Minimizing f-Divergences by Interpolating Velocity Fields

Song Liu[1] (song.liu@bristol.ac.uk), Jiahao Yu[1], Jack Simons[1], Mingxuan Yi[1], Mark Beaumont[1]
[1]School of Mathematics, University of Bristol, UK

Supported by EPSRC CDT in Computational Statistics and Data Science, Alumni and Friends.

1. Motivations

- Many tasks can be formulated as **minimizing statistical discrepancies** between a particle distribution q and a target distribution p:
  - Variational Inference, Generative Modeling ...
- f-divergences are common choices of such statistical discrepancies:
  - **Definition:** $\mathrm{D}_f[p,q] := \int q(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}$
  - Examples: Forward KL, Backward KL, Pearson's χ^2, and Neyman's χ^2...
- How to **minimize these divergences by moving** q**'s particles in sample space** ($\mathbb{R}^d$)?
  - Particle movement is governed by velocity field.

We show that velocity field induced by the Wasserstein Gradient Flow can be effectively estimated via interpolation techniques.

2. Wasserstein Gradient Flow

- **Wasserstein Gradient Flow** (WGF) of a functional objective $\mathcal{F}(q_t)$ is a curve in a probability space $\mathcal{P}(\mathbb{R}^d)$

$$q_t : \mathbb{R}^+ \to \mathcal{P}(\mathbb{R}^d).$$

  - As $t \to \infty$, $\mathcal{F}(q_t)$ is reduced.
- Let $\mathcal{F}(q_t)$ be the f-divergence $D_f[p, q_t]$, WGF q_t induces the following particle moving ODE (Yi et al., 2023, Gao et al., 2019, Ansari 2021):

$$\mathrm{d}\boldsymbol{x}_t = \nabla(h \circ r_t)(\boldsymbol{x}_t)\mathrm{d}t.$$

  - where $h(r_t) = r_t f'(r_t) - f(r_t), r_t := \frac{p}{q_t}$.

In plain words, moving particles $\boldsymbol{x}_t$ according to the velocity field $\nabla(h \circ r_t)(\boldsymbol{x}_t)$ reduces $D_f[p, q_t]$ over time t.

- In practice, we move particles by the Euler discretization of the above ODE:
  - Draw particles $\boldsymbol{x}_0$ from an initial distribution q_0
  - For time $t = 0, 1 \dots T$:

$$\boldsymbol{x}_{t+1} := \boldsymbol{x}_t + \eta \nabla(h \circ r_t)(\boldsymbol{x}_t)$$

  where η is a small step size.

3. Velocity Field Estimation by Interpolation

- How to compute the velocity field $\nabla(h \circ r_t)(\boldsymbol{x}^\star)$?
  - For backward KL, $h \circ r_t = \log r_t$.
  - We do not know r_t.
- Nadaraya-Watson (NW) Interpolation:
  - Observe $g(\boldsymbol{x})$ at $\{\boldsymbol{x}_i\}_{i=1}^n \sim q$, NW interpolates $g(\boldsymbol{x}^\star)$ by computing:

$$\hat{g}(\boldsymbol{x}^\star) := \widehat{\mathbb{E}}_q[k_\sigma(\boldsymbol{x}, \boldsymbol{x}^\star) g(\boldsymbol{x})] / \widehat{\mathbb{E}}_q[k_\sigma(\boldsymbol{x}, \boldsymbol{x}^\star)].$$

- NW interpolation of the backward KL field is

$$\hat{\boldsymbol{u}}_t(\boldsymbol{x}^\star) := \widehat{\mathbb{E}}_{q_t}[k_\sigma(\boldsymbol{x}, \boldsymbol{x}^\star) \nabla \log r_t(\boldsymbol{x})] / \widehat{\mathbb{E}}_{q_t}[k(\boldsymbol{x}, \boldsymbol{x}^\star)],$$

  - not tractable as we do not know r_t.
  - What if we know the target $p(\boldsymbol{x})$? e.g., Bayesian inference

Due to integration by parts,

$$\mathbb{E}_{q_t}[k^\star_\sigma \nabla \log r_t(\boldsymbol{x})] = \mathbb{E}_{q_t}[k^\star_\sigma \nabla \log p(\boldsymbol{x}) + \nabla k^\star_\sigma].$$

NW estimator of the backward KL velocity field:

$$\hat{\boldsymbol{u}}_t(\boldsymbol{x}^\star) \approx \underbrace{\widehat{\mathbb{E}}_{q_t}[k^\star_\sigma \nabla \log p(\boldsymbol{x}) + \nabla k^\star_\sigma] / \widehat{\mathbb{E}}_{q_t}[k^\star_\sigma]}_{\text{Stein Variational Gradient Descent}}.$$

4. Local Linear Interpolation of Velocity Fields

- How to interpolate if we only have samples $\boldsymbol{x} \sim p$?

Mirror divergence:

Let $D_\phi[p,q]$ and $D_\psi[p,q]$ denote two f-divergences with f being ϕ and ψ respectively. D_ψ is the mirror of D_ϕ if and only if $\psi'(r) \triangleq r\phi'(r) - \phi(r)$, where $\triangleq$ means equal up to a constant.

Suppose h is associated with D_ϕ,

$$h \circ r = \underset{d}{\mathrm{argmax}}\, \mathbb{E}_p[d(\boldsymbol{x})] - \mathbb{E}_q[\psi_{\mathrm{con}}(d(\boldsymbol{x}))], \quad (1)$$

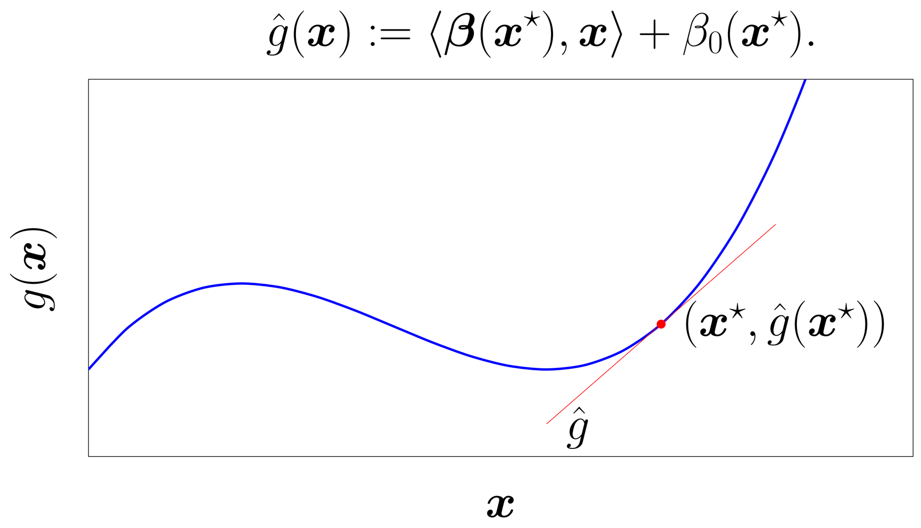
where ψ_{con} is the *convex conjugate* of ψ.

- Now we can $h \circ r$, but how to get $\nabla(h \circ r)$?

Local linear (LL) regression for gradient est.:

Approximate function g at $\boldsymbol{x}^\star$ by a linear model:

$$\hat{g}(\boldsymbol{x}) := \langle \boldsymbol{\beta}(\boldsymbol{x}^\star), \boldsymbol{x} \rangle + \beta_0(\boldsymbol{x}^\star).$$



$\boldsymbol{\beta}(\boldsymbol{x}^\star) \approx \nabla g(\boldsymbol{x}^\star)$ as the gradient of a function is the "slope" of its best local linear fit.

LL interpolation:

- Parameterize the function d in (1) using a linear model $d_{\boldsymbol{w},b}(\boldsymbol{x}) := \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b$.
- Localize (1) at a fixed point $\boldsymbol{x}^\star$ using a kernel $k^\star_\sigma$.

$$(\boldsymbol{w}(\boldsymbol{x}^\star), b(\boldsymbol{x}^\star)) = \underset{\boldsymbol{w} \in \mathbb{R}^d, b \in \mathbb{R}}{\mathrm{argmax}}\, \ell(\boldsymbol{w}, b; \boldsymbol{x}^\star),$$

$$\ell(\boldsymbol{w}, b; \boldsymbol{x}^\star) := \widehat{\mathbb{E}}_p[k^\star_\sigma d_{\boldsymbol{w},b}(\boldsymbol{x})] - \widehat{\mathbb{E}}_q[k^\star_\sigma \psi_{\mathrm{con}}(d_{\boldsymbol{w},b}(\boldsymbol{x}))]$$

5. Estimation Consistency

The consistency of the interpolation depends on the "curvature" of the velocity field:

Assumption 5.1. The velocity fields is well-behaved, i.e.,

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \|\nabla^2(h \circ r)(\boldsymbol{x})\| \le \kappa.$$

and the boundedness of the second order derivative of ψ''_{con}.

Assumption 5.2. $\|\psi''_{\mathrm{con}}\|_\infty \le C_{\psi''_{\mathrm{con}}}$.

Define: $b^* := h(r(\boldsymbol{x}^\star)) - \langle \nabla(h \circ r)(\boldsymbol{x}^\star), \boldsymbol{x}^\star \rangle$.

Theorem 5.3. *Suppose Assumption 5.1 and 5.2 holds and other mild assumptions on the kernel* k_σ *hold, if there exist strictly positive constants* $W, B, \Lambda_{\min}$ *such that,*

$$\|\nabla(h \circ r)(\boldsymbol{x}^\star)\| \le W, \quad |b^*| \le B$$

and for all $\boldsymbol{w} \in \{\boldsymbol{w} | \|\boldsymbol{w}\| < 2W\}$ *and* $b \in \{b | |b| < 2B\}$,

$$\lambda_{\min}\left\{\widehat{\mathbb{E}}_q\left[k^\star_\sigma \nabla^2_{[\boldsymbol{w},b]} \psi_{\mathrm{con}}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b)\right]\right\} \ge \sigma^d \Lambda_{\min},$$

holds with h.p.. Then for all $0 < \sigma < \sigma_0, n > N$,

$$\|\boldsymbol{w}(\boldsymbol{x}^\star) - \nabla(h \circ r)(\boldsymbol{x}^\star)\| \le \frac{\frac{K}{\sqrt{n\sigma^d}} + \kappa C_k C_{\psi''_{\mathrm{con}}} \sigma^2}{\Lambda_{\min}},$$

holds with high probability.

6. Experiments

Transport distribution by minimizing $\mathrm{KL}[q_t, p]$

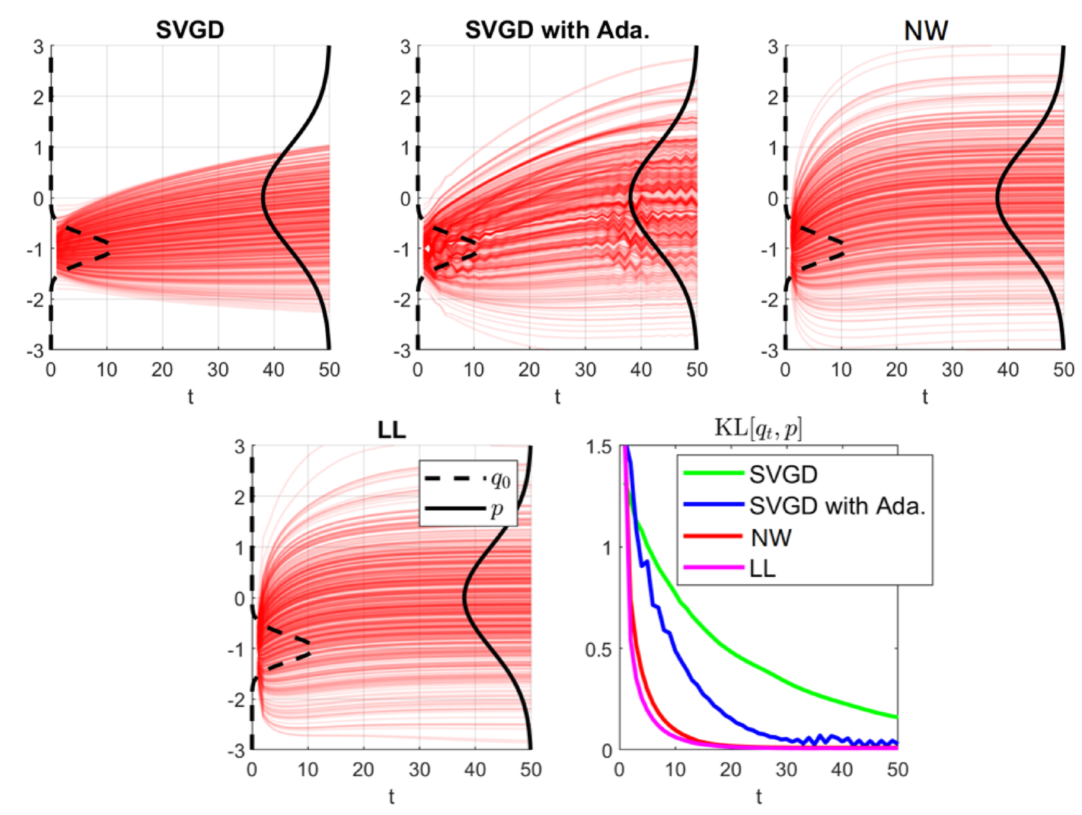


Figure 1: Particle Trajectories of SVGD, SVGD with AdaGrad, NW, LL. Approximated $\mathrm{KL}[q_t,p]$ with different methods.

Domain adaptation by minimizing $\mathrm{KL}[q_t, p]$

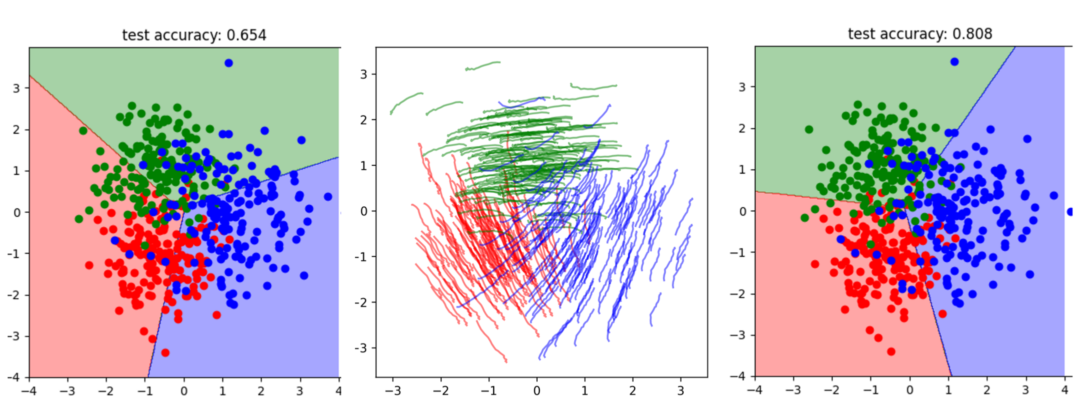


Figure 2: Left: the source classifier (represented by colored areas) misclassifies many testing points (colored dots). Middle: WGF moves particles to align the source and target samples. Lines are trajectories of sample movements in each class. Right: the retrained classifier on the transported source samples gives a much better prediction.

References

[1] Y. Gao, Y. Jiao, Y. Wang, Y. Wang, C. Yang, and S. Zhang. Deep generative learning via variational gradient flow. In *International Conference on Machine Learning (ICML 2019)*, pages 2093–2101, 2019.

[2] M. Yi, Z. Zhu, and S. Liu. Monoflow: Rethinking divergence gans via the perspective of wasserstein gradient flows. In *International Conference on Machine Learning (ICML 2023)*, pages 39984–40000, 2023.