Sampling in Discrete and Constrained Domains

Ruqi Zhang

Purdue University

SPIGM@ICML

Joint work with Qiang Liu, Xingchao Liu, Xin T. Tong

The task of sampling is ubiquitous in ML

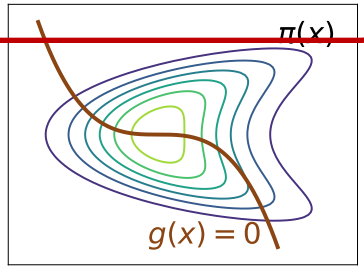
Obtain samples from a target distribution $\pi(\theta)$

- Probabilistic inference: $\pi(\theta)$ is a parameter distribution (e.g. the posterior of deep neural network weights)
- Generative modeling: $\pi(\theta)$ is a data distribution (e.g. energy-based models, diffusion models)
- Representation learning: $\pi(\theta)$ is a latent variable distribution (e.g. restricted Boltzmann machine)
- ……

Sampling beyond unconstrained continuous domains

- Sampling in unconstrained continuous domains is relatively well-studied
- Many powerful samplers, e.g. Langevin dynamics, Hamiltonian Monte Carlo
- However, sampling in domains with complicated structures is challenging
  - Discrete: lack of continuity; combinatorially large search space
  - Constrained: a implicitly-defined submanifold

Focus for today's talk

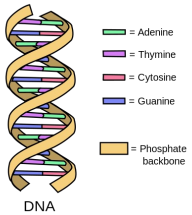


Discrete data and models

- Discrete data

Text

- beginning in december 1934 , training exercises were conducted for the tetrarchs and their crews using hamilcar gliders
- beginning in march 1946 , training exercises were conducted by the tetrarchs and their crews with hamilcar gliders .
- beginning in may 1926 , training exercises were conducted between the tetrarchs and their crews using hamilcar gliders .
- beginning in late 1942 , training exercises were conducted with the tetrarchs and their crews onboard hamilcar gliders .
- beginning in september 1961 , training exercises were conducted between the tetrarchs and their crews in hamilcar gliders .



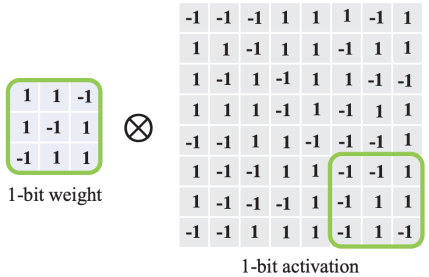
Genome

	A	B	C	D	E	F	G
1	Region	Gender	Style	Ship Date	Units	Price	Cost
2	East	Boy	Tee	1/31/2005	12	11.04	10.42
3	East	Boy	Golf	1/31/2005	12	13	12.6
4	East	Boy	Fancy	1/31/2005	12	11.96	11.74
5	East	Girl	Tee	1/31/2005	10	11.27	10.56
6	East	Girl	Golf	1/31/2005	10	12.12	11.95
7	East	Girl	Fancy	1/31/2005	10	13.74	13.33
8	West	Boy	Tee	1/31/2005	11	11.44	10.94
9	West	Boy	Golf	1/31/2005	11	12.63	11.73
10	West	Boy	Fancy	1/31/2005	11	12.06	11.51
11	West	Girl	Tee	1/31/2005	15	13.42	13.29
12	West	Girl	Golf	1/31/2005	15	11.48	10.67

Tabular Data

- Discrete models

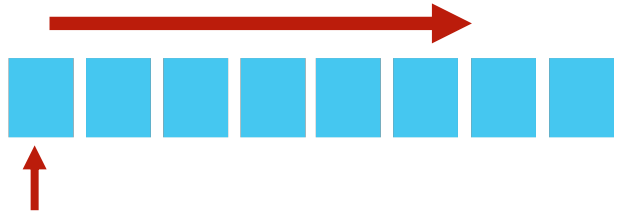
Binary neural networks



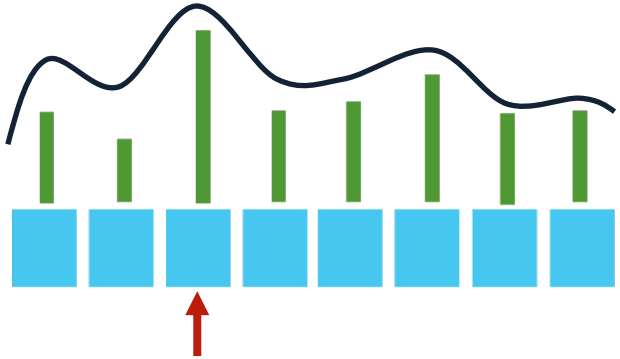
[Qin et al. 2020]

Discrete Samplers

- Gibbs sampling



- Gibbs with Gradients

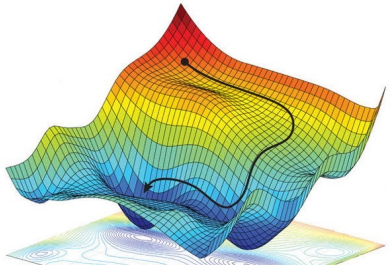


Only update one dim: suffer from high-dimensional and highly correlated distributions!

Oops I Took A Gradient: Scalable Sampling for Discrete Distributions. Grathwohl et al., ICML 2021

Continuous Sampler: Langevin Dynamics

$$\theta' = \theta + \frac{\alpha}{2}\nabla U(\theta) + \sqrt{\alpha}\xi, \qquad \xi \sim \mathcal{N}(0, I)$$



- Gradients guide the sampler to efficiently explore high probability regions
- Cheaply update all coordinates in parallel in a single step

What is the analog of Langevin dynamics in discrete domains?

Our Method: Discrete Langevin Proposal

$$q(\theta'|\theta) = \frac{\exp\left(-\frac{1}{2\alpha}\left\|\theta' - \theta - \frac{\alpha}{2}\nabla U(\theta)\right\|_2^2\right)}{Z_\Theta(\theta)}$$

- Langevin proposal is applicable to any kind of spaces
  - When $\Theta = \mathbb{R}^d$, recover the Gaussian proposal
  - When Θ is a discrete domain, obtain a gradient-based discrete proposal
- Coordinatewise factorization $q(\theta'|\theta) = \prod_{i=1}^{d} q_i(\theta_i'|\theta)$

$$q_i(\theta_i'|\theta) = \text{Categorical}\left(\text{Softmax}\left(\frac{1}{2}\nabla U(\theta)_i(\theta_i' - \theta_i) - \frac{(\theta_i' - \theta_i)^2}{2\alpha}\right)\right)$$

cheaply computed in parallel

Discrete Langevin Proposal (DLP)

A Langevin-like Sampler for Discrete Distributions. Zhang et al., ICML 2022

Visualization of Discrete Langevin Proposal

$$q_i(\theta_i'|\theta) = \text{Categorical}\left(\text{Softmax}\left(\frac{1}{2}\nabla U(\theta)_i(\theta_i' - \theta_i) - \frac{(\theta_i' - \theta_i)^2}{2\alpha}\right)\right)$$

update all coordinates based on gradient info in parallel

Samplers: *discrete unadjusted Langevin algorithm* (DULA)

discrete Metropolis-adjusted Langevin algorithm (DMALA)

Convergence Analysis

Theorem (informal): *The asymptotic bias of DULA's stationary distribution is zero for log-quadratic distributions and is small for distributions that are close to being log-quadratic*

Other Variants

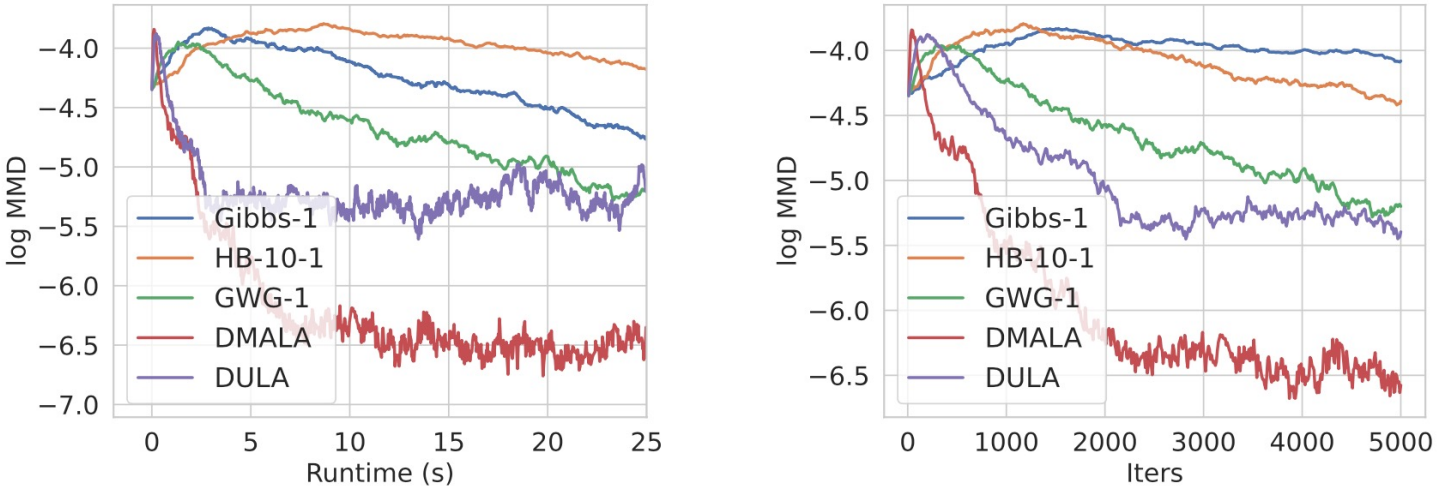
- With stochastic gradients

Theorem (informal): *When the variance of the stochastic gradient or the stepsize decreases, the stochastic DLP in expectation will be closer to the full-batch DLP*

- With preconditioners

$$q_i(\theta_i'|\theta) \propto \exp\left(\frac{1}{2}\nabla U(\theta)_i(\theta_i' - \theta_i) - \frac{(\theta_i - \theta_i')^2}{2\alpha g_i}\right)$$

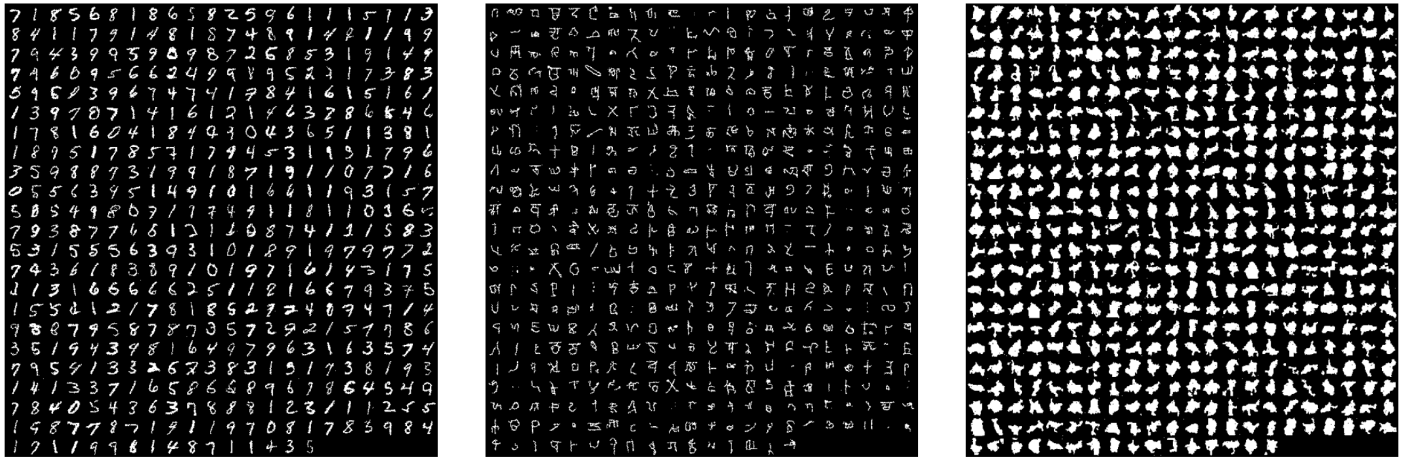
Experiments: Restricted Boltzmann Machines



- DULA and DMALA converge faster to the target distribution

Experiments: Deep Energy-based Models

Dataset	VAE (Conv)	EBM (Gibbs)	EBM (GWG)	EBM (DULA)	EBM (DMALA)
Static MNIST	-82.41	-117.17	-80.01	-80.71	**-79.46**
Dynamic MNIST	-80.40	-121.19	-80.51	-81.29	**-79.54**
Omniglot	-97.65	-142.06	-94.72	-145.68	**-91.11**
Caltech Silhouettes	-106.35	-163.50	-96.20	-100.52	**-87.82**



Generated images

Experiments: Language Models

Infilling Task: he had not , after all , [MASK] me the chance but [MASK] abandoned me [MASK] .

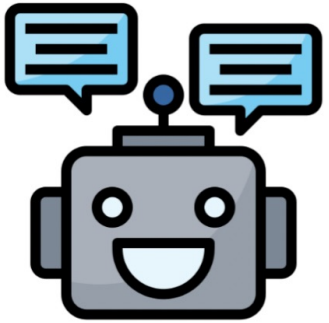
Gibbs Results:	**GWG Results**:	**DMALA Results**:
given me the chance but had abandoned me instead	given me the chance but had abandoned me instead	shown me the chance but had abandoned me anyway
given me the chance but had abandoned me instead	given me the chance but had abandoned me himself	shown me the chance but not abandoned me immediately
given me the chance but had abandoned me instead	offered me the chance but had abandoned me completely	gives me the chance but also abandoned me perhaps
given me the chance but had abandoned me completely	gave me the chance but had abandoned me anyway	grants me the chance but really abandoned me entirely
given me the chance but had abandoned me anyway	given me the chance but he abandoned me instead	offered me the chance but yet abandoned me instead

Model	Methods	Self-BLEU (↓)	Unique n-grams (%) (↑)						Corpus BLEU (↑)
			Self		WT103		TBC		
			$n=2$	$n=3$	$n=2$	$n=3$	$n=2$	$n=3$	
Bert-Base	Gibbs	86.84	10.98	16.08	18.57	32.21	21.22	33.05	**23.82**
	GWG	81.97	15.12	21.79	22.76	37.59	24.72	37.98	22.84
	DULA	**72.37**	**23.33**	**32.88**	27.74	**45.85**	30.02	**46.75**	21.82
	DMALA	72.59	23.26	32.64	**27.99**	45.77	**30.32**	46.49	21.85
Bert-Large	Gibbs	88.78	9.31	13.74	17.78	30.50	20.48	31.23	22.57
	GWG	86.50	11.03	16.13	19.25	33.20	21.42	33.54	**23.08**
	DULA	77.96	17.97	26.64	23.69	41.30	26.18	42.14	21.28
	DMALA	**76.27**	**19.83**	**28.48**	**25.38**	**42.94**	**27.87**	**43.77**	21.73

Constraints are everywhere in ML

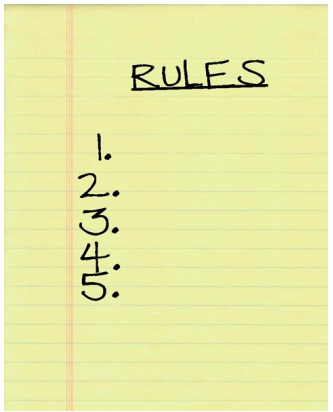
Fairness

Privacy



Interpretability

Safety



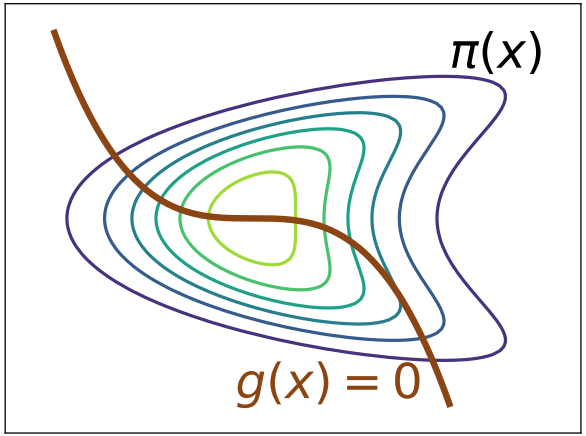
Logic rule

Problem Setup

- Consider sampling with an equality constraint:

$$\text{sample } \pi(x) \text{ on } \mathcal{G}_0 = \{x \in \mathbb{R}^d : g(x) = 0\}$$

where $g(x)$ can be any differentiable function



Variational View: Sampling as Optimization

- Transform the constrained sampling problem into a constrained functional minimization problem

$$\min_{q \in \mathcal{P}} \mathrm{KL}(q \mid\mid \pi), \quad \text{s.t.} \quad q(g(x) = 0) = 1$$

- The velocity field v_t that solves above problem is

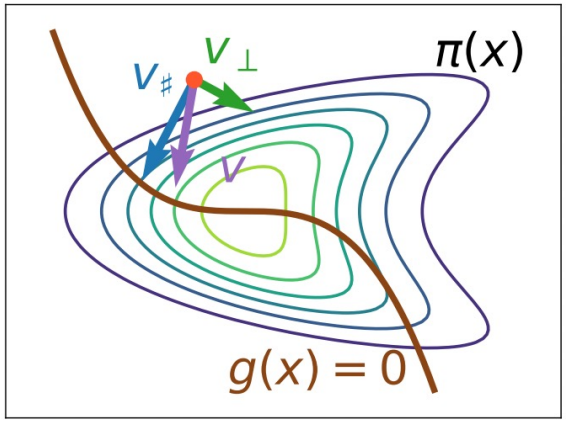
$$v_t = \arg\max_{v \in \mathcal{H}} \mathbb{E}_{q_t}[(s_\pi - s_{q_t})^\top v] - \frac{1}{2}\|v\|_{\mathcal{H}}^2, \quad \text{s.t.} \quad v_t(x)^\top \nabla g(x) = -\psi(g(x))$$

where $\psi(x) = \alpha \mathrm{sign}(x)|x|^{1+\beta}$

Sampling in Constrained Domains with Orthogonal-Space Variational Gradient Descent.
Zhang et al., NeurIPS 2022

Orthogonal-Space Variational Gradient Descent

- v_t can be decomposed as $v_t = v_\sharp + v_\perp$

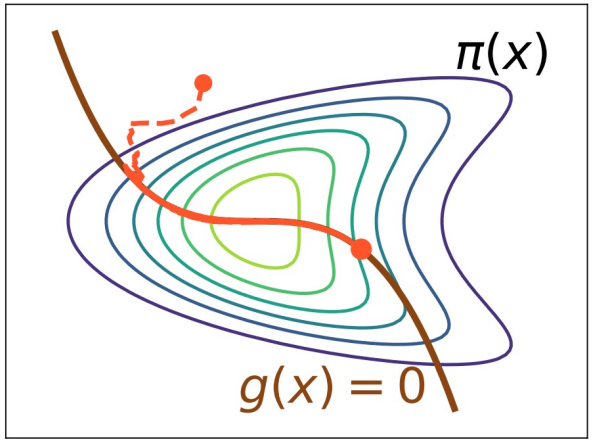


- $v_\sharp(x)$: drives the sampler towards the manifold following ∇g
- $v_\perp$: makes the sampler explore the manifold following the density π

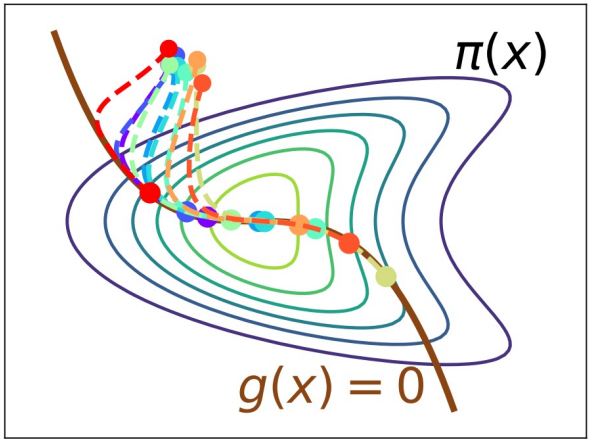
Convergence Analysis

Theorem (informal): *O-Gradient converges to the target constrained distribution with rate O(1/the number of iterations) under mild conditions*

Practical Algorithms: O-Langevin and O-SVGD



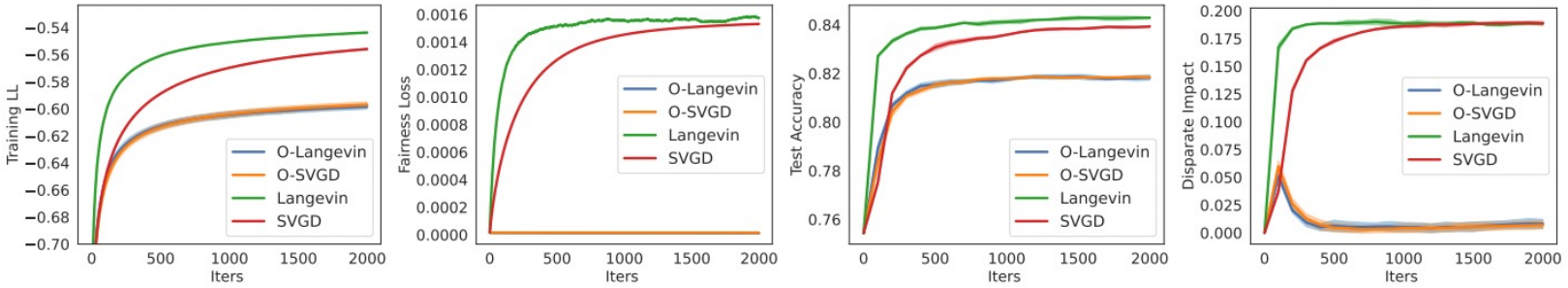
O-Langevin



O-SVGD

Income Classification with Fairness Constraint

- Predict whether an individual's annual income is greater than 50, 000 unfavorably in terms of the gender



(a) Training Curve

(b) Testing Curve

Prior-Agnostic Bayesian Neural Networks

- To avoid a bad prior, sample from the posterior with the constraint of a reasonably high data fitness

	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

Sampling in many other scenarios

- Low precision

  Low-Precision Stochastic Gradient Langevin Dynamics. ICML 2022

- Privacy

  DP-Fast MH: Private, Fast, and Accurate Metropolis-Hastings for Large-Scale Bayesian Inference. ICML 2023

- Data distribution shifts

  Long-tailed Classification from a Bayesian-decision-theory Perspective. AABI 2023

- ……

Takeaways

- Sampling is a ubiquitous task in ML, ranging from probabilistic inference to generative modeling and representation learning
- Sampling in discrete domains can be efficient using a discrete version of Langevin dynamics
- Sampling in constrained domains can be formulated as a functional optimization solved using a special gradient flow

Thank you!