WASSERSTEIN INFORMATION MATRIX

WUCHEN LI AND JIAXI ZHAO

Abstract. We study information matrices for statistical models by the L^2-Wasserstein metric. We call them Wasserstein information matrices (WIMs), which are analogs of classical Fisher information matrices. We introduce Wasserstein score functions and study covariance operators in statistical models. Using them, we establish Wasserstein-Cramer-Rao bounds for estimations and explore their comparisons with classical results. We next consider the asymptotic behaviors and efficiency of estimators. We derive the online asymptotic efficiency for Wasserstein natural gradient. Besides, we study a Poincaré efficiency for Wasserstein natural gradient of maximal likelihood estimation. Several analytical examples of WIMs are presented, including location-scale families, independent families, and rectified linear unit (ReLU) generative models.

1. Introduction

Fisher information matrix plays essential roles in statistics, physics, and differential geometry with applications in machine learning [1, 2, 5, 8, 10]. In statistics, it is a fundamental quantity for the estimation theory, including both design and analysis of estimators. In particular, the maximal likelihood principle is a well-known example. It connects the Fisher information matrix to another concept, named score functions. They frequently arise in statistical efficiency and sufficiency problems, especially for Cramer-Rao bound and Fisher-efficiency.

Fisher information matrix is also named Fisher-Rao metric in information geometry [3]. It uses the Fisher information matrix to study divergence functions and their invariance properties [3]. Furthermore, the Fisher information matrix is also useful for statistical learning problems. In particular, the natural gradient method [2] rectifies the gradient direction by the Fisher information matrix. It is shown that the Fisher natural gradient method is asymptotically online Fisher-efficient.

On the other hand, optimal transport introduces the other metric in probability space [23, 24], often named Wasserstein metric [11, 20]. Different from information geometry, it encodes the geometry of sample space into the definition of metric in probability space. Nowadays, it is known that the Wasserstein metric intrinsically connects the Kullback-Leibler (KL) divergence with Fisher information functional [20], known as de Bruijn identities [26]. Many concentration inequalities such as log-Sobolev inequalities and Poincaré inequalities arise naturally [21].

Key words and phrases. Wasserstein information matrix; Wasserstein score function; Wasserstein-Cramer-Rao inequality; Wasserstein online efficiency; Poincaré efficiency.

Despite various studies of optimal transport in full probability space, not much is known in parametric statistical models, which play crucial roles in parametric statistics. Fundamental questions arise: *Is there a statistical theory based on optimal transport? Compared to Fisher information matrices and Fisher statistics, what are counterparts of information matrices, score functions, Cramer-Rao bounds, and online efficiencies of natural gradient methods in Wasserstein statistics? Moreover, can this theory provide statistical tools for machine learning models, especially for generative models?*

In this paper, following key ideas in [12], we positively answer the above questions by introducing a Wasserstein information matrix (WIM). We derive the WIM by pulling back the Wasserstein metric from full probability space to finite-dimensional parametric statistical models [15, 16]. We show that the WIM defines Wasserstein score functions with a Wasserstein covariance operator of estimators. Based on them, a Wasserstein-Cramer-Rao bound is derived. Furthermore, combining WIM with Wasserstein score functions, we recover an asymptotic efficiency property of the online Wasserstein natural gradient methods.

Meanwhile, by comparing both Wasserstein and Fisher information matrices, we naturally prove several concentration inequalities such as log-Sobolev inequalities and Poincaré inequalities within statistical models. Extending the study in full probability space, we further decompose a Hessian term and study the Ricci-Information-Wasserstein (RIW) criterion for log-Sobolev inequalities and Poincaré inequalities in statistical models. Here we provide several examples in analytic probability families. Those functional inequalities turn out to be essential in a new efficiency property named Poincaré efficiency. This is concerned with dynamics where the Wasserstein natural gradient works on Fisher score functions (related to maximal likelihood estimators). We prove convergence rate analysis for these dynamics. Several numerical experiments are provided to confirm our conclusions.

Lastly, we demonstrate that the WIM provides a clear statistical theory for complicated models coming from machine learning approaches, especially implicit generative models. For example, we carefully study a one-dimensional probability family generated by push-forward maps based on the ReLU function. We demonstrate that the WIM still exists in this family while the classical Fisher information matrix does not exist. In other words, it is suitable to introduce a statistical theory based on WIMs. It can be a theoretical background for machine learning implicit models.

In literature, there have been lots of works attempting to use tools from optimal transport and information geometry to study statistical problems. [6] designs new estimators for parametric inference using Wasserstein distance. This idea is utilized in approximating Bayesian computation. The authors apply Wasserstein distance to measure the similarity between synthetic and observed data sets. Compared to them, we focus on the study of estimation and efficiency of WIMs. We expect it could have potential properties in Wasserstein estimators. In [7], they design a generalized information matrix based on a maximum mean discrepancy. Compared to them, we majorly focus on information matrices generated by the Wasserstein metric and study related statistical properties. Most closely, [22] defines a Wasserstein covariance by applying a closed-form formula for one-dimensional Wasserstein metric. This is a canonical definition. Our approach further extends this

idea into general parametric models. We start by introducing the WIM in parametric models. Using it, we define Wasserstein score functions as well as the Wasserstein covariance operator. We further establish the Wasserstein-Cramer-Rao bound and associated statistical efficiency properties. Also, [25] defines several new divergence functions by combining knowledge from both optimal transport and information geometry. Here we focus on statistical properties of WIMs in statistical models. Furthermore, Wasserstein natural gradient method has been widely studied in optimization techniques with machine learning applications [4, 9, 18, 13, 17]. Here we focus on statistical theory and study its associated online efficiency. Compared with classical online Fisher-efficiency results in [2, 19], our results can deal with general information matrices. In particular, for WIMs, we discover a new efficiency property named Poincaré efficiency. It relies on a comparison between Wasserstein and Fisher information matrices, demonstrate its connection with Poincaré inequalities.

The paper is organized as follows. In section 2, we establish the definition of the WIM. We present it analytically for several well-known probability families. We provide an explicit example of WIMs for ReLU generative models. Under this model, we show that the WIM exists while the Fisher information matrix does not exist. In section 3, with the introduction of the Wasserstein covariance, the Wasserstein-Cramer-Rao inequality is established. In section 4, we introduce and discuss both Wasserstein efficiency and Poincaré efficiency.

Probability Family	Wasserstein information matrix	Fisher information matrix
Uniform: $p(x;a,b)=\frac{1}{b-a}\mathbf{1}_{(a,b)}(x)$	$G_W(a,b)=\frac{1}{3}\begin{pmatrix}1 & \frac{1}{2}\\ \frac{1}{2} & 1\end{pmatrix}$	$G_F(a,b)$ not well-defined
Gaussian: $p(x;\mu,\sigma)=\frac{e^{-\frac{1}{2\sigma^2}(x-\mu)^2}}{\sqrt{2\pi}\sigma}$	$G_W(\mu,\sigma)=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$	$G_F(\mu,\sigma)=\begin{pmatrix}\frac{1}{\sigma^2} & 0\\ 0 & \frac{2}{\sigma^2}\end{pmatrix}$
Exponential: $p(x;m,\lambda)=\lambda e^{-\lambda(x-m)}$	$G_W(m,\lambda)=\begin{pmatrix}1 & \frac{1}{\lambda^2}\\ \frac{1}{\lambda^2} & \frac{2}{\lambda^4}\end{pmatrix}$	$G_F(m,\lambda)$ not well-defined
Laplacian: $p(x;m,\lambda)=\frac{\lambda}{2}e^{-\lambda\lvert x-m\rvert}$	$G_W(m,\lambda)=\begin{pmatrix}1 & 0\\ 0 & \frac{2}{\lambda^4}\end{pmatrix}$	$G_F(m,\lambda)=\begin{pmatrix}\lambda^2 & 0\\ 0 & \frac{1}{\lambda^2}\end{pmatrix}$
Location-scale: $p(x;m,\lambda)=\frac{1}{\lambda}p(\frac{x-p}{\lambda})$	$G_W(\lambda,m)=\begin{pmatrix}\frac{\mathbb{E}_{\lambda,m}x^2-2m\mathbb{E}_{\lambda,m}x+m^2}{\lambda^2} & 0\\ 0 & 1\end{pmatrix}$	$G_F(\lambda,m)=\begin{pmatrix}\frac{1}{\lambda^2}\left(1+\int_{\mathbb{R}}\left(\frac{(x-m)^2p'^2}{\lambda^2p}+\frac{(x-m)p'}{\lambda}\right)dx\right) & \int_{\mathbb{R}}\frac{(x-m)p'^2}{\lambda^3p}dx\\ \int_{\mathbb{R}}\frac{(x-m)p'^2}{\lambda^3p}dx & \frac{1}{\lambda^2}\int_{\mathbb{R}}\frac{p'^2}{p}dx\end{pmatrix}$
Independent: $p(x,y;\theta)=p(x;\theta)p(y;\theta)$	$G_W(x,y;\theta)=G_W^1(x;\theta)+G_W^2(y;\theta)$	$G_F(x,y;\theta)=G_F^1(x;\theta)+G_F^2(y;\theta)$
ReLU push-forward: $p(x;\theta)=f_{\theta*}p(x)$, f_θ θ-parameterized ReLUs, Ex. 8.	$G_W(\theta)=F(\theta)$, F cdf of $p(x)$, $F(y)=\int_{-\infty}^{y}p(x)dx$.	$G_F(\theta)$ not well-defined

TABLE 1. In this table, we present Wasserstein, Fisher information matrices for probability families.

2. Wasserstein information matrix and score functions

In this section, we present Wasserstein information matrices (WIMs) and score functions. Several analytical studies are presented.

Given a sample space $\mathcal{X}\subset\mathbb{R}^n$, let $\mathcal{P}(\mathcal{X})$ denote the space of probability distributions over $\mathcal{X}$. Given a metric tensor g on $\mathcal{P}(\mathcal{X})$, we call $(\mathcal{P}(\mathcal{X}),g)$ density manifold. Consider a

Family	Entropy functional	Fisher-information functional	Log-Sobolev inequality(LSI(α))
Gaussian	$\widetilde{H}(p_{\mu,\sigma}) = -\frac{1}{2}\log 2\pi - \log\sigma - \frac{1}{2}$, $\widetilde{H}(p_{\mu,\sigma}\|p_{\mu_*,\sigma_*}) = -\log\sigma + \log\sigma_* - \frac{1}{2} + \frac{\sigma^2+(\mu-\mu_*)^2}{2\sigma_*^2}$.	$\widetilde{I}(p_{\mu,\sigma}) = \frac{1}{\sigma^2}$, $\widetilde{I}(p_{\mu,\sigma}\|p_{\mu_*,\sigma_*}) = \frac{(\mu-\mu_*)^2}{4\sigma_*^4} + \left(-\frac{1}{\sigma}+\frac{\sigma}{\sigma_*^2}\right)^2$.	$\widetilde{H}(p_{\mu,\sigma}\|p_{\mu_*,\sigma_*}) < \frac{1}{2\alpha}\widetilde{I}(p_{\mu,\sigma}\|p_{\mu_*,\sigma_*})$, $\mu,\sigma>0$.
Laplacian	$\widetilde{H}(p_{m,\lambda}) = -1+\log\lambda-\log 2$, $\widetilde{H}(p_{m,\lambda}\|p_{m_*,\lambda_*}) = -1+\log\lambda-\log\lambda_* + \lambda_*\|m-m_*\| + \frac{\lambda_* e^{-\lambda\|m-m_*\|}}{\lambda}$.	$\widetilde{I}(p_{\lambda,m}) = \frac{\lambda^2}{2}$, $\widetilde{I}(p_{m,\lambda}\|p_{m_*,\lambda_*}) = \lambda_*^2\left(1-e^{-\lambda\|m-m_*\|}\right)^2$ $+\frac{\left((\lambda\|m-m_*\|+1)\lambda_* e^{-\lambda\|m-m_*\|}-\lambda\right)^2}{2}$.	$\widetilde{H}(p_{\lambda,m}\|p_{\lambda_*,m_*}) < \frac{1}{2\alpha}\widetilde{I}(p_{\lambda,m}\|p_{\lambda_*,m_*})$, $m\in\mathbb{R},\lambda>0$.

TABLE 2. In this table, we continue to list the entropy functional, the Fisher information functionals, log-Sobolev inequalities for probability families.

parameter space $\Theta \subset \mathbb{R}^d$ and a parameterization function

$$p\colon \Theta \to \mathcal{P}(\mathcal{X}), \quad \theta \mapsto p_\theta$$

which can also be viewed as $p\colon \mathcal{X}\times\Theta\to\mathbb{R}$. Here Θ is named a statistical model. Denote $\langle f,h\rangle = \int_{\mathcal{X}} f(x)h(x)dx$ for the $L^2(\mathcal{X})$ inner product, where dx refers to the Lebesgue measure on $\mathcal{X}$. And we denote by $(v,w) = v\cdot w$ the (pointwise) Euclidean inner product of two vectors.

2.1. **Information matrix.** We first review metric tensors on parameter space and connect them with information matrices.

Definition 1 (Statistical information matrix)**.** *Consider the density manifold $(\mathcal{P}(\mathcal{X}), g)$ with a metric tensor g, and a smoothly parametrized statistical model p_θ with parameter $\theta\in\Theta\subset\mathbb{R}^d$. Then the pull-back metric $G\in\mathbb{R}^{d\times d}$ of g onto this parameter space Θ is given by*

$$G(\theta) = \left\langle \nabla_\theta p_\theta, g(p_\theta)\nabla_\theta p_\theta \right\rangle.$$

Denote $G(\theta) = (G(\theta)_{ij})_{1\le i,j\le d}$, then

$$G(\theta)_{ij} = \int_{\mathcal{X}} \frac{\partial}{\partial\theta_i}p(x;\theta)\Big(g(p_\theta)\frac{\partial}{\partial\theta_j}p\Big)(x;\theta)dx.$$

Here we name g statistical metric, and call G statistical information matrix.

In geometry, using this metric tensor g to rise(resp. lower) indices, there exists a canonical isomorphism from tangent(resp. cotangent) space to cotangent(resp. tangent) space, namely:

$$g(p) : T_p\mathcal{P}(\mathcal{X}) \simeq T_p^*\mathcal{P}(\mathcal{X}), \qquad f\mapsto [g(p)(f)],$$
$$g(p)^{-1} : T_p^*\mathcal{P}(\mathcal{X}) \simeq T_p\mathcal{P}(\mathcal{X}), \qquad [f]\mapsto g(p)^{-1}(f).$$

Thus the metric tensor can actually be viewed as an operator between these two spaces. The above tangent space $T_p\mathcal{P}(\mathcal{X})$ is identified with the function space:

$$T_p\mathcal{P}(\mathcal{X}) \simeq C_0(\mathcal{X}) = \{f\in C(\mathcal{X}) | \int_{\mathcal{X}} f dx = 0\},$$

where $C(\mathcal{X})$ is the function space of continuous function on the space $\mathcal{X}$. And its dual space $C(\mathcal{X})/\mathbb{R}$, i.e. $f,g\in C(\mathcal{X}), f\sim g$ if $f = g+a, a\in\mathbb{R}$, can be identified with the

cotangent space of the density manifold: $T_p^*\mathcal{P}(\mathcal{X}) \simeq C(\mathcal{X})/\mathbb{R}$. We use $[f]$ to represent the equivalent class of this function in $C(\mathcal{X})/\mathbb{R}$. And the pairing between tangent spaces and cotangent spaces is merely $\langle f, h\rangle = \int_{\mathcal{X}} fhdx$, where we abuse the symbol $\langle, \rangle$ for the inner product. We note here that tangent spaces of a statistical model Θ can be viewed as subspaces of that of density manifold, i.e. we have inclusion:

$$T_p\Theta \hookrightarrow T_p\mathcal{P}(\mathcal{X}).$$

While taking the dual of this inclusion we get projection from the cotangent space of the density manifold to that of the statistical model:

$$T_p^*\mathcal{P}(\mathcal{X}) \to T_p^*\Theta.$$

An approach in information geometry is that one can reinterpret the metric tensor in the dual coordinates, i.e. cotangent space.

Definition 2 (Score function)**.** *Denote $\Phi_i : \mathcal{X} \times \Theta \to \mathbb{R}, i = 1, ..., n$ satisfying*

$$\Phi_i(x;\theta) = \left[g(p)\left(\frac{\partial}{\partial\theta_i}p(x;\theta)\right)\right].$$

We call Φ_is score functions associated with the statistical information matrix G and are equivalent classes in $C(\mathcal{X})/\mathbb{R}$. The representatives in equivalent classes are determined by the following normalization condition:

$$\mathbb{E}_{p_\theta}\Phi_i = 0, \qquad i = 1, ..., n. \tag{1}$$

Then the statistical information matrix satisfies

$$G(\theta)_{ij} = \int_{\mathcal{X}} \Phi_i(x;\theta)\Big(g(p_\theta)^{-1}\Phi_j\Big)(x;\theta)dx.$$

Remark 1. The normalization condition is an enforced condition. It fixes a representative for the score function in the equivalent class. And we assume that score functions are always integrable w.r.t. p_θ.

In above, there are two formulations of metric tensor, which use the following fact $g(p)^{-1} = g(p)^{-1}g(p)g(p)^{-1}$. Thus

$$\begin{aligned} G(\theta)_{ij} &= \Big\langle \nabla_{\theta_i}p_\theta, g(p_\theta)\nabla_{\theta_j}p_\theta \Big\rangle \\ &= \Big\langle g(p_\theta)\nabla_{\theta_i}p_\theta, g(p_\theta)^{-1}g(p_\theta)\nabla_{\theta_j}p_\theta \Big\rangle \\ &= \Big\langle \Phi_i, g(p_\theta)^{-1}\Phi_j \Big\rangle. \end{aligned}$$

Example 1. One important choice of metric is the Fisher-Rao metric:

$$\begin{aligned} g(p) : T_p\mathcal{P}(\mathcal{X}) \simeq T_p^*\mathcal{P}(\mathcal{X}), \qquad f \mapsto \left[\frac{f}{p}\right], \\ g(p)^{-1} : T_p^*\mathcal{P}(\mathcal{X}) \simeq T_p\mathcal{P}(\mathcal{X}), \qquad [f] \mapsto p\left(f - E_p f\right). \end{aligned}$$

In this case, the statistical information matrix satisfies

$$G_F(\theta)_{ij} = \langle \frac{\partial}{\partial\theta_i}p_\theta, \frac{1}{p_\theta}\frac{\partial}{\partial\theta_j}p_\theta\rangle = \int_{\mathcal{X}} \frac{\frac{\partial}{\partial\theta_i}p(x;\theta)\frac{\partial}{\partial\theta_j}p(x;\theta)}{p(x;\theta)}dx.$$

And score functions of Fisher information matrix form

$$\Phi_i^F(x;\theta) = \frac{1}{p(x;\theta)}\frac{\partial}{\partial\theta_i}p(x;\theta) = \frac{\partial}{\partial\theta_i}\log p(x;\theta),$$

where the normalization condition holds automatically. In terms of score functions, the Fisher information matrix forms

$$\begin{aligned} G_F(\theta)_{ij} &= \int_{\mathcal{X}} \Phi_i^F(x;\theta)\Big(g_F(p)^{-1}\Phi_j^F\Big)(x;\theta)dx \\ &= \int_{\mathcal{X}} \frac{\partial}{\partial\theta_i}\log p(x;\theta)\frac{\partial}{\partial\theta_j}\log p(x;\theta)p(x;\theta)dx \\ &= \mathbb{E}_{p_\theta}\left(\frac{\partial}{\partial\theta_i}\log p(x;\theta)\frac{\partial}{\partial\theta_j}\log p(x;\theta)\right). \end{aligned}$$

In literature, $\Phi_i^F(x;\theta) = \frac{\partial}{\partial\theta_i}\log p(x;\theta)$ is named (Fisher) *score function*; while $G_F(\theta)$ is the *Fisher information matrix*. They play important roles in estimation, efficiency and Cramer-Rao bound.

Remark 2. The definition of Fisher score functions can be given in classical statistics as the gradient of the log-likelihood function w.r.t. parameters. Here we view it as an object on cotangent space associated with the Fisher-Rao metric on statistical models. That is, we have a family of canonical tangent vector fields $\frac{\partial}{\partial\theta_i}p_\theta$ on statistical models. Whenever there is a metric $g(p_\theta)$ on this manifold, we can define score functions associated with it as:

$$\Phi_i = g(p_\theta)\frac{\partial}{\partial\theta_i}p_\theta.$$

From above fact, we observe that statistical concepts are related to the metric tensor in density manifold pull-back onto parameter space. In particular, classical statistics relates to the Fisher-Rao metric. The pull-back metric tensor forms an information matrix while dual variables define score functions. In this paper, we derive these notations in the other important statistical metric, known as the Wasserstein metric.

2.2. **Wasserstein information matrix.** The other statistical metric, namely Wasserstein metric tensor forms

$$g_W(p) = (-\Delta_p)^{-1}, \quad \text{where } \Delta_p = \nabla\cdot(p\nabla).$$

Here Δ_p is an elliptic operator weighted on a probability density p. When p satisfies suitable conditions, standard PDE theory guarantees that the operators Δ_p^{-1} and Δ_p are an inverse to each other between function spaces:

$$\begin{aligned} \Delta_p^{-1} &: C_0(\mathcal{X}) \to C(\mathcal{X})/\mathbb{R}; \\ \Delta_p &: C(\mathcal{X})/\mathbb{R} \to C_0(\mathcal{X}). \end{aligned}$$

The pull-back G_W of g_W is given by

$$G_W(\theta)_{ij} = \langle\frac{\partial}{\partial\theta_i}p_\theta, (-\Delta_{p_\theta})^{-1}\frac{\partial}{\partial\theta_j}p_\theta\rangle.$$

Similar to the Fisher information matrix, we can rewrite G_W by dual coordinates. Denote

$$\Phi_i^W(x;\theta) = (-\Delta_{p_\theta})^{-1}\frac{\partial}{\partial\theta_i}p(x;\theta).$$

Then

$$\begin{aligned} G_W(\theta)_{ij} &= \langle \frac{\partial}{\partial\theta_i}p_\theta, (-\Delta_{p_\theta})^{-1}\frac{\partial}{\partial\theta_j}p_\theta\rangle \\ &= \langle \Phi_i^W, (-\Delta_{p_\theta})\Phi_j^W\rangle \\ &= \int_{\mathcal{X}} (\nabla_x\Phi_i^W(x;\theta), \nabla_x\Phi_j^W(x;\theta))p(x;\theta)dx, \end{aligned}$$

where the last equality holds by integration by parts w.r.t. x.

We summarize the above fact into the following definition.

Definition 3 (Wasserstein information matrix & score function)**.** *Denote $G_W(\theta) \in \mathbb{R}^{d\times d}$:*

$$G_W(\theta)_{ij} = \mathbb{E}_{p_\theta}\left[\nabla_x\Phi_i^W(x;\theta)\cdot\nabla_x\Phi_j^W(x;\theta)\right],$$

where $\cdot$ refers to the inner product of vector and $\Phi_i^W : \mathcal{X}\times\Theta\to\mathbb{R}$ satisfies

$$-\nabla_x\cdot(p(x;\theta)\nabla_x\Phi_i^W(x;\theta)) = \frac{\partial}{\partial\theta_i}p(x;\theta), \quad \mathbb{E}_{p_\theta}\Phi_i^W = 0, \quad i = 1,2,...,d.$$

We name functions $\Phi_i^W(x;\theta) = \left((-\Delta_{p_\theta})^{-1}\frac{\partial}{\partial\theta_i}p_\theta\right)(x;\theta)$ Wasserstein score functions*, and call the matrix $G_W(\theta)$ the* Wasserstein information matrix*.*

Remark 3. This definition of information matrices is motivated by an intrinsic connection among distances, divergence functions, and metrics. Specifically, given a smooth family of probability densities $p(x;\theta)$ and a given perturbation $\Delta\theta \in T_\theta\Theta$, consider following Taylor expansions in term of $\Delta\theta$:

$$\begin{aligned} H(p(\theta)\|p(\theta+\Delta\theta)) &= \frac{1}{2}\Delta\theta^{\mathsf{T}}G_F(\theta)\Delta\theta + o((\Delta\theta)^2), \\ W_2(p(\theta+\Delta\theta), p(\theta))^2 &= \Delta\theta^{\mathsf{T}}G_W(\theta)\Delta\theta + o((\Delta\theta)^2). \end{aligned} \tag{2}$$

Here H represents the Kullback–Leibler (KL) divergence or the relative entropy functional

$$H(p(\theta)\|p(\theta+\Delta\theta)) = \int_{\mathcal{X}} p(x;\theta)\log\frac{p(x;\theta)}{p(x;\theta+\Delta\theta)}dx.$$

While W_2^2 denotes the squared L^2-Wasserstein distance defined by

$$W_2(p(\theta), p(\theta+\Delta\theta))^2 = \inf_{\pi\in\Pi(p(\theta),p(\theta+\Delta\theta))}\left\{\int_{\mathcal{X}\times\mathcal{X}} d_{\mathcal{X}}(x,y)^2\, d\pi(x,y)\right\}, \tag{3}$$

where $\Pi(p(\theta), p(\theta+\Delta\theta))$ refers to the set of couplings between $p(\theta)$, $p(\theta+\Delta\theta)$ and $d_{\mathcal{X}}$ is a distance function defined in $\mathcal{X}$. Thus our approach parallels classical Fisher statistics. The Fisher information matrix approximates the KL divergence, which relates to the Fisher distance in Fisher geometry [1, 5], while WIM approximates the Wasserstein distance in Wasserstein geometry. Meanwhile, our approach can be viewed as exploring another aspect, namely metric aspect, of the Wasserstein statistics. For example, it can be related to the study of Wasserstein estimators [6].

We next study several basic properties of WIMs and score functions. We first illustrate a relation between Wasserstein and Fisher score functions.

Proposition 4 (Poisson equation)**.** *Wasserstein score functions $\Phi_i^W(x;\theta)$ satisfy the following Poisson equation*

$$\nabla_x \log p(x;\theta)\cdot\nabla_x\Phi_i^W(x;\theta)+\Delta_x\Phi_i^W(x;\theta)=-\frac{\partial}{\partial\theta_i}\log p(x;\theta). \tag{4}$$

Proof. Notice the fact that

$$\begin{aligned}\Big(\Delta_{p_\theta}\Big)\Phi^W(x;\theta)&=\nabla_x\cdot(p(x;\theta)\nabla_x\Phi^W(x;\theta))\\&=\nabla_x p(x;\theta)\cdot\nabla_x\Phi^W(x;\theta)+p(x;\theta)\Delta_x\Phi^W(x;\theta).\end{aligned}$$

Then the Wasserstein score function $\Phi_i^W(x)$ satisfies

$$\nabla_x p(x;\theta)\cdot\nabla_x\Phi^W(x;\theta)+p(x;\theta)\Delta_x\Phi^W(x;\theta)=-\frac{\partial}{\partial\theta_i}p(x;\theta).$$

Divide the above equation on both sides by $p(x;\theta)$:

$$\frac{1}{p(x;\theta)}\Big\{\nabla_x p(x;\theta)\cdot\nabla_x\Phi^W(x;\theta)+p(x;\theta)\Delta_x\Phi^W(x;\theta)\Big\}=-\frac{1}{p(x;\theta)}\frac{\partial}{\partial\theta_i}p(x;\theta),$$

i.e.

$$\frac{1}{p(x;\theta)}\nabla_x p(x;\theta)\cdot\nabla_x\Phi^W(x;\theta)+\Delta_x\Phi^W(x;\theta)=-\frac{1}{p(x;\theta)}\frac{\partial}{\partial\theta_i}p(x;\theta).$$

Since $\frac{1}{p(x;\theta)}\nabla_x p(x;\theta)=\nabla_x\log p(x;\theta)$ and $\frac{1}{p(x;\theta)}\frac{\partial}{\partial\theta_i}p(x;\theta)=\frac{\partial}{\partial\theta_i}\log p(x;\theta)$, we prove the property (4). □

We then demonstrate that Wasserstein score functions and information matrices can also be decomposed into a summation of separable functions in independent models.

Proposition 5 (Separability)**.** *If $p(x;\theta)$ is an independence model, i.e.*

$$p(x,\theta)=\Pi_{k=1}^n p_k(x_k;\theta),\quad x_k\in\mathcal{X}_k,\quad x=(x_1,\cdots,x_n).$$

Then there exists a set of one-dimensional functions $\Phi^{W,k}\colon\mathcal{X}_k\times\Theta_k\to\mathbb{R}$, such that

$$\Phi^W(x;\theta)=\sum_{k=1}^n\Phi^{W,k}(x_k;\theta). \tag{5}$$

In addition, the WIM is separable:

$$G_W(\theta)=\sum_{k=1}^n G_W^k(\theta),$$

where $G_W^k(\theta)=\mathbb{E}_{p_\theta}\left(\nabla_{x_k}\Phi^{W,k}(x;\theta),\nabla_{x_k}\Phi^{W,k}(x;\theta)\right)$.

Proof. The proof follows from proposition 4. Suppose one can write the solution in form of (5), then equation (4) forms

$$\sum_{k=1}^n\Big\{\nabla_{x_k}\log p_k(x_k;\theta_k)\nabla_{x_k}\Phi^{W,k}(x_k;\theta)+\Delta_{x_k}\Phi^{W,k}(x_k;\theta)-\frac{\partial}{\partial\theta_i}\log p_k(x_k;\theta)\Big\}=0.$$

From the separable method for solving the Poisson equation, we derive

$$\nabla_{x_k} \log p_k(x_k;\theta_k)\nabla_{x_k}\Phi^{W,k}(x_k;\theta) + \Delta_{x_k}\Phi^{W,k}(x_k;\theta_k) - \frac{\partial}{\partial\theta_i}\log p_k(x_k;\theta) = 0.$$

We finish the first part of the proof. In addition,

$$\begin{aligned} G_W(\theta) &= \mathbb{E}_{p_\theta}\left(\nabla_x\Phi^W(x;\theta), \nabla_x\Phi^W(x;\theta)\right) \\ &= \mathbb{E}_{p_\theta}\left(\sum_k (\nabla_{x_k}\Phi^{W,k}(x;\theta), \nabla_{x_k}\Phi^{W,k}(x;\theta)\right) \\ &= \sum_k \mathbb{E}_{p_\theta}\left(\nabla_{x_k}\Phi^{W,k}(x;\theta), \nabla_{x_k}\Phi^{W,k}(x;\theta)\right) \\ &= \sum_k G_W^k(\theta). \end{aligned}$$

□

We next list some analytical solutions for WIMs and score functions in 1-d case. See related studies in [23] (c.f. Ch 2.2).

Proposition 6 (One-dimensional sample space)**.** *If $\mathcal{X} \subset \mathbb{R}^1$, Wasserstein score functions satisfy*

$$\Phi_i^W(x;\theta) = -\int_{\mathcal{X}\cap(\infty,x]} \frac{1}{p(z;\theta)}\frac{\partial}{\partial\theta_i}F(z;\theta)dz, \tag{6}$$

where $F(x;\theta) = \int_{\mathcal{X}\cap(\infty,x]} p(y;\theta)dy$ is the cumulative distribution function. And the WIM satisfies

$$G_W(\theta)_{ij} = \mathbb{E}_{p_\theta}\left(\frac{\frac{\partial}{\partial\theta_i}F(x;\theta)\frac{\partial}{\partial\theta_j}F(x;\theta)}{p(x;\theta)^2}\right).$$

If the dimension of sample space $\mathcal{X}$ is larger than 1, exact solutions of Wasserstein score functions and information matrices depend on solutions of Poisson equation (4). We leave the derivation of general formulas for interested readers.

2.3. **Analytic examples.** We present several analytical examples of the WIM in one-dimensional sample space. The derivation is given in section A.

Example 2 (Gaussian distribution)**.** Consider Gaussian distribution families with mean value μ and standard variance $\sigma > 0$, i.e. $p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$. Wasserstein score functions satisfy

$$\Phi_\mu^W(x;\mu,\sigma) = x-\mu, \quad \Phi_\sigma^W(x;\mu,\sigma) = \frac{(x-\mu)^2-\sigma^2}{2\sigma}.$$

And the WIM satisfies

$$G_W(\mu,\sigma) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Example 3 (Exponential distribution)**.** Consider exponential distribution families $Exp(m, \lambda)$, i.e. $p(x; m, \lambda) = \mathbf{1}_{[m,\infty)}(x)\lambda e^{-\lambda(x-m)}$, where the function $\mathbf{1}_C$ is the indicator function for a set $C \subset \mathbb{R}$. Wasserstein score functions satisfy

$$\Phi_\lambda^W(x; m, \lambda) = \frac{(x-m)^2 - \frac{2}{\lambda^2}}{2\lambda}, \qquad \Phi_m^W(x; m, \lambda) = x - m - \frac{1}{\lambda}.$$

And the WIM satisfies

$$G_W(m, \lambda) = \begin{pmatrix} 1 & \frac{1}{\lambda^2} \\ \frac{1}{\lambda^2} & \frac{2}{\lambda^4} \end{pmatrix}.$$

Example 4 (Laplacian distribution)**.** Consider Laplacian distribution families $La(m, \lambda)$, i.e. $p(x; m, \lambda) = \frac{\lambda}{2}e^{-\lambda|x-m|}$. Wasserstein score functions satisfy

$$\Phi_\lambda^W(x; m, \lambda) = \frac{(x-m)^2 - \frac{2}{\lambda^2}}{2\lambda}, \qquad \Phi_m^W(x; m, \lambda) = x - m.$$

Notice that score functions for exponential families and Laplacian families have similar formulas. And the WIM satisfies

$$G_W(m, \lambda) = \begin{pmatrix} 1 & 0 \\ 0 & \frac{2}{\lambda^4} \end{pmatrix}.$$

We will show below that the Laplacian family has an advantage that densities within this family have the same support. Thus this model is convenient for us to compare the WIM with the Fisher information matrix. See details in section B.3.

Example 5 (Uniform distribution)**.** Consider uniform distribution families within interval $[a, b]$, i.e. $p(x; a, b) = \frac{1}{b-a}\mathbf{1}_{[a,b]}(x)$. Wasserstein score functions satisfy

$$\Phi_a^W(x; a, b) = \frac{x(a+b-x)}{(b-a)} - \frac{b^2 + a^2 + 4ab}{6},$$
$$\Phi_b^W(x; a, b) = \frac{b(x-2a)}{(b-a)} - \frac{b^2 - 3ab}{2}.$$

And the WIM satisfies

$$G_W(a, b) = \frac{1}{3}\begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}.$$

Example 6 (Wigner semicircle distribution)**.** Consider semicircle distribution families, i.e. $p(x; m, R) = \mathbf{1}_{[-R+m,R+m]}(x)\frac{2}{\pi R^2}\sqrt{R^2 - (x-m)^2}$. Wasserstein score functions satisfy

$$\Phi_R^W(x; m, R) = \frac{1}{R}\left(\frac{(x-m)^2}{2} - \frac{R^2}{8}\right), \qquad \Phi_p^W(x; m, R) = x - m.$$

And the WIM satisfies

$$G_W(m, R) = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{4} \end{pmatrix}.$$

Example 7 (Independent model)**.** Consider an independent model as follow: suppose $X \sim p_1(x; \theta)$, and $Y \sim p_2(x; \theta)$, and $(X, Y) \sim p(x, y; \theta)$, then

$$p(x, y; \theta) = p_1(x; \theta)p_2(y; \theta).$$

Denote Wasserstein score functions (resp. WIM) for statistical model $X \sim p_1(x;\theta), Y \sim p_2(x;\theta)$ as $\Phi_1^W(x;\theta), \Phi_2^W(x;\theta)(G_W^1(x;\theta), G_W^2(x;\theta))$ respectively. Then, Wasserstein score functions for this model $(X,Y) \sim p(x,y;\theta)$ satisfy

$$\Phi^W(x,y;\theta) = \Phi_1^W(x;\theta) + \Phi_2^W(y;\theta),$$

because of the additivity of expectation $\mathbb{E}_{p_\theta}\Phi^W(x,y;\theta) = \mathbb{E}_{p_\theta}\Phi_1^W(x;\theta) + \mathbb{E}_{p_\theta}\Phi_2^W(y;\theta) = 0$. And the WIM satisfies

$$G_W(x,y;\theta) = G_W^1(x;\theta) + G_W^2(y;\theta).$$

The proof follows directly from proposition 5.

In above discussions, all examples are based on location-scale families, which will be derived carefully in section A.2. We show that location-scale families are totally geodesic submanifolds in Wasserstein geometry.

2.4. **WIM in generative models.** In this section, we study the WIM for generative models using ReLU function, which is given by

$$\sigma(x) = \begin{cases} 0, & x \le 0, \\ x, & x > 0. \end{cases}$$

Generative models are powerful in machine learning [14]. It applies the reparameterization trick (known as push-forward relation) to conduct efficient sampling. In practice, one often applies the ReLU as a push-forward function (7). For this reason, we call this kind of models ReLU push-forward family. The push-forward measure f_*p is defined as

$$\int_A f_* p dx = \int_{f^{-1}(A)} p dx, \quad \forall A \subset \mathbb{R}. \tag{7}$$

To keep derivations simple, we consider one-dimensional cases with a given distribution $p_0(x)$, $x \in \mathbb{R}$. And its cumulative distribution function is denoted by $F_0(x)$.

Example 8 (ReLU push-forward family)**.** We use a family of ReLU functions f_θ parameterized by θ to generate a push-forward family

$$p : \Theta \simeq \mathbb{R} \to \mathcal{P}(\mathbb{R}) : \quad \theta \mapsto p_\theta,$$

$$p_\theta(x) = p(x;\theta) = (f_{\theta *} p_0)(x), \quad f_\theta(x) = \sigma(x-\theta) = \begin{cases} 0, & x \le \theta, \\ x-\theta, & x > \theta. \end{cases}$$

The WIM of p_θ satisfies

$$G_W(\theta) = 1 - F_0(\theta). \tag{8}$$

We can also consider another family of ReLU maps to push forward the source distribution. This family is given by

$$p : \Theta \simeq \mathbb{R} \to \mathcal{P}(\mathbb{R}) : \quad \theta \mapsto p_\theta$$

$$p_\theta(x) = p(x;\theta) = (h_{\theta *} p_0)(x), \qquad h_\theta(x) = \sigma(x-\theta) + \theta = \begin{cases} \theta, & x \le \theta, \\ x, & x > \theta. \end{cases}$$

The WIM of p_θ satisfies

$$G_W(\theta) = F_0(\theta). \tag{9}$$

A figure illustrating these two families is provided below.

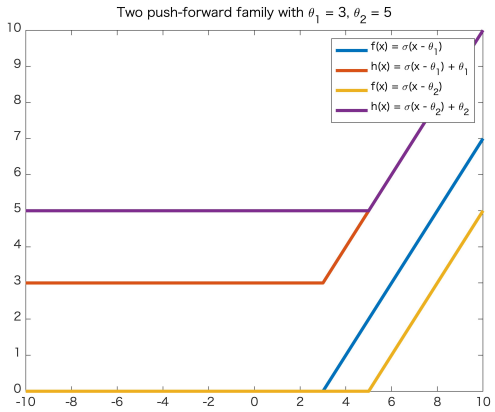

FIGURE 1. This figure plots two examples of push-forward families we described above with parameters chosen as $\theta_1 = 3, \theta_2 = 5$.

Remark 4. To calculate the WIMs of this model, we cannot use previous approaches of score functions, since it is not smooth enough. Instead, we utilize the idea stated in remark 3. Namely, we use the relation (2) between Wasserstein distance and the WIM to compute the latter.

Proof. Consider the following two push-forward distributions given by

$$\begin{aligned}\left(f_{\theta+\Delta\theta *}p_0\right)(x) &= F_0\left(\theta+\Delta\theta\right)\delta_0 + p_0\left(\cdot+\theta+\Delta\theta\right)_{[0,\infty)},\\ \left(f_{\theta *}p_0\right)(x) &= F_0\left(\theta\right)\delta_0 + p_0\left(\cdot+\theta\right)_{[0,\infty)},\end{aligned}$$

where δ_0 refers to the Dirac measure concentrating at point 0. And $p_0\left(\cdot+\theta\right)_{[0,\infty)}$ represents the measure $\widetilde{p}\left(x\right) = p_0\left(x+\theta\right)$ restricting to the interval $[0,\infty)$. Using monotonicity of transportation plan in 1-d, we conclude that its restriction on $(0,\infty)$ transports measure on x to $x+\Delta\theta$. And it remains to transport the Dirac measure centered at 0 to the remained place. The transportation cost is given by

$$\begin{aligned}W_2^2\left(f_{\theta *}p_0, f_{\theta+\Delta\theta *}p_0\right) &= \int_0^\infty p_0\left(x+\theta+\Delta\theta\right)\left(\Delta\theta\right)^2 dx + \int_0^{\Delta\theta} x^2 p_0\left(x+\theta\right)dx\\ &= \left(\Delta\theta\right)^2\left(1-F_0\left(\theta+\Delta\theta\right)\right) + O\left(\left(\Delta\theta\right)^3\right),\end{aligned} \tag{10}$$

where the third equality holds by

$$\int_0^{\Delta\theta} p_0\left(x+\theta\right)dx = O\left(\Delta\theta\right).$$

Notice in formula (10), we decompose the transportation cost into two parts: the first one is concerned with the cost on the right part of 0, while the second one considers transporting Dirac measure at 0 to the remained part. Since the WIM is an infinitesimal approximation of the Wasserstein distance, i.e. equation (2). The conclusion (8) follows.

For the other family, derivations follow the same method as before. Specifically, we have

$$\begin{aligned} W_2^2\left(h_{\theta*}p_0, h_{\theta+\Delta\theta*}p_0\right) = & \int_0^{\Delta\theta} x^2 p_0\left(x+\theta\right) dx + \left(\Delta\theta\right)^2 F_0\left(\theta\right) \\ & = \left(\Delta\theta\right)^2 F_0\left(\theta\right) + O\left(\left(\Delta\theta\right)^3\right), \end{aligned}$$

where we again decompose the transportation cost into two parts. The first one is absolutely continuous w.r.t the Lebesgue measure, while the second one contains a Dirac measure. □

Here we notice that density functions in ReLU push-forward family can be singular. Thus the Fisher information matrix, which depends on an explicit formula of density functions, namely

$$G_F(\theta)_{ij} = \int_{\mathcal{X}} \frac{\partial}{\partial\theta_i} \log p(x;\theta) \frac{\partial}{\partial\theta_j} \log p(x;\theta) p(x;\theta) dx$$

fails to exist in these models. On the contrary, as we have shown in the above example, the WIM still exists. This property shows that the WIM can provide statistical studies for generative models, while the Fisher information matrix in classical statistics can not.

3. Wasserstein estimation

In this section, we define the Wasserstein covariance and establish the Wasserstein-Cramer-Rao bound. Based on these concepts, we introduce a notion of efficiency in Wasserstein statistics. Several examples based on the previous section are provided.

3.1. **Estimation and efficiency.**

Following the spirit under which we introduce information matrices in section 2, we generalize the definition of covariance matrix for a given metric tensor g on probability space.

Denote $\langle f, h\rangle_g$ the inner product of cotangent vectors f, h in the metric g.

$$\langle f, h\rangle_g = \langle f, g(p)^{-1} h\rangle.$$

Definition 7 (Information covariance matrix)**.** *Given a statistical model Θ with metric g, and statistics T, $\widetilde{T}$ which are of dimension m, n respectively, the information covariance matrix of T,$\widetilde{T}$ associated to metric g is defined as:*

$$\mathrm{Cov}_\theta^g[T, \widetilde{T}]_{ij} = \langle T_i, \widetilde{T}_j\rangle_g,$$

where T_i, $\widetilde{T}_j$ are random variables as function of x. Denote the information variance as:

$$\mathrm{Var}_\theta^g[T] = \langle T, T\rangle_g.$$

Remark 5. Here the inner product $\langle T_i, \widetilde{T}_j\rangle_g$ is obtained by viewing the statistics as cotangent vectors on density manifold $\mathcal{P}(\mathcal{X})$.

Example 9 (Fisher covariance)**.** Given two statistics T_1, T_2, we view them as cotangent vectors in space $C(\mathcal{X})/\mathbb{R}$. Hence their Fisher inner product is defined as

$$\langle T_1, T_2\rangle_{g_F} = \int_{\mathcal{X}} \left(T_1 - \mathbb{E}_{p_\theta}\left[T_1\right]\right)\left(T_2 - \mathbb{E}_{p_\theta}\left[T_2\right]\right) p_\theta dx.$$

Here choosing the function $T_1 - \mathbb{E}_{p_\theta}[T_1]$ as the representative of $[T_1]$ is consistent with the normalization requirement (1). Thus Fisher covariance (resp. variance) reduces to the original definition of the covariance (resp. variance) in probability theory.

And the classical Cramer-Rao bound is given by

$$\mathrm{Cov}_\theta^F[T(x)] \succeq \nabla_\theta \mathbb{E}_{p_\theta}[T(x)]^\mathsf{T} G_F(\theta)^{-1} \nabla_\theta \mathbb{E}_{p_\theta}[T(x)],$$

where $G_F(\theta)$ is the Fisher information matrix. In 1-d cases, the above forms

$$\mathrm{Var}_\theta[T(x)] \geq \frac{\left(\nabla_\theta \mathbb{E}_{p_\theta} T(x)\right)^2}{G_F(\theta)}.$$

We next focus on the Wasserstein covariance operator.

Definition 8 (Wasserstein covariance)**.** *Given a statistical model Θ, denote the Wasserstein covariance as follows:*

$$\mathrm{Cov}_\theta^W[T_1, T_2] = \mathbb{E}_{p_\theta}\left(\nabla_x T_1(x), \nabla_x T_2(x)^\mathsf{T}\right),$$

where T_1, T_2 are random variables as functions of x and the expectation is taken w.r.t. $x \sim p_\theta$. Denote the Wasserstein variance:

$$\mathrm{Var}_\theta^W[T] = \mathbb{E}_{p_\theta}\left(\nabla_x T(x), \nabla_x T(x)^\mathsf{T}\right).$$

Theorem 9 (Wasserstein-Cramer-Rao inequality)**.** *Given any set of statistics $T = (T_1, ..., T_m) : \mathcal{X} \to \mathbb{R}^m$, where m is the number of the statistics, define two matrices $\mathrm{Cov}_\theta^W[T(x)]$, $\nabla_\theta \mathbb{E}_{p_\theta}[T(x)]^\mathsf{T}$ as below:*

$$\mathrm{Cov}_\theta^W[T(x)]_{ij} = \mathrm{Cov}_\theta^W[T_i, T_j], \qquad \nabla_\theta \mathbb{E}_{p_\theta}[T(x)]_{ij}^\mathsf{T} = \frac{\partial}{\partial \theta_j}\mathbb{E}_{p_\theta}[T_i(x)],$$

then

$$\mathrm{Cov}_\theta^W[T(x)] \succeq \nabla_\theta \mathbb{E}_{p_\theta}[T(x)]^\mathsf{T} G_W(\theta)^{-1} \nabla_\theta \mathbb{E}_{p_\theta}[T(x)],$$

where the notion $\succeq$ refers to that the difference of two matrices is positive semi-definite.

Proposition 10 (Covariance property)**.** *Given the Wasserstein score function $\Phi_i^W(x;\theta)$ and any smooth statistic $f\colon \mathcal{X} \to \mathbb{R}$, then*

$$\frac{\partial}{\partial \theta_i}\mathbb{E}_{p_\theta} f(x) = \mathbb{E}_{p_\theta}(\nabla_x \Phi_i^W(x), \nabla_x f(x)) = \langle \Phi_i^W, f \rangle_{g_W}.$$

Proof. Notice the fact that

$$\begin{aligned}
\frac{\partial}{\partial \theta_i}\mathbb{E}_{p_\theta} f(x) =& \frac{\partial}{\partial \theta_i}\int_{\mathcal{X}} f(x) p(x;\theta) dx \\
=& \int_{\mathcal{X}} f(x) \frac{\partial}{\partial \theta_i} p(x;\theta) dx \\
=& \int_{\mathcal{X}} f(x)\Big(- \nabla_x \cdot (p(x;\theta) \nabla_x \Phi_i^W(x;\theta))\Big) dx \\
=& \int_{\mathcal{X}} \nabla_x f(x) \cdot \nabla_x \Phi_i^W(x;\theta) p(x;\theta) dx,
\end{aligned}$$

where the third equality comes from the definition of Wasserstein score functions, while the last equality holds by integration by parts formula in spatial domain. □

Remark 6. This property is in contrast to Fisher score functions

$$\frac{\partial}{\partial\theta_i}\mathbb{E}_{p_\theta} f(x) = \mathbb{E}_{p_\theta}\left(f(x)\frac{\partial}{\partial\theta_i}\log p(x;\theta)\right)$$
$$= \mathrm{Cov}_\theta[(f(x), \frac{\partial}{\partial\theta_i}\log p(x;\theta))] = \langle \Phi_i^F, f\rangle_{g_F}.$$

This is merely a dual relation between tangent and cotangent space in the density manifold.

Proof of Theorem 9. By the definition of semi-positive matrix, it suffices to prove that for arbitrary $v \in \mathbb{R}^m$, we have:

$$v^\mathsf{T}\mathrm{Cov}_\theta^W[T(x)]v \geq v^\mathsf{T}\nabla_\theta\mathbb{E}_{p_\theta}[T(x)]^\mathsf{T} G_W(\theta)^{-1}\nabla_\theta\mathbb{E}_{p_\theta}[T(x)]v.$$

Here we define $T_v = v^\mathsf{T}T$ as the statistic associated to the vector v. Then the LHS of above formula equals to the variance of T_v:

$$v^\mathsf{T}\mathrm{Cov}_\theta^W[T(x)]v = \mathrm{Var}_\theta^W[T_v].$$

As we have mentioned before, score functions Φ_i^Ws, as a set of basis, span a linear space $V^*_{p(x;\theta)}\Theta$ of the cotangent space $T^*_{p(x;\theta)}\mathcal{P}(\mathcal{X})$ at each point $p(x;\theta)$ on the density manifold. Meanwhile, the statistic $T_v\colon \mathcal{X} \to \mathbb{R}$ can be viewed as a cotangent vector field on this statistical model. Now, the subspace $V^*_{p(x;\theta)}\Theta$ at each point θ is a finite-dimensional subspace of the Hilbert space $T^*_{p(x;\theta)}\mathcal{P}(\mathcal{X})$ endowed with the Wasserstein inner product. Thus it is a closed linear subspace. By elementary theory of functional analysis, we have orthogonal projection operator $\mathbf{P}$:

$$\mathbf{P}\colon T^*_{p(x;\theta)}\mathcal{P}(\mathcal{X}) \to V^*_{p(x;\theta)}\Theta.$$

Since Φ_i^Ws span the whole subspace, we have:

$$\langle \Phi_i^W, v - \mathbf{P}v\rangle_{g_W} = 0, \qquad \forall v \in T^*_{p(x;\theta)}\mathcal{P}(\mathcal{X}).$$

Now, back to the theorem, we have:

$$\mathrm{Var}_\theta^W[T_v] = \mathbb{E}_{p_\theta}\left[(\nabla_x T_v(x), \nabla_x T_v(x)^\mathsf{T})\right] = \langle T_v, T_v\rangle_{g_W} \geq \langle \mathbf{P}T_v, \mathbf{P}T_v\rangle_{g_W},$$

where the last inequality holds by the property of the orthogonal projection operator.

Now, since $\mathbf{P}$ is the projection onto the subspace $V^*_{p(x;\theta)}\Theta$ with a set of basis Φ_i^W, at each point θ, we can write the cotangent vector $\mathbf{P}T_v$ as a linear combination of Wasserstein score functions:

$$\mathbf{P}T_v = \sum_{i=1}^d t_i^\theta \Phi_i^W,$$

where the superscript of t_i^θ indicates the dependency on point θ. Now, plugging this linear combination into the Wasserstein metric, we get:

$$\begin{aligned}
\langle \mathbf{P}T_v, \mathbf{P}T_v\rangle_{g_W} &= \sum_{i=1}^{d} t_i^\theta \langle \mathbf{P}T_v, \Phi_i^W\rangle_{g_W} \\
&= \sum_{i,k=1}^{d} t_k^\theta \delta_i^k \langle \mathbf{P}T_v, \Phi_i^W\rangle_{g_W} \\
&= \sum_{i,j,k=1}^{d} t_k^\theta g_{kj} g^{ij} \langle \mathbf{P}T_v, \Phi_i^W\rangle_{g_W} \\
&= \sum_{i,j=1}^{d} \langle \mathbf{P}T_v, \Phi_i^W\rangle_{g_W} \left(G_W(\theta)^{-1}\right)_{ij} \langle \mathbf{P}T_v, \Phi_i^W\rangle_{g_W} \\
&= \nabla_\theta \mathbb{E}_{p_\theta}[T_v(x)]^\mathsf{T} G_W(\theta)^{-1} \nabla_\theta \mathbb{E}_{p_\theta}[T_v(x)],
\end{aligned}$$

where $G_W(\theta)^{-1}$ is the inverse matrix of the WIM, g_{kj}, g^{ij} are elements of matrix G_W, G_W^{-1} respectively, and the third equality holds by the fact $\sum_j g_{kj} g^{ij} = \delta_i^k$. The last equality is guaranteed by proposition 10. Combining the above calculation and the comparison between the inner product of T_v and $\mathbf{P}T_v$, we obtain the desired result. □

Given the above theorem, we can define the Wasserstein efficiency as follows.

Definition 11. *For an estimator $T(x)$, it is Wasserstein efficient if and only if it attains the Wasserstein-Cramer-Rao bound, namely:*

$$\mathrm{Var}_\theta^W[T(x)] = \nabla_\theta \mathbb{E}_{p_\theta}[T(x)]^\mathsf{T} G_W(\theta)^{-1} \nabla_\theta \mathbb{E}_{p_\theta}[T(x)].$$

Remark 7. From the above derivation, a sufficient and necessary condition for a statistic to be efficient is that, it can be written as a linear combination of score functions. Notice this criterion is valid for any metrics, including both Fisher and Wasserstein metrics. This is a purely geometric condition and we seek below in various statistical models to find out its statistical significance.

Remark 8. As shown in the above theorem, if we denote the Fisher-Rao metric as $g_F(p) = \frac{1}{p}$, we then derive the classical Cramer-Rao bound:

$$\mathrm{Cov}_\theta(T(x), T(x)) \geq \nabla_\theta \mathbb{E}_{p_\theta}[T(x)]^\mathsf{T} G_F(\theta)^{-1} \nabla_\theta \mathbb{E}_{p_\theta}[T(x)].$$

Here the Fisher-Rao metric corresponds to the classical covariance operator

$$\mathrm{Cov}_\theta(T(x), T(x)) = \mathbb{E}_{p_\theta}[(T(x) - \mathbb{E}_{p_\theta} T(x), T(x) - \mathbb{E}_{p_\theta} T(x))],$$

which depends on the expectation of statistics $\mathbb{E}_{p_\theta} T(X)$. Furthermore, given any information matrix on a statistical model, we have an associated Cramer-Rao bound.

3.2. **Analytic examples.**

Example 10 (Gaussian distribution)**.** Recall that given a Gaussian distribution with mean value μ and standard variance σ, Wasserstein score functions satisfy

$$\Phi^W_\mu(x;\mu,\sigma) = x-\mu, \quad \Phi^W_\sigma(x;\mu,\sigma) = \frac{(x-\mu)^2-\sigma^2}{2\sigma},$$

with the WIM

$$G_W(\mu,\sigma) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Thus by the criterion, we know that efficient statistics in Wasserstein cases are exactly those which can be written as linear combinations of score functions. Since statistics only depend on samples x_i and do not depend on parameters μ, σ, they must be of the form:

$$ax^2+bx+c = 2a\sigma\Phi^W_\sigma + (2a\mu+b)\Phi^W_\mu + c + a\mu^2 + b\mu - a\sigma^2,$$

since Wasserstein cotangent vectors are determined up to a constant. Wasserstein efficient statistics are degree 2 polynomials of x.

While score functions for Fisher case are given by:

$$\Phi^F_\mu(x;\mu,\sigma) = \frac{x-\mu}{\sigma^2}, \quad \Phi^F_\sigma(x;\mu,\sigma) = \frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma}.$$

And the Fisher information matrix satisfies

$$G_F(\mu,\sigma) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}.$$

Thus we conclude that although we have different score functions in Wasserstein and Fisher-Rao cases, it turns out that efficient statistics associated with these two information matrices coincide. But still, Fisher and Wasserstein information matrices provide us with different Cramer-Rao bounds. The Fisher-Cramer-Rao bound is better when we have prior knowledge that σ is large while worse if σ is small.

Example 11 (Exponential distribution)**.** Given an exponential distribution, Wasserstein score functions satisfy

$$\Phi^W_\lambda(x;m,\lambda) = \frac{(x-m)^2 - \frac{2}{\lambda^2}}{2\lambda}, \qquad \Phi^W_m(x;m,\lambda) = x - m - \frac{1}{\lambda},$$

and the WIM satisfies

$$G_W(m,\lambda) = \begin{pmatrix} 1 & \frac{1}{\lambda^2} \\ \frac{1}{\lambda^2} & \frac{2}{\lambda^4} \end{pmatrix}.$$

Similarly to Gaussian cases, Wasserstein sufficient statistics are also those which can be written as quadratic functions of variables x.

While the counterpart for Fisher case reads:

$$\Phi^F_\lambda(x;m,\lambda) = m - x + \frac{1}{\lambda}, \quad \Phi^F_m(x;m,\lambda) \text{ not well defined.}$$

Meanwhile, the Fisher information matrix is also ill-behaved, in contrast to the well-definedness of both Wasserstein score functions and WIMs. This example provides a situation where Wasserstein statistics are better than the classical Fisher statistics.

4. Wasserstein natural gradient works efficiently

In this section, we study Wasserstein dynamics in terms of sampling and estimation processes. As a consequence, we prove asymptotic efficiencies of the Wasserstein natural gradient algorithm. And we refer it as Wasserstein efficiency. Meanwhile, another efficiency that we named Poincaré efficiency is introduced and connected to Poincaré inequalities and log-Sobolev inequalities, which are discussed in section B.

In the beginning, we review the natural gradient algorithm. We aim to estimate an un-known distribution in a probability family $p(x;\theta)$ with unknown parameters $\theta \in \Theta$. Assume that an optimal parameter $p(x;\theta_*)$ coincides with the target distribution. Given a set of i.i.d. samples $x_i, i = 1, 2, ...$ from this distribution, we utilize a general online natural gradient algorithm to solve this problem:

$$\theta_{t+1} = \theta_t - \frac{1}{t}\nabla_\theta^W l(x_t, \theta_t). \tag{11}$$

In the above formula, θ_t is an updating state variable, $\frac{1}{t}$ in the RHS is an adaptive factor. And ∇_θ^W is the Riemannian (natural) gradient of the loss function l w.r.t. θ in Wasserstein metric. It can also be understood as using WIM as a preconditioner to get a new gradient direction, i.e. $\nabla_\theta^W l = G_W^{-1}\nabla_\theta l$, with $\nabla_\theta l$ being the Euclidean gradient. We first pose here a definition of the efficiency of the natural gradient algorithm, which generalizes the notion discussed in [2]. Denote the Wasserstein covariance matrix of estimator θ_t by:

$$V_t = \mathbb{E}_{p_{\theta_*}}\left(\nabla_x(\theta_t - \theta_*)\cdot\nabla_x(\theta_t - \theta_*)^T\right),$$

where $\nabla_x(\theta_t - \theta_*)$ is the matrix given by

$$\nabla_x(\theta_t - \theta_*) = \begin{pmatrix} \frac{\partial(\theta_t-\theta_*)_1}{\partial x_1} & \frac{\partial(\theta_t-\theta_*)_1}{\partial x_2} & \cdots & \frac{\partial(\theta_t-\theta_*)_1}{\partial x_n} \\ \frac{\partial(\theta_t-\theta_*)_2}{\partial x_1} & \frac{\partial(\theta_t-\theta_*)_2}{\partial x_2} & \cdots & \frac{\partial(\theta_t-\theta_*)_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial(\theta_t-\theta_*)_n}{\partial x_1} & \frac{\partial(\theta_t-\theta_*)_n}{\partial x_2} & \cdots & \frac{\partial(\theta_t-\theta_*)_n}{\partial x_n} \end{pmatrix},$$

and the multiplication is simply in matrix sense. It turns out that the element of the covariance matrix is given by

$$\mathbb{E}_{p_{\theta_*}}\left(\nabla_x(\theta_t - \theta_*)\cdot\nabla_x(\theta_t - \theta_*)^T\right)_{ij} = \mathbb{E}_{p_{\theta_*}}\left(\nabla_x(\theta_t - \theta_*)_i \cdot \nabla_x(\theta_t - \theta_*)_j\right)$$

Here $\cdot$ refers to inner product of gradient vectors. And subscripts $p(\cdot;\theta_*)$ refer to take expectation on the set of samples $x_i \sim p(\cdot;\theta_*), i = 1, 2, ...$ Notice that in this algorithm, we obtain the t-th estimator θ_t via $t-1$ iterations of the above equation (11). Then we actually have $\theta_t = \theta_t(x_1, x_2..., x_{t-1})$. Hence intuitively, the Wasserstein-Cramer-Rao bound for θ_t is given by $\frac{1}{t-1}G_W^{-1}(\theta_*)$. It inspires the following definition:

Definition 12. *The Wasserstein natural gradient is asymptotic efficient if*

$$V_t = \frac{1}{t}G_W^{-1}(\theta_*) + O(\frac{1}{t^2}).$$

Remark 9. This definition is similar to the definition of classical Fisher efficiency except that we substitute the Fisher information matrix by the WIM. This also indicates the importance of studying information matrices. And it will be shown that this quantity characterizes convergence rates of dynamics in statistical inference problems.

We first state a general updating equation for this dynamics. Then, we specify two different loss functions, namely, Fisher scores and Wasserstein scores. And we discuss convergence properties of these two cases separately.

Theorem 13 (Variance updating equation of the Wasserstein natural gradient)**.** *For any function $f(x,\theta)$ that satisfies the condition $\mathbb{E}_{p_\theta} f(x,\theta)=0$, consider here the asymptotic behavior of the Wasserstein dynamics $\theta_{t+1}=\theta_t-\frac{1}{t}G_W^{-1}(\theta_t)f(x_t,\theta_t)$. That is, assume priorly $\mathbb{E}_{p_{\theta_*}}\left[(\theta_t-\theta_*)^2\right]$,*
$\mathbb{E}_{p_{\theta_*}}\left[|\nabla_x(\theta_t-\theta_*)|^2\right]=o(1),\ \forall t$. *Then, the Wasserstein covariance matrix V_t updates according to the following equation:*

$$V_{t+1}=V_t+\frac{1}{t^2}G_W^{-1}(\theta_*)\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(f(x_t,\theta_*)\right)\cdot\nabla_x\left(f(x_t,\theta_*)^T\right)\right]\left(G_W^{-1}(\theta_*)\right)$$
$$-\frac{2V_t}{t}\mathbb{E}_{p_{\theta_*}}\left[\nabla_\theta f(x_t,\theta_*)\right]G_W^{-1}(\theta_*)+o(\frac{V_t}{t})+o\left(\frac{1}{t^2}\right).$$

Remark 10. In general, it will be shown that such a simple updating equation merely attributes to properties of information matrices. Specifically, any statistical information matrices with separability property w.r.t. independent variables have this form of updating equation. For the WIM, this is already established in proposition 5. And for the Fisher information matrix, this is only a property of expectation under independent variables. Further results such as efficiency of the natural gradient can be established with the same procedure below.

The proof is technical and we leave it to section C.1. Here, we show several important cases of Theorem 13.

4.1. **Wasserstein natural gradient for Wasserstein scores.** In Fisher case studied by [2], we have:

$$\nabla_\theta^F l(x_t,\theta_t)=G_F^{-1}\Phi^F(x_t,\theta_t),$$

with $l(x_t,\theta_t)$ the log-likelihood function. Thus a natural generalization to Wasserstein geometry is:

$$\nabla_\theta^W f(x_t,\theta_t)=G_W^{-1}\Phi^W(x_t,\theta_t). \tag{12}$$

Concerned with this dynamics, we have the following corrolary.

Corollary 14 (Wasserstein natural gradient efficiency)**.** *For the dynamics*

$$\theta_{t+1}=\theta_t-\frac{1}{t}G_W^{-1}(\theta_t)\Phi^W(x_t,\theta_t),$$

the Wasserstein covariance updates according to

$$V_{t+1}=V_t+\frac{1}{t^2}G_W^{-1}(\theta_*)-\frac{2}{t}V_t+o\left(\frac{1}{t^2}\right)+o(\frac{V_t}{t}).$$

Then, the online Wasserstein natural gradient algorithm is Wasserstein efficient, that is:

$$V_t=\frac{1}{t}G_W^{-1}(\theta_*)+O\left(\frac{1}{t^2}\right). \tag{13}$$

Proof of Corollary 14. If we choose function $f(x,t)$ to be Wasserstein scores Φ_i^W, we will have following simplification:

$$\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(\Phi^W(x_t,\theta_*)\right)\cdot\nabla_x\left(\Phi^W(x_t,\theta_*)^T\right)\right]=G_W(\theta_*),$$

since Φ^W is the dual coordinate of the statistical model. We also have:

$$\mathbb{E}_{p_{\theta_*}}\left[\nabla_\theta\Phi^W(x_t,\theta_*)\right]=-G_W(\theta_*),$$

which is given by differentiating $\mathbb{E}_{p_{\theta_*}}\left[\Phi^W(x_t,\theta_*)\right]=\mathbf{0}$ by θ:

$$\begin{aligned}\mathbf{0}&=\nabla_\theta\mathbb{E}_{p_{\theta_*}}\left[\Phi^W(x_t,\theta_*)\right]\\&=\nabla_\theta\left[\int_{\mathcal{X}}p(x;\theta_*)\Phi^W(x,\theta_*)dx\right]\\&=\int_{\mathcal{X}}\nabla_\theta p(x;\theta_*)\Phi^W(x,\theta_*)dx+\int_{\mathcal{X}}p(x;\theta_*)\nabla_\theta\Phi^W(x,\theta_*)dx\\&=G_W+\int_{\mathcal{X}}p(x;\theta_*)\nabla_\theta\Phi^W(x,\theta_*)dx,\end{aligned}$$

where the last equality holds because of the pairing between tangent vector and cotangent vector. And the final updating equation for the Wasserstein covariance reduces to:

$$V_{t+1}=V_t+\frac{1}{t^2}G_W^{-1}(\theta_*)-\frac{2}{t}V_t+O\left(\frac{1}{t^3}\right)+o(\frac{V_t}{t}).$$

To further solve this updating equation, we expand $V_t=\frac{x}{t}+\frac{y}{t^2}+o\left(\frac{1}{t^2}\right)$ with constant x,y to be determined and plug into the equation(we ignore the term that is of order $o\left(\frac{1}{t^2}\right)$):

$$\frac{x}{t+1}+\frac{y}{(t+1)^2}+o\left(\frac{1}{t^2}\right)=\frac{x}{t}+\frac{y}{t^2}+o\left(\frac{1}{t^2}\right)+\frac{1}{t^2}G_W^{-1}\left(\theta_*\right)-\frac{2x}{t^2}+o\left(\frac{1}{t^2}\right),$$

which is equivalent to:

$$\left(\frac{x}{t+1}-\frac{x}{t}\right)+\left(\frac{y}{(t+1)^2}-\frac{y}{t^2}\right)+\frac{2x}{t^2}-\frac{1}{t^2}G_W^{-1}\left(\theta_*\right)+o\left(\frac{1}{t^2}\right)=0.$$

And we conclude that:

$$x=G_W^{-1}\left(\theta_*\right).$$

Thus, we asymptotically have following estimation on the Wasserstein covariance concerned with this dynamics:

$$V_t=\frac{1}{t}G_W^{-1}\left(\theta_*\right)+o\left(\frac{1}{t}\right).$$

□

Remark 11. At first, such a generalization to Wasserstein metric may seem unreasonable. We only use a fact that both of them are metrics on probability spaces. Different from Fisher scores $\Phi^F=\nabla_\theta l\left(x;\theta\right)$, Wasserstein scores Φ^W can not be written as gradients of some functions w.r.t. θ. There is no such "loss functions". However, a key insight here is

that, if in a second we assume that the statistical model Θ is exactly the density manifold $G_W(p_\theta) = g_W(p_\theta), G_F(p_\theta) = g_F(p_\theta)$:

$$\begin{aligned} G_W^{-1}(p_\theta)\Phi^W(x,\theta) &= g_W(p_\theta)g_W^{-1}(p_\theta)\frac{\partial}{\partial\theta}p(x;\theta) = \nabla_\theta p(x_t,\theta_t) \\ &= g_F(p_\theta)g_F^{-1}(p_\theta)\frac{\partial}{\partial\theta}p(x;\theta) = G_F^{-1}(p_\theta)\Phi^F(x,\theta). \end{aligned}$$

Then both two dynamics can be written in the following way:

$$\theta_{t+1} = \theta_t - \frac{1}{t}\nabla_\theta p(x_t,\theta_t).$$

4.2. **Wasserstein natural gradient for Fisher scores.** Another phenomenon appears when we consider the Wasserstein natural gradient applies to Fisher scores. Specifically, we use log-likelihood function as a loss function and apply WIM as a preconditioner. The dynamics concerned in this case is given by:

$$\theta_{t+1} = \theta_t - \frac{1}{t}\nabla_\theta^W l(x_t,\theta_t).$$

The Wasserstein natural gradient is simply $\nabla_\theta^W l(x_t,\theta_t) = G_W^{-1}\nabla_\theta l(x_t,\theta_t)$. We comment here that $\nabla_\theta l(x_t,\theta_t) = \Phi^F(x_t,\theta_t)$ is both the Euclidean gradient of log-likelihood function l w.r.t. θ and the Fisher score. And the convergence analysis is shown in the following corollary:

Corollary 15 (Poincaré efficiency)**.** *For the dynamics*

$$\theta_{t+1} = \theta_t - \frac{1}{t}\nabla_\theta^W l(x_t,\theta_t),$$

the Wasserstein covariance updates according to

$$\begin{aligned} V_{t+1} = V_t &+ \frac{1}{t^2}G_W^{-1}(\theta_*)\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(\nabla_\theta l(x_t,\theta_*)\right)\cdot\nabla_x\left(\nabla_\theta l(x_t,\theta_*)^T\right)\right]G_W^{-1}(\theta_*) \\ &- \frac{2}{t}V_t G_F(\theta_*)G_W^{-1}(\theta_*) + O\left(\frac{1}{t^3}\right) + o\left(\frac{V_t}{t}\right). \end{aligned}$$

Now suppose that $\alpha = sup\{a|G_F \succeq aG_W\}$. Then the dynamics is characterized by the following formula

$$V_t = \begin{cases} O\left(t^{-2\alpha}\right), & 2\alpha \le 1, \\ \frac{1}{t}\left(2G_F G_W^{-1} - \mathbf{I}\right)^{-1}G_W^{-1}(\theta_*)\mathfrak{I}\left(G_W^{-1}(\theta_*)\right) + O\left(\frac{1}{t^2}\right), & 2\alpha > 1, \end{cases} \tag{14}$$

where

$$\mathfrak{I} = \mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(\nabla_\theta l(x_t,\theta_*)\right)\cdot\nabla_x\left(\nabla_\theta l(x_t,\theta_*)^T\right)\right],$$

where elements of this matrix is given by

$$\mathfrak{I}_{ij} = \mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(\nabla_{\theta_i} l(x_t,\theta_*)\right)\cdot\nabla_x\left(\nabla_{\theta_j} l(x_t,\theta_*)^T\right)\right]$$

Proof of Corollary 15. The result is obtained once we observe that:

$$\mathbb{E}_{p_{\theta_*}}\left[\nabla_\theta\Phi^F(x_t,\theta_*)\right] = -G_F(\theta_*),$$

which follows exactly the same philosophy of the previous case. We conclude that the Wasserstein covariance updates according to:

$$V_{t+1} = V_t + \frac{1}{t^2} G_W^{-1}(\theta_*) \mathbb{E}_{p_{\theta_*}} \left[\nabla_x \left(\nabla_\theta l(x_t, \theta_*) \right) \cdot \nabla_x \left(\nabla_\theta l(x_t, \theta_*)^T \right) \right] \left(G_W^{-1}(\theta_*) \right) - \frac{2}{t} V_t G_F(\theta_*) G_W^{-1}(\theta_*) + O\left(\frac{1}{t^3} \right) + o\left(\frac{V_t}{t} \right).$$

Next, we solve this dynamics asymptotically. We denote $G_F(\theta_*)G_W^{-1}(\theta_*) = B$ and $G_W^{-1}(\theta_*)\mathbb{E}_{p_{\theta_*}} \left[\nabla_x \left(\nabla_\theta l(x_t, \theta_*) \right) \cdot \nabla_x \left(\nabla_\theta l(x_t, \theta_*)^T \right) \right] \left(G_W^{-1}(\theta_*) \right) = C$.

Now by elementary linear algebra, we know that the matrix $B = G_F(\theta_*)G_W^{-1}(\theta_*)$ is similar to the matrix $G_W^{-\frac{1}{2}} G_F G_W^{-\frac{1}{2}}$. Hence their eigenvalues coincide. While the definition of α translates to that the least eigenvalue of the symmetric matrix $G_W^{-\frac{1}{2}} G_F G_W^{-\frac{1}{2}}$ is exactly α. Thus we conclude that the least eigenvalues of the matrix B are also α. Suppose first that $2\alpha < 1$, we consider the following expansion of matrix V_t:

$$V_t = \frac{A_1}{t^q} + \frac{A_2}{t^{q+1}} + o\left(\frac{1}{t^{q+1}} \right), \qquad A_1, A_2 = O(1).$$

And plug the above equation to both sides of the updating equation, we find:

$$\frac{A_1}{(t+1)^q} + \frac{A_2}{(t+1)^{q+1}} + o\left(\frac{1}{t^{q+1}} \right) = \frac{A_1}{t^q} + \frac{A_2}{t^{q+1}} + o\left(\frac{1}{t^{q+1}} \right) - \frac{2A_1 B}{t^{q+1}} + \frac{C}{t^2} + O(\frac{1}{t^3}).$$

Using the Lagrange's mean value theorem, we have:

$$\frac{A}{t^q} - \frac{A}{(t+1)^q} = \frac{qA}{(t+\upsilon)^{q+1}} = \frac{qA}{t^{q+1}} + o(\frac{1}{t^q}), \qquad \upsilon \in [0,1].$$

Substituting back to the equation, we get:

$$\mathbf{0} = \frac{A_1 \left(q\mathbf{I} - 2B \right)}{t^{q+1}} + \frac{C}{t^2} + o(\frac{1}{t^{q+1}}) + O(\frac{1}{t^3}).$$

We conclude that we cannot have q strictly greater than 1, for then the most significant term in the RHS will be $\frac{C}{t^2} \neq \mathbf{0}$ which contradicts to the LHS. Thus if we have $q < 2\alpha < 1$, the matrix $q\mathbf{I} - 2B$ will be negative definite, and we cannot have $A_1 \left(q\mathbf{I} - 2B \right) = \mathbf{0}$ unless A_1 equals to 0. Consequently, the index q should be greater than or equal to 2α. And we have that asymptotically

$$V_t = O\left(\frac{1}{t^{2\alpha}} \right).$$

While for the situation such that $2\alpha > 1$, we expand $V_t = \frac{A_1}{t} + \frac{A_2}{t^2} + o\left(\frac{1}{t^2} \right)$ with constant A_1, A_2 to be determined:

$$\frac{A_1}{t+1} + \frac{A_2}{(t+1)^2} + o\left(\frac{1}{t^2} \right) = \frac{A_1}{t} + \frac{A_2}{t^2} + o\left(\frac{1}{t^2} \right) + \frac{C}{t^2} - \frac{2A_1 B}{t^2} + o\left(\frac{1}{t^2} \right).$$

The constant A_1 can be fixed by considering the coefficient of the term $\frac{1}{t^2}$ for both sides with conclusion:

$$A_1 = \left(2B - \mathbf{I} \right)^{-1} C.$$

Here, invertibility of the matrix $2B - \mathbf{I}$ is guaranteed by the fact that eigenvalues of $2B$ are all greater than 1, thus the matrix $2B - \mathbf{I}$ is indeed positive definite. □

The convergence behavior of this dynamics relies largely on the least significant eigenvalue of the matrix $G_F G_W^{-1}$. This is in great similarity with the RIW condition for Poincaré inequality in statistical models [16]. This inspires us to name such efficiency Poincaré efficiency. For detailed discussions and calculations on Poincaré inequalities in statistical models, please refer to the section B. We also illustrate some results of two efficiencies in Gaussian family, whose proof and numerical experiments are delayed to the section C.2.

Example 12 (Gaussian distribution)**.** Suppose we have following dynamics in a Gaussian model $p(x;\mu,\sigma)$

$$\theta_{t+1} = \theta_t - \frac{1}{t}\nabla_\theta^W l(x_t,\theta_t), \quad x_t \sim p(x;\mu_*,\sigma_*).$$

The asymptotic behavior of the covariance matrix for the online Wasserstein natural gradient algorithm is given by

$$V_t = \begin{cases} O\left(t^{-\frac{2}{\sigma_*^2}}\right), & \frac{1}{\sigma_*^2} \le \frac{1}{2}, \\ \frac{1}{t}\begin{pmatrix} \frac{1}{(2-\sigma_*^2)\sigma_*^2} & 0 \\ 0 & \frac{4}{(4-\sigma_*^2)\sigma_*^2} \end{pmatrix} + O(\frac{1}{t^2}), & \frac{1}{\sigma_*^2} > \frac{1}{2}. \end{cases}$$

5. Discussions

In this paper, we introduce the Wasserstein information matrix in statistical models. Similar to the study in information geometry, we turn the geometric aspect of the Wasserstein metric into statistics. Here we generalize the classical concepts such as score function, covariance operator, Cramer-Rao bound, and estimation to the Wasserstein statistics. Several explicit computable examples are provided, including the location-scale family, and the ReLU push-forward family. Also, by comparing both Wasserstein and Fisher information matrices, some new efficiency concepts, such as Wasserstein efficiency and Poincaré efficiency have been introduced.

In the future, several natural questions between Fisher and Wasserstein statistics arise. For example, similar to the relation with Fisher information matrices and maximal likelihood estimators, what is the relation between the WIM and the Wasserstein distance estimator? Is there a canonical Wasserstein divergence function for the WIM? What is the corresponding Wasserstein maximal likelihood estimator? Meanwhile, we will apply Wasserstein natural gradient to study stochastic gradient descent algorithms in statistical learning problems. Lastly and most importantly, we have shown that the Wasserstein statistics provide the rigorous statistical advantages in generative models than classical Fisher statistics. We will study the properties of WIMs in clear statistical terms for machine learning models.

References

[1] S. Amari. *Differential-Geometrical Methods in Statistics*. Number 28 in Lecture Notes in Statistics. Springer-Verlag, Berlin ; New York, corr. 2nd print edition, 1990.

[2] S. Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2):251–276, 1998.
[3] S. Amari. *Information Geometry and Its Applications*. Number volume 194 in Applied mathematical sciences. Springer, Japan, 2016.
[4] M. Arbel, A. Gretton, W. Li, and G. Montufar. Kernelized Wasserstein Natural Gradient. *arXiv:1910.09652 [cs, stat]*, 2019.
[5] N. Ay, J. Jost, H. Vân Lê, and L. Schwachhöfer. *Information geometry*, volume 64. Springer, 2017.
[6] E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. On parameter estimation with the Wasserstein distance. *arXiv e-prints*, Jan. 2017.
[7] F.-X. Briol, A. Barp, A. B. Duncan, and M. Girolami. Statistical inference for generative models with maximum mean discrepancy. *arXiv preprint arXiv:1906.05944*, 2019.
[8] G. Casella and R. L. Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.
[9] Y. Chen and W. Li. Natural gradient in wasserstein statistical manifold. *arXiv preprint arXiv:1805.08380*, 2018.
[10] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, N.J, 2nd ed edition, 2006.
[11] J. D. Lafferty. The density manifold and configuration space quantization. *Transactions of the American Mathematical Society*, 305(2):699–741, 1988.
[12] W. Li. Geometry of probability simplex via optimal transport. *arXiv:1803.06360 [math]*, 2018.
[13] W. Li, A. T. Lin, and G. Montúfar. Affine natural proximal learning. *Geometric science of information*, 2019, 2019.
[14] W. Li, S. Liu, H. Zha, and H. Zhou. Parametric fokker-planck equation. In F. Nielsen and F. Barbaresco, editors, *Geometric Science of Information*, pages 715–724, Cham, 2019. Springer International Publishing.
[15] W. Li and G. Montúfar. Natural gradient via optimal transport. *Information Geometry*, 2018.
[16] W. Li and G. Montúfar. Ricci curvature for parametric statistics via optimal transport. *arXiv:1807.07095 [cs, math, stat]*, 2018.
[17] A. T. Lin, W. Li, S. Osher, and G. Montufar. Wasserstein proximal of GANs, 2019.
[18] A. Mallasto, T. D. Haije, and A. Feragen. A formalization of the natural gradient method for general similarity measures. *arXiv preprint arXiv:1902.08959*, 2019.
[19] Y. Ollivier. Online natural gradient as a Kalman filter. *Electronic Journal of Statistics*, 12(2):2930–2961, 2018.
[20] F. Otto. The geometry of dissipative evolution equations the porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.
[21] F. Otto and C. Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.
[22] A. Petersen and H.-G. Müller. Wasserstein covariance for multiple random densities. *Biometrika*, 106(2):339–351, 2019.
[23] C. Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.
[24] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.
[25] T.-K. L. Wong. Logarithmic divergences from optimal transport and Rényi geometry. *Information Geometry*, 1(1):39–78, 2018.
[26] S. Zozor and J.-M. Brossier. debruijn identities: From shannon, kullback-leibler and fisher to generalized φ-entropies, φ-divergences and φ-fisher informations. In *AIP Conference Proceedings*, volume 1641, pages 522–529. AIP, 2015.

Appendix A. Proofs in section 2

A.1. **WIMs and score functions in analytic examples.**

Proof of WIMs in Gaussian families. Since we have

$$\log p(x;\mu,\sigma) = -\frac{(x-\mu)^2}{2\sigma^2} - \log\sigma - \log\sqrt{2\pi},$$

taking derivative, we get

$$\begin{aligned}
&\nabla_x \log p(x;\mu,\sigma) = -\frac{x-\mu}{\sigma^2},\\
&\frac{\partial}{\partial\mu}\log p(x;\mu,\sigma) = \frac{x-\mu}{\sigma^2}, \quad \frac{\partial}{\partial\sigma}\log p(x;\mu,\sigma) = \frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma}.
\end{aligned}$$

In this case, the Possion equation for Wasserstein score functions $(\Phi_\mu^W, \Phi_\sigma^W)$ forms

$$\begin{cases}
-\dfrac{x-\mu}{\sigma^2}\cdot\dfrac{d}{dx}\Phi_\mu^W + \dfrac{d^2}{dx^2}\Phi_\mu^W = -\dfrac{x-\mu}{\sigma^2},\\
-\dfrac{x-\mu}{\sigma^2}\cdot\dfrac{d}{dx}\Phi_\sigma^W + \dfrac{d^2}{dx^2}\Phi_\sigma^W = -\dfrac{(x-\mu)^2}{\sigma^3} + \dfrac{1}{\sigma}.
\end{cases}$$

We simply check that $\Phi_\mu^W(x) = x-\mu$ and $\Phi_\sigma^W(x) = \frac{(x-\mu)^2-\sigma^2}{2\sigma}$ are solutions, and they also satisfy the normalization condition $\mathbb{E}_{p_\theta}\Phi_i^W = 0$. Thus

$$\begin{aligned}
G_W(\mu,\sigma)_{\mu\mu} &= \mathbb{E}_{p_{\mu,\sigma}}\left(\frac{d}{dx}\Phi_\mu^W, \frac{d}{dx}\Phi_\mu^W\right) = \mathbb{E}_{p_{\mu,\sigma}} 1 = 1,\\
G_W(\mu,\sigma)_{\mu\sigma} &= \mathbb{E}_{p_{\mu,\sigma}}\left(\frac{d}{dx}\Phi_\mu^W, \frac{d}{dx}\Phi_\sigma^W\right) = \mathbb{E}_{p_{\mu,\sigma}}\left(1\cdot(-\frac{X-\mu}{2\sigma})\right) = 0,\\
G_W(\mu,\sigma)_{\sigma\sigma} &= \mathbb{E}_{p_{\mu,\sigma}}\left(\frac{d}{dx}\Phi_\sigma^W, \frac{d}{dx}\Phi_\sigma^W\right) = \mathbb{E}_{p_{\mu,\sigma}}\left(\frac{X-\mu}{\sigma}\cdot\frac{X-\mu}{\sigma}\right) = 1.
\end{aligned}$$

□

Proof of WIMs in exponential families. We derive results using the closed-form solution in 1-d. The cumulative distribution function satisfies

$$F(x;m,\lambda) = \begin{cases} 1-e^{-\lambda(x-m)} & x\geq m,\\ 0 & x<m.\end{cases}$$

Thus

$$\frac{\partial}{\partial\lambda}F(x;m,\lambda) = \begin{cases}(x-m)e^{-\lambda(x-m)} & x\geq m,\\ 0 & x<m.\end{cases}$$

$$\frac{\partial}{\partial m}F(x;m,\lambda) = \begin{cases}\lambda e^{-\lambda(x-m)} & x\geq m,\\ 0 & x<m.\end{cases}$$

Then

$$\begin{aligned}\Phi_\lambda^W(x;m,\lambda) &= -\int_m^x \frac{1}{p(y;m,\lambda)}\frac{\partial}{\partial\lambda}F(y;m,\lambda)dy + C_1\\ &= -\int_m^x \frac{(y-m)}{\lambda}dy + C_1 = \frac{(x-m)^2}{2\lambda} + C_1,\\ \Phi_m^W(x;m,\lambda) &= -\int_m^x \frac{1}{p(y;m,\lambda)}\frac{\partial}{\partial m}F(y;m,\lambda)dy + C_2\\ &= -\int_m^x dy + C_2 = (x-m) + C_2.\end{aligned}$$

Using the normalization condition, we can decide integration constants appearing above. And inner products between score functions follow:

$$\begin{aligned}G_W(m,\lambda)_{\lambda\lambda} &= \mathbb{E}_{p_{m,\lambda}}\left(\frac{d}{dx}\Phi_\lambda^W, \frac{d}{dx}\Phi_\lambda^W\right)\\ &= \int_m^\infty \frac{(x-m)}{\lambda}\cdot\frac{(x-m)}{\lambda}\cdot\lambda e^{-\lambda(x-m)}dx\\ &= \int_m^\infty \frac{(x-m)^2}{\lambda}e^{-\lambda(x-m)}dx = \frac{2}{\lambda^4},\\ G_W(m,\lambda)_{\lambda m} &= \mathbb{E}_{p_{m,\lambda}}\left(\frac{d}{dx}\Phi_\lambda^W, \frac{d}{dx}\Phi_m^W\right) = \int_m^\infty \frac{(x-m)}{\lambda}\cdot\lambda e^{-\lambda(x-m)}dx = \frac{1}{\lambda^2},\\ G_W(m,\lambda)_{mm} &= \mathbb{E}_{p_{m,\lambda}}\left(\frac{d}{dx}\Phi_m^W, \frac{d}{dx}\Phi_m^W\right) = \int_m^\infty \lambda e^{-\lambda(x-m)}dx = 1.\end{aligned}$$

□

Proof of WIMs in uniform families. The cumulative distribution function satisfies

$$F(x;a,b) = \begin{cases}1 & x > b,\\ \frac{x-a}{b-a} & a \le x \le b,\\ 0 & x < a.\end{cases}$$

Thus when $x \in [a,b]$,

$$\frac{\partial}{\partial a}F(x;a,b) = \frac{x-b}{(b-a)^2}, \quad \frac{\partial}{\partial b}F(x;a,b) = \frac{a-x}{(b-a)^2}.$$

Then

$$\begin{aligned}\Phi_a^W(x;a,b) &= -\int_a^x \frac{1}{p(y;a,b)}\frac{\partial}{\partial a}F(y;a,b)dy + C_1 = \frac{(x-a)(a-2b+x)}{2(b-a)} + C_1,\\ \Phi_b^W(x;a,b) &= -\int_a^x \frac{1}{p(y;a,b)}\frac{\partial}{\partial b}F(y;a,b)dy + C_2 = \frac{(a-x)^2}{2(b-a)} + C_2,\end{aligned}$$

where integration constants C_1, C_2 can be decided via the normalization condition. Thus

$$\begin{aligned}
G_W(a,b)_{aa} &= \mathbb{E}_{p_{a,b}}\left(\frac{d}{dx}\Phi_a^W, \frac{d}{dx}\Phi_a^W\right) = \frac{1}{3},\\
G_W(a,b)_{ab} &= \mathbb{E}_{p_{a,b}}\left(\frac{d}{dx}\Phi_a^W, \frac{d}{dx}\Phi_b^W\right) = \frac{1}{6},\\
G_W(a,b)_{bb} &= \mathbb{E}_{p_{a,b}}\left(\frac{d}{dx}\Phi_b^W, \frac{d}{dx}\Phi_b^W\right) = \frac{1}{3}.
\end{aligned}$$

□

Proof of the WIM in semicircle families. The cumulative distribution function satisfies

$$\begin{aligned}
F(x+m;m,R) =& \int_{-R}^{x} \frac{2}{\pi R^2}\sqrt{R^2-y^2}dy\\
=& \int_{-\frac{\pi}{2}}^{\arcsin(\frac{x}{R})} \frac{2}{\pi R^2}\sqrt{R^2-R^2\sin^2 t}\,td(R\sin t)\\
=& \int_{-\frac{\pi}{2}}^{\arcsin(\frac{x}{R})} \frac{2}{\pi R^2}R^2(\cos t)^2dt\\
=& \int_{-\frac{\pi}{2}}^{\arcsin(\frac{x}{R})} \frac{1}{2\pi}\frac{\cos(2t)+1}{2}dt\\
=& \frac{1}{\pi}\Big(\frac{\sin(2t)}{2}+t\Big)\Big|_{-\frac{\pi}{2}}^{\arcsin\frac{x}{R}}\\
=& \frac{1}{\pi}\Big\{\frac{x\sqrt{R^2-x^2}}{R^2}+\arcsin\frac{x}{R}+\frac{\pi}{2}\Big\},
\end{aligned}$$

where we use a transformation $y = R\sin t$. Thus

$$\begin{aligned}
\frac{\partial}{\partial R}F(x+m;m,R) =& \frac{1}{\pi}\Big\{\frac{(x\sqrt{R^2-x^2})'R^2-2Rx\sqrt{R^2-x^2}}{R^4}+(\arcsin\frac{x}{R})'\Big\}\\
=& \frac{1}{\pi}\Big\{\frac{xR(R^2-x^2)^{-\frac{1}{2}}R^2-2Rx\sqrt{R^2-x^2}}{R^4}-\frac{x}{R\sqrt{R^2-x^2}}\Big\}\\
=& -\frac{2x\sqrt{R^2-x^2}}{\pi R^3}.
\end{aligned}$$

Thus

$$\begin{aligned}
\Phi_R^W(x+m;m,R) =& -\int_{-R}^{x}\frac{1}{p(y;m,R)}\frac{\partial}{\partial R}F(y;m,R)dy+C\\
=& \int_{-R}^{x}\frac{y}{R}dy+C\\
=& \frac{1}{R}(\frac{x^2}{2}-\frac{R^2}{2})+C.
\end{aligned}$$

The calculation of the score function associated with the parameter p is the same as before. And we conclude

$$\Phi_p^W(x;m,R) = x-m.$$

Thus

$$G_W(m,R)_{mm} = \mathbb{E}_{p_{m,R}}\left(\frac{d}{dx}\Phi_m^W, \frac{d}{dx}\Phi_m^W\right) = 1,$$
$$G_W(m,R)_{mR} = \mathbb{E}_{p_{m,R}}\left(\frac{d}{dx}\Phi_R^W, \frac{d}{dx}\Phi_m^W\right) = \frac{1}{R^2}\mathbb{E}_{m_R}(x-m) = 0,$$
$$G_W(m,R)_{RR} = \mathbb{E}_{p_{m,R}}\left(\frac{d}{dx}\Phi_R^W, \frac{d}{dx}\Phi_R^W\right) = \frac{1}{R^2}\mathbb{E}_{m_R}(x-m)^2 = \frac{1}{4}.$$

□

A.2. **The location-scale family.**

Example 13 (Location-scale families)**.** Consider a location-scale family as following: given a probability density function $p(x)$ with $\int_{\mathbb{R}} p(x)dx = 1$, we define density functions of a location-scale family with a location parameter m, and a scale parameter λ as

$$p(x;m,\lambda) = \frac{1}{\lambda}p(\frac{x-m}{\lambda}), \qquad \lambda > 0.$$

Most of previously discussed examples belong to this family, except that we do not use location and scale parameters in their parameterizations. We present some geometric formulas in this setting. We further require the original density function to be symmetric according to the location parameter m, i.e. $p(x) = p(2m-x)$. Notice that a simple corollary of this assumption is $\mathbb{E}_{p_{m,\lambda}} x = m$.

We use the closed-form solution for 1-d model to calculate the score function associated with the location parameter m. Thus we have:

$$\begin{aligned}\frac{\partial}{\partial m}F(x;m,\lambda) =& \frac{\partial}{\partial m}\int_{-\infty}^{x} p(y;m,\lambda)dy = \frac{\partial}{\partial m}\int_{-\infty}^{x}\frac{1}{\lambda}p(\frac{y-m}{\lambda})dy\\ =& -\frac{\partial}{\partial x}\int_{-\infty}^{x}\frac{1}{\lambda}p(\frac{y-m}{\lambda})dy = -p(x;m,\lambda).\end{aligned}$$

Consequently, the score function associated to the parameter m satisfies

$$\Phi_m^W(x;m,\lambda) = -\int_m^x \frac{1}{p(y;m,\lambda)}\frac{\partial}{\partial m}F(y;m,\lambda)dy + C_1 = (x-m) + C_1,$$

where the integration constant C_1 is determined to be 0. Thus we have

$$G_W(m,\lambda)_{mm} = \mathbb{E}_{p_{m,\lambda}}\left(\frac{d}{dx}\Phi_m^W, \frac{d}{dx}\Phi_m^W\right) = 1.$$

For the scaling parameter λ, we use a method of optimal transportation map to determine its score function. Namely, for two smooth distributions p_1, p_2 which are absolutely continuous w.r.t. each other, their Wasserstein distance can be obtained by an optimal transportation map f, i.e.

$$f_* p_1 = p_2, \quad W_2^2(p_1,p_2) = \int_{\mathcal{X}} (f(x)-x)^2 p_1(x)\,dx.$$

Assume we have a tangent vector $\frac{\partial p}{\partial \theta}$ and a smooth path $p(t) \subset \mathcal{P}(\mathcal{X}), t \in [-\epsilon,\epsilon]$ with $p(0) = p_0, p'(0) = \frac{\partial p}{\partial \theta}$. Denote the optimal transportation map between $p(0), p(\theta)$ as

$f(x,\theta)$. Then we have following relation between optimal transportation maps and the score function associated with tangent vector $\frac{\partial p}{\partial \theta}$

$$\frac{d}{dx}\Phi^W(x) = \lim_{\Delta\theta\to 0}\frac{f(x,\Delta\theta)-x}{\Delta\theta}.$$

First, we show that the optimal transportation map between distributions $p(x;m_1,\lambda_1)$ and $p(x;m_2,\lambda_2)$ is given by a linear map:

$$l(x) = m_2 + \frac{(x-m_1)\lambda_2}{\lambda_1}.$$

As we are working in a location-scale family, it is easy to show that this map pushes $p(x;m_1,\lambda_1)$ forward to $p(x;m_2,\lambda_2)$, i.e. $l_* p_{m_1,\lambda_1} = p_{m_2,\lambda_2}$. Then, we have

$$l(x) = \nabla_x\left(m_2(x-m_1) + \frac{(x-m_1)^2\lambda_2}{2\lambda_1}\right).$$

The function in the bracket is a convex function. Therefore, $l(x)$ is exactly the optimal transportation map between these two distributions.

To calculate the score function correspondent to the tangent vector $\frac{\partial}{\partial\lambda}$, we consider following infinitesimal optimal transportation $p(x;m_1,\lambda_1)\to p(x;m_1,\lambda_1+d\lambda)$. By discussions above, the optimal transportation map is given by

$$l(x) = m_1 + \frac{(x-m_1)(\lambda_1+d\lambda)}{\lambda_1} = x + (x-m_1)\frac{d\lambda}{\lambda_1}.$$

Thus the gradient of the score function is given by

$$\frac{d}{dx}\Phi^W_\lambda(x;m_1,\lambda_1) = \frac{l(x)-x}{d\lambda} = \frac{(x-m_1)}{\lambda_1}.$$

The inner product of this tangent vector is given by

$$\begin{aligned} G_W(m,\lambda)_{\lambda\lambda} =& \mathbb{E}_{p_{m,\lambda}}\left(\frac{d}{dx}\Phi^W_\lambda, \frac{d}{dx}\Phi^W_\lambda\right)\\ =& \int_{\mathbb{R}}\left(\frac{x-m}{\lambda}\right)^2 p(x;m,\lambda)dx\\ =& \frac{\mathbb{E}_{p_{m,\lambda}}x^2 - 2m\mathbb{E}_{p_{m,\lambda}}x + m^2}{\lambda^2}. \end{aligned}$$

The gradient of the score function associated to the parameter λ (resp. m) is odd (resp. even) function when viewing as a function of $x-m$. We conclude that the integration of their product is zero:

$$G_W(m,\lambda)_{\lambda m} = \mathbb{E}_{p_{m,\lambda}}\left(\frac{d}{dx}\Phi^W_\lambda, \frac{d}{dx}\Phi^W_m\right) = \mathbb{E}_{p_{m,\lambda}}(x-m) = 0.$$

Consequently, WIMs of location-scale families are diagonal matrices, i.e.

$$G_W(m,\lambda) = \begin{pmatrix} 1 & 0\\ 0 & \frac{\mathbb{E}_{p_{m,\lambda}}x^2 - 2m\mathbb{E}_{p_{m,\lambda}}x + m^2}{\lambda^2}\end{pmatrix}.$$

We next explain above closed-form solutions of WIMs by following proposition.

Proposition 16. *A location-scale family $p(x; m, \lambda)$ is a totally geodesic family in density manifold under Wasserstein metric.*

Proof. It suffices to prove that for any two densities $\rho_1 = p(x; m_1, \lambda_1)$ and $\rho_2 = p(x; m_2, \lambda_2)$, a geodesic connecting them lies within this family. We compute the optimal transport map T associated with these two measures ρ_1, ρ_2, that is:

$$T = \text{argmin}_{T_*\rho_1=\rho_2} \int_{\mathbb{R}} (T(x) - x)^2 \rho_1(x) dx,$$

where T is a map that pushes density ρ_1 forward to density ρ_2. It is known that a sufficient and necessary condition for an optimal map in 1-d case is that it is a monotone map, i.e. $(T(x) - T(y))(x - y) \geq 0$. And in a location-scale family, such map has a closed-form solution, namely:

$$T(x) = \frac{\lambda_2 (x - m_1)}{\lambda_1} + m_2.$$

The geodesic $\gamma(t) : [0, 1] \to \mathcal{P}(\mathbb{R})$ between ρ_1 and ρ_2 follows easily as below by the classical theory of optimal transport

$$\gamma(t) = (tx + (1 - t)T(x))_* \rho_1,$$

where the push-forward map has a closed-form solution:

$$\begin{aligned} tx + (1 - t)T(x) &= tx + (1 - t)\frac{\lambda_2 (x - m_1)}{\lambda_1} + (1 - t)m_2 \\ &= \frac{(t\lambda_1 + (1 - t)\lambda_2)(x - m_1)}{\lambda_1} + (1 - t)m_2 + tm_1. \end{aligned}$$

And by the same argument, $\gamma(t)$ lies in this location-scale family with parameters given by

$$\lambda_t = t\lambda_1 + (1 - t)\lambda_2, \qquad m_t = (1 - t)m_2 + tm_1.$$

Thus we show that geodesics between any two densities in a location-scale family lie in this family. In other words, location-scale families are totally geodesic submanifolds in density manifold. □

Remark 12. This result on totally geodesic of location-scale families is a generalization of the same result on Gaussian families in 1-d. Both proofs of these two cases rely on the fact that optimal transport maps in these families are linear.

Remark 13. For location-scale families, we also formulate its Fisher scores and Fisher information matrices for comparisons:

$$\begin{aligned}
\Phi_m^F(x;m,\lambda) &= \frac{p'}{\lambda p}, \quad \Phi_\lambda^F(x;m,\lambda) = -\frac{1}{\lambda} - \frac{(x-m)p'}{\lambda^2 p},\\
G_F(m,\lambda)_{\lambda\lambda} &= \int_{\mathbb{R}} p\left(\partial_\lambda \log p\right)^2 dx = \int_{\mathbb{R}} p\left(-\frac{1}{\lambda} - \frac{(x-m)p'}{\lambda^2 p}\right)^2 dx\\
&= \frac{1}{\lambda^2}\left(1 + \int_{\mathbb{R}}\left(\frac{(x-m)^2 p'^2}{\lambda^2 p} + \frac{(x-m)\,p'}{\lambda}\right) dx\right),\\
G_F(m,\lambda)_{mm} &= \int_{\mathbb{R}} p\left(\partial_m \log p\right)^2 dx = \frac{1}{\lambda^2}\int_{\mathbb{R}} \frac{p'^2}{p} dx,\\
G_F(m,\lambda)_{m\lambda} &= \int_{\mathbb{R}} p\left(\partial_m \log p\right)\left(\partial_\lambda \log p\right) dx\\
&= \int_{\mathbb{R}} p\left(-\frac{p'}{\lambda p}\right)\left(-\frac{1}{\lambda} - \frac{(x-m)p'}{\lambda^2 p}\right) dx\\
&= \int_{\mathbb{R}} \frac{(x-m)p'^2}{\lambda^3 p} dx.
\end{aligned} \tag{15}$$

WIMs and Fisher information matrices are given by

$$\begin{aligned}
G_W(m,\lambda) &= \begin{pmatrix} 1 & 0 \\ 0 & \frac{\mathbb{E}_{p_{m,\lambda}} x^2 - 2m\mathbb{E}_{p_{m,\lambda}} x + m^2}{\lambda^2} \end{pmatrix},\\
G_F(m,\lambda) &= \frac{1}{\lambda^2}\begin{pmatrix} \int_{\mathbb{R}} \frac{p'^2}{p} dx & \int_{\mathbb{R}} \frac{(x-m)p'^2}{\lambda p} dx \\ \int_{\mathbb{R}} \frac{(x-m)p'^2}{\lambda p} dx & 1 + \int_{\mathbb{R}}\left(\frac{(x-m)^2 p'^2}{\lambda^2 p} + \frac{(x-m)p'}{\lambda}\right) dx \end{pmatrix},
\end{aligned}$$

which illustrates that WIMs are simpler than Fisher information matrices in location-scale families.

Appendix B. Functional inequalities via information matrices

In this section, we explore connections between information matrices and functional inequalities such as log-Sobolev inequalities (LSIs) and Poincaré inequalities (PIs) in statistical models. In section 4, we show that these inequalities are important for the study of statistical efficiency properties.

B.1. **Classical functional inequalities.** Before working in statistical models, we first give a summary of relations among PIs, LSIs and dynamical quantities on density manifold.

Consider the relative entropy (KL-divergence) defined on density manifold:

$$H(\mu|\nu) = \int_{\mathcal{X}} \log\frac{\mu(x)}{\nu(x)}\mu(x)dx, \qquad \mu \in \mathcal{P}(\mathcal{X}).$$

Here, we use a notation $H(\cdot|\cdot)$ in order to be consistent with literature. We recall the definition of log-Sobolev inequality as below.

Definition 17 (Log-Sobolev inequality)**.** *A probability measure ν is said to satisfy a log-Sobolev inequality with constant $\alpha > 0$ (in short: LSI(α)) if we have:*

$$H(\mu|\nu) < \frac{1}{2\alpha} I(\mu|\nu), \qquad \mu \in \mathcal{P}(\mathcal{X}),$$

where the quantity $I(\mu|\nu)$ is the so-called Fisher-information functional

$$I(\mu|\nu) = \int_{\mathcal{X}} \left| \nabla_x \log \frac{\mu(x)}{\nu(x)} \right|^2 \mu(x) dx, \qquad \mu \in \mathcal{P}(\mathcal{X}).$$

Remark 14. If we assume that μ is absolutely continuous w.r.t. the reference measure ν and define function h on $\mathcal{X}$ as:

$$\mu(x) = \frac{h(x)\nu(x)}{\int_{\mathcal{X}} h(x)\nu(x)dx},$$

then above definition of LSI translates to:

$$\begin{aligned}
&\left(H(\mu|\nu) \int_{\mathcal{X}} h(x)\nu(x)dx \right) \\
=&\int_{\mathcal{X}} h(x) \log h(x)\nu(x)dx - \left(\int_{\mathcal{X}} h(x)\nu(x)dx \right) \log \left(\int_{\mathcal{X}} h(x)\nu(x)dx \right) \\
\leq& \frac{1}{2\alpha} \int_{\mathcal{X}} \frac{|\nabla_x h(x)|^2}{h(x)} \nu(x)dx = \frac{1}{2\alpha} \left(I(\mu|\nu) \int_{\mathcal{X}} h(x)\nu(x)dx \right).
\end{aligned}$$

The middle inequality is a more familiar definition of LSI(α). By linearizing above formula with $h = 1 + \epsilon f, \epsilon \to 0$, we get the classical definition of PI(α)

$$\int_{\mathcal{X}} f^2(x)\nu(x)dx \leq \frac{1}{\alpha} \int_{\mathcal{X}} |\nabla_x f(x)|^2 \nu(x)dx, \qquad \int_{\mathcal{X}} f(x)\nu(x)dx = 0.$$

Definition 18 (Poincaré inequalities)**.** *A probability measure ν is said to satisfy a Poincaré inequalities with constant $\alpha > 0$ (in short: PI(α)) if we have:*

$$\int_{\mathcal{X}} f^2(x)\nu(x)dx \leq \frac{1}{\alpha} \int_{\mathcal{X}} |\nabla_x f(x)|^2 \nu(x)dx, \qquad \forall f, \ s.t. \int_{\mathcal{X}} f(x)\nu(x)dx = 0.$$

A sufficient criterion that guarantees LSIs and PIs is related to information matrices (operators or metrics in infinite dimension case) G_W.

Proposition 19. *Denote* $\mathrm{Hess}_W H(\mu|\nu), G_W(\mu)$ *two bi-linear forms correspondent to Hessian of the relative entropy and Wasserstein metric.*

(1) Suppose $\mathrm{Hess}_W H(\mu|\nu) - 2\alpha G_W(\mu)$ *is a semi-positive definite bi-linear form on the Hilbert space $T_\mu \mathcal{P}(\mathcal{X})$, $\forall \mu \in \mathcal{P}(\mathcal{X})$. Then LSI($\alpha$) holds for ν.*

(2) Suppose $\mathrm{Hess}_W H(\nu|\nu) - 2\alpha G_W(\nu)$ *is a semi-positive definite bi-linear form on the Hilbert space $T_\nu \mathcal{P}(\mathcal{X})$. Then PI($\alpha$) holds for ν.*

Proof. First, we prove the result concerned with LSIs. We compute the gradient of the relative entropy w.r.t. Wasserstein metric, which is given by:

$$\mathrm{grad}_W H(\mu|\nu) = -\nabla \cdot \left(\mu \nabla \frac{\delta}{\delta \mu} H(\mu|\nu) \right) = -\nabla \cdot \left(\mu \nabla \log \frac{\mu(x)}{\nu(x)} \right),$$

where $\frac{\delta}{\delta\mu}$ refers to the L^2 functional derivative. Thus it is easy to obtain the relative entropy dissipation along the gradient flow as:

$$\begin{aligned} &\frac{d}{dt}H(\mu|\nu) \\ =& -g_W\left(\mathrm{grad}_W H(\mu|\nu), \mathrm{grad}_W H(\mu|\nu)\right) \\ =& -\int_{\mathcal{X}} \left|\nabla_x \log\frac{\mu(x)}{\nu(x)}\right|^2 \mu(x)dx = -I(\mu|\nu). \end{aligned} \tag{16}$$

Using the assumption, we have:

$$\begin{aligned} \frac{d^2}{dt^2}H(\mu_t|\nu) =& \ \mathrm{Hess}_W H(\mu_t|\nu)\left(\mathrm{grad}_W H(\mu_t|\nu), \mathrm{grad}_W H(\mu_t|\nu)\right) \\ \geq& \ 2\alpha G_W(\mu_t)\left(\mathrm{grad}_W H(\mu_t|\nu), \mathrm{grad}_W H(\mu_t|\nu)\right) \\ =& -2\alpha\frac{d}{dt}H(\mu_t|\nu), \end{aligned}$$

from which LSI(α) holds via integrating the above formula, i.e.

$$\begin{aligned} &I(\mu_t|\nu) = I(\mu_t|\nu) - I(\nu|\nu) \\ =& \int_t^\infty \left(\frac{d^2}{dt^2}H(\mu_\tau|\nu)\right)d\tau \\ \geq& \ 2\alpha\int_t^\infty \left(-\frac{d}{dt}H(\mu_\tau|\nu)\right)d\tau \\ =& \ 2\alpha\left(H(\mu_t|\nu) - H(\nu|\nu)\right) \\ =& \ 2\alpha H(\mu_t|\nu), \end{aligned}$$

where we use the fact that this gradient flow μ_t converges to ν and $H(\nu|\nu) = I(\nu|\nu) = 0$.

To prove the conclusion of Poincaré inequalities, we consider a path in density manifold, i.e $\rho(\epsilon) = \nu(1+\epsilon f), \int_{\mathcal{X}} f(x)\nu(x)dx = 0$. Since we have

$$\begin{aligned} H\left(\rho(\epsilon)|\nu\right) &= \frac{\epsilon^2}{2}\int_{\mathcal{X}} f^2(x)\nu(x)dx + o\left(\epsilon^2\right), \\ -\frac{d}{dt}H(\rho(\epsilon)|\nu) = I\left(\rho(\epsilon)|\nu\right) &= \epsilon^2\int_{\mathcal{X}} |\nabla_x f(x)|^2\nu(x)dx + o\left(\epsilon^2\right). \end{aligned}$$

Consequently, we obtain

$$\begin{aligned}
&\frac{\int_{\mathcal{X}} f^2(x)\nu(x)dx}{\int_{\mathcal{X}} |\nabla_x f(x)|^2 \nu(x)dx} \\
=& \frac{1}{2}\lim_{\epsilon\to 0} -\frac{H\left(\rho\left(\epsilon\right)|\nu\right)}{\frac{d}{d\epsilon}H(\rho\left(\epsilon\right)|\nu)} \\
=& \frac{1}{2}\lim_{\epsilon\to 0} -\frac{\frac{d}{d\epsilon}H\left(\rho\left(\epsilon\right)|\nu\right)}{\frac{d^2}{d\epsilon^2}H(\rho\left(\epsilon\right)|\nu)} \\
=& \frac{1}{2}\lim_{\epsilon\to 0} \frac{G_W\left(\rho\left(0\right)\right)\left(\frac{d}{d\epsilon}\rho\left(0\right), \frac{d}{d\epsilon}\rho\left(0\right)\right)}{\mathrm{Hess}_W H(\rho\left(0\right)|\nu)\left(\frac{d}{d\epsilon}\rho\left(0\right), \frac{d}{d\epsilon}\rho\left(0\right)\right)} \\
\leq& \frac{1}{\alpha},
\end{aligned}$$

where we use L'Hopital's rule in second equality and the third equality holds because of the assumption that $\mathrm{Hess}_W H(\nu|\nu) - 2\alpha G_W(\nu)$ is semi-definite. □

Remark 15. With the help of (16), readers can recognize that LSI guarantees a global exponential convergence of the gradient flow of the relative entropy $H\left(\cdot|\nu\right)$. Indeed, suppose μ_t is a gradient flow of $H(\cdot|\nu)$ starting from μ_0, then we have:

$$H(\mu_t|\nu) \leq e^{-2\alpha t} H(\mu_0|\nu), \qquad \mu_0 \in \mathcal{P}(\mathcal{X}) \text{ (LSI}(\alpha)\text{)}.$$

While intuitively speaking, a PI can be viewed as an infinitesimal version of a LSI, that is to consider the dynamics in a neighborhood of the optimal value.

B.2. **LSIs and PIs in families.** Now, it is clear that PIs and LSIs are related to density manifold. Here, we attempt to find those counterparts in statistical models, i.e. submanifolds.

Now, we fix a model $\Theta \subset \mathcal{P}(\mathcal{X})$ with metric given by G_W. The relative entropy $H\left(\cdot|\nu\right)$ is indeed a restriction of global functional to this family. And we furthermore require the reference measure ν to lie in this family, i.e. $\nu = p_{\theta_*}, \theta_* \in \Theta$. We use $\widetilde{\ }$ to distinguish constraint cases (statistical models) from the global situation (density manifold). Recall that the Fisher information functional is merely the relative entropy dissipation along a gradient flow. Thus we have

$$\begin{aligned}
\widetilde{I}(p_{\theta_t}|p_{\theta_*}) =& -\frac{d}{dt}\widetilde{H}(p_{\theta_t}|p_{\theta_*}) \\
=& \, g_W\left(\mathrm{grad}_W\, \widetilde{H}(p_\theta|p_{\theta_*}), \mathrm{grad}_W\, \widetilde{H}(p_\theta|p_{\theta_*})\right) \\
=& \, \nabla_\theta \widetilde{H}^T \left(\widetilde{G}_W^{-1}\right)^T \widetilde{G}_W \widetilde{G}_W^{-1} \nabla_\theta \widetilde{H} \\
=& \, \nabla_\theta \widetilde{H}^T \widetilde{G}_W^{-1} \nabla_\theta \widetilde{H},
\end{aligned} \tag{17}$$

where we use a fact

$$\mathrm{grad}_W\, \widetilde{H}(p_\theta|p_{\theta_*}) = \widetilde{G}_W^{-1}\nabla_\theta \widetilde{H}.$$

Definition 20 (LSI in family)**.** *Consider a statistical model $p : \mathcal{X} \times \Theta \to \mathbb{R}$, a probability measure p_{θ_*} is said to satisfy LSI(α) in Θ with constant $\alpha > 0$ (in short: LSI(α)) if we*

have:

$$\widetilde{H}(p_\theta|p_{\theta_*}) < \frac{1}{2\alpha}\widetilde{I}(p_\theta|p_{\theta_*}), \qquad \theta \in \Theta.$$

Using information matrices, we seek a sufficient condition for LSIs and PIs as proposition 19:

$$\mathrm{Hess}_W\, \widetilde{H}(p_\theta|p_{\theta_*}) \geq 2\alpha \widetilde{G}_W(\theta),$$

where we have to take care that the Hessian on LHS is calculated in a submanifold instead of density manifold. Fisher information matrix also comes into this picture, via a decomposition of the Hessian term $\mathrm{Hess}_W\, \widetilde{H}(p_\theta|p_{\theta_*})$. This point is known as the Ricci-information-Wasserstein (RIW) condition.

Theorem 21 (RIW-condition)**.** *The information matrices criterion for LSI(α) of distribution $p_{\theta*}$ is given by:*

$$G_F(\theta) + \nabla_\theta^2 p_\theta \log \frac{p_\theta}{p_{\theta_*}} - \Gamma^W \nabla_\theta \widetilde{H}(p_\theta|p_{\theta_*}) \geq 2\alpha G_W(\theta),$$

where Γ^Ws are Christoffel symbols in Wasserstein statistical model Θ, while for PI(α) of distribution $p_{\theta*}$ can be written as:

$$G_F(\theta) + \nabla_\theta^2 p_\theta \log \frac{p_\theta}{p_{\theta_*}} \geq 2\alpha G_W(\theta).$$

Remark 16. It can be seen that the condition for log-Sobolev inequalities is much more complicated than that of Poincaré inequalities. For LSIs require a global convexity of the entropy while PIs only correspond to local behavior at the minimum. The most significant change takes place in the Hessian term of entropy, where Wasserstein Christoffel symbols come in.

B.3. **Examples in 1-d Family.**] Both LSIs and PIs can be proved by using Wasserstein and Fisher information matrices. Previously, we have done geometric computations on metric tensor and Hessian of the entropy. This prepares ingredients for us to establish inequalities in families of probability distributions. In this section, we utilize previous calculations to obtain concrete bounds on these functional inequalities.

Example 14 (Gaussian distribution)**.** Recall that for a Gaussian distribution with mean value μ and standard variance σ, the Wasserstein and Fisher information matrices are given by:

$$G_W(\mu,\sigma) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad G_F(\mu,\sigma) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}.$$

The entropy and the relative entropy defined on this model are provided by:

$$\begin{aligned} \widetilde{H}(\mu,\sigma) &= -\frac{1}{2}\log 2\pi - \log\sigma - \frac{1}{2}, \\ \widetilde{H}(\mu,\sigma|p_*) &= -\log\sigma + \log\sigma_* - \frac{1}{2} + \frac{\sigma^2 + (\mu-\mu_*)^2}{2\sigma_*^2}, \qquad p_* \sim p_{\mu_*,\sigma_*}. \end{aligned}$$

We can calculate Wasserstein gradients associated with these two functionals:

$$\nabla^W_{\mu,\sigma}\widetilde{H}(\mu,\sigma) = \begin{pmatrix} 0 \\ -\frac{1}{\sigma} \end{pmatrix}, \quad \nabla^W_{\mu,\sigma}\widetilde{H}(\mu,\sigma|p_*) = \begin{pmatrix} \frac{\mu-\mu_*}{\sigma_*^2} \\ -\frac{1}{\sigma}+\frac{\sigma}{\sigma_*^2} \end{pmatrix},$$

with the correspondent Fisher information functionals as:

$$\widetilde{I}(\mu,\sigma) = \frac{1}{\sigma^2},$$
$$\widetilde{I}(\mu,\sigma|p_*) = \frac{(\mu-\mu_*)^2}{\sigma_*^4} + \left(-\frac{1}{\sigma}+\frac{\sigma}{\sigma_*^2}\right)^2.$$

Thus, the LSI(α) for Gaussian p_{μ_*,σ_*} is given by

$$\frac{(\mu-\mu_*)^2}{\sigma_*^4} + \left(-\frac{1}{\sigma}+\frac{\sigma}{\sigma_*^2}\right)^2 \geq 2\alpha\left(-\log\sigma+\log\sigma_* - \frac{1}{2} + \frac{\sigma^2+(\mu-\mu_*)^2}{2\sigma_*^2}\right).$$

Next, we move onto the derivation of the RIW condition. It suffices to consider a relation between $\widetilde{G}_W, \mathrm{Hess}_W\,\widetilde{H}$ at each point in a statistical model. Recall the formula for Hessian in Riemannian geometry:

$$(\mathrm{Hess}\, f)_{ij} = \partial_i\partial_j f - \Gamma_{ij}^{k(W)}\partial_k f,$$

where Γ^Ws are Christoffel symbols in Wasserstein geometry. In Wasserstein Gaussian model where the metric is Euclidean, Christoffel symbols vanish, i.e. $\Gamma^W = 0$. Thus we have:

$$\mathrm{Hess}_W\,\widetilde{H}\,(\mu,\sigma) = \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{\sigma^2} \end{pmatrix}, \qquad \mathrm{Hess}_W\,\widetilde{H}\,(\mu,\sigma|p_*) = \begin{pmatrix} \frac{1}{\sigma_*^2} & 0 \\ 0 & \frac{1}{\sigma^2}+\frac{1}{\sigma_*^2} \end{pmatrix}.$$

For a gradient flow of the relative entropy w.r.t. a Gaussian p_{μ_*,σ_*}, we conclude that

$$\mathrm{Hess}_W\,\widetilde{H}(\mu,\sigma|p_{\theta_*}) \geq \left(\frac{1}{\sigma_*^2}\right) G_W(\mu,\sigma),$$

since $G_W(\mu,\sigma)$ is exactly an identity matrix. In other words, the Gaussian p_{μ_*,σ_*} satisfies a LSI$\left(\frac{1}{2\sigma_*^2}\right)$ in a Gaussian model. Notice this result coincides with the one in global case.

Next, for the gradient flow of the entropy $\widetilde{H}\,(\cdot)$, we do not have a satisfying constant α such that the Hessian condition proposition 19 holds. For $\mathrm{Hess}_W\,\widetilde{H}(\mu,\sigma)$ matrix has an eigenvalue 0. Despite of this, we have:

$$\mathrm{grad}_W\,\widetilde{H}(\mu,\sigma) = G_W^{-1}\nabla_{\mu,\sigma}\widetilde{H}(\mu,\sigma) = \nabla_{\mu,\sigma}\widetilde{H}(\mu,\sigma),$$

whose μ component always vanishes. Thus the gradient direction of $\widetilde{H}(\cdot)$ always coincides with σ direction, in which we have eigenvalue's bound: $\mathrm{eig}_\sigma(\widetilde{H}) \geq \frac{1}{\sigma^2}\,\mathrm{eig}_\sigma(G_W)$. This refers to that eigenvalues of two matrices correspond to direction $\frac{\partial}{\partial\mu}$ have a bound. For LSIs, if the range of σ is the whole $\mathbb{R}$, then it is easy to see there will not exist a satisfying constant $\alpha > 0$ for LSI(α) to hold, i.e. $\frac{1}{\sigma^2} \geq 2\alpha,\ \forall\sigma\in\mathbb{R}$. However, if we restrict the range of σ to a bounded region such as $[-M, M]$, then LSI($\frac{1}{2M^2}$) will hold.

Remark 17. Above calculation on gradient flows of the entropy does not establish LSI(α) for any specific distribution. It merely provides an example of using Hessian condition to study dynamical behaviors.

Example 15 (Laplacian distribution)**.** Consider the case of Laplacian distribution, where

$$G_W(m,\lambda)=\begin{pmatrix}1 & 0\\ 0 & \frac{2}{\lambda^4}\end{pmatrix},\quad G_F(m,\lambda)=\begin{pmatrix}\lambda^2 & 0\\ 0 & \frac{1}{\lambda^2}\end{pmatrix},$$

from which we can calculate the Christoffel symbol as:

$$\Gamma_{22}^{2(W)}(m,\lambda)=\frac{g_{22}^{-1}}{2}(\partial_2 g_{22}+\partial_2 g_{22}-\partial_2 g_{22})=\frac{g_{22}^{-1}}{2}\partial_2 g_{22}=\frac{\lambda^4}{4}\cdot\left(-\frac{8}{\lambda^5}\right)=-\frac{2}{\lambda},$$
$$\Gamma_{ij}^{k(W)}(m,\lambda)=0\qquad\text{otherwise.}$$

Following the same procedure we have done before, the entropy and the relative entropy w.r.t. p_{m_*,λ_*} defined on this model is provided by:

$$\widetilde{H}(m,\lambda)=-1+\log\lambda-\log 2,$$
$$\widetilde{H}(m,\lambda|p*)=-1+\log\lambda-\log\lambda_*+\lambda_*\left|m-m_*\right|+\frac{\lambda_* e^{-\lambda|m-m_*|}}{\lambda},$$

from which we can calculate Wasserstein gradients associated with two functionals:

$$\nabla^W_{m,\lambda}\widetilde{H}(m,\lambda)=\begin{pmatrix}0\\ \dfrac{1}{\lambda}\end{pmatrix},$$
$$\nabla^W_{m,\lambda}\widetilde{H}(m,\lambda|p_*)=\begin{cases}\begin{pmatrix}\lambda_*\left(1-e^{-\lambda(m-m_*)}\right)\\ -\dfrac{\left(\lambda\left(m-m_*\right)+1\right)\lambda_* e^{-\lambda(m-m_*)}-\lambda}{\lambda^2}\end{pmatrix}, & m>m_*,\\ \begin{pmatrix}-\lambda_*\left(1-e^{-\lambda(m_*-m)}\right)\\ -\dfrac{\left(\lambda\left(m_*-m\right)+1\right)\lambda_* e^{-\lambda(m_*-m)}-\lambda}{\lambda^2}\end{pmatrix}, & m<m_*,\end{cases}$$

with the Fisher information functionals as:

$$\widetilde{I}(m,\lambda)=\frac{\lambda^2}{2},$$
$$\widetilde{I}(m,\lambda|p_*)=\lambda_*^2\left(1-e^{-\lambda|m-m_*|}\right)^2+\frac{\left(\left(\lambda\left|m-m_*\right|+1\right)\lambda_* e^{-\lambda|m-m_*|}-\lambda\right)^2}{2}.$$

Notice that the value of $\nabla^W_{m,\lambda}\widetilde{H}(m,\lambda|p_*)$ is not well-defined at point $m=m_*$. However, what we considered is integral on the whole $\mathbb{R}$. Thus we can simply ignore its value at $m=m_*$. As before, LSI(α) is given by

$$\lambda_*^2\left(1-e^{-\lambda|m_*-m|}\right)^2+\frac{\left(\left(\lambda\left|m_*-m\right|+1\right)\lambda_* e^{-\lambda|m_*-m|}-\lambda\right)^2}{2}$$
$$\geq 2\alpha\left(-1+\log\lambda-\log\lambda_*+\lambda_*\left|m-m_*\right|+\frac{\lambda_* e^{-\lambda|m-m_*|}}{\lambda}\right).$$

And we find Hessians of the entropy and the relative entropy in (Θ, G_W) are given by:

$$\begin{aligned}
\mathrm{Hess}_W \widetilde{H}(m,\lambda) &= \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{\lambda^2} \end{pmatrix}, \\
\mathrm{Hess}_W \widetilde{H}(m,\lambda|p_*) &= \begin{pmatrix} \lambda\lambda_* e^{-\lambda|m-m_*|} & 0 \\ 0 & \frac{1}{\lambda^2} + \frac{\lambda_* e^{-\lambda|m-m_*|}(m_*-m)^2}{\lambda^3} \end{pmatrix}.
\end{aligned}$$

Following the same analysis, we conclude that for gradient flows of the entropy $\widetilde{H}(m,\lambda)$, a LSI($\frac{\lambda^2}{4}$) holds. While for the relative entropy $H(m,\lambda|p_{m_*,\lambda_*})$, Hessian condition can be written as

$$\begin{pmatrix} \lambda\lambda_* e^{-\lambda|m-m_*|} & 0 \\ 0 & \frac{1}{\lambda^2} + \frac{\lambda_* e^{-\lambda|m-m_*|}(m_*-m)^2}{\lambda^3} \end{pmatrix} \geq \alpha \begin{pmatrix} 1 & 0 \\ 0 & \frac{2}{\lambda^4} \end{pmatrix},$$

which can be reformulated as

$$\alpha = \min_{m,\lambda} \left\{ \lambda\lambda_* e^{-\lambda|m-m_*|}, \frac{1}{2}\left(\lambda^2 + \lambda_* e^{-\lambda|m-m_*|}\lambda(m_*-m)^2\right) \right\}.$$

From above formula, we conclude that in order to find a satisfying constant, it suffices to restrict the region of $m \in [-M, M], \lambda \in [N, \infty)$. The distribution $\mathrm{La}(m_*, \lambda_*)$ satisfies a LSI(α) in Laplacian family with α given above.

Example 16 (Independent model)**.** For an independent family $p(x,y;\theta) = p_1(x;\theta)\, p(y;\theta)$, we have

$$G_W = G_W^1 + G_W^2, \qquad G_F = G_F^1 + G_F^2.$$

The entropy and the relative entropy also have this separability property

$$\begin{aligned}
\widetilde{H}(\theta) &= \widetilde{H}_1(\theta) + \widetilde{H}_2(\theta), \\
\widetilde{H}(\theta|p_*) &= \widetilde{H}_1(\theta|p_{1*}) + \widetilde{H}_2(\theta|p_{2*}), \\
\nabla_\theta \widetilde{H}(\theta|p_*) &= \nabla_\theta \widetilde{H}_1(\theta|p_{1*}) + \nabla_\theta \widetilde{H}_2(\theta|p_{2*}).
\end{aligned}$$

The Fisher information functional is given by

$$\begin{aligned}
& I(p_\theta|p_*) \\
=\ & \left(\nabla_\theta \widetilde{H}_1(\theta|p_{1*}) + \nabla_\theta \widetilde{H}_2(\theta|p_{2*})\right)^T \left(G_W^1 + G_W^2\right)^{-1} \left(\nabla_\theta \widetilde{H}_1(\theta|p_{1*}) + \nabla_\theta \widetilde{H}_2(\theta|p_{2*})\right),
\end{aligned}$$

with LSI(α) given by

$$\begin{aligned}
& \left(\nabla_\theta \widetilde{H}_1(\theta|p_{1*}) + \nabla_\theta \widetilde{H}_2(\theta|p_{2*})\right)^T \left(G_W^1 + G_W^2\right)^{-1} \left(\nabla_\theta \widetilde{H}_1(\theta|p_{1*}) + \nabla_\theta \widetilde{H}_2(\theta|p_{2*})\right) \\
\geq\ & 2\alpha \nabla_\theta \widetilde{H}_1(\theta|p_{1*}) + \nabla_\theta \widetilde{H}_2(\theta|p_{2*}).
\end{aligned}$$

In conclusion, above examples introduce another way to prove functional inequalities as well as convergence rates of dynamics in probability families.

Appendix C. Proofs in section 4

C.1. **Proof of Theorem 13.**

Proof of Theorem 13. First, we postulate that ∇_x refers to the gradient w.r.t. x variable while ∇_θ refers to the gradient w.r.t. θ variable. We expand the function $f(x_t, \theta_t)$

$$f(x_t, \theta_t) = f(x_t, \theta_*) + \nabla_\theta f(x_t, \theta_*)(\theta_t - \theta_*) + O\left(|\theta_t - \theta_*|^2\right).$$

By substrating θ_* in both sides of the updating equation and plugging in the expansion above, we get:

$$\begin{aligned}\theta_{t+1} - \theta_* = &\ (\theta_t - \theta_*) - \frac{1}{t} G_W^{-1}(\theta_t)\left(f(x_t, \theta_*) + \nabla_\theta f(x_t, \theta_*)(\theta_t - \theta_*)\right.\\ &\left. + O\left(|\theta_t - \theta_*|^2\right)\right).\end{aligned}$$

Then, taking Wasserstein covariances of both sides, we get:

$$\begin{aligned}V_{t+1} = V_t &+ \frac{1}{t^2}\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(G_W^{-1}(\theta_t) f(x_t, \theta_t)\right) \cdot \nabla_x\left(f(x_t, \theta_t)^T G_W^{-1}(\theta_t)\right)\right]\\ &- \frac{2}{t}\mathbb{E}_{p_{\theta_*}}\left[\nabla_x(\theta_t - \theta_*) \cdot \nabla_x\left(f(x_t, \theta_*)^T G_W^{-1}(\theta_t)\right)\right] + o\left(\frac{V_t}{t}\right)\\ &- \frac{2}{t}\mathbb{E}_{p_{\theta_*}}\left[\nabla_x(\theta_t - \theta_*) \cdot \nabla_x\left((\theta_t - \theta_*)^T \nabla_\theta f(x_t, \theta_*)^T G_W^{-1}(\theta_t)\right)\right],\end{aligned}$$

where the last term corresponds to an expansion term $O\left(|\theta_t - \theta_*|^2\right)$ and we use an assumption that $\mathbb{E}_{p_{\theta_*}}\left[(\theta_t - \theta_*)^2\right], \mathbb{E}_{p_{\theta_*}}\left[|\nabla_x(\theta_t - \theta_*)|^2\right] = o(1)$. In above formula, we eliminate transpose symbols ^{T} on metric tensor G_W because of its symmetry. For the second term on the RHS, we have:

$$\begin{aligned}&\frac{1}{t^2}\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(G_W^{-1}(\theta_t) f(x_t, \theta_t)\right) \cdot \nabla_x\left(f(x_t, \theta_t)^T G_W^{-1}(\theta_t)\right)\right]\\ =&\ \frac{1}{t^2}\mathbb{E}_{p_{\theta_*}}\left[G_W^{-1}(\theta_*)\nabla_x\left(f(x_t, \theta_*)\right) \cdot \nabla_x\left(f(x_t, \theta_*)^T\right) G_W^{-1}(\theta_*)\right] + o\left(\frac{1}{t^2}\right)\\ =&\ \frac{1}{t^2} G_W^{-1}(\theta_*)\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(f(x_t, \theta_*)\right) \cdot \nabla_x\left(f(x_t, \theta_*)^T\right)\right] G_W^{-1}(\theta_*) + o\left(\frac{1}{t^2}\right),\end{aligned}$$

where we use the following fact

$$\begin{aligned}&\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(G_W^{-1}(\theta_t) f(x_t, \theta_t)\right) \cdot \nabla_x\left(f(x_t, \theta_t)^T G_W^{-1}(\theta_t)\right)\right]\\ &- \mathbb{E}_{p_{\theta_*}}\left[G_W^{-1}(\theta_*)\nabla_x\left(f(x_t, \theta_*)\right) \cdot \nabla_x\left(f(x_t, \theta_*)^T\right) G_W^{-1}(\theta_*)\right]\\ =&\ O\left(\mathbb{E}_{p_{\theta_*}}|\theta_t - \theta_*|\right) = o(1).\end{aligned}$$

And the third term in the RHS can be reduced according to:

$$\begin{aligned}
&-\frac{2}{t}\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(\theta_t-\theta_*\right)\cdot\nabla_x\left(f(x_t,\theta_*)^T G_W^{-1}(\theta_t)\right)\right]\\
=&-\frac{2}{t}\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(\theta_t-\theta_*\right)\cdot\nabla_x\left(f(x_t,\theta_*)^T\right) G_W^{-1}(\theta_t)\right]\\
&-\frac{2}{t}\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(\theta_t-\theta_*\right)\cdot f(x_t,\theta_*)^T\nabla_x G_W^{-1}(\theta_t)\right]\\
=&\,0,
\end{aligned}$$

where the first term vanishes because $\nabla_x\left(\theta_t-\theta_*\right)$ only has non-vanishing components at $x_1,...,x_{t-1}$ while $\nabla_x\left(f(x_t,\theta_*)^T\right)$ only has a non-vanishing component at x_t. Consequently their inner product vanishes everywhere. While the second term vanishes by considering each element of this matrix, we have:

$$\begin{aligned}
&\left(\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(\theta_t-\theta_*\right)\cdot f(x_t,\theta_*)^T\nabla_x G_W^{-1}(\theta_t)\right]\right)_{ij}\\
=&\frac{2}{t}\mathbb{E}_{p_{\theta_*}}\nabla_x\left(\theta_t-\theta_*\right)_i\cdot\left(f(x_t,\theta_*)^T\nabla_x G_W^{-1}(\theta_t)\right)_j\\
=&\frac{2}{t}\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(\theta_t-\theta_*\right)_i\cdot\nabla_x\left(G_W^{-1}(\theta_t)_{kj}\right)f(x_t,\theta_*)_k^T\right]\\
=&\frac{2}{t}\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(\theta_t-\theta_*\right)_i\cdot\nabla_x\left(G_W^{-1}(\theta_t)_{kj}\right)\right]\mathbb{E}_{p_{\theta_*}}f(x_t,\theta_*)_k^T\\
=&\,0,
\end{aligned}$$

where the third equality is guaranteed by the fact that $\theta_t-\theta_*$ is independent to $\nabla_\theta f(x_t,\theta_*)$ since θ_t, x_t are mutually independent. While the last equality holds by an assumption:

$$\mathbb{E}_{p_{\theta_*}}f(x_t,\theta_*)=0.$$

For the last term, same as the analysis of the third term, we find:

$$\begin{aligned}
&-\frac{2}{t}\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(\theta_t-\theta_*\right)\cdot\nabla_x\left(\left(\theta_t-\theta_*\right)^T\nabla_\theta f(x_t,\theta_*)^T G_W^{-1}(\theta_t)\right)\right]\\
=&-\frac{2}{t}\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(\theta_t-\theta_*\right)\cdot\nabla_x\left(\left(\theta_t-\theta_*\right)^T\right)\nabla_\theta f(x_t,\theta_*)^T G_W^{-1}(\theta_t)\right]\\
&-\frac{2}{t}\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(\theta_t-\theta_*\right)\cdot\left(\theta_t-\theta_*\right)^T\nabla_x\left(\nabla_\theta f(x_t,\theta_*)^T\right)G_W^{-1}(\theta_t)\right]\\
&-\frac{2}{t}\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(\theta_t-\theta_*\right)\cdot\left(\theta_t-\theta_*\right)^T\nabla_\theta f(x_t,\theta_*)^T\nabla_x\left(G_W^{-1}(\theta_t)\right)\right]\\
=&-\frac{2}{t}\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(\theta_t-\theta_*\right)\cdot\nabla_x\left(\left(\theta_t-\theta_*\right)^T\right)\nabla_\theta f(x_t,\theta_*)^T G_W^{-1}(\theta_*)\right]+o(\frac{V_t}{t})\\
&-\frac{2}{t}\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(\theta_t-\theta_*\right)\cdot\left(\theta_t-\theta_*\right)^T\nabla_\theta f(x_t,\theta_*)^T\nabla_x\left(G_W^{-1}(\theta_t)\right)\right],
\end{aligned}$$

where we again use the independent relation between $(\theta_t-\theta_*)$ and $f(x_t,\theta_*)$. The additional term appearing above, with the help that $\mathbb{E}_{p_{\theta_*}}\left[\nabla_\theta f(x_t,\theta_*)\right]=O(1)$, $\nabla_x\left(G_W^{-1}(\theta_t)\right)=$

$\nabla_x\theta_t\nabla_\theta\left(G_W^{-1}(\theta_t)\right) = O(\nabla_x(\theta_t-\theta_*))$, can be further reduced to the form below:

$$\begin{aligned}
&\frac{2}{t}\mathbb{E}_{p_{\theta_*}}\left[\nabla_x(\theta_t-\theta_*)\cdot(\theta_t-\theta_*)^T\nabla_\theta f(x_t,\theta_*)\nabla_x\left(\left(G_W^{-1}(\theta_t)\right)\right)\right]\\
=&\frac{2}{t}\mathbb{E}_{p_{\theta_*}}\left[\nabla_x(\theta_t-\theta_*)\cdot(\theta_t-\theta_*)^T O(1)\nabla_x(\theta_t-\theta_*)O(1)\right]\\
\leq&\frac{O(1)}{t}\sqrt{\mathbb{E}_{p_{\theta_*}}\left[\left|\nabla_x(\theta_t-\theta_*)\right|^2\right]\mathbb{E}_{p_{\theta_*}}\left[(\theta_t-\theta_*)^2\right]\mathbb{E}_{p_{\theta_*}}\left[\left|\nabla_x(\theta_t-\theta_*)\right|^2\right]}\\
=&\,o\left(\frac{V_t}{t}\right).
\end{aligned}$$

And the last term finally reduces to:

$$\begin{aligned}
&-\frac{2}{t}\mathbb{E}_{p_{\theta_*}}\left[\nabla_x(\theta_t-\theta_*)\cdot\nabla_x\left((\theta_t-\theta_*)^T\right)\nabla_\theta f(x_t,\theta_*)\left(G_W^{-1}(\theta_*)\right)\right]+o(\frac{V_t}{t})\\
=&-\frac{2V_t}{t}\mathbb{E}_{p_{\theta_*}}\left[\nabla_\theta f(x_t,\theta_*)\right]G_W^{-1}(\theta_*)+o(\frac{V_t}{t}).
\end{aligned}$$

Combining all the terms we have in hand, we derive the following updating equation for Wasserstein covariances during a natural gradient descent:

$$\begin{aligned}
V_{t+1} = V_t &+ \frac{1}{t^2}G_W^{-1}(\theta_*)\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(f(x_t,\theta_*)\right)\cdot\nabla_x\left(f(x_t,\theta_*)^T\right)\right]\left(G_W^{-1}(\theta_*)\right)\\
&-\frac{2V_t}{t}\mathbb{E}_{p_{\theta_*}}\left[\nabla_\theta f(x_t,\theta_*)\right]G_W^{-1}(\theta_*)+o(\frac{V_t}{t})+o\left(\frac{1}{t^2}\right)+O(\frac{V_t}{t^2}).
\end{aligned}$$

□

Remark 18. The most frequently used tools in this proof is a separability property, c.f. proposition 5. The key observation here is that, for two statistics T_1, T_2 which depend on (independent) different variables, such as $T_1 = T_1(x_1, ..., x_{t-1})$, $T_2 = T_2(x_t, ..., x_{t+n})$ are "orthogonal" in both Wasserstein and Fisher metrics. Specifically, consider gradients of T_1, T_2 w.r.t. x, since they depend on different variables, thus

$$\mathrm{Cov}^W[T_1,T_2] = \mathbb{E}_{p_{\theta_*}}[\nabla_x T_1\cdot\nabla_x T_2] = 0.$$

This type of separability is a direct analog of the one in Fisher-Rao geometry:

$$\mathrm{Cov}^F[T_1,T_2] = \mathbb{E}_{p_{\theta_*}}[T_1T_2] = \mathbb{E}_{p_{\theta_*}}[T_1]\cdot\mathbb{E}_{p_{\theta_*}}[T_2] = 0.$$

C.2. **Examples and numerical experiments of two efficiencies.**

Example 17 (Gaussian distribution)**.** Consider the Gaussian distribution with mean value μ and standard variance σ:

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

The WIM satisfies

$$G_W(\mu,\sigma) = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}.$$

The Fisher information matrix satisfies

$$G_F(\mu,\sigma) = \begin{pmatrix}\frac{1}{\sigma^2} & 0\\ 0 & \frac{2}{\sigma^2}\end{pmatrix}.$$

Further, the matrix $G_F G_W^{-1}$ is given by:

$$G_F(\mu,\sigma)G_W^{-1}(\mu,\sigma) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}.$$

And optimal parameters are given by μ_*, σ_*. Thus we have following conclusions on efficiency of the Fisher, Wasserstein natural gradients and Wasserstein natural gradient on Fisher score (maximal likelihood estimator).

The Wasserstein natural gradient is asymptotically efficient with an asymptotic Wasserstein covariance given by:

$$V_t = \frac{1}{t}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + O\left(\frac{1}{t^2}\right).$$

The Fisher natural gradient is asymptotic efficient with an asymptotic classical covariance given by:

$$V_t = \frac{1}{t}\begin{pmatrix} \sigma_*^2 & 0 \\ 0 & \frac{\sigma_*^2}{2} \end{pmatrix} + O\left(\frac{1}{t^2}\right).$$

An interesting thing here is that the covariance matrix appears in the Wasserstein efficiency is independent of the optimal value. While in Fisher case, the asymptotic behavior depends a lot on the optimal parameter we obtain.

For the last case, in Gaussian family, two metric tensors G_F, G_W can be simultaneously diagonalized, thus the situation is even simpler. We denote the least significant eigenvalue of $G_F G_W^{-1}$ as α:

$$\alpha = \frac{1}{\sigma_*^2}.$$

Further more, we have to figure out the term

$$\mathbb{E}_{p_{\mu_*,\sigma_*}}\left[\nabla_x\left(\nabla_{\mu_*,\sigma_*} l(x_t,\mu_*,\sigma_*)\right)\cdot\nabla_x\left(\nabla_{\mu_*,\sigma_*} l(x_t,\mu_*,\sigma_*)^T\right)\right],$$

that appears in the final result. In Gaussian, since we have Fisher scores $\nabla_{\mu_*,\sigma_*} l(x;\theta) = \Phi^F(x,\mu_*,\sigma_*)$ as:

$$\Phi^F_\mu(x;\mu,\sigma) = \frac{x-\mu}{\sigma^2}, \quad \Phi^F_\sigma(x;\mu,\sigma) = \frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma}.$$

Via calculation, we have

$$\begin{aligned}
\mathbb{E}_{p_{\mu_*,\sigma_*}}\left[\nabla_x\Phi^F_\mu(x;\mu_*,\sigma_*)\cdot\nabla_x\left(\Phi^F_\mu(x;\mu_*,\sigma_*)^T\right)\right] &= \mathbb{E}_{p_{\mu_*,\sigma_*}}\left[\frac{1}{\sigma_*^4}\right] = \frac{1}{\sigma_*^4}, \\
\mathbb{E}_{p_{\mu_*,\sigma_*}}\left[\nabla_x\Phi^F_\mu(x;\mu_*,\sigma_*)\cdot\nabla_x\left(\Phi^F_\sigma(x;\mu_*,\sigma_*)^T\right)\right] &= \mathbb{E}_{p_{\mu_*,\sigma_*}}\left[\frac{1}{\sigma_*^2}\cdot\frac{2(x-\mu_*)}{\sigma_*^3}\right] = 0, \\
\mathbb{E}_{p_{\mu_*,\sigma_*}}\left[\nabla_x\Phi^F_\sigma(x;\mu_*,\sigma_*)\cdot\nabla_x\left(\Phi^F_\sigma(x;\mu_*,\sigma_*)^T\right)\right] &= \mathbb{E}_{p_{\mu_*,\sigma_*}}\left[\frac{4\left(x-\mu_*\right)^2}{\sigma_*^6}\right] = \frac{4}{\sigma_*^4},
\end{aligned}$$

we conclude the middle term is given by

$$\mathfrak{I} = \mathbb{E}_{p_{\mu_*,\sigma_*}}\left[\nabla_x\left(\nabla_{\mu_*,\sigma_*} l(x_t,\mu_*,\sigma_*)\right)\cdot\nabla_x\left(\nabla_{\mu_*,\sigma_*} l(x_t,\mu_*,\sigma_*)^T\right)\right] = \begin{pmatrix} \frac{1}{\sigma_*^4} & 0 \\ 0 & \frac{4}{\sigma_*^4} \end{pmatrix}.$$

And when we have $\frac{2}{\sigma_*^2} > 1$, the inverse matrix of $2B - \mathbf{I}$ is given by

$$(2B - \mathbf{I})^{-1} = \begin{pmatrix} \frac{\sigma_*^2}{2-\sigma_*^2} & 0 \\ 0 & \frac{\sigma_*^2}{4-\sigma_*^2} \end{pmatrix}.$$

Consequently, the term appearing in the asymptotic behavior of the Poincaré efficiency is given by

$$\begin{aligned} &\frac{1}{t}\left(2G_F G_W^{-1} - \mathbf{I}\right)^{-1} G_W^{-1}(\theta_*)\mathfrak{I}\left(G_W^{-1}(\theta_*)\right) \\ =& \begin{pmatrix} \frac{\sigma_*^2}{2-\sigma_*^2} & 0 \\ 0 & \frac{\sigma_*^2}{4-\sigma_*^2} \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma_*^4} & 0 \\ 0 & \frac{4}{\sigma_*^4} \end{pmatrix} \\ =& \begin{pmatrix} \frac{1}{(2-\sigma_*^2)\sigma_*^2} & 0 \\ 0 & \frac{4}{(4-\sigma_*^2)\sigma_*^2} \end{pmatrix}. \end{aligned}$$

Thus the asymptotic behavior the Wasserstein covariance in the Wasserstein natural gradient of Fisher scores is given by:

$$V_t = \begin{cases} O\left(t^{-\frac{2}{\sigma_*^2}}\right), & \frac{1}{\sigma_*^2} \leq \frac{1}{2}, \\ \frac{1}{t}\begin{pmatrix} \frac{1}{(2-\sigma_*^2)\sigma_*^2} & 0 \\ 0 & \frac{4}{(4-\sigma_*^2)\sigma_*^2} \end{pmatrix} + O(\frac{1}{t^2}), & \frac{1}{\sigma_*^2} > \frac{1}{2}. \end{cases}$$

We verify our theory by following numerical experiments. In two cases, we verify two kinds of efficiency, namely the Wasserstein-Cramer-Rao efficiency and the Poincaré efficiency respectively. In the first experiment, we verify the constant G_W^{-1} appearing in asymptotic efficiency of the Wasserstein natural gradient. While for the other situation we verify the asymptotic exponential index α showing up in Poincaré efficiency.

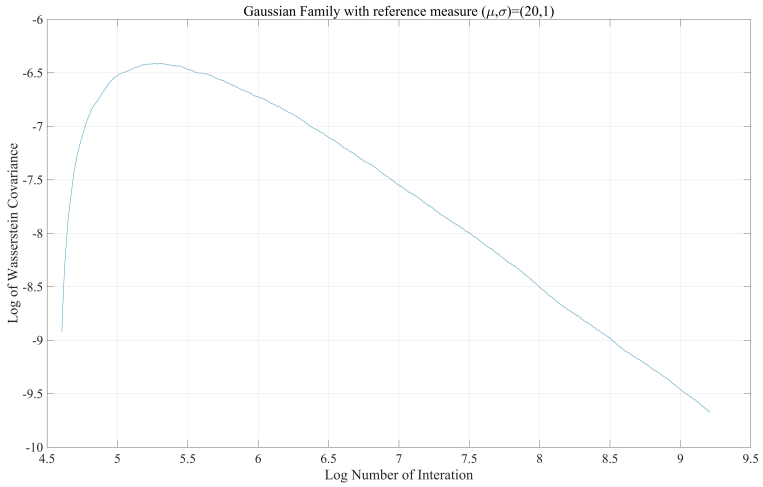

FIGURE 2. The Wasserstein-Cramer-Rao Type Convergence Rate. Here x-axis represents the logarithm of iteration t while y-axis represents the logarithm of Wasserstein covariance V_t. We take the reference measure in KL-divergence to be Gaussian $\mathcal{N}(20, 1)$ where the parameter $\mu_* = 20$ is the optimal point we aim to estimate. Since we have $\frac{1}{\sigma_*^2} = 1 > \frac{1}{2}$, the Cramer-Rao type convergence holds.

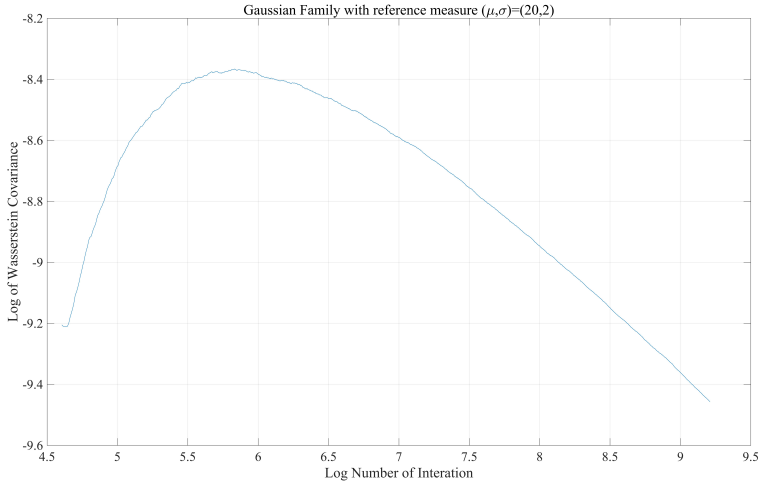

FIGURE 3. Poincaré Type Convergence Rate. Here x-axis represents the logarithm of iteration t while y-axis represents the logarithm of Wasserstein covariance V_t. We take the reference measure in KL-divergence to be Gaussian $\mathcal{N}(20, 1)$ where the parameter $\mu_* = 20$ is the optimal point we aim to estimate. Since we have $\frac{1}{\sigma_*^2} = \frac{1}{4} < \frac{1}{2}$, the Poincaré type convergence holds.

E-mail address: wcli@math.ucla.edu,zjx98math@gmail.com