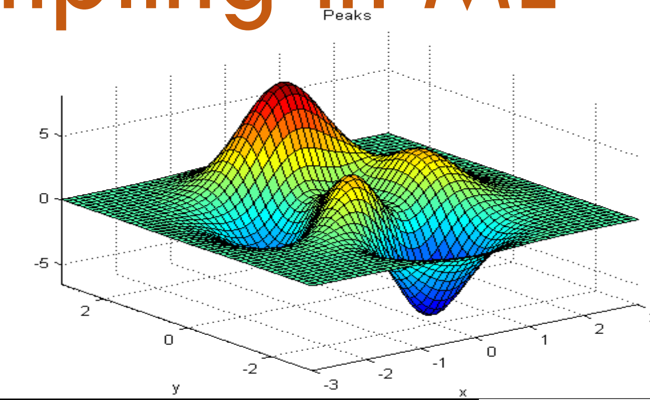
PHYSICS-INSPIRED ALGORITHMS

Nisheeth Vishnoi

Yale

Optimization and Sampling in ML



Given access to $f: \mathbb{R}^d \to \mathbb{R}$

Optimize $\min_{\theta} f(\theta)$

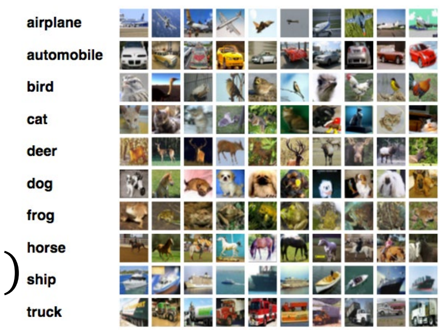
Sample θ with prob. $\propto e^{-f(\theta)}$

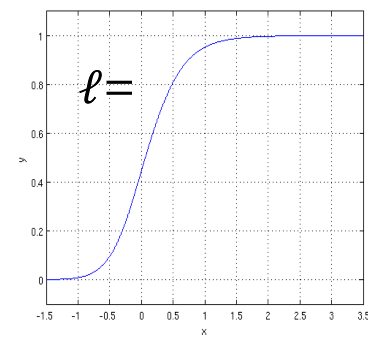
Typically, harder than optimization – but ***robust***

Availability of **large, real-world datasets** has given rise to complex objective functions in high dimensions

e.g. $f(\theta) = \sum_i \ell(\theta; x_i, y_i)$

$$f(\theta) = -\sum_i y_i \log \ell(\theta^\top x_i) + (1 - y_i) \log \ell(-\theta^\top x_i)$$





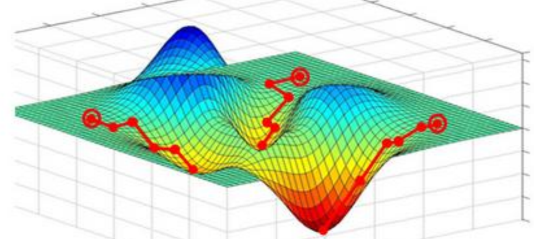
Two facets:

1. Develop methods

(associate a physical meaning and search for the right equations of motion)

2. Prove guarantees, tune parameters

(search for potential functions, "beyond worst case" assumptions on data)



$$\theta^{t+1} = \theta^t + \eta^t G_f(\theta^t)$$

Physics viewpoint has been helpful in both!

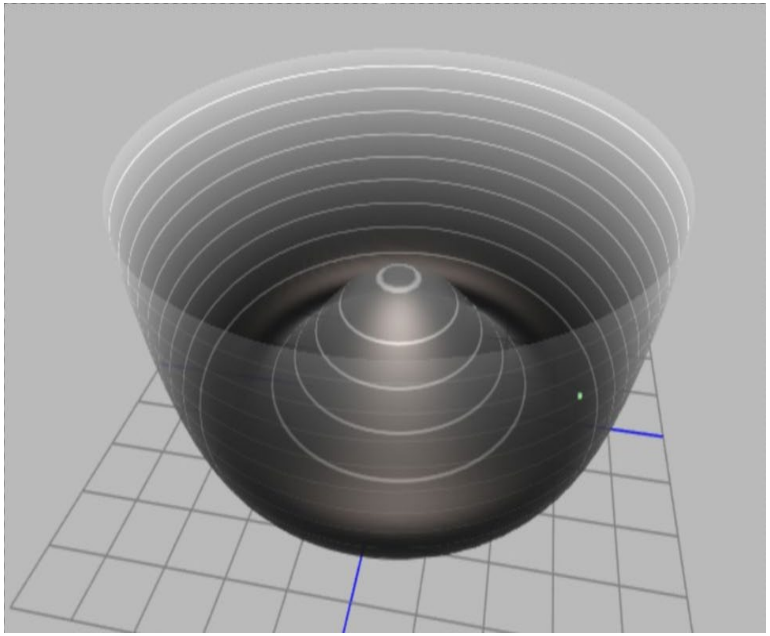
HAMILTONIAN DYNAMICS
&
SAMPLING
from
CONTINUOUS DISTRIBUTIONS

Sampling from Continuous Distributions

Given access to $f: \mathbb{R}^d \to \mathbb{R}$

Sample $\boldsymbol{\theta}$ with prob. $\pi(\theta) \propto e^{-f(\theta)}$

Statistics, TCS, Optimization (vis annealing), Bayesian inference, Molecular dynamics ..



Iterative methods: MCMC+Metropolis

Propose: $\theta^{k+1} = \theta^k + G_f(\theta^k)$

Accept/Reject

Number of gradient (or function) evaluations to sample from smooth, strongly logconcave π (for smoothness/convexity $= \Theta(1)$):

- Random Walk Metropolis: d^2 [Gelman et al. '97]
- Unadjusted Langevin: d [Durmus, Moulines, '16]
- Underdamped Langevin: $d^{1/2}$ [Cheng et al. '17]

Beyond $d^{1/2}$?

Hamiltonian Monte Carlo

[Duane et al. '87] ***No Accept/Reject step!***

Define: $H(\theta, v) = f(\theta) + \frac{1}{2}\|v\|^2$

In step i, sample $V_i \sim N(0, I_d)$

Obtain Θ_{i+1} by simulating *Hamiltonian Dynamics* starting at (Θ_i, V_i) for time T

$$\frac{d\theta(t)}{dt} = v(t)$$

$$\frac{dv(t)}{dt} = -\nabla f(\theta(t))$$

Fact: Invariant distribution $\propto e^{-f(\theta)} e^{-\frac{1}{2}\|v\|^2}$

2nd-order Leapfrog integrator

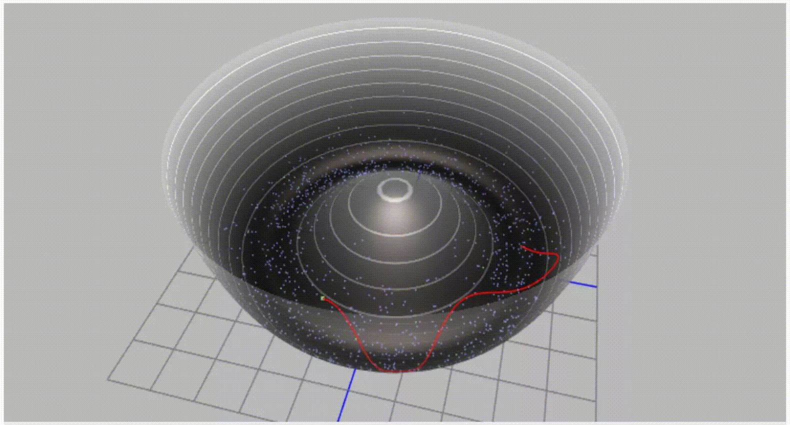
Let $(\theta_0, v_0) = (\Theta_i, V_i)$

For j $= 0, \ldots, \frac{T}{\eta} - 1$, **do**

$$\theta_{j+1} = \theta_j + \eta v_j - \frac{1}{2}\eta^2 \nabla f(\theta_j)$$

$$v_{j+1} = v_j - \eta \nabla f(\theta_j) - \frac{1}{2}\eta^2 \frac{\nabla f(\theta_{j+1}) - \nabla f(\theta_j)}{\eta}$$

$\Theta_{i+1} = \theta_{\frac{T}{\eta}}$

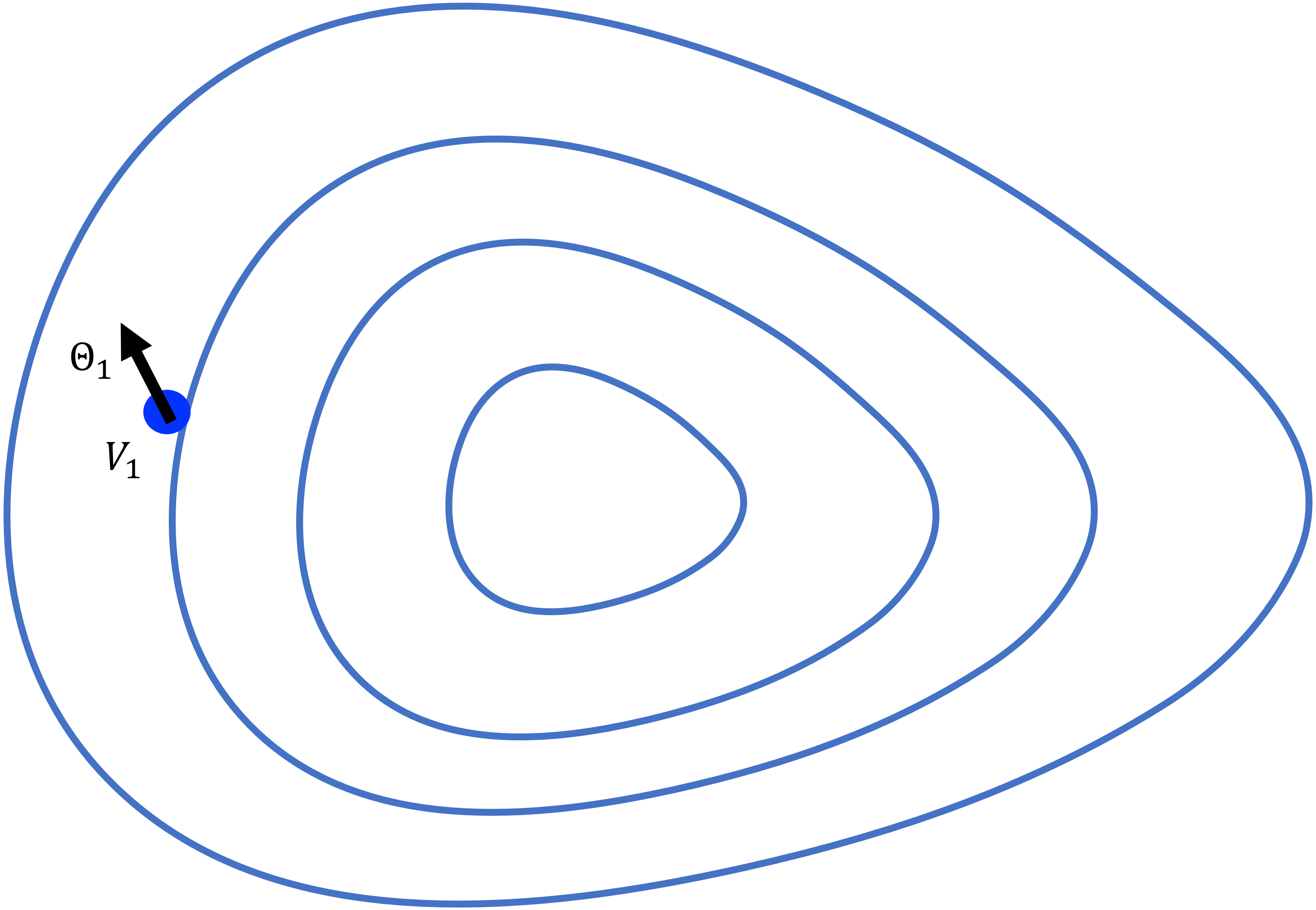


Widely deployed in practice – convergence bounds/tuning parameters?

(Informal) Conjecture: [Creutz, 1988]

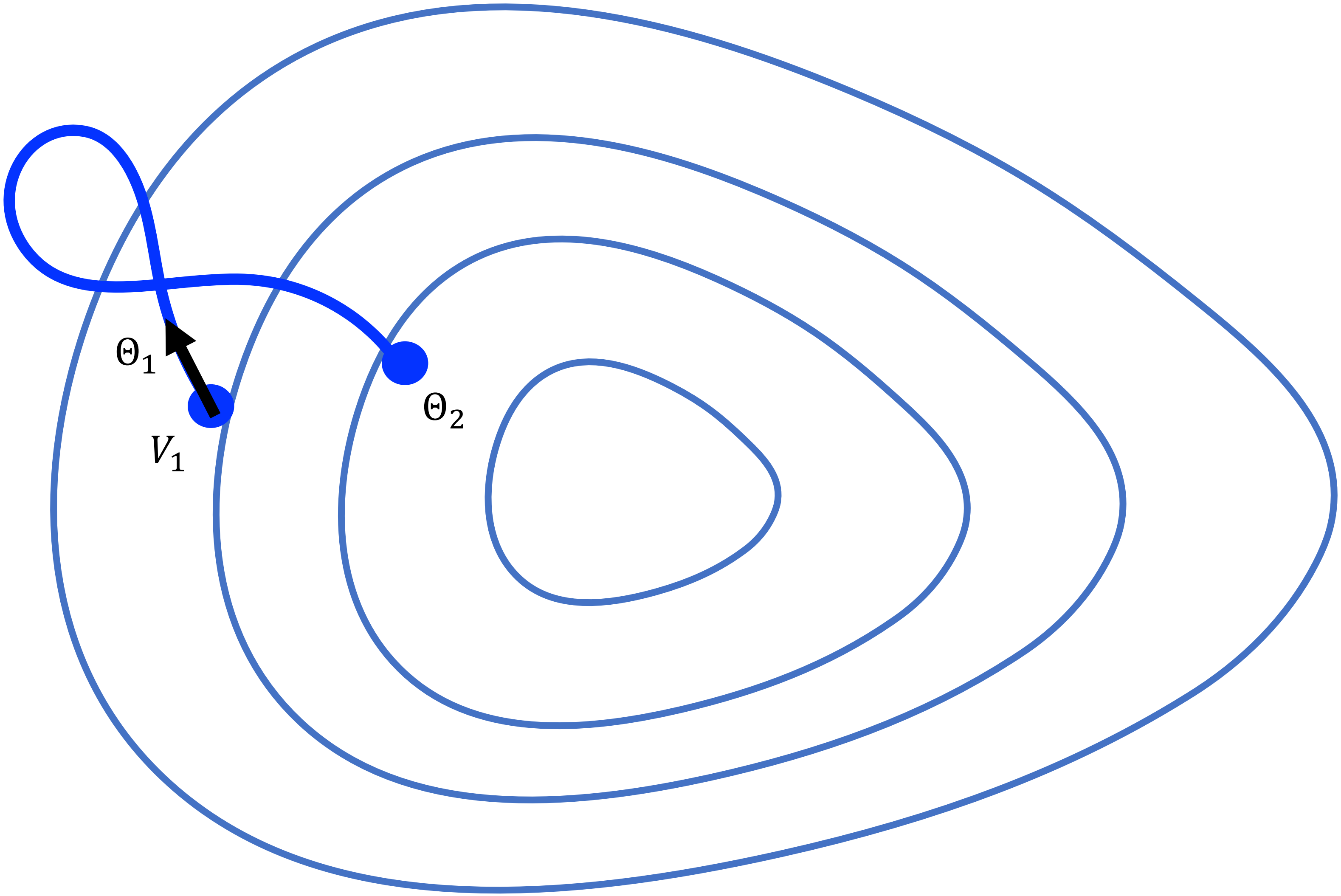
$d^{1/4}$ gradient evaluations are sufficient for 2nd-order HMC to sample from O(1)-smooth, O(1)-strongly convex π with bounded higher-order derivatives

HMC



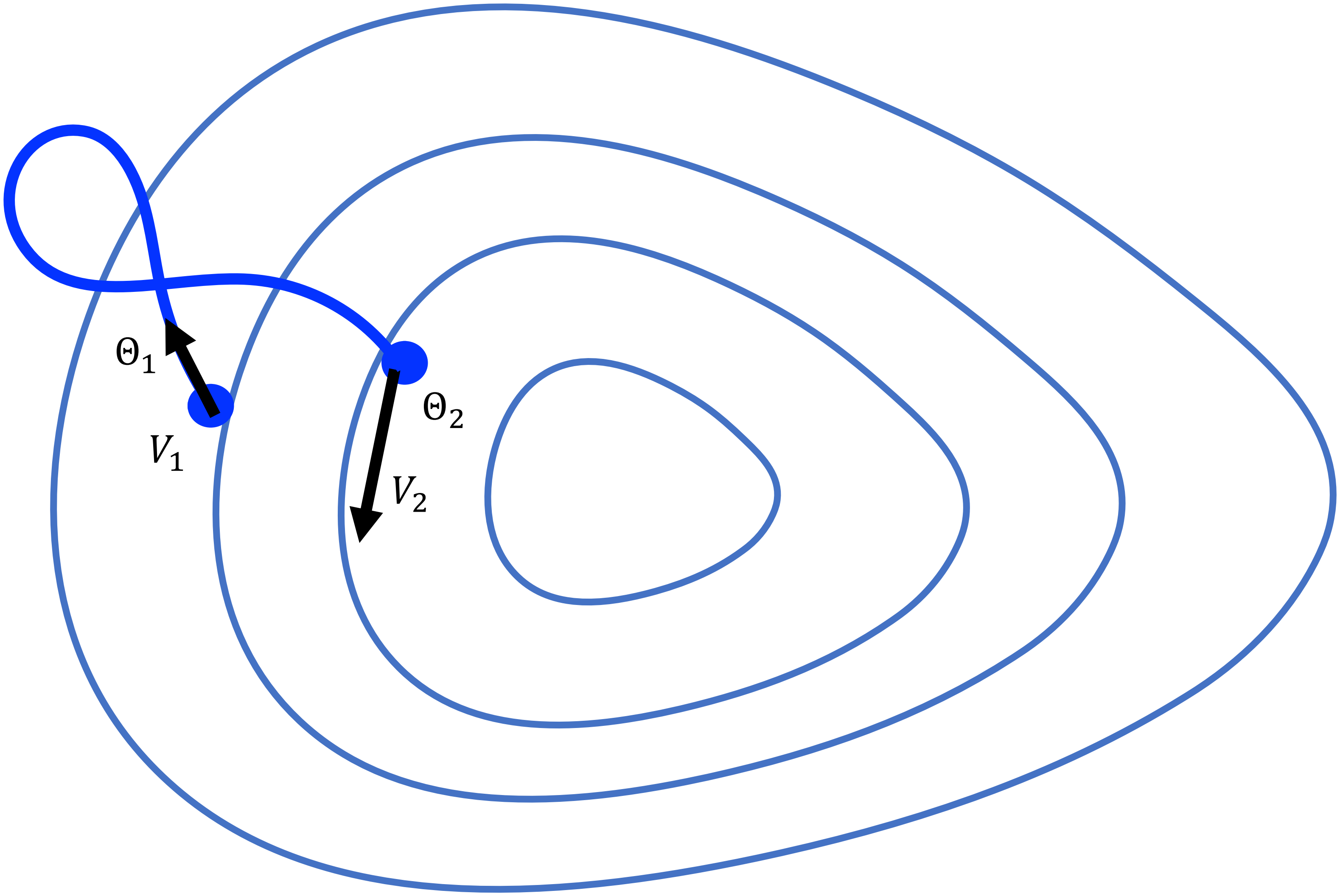
Step 1: sample $V_1 \sim N(O, I_d)$

HMC



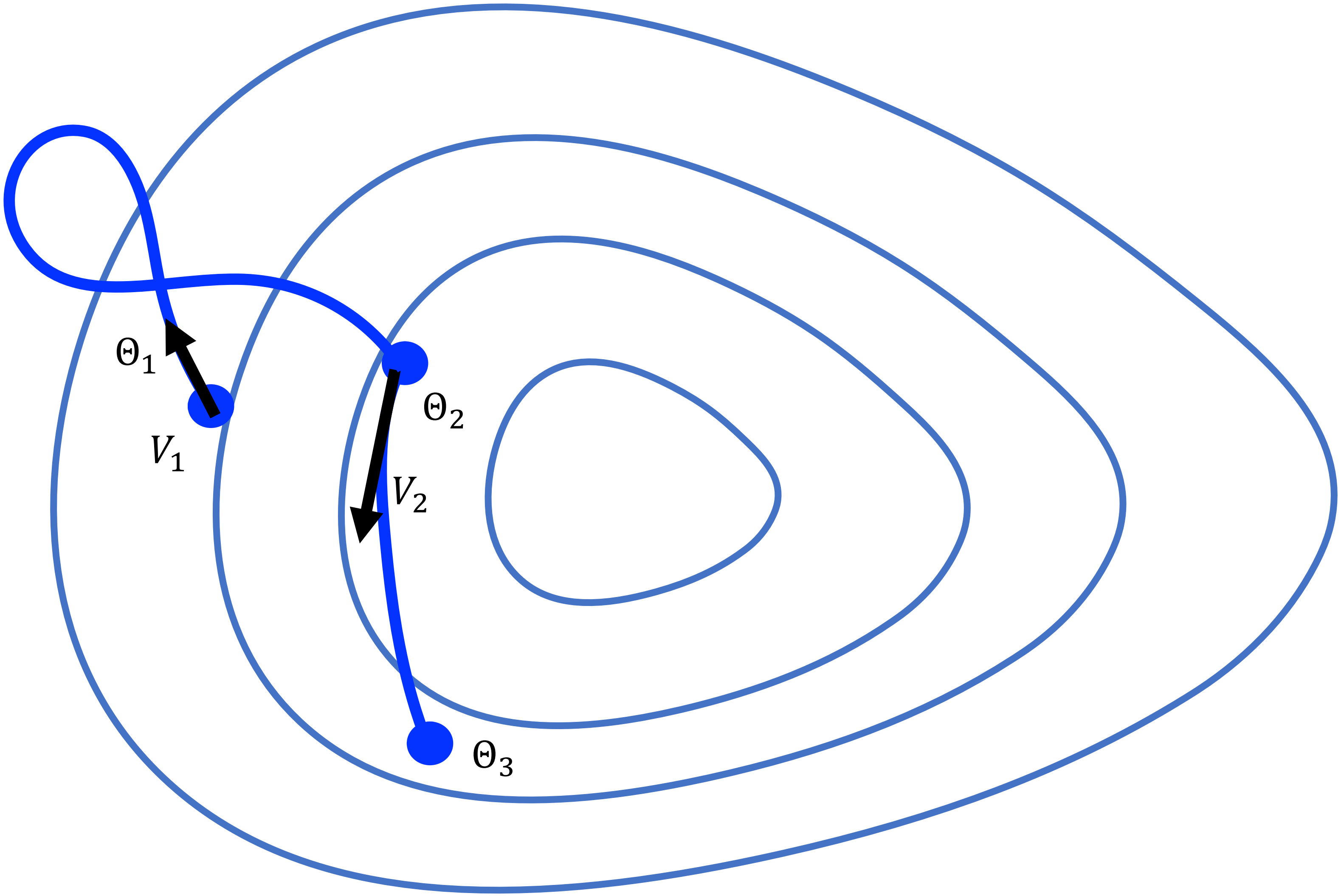
Step 2: Compute Hamiltonian trajectory for fixed time T

HMC



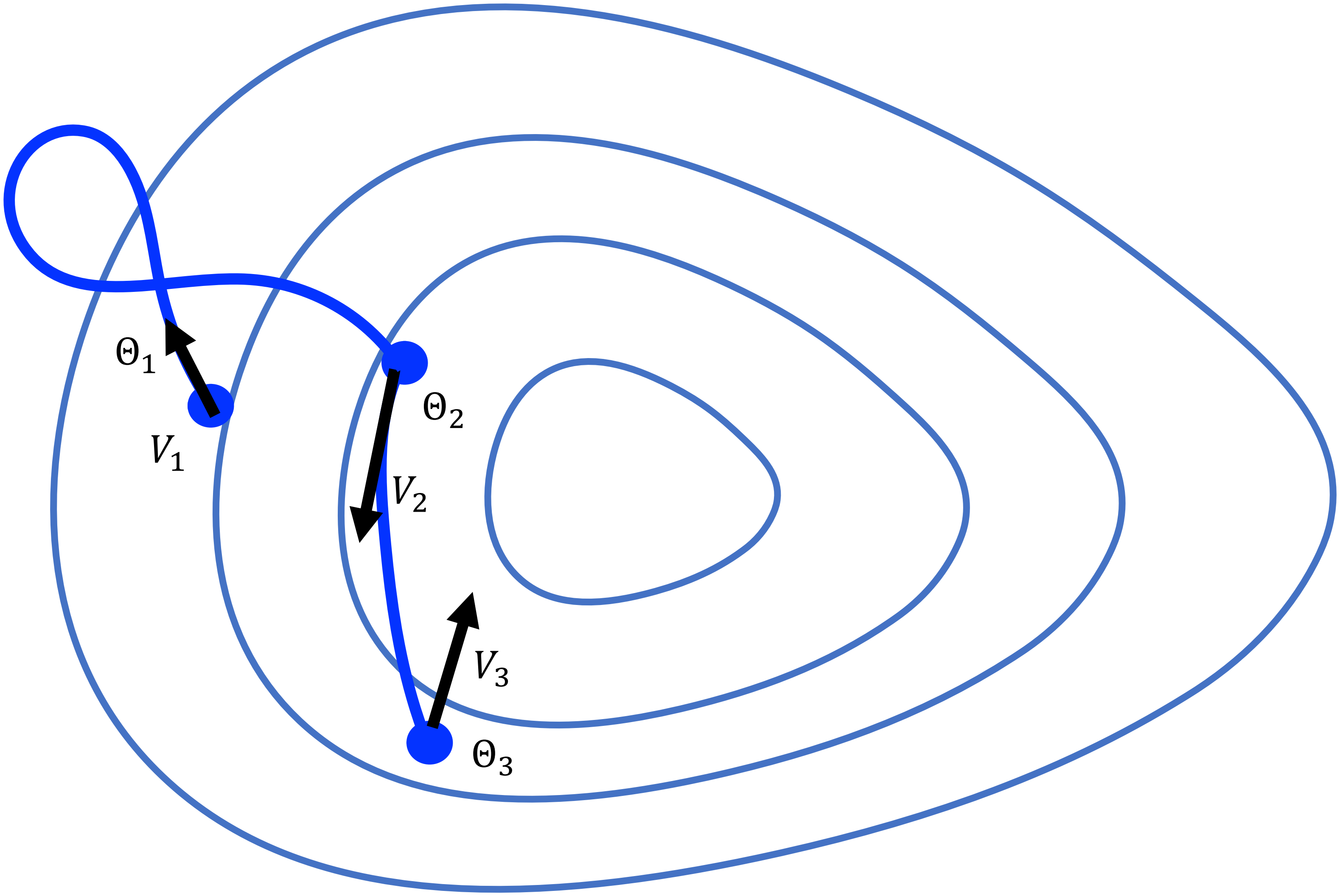
Step 3: Throw out old momentum and sample new independent momentum

HMC



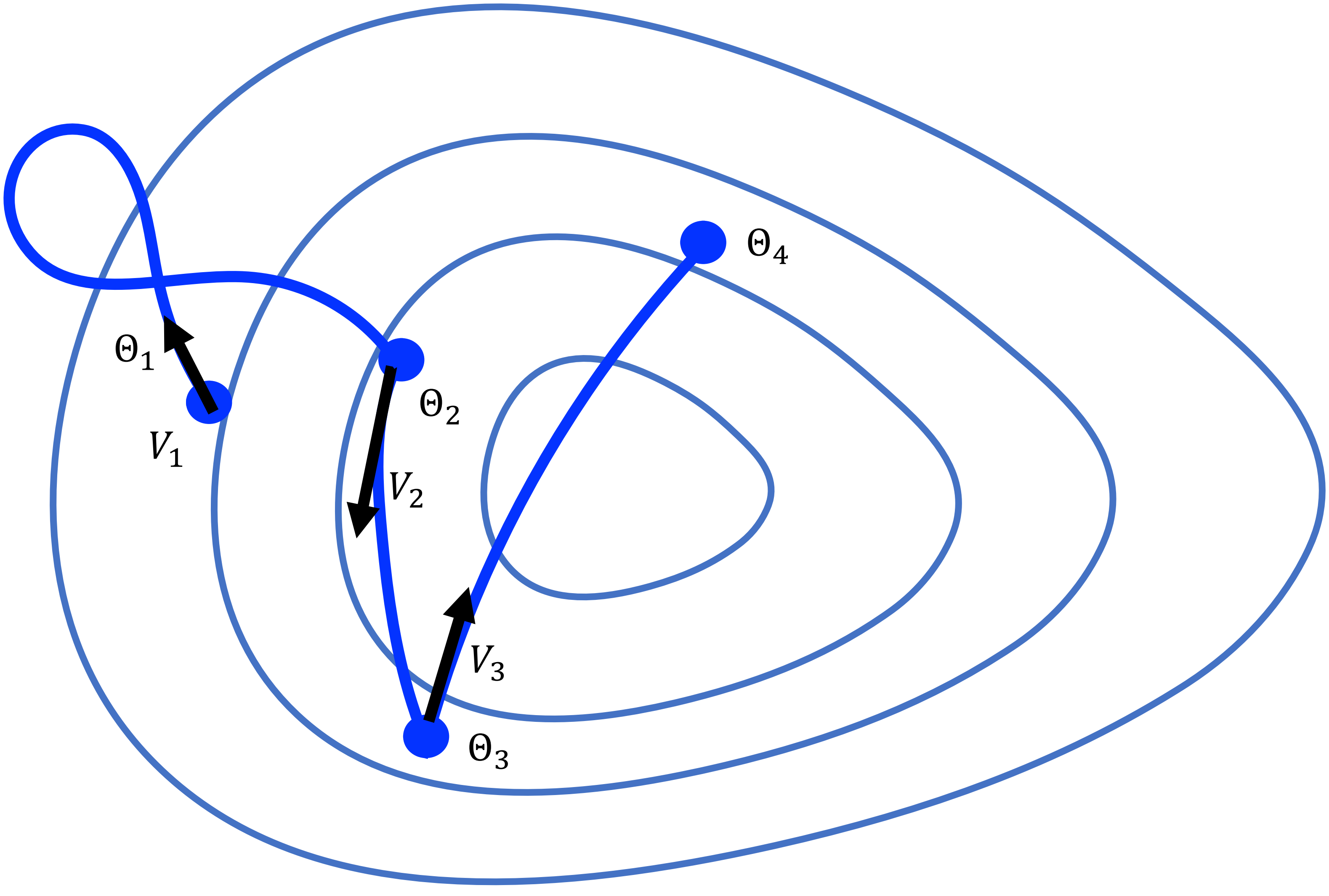
steps 4,5,…: iteratively repeat steps 1 and 2

HMC

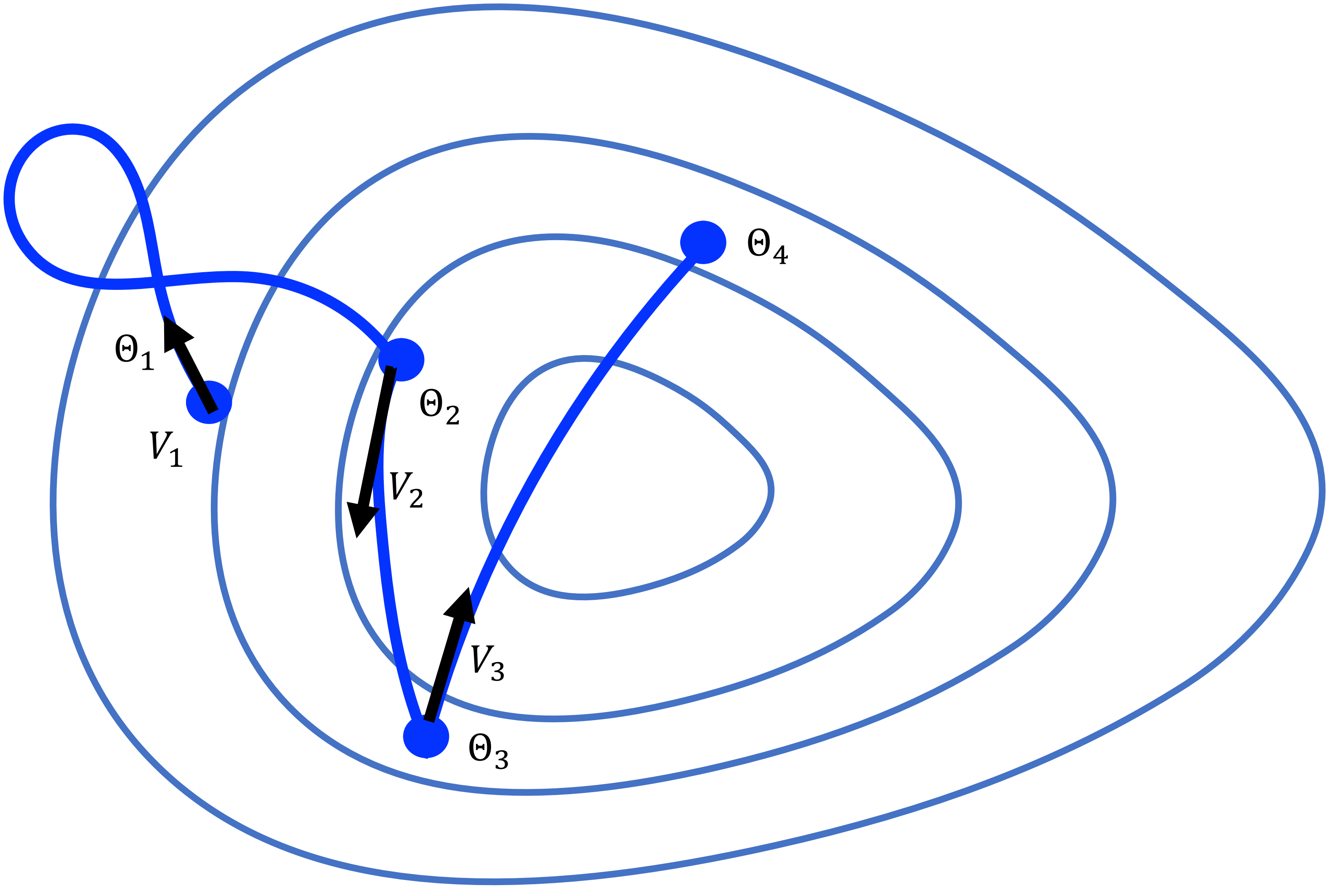


Steps 4,5,…: iteratively repeat steps 1 and 2

HMC



Steps 4,5,…: iteratively repeat steps 1 and 2

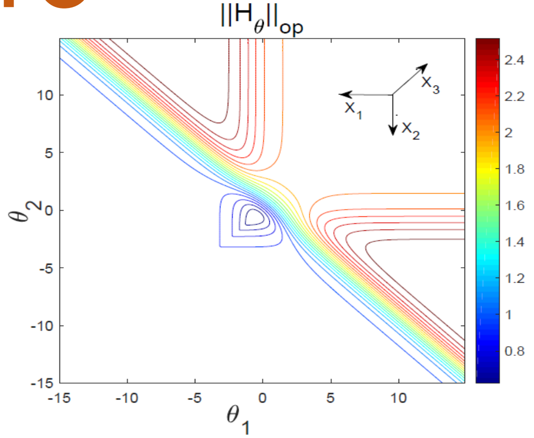


Θ_1
V_1
Θ_2
V_2
V_3
Θ_3
Θ_4

Confirming Creutz's Conjecture

[Mangoubi-V. NeurIPS '18] Strongly convex f + *regularity conditions*

HMC with Leapfrog Integrator requires (roughly) $d^{1/4}$ gradient evaluations



Bit more formally: *Suppose that*

1. $\frac{1}{10} I \preccurlyeq \nabla^2 f(\theta) \preccurlyeq 10\, I$
2. $\nabla^2 f$ satisfies a Lipschitz condition for $L_\infty, r > 0$ and $x_1, \ldots, x_r \in \mathbb{S}^d$:

$\left\|\left(\nabla^2 f(\theta_1) - \nabla^2 f(\theta_2)\right) v\right\|_2 \leq L_\infty \left\|\mathrm{X}^\top(\theta_1 - \theta_2)\right\|_\infty \times \left\|\mathrm{X}^\top v\right\|_\infty$,

where $\mathrm{X} := [x_1, \ldots, x_r]$

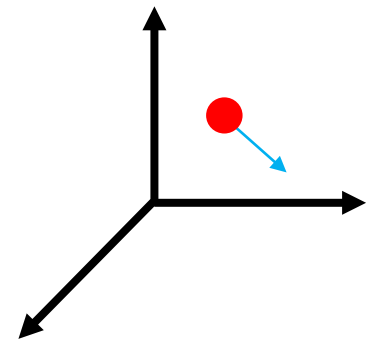
Then Leapfrog HMC requires $\tilde{O}(\max\left(d^{\frac{1}{4}}, \sqrt{L_\infty}\right) \varepsilon^{-1/2})$ gradient calls to obtain a sample ε-close (in Wasserstein-2 metric) to π

Application of our result: *Fast sampling from logistic "ridge" regression*

$$f(\theta) = \|\theta\|^2 - \sum_i y_i \log \ell(\theta^\top x_i) + (1 - y_i) \log \ell(-\theta^\top x_i)$$

$$L_\infty = \sqrt{C}, \text{ where } \textbf{\textit{coherence}}\ C := \max_{i \in [r]} \sum_{j=1}^{r} \left|x_i^\top x_j\right|$$

Hamiltonian Dynamics



$\theta(t), v(t) \in \mathbb{R}^d$

Setting:

- Particle with position $\theta(t)$ and momentum/velocity $v(t)$
- Moves according to classical physics laws in a potential well f

Hamiltonian: $H(\theta, v) = f(\theta) + \frac{1}{2}\|v\|^2$

Hamilton's Equations:

- ***Momentum:*** $\frac{d\theta(t)}{dt} = \frac{\partial H}{\partial v} = v(t)$

- ***Force:*** $\frac{dv(t)}{dt} = -\frac{\partial H}{\partial \theta} = -\nabla f(\theta(t))$

Properties:

- **Time Reversible**

- **Preserves Hamiltonian (Energy):**

$$\frac{dH}{dt} = \sum_i \frac{d\theta_i}{dt}\frac{\partial H}{\partial \theta_i} + \frac{dv_i}{dt}\frac{\partial H}{\partial v_i} = 0$$

- **Preserves Volume:**

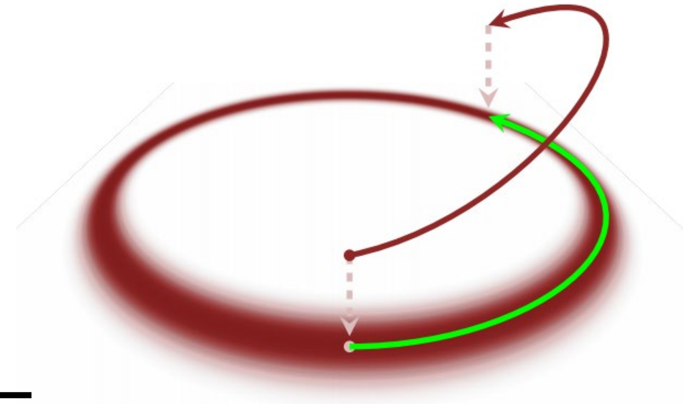
Vector field F in $\mathbb{R}^d \times \mathbb{R}^d$ at (θ, v)

$$\frac{d\theta}{dt}, \frac{dv}{dt}$$

Check: $div\ F = \sum_i \frac{\partial}{\partial \theta_i}\frac{d\theta_i}{dt} + \frac{\partial}{\partial v_i}\frac{dv_i}{dt} = 0$

Correctness of continuous-time HMC

Correct: Time reversible, energy-preserving, volume preserving (in "phase space")



Proof: Two steps in the HMC chain. Sufficient to that $e^{-H(\cdot,\cdot)}$ is invariant

Refresh Velocity: Only v is changing, independent of θ and sampled from the right marginal. Hence, $e^{-H(\theta,v)} = e^{-f(\theta)}e^{-\frac{1}{2}\|v\|^2}$ is invariant

Simulate Hamiltonian dynamics:
Partition the phase space into infinitesimal cubes and let C be one cube and (θ, v) be a point in C. The probability of being in C is proportional to $e^{-H(\theta,v)} \times \mathrm{vol}(C)$

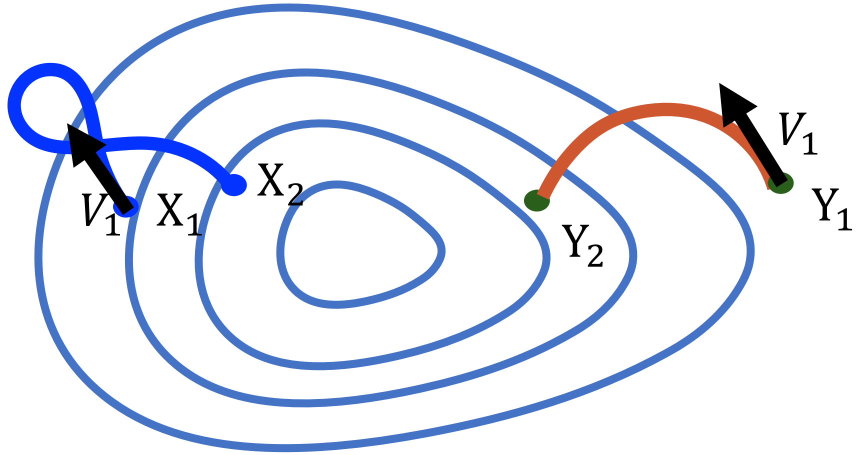
Since the Hamiltonian flow conserves the Hamiltonian (energy) and the volume, the probability of being in the image of C is also conserved (uses *time-reversibility* of Hamiltonian dynamics)

Coupling Bounds for Idealized HMC

- Two chains $X_1, X_2, \dots$ and $Y_1, Y_2, \dots$ with same transition kernel
- Choose a coupling so that $\|X_{i+1} - Y_{i+1}\| \leq c\, \|X_i - Y_i\|$ for $c < 1$

Momentum (velocity) allows HMC to take long steps …

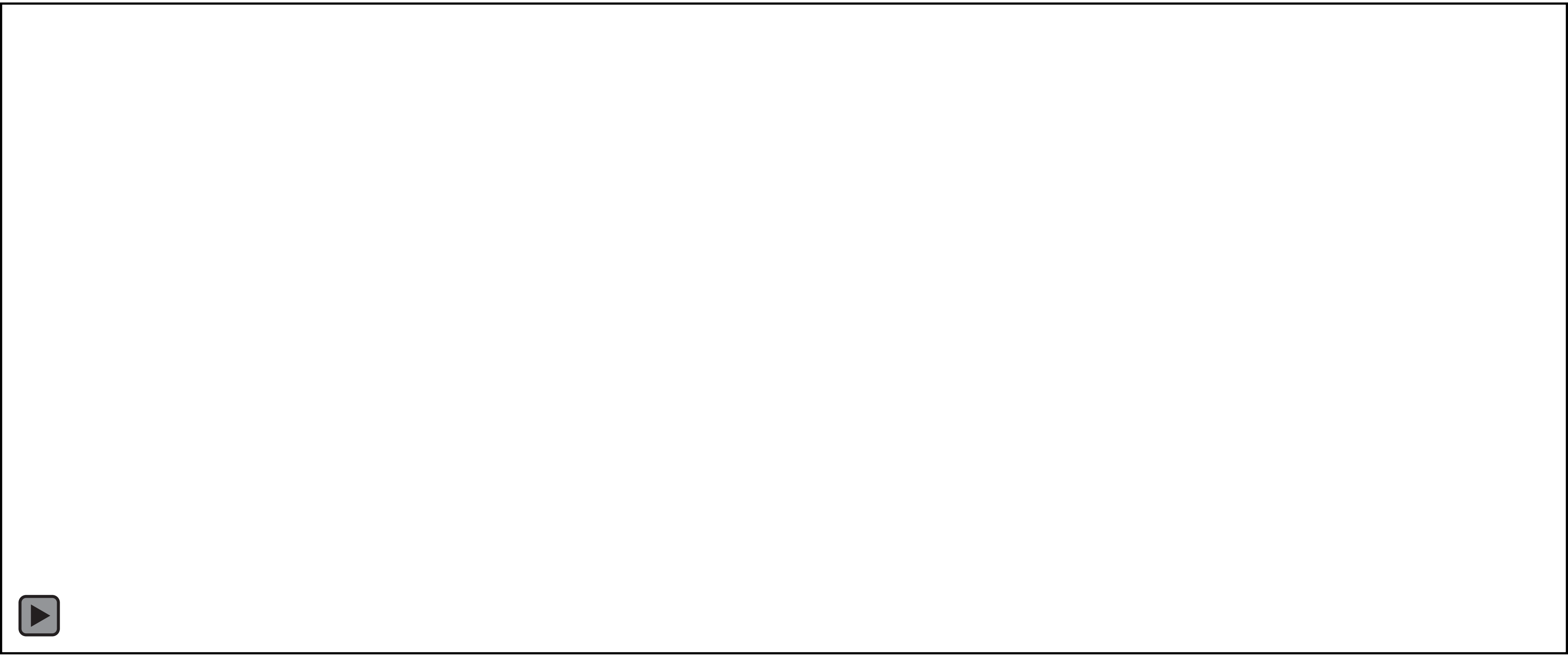
Can we couple the momentum of the two (ideal) HMC chains in a way that leads to large (dimension independent) contractions over these long steps?



Hamiltonian trajectories contract for strongly convex potentials; c independent of d

Exercise: $f(\theta) = \sum_i c_i \theta_i^2, \frac{1}{10} \leq c_i \leq 10$

Example: Coupled Pendulums



- pendulums kicked with the same initial velocity
- distance between pendulums contracts for a long time
- difference between velocities increases during this time

Leapfrog Integrator

2nd-order Leapfrog integrator

For j $= 0, \ldots, \frac{T}{\eta} - 1$, do

$$\theta_{j+1} = \theta_j + \eta v_j - \frac{1}{2}\eta^2 \nabla f(\theta_j)$$

$$v_{j+1} = v_j - \eta \nabla f(\theta_j) - \frac{1}{2}\eta^2 \frac{\nabla f(\theta_{j+1}) - \nabla f(\theta_j)}{\eta}$$

$$\approx \eta^2 \nabla^2 f(\theta_j) v_j = \eta^2 H(\theta_j) v_j$$

- ***Symplectic integrator***: Approximately conserves target measure
    - volume is conserved in phase space
    - a *perturbed* Hamiltonian is conserved
- Only one gradient call per iteration

Bound numerical error for a given discretization η?

Our Lipschitz Hessian Condition

Suppose "Euclidean Lipschitz" Hessian

$$\left\|\left(H(\theta_1) - H(\theta_2)\right)v\right\|_2 \le L_2 \cdot \|\theta_1 - \theta_2\|_2 \cdot \|v\|_2$$

Turns out that numerical error: $\left\|\eta\left(H(\theta + \eta v) - H(\theta)\right)v\right\|_2 \le L_2 \cdot \eta^2 \cdot \|v\|_2^2$

Here $v \sim N(0, I_d)$, so $\|v\|_2 \approx \sqrt{d}$, so $\eta \approx 1/\sqrt{d}$ leading to no better than $\sqrt{d}$ bound!

Idea: Use a different norm ..

Infinity Lipschitz Hessian

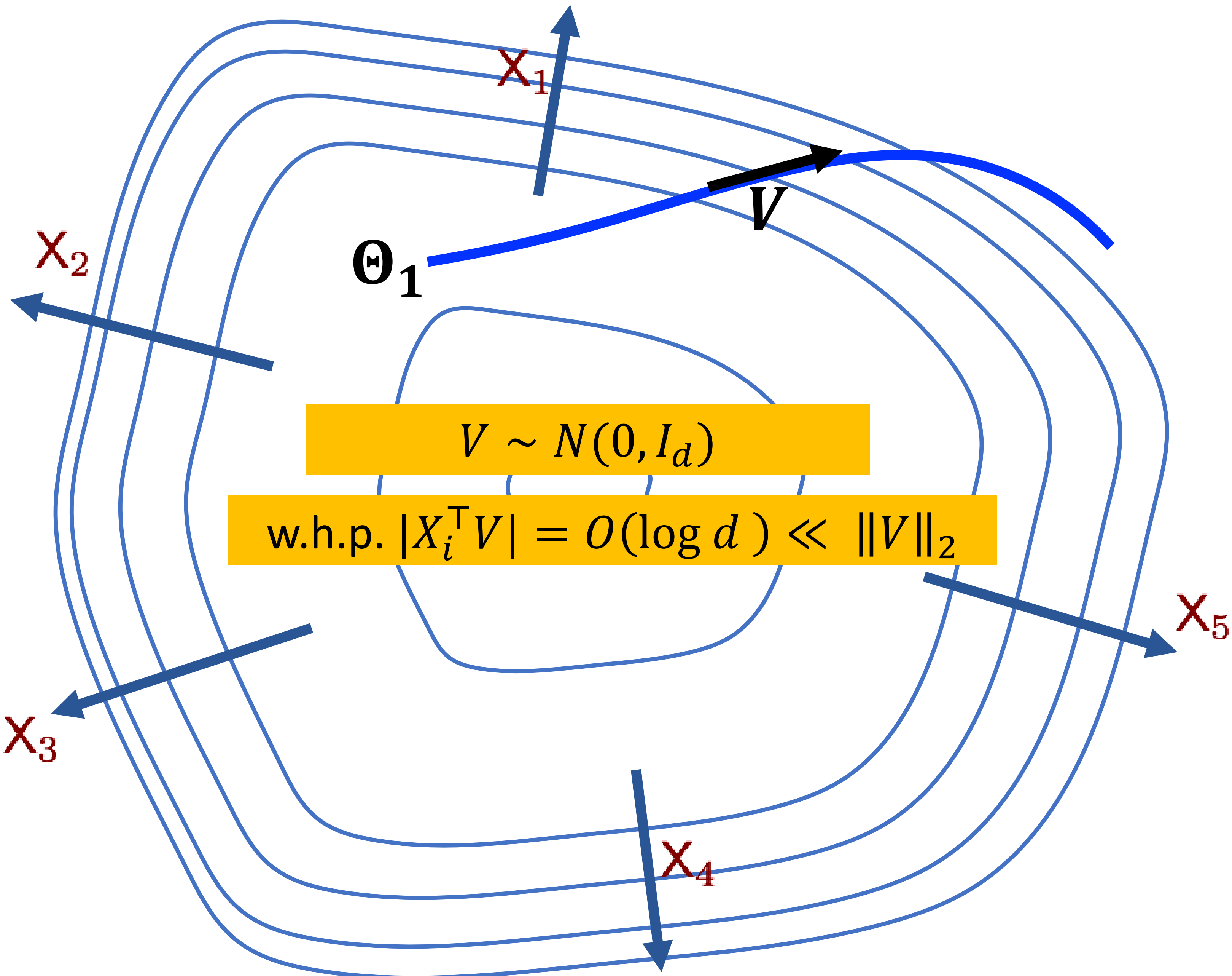
$$\left\|\left(H(\theta_1) - H(\theta_2)\right)v\right\|_2 \le L_\infty \cdot \|\theta_1 - \theta_2\|_\infty \cdot \|v\|_\infty$$

Positive: $\|v\|_\infty \approx \sqrt{\log d}$, **Negative:** L_∞ is large unless f is separable

Idea: transform the norm to align with the "data vectors"

We use: $\left\|X^\top v\right\|_\infty$ where $X := [x_1, \ldots, x_r]$

Intuition



Concluding the Proof (for $d = r$)

We bound (inductively on j) the errors $\|\theta_j - \theta(\eta j)\|_2$ and $\|v_j - v(\eta j)\|_2$ by $O(\eta j \varepsilon)$, where $(\theta(t), v(t))$ is the continuous solution to Hamilton's equations with initial conditions in that phase. Since $\eta j \leq T = O(1)$, $O(\eta j \varepsilon) = O(\varepsilon)$

- The error in the quadratic term of the velocity update is roughly $\|(\eta^2 H(\theta + \eta v_j) - \eta^2 H(\theta))v_j\|_2 \quad \leq \quad \eta^3 L_\infty \sqrt{d} \|\mathrm{X}^\top v_j\|_\infty^2$

- The invariance property of Hamiltonian mechanics implies v_j is roughly $N(0, I_d)$ at every point on the exact trajectory if HMC has a warm start

- Thus, $\|\mathrm{X}^\top v_j\|_\infty = O(\log(d))$ w.h.p., since by inductive assumption $\|v_j - v(\eta j)\|_2 = O(\eta j \varepsilon) = O(1)$

- After T/η iterations, the errors sum to $\tilde{O}(\eta^2 L_\infty \sqrt{d})$. Choosing η to have error ε, # of gradients is $T/\eta = \widetilde{\Theta}(\varepsilon^{-1/2} d^{1/4} L_\infty^{1/2})$

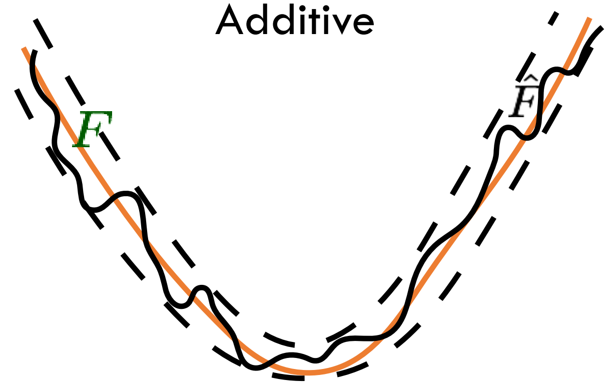
LANGEVIN DYNAMICS, SIMULATED ANNEALING AND NOISY CONVEX OPTIMIZATION

Optimizing using Noisy Oracles

Input: Noisy approximation $\hat{F}$ to a convex function $F\colon \mathbb{R}^d \to \mathbb{R}$ with global minimum $\theta^\star$

Goal: Find $\hat{x}$, s.t. $F\left(\hat{\theta}\right) - F(\theta^\star) < \varepsilon$ for given $\varepsilon > 0$

Additive

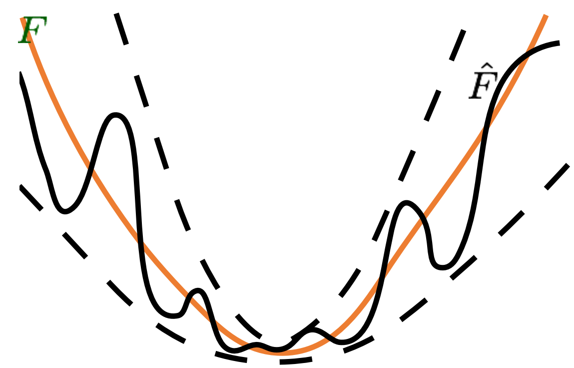


[Applegate-Kannan '91, Zhang et al. '17]

Applications:

- Optimizing F when an accurate value of F is expensive to compute
- Optimizing non-convex functions which are close to a convex function

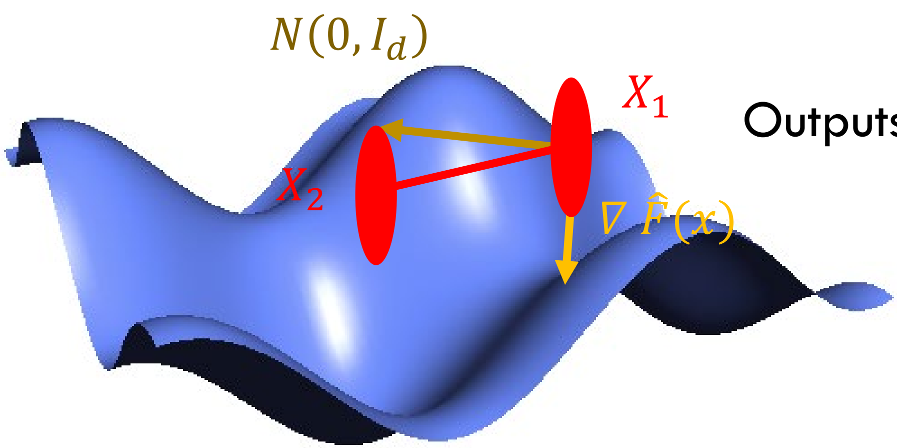
Additive & Multiplicative



[Belloni-Liang-Narayanan-Rakhlin '15]

Algorithm: Langevin Dynamics

(Arises in statistical physics)

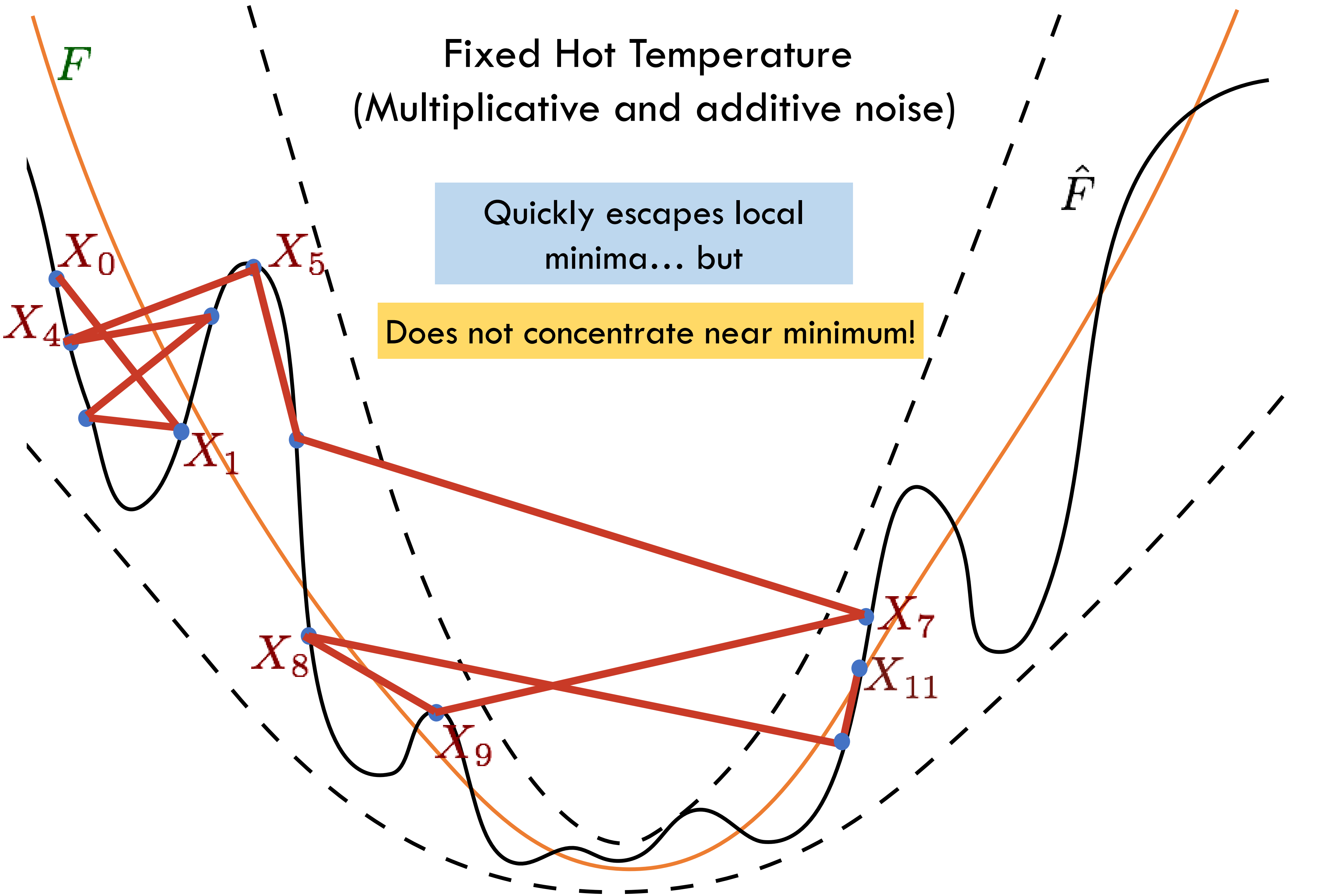


Outputs $\theta \propto e^{-\hat{F}(\theta)/T}$

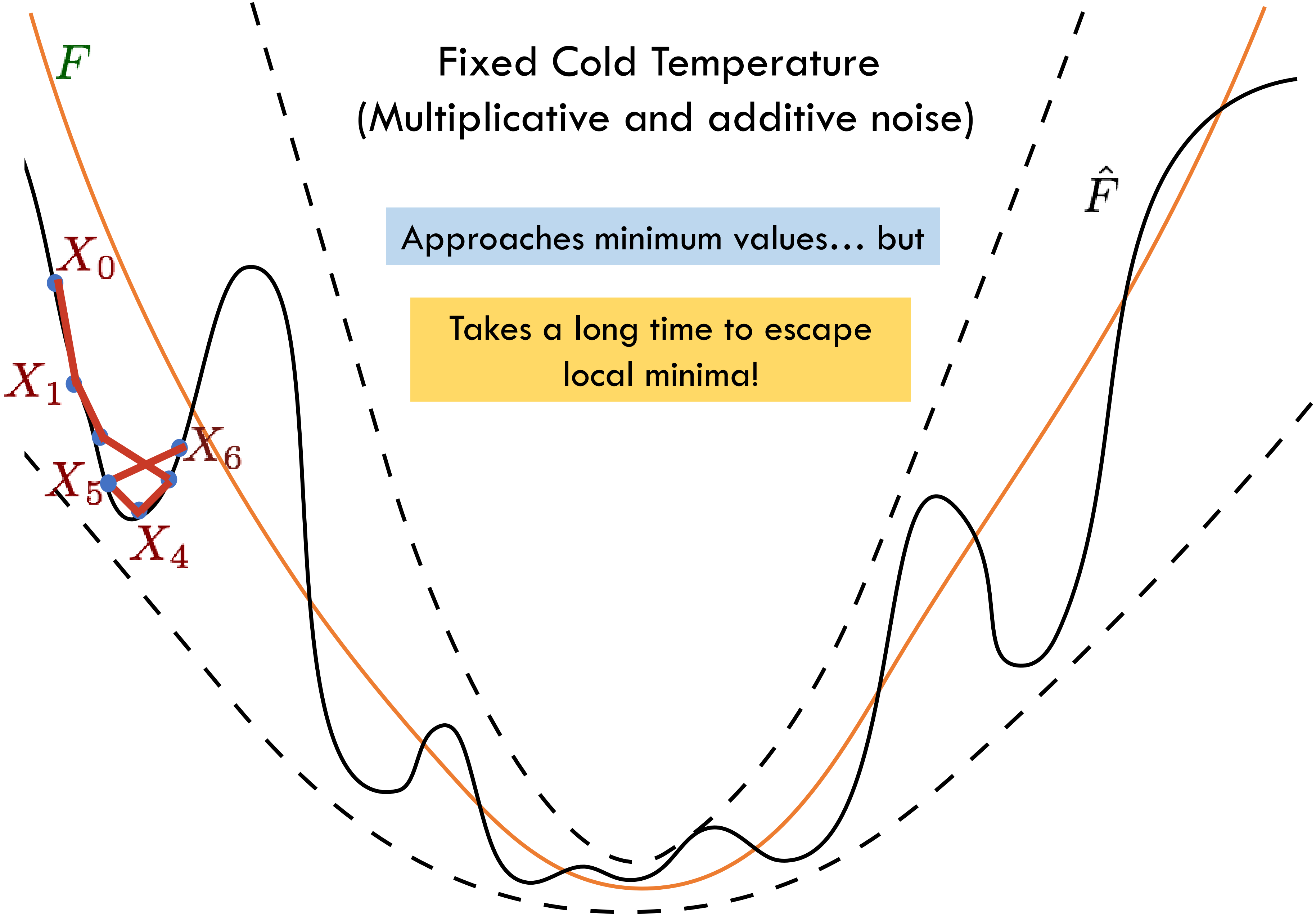
$$X_{i+1} = X_i - \eta \nabla \hat{F}(X_i) + \sqrt{2\eta \boldsymbol{T}}\, N(0, I_d)$$

High $\boldsymbol{T}$: escape local minima quickly

Low $\boldsymbol{T}$: limiting distribution concentrates near minimum value

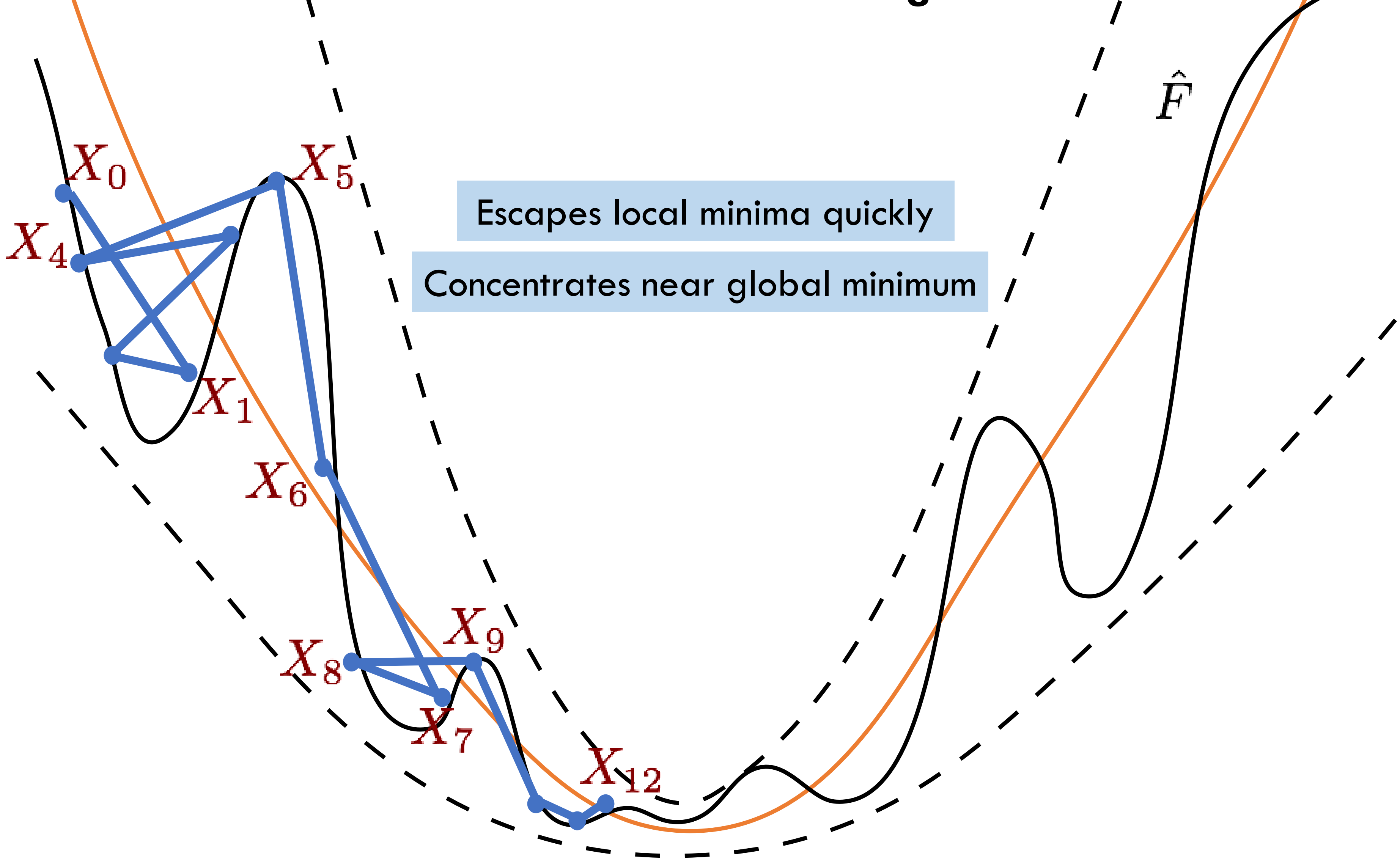

Fixed Hot Temperature
(Multiplicative and additive noise)
Quickly escapes local minima… but
Does not concentrate near minimum!
F
$\hat{F}$
X_0
X_5
X_4
X_1
X_7
X_8
X_{11}
X_9

Fixed Cold Temperature
(Multiplicative and additive noise)



Our Approach and Result [Mangoubi-V. COLT '18]

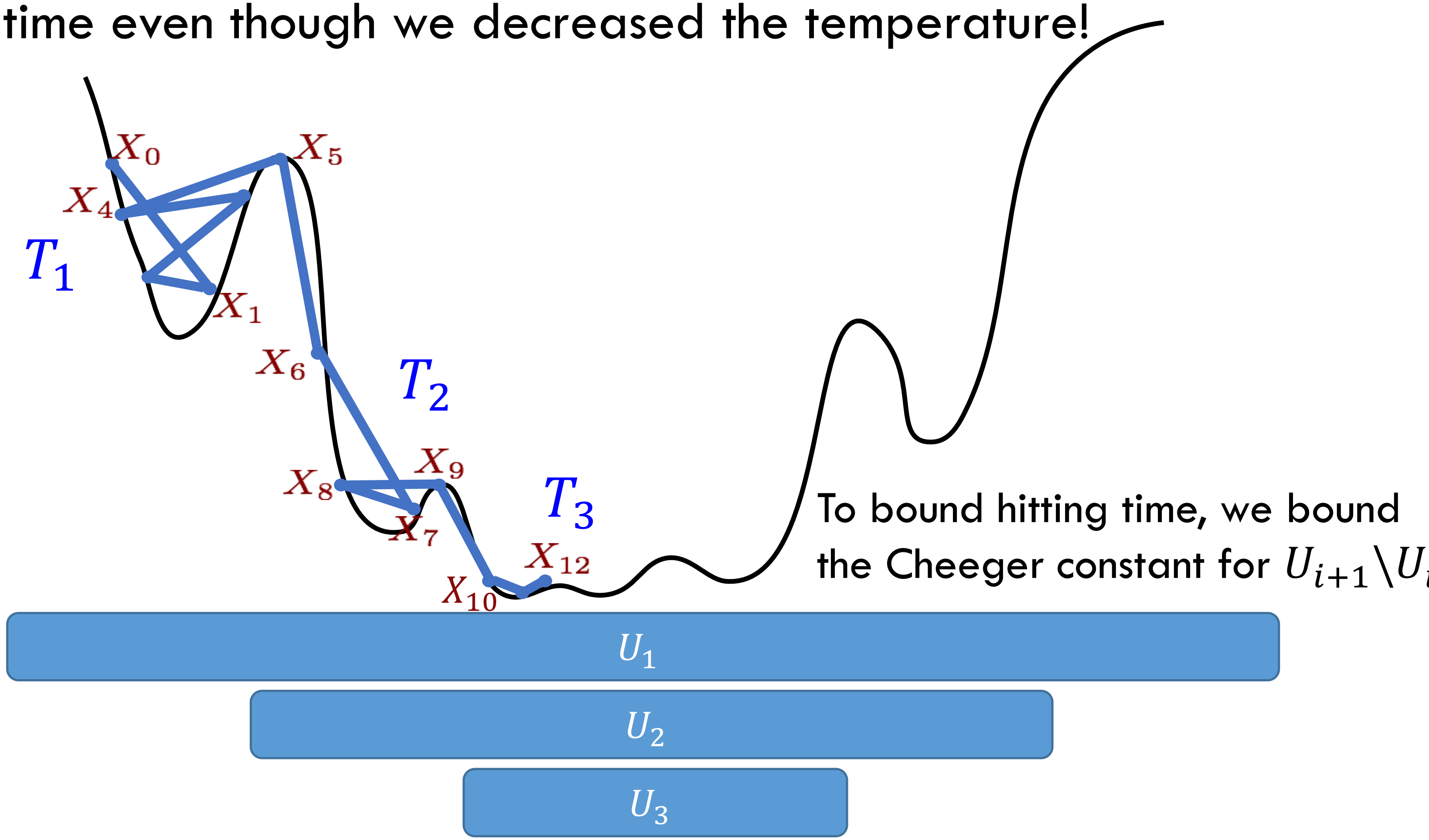
Simulated Annealing



Theorem: Suppose the approximation $\hat{F}$ has additive noise $< O(\frac{\varepsilon}{d})$ and multiplicative noise $< O(\frac{1}{d})$. Then, our algorithm can minimize F to accuracy ε in time that is polynomial in d

Proof Strategy

1. At epoch $i+1$: show that our algorithm remains inside U_i w.h.p.
2. Noise is lower than at previous i, since multiplicative noise decreases
3. A lower noise implies shallower local minima and therefore a fast hitting time even though we decreased the temperature!



Conclusion

- Physics and physical systems (have been) and can be a great source of ideas and insights for
  - **rigorous** algorithm design
  - **tuning** parameters
  - identifying the right **potential** functions
  - obtaining "**beyond worst case**" assumptions on functions/data

- **Other recent physics-inspired optimization and sampling algo.**
  - Online sampling with O(1) update time [H.Lee-Mangoubi-**V.** NeurIPS '19]
  - Langevin dynamics for non-convex potentials [Mangoubi-**V.** COLT '19]
  - Sampling from polytopes [Mangoubi-**V.** FOCS '19]

Thanks! Questions?