A Langevin-like Sampler for Discrete Distributions

Ruqi Zhang
UT Austin/Purdue

Xingchao Liu
UT Austin

Qiang Liu
UT Austin

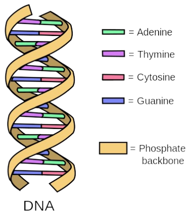
Discrete variables are ubiquitous

Discrete variables are ubiquitous

- 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

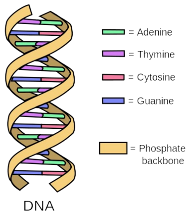
Tabular Data

Discrete variables are ubiquitous

- 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 .



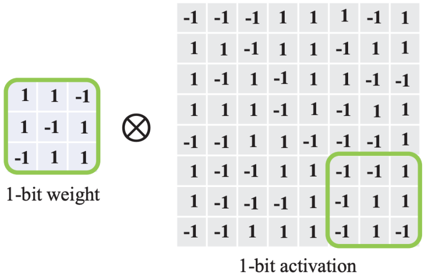
Genome

	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



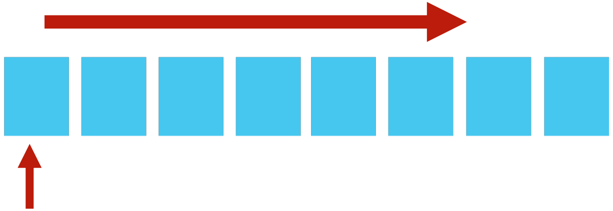
[Qin et al. 2020]

Discrete Samplers

- Gibbs sampling

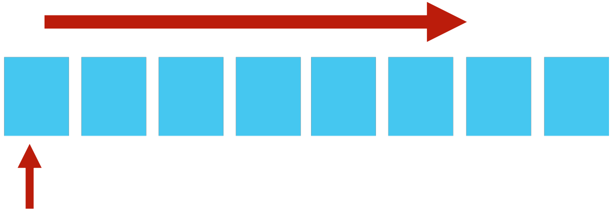
Discrete Samplers

- Gibbs sampling

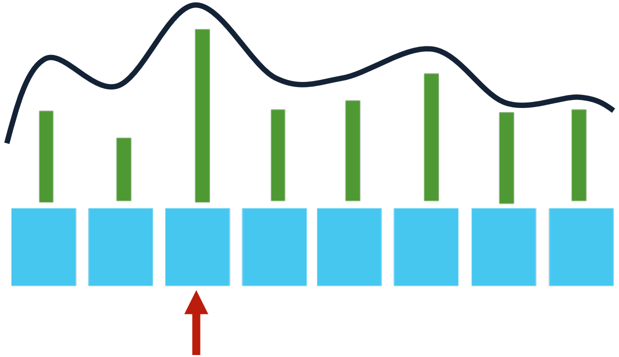


Discrete Samplers

- Gibbs sampling



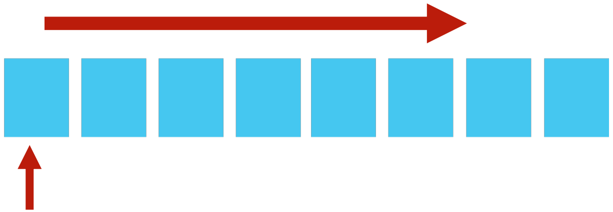
- Gibbs with Gradients



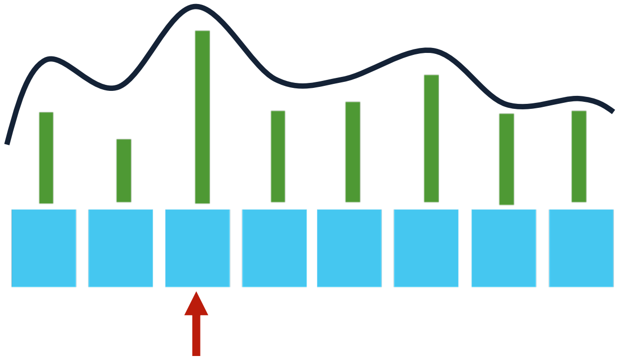
Oops I Took A Gradient: Scalable Sampling for Discrete Distributions. Grathwohl et al., 2021

Discrete Samplers

- Gibbs sampling



- Gibbs with Gradients

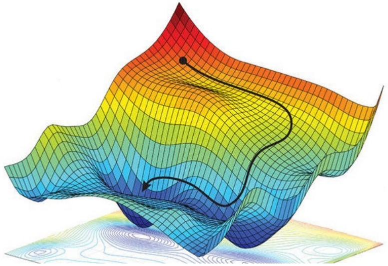


Only update one dim: suffer from high-dimensional and highly correlated distributions!

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

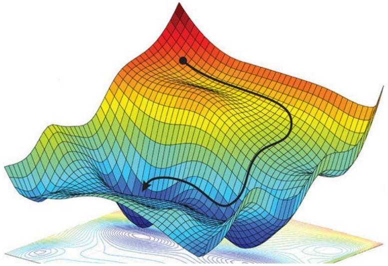
Continuous Sampler: Langevin algorithm

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



Continuous Sampler: Langevin algorithm

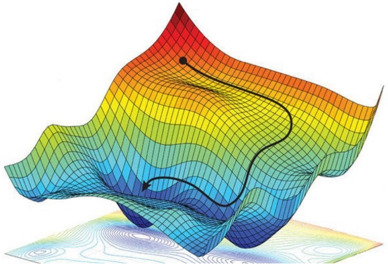
$$\theta' = \theta + \frac{\alpha}{2}\nabla U(\theta) + \sqrt{\alpha}\xi, \qquad \xi \sim \mathcal{N}(0, I)$$



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

Continuous Sampler: Langevin algorithm

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



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

What is the analogue of the Langevin algorithm in discrete domains?

Our Method: Discrete Langevin Proposal

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

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

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

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

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

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)

Visualization of Discrete Langevin Proposal

Visualization of Discrete Langevin Proposal

Visualization of Discrete Langevin Proposal

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

Visualization of Discrete Langevin Proposal

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

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

Visualization of Discrete Langevin Proposal

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

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

Other Variants

- With stochastic gradients

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*

Other Variants

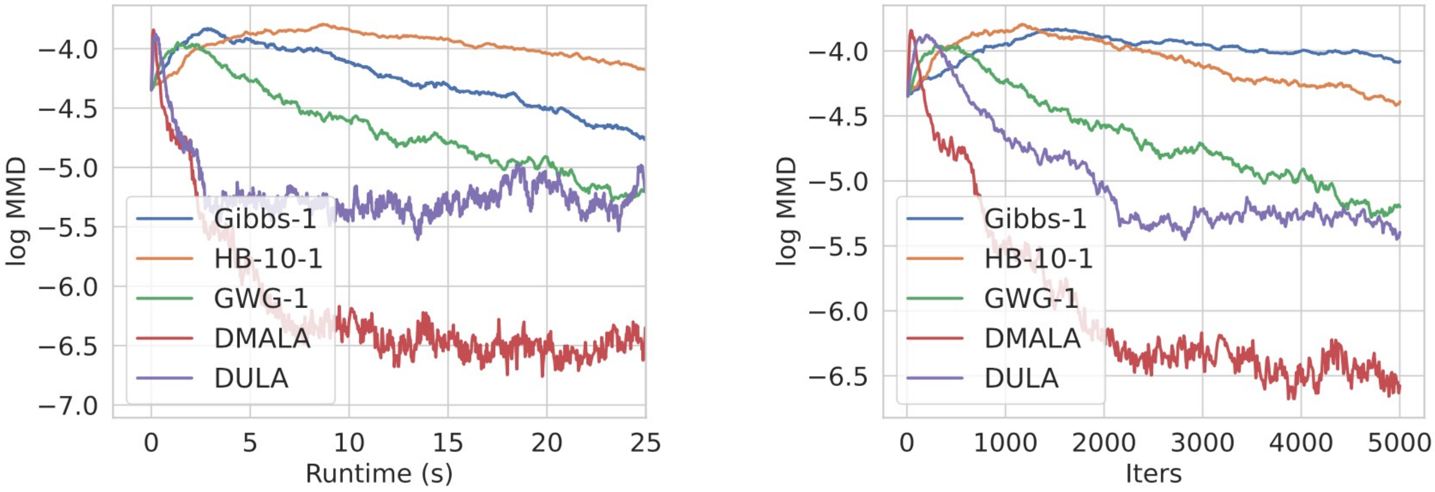
- With stochastic gradients

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

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

Sampling From Restricted Boltzmann Machines



- DULA and DMALA converge faster to the target distribution

Summary

- We propose Discrete Langevin Proposal (DLP) for discrete distributions

Summary

- We propose Discrete Langevin Proposal (DLP) for discrete distributions
- We develop several variants with DLP, including unadjusted, Metropolis-adjusted, stochastic, and preconditioned versions

Summary

- We propose Discrete Langevin Proposal (DLP) for discrete distributions
- We develop several variants with DLP, including unadjusted, Metropolis-adjusted, stochastic, and preconditioned versions
- We prove the asymptotic convergence of DLP under log-quadratic and general distributions

Summary

- We propose Discrete Langevin Proposal (DLP) for discrete distributions
- We develop several variants with DLP, including unadjusted, Metropolis-adjusted, stochastic, and preconditioned versions
- We prove the asymptotic convergence of DLP under log-quadratic and general distributions
- We provide a thorough empirical evaluation including deep EBMs, binary DNNs and text generation

arXiv.org https://arxiv.org/abs/2206.09914

https://github.com/ruqizhang/discrete-langevin