, \ldots, x_{t-1}} \frac{\partial}{\partial \theta_s} \ln p(x_s | x_{\mathrm{pa}(s)}; \theta_s) \prod_{i=1}^{t-1} p(x_i | x_{\mathrm{pa}(i)}; \theta) \sum_{x_t} p(x_t | x_{\mathrm{pa}(t)}; \theta_t) \frac{\partial}{\partial \theta_t} \ln p(x_t | x_{\mathrm{pa}(t)}; \theta_t) \\
&= \sum_{x_1, \ldots, x_{t-1}} \frac{\partial}{\partial \theta_s} \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} \sum_{x_t} p(x_t | x_{\mathrm{pa}(t)}; \theta_t) \\
&= \sum_{x_1, \ldots, x_{t-1}} \frac{\partial}{\partial \theta_s} \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} 1 \\
&= 0
\end{aligned}
$$

□