Max-Planck-Institut für Mathematik in den Naturwissenschaften Leipzig

Affine Natural Proximal Learning

by

Wuchen Li, Alex Tong Lin, and Guido Montúfar

Preprint no.: 6 2021

Affine Natural Proximal Learning

Wuchen Li, Alex Tong Lin, and Guido Montúfar

Department of Mathematics, UCLA, Los Angeles, CA 90095

Abstract. We revisit the natural gradient method for learning in statistical manifolds. We consider the proximal formulation and obtain a closed form approximation of the proximity term over an affine subspace of functions in the Legendre dual formulation. We consider two important types of statistical metrics, namely the Wasserstein and Fisher-Rao metrics, and introduce numerical methods for high dimensional parameter spaces.

Keywords: Optimal transport; Information geometry; Proximal operator.

1 Introduction

Learning algorithms usually proceed by minimizing a loss function that measures the discrepancy between a data distribution and a model distribution. Given a parametric model and a metric in probability space, the loss can be minimized by the Riemannian gradient descent method, also known as the natural gradient method. An important metric in this context is the Fisher-Rao information metric [4,17], which induces the Fisher-Rao natural gradient [1]. Another important metric is the Wasserstein metric [14,18], which induces the Wasserstein natural gradient [7,8,11,13]. Natural gradient methods have numerous applications in learning; see, e.g., [2,3,9,12,15,16].

In spite of having numerous theoretical advantages, applying natural gradient methods is often challenging. In particular, machine learning models usually have many parameters, making the direct computation of the parameter updates too costly. Each update requires to compute the Jacobi matrix of the model and the inverse of the metric tensor in parameter space. An alternative, implicit, way to formulate the update is via a proximal operator. Recently [10] proposed proximal methods as an approach to natural gradients and demonstrated their viability in state of the art generative modeling. The idea is to compute the proximity penalty in closed form over an approximation space. This results in a tractable iterative regularization for the parameter updates.

We develop this idea to obtain a general natural proximal method, and provide explicit formulas for the Fisher-Rao and the Wasserstein metrics. These serve three purposes: (i) The proximal operator and its approximation can enable efficient and effective expressions for the time discretized parameter updates of the natural gradient flow. (ii) The proximal method, as an implicit method, naturally regularizes the objective function, and can be used to optimize non-smooth objective functions. (iii) The metric regularization is expressed in terms of statistics, such as mean and variance, and can be estimated from samples.

2 Natural proximal gradient

We review the natural gradient flow in a statistical manifold with Wasserstein and Fisher-Rao metrics, present the natural proximal operators, and introduce a systematic approximation which is suitable for estimation from samples.

2.1 Natural gradients flows

Learning problems are often formulated as the minimization of a loss function, as $\min_{\theta\in\Theta} F(\theta)$, where $\Theta \in \mathbb{R}^d$ is the parameter of the hypothesis class, and $F\colon \Theta \to \mathbb{R}$ is the loss function. As the hypothesis class, we consider a parametrized probability model $\rho\colon \Theta \to \mathcal{P}(\Omega)$, where Ω the sample space, which is a discrete or continuous set on which the distributions are supported. The loss is usually a divergence (sometimes distance) function between the empirical data distribution $\hat{\rho}_{\text{data}}$ and the model distribution ρ_θ.

To find a minimizer, the gradient flow approach is often considered. This flow follows the steepest descent direction of the loss function with respect to a given Riemannian metric. In general, this is defined by

$$\dot{\theta}(t) = -G(\theta(t))^{-1}\nabla_\theta F(\theta(t)), \tag{1}$$

where $G(\theta) \in \mathbb{R}^{d\times d}$ is the matrix representation of the Riemannian metric tensor (for our choice of coordinates), and $\nabla_\theta = (\frac{\partial}{\partial\theta_1}, \dots, \frac{\partial}{\partial\theta_d})^\top$ is the standard (Euclidean) gradient operator. In the context of probability distributions, the metric $G(\theta)$ is pulled back from a natural metric structure on probability space. This implies that for any choice of the parametrization, (1) defines the same flow of probability distributions. Hence it is said to be parametrization invariant.

We will focus on two important statistical metrics on probability space: the Wasserstein metric and the Fisher-Rao metric. These metrics induce the following metric tensors in parameter space. We write $(\cdot,\cdot)$ for the Euclidean or L^2 inner product on the sample space Ω (which might be continuous or discrete).

Definition 1 (Statistical metric tensor on parameter space). *Consider the probability space $(\mathcal{P}(\Omega), g)$ with metric tensor g, and a smoothly parametrized probability model ρ_θ with parameter $\theta \in \Theta$. Then the pull-back G of g is given by*

$$G(\theta) = \Big(\nabla_\theta\rho_\theta, g(\rho_\theta)\nabla_\theta\rho_\theta\Big).$$

(i) If $g_\theta = -(\Delta_{\rho_\theta})^{-1}$, with Δ_{ρ_θ} being the weighted elliptic operator, then $G(\theta)$ is the Wasserstein metric tensor, given by

$$G_W(\theta)_{ij} = \Big(\nabla_{\theta_i}\rho_\theta, (-\Delta_{\rho_\theta})^{-1}\nabla_{\theta_j}\rho_\theta\Big),$$

(ii) If $g_\theta = \frac{1}{\rho_\theta}$, then $G(\theta)$ is the Fisher-Rao metric tensor, given by

$$G_{FR}(\theta)_{ij} = \Big(\nabla_{\theta_i}\rho_\theta, \frac{1}{\rho_\theta}\nabla_{\theta_j}\rho_\theta\Big).$$

Given a metric tensor on parameter space, the standard approach for numerical computation of the gradient flow (1) is the forward Euler method, i.e.,

$$\theta^{k+1} = \theta^k - hG(\theta^k)^{-1}\nabla_\theta F(\theta^k),$$

where $h > 0$ is a step-size. This is known as the natural gradient descent method [2]. In practice, we need to compute the matrix $G(\theta)$ and its inverse at each parameter update, which is difficult in high dimensional parameter spaces.

2.2 Natural proximal operators

We next present another way to approximate the gradient flow, known as the backward Euler or proximal operator method. The proximal operator refers to

$$\theta^{k+1} = \mathrm{Prox}_{hF}(\theta^k) = \arg\min_\theta \; F(\theta) + \frac{D(\theta, \theta^k)}{2h}, \tag{2}$$

where D is a proximity term that penalizes the distance from the current point, and h adjusts the strength. When h is infinity, the proximal operator returns the global minimizer of F. The proximity term is given by the metric function:

$$\begin{aligned} D(\theta, \theta^k) &= \inf_{\theta(t)} \Big\{ \int_0^1 \dot\theta(t)^\top G(\theta(t))\dot\theta(t)dt \colon \theta_0 = \theta,\ \theta_1 = \theta^k \Big\} \\ &= \inf_{\theta(t)} \Big\{ \int_0^1 (\partial_t \rho_{\theta(t)}, g(\rho_{\theta(t)})\partial_t \rho_{\theta(t)})dt \colon \theta_0 = \theta,\ \theta_1 = \theta^k \Big\}. \end{aligned} \tag{3}$$

In rare cases, the proximal operator (2) can be written explicitly.

We shall approximate D in a way that allows for a more friendly computation of the proximal operator. Consider the iterative proximal update

$$\theta^{k+1} = \arg\min_\theta \; F(\theta) + \frac{1}{2h}\Big(\rho_\theta - \rho_{\theta^k}, g(\rho_{\tilde\theta})(\rho_\theta - \rho_{\theta^k})\Big), \tag{4}$$

where $\tilde\theta = \frac{\theta+\theta^k}{2}$. Here the D term in (2) is replaced by a mid-point expression, which is exact up to the order $o(\|\theta - \theta^k\|^2)$. This new proximal operator corresponds to a numerical method known as the semi-backward Euler method. Both (2) and (4) are time discretizations of (1) with first order accuracy. We shall focus on (4), and derive a tractable approximation of the regularization term.

3 Affine space approximation of the metric

Consider the proximity term

$$\tilde D(\theta, \theta^k) = \Big(\rho_\theta - \rho_{\theta^k}, g(\rho_{\tilde\theta})(\rho_\theta - \rho_{\theta^k})\Big). \tag{5}$$

In the following we derive an explicit and computer friendly approximation. To this end, we first consider

$$\frac{1}{2}\tilde D(\theta, \theta^k) = \sup_{\Phi\colon\, \Omega\to\mathbb{R}} (\Phi, \rho_\theta - \rho_{\theta^k}) - \frac{1}{2}\Big(\Phi, g(\rho_{\tilde\theta})^\dagger \Phi\Big), \tag{6}$$

where the maximizer $\Phi = g(\rho_{\tilde{\theta}})(\rho_\theta - \rho_{\theta^k})$ recovers the previous formula. This corresponds to a expressing (5) in terms of its Legendre dual between tangent space and cotangent space in probability space; for a discussion see [6].

Now we restrict the optimization domain (i.e., the set of functions $\Phi\colon \Omega \to \mathbb{R}$) to an affine space of functions of the form

$$\mathcal{F}_\Psi = \Big\{\Phi(x) = \sum_{j=1}^n \xi_j \psi_j(x) = \xi^\top \Psi(x)\colon \xi \in \mathbb{R}^n\Big\},$$

where $\xi = (\xi_j)_{j=1}^n$ is a parameter vector and $\Psi = (\psi_j)_{j=1}^n$ collects a choice of basis functions $\psi_j\colon \Omega \to \mathbb{R}$. This results in following optimization problems:

(i) For the Wasserstein metric, we have

$$\frac{1}{2}\tilde{D}_\Psi^W(\theta, \theta^k) = \sup_{\Phi = \xi^\top \Psi} \mathbb{E}_\theta[\Phi] - \mathbb{E}_{\theta^k}[\Phi] - \frac{1}{2}\mathbb{E}_{\tilde{\theta}}[\|\nabla\Phi\|^2];$$

(ii) For the Fisher-Rao metric, we have

$$\frac{1}{2}\tilde{D}_\Psi^{FR}(\theta, \theta^k) = \sup_{\Phi = \xi^\top \Psi} \mathbb{E}_\theta[\Phi] - \mathbb{E}_{\theta^k}[\Phi] - \frac{1}{2}\mathbb{E}_{\tilde{\theta}}\Big[(\Phi - \mathbb{E}_{\tilde{\theta}}[\Phi])^2\Big].$$

These are quadratic semi-definite programs in ξ. In practice, if using small sample estimates for the expectations, one can add a regularization $-\lambda\|\xi\|^2$, with a small $\lambda > 0$, to ensure strict definiteness and existence of a solution. We proceed to solve these problems. We write $\mathbb{E}_\theta[\psi] = \mathbb{E}_{x\sim\rho_\theta}[\psi(x)]$ and $\partial_l = \frac{\partial}{\partial x_l}$ for the partial derivative w.r.t. the lth sample space variable.[1]

Theorem 1 (Affine space approximation). *Given a basis Ψ, the proximity term $\tilde{D}$ within the affine function space $\mathcal{F}_\Psi = \{\xi^\top\Psi\colon \xi \in \mathbb{R}^n\}$ is given by*

$$\tilde{D}_\Psi(\theta, \theta^k) = (\mathbb{E}_\theta[\Psi] - \mathbb{E}_{\theta^k}[\Psi])^\top \Big(\Psi, g(\rho_\theta)^\dagger \Psi\Big)^\dagger (\mathbb{E}_\theta[\Psi] - \mathbb{E}_{\theta^k}[\Psi]).$$

(i) For the Wasserstein metric, we have

$$\tilde{D}_\Psi^W(\theta, \theta^k) = (\mathbb{E}_\theta[\Psi] - \mathbb{E}_{\theta^k}[\Psi])^\top \Big(\mathfrak{C}^W(\tilde{\theta})\Big)^{-1} (\mathbb{E}_\theta[\Psi] - \mathbb{E}_{\theta^k}[\Psi]),$$

where $\mathfrak{C}^W(\tilde{\theta}) = \mathbb{E}_{\tilde{\theta}}[\sum_l \Big(\partial_l\Psi\Big)\Big(\partial_l\Psi\Big)^\top]$.

(ii) For the Fisher-Rao metric, we have

$$\tilde{D}_\Psi^{FR}(\theta, \theta^k) = (\mathbb{E}_\theta[\Psi] - \mathbb{E}_{\theta^k}[\Psi])^\top \Big(\mathfrak{C}^{FR}(\tilde{\theta})\Big)^{-1} (\mathbb{E}_\theta[\Psi] - \mathbb{E}_{\theta^k}[\Psi]),$$

where $\mathfrak{C}^{FR}(\tilde{\theta}) = \mathbb{E}_{\tilde{\theta}}[\Big(\Psi(x) - \mathbb{E}_{\tilde{\theta}}[\Psi]\Big)\Big(\Psi(x) - \mathbb{E}_{\tilde{\theta}}[\Psi]\Big)^\top]$.

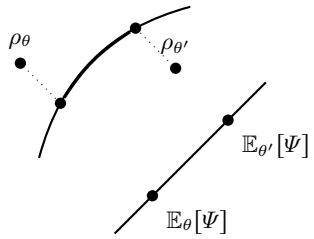

Fig. 1. Illustration of the proximity term over an affine space. Intuitively, the metric between two distributions is measured along a chosen set of statistics.

Remark 1. The matrix $\mathfrak{C}$ has size $n \times n$, corresponding to the dimension of Ψ. For the Fisher-Rao metric, it is the covariance of the basis functions Ψ w.r.t. $\rho_{\tilde{\theta}}$. This corresponds to the Fisher-Rao matrix when the basis is a sufficient statistics of the model. See Fig. 1. Similar observations apply for the Wasserstein metric.

Remark 2. In the case of implicit generative models (used in GANs), where ρ_θ is expressed as the push-forward measure of a latent variable z by a parametrized family of functions $\mathfrak{g}_\theta$, we obtain

$$\tilde{D}(\theta, \theta^k) = (\mathbb{E}_z[\Psi(\mathfrak{g}_\theta(z))] - \mathbb{E}_z[\Psi(\mathfrak{g}_{\theta^k}(z))])^\top \mathbb{E}_z[C(\mathfrak{g}_{\tilde{\theta}}(z))]^{-1} (\mathbb{E}_z[\Psi(\mathfrak{g}_\theta(z))] - \mathbb{E}_z[\Psi(\mathfrak{g}_{\theta^k}(z))]),$$

where C is the corresponding term inside the expectation in Theorem 1.

Proof. (i) For the constrained Wasserstein metric, the gradient of Φ w.r.t. the sample space variable x is $\nabla\Phi(x) = (\sum_{i=1}^n \xi_i \partial_l \psi_i(x))_l$. The squared norm is then

$$\|\nabla\Phi(x)\|^2 = \sum_l (\sum_i \xi_i \partial_l \psi_i(x))^2 = \sum_l \sum_i \xi_i \partial_l \psi_i(x) \sum_j \xi_j \partial_l \psi_j(x) = \xi^\top C^W(x) \xi,$$

where $C^W_{ij}(x) = \sum_l \partial_l \psi_i(x) \partial_l \psi_j(x)$. Now we consider the distance

$$\begin{aligned} \frac{1}{2}\tilde{D}^W_\Psi(\theta, \theta^k) &= \sup_{\Phi = \xi^\top \Psi} \Big(\Phi, \rho_\theta - \rho_{\theta^k} \Big) - \frac{1}{2} \Big((\nabla\Phi)^2, \rho_{\tilde{\theta}} \Big) \\ &= \sup_\xi \xi^\top (\mathbb{E}_\theta[\Psi] - \mathbb{E}_{\theta^k}[\Psi]) - \frac{1}{2} \xi^\top \mathbb{E}_{\tilde{\theta}}[C^W] \xi. \end{aligned}$$

In turn, by first order optimality conditions, at the maximizer we have

$$\xi^* = (\mathbb{E}_{\tilde{\theta}}[C^W])^{-1} (\mathbb{E}_\theta[\Psi] - \mathbb{E}_{\theta^k}[\Psi]).$$

Thus $\tilde{D}^W_\Psi(\theta, \theta^k) = (\mathbb{E}_\theta[\Psi] - \mathbb{E}_{\theta^k}[\Psi])(\mathbb{E}_{\tilde{\theta}}[C^W])^{-1}(\mathbb{E}_\theta[\Psi] - \mathbb{E}_{\theta^k}[\Psi])$.

[1] If the sample space is discrete, we use the discrete differential operator. For an edge weighted graph $G = (V, E, \omega)$, the gradient of $\Phi \in \mathbb{R}^{|V|}$ is $\nabla\Phi = (\omega_{ij}(\Phi_i - \Phi_j))_{(i,j)\in E} \in \mathbb{R}^{|E|}$, and $\mathbb{E}_\theta[\|\nabla\Phi\|^2] = \frac{1}{2}\sum_{i\in V} p_i(\theta) \sum_{j\in V} \omega_{ij}(\Phi_i - \Phi_j)^2$. For details see [7].

(ii) For the Fisher-Rao metric, the term $\|\Phi(z) - \mathbb{E}_{\tilde{\theta}}[\Phi]\|^2$ equals

$$\|\xi^\top \Psi(z) - \xi^\top \mathbb{E}_{\tilde{\theta}}[\Psi]\|^2 = \xi^\top (\Psi(z) - \mathbb{E}_{\tilde{\theta}}[\Psi])(\Psi(z) - \mathbb{E}_{\tilde{\theta}}[\Psi])^\top \xi = \xi^\top C^{FR}(z)\xi,$$

where $C^{FR}(z) = (\Psi(z) - \mathbb{E}_{\tilde{\theta}}[\Psi])(\Psi(z) - \mathbb{E}_{\tilde{\theta}}[\Psi])^\top$. ▫

Example 1 (Order-1 approximation). For the metric approximation with the space of linear functions, $\mathcal{F}_1 = \Big\{\Phi(x) = a^\top x + b\colon a \in \mathbb{R}^m,\ b \in \mathbb{R}\Big\}$, we have:

(i)
$$\tilde{D}_1^W(\theta, \theta^k) = (\mathbb{E}_\theta[x] - \mathbb{E}_{\theta^k}[x])^\top (\mathbb{E}_\theta[x] - \mathbb{E}_{\theta^k}[x]).$$

(ii)
$$\tilde{D}_1^{FR}(\theta, \theta^k) = (\mathbb{E}_\theta[x] - \mathbb{E}_{\theta^k}[x])^\top \Big(\mathbb{E}_{\tilde{\theta}}\Big[(x - \mathbb{E}_{\tilde{\theta}}x)(x - \mathbb{E}_{\tilde{\theta}}x)^\top\Big]\Big)^{-1} (\mathbb{E}_\theta[x] - \mathbb{E}_{\theta^k}[x]).$$

Example 2 (Order-2 approximation). For the space of quadratic functions, $\mathcal{F}_2 = \Big\{\Phi(x) = \frac{1}{2}x^\top Q x + a^\top x + b\colon Q \in \mathbb{R}^{m\times m},\ a \in \mathbb{R}^m,\ b \in \mathbb{R}\Big\}$, we have:

(i)
$$\tilde{D}_2^W(\theta, \theta^k) = \Big(\mathbb{E}_\theta\left[\tfrac{x}{\frac{x\otimes x}{2}}\right] - \mathbb{E}_{\theta^k}\left[\tfrac{x}{\frac{x\otimes x}{2}}\right]\Big)^\top \mathbb{E}_{\tilde{\theta}}\begin{bmatrix} I_m & x^\top \otimes I_m \\ x\otimes I_m & I_m \otimes x x^\top \end{bmatrix}^{-1} \Big(\mathbb{E}_\theta\left[\tfrac{x}{\frac{x\otimes x}{2}}\right] - \mathbb{E}_{\theta^k}\left[\tfrac{x}{\frac{x\otimes x}{2}}\right]\Big).$$

(ii)
$$\tilde{D}_2^{FR}(\theta, \theta^k) = \Big(\mathbb{E}_\theta\left[\tfrac{x}{\frac{x\otimes x}{2}}\right] - \mathbb{E}_{\theta^k}\left[\tfrac{x}{\frac{x\otimes x}{2}}\right]\Big)^\top \Big(\mathfrak{C}^{FR}(\tilde{\theta})\Big)^{-1} \Big(\mathbb{E}_\theta\left[\tfrac{x}{\frac{x\otimes x}{2}}\right] - \mathbb{E}_{\theta^k}\left[\tfrac{x}{\frac{x\otimes x}{2}}\right]\Big),$$

where $\otimes$ is the Kronecker product (e.g., $x \otimes x$ is an $m^2 \times 1$ vector), and

$$\mathfrak{C}^{FR} = \mathbb{E}_{\tilde{\theta}}\left[\Big(\left[\tfrac{x}{\frac{x\otimes x}{2}}\right] - \mathbb{E}_{\tilde{\theta}}\left[\tfrac{x}{\frac{x\otimes x}{2}}\right]\Big)\Big(\left[\tfrac{x}{\frac{x\otimes x}{2}}\right] - \mathbb{E}_{\tilde{\theta}}\left[\tfrac{x}{\frac{x\otimes x}{2}}\right]\Big)^\top\right].$$

4 Numerical examples

The optimization loop can be implemented as shown in Algorithm 1. Here the proximal operator is computed by a short gradient iteration. In practice we can replace the expectations by sample averages, $\mathbb{E}_\theta[f] \approx \frac{1}{N}\sum_{i=1}^N f(x^{(i)})$, with $x^{(i)}$ i.i.d. from ρ_θ. For the basis Ψ we can choose low order polynomials, as in Examples 1 and 2, but even random functions worked well in our experiments. The optimal choice will balance low dimension and relevant statistics for the model under consideration. Orthogonality tends to be beneficial.

4.1 Maximum likelihood estimation for hierarchical models

We consider binary k-interaction models, which are exponential families $\rho_\theta(x) = \exp(\theta^\top A(x))/Z(\theta)$, $x \in \{0,1\}^m$, with sufficient statistics $A_\lambda(x) = \prod_{i\in\lambda}(-1)^{x_i}$, for $\lambda \subseteq \{1,\dots,m\}$, $|\lambda| \leqslant k$. We use $\Psi_j(x) = (-1)^{x_j}$, $j \in \{1,\dots,m\}$, which are sufficient statistics for the 1-interaction model (independence model). We draw target distributions uniformly from the simplex and compute the MLEs. We compare Euclidean, Fisher-Rao, Wasserstein, and proximals. For each problem and method we run grid search over the step size α and proximal strength h, which are kept fixed during optimization. The results are shown in Fig. 2.

Algorithm 1 Natural gradient with affine space proximal approximation.

Require: Loss F, basis of affine space Ψ, proximal step-size h, step-size α

for $t = 0$ **to** max outer iterations **do**

$\mathfrak{C}(\theta) = \mathrm{cov}_\theta[\Psi]^{-1}$ (Fisher-Rao); $\mathfrak{C}(\theta) = \mathbb{E}_\theta[\sum_l (\partial_l \Psi)(\partial_l \Psi)^\top]^{-1}$ (Wasserstein)

for $t' = 0$ **to** max inner iterations **do**

$\nabla_{\theta'} D(\theta, \theta') \leftarrow \frac{1}{2} \nabla_{\theta'} \mathbb{E}_{\theta'}[\Psi^\top] \mathfrak{C}(\theta)(\mathbb{E}_{\theta'}[\Psi] - \mathbb{E}_\theta[\Psi])$

$\theta' \leftarrow \theta' - \alpha(\nabla_{\theta'} F(\theta') + \frac{1}{2h} \nabla D_{\theta'}(\theta, \theta'))$

$\theta \leftarrow \theta'$

4.2 Classification on CIFAR-10

Here we present an image classification task on the CIFAR-10 dataset [5] using the Wasserstein proximal method. We use a simple CNN with two convolutional layers followed by two fully-connected layers, with ReLU activations. In this experiment F is the categorical cross-entropy loss and $D = \tilde{D}_\Psi^W$ is the Order 1 or Order 2 Wasserstein approximation. The specific details of our experiments can be found in Appendix A (online). Fig. 2 provides the results, where we give curves for the validation error per epoch. As a baseline, we also give results when performing SGD many times per epoch, but without regularization. We see that the best result comes from the Order 2 Wasserstein distance approximation.

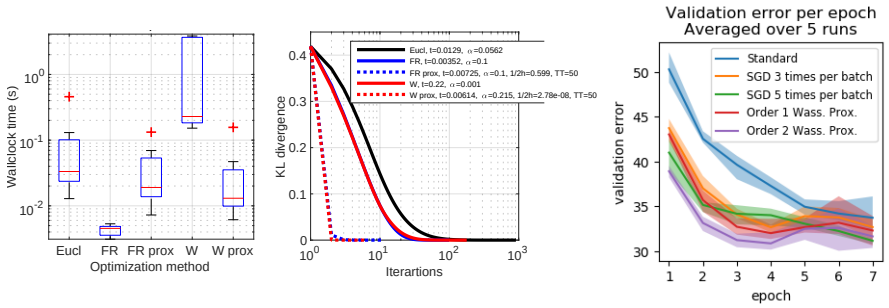

Fig. 2. Left: MLE wall-clock computation times until the KL-divergence is within 10^{-9} of optimal, for 4 binary variables and Ψ the independence model, and typical optimization curves. Right: The learning curves for the image classification task on CIFAR-10. Each experiment was averaged over 5 runs. The bold lines represent the average, and the envelopes are the minimum and maximum achieved.

5 Discussion

We studied sampling–friendly implementations of the natural gradient based on the proximal operator. We approximate the proximity penalty by an affine space restriction in the Legendre dual formulation. This gives rise to a lower

dimensional metric, expressed in expectation parameters, which can be estimated from samples. We cover both Fisher-Rao and Wasserstein metrics. Especially for the Wasserstein proximal, our method offers significant savings in computation time and provide improvement in validation error (in CIFAR-10 classification).

Acknowledgement This project has received funding from AFOSR MURI FA9550-18-1-0502 and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 757983).

References

1. S. Amari. *Differential-geometrical methods in statistics*. Lecture notes in statistics. Springer-Verlag, 1985.
2. S. Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2):251–276, 1998.
3. G. Desjardins, K. Simonyan, R. Pascanu, and K. Kavukcuoglu. Natural neural networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *NIPS 28*, pages 2071–2079. Curran Associates, Inc., 2015.
4. R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A 222*, pages 309–368, 1922.
5. A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.
6. W. Li. Geometry of probability simplex via optimal transport. *arXiv:1803.06360 [math]*, 2018.
7. W. Li and G. Montúfar. Natural gradient via optimal transport. *Information Geometry*, 1(2):181–214, Dec 2018.
8. W. Li and G. Montúfar. Ricci curvature for parametric statistics via optimal transport. *arXiv:1807.07095 [cs, math, stat]*, 2018.
9. T. Liang, T. A. Poggio, A. Rakhlin, and J. Stokes. Fisher-Rao metric, geometry, and complexity of neural networks. *CoRR*, abs/1711.01530, 2017.
10. A. Lin, W. Li, S. Osher, and G. Montúfar. Wasserstein proximal of GANs. In *CAM reports*, 2018.
11. L. Malagò, L. Montrucchio, and G. Pistone. Wasserstein Riemannian geometry of Gaussian densities. *Information Geometry*, 1(2):137–179, Dec 2018.
12. J. Martens and R. Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *ICML 32*, volume 37 of *PMLR*, pages 2408–2417, 2015.
13. K. Modin. Geometry of matrix decompositions seen through optimal transport and Information Geometry. *Journal of Geometric Mechanics*, 9(3):335–390, 2017.
14. F. Otto. The geometry of dissipative evolution equations the porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.
15. H. Park, S. Amari, and K. Fukumizu. Adaptive natural gradient learning algorithms for various stochastic models. *Neural Networks*, 13(7):755 – 764, 2000.
16. J. Peters and S. Schaal. Natural actor-critic. *Neurocomputing*, 71(7):1180 – 1190, 2008. Progress in Modeling, Theory, and Application of Computational Intelligenc.
17. C. R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc. 37*, pages 81–89, 1945.
18. C. Villani. *Optimal Transport: Old and New*. Number 338 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009.

Appendix for image classification on CIFAR-10

Here is the detailed version of our experiments, for image classification on CIFAR-10. We use a simple CNN with two convolutional layers (each with 32 filters, with a kernel size of 3×3, a stride of 1, and zero padding), followed by two fully-connected layers eaching having 512 nodes. For the optimizer, we use standard stochastic gradient descent (SGD) with momentum value 0.95 and learning rate of 0.001.

For the Wasserstein distance, if we denote the (deterministic) output of our neural network as $f(x, \theta)$ (the log probability vector), the loss function as $L(y, f(x, \theta))$ where x is the image and y the label, and the dataset as $\mathcal{D}$, then the Order 1 and Order 2 approximations for the Wasserstein distance on image classification on CIFAR-10 are: Order 1 approximation:

$$\tilde{D}_1^W(\theta, \theta^k) = \|\mathbb{E}_{x\sim\mathcal{D}}[f(x,\theta)] - \mathbb{E}_{x\sim\mathcal{D}}[f(x,\theta^k)]\|^2, \tag{7}$$

and the Order 2 approximation:

$$\begin{aligned}\tilde{D}_2^W(\theta, \theta^k) = &\|\mathbb{E}_{x\sim\mathcal{D}}[f(x,\theta)] - \mathbb{E}_{x\sim\mathcal{D}}[f(x,\theta^k)]\|^2 \\ &+ \mathrm{Tr}\Big(\mathrm{Var}_{x\sim\mathcal{D}}[f(x,\theta)] + \mathrm{Var}_{x\sim\mathcal{D}}[f(x,\theta^k)] \\ &- 2\left(\mathrm{Var}_{x\sim\mathcal{D}}[f(x,\theta^k)]^{1/2}\mathrm{Var}_{x\sim\mathcal{D}}[f(x,\theta)]\mathrm{Var}_{x\sim\mathcal{D}}[f(x,\theta^k)]^{1/2}\right)\Big)\end{aligned} \tag{8}$$

We present our experiments on 5 different settings: (1) Standard learning with no regularization, (2) performing SGD 3 times per batch, (3) performing SGD 5 times per batch, (4) using the Order 1 Wasserstein Proximal (with $m = 3$ and $h = 2$), (5) and using the Order 2 Wasserstein proximal (with $m = 5$ and $h = 1$). From Fig. 2, we see that using the Order 2 Wasserstein proximal provides the best results. We note that performing SGD a number of times per batch is presented as a baseline, as we experimentally found that they also provided improvements in validation error per epoch (but they are not the best as we can see from Fig. 2).

Algorithm 2 Wasserstein Proximal Natural Gradient for Neural Networks

Require: Loss function L, neural network $f(x, \theta)$, Order 1 or 2 Wasserstein distance approximation D, and data-label pairs $\{(x, y)\}$ from dataset $\mathcal{D}$.

Require: m number of gradient descent steps, and h strength of the proximal term

while stopping criteria not met **do**

Sample a mini-batch of image-label pairs $\{(x_b, y_b)\}_{b=1}^B \in \mathcal{D}$

Approximately solve (by performing SGD m times)

$$\theta^{k+1} \leftarrow \mathrm{argmin}_\theta \left\{ \frac{1}{B}\sum_{b=1}^{B} L(y, f(x,\theta)) + \frac{1}{2h} D(\theta, \theta^k) \right\}$$