KERNELIZED WASSERSTEIN NATURAL GRADIENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Many machine learning problems can be expressed as the optimization of some cost functional over a parametric family of probability distributions. It is often beneficial to solve such optimization problems using natural gradient methods. These methods are invariant to the parametrization of the family, and thus can yield more effective optimization. Unfortunately, computing the natural gradient is challenging as it requires inverting a high dimensional matrix at each iteration. We propose a general framework to approximate the natural gradient for the Wasserstein metric, by leveraging a dual formulation of the metric restricted to a Reproducing Kernel Hilbert Space. Our approach leads to an estimator for gradient direction that can trade-off accuracy and computational cost, with theoretical guarantees. We verify its accuracy on simple examples, and show the advantage of using such an estimator in classification tasks on `Cifar10` and `Cifar100` empirically.

1 INTRODUCTION

The success of machine learning algorithms relies on the quality of an underlying optimization method. Many of the current state-of-the-art methods rely on variants of Stochastic Gradient Descent (SGD) such as AdaGrad (Duchi et al., 2011), RMSProp (Hinton et al., 2012), and Adam (Kingma and Ba, 2014). While generally effective, the performance of such methods remains sensitive to the curvature of the optimization objective. When the Hessian matrix of the objective at the optimum has a large condition number, the problem is said to have a pathological curvature (Martens, 2010; Sutskever et al., 2013). In this case, the first-order optimization methods tend to have poor performance. It can be not alleviated by using adaptive step sizes on individual parameters. However, the curvature depends on the parametrization of the model. One strategy is to find an alternative parametrization of the same model that has a better-behaved curvature and is thus easier to optimize with standard first-order optimization methods. Designing good network architectures (Simonyan and Zisserman, 2014; He et al., 2015) along with normalization techniques (LeCun et al., 2012; Ioffe and Szegedy, 2015; Salimans and Kingma, 2016) is often critical for the success of such optimization methods.

The natural gradient method (Amari, 1998) takes a related but different perspective. Rather than re-parametrizing the model, the natural gradient method tries to make the optimizer itself invariant to re-parameterizations by directly operating on the manifold of probability distributions. This requires endowing the parameter space with a suitable notion of proximity formalized by a metric. An important metric in this context is the Fisher information metric (Fisher and Russell, 1922; Rao, 1992), which induces the Fisher-Rao natural gradient (Amari, 1985). Another important metric in probability space is the Wasserstein metric (Villani, 2009; Otto, 2001), which induces the Wasserstein natural gradient (Li and Montufar, 2018a;b); see similar formulations in Gaussian families (Malagò et al., 2018; Modin, 2017). In spite of their numerous theoretical advantages, applying natural gradient methods is challenging in practice. Indeed, each parameter update requires inverting the the metric tensor. This becomes infeasible for current deep learning models, which typically have millions of parameters. This has motivated research into finding efficient algorithms to estimate the natural gradient (Martens and Grosse, 2015; Grosse and Martens, 2016; George et al., 2018; Heskes, 2000; Bernacchia et al., 2018). Such algorithms often address the case of the Fisher metric and either exploit a particular structure of the parametric family or rely on a low rank decomposition of the information matrix. Recently, Li et al. (2019) proposed to estimate the metric based on a dual formulation and used this estimate in a proximal method. While this avoids explicitly computing the natural gradient, the proximal method also introduces an additional optimization problem to be solved at each update of the model's parameters. The quality of the solver will thus depend on the accuracy of this additional optimization.

In this paper, we use the dual formulation of the metric to directly obtain a closed form expression of the natural gradient as a solution to a convex functional optimization problem. We focus on the Wasserstein metric as it has the advantage of being well defined even when the model doesn't admit a density. The expression remains valid for general metrics including the Fisher-Rao metric. We leverage recent work on Kernel methods (Sriperumbudur et al., 2017; Strathmann et al., 2015; Arbel and Gretton, 2017; Sutherland et al., 2017) to compute an estimate of the natural gradient by restricting the functional space appearing in the dual formulation to a Reproducing Kernel Hilbert Space. We demonstrate empirically the accuracy of our estimator on toy examples, and show how it can be effectively used to approximate the trajectory of the natural gradient descent algorithm. We also analyze the effect of the dimensionality of the model on the accuracy of the proposed estimator. Finally, we illustrate the benefits of our proposed estimator for solving classification problems when the model has an ill-conditioned parametrization.

The paper is organized as follows. In Section 2, after a brief description of natural gradients, we discuss Legendre duality of metrics, and provide details on the Wasserstein natural gradient. In Section 3, we present our kernel estimator of the natural gradient. In Section 4 we present experiments to evaluate the accuracy of the proposed estimator and demonstrate its effectiveness in supervised learning tasks.

2 Natural Gradient Descent

We first briefly recall the natural gradient descent method in Section 2.1, and its relation to metrics on probability distribution spaces. We next present Legendre dual formulations for metrics in Section 2.2 where we highlight the Fisher-Rao and Wasserstein metrics as important examples.

2.1 General Formulation

It is often possible to formulate learning problems as the minimization of some cost functional $\rho \mapsto \mathcal{F}(\rho)$ over probability distributions ρ from a parametric model $\mathcal{P}_\Theta$. The set $\mathcal{P}_\Theta$ contains probability distributions defined on an open sample space $\Omega \subset \mathbb{R}^d$ and parametrized by some vector $\theta \in \Theta$, where Θ is an open subset of $\mathbb{R}^q$. The learning problem can thus be formalized as finding an optimal value θ^* that locally minimizes a loss function $\mathcal{L}(\theta) := \mathcal{F}(\rho_\theta)$ defined over the parameter space Θ. One convenient way to solve this problem approximately is by gradient descent, which uses the *Euclidean gradient* of $\mathcal{L}$ w.r.t. the parameter vector θ to produce a sequence of updates θ_t according to the following rule:

$$\theta_{t+1} = \theta_t - \gamma_t \nabla \mathcal{L}(\theta_t).$$

Here the step-size γ_t is a positive real number. The *Euclidean gradient* can be viewed as the direction in parameter space that leads to the highest decrease of some *linear model* $\mathcal{M}_t$ of the cost function $\mathcal{L}$ per unit of change of the parameter. More precisely, the *Euclidean gradient* is obtained as the solution of the optimization problem:

$$\nabla \mathcal{L}(\theta_t) = -\operatorname*{argmin}_{u \in \mathbb{R}^q} \mathcal{M}_t(u) + \frac{1}{2}\|u\|^2. \tag{1}$$

The linear model $\mathcal{M}_t$ is an approximation of the cost function $\mathcal{L}$ in the neighborhood of θ_t and is simply obtained by a first order expansion: $\mathcal{M}_t(u) = \mathcal{L}(\theta_t) + \nabla \mathcal{L}(\theta_t)^\top u$. The quadratic term $\|u\|^2$ penalizes the change in the parameter and ensures that the solution remains in the neighborhood where the linear model is still a good approximation of the cost function.

This particular choice of quadratic term is what defines the *Euclidean gradient* descent algorithm, which can often be efficiently implemented for neural network models using *back-propagation*. The performance of this algorithm is highly dependent on the parametrization of the model $\mathcal{P}_\Theta$, however (Martens, 2010; Sutskever et al., 2013). To obtain an algorithm that is robust to parametrization, one can take advantage of the structure of the cost function $\mathcal{L}(\theta)$ which is obtained as the composition of the functional $\mathcal{F}$ and the model $\theta \mapsto \rho_\theta$. This offers the possibility to use a different quadratic term to penalize the change in the model ρ_θ regardless of how it is parameterized.

The Fisher information matrix $\theta \mapsto G_F(\theta)$ (Amari, 1985) is one possible way to construct the quadratic term, leading to the *Fisher-Rao natural gradient*. We recall that $G_F(\theta)$ is well defined when the probability distributions in $\mathcal{P}_\Theta$ all have positive densities, and certain additional differentiability and integrability assumptions on ρ_θ are satisfied. In fact, G_F has an interpretation in Riemannian geometry as the pull-back of a metric tensor g^F defined over the set of probability distributions with positive densities and known as the *Fisher-Rao metric* (see Definition 4 in Appendix B.1; see also Holbrook et al. 2017):

Definition 1 (Fisher information matrix)**.** *Assume $\theta \mapsto \rho_\theta(x)$ is differentiable for all x on Ω and that $\int \frac{\|\nabla \rho_\theta(x)\|^2}{\rho_\theta(x)} \mathrm{d}x < \infty$. Then the Fisher information matrix is defined as the pull-back of the Fisher-Rao metric g^F:*

$$G_F(\theta)_{ij} = g^F_{\rho_\theta}(\partial_i \rho_\theta, \partial_j \rho_\theta) := \int f_i(x) f_j(x) \rho_\theta(x) dx,$$

where the functions f_i on Ω are given by: $f_i = \frac{\partial_i \rho_\theta}{\rho_\theta}$.

The *Fisher-Rao natural gradient* $\nabla^{G_F} \mathcal{L}(\theta_t)$ is obtained by replacing the quadratic term $\|u\|^2$ in (1) by $\frac{1}{2} u^\top G_F(\theta_t) u$. It corresponds to the direction with the largest decrease in the linear model $\mathcal{M}_t$ per unit of change in ρ_θ as measured by the *Fisher-Rao metric* g^F. The resulting sequence of probability densities $(\rho_{\theta_t})_{t\geq 0}$ is invariant to parametrization in the limit when the step-size γ_t tends to 0.

This approach can be extended to any metric tensor g defined on a suitable space of probability distributions containing $\mathcal{P}_\Theta$. In this case, the induced metric on the space of parameters Θ has a matrix representation $G(\theta) \in \mathbb{R}^q \times \mathbb{R}^q$ called the *information matrix* and the corresponding *natural gradient* $\nabla^G_\theta \mathcal{L}(\theta)$ is then obtained by solving:

$$\nabla^G \mathcal{L}(\theta_t) = -\underset{u \in \mathbb{R}^q}{\operatorname{argmin}} \mathcal{M}_t(u) + \frac{1}{2} u^\top G(\theta_t) u. \tag{2}$$

From (2), it is possible to express the *natural gradient* by means of the *Euclidean gradient*: $\nabla^G \mathcal{L}(\theta_t) = G(\theta_t)^{-1} \nabla \mathcal{L}(\theta_t)$. The parameter updates are then obtained by the new update rule:

$$\theta_{t+1} = \theta_t - \gamma_t G(\theta_t)^{-1} \nabla \mathcal{L}(\theta_t). \tag{3}$$

Similarly to the *Fisher-Rao natural gradient*, the generalized *natural gradient* leads to a descent algorithm which is invariant (in the sense of the traversed hypotheses) to parametrization in the limit of small step-sizes γ_t. A detailed discussion of parametrization invariant gradient methods is given by Ollivier et al. (2011). While the particular case when the metric g is chosen to be the Fisher-Rao metric leads to the well-know *Fisher-Rao natural gradient*, other choices for the metric g are also possible. Recently, Li and Montufar (2018a); Chen and Li (2018) proposed to use the Wasserstein 2 metric (Otto and Villani, 2000; Lafferty and Wasserman, 2008) denoted g^W (Definition 5 of Appendix B.2) which leads to the *Wasserstein information matrix* $G_W(\theta)$:

Definition 2 (Wasserstein information matrix)**.** *The Wasserstein information matrix is defined as the pull-back of the Wasserstein 2 metric g^W:*

$$G_W(\theta)_{ij} = g^W_{\rho_\theta}(\partial_i \rho_\theta, \partial_j \rho_\theta) := \int \phi_i(x)^\top \phi_j(x) \rho_\theta(x) dx,$$

where ϕ_i are vector valued functions on Ω that are solutions to the partial differential equations with Neumann boundary condition:

$$\partial_i \rho_\theta = -div(\rho_\theta \phi_i), \qquad \forall 1 \leq i \leq q.$$

Moreover, ϕ_i are required to be in the closure of the set of gradients of smooth and compactly supported functions in $L_2(\rho_\theta)^d$. In particular, when ρ_θ has a density, $\phi_i = \nabla_x f_i$, for some real valued function f_i on Ω.

The partial derivatives $\partial_i \rho_\theta$ should be understood in distribution sense, as discussed in more detail in Section 2.2. This allows to define the *Wasserstein natural gradient* even when the model ρ_θ does not admit a density. From now on, we will focus on the above two cases of the metric $G(\theta)$, namely $G_F(\theta)$ and $G_W(\theta)$. When the dimension of the parameter space is high, directly using equation (3) becomes impractical as it requires storing and inverting the matrix $G(\theta)$. In Section 2.2 we will see how equation (2) can be exploited along with Legendre duality to get a expression for the natural gradient that can be efficiently approximated using kernel methods.

2.2 Legendre Duality for Metrics

In this section we provide an expression for the *natural gradient* defined in (2) as a solution of a saddle-point optimization problem. It exploits Legendre duality for metrics to express the quadratic term $u^\top G(\theta) u$ as a solution to a functional optimization problem over $C_c^\infty(\Omega)$, the set of smooth and compactly supported functions on Ω. The starting point is to extend the notion of gradient $\nabla \rho_\theta$ which appears in Definitions 1 and 2 to the distributional sense of Definition 3 below.

Definition 3. *Given a parametric family $\mathcal{P}_\Theta$ of probability distributions, the distributional gradient $\nabla\rho_\theta$ at point θ of ρ_θ is defined as the linear map $\nabla\rho_\theta : C_c^\infty(\Omega) \rightarrow \mathbb{R}^q$ whose components are given by:*

$$(\nabla\rho_\theta(f))_i = \lim_{\epsilon\rightarrow 0}\epsilon^{-1}\Big(\int f(x)\mathrm{d}\rho_{\theta+\epsilon e_i}(x) - \int f(x)\mathrm{d}\rho_\theta(x)\Big), \quad \leq 1 \leq i \leq q,$$

where $(e_i)_{1\leq i\leq q}$ is an orthonormal basis of $\mathbb{R}^q$.

When the model $\mathcal{P}_\Theta$ has a density that is differentiable w.r.t. θ, $\nabla\rho_\theta(f)$ is simply given by $\int f(x)\nabla_\theta\rho(x)\mathrm{d}x$ and one recovers the usual gradient $\nabla\rho_\theta(x)$. In this case, the *Fisher information matrix* admits a dual formulation provided in Proposition 1 with a proof in Appendix C.1.

Proposition 1. *Under the same assumptions as in Definition 1, the Fisher information matrix satisfies:*

$$\frac{1}{2}u^\top G_F(\theta)u = \sup_{\substack{f\in C_c^\infty(\Omega)\\ \int f(x)\mathrm{d}\rho_\theta(x)=0}} \nabla\rho_\theta(f)^\top u - \frac{1}{2}\int f(x)^2\rho_\theta(x)\mathrm{d}x. \tag{4}$$

Another important case is when $\mathcal{P}_\Theta$ is defined as an implicit model. In this case, any sample x from a distribution ρ_θ in $\mathcal{P}_\Theta$ is obtained as $x = h_\theta(z)$, where z is a sample from a fixed latent distribution ν defined over a latent space $\mathcal{Z}$ and $(\theta, z) \mapsto h_\theta(z)$ is a deterministic function with values in Ω. This can be written in a more compact way as the push-forward of ν by the function h_θ:

$$\mathcal{P}_\Theta := \{\rho_\theta := (h_\theta)_\# \nu \,|\, \theta \in \Omega\}. \tag{5}$$

A different expression for $\nabla\rho_\theta$ is obtained in the case of implicit models when $\theta \mapsto h_\theta(z)$ is differentiable for ν-almost all z and when ∇h_θ is square integrable under ν:

$$\nabla\rho_\theta(f) = \int \nabla h_\theta(z)^\top \nabla_x f(h_\theta(z))\mathrm{d}\nu(z). \tag{6}$$

Equation (6) is also known as the re-parametrization trick (Kingma et al., 2015) and allows to derive a dual formulation of the *Wasserstein information matrix* in the case of implicit models. Proposition 2 below provides such formulation under mild assumptions stated in Appendix A.2 along with a proof in Appendix C.1.

Proposition 2. *Assume $\mathcal{P}_\Theta$ is defined by (5) such that $\nabla\rho_\theta$ is given by (6). Under Assumptions (**B**) and (**C**), the Wasserstein information matrix satisfies:*

$$\frac{1}{2}u^\top G_W(\theta)u = \sup_{\substack{f\in C_c^\infty(\Omega)\\ \int f(x)\mathrm{d}\rho_\theta(x)=0}} \nabla\rho_\theta(f)^\top u - \frac{1}{2}\int \|\nabla_x f(h_\theta(z))\|^2\mathrm{d}\nu(z). \tag{7}$$

The similarity between the variational formulations provided in Propositions 1 and 2 is worth noting. However, while (7) is well defined, the expression in (4) can be infinite when $\nabla\rho_\theta$ is given by (6). Indeed, if the ρ_θ doesn't admit a density, it is always possible to find an admissible function $f \in C_c^\infty(\Omega)$ with bounded second moment under ρ_θ but for which $\nabla\rho_\theta(f)$ is arbitrarily large. This is avoided in (7) since the quadratic term penalizes the second moment of the gradient of functions instead.

3 Kernelized Wasserstein Natural Gradient

In this section we propose an estimator for the Wasserstein natural gradient using kernel methods and exploiting the formulation in (7). We restrict to the case of the Wasserstein natural gradient, denoted by $\nabla^W\mathcal{L}(\theta)$, as it is well defined for implicit models, but a similar approach can be used for the Fisher-Rao natural gradient in the case of models with densities. We first start by presenting the *kernelized Wasserstein natural gradient* (KWNG) in Section 3.1, then we introduce an efficient estimator for KWNG in Section 3.2. In Section 3.3 we provide statistical guarantees and discuss practical considerations in Section 3.4.

3.1 General Formulation and Minimax Theorem

Consider a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ which is a Hilbert space endowed with an inner product $\langle .,. \rangle_\mathcal{H}$ along with its norm $\|.\|_\mathcal{H}$. $\mathcal{H}$ has the additional property that there exists a symmetric positive semi-definite kernel $k:\Omega\times\Omega\mapsto\mathbb{R}$ such that $k(x,.)\in\mathcal{H}$ for all $x\in\Omega$ and satisfying the *Reproducing property* for all functions f in $\mathcal{H}$:

$$f(x)=\langle f,k(x,.)\rangle_{\mathcal{H}}, \qquad \forall x\in\Omega. \tag{8}$$

The above property is central in all kernel methods as it allows to obtain closed form expressions for some class of functional optimization problems. In order to take advantage of such property for estimating the natural gradient, we consider a new saddle problem obtained by restricting (7) to functions in the RKHS $\mathcal{H}$ and adding some regularization terms:

$$\min_{u\in\mathbb{R}^q} \sup_{\substack{f\in\mathcal{H} \\ \int f(x)\mathrm{d}\rho_\theta(x)=0}} \mathcal{U}_\theta(f)^\top u - \frac{1}{2}\int \|\nabla_x f(x)\|^2 \mathrm{d}\rho_\theta(x) + \frac{1}{2}(\epsilon u^\top D(\theta) u - \lambda\|f\|^2_{\mathcal{H}}) \tag{9}$$

where we introduced $\mathcal{U}_\theta(f):=\nabla\mathcal{L}(\theta)+\nabla\rho_\theta(f)$ for simplicity. The *kernelized Wasserstein natural gradient* is obtained by solving (9) and is denoted by $\widetilde{\nabla}^W\mathcal{L}(\theta)$. Here, λ and ϵ are non-negative real numbers while $D(\theta)$ is a diagonal matrix in $\mathbb{R}^q$ with positive diagonal elements whose choice will be discussed in Section 3.4. The additional regularization terms allow to use a version of the minimax theorem (Ekeland and Témam, 1999, Proposition 2.3, Chapter VI) to exchange the order of the supremum and minimum. This leads to a new expression for the kernelized natural gradient which is provided in Proposition 3.

Proposition 3. *The kernelized natural gradient is given by:*

$$\widetilde{\nabla}^W\mathcal{L}(\theta)=\frac{1}{\epsilon}D(\theta)^{-1}\mathcal{U}_\theta(f^*), \tag{10}$$

where f^* is the unique solution to the quadratic optimization problem:

$$\inf_{\substack{f\in\mathcal{H} \\ \int f(x)\mathrm{d}\rho_\theta(x)=0}} \mathcal{J}(f):=\int \|\nabla_x f(x)\|^2 \mathrm{d}\rho_\theta(x) + \frac{1}{\epsilon}\mathcal{U}_\theta(f)^\top D(\theta)^{-1}\mathcal{U}_\theta(f)+\lambda\|f\|^2_{\mathcal{H}}. \tag{11}$$

Proposition 3 allows to compute the kernelized natural gradient directly provided that the functional optimization (11) can be solved. This circumvents the direct computation and inversion of the metric as suggested by (9). In Section 3.2, we propose a method to efficiently compute an approximate solution to (11) using Nyström projections. We also show in Section 3.3 that restricting the space of functions to $\mathcal{H}$ can still lead to a good approximation of the Wasserstein natural gradient provided that $\mathcal{H}$ enjoys some denseness properties.

3.2 Nyström Methods for the Kerenalized Natural Gradient

We are interested now in finding an approximate solution to (11) which will allow to compute an estimator for the Wasserstein natural gradient using Proposition 3. Here we consider N samples $(Z_n)_{1\leq n\leq N}$ from the latent distribution ν which are used to produce N samples $(X_n)_{1\leq n\leq N}$ from ρ_θ using the map h_θ, i.e., $X_n=h_\theta(Z_n)$. We also assume we have access to an estimate of the *Euclidean gradient* $\nabla\mathcal{L}(\theta)$ which is denoted by $\widehat{\nabla\mathcal{L}(\theta)}$. This allows to compute an empirical version of the cost function in (11),

$$\hat{\mathcal{J}}(f):=\frac{1}{N}\sum_{n=1}^N\|\nabla_x f(X_n)\|^2+\frac{1}{\epsilon}\widehat{\mathcal{U}_\theta}(f)^\top D(\theta)^{-1}\widehat{\mathcal{U}_\theta}(f)+\lambda\|f\|^2_{\mathcal{H}}, \tag{12}$$

where $\widehat{\mathcal{U}_\theta}(f)$ is given by: $\widehat{\mathcal{U}_\theta}(f)=\widehat{\nabla\mathcal{L}(\theta)}+\frac{1}{N}\sum_{n=1}^N\nabla h_\theta(Z_n))^\top\nabla_x f(X_n)$. Equation (12) is similar to another functional arising in the context of *score* estimation for *infinite dimensional exponential families* (Sriperumbudur et al., 2017; Strathmann et al., 2015; Arbel and Gretton, 2017). It can be shown using the generalized Representer Theorem (Schölkopf et al., 2001) that the optimal function minimizing (12) is a linear combination of functions of the form $x\mapsto\partial_i k(X_n,x)$ with $1\leq n\leq N$ and $1\leq i\leq d$ and $\partial_i k(y,x)$ denotes the partial derivative of k w.r.t. y_i. This requires to solve a system of size $Nd\times Nd$ which can be prohibitive when both N and d are large. Nyström methods provide a way to improve such computational cost by further restricting the optimal solution to belong to a finite dimensional subspace $\mathcal{H}_M$ of $\mathcal{H}$ called the *Nyström subspace*. In the context of *score* estimation, Sutherland et al. (2017) proposed to use a subspace formed by linear combinations of the *basis functions* $x\mapsto\partial_i k(Y_m,x)$:

$$span\{x\mapsto\partial_i k(Y_m,x)\,|\,1\leq m\leq M;\quad 1\leq i\leq d\}, \tag{13}$$

where $(Y_m)_{1\leq m\leq M}$ are *basis points* drawn uniformly from $(X_n)_{1\leq n\leq N}$ with $M\leq N$. This further reduces the computational cost when $M \ll N$ but still has a cubic dependence in the dimension d since all partial derivatives of the kernel are considered to construct (13). Here, we propose to randomly sample one component of $(\partial_i k(Y_m,.))_{1\leq i\leq d}$ for each basis point Y_m. Hence, we consider M indices $(i_m)_{1\leq m\leq M}$ uniformly drawn form $\{1,...,d\}$ and define the *Nyström subspace* $\mathcal{H}_M$ to be:

$$\mathcal{H}_M := span\{x\mapsto \partial_{i_m} k(Y_m,x) | 1\leq m\leq M\}.$$

An estimator for the *kernelized Wasserstein natural gradient* (KWNG) is then given by:

$$\widehat{\nabla^W\mathcal{L}(\theta)} = \frac{1}{\epsilon} D(\theta)^{-1}\widehat{\mathcal{U}_\theta}(\hat{f}^*), \qquad \hat{f}^* := \underset{\substack{f\in\mathcal{H}_M\\ \int f(x)\mathrm{d}\rho_\theta(x)=0}}{\operatorname{argmin}} \hat{\mathcal{J}}(f). \tag{14}$$

By definition of the Nyström subspace $\mathcal{H}_M$, the optimal solution $\hat{f}^*$ is necessarily of the form: $\hat{f}^*(x) = \sum_{m=1}^M \alpha_m \partial_{i_m} k(Y_m,x)$, where the coefficients $(\alpha_m)_{1\leq m\leq M}$ are obtained by solving a finite dimensional quadratic optimization problem. This allows to provide a closed form expression for (15) in Proposition 4.

Proposition 4. *The estimator in (14) is given by:*

$$\widehat{\nabla^W\mathcal{L}(\theta)} = \frac{1}{\epsilon}\Big(D(\theta)^{-1} - D(\theta)^{-1}T^\top\Big(TD(\theta)^{-1}T^\top + \lambda\epsilon K + \frac{\epsilon}{N}CC^\top\Big)^\dagger TD(\theta)^{-1}\Big)\widehat{\nabla\mathcal{L}(\theta)}, \tag{15}$$

where C and K are matrices in $\mathbb{R}^{M\times Nd}$ and $\mathbb{R}^{M\times M}$ given by

$$C_{m,(n,i)} = \partial_{i_m}\partial_{i+d} k(U_m,X_n), \qquad K_{m,m'} = \partial_{i_m}\partial_{i_{m'}+d} k(Y_m,Y_{m'}), \tag{16}$$

while T is a matrix in $\mathbb{R}^{M\times q}$ obtained as the Jacobian of $\theta\mapsto\tau(\theta)\in\mathbb{R}^M$, i.e., $T:=\nabla\tau(\theta)$, with

$$(\tau(\theta))_m = \frac{1}{N}\sum_{n=1}^N \partial_{i_m} k(Y_m,h_\theta(Z_n)).$$

In (16), we used the notation $\partial_{i+d}k(y,x)$ for the partial derivative of k w.r.t. x_i. A proof of Proposition 4 is provided in Appendix C.2 and relies on the reproducing property (8) and its generalization for partial derivatives of functions. The estimator in Proposition 4 is in fact a low rank approximation of the natural gradient obtained from the dual representation of the metric (7). While low-rank approximations for the Fisher-Rao natural gradient were considered in the context of variational inference and for a Gaussian variational posterior (Mishkin et al., 2018), (15) can be applied as a plug-in estimator for any family $\mathcal{P}_\Theta$ obtained as an implicit model. We next provide theoretical guarantees for our method in Section 3.3, and discuss the computational cost of the proposed estimator, along with some practical considerations, in Section 3.4.

3.3 Theory

In this section we are interested in the behavior of the estimator in the limit of large N and M. We distinguish two cases: the *well-specified* case and the *miss-specified* case.

Well-specified case. Here, we assume that the vector valued functions $(\phi_i)_{1\leq i\leq q}$ involved in Definition 2 can be expressed as gradients of functions in $\mathcal{H}$. More precisely:

Assumption 1. *For all $1\leq i\leq q$, there exits functions $f_i\in\mathcal{H}$ such that $\phi_i = \nabla f_i$. Additionally, f_i are of the form $f_i = A^\alpha v_i$ for some fixed $\alpha\geq 0$, with $v_i\in\mathcal{H}$ and A being the differential covariance operator defined on $\mathcal{H}$ by:*

$$A: f\mapsto \int \sum_{i=1}^d \partial_i k(h_\theta(z),.)\partial_i f(h_\theta(z))\mathrm{d}\nu(z). \tag{17}$$

The parameter α characterizes the smoothness of f_i and therefore controls the statistical complexity of the estimation problem. Using a similar analysis as Sutherland et al. (2017) we obtain the following convergence rates for the estimator in Proposition 4:

Theorem 5. *Let δ be such that $0 \leq \delta \leq 1$ and $b := \min(1, \alpha + \frac{1}{2})$. Under Assumption 1 and Assumptions **(A)** to **(G)** listed in Appendix A.2, for N large enough, $M \sim (dN^{\frac{1}{2b+1}} \log(N))$, $\lambda \sim N^{-\frac{1}{2b+1}}$ and $\epsilon \lesssim N^{-\frac{b}{2b+1}}$, it holds with probability at least $1-\delta$ that:*

$$\|\widehat{\nabla^W \mathcal{L}(\theta)} - \nabla^W \mathcal{L}(\theta)\|^2 = \mathcal{O}\Big(N^{-\frac{2b}{2b+1}}\Big).$$

A proof of Theorem 5 is provided in Appendix C.3 and relies on the same techniques used by Rudi et al. (2015); Sutherland et al. (2017). The additional assumptions are easy to satisfy in practice by choosing a suitable model $\mathcal{P}_\Theta$ and kernel k. In the worst case where $\alpha = 0$, the proposed estimator needs at most $M \sim (d\sqrt{N}\log(N))$ to achieve a convergence rate of $N^{-\frac{1}{2}}$. The smoothest case requires only $M \sim (dN^{\frac{1}{3}}\log(N))$ to achieve a rate of $N^{-\frac{2}{3}}$. Thus, the proposed estimator enjoys the same statistical properties as the ones proposed by Sriperumbudur et al. (2017); Sutherland et al. (2017) while maintaining a computational advantage[1].

Miss-specified case. We extend the previous analysis to the case where Assumption 1 doesn't hold. Instead, we rely on the following weaker assumption:

Assumption 2. *There exists two constants $C > 0$ and $c \geq 0$ such that for all $\kappa > 0$ and all $1 \leq i \leq q$, there is a function f_i^κ satisfying:*

$$\|\phi_i - \nabla f_i^\kappa\|_{L_2(\rho_\theta)} \leq C\kappa, \qquad \|f_i^\kappa\|_{\mathcal{H}} \leq C\kappa^{-c}. \tag{18}$$

The left inequality in (18) represents the accuracy of the approximation of ϕ_i by gradients of functions in $\mathcal{H}$ while the right inequality represents the complexity of such approximation. Thus, the parameter c characterizes the difficulty of the problem: a higher value of c means that a more accurate approximation of ϕ_i comes at a higher cost in terms of its complexity. Theorem 6 provides convergences rates for the same estimator Proposition 4 under Assumption 2:

Theorem 6. *Let δ be such that $0 \leq \delta \leq 1$ and $b := \frac{1}{2+c}$. Under Assumption 2 and Assumptions **(A)** to **(G)** listed in Appendix A.2, for N large enough, $M \sim (dN^{\frac{1}{2b+1}} \log(N))$, $\lambda \sim N^{\frac{1}{2b+1}}$ and $\epsilon \lesssim N^{-\frac{b}{2b+1}}$, it holds with probability at least $1-\delta$ that:*

$$\|\widehat{\nabla^W \mathcal{L}(\theta)} - \nabla^W \mathcal{L}(\theta)\|^2 = \mathcal{O}\Big(N^{-\frac{2}{4+c}}\Big).$$

A proof of Theorem 6 is also provided in Appendix C.3. Here, in the case where $c = 0$, we recover the same convergence rate as in Theorem 5 for $\alpha = 0$. While $c = 0$ represents the best case under Assumption 2, $\alpha = 0$ corresponds to the worst case under Assumption 1. Hence, Theorem 6 is a consistent extension of Theorem 5 to the miss-specified case. For harder problems where $c > 0$ more basis points are required, with M required to be of order $dN\log(N)$ in the limit when $c \rightarrow \infty$ in which case the Nyström approximation loses its computational advantage.

3.4 Practical Considerations

Computational cost. (15) has a computational cost that is controlled by the number of basis points M with the main contributions due to the cost of computing T, C and solving an $M \times M$ linear system. This gives an overall cost of $O(dNM^2 + qM^2 + M^3)$. In practice, M can be chosen to be small ($M \leq 20$) while N corresponds to the number of samples in a mini-batch. Hence, in a typical deep learning model, most of the computational cost is due to computing T as the typical number of parameters q is of the order of millions. In fact, T can be computed using automatic differentiation and would require performing M backward passes on the model to compute the gradient for each component of τ. Overall, the proposed estimator can be efficiently implemented and used for typical deep learning problems as shown in Section 4.

Choice of damping term. So far, we only required $D(\theta)$ to be a diagonal matrix with positive coefficients. While a natural choice would be the identity matrix, this doesn't necessarily represent the best choice. As discussed by Martens and Sutskever (2012, Section 8.2), using the identity breaks the

[1] The estimator proposed by Sutherland et al. (2017) also requires M to grow linearly with the dimension d although such dependence doesn't appear explicitly in the statement of Sutherland et al. 2017, Theorem 2.

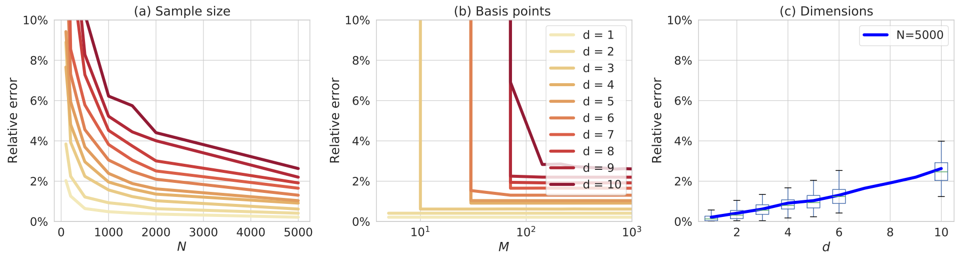

Figure 1: Evolution of the relative error of KWNG for the multivariate normal model averaged over 100 runs. For each run, a random value for the parameter θ and for the Euclidean gradient $\nabla\mathcal{L}(\theta)$ is sampled from a centered Gaussian with variance 0.1. In all cases, $\lambda = \epsilon = 10^{-10}$. Left (a): Relative error as the sample size N increases. The dimension d of the sample space varies from $d = 1$ (yellow) to $d = 10$ (dark red) and the number of basis points is set to $M = \lfloor d\sqrt{N} \rfloor$. Middle (b) Relative error as M increases with dimension d varying from $d = 1$ (yellow) to $d = 10$ (dark red) and N fixed to 5000. Right (c) Box-plot of the relative error as d increases with $N = 5000$ and $M = \lfloor d\sqrt{N} \rfloor$.

self-rescaling properties enjoyed by the natural gradient. Instead, we consider a scale-sensitive choice by setting $(D(\theta))_i = \|T_{.,i}\|^2$ where T is defined in Proposition 4. Moreover, while Theorems 5 and 6 suggest to choose $\epsilon \lesssim N^{-\frac{b}{2b+1}}$, this is only valid in the large sample-size limit $N \rightarrow \infty$. When the sample-size is limited, as it is often the case when N is the size of a mini-batch, larger values for ϵ might be required. That is to prevent the KWNG from over-estimating the step-size in the directions of low curvature. Indeed, these directions are rescaled by the inverse of the smallest eigenvalues of the information matrix which are harder to estimate accurately. To adjust ϵ dynamically during training, we use the Levenberg-Marquardt heuristic as Martens and Sutskever (2012) which seems to perform well in practice; see Section 4.

4 Experiments

This section presents an empirical evaluation of the Kernelized Wasserstein Natural Gradient (KWNG) as an estimator of the Wasserstein Natural Gradient (WNG). In all experiments, the kernel k is chosen to be a Gaussian kernel with an adaptive bandwidth of the form $\sigma = \sigma_0 \sigma_{N,M}$, where $\sigma_0 = 10^{3.3}$ and $\sigma_{N,M}$ is equal to the average square distance between samples $(X_n)_{1 \leq n \leq N}$ and the basis points $(Y_m)_{1 \leq m \leq M}$.

4.1 Convergence on Synthetic Models

To empirically assess the accuracy of KWNG, we consider two choices for the parametric model $\mathcal{P}_\Theta$: the multivariate normal model and the multivariate log-normal model. Both have the advantage that the Wasserstein information matrix can be computed in closed form (Chen and Li, 2018; Malagò et al., 2018). While this choice is essential to obtain closed form expressions for WNG, the proposed estimator is agnostic to such choice of family. We also assume we have access to the exact Euclidean Gradient (EG) which is used to compute both of WNG and KWNG.

Figure 1 shows the evolution of the the relative error w.r.t. the sample-size N, the number of basis points M and the dimension d in the case of the multivariate normal model. As expected from the consistency results provided in Section 3.3, the relative error decreases as the samples size N increases with the problem becoming harder as the dimension d increases. The behavior in the number of basis points M shows a clear threshold beyond which the estimator becomes consistent and where increasing M doesn't decrease the relative error anymore. This threshold increases with the dimension d as discussed in Section 3.3. In practice, using the rule $M = \lfloor d\sqrt{N} \rfloor$ seems to be a good heuristic as shown in Figure 1 (a). All these observations persist in the case of the log-normal model as shown in Figure 5 of Appendix D.1.

We also compare the optimization trajectory obtained using KWNG with the trajectories of both the exact WNG and EG in a simple setting. In this case, $\mathcal{P}_\Theta$ is the multivariate normal family and the loss function $\mathcal{L}(\theta)$ is the squared Wasserstein 2 distance between ρ_θ and a fixed target distribution ρ_{θ^*}. Figure 2 (a), shows the evolution of the loss function at every iteration. There is a clear advantage of using the WNG over EG as larger step-sizes are allowed leading to faster convergence. Moreover, KWNG maintains this

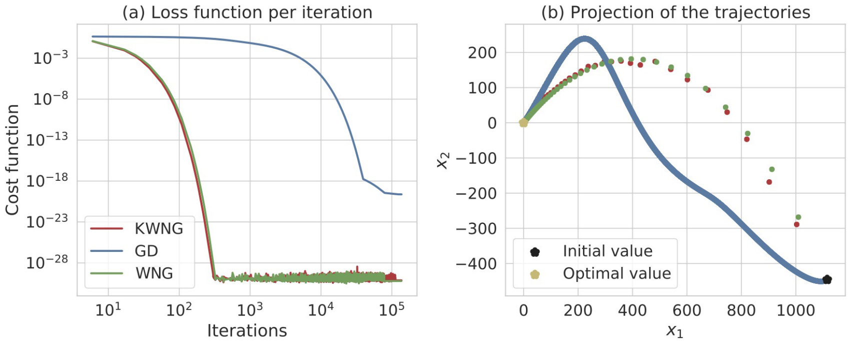

Figure 2: Left (a): Training error per iteration for KWNG, WNG, and EG. Right (b): projection of the sequence of updates obtained using KWNG, WNG and EG along the first two PCA directions of the WNG trajectory. The dimension of the sample space is fixed to $d = 10$. Exact valued for the gradient are used for EG and WNG. For KWNG, $N = 128$ samples and $M = 100$ basis points are used. The regularization parameters are set to: $\lambda = \epsilon = 10^{-10}$. An optimal step-size γ_t is used: $\gamma_t = 0.1$ for both KWNG and WNG while $\gamma_t = 0.0001$ for EG.

properties while being agnostic to the choice of the model. Figure 2 (b) shows the projected dynamics of the three methods along the two PCA directions of the WNG trajectory with highest variance. The dynamics of WNG seems to be well approximated by the one obtained using KWNG.

4.2 Approximate Invariance to Parametrization

We illustrate now the approximate invariance to parametrization of the KWNG and show its benefits for training training deep neural networks when the model is ill-conditioned. We consider a classification task on two datasets `Cifar10` and `Cifar100` with a Residual Network He et al. (2015). To use the KWNG estimator, we view the input RGB image as a latent variable z with probability distribution ν and the output logits of the network $x := h_\theta(z)$ as a sample from the model distribution $\rho_\theta \in \mathcal{P}_\Theta$ where θ denotes the weights of the network. The loss function $\mathcal{L}$ is given by:

$$\mathcal{L}(\theta) := \int y(z)^\top \log(SM(Uh_\theta(z)))\mathrm{d}\nu(z),$$

where SM is the Softmax function, $y(z)$ denotes the one-hot vector representing the class of the image z and U is a fixed invertible diagonal matrix which controls how well the model is conditioned. We consider two cases, the *Well-conditioned* case (WC) in which U is the identity and the *Ill-conditioned* case (IC) where U is chosen to have a condition number equal to 10^7. We compare the performance of the proposed method with several variants of SGD: plain SGD, SGD + Momentum, and SGD + Momentum + Weight decay. We also compare with KFAC optimizer (Martens and Grosse, 2015; Grosse and Martens, 2016) and eKFAC (George et al., 2018) which implements a fast approximation of the empirical Fisher Natural Gradient. Details of the experiments are provided in Appendix D.2. Figure 3 shows the training and test accuracy at each epoch on `Cifar10` in both (WC) and (IC) cases. While all methods achieve a similar test accuracy in the (WC) case on both datasets, methods based on the Euclidean gradient seem to suffer a drastic drop in performance in the (IC) case. This doesn't happen for KWNG (red line) which achieves a similar test accuracy as in (WC) case. Moreover, a speed-up in convergence can be obtained by increasing the number of basis points M (brown line). On `Cifar100`, KWNG is also less affected by the ill-conditioning as shown in Figure 4, albeit to a lower extent. Indeed, the larger number of classes in `Cifar100` makes the estimation of KWNG harder as discussed in Section 4.1. In this case, increasing the batch-size can substantially improve the training accuracy (pink line). Moreover, methods that are used to improve optimization using the Euclidean gradient can also be used for KWNG. For instance, using Momentum leads to an improved performance in the (WC) case (grey line). Interestingly, KFAC seems to also suffer a drop in performance in the (IC) case. This might result from the use of an isotropic damping term $D(\theta) = I$ which would be harmful in this case. Thus, we expect an increased performance for KFAC if a scale-sensitive damping can be used instead as for KWNG.

References

Amari, S.-i. (1985). *Differential-Geometrical Methods in Statistics*. Lecture Notes in Statistics. Springer-Verlag, New York.

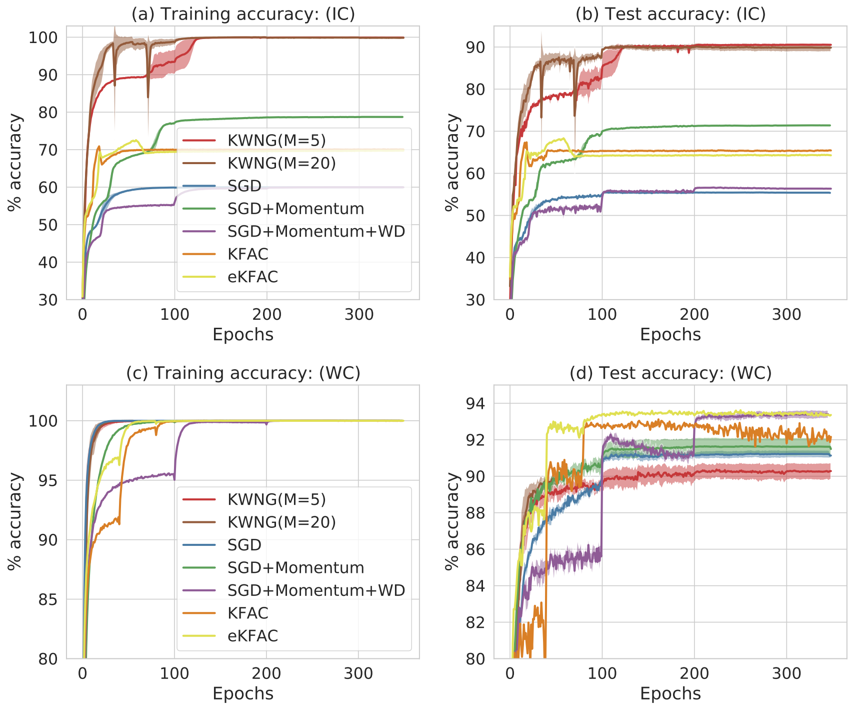

Figure 3: Training accuracy (left) and test accuracy (right) for classification on `Cifar10` in both the ill-conditioned case (top) and well-conditioned case (bottom) for different optimization methods. Results are averaged over 5 independent runs except for KFAC and eKFAC.

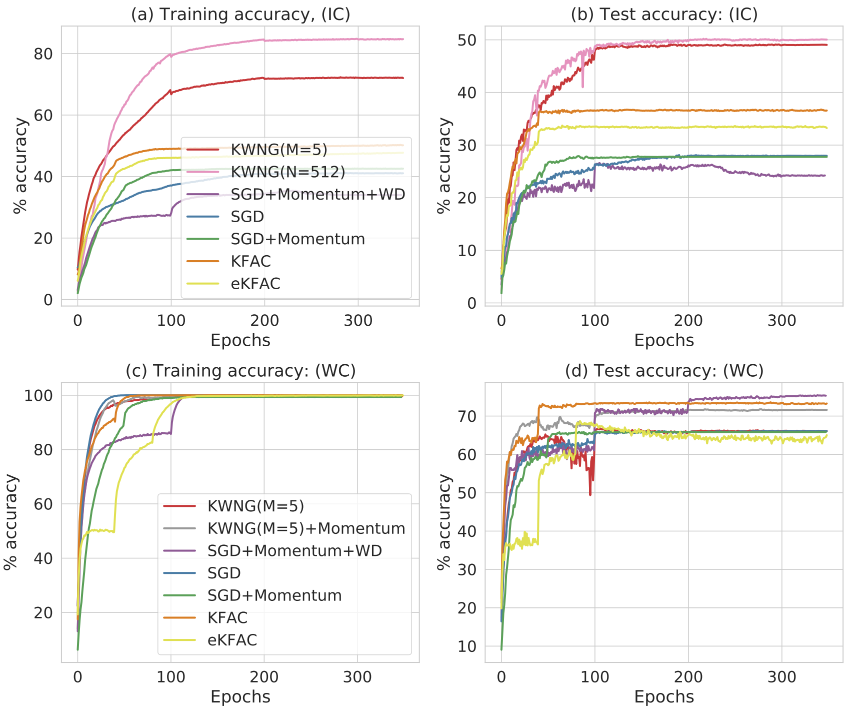

Figure 4: Training accuracy (left) and test accuracy (right) for classification on `Cifar100` in both the ill-conditioned case (top) and well-conditioned case (bottom) for different optimization methods.

Amari, S.-i. (1998). Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2):251–276.

Ambrosio, L., Gigli, N., and Savaré, G. (2004). Gradient flows with metric and differentiable structures, and applications to the Wasserstein space. *Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti Lincei. Matematica e Applicazioni*, 15(3-4):327–343.

Arbel, M. and Gretton, A. (2017). Kernel Conditional Exponential Family. *arXiv:1711.05363 [stat]*. arXiv: 1711.05363.

Benamou, J.-D. and Brenier, Y. (2000). A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393.

Bernacchia, A., Lengyel, M., and Hennequin, G. (2018). Exact natural gradient in deep linear networks and its application to the nonlinear case. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 5941–5950. Curran Associates, Inc.

Chen, Y. and Li, W. (2018). Natural gradient in Wasserstein statistical manifold. *arXiv:1805.08380 [cs, math]*. arXiv: 1805.08380.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Ekeland, I. and Témam, R. (1999). *Convex Analysis and Variational Problems*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics.

Fisher, R. A. and Russell, E. J. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368.

George, T., Laurent, C., Bouthillier, X., Ballas, N., and Vincent, P. (2018). Fast Approximate Natural Gradient Descent in a Kronecker-factored Eigenbasis. *arXiv:1806.03884 [cs, stat]*. arXiv: 1806.03884.

Grosse, R. and Martens, J. (2016). A Kronecker-factored Approximate Fisher Matrix for Convolution Layers. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 573–582. JMLR.org. event-place: New York, NY, USA.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Heskes, T. (2000). On "Natural" Learning and Pruning in Multilayered Perceptrons. *Neural Computation*, 12(4):881–901.

Hinton, G., Srivastava, N., and Swersky, K. (2012). Lecture 6a overview of mini–batch gradient descent.

Holbrook, A., Lan, S., Streets, J., and Shahbaba, B. (2017). The nonparametric Fisher geometry and the chi-square process density prior. *arXiv:1707.03117 [stat]*. arXiv: 1707.03117.

Ioffe, S. and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 448–456. JMLR.org. event-place: Lille, France.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. arXiv: 1412.6980.

Kingma, D. P., Salimans, T., and Welling, M. (2015). Variational Dropout and the Local Reparameterization Trick. *ArXiv*, abs/1506.02557. arXiv: 1506.02557.

Lafferty, J. and Wasserman, L. (2008). Rodeo: sparse, greedy nonparametric regression. *The Annals of Statistics*, 36(1):28–63.

LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R. (2012). Efficient BackProp. In Montavon, G., Orr, G. B., and Müller, K.-R., editors, *Neural Networks: Tricks of the Trade: Second Edition*, Lecture Notes in Computer Science, pages 9–48. Springer Berlin Heidelberg, Berlin, Heidelberg.

Li, W., Lin, A. T., and Montufar, G. (2019). Affine natural proximal learning.

Li, W. and Montufar, G. (2018a). Natural gradient via optimal transport. *arXiv:1803.07033 [cs, math]*. arXiv: 1803.07033.

Li, W. and Montufar, G. (2018b). Ricci curvature for parametric statistics via optimal transport. *arXiv:1807.07095 [cs, math, stat]*. arXiv: 1807.07095.

Malagò, L., Montrucchio, L., and Pistone, G. (2018). Wasserstein Riemannian Geometry of Positive Definite Matrices. *arXiv:1801.09269 [math, stat]*. arXiv: 1801.09269.

Martens, J. (2010). Deep Learning via Hessian-free Optimization. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 735–742, USA. Omnipress. event-place: Haifa, Israel.

Martens, J. and Grosse, R. (2015). Optimizing Neural Networks with Kronecker-factored Approximate Curvature. *arXiv:1503.05671 [cs, stat]*. arXiv: 1503.05671.

Martens, J. and Sutskever, I. (2012). Training Deep and Recurrent Networks with Hessian-Free Optimization. In Montavon, G., Orr, G. B., and Müller, K.-R., editors, *Neural Networks: Tricks of the Trade: Second Edition*, Lecture Notes in Computer Science, pages 479–535. Springer Berlin Heidelberg, Berlin, Heidelberg.

Mishkin, A., Kunstner, F., Nielsen, D., Schmidt, M., and Khan, M. E. (2018). SLANG: Fast Structured Covariance Approximations for Bayesian Deep Learning with Natural Gradient. *arXiv:1811.04504 [cs, stat]*. arXiv: 1811.04504.

Modin, K. (2017). Geometry of Matrix Decompositions Seen Through Optimal Transport and Information Geometry. *Journal of Geometric Mechanics*, 9(3):335–390. arXiv: 1601.01875.

Ollivier, Y., Arnold, L., Auger, A., and Hansen, N. (2011). Information-geometric optimization algorithms: A unifying picture via invariance principles. *J. Mach. Learn. Res.*, 18:18:1–18:65.

Otto, F. (2001). The Geometry of Dissipative Evolution Equations: The Porous Medium Equation. *Communications in Partial Differential Equations*, 26(1-2):101–174.

Otto, F. and Villani, C. (2000). Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality. *Journal of Functional Analysis*, 173(2):361–400.

Rao, C. R. (1992). Information and the Accuracy Attainable in the Estimation of Statistical Parameters. In Kotz, S. and Johnson, N. L., editors, *Breakthroughs in Statistics: Foundations and Basic Theory*, Springer Series in Statistics, pages 235–247. Springer New York, New York, NY.

Rudi, A., Camoriano, R., and Rosasco, L. (2015). Less is more: Nyström computational regularization.

Salimans, T. and Kingma, D. P. (2016). Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks. *arXiv:1602.07868 [cs]*. arXiv: 1602.07868.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In Helmbold, D. and Williamson, B., editors, *Computational Learning Theory*, pages 416–426, Berlin, Heidelberg. Springer Berlin Heidelberg.

Simonyan, K. and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*. arXiv: 1409.1556.

Sriperumbudur, B., Fukumizu, K., Kumar, R., Gretton, A., and Hyvärinen, A. (2017). Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition.

Strathmann, H., Sejdinovic, D., Livingstone, S., Szabo, Z., and Gretton, A. (2015). Gradient-free hamiltonian monte carlo with efficient kernel exponential families. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 955–963. Curran Associates, Inc.

Sutherland, D. J., Strathmann, H., Arbel, M., and Gretton, A. (2017). Efficient and principled score estimation.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pages 1139–1147.

Villani, C. (2009). Optimal transport: Old and new. Technical report.

A Preliminaries

A.1 Notation

We recall that Ω is an open subset of $\mathbb{R}^d$ while Θ is an open subset of parameters in $\mathbb{R}^q$. Let $\mathcal{Z} \subset \mathbb{R}^p$ be a latent space endowed with a probability distribution ν over $\mathcal{Z}$. Additionally, $(\theta,z) \mapsto h_\theta(z) \in \Omega$ is a function defined over $\Theta \times \mathcal{Z}$. We consider a parametric set of probability distributions $\mathcal{P}_\Theta$ over Ω defined as the implicit model:

$$\mathcal{P}_\Theta := \{\rho_\theta := (h_\theta)_{\#}\nu \quad ; \quad \theta \in \Theta\},$$

where by definition, $\rho_\theta = (h_\theta)_{\#}\nu$ means that any sample x from ρ_θ can be written as $x = h_\theta(z)$ where z is a sample from ν. We will write B to denote the jacobian of h_θ w.r.t. θ viewed as a linear map from $\mathbb{R}^q$ to $L_2(\nu)^d$ without explicit reference to θ:

$$Bu(z) = \nabla h_\theta(z).u; \qquad \forall u \in \mathbb{R}^q.$$

As in the main text, $\mathcal{L} : \Theta \to \mathbb{R}$ is a loss functions which is assumed to be of the form $\mathcal{L} = \mathcal{F}(\rho_\theta)$, with $\mathcal{F}$ a real valued functional over the set of probability distributions. $\nabla\mathcal{L}(\theta)$ denotes the euclidean gradient of $\mathcal{L}$ w.r.t θ while $\widehat{\nabla\mathcal{L}(\theta)}$ is an estimator of $\nabla\mathcal{L}(\theta)$ using N samples from ρ_θ.

We also consider a Reproducing Kernel Hilbert Space $\mathcal{H}$ of functions defined over Ω with inner product $\langle .,. \rangle_{\mathcal{H}}$ nad norm $\|.\|_{\mathcal{H}}$ and with a kernel $k : \Omega \times \Omega \to \mathbb{R}$. The reproducing property for the derivatives (Steinwart and Christmann, 2008, Lemma 4.34) will be important $\partial_i f(x) = \langle f, \partial_i k(x,.) \rangle_{\mathcal{H}}$ for all $x \in \Omega$. It holds as long as k is differentiable.

$C_b^\infty(\Omega)$ denotes the space of smooth bounded real valued functions on Ω, and $C_c^\infty(\Omega) \subset C_b^\infty(\Omega)$ denotes the subset of compactly supported functions. For any measured space $\mathcal{Z}$ with probability distribution ν, we denote by $L_2(\nu)$ the space of real valued and square integrable functions under ν and by $L_2(\nu)^d$ the space of square integrable vector valued functions under ν and with values in $\mathbb{R}^d$.

A.2 Assumptions

We make the following set of assumptions:

- **(A)** Ω is a non-empty open subset of $\mathbb{R}^d$.
- **(B)** There exists positive constants ζ and σ such that $\int \|z\|^p \mathrm{d}\nu(z) \leq \frac{1}{2} p! \zeta^{p-2} \sigma^2$ for any $p \geq 2$.
- **(C)** For all $\theta \in \Theta$ there exists $C(\theta)$ such that $\|\nabla_\theta h_\theta(z)\| \leq C(\theta)(1 + \|z\|)$ for all $z \in \mathcal{Z}$.
- **(D)** For all $0 \leq \delta \leq 1$, it holds with probability at least $1-\delta$ that $\|\widehat{\nabla\mathcal{L}(\theta)} - \nabla\mathcal{L}(\theta)\| \lesssim N^{-\frac{1}{2}}$.
- **(E)** k is twice continuously differentiable on $\Omega \times \Omega$.
- **(F)** For all $\theta \in \Theta$ it holds that $\int \partial_i \partial_{i+d} k(x,x) \mathrm{d}p_\theta(x) < \infty$ for all $1 \leq i \leq d$.
- **(G)** The following quantity is finite: $\kappa^2 = \sup_{\substack{x \in \Omega \\ 1 \leq i \leq q}} \partial_i \partial_{i+q} k(x,x)$.

A.3 Operators definition

Differential operators. We introduce the linear L operator and its adjoint $L^\top$:

$$\begin{aligned} L : \mathcal{H} &\to L_2(\nu)^d & \qquad L^\top : L_2(\nu)^d &\to \mathcal{H} \\ f &\mapsto (\partial_i f \circ h_\theta)_{1 \leq i \leq d} & v &\mapsto \int \sum_{i=1}^d \partial_i k(h_\theta(z),.) v_i(z) \mathrm{d}\nu(z) \end{aligned}$$

This allows to obtain the linear operator A defined in the main text (17) by composition $A := L^\top L$. We recall here another expression for A in terms of outer product $\otimes$ and its regularized version for a given $\lambda > 0$,

$$A = \int \sum_{i=1}^d \partial_i k(h_\theta(z),.) \otimes \partial_i k(h_\theta(z),.) \mathrm{d}\nu(z) \qquad A_\lambda := A + \lambda I.$$

It is easy to see that A is a symmetric positive operator. Moreover, it was established in Sriperumbudur et al. (2017) that A is also a compact operator under Assumption (F).

Assume now we have access to N samples $(Z_n)_{1\leq n\leq N}$ as in the main text. We define the following objects:

$$\hat{A} := \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{d}\partial_i k(h_\theta(Z_n),.)\otimes\partial_i k(h_\theta(Z_n),.), \qquad \hat{A}_\lambda := \hat{A}+\lambda I.$$

Furthermore, if v is a continuous function in $L_2(\nu)^d$, then we can also consider an empirical estimator for $L^\top v$:

$$\widehat{L^\top v} := \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{d}\partial_i k(h_\theta(Z_n),.)v_i(Z_n).$$

Subsampling operators. We consider the operator Q_M defined from $\mathcal{H}$ to $\mathbb{R}^M$ by:

$$Q_M := \frac{\sqrt{q}}{\sqrt{M}}\sum_{m=1}^{M} e_m\otimes\partial_{i_m}k(Y_m,.) \tag{19}$$

where $(e_m)_{1\leq m\leq M}$ is an orthonormal basis of $\mathbb{R}^M$. Q_M admits a singular value decomposition of the form $Q_M = U\Sigma V^\top$, with $VV^\top := P_M$ being the orthogonal projection operator on the Nyström subspace $\mathcal{H}_M$. Similarly to Rudi et al. (2015); Sutherland et al. (2017), we define the projected inverse function $\mathcal{G}_M(C)$ as:

$$\mathcal{G}_M(C) = V(V^\top CV)^{-1}V^\top.$$

We recall here some properties of $\mathcal{G}_M$ from (Sutherland et al., 2017, Lemma 1):

Lemma 7. *Let $A:\mathcal{H}\to\mathcal{H}$ be a positive operator, and define $A_\lambda = A+\lambda I$ for any $\lambda > 0$. The following holds:*

1. $\mathcal{G}_M(A)P_M = \mathcal{G}_M(A)$
2. $P_M\mathcal{G}_M(A) = \mathcal{G}_M(A)$
3. $\mathcal{G}_M(A_\lambda)A_\lambda P_M = P_M$
4. $\mathcal{G}_M(A_\lambda) = (P_M A P_M + \lambda I)^{-1}P_M$
5. $\|A_\lambda^{\frac{1}{2}}\mathcal{G}_M(A_\lambda)A_\lambda^{\frac{1}{2}}\|$

Estimators of the Wasserstein information matrix. Here we would like to express the estimator in Proposition 4 in terms of the operators introduced previously. We have the following proposition:

Proposition 8. *The estimator defined in Proposition 4 admits the following representation:*

$$\widehat{\nabla^W\mathcal{L}(\theta)} = (\epsilon D(\theta) + G_{M,N})^{-1}\widehat{\nabla\mathcal{L}(\theta)}$$

where $G_{M,N}$ is given by:

$$G_{M,N} := (\widehat{L^\top B})^\top\mathcal{G}_M(\hat{A}_\lambda)\widehat{L^\top B}.$$

Proof. This a direct consequence of the minimax theorem (Ekeland and Témam, 1999, Proposition 2.3, Chapter VI) and applying (Sutherland et al., 2017, Lemma 3). □

The matrix $G_{M,N}$ is in fact an estimator of the Wasserstein information matrix defined in Definition 2. We will also need to consider the following population version of $G_{M,N}$ defined as :

$$G_M := (L^\top B)^\top\mathcal{G}_M(A_\lambda)L^\top B \tag{20}$$

B Background in Information Geometry

B.1 Fisher-Rao Statistical Manifold

In this section we briefly introduce the non-parametric Fisher-Rao metric defined over the set $\mathcal{P}$ of probability distributions with positive density. More details can be found in Holbrook et al. (2017). By abuse of notation, an element $\rho \in \mathcal{P}$ will be identified with its density which will also be denoted by ρ. Consider $\mathcal{T}_\rho$, the set of real valued functions f defined over Ω and satisfying

$$\int \frac{f(x)^2}{\rho(x)} \mathrm{d}x < \infty; \qquad \int f(x)\rho(x)\mathrm{d}x = 0.$$

We have the following definition for the Fisher-Rao metric:

Definition 4 (Fisher-Rao metric)**.** *The Fisher-Rao metric g^F is defined for all $\rho \in \mathcal{P}$ as an inner product over $\mathcal{T}_\rho$ of the form:*

$$g^F_\rho(f,g) := \int \frac{1}{\rho(x)} f(x)g(x)\mathrm{d}x, \qquad \forall f,g \in \mathcal{T}_\rho$$

Note that the choice of the set $\mathcal{T}_\rho$ is different from the one considered in Holbrook et al. (2017) which replaces the integrability condition by a smoothness one. In fact, it can be shown that these choices result in the same metric by a density argument.

B.2 Wasserstein Statistical Manifold

In this section we review the theory of Wasserstein statistical manifold introduced in Li and Montufar (2018a); Chen and Li (2018). By analogy to the Fisher-Rao metric which allows to endow the parametric model $\mathcal{P}_\Theta$ with the structure of a Riemannian manifold, it is also possible to use a different metric that is derived from the Wasserstein 2 distance. We first start by briefly introducing the Wasserstein 2 distance. Given two probability distributions ρ and ρ', we consider the set of all joint probability distributions $\Pi(\rho,\rho')$ between ρ and ρ' usually called the set of *couplings* between ρ and ρ'. Any coupling π defines a way of transporting mass from ρ to ρ'. The cost of such transport can be measured as the expected distance between an element of mass of ρ at location x that is mapped to an element of mass of ρ' at location y using the coupling π:

$$\int \|x-y\|^2 \mathrm{d}\pi(x,y)$$

The squared Wasserstein 2 distance between ρ and ρ' is defined as the smallest transport cost over all possible couplings:

$$W_2^2(\rho,\rho') = \inf_{\pi \in \Pi(\rho,\rho')} \int \|x-y\|^2 \mathrm{d}\pi(x,y).$$

A dynamical formulation of W_2 was provided by the celebrated Benamou-Brenier formula in Benamou and Brenier (2000):

$$W_2^2(\rho,\rho') = \inf_{\phi_t} \int_0^1 \int \|\phi_l(x)\|^2 \mathrm{d}\rho_l(x)\mathrm{d}l$$

where the infimum is taken over the set of vector fields $\phi : [0,1] \times \Omega \rightarrow \mathbb{R}^d$. Each vector field of the potential determines a corresponding probability distribution ρ_l as the solution of the continuity equation:

$$\partial_l \rho_l + div(\rho_l \phi_l) = 0, \qquad \rho_0 = \rho, \rho_1 = \rho'. \tag{21}$$

When Ω is a compact set, a Neumann condition is added on the boundary of Ω to ensure that the total mass is conserved. Such formulation suggests that $W_2(\rho,\rho')$ corresponds in fact to the shortest path from ρ to ρ'. Indeed, given a path ρ_l from ρ to ρ', the infinitesimal displacement direction is given by the distribution $\partial_l \rho_l$. The length $|\partial_l \rho_l|$ of this direction is measured by: $|\partial_l \rho_l|^2 := \int \|\phi_l(x)\|^2 \mathrm{d}\rho_l(x)$ Hence, $W_2^2(\rho,\rho')$ can be written as:

$$W_2^2(\rho\rho') = \inf_{\rho_l} \int_0^1 |\partial_l \rho_l|^2 \mathrm{d}l.$$

In fact, $\partial_l \rho_l$ can be seen as an element in the tangent space $T_{\rho_l}\mathcal{P}_2$ to $\mathcal{P}_2$ at point ρ_l. To ensure that (21) is well defined, $T_\rho \mathcal{P}_2$ can be defined as the set of distributions σ satisfying $\sigma(1) = 0$.

$$|\sigma(f)| \leq C \|\nabla f\|_{L_2(\rho)}, \qquad \forall f \in C_c^\infty(\Omega) \tag{22}$$

for some positive constant C. Indeed, the condition in (22) guarantees the existence of a vector field ϕ_σ that is a solution to the PDE: $\sigma = -div(\rho\phi_\sigma)$.

Moreover, $|\partial_l \rho_l|^2$ can be seen as an inner product of $\partial_l \rho_l$ with itself in $T_{\rho_l}\mathcal{P}_2$. This inner product defines in turn a metric tensor g^W on $\mathcal{P}_2$ called the Wasserstein metric tensor (see Otto and Villani (2000); Ambrosio et al. (2004)):

Definition 5. *The Wasserstein metric g^W is defined for all $\rho \in \mathcal{P}_2$ as the inner product over $T_\rho \mathcal{P}_2$ of the form:*

$$g_\rho^W(\sigma,\sigma') := \int \phi_\sigma(x)^\top \phi_{\sigma'}(x) \mathrm{d}\rho(x), \qquad \forall \sigma, \sigma' \in T_\rho \mathcal{P}_2$$

where ϕ_σ and $\phi_{\sigma'}$ are solutions to the partial differential equations:

$$\sigma = -div(\rho\phi_\sigma), \qquad \sigma' = -div(\rho\phi_{\sigma'}).$$

Moreover, ϕ_σ and $\phi_{\sigma'}$ are required to be in the closure of gradient of smooth and compactly supported functions w.r.t. $L_2(\rho)^d$.

Definition 5 allows to endow $\mathcal{P}_2$ with a formal Riemannian structure with W_2 being its geodesic distance:

$$W_2^2(\rho,\rho') = \inf_{\rho_l} \int_0^1 g_{\rho_l}(\partial_l \rho_l, \partial_l \rho_l) \mathrm{d}l.$$

C Proofs

C.1 Preliminary results

We fist provide a proof of the dual formulation for the Fisher information matrix.

Proof of Proposition 1 . Consider the optimization problem:

$$\sup_{\substack{f \in C_c^\infty(\Omega) \\ \int f(x)\mathrm{d}\rho_\theta(x) = 0}} \left(\int f(x) \nabla \rho_\theta(x) \mathrm{d}x \right)^\top u - \frac{1}{2} \int f(x)^2 \rho_\theta(x) \mathrm{d}x \tag{23}$$

Recalling that the set of smooth and compactly supported functions $C_c^\infty(\infty)$ is dense in $L_2(\rho_\theta)$ and that the objective function in (23) is continuous and coercive in f, it follows that (23) admits a unique solution f^* in $L_2(\rho_\theta)$ which satisfies the optimality condition:

$$\int f(x) (\nabla \rho_\theta(x))^\top u \mathrm{d}x = \int f(x) f^*(x) \rho_\theta(x) \mathrm{d}x \qquad \forall f \in L_2(\rho_\theta)$$

Hence, it is easy to see that $f^* = (\nabla \rho_\theta)^\top u / \rho_\theta$ and that the optimal value of (23) is given by:

$$\frac{1}{2} \int \frac{((\nabla \rho_\theta(x))^\top u)^2}{\rho_\theta(x)} \mathrm{d}x.$$

This is equal to $u^\top G_F(\theta) u$ by Definition 1. □

The next proposition ensures that the Wasserstein information matrix defined in Definition 2 is well-defined and has a dual formulation.

Proposition 9. *Consider the model defined in (5) and let $(e_s)_{1 \leq s \leq q}$ be an orthonormal basis of $\mathbb{R}^q$. Under Assumptions **(B)** and **(C)**, there exists an optimal solution $\Phi = (\phi_s)_{1 \leq s \leq q}$ with ϕ_s in $L_2(\rho_\theta)^d$ satisfying the PDE:*

$$\partial_s \rho_\theta = -div(\rho_\theta \phi_s)$$

The elliptic equations also imply that $L^\top \nabla h_\theta = L^\top(\Phi \circ h_\theta)$. Moreover, the Wasserstein information matrix $G_W(\theta)$ on $\mathcal{P}_\Theta$ at point θ can be written as $G_W(\theta) = \Phi^\top \Phi$ where the inner-product is in $L_2(\rho_\theta)^d$ and satisfies:

$$\frac{1}{2}u^\top G_W(\theta)u = \sup_{\substack{f \in C_c^\infty(\Omega) \\ \int f(x)\mathrm{d}\rho_\theta(x) = 0}} \nabla \rho_\theta(f)^\top u - \frac{1}{2}\int \|\nabla_x f(h_\theta(z))\|^2 \mathrm{d}\nu(z).$$

for all $u \in \mathbb{R}^q$.

Proof. Let $(e_s)_{1 \le s \le q}$ be an orthonormal basis of $\mathbb{R}^q$. For all $1 \le s \le q$, we will establish the existence of an optimal solution ϕ_s in $L_2(\rho_\theta)^d$ satisfying the PDE:

$$\partial_s \rho_\theta = -div(\rho_\theta \phi_s) \tag{24}$$

Consider the variational problem:

$$\sup_{\phi \in \mathcal{S}} \int \phi(h_\theta(z))^\top \partial_{\theta_s} h_\theta(z) - \frac{1}{2}\|\phi\|^2_{L_2(\rho_\theta)} \tag{25}$$

where $\mathcal{S}$ is a Hilbert space obtained as the closure in $L_2(\rho_\theta)^d$ of functions of the form $\phi = \nabla_x f$ with $f \in C_c^\infty(\Omega)$:

$$\mathcal{S} := \overline{\{\nabla_x f \quad | \quad f \in C_c^\infty(\Omega)\}}_{L_2(\rho_\theta^d)}.$$

We have by Assumption (C) that:

$$\int \phi(h_\theta(z))^\top \partial_{\theta_s} h_\theta(z)\mathrm{d}\nu(z) \le \le C(\theta)\sqrt{\int (1+\|z\|^2)\mathrm{d}\nu(z)} \int \|\phi\|_{L_2(\rho_\theta)}.$$

Moreover, by Assumption (B), we know that $\sqrt{\int (1+\|z\|^2)\mathrm{d}\nu(z)} < \infty$. This implies that the objective in (25) is continuous in ϕ while also being convex.s It follows that (25) admits a unique solution $\phi_s^* \in \mathcal{S}$ which satisfies for all $\phi \in \mathcal{S}$:

$$\int \phi(h_\theta(z))^\top \phi_s^*(h_\theta(z))\mathrm{d}\nu(z) = \int \phi(h_\theta(z))^\top \partial_{\theta_s} h_\theta(z))\mathrm{d}\nu(z)$$

In particular, for any $f \in C_c^\infty(\Omega)$, it holds that:

$$\int \nabla_x f(h_\theta(z))^\top \phi_s^*(h_\theta(z))\mathrm{d}\nu(z) = \int \nabla_x f(h_\theta(z))^\top \partial_{\theta_s} h_\theta(z))\mathrm{d}\nu(z)$$

which is equivalent to (24) and implies directly that $L^T \nabla h_\theta = L^T \Phi \circ h_\theta$ where $\Phi := (\phi_s^*)_{1 \le s \le q}$. The variational expression for $\frac{1}{2}u^\top G_W u$ follows by noting that (25) admits the same optimal value as

$$\sup_{\substack{f \in C_c^\infty(\Omega) \\ \int f(x)\mathrm{d}\rho_\theta(x) = 0}} \nabla \rho_\theta(f)^\top u - \frac{1}{2}\int \|\nabla_x f(h_\theta(z))\|^2 \mathrm{d}\nu(z).$$

That is because $\mathcal{S}$ is by definition the closure in $L_2(\rho_\theta)^d$ of the set of gradients of smooth and compactly supported functions on Ω. □

Proof of Proposition 2. This is a consequence of Proposition 9. □

C.2 Expression of the Estimator

We provide here a proof of Proposition 4

Proof of Proposition 4. Here, to simplify notations, we simply write D instead of $D(\theta)$. First consider the following optimization problem:

$$\inf_{f \in \mathcal{H}_M} \frac{1}{N}\sum_{n=1}^{N} \|\nabla f(X_n)\|^2 + \lambda \|f\|^2_{\mathcal{H}} + \frac{1}{\epsilon}\mathcal{R}(f)^\top D^{-1}\mathcal{R}(f) + \frac{2}{\epsilon}\mathcal{R}(f)^\top D^{-1}\widehat{\nabla \mathcal{L}(\theta)}$$

with $\mathcal{R}(f)$ given by $\mathcal{R}(f)=\frac{1}{N}\sum_{n=1}^{N}\nabla f(X_n)^\top B(Z_n)$. Now, recalling that any $f\in\mathcal{H}_M$ can be written as $f=\sum_{m=1}^{M}\alpha_m\partial_{i_m}k(Y_m,.)$, and using the reproducing property $\partial_i f(x)=\langle f,\partial_i k(x,.)\rangle_{\mathcal{H}}$ (Steinwart and Christmann, 2008, Lemma 4.34) , it is easy to see that:

$$\frac{1}{N}\sum_{n=1}^{N}\|\nabla f(X_n)\|^2=\frac{1}{N}\sum_{\substack{1\leq n\leq N\\1\leq i\leq d}}(\sum_{m=1}^{M}\alpha_m\partial_{i_m}\partial_{i+d}k(Y_m,X_n))^2.$$

$$\|f\|_{\mathcal{H}}^2=\sum_{1\leq m,m'\leq M}\alpha_m\alpha_{m'}\partial_{i_m}\partial_{i_{m'}+d}k(Y_m,Y_{m'})$$

$$\mathcal{R}(f)=\frac{1}{N}\sum_{\substack{1\leq n\leq N\\1\leq i\leq d\\1\leq m\leq M}}\alpha_m\partial_{i_m}\partial_{i+d}k(Y_m,X_n)B_i(Z_n)$$

The above can be expressed in matrix form using the matrices defined in Proposition 4:

$$\frac{1}{N}\sum_{n=1}^{N}\|\nabla f(X_n)\|^2=\alpha^\top CC^\top\alpha;\qquad\|f\|_{\mathcal{H}}^2=\alpha^\top K\alpha;\qquad\mathcal{R}(f)=\alpha^\top CB.$$

Hence the optimal solution $\hat{f}^*$ is of the form $\hat{f}^*=\sum_{m=1}^{M}\alpha_m^*\partial_{i_m}k(Y_m,.)$, with α^* obtained as a solution to the finite dimensional problem in $\mathbb{R}^M$:

$$\min_{\alpha\in\mathbb{R}^M}\alpha^\top(\epsilon CC^\top+\epsilon\lambda K+CBD^{-1}B^\top C^\top)\alpha+2\alpha^\top CBD^{-1}\widehat{\nabla\mathcal{L}(\theta)}$$

It is easy to see that α^* are given by:

$$\alpha^*=-(\epsilon CC^T+\epsilon\lambda K+CBD^{-1}B^TC^T)^\dagger CBD^{-1}\widehat{\nabla\mathcal{L}(\theta)}.$$

Now recall that the estimator in Proposition 4 is given by: $\widehat{\nabla^W\mathcal{L}(\theta)}=\frac{1}{\epsilon}D^{-1}\mathcal{U}_\theta(\hat{f}^*)$. Hence, $\frac{1}{\epsilon}D^{-1}(\widehat{\nabla\mathcal{L}(\theta)}-B^TC^T\alpha^*)$ The desired expression is obtained by noting that $CB=T$ using the chain rule.

□

C.3 Consistency Results

Now we provide a proof for Theorems 5 and 6 which are direct consequences of on Proposition 10 and rely on Proposition 11.

Proof of Theorem 5 . The proof is a direct consequence of Proposition 10 under Assumption 1. □

Proof of Theorem 6. The proof is a direct consequence of Proposition 10 under Assumption 2. □

Proposition 10. *Under Assumptions (A) to (G) and for $0\leq\delta\leq1$ and N large enough, it holds with probability at least $1-\delta$:*

$$\|\widehat{\nabla^W\mathcal{L}}-\nabla^W\mathcal{L}\|=\mathcal{O}(N^{-\frac{b}{2b+1}})$$

provided that $M\sim dN^{\frac{1}{2b+1}}\log N$, $\lambda\sim N^{\frac{1}{2b+1}}$ and $\epsilon\lesssim N^{-\frac{b}{2b+1}}$ where $b:=\min(1,\alpha+\frac{1}{2})$ when Assumption 1 holds and $b=\frac{1}{2+c}$ when Assumption 2 holds instead.

Proof. Here for simplicity we assume that $D(\theta)=I$ without loss of generality and we omit the dependence in θ and write $\nabla^W\mathcal{L}$ and $\nabla\mathcal{L}$ instead of $\nabla^W\mathcal{L}(\theta)$ and $\nabla\mathcal{L}(\theta)$ and $\nabla^W\mathcal{L}(\theta)$. We also define $\hat{G}_\epsilon=\epsilon I+G_{M,N}$ and $G_\epsilon=\epsilon I+G_W$. By Proposition 8, we know that $\widehat{\nabla^W\mathcal{L}}=\hat{G}_\epsilon^{-1}\widehat{\nabla\mathcal{L}}$. We use the following decomposition:

$$\|\widehat{\nabla^W\mathcal{L}}-\nabla^W\mathcal{L}\|\leq\|\hat{G}_\epsilon^{-1}(\widehat{\nabla\mathcal{L}}-\nabla\mathcal{L})\|+\|\hat{G}_\epsilon^{-1}(G_{M,N}-G_W)G_W^{-1}\nabla\mathcal{L}\|+\epsilon\|\hat{G}_\epsilon^{-1}G_W^{-1}\nabla\mathcal{L}\|$$

To control the norm of $\hat{G}_\epsilon^{-1}$ we write $\hat{G}_\epsilon^{-1} = G_\epsilon^{-\frac{1}{2}}(H + I)^{-1}G_\epsilon^{-\frac{1}{2}}$, where H is given by $H := G_\epsilon^{-\frac{1}{2}}(G_{M,N} - G_W)G_\epsilon^{-\frac{1}{2}}$. Hence, provided that $\mu := \lambda_{\max}(H)$, the highest eigenvalue of H, is smaller than 1, it holds that:

$$\|(H+I)^{-1}\| \leq (1-\mu)^{-1}.$$

Moreover, since G_W is positive definite, its smallest eigenvalue η is strictly positive. Hence, $\|G_\epsilon^{-1}\| \leq (\eta+\epsilon)^{-1}$. Therefore, we have $\|\hat{G}_\epsilon^{-1}\| \leq (\eta+\epsilon)(1-\mu))^{-1}$, which implies:

$$\|\widehat{\nabla^W \mathcal{L}} - \nabla^W \mathcal{L}\| \leq (\eta+\epsilon)^{-1}\left(\frac{\|\widehat{\nabla\mathcal{L}} - \nabla\mathcal{L}\|}{1-\mu} + \eta^{-1}\|\nabla\mathcal{L}\|\|G_{M,N} - G_W\| + \epsilon\eta^{-1}\|\nabla\mathcal{L}\|\right).$$

Let $0 \leq \delta \leq 1$. We have by Assumption (D) that $\|\widehat{\nabla\mathcal{L}} - \nabla\mathcal{L}\| = \mathcal{O}(N^{-\frac{1}{2}})$ with probability at least $1-\delta$. Similarly, by Proposition 11 and for N large enough, we have with probability at least $1-\delta$ that $\|G_{M,N} - G_W\| = \mathcal{O}(N^{-\frac{b}{2b+1}})$ where b is defined in Proposition 11. Moreover, for N large enough, one can ensure that $\mu \leq \frac{1}{2}$ so that the following bound holds with probability at least $1-\delta$:

$$\|\widehat{\nabla^W \mathcal{L}} - \nabla^W \mathcal{L}\| \lesssim (\eta+\epsilon)^{-1}\left(2N^{-\frac{1}{2}} + \eta^{-1}\|\nabla\mathcal{L}\|(N^{-\frac{b}{2b+1}} + \epsilon)\right).$$

Thus by setting $\epsilon \lesssim N^{-\frac{b}{2b+1}}$ we get the desired convergence rate. □

Proposition 11. *For any $0 \leq \delta \leq 1$, we have with probability as least $1-\delta$ and for N large enough that:*

$$\|G_{M,N} - G_W\| = \mathcal{O}(N^{-\frac{b}{2b+1}}).$$

provided that $M \sim dN^{\frac{1}{2b+1}}\log N$ where $b := \min(1, \alpha + \frac{1}{2})$ when Assumption 1 holds and $b = \frac{1}{2+c}$ when Assumption 2 holds instead.

Proof. To control the error $\|G_{M,N} - G_W\|$ we decompose it into an estimation error $\|G_{M,N} - G_M\|$ and approximation error $\|G_M - G_W\|$:

$$\|G_{M,N} - G_W\| \leq \|G_M - G_W\| + \|G_M - G_{M,N}\|$$

were G_M is defined in (20) and is obtained by taking the number of samples N to infinity while keeping the number of basis points M fixed.

The estimation error $\|G_M - G_{M,N}\|$ is controlled using Proposition 12 where, for any $0 \leq \delta \leq 1$, we have with probability at least $1-\delta$ and as long as $N \geq M(1, \lambda, \delta)$:

$$\|G_{M,N} - G_M\| \leq \frac{\|B\|}{\sqrt{N\lambda}}(a_{N,\delta} + \sqrt{2\gamma_1\kappa} + 2\gamma_1\frac{\lambda+\kappa}{\sqrt{N\lambda}}) + \frac{1}{N\lambda}a_{N,\delta}^2.$$

In the limit where $N \to \infty$ and $\lambda \to 0$, only the dominant terms in the above equation remain which leads to an error $\|G_{M,N} - G_M\| = \mathcal{O}((N\lambda)^{-\frac{1}{2}})$. Moreover, the condition on N can be expressed as $\lambda^{-1}\log\lambda^{-1} \lesssim N$.

To control the error approximation error $\|G_M - G_W\|$ we consider two cases: the *well-specified* case and the *miss-specified* case.

- *Well-specified* case. Here we work under Assumption 1 which allows to use Proposition 14. Hence, for any $0 \leq \delta \leq 1$ and if $M \geq M(d, \lambda, \delta)$, it holds with probability at least $1-\delta$:

$$\|G_M - G_W\| \lesssim \lambda^{\min(1, \alpha + \frac{1}{2})}$$

- *Miss-specified* case. Here we work under Assumption 2 which allows to use Proposition 13. Hence, for any $0 \leq \delta \leq 1$ and if $M \geq M(d, \lambda, \delta)$, it holds with probability at least $1-\delta$:

$$\|G_M - G_W\| \lesssim \lambda^{\frac{1}{2+c}}$$

Let's set $b := \min(1, \alpha + \frac{1}{2})$ for the well-specified case and $b = \frac{1}{2+c}$ for the miss-specified case. In the limit where $M \to \infty$ and $\lambda \to 0$ the condition on M becomes: $M \sim d\lambda^{-1}\log\lambda^{-1}$. Hence, when $M \sim d\lambda^{-1}\log\lambda^{-1}$ and $\lambda^{-1}\log\lambda^{-1} \lesssim N$ it holds with probability as least $1-\delta$ that

$$\|G_{M,N} - G_W\| = \mathcal{O}(\lambda^b + (\lambda N)^{-\frac{1}{2}}).$$

One can further choose λ of the form $\lambda = N^{-\theta}$. This implies a condition on M of the form $dN^\theta \log(N) \lesssim M$ and $N^\theta \log(N) \lesssim N$. After optimizing over θ to get the tightest bound, the optimal value is obtained when $\theta = 1/(2b+1)$ and the requirement on N is always satisfied once N is large enough. Moreover, one can choose $M \sim dN^{\frac{1}{2b+1}}\log N$ so that the requirement on M is satisfied for N large enough. In this case we get the following convergence rate:

$$\|G_{M,N} - G_W\| = \mathcal{O}(N^{-\frac{b}{2b+1}}).$$

□

Proposition 12. *For any $0 \leq \delta \leq 1$, provided that $N \geq M(1,\lambda,\delta)$, we have with probability as least $1-\delta$:*

$$\|G_{M,N} - G_M\| \leq \frac{\|B\|}{\sqrt{N\lambda}}\left(2a_{N,\delta} + \sqrt{2\gamma_1\kappa} + 2\gamma_1\frac{\lambda+\kappa}{\sqrt{N\lambda}}\right) + \frac{1}{N\lambda}a_{N,\delta}^2.$$

with:

$$a_{N,\delta} := \sqrt{2\sigma_1^2 \log\frac{2}{\delta}} + \frac{2a\log\frac{2}{\delta}}{\sqrt{N}}$$

Proof. For simplicity, we define $E = \widehat{L^\top B} - L^\top B$. By definition of $G_{M,N}$ and G_M we have the following decomposition:

$$\begin{aligned} G_{M,N} - G_M = & \underbrace{E^\top \mathcal{G}_M(\hat{A}_\lambda)E}_{\mathfrak{E}_0} + \underbrace{E^\top \mathcal{G}_M(\hat{A}_\lambda)L^\top B}_{\mathfrak{E}_1} + \underbrace{B^\top L\mathcal{G}_M(\hat{A}_\lambda)E}_{\mathfrak{E}_2} \\ & - \underbrace{B^\top L\mathcal{G}_M(A_\lambda)P_M(\hat{A}-A)P_M\mathcal{G}_M(\hat{A}_\lambda)L^\top B}_{\mathfrak{E}_3} \end{aligned}$$

The first three terms can be upper-bounded in the following way:

$$\begin{aligned} \|\mathfrak{E}_0\| &= \|E^\top \hat{A}_\lambda^{-\frac{1}{2}} \hat{A}_\lambda^{\frac{1}{2}} \mathcal{G}_M(\hat{A}_\lambda) \hat{A}_\lambda^{\frac{1}{2}} \hat{A}_\lambda^{-\frac{1}{2}} E\| \\ &\leq \|E\|^2 \underbrace{\|\hat{A}_\lambda^{-1}\|}_{\leq 1/\lambda} \underbrace{\|\hat{A}_\lambda^{\frac{1}{2}} \mathcal{G}_M(\hat{A}_\lambda) \hat{A}_\lambda^{\frac{1}{2}}\|}_{\leq 1} \\ \|\mathfrak{E}_1\| = \|\mathfrak{E}_2\| &= \|E^\top A_\lambda^{-\frac{1}{2}} A_\lambda^{\frac{1}{2}} \mathcal{G}_M(A_\lambda) A_\lambda^{\frac{1}{2}} A_\lambda^{-\frac{1}{2}} L^\top B\| \\ &\leq \|B\|\|E\| \underbrace{\|\hat{A}_\lambda^{-\frac{1}{2}}\|}_{\leq 1/\sqrt{\lambda}} \underbrace{\|\hat{A}_\lambda^{\frac{1}{2}} \mathcal{G}_M(\hat{A}_\lambda) \hat{A}_\lambda^{\frac{1}{2}}\|}_{\leq 1} \underbrace{\|A_\lambda^{-\frac{1}{2}} L^\top\|}_{\leq 1} \|\hat{A}_\lambda^{-\frac{1}{2}} A_\lambda^{\frac{1}{2}}\| \end{aligned}$$

For the last term $\mathfrak{E}_3$, we first recall that by definition of $\mathcal{G}_M(A_\lambda)$ we have:

$$\mathcal{G}_M(A_\lambda)P_M(\hat{A}-A)P_M\mathcal{G}_M(A_\lambda) = \mathcal{G}_M(A_\lambda)(\hat{A}-A)\mathcal{G}_M(A_\lambda).$$

Therefore, one can write:

$$\begin{aligned} \|\mathfrak{E}_3\| &= \|B^\top L A_\lambda^{-\frac{1}{2}} A_\lambda^{\frac{1}{2}} \mathcal{G}_M(A_\lambda) A_\lambda^{\frac{1}{2}} A_\lambda^{-\frac{1}{2}} (\hat{A}-A) A_\lambda^{-\frac{1}{2}} A_\lambda^{\frac{1}{2}} \hat{A}_\lambda^{-\frac{1}{2}} \hat{A}_\lambda^{\frac{1}{2}} \mathcal{G}_M(\hat{A}_\lambda) \hat{A}_\lambda^{\frac{1}{2}} \hat{A}_\lambda^{-\frac{1}{2}} A_\lambda^{\frac{1}{2}} A_\lambda^{-\frac{1}{2}} L^\top B\| \\ &\leq \|B\|^2 \underbrace{\|L A_\lambda^{-\frac{1}{2}}\|^2}_{\leq 1} \underbrace{\|A_\lambda^{\frac{1}{2}} \mathcal{G}_M(A_\lambda) A_\lambda^{\frac{1}{2}}\|}_{\leq 1} \underbrace{\|\hat{A}_\lambda^{\frac{1}{2}} \mathcal{G}_M(\hat{A}_\lambda) \hat{A}_\lambda^{\frac{1}{2}}\|}_{\leq 1} \|A_\lambda^{\frac{1}{2}} \hat{A}_\lambda^{\frac{1}{2}}\|^2 \|A_\lambda^{-\frac{1}{2}} (\hat{A}-A) A_\lambda^{-\frac{1}{2}}\| \\ &\leq \|B\|^2 \|A_\lambda^{\frac{1}{2}} \hat{A}_\lambda^{\frac{1}{2}}\|^2 \|A_\lambda^{-\frac{1}{2}} (\hat{A}-A) A_\lambda^{-\frac{1}{2}}\| \end{aligned}$$

We recall now (Rudi et al., 2015, Proposition 7.) which allows to upper-bound $\|A_\lambda^{\frac{1}{2}} \hat{A}_\lambda^{\frac{1}{2}}\|$ by $(1-\eta)^{-\frac{1}{2}}$ where $\eta = \lambda_{\max}(A_\lambda^{\frac{1}{2}}(A-\hat{A})A_\lambda^{\frac{1}{2}})$ provided that $\eta < 1$. Moreover, (Rudi et al., 2015, Proposition 8.) allows to control both η and $\|A_\lambda^{-\frac{1}{2}}(\hat{A}-A)A_\lambda^{-\frac{1}{2}}\|$ under Assumption (G). Indeed, for any $0\leq\delta\leq 1$ and provided that $0<\lambda\leq\|A\|$ it holds with probability $1-\delta$ that:

$$\|A_\lambda^{-\frac{1}{2}}(\hat{A}-A)A_\lambda^{-\frac{1}{2}}\|\leq 2\gamma_1\frac{1+\kappa/\lambda}{3N}+\sqrt{\frac{2\gamma_1\kappa}{N\lambda}};\qquad \eta\leq\frac{2\gamma_2}{3N}+\sqrt{\frac{2\gamma_2\kappa}{N\lambda}}$$

where γ_1 and γ_2 are given by:

$$\gamma_1=\log(\frac{8Tr(A)}{\lambda\delta});\qquad \gamma_2=\log(\frac{4Tr(A)}{\lambda\delta}).$$

Hence, for $N\geq M(1,\lambda,\delta)$ we have that $(1-\eta)^{-\frac{1}{2}}\leq 2$ and one can therefore write:

$$\|\mathfrak{E}_3\|\leq 4\|B\|^2(2\gamma_1\frac{1+\kappa/\lambda}{3N}+\sqrt{\frac{2\gamma_1\kappa}{N\lambda}})$$
$$\|\mathfrak{E}_1\|=\|\mathfrak{E}_1\|\leq\frac{2\|B\|}{\sqrt{\lambda}}\|E\|$$

The error $\|E\|$ is controlled by Proposition 17 where it holds with probability greater or equal to $1-\delta$ that:

$$\|E\|\leq\frac{1}{\sqrt{N}}(\sqrt{2\sigma_1^2\log\frac{2}{\delta}}+\frac{2a\log\frac{2}{\delta}}{\sqrt{N}}):=\frac{1}{\sqrt{N}}a_{N,\delta}.$$

Finally, we have shown that provided that $N\geq M(1,\lambda,\delta)$ then with probability greater than $1-\delta$ one has:

$$\|G_{M,N}-G_M\|\leq\frac{\|B\|}{\sqrt{N\lambda}}(2a_{N,\delta}+\sqrt{2\gamma_1\kappa}+2\gamma_1\frac{\lambda+\kappa}{\sqrt{N\lambda}})+\frac{1}{N\lambda}a_{N,\delta}^2.$$

□

Proposition 13. *Let* $0\leq\lambda\leq\|A\|$ *and define* $M(d,\lambda,\delta):=\frac{128}{9}\log\frac{4Tr(A)}{\lambda\delta}(d\kappa\lambda^{-1}+1)$. *Under Assumption 2 and Assumption (G), for any* $\delta\geq 0$ *such that* $M\geq M(d,\lambda,\delta)$ *the following holds with probability* $1-\delta$*:*

$$\|G_M-G_W\|\lesssim\lambda^{\frac{1}{2+c}}$$

Proof. We consider the error $\|G_M-G_W\|$. Recall that G_W is given by $G_W=\Phi^\top\Phi$ with Φ defined in Proposition 9. Let κ be a positive real number, we know by Assumption 2 that there exists $F^\kappa:=(f_s^\kappa)_{1\leq s\leq q}$ with $f_s^\kappa\in\mathcal{H}$ such that $\|\Phi-F^\kappa\|_{L_2(\rho_\theta)}\leq C\kappa$ and $\|f_s^\kappa\|_{\mathcal{H}}\leq C\kappa^{-c}$ for some fixed positive constant C. Therefore, we use F^κ to control the error $\|G_M-G_W\|$. Let's call $E=\Phi\circ h_\theta-LF^\kappa$ We consider the following decomposition:

$$\begin{aligned}G_M-G_W=&(L^\top\Phi\circ h_\theta)^\top\mathcal{G}_M(A_\lambda)L^\top\Phi\circ h_\theta-\Phi^\top\Phi\\ =&\underbrace{E^\top L\mathcal{G}_M(A_\lambda)L^\top E}_{\mathfrak{E}_1}-\underbrace{E^\top E}_{\mathfrak{E}_2}\\ &+\underbrace{F_\kappa^\top\big(L^\top L\mathcal{G}_M(A_\lambda)-I\big)L^\top\Phi\circ h_\theta}_{\mathfrak{E}_3}+\underbrace{E^\top L\big(\mathcal{G}_M(A_\lambda)L^\top L-I\big)F^\kappa}_{\mathfrak{E}_4}\end{aligned}$$

First we consider the term $\mathfrak{E}_1$ one simply has:

$$\|\mathfrak{E}_1\|\leq\kappa^2\underbrace{\|LA_\lambda^{-\frac{1}{2}}\|}_{\leq 1}\underbrace{\|A_\lambda^{\frac{1}{2}}\mathcal{G}_M(A_\lambda)A_\lambda^{\frac{1}{2}}\|}_{\leq 1}\underbrace{\|A_\lambda^{-\frac{1}{2}}L^\top\|}_{\leq 1}\leq\kappa^2$$

The second term also satisfies $\|\mathfrak{E}_1\|\leq\kappa^2$ by definition of F_κ. For the last two terms $\mathfrak{E}_3$ and $\mathfrak{E}_4$ we use Lemma 15 which allows to control the operator norm of $L(\mathcal{G}_M(A_\lambda)L^\top L-I)$. Hence, for any $\delta\geq 0$ and M such that $M\geq M(d,\lambda,\delta)$ and for $\kappa\leq 1$ it holds with probability $1-\delta$ that:

$$\|\mathfrak{E}_3\|\lesssim\sqrt{\lambda}\kappa^{-c};\qquad \|\mathfrak{E}_4\|\lesssim\sqrt{\lambda}\kappa^{-c}$$

We have shown so far that $\|G_M-G_W\|\lesssim(\kappa^2+2\kappa^{-c}\sqrt{\lambda})$. One can further optimize over κ on the interval $[0,1]$ to get a tighter bound. The optimal value in this case is $\kappa^*=\min(1,(c\lambda^{\frac{1}{2}})^{\frac{1}{2+c}})$. By considering $\lambda>0$ such that $(c\lambda^{\frac{1}{2}})^{\frac{1}{2+c}})\leq 1$, it follows directly that $\|G_M-G_W\|\lesssim\lambda^{\frac{1}{2+c}}$ which shows the desired result. □

Proposition 14. *Let $0 \leq \lambda \leq \|A\|$ and define $M(d,\lambda,\delta) := \frac{128}{9}\log\frac{4Tr(A)}{\lambda\delta}(d\kappa\lambda^{-1}+1)$. Under Assumption 1 and Assumption **(G)**, for any $\delta \geq 0$ such that $M \geq M(d,\lambda,\delta)$ the following holds with probability $1-\delta$:*

$$\|G_M - G_W\| \lesssim \lambda^{\min(1,\alpha+\frac{1}{2})}$$

Proof. Recall that G_W is given by $G_W = \Phi^\top \Phi$ with Φ defined in Proposition 9. By Assumption 1, we have that $\Phi = \nabla(A^\alpha V)$ with $V := (v_s)_{1\leq s\leq q} \in \mathcal{H}^q$. Hence, one can write

$$\begin{aligned} G_M - G_W &= (L^\top \Phi \circ h_\theta)^\top \mathcal{G}_M(A_\lambda) L^\top \Phi \circ h_\theta - \Phi^\top \Phi \\ &= V^\top (A^\alpha (A\mathcal{G}_M(A_\lambda)A - A) A^\alpha V \end{aligned}$$

we can therefore directly apply Lemma 15 and get $\|G_M - G_W\| \lesssim \lambda^{\min(1,\alpha+\frac{1}{2})}$ with probability $1-\delta$ for any $\delta \geq 0$ such that $M \geq M(d,\lambda,\delta)$. □

Lemma 15. *Let $0 \leq \lambda \leq \|A\|$, $\alpha \geq 0$ and define $M(d,\lambda,\delta) := \frac{128}{9}\log\frac{4Tr(A)}{\lambda\delta}(d\kappa\lambda^{-1}+1)$. Under Assumption **(G)**, for any $\delta \geq 0$ such that $M \geq M(d,\lambda,\delta)$ the following holds with probability $1-\delta$:*

$$\|L(\mathcal{G}_M(A_\lambda)L^\top L - I)A^\alpha\| \lesssim \lambda^{min(1,\alpha+\frac{1}{2})}$$

Proof. We have the following identities:

$$\begin{aligned} L(\mathcal{G}_M(A_\lambda)L^\top L - I)A^\alpha &= L(\mathcal{G}_M(A_\lambda)A_\lambda - I - \lambda\mathcal{G}_M(A_\lambda))A^\alpha \\ &= \underbrace{LA_\lambda^{-\frac{1}{2}}A_\lambda^{\frac{1}{2}}(\mathcal{G}_M(A_\lambda)A_\lambda P_M - I)A^\alpha}_{\mathfrak{E}_1} - \underbrace{\lambda LA_\lambda^{-\frac{1}{2}}A_\lambda^{\frac{1}{2}}\mathcal{G}_M(A_\lambda)A_\lambda^{\frac{1}{2}}A_\lambda^{-\frac{1}{2}}A^\alpha}_{\mathfrak{E}_3} \\ &\quad + \underbrace{LA_\lambda^{-\frac{1}{2}}A_\lambda^{\frac{1}{2}}\mathcal{G}_M(A_\lambda)A_\lambda^{\frac{1}{2}}A_\lambda^{\frac{1}{2}}(I-P_M)A^\alpha}_{\mathfrak{E}_2}. \end{aligned}$$

For the first $\mathfrak{E}_1$ we use (Sutherland et al., 2017, Lemma 1 (iii)) which implies that $\mathcal{G}_M(A_\lambda)A_\lambda P_M = P_M$. Thus $\mathfrak{E}_1 = LA_\lambda^{-\frac{1}{2}}A_\lambda^{\frac{1}{2}}(P_M - I)A^\alpha$. Moreover, by Lemma 16 we have that $\|A_\lambda^{\frac{1}{2}}(I-P_M)\| \leq 2\sqrt{\lambda}$ with probability $1-\delta$ for $M > M(d,\lambda,\delta)$. Therefore, recalling that $(I-P_M)^2 = I - P_M$ since P_M is a projection, one can further write:

$$\begin{aligned} \|\mathfrak{E}_1\| &\leq \underbrace{\|LA_\lambda^{-\frac{1}{2}}\|}_{\leq 1}\underbrace{\|\|A_\lambda^{\frac{1}{2}}(P_M - I)\|^2}_{\leq \lambda}\|A_\lambda^{-\frac{1}{2}}A^\alpha\| \\ \|\mathfrak{E}_2\| &\leq \underbrace{\|LA_\lambda^{-\frac{1}{2}}\|}_{\leq 1}\underbrace{\|\|A_\lambda^{\frac{1}{2}}\mathcal{G}_M(A_\lambda)A_\lambda^{\frac{1}{2}}\|}_{\leq 1}\underbrace{\|\|A_\lambda^{\frac{1}{2}}(P_M - I)\|^2}_{\leq 4\lambda}\|A_\lambda^{-\frac{1}{2}}A^\alpha\| \\ \|\mathfrak{E}_3\| &\leq \lambda\underbrace{\|LA_\lambda^{-\frac{1}{2}}\|}_{\leq 1}\underbrace{\|\|A_\lambda^{\frac{1}{2}}\mathcal{G}_M(A_\lambda)A_\lambda^{\frac{1}{2}}\|}_{\leq 1}\|A_\lambda^{-\frac{1}{2}}A^\alpha\| \end{aligned}$$

It remains to note that $\|A_\lambda^{-\frac{1}{2}}A^\alpha\| \leq \lambda^{\alpha-\frac{1}{2}}$ when $0 \leq \alpha \leq \frac{1}{2}$ and that $\|A_\lambda^{-\frac{1}{2}}A^\alpha\| \leq \|A\|^{\alpha-\frac{1}{2}}$ for $\alpha > \frac{1}{2}$ which allows to conclude. □

C.4 AUXILIARY RESULTS

Lemma 16. *Let $0 \leq \lambda \leq \|A\|$. Under Assumption **(G)**, for any $\delta \geq 0$ such that $M \geq M(d,\lambda,\delta) := \frac{128}{9}\log\frac{4Tr(A)}{\lambda\delta}(\kappa\lambda^{-1}+1)$ the following holds with probability $1-\delta$:*

$$\|A_\lambda^{\frac{1}{2}}(I - P_M)\| \leq 2\sqrt{\lambda}$$

Proof. The proof is an adaptation of the results in Rudi et al. (2015); Sutherland et al. (2017). Here we recall Q_M defined in (19). Its transpose $Q_M^\top$ sends vectors in $\mathbb{R}^M$ to elements in the span of the Nyström basis points, hence P_M and $Q_M^\top$ have the same range, i.e.: $range(P_M)=\overline{range}(Q_M^\top)$. We are in position to apply (Rudi et al., 2015, Proposition 3.) which allows to find an upper-bound on $A_\lambda^{\frac{1}{2}}(P_M-I)$ in terms of Q_M:

$$\|A_\lambda^{\frac{1}{2}}(P_M-I)\|\leq\sqrt{\lambda}\|A_\lambda^{\frac{1}{2}}(Q_M^\top Q_M+\lambda I)^{-\frac{1}{2}}\|.$$

For simplicity we write $\hat{A}_M:=Q_M^\top Q_M$ and $E_2:=A_\lambda^{-\frac{1}{2}}(A-\hat{A}_M)A_\lambda^{-\frac{1}{2}}$. We also denote by $\beta=\lambda_{max}(E_2)$ the highest eigenvalue of E_2. We can therefore control $\|A_\lambda^{\frac{1}{2}}(\hat{A}_M+\lambda I)^{-\frac{1}{2}}\|$ in terms of β using (Rudi et al., 2015, Proposition 7) provided that $\beta<1$:

$$\|A_\lambda^{\frac{1}{2}}(P_M-I)\|\leq\sqrt{\lambda}\frac{1}{\sqrt{1-\beta}}.$$

Now we need to make sure that $\beta<1$ for M large enough. To this end, we will apply (Rudi et al., 2015, Proposition 8.) to $\hat{A}_M$. Denote by $v_m=\sqrt{d}\partial_{i_m}k(Y_m,.)$. Hence, by definition of $\hat{A}_M$ it follows that $\hat{A}_M=\frac{1}{M}\sum_{m=1}^M v_m\otimes v_m$. Moreover, $(v_m)_{1\leq m\leq M}$ are independent and identically distributed and satisfy:

$$\mathbb{E}[v_m\otimes v_m]=\int\sum_{i=1}^q\partial_i k(y,.)\otimes\partial_i k(y,.)\mathrm{d}p_\theta(y)=A.$$

We also have by Assumption (G) that $\langle v_m,A_\lambda^{-1}v_m\rangle\leq\frac{d\kappa}{\lambda}$ almost surely and for all $\lambda>0$. We can therefore apply (Rudi et al., 2015, Proposition 8.) which implies that for any $1\geq\delta\geq0$ and with probability $1-\delta$ it holds that:

$$\beta\leq\frac{2\gamma}{3M}+\sqrt{\frac{2\gamma d\kappa}{M\lambda}}$$

with $\gamma=\log\frac{4Tr(A)}{\lambda\delta}$ provided that $\lambda\leq\|A\|$. Thus by choosing $M\geq\frac{128\gamma}{9}(d\kappa\lambda^{-1}+1)$ we have that $\beta\leq\frac{3}{4}$ with probability $1-\delta$ which allows to conclude. □

Proposition 17. *There exist $a>0$ and $\sigma_1>0$ such that for any $0\leq\delta\leq1$, it holds with probability greater of equal than $1-\delta$ that:*

$$\|\widehat{L^\top B}-L^\top B\|\leq\frac{2a\log\frac{2}{\delta}}{N}+\sqrt{\frac{2\sigma_1^2\log\frac{2}{\delta}}{N}}$$

Proof. denote by $v_n=\sum_{i=1}^d\partial_i k(X_n,.)B_i(Z_n)$, we have that $\mathbb{E}[v_n]=L^\top B$. We will apply Bernstein's inequality for sum of random vectors. For this we first need to find $a>0$ and $\sigma_1>0$ such that $\mathbb{E}[\|z_n-L^\top B\|_{\mathcal{H}}^p]\leq\frac{1}{2}p!\sigma_1^2a^{p-2}$. To simplify notations, we write x and x' instead of $h_\theta(z)$ and $h_\theta(z')$. We have that:

$$\begin{aligned}\mathbb{E}[\|z_n-L^\top B\|_{\mathcal{H}}^p]=&\int\left\|\sum_{i=1}^d\partial_i k(x,.)B_i(z)-\int\sum_{i=1}^d\partial_i k(x',.)B_i(z')\mathrm{d}\nu(z')\right\|^p\mathrm{d}\nu(z)\\ \leq&2^{p-1}\underbrace{\int\left\|\sum_{i=1}^d\int(\partial_i k(x,.)-\partial_i k(x',.))B_i(z)\mathrm{d}\nu(z')\right\|^p\mathrm{d}\nu(z)}_{\mathfrak{E}_1}\\ &+2^{p-1}\underbrace{\int\left\|\int\sum_{i=1}^d\partial_i k(x,.)(B_i(z)-B_i(z'))\mathrm{d}\nu(z')\right\|^p\mathrm{d}\nu(z)}_{\mathfrak{E}_2}\end{aligned}$$

We used the convexity of the norm and the triangular inequality to get the last line. We introduce the notation $\gamma_i(x):=\partial_i k(x,.)-\int\partial_i k(h_\theta(z'),.)\mathrm{d}\nu(z')$ and by $\Gamma(x)$ we denote the matrix whose components are given by $\Gamma(x)_{ij}:=\langle\gamma_i(x),\gamma_j(x)\rangle_{\mathcal{H}}$. The first term $\mathfrak{E}_1$ can be upper-bounded as follows:

$$\begin{aligned}\mathfrak{E}_1=&\int\left|Tr(B(z)B(z)^\top\Gamma(x))\right|^{\frac{p}{2}}\\ \leq&\int\left|\|B(z)\|^2Tr(\Gamma(x)^2)^{\frac{1}{2}}\right|^{\frac{p}{2}}.\end{aligned}$$

Moreover, we have that $Tr(\Gamma(x)^2)^{\frac{1}{2}} = (\sum_{1\leq i,j\leq d}\langle\gamma_i(x),\gamma_j(x)\rangle_{\mathcal{H}}^2)^{\frac{1}{2}} \leq \sum_{i=1}^d \|\gamma_i(x)\|^2$. We further have that $\|\gamma_i(x)\| \leq \partial_i\partial_{i+d}k(x,x)^{\frac{1}{2}} + \int \partial_i\partial_{i+d}k(h_\theta(z),h_\theta(z))^{\frac{1}{2}}\,\mathrm{d}\nu(z)$ and by Assumption (G) it follows that $\|\gamma_i(x)\| \leq 2\sqrt{\kappa}$. Hence, one can directly write that: $\mathfrak{E}_1 \leq (2\sqrt{\kappa d})^p \int \|B(z)\|^p\,\mathrm{d}\nu(z)$. Recalling Assumptions (B) and (C) we get:

$$\mathfrak{E}_1 \leq 2^{p-1}(2\sqrt{\kappa d})^p C(\theta)^p(1+\frac{1}{2}p!\zeta^{p-2}\sigma^2)$$

Similarly, we will find an upper-bound on $\mathfrak{E}_2$. To this end, we introduce the matrix $Q(x',x")$ whose components are given by $Q(x',x")_{i,j} = \partial_i\partial_{i+d}k(x',x")$. One, therefore has:

$$\begin{aligned}\mathfrak{E}_2 &= \int\left|\int\int Tr((B(z)-B(z'))(B(z)-B(z"))^\top Q(x',x")\mathrm{d}\nu(z')\mathrm{d}\nu(z")\right|^{\frac{p}{2}}\mathrm{d}\nu(z)\\ &\leq \int\left|\int\int \|B(z)-B(z')\|\|B(z)-B(z")\| Tr(Q(x',x")^2)^{\frac{1}{2}}\mathrm{d}\nu(z')\mathrm{d}\nu(z")\right|^{\frac{p}{2}}\mathrm{d}\nu(z)\end{aligned}$$

Once again, we have that $Tr(Q(x',x")^2)^{\frac{1}{2}} \leq (\sum_{i=1}^d \partial_i\partial_{i+d}k(x',x'))^{\frac{1}{2}}(\sum_{i=1}^d \partial_i\partial_{i+d}k(x",x"))^{\frac{1}{2}} \leq d\kappa$ thanks to Assumption (G). Therefore, it follows that:

$$\begin{aligned}\mathfrak{E}_2 &\leq (\sqrt{d\kappa})^p \int|\int \|B(z)-B(z')\|\mathrm{d}\nu(z)|^p\mathrm{d}\nu(z)\\ &\leq 3^{p-1}(\sqrt{d\kappa})^p C(\theta)^p(2^p + \int\|z\|^p\mathrm{d}\nu(z) + \left(\int\|z\|\mathrm{d}\nu(z)\right)^p)\\ &\leq 3^{p-1}(\sqrt{d\kappa})^p C(\theta)^p(2^p + \frac{1}{2}p!\zeta^{p-2}\sigma^2 + \left(\int\|z\|\mathrm{d}\nu(z)\right)^p).\end{aligned}$$

The second line is a consequence of Assumption (C) while the last line is due to Assumption (B). These calculations, show that it is possible to find constants a and σ_1 such that $\mathbb{E}[\|z_n - L^\top B\|_{\mathcal{H}}^p] \leq \frac{1}{2}p!\sigma_1^2 a^{p-2}$. Hence one concludes using Bernstein's inequality for a sum of random vectors (see for instance Rudi et al., 2015, Proposition 11). □

D Experimental details

D.1 Natural Wasserstein gradient for the Multivariate Normal model

Multivariate Gaussian. Consider a multivariate gaussian with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^d \times \mathbb{R}^d$ parametrized using its lower triangular components $s = T(\Sigma)$. We denote by $\Sigma = T^{-1}(s)$ the inverse operation that maps any vector $s \in \mathbb{R}^{\frac{d(d+1)}{2}}$ to its corresponding symmetric matrix in $\mathbb{R}^d \times \mathbb{R}^d$. The concatenation of the mean μ and s will be denoted as θ : $\theta = (\mu, s)$. Given two parameter vectors $u = (m, T(S))$ and $v = (m', T(S'))$ where m and m' are vectors in $\mathbb{R}^d$ and S and S' are symmetric matrices in $\mathbb{R}^d \times \mathbb{R}^d$ the metric evaluated at u and v is given by:

$$u^\top G(\theta)v = m^\top m' + Tr(A\Sigma A')$$

where A and A' are symmetric matrices that are solutions to the Lyapunov equation:

$$S = A\Sigma + \Sigma A, \qquad S' = A'\Sigma + \Sigma A'.$$

A and A' can be computed in closed form using standard routines making the evaluation of the metric easy to perform. Given a loss function $\mathcal{L}(\theta)$ and gradient direction $\nabla_\theta\mathcal{L}(\theta) = \nabla_\mu\mathcal{L}(\theta), \nabla_s\mathcal{L}(\theta)$, the corresponding natural gradient $\nabla_\theta^W\mathcal{L}(\theta)$ can also be computed in closed form:

$$\nabla_\theta^W\mathcal{L}(\theta) = (\nabla_\mu\mathcal{L}(\theta), T(\Sigma(A + diag(A)) + (A + diag(A))\Sigma)),$$

where $A = T^{-1}(\nabla_s\mathcal{L}(\theta))$. To use the estimator proposed in Proposition 4 we take advantage of the parametrization of the Gaussian distribution as a push-forward of a standard normal vector:

$$X \sim \mathcal{N}(\mu, \Sigma) \iff X = \Sigma^{\frac{1}{2}}Z + \mu, Z \sim \mathcal{N}(0, I_d)$$

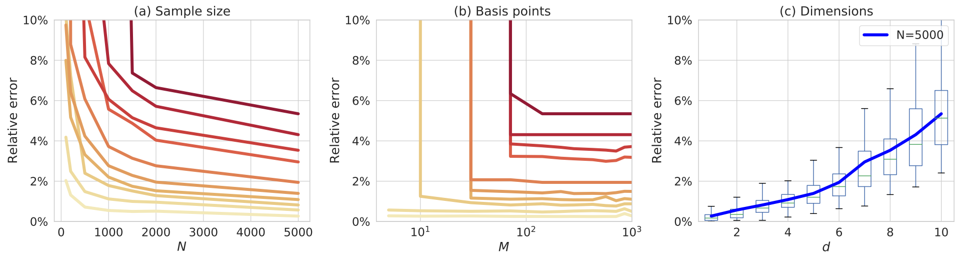

Figure 5: Evolution of the relative error of KWNG for the multivariate log-normal model averaged over 100 runs. For each run, a random value for the parameter θ and for the Euclidean gradient $\nabla\mathcal{L}(\theta)$ is sampled from a centered Gaussian with variance 0.1. In all cases, $\lambda = \epsilon = 10^{-10}$. Left (a): Relative error as the sample size N increases. The dimension d of the sample space varies from $d = 1$ (yellow) to $d = 10$ (dark red) and the number of basis points is set to $M = \left\lfloor d\sqrt{N} \right\rfloor$. Middle (b) Relative error as M increases with dimension d varying from $d = 1$ (yellow) to $d = 10$ (dark red) and N fixed to 5000. Right (c) Box-plot of the relative error as d increases with $N = 5000$ and $M = \left\lfloor d\sqrt{N} \right\rfloor$.

	Kernel size	Output shape
z		$32\times32\times3$
Conv	3×3	64
Residual block	$[3\times3]\times2$	64
Residual block	$[3\times3]\times2$	128
Residual block	$[3\times3]\times2$	256
Residual block	$[3\times3]\times2$	512
Linear	-	Number of classes

Table 1: Network architecture.

D.2 Classification on Cifar10 and Cifar100

Architecture. We use a residual network with one convolutional layer followed by 8 residual blocks and a final fully connected layer. Each residual block consists of two 3×3 convolutional layers each and ReLU nonlinearity. We use batch normalization for all methods. Details of the intermediate output shapes and kernel size are provided in Table 1.

Hyper-parameters. For all methods, we used a batch-size of 128. The optimal step-size γ was selected in $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ for each method. In the case of SGD with momentum, we used a momentum parameter of 0.9 and a weight decay of either 0 or 5×10^{-4}. For KFAC and EKFAC, we used a damping coefficient of 10^{-3} and a frequency of reparametrization of 100 updates. For KWGN we set $M = 5$, $\lambda = 10^{-5}$ and adjust ϵ using an adaptive scheme based on the Levenberg-Marquardt dynamics as in (Martens and Grosse, 2015, Section 6.5). More precisely, we use the following update equation for ϵ after every 5 iterations of the optimizer:

$$\begin{aligned} &\epsilon \leftarrow \omega\epsilon, && \text{if } r > \frac{3}{4} \\ &\epsilon \leftarrow \omega^{-1}\epsilon, && \text{if } r < \frac{1}{4}. \end{aligned}$$

Here, r is the reduction ratio: $r = 2\frac{\mathcal{L}(\theta_t) - \mathcal{L}(\theta_{t+1})}{\nabla^W\mathcal{L}(\theta)^\top\nabla\mathcal{L}(\theta)^\top}$ and ω is the decay constant chosen to $\omega = 0.85$.