Wasserstein Gradient Flows of MMD Functionals with Distance Kernel and Cauchy Problems on Quantile Functions

joint work with

Richard Duong, TU Berlin

Robert Beinert, TU Berlin

Johannes Hertrich, UCL

Gabriele Steidl, TU Berlin

University of South Carolina, Columbia, 30.08.2024.

Joint ACM and RTG data science seminar (Changhui Tan, Siming He, Wuchen Li).

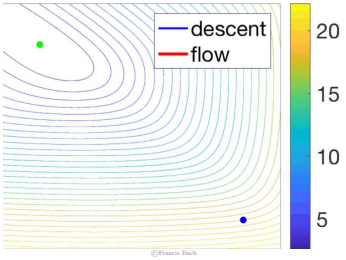
Problem: Target measure $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ is unknown, only **samples** are available.

Goal: Recover ν.

Solution: *minimize* the metric **kernel discrepancy**

$$\mathcal{F}_\nu \coloneqq \mathrm{MMD}_K(\cdot, \nu)^2 \colon \mathcal{P}_2(\mathbb{R}^d) \to [0, \infty)$$

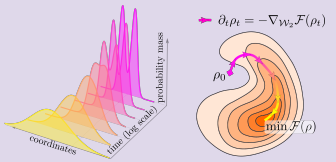
(which can be estimated using samples) to ν by finding curve of measures $(\gamma_t)_{t>0} \subset \mathcal{P}_2(\mathbb{R}^d)$ along which $\mathcal{F}_\nu$ *decreases* "the fastest".



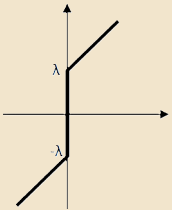
One way to construct $(\gamma_t)_{t>0}$: **Wasserstein gradient flows**.
Different kernels $K \rightsquigarrow$ very different behavior of $(\gamma_t)_{t>0}$.

Aim of this preprint: study behavior of $(\gamma_t)_{t>0}$ for irregular kernel $K(x, y) \coloneqq -|x - y|$ for $d = 1$.

1. Optimal transport and Wasserstein gradient flow

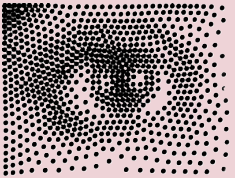


2. Maximal monotone inclusions on Hilbert spaces

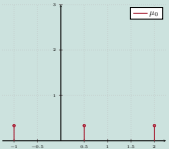


$$(I + \lambda \partial F)^{-1}$$

3. Negative distance kernel, Max. Mean Discrepancy



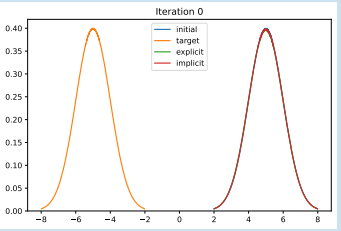
4. Wasserstein gradient flow of the MMD with negative distance kernel



5. Invariant subsets & smoothing properties

$$\gamma_0 = \delta_x,$$
$$\gamma_t \sim \mathcal{U}[a_t, b_t], \ t > 0.$$

6. Numerical results

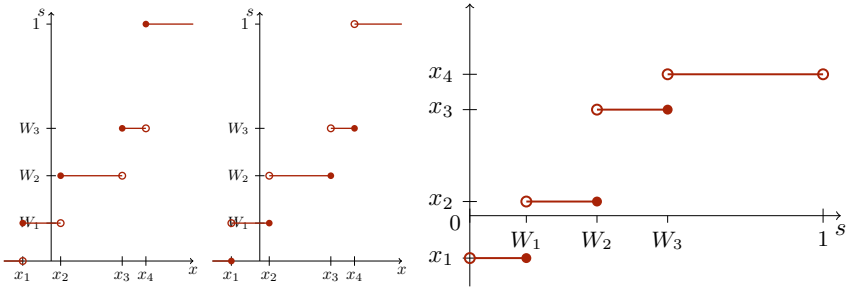


The **cumulative distribution functions** (CDFs) of $\nu \in \mathcal{P}(\mathbb{R})$ are given by

$$R_\nu^+ \colon \mathbb{R} \to [0,1], \quad R_\nu^+(x) \coloneqq \nu\big((-\infty, x]\big), \qquad R_\nu^- \colon \mathbb{R} \to [0,1], \quad R_\nu^-(x) \coloneqq \nu\big((-\infty, x)\big).$$

and its (generalized inverse, the) **quantile function** by

$$Q_\nu \colon (0,1) \to \mathbb{R}, \quad Q_\nu(s) \coloneqq \min\{x \in \mathbb{R} : R_\nu^+(x) \geq s\}.$$



Left to right: R_ν^+, R_ν^-, Q_ν, for $\nu = \sum_{k=1}^4 w_k \delta_{x_k}$, $W_j \coloneqq \sum_{k=1}^j w_k$.

Consider the subset of probability measures $\mathcal{P}(\mathbb{R})$,

$$\mathcal{P}_2(\mathbb{R}) := \left\{ \mu \in \mathcal{P}(\mathbb{R}) : \int_{\mathbb{R}} x^2 \, \mathrm{d}\mu(x) < \infty \right\}.$$

On $\mathcal{P}_2(\mathbb{R})$, the **Wasserstein-2 metric** is

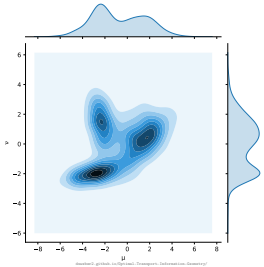
$$W_2(\mu,\nu)^2 = \min_{\pi \in \Gamma(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} (x-y)^2 \, \mathrm{d}\pi(x,y), \quad \mu,\nu \in \mathcal{P}_2(\mathbb{R}),$$

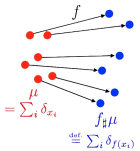
$$\Gamma(\mu,\nu) := \{\pi \in \mathcal{P}(\mathbb{R}\times\mathbb{R}) : (P_1)_{\#}\pi = \mu, (P_2)_{\#}\pi = \nu\}$$

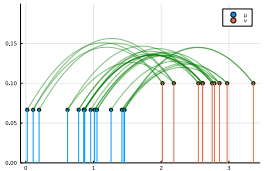
with the **projections** $P_i(x_1,x_2) := x_i$.

The **push-forward** $\#$ acts as $(f_{\#}\sigma)(A) := \sigma(f^{-1}(A))$ for all measurable sets $A \subset \mathbb{R}$.

The optimal plan $\hat{\pi}$ is unique (in 1D).



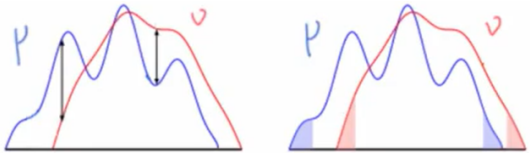




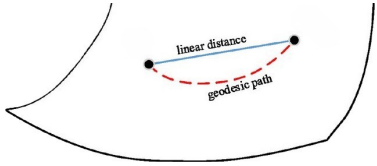
Push-forward $\hat{\pi}$ realizes monotone rearrangement.

Geodesic (= shortest constant-speed path) between μ and ν is

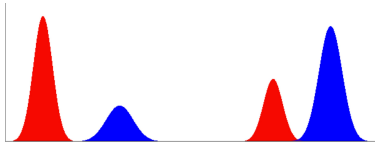
$$\gamma_t := \big((1-t)P_1 + tP_2\big)_{\#}\hat{\pi}, \quad t \in [0,1], \tag{1}$$



Vertical (L_2, $g_t := (1-t)f_\mu + tf_\nu$) vs. horizontal ($W_2$) mass displacement.



Geodesic on a manifold.



W_2 geodesic

Definition (W_2-Geodesic convexity)

$\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$ is **convex along geodesics** if

$$\mathcal{F}(\gamma_t) \leq (1-t)\,\mathcal{F}(\mu) + t\,\mathcal{F}(\nu), \quad \forall \mu, \nu \in \mathcal{P}_2(\mathbb{R}),\ \gamma_t \text{ from (1)}.$$

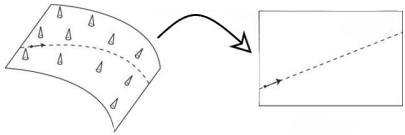
Theorem (Isometric embedding of $(\mathcal{P}_2(\mathbb{R}), W_2)$)

Let $\mathcal{C}(0,1) \coloneqq \{Q_\mu : \mu \in \mathcal{P}_2(\mathbb{R})\} \subset L_2(0,1)$ be the set of quantile functions.

The map

$$\mathcal{P}_2(\mathbb{R}) \to \mathcal{C}(0,1), \qquad \mu \mapsto Q_\mu$$

*is an **isometric isomorphism** with inverse $Q \mapsto Q_{\#}\Lambda_{(0,1)}$, where Λ is the Lebesgue measure.*

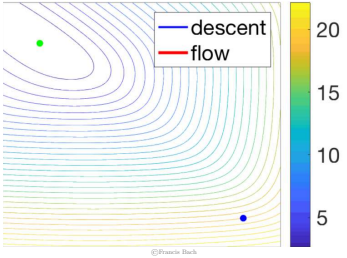


Isometric embedding.

KEY IDEA: instead of working with $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}) \to (-\infty, \infty]$, find

$$F\colon L_2(0,1) \to (-\infty, \infty] \quad \text{satisfying} \quad F(Q_\mu) = \mathcal{F}(\mu).$$

Given: Hilbert space H, energy functional $f\colon H \to (-\infty, \infty]$ (convex, bounded below).



Idea: minimize $f \rightsquigarrow$ find an *absolutely continuous* curve $t \mapsto x(t)$ such that $t \mapsto f(x(t))$ decreases "the fastest".

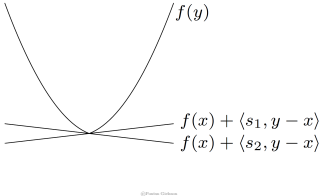
$$\frac{\mathrm{d}}{\mathrm{d}t} x(t) = -\nabla f(x(t))$$

If f is not differentiable, there might be more than one descent direction.

Definition

The **subdifferential** of a **convex** function $F\colon H \to \mathbb{R}$ is

$$\partial F(u) \coloneqq \big\{v \in H : F(w) \geq F(u) + \langle v, w - u\rangle \ \forall w \in H\big\}.$$

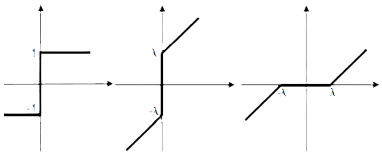


Theorem (Properties of the resolvent)

Let $F\colon H \to \mathbb{R}$ be convex and lower semicontinuous, $F \not\equiv \infty$. Then ∂F is maximal monotone, hence $\forall \varepsilon > 0$, the ***resolvent***

$$J_\varepsilon^{\partial F} \coloneqq (I + \varepsilon \partial F)^{-1} : H \to H$$

is single-valued.



∂F, $I + \lambda \partial F$, $(I + \lambda \partial F)^{-1}$, where $F \coloneqq |\cdot|$.

This theorem tells us that gradient flows in Hilbert spaces exist and are unique.

Theorem (Existence, regularity of strong solutions to Cauchy problems (Brezis '67))

Let $f\colon H \to \mathbb{R}$ s.t. $\partial f\colon H \to 2^H$ is maximal monotone and $g_0 \in \operatorname{dom}(\partial f)$. Then $\exists! g\colon [0,\infty) \to H$ s.t.

- *$g(0) = g_0$ and $\frac{\mathrm{d}g}{\mathrm{d}t}(t) \in -(\partial f)\big(g(t)\big)$ for almost all $t > 0$, and $g(t) \in \operatorname{dom}(\partial f)$ for all $t > 0$*
- *g is Lipschitz continuous on $[0,\infty)$.*
- *g is given by the exponential formula*

$$g(t) = \lim_{n\to\infty} \left(J^{\partial f}_{\frac{t}{n}}\right)^n (g_0) \qquad \text{uniformly in } t \text{ on compact intervals.}$$

For $\mathcal{P}_2(\mathbb{R})$ instead of H: what are the analogs of the tangent vector $\frac{\mathrm{d}}{\mathrm{d}t}x(t)$ and the subdifferential ∂f?

Definition (Fréchet subdifferential in Wasserstein space)

The (reduced) Fréchet subdifferential of $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$ at μ is

$$\partial \mathcal{F}(\mu) \coloneqq \left\{ \xi \in L^2(\mathbb{R}; \mu) : \mathcal{F}(\nu) - \mathcal{F}(\mu) \geq \int_{\mathbb{R}\times\mathbb{R}} \xi(x_1)(x_2 - x_1)\, \mathrm{d}\hat{\pi}(x, y) + o\big(W_2(\mu, \nu)\big) \right\}$$

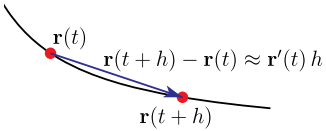
A curve $\gamma\colon (0, \infty) \to \mathcal{P}_2(\mathbb{R})$ is **absolutely continuous** if $\exists$ L^2-Borel **velocity field** $(v_t\colon \mathbb{R} \to \mathbb{R})_{t>0}$ s.t.

$$\partial_t \gamma_t + \nabla \cdot (v_t \gamma_t) = 0, \qquad (t, x) \in (0, \infty) \times \mathbb{R}, \text{ weakly.} \qquad \text{(Continuity Eq.)}$$

Definition (Wasserstein gradient flow)

A absolutely continuous curve $\gamma\colon (0, \infty) \to \mathcal{P}_2(\mathbb{R})$ with velocity field $\big(v_t \in \mathrm{T}_{\gamma_t} \mathcal{P}_2(\mathbb{R})\big)_{t>0}$ is a *Wasserstein gradient flow with respect to* $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$ if

$$v_t \in -\partial \mathcal{F}(\gamma_t), \quad \text{for a.e. } t > 0.$$



Tangent vector.

Theorem (Ambrosio, Gigli, Savaré (2005))

Let $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$ be bounded from below, lower semicontinuous, ***geodesically convex****.*

Existence and uniqueness. *Then $\exists!$* ***Wasserstein gradient flow*** *$\gamma\colon (0,\infty) \to \mathcal{P}_2(\mathbb{R})$ with respect to $\mathcal{F}$ starting at $\gamma(0+) = \mu_0 \in \mathcal{P}_2(\mathbb{R})$.*

Approximation scheme. *The piecewise constant curves constructed from the iterates*

$$\mu_{n+1} \coloneqq \operatorname*{argmin}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathcal{F}(\mu) + \frac{1}{2\tau} W_2^2(\mu_n, \mu) \right\}, \qquad \tau > 0, \tag{Minimizing movement scheme}$$

i.e., γ_τ defined by $\gamma_\tau|_{(n\tau,(n+1)\tau]} \coloneqq \mu_n$, $n \in \mathbb{N}$, converge locally uniformly to γ as $\tau \downarrow 0$.

Convergence speed. *If $\bar{\mu}$ is a minimizer of $\mathcal{F}$, then*

$$\mathcal{F}(\gamma_t) - \mathcal{F}(\bar{\mu}) \leq \frac{1}{2t} W_2^2(\mu_0, \bar{\mu}). \tag{Sublinear convergence rate}$$

We establish a correspondence between $L_2(0,1)$-gradient flows of F and Wasserstein gradient flows of $\mathcal{F}$.

Theorem (Quantile function reformulation [DSBHS24])

Let $F\colon L_2(0,1) \to (-\infty, \infty]$ be convex and lsc, $F \not\equiv \infty$ with $F(Q_\mu) = \mathcal{F}(\mu)$ for all $\mu \in \mathcal{P}_2(\mathbb{R})$. Assume $J_\varepsilon^{\partial F}$ maps $\mathcal{C}(0,1)$ into itself $\forall \varepsilon > 0$. Then, $\forall$ initial datum $g_0 \in \mathcal{C}(0,1) \cap \operatorname{dom}(\partial F)$,

$$\begin{cases} \partial_t g(t) + \partial F(g(t)) \ \ni \ 0, \quad t \in (0,\infty), \\ g(0) = g_0, \end{cases} \tag{Cauchy Problem}$$

has a unique strong solution g.

*The curve $\gamma_t \coloneqq (g(t))_\# \Lambda_{(0,1)}$ has quantile functions $Q_{\gamma_t} = g(t)$ and is a **Wasserstein gradient flow of** $\mathcal{F}$ with $\gamma(0+) = (g_0)_\# \Lambda_{(0,1)}$.*

Consider the **negative distance kernel** $K(x, y) \coloneqq -|x - y|$.

K is only **conditionally positive definite**.

Motivation: electrostatic principles, interacting species, dithering.

Definition (MMD)

The **maximum mean discrepancy** of with respect to K is

$$\mathcal{P}_2(\mathbb{R}) \times \mathcal{P}_2(\mathbb{R}) \to [0, \infty),$$
$$(\mu, \nu) \mapsto \frac{1}{2} \operatorname{MMD}_K(\mu, \nu)^2 = \int_{\mathbb{R} \times \mathbb{R}} K(x, y) \, \mathrm{d}(\mu - \nu)(x) \, \mathrm{d}(\mu - \nu)(y).$$

Dithering

$$\mathcal{F}_\nu(\mu) \coloneqq -\frac{1}{2} \underbrace{\int_{\mathbb{R} \times \mathbb{R}} |x - y| \, \mathrm{d}\mu(x) \, \mathrm{d}\mu(y)}_{\text{interaction}} + \underbrace{\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y| \, \mathrm{d}\mu(x) \, \mathrm{d}\nu(y)}_{\text{potential}}. \tag{2}$$

We now apply this theorem to the MMD with the negative distance kernel, i.e. to

$$\mathcal{F}_\nu \coloneqq \frac{1}{2}\,\mathrm{MMD}_K(\cdot,\nu).$$

Lemma (Properties of F_ν)

The functional

$$F_\nu\colon L_2(0,1)\to\mathbb{R},\qquad u\mapsto \int_0^1 \Big((1-2s)\big(u(s)-Q_\nu(s)\big)+\int_0^1 |u(s)-Q_\nu(t)|\,\mathrm{d}t\Big)\,\mathrm{d}s$$

is convex and continuous, we have $F_\nu(Q_\mu)=\mathcal{F}_\nu(\mu)\ \forall \mu\in\mathcal{P}_2(\mathbb{R})$, and

$$\partial F_\nu(u)=\Big\{f\in L_2(0,1): f(s)\in 2\big[R_\nu^-\big(u(s)\big),R_\nu^+\big(u(s)\big)\big]-2s \text{ for a.e. } s\in(0,1)\Big\},\quad u\in L_2(0,1)$$

and $J_\varepsilon^{\partial F_\nu}$ maps $\mathcal{C}(0,1)$ into itself $\forall\varepsilon>0$.

Proof. By elementary means, nothing fancy happens here.

The lower Lipschitz constant of $g \in \mathcal{C}(0,1)$ is

$$L_{\text{low}}(g) := \max\left\{ L \geq 0 : \frac{g(s_1) - g(s_2)}{s_1 - s_2} \geq L \ \forall s_1, s_2 \in (0,1) \right\} \geq 0,$$

If $\mu = \delta_x$, then $Q_\mu \equiv x$ and $L_{\text{low}}(Q_\mu) = 0$.

Theorem (Time evolution of CDF's lower Lipschitz constants)

Let $\nu \in \mathcal{P}_2(\mathbb{R})$ with $L_{\text{low}}(Q_\nu) > 0$, and $g_0 = Q_{\mu_0} \in \mathcal{C}(0,1)$. We have

$$\text{Lip}(R_{\gamma_t}) \leq \left(L_{\text{low}}(g_0) \cdot e^{-\frac{2t}{L_{\text{low}}(Q_\nu)}} + L_{\text{low}}(Q_\nu) \cdot (1 - e^{-\frac{2t}{L_{\text{low}}(Q_\nu)}}) \right)^{-1} < \infty.$$

Theorem (Continuity is preserved & monotonicity of the support)

Let $\nu \in \mathcal{P}_2(\mathbb{R})$, g_0 be continuous and g the solution of the Cauchy problem starting in g_0.

- *$g(t)$ is continuous for all $t \geq 0$.*
- *The ranges fulfill $\overline{g(t_1)(0,1)} \subseteq \overline{g(t_2)(0,1)}$ for all $0 \leq t_1 \leq t_2$.*

Corollary (Point measure target)

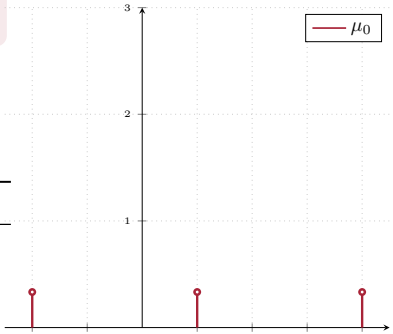
Let $\nu \coloneqq \sum_{j=1}^{n} w_j \delta_{x_j}$. Then F_ν flow is given by

$$[g(t)](s) \coloneqq \begin{cases} Q_{\mu_0}(s) + 2\,(s - R_{s,0})\,t, & t \in [t_{s,0}, t_{s,1}), \\ x_{s,j} + 2\,(s - R_{s,j})\,(t - t_{s,j}), & t \in [t_{s,j}, t_{s,j+1}), \\ Q_\nu(s), & t \geq t_{s,|\ell_s - k_s|}, \end{cases}$$

where

$$t_{s,0} \coloneqq 0, \quad t_{s,1} \coloneqq \frac{x_{s,1} - Q_{\mu_0}(s)}{2(s - R_{s,0})}, \quad t_{s,j+1} \coloneqq t_{s,j} + \frac{x_{s,j+1} - x_{s,j}}{2(s - R_{s,j})},$$

	$Q_{\mu_0}(s) \leq Q_\nu(s)$	$Q_{\mu_0}(s) \geq Q_\nu(s)$	
ℓ_s	$W_{\ell_s - 1} < s < W_{\ell_s}$	$W_{\ell_s - 1} < s < W_{\ell_s}$	
k_s	$x_{k_s} \leq Q_{\mu_0}(s) < x_{k_s + 1}$	$x_{k_s - 1} < Q_{\mu_0}(s) \leq x_{k_s}$	
$x_{s,j}$	$x_{k_s + j}$	$x_{k_s - j}$	$j \leq \lvert \ell_s - k_s \rvert$
$R_{s,j}$	$W_{k_s + j}$	$W_{k_s - j - 1}$	$j \leq \lvert \ell_s - k_s \rvert - 1$



Let $\tau > 0$. The minimizing movement (or JKO) scheme,

$$\mu_{n+1} \coloneqq \operatorname*{argmin}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \Big\{ \mathcal{F}_\nu(\mu) + \frac{1}{2\tau} W_2^2(\mu_n, \mu) \Big\},$$

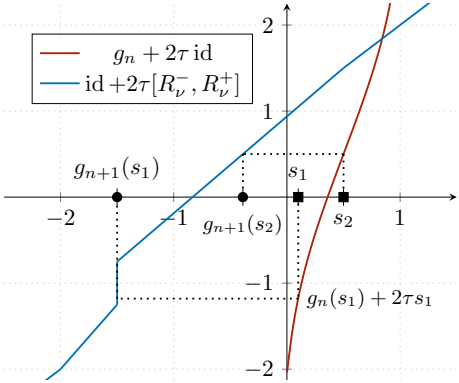
can be rewritten using the isometry $\mathcal{P}_2(\mathbb{R}) \to \mathcal{C}(0,1)$ as

$$g_{n+1} = \operatorname*{argmin}_{g \in \mathcal{C}(0,1)} \Big\{ F_\nu(g) + \frac{1}{2\tau} \int_0^1 |g - g_n|^2 \, \mathrm{d}s \Big\}$$
$$\overset{F_\nu \in \Gamma_0(L_2(0,1))}{=} (I + \tau \partial F_\nu)^{-1}(g_n),$$

which is equivalent to

$$g_n(s) + 2\tau s \in g_{n+1}(s) + 2\tau \big[R_\nu^-(g_{n+1}(s)), R_\nu^+(g_{n+1}(s)) \big] \quad (3)$$

for all $s \in (0,1)$.



Implicit Euler step visualized.

$(g_n)_{\#}\Lambda_{(0,1)}$ approximates WGF for $t \in (n\tau, (n+1)\tau]$.

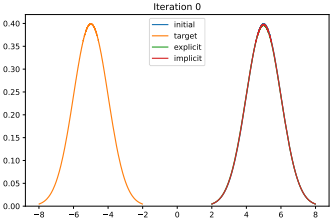
$(g_n)_{n \in \mathbb{N}} \to Q_\nu$ weakly in $L_2(0,1)$ and $\mu_n \coloneqq (g_n)_{\#}\Lambda_{(0,1)} \to \nu$ narrowly for fixed $\tau > 0$.

If $R_\nu^+ = R_\nu^- =: R_\nu$, we can also use explicit Euler discretization

$$g_{n+1} = g_n - \tau \nabla F_\nu(g_n) = g_n - 2\tau(R_\nu \circ g_n - \mathrm{id}),$$

Advantage: we don't have to solve an inclusion in each step.

Disadvantage: weaker convergence guarantees, might not preserve $\mathcal{C}(0,1)$ (iterates not monotone).



$\mu_0 = \mathcal{N}(-5,1)$, $\nu = \mathcal{N}(5,1)$.

- Extension to higher dimensions. Lagrangian reformulation now involves flow map.
- Deterministic particle approximations and mean field limits.
- Convergence of explicit scheme, distance of implicit and explicit iterates.
- Generalize Lipschitz properties to Hölder properties.
- Applying similar techniques to non-convex functional and then regularize.

- Reformulation as **maximal monotone** inclusion Cauchy problem in $L_2(0,1)$ via **quantile functions**.
- Comprehensive description of solutions' behavior, qualitative description of **instantaneous measure-to-L^∞ regularization**.
- Implicit Euler is simple.

Thank you for your attention!

I am happy to take any questions.

Paper link: `https://arxiv.org/abs/2408.07498`

My website: `viktorajstein.github.io`

References I

[AGS08] L. Ambrosio, N. Gigli, and G. Savare, *Gradient flows*, 2nd ed., Lectures in Mathematics ETH Zürich, Birkhäuser, Basel, 2008.

[Bre73] Haim Brezis, *Operateurs maximaux monotones*, North-Holland Mathematics Studies, 1973 (French).

[CL71] M. G. Crandall and T. M. Liggett, *Generation of semi-groups of nonlinear transformations on general banach spaces*, American Journal of Mathematics **93** (1971), no. 2, 265–298.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto, *The variational formulation of the Fokker–Planck equation*, SIAM Journal on Mathematical Analysis **29** (1998), no. 1, 1–17.

Interpolating between OT and KL regularized OT using Rényi Divergences

Rényi divergence $\notin \{f\text{-div.}, \text{Bregman div.}\}$, $\alpha \in (0,1)$

$$R_\alpha(\mu \mid \nu) \coloneqq \frac{1}{\alpha - 1} \ln \left[\int_X \left(\frac{\mathrm{d}\mu}{\mathrm{d}\tau} \right)^\alpha \left(\frac{\mathrm{d}\nu}{\mathrm{d}\tau} \right)^{1-\alpha} \mathrm{d}\tau \right],$$

$$\mathrm{OT}_{\varepsilon,\alpha}(\mu,\nu) \coloneqq \min_{\pi \in \Pi(\mu,\nu)} \langle c, \pi \rangle + \varepsilon R_\alpha(\pi \mid \mu \otimes \nu)$$

is a metric, where $\varepsilon > 0$, $\mu, \nu \in \mathcal{P}(X)$, X compact.

$$\mathrm{OT}(\mu,\nu) \xleftarrow[\text{or } \varepsilon \to 0]{\alpha \searrow 0} \mathrm{OT}_{\varepsilon,\alpha}(\mu,\nu) \xrightarrow{\alpha \nearrow 1} \mathrm{OT}_\varepsilon^{\mathrm{KL}}(\mu,\nu).$$

In the works: **debiased** Rényi-Sinkhorn divergence

$$\mathrm{OT}_{\varepsilon,\alpha}(\mu,\nu) - \frac{1}{2}\mathrm{OT}_{\varepsilon,\alpha}(\mu,\mu) - \frac{1}{2}\mathrm{OT}_{\varepsilon,\alpha}(\nu,\nu).$$