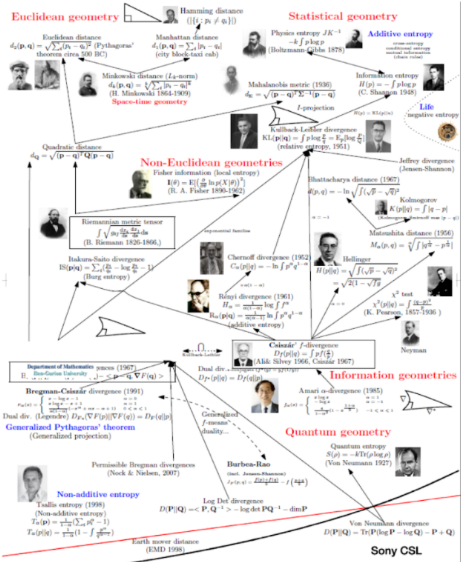
Wasserstein information matrix

Jiaxi Zhao (PKU)

joint work with Wuchen Li (UCLA)

June 2020

Distances on probability space



Euclidean geometry
Hamming distance
$(|\{i : p_i \neq q_i\}|)$
Statistical geometry
Euclidean distance
$d_2(\mathbf{p},\mathbf{q}) = \sqrt{\sum_i (p_i - q_i)^2}$ (Pythagoras' theorem circa 500 BC)
Manhattan distance
$d_1(\mathbf{p},\mathbf{q}) = \sum_i |p_i - q_i|$
(city block-taxi cab)
Physics entropy JK^{-1}
$-k\int p \log p$
(Boltzmann-Gibbs 1878)
Additive entropy
cross-entropy
conditional entropy
mutual information
(chain rules)
Minkowski distance (L_k-norm)
$d_k(\mathbf{p},\mathbf{q}) = \sqrt[k]{\sum_i |p_i - q_i|^k}$
(H. Minkowski 1864-1909)
Space-time geometry
Mahalanobis metric (1936)
$d_\Sigma = \sqrt{(\mathbf{p}-\mathbf{q})^T \Sigma^{-1} (\mathbf{p}-\mathbf{q})}$
Information entropy
$H(p) = -\int p \log p$
(C. Shannon 1948)
I-projection
$H(p) = \mathrm{KL}(p\|u)$
Life
negative entropy
Kullback-Leibler divergence
$\mathrm{KL}(\mathbf{p}\|\mathbf{q}) = \int p \log \frac{p}{q} = \mathrm{E}_p[\log \frac{P}{Q}]$
(relative entropy, 1951)
Quadratic distance
$d_{\mathbf{Q}} = \sqrt{(\mathbf{p}-\mathbf{q})^T \mathbf{Q} (\mathbf{p}-\mathbf{q})}$
Non-Euclidean geometries
Fisher information (local entropy)
$\mathbf{I}(\theta) = \mathrm{E}[\left(\frac{\partial}{\partial\theta} \ln p(X|\theta)\right)^2]$
(R. A. Fisher 1890-1962)
Jeffrey divergence
(Jensen-Shannon)
Bhattacharya distance (1967)
$d(p,q) = -\ln \sqrt{\int (\sqrt{p} - \sqrt{q})^2}$
Kolmogorov
$K(p\|q) = \int |q - p|$
(Kolmogorov-Smirnoff max $|p-q|$)
$\alpha = -1$
Riemannian metric tensor
$\int \sqrt{g_{ij} \frac{dx_i}{ds} \frac{dx_j}{ds}} ds$
(B. Riemann 1826-1866,)
exponential families
Matsushita distance (1956)
$M_\alpha(p,q) = \sqrt[\alpha]{\int |q^{\frac{1}{\alpha}} - p^{\frac{1}{\alpha}}|}$
Itakura-Saito divergence
$\mathrm{IS}(\mathbf{p}|\mathbf{q}) = \sum_i (\frac{p_i}{q_i} - \log \frac{p_i}{q_i} - 1)$
(Burg entropy)
Chernoff divergence (1952)
$C_\alpha(p\|q) = -\ln \int p^\alpha q^{1-\alpha}$
Hellinger
$H(p\|q) = \sqrt{\int (\sqrt{p} - \sqrt{q})^2}$
$= \sqrt{2(1 - \sqrt{fg}}$
$\times\alpha(1-\alpha)$
Rényi divergence (1961)
$H_\alpha = \frac{1}{\alpha(1-\alpha)} \log \int f^\alpha$
$\mathrm{R}_\alpha(\mathbf{p}|\mathbf{q}) = \frac{1}{\alpha(\alpha-1)} \ln \int p^\alpha q^{1-\alpha}$
(additive entropy)
$\alpha = 0$
χ^2 test
$\chi^2(p\|q) = \int \frac{(q-p)^2}{p}$
(K. Pearson, 1857-1936)
Kullback-Leibler
Csiszár' f-divergence
$D_f(p\|q) = \int p f(\frac{q}{p})$
(Ali& Silvey 1966, Csiszár 1967)
Neyman
Department of Mathematics
Ben-Gurion University
B. ...ences (1967)
...)$- < \mathbf{p} - \mathbf{q}, \nabla F(\mathbf{q}) >$
Dual div. conjugate ($f^*(y) = yf(1/y)$)
$D_{f^*}(p\|q) = D_f(q\|p)$
Information geometries
Bregman-Csiszár divergence (1991)
$F_\alpha(x) = \begin{cases} x - \log x - 1 & \alpha = 0 \\ x \log x - x + 1 & \alpha = 1 \\ \frac{1}{\alpha(1-\alpha)}(-x^\alpha + \alpha x - \alpha + 1) & 0 < \alpha < 1 \end{cases}$
Dual div. (Legendre) $D_{F_*}(\nabla F(p)\|\nabla F(q)) = D_F(q\|p)$
Amari α-divergence (1985)
$f_\alpha(x) = \begin{cases} x \log x & \alpha = 1 \\ -\log x & \alpha = -1 \\ \frac{4}{1-\alpha^2}(1 - x^{\frac{1+\alpha}{2}}) & -1 < \alpha < 1 \end{cases}$
∇
∇^*
Generalized Pythagoras' theorem
(Generalized projection)
Generalized
f-means
duality...
Quantum geometry
Quantum entropy
$S(\rho) = -k\mathrm{Tr}(\rho \log \rho)$
(Von Neumann 1927)
Permissible Bregman divergences
(Nock & Nielsen, 2007)
Burbea-Rao
(incl. Jensen-Shannon)
$J_F(p;q) = \frac{f(p)+f(q)}{2} - f\left(\frac{p+q}{2}\right)$
Non-additive entropy
Tsallis entropy (1998)
(Non-additive entropy)
$T_\alpha(\mathbf{p}) = \frac{1}{1-\alpha}(\sum_i p_i^\alpha - 1)$
$T_\alpha(p\|q) = \frac{1}{1-\alpha}(1 - \int \frac{p^\alpha}{q^{\alpha-1}})$
Log Det divergence
$D(\mathbf{P}\|\mathbf{Q}) = < \mathbf{P}, \mathbf{Q}^{-1} > - \log\det \mathbf{PQ}^{-1} - \dim\mathbf{P}$
Von Neumann divergence
$D(\mathbf{P}\|\mathbf{Q}) = \mathrm{Tr}(\mathbf{P}(\log \mathbf{P} - \log \mathbf{Q}) - \mathbf{P} + \mathbf{Q})$
Earth mover distance
(EMD 1998)
Sony CSL

Information matrix

Information matrix (a.k.a Fisher information matrix, Fisher-Rao metric) plays important roles in information science, statistics and machine learning:

1. Population Games via Replicator Dynamics (Shahshahani, Smith);

2. Machine learning: Natural gradient (Amari); ADAM (Kingma 2014); Stochastic relaxation (Malago) and many more in book *Information geometry* (Ay et.al.).

3. **Statistics**: Kullback–Leibler (KL) divergence $\leftrightarrow$ Likelihood principle, Fisher information matrix $\leftrightarrow$ Cramer-Rao bound,

Learning

Given a data measure $\rho_{\mathrm{data}}(x) = \frac{1}{N}\sum_{i=1}^{N}\delta_{X_i}(x)$ and a parameterized model $\rho(x;\theta)$. Machine learning problems often refer to

$$\min_{\rho_\theta \in \rho(\Theta)} \quad D(\rho_{\mathrm{data}}, \rho_\theta).$$

One typical choice of D is the Kullback–Leibler divergence (relative entropy)

$$D(\rho_{\mathrm{data}}, \rho_\theta) = \int_\Omega \rho_{\mathrm{data}}(x) \log \frac{\rho_{\mathrm{data}}(x)}{\rho(x;\theta)} dx.$$

Natural gradient

The natural gradient method refers to

$$\theta^{k+1} = \theta^k - h G_F(\theta^k)^{-1} \nabla_\theta D(\rho_{\mathrm{data}}, \rho_{\theta^k}),$$

where $h > 0$ is a stepsize and

$$G_F(\theta) = \mathbb{E}_{X \sim \rho_\theta} (\nabla_\theta \log \rho(X;\theta))^T (\nabla_\theta \log \rho(X;\theta))$$

is the Fisher information matrix and $\nabla_\theta \log \rho(X;\theta)$ is named score function.

Why natural gradient and Fisher information matrix?

1. Parameterization invariant;
2. Pre-conditioners for KL divergence related learning problems;
3. Online Cramer-Rao bound.

Optimal transport

In recent years, optimal transport (a.k.a Earth mover's distance, Monge-Kantorovich problem, Wasserstein metric) has witnessed a lot of applications:

1. Population Games via Fokker-Planck Equations (Degond et. al. 2014, Li et.al. 2016);

2. Machine learning: Wasserstein Training of Boltzmann Machines (Cuturi et.al. 2015); Learning from Wasserstein Loss (Frogner et.al. 2015); Wasserstein GAN (Bottou et.al. 2017); Wasserstein statistics, and many more in NIPS 2015, 2016, 2017, 2018, 2019.

Why optimal transport?

Optimal transport provides a particular distance (W) among histograms, which relies on the distance on sample spaces (ground cost c).

$$W_c\left(\rho,\mu\right)=\inf_{\pi\in\prod(\rho,\mu)}\int c\left(x,y\right)d\pi\left(x,y\right),$$

$$W_p^p\left(\rho,\mu\right)=\inf_{\pi\in\prod(\rho,\mu)}\int \left|x-y\right|^p d\pi\left(x,y\right)\ \text{(Wasserstein-p distance).}$$

Denote $X_0\sim\rho_0=\delta_{x_0}$, $X_1\sim\rho_1=\delta_{x_1}$. Compare

$$W(\rho_0,\rho_1)=\inf_{\pi\in\Pi(\rho_0,\rho_1)}\mathbb{E}_{(X_0,X_1)\sim\pi}c(X_0,X_1)=c(x_0,x_1);$$

Vs

$$\mathrm{TV}(\rho_0,\rho_1)=\int_\Omega|\rho_0(x)-\rho_1(x)|dx=2;$$

Vs

$$\mathrm{KL}(\rho_0\|\rho_1)=\int_\Omega\rho_0(x)\log\frac{\rho_0(x)}{\rho_1(x)}dx=\infty.$$

Wasserstein Loss function

Given a data distribution ρ_0 and a probability model ρ_θ. Consider

$$\min_{\theta \in \Theta} W(\rho_0, \rho_\theta).$$

This is a double minimization problem, i.e.

$$W(\rho_0, \rho_\theta) = \min_{\pi \in \Pi(\rho_0, \rho_\theta)} \mathbb{E}_{(X_0, X_1) \sim \pi} c(X_0, X_1).$$

Many applications, such as Wasserstein GAN, Wasserstein Loss, are built on the above formulation.

Goals

Main Question:

Instead of looking at the Wasserstein metric, we propose the optimal transport induced **information matrix in probability models**, and study its properties on statistics and machine learning problems.

Related studies

1. Wasserstein covariance (Petersen, Muller.)
2. Wasserstein minimal distance estimator (Bernton, Jacob, Gerber, and Robert.)
3. Wasserstein natural gradient (Li, Montufar, Chen, Lin, Abel et.al.)
4. Statistical inference for generative models with maximum mean discrepancy (Briol, Barp, Duncan, Girolami.)

Problem formulation

1. Mapping formulation: Monge problem (1781): Monge-Ampére equation ;

2. Statical formulation: Kantorovich problem (1940): Linear programming ;

3. Dynamical formulation: Density optimal control (Nelson, Lafferty, Otto, Villani).

In this talk, we will apply density optimal control into learning problems.

Density manifold

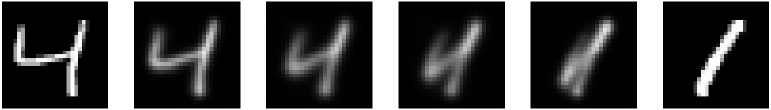
Optimal transport has an optimal control reformulation by the dual of dual of linear programming:

$$\inf_{\rho_t} \int_0^1 g_W(\partial_t \rho_t, \partial_t \rho_t) dt = \int_0^1 \int_\Omega (\nabla \Phi_t, \nabla \Phi_t) \rho_t dx dt,$$

under the dynamical constraint, i.e. continuity equation:

$$\partial_t \rho_t + \nabla \cdot (\rho_t \nabla \Phi_t) = 0, \quad \rho_0 = \rho^0, \quad \rho_1 = \rho^1.$$

Here, $(\mathcal{P}(\Omega), g_W)$ forms an infinite-dimensional Riemannian manifold[1].



[1] John D. Lafferty: The density manifold and configuration space quantization, 1988.

Information matrix

In parametric statistics, we study the finite dimensional statistical models parametrized by several parameters. Thus we need to establish a metric (information matrix) on these spaces.

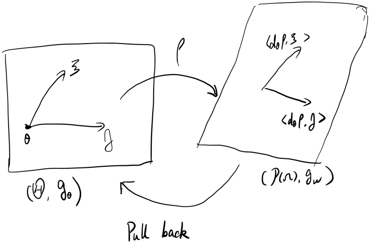

ξ
θ
η
(Θ, g_θ)
ρ
$\langle d_\theta\rho, \xi\rangle$
$\langle d_\theta\rho, \eta\rangle$
$(\mathcal{P}(\Omega), g_W)$
Pull back

Figure: Pull-back of the metric

Statistical information matrix

Definition (Statistical Information Matrix)

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

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

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

Statistical information matrix

Definition (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:

$$\mathbb{E}_{x\sim p_\theta} \Phi_i(x;\theta) = 0, \qquad i = 1, ..., n.$$

Then the statistical metric tensor satisfies

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

Statistical information matrix

Formulating information matrices as expectations:

Fisher Information Matrix

$$G_F(\theta)_{ij} = \mathbb{E}_{p_\theta}\left(\frac{\partial}{\partial\theta_i}\log p(x;\theta)\frac{\partial}{\partial\theta_j}\log p(x;\theta)\right).$$

Wasserstein Information Matrix

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

Poisson equation

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

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

Separability

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

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

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

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

In addition, the Wasserstein information matrix is separable:

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

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

One dimensional sample space

If $\mathcal{X} \subset \mathbb{R}^1$, the Wasserstein score functions satisfy

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

where $F(x;\theta) = \int p(y;\theta)dy$ is the cumulative distribution function. And the Wasserstein information matrix satisfies

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

Analytic examples of Wasserstein information matrix

Classical statistical models:

Gaussian family:

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2},$$
$$G_W(\mu,\sigma) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Laplacian family:

$$p(x;m,\lambda) = \frac{\lambda}{2} e^{-\lambda|x-m|},$$
$$G_W(\mu,\sigma) = \begin{pmatrix} 1 & 0 \\ 0 & \frac{2}{\lambda^4} \end{pmatrix}.$$

Analytic examples of Wasserstein information matrix

Generative models:

Continuous 1-d generative family:

$$p(\cdot;\theta) = f_{\theta *} p_0(\cdot), \ p_0 \text{ a given distribution},$$
$$G_W(\theta) = \int |\nabla_\theta f_\theta(x)|^2 p_0(x)\,dx,$$

where the push-forward distribution is defined as

$$\int_A p_0\,dx = \int_{f_\theta^{-1}(A)} f_{\theta *} p_0\,dx,$$

Analytic examples of Wasserstein information matrix

Generative models with ReLU family:

$$f_\theta(x) = \ \sigma(x-\theta) = \begin{cases} 0, & x \leq \theta, \\ x-\theta, & x > \theta. \end{cases}$$

$$G_W(\theta) = F_0(\theta),\ F_0 \text{ cumulative distribution function of } p_0.$$

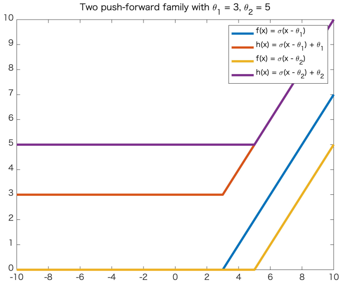


Figure: This figure plots two examples of the push-forward family with $\theta_1 = 3, \theta_2 = 5$.

Statistical Information Matrix

Probability Family	Wasserstein information matrix	Fisher information matrix
Uniform: $p(x;a,b)=\frac{1}{b-a}\mathbf{1}_{(a,b)}(x)$	$G_W(a,b)=\frac{1}{3}\begin{pmatrix}1 & \frac{1}{2}\\ \frac{1}{2} & 1\end{pmatrix}$	$G_F(a,b)$ not well-defined
Gaussian: $p(x;\mu,\sigma)=\frac{e^{-\frac{1}{2\sigma^2}(x-\mu)^2}}{\sqrt{2\pi}\sigma}$	$G_W(\mu,\sigma)=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$	$G_F(\mu,\sigma)=\begin{pmatrix}\frac{1}{\sigma^2} & 0\\ 0 & \frac{2}{\sigma^2}\end{pmatrix}$
Exponential: $p(x;m,\lambda)=\lambda e^{-\lambda(x-m)}$	$G_W(m,\lambda)=\begin{pmatrix}1 & \frac{1}{\lambda^2}\\ \frac{1}{\lambda^2} & \frac{2}{\lambda^4}\end{pmatrix}$	$G_F(m,\lambda)$ not well-defined
Laplacian: $p(x;m,\lambda)=\frac{\lambda}{2}e^{-\lambda\|x-m\|}$	$G_W(m,\lambda)=\begin{pmatrix}1 & 0\\ 0 & \frac{2}{\lambda^4}\end{pmatrix}$	$G_F(m,\lambda)=\begin{pmatrix}\lambda^2 & 0\\ 0 & \frac{1}{\lambda^2}\end{pmatrix}$
Location-scale: $p(x;m,\lambda)=\frac{1}{\lambda}p(\frac{x-p}{\lambda})$	$G_W(\lambda,m)=\begin{pmatrix}\frac{\mathbb{E}_{\lambda,m}x^2-2m\mathbb{E}_{\lambda,m}x+m^2}{\lambda^2} & 0\\ 0 & 1\end{pmatrix}$	$G_F(\lambda,m)=\begin{pmatrix}\frac{1}{\lambda^2}\left(1+\int_{\mathbb{R}}\left(\frac{(x-m)^2p'^2}{\lambda^2 p}+\frac{(x-m)p'}{\lambda}\right)dx\right) & \int_{\mathbb{R}}\frac{(x-m)p'^2}{\lambda^3 p}dx\\ \int_{\mathbb{R}}\frac{(x-m)p'^2}{\lambda^3 p}dx & \frac{1}{\lambda^2}\int_{\mathbb{R}}\frac{p'^2}{p}dx\end{pmatrix}$
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..	$G_W(\theta)=F(\theta)$, F cdf of $p(x)$	$G_F(\theta)$ not well-defined

Table: In this table, we present Wasserstein, Fisher information matrices for various probability families.

Classical (Fisher) statistics & Wasserstein statistics

We develop a parallel Wasserstein statistics following the classical statistics approach

1. Covariance $\xrightarrow{\text{inner product}}$ Wasserstein covariance
2. Cramer-Rao bound $\xrightarrow{\text{cotangent space}}$ Wasserstein-Cramer-Rao bound
3. Natural gradient efficiency $\xrightarrow{\text{separability}}$ Wasserstein Natural gradient efficiency

Wasserstein covariance

Definition (Wasserstein covariance)

Given a statistical model Θ, denote the Wasserstein covariance as follows:

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

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:

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

Problem

How does the Wasserstein covariance describe the variance of an estimator for Wasserstein stochastic gradient descent?

Wasserstein-Cramer-Rao bound

Theorem (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* $\mathrm{Cov}_\theta^W[T(x)]$, $\nabla_\theta \mathbb{E}_{p_\theta}[T(x)]^T$ *as below:*

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

then

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

Cramer-Rao bound: Fisher vs Wasserstein

Gaussian:

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

Exponential:

$$G_W(m,\lambda)=\begin{pmatrix}1 & \frac{1}{\lambda^2}\\ \frac{1}{\lambda^2} & \frac{2}{\lambda^4}\end{pmatrix},\ G_F \text{ not well-defined.}$$

Comparison: G_W is well-defined for wider range of families.

Tighter bound on the variance of an estimator

Wasserstein natural gradient

Given a Loss function $\mathcal{F}\colon \mathcal{P}(\Omega) \to \mathbb{R}$ and probability model $\rho(\cdot, \theta)$, the associated gradient flow on a Riemannian manifold is defined by

$$\frac{d\theta}{dt} = -\nabla_g \mathcal{F}(\rho(\cdot, \theta)).$$

Here ∇_g is the Riemannian gradient operator satisfying

$$g_\theta(\nabla_g \mathcal{F}(\rho(\cdot, \theta)), \xi) = \nabla_\theta \mathcal{F}(\rho(\cdot, \theta)) \cdot \xi$$

for any tangent vector $\xi \in T_\theta \Theta$, where ∇_θ represents the w.r.t. θ (Euclidean gradient).

Wasserstein natural gradient

The gradient flow of Loss function $\mathcal{F}(\rho(\cdot,\theta))$ in (Θ, g_θ) satisfies

$$\frac{d\theta}{dt} = -G_W(\theta)^{-1}\nabla_\theta \mathcal{F}(\rho(\cdot,\theta)).$$

If $\rho(\cdot;\theta) = \rho(x)$ is an identity map, then we recover the Wasserstein gradient flow in full probability space:

$$\partial_t \rho_t = \nabla \cdot (\rho_t \nabla \frac{\delta}{\delta\rho}\mathcal{F}(\rho_t)).$$

Online natural gradient algorithm

We sample from the unknown distribution once in each step, and use sample x_t to get new estimator θ_{t+1}

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

In order to analysis the convergence of this online algorithm, we define the Wasserstein covariance matrix V_t to be

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

where θ_* is the optimal value.

Definition (Natural gradient efficiency)

The Wasserstein natural gradient is asymptotic efficient if

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

Covariance update

Theorem (Variance updating equation of the Wasserstein Natural Gradient)

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

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

Wasserstein online efficiency

Corollary (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:

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

Poincare online efficiency

Corollary (Poincare Efficiency)

For the dynamics

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

where $l(x_t, \theta_t) = \log p\left(x_t, \theta_t\right)$ *is the log-likelihood function. The Wasserstein covariance updates according to*

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

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

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

where

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

Functional inequalities via Wasserstein and Fisher information matrices

Theorem (RIW-condition[a] [b])

[a] Li, Geometry of probability simplex via optimal transport, 2018.

[b] Li, Montufar, Ricci curvature for parameter statistics, 2018.

The information matrix criterion for LSI 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:*

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

Numerical experiments

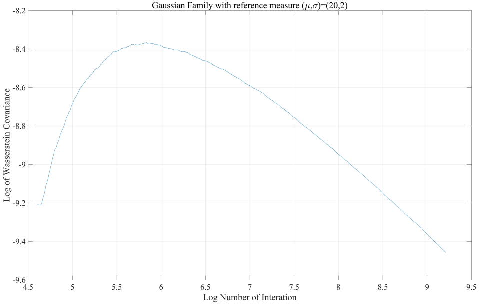


Figure: Poincaré Type Convergence Rate. X-axis is logarithm of iteration t while y-axis is logarithm of the Wasserstein covariance V_t. The reference Gaussian is chosen to be $\mathcal{N}(20, 2)$ where the parameter $\mu_* = 20$ is the critical point. Since we have $\frac{1}{\sigma_*^2} = \frac{1}{4} < \frac{1}{2}$, 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 (9) holds.

Future works

Design Wasserstein divergence and Wasserstein likelihood principle.

Apply Wasserstein geometry on ReLU family to analyze the design architecture of generative models in machine learning.

Conduct analysis and scientific computation in the framework of Wasserstein information geometry for Gaussian mixture models.