How do we compute posterior expectations in practice?

How do we compute posterior expectations in practice?

- Construct a Markov chain that explores the typical set

How do we compute posterior expectations in practice?

- Construct a Markov chain that explores the typical set
- Anything you would want to do if you could write it analytically, you can do to any accuracy with the draws (history) of the chain

How do we compute posterior expectations in practice?

- Construct a Markov chain that explores the typical set

- Anything you would want to do if you could write it analytically, you can do to any accuracy with the draws (history) of the chain

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N} f(\theta_n) \to E_{\pi}[f]$$

How do we compute posterior expectations in practice?

- Construct a Markov chain that explores the typical set

- Anything you would want to do if you could write it analytically, you can do to any accuracy with the draws (history) of the chain

$$\lim_{\boxed{N \to \infty}} \frac{1}{N} \sum_{n=0}^{N} f(\theta_n) \to E_{\pi}[f]$$

With Stan, we aim to provide an MCMC implementation that works robustly for as many target distributions as possible

With Stan, we aim to provide an MCMC implementation that works robustly for as many target distributions as possible

- Gibbs, RW Metrop are usually very inefficient, hard to diagnose

With Stan, we aim to provide an MCMC implementation that works robustly for as many target distributions as possible

- Gibbs, RW Metrop are usually very inefficient, hard to diagnose

- To explore complicated high-dimensional spaces we need to leverage what we know about the **geometry** of the typical set

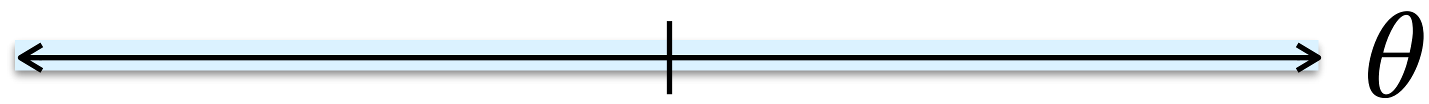
With Stan, we aim to provide an MCMC implementation that works robustly for as many target distributions as possible

- Gibbs, RW Metrop are usually very inefficient, hard to diagnose

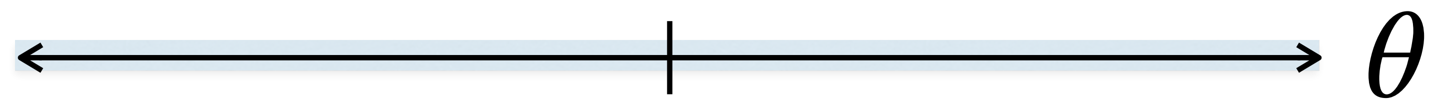
- To explore complicated high-dimensional spaces we need to leverage what we know about the **geometry** of the typical set

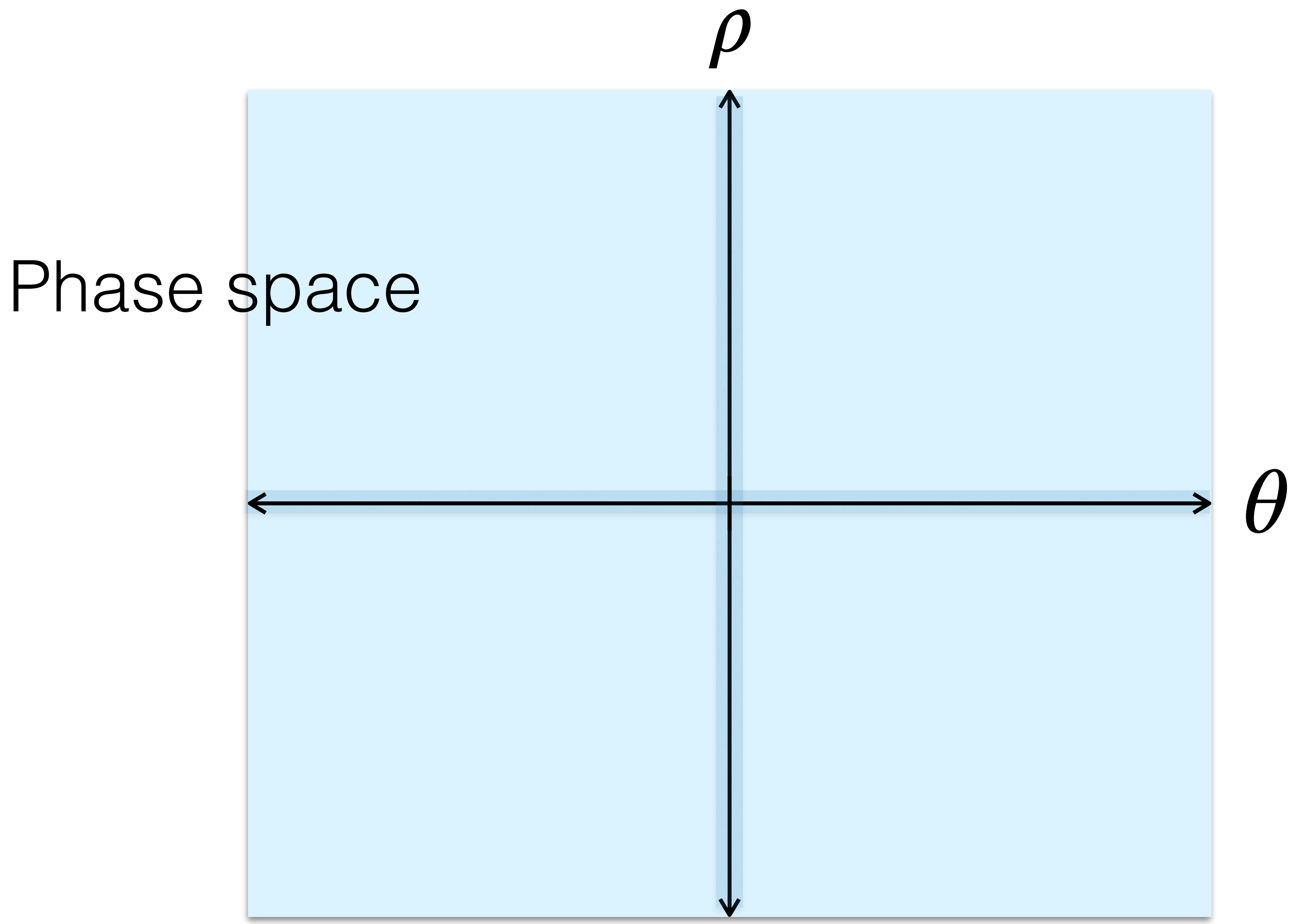
- **Hamiltonian Monte Carlo**

Parameter space

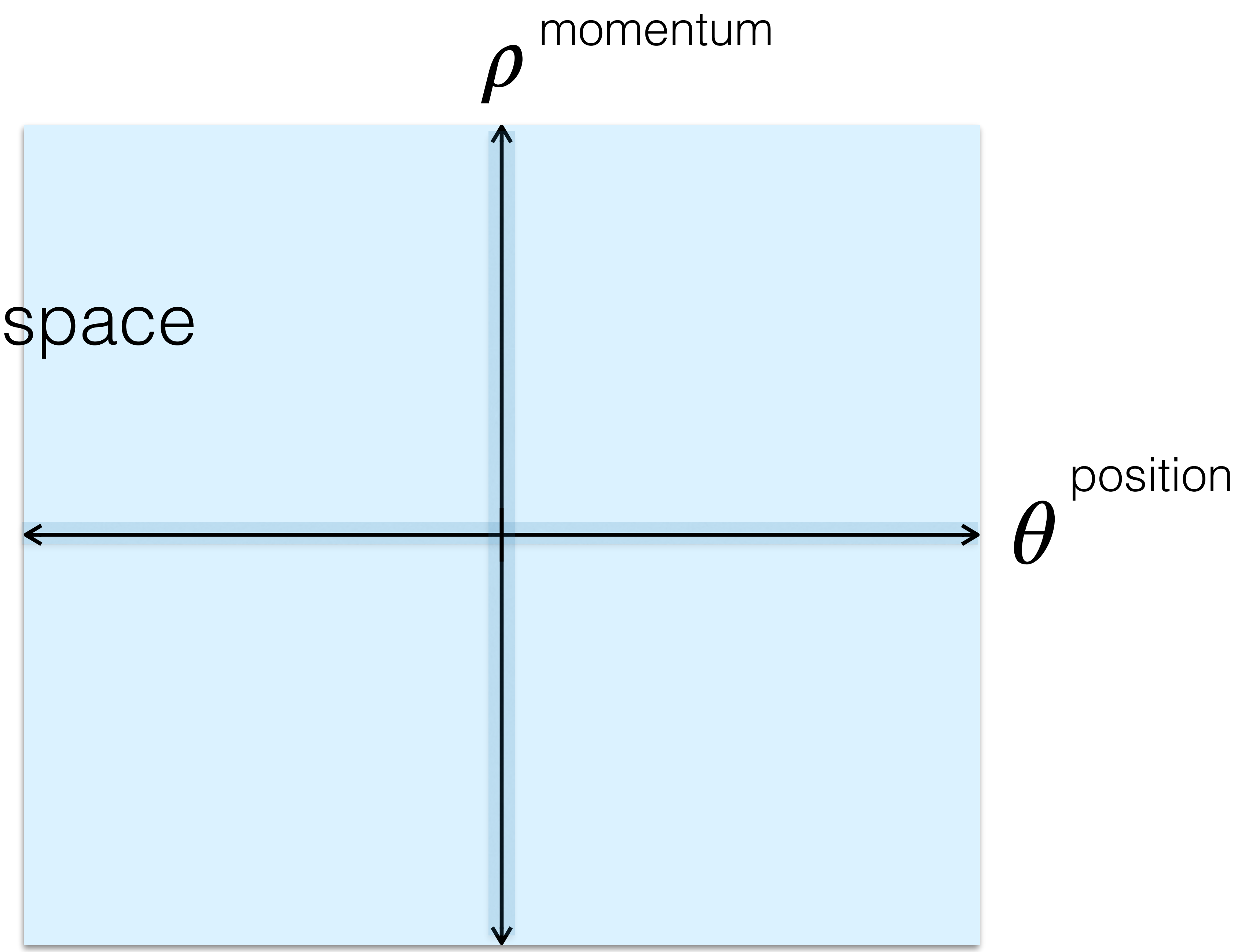


Parameter space




Phase space
ρ
θ

Phase space



At this point, let's abstract away the distinction between prior and posterior and just let

$$\pi(\theta)$$

refer to the target distribution in parameter space, and

At this point, let's abstract away the distinction between prior and posterior and just let

$$\pi(\theta)$$

refer to the target distribution in parameter space, and

$$\pi(\theta, \rho)$$

refer to the joint distribution of the parameters and momenta in phase space.

We can explore the the typical set of the target distribution by simulating **Hamiltonian dynamics** in phase space

Hamiltonian Function

$$
\begin{aligned}
\pi(\theta, \rho) &= \exp\{-H(\theta, \rho)\} \\
H(\theta, \rho) &= -\log \pi(\theta, \rho) \\
&= -\log \pi(\rho|\theta) - \log \pi(\theta) \\
&= \underbrace{K(\rho, \theta)}_{\text{kinetic}} + \underbrace{V(\theta)}_{\text{potential}}
\end{aligned}
$$

We can explore the the typical set of the target distribution by simulating **Hamiltonian dynamics** in phase space

Hamiltonian Function

$$
\begin{aligned}
\pi(\theta, \rho) &= \exp\{-H(\theta, \rho)\} \\
H(\theta, \rho) &= -\log \pi(\theta, \rho) \\
&= -\log \pi(\rho | \theta) - \log \pi(\theta) \\
&= \underbrace{K(\rho, \theta)}_{\text{kinetic}} + \underbrace{V(\theta)}_{\text{potential}}
\end{aligned}
$$

Hamilton's Equations

$$
\begin{aligned}
\frac{d\theta}{dt} &= +\frac{\partial H}{\partial \rho} = \frac{\partial K}{\partial \rho} \\
\frac{d\rho}{dt} &= -\frac{\partial H}{\partial \theta} = -\frac{\partial K}{\partial \theta} - \frac{\partial V}{\partial \theta}
\end{aligned}
$$

We can explore the the typical set of the target distribution by simulating **Hamiltonian dynamics** in phase space

Hamiltonian Function

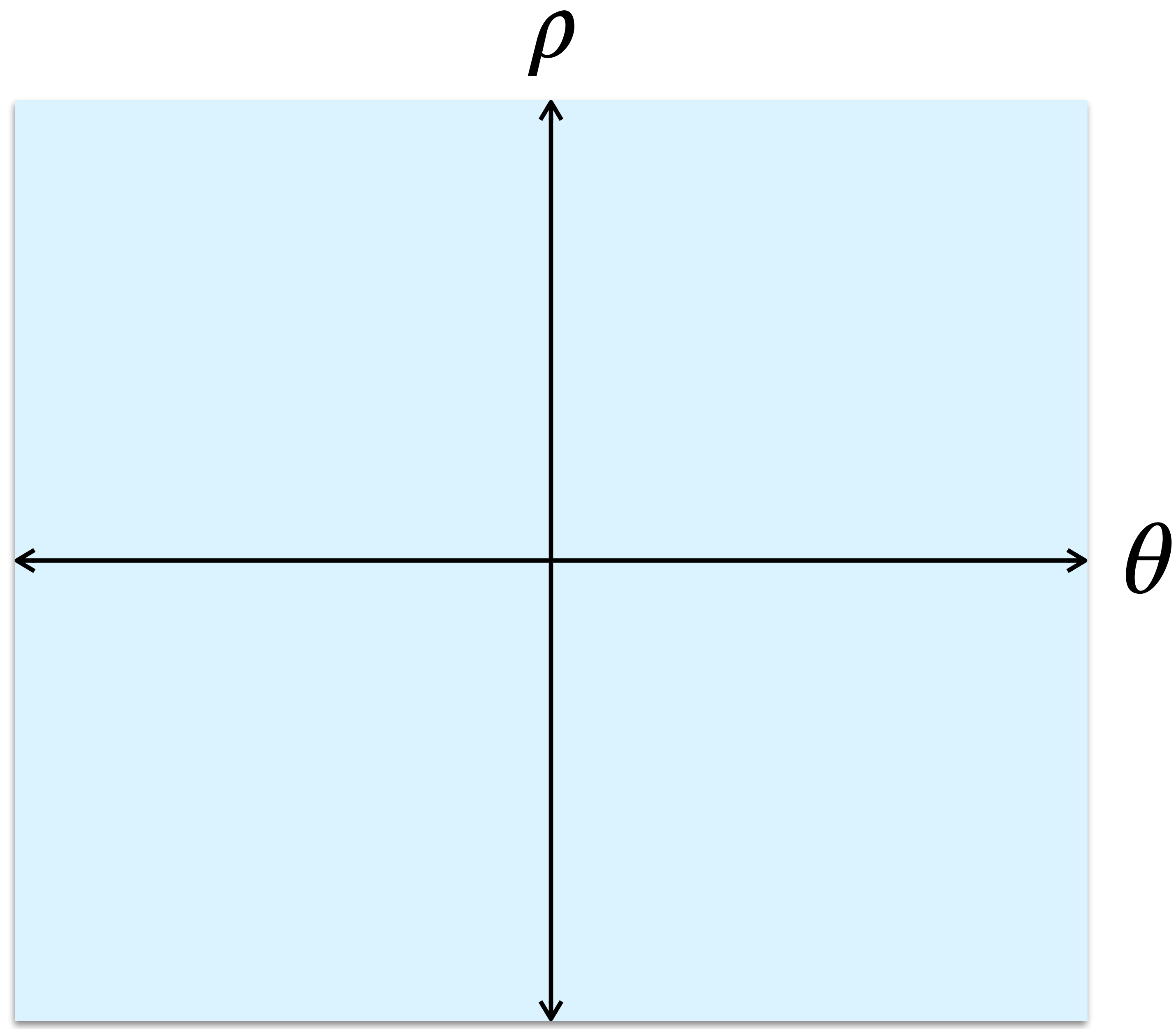
$$\begin{aligned}
\pi(\theta, \rho) &= \exp\{-H(\theta, \rho)\} \\
H(\theta, \rho) &= -\log \pi(\theta, \rho) \\
&= -\log \pi(\rho | \theta) - \log \pi(\theta) \\
&= \underbrace{K(\rho, \theta)}_{\text{kinetic}} + \underbrace{V(\theta)}_{\text{potential}}
\end{aligned}$$

Hamilton's Equations

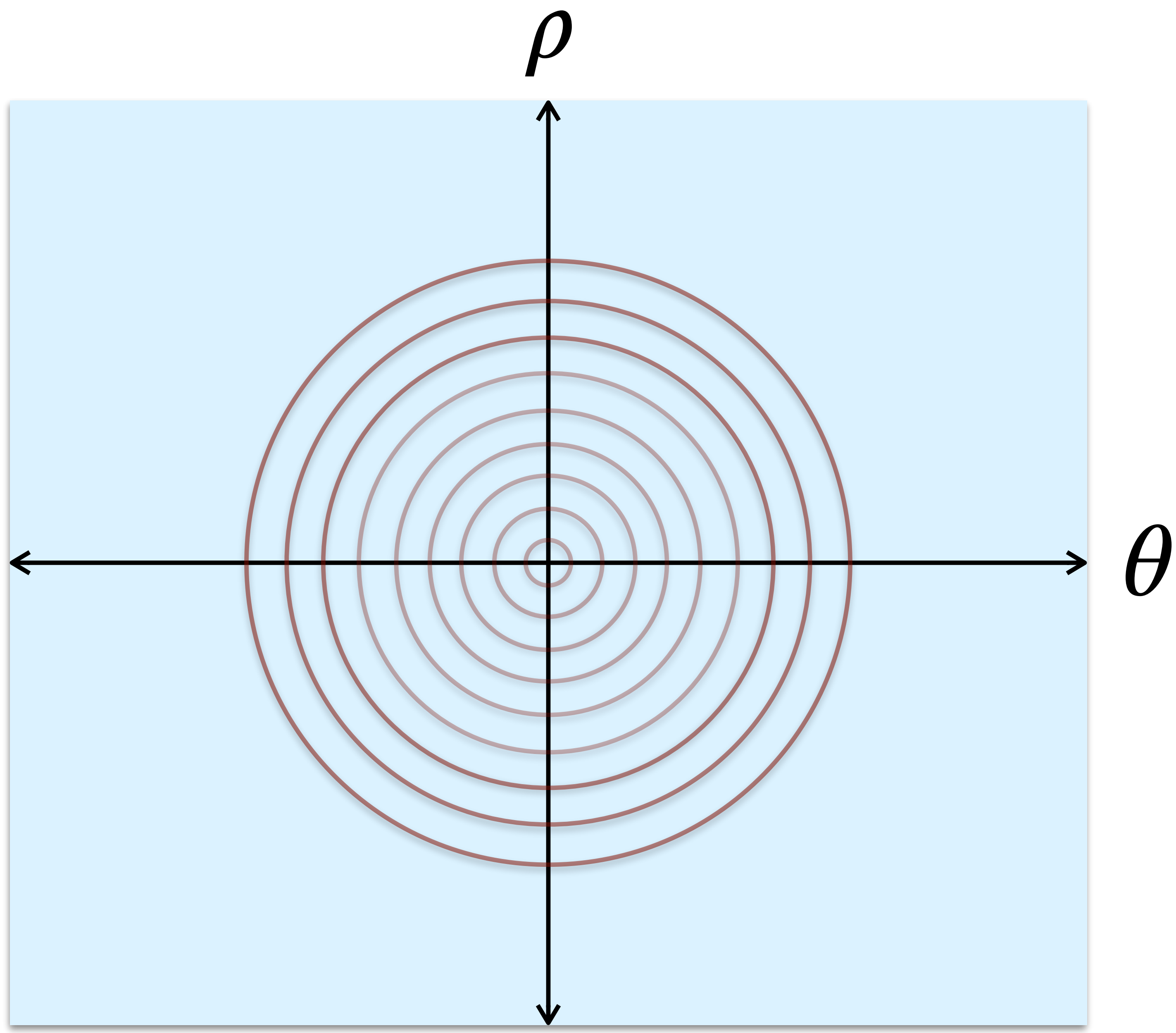
$$\begin{aligned}
\frac{d\theta}{dt} &= +\frac{\partial H}{\partial \rho} = \frac{\partial K}{\partial \rho} \\
\frac{d\rho}{dt} &= -\frac{\partial H}{\partial \theta} = -\frac{\partial K}{\partial \theta} - \frac{\partial V}{\partial \theta}
\end{aligned}$$

gradient of (log) target dist.

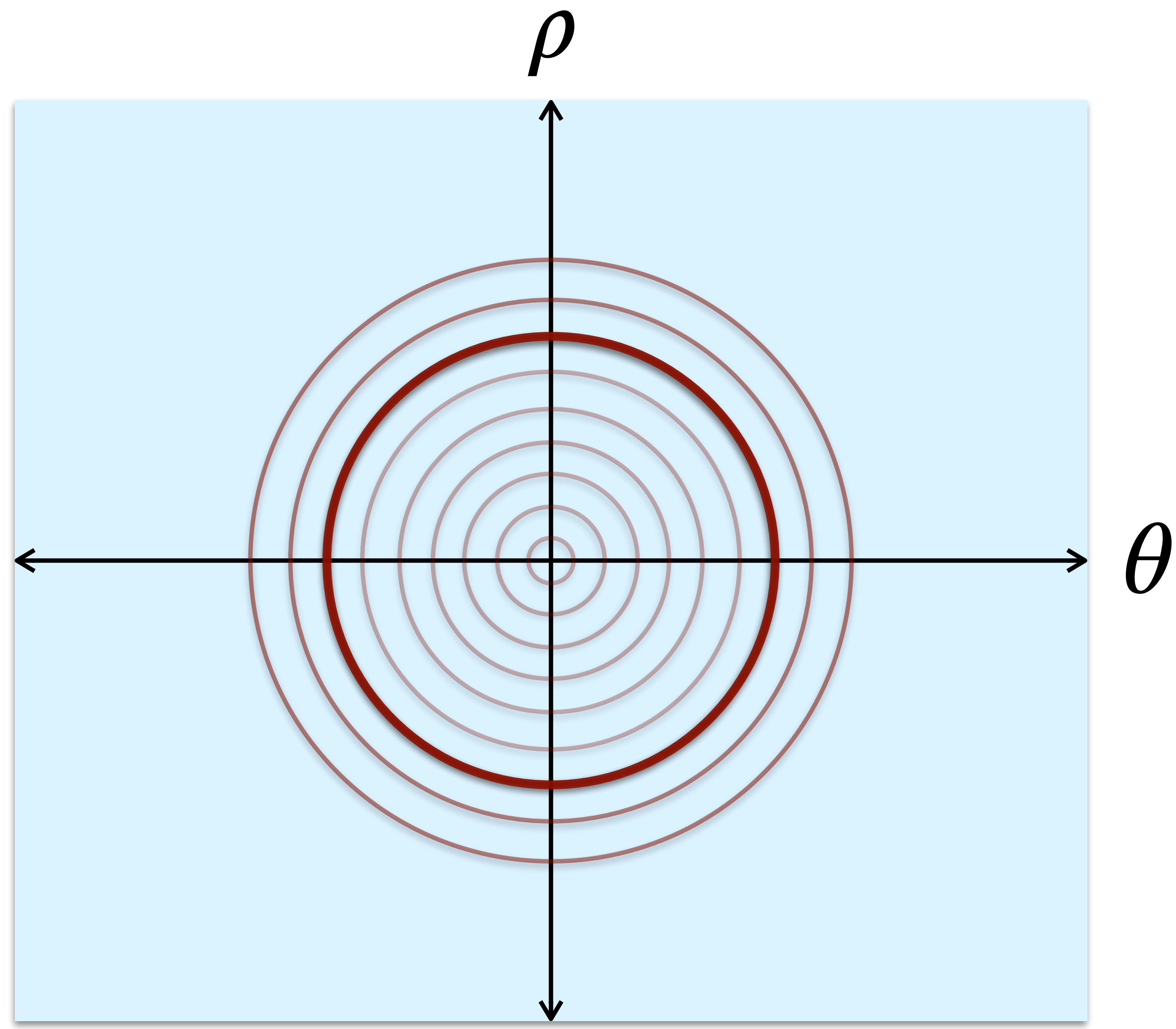
Phase space decomposes into concentric **energy** level sets



Phase space decomposes into concentric **energy** level sets

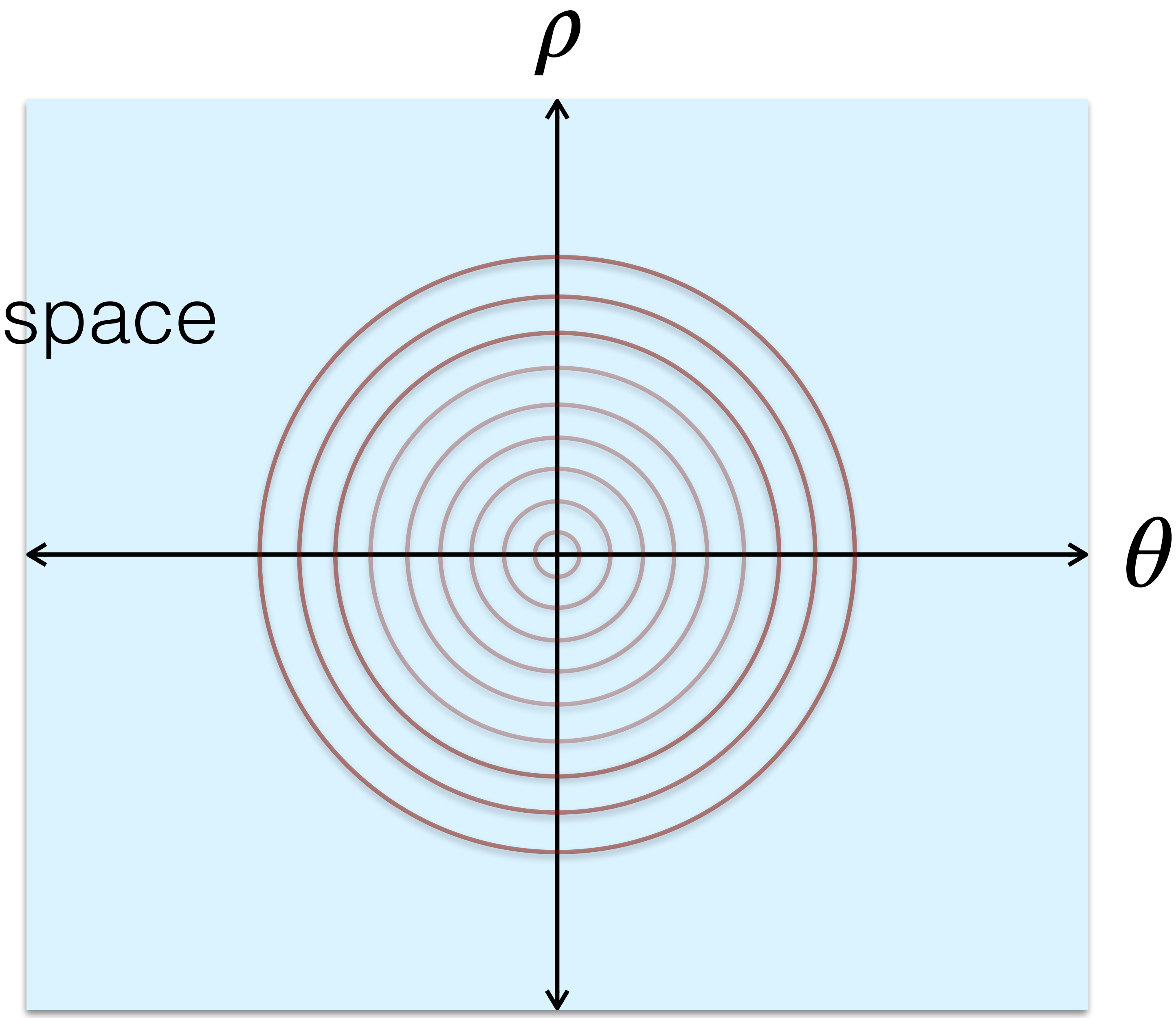


Phase space decomposes into concentric **energy** level sets



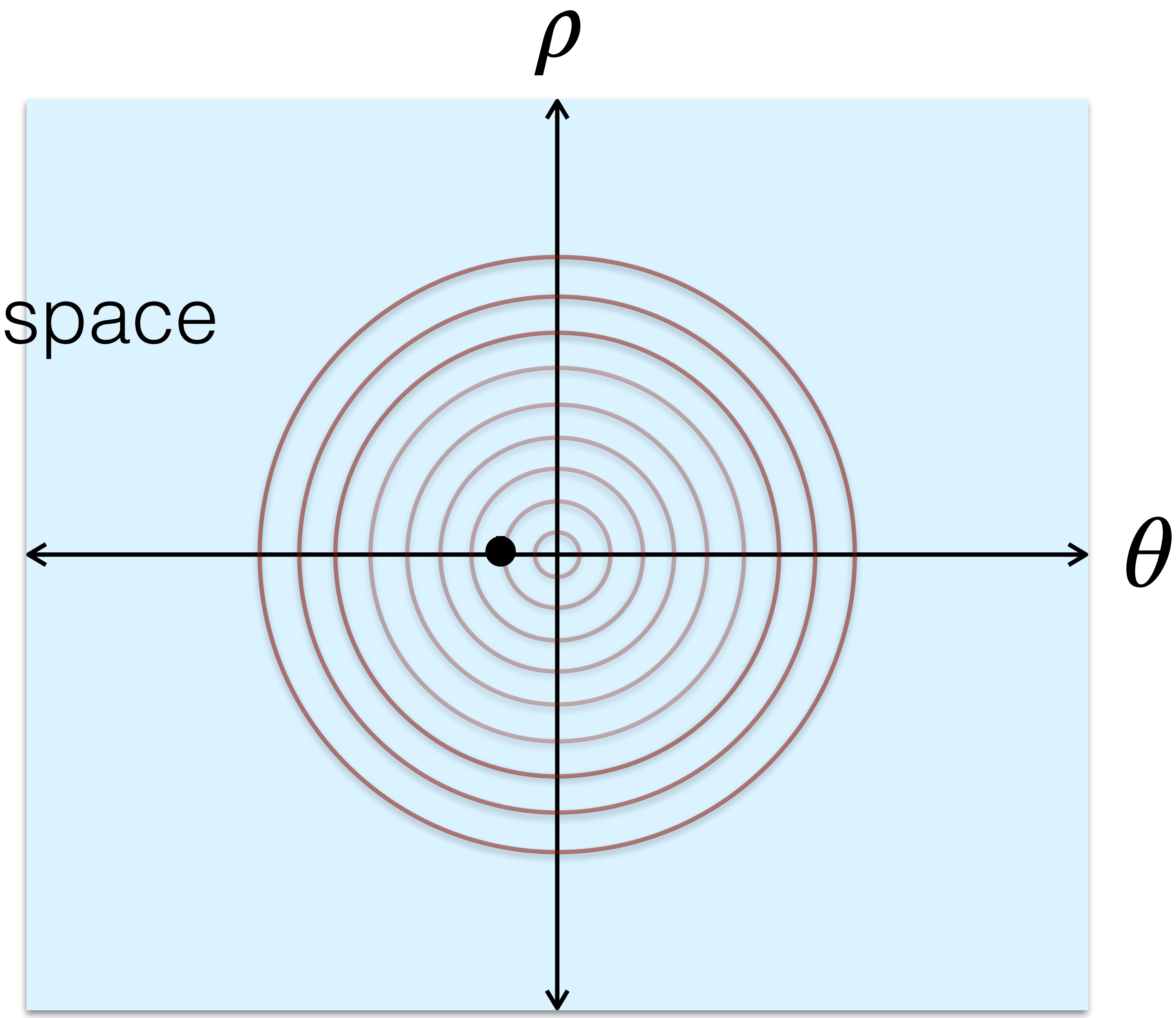
$$H^{-1}(E) = \{\theta, \rho \mid H(\theta, \rho) = E\}$$

Pick an initialization point

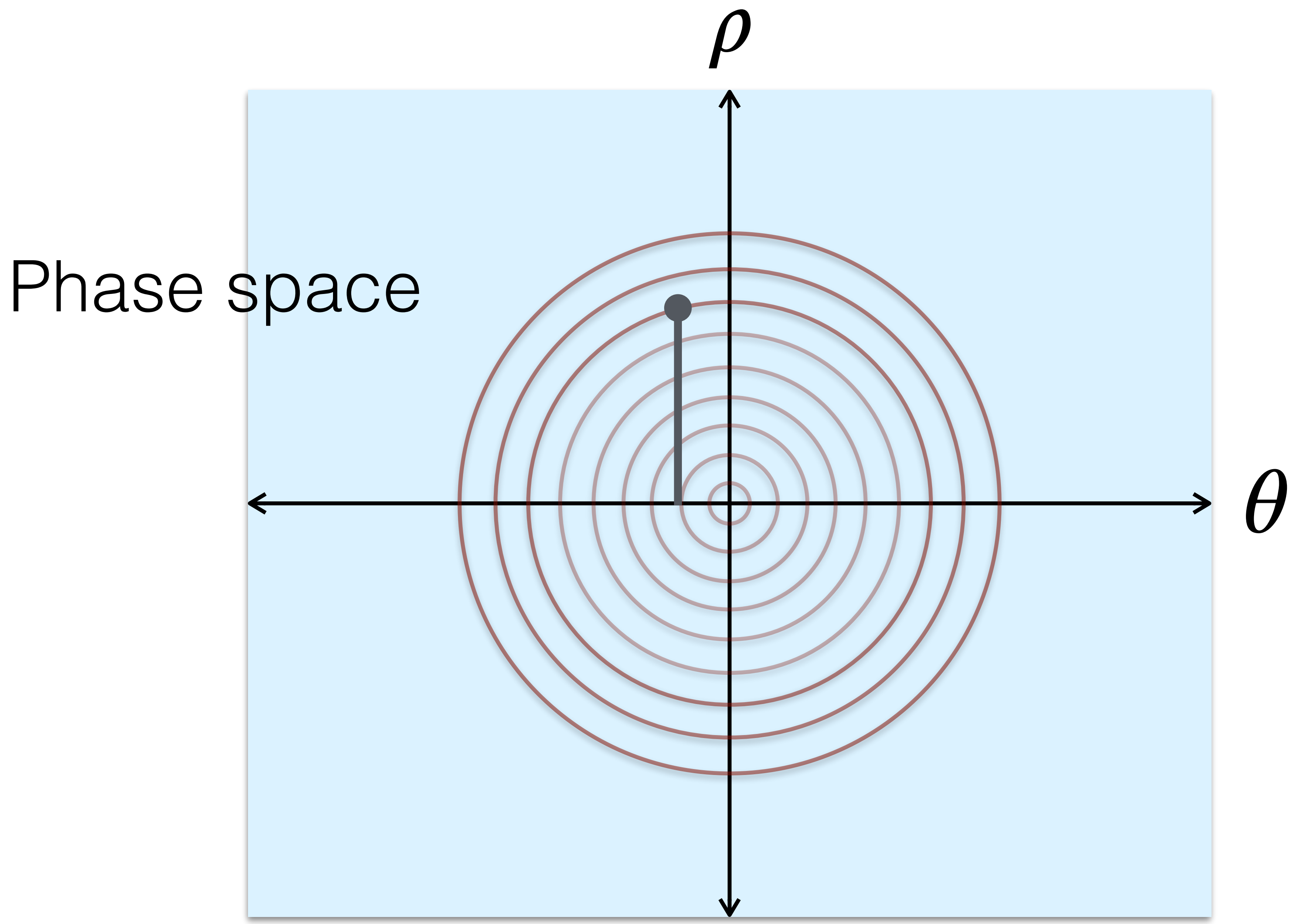


Pick an initialization point

Phase space

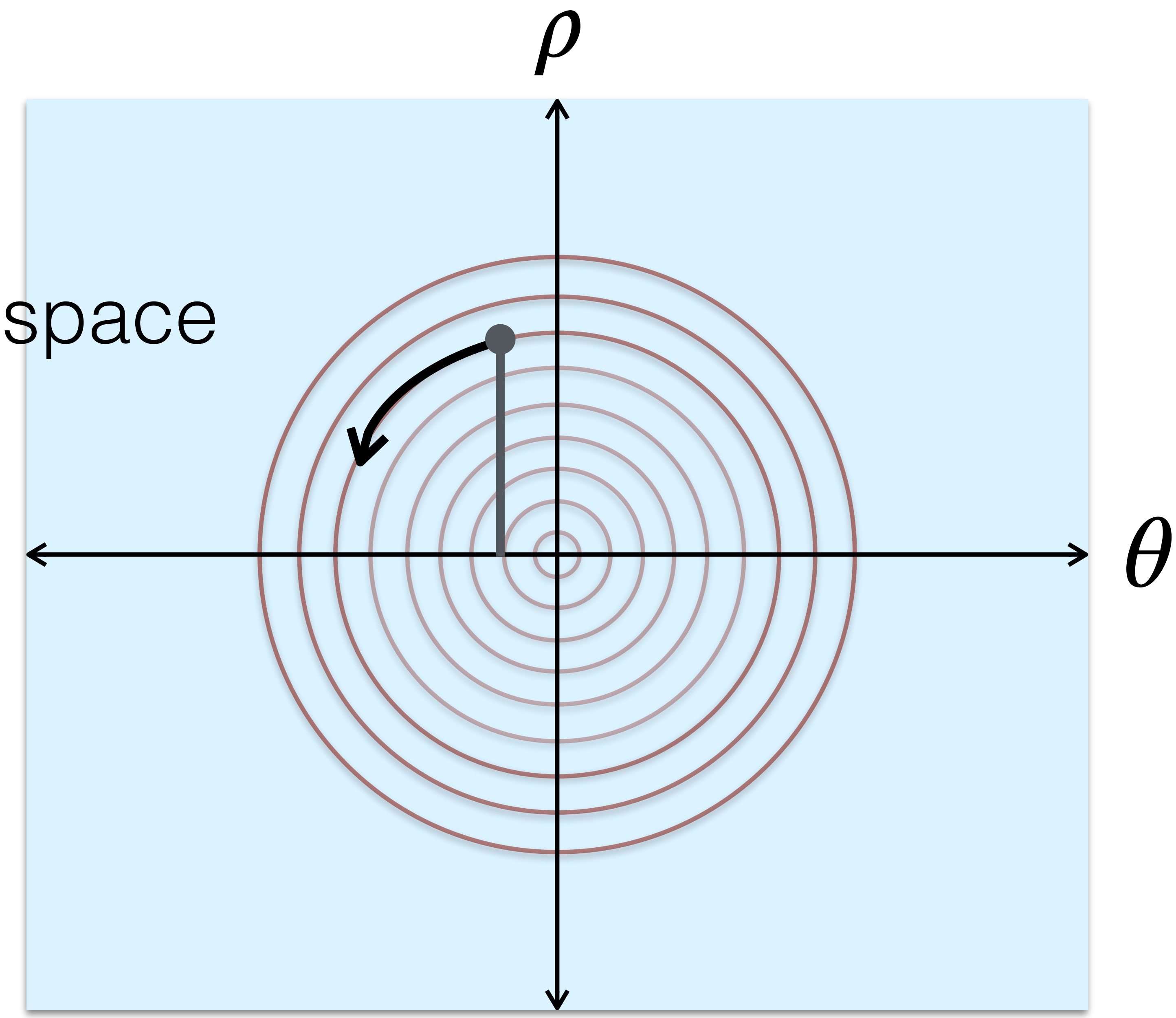


Sample momenta to lift into phase space



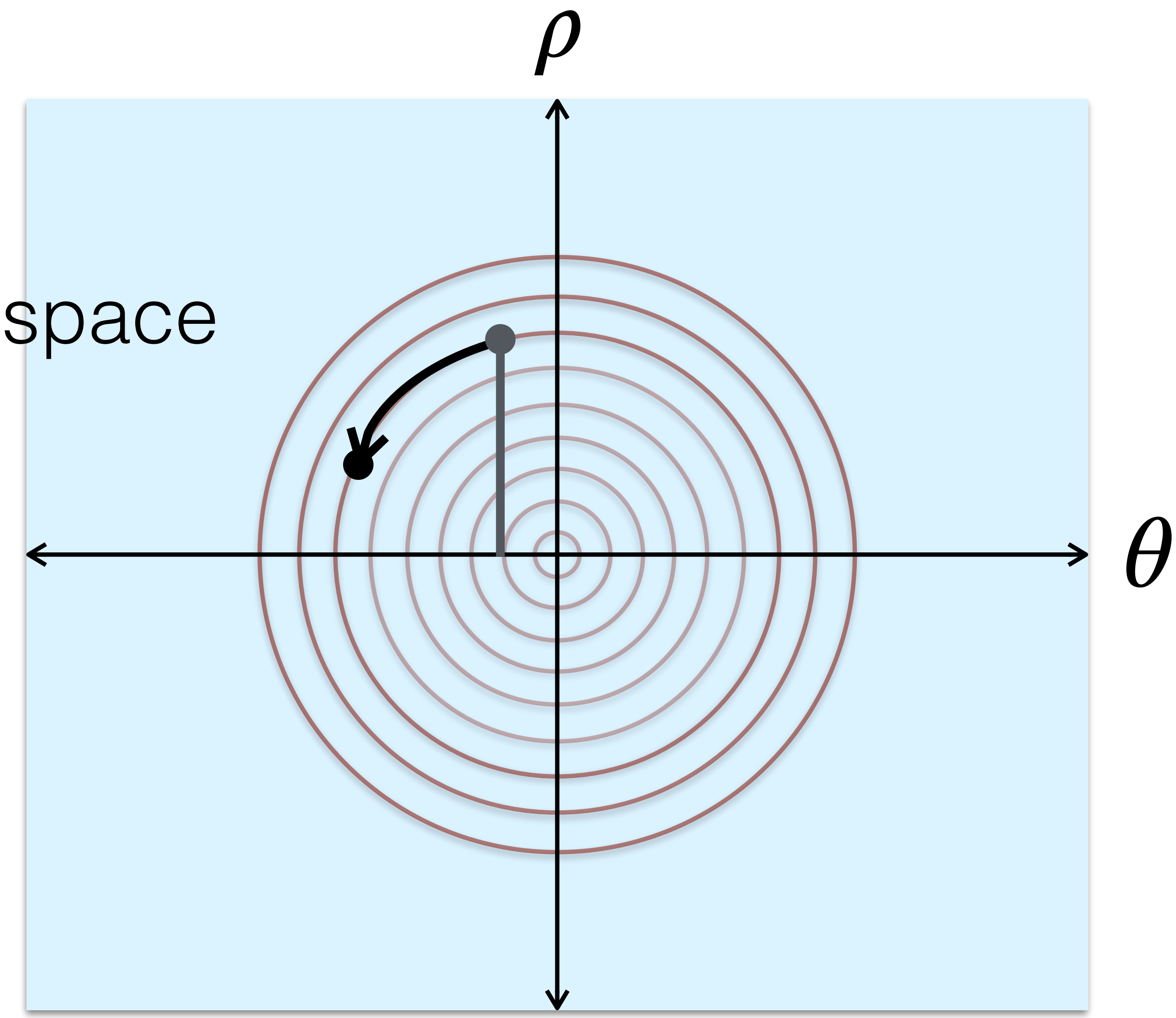
Deterministic exploration within energy levels sets

Phase space

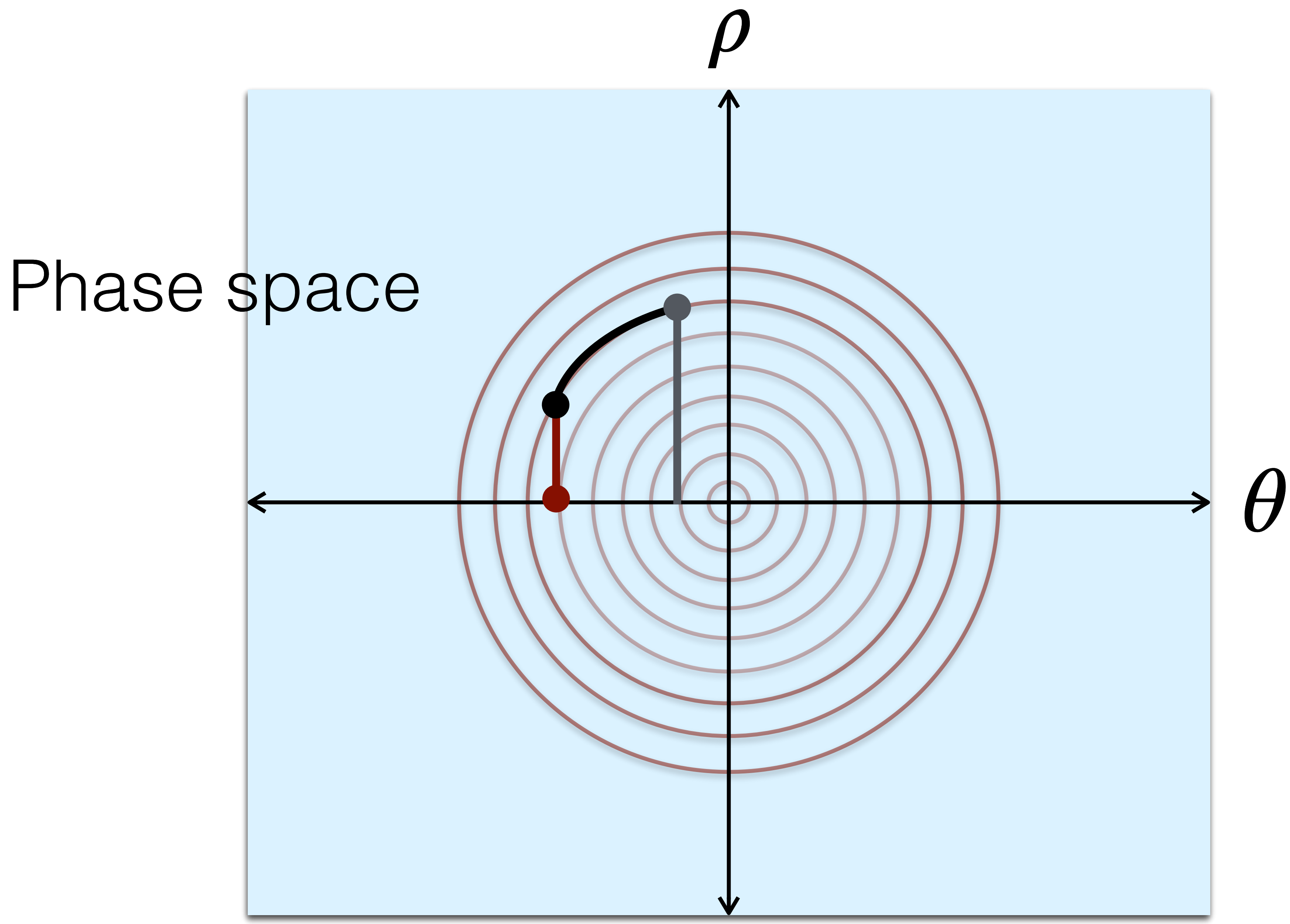


Deterministic exploration within energy levels sets

Phase space

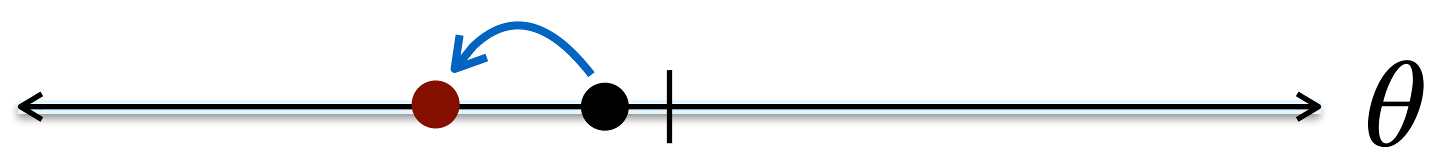


Project back down to parameter space

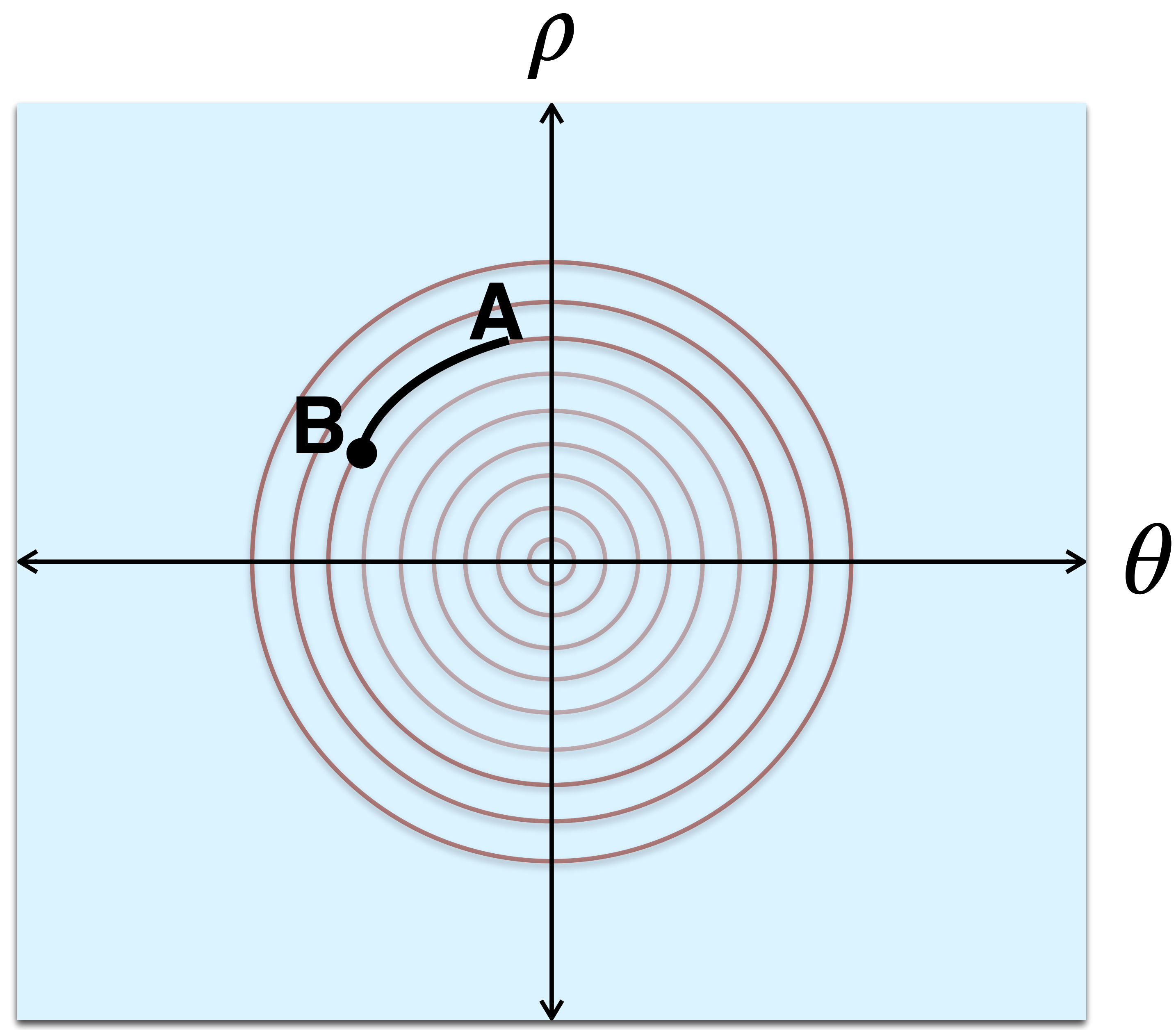


The auxiliary momenta are discarded and we are left with a point in the typical set of the target distribution

Parameter space



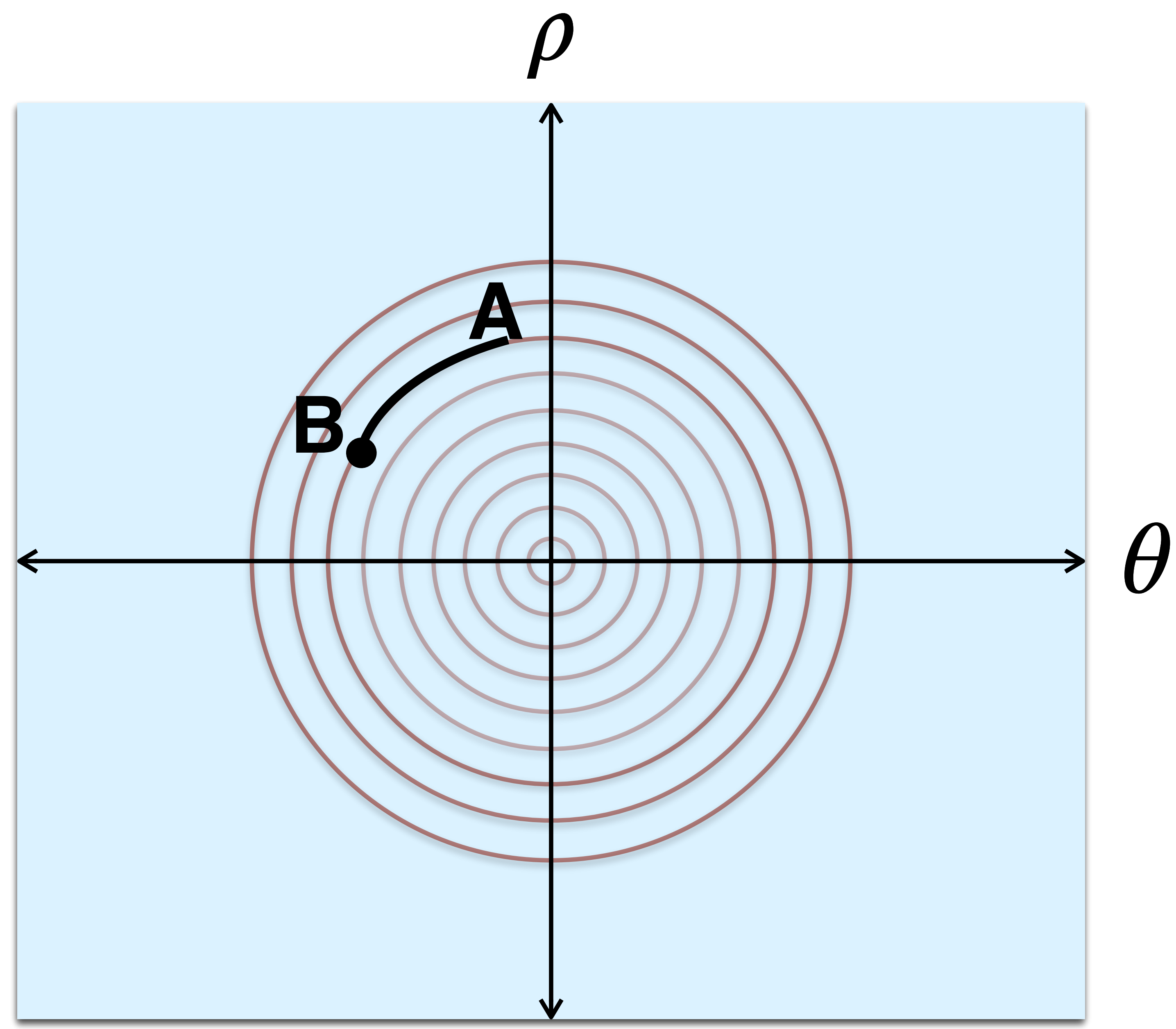
In that middle step, how did we actually get from point **A** to point **B**?



In that middle step, how did we actually get from point **A** to point **B**?

- Integrating Hamilton's equations

$$\frac{d\theta}{dt} = +\frac{\partial H}{\partial \rho} = \frac{\partial K}{\partial \rho}$$
$$\frac{d\rho}{dt} = -\frac{\partial H}{\partial \theta} = -\frac{\partial K}{\partial \theta} - \frac{\partial V}{\partial \theta}$$

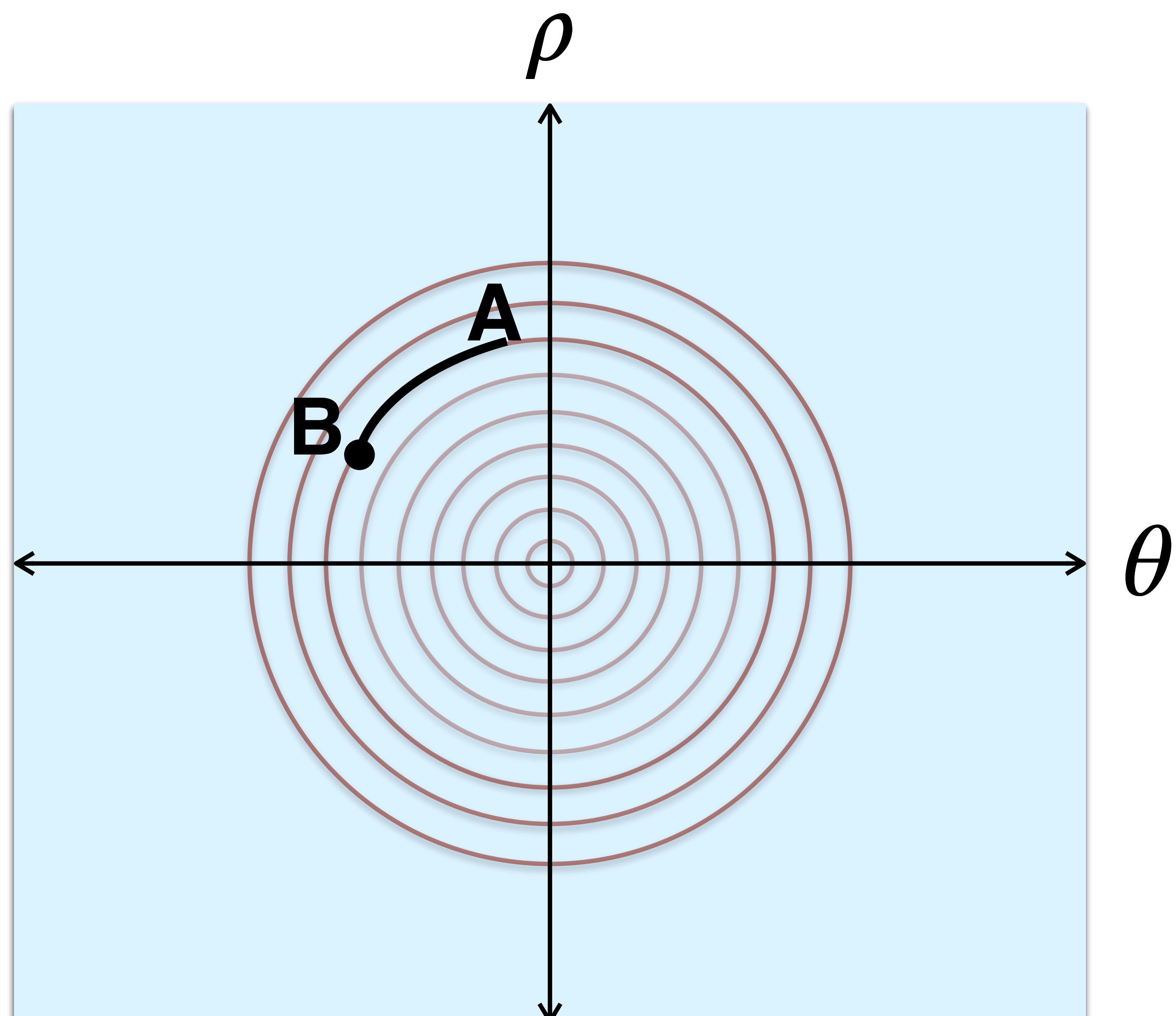


In that middle step, how did we actually get from point **A** to point **B**?

- Integrating Hamilton's equations

$$\frac{d\theta}{dt} = +\frac{\partial H}{\partial \rho} = \frac{\partial K}{\partial \rho}$$
$$\frac{d\rho}{dt} = -\frac{\partial H}{\partial \theta} = -\frac{\partial K}{\partial \theta} - \frac{\partial V}{\partial \theta}$$

- Discrete-time approximation
  - Symplectic integrator
  - Informed by geometry of the system



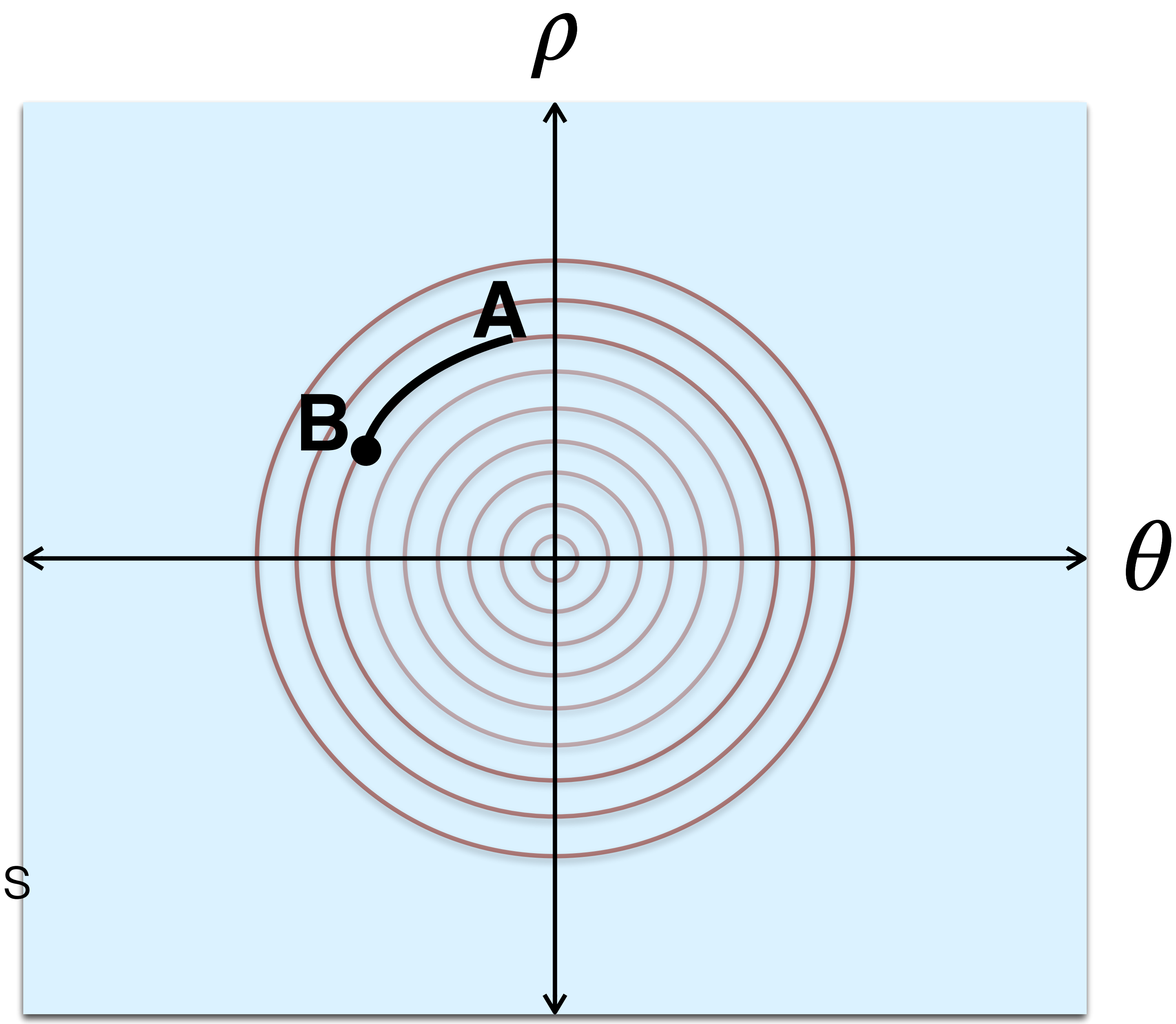
In that middle step, how did we actually get from point **A** to point **B**?

- Integrating Hamilton's equations

$$\frac{d\theta}{dt} = +\frac{\partial H}{\partial \rho} = \frac{\partial K}{\partial \rho}$$
$$\frac{d\rho}{dt} = -\frac{\partial H}{\partial \theta} = -\frac{\partial K}{\partial \theta} - \frac{\partial V}{\partial \theta}$$

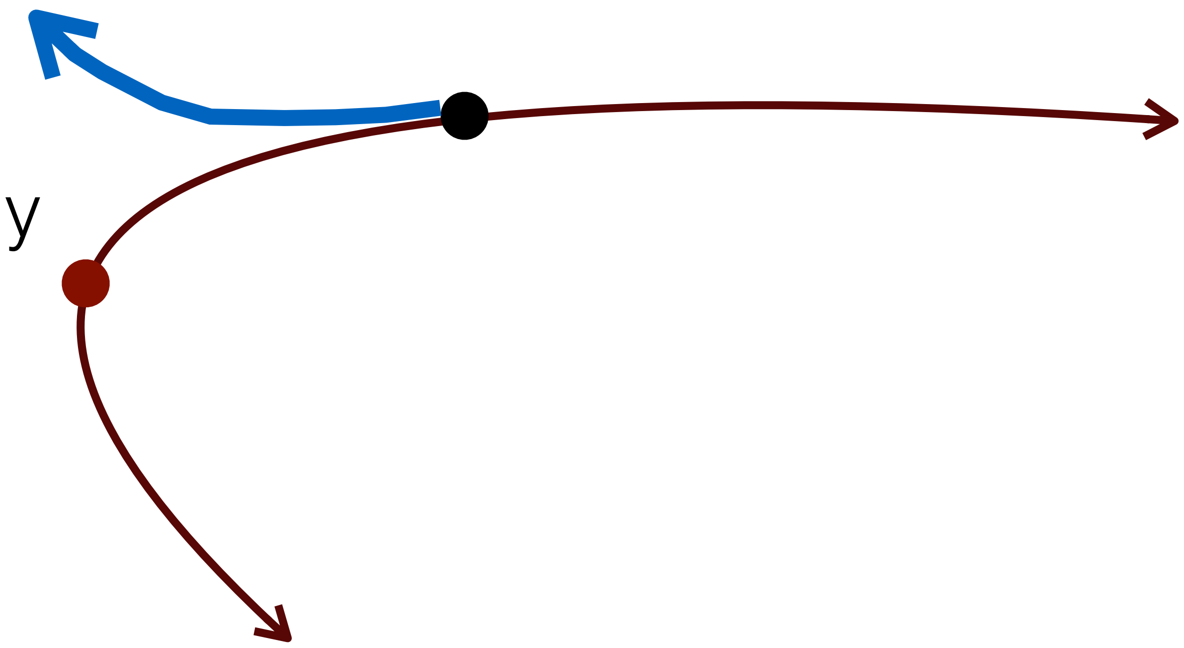
- Discrete-time approximation
  - Symplectic integrator
  - Informed by geometry of the system

- Motivates new MCMC diagnostics
  - Divergent transitions
  - Comparison of marginal & conditional energy distributions



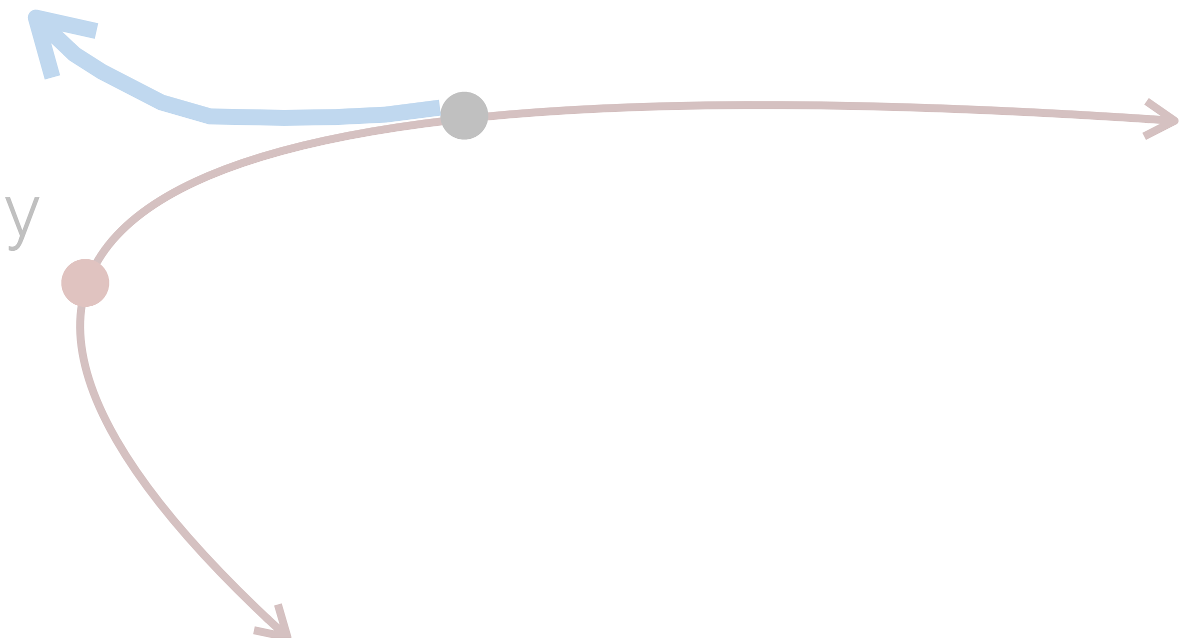
Most numerical integrators suffer from drift

- Trajectories deviate from typical set
- The size of the error increases rapidly with dimension



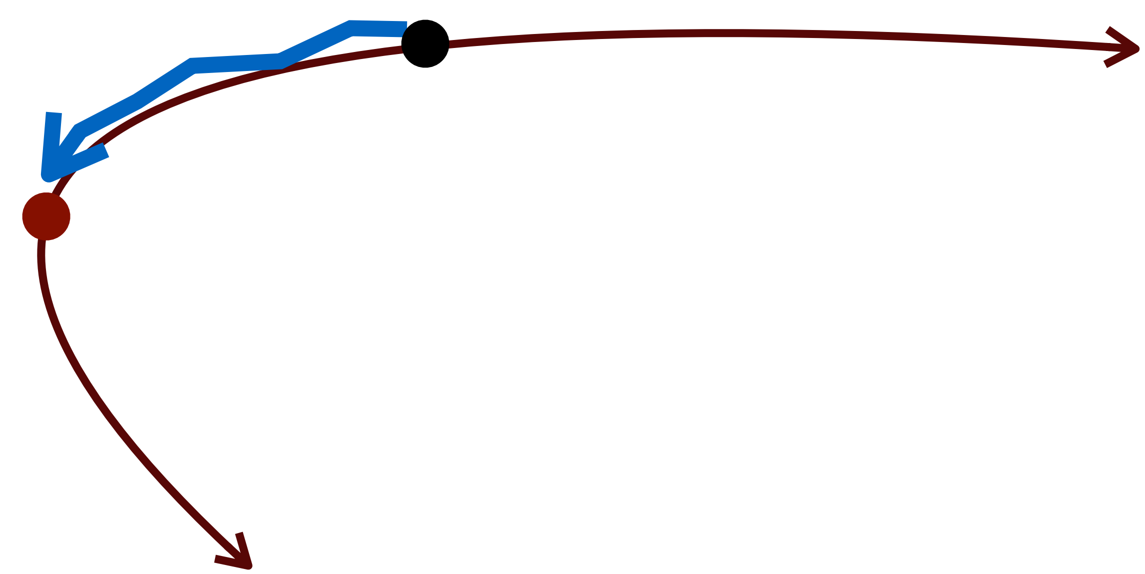
Most numerical integrators suffer from drift

- Trajectories deviate from typical set
- The size of the error increases rapidly with dimension



Symplectic integrators preserve volume in phase space

- Trajectories oscillate around exact energy level set, even for long integration times
- Scale well to higher dimensions
- Pathologies easy to diagnose (divergence)



Comparison of algorithms on **highly correlated** 250-dimensional Gaussian distribution

Comparison of algorithms on **highly correlated** 250-dimensional Gaussian distribution

- Do **1,000,000** draws with both Random Walk Metropolis and Gibbs, thinning by 1000

Comparison of algorithms on **highly correlated** 250-dimensional Gaussian distribution

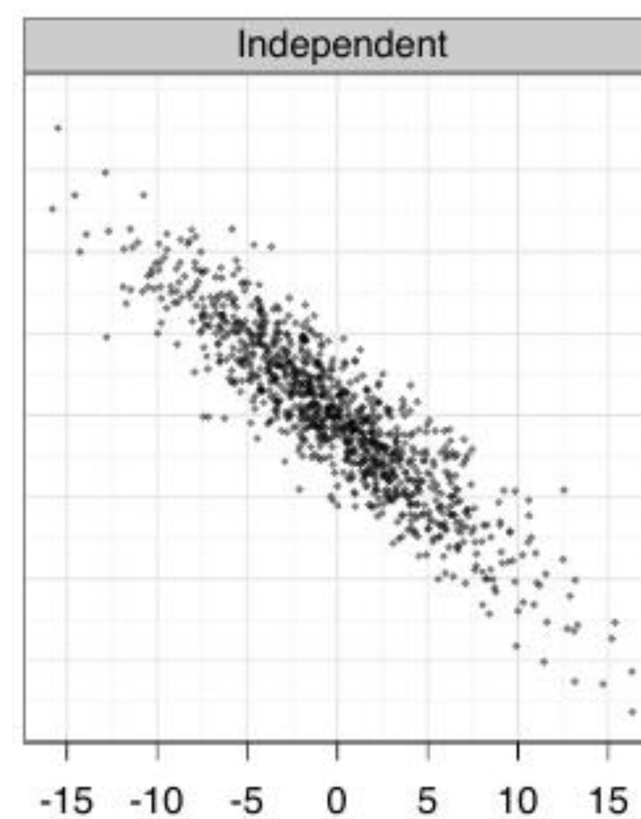
- Do **1,000,000** draws with both Random Walk Metropolis and Gibbs, thinning by 1000
- Do **1,000** draws using Stan’s NUTS algorithm (no thinning)

Comparison of algorithms on **highly correlated** 250-dimensional Gaussian distribution

- Do **1,000,000** draws with both Random Walk Metropolis and Gibbs, thinning by 1000
- Do **1,000** draws using Stan’s NUTS algorithm (no thinning)
- Do 1,000 independent draws (we can do this for multivariate normal)

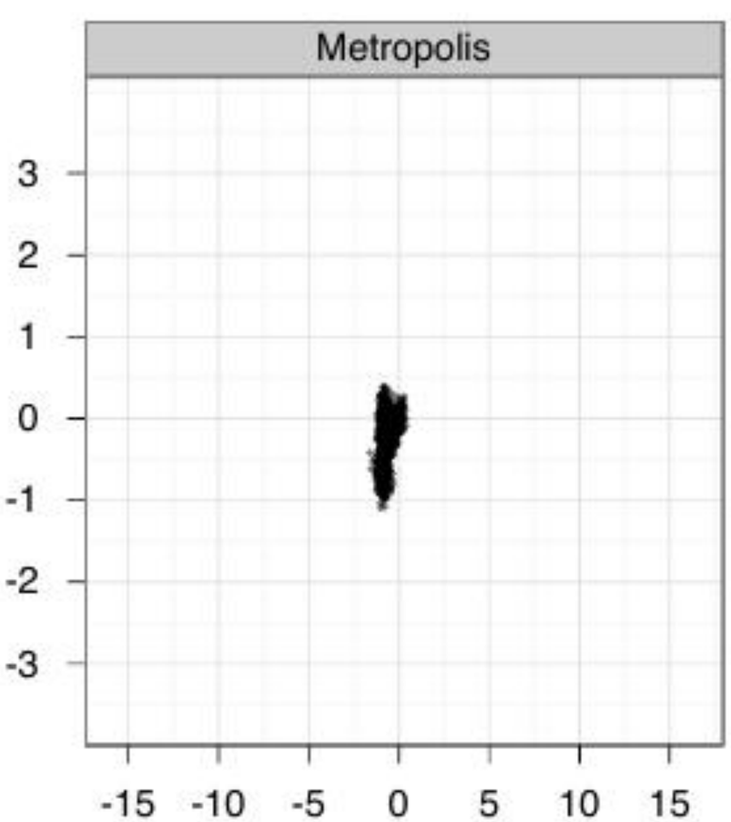
Comparison of algorithms on **highly correlated** 250-dimensional Gaussian distribution

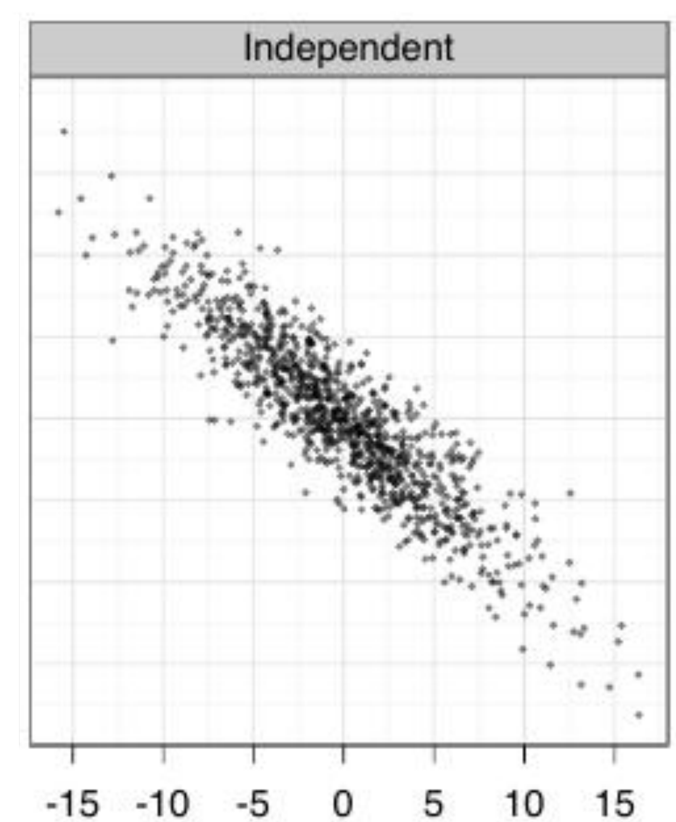
- Do **1,000,000** draws with both Random Walk Metropolis and Gibbs, thinning by 1000
- Do **1,000** draws using Stan's NUTS algorithm (no thinning)
- Do 1,000 independent draws (we can do this for multivariate normal)



Comparison of algorithms on **highly correlated** 250-dimensional Gaussian distribution

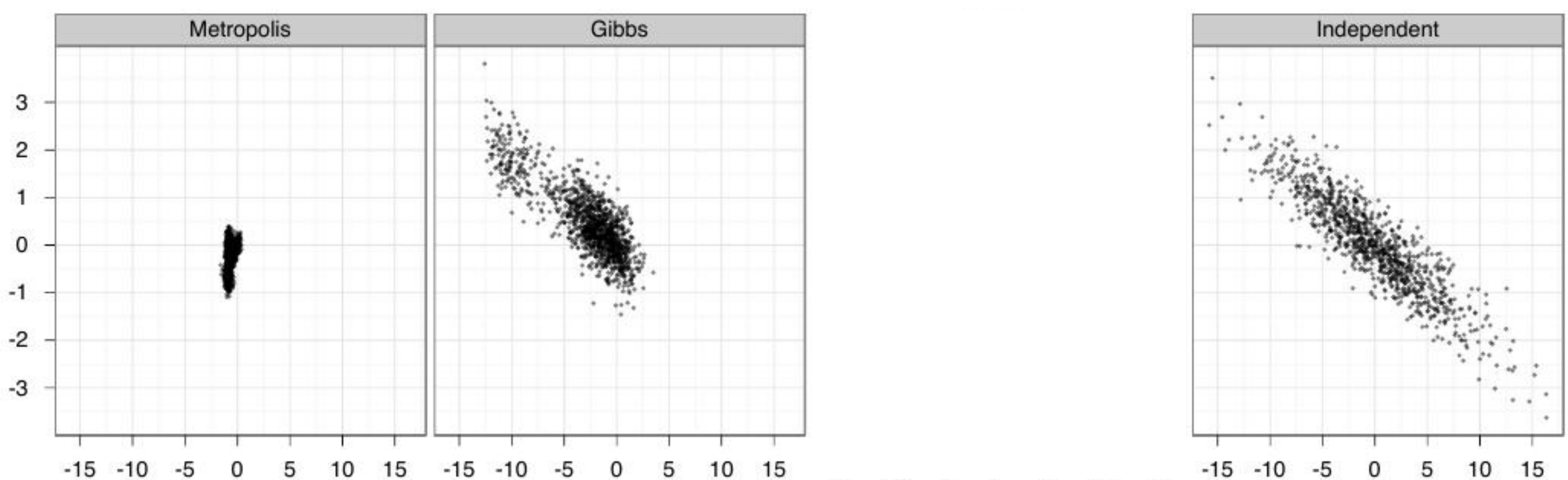
- Do **1,000,000** draws with both Random Walk Metropolis and Gibbs, thinning by 1000
- Do **1,000** draws using Stan's NUTS algorithm (no thinning)
- Do 1,000 independent draws (we can do this for multivariate normal)





Comparison of algorithms on **highly correlated** 250-dimensional Gaussian distribution

- Do **1,000,000** draws with both Random Walk Metropolis and Gibbs, thinning by 1000
- Do **1,000** draws using Stan's NUTS algorithm (no thinning)
- Do 1,000 independent draws (we can do this for multivariate normal)



Comparison of algorithms on **highly correlated** 250-dimensional Gaussian distribution

- Do **1,000,000** draws with both Random Walk Metropolis and Gibbs, thinning by 1000
- Do **1,000** draws using Stan's NUTS algorithm (no thinning)
- Do 1,000 independent draws (we can do this for multivariate normal)