Natural gradient in Wasserstein statistical manifold[1]

undergraduate thesis

Yifan Chen

Department of Mathematical Sciences
Tsinghua University

June 4, 2018

[1] arxiv:1805.08380

Introduction

The statistical distance finds wide applications in machine learning

$$\text{minimize} \quad d(\rho, \rho_e) \quad \text{s.t.} \quad \rho \in \mathcal{P}_\theta.$$

where

- $\mathcal{P}_\theta$: a parameterized subset of probability density space;
- ρ_e : a given target density (often empirical distribution);
- d : quantifies the difference between ρ and ρ_e

Example

- $\mathcal{P}_\theta$: exponential family

$$\rho(x,\theta) = \frac{1}{Z(\theta)} \exp(\sum f_i(x)\theta_i + r(x)), \quad x \in \Omega \subset \mathbb{R}^n$$

other: mixture model, neural network, etc.

- d : KL divergence

$$\min_{\rho(\cdot,\theta)\in\mathcal{P}_\theta} \mathrm{KL}(\rho_e||\rho(\cdot,\theta)) = \int_\Omega \rho_e(x) \log \frac{\rho_e(x)}{\rho(x,\theta)} dx.$$

where, $\rho_e(x) = \frac{1}{N}\sum_{i=1}^N \delta_{X_i}$ empirical distribution.
rewrite:

$$\min_\theta \frac{1}{N} \sum_{i=1}^N -\log \rho(X_i,\theta) = L(\theta),$$

i.e. maximum likelihood method in statistics

Natural gradient

- natural gradient descent

$$\theta^{n+1} = \theta^n - \tau G_F(\theta^n)^{-1} \nabla_\theta \mathrm{KL}(\rho_e || \rho(\cdot, \theta^n)),$$

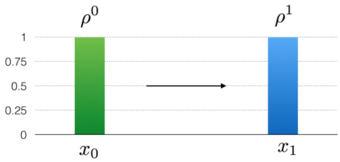
where

$$\begin{aligned} G_F(\theta) &= \int_{\mathbb{R}} \rho(x,\theta)(\nabla_\theta \log \rho(x,\theta))^T \nabla_\theta \log \rho(x,\theta) dx \\ &= \nabla^2_{\theta'} \mathrm{KL}(\rho(\cdot,\theta) || \rho(\cdot,\theta'))|_{\theta'=\theta}. \end{aligned}$$

- advantage:
  - asymptotically Newton preconditioning
  - parameterization invariant
  - Fisher efficient based on Cramer-Rao bound
- (Θ, G_F) (Fisher-Rao) statistical manifold

Ground cost related distance

Optimal transport provides the Wasserstein distance among histograms, relying on the structure of **sample space** (ground cost c)



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

$$W(\rho^0, \rho^1) = c(x_0, x_1)$$

vs.

$$\mathrm{TV}(\rho_0, \rho_1) = \int_\Omega |\rho^0(x) - \rho^1(x)| dx = 2$$

vs.

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

Optimal transport

- Monge's problem

$$\inf_T \int_\Omega \|x - T(x)\|^2 \rho^0(x)dx \quad \text{s.t.} \int_A \rho^1(x)dx = \int_{T^{-1}(A)} \rho^0(x)dx, \forall \text{ Borel } A$$

- Kantorovich's problem

$$\min_{\pi \in \Pi(\rho^0, \rho^1)} \int_{\Omega \times \Omega} \|x - y\|^2 \pi(x, y) dxdy,$$

Π: joint probability measures on $\Omega \times \Omega$ with marginals ρ^0, ρ^1.

- Dynamical formulation, known as Benamou-Brenier formula

$$\inf_{\Phi_t} \int_0^1 \int_\Omega \|\nabla \Phi(t, x)\|^2 \rho(t, x) dxdt$$

$$\text{s.t. } \frac{\partial \rho(t, x)}{\partial t} + \nabla \cdot (\rho(t, x) \nabla \Phi(t, x)) = 0, \rho(0, x) = \rho^0(x), \rho(1, x) = \rho^1(x)$$

Benamou-Brenier formula gives a Riemannian differential structure on density space:

- L^2-Wasserstein metric tensor on the tangent space of densities $g_\rho\colon\ T_\rho\mathcal{P}_2(\Omega)\times T_\rho\mathcal{P}_2(\Omega)\to\mathbb{R}$:

$$g_\rho(\sigma_1,\sigma_2)=\int_\Omega \nabla\Phi_1(x)\cdot\nabla\Phi_2(x)\rho(x)dx,$$

  where $\sigma_1=V_{\Phi_1}:=-\nabla\cdot(\rho(x)\nabla\Phi_1(x))$, $\sigma_2=V_{\Phi_2}$ with $\Phi_1(x)$, $\Phi_2(x)\in C^\infty(\Omega)/\mathbb{R}$.

- Wasserstein metric as geodesic distance

$$\begin{aligned}&(W_2(\rho^0,\rho^1))^2\\ =&\inf_{\Phi_t}\ \Big\{\int_0^1 g_{\rho_t}(V_{\Phi_t},V_{\Phi_t})dt\colon \partial_t\rho_t=V_{\Phi_t},\ \rho(0,x)=\rho^0,\ \rho(1,x)=\rho^1\Big\}.\end{aligned}$$

Remark

Many works use this Riemannian structure and establish connections to optimization
e.g. $V(x)$ σ-strongly convex, $\rho^*(x) \sim \exp(-V(x))$

Table: Comparison

	Optimization in density space	in Euclidean space
loss	$\mathrm{KL}(\rho, \rho^*)$	$f(x) - f(x^*)$
distance	$W_2^2(\rho, \rho^*)$	$\|x - x^*\|^2$
gradient	$\int \|\nabla \log \frac{\rho(x)}{\rho^*(x)}\|^2 \rho(x) dx$	$\|\nabla f(x)\|^2$
iteration	$dX_t = -\nabla V(X_t) + \sqrt{2} B_t$	$dX_t = -\nabla V(X_t)$
density	$\partial_t \rho_t = \nabla \cdot (\rho \nabla \log \frac{\rho(x)}{\rho^*(x)})$	$\partial_t \rho_t = \nabla \cdot (\rho \nabla V)$

Remark

a line of work proves sampling complexity of discrete Langevin dynamics with strongly convex assumptions

- Dalalyan(2017) proved complexity $O(\frac{d}{\epsilon^2})$ in total variation distance, Moulines(2016) in Wasserstein metric, and Bartlett(2017) in KL divergence.
- Bartlett and Jordan(2018) proved a better convergence rate $O(\frac{\sqrt{d}}{\epsilon})$ using underdamped Langevin (COLT 2018)
- Bernton(2018) and Wibisono(2018) both consider sampling as optimization in density space, using the Wasserstein geometry (COLT 2018)

a mean-field view of the landscape of two-layers neural networks

- Mei(2018) considers a density over parameters in two-layer networks, as a mean-field limit

Overview of our work

- **goal:** Wasserstein geometry in parametrized densities $\mathcal{P}_\theta$
- **benefits:** boost computation, e.g. natural gradient descent
- **approach:** "pull-back" of metric tensor

Wasserstein statistical manifold

- Statistical model: triple (Ω, Θ, ρ)
- L^2-Wasserstein metric tensor in parameter space: the inner product g_θ on $T_\theta(\Theta)$ is defined as

$$g_\theta(\xi, \eta) = \int_\Omega \rho(x, \theta) \nabla\Phi_\xi(x) \cdot \nabla\Phi_\eta(x) dx,$$

where ξ, η are tangent vectors in $T_\theta(\Theta)$, Φ_ξ and Φ_η satisfy

$$\langle \nabla_\theta \rho(x, \theta), \xi \rangle = -\nabla \cdot (\rho \nabla\Phi_\xi(x))$$

and

$$\langle \nabla_\theta \rho(x, \theta), \eta \rangle = -\nabla \cdot (\rho \nabla\Phi_\eta(x))$$

.

Gradient flow

- metric tensor

$$\begin{aligned} g_\theta(\xi,\eta) &= \int_\Omega \rho(x,\theta)\nabla\Phi_\xi(x)\cdot\nabla\Phi_\eta(x)dx \\ &= \int_\Omega \langle\nabla_\theta\rho(x,\theta),\xi\rangle\cdot\Phi_\eta(x)dx \\ &= \int_\Omega \langle\nabla_\theta\rho(x,\theta),\xi\rangle(-\Delta_\theta)^{-1}\langle\nabla_\theta\rho(x,\theta),\eta\rangle dx \\ &= \xi^T G_W(\theta)\eta \end{aligned}$$

  where: $-\Delta_\theta = -\nabla\cdot(\rho(x,\theta)\nabla)$,
  $(G_W)_{ij}(\theta) = \int_\Omega \partial_{\theta_i}\rho(x,\theta)(-\Delta_\theta)^{-1}\partial_{\theta_j}\rho(x,\theta)dx$

- gradient flow of function $R\in C^1(\Theta)$ in (Θ, g_θ)

$$\frac{d\theta}{dt} = -G_W(\theta)^{-1}\nabla_\theta R(\theta).$$

Continuous sample space in 1D

For 1D densities, the Wasserstein metric tensor $g_\theta(\xi,\eta) = \langle \xi, G_W(\theta)\eta\rangle$ such that

$$G_W(\theta) = \int_{\mathbb{R}} \frac{1}{\rho(x,\theta)} (\nabla_\theta F(x,\theta))^T \nabla_\theta F(x,\theta) dx$$

where F: cdf of $\rho(\theta)$

Compare it with Fisher-Rao metric tensor (or Fisher information matrix)

$$\begin{aligned} G_F(\theta) &= \int_{\mathbb{R}} \rho(x,\theta)(\nabla_\theta \log \rho(x,\theta))^T \nabla_\theta \log \rho(x,\theta) dx \\ &= \int_{\mathbb{R}} \frac{1}{\rho(x,\theta)} (\nabla_\theta \rho(x,\theta))^T \nabla_\theta \rho(x,\theta) dx, \end{aligned}$$

Wasserstein natural gradient descent

- given objective $R(\rho(\cdot, \theta))$, the Wasserstein natural gradient descent:

$$\theta^{n+1} = \theta^n - \tau G_W(\theta^n)^{-1} \nabla_\theta R(\rho(\cdot, \theta^n)),$$

- (informal theorem) if $R(\rho(\cdot, \theta)) = \frac{1}{2}\left(W_2(\rho(\cdot, \theta), \rho^*)\right)^2$, then

$$\nabla_\theta^2 R(\rho(\cdot, \theta)) = \int_\Omega (T(x, \theta) - x) \nabla_\theta^2 F(x, \theta) dx + \int_\Omega \frac{T'(x, \theta)}{\rho(x, \theta)} (\nabla_\theta F(x, \theta))^T \nabla_\theta F(x, \theta) dx$$

  where, T optimal transport map

- hence if $\rho^* \in \mathcal{P}_\theta$, then

$$\lim_{\theta \to \theta^*} G_W(\theta) = \nabla_\theta^2 R(\rho(\cdot, \theta^*))$$

- another preconditioner:

$$\bar{G}_W(\theta) := \int \frac{T'(x, \theta)}{\rho(x, \theta)} (\nabla_\theta F(x, \theta))^T \nabla_\theta F(x, \theta) dx$$

Numerical examples

Consider the Gaussian mixture model $a\mathcal{N}(\mu_1, \sigma_1) + (1-a)\mathcal{N}(\mu_2, \sigma_2)$ with density functions:

$$\rho(x, \theta) = \frac{a}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + \frac{1-a}{\sigma_2\sqrt{2\pi}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}},$$

where $\theta = (a, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ and $a \in [0, 1]$.

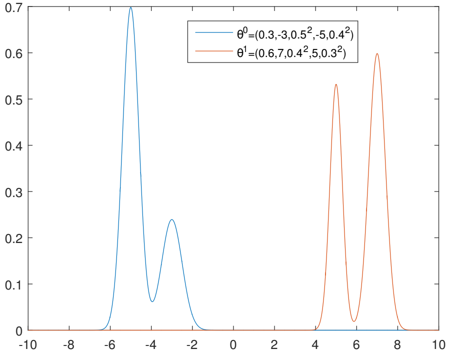


Figure: Densities of Gaussian mixture distribution

Geodesics

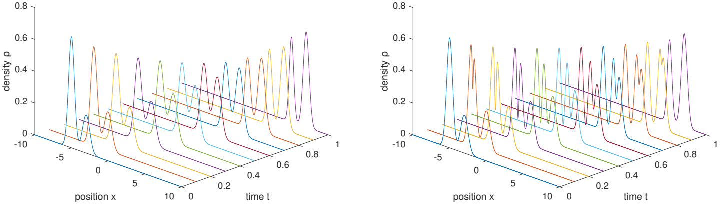


Figure: Geodesic of Gaussian mixtures; left: in the Wasserstein statistical manifold; right: in the whole density space

Natural gradient

Consider the Gaussian mixture fitting problem: given N data points $\{x_i\}_{i=1}^N$ obeying the distribution $\rho(x;\theta^1)$ (unknown), we want to infer θ^1 by using these data points, which leads to a minimization as:

$$\min_\theta W_2^2\left(\rho(\cdot;\theta), \frac{1}{N}\sum_{i=1}^N \delta_{x_i}(\cdot)\right)$$

We perform the following five iterative algorithms to solve the optimization problem:

$$\text{Gradient descent (GD)}: \quad \theta_{n+1} = \theta_n - \tau\nabla_\theta(\frac{1}{2}W^2)|_{\theta_n}$$

$$\text{GD with diag-preconditioning}: \quad \theta_{n+1} = \theta_n - \tau P^{-1}\nabla_\theta(\frac{1}{2}W^2)|_{\theta_n}$$

$$\text{Wasserstein GD}: \quad \theta_{n+1} = \theta_n - \tau G_W(\theta_n)^{-1}\nabla_\theta(\frac{1}{2}W^2)|_{\theta_n}$$

$$\text{Modified Wasserstein GD}: \quad \theta_{n+1} = \theta_n - \tau\left(\bar{G}_W(\theta_n)\right)^{-1}\nabla_\theta(\frac{1}{2}W^2)|_{\theta_n}$$

$$\text{Fisher-Rao GD}: \quad \theta_{n+1} = \theta_n - \tau G_F(\theta_n)^{-1}\nabla_\theta(\frac{1}{2}W^2)|_{\theta_n}$$

Optimization results

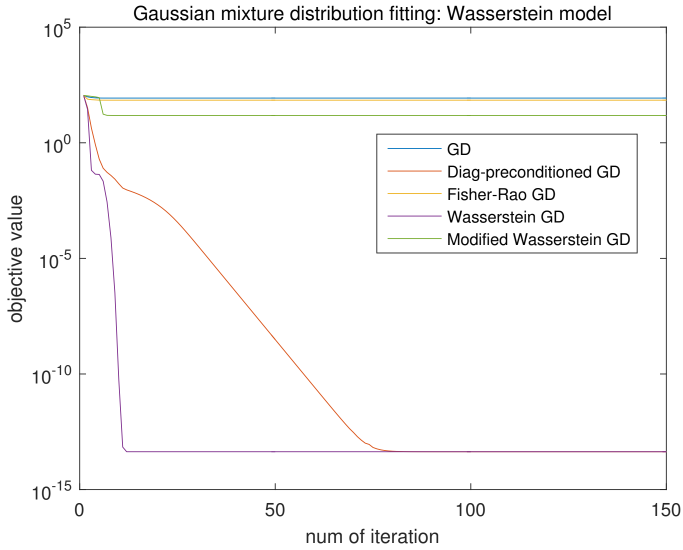


Figure: objective value

Summary

- Wasserstein metric tensor in parametric statistical models
- Wasserstein natural gradient accelerates Wasserstein metric modeled optimization
- interplay between density evolution (Eulerian) and particle moving (Lagrangian)

Thanks for your attention!