From Distributional Ambiguity to Gradient Flows

Wasserstein, Fisher-Rao, and Kernel Approximation

Jia-Jie Zhu

Weierstrass Institute for Applied Analysis and Stochastics
Berlin, Germany

Weierstraß-Institut für
Angewandte Analysis und Stochastik

Leibniz Gemeinschaft

FoDL SPP2298

Berlin Mathematics Research Center MATH+

DFG Deutsche Forschungsgemeinschaft

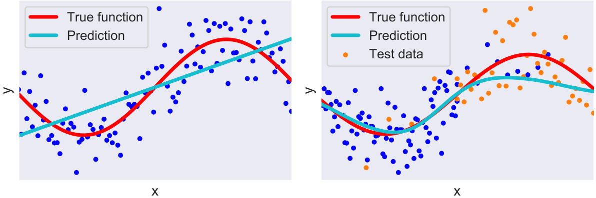
November 28th, 2024
CDM Seminar. École Polytechnique Fédérale de Lausanne, Switzerland

Motivation

Robust learning under distribution shifts [Z. et al. AISTATS 2021, AISTATS 2023, ...]

Empirical risk minimization $\min_\theta \dfrac{1}{N}\sum_{i=1}^{N} \ell(\theta, [x_i, y_i])$

$x_i, y_i \sim P_0$: data sample. θ: learning parameter e.g. DNN weights.



What if the test data **distribution shifts** from P_0?

Distributionally robust optimization (DRO) $\min_\theta \sup_{\mu \in \mathcal{A} \subset \mathcal{P}} \mathbb{E}_\mu \ell(\theta, [X, Y])$

$\mathcal{A} = \left\{ \mu \in \mathcal{P} \middle| \, \mathrm{D}(\mu | \hat{P}_N) \le \epsilon \right\}$

$\hat{P}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$: empirical dist.

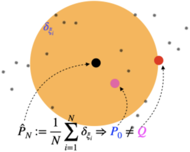
D: divergence between measures

Wasserstein DRO [Esfahani & Kuhn 2018; Sinha et al. 2017] loss ℓ: (p/w) quadratic, logistic, etc.

Kernel DRO [Z. et al. AISTATS 2021, 22...] for general nonlinear loss in ML

Kernel distributionally robust optimization (DRO)

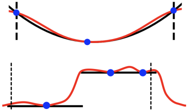
$$\min_{\theta} \sup_{\mathrm{MMD}(\mu,\hat{P})\leq\epsilon} \mathbb{E}_{\mu}\ell(\theta,[X,Y])$$



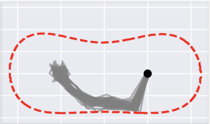
Theorem [Z et al., 2021] DRO problem is equivalent to the dual kernel learning problem

$$\min_{\theta, f\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^{N} f(\xi_i) + \epsilon\|f\|_{\mathcal{H}}$$

$$\text{s.t. } \ell(\theta,\cdot) \leq f, \forall\, x,y \text{ a.e.}$$



Geometric intuition: dual f as robust surrogate; **flatten the curve**



Stochastic control
Nemmour et al **Z**. IEEE CDC'22

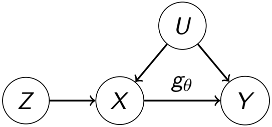
Adversarial robustness
Z et al. AISTATS'22

Q: is $\mathbb{E}_{\mu}\ell(\theta,[X,Y])$ linear? Convex? Along the geodesics? We need math foundation.

Causal inference as measure optimization [Kremer, Z. et al. ICML 2022, ICML 2023]

Conditional moment restriction (CMR)

find θ s.t. $\mathbb{E}[Y - g_\theta(X) | Z = z] = 0$ for z a.e.

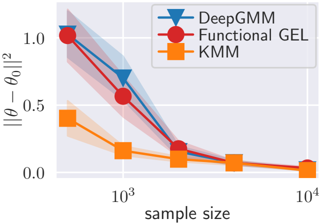


Empirical Likelihood / Kernel Method of Moment [Kremer & Z et al., 2022, Kremer et al. & Z, 2023]; cf. [Owen, 1988; Qin and Lawless, 1994; Bierens, 1982]

$$\inf_{\theta, Q \in \mathcal{P}} \alpha \, \mathrm{MMD}^2(Q, \hat{P}) + \beta D_\varphi(Q|\omega)$$

$$\text{s.t.} \quad \mathbb{E}_Q \left[(Y - g_\theta(X))^T h(Z) \right] = 0, \quad \forall h \in \mathcal{H}$$

Nonlinear/deep instrumental variable (IV) regression

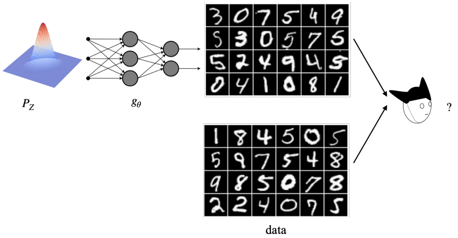


Topic of DFG project SPP 2298 "Theory Foundation of Deep Learning"

Gradient Flows of Probability Measures

Deep generative models

Previouis view of DGM (static) [Goodfellow et al. 2014...]



$$\min_{\theta} \mathrm{D}(\pi_{\mathrm{data}} | g_{\theta\#} P_Z)$$

New view of DGM (dynamic) Simulate an S/O/PDE [Chen et al. 2018, Song et al. 2021]

$$\dot{X}_t = -\nabla \xi_t(X_t), \text{ for some learned } \nabla \xi_t, \text{ e.g. NN}$$

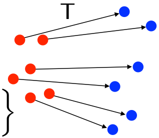
Perspective: flow and evolution of prob. measures

Wasserstein distance and optimal transport

"Euclidean distance" between probability measures

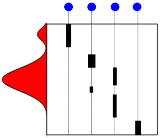
p-th order **Kantorovich-Wasserstein distance** between measures μ_0, μ_1 on $X \subset \mathbb{R}^d$ with p finite moments is defined through the Monge problem

$$W_p^p(\mu_0, \mu_1) := \min \left\{ \int |x - T(x)|^p \, \mathrm{d}\mu_0(x) \middle| \; T_\# \mu_0 = \mu_1 \right\}$$



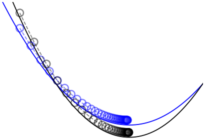
the Kantorovich problem

$$W_p^p(\mu_0, \mu_1) := \min \left\{ \int |x_0 - x_1|^p \, \mathrm{d}\Pi \middle| \; \pi_\#^{(1)}\Pi = \mu_0, \pi_\#^{(2)}\Pi = \mu_1 \right\}$$



[Peyré and Cuturi, 2019]

From gradient descent to gradient flow



Optimization problem in $\mathbb{R}^d$: $\min_{x \in \mathbb{R}^d} F(x)$

$$\text{Gradient descent} \quad x_{k+1} = x_k - \tau \cdot \nabla F(x_k)^\top$$

$$\text{Prox. step (implicit)/ JKO} \quad x_{k+1} \in \operatorname*{argmin}_x \left(F(x) + \frac{1}{2\tau} \|x - x_k\|^2 \right)$$

$$\tau \to 0 \text{ continuous time: ODE} \quad \dot{x}(t) = -\nabla F(x(t))^\top$$

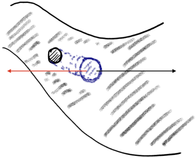
is the *gradient-flow equation* of the energy $F(x)$ in the **space** $\mathbb{R}^d$ with the Eulidean geometry described by $\|x\|^2$.

From Euclidean gradient descent to Wasserstein gradient flow

Generalizing the Euclidean geometry $\left(\mathbb{R}^d, \|\cdot\|\right)$ to $(\mathcal{P}, W_2)$

JKO: Wasserstein gradient flow [Otto, 1996, 2001]

$$\mu^{k+1} \in \operatorname*{argmin}_{\mu\in\mathcal{P}} F(\mu) + \frac{1}{2\tau} {W_2}^2(\mu, \mu^k)$$



Continuous-time ($\tau \to 0$) **gradient flow equation**

$$\partial_t \mu = -\mathrm{div}\left(\mu \nabla \frac{\delta F}{\delta \mu}[\mu]\right)$$

PDE has a **gradient structure**: $\begin{cases} \textbf{Measure Space}: & \mathcal{P} \text{ or } \mathcal{M}^+ \\ \text{Energy functional}: & F \text{ (e.g. KL)} \\ \text{Dissipation Geometry}: & W_2 \text{ or He} \end{cases}$

The merit of the right gradient flow formulation of a dissipative evolution equation is that it **separates energetics and kinetics**: The energetics endow the state space with a functional, the kinetics endow the **state space** with a (Riemannian) geometry via the metric tensor. [Otto 2001]

Inference via interacting particle systems: Langevin MC

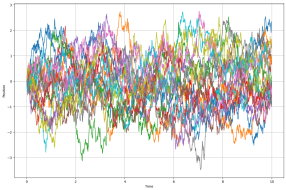
Goal: to sample from $\pi(x) = \dfrac{1}{\int e^{-V(x)}\,\mathrm{d}x} e^{-V(x)}$

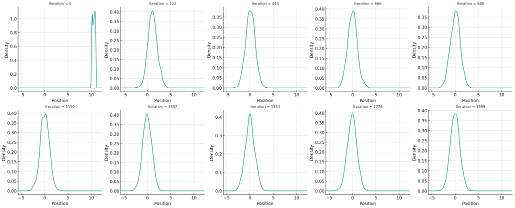
Langevin SDE

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$$

Fokker-Planck PDE

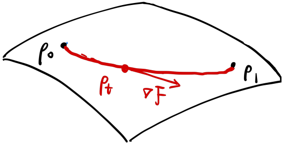
$$\partial_t \mu = \Delta\mu + \mathrm{div}\,(\mu \nabla V)$$





GF & OPT perspective

$$\min_{\mu \in A \subset \mathcal{P}} \mathrm{D_{KL}}(\mu|\pi) \text{ in } (\mathcal{P}, W_2)$$



In a series of papers jointly with A. Mielke, we provide rigorous analysis of various gradient flows beyond the W_2 setting of [Bakry and Émery, 1985] e.g. log-Sobolev.

Information divergence and Hellinger (Fisher-Rao) distance

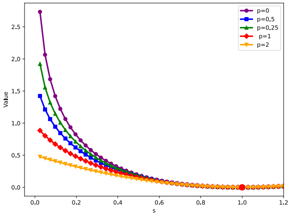
φ-divergence energy [Csiszár, 1967] $\quad \mathrm{D}_\varphi(\mu|\nu) := \int \varphi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right) \, \mathrm{d}\nu$

$$\varphi_p(s) := \frac{1}{p(p-1)}\left(s^p - p(s-1) - 1\right)$$

$$p = 2 : \chi^2, \quad p = \frac{1}{2} : \text{Hellinger}$$

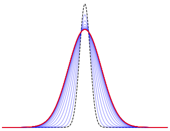
$$p \to 1 : \text{KL}, \;\; \varphi_1(s) := \varphi_{\text{KL}} = s\log s - s + 1$$

$$p \to 0 : \text{rev. KL}, \;\; \varphi_0(s) := s - 1 - \log s$$



Hellinger distance over $\mathcal{M}^+$

$$\mathrm{He}^2(\mu_0, \mu_1) = 4 \cdot \int \left(\sqrt{\mu_0} - \sqrt{\mu_1}\right)^2$$



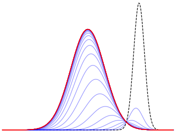
See [Gallouët and Monsaingeon, 2017, Laschos and Mielke, 2019, Z and Mielke, 2024] for other formulations.

Gradient flows over $\mathcal{M}^+$: Hellinger / Fisher-Rao

Wasserstein/diffusion: mass-preserving

Birth-death process $2\mathrm{H_2O} \rightleftharpoons 2\mathrm{H_2} + 1\mathrm{O_2}$

Hellinger gradient flows $(\mathcal{M}^+, F, \mathrm{He})$



$$\min_{\mu \in \mathcal{M}^+} F(\mu) + \frac{1}{2\tau} \mathrm{He}^2(\mu, \mu^k)$$

continuous-time $\tau \to 0 : \dot{\mu} = -\mu \cdot \frac{\delta F}{\delta \mu}[\mu]$

Example

- [Z and Mielke, 2024] Convergence analysis of KL-inference

$$\min_{\mu \in \mathcal{M}^+} \mathrm{D_{KL}}(\mu|\pi) \text{ in } (\mathcal{P}, \mathrm{He})$$

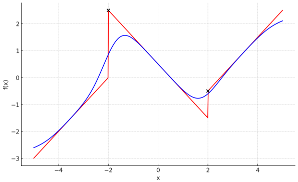
- variational inference via natural gradient: (spherical) Hellinger metric tensor gives the Fisher information matrix [Amari, 1998, Khan and Nielsen, 2018]
- entropic mirror descent in optimization [Nemirovskij and Yudin, 1983, Beck and Teboulle, 2003]

Kernel Approximation

Kernel methods and MMD

$\mathcal{H}$ is the **reproducing kernel Hilbert space** (RKHS), which satisfies

$$f(x) = \langle f, k(x, \cdot)\rangle_{\mathcal{H}}, \forall f \in \mathcal{H}, x \in \mathcal{X}$$

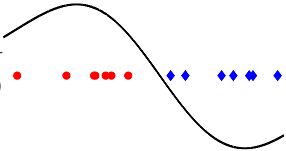


Integral operator $\mathcal{K}_\rho : L^2(\rho) \to L^2(\rho)$:

$$\mathcal{K}_\rho g(x) := \int k\left(x, x'\right) g\left(x'\right) d\rho\left(x'\right)$$

Maximum-mean discrepancy (MMD) [Gretton et al., 2012]

$$\begin{aligned}\mathrm{MMD}(\mu_0, \mu_1) :=& \left\| \int k(x, \cdot) d\mu_0 - \int k(x, \cdot) d\mu_1 \right\|_{\mathcal{H}} \\ =& \sqrt{\int \int k(x, x') \,\mathrm{d}(\mu_0 - \mu_1)(x) \,\mathrm{d}(\mu_0 - \mu_1)(x')} \\ =& \sup_{\|f\|_{\mathcal{H}} \leq 1} \int f \, d(\mu_0 - \mu_1)\end{aligned}$$



MMD as dekernelized Hellinger distance

The "MMD paper" [Gretton et al., 2012] has now $> 5k$ citations.
Dyanmic formulation of MMD: "straight line" geodesics

$$\mathrm{MMD}^2(\mu,\nu) = \min\left\{ \int_0^1 \|\xi_t\|_{\mathcal{H}}^2 \,\mathrm{d}t \;\middle|\; \dot{u} = -\mathcal{K}^{-1}\xi_t, u(0) = \mu, u(1) = \nu \right\}.$$

The integral operator
$\mathcal{K}_\rho := g(x) := \int k\left(x, x'\right) g\left(x'\right) d\rho\left(x'\right), \quad g \in L^2_\rho, L^2(\rho) \to L^2(\rho)$ is compact, positive, self-adjoint, and nuclear.

Theorem (MMD = de-kernelized Hellinger)

The dynamic formulation of the kernelized squared MMD coincides with that of the squared Hellinger distance

The Riemannian metric tensors are related by $\mathbb{G}_{\mathrm{MMD}} = \mathcal{K}_\mu \circ \mathbb{G}_{\mathrm{He}}(\mu)$.

Gradient flow geometries obtained by kernelization

$$\boxed{\text{Wasserstein}} \xrightarrow{\text{kernelize } \nabla \frac{\delta F}{\delta \mu}[\mu]} \boxed{\text{Stein}}$$

Theorem [Z and Mielke, 2024] The Riemannian metric tensors of Hellinger satisfy $\mathbb{G}_{\mathsf{MMD}} = \mathcal{K}_\mu \circ \mathbb{G}_{\mathsf{He}}(\mu)$, i.e.,

MMD=(de-)kernelized Hellinger

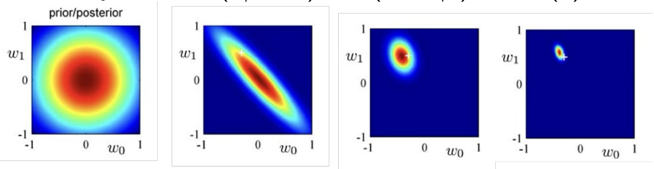
$$\boxed{\text{MMD}} \xrightarrow{\text{kernelize } \frac{\delta F}{\delta \mu}[\mu]} \boxed{\text{Hellinger}}$$

Statistical Inference via Gradient Flows

Bayesian inference and probabilistic ML

Infer posterior distribution π of the model parameters θ given data,

$$\text{Bayes rule: } \pi(\theta|\mathrm{Data}) \propto P(\mathrm{Data}|\theta) \cdot \mathrm{Prior}(\theta)$$



In practice, the exact π is intractable: **approximate inference** [Jordan et al., 1999, Wainwright and Jordan, 2008]

$$\min_{\mu \in A \subset \mathcal{P}} \mathrm{D_{KL}}(\mu | \pi(\theta|\mathrm{Data})).$$

Gaussian **variational inference**: $\mu \in \mathcal{N}^d$; also Laplace approx.

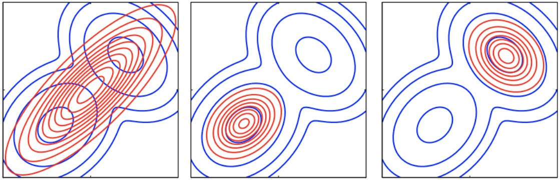
Sampling / MCMC: generate samples $\theta^i \sim \pi$, $\frac{1}{N}\sum_{i=1}^{N} \delta_{\theta^i} \to \pi$

Inference with forward and reverse KL

$$\min_{\mu \in \mathcal{N}^d \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\pi|\mu) \quad \text{vs.} \quad \min_{\mu \in \mathcal{N}^d \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\mu|\pi)$$

forward / inclusive — reverse / exclusive

mode-covering — mode-seeking



[Bishop 2006]

Forward (incl.) KL inference as kernelized Wasserstein flows

$$\min_{\mu \in A \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\pi|\mu).$$

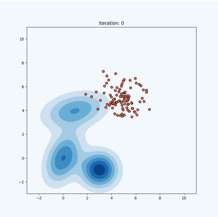
Preferable to $\mathrm{D}_{\mathrm{KL}}(\mu|\pi)$.Existing algorithms are based on heuristics [Minka 2013; Naesseth et al. 2020; Jerfel et al. 2021; McNamara et al. 2024; Zhang et al. 2022; ...]

Wasserstein gradient flow [Z. 2024]: $\dot{\mu} = \mathrm{div}\left(\mu \nabla \left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)\right)$

Not implementable due to $\nabla(1 - \mathrm{d}\pi / \mathrm{d}\mu)$.

Kernel approx. [Z. 2024; Gladin et al. & Z. NeurIPS 2024]:

$$\dot{\mu} = \mathrm{div}\left(\mu \nabla \int k(z, \cdot)\left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(z)\right) \mathrm{d}\mu(z)\right)$$



Theorem. The PDE has gradient structure:

$$\begin{cases} \text{Energy functional}: & \frac{1}{2}\,\mathsf{MMD}^2(\cdot, \pi) \\ \text{Geometry}: & \text{Wasserstein} \end{cases}$$

$\mathsf{MMD}^2(\mu, \pi) = \mathbb{E}_{x,y\sim\mu} k(x, y) + \mathbb{E}_{x,y\sim\pi} k(x, y) - 2\mathbb{E}_{x\sim\mu, y\sim\pi} k(x, y)$

Unbalanced transport gradient flows of forward (incl.) KL inference [Gladin et al. & Z. NeurIPS 2024; Z. 2024]

[Z. 2024] precise connection between the gradient flows of:

$$\min_{\mu\in A\subset\mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\pi|\mu) \quad \text{and} \quad \min_{\mu\in A\subset\mathcal{P}} \mathrm{MMD}^2(\mu,\pi)$$

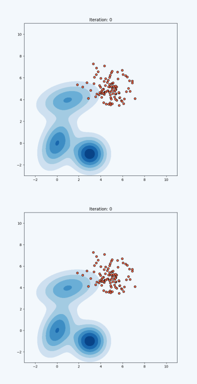
[Arbel et al. 2019] studied the latter without guarantee of convergence; [Chizat, 2022, Hagemann et al., 2023, Neumayer et al., 2024, Chen et al., 2024]

[Z. 2024] Wasserstein-Fisher-Rao flow: reaction-diffusion eq:

$$\dot{\mu} = \underbrace{\alpha\cdot\operatorname{div}\left(\mu\nabla\left(1-\frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)\right)}_{\text{Wasserstein: transport}} \underbrace{-\beta\cdot\mu\cdot\left(1-\frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)}_{\text{Fisher-Rao: birth-death}}$$

Interaction-force transport [Z. 2024, Gladin et al. & Z. NeurIPS 2024] with global convergence guarantee

$$\dot{\mu} = \alpha\cdot\operatorname{div}\left(\mu\nabla\int k(x,\cdot)\,\mathrm{d}\left(\mu-\pi\right)(x)\right) - \beta\cdot(\mu-\pi)$$



JKO splitting scheme

The PDE

$$\dot{\mu} = \alpha \cdot \mathsf{div}\left(\mu \nabla \int k(x,\cdot)\,\mathrm{d}(\mu-\pi)(x)\right) - \beta \cdot (\mu - \pi)$$

can be simulated using the JKO scheme

$$\mu^{\ell+\frac{1}{2}} \leftarrow \operatorname*{argmin}_{\mu \in \mathcal{P}} F(\mu) + \frac{1}{2\tau} W_2^2(\mu, \mu^{\ell}), \qquad \text{(Wasserstein step)}$$

$$\mu^{\ell+1} \leftarrow \operatorname*{argmin}_{\mu \in \mathcal{P}} F(\mu) + \frac{1}{2\eta} \mathsf{MMD}^2(\mu, \mu^{\ell+\frac{1}{2}}), \qquad \text{(MMD step)}$$

for $F(\mu) = \frac{1}{2}\,\mathsf{MMD}^2(\mu, \pi)$.

Insight on the variational principle: kernel methods vs information geometry

Theorem [Z, 2024] Suppose the kernel k is bounded and integrally strictly positive definite. Then, the solutions of the following variational problems coincide:

$$\min_{\mu \in \mathcal{P}} \frac{1}{2} \mathsf{MMD}^2(\mu, \pi) + \frac{1}{2\eta} \mathsf{MMD}^2(\mu, \mu^l).$$

$$\operatorname*{argmin}_{\mu \in \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\pi|\mu) + \frac{1}{\eta} \mathrm{D}_{\mathrm{KL}}(\mu^l|\mu).$$

Thank you!

This talk is based on the following papers

- **Z**. Inclusive KL Minimization: A Wasserstein-Fisher-Rao Gradient Flow Perspective. arXiv preprint
- Gladin-Dvurechensky-Mielke-**Z**. Interaction-Force Transport Gradient Flows. *NeurIPS 2024*
- **Z**-Mielke. Kernel Approximation of Fisher-Rao Gradient Flows. arXiv preprint
  Kremer-Nemmour-Schölkopf-**Z**. Estimation Beyond Data Reweighting: Kernel Method of Moments. *ICML 2023*.
- **Z**-Jitkrittum-Diehl-Schölkopf. Kernel Distributionally Robust Optimization. *AISTATS 2021*.

For more information, see my website: https://jj-zhu.github.io/ ; PhD position (Berlin) available