Journal of Scientific Computing (2021) 86:26
https://doi.org/10.1007/s10915-020-01367-x

TECHNICAL NOTE

Natural Gradient for Combined Loss Using Wavelets

Lexing Ying[1]

Received: 29 June 2020 / Revised: 3 November 2020 / Accepted: 11 November 2020 /
Published online: 7 January 2021

Abstract
Natural gradients have been widely used in the optimization of loss functionals over probability space, with important examples such as Fisher–Rao gradient descent for Kullback–Leibler divergence, Wasserstein gradient descent for transport-related functionals, and Mahalanobis gradient descent for quadratic loss functionals. This note considers the situation in which the loss is a convex linear combination of these examples. We propose a new natural gradient algorithm by utilizing compactly supported wavelets to diagonalize approximately the Hessian of the combined loss. Numerical results are included to demonstrate the efficiency of the proposed algorithm.

Keywords Natural gradient · Fisher–Rao metric · Wasserstein metric · Mahalanobis metric · Compactly supported wavelet · Diagonal approximation

Mathematics Subject Classification 65Z05 · 82B28 · 82B80

1 Introduction

Many problems in partial differential equations and machine learning can be formulated as optimization problems over probability densities. For a domain Ω, let $E(p)$ be a loss or energy functional defined for the probability densities p over Ω. The goal is to find p^* that minimizes $E(p)$. A common approach, especially for $E(p)$ with a unique minimum, is to follow the gradient descent (GD) dynamics. However, depending the metric used in the gradient calculation, different gradient descent algorithms exhibit drastically different convergence behavior. The term *natural gradient* refers to the practice of choosing an appropriate metric depending on the loss functional $E(p)$ as well as the probability space. Below are several well-known examples of natural gradient.

The work of L.Y. is partially supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Scientific Discovery through Advanced Computing (SciDAC) program and also by the National Science Foundation under Award DMS-1818449.

✉ Lexing Ying
lexing@stanford.edu

[1] Department of Mathematics, Stanford University, Stanford, CA 94305, USA

- Wasserstein GD that scales the Euclidean gradient $\frac{\delta E}{\delta p}(p)$ with the metric $-\nabla \cdot (p\nabla)$. Wasserstein GD is typically effective for a loss $E(p)$ that behaves like the square of the 2nd Wasserstein distance.
- Fisher–Rao GD that scales the Euclidean gradient $\frac{\delta E}{\delta p}(p)$ with the diagonal tensor $\mathrm{diag}(p)$. Fisher–Rao GD is quite effective for a loss $E(p)$ such as the Kullback–Leibler divergence $\int p(x) \ln \frac{p(x)}{\mu(x)} \mathrm{d}x$.
- Mahalanobis GD that scales the Euclidean gradient $\frac{\delta E}{\delta p}(p)$ with a positive definite metric B. Mahalanobis GD is efficient for a quadratic loss of the form $\frac{1}{2}(p-\mu, A(p-\mu))$ with $B \approx A^{-1}$. In this note, we consider the case that A is a positive semidefinite pseudo-differential operator, for example $A = -\Delta$.

A general principle from these examples is that, for a natural gradient to be effective, the metric used at the density p should be an approximate inverse of the Hessian of the loss $E(p)$ at p. In each of these three examples, an approximate inverse of the Hessian can be derived quite explicitly.

1.1 Problem Statement

In several problems from kinetic theory and statistical machine learning, one is faced with a loss or energy functional $E(p)$ that is a linear combination of these three forms mentioned above, i.e.,

$$E(p) = \alpha_1 E_1(p) + \alpha_2 E_2(p) + \alpha_3 E_3(p),$$

where $\alpha_1, \alpha_2, \alpha_3 \geq 0$ and E_1, E_2, and E_3 are of the Wasserstein, Fisher–Rao, and Mahalanobis types, respectively, i.e.,

$$\frac{\delta^2 E_1}{\delta p^2}(p) \approx (-\nabla \cdot (p\nabla))^+, \quad \frac{\delta^2 E_2}{\delta p^2}(p) \approx \mathrm{diag}\left(\frac{1}{p}\right), \quad \frac{\delta^2 E_3}{\delta p^2}(p) \approx A$$

where $(\cdot)^+$ stands for pseudo-inverse. As a result, the Hessian of $E(p)$ has the following approximation

$$\frac{\delta^2 E}{\delta p^2}(p) = \alpha_1 \frac{\delta^2 E_1}{\delta p^2}(p) + \alpha_2 \frac{\delta^2 E_2}{\delta p^2}(p) + \alpha_3 \frac{\delta^2 E_3}{\delta p^2}(p).$$

None of three natural gradients listed above is effective for this combined loss functional, since the inverse of $\frac{\delta^2 E}{\delta p^2}(p)$ looks quite different from $-\nabla \cdot (p\nabla)$, $\mathrm{diag}(p)$, or A^{-1}.

An immediate question is design an efficient natural gradient (or even an approximate one) for the combined loss $E(p)$. Due to the efficiency considerations, we prefer this natural gradient to have the following features.

- It utilizes the Hessian information of $E_1(p)$, $E_2(p)$, and $E_3(p)$ in the design of the natural gradient.
- It avoids forming and/or inverting the Hessian $\frac{\delta^2 E}{\delta p^2}(p)$ in order to avoid super-linear costs.
- The computational cost of computing the natural gradient from $\frac{\delta E}{\delta p}(p)$ should be of order $O(n \log^c n)$, where n is the number of degrees of freedom used for discretizing p.

The main idea of our approach is to adopt a basis that diagonalizes each of the three terms $\frac{\delta^2 E_1}{\delta p^2}(p)$, $\frac{\delta^2 E_2}{\delta p^2}(p)$, and $\frac{\delta^2 E_3}{\delta p^2}(p)$ approximately at the same time. Among various choices, compactly supported wavelets emerge as a natural candidate because they approximately

diagonalize (1) differential operators, (2) diagonal scaling by functions with sufficient regularity, and also (3) pseudo-differential operators.

1.2 Related Work

Fisher–Rao metric is essential to many branches of probability and statistics, as it is invariant under diffeomorphisms. The study of Fisher–Rao and related metrics has evolved to become the field of information geometry and we refer to [1,3] for detail discussions. Explicit time-discretization of the Fisher–Rao GD gives rise the mirror descent algorithms [4,5,15], which plays an essential role in online learning and optimization.

Originated from the theory of optimal transport, Wasserstein metric is defined formally as the Hessian of the square of the 2nd Wasserstein distance [18,20,22,23]. Starting from [10,16], it has been shown that many kinetic-type PDEs can be viewed as a Wasserstein GD of free energies defined on probability spaces [6]. In recent years, a parametric version of the Wasserstein metric has been applied to various applied problems from statistical machine learning [7,11–13].

The quadratic term associated with the Mahalanobis metric appears quite often in partial differential equation models, for example as the Dirichlet energy or as the interacting free energy term in the Keller–Segel models [17].

A recent paper [24] considers the case where the loss function is the sum of the Kullback–Leibler divergence and a quadratic interacting term. By adopting a diagonal approximation of interacting term, it proposes new natural gradient dynamics and develops new mirror descent algorithms.

1.3 Contents

The rest of this note is organized as follows. Section 2 proposes a new metric for the combined loss functional and derives the natural gradient algorithm. In Sect. 3, numerical results in 1D and 2D show that the proposed natural gradient outperforms the existing ones for combined loss functionals. Finally, Sect. 4 ends with some discussions on future work.

1.4 Data Availability Statement

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

2 Algorithm

2.1 Metric Design

Consider the 1D problem with $\Omega = [0, 1]$ with the periodic boundary condition for simplicity. As mentioned above for the loss functional $E(p) = \alpha_1 E_1(p) + \alpha_2 E_2(p) + \alpha_3 E_3(p)$, the Hessian can be approximated as follows.

$$\frac{\delta^2 E}{\delta p^2}(p) = \alpha_1 \frac{\delta^2 E_1}{\delta p^2}(p) + \alpha_2 \frac{\delta^2 E_2}{\delta p^2}(p) + \alpha_3 \frac{\delta^2 E_3}{\delta p^2}(p) \approx \alpha_1 (-\nabla \cdot (p\nabla))^+ + \alpha_2 \mathrm{diag}\left(\frac{1}{p}\right) + \alpha_3 A. \tag{1}$$

For simplicity, assume that the domain Ω is discretized with a uniform grid with n points $S = \left\{\frac{0}{n}, \frac{1}{n}, \ldots, \frac{n-1}{n}\right\}$. A density $p(x)$ for $x \in \Omega = [0, 1]$ can be represented as a vector $p \in \mathrm{R}^n$ with entries denoted by p_s for $s \in S$. We denote by D the discrete differential operator. After the discretization, the Hessian approximation (1) takes the following discrete form

$$\frac{\delta^2 E}{\delta p^2}(p) \approx \alpha_1 \left(D^{\mathsf{T}} \mathrm{diag}(p) D\right)^{+} + \alpha_2 \mathrm{diag}\left(\frac{1}{p}\right) + \alpha_3 A. \tag{2}$$

As mentioned earlier, the key idea is to diagonalize each of the three terms in (2) in a compactly supported orthogonal wavelet basis such as the Daubechies wavelets [8,14]. Let us denote by $W \in \mathrm{R}^{n \times n}$ the matrix such that its j-th column is the j-th vector of the wavelet basis. Therefore, W is the matrix for wavelet reconstruction and its transpose W^{T} is the matrix for wavelet decomposition. Notice that for compactly supported wavelets, W and W^{T} are sparse matrices with only $O(n \log n)$ non-zero entries. Applying W or W^{T} to an arbitrary vector of length n takes only $O(n)$ operations by taking advantages of the filter bank structure of the wavelet basis [14].

Applying the matrices W^{T} to the left and W to the right of (2) leads to

$$W^{\mathsf{T}} \frac{\delta^2 E}{\delta p^2}(p) W \approx \alpha_1 W^{\mathsf{T}} \left(D^{\mathsf{T}} \mathrm{diag}(p) D\right)^{+} W + \alpha_2 W^{\mathsf{T}} \mathrm{diag}\left(\frac{1}{p}\right) W + \alpha_3 W^{\mathsf{T}} A W.$$

The three terms on the right hand side are treated as follows.

- For the first term, consider first its pseudo-inverse $W^{\mathsf{T}} D^{\mathsf{T}} \mathrm{diag}(p) D W$. The diagonal entries of $W^{\mathsf{T}} D^{\mathsf{T}} \mathrm{diag}(p) D W$ at the (i, i) slot is given by

$$\sum_{s \in S} (DW)_{si} p_s (DW)_{si} = \sum_{s \in S} (DW)_{si}^2 p_s.$$

  By defining the matrix H_1 with entries given by $(H_1)_{is} = (DW)_{si}^2$, the whole diagonal of $W^{\mathsf{T}} D^{\mathsf{T}} \mathrm{diag}(p) D W$ can be conveniently written as $H_1 p$, which clearly depends linearly on p. Taking its pseudo-inverse implies that $W^{\mathsf{T}} \left(D^{\mathsf{T}} \mathrm{diag}(p) D\right)^{+} W$ can be diagonally approximated with $\mathrm{diag}\left(\frac{1}{H_1 p}\right)$.

- For the second term, consider first its pseudo-inverse $W^{\mathsf{T}} \mathrm{diag}(p) W$. The diagonal entry $W^{\mathsf{T}} \mathrm{diag}(p) W$ at the (i, i) slot is given by

$$\sum_{s \in S} W_{si} p_s W_{si} = \sum_{s \in S} W_{si}^2 p_s.$$

  By defining the matrix $H_2 p$ with entries given by $(H_1)_{is} = W_{si}^2$, the whole diagonal of $W^{\mathsf{T}} \mathrm{diag}(p) W$ can be written as $H_2 p$, which is again linear in p. Taking its pseudo-inverse shows that $W^{\mathsf{T}} \mathrm{diag}\left(\frac{1}{p}\right) W$ can be diagonally approximated with $\mathrm{diag}\left(\frac{1}{H_2 p}\right)$.

- As opposed to the first two terms, the third term $W^{\mathsf{T}} A W$ is independent of the density p. Its diagonal can be precomputed and will be denoted by $h_3 \in \mathrm{R}^n$.

  Putting the three terms together, we conclude that

$$W^{\mathsf{T}} \frac{\delta^2 E}{\delta p^2}(p) W \approx \mathrm{diag}\left(\frac{\alpha_1}{H_1 p} + \frac{\alpha_2}{H_2 p} + \alpha_3 h_3\right),$$

or equivalently

$$\frac{\delta^2 E}{\delta p^2}(p) \approx W \mathrm{diag}\left(\frac{\alpha_1}{H_1 p} + \frac{\alpha_2}{H_2 p} + \alpha_3 h_3\right) W^{\mathsf{T}}.$$

By inverting this approximation, we reach at the metric for the natural gradient

$$\left(\frac{\delta^2 E}{\delta p^2}(p)\right)^{-1} \approx W \mathrm{diag}\left(\frac{1}{\frac{\alpha_1}{H_1 p} + \frac{\alpha_2}{H_2 p} + \alpha_3 h_3}\right) W^{\mathsf{T}}. \tag{3}$$

With the metric ready, the ODE for the new natural gradient reads

$$\dot{p} = -W \mathrm{diag}\left(\frac{1}{\frac{\alpha_1}{H_1 p} + \frac{\alpha_2}{H_2 p} + \alpha_3 h_3}\right) W^{\mathsf{T}} \frac{\delta E}{\delta p}(p). \tag{4}$$

If we denote the wavelet coefficient vector by $c \in \mathrm{R}^n$, i.e., $c = W^{\mathsf{T}} p$ and $p = Wc$, (4) can be written as

$$\dot{c} = -\mathrm{diag}\left(\frac{1}{\frac{\alpha_1}{H_1 W c} + \frac{\alpha_2}{H_2 W c} + \alpha_3 h_3}\right) \frac{\delta E}{\delta c}(c). \tag{5}$$

In what follows, we simply refer to them as the *combined* gradient descent.

The following two claims show that the objects in (4) and (5) can be computed efficiently.

Claim 1 *The computational cost of forming and storing the matrices H_1 and H_2 is $O(n \log n)$.*

Proof Let us recall the definition of the matrices H_1 and H_2

$$(H_1)_{is} = (DW)_{si}^2, \quad (H_2)_{is} = W_{si}^2.$$

Since the wavelets are compactly supported with a constant size support at the finest scale, applying the differential operator and taking the element-wise square for a wavelet at scale ℓ takes $O(n/2^\ell)$ steps. Summing over the wavelets from all scales gives the following bound for the total cost:

$$\sum_{\ell=1}^{\log_2 n} 2^\ell \cdot \frac{n}{2^\ell} = O(n \log n).$$

Claim 2 *For a density $p \in \mathrm{R}^n$ with $p_i > 0$, the computational cost of applying the metric $W \mathrm{diag} \frac{1}{\left(\frac{\alpha_1}{H_1 p} + \frac{\alpha_2}{H_2 p} + \alpha_3 h_3\right)} W^T$ takes $O(n \log n)$ steps.*

Proof As a consequence from the previous claim, forming $H_1 p$ and $H_2 p$ each takes $O(n \log n)$ steps. Applying the wavelet decomposition operator W^{T} or the reconstruction operator W takes $O(n)$ steps by taking advantages of the filter bank construction. Summing them together gives the $O(n \log n)$ total cost.

2.2 Time Discretization

Let us now describe the time discretization of the natural gradient dynamics (4), i.e., how to actually use (4) to find the minimizer. We adopt a backtracking line search algorithm with Armijo condition [2]. At time step k with the current approximation p^k, we introduce

$$s^k = W \mathrm{diag}\left(\frac{1}{\frac{\alpha_1}{H_1 p^k} + \frac{\alpha_2}{H_2 p^k} + \alpha_3 h_3}\right) W^{\mathsf{T}} \frac{\delta E}{\delta p}(p^k).$$

Starting from $\eta = 1$, one repetitively halves η until

$$E(p^k - \eta s^k) - E(p_k) \leq -\frac{1}{2}\eta s^k \cdot \frac{\delta E}{\delta p}(p^k).$$

Once it is reached, one sets

$$p^{k+1} = p^k - \eta s^k$$

and move on to the next iteration until convergence.

3 Numerical Results

This section presents several numerical examples to illustrate the efficiency of the combined gradient descent (4) for the combined loss functionals.

3.1 1D

Consider first the 1D domain $\Omega = [0, 1]$ with the periodic boundary condition. Let μ be a reference measure. Among the three terms of the combined loss functional $E(p) = \alpha_1 E_1(p) + \alpha_2 E_2(p) + \alpha_3 E_3(p)$, the first term $E_1(p)$ is a functional close to the square of the 2nd Wasserstein distance $W_2(p, \mu)$ between p and μ. Because the exact computation of $W_2^2(p, \mu)$ and its derivative with respect to p is quite non-trivial, we replace $E_1(p)$ with the square of the weighted semi H^{-1}-norm

$$E_1(p) = \frac{1}{2}\|p - \mu\|^2_{\dot{H}^{-1}(\mu)},$$

which is known to be equivalent to the square of the W_2 norm [19]. The $\dot{H}^1(\mu)$ for a signed measure ϵ is defined as

$$\|\epsilon\|_{\dot{H}^{-1}(\mu)} = \min_{\theta:\nabla\cdot(\mu\theta)=\epsilon} \int |\theta|^2 \mathrm{d}\mu,$$

or equivalently

$$\|\epsilon\|_{\dot{H}^{-1}(\mu)} = \sup\left\{(f, \epsilon) : \|f\|_{\dot{H}^1(\mu)} \leq 1\right\}, \quad \|f\|_{\dot{H}^1(\mu)} = \int |\nabla f|^2 \mathrm{d}\mu.$$

After discretization, $E_1(p)$ takes the following simple form

$$E_1(p) = \frac{1}{2}(p - \mu)^{\mathsf{T}}(D^{\mathsf{T}}\mu D)^{+}(p - \mu).$$

The second term $E_2(p)$ is the Kullback–Leibler divergence

$$E_2(p) = \sum_s p_s \log \frac{p_s}{\mu_s}.$$

Finally, the last term $E_3(p)$ is the Dirichlet energy given by

$$E_3(p) = \frac{1}{2}(p - \mu)^{\mathsf{T}}(-\Delta)(p - \mu) = \frac{1}{2}(p - \mu)^{\mathsf{T}} D^{\mathsf{T}} D(p - \mu),$$

so $A = (-\Delta)$. The minimizer of $E(p)$ is equal to μ.

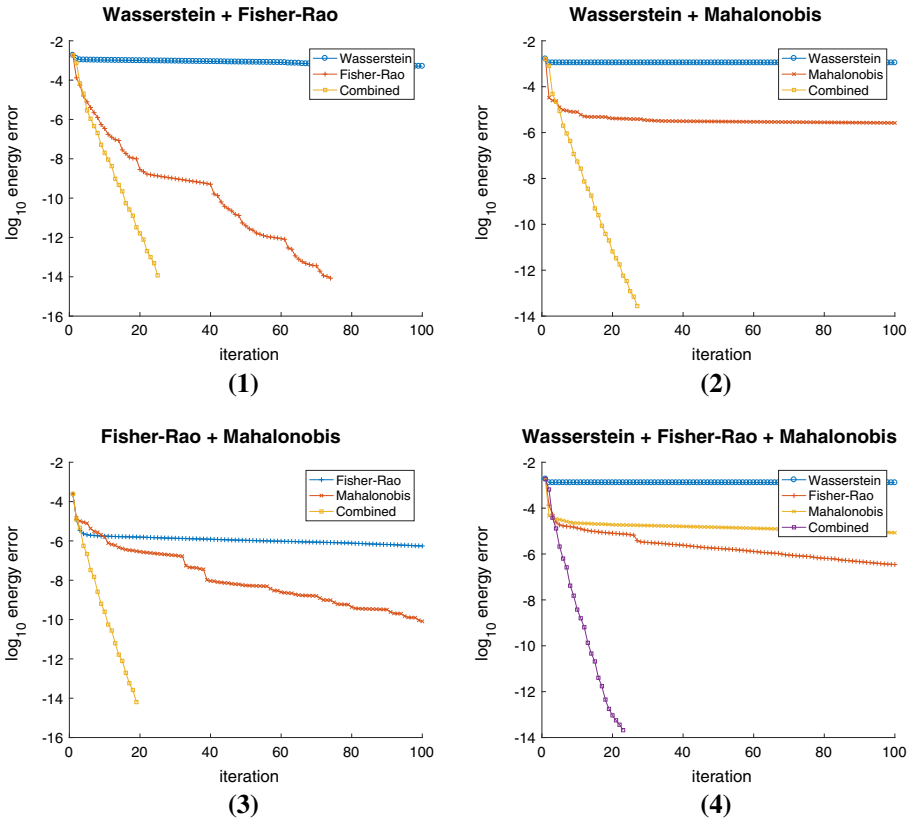

Fig. 1 (1) Wasserstein plus Fisher–Rao terms tested with Wasserstein GD, Fisher–Rao GD, and the combined natural gradient. (2) Wasserstein plus Mahalanobis terms tested with Wasserstein GD, Mahalanobis GD, and the combined natural gradient. (3) Fisher–Rao plus Mahalanobis terms tested with Fisher–Rao GD, Mahalanobis GD, and the combined natural gradient. (4) Wasserstein plus Fisher–Rao plus Mahalanobis terms tested with Wasserstein GD, Fisher–Rao GD, Mahalanobis GD, and the combined natural gradient

The domain is discretized with $n = 512$ grid points. The reference measure $\mu(s) \sim \exp(-V(s))$ with $V(s) = \sin(4\pi s)$ for $s \in S$. The constant factors in front of the three terms are chosen to be 1, 10^{-3}, and 10^{-4}, respectively, in order to balance the contribution from three terms so that none of them dominates. We test with four different linear combinations, with results summarized in Fig. 1.

(1) $(\alpha_1, \alpha_2, \alpha_3) = (1, 10^{-3}, 0)$, i.e., turning off the Mahalanobis term. The combined natural gradient converges much more rapidly compared to the Wasserstein GD and the Fisher–Rao GD.
(2) $(\alpha_1, \alpha_2, \alpha_3) = (1, 0, 10^{-4})$, i.e., turning off the Fisher–Rao term. The combined natural gradient converges much more rapidly compared to the Wasserstein GD and the Mahalanobis GD.
(3) $(\alpha_1, \alpha_2, \alpha_3) = (0, 10^{-3}, 10^{-4})$, i.e., turning off the Wasserstein term. The combined natural gradient converges much more rapidly compared to the Fisher–Rao GD and the Mahalanobis GD.

(4) $(\alpha_1, \alpha_2, \alpha_3) = (1, 10^{-3}, 10^{-4})$. The combined natural gradient converges much more rapidly compared to the Wasserstein GD, the Fisher–Rao GD, and the Mahalanobis GD.

3.2 2D

Consider now the 2D domain $\Omega = [0, 1]^2$ with periodic boundary condition. Among the three terms of the combined loss functional $E(p) = \alpha_1 E_1(p) + \alpha_2 E_2(p) + \alpha_3 E_3(p)$, $E_1(p)$ is again chosen to be the weighted semi H^{-1}-norm

$$E_1(p) = \frac{1}{2}\|p - \mu\|_{\dot{H}^{-1}(\mu)}.$$

After discretization, it takes the following form

$$E_1(p) = \frac{1}{2}(p - \mu)^{\mathsf{T}}(D_1^{\mathsf{T}}\mu D_1 + D_2^{\mathsf{T}}\mu D_2)^{+}(p - \mu)$$

where D_1 and D_2 are the derivative operators in the first and the second directions. $E_2(p)$ is again the Kullback–Leibler divergence

$$E_2(p) = \sum_s p_s \log \frac{p_s}{\mu_s}.$$

Finally, $E_3(p)$ is given by

$$E_3(p) = \frac{1}{2}(p - \mu)^{\mathsf{T}}(-\Delta)(p - \mu) = \frac{1}{2}(p - \mu)^{\mathsf{T}}(D_1^{\mathsf{T}}D_1 + D_2^{\mathsf{T}}D_2)(p - \mu)$$

so $A = (-\Delta)$.

The domain is discretized with $n = 64$ grid point in each direction. $\mu(s_1, s_2) \sim \exp(-V(s_1, s_2))$ with $V(s_1, s_2) = \sin(4\pi s_1)\sin(4\pi s_2)$. The constant factors of the three terms are set to be 1, 3×10^{-4}, and 10^{-4} in order to balance the contribution from them so that no one dominates. We test with four different linear combinations, with the results summarized in Fig. 2.

(1) $(\alpha_1, \alpha_2, \alpha_3) = (1, 3 \times 10^{-4}, 0)$, i.e., turning off the Mahalanobis term. The combined natural gradient converges much more rapidly compared to the Wasserstein GD and the Fisher–Rao GD.
(2) $(\alpha_1, \alpha_2, \alpha_3) = (1, 0, 10^{-4})$, i.e., turning off the Fisher–Rao term. The combined natural gradient converges much more rapidly compared to the Wasserstein GD and the Mahalanobis GD.
(3) $(\alpha_1, \alpha_2, \alpha_3) = (0, 3 \times 10^{-4}, 10^{-4})$, i.e., turning off the Wasserstein term. The combined natural gradient converges much more rapidly compared to the Fisher–Rao GD and the Mahalanobis GD.
(4) $(\alpha_1, \alpha_2, \alpha_3) = (1, 3 \times 10^{-4}, 10^{-4})$. The combined natural gradient converges much more rapidly compared to the Wasserstein GD, the Fisher–Rao GD, and the Mahalanobis GD.

4 Discussions

This note proposes a new natural gradient for minimizing combined loss functionals by using diagonal approximation in the wavelet basis. There are a few open questions. First, so far

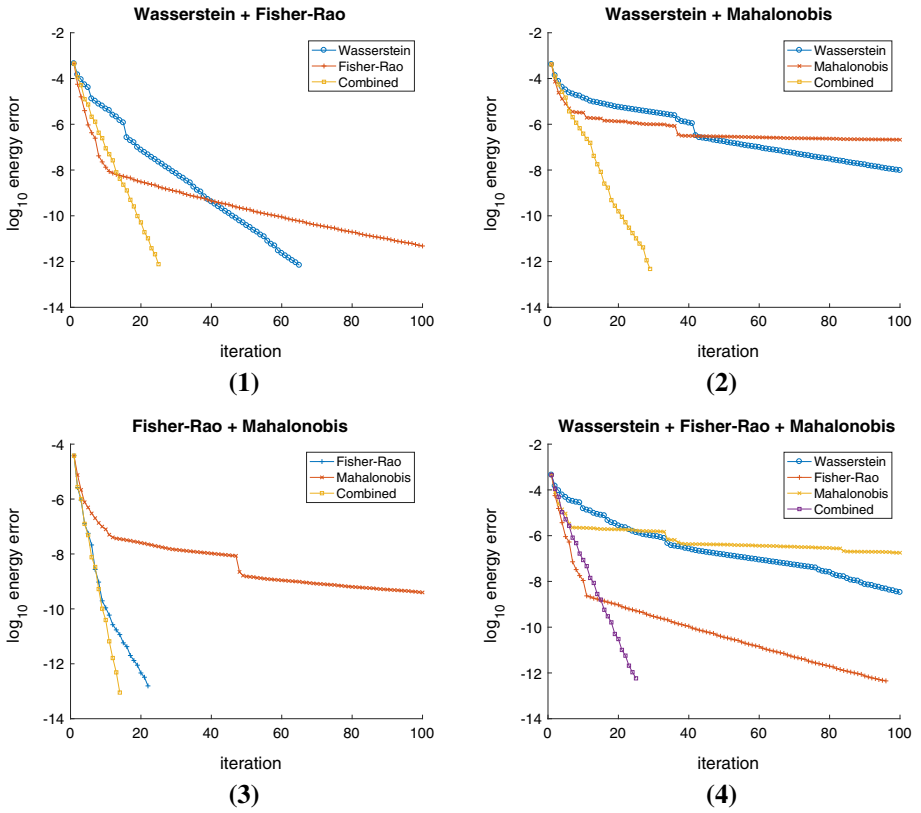

Fig. 2 (1) Wasserstein plus Fisher–Rao terms tested with Wasserstein GD, Fisher–Rao GD, and the combined natural gradient. (2) Wasserstein plus Mahalanobis terms tested with Wasserstein GD, Mahalanobis GD, and the combined natural gradient. (3) Fisher–Rao plus Mahalanobis terms tested with Fisher–Rao GD, Mahalanobis GD, and the combined natural gradient. (4) Wasserstein plus Fisher–Rao plus Mahalanobis terms tested with Wasserstein GD, Fisher–Rao GD, Mahalanobis GD, and the combined natural gradient

we have considered regular domains in one and two dimensions with periodic boundary condition. One direction is to extend this to more general domains using more sophisticated wavelet bases.

Second, we have assumed that the probability density p is non-vanishing everywhere in deriving the interpolating natural gradient metric. It is an important question whether one can remove this condition in order to work with more general probability densities.

Third, the dynamics in the wavelet coefficients (5) enjoys a diagonal metric. It is tempting to ask whether it is possible to design a mirror descent algorithm. Due to the coupling between different wavelet coefficients in the metric computation $H_1 Wc$ and $H_2 Wc$, this seems quite difficult. An interesting observation is that the metric of the coarse scale wavelet coefficients is nearly independent of the values of the fine scale coefficients, while the metric of the fine scale ones depends heavily on the values of the coarse scale ones. This naturally brings the question of whether the combined metric (or even the Wasserstein metric) has an approximate multiscale structure. The wavelet analysis has played an important role understanding the earth mover distance metric W_1 [9,21]. It seems that it might also play a role in understanding the W_2 metric.

Finally, it is possible to apply the general idea of this work to parametric probability models, especially when the number of parameters are large and direct inversion of the Hessian operator should be avoided.

References

1. Amari, S.: Information Geometry and Its Applications, vol. 194. Springer, Berlin (2016)
2. Armijo, L.: Minimization of functions having Lipschitz continuous first partial derivatives. Pac. J. Math. **16**(1), 1–3 (1966)
3. Ay, N., Jost, J., Vân Lê, H., Schwachhöfer, L.: Information Geometry, vol. 64. Springer, Berlin (2017)
4. Beck, A., Teboulle, M.: Mirror descent and nonlinear projected subgradient methods for convex optimization. Oper. Res. Lett. **31**(3), 167–175 (2003)
5. Bubeck, S., et al.: Foundations and trends®. Mach. Learn. **8**(3–4), 231–357 (2015)
6. Carrillo, J.A., McCann, R.J., Villani, C., et al.: Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates. Revista Matematica Iberoamericana **19**(3), 971–1018 (2003)
7. Chen, Y., Li, W.: Natural gradient in wasserstein statistical manifold (2018). arXiv preprint arXiv:1805.08380
8. Daubechies, I.: Ten Lectures on Wavelets, vol. 61. Siam, Philadelphia (1992)
9. Indyk, P., Thaper, N.: Fast image retrieval via embeddings. In: 3rd International Workshop on Statistical and Computational Theories of Vision, p. 5 (2003)
10. Jordan, R., Kinderlehrer, D., Otto, F.: The variational formulation of the Fokker–Planck equation. SIAM J. Math. Anal. **29**(1), 1–17 (1998)
11. Li, W., Lin, A.T., Montúfar, G.: Affine natural proximal learning. In: International Conference on Geometric Science of Information, pp. 705– 714. Springer (2019)
12. Li, W., Montúfar, G.: Natural gradient via optimal transport. Inf. Geom. **1**(2), 181–214 (2018)
13. Li, W., Montúfar, G.: Ricci curvature for parametric statistics via optimal transport. Inf. Geom. **3**, 89–117 (2020)
14. Mallat, S.: A Wavelet Tour of Signal Processing. Elsevier, New York (1999)
15. Nemirovsky, A.S., Yudin, D.B.: Problem Complexity and Method Efficiency in Optimization. A Wiley-Interscience Publication. Wiley, New York. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics (1983)
16. Otto, F.: The geometry of dissipative evolution equations: the porous medium equation. Commun. Part. Diff. Eq. **26**(1–2), 101–174 (2001)
17. Perthame, B.: Transport Equations in Biology. Springer, Berlin (2006)
18. Peyré, G., Cuturi, M., et al.: Computational optimal transport. foundations and trends®. Mach. Learn. **11**(5–6), 355–607 (2019)
19. Peyre, R.: Comparison between w2 distance and - 1 norm, and localization of wasserstein distance. ESAIM Control Optim. Calc. Var. **24**(4), 1489–1501 (2018)
20. Santambrogio, F.: Optimal Transport for Applied Mathematicians, vol. 55, pp. 58–63. Birkäuser, New York (2015)
21. Shirdhonkar, S., Jacobs, D.W.: Approximate earth mover's distance in linear time. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
22. Villani, C.: Topics in Optimal Transportation. American Mathematical Soc., Providence (2003)
23. Villani, C.: Optimal Transport: Old and New, vol. 338. Springer, Berlin (2008)
24. Ying, L.: Mirror descent algorithms for minimizing interacting free energy. J. Sci. Comput. **84**(3), 51 (2020). https://doi.org/10.1007/s10915-020-01303-z