Fast parallel sampling under isoperimetry

Nima Anari
Stanford University

Sinho Chewi
Yale University

Thuy-Duong "June" Vuong
Stanford → Miller Institute

COLT 2024
Edmonton, Canada

Sampling

Input: Distribution $\pi = \exp(-V(\cdot))$ over $\mathbb{R}^d$, oracle access to ∇V

Output: random $x \sim \hat{\pi}$ s.t. $\hat{\pi} \approx \pi$

- Fundamental task in statistics
- Many applications:
  - Generative AI
  - Inference
  - ...

Sampling

Input: Distribution $\pi = \exp(-V(\cdot))$ over $\mathbb{R}^d$, oracle access to ∇V

Assumptions:

- Smooth potential: $||\nabla^2 V||_{OP} \leq \beta$
- Isoperimetry: π satisfies log-Sobolev inequality (LSI)

Guarantee

Input: Distribution $\pi = \exp(-V(\cdot))$ over $\mathbb{R}^d$, oracle access to ∇V

Output: random $x \sim \hat{\pi}$ s.t. $\hat{\pi} \approx \pi$

Measures of distance between distributions:

- KL (Kullback-Leibler): $\hat{\pi} \approx_{KL} \pi$ $\quad \mathcal{D}_{KL}(\rho, \pi) = \mathbb{E}_\rho[\log \frac{\rho}{\pi}]$
- Total variation: $\hat{\pi} \approx_{TV} \pi$ $\quad \mathcal{D}_{TV}(\rho, \pi) = \mathbb{E}_\rho[|1 - \pi/\rho|]$
- $\mathcal{W}_2$ (Wasserstein): $\hat{\pi} \approx_{\mathcal{W}_2} \pi$ $\quad \mathcal{W}_2\,(\rho, \pi) = \mathbb{E}_{(X,Y)\sim\Pi}[||X - Y||^2]$

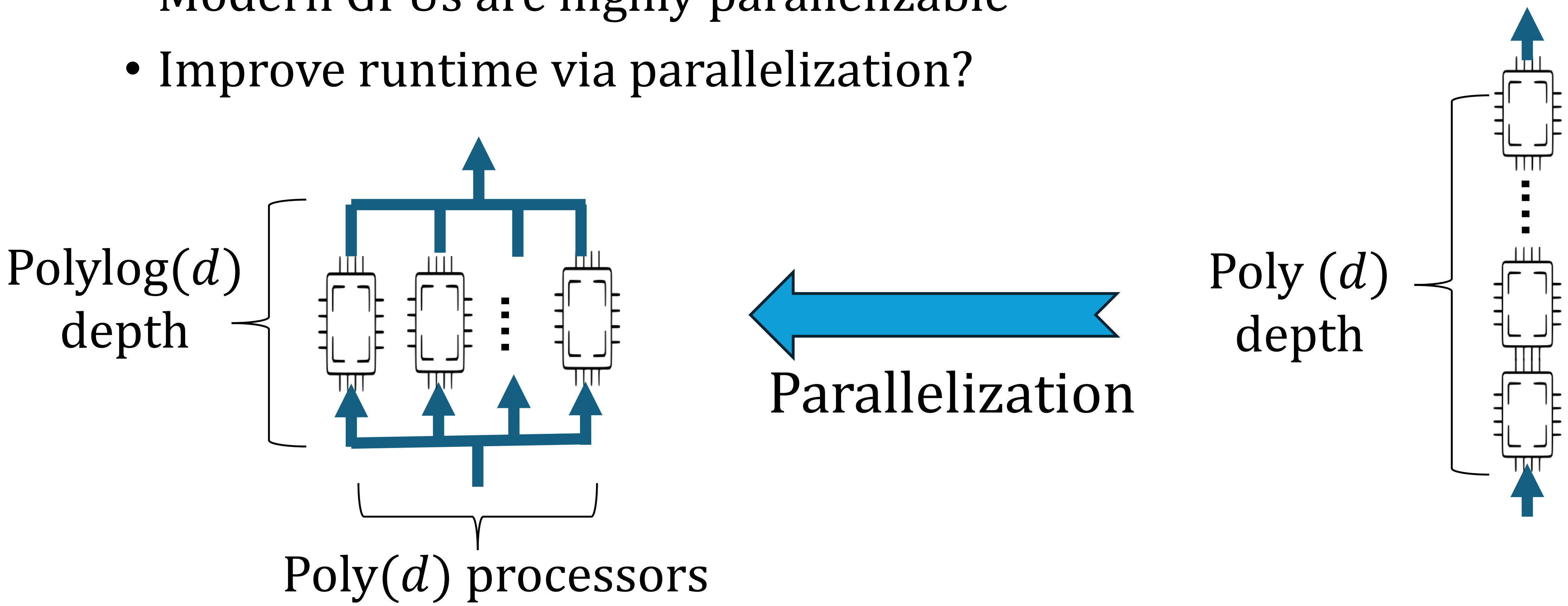
Measures of distance between distributions

- KL (Kullback-Leibler): $\hat{\pi} \approx_{KL} \pi$ $\qquad \mathcal{D}_{KL}(\rho,\pi) = \mathbb{E}_\rho[\log\frac{\rho}{\pi}]$
- Total variation: $\hat{\pi} \approx_{TV} \pi$ $\qquad \mathcal{D}_{TV}(\rho,\pi) = \mathbb{E}_\rho[|1-\pi/\rho|]$
- $\mathcal{W}_2$ (Wasserstein): $\hat{\pi} \approx_{\mathcal{W}_2} \pi$ $\qquad \mathcal{W}_2\,(\rho,\pi) = \mathbb{E}_{(X,Y)\sim\Pi}[||X-Y||^2]$

<u>Fact 0</u>: $\mathcal{D}_{TV}^2 \le \mathcal{D}_{KL}$
<u>Fact 1</u>: When π satisfies LSI, $\mathcal{W}_2^2 \le \mathcal{D}_{KL}$

Parallel computing

- The dimension d is large in modern big-data applications
- Modern GPUs are highly parallelizable
- Improve runtime via parallelization?



Our result: fast parallel sampler

Input: Distribution $\pi = \exp(-V(\cdot))$ over $\mathbb{R}^d$, oracle access to ∇V

Assumptions:

- Smooth potential: $||\nabla^2 V||_{OP} \leq \beta$
- Isoperimetry: π satisfies log-Sobolev inequality (LSI)

Results:

Guarantee	Parallel depth	Total work #(processors+oracle calls)	Algorithm
KL	$O(\log^2 d)$	$\tilde{O}(d)$	Parallel-LMC

Fact 0: $\mathcal{D}_{TV}^2 \leq \mathcal{D}_{KL}$
Fact 1: When π satisfies LSI, $\mathcal{W}_2^2 \leq \mathcal{D}_{KL}$

Our result: fast parallel sampler

Input: Distribution $\pi = \exp(-V(\cdot))$ over $\mathbb{R}^d$, oracle access to ∇V

Assumptions:

- Smooth potential: $||\nabla^2 V||_{OP} \leq \beta$
- Isoperimetry: π satisfies log-Sobolev inequality (LSI)

Results:

Guarantee	Parallel depth	Total work #(processors+oracle calls)	Algorithm
KL	$O(\log^2 d)$	$\tilde{O}(d)$	Parallel-LMC
TV	$O(\log^2 d)$	$\tilde{O}(\sqrt{d})$	Parallel -ULMC

Prior works on parallel sampler

Input: Distribution $\pi = \exp(-V(\cdot))$ over $\mathbb{R}^d$, oracle access to ∇V

<u>Assumptions</u>:

- Smooth potential: $||\nabla^2 V||_{OP} \leq \beta$
- ~~Isoperimetry: π satisfies log-Sobolev inequality (LSI)~~
- Strong log-concavity: $\nabla^2 V \succcurlyeq \alpha I \succ 0$

Implies LSI

Guarantee	Parallel depth	Total work	Algorithm	Authors
$\mathcal{W}_2$	$O(\log^{O(1)} d)$	$\tilde{O}(\sqrt{d})$	Randomized midpoint	[Shen-Lee'19]

Prior works on parallel sampler

Input: Distribution $\pi = \exp(-V(\cdot))$ over $\mathbb{R}^d$, oracle access to ∇V

<u>Assumptions</u>:

- Smooth potential: $||\nabla^2 V||_{OP} \leq \beta$
- Isoperimetry: π satisfies log-Sobolev inequality (LSI)
- Strong log-concavity: $\nabla^2 V \succcurlyeq \alpha I \succ 0$

Implies LSI

Guarantee	Parallel depth	Total work	Algorithm	Authors
KL	$O(\log^2 d)$	$\tilde{O}(d)$	Parallel-LMC	This work
TV	$O(\log^2 d)$	$\tilde{O}(\sqrt{d})$	Parallel -ULMC	This work
$\mathcal{W}_2$	$O(\log^{O(1)} d)$	$\tilde{O}(\sqrt{d})$	Randomized midpoint	[Shen-Lee'19]

Prior works on parallel sampler

Input: Distribution $\pi = \exp(-V(\cdot))$ over $\mathbb{R}^d$, oracle access to ∇V

<u>Assumptions</u>:

- Smooth potential: $||\nabla^2 V||_{OP} \leq \beta$
- ~~Isoperimetry: π satisfies log-Sobolev inequality (LSI)~~
- Strong log-concavity: $\nabla^2 V \succcurlyeq \alpha I \succ 0$

Implies LSI

Guarantee	Parallel depth	Total work	Algorithm	Authors
KL			Parallel-LMC	This work
TV			Parallel -ULMC	This work
$\mathcal{W}_2$		?	Randomized midpoint	[Shen-Lee'19]

Prior works on parallel sampler

Input: Distribution $\pi = \exp(-V(\cdot))$ over $\mathbb{R}^d$, oracle access to ∇V

<u>Assumptions</u>:

- Smooth potential: $||\nabla^2 V||_{OP} \leq \beta$
- ~~Isoperimetry: π satisfies log-Sobolev inequality (LSI)~~
- Strong log-concavity: $\nabla^2 V \succcurlyeq \alpha I \succ 0$c

Implies LSI

Guarantee	Parallel depth	Total work	Algorithm	Authors
KL			Parallel-LMC	This work
TV			Parallel -ULMC	This work
$\mathcal{W}_2$			Randomized midpoint	[Shen-Lee'19]

Application in discrete sampling

Input: Distribution μ over $\{\pm 1\}^d \subseteq \mathbb{R}^d$, oracle access to μ

Output: random $x \sim \hat{\pi}$ s.t. $\hat{\pi} \approx_{TV} \pi$

- [AHLXVY-STOC23] reduces "nice" discrete μ to continuous smooth strongly log-concave π
- $\approx_{TV}$ for continuous $\dashrightarrow \approx_{TV}$ for discrete
- $\approx_{\mathcal{W}_2}$ for continuous $\cancel{\dashrightarrow} \approx_{TV}$ for discrete, except for restricted cases

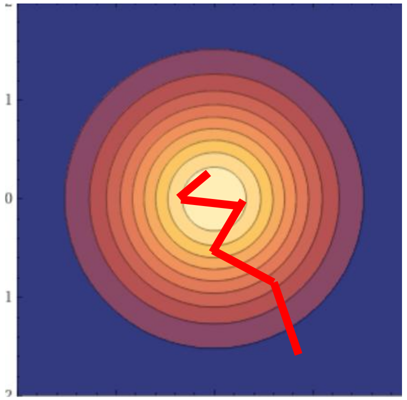
Algorithm

Fast parallel way to discretize continuous diffusions:

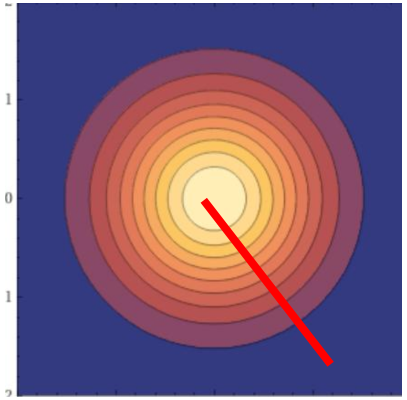
- Langevin diffusions
- Underdamped Langevin diffusions
- DDPM…

Continuous time diffusions

Langevin diffusion: $$dX_t = \underbrace{-\nabla V(X_t)}_{\text{Drift}}dt + \underbrace{\sqrt{2}dB_t}_{\text{Brownian motion}}$$



Gradient descent: $$dX_t = \underbrace{-\nabla V(X_t)}_{\text{Drift}}dt$$



Continuous time diffusions

Langevin diffusion: $$dX_t = \underbrace{-\nabla V(X_t)}_{\text{Drift}}\mathrm{dt} + \underbrace{\sqrt{2}dB_t}_{\text{Brownian motion}}$$

$$Law(X_t) \approx_{KL} \pi \text{ if } t \geq \log d$$

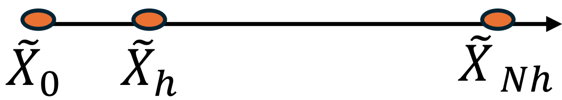
But, we can't implement the continuous time process

Discretization

Langevin diffusion: $$dX_t = \underbrace{-\nabla V(\mathrm{X_t})\mathrm{dt}}_{\text{Drift}} + \underbrace{\sqrt{2}dB_t}_{\text{Brownian motion}}$$

$Law(\tilde{X}_T)$

Langevin Monte-Carlo: (LMC) $$d\tilde{X}_t = -\nabla V\left(\tilde{X}_{\left\lfloor\frac{\mathrm{t}}{\mathrm{h}}\right\rfloor h}\right)\mathrm{dt} + \sqrt{2}dB_t$$

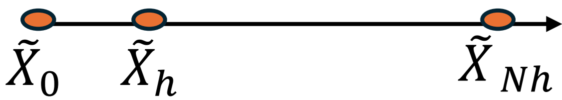


$$\tilde{X}_{(n+1)h} - \tilde{X}_{nh} = -\nabla V(\tilde{X}_{nh})\mathrm{h} \quad + \sqrt{2}(\mathrm{B}_{(n+1)h} - B_h)$$

Discretization

Langevin diffusion: $$dX_t = \underbrace{-\nabla V(\mathrm{X_t})\mathrm{dt}}_{\text{Drift}} + \underbrace{\sqrt{2}dB_t}_{\text{Brownian motion}}$$

Langevin Monte-Carlo: (LMC) $$d\tilde{X}_t = -\nabla V\left(\tilde{X}_{\left\lfloor \frac{\mathrm{t}}{\mathrm{h}} \right\rfloor h}\right)\mathrm{dt} + \sqrt{2}dB_t$$

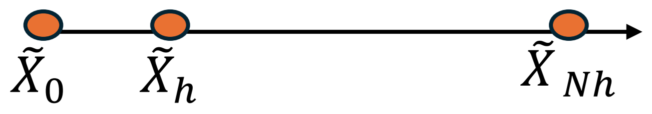


$$Law(X_T) \approx_{KL} \pi \text{ if } \mathrm{T} \geq \log d$$

$$Law(\tilde{X}_T) \approx_{KL} Law(X_T) \approx_{KL} \pi \text{ if } \mathrm{T} \geq \log d \text{ and } h \leq \frac{1}{d} \Rightarrow \mathrm{N} = \frac{\mathrm{T}}{\mathrm{h}} = \tilde{\Omega}(d)$$

Discretization—in parallel

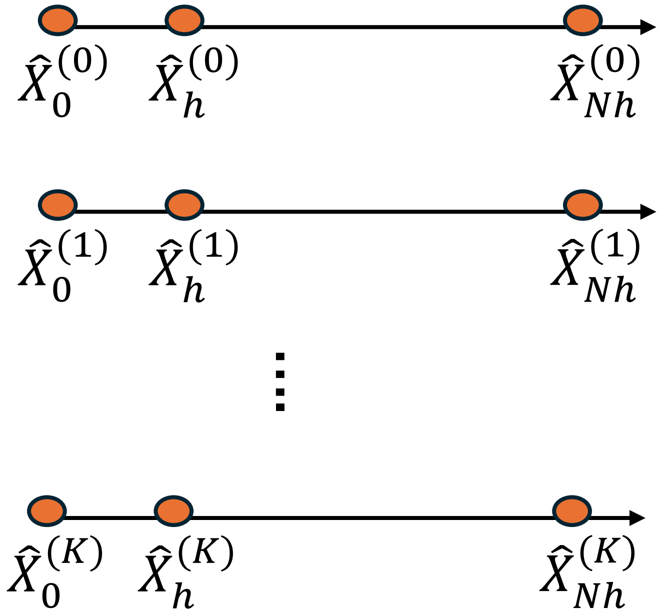
LMC: $$d\tilde{X}_t = -\nabla V\left(\tilde{X}_{\left\lfloor \frac{\mathrm{t}}{\mathrm{h}} \right\rfloor h}\right) \mathrm{dt} + \sqrt{2}\, dB_t$$



Parallel-LMC:

$$d\hat{X}_t^{(r)} = -\nabla V\left(\hat{X}^{(r-1)}_{\left\lfloor \frac{\mathrm{t}}{\mathrm{h}} \right\rfloor h}\right) \mathrm{dt} + \sqrt{2}\, dB_t$$

$$\hat{\mathrm{X}}^{(\mathrm{r})}_{nh} - \hat{\mathrm{X}}^{(\mathrm{r})}_{0} = -\mathrm{h} \sum \nabla \mathrm{V}\left(\hat{X}^{(r-1)}_{\mathrm{ih}}\right) \quad + \sqrt{2}\mathrm{B}_{nh}$$



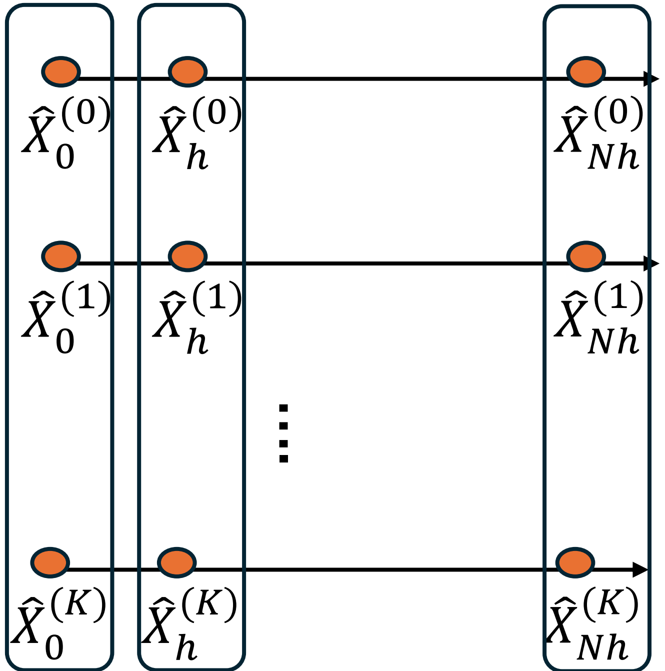
Discretization—in parallel

LMC: $$d\tilde{X}_t = -\nabla V\left(\tilde{X}_{\lfloor \frac{t}{h} \rfloor h}\right) \mathrm{dt} + \sqrt{2} dB_t$$

Parallel-LMC:

$$d\hat{X}_t^{(r)} = -\nabla V\left(\hat{X}^{(r-1)}_{\lfloor \frac{t}{h} \rfloor h}\right) \mathrm{dt} + \sqrt{2} dB_t$$

$$\hat{\mathrm{X}}^{(\mathrm{r})}_{(\mathrm{n}+1)\mathrm{h}} - \hat{\mathrm{X}}^{(\mathrm{r})}_0 = -\mathrm{h} \sum \nabla \mathrm{V}\left(\hat{X}^{(r-1)}_{\mathrm{ih}}\right) \quad + \sqrt{2}\mathrm{B}_{(\mathrm{n}+1)\mathrm{h}}$$



Analysis

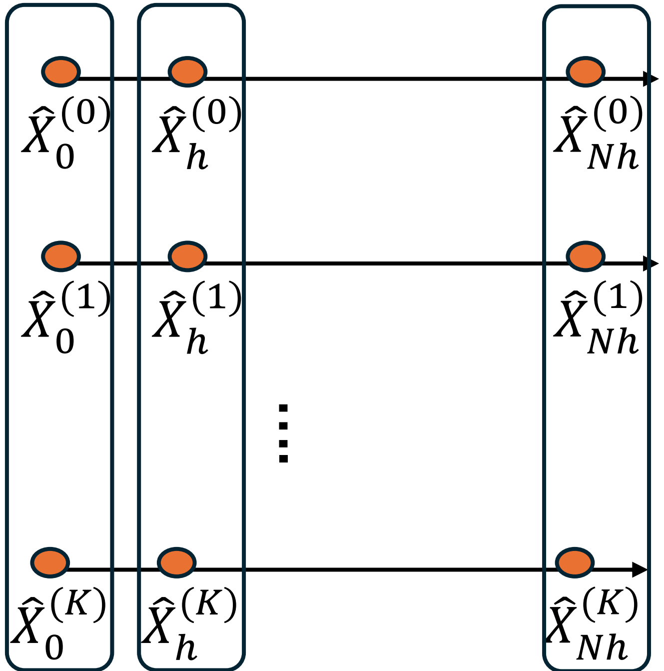
$$Law\left(\hat{X}_t^{(K)}\right) \approx_{KL} Law(X_t) \text{ if } h \simeq \frac{1}{d}, \mathrm{K} \simeq \log d\,, t = 0.1$$

$$dist\left(\hat{X}_t^{(r)}, X_t\right) \leq 0.8\, dist\left(\hat{X}_t^{(r-1)}, X_t\right) + \ldots$$

Parallel-LMC:

$$d\hat{X}_t^{(r)} = -\nabla V\left(\hat{X}^{(r-1)}_{\lfloor \frac{\mathrm{t}}{\mathrm{h}} \rfloor h}\right) \mathrm{dt} + \sqrt{2} dB_t$$

$$\hat{\mathrm{X}}^{(\mathrm{r})}_{(\mathrm{n}+1)\mathrm{h}} - \hat{\mathrm{X}}^{(\mathrm{r})}_{0} = -\mathrm{h} \sum \nabla \mathrm{V}\left(\hat{X}^{(r-1)}_{\mathrm{ih}}\right) \quad + \sqrt{2}\mathrm{B}_{(\mathrm{n}+1)\mathrm{h}}$$



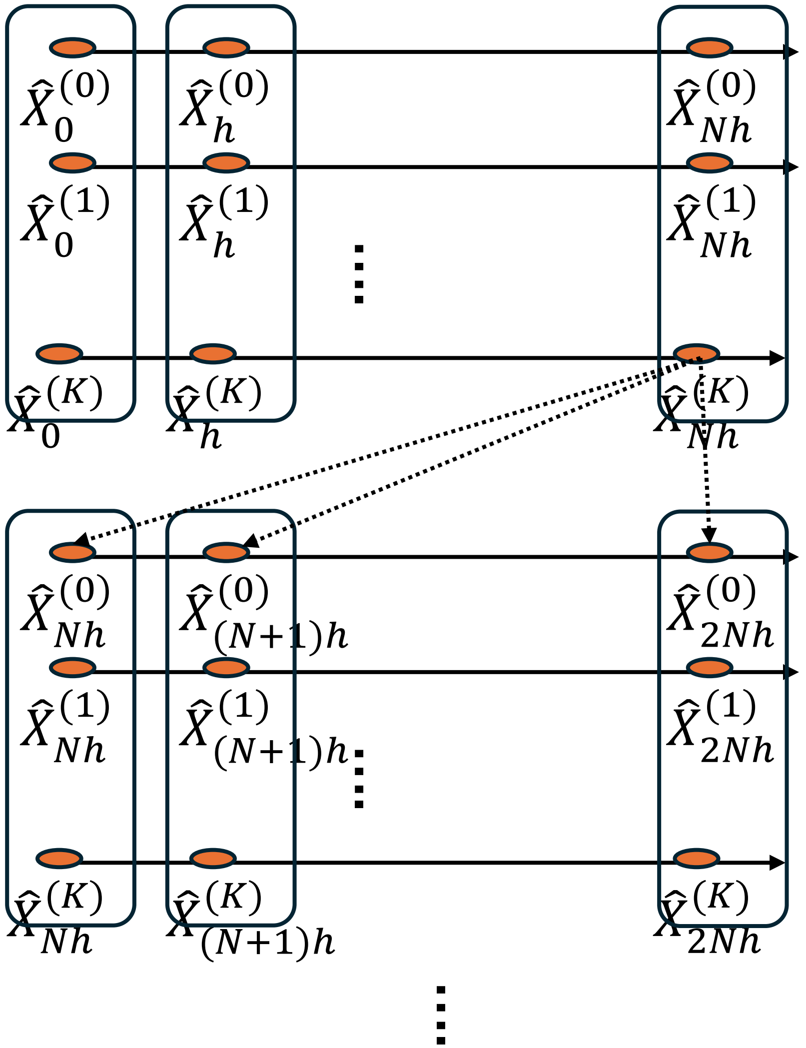
Analysis

$Law\left(\hat{X}_t^{(K)}\right) \approx_{KL} Law(X_t)$ if

$h \simeq \frac{1}{d}, \mathrm{K} \simeq \log d\,, t = 0.1$

Iterate $N' = \log d$ times gives

$Law\left(\hat{X}_{N't}^{(K)}\right) \approx_{KL} Law(X_{N't}) \approx_{KL} \pi$



Analysis

$Law\left(\hat{X}_t^{(K)}\right) \approx_{KL} Law(X_t)$ if

$h \simeq \frac{1}{d}$, K $\simeq \log d$, $t = 0.1$

Iterate $N' = \log d$ times gives

$Law\left(\hat{X}_{N't}^{(K)}\right) \approx_{KL} Law(X_{N't}) \approx_{KL} \pi$

Depth: $KN' \simeq \log^2 d$
Work: $KN'N \simeq \tilde{O}(d)$