WASSERSTEIN INFORMATION MATRIX

WUCHEN LI AND JIAXI ZHAO

Abstract. We study information matrices for statistical models by the L^2-Wasserstein metric. We call it Wasserstein information matrix (WIM), which is an analog of the classical Fisher information matrix. Based on this matrix, we introduce Wasserstein score functions and study covariance operators in statistical models. Using them, we establish a Wasserstein-Cramer-Rao bound for estimation. Also, by a ratio of the Wasserstein and Fisher information matrices, we prove various functional inequalities within statistical models, including both the Log-Sobolev and Poincaré inequalities. These inequalities relate to a new efficiency property named Poincaré efficiency, introduced via a Wasserstein natural gradient of maximal likelihood estimation. Furthermore, online efficiency for Wasserstein natural gradient methods is also established. Several analytical examples and approximations of the WIM are presented, including location-scale families, independent families, Gaussian mixture models, and rectified linear unit (ReLU) generative models.

1. Introduction

The 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 theory of estimation, including the design and analysis of estimators. In particular, the maximal likelihood principle is one of the most important examples. It connects the Fisher information matrix to another key 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 the Fisher-Rao metric in information geometry [3]. It studies divergence functions and their invariance properties by using the Fisher information matrix [3]. Furthermore, the Fisher information matrix is also useful for statistical optimization problems. In particular, the natural gradient method [2] rectifies the gradient direction by the Fisher information matrix. It is shown that the natural gradient method is asymptotically online Fisher-efficient.

On the other hand, optimal transport, in particular, Wasserstein geometry [12, 21], introduces the other metric in probability space [24, 25]. 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 KL divergence with Fisher information functional [21], known as de Bruijn identities [27].

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

Many concentration inequalities such as the Log-Sobolev inequality (LSI) and Poincaré inequality (PI) arise naturally in the Wasserstein geometry [22] on full probability space.

Despite various studies of Wasserstein geometry in global 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 Wasserstein geometry? Compared to Fisher information matrices (Fisher statistics), what are counterparts of information matrices, score functions, Cramer-Rao bound, and online efficiency of natural gradient methods in Wasserstein statistics? Moreover, can this theory provides novel tools for machine learning statistical models, especially for implicit models?*

In this paper, following key ideas in [13], we positively answer the above questions by introducing the Wasserstein information matrix (WIM). We derive it by pulling-back the Wasserstein metric in full probability space to finite-dimensional parametric statistical models [16, 17]. We show that the WIM defines Wasserstein score functions with a Wasserstein covariance operator of an estimator. Based on them, Wasserstein-Cramer-Rao bound is derived. Furthermore, combining WIM with the Wasserstein score function, we recover the asymptotic efficiency property of the online Wasserstein natural gradient methods.

Meanwhile, by comparing the Wasserstein and Fisher information matrices, we naturally prove several concentration inequalities such as LSI and PI within statistical models. Parallel to the study in global space, we further decompose the Hessian term and study the Ricci information Wasserstein (RIW) criterion for the LSI and PI 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. We give convergence rate analysis of 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 deep generative models. For example, we carefully study the one-dimensional probability family generated by a push-forward map based on the ReLU function. We demonstrate that the Wasserstein information matrix 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 Wasserstein information matrix. 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 and design new information matrices. [6] designs new estimators for parametric inference using Wasserstein distance. This idea is utilized in approximating Bayesian computation, which measures the similarity between synthetic and observed data sets by Wasserstein distance. Compared to them, we focus on the study of estimation and efficiency of WIM. We expect it could have potential properties in Wasserstein estimators. In [7], they design a generalized information matrix based on the maximum mean discrepancy. Compared to them, we majorly focus on the information matrix generated by the Wasserstein metric and study the related statistical properties. Most closely, [23] defines the Wasserstein covariance by applying a closed-form

formula in one dimensional Wasserstein metric. This is a canonical definition. Our approach further extends this idea into general parametric models. We start by introducing WIM in parametric models. Using WIM, 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, [26] defines several new divergence functions by combining optimal transport and information geometry. Here we focus on statistical properties of canonical WIM in statistical models. Furthermore, Wasserstein natural gradient method has been widely studied in optimization techniques with machine learning applications [4, 9, 19, 14, 18]. Here we focus on statistical theory and study its associated online efficiency. Compared with classical Fisher–online efficiency result in [2, 20], our result can deal with general information matrices. In particular, for WIM, we discover a new efficiency property, such as Wasserstein efficiency and Poincaré efficiency. They rely on the comparison between Wasserstein and Fisher information matrices, known as Poincaré inequality in sample space.

The paper is organized as follows. In section 2, we establish the Wasserstein information matrix. We present it analytically for several well-known probability families. We provide an explicit example of WIM for the ReLU generative 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 prove LSI and PI inequalities in statistical models. In section 5, we introduce and discuss Wasserstein efficiency and Poincaré efficiency. In section 6, we carefully study a Gaussian mixture model to provide an approximate analysis of WIM.

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: $p(x;m,\lambda)=\lambda e^{-\lambda(x-m)}$	$G_W(m,\lambda)=\begin{pmatrix}1 & 0\\ 0 & \frac{2}{\lambda^4}\end{pmatrix}$	$G_F(m,\lambda)$ 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, Ex9.	$G_W(\theta)=F(\theta)$, F cdf of $p(x)$	$G_F(\theta)$ not well-defined

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

Family	Fisher-information functional	Log-Sobolev inequality(LSI(α))
Uniform	not well-defined	not well-defined
Gaussian	$\widetilde{I}(p_{\mu,\sigma}) = \frac{1}{\sigma^2}$, $\widetilde{I}(p_{\mu,\sigma}\|p_*) = \frac{(\mu-\mu_*)^2}{4\sigma_*^4} + \left(-\frac{1}{\sigma} + \frac{\sigma}{\sigma_*^2}\right)^2$.	$\frac{1}{\sigma^2} \geq 2\alpha \left\| -\frac{1}{2}\log 2\pi - \log\sigma - \frac{1}{2} \right\|$, $\frac{(\mu-\mu_*)^2}{\sigma_*^4} + \left(-\frac{1}{\sigma} + \frac{\sigma}{\sigma_*^2}\right)^2$ $\geq 2\alpha \left(-\log\sigma + \log\sigma_* - \frac{1}{2} + \frac{\sigma^2 + (\mu-\mu_*)^2}{2\sigma_*^2} \right)$.
Exponential	not well-defined	not well-defined
Laplacian	$\widetilde{I}(p_{\lambda,m}) = \frac{\lambda^2}{2}$, $\widetilde{I}(p_{m,\lambda}\|p_*) = \lambda_*^2 \left(1 - e^{-\lambda\|m-m_*\|}\right)^2$ $+ \frac{\left(\left(\lambda\|m-m_*\|+1\right)\lambda_* e^{-\lambda\|m-m_*\|} - \lambda\right)^2}{2}$.	$\lambda_*^2 \left(1 - e^{-\lambda\|m-m_*\|}\right)^2$ $+ \frac{\left(\left(\lambda\|m-m_*\|+1\right)\lambda_* e^{-\lambda\|m-m_*\|} - \lambda\right)^2}{2}$. $\geq 2\alpha \left(-1 + \log\lambda - \log\lambda_* + \lambda_*\|m-m_*\| + \frac{\lambda_* e^{-\lambda\|m-m_*\|}}{\lambda}\right)$.

TABLE 2. In this table, we continue the list to include the Fisher information functional, Log-Sobolev inequality for various probability families.

Family	RIW condition for LSI(α)	RIW condition for PI(α)
Uniform	not well-defined	not well-defined
Gaussian	$\begin{pmatrix} \frac{1}{\sigma_*^2} & 0 \\ 0 & \frac{1}{\sigma_*^2} + \frac{1}{\sigma^2} \end{pmatrix} \succeq 2\alpha \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} \frac{1}{\sigma_*^2} & 0 \\ 0 & \frac{2}{\sigma_*^2} \end{pmatrix} \succeq 2\alpha \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
Exponential	not well-defined	not well-defined
Laplacian	$\begin{pmatrix} \lambda\lambda_* e^{-\lambda\|m-m_*\|} & 0 \\ 0 & \frac{1}{\lambda^2} + \frac{\lambda_* e^{-\lambda\|m-m_*\|}\lambda^2(m_*-m)^2}{\lambda^3} \end{pmatrix}$ $\succeq 2\alpha \begin{pmatrix} 1 & 0 \\ 0 & \frac{2}{\lambda^4} \end{pmatrix}$	$\begin{pmatrix} \lambda_*^2 & 0 \\ 0 & \frac{1}{\lambda_*^2} \end{pmatrix} \succeq 2\alpha \begin{pmatrix} 1 & 0 \\ 0 & \frac{2}{\lambda_*^4} \end{pmatrix}$

TABLE 3. In this table, we present the RIW condition for LSI and PI in various probability families.

2. Wasserstein Information Matrix

In this section, we present the Wasserstein information matrix 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 parameter space $\Theta \subset \mathbb{R}^d$ and a parameterization function

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

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 the metric tensor on parameter space and connect it with the information matrix.

Definition 1 (Statistical Information Matrix)**.** *Consider the density manifold $(\mathcal{P}(\Omega), g)$ 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 G of g onto the parameter space Θ is given by*

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

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 the statistical metric, and call G the statistical information matrix.

In geometry, using this metric tensor g to rising(resp. lowing) the index, there exists a canonical isomorphism from tangent(resp. cotangent) space to cotangent(resp. tangent) space, namely:

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

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 the tangent space and cotangent space is merely:

$$\langle f, h\rangle = \int_{\mathcal{X}} fh dx,$$

where we use the symbol $\langle\cdot,\cdot\rangle$ for the inner product. We note here that the tangent space of a statistical model Θ can be viewed as a subspace of that of the 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*

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

They are the score functions associated with the statistical information matrix G and are equivalent classes in $C(\mathcal{X})/\mathbb{R}$. The representatives in the 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 metric tensor 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. Notice that this condition (1) makes no sense when Φ_i is not integrable w.r.t. p_θ. We comment here that this is not the case. At least for Fisher scores, they always satisfy this condition.

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

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

From the above fact, we observe that the statistical concept is related to the metric tensor in the density manifold pulled-back onto parameter space. In particular, classical statistics relates to the Fisher-Rao metric. The pull-back metric tensor forms the information matrix while the dual variables define the 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.** In this section, we present the Wasserstein Information matrix and Wasserstein score function.

The 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 probability function p. When p is under suitable conditions, standard PDE theory guarantees that the operators Δ_p^{-1} and Δ_p are inverse to each other between the function spaces:

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

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 previous section, we can rewrite G_W by the duality coordinates. Denote

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

Then

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

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

We summarize the above fact into the following definition.

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

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

where $\Phi_i^W : \mathcal{X}\times\Theta \to \mathbb{R}$ *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 the function $\Phi_i^W(x;\theta) = \left((-\Delta_{p_\theta})^{-1}\frac{\partial}{\partial\theta_i}p_\theta\right)(x;\theta)$ *the* Wasserstein score function*, and call the matrix* $G_W(\theta)$ *the* Wasserstein information matrix*.*

Remark 3. Our definition of information matrices are motivated by the intrinsic connection between distance 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 the following Taylor expansions in term of $\Delta\theta$:

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

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

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

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 the distance function defined in $\mathcal{X}$. Thus our approach parallels the classical Fisher statistics. 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 Wassertsein geometry. Meanwhile, our approach can be viewed as exploring another aspect, namely metric aspect, of the Wasserstein statistics. For example, it is related to the study of Wasserstein estimators [6].

We next study several basic properties of the Wasserstein information matrix and score functions. We first illustrate the relation between Wasserstein and Fisher score functions.

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

$$\nabla_x \log p(x;\theta)\cdot\nabla_x\Phi_i^W(x;\theta) + \Delta_x\Phi_i^W(x;\theta) = -\frac{\partial}{\partial\theta_i}\log p(x;\theta). \tag{3}$$

Proof. Notice the fact that

$$\left(\Delta_{p_\theta}\right)\Phi^W(x;\theta) = \nabla_x\cdot(p(x;\theta)\nabla_x\Phi^W(x;\theta)) = \nabla_x p(x;\theta)\cdot\nabla_x\Phi^W(x;\theta) + p(x;\theta)\Delta_x\Phi^W(x;\theta).$$

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 (3). □

We then demonstrate that the Wasserstein score function and matrix 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

$$\Phi^W(x;\theta)=\sum_{k=1}^n\Phi^{W,k}(x_k;\theta).\tag{4}$$

In addition, the Wasserstein information matrix 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 (4), then equation (3) 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 demonstrate some analytical solutions for the Wasserstein information matrix and score functions.

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

$$\Phi_i^W(x;\theta) = -\int_{\mathcal{X}\cap(\infty,x]} \frac{1}{p(z;\theta)} \frac{\partial}{\partial \theta_i} F(z;\theta) dz, \tag{5}$$

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

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

Proof. The Poisson equation (3) in one dimensional space has an explicit solution. Notice the fact that (3) is equivalent to the following second order ODE

$$\frac{d^2}{dx^2}\Phi_i^W(x) + \frac{d}{dx}\log p(x;\theta)\frac{d}{dx}\Phi_i^W(x;\theta) = -\frac{\partial}{\partial \theta_i}\log p(x;\theta).$$

Multiplying $p(x;\theta)$ on the both sides of the equation, we obtain

$$\frac{d}{dx}\Big(p(x;\theta)\frac{d}{dx}\Phi_i^W(x;\theta)\Big) = -\frac{\partial}{\partial \theta_i} p(x;\theta).$$

Applying the integration on both sides w.r.t. x,

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

Dividing p on both sides and then integrate w.r.t. x, we derive

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

Thus the Wasserstein information matrix forms

$$\begin{aligned} G_W(\theta)_{ij} =& \int_{\mathcal{X}} \frac{d}{dx}\Phi_i^W(x;\theta), \frac{d}{dx}\Phi_j^W(x;\theta) dx \\ =& \int_{\mathcal{X}} \frac{1}{p(x;\theta)}\frac{\partial}{\partial \theta_i} F(x;\theta) \frac{1}{p(x;\theta)}\frac{\partial}{\partial \theta_i} F(x;\theta) p(x;\theta) dx, \end{aligned}$$

which finishes the proof. □

Remark 4. We can also represent the Wasserstein score function in terms of the Fisher score function. In other words, denote

$$\Phi_i^W(x;\theta) = C + \int_{\mathcal{X}\cap(\infty,x]} \frac{\int_{\mathcal{X}\cap(\infty,z]} p(x';\theta)\frac{\partial}{\partial\theta_i}\log p(x';\theta)dx'}{p(x;\theta)}dz.$$

The Wasserstein information matrix satisfies

$$G_W(\theta)_{ij} = \int_{\mathcal{X}} \frac{\int_{\mathcal{X}\cap(\infty,x]} p(x';\theta)\frac{\partial}{\partial\theta_i}\log p(x';\theta)dx'}{p(x;\theta)} \frac{\int_{\mathcal{X}\cap(\infty,x]} p(x';\theta)\frac{\partial}{\partial\theta_j}\log p(x';\theta)dx'}{p(x;\theta)} p(x;\theta)dx.$$

We remark that the Wasserstein score function is the average of the cumulative Fisher score function, and the Wasserstein information matrix is the covariance of the density average of the cumulative Fisher score function.

If the dimension of sample space $\mathcal{X}$ is larger than 1, the exact solution of the Wasserstein score function and information matrix depends on the solution of Poisson equation (3). We leave the derivation of general formulas for interested readers.

2.3. **Examples.** We present several analytical examples of Wasserstein information matrices in one dimensional sample space.

Example 2 (Gaussian Distribution)**.** Consider a Gaussian distribution with mean value μ and standard variance $\sigma > 0$:

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

The 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 Wasserstein information matrix satisfies

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

Proof. Since

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

then

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

In this case, the Possion equation (3) for Wasserstein score functions $(\Phi_\mu^W, \Phi_\sigma^W)$ forms

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

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

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

□

Example 3 (Exponential distribution)**.** Consider the exponential distribution $Exp(m,\lambda)$:

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

The 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 Wasserstein information matrix satisfies

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

Proof. We derive the result by (5). 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

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

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

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 (1), we can decide the integration constants appear above. And inner products between the score functions follow:

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

□

Example 4 (Laplacian distribution)**.** Consider a Laplacian distribution $La(m,\lambda)$:

$$p(x;m,\lambda) = \frac{\lambda}{2}e^{-\lambda|x-m|}.$$

The Wasserstein score functions satisfy

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

Notice that the score functions for exponential family and Laplacian family have similar formulas. And the Wasserstein information matrix satisfies

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

Because the calculation within the Laplacian family is similar to the one we have done in the exponential distribution, we omit the proof here. And we will show below that the Laplacian family has advantages that densities within this family have the same support. Thus it is convenient for us to compare Wasserstein information matrix with Fisher information matrix.

Example 5 (Uniform distribution)**.** Consider the uniform distribution within interval $[a,b]$:

$$p(x;a,b) = \begin{cases} \frac{1}{b-a} & x\in[a,b]; \\ 0 & \text{otherwise.} \end{cases}$$

The Wasserstein score functions satisfy

$$\begin{aligned}
\Phi^W_a(x;a,b) &= \frac{x(a+b-x)}{(b-a)} - \frac{b^2+a^2+4ab}{6}, \\
\Phi^W_b(x;a,b) &= \frac{b(x-2a)}{(b-a)} - \frac{b^2-3ab}{2}.
\end{aligned}$$

And the Wasserstein information matrix satisfies

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

Proof. 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 the integration constants C_1,C_2 can be decided via the normalization condition (1). Thus

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

□

Example 6 (Wigner semicircle distribution)**.** Consider a Semicircle distribution

$$p(x;m,R)=\begin{cases}\frac{2}{\pi R^2}\sqrt{R^2-(x-m)^2}, & x\in[-R+m,R+m];\\ 0, & \text{otherwise.}\end{cases}$$

The Wasserstein score functions satisfy

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

And the Wasserstein information matrix satisfies

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

Proof. The cumulative distribution function satisfies

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

where we use the 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 to the parameter p is the same as before. And we conclude

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

Thus

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

□

Example 7 (Independence model)**.** Consider an independent model as follows: 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 the 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, the Wasserstein score functions for 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 the 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 Wasserstein information matrix satisfies

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

The proof follows directly from proposition 5.

Example 8 (Location-Scale Family)**.** Consider a location-Scale family as follows: given a probability density function $p(x)$ with $\int_{\mathbb{R}} p(x)dx = 1$, we define the density function of a location-scale family with location parameter m, scale parameter λ

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

Most of the previously discussed examples belong to this family, except that we do not use the location and scale parameter 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 $\mathbb{E}_{p_{m,\lambda}}x = m$.

Since m is the location parameter, the tangent vector $\frac{\partial}{\partial m}$ corresponds to the flow of shifting the density function with the shape of the function unchanged. 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 argue by pointing out that the transformation between the distribution $p(x;m_1,\lambda_1)$ and $p(x;m_2,\lambda_2)$ is actually given by the linear transform:

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

First, as we are working in the location-scale family, it is easy to show that this map pushes the density $p(x; m_1, \lambda_1)$ forward to $p(x; m_2, \lambda_2)$. That is,

$$\int_A p_{m_2,\lambda_2} dx = \int_{l^{-1}(A)} p_{m_1,\lambda_1} dx, \tag{6}$$

$\forall A \subset \mathbb{R}$, denoted by $l_* p_{m_1,\lambda_1} = p_{m_2,\lambda_2}$. 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 on the RHS is a convex function. Therefore, $l(x)$ is exactly the optimal transportation map between these two distributions (c.f. [24] Ch2.2).

To calculate the velocity field corresponds to the tangent vector $\frac{\partial}{\partial \lambda}$, we consider the following infinitesimal optimal transportation $p(x; m_1, \lambda_1) \to p(x; m_1, \lambda_1 + d\lambda)$. By the discussion 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 velocity field is given by

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

The inner product of this tangent vector is given by

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

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:

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

Consequently, the Wasserstein metric is diagonalized within the canonical parameterization of the location-scale family.

For location-scale family, we also formulate its Fisher score and Fisher information matrix for comparisons:

$$
\begin{aligned}
&\Phi_m^F(x;m,\lambda)=\frac{p'}{\lambda p},\quad \Phi_\lambda^F(x;m,\lambda)=-\frac{1}{\lambda}-\frac{(x-m)p'}{\lambda^2 p},\\
&G_F(m,\lambda)_{\lambda\lambda}=\int_{\mathbb{R}} p\left(\partial_\lambda \log p\right)^2 dx=\int_{\mathbb{R}} p\left(-\frac{1}{\lambda}-\frac{(x-m)p'}{\lambda^2 p}\right)^2 dx\\
&\qquad\qquad=\frac{1}{\lambda^2}\left(1+\int_{\mathbb{R}}\left(\frac{(x-m)^2 p'^2}{\lambda^2 p}+\frac{(x-m)\,p'}{\lambda}\right)dx\right),\\
&G_F(m,\lambda)_{mm}=\int_{\mathbb{R}} p\left(\partial_m \log p\right)^2 dx=\frac{1}{\lambda^2}\int_{\mathbb{R}}\frac{p'^2}{p}dx,\\
&G_F(m,\lambda)_{m\lambda}=\int_{\mathbb{R}} p\left(\partial_m \log p\right)\left(\partial_\lambda \log p\right) dx=\int_{\mathbb{R}} p\left(-\frac{p'}{\lambda p}\right)\left(-\frac{1}{\lambda}-\frac{(x-m)p'}{\lambda^2 p}\right)dx\\
&\qquad\qquad=\int_{\mathbb{R}}\frac{(x-m)p'^2}{\lambda^3 p}dx.
\end{aligned}
\tag{7}
$$

We next explain the above closed-form solution of WIM by the following proposition.

Proposition 7. *The location-scale family $p(x;m,\lambda)$ is a totally geodesic family in the density manifold under the Wasserstein metric.*

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

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

where T is the mapping function that push density ρ_1 forward to density ρ_2. It is known that the sufficient and necessary condition for an optimal map in 1-d case is that it is a monotone map, i.e. $\left(T(x)-T(y)\right)(x-y)\geq 0$(c.f. [24] Ch2.2). And in location-scale family, such map actually has a closed-form, namely:

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

Readers can check easily this map actually pushes measure ρ_1 forward to measure ρ_2 as desired. And it also satisfies monotonicity because it is linear with first-degree-coefficient $\frac{\lambda_2}{\lambda_1}>0$. Thus by the characterization in 1-d of optimal map, we conclude that T is actually optimal transportation plan between ρ_1 and ρ_2. Notice that this is another method to obtain the conclusion which has been proved in computation of WIM.

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 on Wasserstein geometry

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

where the push-forward map has closed-form:

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

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

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

Thus we show that the geodesic between any two densities in a location-scale family actually lies in this family. Any Location-scale family $p(x; m, \lambda)$ is a totally geodesic submanifold in the density manifold. □

Remark 5. This result on the 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.

In above discussions, all examples are based on location-scale families. We show that they are totally geodesic submanifold in Wasserstein geometry. In section 6, we shall study general models that are not totally geodesic submanifold, e.g. the Gaussian mixture family.

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

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

Here the generative models are powerful in machine learning [15]. It applies the reparameterization trick (known as push-forward relation) to do efficient sampling. In practice, one often applies the ReLU as the pushforward function (6). For this reason, we call the model ReLU push-forwarded family.

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

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

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

And the WIM of this family is given by

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

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 conclusion is

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

A figure illustrating these two families is provided below.

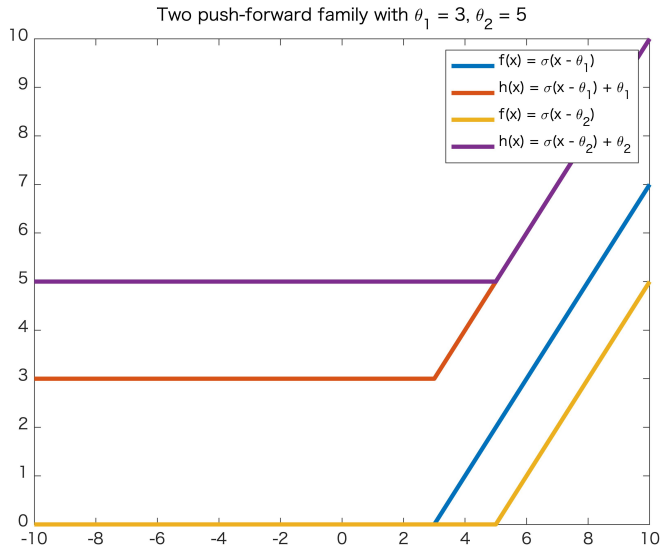

FIGURE 1. This figure plots two example of the push-forward family we descibed above with parameter chosen as $\theta_1 = 3, \theta_2 = 5$.

Proof. Here we don't have an explicit formula (2) to compute WIM, since push-forward distributions are no longer absolute continuous w.r.t. the Lebesgue measure. However, the linear programming formulation of the squared Wasserstein distance is still available.

Consider the following two push-forward distributions given by

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

where δ_0 refers to the Dirac measure concentrating at the point 0. And $p_0(\cdot+\theta)_{[0,\infty)}$ represents the measure $\widetilde{p}(x) = p_0(x+\theta)$ restricting to the interval $[0,\infty)$. Using the 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

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

where the third equality holds by the fact that

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

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. And the conclusion (8) follows.

For the other family, the derivation follows the same method 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 Dirac measure. □

Here we notice that the density function in ReLU push-forward family can be singular. Thus the Fisher information matrix, which depends on an explicit formula of the density function, 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 Wasserstein information matrix can provide systematical statistics 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 the information matrix 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 the cotangent vector f, h in the metric g.

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

Definition 8 (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 6. Here the inner product $\langle T_i, \widetilde{T}_j\rangle_g$ is obtained by viewing the statistics as cotangent vectors on the density manifold $\mathcal{P}(\mathcal{X})$.

Example 10 (Fisher Covariance)**.** Given two statistics T_1, T_2, we view them as cotangent vector in the space $C(\mathcal{X})/\mathbb{R}$. Hence their Fisher inner product is defined by

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

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

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

We next focus on the Wasserstein covariance operator.

Definition 9 (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 10 (Wasserstein-Cramer-Rao inequalities)**.** *Given any set of statistics $T = (T_1, ..., T_n) : \mathcal{X} \to \mathbb{R}^n$, where n 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

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

Proposition 11 (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 function, while the last equality holds by the integration by parts formula in spatial domain. □

Remark 7. This property is in contrast to the Fisher score function

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

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

Proof of Theorem 10. By the definition of semi-positive matrix, it suffices to prove that for arbitrary $v\in\mathbf{R}^n$, 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 the 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 an 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} with 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 11. 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 12. *For an estimator $T(x)$, it is Wasserstein efficient if and only if it attains the WCR 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 8. From the above derivation, the sufficient and necessary condition for a statistic to be efficient is that, it can be written as the linear combination of the score functions. Notice this criterion is valid for any metrics, including both the Fisher metric and Wasserstein metric. This is a purely geometric condition and we seek below in various statistical models to find out its statistical significance.

Remark 9. As shown in the above theorem, if we denote $G(p) = \frac{1}{p}$, 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 function of statistics $\mathbb{E}_{p_\theta} T(X)$. Furthermore, given any information matrix on a statistical model, we have the correspondent Cramer-Rao bound.

3.2. **Examples.**

Example 11 (Gaussian Distribution)**.** Given a 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 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 the efficient statistics in Wasserstein case are merely those which can be written as the linear combination of the score functions. Since statistics only depend on the sample point x_i and do not depend on the parameter μ,σ, they must be of the form:

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

Because Wasserstein cotangent vector is determined up to a constant. The Wasserstein efficient statistics are degree 2 polynomials of x.

While the 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 case, it turns out that the efficient statistics associated with these two information matrix coincide. But still, Fisher and Wasserstein information matrix provide us different Cramer-Rao bounds. The Fisher-Rao bound is better when we have prior knowledge that σ is pretty large while worse if σ is assumed to be small.

Example 12 (Exponential distribution)**.** Given an exponential distribution:

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

The Wasserstein score functions satisfy

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

with WIM

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

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

While the counterpart for Fisher case reads:

$$\Phi_\lambda^F(x;m,\lambda) = m - x + \frac{1}{\lambda}, \quad \Phi_m^F(x;m,\lambda) \text{ not well defined.}$$

Meanwhile, the Fisher information matrix is also ill-behaved, in contrast to the well-definedness of both the Wasserstein score and WIM. This example provides a situation where Wasserstein statistics are better than the classical Fisher statistics, at least at range to define efficiency for estimators.

Remark 10. Following the proposition 7, we can show that for any location-scale families, the efficient Wasserstein statistics are quadratic polynomial functions on x.

4. Functional inequalities via Information Matrices

In this section, we explore connections between information matrices(IMs) and functional inequalities in statistical models. Later on, we will show these inequalities are important for the study of statistical efficiency properties.

4.1. **Classical Functional Inequalities.**

Before working in statistical model, we first give a summary on relation between the PI, LSI and dynamical quantities on the density manifold; see related studies in [22].

We consider the relative entropy functional(KL-divergence) defined on the 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 the notation $H(\cdot|\cdot)$ in order to be consistent with the literature. We recall the classical definitions of the Log-Sobolev inequality and relative Fisher information functional as below.

Definition 13 (Log-Sobolev Inequality & Relative Fisher Information Functional)**.** *The probability measure ν is said to satisfy the 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 $I(\mu|\nu)$ shows up above is the so-called relative Fisher information functional

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

The most well-known and efficient criterion that guarantees LSI and PI is related to the information matrix(operator or metric in infinite dimension case) G_W we have studied before.

Proposition 14. *Denote* $\mathrm{Hess}_W H(\mu|\nu), G_W(\mu)$ *two bi-linear forms correspond 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,* $\forall \mu \in \mathcal{P}(\mathcal{X})$ *on the Hilbert space* $T_\mu \mathcal{P}(\mathcal{X})$*. Then LSI*($\alpha$) *holds.*

(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.*

Proof. We prove the first conclusion and the second follows easily since PI is the linearization of LSI. We compute gradient of the relative entropy functional w.r.t. the 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 the $\frac{\delta}{\delta\mu}$ refers to the L^2 functional derivative. Thus it is easy to obtain the relative entropy dissipation along the gradient flow as:

$$\frac{d}{dt}H(\mu|\nu) = -g_W\left(\mathrm{grad}_W H(\mu|\nu), \mathrm{grad}_W H(\mu|\nu)\right) = -\int_{\mathcal{X}}\left|\nabla_x\log\frac{\mu(x)}{\nu(x)}\right|^2\mu(x)dx = -I(\mu|\nu). \tag{11}$$

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 the LSI(α) holds via integration of the above formula. □

Remark 11. With the help of (11), readers can recognize that LSI guarantees the global exponential convergence of the gradient flow of the relative entropy functional $H(\cdot|\nu)$. Indeed, suppose μ_t is the gradient flow of $H(\cdot|\nu)$ starts from μ_0, then we have:

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

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

Remark 12. 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 the above definition of LSI translates to:

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

which is the more familiar definition of LSI(α), and linearize the above formula with $h = e^{\epsilon f}, \epsilon\to 0$, we get the classical definition of PI(α)

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

4.2. **LSI and PI in family.** Now, it is clear that PI and LSI are related to dynamical quantities on the density manifold. Consequently, the density manifold is not the only setting we can study such an important connection. Thus, we attempt to find those counterparts in statistical models.

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

$$\begin{aligned}\widetilde{I}(p_{\theta_t}|p_{\theta_*}) &= -\frac{d}{dt}\widetilde{H}(p_{\theta_t}|p_{\theta_*}) \\ &= g_W\left(\operatorname{grad}_W \widetilde{H}(p_\theta|p_{\theta_*}), \operatorname{grad}_W \widetilde{H}(p_\theta|p_{\theta_*})\right) \\ &= \nabla_\theta \widetilde{H}^T \left(\widetilde{G}_W^{-1}\right)^T \widetilde{G}_W \widetilde{G}_W^{-1} \nabla_\theta \widetilde{H} = \nabla_\theta \widetilde{H}^T \widetilde{G}_W^{-1} \nabla_\theta \widetilde{H},\end{aligned} \tag{12}$$

where we use the fact with the WIM $\widetilde{G}_W$

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

Definition 15 (LSI in family)**.** *Consider a statistical model $p : \mathcal{X} \times \Theta \to \mathbb{R}$, the 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 the sufficient condition for LSI and PI in proposition 14:

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

where we have to take care that the Hessian on LHS is calculated in the submanifold instead of the density manifold. An interesting point which shows up in the parameterized case is that the Fisher information matrix also comes into this picture, via a decomposition of the Hessian term. This point is known as the Ricci-information-Wasserstein (RIW) condition [17]:

Theorem 16 (RIW-condition)**.** *The information matrix criterion for LSI in proposition 14 can be written as:*

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

where Γ^W is the Christoffel symbol in Wasserstein statistical model Θ, while for PI 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 13. It can be seen that the condition for the Log-Sobolev inequality is much trickier than that of the Poincaré inequality. For LSI requires the global convexity of the entropy functional while PI only corresponds to local behavior at the minimum. The most significant change takes place in the Hessian term of entropy functional, where the Wasserstein Christoffel symbols come in; see related studies in the density manifold [13].

4.3. **Examples in 1-d Family.** The LSI and PI can be proved with the help of the Wasserstein and Fisher information matrix. Previously, we have done geometric computations on metric tensor and Hessian of the entropy functional. This prepares the ingredient for us to establish inequality in the family of probability distributions. In this section, we utilize the calculation before to obtain concrete bounds on these functional inequalities.

Example 13 (Gaussian Distribution)**.** Consider a 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 Wasserstein and Fisher information matrices are given by:

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

The entropy functional and relative entropy functional defined on this model is provided by:

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

from which we can calculate the Wasserstein gradient associated with two functionals:

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

with the Fisher information functional followed by:

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

Thus, the LSI(α) is given by

$$\begin{aligned}\frac{1}{\sigma^2}&\geq 2\alpha\left|-\frac{1}{2}\log 2\pi-\log\sigma-\frac{1}{2}\right|,\\ \frac{(\mu-\mu_*)^2}{\sigma_*^4}+\left(-\frac{1}{\sigma}+\frac{\sigma}{\sigma_*^2}\right)^2&\geq 2\alpha\left(-\log\sigma+\log\sigma_*-\frac{1}{2}+\frac{\sigma^2+(\mu-\mu_*)^2}{2\sigma_*^2}\right).\end{aligned}$$

Next, we move onto the derivation of RIW condition. Instead of establish LSI(α) for a specific distribution ν, we consider the relation between $\widetilde{G}_W, \mathrm{Hess}_W\,\widetilde{H}$ at each point. Recall the formula for Hess in Riemmanian geometry:

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

where Γ^W is the Christoffel symbol in Wasserstein geometry. In Wasserstein Gaussian model where the metric is Euclidean, Christoffel symbols vanish $\Gamma^W = 0$, 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 the gradient flow w.r.t. the entropy functional $\widetilde{H}(\cdot)$, we do not have satisfying constant α such that the Hess condition proposition 14 holds. For the $\mathrm{Hess}_W \widetilde{H}(\cdot)$ matrix has an eigenvalue 0. Despite of this, we have:

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

whose μ component always vanishes. Thus the gradient direction of $\widetilde{H}(\cdot)$ is always the σ direction, in which we have eigenvalue bound: $\mathrm{eig}_\sigma(H) \geq \frac{1}{\sigma^2}\mathrm{eig}_\sigma(G)$. This refers to that the eigenvalue of two matrix corresponds to direction $\frac{\partial}{\partial\mu}$ has the above bound. For the LSI, if the range of σ is the whole $\mathbb{R}$, then it is easy to see there will not exist satifying constant $\alpha > 0$ for the LSI(α) to hold. However, if we restrict the range of σ to a bounded region in $[-M, M]$, then LSI($\frac{1}{2M^2}$) will hold.

Remark 14. The following calculation on the gradient flow of entropy functional does not establish LSI(α) for any specific distribution. It merely provides an example of using Hessian condition to study the dynamical behavior.

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

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

In other words, the Gaussian p_{μ_*,σ_*} satisfies LSI$\left(\frac{1}{2\sigma_*^2}\right)$ in the Gaussian model.

Example 14 (Laplacian Distribution)**.** Consider the case of Laplacian distribution, where

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

from which we can calculate the Christoffel symbol as:

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

Following the same procedure we have done before, the entropy functional and relative entropy functional defined on this model is provided by:

$$H(m,\lambda) = -1 + \log\lambda - \log 2;$$
$$H(m,\lambda|p*) = -1 + \log\lambda - \log\lambda_* + \lambda_*|m-m_*| + \frac{\lambda_* e^{-\lambda|m-m_*|}}{\lambda}, \qquad p_* \sim p_{m_*,\lambda_*}.$$

from which we can calculate the Wasserstein gradient associated with two functionals:

$$\nabla^W_{m,\lambda} H(m,\lambda) = \begin{pmatrix} 0 \\ \dfrac{1}{\lambda} \end{pmatrix},$$

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

which can be further reduced to:

$$\nabla^W_{m,\lambda} H(m,\lambda) = \begin{pmatrix} 0 \\ \dfrac{1}{\lambda} \end{pmatrix},$$

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

with the Fisher information functional followed by:

$$\widetilde{I}(p_{m,\lambda}) = \frac{\lambda^2}{2},$$

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

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

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

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

$$\mathrm{Hess}_W \widetilde{H}(m,\lambda) = \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{\lambda^2} \end{pmatrix},$$

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

Following the same analysis, we conclude that for gradient flow of the entropy functional $H(\cdot)$, the LSI($\frac{\lambda^2}{4}$) holds. While for the relative entropy functional $H(\cdot|p_{m_*,\lambda_*})$, 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

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

From the 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 La$(m_*, \lambda_*)$ satisfies LSI(α) in Laplacian family with α given above.

Example 15 (Separable family)**.** For the separable 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 relative entropy also have this separability property

$$\begin{aligned} H(\theta) &= H_1(\theta) + H_2(\theta), \\ H(\theta|p_*) &= H_1(\theta|p_{1*}) + H_2(\theta|p_{2*}), \\ \nabla_\theta H(\theta|p_*) &= \nabla_\theta H_1(\theta|p_{1*}) + \nabla_\theta H_2(\theta|p_{2*}). \end{aligned}$$

The Fisher information functional is given by

$$I(p_\theta|p_*) = \left(\nabla_\theta H_1(\theta|p_{1*}) + \nabla_\theta H_2(\theta|p_{2*})\right)^T \left(G_W^1 + G_W^2\right)^{-1} \left(\nabla_\theta H_1(\theta|p_{1*}) + \nabla_\theta H_2(\theta|p_{2*})\right),$$

with LSI

$$\begin{aligned} &\left(\nabla_\theta H_1(\theta|p_{1*}) + \nabla_\theta H_2(\theta|p_{2*})\right)^T \left(G_W^1 + G_W^2\right)^{-1} \left(\nabla_\theta H_1(\theta|p_{1*}) + \nabla_\theta H_2(\theta|p_{2*})\right) \\ \geq\ & 2\alpha \nabla_\theta H_1(\theta|p_{1*}) + \nabla_\theta H_2(\theta|p_{2*}). \end{aligned}$$

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

5. Wasserstein natural gradient works efficiently

The LSI and PI discussed in the previous section are naturally connected to the convergence property of dynamical systems via Wasserstein geometry. In this section, we study Wasserstein dynamics in terms of sampling and estimation processes. As a consequence, we prove the asymptotic efficiency 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 the PI and LSI.

In the beginning, we state the natural gradient algorithm. We aim to estimate an un-known distribution in the probability family $p(x;\theta)$ with unknown parameter $\theta \in \Theta$. Assume that an optimal parameter $p(x;\theta_*)$ coincides with the unknown 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:

$$\theta_{t+1} = \theta_t - \frac{1}{t}\nabla_\theta^W l(x_t, \theta_t). \tag{13}$$

In the above formula, the θ_t is the updating state variable, $\frac{1}{t}$ in the RHS is the adaptive factor. And ∇_θ^W is the natural (Riemannian) gradient of the loss function l w.r.t. θ in the 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 $\nabla_\theta l$ the Euclidean gradient. We first pose here the definition of the efficiency of the natural gradient algorithm, which generalizes the notion discussed in [2]. Let us denote the Wasserstein covariance matrix of estimator θ_t by:

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

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

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

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

$$\begin{aligned}\mathbb{E}_{p_{\theta_*}}\left(\nabla(\theta_t - \theta_*)\cdot\nabla(\theta_t - \theta_*)^T\right)_{ij} &= \mathbb{E}_{p_{\theta_*}}\left(\nabla(\theta_t - \theta_*)_i \cdot \nabla(\theta_t - \theta_*)_j\right) \\ &= \mathrm{Cov}^W\left[(\theta_t - \theta_*)_i, (\theta_t - \theta_*)_j\right].\end{aligned}$$

Now the $\cdot$ refers to inner product of the vector. This shows the power of this estimator. And the subscript $p(\cdot;\theta_*)$ refers 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 (13). Then we actually have $\theta_t = \theta_t(x_1, x_2..., x_{t-1})$. Hence intuitively, the Cramer-Rao bound for θ_t we derived above is given by

$$CR = \frac{1}{t-1}G_W^{-1}(\theta_*).$$

It inspires the following definition:

Definition 17. *The Wasserstein natural gradient is asymptotic efficient if*

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

Remark 15. This definition is similar to the definition of classical Fisher efficient except that we substitute the Fisher information matrix by WIM. This also indicates the importance of studying the information matrix. And it will be shown that this quantity characterizes the convergence rate of dynamics in statistical inference.

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

Theorem 18 (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:*

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

Remark 16. More generally, it will be shown that such a simple updating equation merely attributes to the property of metric. Specifically, any statistical metric with separability property w.r.t. the independent variables have this form of updating equation. For the Wasserstein metric, this is already established in proposition 5. And for Fisher-Rao metric, this is merely the property of expectation under independent variables. Further results such as efficiency of the natural gradient can be established with the same procedure below.

Before working on the proof, we first show several important cases of this general theorem.

5.1. **Natural Gradient for Wasserstein Scores in Wasserstein Geometry.** 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:

$$\nabla_\theta^W f(x_t,\theta_t)=G_W^{-1}\Phi^W(x_t,\theta_t). \tag{14}$$

Concerned with this dynamics, we have the following characterization theorem.

Corollary 19 (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:

$$V_t=\frac{1}{t}G_W^{-1}(\theta_*)+O\left(\frac{1}{t^2}\right). \tag{15}$$

Proof of Corollary 19. 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:

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

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 17. The updating equation cannot be further simplified if we choose another normalized condition:

$$\int\Phi_j^W dx=\mathbf{0},\qquad j=1,...,n,$$

which is more natural in the perspective of unnormalized optimal transport(unnormalized Wasserstein geometry)[11]. Two main difficulties appear here, the first is that: we can no longer simplify the term:

$$\frac{2}{t}\mathbb{E}_{p_{\theta_*}}\left[\nabla_x\left(\theta_t-\theta_*\right)\cdot\Phi^W(x_t,\theta_*)^T\nabla_xG_W^{-1}(\theta_t)\right].$$

The other is: we cannot have the simple relation:

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

which is another key fact to establish the efficiency result.

Remark 18. At first, such a generalization to the Wasserstein metric may seem unreasonable. We only use the fact that both of them are metric on the probability space. Different from the Fisher score $\Phi^F = \nabla_\theta l$, the Wasserstein score Φ^W can not be written as the gradient of some functions w.r.t. θ. There is no such a "loss function". However, the 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:

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

5.2. **Natural Gradient for Fisher Scores in Wasserstein Geometry.** Another phenomenon appears when we consider the Wasserstein natural gradient applies to the Fisher score. 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 20 (Poincare 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

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

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

$$V_t = \begin{cases} O\left(t^{-2\alpha}\right), & 2\alpha \leq 1, \\ \frac{1}{t} \left(2 G_F G_W^{-1} - \mathbf{I} \right)^{-1} G_W^{-1}(\theta_*) \mathfrak{I} \left(G_W^{-1}(\theta_*) \right) + O\left(\frac{1}{t^2} \right), & 2\alpha > 1, \end{cases} \tag{16}$$

where

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

Proof of Corollary 20. 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:

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

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 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 condition determines 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 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:

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

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

Remark 19. As the conclusion revealed, 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 (c.f. Theorem 16) we introduced in last section. This inspires us to name such efficiency Poincaré efficiency.

We now come back to the proof of theorem 18 on the updating equation of covariance matrix.

Proof of Theorem 18. First, we postulate that ∇_x refers to gradient w.r.t. x variable while ∇_θ refers to θ-gradient. 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_* to both sides of the updating equation and plugging in the expansion above, we get:

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

Then, taking the Wasserstein covariance of the both sides, we get:

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

where the last term corresponds to the $O\left(|\theta_t - \theta_*|^2\right)$ expansion and we use the assumption that $\mathbb{E}_{p_{\theta_*}} \left[(\theta_t - \theta_*)^2 \right], \mathbb{E}_{p_{\theta_*}} \left[|\nabla_x (\theta_t - \theta_*)|^2 \right] = o(1)$. On the above formula, we eliminate the transpose symbol ^{T} on the 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 component at $x_1,...,x_{t-1}$ while $\nabla_x\left(f(x_t,\theta_*)^T\right)$ only has 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 inequality is guaranteed by the fact that $\theta_t-\theta_*$ is independent to $\nabla_\theta f(x_t,\theta_*)$ by the independence assumption on the sampling. While the last inequality holds by the assumption on $f(x_t,\theta_*)$

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

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 appears above, with the help that $\mathbb{E}_{p_{\theta_*}}\left[\nabla_\theta f(x_t,\theta_*)\right]=O(1)$, $\nabla_x\left(G_W^{-1}(\theta_t)\right)=$

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

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

And the last term finally reduces to:

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

Combining all the terms we have in hand, we derive the following updating equation for the Wasserstein covariance during the 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 20. The most frequently used tools in the proof of the Wasserstein natural gradient is the 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 the Wasserstein and Fisher metric. Specifically, consider the gradient of T_1, T_2 w.r.t. x, since they depend on different variables, thus

$$\partial_{x_i} T_1 \cdot \partial_{x_i} T_2 = 0, \qquad i = 1, ..., n+t,$$

which holds pointwise. And the Wasserstein covariance of these two statistics vanishes because their inner product vanishes everywhere. This type of separability is directly an analog of the one in Fisher-Rao geometry:

$$g_F\left(T_1, T_2\right) = \mathbb{E}_{p_{\theta_*}} T_1 T_2 = \mathbb{E}_{p_{\theta_*}} T_1 \cdot \mathbb{E}_{p_{\theta_*}} T_2 = 0.$$

5.2.1. *Examples.*

Example 16 (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 the optimal parameters are given by μ_*,σ_*. Thus we have following conclusions on efficiency of the Fisher, Wasserstein natural gradient and Wasserstein natural gradient on Fisher score.

The Wasserstein natural gradient is asymptotically efficient with the 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 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 more simple. 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 the Fisher score $\nabla_{\mu_*,\sigma_*}l(x;\theta)=\Phi^F(x,\mu_*,\sigma_*)$ 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

$$\mathbb{E}_{p_{\mu_*,\sigma_*}}\left[\nabla_x\Phi^F_\mu(x;\mu_*,\sigma_*)\cdot\nabla_x\left(\Phi^F_\mu(x;\mu_*,\sigma_*)^T\right)\right]=\mathbb{E}_{p_{\mu_*,\sigma_*}}\left[\frac{1}{\sigma_*^4}\right]=\frac{1}{\sigma_*^4};$$

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

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

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 appears in the asymptotic behavior of the Poincaré dynamics 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 the Fisher score is given by:

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

Numerical experiments that verify our theory are also completed with the results shown below. In two cases, we verify two kinds of efficiency, namely theb Wasserstein Cramer-Rao efficiency and Poincaré efficiency respectively. In the first experiment, we verify the constant G_W^{-1} appears in the asymptotic efficient of the Wasserstein natural gradient. While for the other situation we verify the asymptotic exponential index α shows up in the Poincaré efficiency.

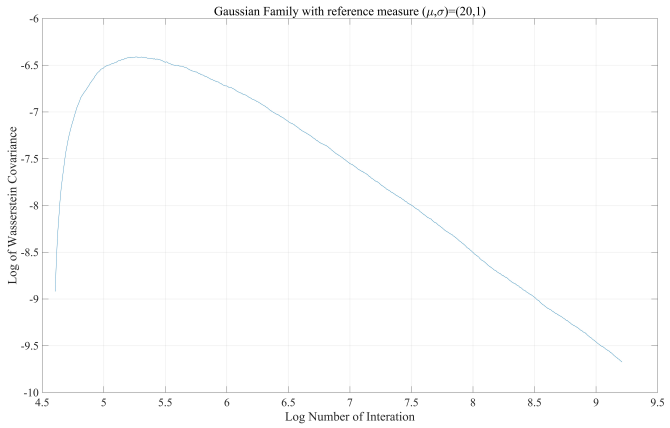

FIGURE 2. The Wasserstein-Cramer-Rao Type Convergence Rate. X-axis is logarithm of iteration t while y-axis is logarithm of Wasserstein covariance V_t. Reference Gaussian is chosen to be $\mathcal{N}(20, 1)$ where the parameter $\mu_* = 20$ is the critical point. Since we have $\frac{1}{\sigma_*^2} = 1 > \frac{1}{2}$, the Cramer-Rao type convergence holds. To satisfy the asymptotic assumption, we start the iteration at step $t = 100$ as well as $\mu = 19.9$. From the figure, we conclude that Cramer-Rao type convergence (15) holds.

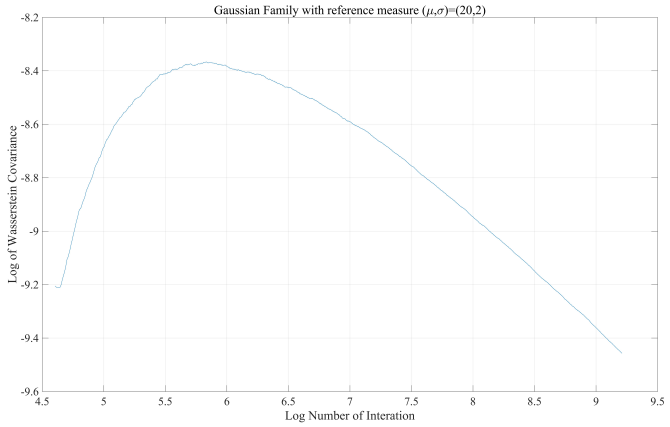

FIGURE 3. Poincaré Type Convergence Rate. X-axis is logarithm of iteration t while y-axis is logarithm of the Wasserstein covariance V_t. The reference Gaussian is chosen to be $\mathcal{N}(20, 2)$ where the parameter $\mu_* = 20$ is the critical point. Since we have $\frac{1}{\sigma_*^2} = \frac{1}{4} < \frac{1}{2}$, the Poincaré type convergence holds. To satisfy the asymptotic assumption, we start the iteration at step $t = 100$ as well as $\mu = 19.9$. From the figure, we conclude that the Poincaré type convergence (16) holds.

Example 17 (Laplacian Distribution)**.** Consider the Laplacian distribution $La(m, \lambda)$:

$$p(x; m, \lambda) = \frac{\lambda}{2} e^{-\lambda|x-m|}.$$

The WIM satisfies

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

The Fisher score functions are given by

$$\begin{aligned} \Phi_\lambda^F(x; m, \lambda) =& \frac{1}{\lambda} - |x - m|, \\ \Phi_m^F(x; m, \lambda) =& \begin{cases} -\lambda & x < m, \\ \lambda & x > m. \end{cases} \end{aligned}$$

We comment that the Fisher score Φ_m^F is not well-defined at the point $x = m$. But this is not a problem since what we want is the integral of them on the real line(the function Φ_m^F is obviously bounded on the real line). Consequently, the Fisher information matrix satisfies

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

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

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

And the optimal parameters again are given as λ_*, m_*. Thus we have following conclusions on efficiency of the Fisher, Wasserstein natural gradient and Wasserstein natural gradient on Fisher score.

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

$$V_t = \frac{1}{t} \begin{pmatrix} 1 & 0 \\ 0 & \frac{\lambda_*^4}{2} \end{pmatrix} + O\left(\frac{1}{t^2}\right).$$

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

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

For the last case, same as before, we denote the least significant eigenvalue of $G_F G_W^{-1}$ as α:

$$\alpha = \frac{\lambda_*^2}{2}.$$

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

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

With the help of explicit formula for the Fisher score, we deduce the formula of the addition term:

$$\begin{aligned}
&\mathbb{E}_{p_{\lambda_*,m_*}}\left[\nabla_x\Phi^F_\lambda(x;\lambda_*,m_*)\cdot\nabla_x\left(\Phi^F_\lambda(x;\lambda_*,m_*)^T\right)\right] = \mathbb{E}_{p_{\lambda_*,m_*}}1 = 1;\\
&\mathbb{E}_{p_{\lambda_*,m_*}}\left[\nabla_x\Phi^F_\lambda(x;\lambda_*,m_*)\cdot\nabla_x\left(\Phi^F_m(x;\lambda_*,m_*)^T\right)\right]\\
=\ &\mathbb{E}_{p_{\lambda_*,m_*}}\left[\nabla_x\Phi^F_\lambda(x;\lambda_*,m_*)\cdot 2\lambda\delta(x-m)\right] \text{ not well-defined;}\\
&\mathbb{E}_{p_{\lambda_*,m_*}}\left[\nabla_x\Phi^F_m(x;\lambda_*,m_*)\cdot\nabla_x\left(\Phi^F_m(x;\lambda_*,m_*)^T\right)\right] \text{not well-defined.}
\end{aligned}$$

Again, for the first integral, we omit the value of the function $\nabla_x\Phi^F_\lambda(x;\lambda_*,m_*)$ at the point $x=m$(which is assumed to be bounded). For the second and third integral, the integrands contain the multiple of Dirac function which is not well-defined. This can be overcomed by smoothing the Laplacian family.

6. Analysis of WIM in Gaussian mixture Family

In this section, we provide a detailed study of the 1-d Gaussian mixture model(GMM). Since an 1-d GMM does not fall into the class of location-scale family, the closed-form solution of transportation map does not exist, and approximate analysis of WIM becomes essential.

We first set the background of this statistical model. Given a family of Gaussian distribution $\rho_i = N(\mu_i,\sigma), i=1,2,...,N$ with same variance σ and their means are in increasing order $\mu_1<\mu_2<\cdots<\mu_N$:

$$\begin{aligned}
&p:\Theta\to\mathcal{P}(\mathcal{X}), \qquad \Theta=\{(\theta_1,\theta_2,...\theta_{N-1})\in\mathbb{R}^{N-1}|1=\theta_0\geq\theta_1\geq\cdots\geq\theta_{N-1}\geq\theta_N=0\},\\
&\theta\mapsto\sum_{i=1}^{N-1}\theta_i\left(\rho_{i+1}-\rho_i\right)+\rho_1.
\end{aligned}\tag{17}$$

To simplify the argument for approximation, we propose following assumption:

assumption 1. For a Gaussian mixture model with components $\rho_i\sim N(\mu_i,\sigma)$, intervals $[\mu_i-\sigma,\mu_i+\sigma]$ are disjoint.

Remark 21. Above assumption is not essential for proof of the result. We simply state it here to make the proof clear. It can be substituted by other form such as intervals $[\mu_i-\alpha\sigma,\mu_i+\alpha\sigma]$ are disjoint for any $\alpha>0$. The proof remains the same with minor changes.

Proposition 21. *GMMs are not totally geodesic submanifolds in the density manifold under the Wasserstein metric.*

Proof of Proposition 21. First, if we suppose that a GMM contains a boundary, then conclusion is almost trivial. Since the geodesic between two adjacent Gaussians ρ_i,ρ_{i+1}(which lies in the boundary of the GMM) are Gaussians with means interpolate between theirs. And these Gaussians are definitely not in this GMM since they are not the component of it.

Even if we consider merelte the interior of this GMM, this conclusion still holds. We consider transportation between the distribution $\widetilde{\rho}_0 = (1-\epsilon)\,\rho_{i-1}+\epsilon\rho_{i-2}$ and $\widetilde{\rho}_1 =$

$(1-\epsilon)\,\rho_i + \epsilon\rho_{i+1}$ where ϵ is a small number. Suppose the geodesic $\widetilde{\rho}_t$ connects them lies in this GMM. We focus on $\widetilde{\rho}_{\frac{1}{2}}$ and consider the quantiles of them. We use $\widetilde{F}_i$ to denote the cumulative distribution function of ρ_i.

Denote $x_i = \widetilde{F}_i^{-1}\left(\frac{1}{2}\right), i = 0, \frac{1}{2}, 1$. Since we choose ϵ sufficiently small, we should have

$$x_0 = \mu_{i-1} + O(\epsilon),\ x_1 = \mu_i + O(\epsilon).$$

In differential formulation of optimal transport (c.f. [24] Ch 8), particles move with constant speed along geodesic. Thus we have

$$x_0 + x_1 = 2x_{\frac{1}{2}},\ x_{\frac{1}{2}} = \frac{\mu_{i-1}+\mu_i}{2} + O(\epsilon).$$

It is obvious that the only two components that is close to $x_{\frac{1}{2}}$ is μ_{i-1}, μ_i.

Now, we consider $\widetilde{x}_i = \widetilde{F}_i^{-1}\left(\frac{1}{2}+\delta\right), i = 0, \frac{1}{2}, 1$, where δ is again a small number. Samely, we have

$$\widetilde{x}_0 + \widetilde{x}_1 = 2\widetilde{x}_{\frac{1}{2}}.$$

Substracting it with above formula, we have

$$(\widetilde{x}_0 - x_0) + (\widetilde{x}_1 - x_1) = 2\left(\widetilde{x}_{\frac{1}{2}} - x_{\frac{1}{2}}\right).$$

Then, taking $\delta \to 0$, we will have

$$(\widetilde{x}_0 - x_0) = \frac{\delta}{\widetilde{\rho}_0(x_0)} + o(\delta).$$

Consequently,

$$\frac{\delta}{\widetilde{\rho}_0(x_0)} + \frac{\delta}{\widetilde{\rho}_1(x_1)} + o(\delta) = \frac{2\delta}{\widetilde{\rho}_{\frac{1}{2}}(x_{\frac{1}{2}})} + o(\delta).$$

We conclude that

$$\rho_{\frac{1}{2}}(x_{\frac{1}{2}}) = \frac{2\widetilde{\rho}_1(x_1)\,\widetilde{\rho}_0(x_0)}{\widetilde{\rho}_0(x_0) + \widetilde{\rho}_1(x_1)}.$$

While for now, we merely need more simple conclusion

$$\widetilde{\rho}_{\frac{1}{2}}(x_{\frac{1}{2}}) \geq \min\{\widetilde{\rho}_1(x_1), \widetilde{\rho}_0(x_0)\}. \tag{18}$$

Recall that we have $\epsilon = o(1)$. Thus x_0, x_1 are close to μ_{i-1}, μ_i, while $x_{\frac{1}{2}}$ is around the middle of them. We have

$$\rho_{i-1}(x_{i-1}), \rho_i(x_i) > \rho_j(x_{\frac{1}{2}}), \quad \forall j = 1, 2, ..., n.$$

Therefore, $\rho_{\frac{1}{2}}(x_{\frac{1}{2}})$, as a linear combination of $\rho_j(x_{\frac{1}{2}})$, is strictly smaller than $\rho_1(x_1), \rho_0(x_0)$, contradicting to the established result (18). We conclude that this GMM is not a totally geodesic submanifold in the density manifold. □

Remark 22. Since a differential formulation is valid for any dimension, techniques to prove the result is suitable for high dimensional cases. We remark here that GMMs are not totally geodesic families for any dimension.

6.1. **The Metric.** At the very beginning, we directly give following characterizations of WIM for GMMs and postpone detailed study to the next section. For simplification consideration, we provide here another parameterization of a GMM

$$\sum_{i=1}^{N-1} \theta_i \left(\rho_{i+1} - \rho_i\right) + \rho_1 = \sum_{i=1}^{N} p_i \rho_i.$$

with the change of parameters given by

$$p_1 = 1 - \theta_1, p_N = \theta_{N-1}, p_i = \theta_{i-1} - \theta_i, i = 2, ..., N-1.$$

A remark here is that parameters p_is do not characterize the GMM freely because of explicit dependence on their sum. And following conclusions on metric are all concerned with the θ_i parameters.

Proposition 22. *Given a GMM, the Wasserstein metric tensor is approximately diagonal in following sense:*

$$\begin{aligned} G_W &= \widetilde{G_W} + \Delta, \quad \widetilde{G_W} = \mathrm{Diag}(a_1, a_2, \cdots, a_{N-1}), \\ a_i &= \int_{\mu_i}^{\mu_{i+1}} \frac{1}{p_{i+1}\rho_{i+1} + p_i \rho_i} dx, \quad i = 1, ..., N-1, \\ |\Delta_{ij}| &= O\left(\frac{\sigma^2}{\min_i p_i}\right), \quad |\Delta_{ii}| = O\left(\frac{\sigma}{\min_l p_l}\right). \end{aligned} \tag{19}$$

Δ *is the error term of this approximation. Furthermore, simple estimation of Laplacian asymptotic shows that*

$$a_i = \int_{\mu_i}^{\mu_{i+1}} \frac{1}{p_{i+1}\rho_{i+1} + p_i \rho_i} dx = \Theta\left(\frac{e^{\frac{1}{2}\left(\frac{\mu_{i+1}-\mu_i}{2\sigma}\right)^2}}{\frac{\mu_{i+1}-\mu_i}{2\sigma}}\right), \quad i = 1, ..., N-1.$$

Suppose we further have that $[\mu_i - 3\sigma, \mu_i + 3\sigma]$ *are mutually disjoint. Then we conclude that* $a_i = \Theta\left(e^{\frac{9}{2}}\right) \gg \Delta_{ij}, \Delta_{ii}$. *The error term is shown to be significantly smaller than the main part* $\widetilde{G}_W$. *Under this assumption, the Wasserstein information matrix is a diagonally dominant matrix.*

During the proof, we use intensively the closed-form solution for Wasserstein score functions in 1-d, namely

$$g_{ij} = \int \frac{\partial_{\theta_i} F \partial_{\theta_j} F}{\rho_\theta} dx = \int \frac{\left(F_i - F_{i-1}\right)\left(F_j - F_{j-1}\right)}{\rho_\theta} dx.$$

The key idea of the proof is that above integrals can be decomposed into several parts which are easy to be evaluated. An example that illustrates the idea of decomposition is shown below.

To analysis WIM, we separately consider two cases. The first is crossing term and the other is diagonal term.

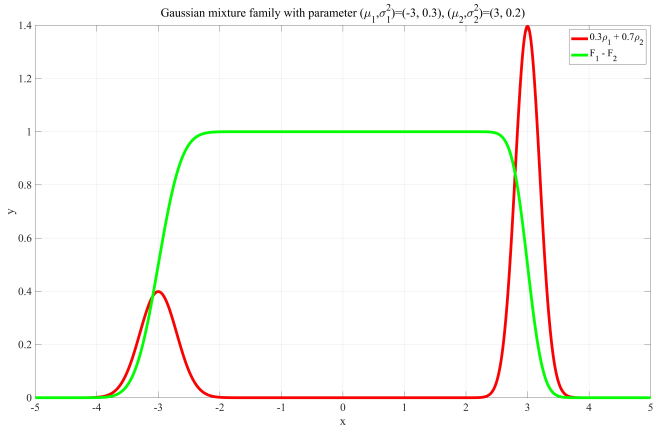

FIGURE 4. This figure plots two functions $F_1 - F_2, \rho_\theta$ with parameter given above. F_i is cumulative distribution function of density $\rho_i, \forall i = 1, 2$. $F_1 - F_2$ highly concentrates in region between the means of two distribution, namely $[-3, 3]$. Consequently, decomposition trick is simply dividing integral region $\mathbb{R}$ to three regions $(-\infty, \mu_1], [\mu_1, \mu_2], [\mu_2, \infty)$. On region $[\mu_1, \mu_2]$, we can treat $F_1 - F_2$ as constant 1. While at regions $(-\infty, \mu_1], [\mu_2, \infty)$, we can use ρ_1, ρ_2 to bound $F_1 - F_2$. Rigid formulation follows. Furthermore, if the integral has the multiplication of $F_i - F_{i-1}, F_j - F_{j-1}$ as nominator, we have to decompose the integral region to five parts to apply approximation.

6.2. **Analysis of Crossing Terms.** First we consider crossing term $\int \frac{(F_i - F_{i-1})(F_j - F_{j-1})}{\rho} dx$ with $i < j$. Using our decomposition trick, we first bound this term at two edges, namely, $(-\infty, \mu_i], [\mu_j, \infty)$. Before that, we need a comparison between the Gaussian pdf and cdf;

Lemma 1. For the Gaussian $\rho(x) \sim \mathcal{N}(\mu, \sigma)$, following inequalities between pdf and cdf hold:
For $x \leq 0$(resp. $x \geq 0$):

$$\lim_{x \to -\infty} \frac{F(x)}{\frac{\sigma^2 \rho(x)}{|x-\mu|}} = 1, \qquad \left(\text{resp.} \quad \lim_{x \to \infty} \frac{1 - F(x)}{\frac{\sigma^2 \rho(x)}{|x-\mu|}} = 1 \right)$$

$$F(x) \leq \min\{\sqrt{\frac{\pi}{2}}\sigma\rho(x), \ \frac{\sigma^2 \rho(x)}{|x-\mu|}\} \qquad \left(\text{resp.} \quad 1 - F(x) \leq \min\{\sqrt{\frac{\pi}{2}}\sigma, \ \frac{\sigma^2 \rho(x)}{|x-\mu|}\} \right).$$

The proof is simple via expansion of the integration of cdf. The following simple lemma quantifies the ratio of density functions for different components under assumption 1. The proof is omitted.

Lemma 2. For a GMM, suppose $i < j$, then we have following bounds:

$$\rho_j(x) \leq \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\mu_i - \mu_j)^2}{2\sigma^2}}, \quad \frac{\rho_j(x)}{\rho_i(x)} \leq e^{-\frac{(\mu_i - \mu_j)^2}{2\sigma^2}} \qquad x \in (-\infty, \mu_i]$$

Now, we can analyze the integral on the edge.

Lemma 3. The integration on the edge part is bounded by

$$|\int_{-\infty}^{\mu_{i-1}} \frac{(F_i - F_{i-1})(F_j - F_{j-1})}{\rho} dx| \leq \frac{e^{-\frac{\left(\mu_{i-1}-\mu_{j-1}\right)^2}{2\sigma^2}}\sigma^2}{p_{i-1}\sqrt{2\pi}} \left(\frac{\sqrt{\frac{\pi}{2}}\sigma}{\mu_{j-1}-\mu_{i-1}} + 1\right), \quad i < j.$$

Proof. Using the order on the mean of each component, we have:

$$F_i \leq F_{i-1},\ F_j \leq F_{j-1}, \qquad x \in (-\infty, \mu_{i-1}].$$

Consequently,

$$\left|\frac{F_i - F_{i-1}}{\rho}\right| \leq \left|\frac{F_{i-1}}{p_{i-1}\rho_{i-1}}\right| \qquad x \in (-\infty, \mu_{i-1}).$$

Then, the integral concerned can be decomposed as:

$$\begin{aligned}
&\left|\int_{-\infty}^{\mu_{i-1}} \frac{(F_i - F_{i-1})(F_j - F_{j-1})}{\rho} dx\right| \\
\leq& \int_{-\infty}^{\mu_{i-1}} \left|\frac{F_{i-1}}{p_{i-1}\rho_{i-1}}\right| |F_{j-1}|\, dx \\
\leq& \frac{1}{p_{i-1}} \int_{-\infty}^{\mu_{i-1}-\sigma} \frac{\sigma^2\rho_{i-1}(x)}{\rho_{i-1}(x)(\mu_{i-1}-x)} \frac{\sigma^2\rho_{j-1}(x)}{(\mu_{j-1}-x)} dx \\
&+ \frac{1}{p_{i-1}} \int_{\mu_{i-1}-\sigma}^{\mu_{i-1}} \frac{\sqrt{\frac{\pi}{2}}\sigma\rho_{i-1}(x)}{\rho_{i-1}(x)} \frac{\sigma^2\rho_{j-1}(x)}{(\mu_{j-1}-x)} dx \\
\leq& \frac{e^{-\frac{\left(\mu_{i-1}-\mu_{j-1}\right)^2}{2\sigma^2}}}{p_{i-1}\sqrt{2\pi}} \left(\int_{-\infty}^{\mu_{i-1}-\sigma} \frac{\sigma^3}{(x-\mu_{i-1})^2} dx + \int_{\mu_{i-1}-\sigma}^{\mu_{i-1}} \frac{\sqrt{\frac{\pi}{2}}\sigma^2}{(x-\mu_{j-1})} dx\right) \\
\leq& \frac{e^{-\frac{\left(\mu_{i-1}-\mu_{j-1}\right)^2}{2\sigma^2}}}{p_{i-1}\sqrt{2\pi}} \left(\sigma^2 + \frac{\sqrt{\frac{\pi}{2}}\sigma^3}{\mu_{j-1}-\mu_{i-1}}\right) \\
=& \frac{e^{-\frac{\left(\mu_{i-1}-\mu_{j-1}\right)^2}{2\sigma^2}}\sigma^2}{p_{i-1}\sqrt{2\pi}} \left(\frac{\sqrt{\frac{\pi}{2}}\sigma}{\mu_{j-1}-\mu_{i-1}} + 1\right),
\end{aligned}$$

where in the second step we use previous estimation on Gaussian cdf and pdf. In the third step, we use bounds on density functions and simple inequalities for integrals. □

Remark 23. The other-side-term concerned with integral on $[\mu_j, \infty)$ can be bounded by the same technique, and only the result is shown here.

$$\left|\int_{\mu_j}^{\infty} \frac{(F_i - F_{i-1})(F_j - F_{j-1})}{\rho} dx\right| \leq \frac{e^{-\frac{\left(\mu_i-\mu_j\right)^2}{2\sigma^2}}\sigma^2}{p_j\sqrt{2\pi}} \left(\frac{\sqrt{\frac{\pi}{2}}\sigma}{\mu_j-\mu_i} + 1\right), \quad i < j.$$

Next we consider the term at middle, that is, integration on $[\mu_i, \mu_{j-1}]$.

Lemma 4. Assume $j - i > 1$, we have:

$$\left| \int_{\mu_i}^{\mu_{j-1}} \frac{(F_i - F_{i-1})(F_j - F_{j-1})}{\rho} dx \right| \leq \frac{\sigma^3}{\mu_{j-1} - \mu_i - \sigma} \left(\frac{1}{2p_i} + \frac{1}{2p_{j-1}} \right) + \frac{\sigma^2}{\sqrt{2\pi} p_i}. \tag{20}$$

Proof. First we conclude that

$$|F_i - F_{i-1}| = |(1 - F_i) - (1 - F_{i-1})| \leq 1 - F_i, \ |F_j - F_{j-1}| \leq F_{j-1}, \qquad x \in [\mu_i, \mu_{j-1}].$$

Using the same trick as in the previous lemma, we get:

$$\begin{aligned}
&\left| \int_{\mu_i}^{\mu_{j-1}} \frac{(F_i - F_{i-1})(F_j - F_{j-1})}{\rho} dx \right| \\
\leq &\int_{\mu_i}^{\mu_i + \sigma} \left| \frac{(1 - F_i)}{p_i \rho_i} \right| |F_{j-1}| \, dx + \int_{\mu_i + \sigma}^{\mu_{j-1} - \sigma} \left| \frac{(1 - F_i)}{p_i \rho_i} \right| |F_{j-1}| \, dx \\
&+ \int_{\mu_{j-1} - \sigma}^{\mu_{j-1}} \left| \frac{F_{j-1}}{p_{j-1} \rho_{j-1}} \right| |(1 - F_i)| \, dx \\
\leq &\frac{1}{p_i} \int_{\mu_i}^{\mu_i + \sigma} \frac{\sqrt{\frac{\pi}{2}} \sigma \rho_i(x)}{\rho_i(x)} \frac{\sigma^2 \rho_{j-1}(x)}{(\mu_{j-1} - x)} dx \\
&+ \frac{1}{p_i} \int_{\mu_i + \sigma}^{\mu_{j-1} - \sigma} \frac{\sigma^2 \rho_i(x)}{\rho_i(x)(x - \mu_i)} \frac{\sigma^2 \rho_{j-1}(x)}{(\mu_{j-1} - x)} dx \\
&+ \frac{1}{p_{j-1}} \int_{\mu_{j-1} - \sigma}^{\mu_{j-1}} \frac{\sqrt{\frac{\pi}{2}} \sigma \rho_{j-1}(x)}{\rho_{j-1}(x)} \frac{\sigma^2 \rho_i(x)}{(x - \mu_i)} dx \\
\leq &\frac{\sigma^3}{\mu_{j-1} - \mu_i - \sigma} \left(\frac{1}{2p_i} + \frac{1}{2p_{j-1}} \right) + \frac{\sigma^2}{\sqrt{2\pi} p_i}
\end{aligned}$$

□

Remark 24. Notice that this bound is significantly larger than previous one, the factor $e^{-\frac{(\mu_i - \mu_j)^2}{2\sigma^2}}$ disappears in the formula. We remark here this bound can actually be improved by further partition of integration regions.

At last, we consider integration on $[\mu_{i-1}, \mu_i]$ where one of the factors of nominator becomes close to 1.

Lemma 5.

$$\left| \int_{\mu_{i-1}}^{\mu_i} \frac{(F_i - F_{i-1})(F_j - F_{j-1})}{\rho} dx \right| \leq \frac{\sigma^2}{p_i} \left(\sqrt{\frac{\pi}{2}} + \log \left(\frac{\mu_{j-1} - \mu_{i-1}}{\sigma + \mu_{j-1} - \mu_i} \right) \right), \quad i < j.$$

Proof.

$$\begin{aligned}
&\left|\int_{\mu_{i-1}}^{\mu_i} \frac{(F_i - F_{i-1})(F_j - F_{j-1})}{\rho} dx\right| \\
\leq& \int_{\mu_{i-1}}^{\mu_i} \left|\frac{F_{j-1}}{\rho}\right| dx \\
\leq& \int_{\mu_{i-1}}^{\mu_i - \sigma} \left|\frac{\sigma^2 \rho_{j-1}(x)}{p_i \rho_i(x)(\mu_{j-1} - x)}\right| dx + \int_{\mu_i - \sigma}^{\mu_i} \left|\frac{\sqrt{\frac{\pi}{2}}\sigma \rho_{j-1}(x)}{p_i \rho_i(x)}\right| dx \\
\leq& \frac{\sigma^2}{p_i}\left(\sqrt{\frac{\pi}{2}} + \log\left(\frac{\mu_{j-1} - \mu_{i-1}}{\sigma + \mu_{j-1} - \mu_i}\right)\right),
\end{aligned}$$

where we use the fact $\rho_i(x) \geq \rho_{j-1}(x), x \in [\mu_{i-1}, \mu_i]$. □

Remark 25. It is obvious that above bound can be significantly lowered if we consider terms with $j - i > 1$. Technique coincides with that of previous lemma. But for the case of $j - i = 1$, we cannot lower the bound anymore, which illustrates that the matrix G_W is in some sense a tridiagonal matrix.

We combine all the results in following proposition.

Proposition 23. *For $i < j$, we have following bound on crossing term in metric matrix:*

$$\begin{aligned}
g_{ij} =& \left|\int_{\mathbb{R}} \frac{(F_i - F_{i-1})(F_j - F_{j-1})}{\rho} dx\right| \\
\leq& \frac{e^{-\frac{\left(\mu_{i-1} - \mu_{j-1}\right)^2}{2\sigma^2}} \sigma^2}{p_{i-1}\sqrt{2\pi}} \left(\frac{\sqrt{\frac{\pi}{2}}\sigma}{\mu_{j-1} - \mu_{i-1}} + 1\right) + \frac{e^{-\frac{\left(\mu_i - \mu_j\right)^2}{2\sigma^2}} \sigma^2}{p_j\sqrt{2\pi}} \left(\frac{\sqrt{\frac{\pi}{2}}\sigma}{\mu_j - \mu_i} + 1\right) \\
&+ \frac{\sigma^3}{\mu_{j-1} - \mu_i - \sigma}\left(\frac{1}{2p_i} + \frac{1}{2p_{j-1}}\right) + \frac{\sigma^2}{\sqrt{2\pi}p_i} + \frac{\sigma^2}{p_i}\left(\sqrt{\frac{\pi}{2}} + \log\left(\frac{\mu_{j-1} - \mu_{i-1}}{\sigma + \mu_{j-1} - \mu_i}\right)\right) \\
&+ \frac{\sigma^2}{p_{j-1}}\left(\sqrt{\frac{\pi}{2}} + \log\left(\frac{\mu_j - \mu_i}{\sigma + \mu_{j-1} - \mu_i}\right)\right) \\
=& O\left(\frac{\sigma^2}{\min_l p_l}\right).
\end{aligned} \tag{21}$$

6.3. **Analysis of Diagonal Terms.** In this section, we approach diagonal terms which are dominant terms appear in metric tensor. It will be shown that the integration can be simplified to following simple form.

$$\int_{\mathbb{R}} \frac{(F_i - F_{i-1})^2}{\rho_\theta} dx \Longrightarrow \int_{\mu_{i-1}}^{\mu_i} \frac{1}{p_{i-1}\rho_{i-1} + p_i \rho_i} dx. \tag{22}$$

We further decompose the difference between these two integrals to several parts

$$\begin{aligned}
&\left|\int_{\mathbb{R}} \frac{(F_i - F_{i-1})^2}{\rho_\theta} dx - \int_{\mu_{i-1}}^{\mu_i} \frac{1}{p_{i-1}\rho_{i-1} + p_i\rho_i} dx\right| \\
\leq &\left|\int_{\mu_{i-1}}^{\mu_i} \frac{(F_i - F_{i-1})^2}{\rho_\theta} dx - \int_{\mu_{i-1}}^{\mu_i} \frac{(F_i - F_{i-1})^2}{p_{i-1}\rho_{i-1} + p_i\rho_i} dx\right| \\
&+ \left|\int_{-\infty}^{\mu_{i-1}} \frac{(F_i - F_{i-1})^2}{\rho} dx\right| + \left|\int_{\mu_i}^{\infty} \frac{(F_i - F_{i-1})^2}{\rho} dx\right| \\
&+ \left|\int_{\mu_{i-1}}^{\mu_i} \frac{1}{p_{i-1}\rho_{i-1} + p_i\rho_i} dx - \int_{\mu_{i-1}}^{\mu_i} \frac{(F_i - F_{i-1})^2}{p_{i-1}\rho_{i-1} + p_i\rho_i} dx\right|,
\end{aligned}$$

and consider different parts separately.

Lemma 6.

$$\begin{aligned}
\left|\int_{-\infty}^{\mu_{i-1}} \frac{(F_i - F_{i-1})^2}{\rho} dx\right| &\leq \frac{\sigma^2}{p_{i-1}} \left(\frac{\pi}{2} + \frac{1}{\sqrt{2\pi}}\right), \\
\left|\int_{\mu_i}^{\infty} \frac{(F_i - F_{i-1})^2}{\rho} dx\right| &\leq \frac{\sigma^2}{p_i} \left(\frac{\pi}{2} + \frac{1}{\sqrt{2\pi}}\right).
\end{aligned}$$

The proof is similar to that of lemma 3 and is omitted here.

Lemma 7.

$$\left|\int_{\mu_{i-1}}^{\mu_i} \frac{1}{p_{i-1}\rho_{i-1} + p_i\rho_i} dx - \int_{\mu_{i-1}}^{\mu_i} \frac{(F_i - F_{i-1})^2}{p_{i-1}\rho_{i-1} + p_i\rho_i} dx\right| \leq (\mu_i - \mu_{i-1}) \sqrt{2\pi}\sigma \left(\frac{1}{p_i} + \frac{1}{p_{i-1}}\right),$$

Proof. We have

$$\begin{aligned}
&\left|\int_{\mu_{i-1}}^{\mu_i} \frac{1}{p_{i-1}\rho_{i-1} + p_i\rho_i} dx - \int_{\mu_{i-1}}^{\mu_i} \frac{(F_i - F_{i-1})^2}{p_{i-1}\rho_{i-1} + p_i\rho_i} dx\right| \\
= &\left|\int_{\mu_{i-1}}^{\mu_i} \frac{(F_i - F_{i-1} - 1)(F_i - F_{i-1} + 1)}{p_{i-1}\rho_{i-1} + p_i\rho_i} dx\right| \\
\leq &\left|\int_{\mu_{i-1}}^{\mu_i} \frac{2(F_i - F_{i-1} + 1)}{p_{i-1}\rho_{i-1} + p_i\rho_i} dx\right| \\
\leq &\left|\int_{\mu_{i-1}}^{\mu_i} \frac{2(1 - F_{i-1})}{p_{i-1}\rho_{i-1} + p_i\rho_i} dx\right| + \left|\int_{\mu_{i-1}}^{\mu_i} \frac{2F_i}{p_{i-1}\rho_{i-1} + p_i\rho_i} dx\right|,
\end{aligned}$$

where we use the fact that

$$|F_i - F_{i-1} + 1| \leq 2,$$

and absolute value inequality. We just show the bound for one and the other follows equivalently.

$$\left|\int_{\mu_{i-1}}^{\mu_i} \frac{2F_i}{p_{i-1}\rho_{i-1} + p_i\rho_i} dx\right| \leq \frac{1}{p_i} \left|\int_{\mu_{i-1}}^{\mu_i} \frac{2F_i}{\rho_i} dx\right| \leq \frac{\sqrt{2\pi}\sigma(\mu_i - \mu_{i-1})}{p_i},$$

where we use the fact that

$$F(x) \leq \sqrt{\frac{\pi}{2}}\sigma\rho(x).$$

And for the other side, the following holds

$$\left|\int_{\mu_{i-1}}^{\mu_i} \frac{2\left(1-F_{i-1}\right)}{p_{i-1}\rho_{i-1}+p_i\rho_i}dx\right| \leq \frac{\left(\mu_i-\mu_{i-1}\right)\sqrt{2\pi}\sigma}{p_{i-1}}.$$

And the conclusion follows. □

Lemma 8.

$$\begin{aligned}&\left|\int_{\mu_{i-1}}^{\mu_i} \frac{\left(F_i-F_{i-1}\right)^2}{\rho_\theta}dx - \int_{\mu_{i-1}}^{\mu_i} \frac{\left(F_i-F_{i-1}\right)^2}{p_{i-1}\rho_{i-1}+p_i\rho_i}dx\right| \\ \leq\ & e^{-\frac{\left(\mu_i-\mu_{i-1}\right)^2}{2\sigma^2}} \frac{1}{\min\left\{p_{i-1},p_i\right\}} \int_{\mu_{i-1}}^{\mu_i} \frac{1}{p_{i-1}\rho_{i-1}+p_i\rho_i}dx.\end{aligned}$$

Proof. On the interval $[\mu_{i-1},\mu_i]$, we notice that densities of other components will be significantly smaller than the value of $i-1,i$ component. Specifically, we have

$$\max_{j\neq i,i-1}\left\{\rho_j\left(x\right)\right\} \leq e^{-\frac{\left(\mu_i-\mu_{i-1}\right)^2}{2\sigma^2}} \max\left\{\rho_{i-1}\left(x\right),\rho_i\left(x\right)\right\}, \quad x\in[\mu_{i-1},\mu_i].$$

Thus, we conclude

$$\sum_{j\neq i-1,i}\rho_j\left(x\right) \leq e^{-\frac{\left(\mu_i-\mu_{i-1}\right)^2}{2\sigma^2}} \frac{\sum_{j\neq i-1,i}p_j}{\min\left\{p_{i-1},p_i\right\}}\left(p_{i-1}\rho_{i-1}\left(x\right)+p_i\rho_i\left(x\right)\right), \qquad x\in[\mu_{i-1},\mu_i].$$

Plugging above formula to the formula we aim to bound,

$$\begin{aligned}&\left|\int_{\mu_{i-1}}^{\mu_i} \frac{\left(F_i-F_{i-1}\right)^2}{\rho_\theta}dx - \int_{\mu_{i-1}}^{\mu_i} \frac{\left(F_i-F_{i-1}\right)^2}{p_{i-1}\rho_{i-1}+p_i\rho_i}dx\right| \\ =\ & \left|\int_{\mu_{i-1}}^{\mu_i} \frac{\left(F_i-F_{i-1}\right)^2\left(\rho_\theta-p_{i-1}\rho_{i-1}-p_i\rho_i\right)}{\rho_\theta\left(p_{i-1}\rho_{i-1}+p_i\rho_i\right)}dx\right| \\ \leq\ & \left|\int_{\mu_{i-1}}^{\mu_i} \frac{\left(F_i-F_{i-1}\right)^2\frac{\left(\rho_\theta-p_{i-1}\rho_{i-1}-p_i\rho_i\right)}{\left(p_{i-1}\rho_{i-1}+p_i\rho_i\right)}}{\left(p_{i-1}\rho_{i-1}+p_i\rho_i\right)}dx\right| \\ \leq\ & \left|\int_{\mu_{i-1}}^{\mu_i} \frac{\left(F_i-F_{i-1}\right)^2 e^{-\frac{\left(\mu_i-\mu_{i-1}\right)^2}{2\sigma^2}}\frac{\sum_{j\neq i-1,i}p_j}{\min\left\{p_{i-1},p_i\right\}}}{\left(p_{i-1}\rho_{i-1}+p_i\rho_i\right)}dx\right| \\ \leq\ & e^{-\frac{\left(\mu_i-\mu_{i-1}\right)^2}{2\sigma^2}} \frac{1}{\min\left\{p_{i-1},p_i\right\}}\left|\int_{\mu_{i-1}}^{\mu_i} \frac{\left(F_i-F_{i-1}\right)^2}{p_{i-1}\rho_{i-1}+p_i\rho_i}dx\right| \\ \leq\ & e^{-\frac{\left(\mu_i-\mu_{i-1}\right)^2}{2\sigma^2}} \frac{1}{\min\left\{p_{i-1},p_i\right\}} \int_{\mu_{i-1}}^{\mu_i} \frac{1}{p_{i-1}\rho_{i-1}+p_i\rho_i}dx.\end{aligned}$$

□

Above proposition establishes the result that this error term is significantly small compared with the main part.

Combining all the results in hand, we conclude with following characterization of diagonal term for a GMM.

Proposition 24 (Non-asymptotic Analysis of Diagonal Term)**.** *The diagonal term of the Wasserstein metric in a GMM can be estimated as*

$$\begin{aligned}
&\left|\int_{\mathbb{R}} \frac{(F_i - F_{i-1})^2}{\rho_\theta} dx - \int_{\mu_{i-1}}^{\mu_i} \frac{1}{p_{i-1}\rho_{i-1} + p_i\rho_i} dx\right| \\
&\leq \sigma^2 \left(\frac{\pi}{2} + 1\right)\left(\frac{1}{p_{i-1}} + \frac{1}{p_i}\right) + (\mu_i - \mu_{i-1})\sqrt{2\pi}\sigma\left(\frac{1}{p_i} + \frac{1}{p_{i-1}}\right) \\
&\quad + e^{-\frac{(\mu_i - \mu_{i-1})^2}{2\sigma^2}} \frac{1}{\min\{p_{i-1}, p_i\}} \int_{\mu_{i-1}}^{\mu_i} \frac{1}{p_{i-1}\rho_{i-1} + p_i\rho_i} dx \\
&= O\left(\frac{\sigma}{\min_l p_l}\right).
\end{aligned} \tag{23}$$

Proof of Proposition 24. Using the estimation

$$\int_{\mu_{i-1}}^{\mu_i} \frac{1}{p_{i-1}\rho_{i-1} + p_i\rho_i} dx = O\left(\frac{e^{\frac{1}{2}\left(\frac{\mu_i - \mu_{i-1}}{2\sigma}\right)^2}}{\frac{\mu_i - \mu_{i-1}}{2\sigma}}\right),$$

we conclude

$$e^{-\frac{d^2}{2\sigma^2}} \frac{\sum_{j \neq i-1,i} p_j}{\min\{p_{i-1}, p_i\}} \int_{\mu_{i-1}}^{\mu_i} \frac{1}{p_{i-1}\rho_{i-1} + p_i\rho_i} dx = o(\sigma).$$

□

7. Discussions

In this paper, we introduce the Wasserstein information matrix in statistical models. Following 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, the ReLU push-forward family and the Gaussian mixture family. Also, by comparing both Wasserstein and Fisher information matrices, several 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 the Fisher information matrix and maximal likelihood estimator, what is the relation between the Wasserstein information matrix and the Wasserstein distance estimator? Is there a canonical Wasserstein divergence function for the Wasserstein information matrix? What is the corresponding Wasserstein maximal likelihood estimator? Meanwhile, we shall study the efficiency property in general statistical problems and algorithms including stochastic gradient descent by using Wasserstein

natural gradient. Lastly and most importantly, we have shown that the Wasserstein statistics provide the rigorous statistical advantages in implicit generative models than classical Fisher statistics. We will study properties of Wasserstein geometry in clear statistical terms over machine learning models.

Acknowledgement This project has received funding from AFOSR MURI FA9550-18-1-0502. Jiaxi Zhao is partially supported by the elite undergraduate training program of School of Mathematical Sciences at Peking University.

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] W. Gangbo, W. Li, S. Osher, and M. Puthawala. Unnormalized optimal transport. *arXiv preprint arXiv:1902.03367*, 2019.
[12] J. D. Lafferty. The density manifold and configuration space quantization. *Transactions of the American Mathematical Society*, 305(2):699–741, 1988.
[13] W. Li. Geometry of probability simplex via optimal transport. *arXiv:1803.06360 [math]*, 2018.
[14] W. Li, A. T. Lin, and G. Montúfar. Affine natural proximal learning. *Geometric science of information*, 2019, 2019.
[15] 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.
[16] W. Li and G. Montúfar. Natural gradient via optimal transport. *Information Geometry*, 2018.
[17] W. Li and G. Montúfar. Ricci curvature for parametric statistics via optimal transport. *arXiv:1807.07095 [cs, math, stat]*, 2018.
[18] A. T. Lin, W. Li, S. Osher, and G. Montufar. Wasserstein proximal of GANs, 2019.
[19] 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.
[20] Y. Ollivier. Online natural gradient as a Kalman filter. *Electronic Journal of Statistics*, 12(2):2930–2961, 2018.
[21] F. Otto. The geometry of dissipative evolution equations the porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.
[22] 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.
[23] A. Petersen and H.-G. Müller. Wasserstein covariance for multiple random densities. *Biometrika*, 106(2):339–351, 2019.
[24] C. Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

[25] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.
[26] T.-K. L. Wong. Logarithmic divergences from optimal transport and Rényi geometry. *Information Geometry*, 1(1):39–78, 2018.
[27] 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.

E-mail address: wcli@math.ucla.edu,Zjx98math@gmail.com