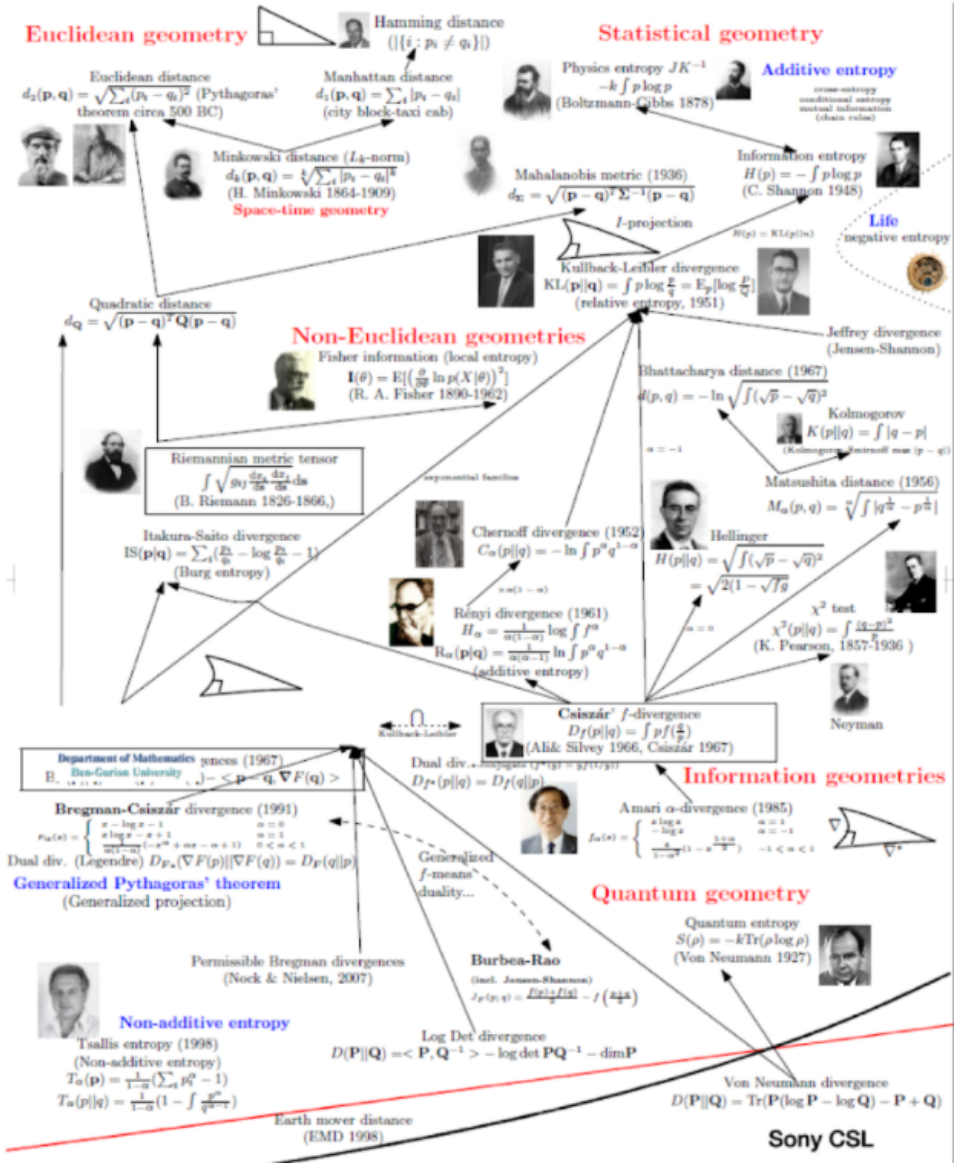
Scaling limits of Wasserstein Gaussian mixture models

Jiaxi Zhao (NUS)

joint work with Wuchen Li (USC)

SCLA workshop Dec 2023

Distances on probability space



Statistics and machine learning

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

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

The loss function is chosen to be a distance or a divergence on probability spaces which measures the difference between distributions. One typical choice of D is the Kullback–Leibler divergence (relative entropy)

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

Scientific computing

Distance and geometric structure in probability space play important roles in scientific computation, both theoretic and computational:

1. PDE and especially mean-field games (Osher, Gangbo, Li).
2. Accelerated algorithms (Li et. al. 2019, 2020).
3. Deep learning approach in scientific computation (Osher et. al. 2020).

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.

Gaussian mixture model (GMM)

Geometric structures on GMM:

1. Information geometry (Fisher-Rao metric): mixture family, mixture connection, duality, etc (N. Ay et. al., 2017).

2. Wasserstein geometry: some researchers already consider optimal transport in GMM (Chen et. al. 2017). But they focus on the perspective of distance.

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. Mean field games (Lions, Gangbo, Osher);
2. Population Games via Fokker-Planck Equations (Li et.al. 2016, Degond et. al. 2014);
3. Machine learning: Wasserstein Training of Boltzmann Machines (Cuturi et.al. 2015); Learning from Wasserstein Loss (Li et.al. 2018); Wasserstein GAN; Wasserstein statistics, and many more in NIPS 2015, 2016, 2017, 2018, 2019.

Why optimal transport?

Optimal transport provides a particular distance (W) among distributions, which relies on the distance on sample spaces Ω (ground cost $c : \Omega \times \Omega \to \mathbb{R}_{\geq 0}$).

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

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

Denote $X_0 \sim \rho^0 = \delta_{x_0}$, $X_1 \sim \rho^1 = \delta_{x_1}$. Compare

$$W_c(\rho^0, \rho^1) = \inf_{\pi \in \Pi(\rho^0,\rho^1)} \mathbb{E}_{(X_0,X_1)\sim\pi} c(X_0, X_1) = c(x_0, x_1);$$

Vs

$$\mathrm{TV}(\rho^0, \rho^1) = \int_\Omega |\rho^0(x) - \rho^1(x)| dx = 2;$$

Vs

$$\mathrm{KL}(\rho^0 \| \rho^1) = \int_\Omega \rho^0(x) \log \frac{\rho^0(x)}{\rho^1(x)} dx = \infty.$$

Dynamical formulation

Wasserstein-2 distance has a dynamical formulation (Brenier et. al. 2000), namely it minimizes the kinetic energy of an evolution between μ and ν:

$$W_2^2(\nu,\mu) = \inf_{\rho_t} \int_0^1 g_W(\partial_t\rho_t, \partial_t\rho_t)dt = \int_0^1 \int_\Omega \nabla\Phi_t \cdot \nabla\Phi_t \rho_t dxdt,$$
$$s.t.\ \rho_0 = \nu, \rho_1 = \mu,$$
$$\partial_t\rho_t + \nabla\cdot(\rho_t\nabla\Phi_t) = 0.$$

This dynamical formulation induces a metric structure on probability space (density manifold) $(\mathcal{P}(\Omega), g_W)$, which is an infinite-dimensional Riemannian manifold. This requires to solve a PDE in high dimension while it has an explicit formula in 1-d sample space.

Review of optimal transport on graph

Although it had been deeply understood when sample spaces are Riemannian manifolds, defining a canonical theory of Wasserstein geometry on graphs remains a question.

Several groups already establish some results on this direction:

1. Heat flow on graph (Li et.al. 2017, 2018), under the choice of weight:

$$\nabla \Phi_{ij} = \sqrt{w_{ij}} \left(\Phi_i - \Phi_j \right),$$

$$\left(\nabla \cdot (\rho \nabla \Phi) \right)_i = - \sum_{j \in N(i)} \sqrt{w_{ij}} \left(\Phi_i - \Phi_j \right) \left(\frac{\rho_i + \rho_j}{2} \right).$$

2. Geodesic in probability simplex. (W. Gangbo et. al. 2017)

3. Markov jump process as gradient flows. (J. Maas et. al. 2011, 2018)

Goals

Main Question:

Is there a natural definition and a dynamical formulation of optimal transport on graphs or discrete spaces? i.e. is there a natural choice of weight for graph Laplacian? Can we use this to design numerical schemes?

More related studies

- Wasserstein natural gradient (Li, Osher, Montufar et. al.)
- Information geometry (S. Amari et. al.)
- Gradient flows on graphs (J. Maas et. al.)
- Wasserstein statistics (W. Li, J. Zhao et. al.)

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.

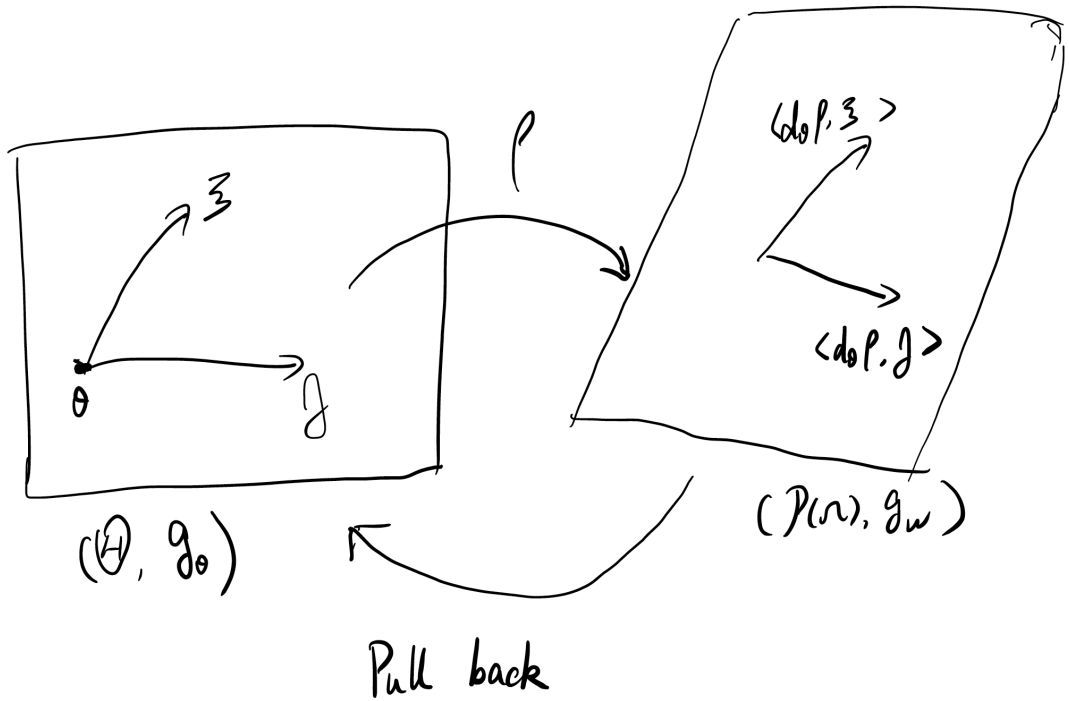


Figure: Pull-back of the metric

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.

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

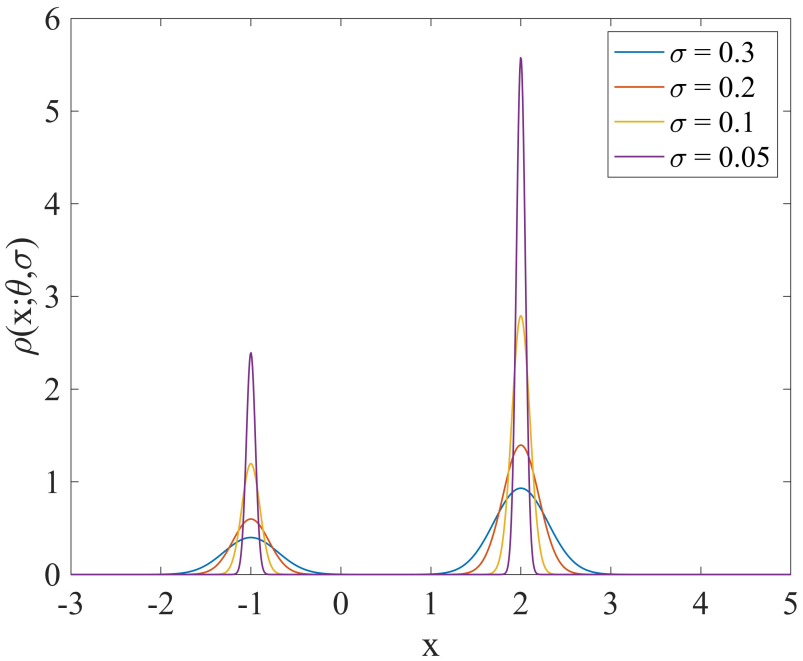


Figure: The figure plots the density function of a family of Gaussian mixture given by $\rho(x) \sim 0.3 * \mathcal{N}(-1, \sigma) + 0.7 * \mathcal{N}(2, \sigma)$.

Fisher-Rao pull-back metric on GMM

We first consider the Fisher-Rao metric as a toy example.

Explicit formula for Fisher-Rao metric in one-dimensional sample space reads

$$g_F\left(\partial_i \rho(x;\theta), \partial_j \rho(x;\theta)\right) = G_F(\theta)_{ij} = \mathbb{E}_{\rho_\theta}\left(\frac{\frac{\partial}{\partial\theta_i}\rho(x;\theta)\frac{\partial}{\partial\theta_j}\rho(x;\theta)}{\rho(x;\theta)^2}\right).$$

In cases of GMM, it is given by

$$G_F(\theta)_{ij} = \mathbb{E}_{p_\theta}\left[\frac{\left(\rho_{i+1}(x;\theta)-\rho_i(x;\theta)\right)\left(\rho_{j+1}(x;\theta)-\rho_j(x;\theta)\right)}{\rho(x;\theta)^2}\right].$$

Scaling approximation Fisher-Rao metric

Theorem (Scaling Fisher-Rao metric)

For a 1-d homogeneous GMM, the above scaling limit of Fisher information matrices is given by

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

This coincides with original definition of Fisher-Rao metric on graph!

Scaling approximation Wasserstein metric

Sketch of proof:

$$\lim_{\sigma\to 0}\int \frac{\rho_i(x)\,\rho_j(x)}{\rho_\theta(x)}dx$$
$$= \lim_{\sigma\to 0}\mathbb{E}_{x\sim\rho_i}\frac{\rho_j(x)}{\rho_\theta(x)}$$
$$= \mathbb{E}_{x\sim\delta_{\mu_i}}\lim_{\sigma\to 0}\frac{\rho_j(x)}{\rho_\theta(x)}$$
$$= \lim_{\sigma\to 0}\frac{\rho_j(\mu_i)}{\rho_\theta(\mu_i)} = \frac{\delta_{ij}}{p_i}.$$

Wasserstein pull-back metric on GMM

Explicit formula for Wasserstein metric in one-dimensional sample space reads

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

where $F(x;\theta)$ is the cumulative distribution function of the density function $\rho(x;\theta)$. In case of GMM, it is given by

$$G_W(\theta)_{ij} = \mathbb{E}_{p_\theta}\left[\frac{(F_{i+1}(x;\theta)-F_i(x;\theta))\,(F_{j+1}(x;\theta)-F_j(x;\theta))}{\rho(x;\theta)^2}\right],$$

where $F_i(x;\theta)$ is the cumulative distribution function of the density function $\rho_i(x;\theta)$.

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

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

Scaling approximation Wasserstein metric

Sketch of proof:

$$
\begin{aligned}
\left(G_W(\theta)\right)_{i-1,i-1} &= \int \frac{\left(F_i - F_{i-1}\right)^2}{\rho_\theta} dx \\
&\overset{\Delta_1}{\Longrightarrow} \int_{\mu_{i-1}}^{\mu_i} \frac{1}{p_{i-1}\rho_{i-1} + p_i\rho_i} dx \\
&\overset{\Delta_2}{\Longrightarrow} \text{Laplacian asymptotics.}
\end{aligned}
$$

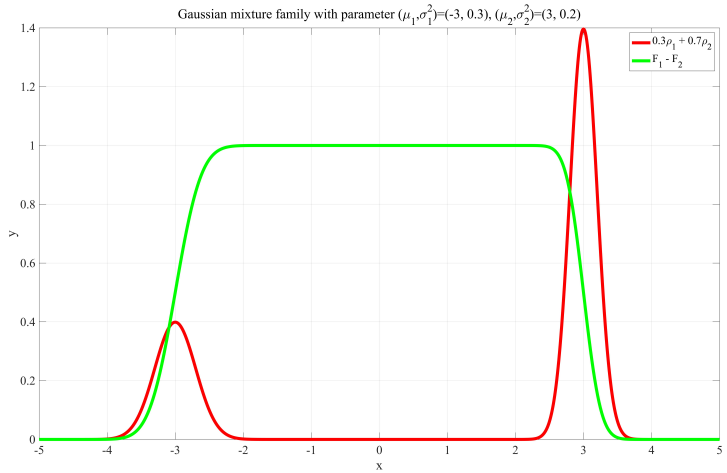


Figure: This figure plots an example of the function $\partial_{\theta_i} F_\theta(x)$ for a GMM.

Scaling approximation Wasserstein metric

This scaling approximate metric is natural in the following senses:

1. In Fisher-Rao case, this scaling approximate metric coincides with original definition.
2. It can be derived as the scaling approximate of many general mixture models (Laplacian mixture models). Intuitively, it is the scaling approximation of most mixture models whose components decay exponentially or in other word, obey some concentration inequalities.
3. It is natural in the sense of gradient flow.

Inhomogeneous GMM

What if the gap between adjacent components are not the same? $\mu_{i+1} - \mu_i \neq \mu_{j+1} - \mu_j$

Homogeneous lattice $\Longleftrightarrow$ GMM with same variances for components.

Inhomogeneous lattice $\Longleftrightarrow$ GMM with different variances for components.

Theorem (informal)

For a 1-d inhomogeneous GMM with difference between adjacent components given by $\mu_{i+1} - \mu_i = d_i$, choose the variances of components as $\sigma_i = s_i\sigma$ such that the condition below is satisfied

$$\frac{d_1}{s_1 + s_2} = \frac{d_2}{s_2 + s_3} = \cdots = \frac{d_{N-1}}{s_{N-1} + s_N} = d. \tag{1}$$

A scaling limit of Wasserstein information matrices is again diagonal.

Structure of variances σ_i characterized geometric structures of the

Extended GMM

Now, we discuss an extended GMM, which is able to be tackled in our framework. Recall GMMs as:

$$\rho_\theta = \sum_{i=1}^{N-1} \theta_i \left(\rho_{i+1} - \rho_i\right) + \rho_1, \qquad \theta_{i-1} \geq \theta_i \geq 0, \quad \forall i = 2, ..., N,$$

$$\rho_i \sim \mathcal{N}\left(\mu_i, \sigma\right), \quad \forall i = 1, ..., N.$$

In ordinary GMMs, the μ_is are constants while only θ_is are parameters. In extended GMMs, both these two sets are parameters, namely, the means of each components can vary.

Extended GMMs and Wasserstein information geometry

Theorem (Wasserstein information geometry of extended GMMs)

For the extended GMM, we have the following relationship between the Wasserstein information matrix(WIM) of the set of tangent vectors $\frac{\partial}{\partial \mu_i} s$ *and Fisher information matrix(FIM) of the set of tangent vectors* $\frac{\partial}{\partial \theta_i} s$

$$G_F = \Sigma G_W \Sigma^T,$$

where $(G_F)_{ij} = g_F\left(\frac{\partial}{\partial \theta_i}, \frac{\partial}{\partial \theta_j}\right)$, $(G_W)_{ij} = g_W\left(\frac{\partial}{\partial \mu_i}, \frac{\partial}{\partial \mu_j}\right)$ *and the matrix* $\Sigma \in \mathbb{R}^{N-1,N}$ *appears above is given by*

$$\Sigma = \begin{pmatrix} -\frac{1}{p_1} & \frac{1}{p_2} & 0 & \cdots & 0 & 0 \\ 0 & -\frac{1}{p_2} & \frac{1}{p_2} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & -\frac{1}{p_{N-2}} & \frac{1}{p_{N-1}} & 0 \\ 0 & 0 & 0 & \cdots & -\frac{1}{p_{N-1}} & \frac{1}{p_N} \end{pmatrix}.$$

Extended GMMs and Wasserstein information geometry

This result on the connection between Wasserstein information matrix and Fisher information matrix could be understood simply using the language of score functions.

Since WIM and FIM play essential roles in statistical theory (Li et. al. 2019), it will be interesting to find statistical illustrations and corollaries of this result.

Scaling approximation Wasserstein metric

Theorem (Scaling Wasserstein metric for extended GMMs)

The Wasserstein metric in 1-d extended homogeneous GMM is given by following in block form

$$\lim_{\sigma\to 0} G_W^{(ext)}(\theta,\mu;\sigma) = \begin{pmatrix} \left(G_{\widetilde{W}}^{(ext)}\right)_{\theta\theta} & \left(G_{\widetilde{W}}^{(ext)}\right)_{\theta\mu} \\ \left(G_{\widetilde{W}}^{(ext)}\right)_{\mu\theta} & \left(G_{\widetilde{W}}^{(ext)}\right)_{\mu\mu} \end{pmatrix},$$

$$\left(G_{\widetilde{W}}^{(ext)}\right)_{\theta\theta} = K(\sigma) \begin{pmatrix} \frac{1}{\sqrt{p_1 p_2}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{p_2 p_3}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{p_{N-1} p_N}} \end{pmatrix},$$

$$\left(G_{\widetilde{W}}^{(ext)}\right)_{\mu\mu} = \begin{pmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_N \end{pmatrix}.$$

Scaling approximation Wasserstein metric

Theorem (Scaling Wasserstein metric for extended GMMs)

$$\left(\left(G^{(ext)}_{\widetilde{W}}\right)_{\mu\theta}\right)^T$$

$$= \left(G^{(ext)}_{\widetilde{W}}\right)_{\theta\mu} = \begin{pmatrix} \frac{\mu_2-\mu_1}{2} & \frac{\mu_2-\mu_1}{2} & 0 & \cdots & 0 & 0 \\ 0 & \frac{\mu_3-\mu_2}{2} & \frac{\mu_3-\mu_2}{2} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{\mu_{N-1}-\mu_{N-2}}{2} & 0 \\ 0 & 0 & 0 & \cdots & \frac{\mu_N-\mu_{N-1}}{2} & \frac{\mu_N-\mu_{N-1}}{2} \end{pmatrix}.$$

Due to the existence of factor $K(\sigma)$, the scaling metric tensor has a multi-scale phenomenon, i.e. its block form has different orders in different blocks. This properties will the reason for multi-scale phenomenons in gradient flows.

Parametric gradient flows

Given a function $\mathcal{E}(\theta)$ on a parametric space Θ with metric given by $G(\theta)$, we define the gradient flow associated with this function as

$$\dot{\theta} = -G(\theta)^{-1} \nabla_\theta \mathcal{E}(\theta),$$

where $\nabla_\theta \mathcal{E}(\theta)$ is ordinary Euclidean gradient. Since in GMM, parameters θ_is do not carry any physical meaning, we write the gradient flows equations for parameters p_is

$$\dot{p}_i = \dot{\theta}_{i-1} - \dot{\theta}_i = \Big(G(\theta)^{-1} \nabla_\theta \mathcal{E}(\theta) \Big)_i - \Big(G(\theta)^{-1} \nabla_\theta \mathcal{E}(\theta) \Big)_{i-1}.$$

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

We first show the continuous gradient flow on density manifold:

$$\begin{aligned}
\textbf{internal energy:} \quad & \partial_t \rho = -\nabla \cdot \left(\rho \nabla U'(\rho)\right), \\
\textbf{potential energy:} \quad & \partial_t \rho = -\nabla \cdot (\rho \nabla V), \\
\textbf{interaction energy:} \quad & \partial_t \rho = -\nabla \cdot (\rho \nabla (W * \rho)).
\end{aligned}$$

These are well-known facts in the community of optimal transport.

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.

Properties of gradient flows

We establish the following properties for this parametric gradient flows:

1. positivity
2. conservation of mass
3. long time existence
4. consistency

Gradient flows in extended GMM

Based on the scaling metric for extended GMM, we derive following gradient flow equation.

Theorem (Gradient flow in extended GMM)

The gradient flow w.r.t. a potential energy functional (interaction energy functionals) $\mathcal{E}(\rho) = \sum_{i=1}^{N} p_i V(\mu_i)$
$(\mathcal{E}(\rho) = \sum_{1\le i<j\le N} p_i p_j W(|\mu_i - \mu_j|))$ *is provided below*

$$\dot{\theta}_i = -\frac{\sqrt{p_i p_{i+1}}}{K(\sigma)}\left(\partial_{\theta_i}\mathcal{E}(\rho) - \frac{\mu_{i+1}-\mu_i}{2}\left(\frac{\partial_{\mu_i}\mathcal{E}(\rho)}{p_i} + \frac{\partial_{\mu_{i+1}}\mathcal{E}(\rho)}{p_{i+1}}\right)\right), \quad i = 1, 2, \cdots, N-1,$$

$$\dot{\mu}_i = -\frac{1}{p_i}\partial_{\mu_i}\mathcal{E}(\rho) + \frac{1}{K(\sigma)}\left(\sqrt{\frac{p_{i+1}}{p_i}}\frac{\mu_{i+1}-\mu_i}{2}\partial_{\theta_i}\mathcal{E}(\rho) + \sqrt{\frac{p_{i-1}}{p_i}}\frac{\mu_i-\mu_{i-1}}{2}\partial_{\theta_{i-1}}\mathcal{E}(\rho)\right),$$
$$i = 2, 3, \cdots, N-1,$$

where we ignore the equation for end-point parameters μ_1, μ_N *since they are not consistent with this framework.*

Analysis of gradient flows in extended GMMs

There are following interesting aspects in above gradient flow equations:

1. The flows of means parameters μ_is and ratio parameters p_is have different orders. Hence they converge to equilibrium in different time scales.

2. The flow equations of parameters μ_i have two terms of different orders.

3. Comparing the flow equations of parameters p_is with previous flows equations, we find an extra term.

Numerical experiments

$$\dot{p}_i = -\frac{1}{d^2}\left(\sqrt{p_{i-1}p_i}\log\frac{p_i}{p_{i-1}} - \sqrt{p_{i+1}p_i}\log\frac{p_{i+1}}{p_i}\right),$$

where p_i is the mass on i-th lattice and d is the gap, Δt is the stepsize.

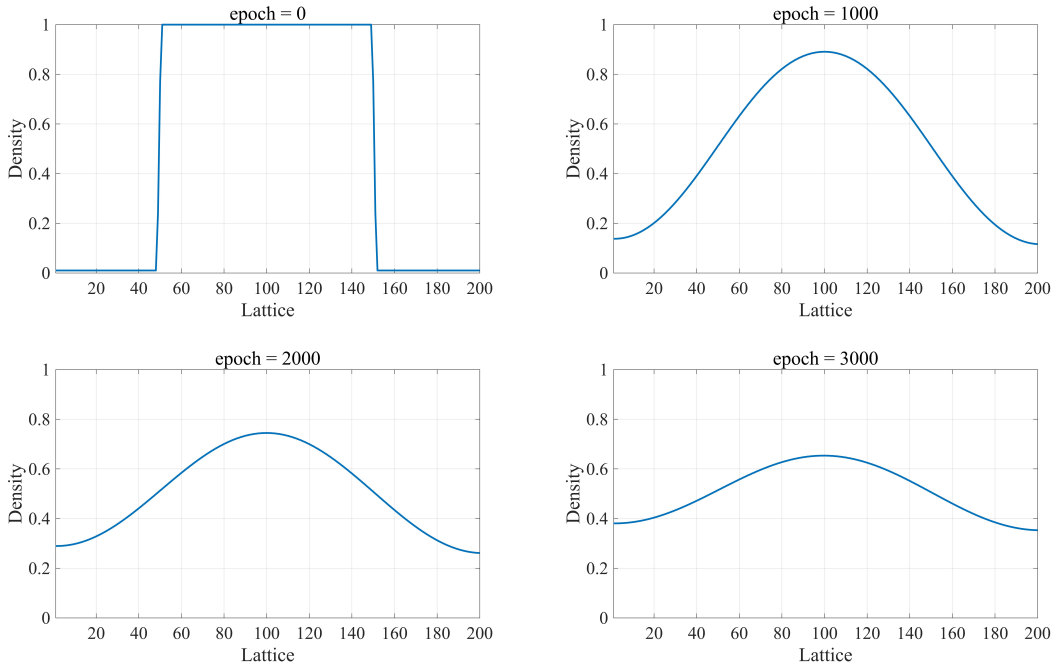


Figure: This figure plots a simulation of the 1-d heat flow via the discretization introduced in this paper. The gap of the lattice is set as $d = 0.01$. And the time step is set as $\Delta t = 0.000005$. The initial distribution is given by $\rho(x; 0) = \mathbf{1}_{[-0.5,0.5]}(x)$ and we consider its restriction to the interval $[-1, 1]$.

Numerical experiments

$$\dot{p}_{ij} = -\frac{1}{d^2}\left(\sqrt{p_{i-1,j}p_{ij}}\log\frac{p_{ij}}{p_{i-1,j}} - \sqrt{p_{i+1,j}p_{ij}}\log\frac{p_{i+1,j}}{p_{ij}} + \sqrt{p_{i,j-1}p_{ij}}\log\frac{p_{ij}}{p_{i,j-1}} - \sqrt{p_{i,j+1}p_{ij}}\log\frac{p_{i,j+1}}{p_{ij}}\right).$$

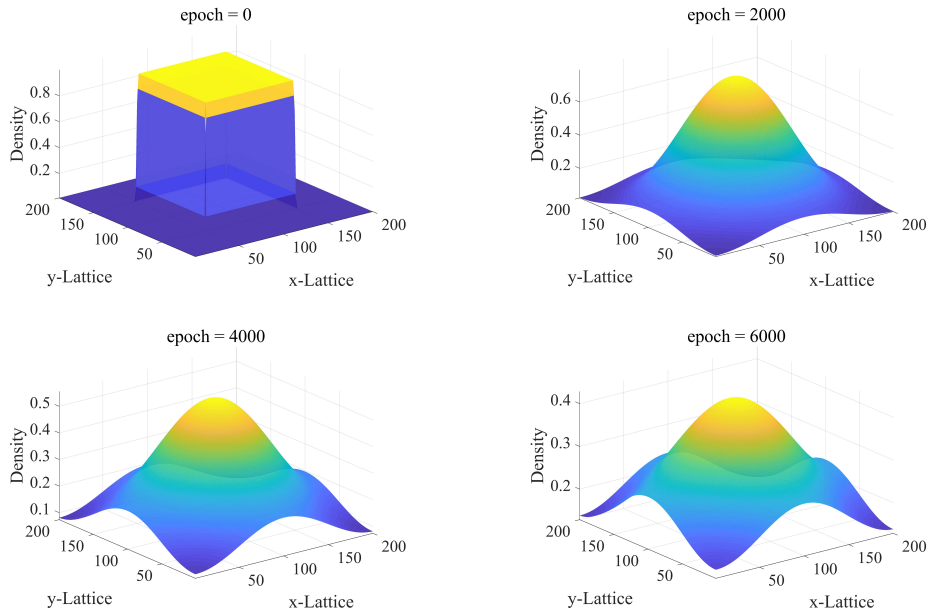


Figure: Simulation of a 2-d heat flow via discretization in this paper. The gap of the lattice and the time step are set as $d = 0.01, \Delta t = 2.5 \times 10^{-6}$ respectively. The initial distribution is given by $\rho(\mathbf{x}; 0) = \mathbf{1}_{[-0.5,0.5]\times[-0.5,0.5]}(\mathbf{x})$ and we consider its restriction to the interval $[-1, 1] \times [-1, 1]$.

Numerical experiments

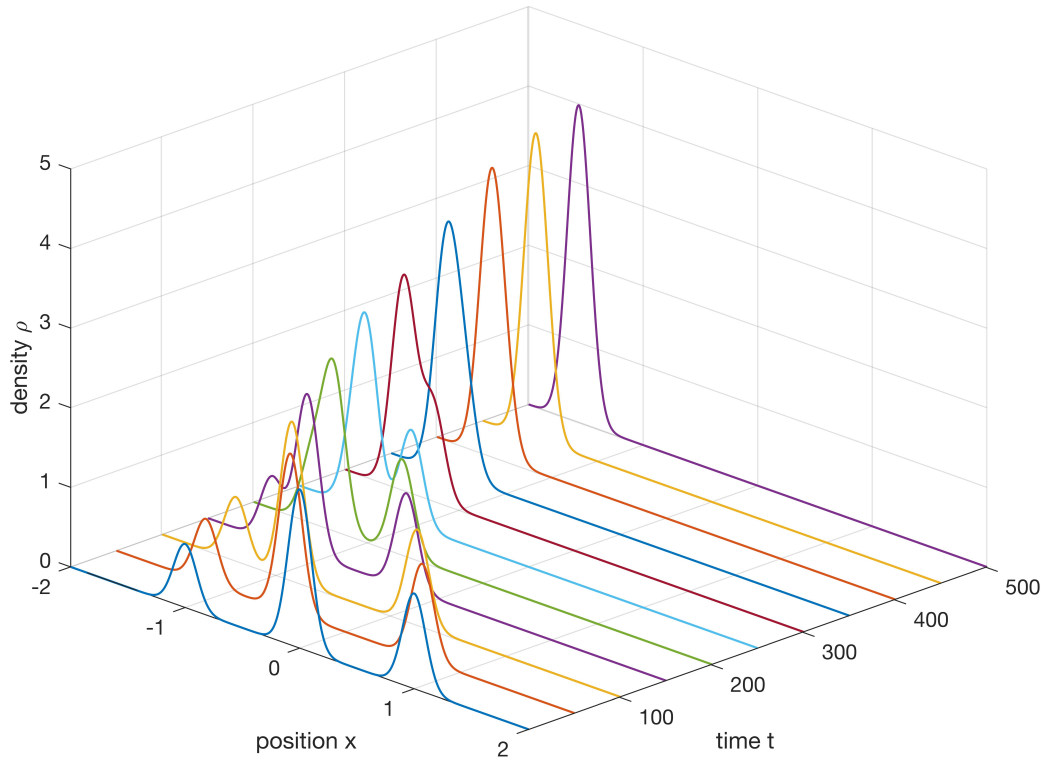


Figure: This figure plots a simulation of the gradient flow associated with a potential energy functional via the discretization in an extended GMM introduced in this paper. And the time step is set as $\Delta t = 0.01$. The initial distribution is given by $\rho(x;0) \sim 0.2 * \mathcal{N}(-1, 0.1) + 0.5 * \mathcal{N}(0, 0.1) + 0.3 * \mathcal{N}(1, 0.1)$. Direct calculation shows that the scaling factor $K(\sigma) \approx 2000$. The potential function is periodic as $V(x) = \sin x$. The simulation illustrates the degeneracy of extended GMMs.

Numerical experiments

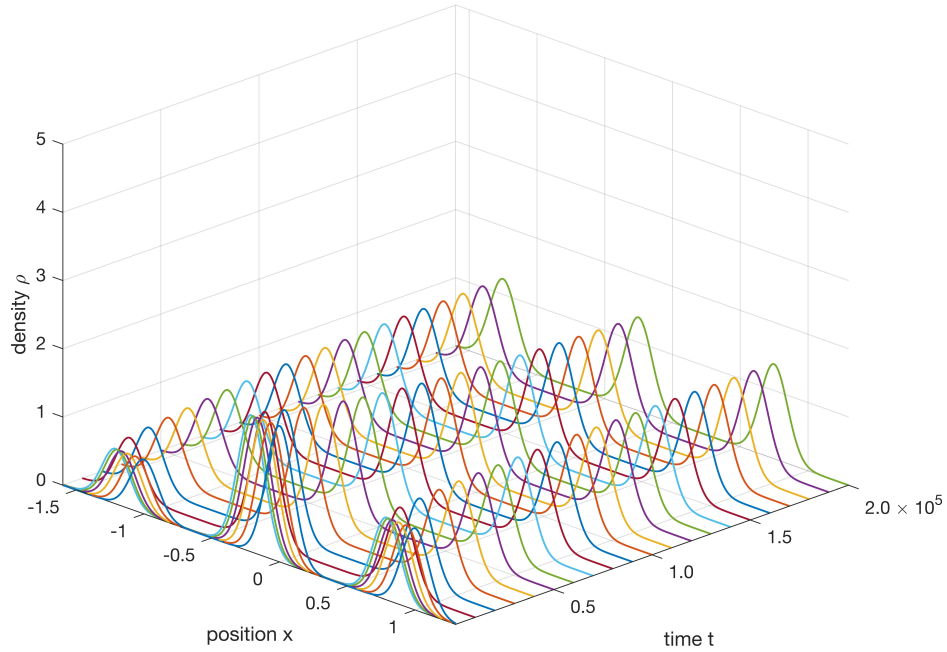


Figure: The functional in this gradient flow is the sum of the entropy and a potential energy, i.e. $\mathcal{E} = \sum_{i=1}^{N} p_i \left(\log p_i + V\left(\mu_i\right)\right)$, where the potential is periodic as $V(x) = \sin 2\pi x$. The continuous flow is drift diffusion, i.e. $\partial_t \rho = \nabla \cdot (\rho \nabla V) + \Delta \rho$. And the time step is set as $\Delta t = 0.01$. The initial distribution is given by $\rho(x; 0) \sim 0.2 * \mathcal{N}(-1, 0.1) + 0.5 * \mathcal{N}(0, 0.1) + 0.3 * \mathcal{N}(1, 0.1)$. Consistent to our analysis, there exist two time scales in this flow: in the short scale, parameters μ_is tend to equilibrium, i.e. minimizor of $V(x)$, while in the large scale parameters p_is tend to equilibrium.

Future works

1. Consider applications in scientific computing, such as mean-field games and machine learning.
2. Establish a correspondent theory for high-dimensional sample spaces.
3. A complete study of the scaling Wasserstein geometry on graphs, i.e. functional inequalities, possible connection with Hamiltonian flows.

Take-home message

1. Scaling limits of Wasserstein GMMs exist and can be viewed as a geometric structure on graph.

2. Geometric structures on graphs further provide ways to discretize PDE preserving certain structure.

WIM and Gaussian mixture model (GMM)

Gaussian mixture models (GMM) are widely used in statistical inference (statistical models) and scientific computation. One parameterization of 1d GMM:

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

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

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

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

We consider the pull-back metric $G_W(\theta)$ on GMM which has the following integration formula in 1d:

$$G_W(\theta)_{ij} = \mathbb{E}_{p_\theta}\left[\frac{(F_{i+1}(x;\theta) - F_i(x;\theta))(F_{j+1}(x;\theta) - F_j(x;\theta))}{\rho(x;\theta)^2}\right], \ F \text{ c.d.f.}$$

Scaling FIM & WIM

Shrinking the components' variances to 0 weakly converges to Dirac mixture model.

Scaling FIM coincides with FIM on graph.

Scaling WIM is diagonal and easy to be inverted.

Theorem (Scaling FIM & WIM)

$$\lim_{\sigma\to 0} G_F(\theta;\sigma) = \begin{pmatrix} \frac{1}{p_1}+\frac{1}{p_2} & -\frac{1}{p_2} & 0 & \cdots & 0 & 0 \\ -\frac{1}{p_2} & \frac{1}{p_2}+\frac{1}{p_3} & -\frac{1}{p_3} & \cdots & 0 & 0 \\ 0 & -\frac{1}{p_3} & \frac{1}{p_3}+\frac{1}{p_4} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\frac{1}{p_{N-1}} & \frac{1}{p_{N-1}}+\frac{1}{p_N} \end{pmatrix}$$

$$\lim_{\sigma\to 0} \frac{G_W(\theta;\sigma)}{\sqrt{2\pi^3}\frac{\sigma^3}{d}e^{\frac{1}{2}\left(\frac{d}{2\sigma}\right)^2}} = \begin{pmatrix} \frac{1}{\sqrt{p_1p_2}} & 0 & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{\sqrt{p_2p_3}} & 0 & \cdots & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{p_3p_4}} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \frac{1}{\sqrt{p_{N-1}p_N}} \end{pmatrix}.$$

Numerical experiments

Restriction of PDE on GMM $\Longleftrightarrow$ spatial discretization of the differential operator. Take heat equation $\partial_t \rho = \Delta \rho$ as an example, the scaling WIM provide the following spatial discretization of 2d Laplacian

$$\dot{p}_{ij} = -\frac{1}{d^2}\left(\sqrt{p_{i-1,j}p_{ij}}\log\frac{p_{ij}}{p_{i-1,j}} - \sqrt{p_{i+1,j}p_{ij}}\log\frac{p_{i+1,j}}{p_{ij}} + \sqrt{p_{i,j-1}p_{ij}}\log\frac{p_{ij}}{p_{i,j-1}} - \sqrt{p_{i,j+1}p_{ij}}\log\frac{p_{i,j+1}}{p_{ij}}\right).$$

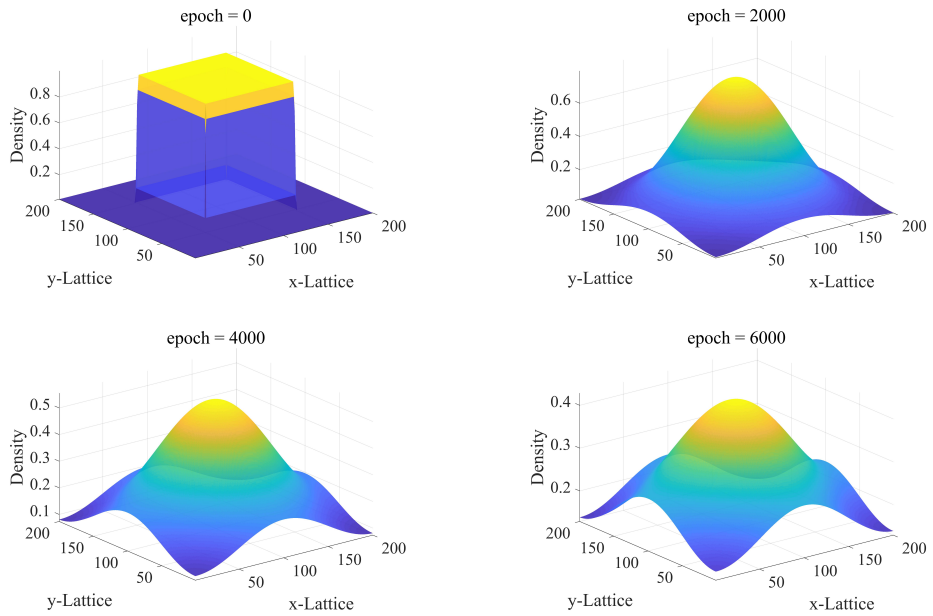


Figure: Simulation of a 2-d heat flow via discretization in this paper. The gap of the lattice and the time step are set as $d = 0.01, \Delta t = 2.5 \times 10^{-6}$ respectively. The initial

Scaling WIM on extended GMM

Allowing mean parameters of GMM μ_is to move, one obtains WIM of the form $G^{(ext)}_{\widetilde{W}}(\theta, \mu; \sigma)$

Theorem (Scaling WIM for extended GMMs)

$$\lim_{\sigma\to 0} G^{(ext)}_{\widetilde{W}}(\theta, \mu; \sigma) = \begin{pmatrix} \left(G^{(ext)}_{\widetilde{W}}\right)_{\theta\theta} & \left(G^{(ext)}_{\widetilde{W}}\right)_{\theta\mu} \\ \left(G^{(ext)}_{\widetilde{W}}\right)_{\mu\theta} & \left(G^{(ext)}_{\widetilde{W}}\right)_{\mu\mu} \end{pmatrix},$$

$\left(G^{(ext)}_{\widetilde{W}}\right)_{\theta\theta}, \left(G^{(ext)}_{\widetilde{W}}\right)_{\mu\mu}$ *both diagonal.*

The gradient flow w.r.t. a energy functional $\mathcal{E}(\rho)$ is

$$\dot{\theta}_i = -\frac{\sqrt{p_i p_{i+1}}}{K(\sigma)}\left(\partial_{\theta_i}\mathcal{E}(\rho) - \frac{\mu_{i+1}-\mu_i}{2}\left(\frac{\partial_{\mu_i}\mathcal{E}(\rho)}{p_i} + \frac{\partial_{\mu_{i+1}}\mathcal{E}(\rho)}{p_{i+1}}\right)\right),$$

$$\dot{\mu}_i = -\frac{1}{p_i}\partial_{\mu_i}\mathcal{E}(\rho) + \frac{\left(\sqrt{\frac{p_{i+1}}{p_i}}\frac{\mu_{i+1}-\mu_i}{2}\partial_{\theta_i}\mathcal{E}(\rho) + \sqrt{\frac{p_{i-1}}{p_i}}\frac{\mu_i-\mu_{i-1}}{2}\partial_{\theta_{i-1}}\mathcal{E}(\rho)\right)}{K(\sigma)},$$

Numerical experiments

Solving PDE on extended GMM $\Longleftrightarrow$ moving mesh method. We consider solving $\partial_t \rho + \nabla(\rho \nabla \sin x) = 0$. One of the minimum is $-\frac{\pi}{2}$, where three components eventually settle down.

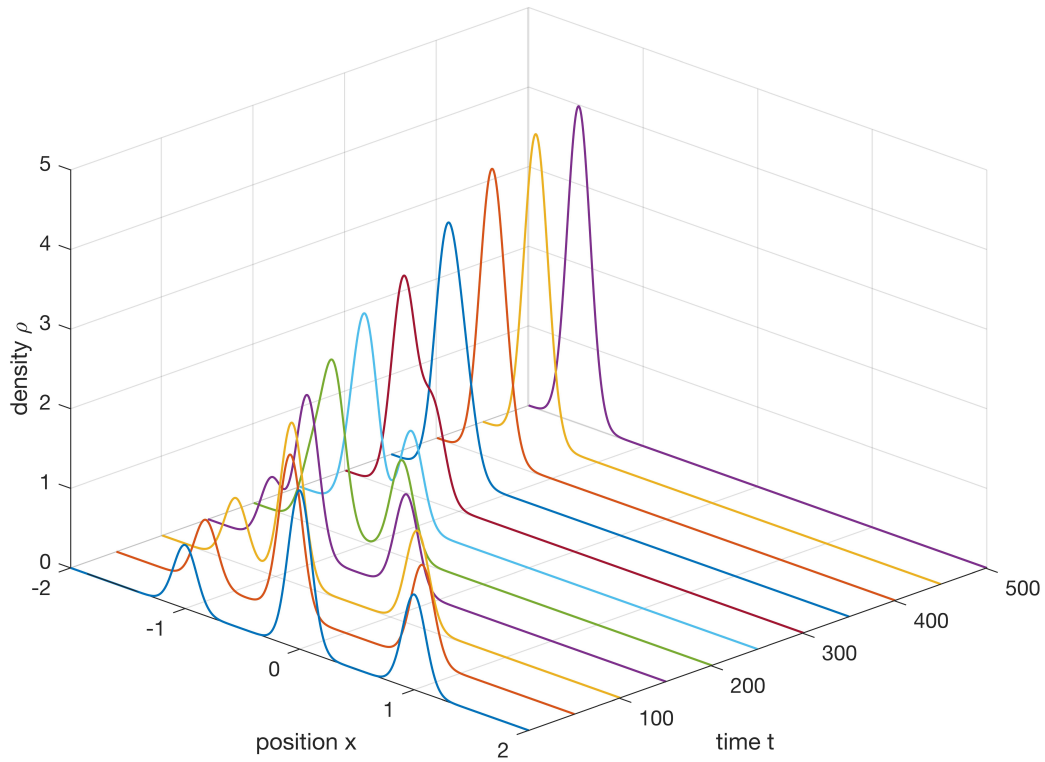


Figure: This figure plots a simulation of the gradient flow associated with a potential energy functional via the discretization in an extended GMM introduced in