A Formalization of the Natural Gradient Method for General Similarity Measures

Anton Mallasto(✉), Tom Dela Haije, and Aasa Feragen

University of Copenhagen, Copenhagen, Denmark
{mallasto,haije,aasa}@di.ku.dk

Abstract. In optimization, the natural gradient method is well-known for likelihood maximization. The method uses the Kullback–Leibler (KL) divergence, corresponding infinitesimally to the Fisher–Rao metric, which is pulled back to the parameter space of a family of probability distributions. This way, gradients with respect to the parameters respect the Fisher–Rao geometry of the space of distributions, which might differ vastly from the standard Euclidean geometry of the parameter space, often leading to faster convergence. The concept of natural gradient has in most discussions been restricted to the KL-divergence/Fisher–Rao case, although in information geometry the local C^2 structure of a general divergence has been used for deriving a closely related Riemannian metric analogous to the KL-divergence case. In this work, we wish to cast natural gradients into this more general context and provide example computations, notably in the case of a Finsler metric and the p-Wasserstein metric. We additionally discuss connections between the natural gradient method and multiple other optimization techniques in the literature.

Keywords: Optimization · Natural gradient · Statistical manifolds

1 Introduction

The natural gradient method [2] in optimization originates from *information geometry* [4], which utilizes the Riemannian geometry of statistical manifolds (the parameter spaces of model families) endowed with the *Fisher–Rao metric*. The natural gradient is used for minimizing the *Kullback–Leibler* (KL) divergence, a *similarity measure* between a model distribution and a target distribution, that can be shown to be equivalent to maximizing model likelihood of given data. The success of natural gradient in optimization stems from accelerating likelihood maximization and providing infinitesimal invariance to reparametrizations of the model, providing robustness towards arbitrary parametrization choices.

In the modern formulation of the natural gradient, a *Riemannian metric* on the statistical manifold is chosen, with respect to which the gradient of the given similarity is computed [4, Sec. 12]. The choice of the Riemannian metric should, however, relate closely to the similarity measure being minimized.

F. Nielsen and F. Barbaresco (Eds.): GSI 2019, LNCS 11712, pp. 599–607, 2019.
https://doi.org/10.1007/978-3-030-26980-7_62

We have illustrated this in Fig. 1, where model selection for Gaussian process regression is carried out by maximizing the prior-likelihood of the data with natural gradients stemming form different metrics. Clearly, the Fisher–Rao metric—which infinitesimally corresponds to the KL-divergence—achieves the fastest convergence.

An example of an approach to choose a related Riemannian metric is the classical Newton's method that derives a metric from the Hessian of a convex objective function, or its absolute value in the non-convex case [7]. Unfortunately, evaluating the Hessian is not feasible in some cases. Instead, we can compute a *local Hessian*, which corresponds to a local second order expansion of the similarity measure [3]. This approach generalizes the natural gradient from the KL-divergence case to general similarity measures, and to avoid confusion with the well-known KL-divergence setting, we refer to this approach as the *formal natural gradient*. We furthermore discuss the similarities between the trust region, proximal, and natural gradient methods in Sect. 3 and provide example computations in Sect. 4.

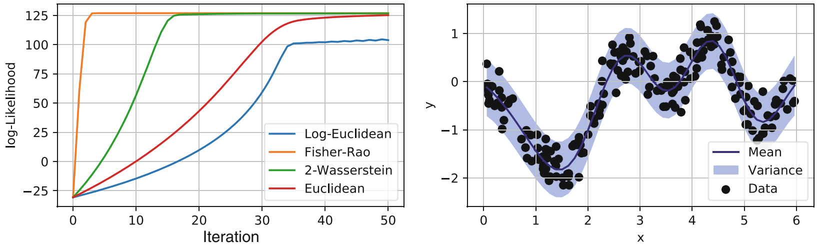

Fig. 1. Maximizing prior likelihood for Gaussian process regression using natural gradients under different metrics on Gaussian distributions. Convergence plots on left. Data and model fit, with optimal exponentiated quadratic kernel parameters, on right.

2 Useful Metrics via Formalizing the Natural Gradient

The natural gradient is computed with respect to a chosen metric on the statistical manifold, which often results from pulling back a metric between distributions. This way, the gradient takes into account how the metric on distributions penalizes movement into different directions. We will now review how the natural gradient is computed given a Riemannian metric. Then, we introduce the formal natural gradient, which derives this metric from the similarity measure.

Statistical Manifold. Let $\mathrm{AC}(X)$ denote the set of absolutely continuous probability distributions on some manifold X. A *statistical manifold* is defined by a triple (X, Θ, ρ), where X is called the *sample space* and $\Theta \subseteq \mathbb{R}^n$ the *parameter space*. Then, $\rho\colon \Theta \to \mathrm{AC}(X)$ maps a parameter to a density, given by $\rho\colon \theta \mapsto \rho_\theta(\cdot)$, for any $\theta \in \Theta$. Abusing terminology, we also call Θ the statistical manifold.

Cost Function. Let a *similarity measure* $c^*\colon \mathrm{AC}(X) \times \mathrm{AC}(X) \to \mathbb{R}_{\geq 0}$ (e.g. a metric or an information divergence) be defined on $\mathrm{AC}(X)$ satisfying $c^*(\rho, \rho') = 0$ if and only if $\rho = \rho'$. Assume c^* to be strictly convex in ρ. Given a target distribution $\rho \in \mathrm{AC}(X)$ and a statistical manifold (X, Θ, ρ), we wish to minimize the *cost function* $c \to \Theta \times \mathrm{AC}(X) \to \mathbb{R}_{\geq 0}$ given by

$$c(\theta, \rho) = c^*(\rho_\theta, \rho). \tag{2.1}$$

If $\rho = \rho_{\theta'}$ for some $\theta' \in \Theta$, then by abuse of notation we write $c(\theta, \theta')$. We finally assume that $\theta \mapsto c(\theta, \theta')$ is C^2 whenever $\theta \neq \theta'$.

Natural Gradient. Assume a Riemannian structure (Θ, g^Θ) on the statistical manifold. The *Riemannian metric* g^Θ induces a *metric tensor* G^Θ, given by $g^\Theta_\theta(u, v) = u^T G^\Theta_\theta v$ and a *distance function* which we denote by d_Θ. The vectors u, v belong to the *tangent space* $T_\theta\Theta$ at θ. It is common intuition that the negative gradient $v = -\nabla_\theta c(\theta, \rho)$ gives the direction of maximal descent for c. However, this is only true on a Euclidean manifold. Consider

$$\hat{v} = \operatorname*{arg\,min}_{v \in T_\theta\Theta : d_\Theta(\theta, \theta+v) = \Delta} c(\theta + v, \rho), \tag{2.2}$$

where $\theta + v$ is to be understood in a chart of Θ, and $\Delta > 0$ defines the radius of the trust region. Linearly approximating the objective and quadratically approximating the constraint, this is solved using Lagrangian multipliers, giving the *natural gradient*

$$\hat{v} = -\frac{1}{\lambda}\left[G^\Theta_\theta\right]^{-1} \nabla_\theta c(\theta, \rho), \tag{2.3}$$

for some Lagrangrian multiplier $\lambda > 0$, which we refer to as the *learning rate.* Below, a similar derivation is carried out in more detail.

Formal Natural Gradient. Traditionally, the natural gradient uses the Fisher–Rao metric when the similarity measure used is the KL-divergence. We will now show, how a trust region formulation with respect to the chosen similarity measure can be used to derive a natural metric under which the natural gradient can be computed, resulting in the *formal natural gradient.* Thus, consider the minimization task

$$\hat{v} := \operatorname*{arg\,min}_{v \in T_\theta\Theta,\ c(\theta+v, \theta) = \Delta} c(\theta + v, \rho). \tag{2.4}$$

We approximate the constraint by the second degree Taylor expansion

$$c(\theta + v, \theta) \approx \frac{1}{2} v^T \left(\nabla^2_{\eta \to \theta} c(\eta, \theta)\right) v, \tag{2.5}$$

where the 0^{th} and 1^{st} degree terms disappear as $c(\theta + v, \theta)$ has a minimum 0 at $v = 0$. We call the symmetric positive definite matrix $H^c_\theta := \nabla^2_{\eta \to \theta} c(\eta, \theta)$ the *local Hessian.* Then, we further approximate the objective function

$$c(\theta + v, \rho) \approx c(\theta, \rho) + \nabla_\theta c(\theta, \rho)^T v. \tag{2.6}$$

Writing the approximate Langrangian $\mathcal{L}(v)$ of (2.4) with a multiplier $\lambda > 0$, we get

$$\mathcal{L}(v) \approx c(\theta, \rho) + \nabla_\theta c(\theta, \rho)^T v + \frac{\lambda}{2} v^T \left(\nabla^2_{\eta \to \theta} c(\eta, \theta)\right) v. \tag{2.7}$$

Thus by the method of Langrangian multipliers, (2.4) is solved as

$$\hat{v} = -\frac{1}{\lambda} \left[H^c_\theta\right]^{-1} \nabla_\theta c(\theta, \rho). \tag{2.8}$$

We refer to $\hat{v}$ as the *formal natural gradient* with respect to c.

Remark 1. We could have just substituted $\eta = \theta$ in the local Hessian if $\nabla^2_\eta c(\eta, \theta)$ was continuous at η. However, when studying Finsler metrics later in this work, the expression has a discontinuity at $\eta = \theta$. Therefore, a direction for a limit has to be chosen, and as a straight-forward candidate we compute the limit from the direction of the gradient.

Metric Interpretation. The local Hessian G^c_θ can be seen as a metric tensor at any $\theta \in \Theta$, inducing an inner product $g^c_\theta \colon T_\theta\Theta \times T_\theta\Theta \to \mathbb{R}$ given by $g^c_\theta(v, u) = v^T H^c_\theta u$. This imposes a *pseudo-Riemannian* structure on Θ, forming the pseudo-Riemannian manifold (Θ, g^c). Therefore, G^c_x provides us a natural metric under which to compute the natural gradient for a general c^*. If ρ has a full rank Jacobian everywhere, then a Riemannian metric is retrieved. Also, there is an obvious *pullback* structure at play. Recall, that the cost is defined by $c(\theta, \theta') = c^*(\rho_\theta, \rho_{\theta'})$. Then, computing the local Hessian yields

$$H^c_\theta = J^T_\theta H^{c^*}_{\rho_\theta} J_\theta, \tag{2.9}$$

where $H^{c^*}_{\rho_\theta} = \nabla^2_{\rho \to \rho_\theta} c^*(\rho, \rho_\theta)$. Thus, H^c results from pulling back the c^* induced metric tensor H^{c^*} on $\mathrm{AC}(X)$ to the statistical manifold Θ. In information geometry, this Riemannian metric is said to be induced by the corresponding divergence (similarity measure) [3]. Therefore, the formal natural gradient is just the Riemannian gradient under the aforementioned induced metric.

Asymptotically Newton's Method. We provide a straightforward result, stating that the local Hessian approaches the actual Hessian in the limit, thus the formal natural gradient method approaches Newton's method. This is well known in the Fisher–Rao case, but for completeness we provide the result for the formal natural gradient.

Proposition 1. *Assume $c(\theta, \rho) = c(\theta, \theta')$ for some $\theta' \in \Theta$, and that c is C^2 in θ. Then, the natural gradient yields asymptotically Newton's method.*

Proof. The Hessian at θ is given by $\nabla^2_\theta c(\theta, \theta')$. Then, as c is C^2 in the first argument, passing the limit $\theta \to \theta'$ yields

$$H^c_\theta = \nabla^2_{\eta \to \theta} c(\eta, \theta) \overset{\theta \to \theta'}{\to} \nabla^2_{\eta \to \theta'} c(\eta, \theta') = \nabla^2_{\eta = \theta'} c(\eta, \theta'), \tag{2.10}$$

where the last expression is the Hessian at θ'.

3 Loved Child has Many Names – Related Methods

In this section, we discuss connections between seemingly different optimization methods. Some of these connections have already been reported in the literature, some are likely to be known to some extent in the community. However, the authors are unaware of previous work drawing out these connections in their full extent. We provide such a discussion, and then present other related connections.

As discussed in [14], *proximal methods* and *trust region methods* are equivalent up to learning rate. Trust region methods employ an l^2-metric constraint

$$x_{t+1} = \underset{x : \|x - x_t\|_2 \leq \varDelta}{\arg\min} \; f(x), \;\; \varDelta > 0, \tag{3.1}$$

whereas proximal methods include a l^2-metric penalization term

$$x_{t+1} = \arg\min_x \left\{ f(x) + \frac{1}{2\lambda} \|x - x_t\|_2^2 \right\}, \;\; \lambda > 0, \tag{3.2}$$

The two can be shown to be equivalent up to learning rate via Lagrangian duality.

Instead of the l^2 metric penalization, *mirror gradient descent* [13] employs a more general *proximity function* $\varPsi \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{>0}$, that is strictly convex in the first argument. Then, the mirror descent step is given by

$$x_{t+1} = \arg\min_x \left\{ \langle x - x_t, \nabla f(x_t) \rangle + \frac{1}{\lambda} \varPsi(x, x_t) \right\}. \tag{3.3}$$

Commonly, $\varPsi$ is chosen to be a *Bregman divergence* D_g, defined by choosing a strictly convex C^2 function g and writing

$$D_g(x, x') = g(x) - g(x') - \langle \nabla g(x'), x - x' \rangle. \tag{3.4}$$

To explain how these methods are related to the natural gradient, assume that we are minimizing a general similarity measure $c(x, y)$ with respect to x, as in Sect. 2. Recall, that we first defined the natural gradient as a *trust region step*. In order to derive an analytical expression for the iteration, we approximated the objective function with the first order Taylor polynomial and the constraints by the local Hessian and then used Lagrangian duality to yield a *proximal expression*, which yields the formal natural gradient when solved. In Sect. 4, we will show how this workflow indeed corresponds to known examples of the natural gradient.

Further Connections. Raskutti and Mukherjee [16] showed, that Bregman divergence proximal mirror gradient descent is equivalent to the natural gradient method on the *dual manifold* of the Bregman divergence. Khan et al. [8], consider a KL divergence proximal algorithm for learning *conditionally conjugate exponential families*, which they show to correspond to a natural gradient step. For exponential families, the KL-divergence corresponds to a Bregman divergence, and so the natural gradient step is on the *primal manifold* of the Bregman divergence. Thus the result seems to conflict with the resut in [16]. However, this can be explained, as the gradient is taken with respect to a different argument of the divergence, i.e., they consider $\nabla_x D_g(x', x)$ and not $\nabla_x D_g(x, x')$. It is intriguing how two different geometries are involved in this choice.

Pascanu and Bengio [15] remarked on the connections between the natural gradient method and Hessian-free optimization [11], Krylov Subspace Descent [17], and TONGA [9]. The main connection between Hessian-free optimization and Krylov subspace descent is the use of *extended Gauss–Newton approximation of the Hessian* [18], which gives a similar square form involving the Jacobian as the *pullback* Fisher–Rao metric on a statistical manifold. The connection was further studied by Martens [12], where an equivalence criterion between the Fisher–Rao natural gradient and extended Gauss–Newton was given.

4 Example Computations

We will now provide example computations for the local Hessian H^c of different similarity measures c, as it is the essential object in computing the natural gradient given in (2.8). We first show that in the cases of KL-divergence and a Riemannian metric, the definition of the formal natural gradient matches the classical definition, as expected. Furthermore, we contribute local Hessians for general f-divergences and Finsler metrics, specifically for the p-Wasserstein metrics.

Natural Gradient of f-Divergences. Let $\rho, \rho' \in \mathrm{AC}(X)$ and $f\colon \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ be a convex function satisfying $f(1) = 0$. Then, the f-divergence from ρ' to ρ is

$$D_f(\rho||\rho') = \int_X \rho(x) f\left(\frac{\rho'(x)}{\rho(x)}\right) dx. \tag{4.1}$$

Now, consider the statistical manifold $(\mathbb{R}^d, \Theta, \rho)$, and compute the local Hessian

$$\left[H_\theta^{D_f}\right]_{ij} = \nabla^2 f(1) \int_X \frac{\partial \log \rho_\theta(x)}{\partial \theta_i} \frac{\partial \log \rho_\theta(x)}{\partial \theta_j} \rho_\theta(x) dx. \tag{4.2}$$

Substituting $f = -\log$ in (4.1) results in the KL-divergence, denoted by $D_{\mathrm{KL}}(\rho||\rho')$. Noticing that $\nabla^2 f(1) = 1$ with this substitution, we can write (4.2) as $H_\theta^{D_f} = \nabla^2 f(1) H_\theta^{D_{\mathrm{KL}}}$, where the local Hessian $H_\theta^{D_{\mathrm{KL}}}$ is also the Fisher–Rao metric tensor at θ, and thus the natural gradient of Amari [2] is retrieved.

Natural Gradient of Riemannian Distance. Let (M, g) be a Riemannian manifold with the induced distance function d_g and the metric tensor at $\rho \in M$ denoted by G_ρ^M. Finally, denote by ρ_θ a submanifold of M parametrized by $\theta \in \Theta$. Then, when $c = \frac{1}{2}d^2$, we compute $G_\theta^{\frac{1}{2}d_g}$ as follows

$$\begin{aligned}\left[H_\theta^{\frac{1}{2}d^2}\right]_{ij} =& \frac{1}{2}\left(\frac{\partial}{\partial \theta_j}\rho_\theta\right)^T \left[\nabla^2_{\rho_\eta \to \rho_\theta} d^2(\rho_\eta, \rho_\theta)\right] \left(\frac{\partial}{\partial \theta_i}\rho_\theta\right) \\ &+ \frac{1}{2}\left[\frac{\partial^2}{\partial \theta_j \partial \theta_i}\rho_\theta\right] \left[\nabla_{\rho_\eta \to \rho_\theta} d^2(\rho_\eta, \rho_\theta)\right],\end{aligned} \tag{4.3}$$

as $\theta' \to \theta$, the second term vanishes. Finally, $\nabla^2_{\rho_\eta \to \rho_\theta} d^2(\rho_\eta, \rho_\theta) = 2G^M_{\rho_\theta}$, thus

$$H_\theta^{\frac{1}{2}d_g} = J_\theta^T G^M_{x_\theta} J_\theta, \tag{4.4}$$

where $J_\theta = \frac{\partial}{\partial\theta}\rho_\theta$ denotes the Jacobian. Therefore, the formal natural gradient corresponds to the traditional coordinate-free definition of a gradient on a Riemannian manifold, when the metric is given by the pullback.

Natural Gradient of Finsler Distance. Let (M, F) denote a Finsler manifold, where $F_\rho \colon T_\rho M \to \mathbb{R}_{\geq 0}$, for any $\rho \in M$, is a *Finsler metric*, satisfying the properties of strong convexity, positive 1-homogeneity and positive definiteness. Then, a distance d_F is induced on M by

$$d_F(\rho, \rho') = \inf_\gamma \int_0^1 F_{\gamma(t)}(\dot\gamma(t))dt, \ \rho, \rho' \in M \tag{4.5}$$

where γ is any continuous, unit-parametrized curve with $\gamma(0) = \rho$ and $\gamma(1) = \rho'$.

The *fundamental tensor* G^F of F at (ρ, v) is defined as $G^F_\rho(v) = \frac{1}{2}\nabla^2_v F^2_\rho(v)$. Then, G^F_ρ is 0-homogeneous as the second differential of a 2-homogeneous function. Therefore, $G^F_\rho(\lambda v) = G^F_\rho(v)$ for any $\lambda > 0$. Furthermore, $G^F_\rho(v)$ is positive-definite when $v \neq 0$. Now, let $u = -J_\theta \nabla_\theta d^2_F(\rho_\theta, \rho')$, and as we can locally write $d^2_F(\rho, \rho') = F^2_{\rho\theta}(v)$ for a suitable v, then

$$H_\theta^{\frac{1}{2}d_F^2} = \frac{1}{2}\nabla^2_{\eta\to\theta} d^2_F(\rho_\eta, \rho_\theta) = \frac{1}{2}\lim_{\lambda\to 0} \nabla^2_{v=\lambda u} F^2_{\rho_\theta}(v) = J_\theta^T G^F_{\rho_\theta}(u) J_\theta. \tag{4.6}$$

Coordinate-free gradient descent on Finsler manifolds has been studied by Bercu [5]. The formal natural gradient differs slightly from this, as we use $v = -J_\theta \nabla_\theta d^2_F(\rho_\theta, \rho')$ in the preconditioning matrix $G^F_{(\rho_\theta, v)}$ (see Remark 1), where as in [5], v is chosen to maximize the descent. Thus the natural gradient descent in the Finsler case approximates the geometry in the direction of the gradient quadratically to improve the descent, but fails to take the entire local geometry into account.

***p*-Wasserstein Metric.** Let $X = \mathbb{R}^n$ and $\rho \in \mathcal{P}_p(X)$ if

$$\int_X d_2^p(x_0, x)\rho(x)dx, \text{ for some } x_0 \in X, \tag{4.7}$$

where d_2 is the Euclidean distance. Then, the p-Wasserstein distance W_p between $\rho, \rho' \in \mathcal{P}_p(X)$ is given by

$$W_p(\rho, \rho') = \left(\inf_{\gamma \in \text{ADM}(\rho,\rho')} \int_{X\times X} d_2^p(x, x')d\gamma(x, x')\right)^{\frac{1}{p}}, \tag{4.8}$$

where $\text{ADM}(\rho, \rho')$ is the set of joint measures with marginal densities ρ and ρ'. The p-Wasserstein distance is induced by a Finsler metric [1], given by

$$F_\rho(v) = \left(\int_X \|\nabla\Phi_v\|_2^p d\rho\right)^{\frac{1}{p}}, \tag{4.9}$$

where $v \in T_\rho \mathcal{P}_p(X)$ and Φ_v satisfies $v(x) = -\nabla \cdot (\rho(x) \nabla_x \Phi_v(x))$ for any $x \in X$, where $\nabla \cdot$ is the divergence operator. Now, choose $v = -J_\theta \nabla_\theta W_p^2(\rho_\theta, \rho)$. Then, through a cumbersome computation, we compute how the local Hessian acts on two tangent vectors $d\theta_1, d\theta_2 \in T_\theta \Theta$

$$\begin{aligned} &H_\theta^{\frac{1}{2}W_p^2}(d\theta_1, d\theta_2) \\ &= (2-p) F_{\rho_\theta}^{2(1-p)}(v) \left(\int_X \|\nabla \Phi_v\|_2^{p-2} \langle \nabla \Phi_{d\theta_1}, \nabla \Phi_v \rangle d\rho_\theta \right) \\ &\quad \times \left(\int_X \|\nabla \Phi_v\|_2^{p-2} \langle \nabla \Phi_{d\theta_2}, \nabla \Phi_v \rangle d\rho_\theta \right) \\ &\quad + F_{\rho_\theta}^{2-p}(v) \int_X \|\nabla \Phi_v\|_2^{p-2} \langle \nabla \Phi_{d\theta_1}, \nabla \Phi_{d\theta_2} \rangle d\rho_\theta \\ &\quad + (p-2) F_{\rho_\theta}^{2-p}(v) \int_X \|\nabla \Phi_v\|_2^{p-4} \langle \nabla \Phi_{d\theta_1}, \nabla \Phi_v \rangle \langle \nabla \Phi_{d\theta_2}, \nabla \Phi_v \rangle d\rho_\theta, \end{aligned} \tag{4.10}$$

where $J_\theta d\theta_i = -\nabla \cdot (\rho_\theta \nabla \Phi_{d\theta_i})$ for $i = 1, 2$. The case $p = 2$ is special, as the 2-Wasserstein metric is induced by a Riemannian metric, whose pullback can be recovered by substituting $p = 2$ in (4.10), yielding

$$H_\theta^{\frac{1}{2}W_2^2}(d\theta_1, d\theta_2) = \int_X \langle \nabla \Phi_{d\theta_1}, \nabla \Phi_{d\theta_2} \rangle d\rho_\theta. \tag{4.11}$$

This yields the natural gradient of W_2^2 as introduced in [6,10].

Acknowledgements. The authors were supported by Centre for Stochastic Geometry and Advanced Bioimaging, and a block stipendium, both funded by a grant from the Villum Foundation. We furthermore wish to thank the anonymous reviewers for their very useful comments.

References

1. Agueh, M.: Finsler structure in the p-Wasserstein space and gradient flows. Comptes Rendus Mathematique **350**(1–2), 35–40 (2012)
2. Amari, S.I.: Natural gradient works efficiently in learning. Neural Comput. **10**(2), 251–276 (1998)
3. Amari, S.i.: Divergence function, information monotonicity and information geometry. In: Workshop on Information Theoretic Methods in Science and Engineering (WITMSE). Citeseer (2009)
4. Amari, S.I.: Information Geometry and Its Applications. Springer, Tokyo (2016). https://doi.org/10.1007/978-4-431-55978-8
5. Bercu, G.: Gradient methods on Finsler manifolds. In: Proceedings of the Workshop on Global Analysis, Differential Geometry and Lie Algebras, pp. 230–233 (2000)
6. Chen, Y., Li, W.: Natural gradient in Wasserstein statistical manifold. arXiv preprint arXiv:1805.08380 (2018)
7. Dauphin, Y.N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., Bengio, Y.: Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In: Advances in Neural Information Processing Systems, pp. 2933–2941 (2014)

8. Khan, M.E., Baqué, P., Fleuret, F., Fua, P.: Kullback-Leibler proximal variational inference. In: Advances in Neural Information Processing Systems, pp. 3402–3410 (2015)
9. Le Roux, N., Manzagol, P.A., Bengio, Y.: Topmoumoute online natural gradient algorithm. In: Advances in Neural Information Processing Systems, pp. 849–856 (2008)
10. Li, W., Montúfar, G.: Natural gradient via optimal transport. Inf. Geom. **1**(2), 181–214 (2018)
11. Martens, J.: Deep learning via Hessian-free optimization. In: ICML, vol. 27, pp. 735–742 (2010)
12. Martens, J.: New insights and perspectives on the natural gradient method. arXiv preprint arXiv:1412.1193 (2014)
13. Nemirovsky, A.S., Yudin, D.B.: Problem complexity and method efficiency in optimization (1983)
14. Parikh, N., Boyd, S., et al.: Proximal algorithms. Found. Trends Optim. **1**(3), 127–239 (2014)
15. Pascanu, R., Bengio, Y.: Revisiting natural gradient for deep networks. arXiv preprint arXiv:1301.3584 (2013)
16. Raskutti, G., Mukherjee, S.: The information geometry of mirror descent. IEEE Trans. Inf. Theory **61**(3), 1451–1457 (2015)
17. Saad, Y.: Krylov subspace methods for solving large unsymmetric linear systems. Math. Comput. **37**(155), 105–126 (1981)
18. Schraudolph, N.N.: Fast curvature matrix-vector products for second-order gradient descent. Neural Comput. **14**(7), 1723–1738 (2002)