Sampling with constraints

Xin T Tong

BIRS workshop

Joint work with Qiang Liu, Xingchao Liu, Ruqi Zhang (UT Austin)

Friday 9$^{\text{th}}$ September, 2022

List of content

- Constrainted sampling
- Review: KL gradient flow without constraint
- Moment Constraints
- Level set constraints
- *Sampling with Trustworthy Constraints: A Variational Gradient Framework* NeurIPS 2021.
- *Sampling in Constrained Domains with Orthogonal-Space Variational Gradient Descent* Under review

Bayesian machine learning

Standard Bayesian problem:

$$\text{Sample} \quad (\ ) \propto p_0(\ ) \exp(\ \ell(\ ))$$

Moment constrained Bayesian problem:

$$\text{Sample } q \qquad \text{s.t. } \mathsf{E}_q[g(\ )]$$

Equality constraint

$$\text{Sample } q \qquad \text{s.t. } g(x) = 0 \text{ for } q\text{-a.s.} x$$

Constraints

Type of constraint functions

- Agnostic learning: $g(\ ) = l(\ )$
- Fairness: $g(\ ) = \mathrm{cov}(\hat{y}(x;\ );z)$
- Montonicity: $g(\ ) = [\ \ @_x\hat{y}(x;\ )]_+$
- Safety: $\mathrm{dist}(\hat{y}(x;\ );S)$

Type of questions:

- What would the solution be?
- How to obtain the distribution?
- Pareto front of l vs g

Existing fairness works: Chakraborty, Ji, Dimitrakakis ....

Review: unconstrained case

Sampling π

Markov Chain Monte Carlo (MCMC)

- Simulate a Markov Chain with being the invariant
- Fairly well understood
- Require well specified
- Iterates tend to be dependents
- MC convergence: $O(1/\ ^2)$

Variational method

- Try to push a density towards .
- Interacting particle system
- Promising on some problems.
- Understanding is much less.
- Potentially can be faster(?)

Variation formulation

Basic formulation:

- Try to minimize KL(q_t;)
- Suppose we have samples from a density q_t.
- We can estimate $E_{q_t}[f]$ for any f.
- Try to push each point x in q_t with $_t(x)$
- Continuity equation: $\frac{d}{dt}q_t =$ r (q_t)
- What is the optimal for reducing KL?
- Solve sampling by optimization methods.

Review: unconstrained Gradient flow

Rate of decay

$$\frac{d}{dt}\mathrm{KL}(q_t; \quad) = \mathsf{E}_{q_t}[h\, r \ \log \qquad r \ \log q_t; \quad i]$$

Try to maximize, write $r \ \log \quad = s$

$$\max_{2H} \mathsf{E}_{q_t}[hs \qquad s_{q_t}; \quad i] \qquad \frac{1}{2}k \ \ k_H^2$$

Langevin Dynamics

If we use $H = L_{q_t}^2$

- We obtain $_t = s \quad s_{q_t}$.
- But how to get s_{q_t}?
- Stein operator $A \ = (s \ + \ r \ )$

$$\frac{d}{dt} q_t = \quad r \quad (\ q_t) = \quad r \quad (s \ q_t) + \Delta q_t = \ r \quad (A \ q_t)$$

- Fokker–Plank equation (FPE) of Langevin dynamics (LD)[Jordan, Kinderlehrer, and Otto 1998]
- Algorithmic implementation (ULA):

$$_{t+1} = \ _t + \ s \ (\ _t) + \ \frac{\sqrt{}}{2} \ _{t+1}.$$

- Can be seen as an MCMC as well.

Convergence

Use

$$\frac{d}{dt}\mathrm{KL}(q_t;\ ) = \quad \mathsf{E}_{q_t} ks \quad s_{q_t} k^2$$

- $\int_0^T \mathsf{E}_{q_t} ks \quad s_{q_t} k^2 \quad \mathrm{KL}(q_0;\ )$
- Fisher divergence $\min_t \ _T \mathsf{E}_{q_t} ks \quad s_{q_t} k^2 = O(1{=}T)$
- If the log-Sobolev inequality (LSI) holds, $ks \quad s_{q_t} k^2 \quad c\mathrm{KL}(q_t;\ )$, $\mathrm{KL}(q_t;\ ) = O(\exp(\ \ ct))$.
- Can be inherited by ULA (Vempala and Wibisono 2019)

Use H =RKHS with kernel k,

- $(x) = \int (s\ (y) \quad r\ \log q_t(y))k(x;y)q_t(y)dy$
- A kernel embedding of A into H
- Limit point meets Stein equation $\mathsf{E}_q\ A\ f = 0$ for $f\ 2\ H$.
- $(x) = \int s\ (y)k(x;y)q_t(y)dy + \int r_y k(x;y)q_t(y)dy$
- Replace q_t with samples from q_t.

$$_{i;t+1} = \ _{i;t} + \frac{}{n} \sum_{j=1}^{n} k(\ _{i;t};\ _{j;t})\, r\ _{j;t} \log\ (\ _{j;t}) + r\ _{j;t} k(\ _{i;t};\ _{j;t}):$$

- Deterministic after initialization.
- Stein Variational Gradient Descent (SVGD) [Liu and Wang 2016]

Convergence

Use

$$\frac{d}{dt}\mathrm{KL}(q_t; \quad) = \quad ks \qquad s_{q_t}k_k^2$$

$$:= \qquad q_t(x)q_t(y)k(x;y)(s \qquad s_{q_t})(x)^T(s \qquad s_{q_t})(y)$$

- Kernel Stein divergence $\min_t \ _T \mathsf{E}_{q_t} ks \qquad s_{q_t}k_k^2 = O(1{=}T)$
- Is there LIS $ks \qquad s_{q_t}k_k^2 \quad c\mathrm{KL}(q_t; \quad)$?
- Actually not corret in general (Gorham and Mackey 2017)

Moment constrained

Formulation

Solve

$$\min_{q} \mathrm{KL}(q; \quad); \quad s{:}t{:} \quad \mathsf{E}_q[g] \quad 0{:}$$

- Ignore the possibility $\mathsf{E}\ [g] \quad 0$; where is the solution.
- Solution: $q =$ / $\exp(\quad g)$ and $\mathsf{E}\ \ [g] = 0$
- Chicken: Checking $\mathsf{E}\ \ [g] = 0$ requires samples from
- Egg: sampling from requires
- Double loop: MCMC or variational, feasible but expensive

Primal dual gradient flow (PDGF)

Reformulate as

$$\min_{q} \max_{0} fL(q; \quad) = \mathrm{KL}(q \; jj \quad) + \quad \mathsf{E}_q[g]g :$$

Gradient ascent on :

$$\frac{d}{dt} \; _t = [\; \mathsf{E}_{q_t}[g]] \; _{t;+}$$

When $H = L^2$, gradient descent on q via :

$$_t = r(\log \quad _t \quad \log q_t) = s \qquad _t r g \quad s_q$$

When H =RKHS, gradient descent on q via :

$$_t(x) = \int (s \; (y) \qquad _t r g(y) + r _y)k(x; y)q_t(y)dy$$

Assume

Convergence

Theorem

Suppose

$$ks_{q_t} \quad s \quad k^2_{q_t} \quad c_1(\mathsf{E}_{q_t}[g] \quad \mathsf{E} \quad [g])^2$$

LD-PDGF finds solutions $ks_{q_t} \quad s \quad k^2_{q_t} = O(1=T)$. *If* g *is convex, satisfies log Sobolev, then linear convergence for* $KL(q_t; \quad)$

For SVGD, $k \quad k^2_{q_t}$ is replaced by kernel Stein discrepancy.

Theorem

Suppose

$$ks_{q_t} \quad s \quad k^2_k \quad c_1(\mathsf{E}_{q_t}[g] \quad \mathsf{E} \quad [g])^2$$

LD-PDGF finds solutions $ks_{q_t} \quad s \quad k^2_k = O(1=T)$.

Constraint Controlled gradient flow (CCGF)

Try to solve

$$\max \mathsf{E}_{q_t}[hs \quad s_{q_t}; \quad i] \quad \frac{1}{2}k \quad k_H^2; \quad s{:}t{:}\frac{d}{dt}\mathsf{E}_{q_t}g = \mathsf{E}_{q_t} \quad ^T r\, g \qquad \mathsf{E}_{q_t}[g]$$

Solve quadratic opt.

$$\min_0 \max \mathsf{E}_{q_t}[hs \qquad s_{q_t}; \quad i] \quad \frac{1}{2}k \quad k_H^2 + \quad (\mathsf{E}_{q_t} \quad ^T r\, g + \quad \mathsf{E}_{q_t}[g])$$

We have $\quad _t = s \qquad _t r\, g \quad s_q$ (LD case)

$$_t = \max \quad \frac{\mathsf{E}_{q_t}[g] + hs \qquad s_{q_t};\ r\, g i_{q_t}}{k\, r\, g k_{q_t}^2}; 0$$

Or $\quad _t(x) = \int^{\mathsf{R}} (s \qquad _t r\, g \quad s_q)(y)k(x;y)q_t(y)dy$ (SVGD case).

$$_t = \max \quad \frac{\mathsf{E}_{q_t}[g] + hs \qquad s_{q_t};\ r\, g i_k}{k\, r\, g k_k^2}; 0$$

Convergence

Theorem

Suppose $_t$ *is bounded by a constant, LD-CCGF finds solutions* ks_{q_t} s $k^2_{q_t} = O(1=T)$. *If* g *is convex,* *satisfies log Sobolev, then linear convergence for* $KL(q_t;$ $)$

For SVGD, k $k^2_{q_t}$ is replaced by kernel Stein discrepancy.

Theorem

Suppose $_t$ *is bounded by a constant, SVGD-CCGF finds solutions* ks_{q_t} s $k^2_k = O(1=T)$.

Algorithm 3 Primal-Dual Method

Initialize the particles $\{\theta_{i,0}\}_{i=1}^n$ and λ_0.
for iteration t **do**
If `Langevin`, update $\theta_{i,t+1} = \theta_{i,t} + h(\nabla \log p_0^*(\theta_{i,t}) - \lambda_t \nabla g(\theta_{i,t})) + \sqrt{2h}\xi_{i,t}$.
If `SVGD`, update

$$\theta_{i,t+1} = \theta_{i,t} + \frac{h}{n}\sum_{j=1}^n [(\nabla \log p_0^*(\theta_{j,t}) - \lambda_t \nabla g(\theta_{j,t}))k_t(\theta_{j,t}, \theta_{i,t})] + \nabla_{\theta_{j,t}} k_t(\theta_{j,t}, \theta_{i,t}).$$

Update λ_t by $\lambda_{t+1} = \max(\lambda_t + \frac{\tilde{h}}{n}\sum_{i=1}^n [g(\theta_{i,t+1})],\ 0)$.
end for

Algorithm 4 Constraint Controlled Method

Initialize the particles $\{\theta_{i,0}\}_{i=1}^n$.
for iteration t **do**
If `Langevin`, update

$$\lambda_t = \max\left(\frac{\sum_{j=1}^n \alpha g(\theta_{j,t}) + [(\nabla \log p_0^*(\theta_{j,t}))^\top \nabla g(\theta_{j,t}) + \nabla^\top \nabla g(\theta_{j,t})]}{\sum_{j=1}^n [\|\nabla g(\theta_{j,t})\|^2]},\ 0\right),$$

update $\theta_{i,t+1} = \theta_{i,t} + h(\nabla \log p_0^*(\theta_{i,t}) - \lambda_t \nabla g(\theta_{i,t})) + \sqrt{2h}\xi_{i,t}$.
If `SVGD`, update

$$\lambda_t = \max\left(\frac{\sum_{i,j=1}^n \alpha g(\theta_{i,t}) + [\nabla g(\theta_{j,t})^\top (\nabla \log p_0^*(\theta_{i,t}) + \nabla_{\theta_{i,t}})k_t(\theta_{i,t}, \theta_{j,t}])]}{\sum_{i,j=1}^n [\nabla g(\theta_{i,t})^\top \nabla g(\theta_{j,t}) k_t(\theta_{i,t}, \theta_{j,t})]},\ 0\right),$$

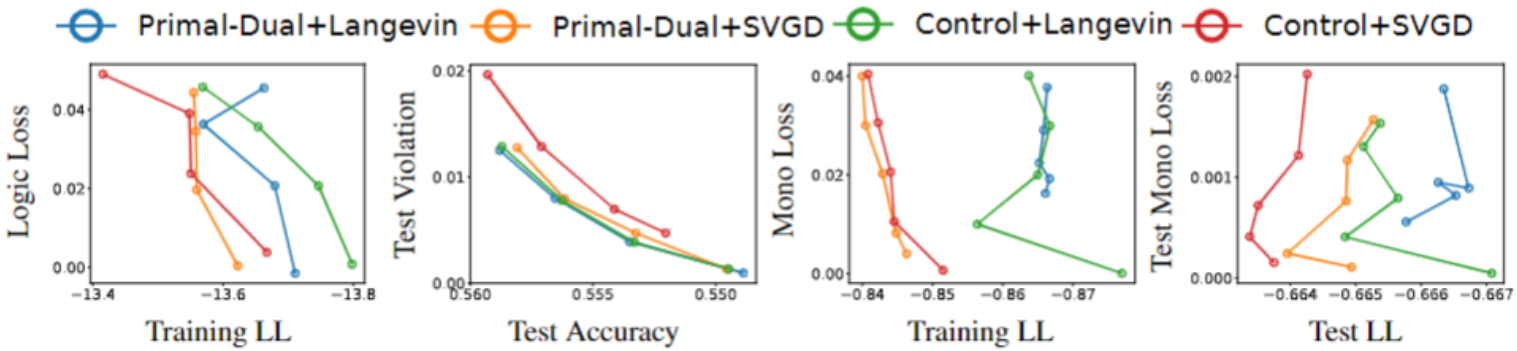
update

$$\theta_{i,t+1} = \theta_{i,t} + \frac{h}{n}\sum_{j=1}^n [(\nabla \log p^*(\theta_{j,t}) - \lambda_t \nabla g(\theta_{j,t}))k_t(\theta_{j,t}, \theta_{i,t}) + \nabla_{\theta_{j,t}} k_t(\theta_{j,t}, \theta_{i,t})].$$

end for

Numerical results

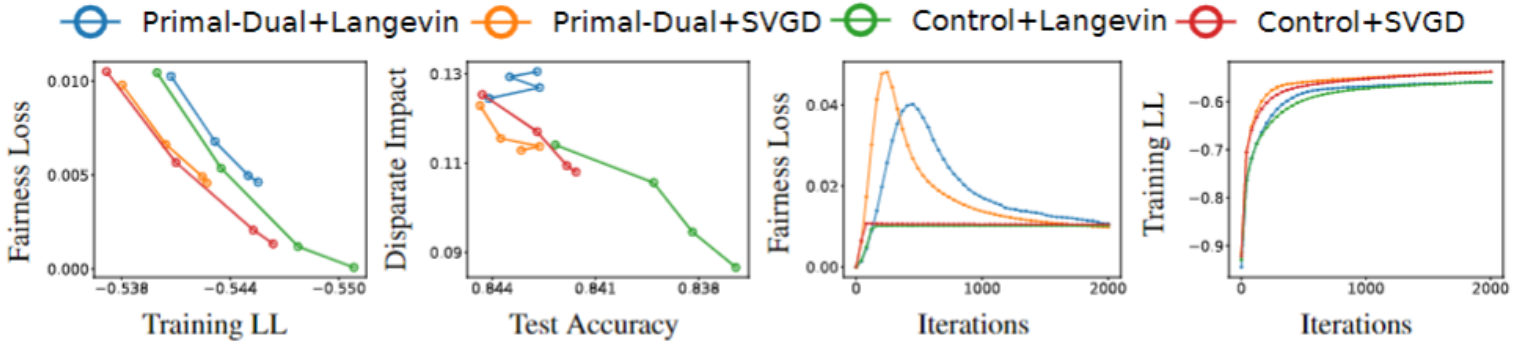
Logic and Montonicity constrained logistic regression.



(a) Logic Constraint on Loan Application (b) Monotonic Constraint on COMPAS

Numerical results

Fairness constrained Neural Network



(a) Trade-off Curve with Different ϵ (b) Training Curve

Equality constrained

Overview

Formulation of problem

- Minimize KL(q;) so that q is supported on $G_0 = \{x : g(x) = 0\}$
- Ill-posed: q is singular w.r.t. .
- Try to sample the conditional measaure $_0( ) = [\ jg = 0]$.
- Haussdorf density $(x)=j\, r\, g(x)j$ on G_0.

Sampling on manifolds

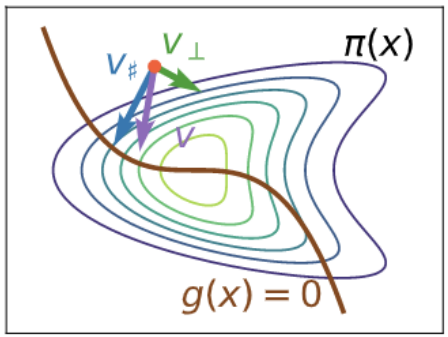
- Several existing MCMC (Girolami, Brubaker, Lelievre...)
- Assume MCMC start and stay on G_0
- Often require explicit knowledge of G_0 (parameterization, geodesic, projection)
- Not so friendly for large scale ML models.

Deriving algorithm

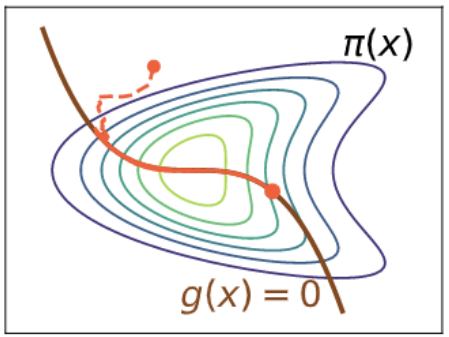
Try to solve

$$\max \mathsf{E}_{q_t}[hs \qquad s_{q_t}; vi] \qquad \frac{1}{2}kvk_H^2;$$

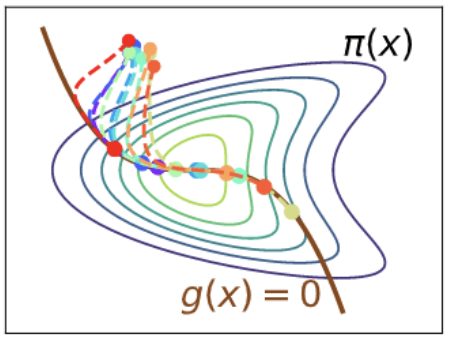
$$s{:}t{:}\frac{d}{dt}g(x_t) = v^T(x)\, r\, g(x) = \qquad (g(x))$$



(a) O-Gradient



(b) O-Langevin



(c) O-SVGD

Deriving algorithm

Along $r\,g$

- Use $\quad(z) = \quad \text{sign}(z)jzj^{1+}$
- The component along $r\,g$: $v_J = \frac{(g(x))\,r\,g(x)}{k\,r\,g(x)k^2}$

Along the orthogonal direction:

- Projection: $D = I \quad \frac{r\,g\,r\,g^T}{k\,r\,gk^2}$
- $v_? = Du$, $\max_u \mathsf{E}_{q_t}[(D(s \quad s_{q_t}))^T u] \quad \frac{1}{2}kDuk_H^2$:
- LD: $v_? = D(s \quad s_{q_t})$
- SVGD:

$$\begin{aligned} v_?(x) &= \int \quad D(x)k(x;y)D(y)(s \quad s_{q_t})(y)q_t(y)dy \\ &= \int \quad k_?(x;y)(s \quad s_{q_t})(y)q_t(y)dy \end{aligned}$$

Implementation

- LD: $v_? = D(s \quad s_{q_t})$ cannot be implemented direclty by $dx_t = (v_J(x_t) + D(x_t)s\ (x_t))dt + \sqrt{2}D(x_t)dW_t$:
- Consider adding a correction drift r

Theorem

When $r(x) = r \quad D(x)$,

$$dx_t = (v_J(x_t) + D(x_t)s\ (x_t))dt + \sqrt{2}D(x_t)dW_t \qquad (1)$$

its FPE mathches the orthogonal density flow. Moreover, i) the value $g(x_t)$ *has deterministic decay* $\frac{d}{dt}g(x_t) = \quad (x_t)$*; ii) for any* f *with* $r\,f\,?\,r\,g = 0$*, the generator of* x_t *matches the Langevin ones* $Lf(x) = r\,f^{>}(x)s\ (x) + \Delta f(x)$.

Convergence

Define orthogonal space (OS) Fisher divergence

$$F_?(q;\ ) = kD(s\quad s_q)k_q^2 \text{ or } kD(s\quad s_q)k_k^2$$

Theorem

Suppose $g(x)$ *is bounded for the initial distribution, and it's "regular",* $KL(q_0;\ ) < 1$ *, then*
$M_T = \max fg(x); x\quad q_T g = O(T^{\ \frac{1}{}})$*; also convergence in OS-Fisher* $\min_t\ _T F_?(q_t;\ ) = O(\log T=T)$.

But is OS-Fisher useful?

Simpler formulation

The distribution $\Pi_z = \quad (\ jg(x) = z)$ is too abstract.

Theorem

Suppose $g]$ *has Lipschitz density. Then the weak limit of* $_{;z}(x) \ / \ (x)\exp(\ \frac{1}{2}(g(x) \quad z)^2)$ *as* $\ ! \ 0$ *concentrates on* $G_z = fx : g(x) = zg$ *and is a version of* $_z$*. Moreover,*

$$\mathsf{E}_{\Pi_z}[A \quad] = 0; \quad 8 \ ? \ r \ g:$$

- This gives a Stein equation $\mathsf{E}_q[A \quad] = 0$
- The tangent bundle of G_z is a subset of $\quad ? \ r \ g$
- $\mathsf{E}_q[A \quad] \quad \overline{F_?(q; \quad)}$ when $k \ k \ = 1$.
- $\mathsf{E}_q[A \quad]$ or $F_?(q; \quad)$ do not require q being on G_z
- This only check the OS directions.
- Checking how far is q away from G_z is easy.

Fisher-¿TV difference

Theorem

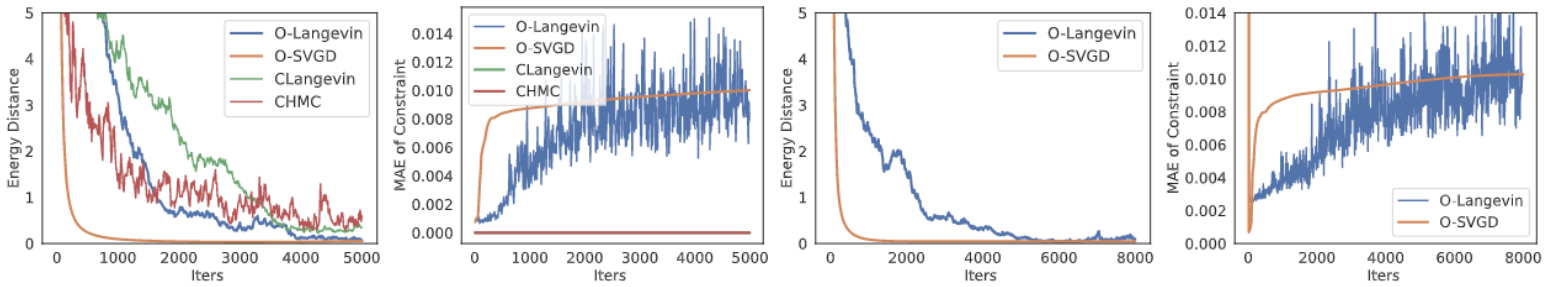
Suppose that Π_z *satisfies -Poincare Inequality for* jzj *, and* q *is supported on* $fx : jg(x)j \quad g$*. Then for any function* f *such that* $jfj \quad 1$*, the following holds*

$$jE_q[f] \quad E_{\Pi_0}[f]j \quad \sqrt{F_?(q; \ )} + \max_{jzj} jE_{\Pi_z}[f] \quad E_{\Pi_0}[f]j:$$

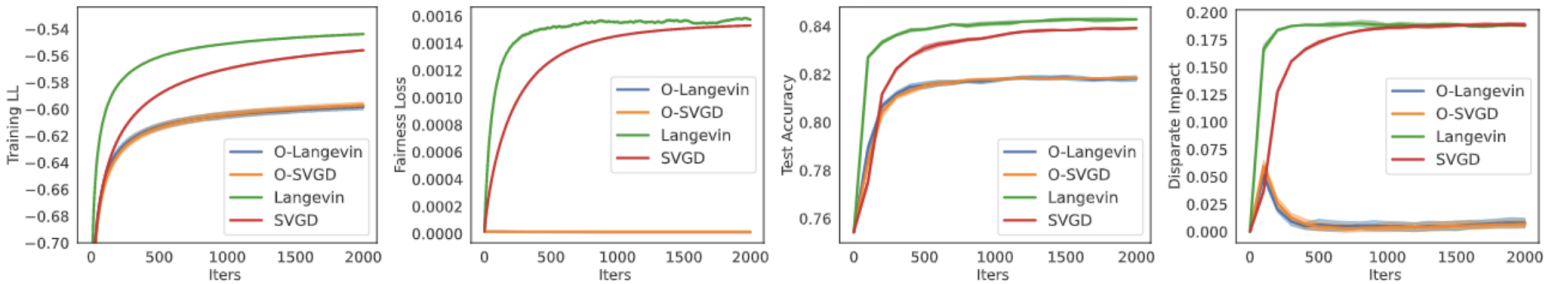
- Decomposition of mean difference/TV
- Only in L^2 case
- Poincare inequality with Euclidean-inheriant distance
- Can be used for q supported on R^d.

Numerical examples

Toy example (Intialized on/off manifold)



Income prediction



Agonostic Bayesian Image classification

	Test Error (↓)	ECE (↓)	AUROC (↑)
SGLD	15.00	2.21	89.41
Tempered SGLD	4.73	0.83	97.63
O-Langevin	**4.46**	0.87	**98.68**
SVGD	6.11	0.93	93.55
O-SVGD	4.92	**0.77**	94.69