Sampling with constraints

Xin T Tong
BIRS workshop
Joint work with Qiang Liu, Xingchao Liu, Ruqi Zhang (UT Austin)

Friday 9th 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

Standard Bayesian problem:

$$\text{Sample } \pi(\theta) \propto p_0(\theta) \exp(-l(\theta))$$

Moment constrained Bayesian problem:

$$\text{Sample } q \approx \pi \text{ s.t. } \mathbb{E}_q[g(\theta)] \leq \epsilon$$

Equality constraint

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

Type of constraint functions

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

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 π being the invariant
- Fairly well understood
- Require well specified π
- Iterates tend to be dependents
- MC convergence: $O(1/\epsilon^2)$

Variational method

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

Variation formulation

Basic formulation:

- Try to minimize $\mathrm{KL}(q_t, \pi)$
- 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 $\phi_t(x)$
- Continuity equation: $\frac{d}{dt} q_t = -\nabla \cdot (\phi q_t)$
- What is the optimal ϕ for reducing KL?
- Solve sampling by optimization methods.

Review: unconstrained Gradient flow

Rate of decay

$$-\frac{d}{dt}\mathrm{KL}(q_t, \pi) = \mathbb{E}_{q_t}[\langle \nabla \log \pi - \nabla \log q_t, \phi \rangle]$$

Try to maximize, write $\nabla \log \pi = s_\pi$

$$\max_{\phi \in \mathcal{H}} \mathbb{E}_{q_t}[\langle s_\pi - s_{q_t}, \phi \rangle] - \frac{1}{2}\|\phi\|_{\mathcal{H}}^2$$

If we use $\mathcal{H} = L^2_{q_t}$

- We obtain $\phi_t = s_\pi - s_{q_t}$.
- But how to get s_{q_t}?
- Stein operator $A_\pi = (s_\pi + \nabla)$

$$\frac{d}{dt} q_t = -\nabla \cdot (\phi q_t) = -\nabla \cdot (s_\pi q_t) + \Delta q_t = \nabla \cdot (A_\pi q_t)$$

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

$$\theta_{t+1} = \theta_t + \eta s_\pi(\theta_t) + \sqrt{2\eta}\xi_{t+1}.$$

- Can be seen as an MCMC as well.

Use

$$\frac{d}{dt}\mathrm{KL}(q_t,\pi) = -\mathbb{E}_{q_t}\|s_\pi - s_{q_t}\|^2$$

- $\int_0^T \mathbb{E}_{q_t}\|s_\pi - s_{q_t}\|^2 \le \mathrm{KL}(q_0,\pi)$
- Fisher divergence $\min_{t\le T}\mathbb{E}_{q_t}\|s_\pi - s_{q_t}\|^2 = O(1/T)$
- If the log-Sobolev inequality (LSI) holds, $\|s_\pi - s_{q_t}\|^2 \ge c\mathrm{KL}(q_t,\pi)$, $\mathrm{KL}(q_t,\pi) = O(\exp(-ct))$.
- Can be inherited by ULA (Vempala and Wibisono 2019)

Use $\mathcal{H}$ =RKHS with kernel k,

- $\phi(x) = \int (s_\pi(y) - \nabla \log q_t(y))k(x,y)q_t(y)dy$
- A kernel embedding of A_π into $\mathcal{H}$
- Limit point meets Stein equation $\mathbb{E}_{q^*} A_\pi f = 0$ for $f \in \mathcal{H}$.
- $\phi(x) = \int s_\pi(y)k(x,y)q_t(y)dy + \int \nabla_y k(x,y)q_t(y)dy$
- Replace q_t with samples from q_t.

$$\theta_{i,t+1} = \theta_{i,t} + \frac{\eta}{n}\sum_{j=1}^{n} k(\theta_{i,t},\theta_{j,t})\nabla_{\theta_{j,t}}\log\pi(\theta_{j,t}) + \nabla_{\theta_{j,t}} k(\theta_{i,t},\theta_{j,t}).$$

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

Use

$$\frac{d}{dt}\mathrm{KL}(q_t, \pi) = -\|s_\pi - s_{q_t}\|_k^2$$
$$:= -\int q_t(x)q_t(y)k(x,y)(s_\pi - s_{q_t})(x)^T(s_\pi - s_{q_t})(y)$$

- Kernel Stein divergence $\min_{t\le T} \mathbb{E}_{q_t}\|s_\pi - s_{q_t}\|_k^2 = O(1/T)$
- Is there LIS $\|s_\pi - s_{q_t}\|_k^2 \ge c\mathrm{KL}(q_t, \pi)$?
- Actually not corret in general (Gorham and Mackey 2017)

Moment constrained

Formulation

Solve

$$\min_q \mathrm{KL}(q, \pi), \quad s.t. \quad \mathbb{E}_q[g] \leq 0.$$

- Ignore the possibility $\mathbb{E}_\pi[g] \leq 0$, where π is the solution.
- Solution: $q = \pi_{\lambda^*} \propto \pi \exp(-\lambda^* g)$ and $\mathbb{E}_{\pi_{\lambda^*}}[g] = 0$
- Chicken: Checking $\mathbb{E}_{\pi_\lambda}[g] = 0$ requires samples from π_λ
- Egg: sampling from π_λ requires λ
- Double loop: MCMC or variational, feasible but expensive

Primal dual gradient flow (PDGF)

Reformulate as

$$\min_{q} \max_{\lambda \geq 0} \left\{ L(q, \lambda) = \mathrm{KL}(q \mid\mid \pi) + \lambda \mathbb{E}_q[g] \right\}.$$

Gradient ascent on λ:

$$\frac{d}{dt}\lambda_t = [\eta \mathbb{E}_{q_t}[g]]_{\lambda_t,+}$$

When $\mathcal{H} = L^2$, gradient descent on q via ϕ:

$$\phi_t = \nabla(\log \pi_{\lambda_t} - \log q_t) = s_\pi - \lambda_t \nabla g - s_q$$

When $\mathcal{H} =$RKHS, gradient descent on q via ϕ:

$$\phi_t(x) = \int (s_\pi(y) - \lambda_t \nabla g(y) + \nabla_y) k(x, y) q_t(y) dy$$

Assume

Convergence

Theorem

Suppose

$$\|s_{q_t} - s_{\pi_{\lambda^*}}\|_{q_t}^2 \geq c_1(\mathbb{E}_{q_t}[g] - \mathbb{E}_{\pi_{\lambda^*}}[g])^2$$

LD-PDGF finds solutions $\|s_{q_t} - s_{\pi_{\lambda^*}}\|_{q_t}^2 = O(1/T)$. *If* g *is convex,* π *satisfies log Sobolev, then linear convergence for* $KL(q_t, \pi_{\lambda^*})$

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

Theorem

Suppose

$$\|s_{q_t} - s_{\pi_{\lambda^*}}\|_{k}^2 \geq c_1(\mathbb{E}_{q_t}[g] - \mathbb{E}_{\pi_{\lambda^*}}[g])^2$$

LD-PDGF finds solutions $\|s_{q_t} - s_{\pi_{\lambda^*}}\|_{k}^2 = O(1/T)$.

Constraint Controlled gradient flow (CCGF)

Try to solve

$$\max_{\phi} \mathbb{E}_{q_t}[\langle s_\pi - s_{q_t}, \phi\rangle] - \frac{1}{2}\|\phi\|^2_{\mathcal{H}}, \quad s.t. \frac{d}{dt}\mathbb{E}_{q_t} g = \mathbb{E}_{q_t}\phi^T \nabla g \leq -\alpha \mathbb{E}_{q_t}[g]$$

Solve quadratic opt.

$$\min_{\lambda \geq 0} \max_{\phi} \mathbb{E}_{q_t}[\langle s_\pi - s_{q_t}, \phi\rangle] - \frac{1}{2}\|\phi\|^2_{\mathcal{H}} + \lambda(\mathbb{E}_{q_t}\phi^T \nabla g + \alpha \mathbb{E}_{q_t}[g])$$

We have $\phi_t = s_\pi - \lambda_t \nabla g - s_q$ (LD case)

$$\lambda_t = \max\left(\frac{\alpha \mathbb{E}_{q_t}[g] + \langle s_\pi - s_{q_t}, \nabla g\rangle_{q_t}}{\|\nabla g\|^2_{q_t}}, 0\right)$$

Or $\phi_t(x) = \int (s_\pi - \lambda_t \nabla g - s_q)(y) k(x, y) q_t(y) dy$ (SVGD case).

$$\lambda_t = \max\left(\frac{\alpha \mathbb{E}_{q_t}[g] + \langle s_\pi - s_{q_t}, \nabla g\rangle_{k}}{\|\nabla g\|^2_{k}}, 0\right)$$

Convergence

Theorem

Suppose λ_t is bounded by a constant, LD-CCGF finds solutions $\|s_{q_t} - s_{\pi_{\lambda^*}}\|^2_{q_t} = O(1/T)$. If g is convex, π satisfies log Sobolev, then linear convergence for $KL(q_t, \pi_{\lambda^*})$

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

Theorem

Suppose λ_t is bounded by a constant, SVGD-CCGF finds solutions $\|s_{q_t} - s_{\pi_{\lambda^*}}\|^2_{k} = O(1/T)$.

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

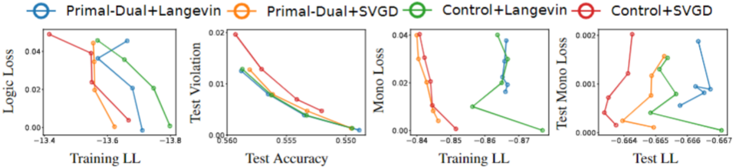
$$\lambda_t = \max\left(\frac{\sum_{i,j=1}^n \alpha g(\theta_{i,t}) + [\nabla g(\theta_{j,t})^\top (\nabla \log p_0^*(\theta_{i,t}) + \nabla_{\theta_{i,t}})k_t(\theta_{i,t}, \theta_{j,t}])]}{\sum_{i,j=1}^n [\nabla g(\theta_{i,t})^\top \nabla g(\theta_{j,t}) k_t(\theta_{i,t}, \theta_{j,t})]},\ 0\right),$$

update

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

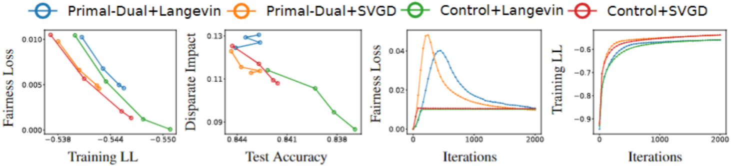
end for

Logic and Montonicity constrained logistic regression.



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

Fairness constrained Neural Network



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

Equality constrained

Overview

Formulation of problem

- Minimize $\mathrm{KL}(q, \pi)$ so that q is supported on $\mathcal{G}_0 = \{x : g(x) = 0\}$
- Ill-posed: q is singular w.r.t. π.
- Try to sample the conditional measaure $\pi_0(\cdot) = \pi[\cdot\,|g = 0]$.
- Haussdorf density $\pi(x)/|\nabla g(x)|$ on $\mathcal{G}_0$.

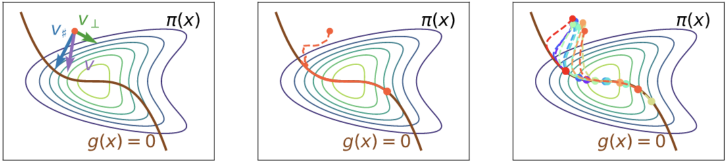
Sampling on manifolds

- Several existing MCMC (Girolami, Brubaker, Lelievre...)
- Assume MCMC start and stay on $\mathcal{G}_0$
- Often require explicit knowledge of $\mathcal{G}_0$ (parameterization, geodesic, projection)
- Not so friendly for large scale ML models.

Try to solve

$$\max_{\phi} \mathbb{E}_{q_t}[\langle s_\pi - s_{q_t}, v\rangle] - \frac{1}{2}\|v\|^2_{\mathcal{H}},$$

$$s.t. \frac{d}{dt} g(x_t) = v^T(x)\nabla g(x) = -\psi(g(x))$$



(a) O-Gradient (b) O-Langevin (c) O-SVGD

Along ∇g

- Use $\psi(z) = \alpha \text{sign}(z)|z|^{1+\beta}$
- The component along ∇g: $v_\sharp = \frac{-\psi(g(x))\nabla g(x)}{\|\nabla g(x)\|^2}$

Along the orthogonal direction:

- Projection: $D = I - \frac{\nabla g \nabla g^T}{\|\nabla g\|^2}$
- $v_\perp = Du$, $\max_u \mathbb{E}_{q_t}[(D(s_\pi - s_{q_t}))^T u] - \frac{1}{2}\|Du\|_{\mathcal{H}}^2$.
- LD: $v_\perp = D(s_\pi - s_{q_t})$
- SVGD:

$$v_\perp(x) = \int D(x)k(x,y)D(y)(s_\pi - s_{q_t})(y)q_t(y)dy$$

$$= \int k_\perp(x,y)(s_\pi - s_{q_t})(y)q_t(y)dy$$

Implementation

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

Theorem

When $r(x) = \nabla \cdot D(x)$,

$$dx_t = (v_\sharp(x_t) + D(x_t)s_\pi(x_t))dt + \sqrt{2}D(x_t)dW_t \tag{1}$$

its FPE mathches the orthogonal density flow. Moreover, i) the value $g(x_t)$ *has deterministic decay* $\frac{d}{dt}g(x_t) = -\psi(x_t)$*; ii) for any* f *with* $\nabla f \perp \nabla g = 0$*, the generator of* x_t *matches the Langevin ones* $\mathcal{L}f(x) = \nabla f^\top(x)s_\pi(x) + \Delta f(x)$.

Define orthogonal space (OS) Fisher divergence

$$F_{\perp}(q,\pi) = \|D(s_\pi - s_q)\|_q^2 \text{ or } \|D(s_\pi - s_q)\|_k^2$$

Theorem

Suppose $g(x)$ is bounded for the initial distribution, and it's "regular", $KL(q_0, \pi) < \infty$, then $M_T = \max\{g(x), x \sim q_T\} = O(T^{-\frac{1}{\beta}})$, also convergence in OS-Fisher $\min_{t \leq T} F_{\perp}(q_t, \pi) = O(\log T/T)$.

But is OS-Fisher useful?

Simpler formulation

The distribution $\Pi_z = \pi(\cdot | g(x) = z)$ is too abstract.

Theorem

Suppose $g\sharp\pi$ has Lipschitz density. Then the weak limit of $\pi_{\eta,z}(x) \propto \pi(x) \exp(-\frac{1}{2\eta}(g(x)-z)^2)$ as $\eta \to 0$ concentrates on $\mathcal{G}_z = \{x : g(x) = z\}$ and is a version of π_z. Moreover,

$$\mathbb{E}_{\Pi_z}[A_\pi \phi] = 0, \quad \forall \phi \perp \nabla g.$$

- This gives a Stein equation $\mathbb{E}_q[A_\pi \phi] = 0$
- The tangent bundle of $\mathcal{G}_z$ is a subset of $\phi \perp \nabla g$
- $\mathbb{E}_q[A_\pi \phi] \leq \sqrt{F_\perp(q, \pi)}$ when $\|\phi\|_\phi = 1$.
- $\mathbb{E}_q[A_\pi \phi]$ or $F_\perp(q, \pi)$ do not require q being on $\mathcal{G}_z$
- This only check the OS directions.
- Checking how far is q away from $\mathcal{G}_z$ is easy.

Theorem

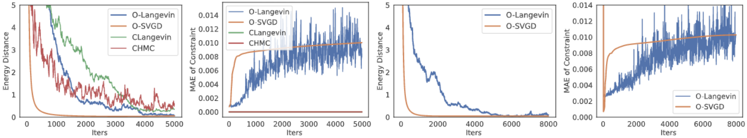
Suppose that Π_z satisfies κ-Poincare Inequality for $|z| \leq \delta$, and q is supported on $\{x : |g(x)| \leq \delta\}$. Then for any function f such that $|f| \leq 1$, the following holds

$$|\mathbb{E}_q[f] - \mathbb{E}_{\Pi_0}[f]| \leq \sqrt{\kappa F_\perp(q, \pi)} + \max_{|z|\leq\delta} |\mathbb{E}_{\Pi_z}[f] - \mathbb{E}_{\Pi_0}[f]|.$$

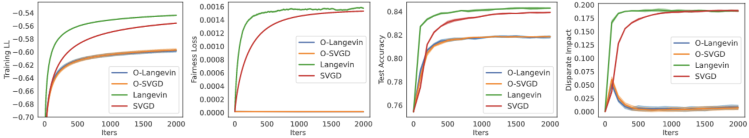
- Decomposition of mean difference/TV
- Only in L^2 case
- Poincare inequality with Euclidean-inheriant distance
- Can be used for q supported on R^d.

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