A Wasserstein Minimum Velocity Approach to Learning Unnormalized Models

Ziyu Wang, Shuyu Cheng, Yueru Li, Jun Zhu, Bo Zhang

wzy196@gmail.com; Tsinghua University

Overview

- EBM: interesting on its own; as score estimator for implicit VI, mutual information estimation, etc
- Optimizing the learning objective (score matching) is nontrivial since it involves second-order derivatives
- We present scalable approximations to a family of learning objectives including score matching, by connecting them to Wasserstein gradient flows
- We derive a CD-1-like approximation to these objectives
- Applications: Riemannian score matching for implicit VAEs and WAEs with manifold-valued prior

EBMs and Score Matching

- EBM:
$$q(x;\theta) := \frac{1}{Z(\theta)}\exp(-\mathcal{E}(x;\theta)),$$
$\mathcal{E}$ parameterized by e.g. NNs.
- MLE intractable: $\nabla_\theta \log q(x;\theta)$ involves $\nabla_\theta \log Z = \mathbb{E}_{q(x;\theta)}(\nabla_\theta \log \mathcal{E})$.
- Score estimation: match
$$D_{Fisher}(p|q) = \mathbb{E}_p\|\nabla \log p - \nabla \log q\|^2$$
which does not depend on Z.
- Hyvarinen (2005):
$$D_{Fisher}(p|q) = \mathbb{E}_p\left[-\Delta\mathcal{E} + \frac{1}{2}\|\nabla\mathcal{E}\|^2\right] + \underbrace{\text{const}}_{\text{only depends on } p}$$
Estimation possible but expensive (involves $\Delta\mathcal{E}$).

Background: Manifold and Flows

- Differential and gradient on *general manifolds*: for $f:\mathcal{M}\to\mathbb{R}$,
$$(df)_{c(t_0)}\left(\frac{dc}{dt}\bigg|_{t_0}\right) = \frac{d}{dt}f(c(t))\bigg|_{t_0},\ \langle \mathrm{grad}_p f, v\rangle = (df)_p$$
for any $c:[0,a]\to\mathcal{M}, p\in\mathcal{M}, v\in\mathcal{T}_p\mathcal{M}$.
- The *2-Wasserstein* space $\mathcal{P}(\mathcal{X})$:
  - Tangent vector $v\in\mathcal{T}_p\mathcal{P}(\mathcal{X}) \Leftrightarrow$ vector field v on $\mathcal{X}$
  - $\langle v, v'\rangle_p = \mathbb{E}_{p(x)}\langle v(x), v'(x)\rangle_x$
  - $(\mathrm{grad}_p \mathrm{KL}_q)(u) = \mathrm{grad}_u \log\frac{p(u)}{q(u)}$
- Gradient flow of $\mathcal{F}:\mathcal{M}\to\mathbb{R}$: $\frac{dc}{dt} = -\mathrm{grad}_p\mathcal{F}$.

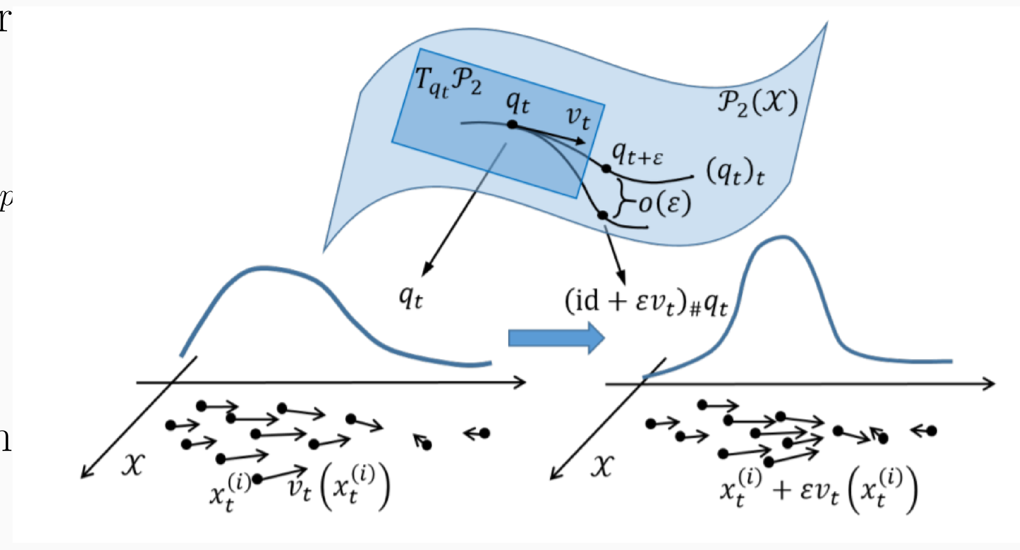


Image from Liu et al (2019)

Background: Sampling Dynamics

Common samplers can be interpreted as simulating the gradient flow of $\mathrm{KL}_p: q\mapsto \mathrm{KL}(q\|p)$, in different spaces of probability measures:

- $\mathcal{P}(\mathcal{X}), \mathcal{X}=\mathbb{R}^d$: Langevin dynamics
$$dx := \mathrm{grad}_x \log p(x)dt + \sqrt{2}dB.$$
- $\mathcal{P}(\mathcal{X}), \mathcal{X}$ general manifold: Riemannian Langevin dynamics
$$dx := V(x)dt + \sqrt{2G^{-1}(x)}dB, \quad V^i(x) := g^{ij}\partial_j\left(\log p(x) - \frac{\log|G(x)|}{2}\right) + \partial_j g^{ij}$$
(p is the density w.r.t. the Hausdorff measure here)
- The $\mathcal{H}$-Wasserstein space: Stein Variational Gradient Descent
- Other examples: birth-death LD, stochastic particle optimization

Score Matching as *Minimum Velocity Learning*

$$D_{Fisher}(p|q) = \|\mathrm{grad}_p \mathrm{KL}_q\|^2$$

where $\|\cdot\|$ is in defined in $\mathcal{P}(\mathcal{X})$.
Interpretation: the **initial velocity** of the Wasserstein gradient flow of KL_q connecting p and q.
Wasserstein MVL: switch from $\mathcal{P}(\mathcal{X})$ to other spaces of probability measures.

Approximation using the MVL Formulation

Let $\mathcal{F}[p] := -\mathbb{E}_p\mathcal{E}, \mathcal{H}[p] := \mathbb{E}_p \log p$ so $\mathrm{KL}_q = \mathcal{H}-\mathcal{F}$.

$$\|\mathrm{grad}_p \mathrm{KL}_q\|^2 = \underbrace{\|\mathrm{grad}_p \mathcal{H}\|^2}_{\text{const}} - 2\langle \mathrm{grad}_p\mathcal{F}, \mathrm{grad}_p \mathrm{KL}_{q^{1/2}}\rangle$$

$$-\langle \mathrm{grad}_p\mathcal{F}, \mathrm{grad}_p \mathrm{KL}_{q^{1/2}}\rangle = (d\mathcal{F})_p(-\mathrm{grad}_p \mathrm{KL}_{q^{1/2}}) = \lim_{\epsilon\to 0}\frac{\mathbb{E}_{\tilde{p}_t}\log q_\theta - \mathbb{E}_p \log q_\theta}{\epsilon}$$

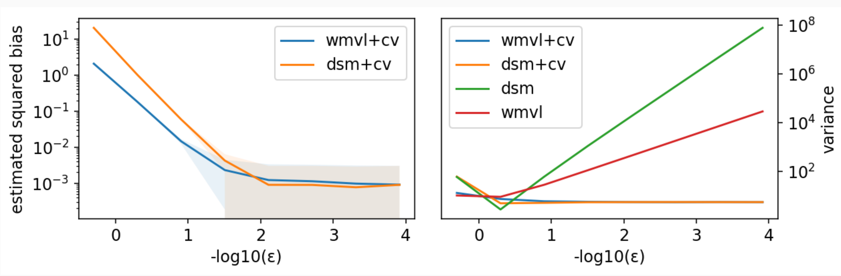
where $\{\tilde{p}_t\}$ is the gradient flow of $\mathrm{KL}_{q^{1/2}}$, and $q^{1/2}\propto \exp(-\mathcal{E}/2)$.
$\Longrightarrow$ Algorithm 0:

1. Simulate $\mathrm{KL}_{q^{1/2}}$ using the corresponding sampling dynamics, for a time of ϵ
2. Return the difference in energy, divided by ϵ

Variance Reduction

Problem: When the sampling dynamics consists of Ito diffusion, the mini-batch estimator
$$\frac{\mathcal{E}(x^+) - \mathcal{E}(x_\epsilon^-)}{\epsilon}$$
has infinite variance as $\epsilon\to 0$.
Solution: subtract the diffusion part from the estimator. For LD the resulted estimator is
$$\frac{1}{\epsilon}\Big(\mathcal{E}(x^+) - (\mathcal{E}(\underbrace{x^+ + \epsilon\log q_{1/2} + \sqrt{2\epsilon}Z}_{x_\epsilon^-}) \underbrace{-\langle\sqrt{2\epsilon}Z, \nabla_{x^+}\mathcal{E}\rangle}_{\text{control variate}})\Big)$$
Side product: the same problem exists in CD-1 for score matching (Hyvarinen (2007)) and denoising score matching; they can be fixed similarly.

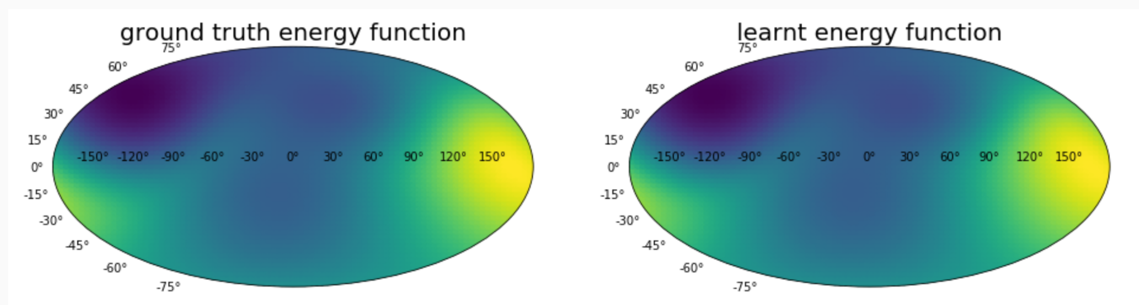
Variance-reduced objective has vanishing bias as $\epsilon\to 0$, and $O(1)$ variance regardless of ϵ.
$\Rightarrow$ Unlike previous work, we can use arbitrarily small ϵ in practice.



Example: Score Matching on Manifolds

- The *Riemannian score matching objective*: same form as D_{Fisher}, but with different metric $\|\cdot\|$.
- Also a MVL objective, with different sampling dynamics (Riemannian LD).
- Final approximator:
$$L_{\text{mvl-rld}} = \frac{2}{\epsilon}\Big(\mathcal{E}(y^-;\theta) - \mathcal{E}(y;\theta) - \underbrace{\sqrt{2\epsilon}\partial_i\mathcal{E}(y)z^i}_{\text{control variate}}\Big), \quad \text{where}$$
$$(y^-)^i = y^i + \epsilon\left(-g^{ij}\partial_j\frac{\mathcal{E}(y;\theta)+\log|G(y)|}{2} + \partial_k g^{ik}\right) + \sqrt{2\epsilon}z^i,$$
is a sample from Riemannian LD, and $z\sim\mathcal{N}(0, G^{-1}(y))$.

Simulation: learning mixture of von-Mises-Fisher on S^2.



VAE and WAE with hyperspherical prior

VAE	$n_z=8$		$n_z=32$	
NLL	Euc.	Sph.	Euc.	Sph.
Explicit	96.47	95.38	90.11	91.16
Implicit	95.71	**94.99**	90.17	**88.63**

WAE	$n_z=8$	
FID	Euc.	Sph.
GAN	25.48	20.40
MVL (Ours)	21.95	**19.13**

Related Work

Unified under our framework (and enhanced):

- CD-1 for score matching (Hyvarinen (2007)): a similar approximator for the *gradient* of the score matching objective wrt θ. Suffers from the infinite variance problem above.
- CD-1 for KSD (Liu and Wang, 2017): a similar approximator for the gradient of KSD using SVGD.

(Movellan, 2007, unpublished): score matching as minimizing the "probability velocity field" in data space.
Other unifying perspectives (that do not lead to scalable approximations): Minimum Probability Flow, Minimum Stein discrepancy estimator
Score matching: scalable approximator (Song et al (UAI 2019)), another connection to diffusion (Lyu (UAI 2009))
Our contribution: generalized derivation using WGF; practical implementation with control variate, and estimator for the original objective instead of its gradient