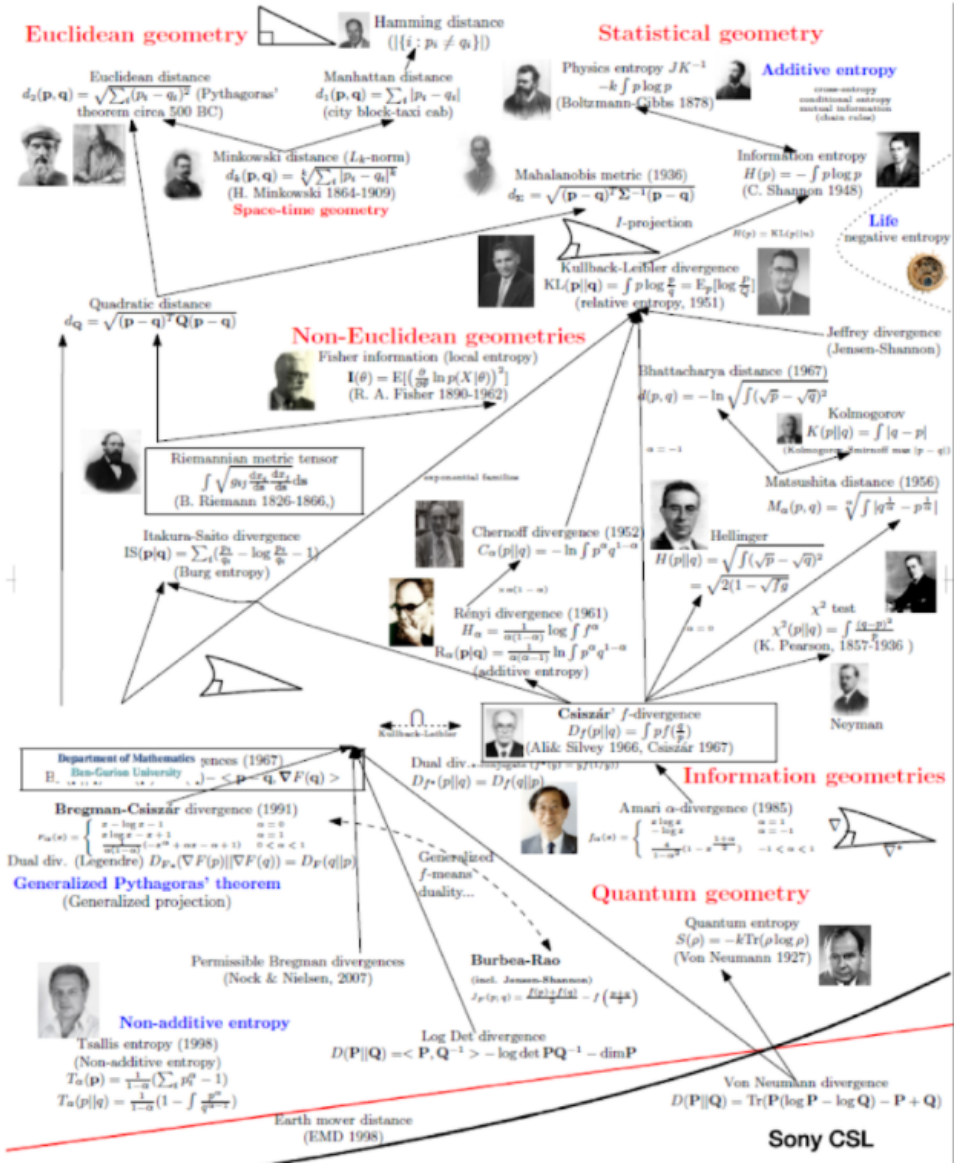
Interview for NUS

Jiaxi Zhao (Stony Brook University)

July 2021

Distances on probability space



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

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

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.

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: $p(x;m,\lambda)=\frac{\lambda}{2}e^{-\lambda\lvert x-m\rvert}$	$G_W(m,\lambda)=\begin{pmatrix}1 & 0\\ 0 & \frac{2}{\lambda^4}\end{pmatrix}$	$G_F(m,\lambda)=\begin{pmatrix}\lambda^2 & 0\\ 0 & \frac{1}{\lambda^2}\end{pmatrix}$
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..	$G_W(\theta)=F(\theta)$, 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.

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

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

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

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

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

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

where

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

Scaling limit of the Gaussian mixture model (GMM)

Gaussian mixture models (GMM) are widely used in statistical inference (statistical models) and scientific computation.

A simple review of GMM:

$$\rho : \Theta \to \mathcal{P}(\mathbb{R}), \quad \Theta \subset \mathbb{R}^{N-1},$$

$$\theta \mapsto \rho_\theta = \sum_{i=1}^{N-1} \theta_i (\rho_{i+1} - \rho_i) + \rho_1 = \sum_{i=1}^{N} p_i \rho_i,$$

$$1 = \theta_0 > \theta_1 > \cdots > \theta_{N-1} > \theta_N = 0,$$

$$\rho_i \sim \mathcal{N}(\mu_i, \sigma^2), \ \mu_1 < \mu_2 < \cdots < \mu_{N-1} < \mu_N,$$

This work is based on the consideration of GMM.

Our approach

We focus on the one-dimensional sample space to utilize the explicit formula for Wasserstein metric.

Step 1: Define a (Fisher-Rao/Wasserstein) pull-back metric $G_W(\theta;\sigma)$ on GMMs, where θ is parameter and σ is the variance.

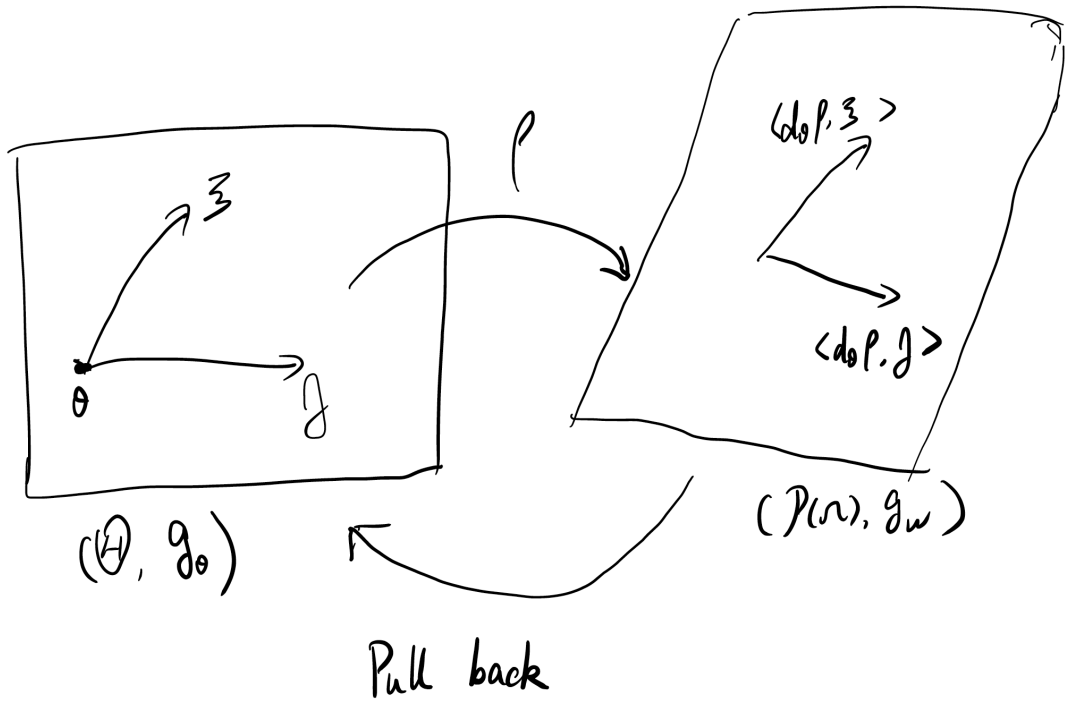


Figure: Pull-back of the metric

Our approach

Step 2: Rescale GMMs to approximate discrete spaces. As the variance of Gaussian tends to 0, distributions weakly converge to a Dirac measure centered at its mean.

$$\lim_{\sigma \to 0} \frac{G_W(\theta;\sigma)}{K(\sigma)} = G_{\widetilde{W}}(\theta).$$

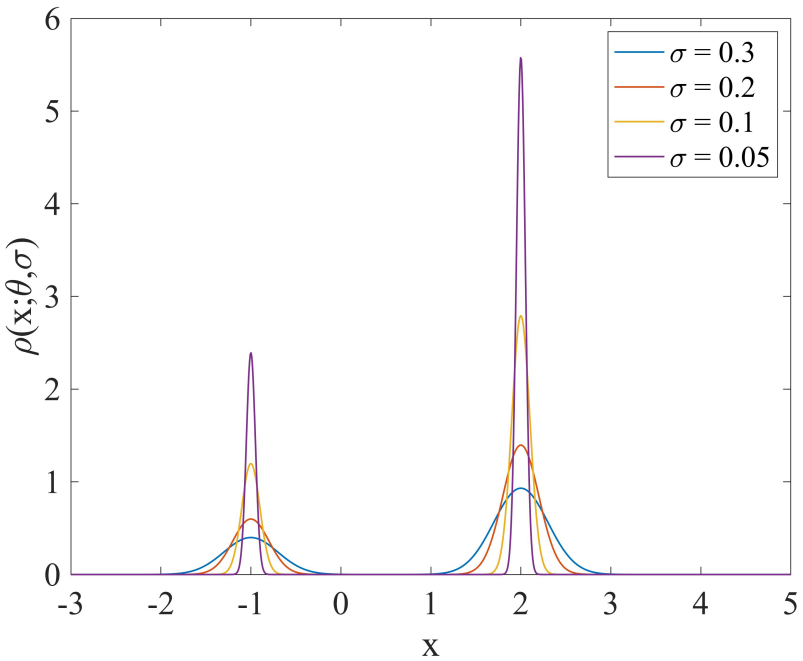


Figure: The figure plots the density function of a family of Gaussian mixture given by $\rho(x) \sim 0.3 * \mathcal{N}(-1,\sigma) + 0.7 * \mathcal{N}(2,\sigma)$.

Scaling approximation Wasserstein metric

Theorem (Scaling Wasserstein metric)

For a 1-d homogeneous GMM with difference between adjacent components given by d, a scaling limit of Wasserstein information matrices is given by

$$G_{\widetilde{W}}(\theta) = \lim_{\sigma \to 0} \frac{G_W(\theta;\sigma)}{K(\sigma)} = \begin{pmatrix} \frac{1}{\sqrt{p_1 p_2}} & 0 & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{\sqrt{p_2 p_3}} & 0 & \cdots & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{p_3 p_4}} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \frac{1}{\sqrt{p_{N-1} p_N}} \end{pmatrix}$$

The scaling factor appearing above is given by

$$K(\sigma) = \sqrt{2\pi^3}\frac{\sigma^3}{d} e^{\frac{1}{2}\left(\frac{d}{2\sigma}\right)^2}.$$

Gradient flows in scaling Wasserstein geometry

Following the discussion in Villani, we derive the gradient flow equations of potential, internal, and interaction energy functional on density manifold and scaling Wasserstein geometry

internal energy:
$$\mathcal{U}(\rho) = \int U(\rho(x))\, dx = \sum_{i=1}^{N} U(p_i),$$

potential energy:
$$\mathcal{V}(\rho) = \int V(x)\, d\rho = \sum_{i=1}^{N} V_i p_i,$$

interaction energy:
$$\mathcal{W}(\rho) = \frac{1}{2}\int\int W(x-y)\,\rho(x)\,\rho(y)\, dxdy = \sum_{i,j=1}^{N} W_{ij} p_i p_j.$$

Gradient flows in scaling Wasserstein geometry

$$
\begin{aligned}
\textbf{internal energy:} \quad \dot{p}_i &= -\sqrt{p_i p_{i-1}}\left(U'(p_i) - U'(p_{i-1})\right) \\
&\quad + \sqrt{p_i p_{i+1}}\left(U'(p_{i+1}) - U'(p_i)\right), \\
\textbf{potential energy:} \quad \dot{p}_i &= -\sqrt{p_i p_{i-1}}\left(V_i - V_{i-1}\right) + \sqrt{p_i p_{i+1}}\left(V_{i+1} - V_i\right), \\
\textbf{interaction energy:} \quad \dot{p}_i &= -\sqrt{p_i p_{i-1}}\left(\sum_{k=1}^{N} W_{ik} p_k - \sum_{k=1}^{N} W_{i-1,k} p_k\right) \\
&\quad + \sqrt{p_i p_{i+1}}\left(\sum_{k=1}^{N} W_{i+1,k} p_k - \sum_{k=1}^{N} W_{ik} p_k\right).
\end{aligned}
$$

We prove the consistency of these numerical discretization, i.e. they converge to their continuous counterpart.

Further questions: Numerical schemes

Viewing this discretization as numerical schemes for the corresponding PDE, is there any advantages comparing to FDM and also JKO scheme? If we denote the parameters μ_i as the grid points, does changing the μ_is correspond to moving mesh method?

Kernelized gradient flow for WGMM

Another approach for the Wasserstein Gaussian Mixture Model is kernelized gradient flow, we solve the gradient in a RKHS.

Why Gaussian mixture model? Some connection with quantum mechanics. Solving Schordinger equations in GMM.

Semi-discrete OT: optimization and generalization

We consider to solve the semi-discrete OT as follows

$$\inf_{b} \sum_{i=1}^{N} \nu_i b_i + \int \max_i \{x \cdot y_i - b_i\} \, d\rho(x) := F(b). \qquad (1)$$

Optimization

Suppose we know the position of each discrete components, i.e. y_is, then this is a strongly convex optimization problem that enjoy second order convergence if we use Newton's method or first order convergence if we use simply gradient descent.

Proposition

Suppose the base measure is the uniform measure on unit ball $B_1^d \subset \mathbb{R}^d$. *Consider the power diagram associated with* $\{(y_1, b_1), (y_2, b_2), \cdots, (y_N, b_N)\}$ *and function* $F(b)$ *as a function of* b. *Suppose we have that the ratio of mass in each diagram is* $\frac{1}{N}$, *then the matrix* $\operatorname{Hess} F(b)$ *has an eigenvalue* 0 *and all the remaining eigenvalue greater than* $C = \dfrac{\left(\left(\frac{1}{N}\right)^{\frac{d-1}{d}} + \left(1-\frac{1}{N}\right)^{\frac{d-1}{d}} - 1\right)}{8N^3}$.

Generalization

For generalization, we leverage the localized Rademacher complexity technique (since the loss function is convex) which very often appears in recent theoretic works including ML, RL, Stats.

Then, we can trade off the optimization and generalization error to obtain an optimal bound.

Theorem

The generated distribution based on this semi-discrete formulation of optimal transport provide a measure whose MMD is close to target one in the sense that

$$d_{\mathcal{H}}^{MMD}\left(\widehat{\nu_b}, \nu\right) \leq \sqrt{\frac{4QNd}{Cn}} + \frac{D}{\sqrt{N}}.$$

Drawback

However, semi-discrete OT has the problem that it does not have generalization ability from the sampling perspective. Wasserstein barycenter may be another choice, but researchers mainly use it for data augmentation.