Non-asymptotic Guarantees for High-Dimensional Sampling

Raaz Dwivedi
EECS Department

Part II:
Theoretical guarantees for Markov Chain Monte Carlo (MCMC)

- **Main question**: How many MCMC iterations (T) are needed to get a desired accuracy?

$$\|\mathbf{P}^\star - \mathbf{P}(X_T)\|_{\mathrm{tv}} \leq \delta$$

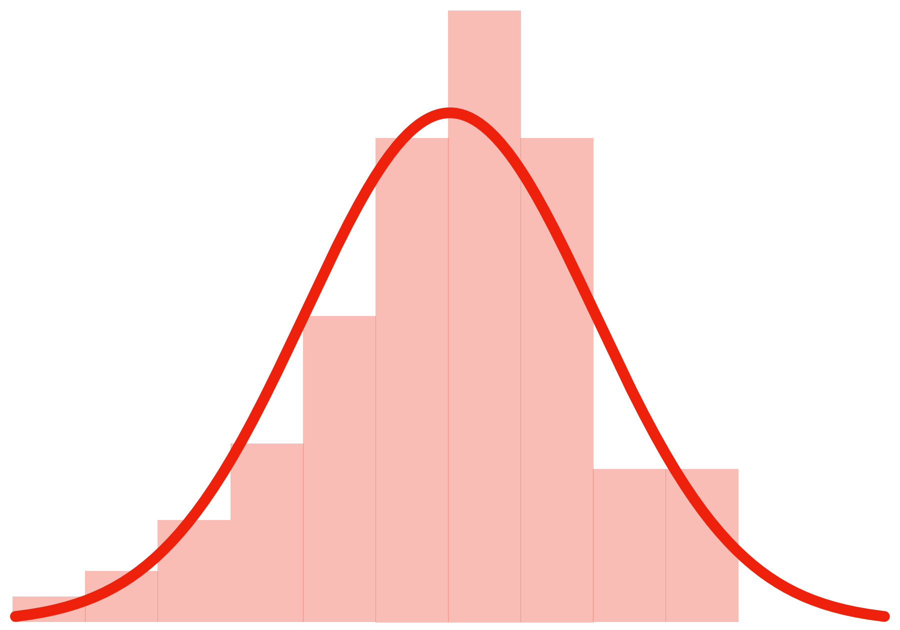
- **Insights en route**: How much does gradient information help for sampling?

Joint work with Chen-Wainwright-Yu

Sampling versus optimization

- Draw samples from the density

$$X \sim p^\star \propto e^{-f}$$

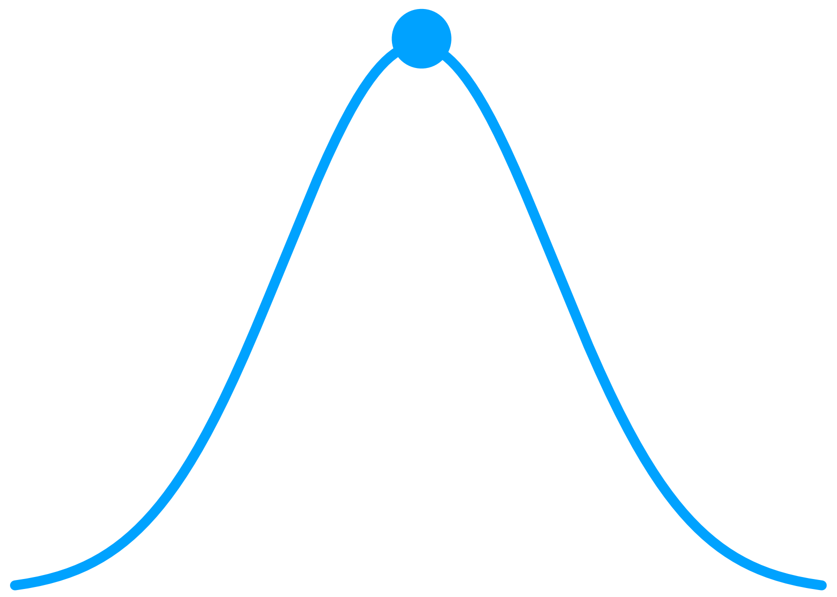


- Unadjusted Langevin algorithm (ULA)

$$X_k = X_{k-1} - h \nabla f(X_{k-1}) + \sqrt{2h}\xi_k$$
$$\xi_k \sim \mathcal{N}(0, I_d)$$

- Find mode of the density (or MAP)

$$x^\star \leftarrow \arg\max p^\star = \arg\min f$$



- Gradient descent

$$x_k = x_{k-1} - h \nabla f(x_{k-1})$$

'81 Parisi '94 Grenander-Miller, '96 Roberts-Tweedie

Langevin algorithms: Origin

- Langevin diffusion

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t$$

- Under mild assumptions, converges to the right limiting distribution

$$\|\mathbf{P}(X_t) - \mathbf{P}^\star\|_{\mathsf{tv}} \to 0 \text{ as } t \to \infty \ (p^\star \propto e^{-f})$$

- ULA updates: Forward Euler discretization of Langevin diffusion

$$X_k = X_{k-1} - h\nabla f(X_{k-1}) + \sqrt{2h}\xi_k$$

Langevin algorithms: Origin

- Langevin diffusion

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t$$

- Under mild assumptions, converges to the right limiting distribution

$$\|\mathbf{P}(X_t) - \mathbf{P}^\star\|_{\mathsf{tv}} \to 0 \text{ as } t \to \infty \ (p^\star \propto e^{-f})$$

- ULA updates: Forward Euler discretization of Langevin diffusion

$$X_k = X_{k-1} - h\nabla f(X_{k-1}) + \sqrt{2h}\xi_k$$

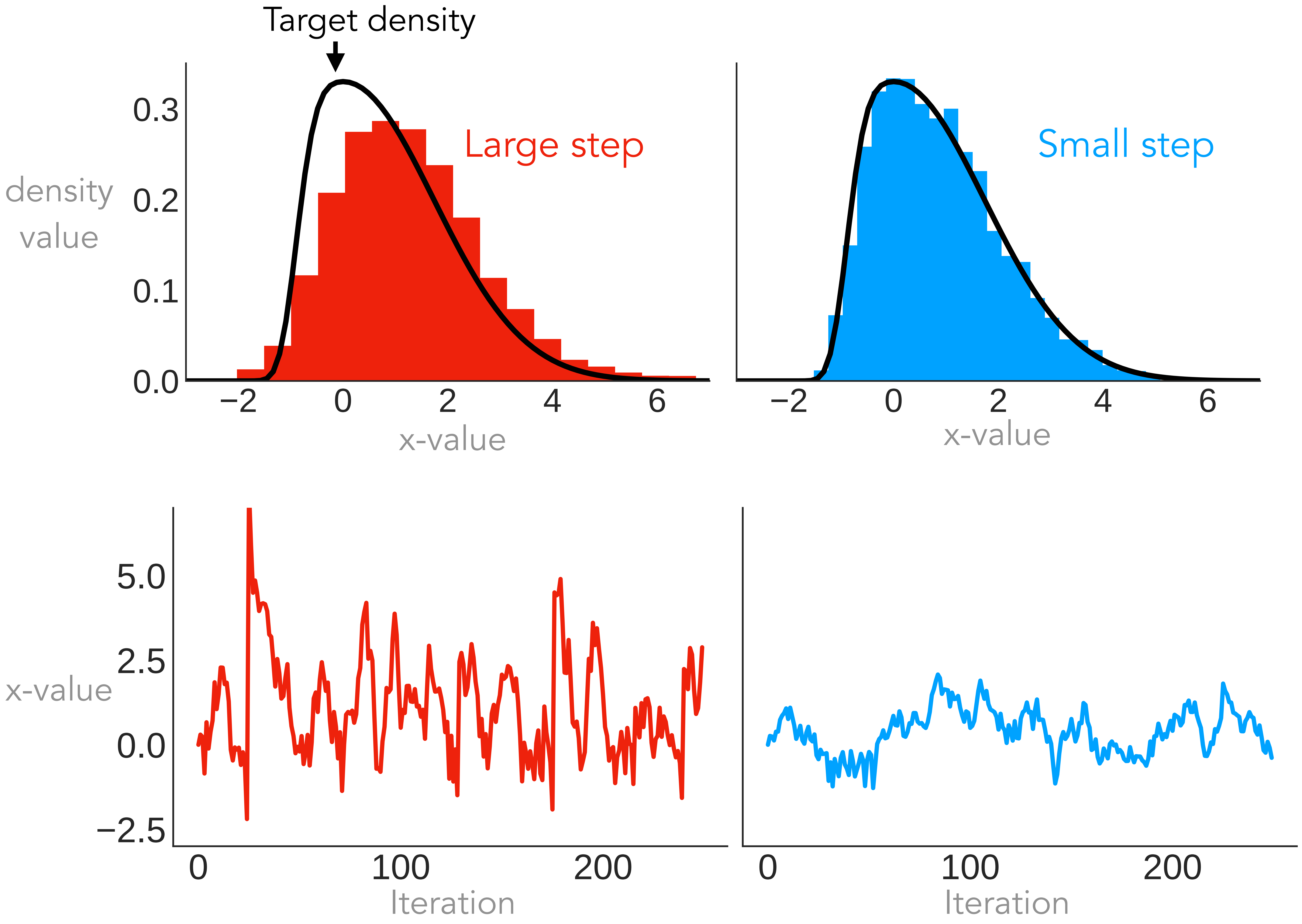
How to choose h?

How many steps to take?

ULA simulation: Trade-offs with step size

Histogram of iterates (upon convergence)

Trace plot of one run

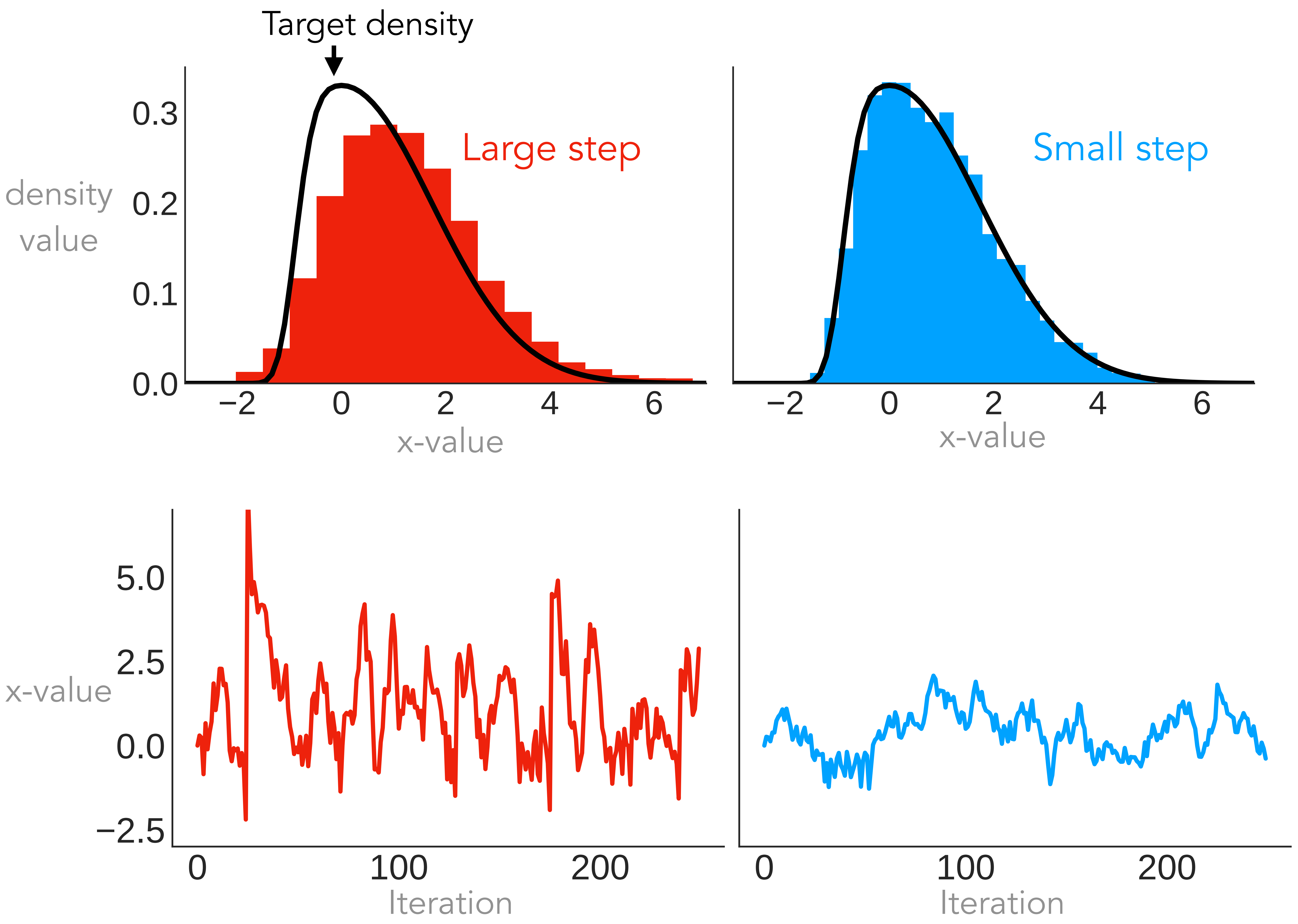


ULA simulation: Trade-offs with step size

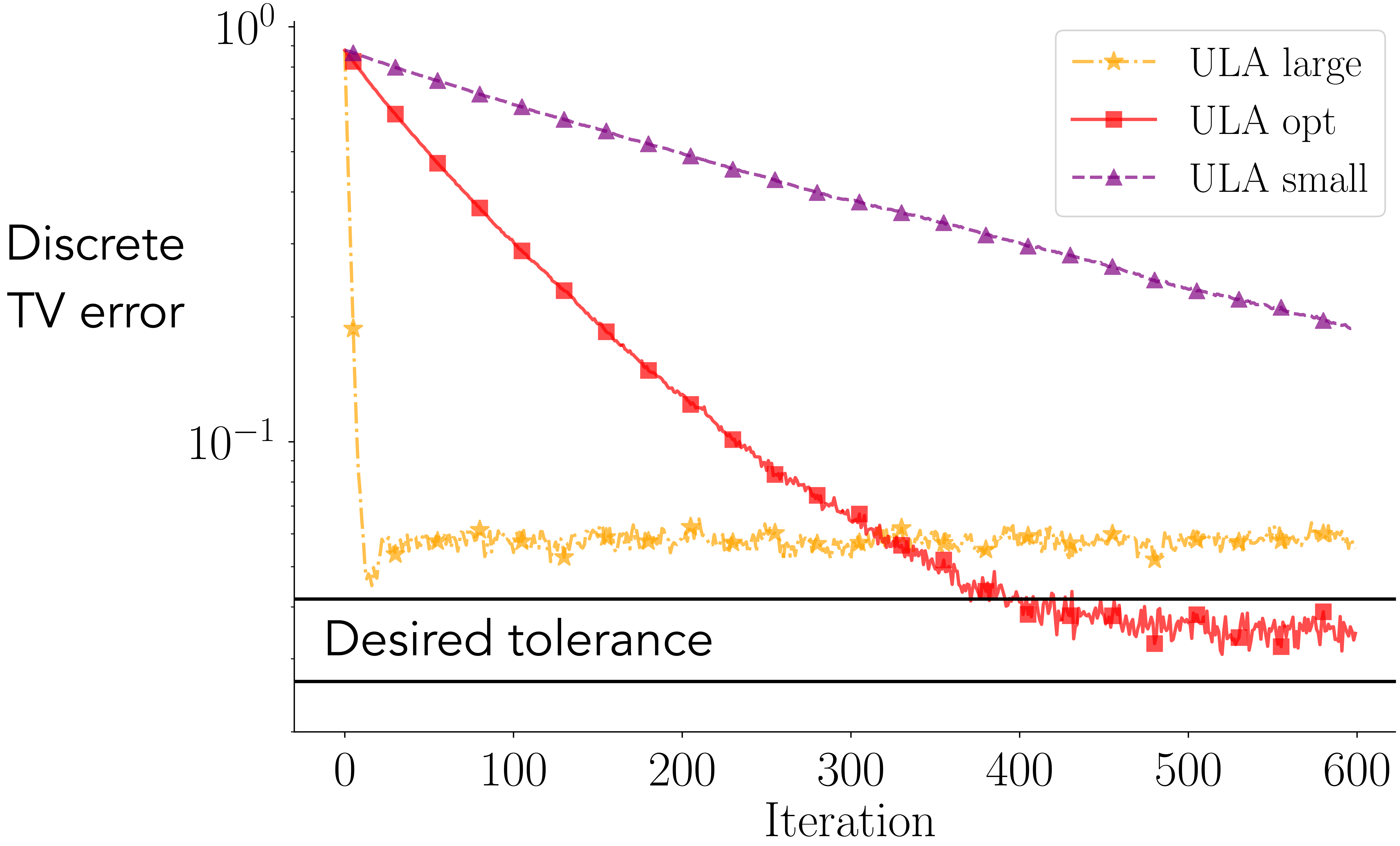
Histogram of iterates (upon convergence)

Large step size ⇒ Large bias & fast mixing

Trace plot of one run



ULA: Bias-mixing trade off with step size



Can we remove the bias? Yes..via accept-reject correction

- Metropolis-adjusted Langevin algorithm (MALA)
  - Use ULA updates as **proposals (Gaussian)**

$$z = x - h\nabla f(x) + \sqrt{2h}\xi$$

  - **Accept** z with probability

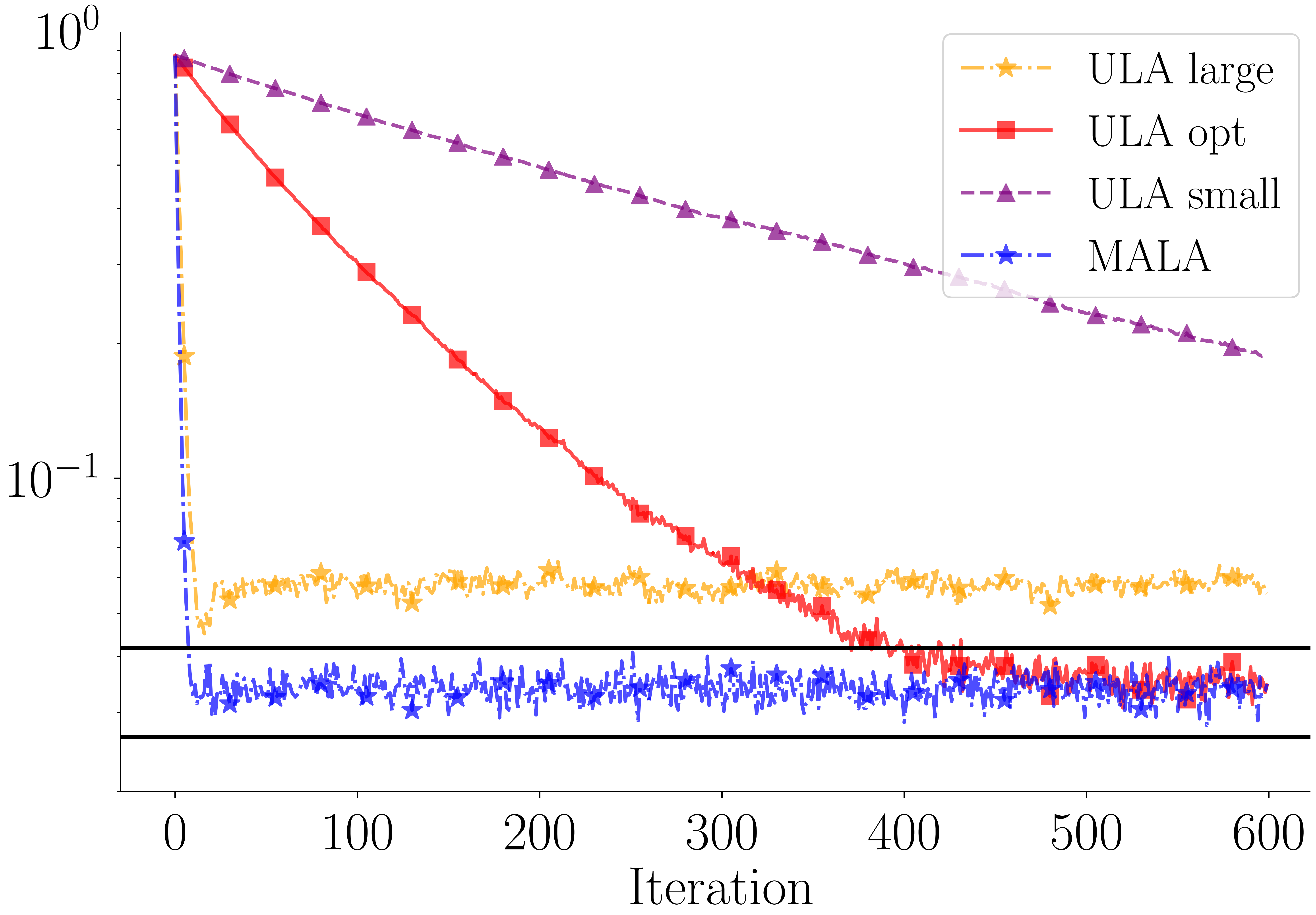
$$\min\left\{1, \frac{e^{-f(z)} \cdot \mathbf{P}_h(z \to x)}{e^{-f(x)} \cdot \mathbf{P}_h(x \to z)}\right\}$$

Ratio of Gaussian proposal densities

  - In case of **rejection**, stay at x

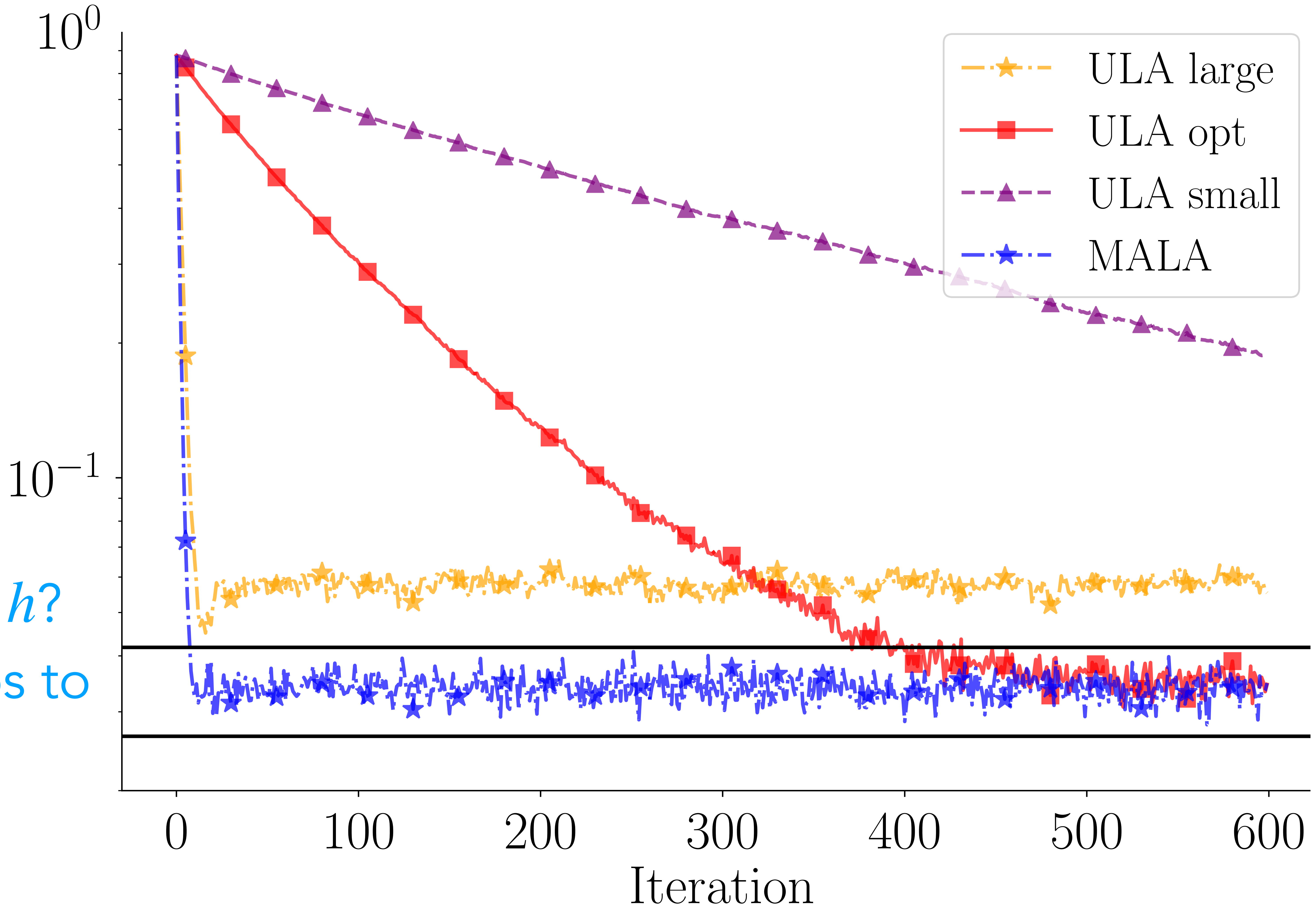
MALA simulation: Fast convergence with no bias

Discrete TV error



MALA simulation: Fast convergence with no bias

Discrete TV error



How to choose h?

How many steps to take?

Several asymptotic and non-explicit guarantees

- Existence, Harris recurrence
  ['95 Meyn-Tweedie, '96 Roberts-Rosenthal, '00 Diaconis-Holmes-Neal,…]

- Weak convergence and diffusion limits as $d \to \infty$
  ['98 Roberts-Rosenthal, '12 Pillai et al., '10 Beskos et al.,…]

- Geometric and uniform ergodicity, Lyapunov coupling
  ['96 Roberts-Tweedie, '04 Roberts-Rosenthal, '09 Bou-Rabee-Hairer, '16 Livingstone et al.,…]

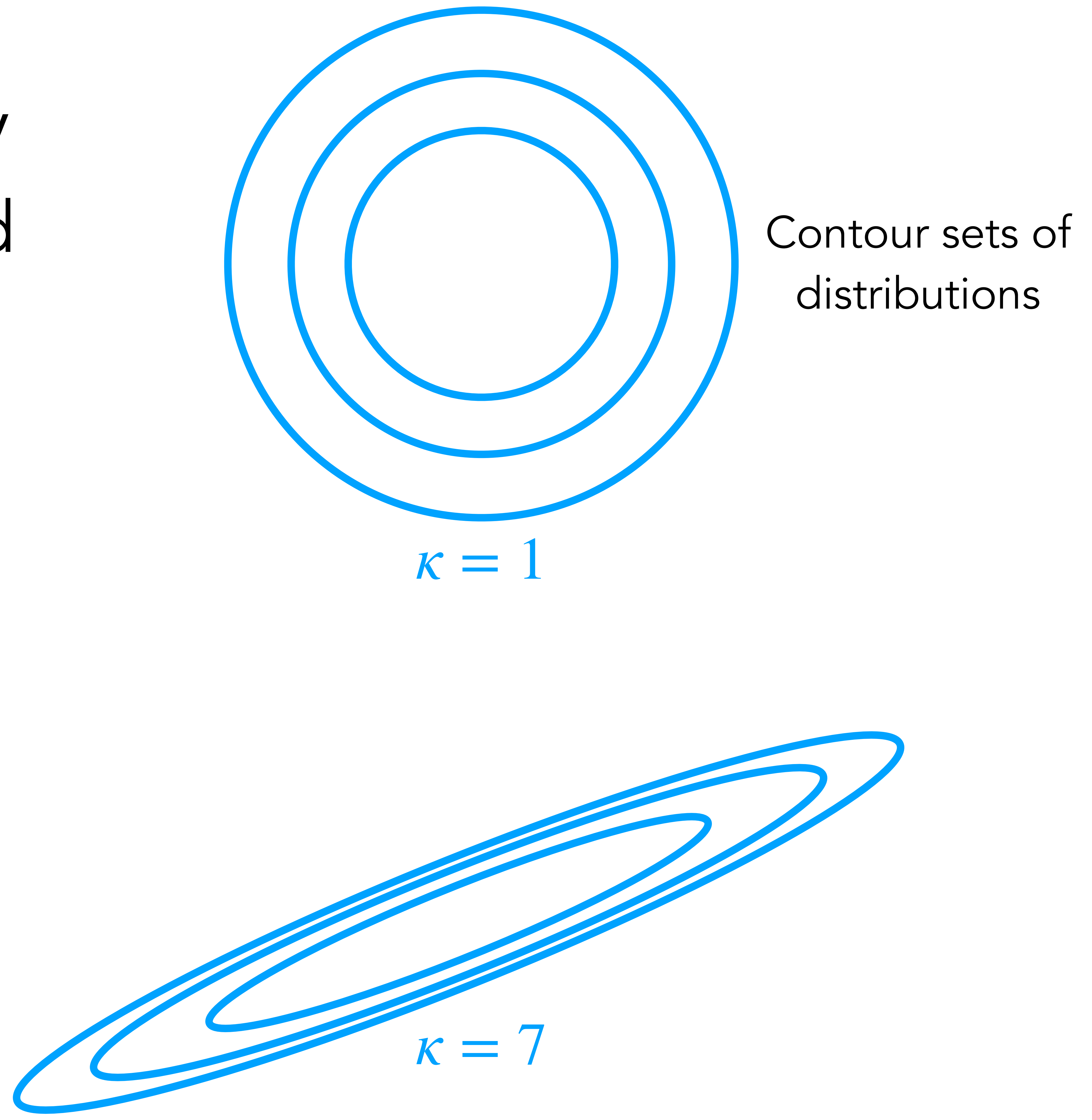
Our goal: Explicit non-asymptotic guarantees

- **Assumption**: Log-concave target density $p^\star \propto e^{-f}$ in $\mathbb{R}^d$ with f strongly convex and smooth

$$m\mathbb{I}_d \preceq \nabla^2 f \preceq L\mathbb{I}_d; \;\; \kappa = L/m$$

- **Mixing-time guarantee**: Bound on iterations T with dimension d, conditioning κ, error δ such that

$$\|\mathbf{P}^\star - \mathbf{P}(X_T)\|_{\mathrm{tv}} \leq \delta$$



Non-asymptotic mixing time for Langevin algorithms

$$\|\mathbf{P}^\star - \mathbf{P}(X_T)\|_{\text{tv}} \leq \delta$$

$p^\star \propto e^{-f}$ with $f : \mathbb{R}^d \to \mathbb{R}$ convex

$$m\mathbb{I}_d \preceq \nabla^2 f \preceq L\mathbb{I}_d; \;\; \kappa = L/m$$

	ULA ['15 Dalalyan]	
Mixing time	$d\kappa^2 \dfrac{\log(1/\delta)}{\delta^2}$	

Non-asymptotic mixing time for Langevin algorithms

$$\|\mathbf{P}^\star - \mathbf{P}(X_T)\|_{\mathrm{tv}} \le \delta$$

$p^\star \propto e^{-f}$ with $f : \mathbb{R}^d \to \mathbb{R}$ convex

$$m\mathbb{I}_d \preceq \nabla^2 f \preceq L\mathbb{I}_d; \;\; \kappa = L/m$$

	ULA ['15 Dalalyan]	MALA [Our work]
Mixing time	$d\kappa^2 \dfrac{\log(1/\delta)}{\delta^2}$	$d\kappa \log(1/\delta)$

Accept-reject helps

- Exponentially better dependence on δ
- Better dependence on κ

Non-asymptotic mixing time for Langevin algorithms

$$\|\mathbf{P}^\star - \mathbf{P}(X_T)\|_{\mathrm{tv}} \leq \delta$$

$p^\star \propto e^{-f}$ with $f : \mathbb{R}^d \to \mathbb{R}$ convex

$$m\mathbb{I}_d \preceq \nabla^2 f \preceq L\mathbb{I}_d; \;\; \kappa = L/m$$

	ULA ['15 Dalalyan]	MALA [Our work]
Mixing time	$d\kappa^2 \frac{\log(1/\delta)}{\delta^2}$	$d\kappa \log(1/\delta)$
Step size	$\frac{\delta^2}{d\kappa L}$	$\frac{1}{dL}$

Accept-reject helps

- Exponentially better dependence on δ
- Better dependence on κ

no bias in MALA allows larger step size and faster mixing

Next: How does gradient information help?

$\|\mathbf{P}^\star - \mathbf{P}(X_T)\|_{\mathrm{tv}} \le \delta$

$p^\star \propto e^{-f}$ with $f : \mathbb{R}^d \to \mathbb{R}$ convex

$m\mathbb{I}_d \preceq \nabla^2 f \preceq L\mathbb{I}_d; \;\; \kappa = L/m$

		Metropolis-adjusted Langevin algorithm (MALA)	
Proposal step		$z = x - h\nabla f(x) + \sqrt{2h}\xi$ one gradient step	
Mixing time		$d\kappa \log(1/\delta)$	
Step size		$\frac{1}{dL}$	

MRW: No gradient leads to slower mixing

$\|\mathbf{P}^\star - \mathbf{P}(X_T)\|_{\text{tv}} \le \delta$

$p^\star \propto e^{-f}$ with $f : \mathbb{R}^d \to \mathbb{R}$ convex

$m\mathbb{I}_d \preceq \nabla^2 f \preceq L\mathbb{I}_d;\ \kappa = L/m$

	Metropolis random walk (MRW)	Metropolis-adjusted Langevin algorithm (MALA)	
Proposal step	$z = x + \sqrt{2h}\xi$ no gradient	$z = x - h\nabla f(x) + \sqrt{2h}\xi$ one gradient step	
Mixing time	$d\kappa^2 \log(1/\delta)$	$d\kappa \log(1/\delta)$	
Step size	$\frac{1}{d\kappa L}$	$\frac{1}{dL}$	

HMC: Multiple gradient steps help mix faster

$\|\mathbf{P}^\star - \mathbf{P}(X_T)\|_{\mathrm{tv}} \leq \delta$ $\quad$ $p^\star \propto e^{-f}$ with $f : \mathbb{R}^d \to \mathbb{R}$ convex

$m\mathbb{I}_d \preceq \nabla^2 f \preceq L\mathbb{I}_d;\ \kappa = L/m$

	Metropolis random walk (MRW)	Metropolis-adjusted Langevin algorithm (MALA)	Metropolis-adjusted Hamiltonian Monte Carlo (HMC)
Proposal step	$z = x + \sqrt{2h}\xi$ no gradient	$z = x - h\nabla f(x) + \sqrt{2h}\xi$ one gradient step	Discretized Hamiltonian dynamics using K gradients per step
Mixing time	$d\kappa^2 \log(1/\delta)$	$d\kappa \log(1/\delta)$	$d^{\frac{2}{3}}\kappa \log(1/\delta)$
Step size	$\frac{1}{d\kappa L}$	$\frac{1}{dL}$	$\frac{1}{d^{\frac{7}{12}}L^{\frac{1}{2}}}$ $(K = d^{\frac{1}{4}})$

Total #gradients = $d^{\frac{11}{12}}\kappa \log(1/\delta)$

HMC: Multiple gradient steps help mix faster

$\|\mathbf{P}^\star - \mathbf{P}(X_T)\|_{\mathrm{tv}} \leq \delta$

$p^\star \propto e^{-f}$ with $f : \mathbb{R}^d \to \mathbb{R}$ convex

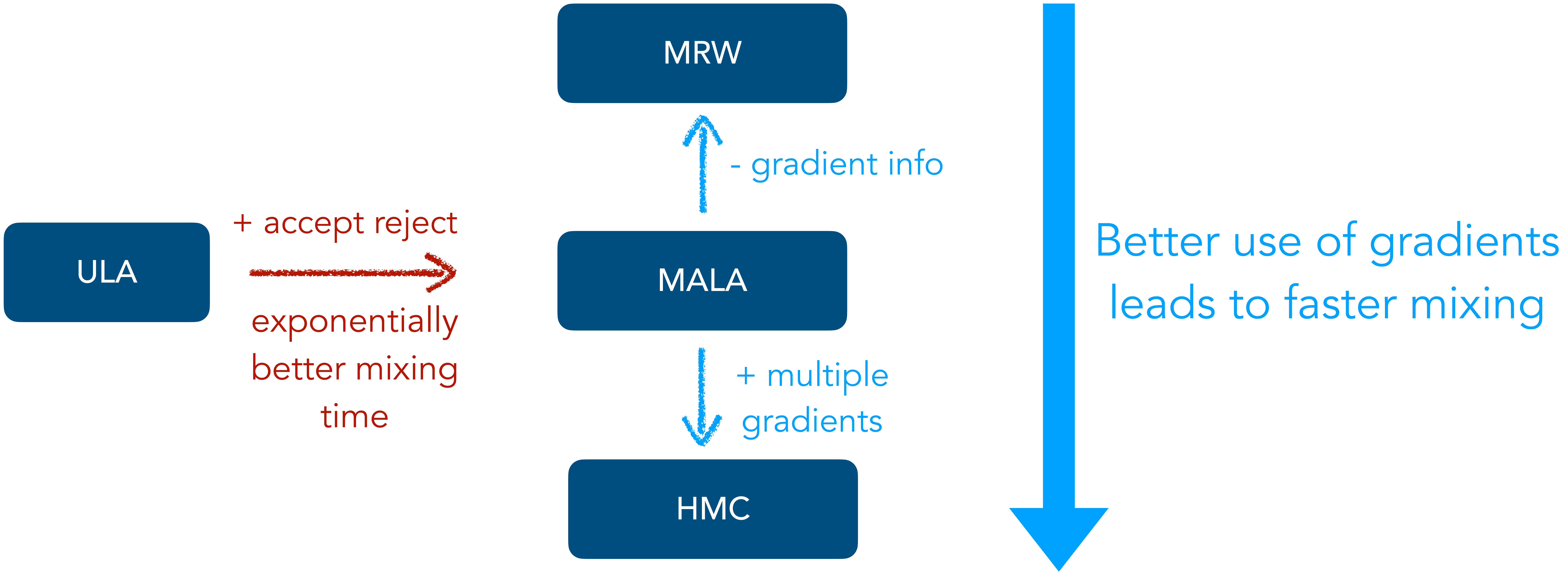
$m\mathbb{I}_d \preceq \nabla^2 f \preceq L\mathbb{I}_d;\ \kappa = L/m$

	Metropolis random walk (MRW)	Metropolis-adjusted Langevin algorithm (MALA)	Metropolis-adjusted Hamiltonian Monte Carlo (HMC)
Proposal step	$z = x + \sqrt{2h}\xi$ no gradient	$z = x - h\nabla f(x) + \sqrt{2h}\xi$ one gradient step	Discretized Hamiltonian dynamics using K gradients per step
Mixing time	$d\kappa^2 \log(1/\delta)$	$d\kappa \log(1/\delta)$	$d^{\frac{2}{3}}\kappa \log(1/\delta)$
Step size	$\frac{1}{d\kappa L}$	$\frac{1}{dL}$	$\frac{1}{d^{\frac{7}{12}}L^{\frac{1}{2}}}$ $(K = d^{\frac{1}{4}})$

Total #gradients = $d^{\frac{11}{12}}\kappa \log(1/\delta)$

Previous HMC bounds either worse than MALA or had $1/\delta^2$ dependence due to no accept-reject step

Summary of MCMC guarantees



Refs: **1.** Log-concave sampling: Metropolis-Hastings algorithms are fast
[**Dwivedi***-Chen*-Wainwright-Yu, JMLR '19]
2. Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients
[Chen-**Dwivedi**-Wainwright-Yu, JMLR '20]