# LECTURE 4

# LOCAL INFERENCE ALGORITHMS

**Saifuddin Syed**

# TARGET DISTRIBUTIONS IN PHYSICS

# TARGET DISTRIBUTIONS IN PHYSICS



- Suppose $\mathbb{X} = \mathbb{R}^d$ and we have target $\mu(\mathrm{d}q) = \mu(q)\mathrm{d}q$

# TARGET DISTRIBUTIONS IN PHYSICS



- Suppose $\mathbb{X} = \mathbb{R}^d$ and we have target $\mu(\mathrm{d}q) = \mu(q)\mathrm{d}q$

$$\mu(q) = \frac{\exp(-U(q))}{Z}, \qquad Z = \int_{\mathbb{R}^d} \exp(-U(q))\mathrm{d}q$$

# TARGET DISTRIBUTIONS IN PHYSICS



- Suppose $\mathbb{X} = \mathbb{R}^d$ and we have target $\mu(\mathrm{d}q) = \mu(q)\mathrm{d}q$

$$\mu(q) = \frac{\exp(-U(q))}{Z}, \qquad Z = \int_{\mathbb{R}^d} \exp(-U(q))\mathrm{d}q$$

  - We refer to $q \in \mathbb{R}^d$ as the **position**

# TARGET DISTRIBUTIONS IN PHYSICS



- Suppose $\mathbb{X} = \mathbb{R}^d$ and we have target $\mu(\mathrm{d}q) = \mu(q)\mathrm{d}q$

$$\mu(q) = \frac{\exp(-U(q))}{Z}, \qquad Z = \int_{\mathbb{R}^d} \exp(-U(q))\mathrm{d}q$$

  - We refer to $q \in \mathbb{R}^d$ as the **position**
  - $U(q)$ is the **potential energy** at $q$

# TARGET DISTRIBUTIONS IN PHYSICS



- Suppose $\mathbb{X} = \mathbb{R}^d$ and we have target $\mu(\mathrm{d}q) = \mu(q)\mathrm{d}q$

$$\mu(q) = \frac{\exp(-U(q))}{Z}, \qquad Z = \int_{\mathbb{R}^d} \exp(-U(q))\mathrm{d}q$$

  - We refer to $q \in \mathbb{R}^d$ as the **position**
  - $U(q)$ is the **potential energy** at $q$

- For example if if $\mu = \mathcal{N}(q_0, \Sigma)$ then,

# TARGET DISTRIBUTIONS IN PHYSICS



- Suppose $\mathbb{X} = \mathbb{R}^d$ and we have target $\mu(\mathrm{d}q) = \mu(q)\mathrm{d}q$

$$\mu(q) = \frac{\exp(-U(q))}{Z}, \qquad Z = \int_{\mathbb{R}^d} \exp(-U(q))\mathrm{d}q$$

  - We refer to $q \in \mathbb{R}^d$ as the **position**
  - $U(q)$ is the **potential energy** at $q$

- For example if if $\mu = \mathcal{N}(q_0, \Sigma)$ then,

$$U(q) = \frac{1}{2}(q - q_0)^\top \Sigma^{-1}(q - q_0)$$

# TARGET DISTRIBUTIONS IN PHYSICS



- Suppose $\mathbb{X} = \mathbb{R}^d$ and we have target $\mu(\mathrm{d}q) = \mu(q)\mathrm{d}q$

$$\mu(q) = \frac{\exp(-U(q))}{Z}, \qquad Z = \int_{\mathbb{R}^d} \exp(-U(q))\mathrm{d}q$$

  - We refer to $q \in \mathbb{R}^d$ as the **position**
  - $U(q)$ is the **potential energy** at $q$

- For example if if $\mu = \mathcal{N}(q_0, \Sigma)$ then,

$$U(q) = \frac{1}{2}(q - q_0)^\top \Sigma^{-1}(q - q_0)$$

  - Gaussians correspond to quadradic potentials

# TARGET DISTRIBUTIONS IN PHYSICS

# TARGET DISTRIBUTIONS IN PHYSICS



- Suppose we have a Metropolis-Hastings chain with a symmetric proposal

# TARGET DISTRIBUTIONS IN PHYSICS



- Suppose we have a Metropolis-Hastings chain with a symmetric proposal

$$\alpha(q, q') = 1 \wedge \frac{\mu(q')}{\mu(q)} = 1 \wedge \exp(-\Delta U)$$

# TARGET DISTRIBUTIONS IN PHYSICS



- Suppose we have a Metropolis-Hastings chain with a symmetric proposal

$$\alpha(q, q') = 1 \wedge \frac{\mu(q')}{\mu(q)} = 1 \wedge \exp(-\Delta U)$$

  - Where $\Delta U = U(q') - U(q)$ is the change in potential energy

# TARGET DISTRIBUTIONS IN PHYSICS



- Suppose we have a Metropolis-Hastings chain with a symmetric proposal

$$\alpha(q, q') = 1 \wedge \frac{\mu(q')}{\mu(q)} = 1 \wedge \exp(-\Delta U)$$

  - Where $\Delta U = U(q') - U(q)$ is the change in potential energy
  - If $\Delta U < 0$ always accept

# TARGET DISTRIBUTIONS IN PHYSICS



- Suppose we have a Metropolis-Hastings chain with a symmetric proposal

$$\alpha(q, q') = 1 \wedge \frac{\mu(q')}{\mu(q)} = 1 \wedge \exp(-\Delta U)$$

  - Where $\Delta U = U(q') - U(q)$ is the change in potential energy
  - If $\Delta U < 0$ always accept
  - If $\Delta U > 0$ reject with probaility $\exp(-\Delta U)$

# TARGET DISTRIBUTIONS IN PHYSICS



- Suppose we have a Metropolis-Hastings chain with a symmetric proposal

$$\alpha(q, q') = 1 \wedge \frac{\mu(q')}{\mu(q)} = 1 \wedge \exp(-\Delta U)$$

  - Where $\Delta U = U(q') - U(q)$ is the change in potential energy
  - If $\Delta U < 0$ always accept
  - If $\Delta U > 0$ reject with probaility $\exp(-\Delta U)$

- MH generates a “particle” with position $q_t$

# TARGET DISTRIBUTIONS IN PHYSICS



- Suppose we have a Metropolis-Hastings chain with a symmetric proposal

$$\alpha(q, q') = 1 \wedge \frac{\mu(q')}{\mu(q)} = 1 \wedge \exp(-\Delta U)$$

  - Where $\Delta U = U(q') - U(q)$ is the change in potential energy
  - If $\Delta U < 0$ always accept
  - If $\Delta U > 0$ reject with probaility $\exp(-\Delta U)$

- MH generates a “particle” with position $q_t$
- The proposal at time $t$ defines proposes a change in position $\Delta q_t = \epsilon p_t$

# TARGET DISTRIBUTIONS IN PHYSICS



- Suppose we have a Metropolis-Hastings chain with a symmetric proposal

$$\alpha(q, q') = 1 \wedge \frac{\mu(q')}{\mu(q)} = 1 \wedge \exp(-\Delta U)$$

  - Where $\Delta U = U(q') - U(q)$ is the change in potential energy
  - If $\Delta U < 0$ always accept
  - If $\Delta U > 0$ reject with probaility $\exp(-\Delta U)$

- MH generates a “particle” with position $q_t$
- The proposal at time $t$ defines proposes a change in position $\Delta q_t = \epsilon p_t$
  - $\epsilon$ is the step size a momentum $p_t$, proposing direction of travel

# TARGET DISTRIBUTIONS IN PHYSICS



- Suppose we have a Metropolis-Hastings chain with a symmetric proposal

$$\alpha(q, q') = 1 \wedge \frac{\mu(q')}{\mu(q)} = 1 \wedge \exp(-\Delta U)$$

  - Where $\Delta U = U(q') - U(q)$ is the change in potential energy
  - If $\Delta U < 0$ always accept
  - If $\Delta U > 0$ reject with probaility $\exp(-\Delta U)$

- MH generates a “particle” with position $q_t$
- The proposal at time $t$ defines proposes a change in position $\Delta q_t = \epsilon p_t$
  - $\epsilon$ is the step size a momentum $p_t$, proposing direction of travel
  - $p_t$ is the momentum indicating direction of travel

# TARGET DISTRIBUTIONS IN PHYSICS



- Suppose we have a Metropolis-Hastings chain with a symmetric proposal

$$\alpha(q, q') = 1 \wedge \frac{\mu(q')}{\mu(q)} = 1 \wedge \exp(-\Delta U)$$

  - Where $\Delta U = U(q') - U(q)$ is the change in potential energy
  - If $\Delta U < 0$ always accept
  - If $\Delta U > 0$ reject with probaility $\exp(-\Delta U)$

- MH generates a “particle” with position $q_t$
- The proposal at time $t$ defines proposes a change in position $\Delta q_t = \epsilon p_t$
  - $\epsilon$ is the step size a momentum $p_t$, proposing direction of travel
  - $p_t$ is the momentum indicating direction of travel

$$q' = q + \Delta q, \quad \Delta q = \epsilon p$$

# PHASE SPACE

# PHASE SPACE



- Define targets distribution $\pi(z)$ over phase space in terms of the **Hamiltonian** $H : \mathbb{R}^{2d} \to \mathbb{R}$

$$\pi(z) \propto \exp(-H(q,p))$$

# PHASE SPACE



- Define targets distribution $\pi(z)$ over phase space in terms of the **Hamiltonian** $H : \mathbb{R}^{2d} \to \mathbb{R}$

$$\pi(z) \propto \exp(-H(q,p))$$

- We will assume $H(q,p)$ is separable

$$H(q,p) = U(q) + K(p)$$

# PHASE SPACE



- Define targets distribution $\pi(z)$ over phase space in terms of the **Hamiltonian** $H : \mathbb{R}^{2d} \to \mathbb{R}$

$$\pi(z) \propto \exp(-H(q,p))$$

- We will assume $H(q,p)$ is separable

$$H(q,p) = U(q) + K(p)$$

  - $U(q)$ is the potential energy encodes the marginal density of the position $q$ (i.e. target)

# PHASE SPACE



- Define targets distribution $\pi(z)$ over phase space in terms of the **Hamiltonian** $H : \mathbb{R}^{2d} \to \mathbb{R}$

$$\pi(z) \propto \exp(-H(q,p))$$

- We will assume $H(q,p)$ is separable

$$H(q,p) = U(q) + K(p)$$

  - $U(q)$ is the potential energy encodes the marginal density of the position $q$ (i.e. target)
  - $K(p)$ is kinetic energy encodes the marginal density of the momentum $p$ (i.e. proposal)

# PHASE SPACE



- Define targets distribution $\pi(z)$ over phase space in terms of the **Hamiltonian** $H : \mathbb{R}^{2d} \to \mathbb{R}$

$$\pi(z) \propto \exp(-H(q,p))$$

- We will assume $H(q,p)$ is separable

$$H(q,p) = U(q) + K(p)$$

  - $U(q)$ is the potential energy encodes the marginal density of the position $q$ (i.e. target)
  - $K(p)$ is kinetic energy encodes the marginal density of the momentum $p$ (i.e. proposal)

$$\pi(z) \propto \exp(-U(q))\exp(-K(p))$$

# PHASE SPACE



- Define targets distribution $\pi(z)$ over phase space in terms of the **Hamiltonian** $H : \mathbb{R}^{2d} \to \mathbb{R}$

$$\pi(z) \propto \exp(-H(q,p))$$

- We will assume $H(q,p)$ is separable

$$H(q,p) = U(q) + K(p)$$

  - $U(q)$ is the potential energy encodes the marginal density of the position $q$ (i.e. target)
  - $K(p)$ is kinetic energy encodes the marginal density of the momentum $p$ (i.e. proposal)

$$\pi(z) \propto \exp(-U(q))\exp(-K(p))$$

- For example, $\mathcal{N}(0,M)$ corresponds to kinetic energy,

$$K(p) = \frac{1}{2} p^\top M^{-1} p$$

# PHASE SPACE



- Define targets distribution $\pi(z)$ over phase space in terms of the **Hamiltonian** $H : \mathbb{R}^{2d} \to \mathbb{R}$

$$\pi(z) \propto \exp(-H(q,p))$$

- We will assume $H(q,p)$ is separable

$$H(q,p) = U(q) + K(p)$$

  - $U(q)$ is the potential energy encodes the marginal density of the position $q$ (i.e. target)
  - $K(p)$ is kinetic energy encodes the marginal density of the momentum $p$ (i.e. proposal)

$$\pi(z) \propto \exp(-U(q))\exp(-K(p))$$

- For example, $\mathcal{N}(0,M)$ corresponds to kinetic energy,

$$K(p) = \frac{1}{2} p^\top M^{-1} p$$

- If $M = mI$ then recover the classical kinetic energy formula

$$K(p) = \frac{1}{2m} \|p\|^2$$

# PHASE SPACE



- Define targets distribution $\pi(z)$ over phase space in terms of the **Hamiltonian** $H : \mathbb{R}^{2d} \to \mathbb{R}$

$$\pi(z) \propto \exp(-H(q,p))$$

- We will assume $H(q,p)$ is separable

$$H(q,p) = U(q) + K(p)$$

  - $U(q)$ is the potential energy encodes the marginal density of the position $q$ (i.e. target)
  - $K(p)$ is kinetic energy encodes the marginal density of the momentum $p$ (i.e. proposal)

$$\pi(z) \propto \exp(-U(q))\exp(-K(p))$$

- For example, $\mathcal{N}(0,M)$ corresponds to kinetic energy,

$$K(p) = \frac{1}{2} p^\top M^{-1} p$$

- If $M = mI$ then recover the classical kinetic energy formula

$$K(p) = \frac{1}{2m} \|p\|^2$$

  - $M$ is referred to the **mass matrix**

# PROPOSALS AS MOMENTUM

# PROPOSALS AS MOMENTUM



- **Example**: RWM $\qquad p = \Sigma W, \qquad W \sim \mathcal{N}(0, I)$

# PROPOSALS AS MOMENTUM



- **Example**: RWM

$$p = \Sigma W, \qquad W \sim \mathcal{N}(0, I)$$

$$q' = q + \epsilon p \sim \mathcal{N}(q, \epsilon\Sigma)$$

# PROPOSALS AS MOMENTUM



- **Example**: RWM

$$p = \Sigma W, \qquad W \sim \mathcal{N}(0, I)$$

$$q' = q + \epsilon p \sim \mathcal{N}(q, \epsilon\Sigma)$$

- **Example:** MALA discretises Langevin dynamics:

$$\mathrm{d}Y_\tau = -\frac{1}{2}\Sigma\nabla U(Y_\tau)\mathrm{d}\tau + \sqrt{\Sigma}\mathrm{d}W_\tau$$

# PROPOSALS AS MOMENTUM



- **Example**: RWM

$$p = \Sigma W, \qquad W \sim \mathcal{N}(0, I)$$

$$q' = q + \epsilon p \sim \mathcal{N}(q, \epsilon\Sigma)$$

- **Example:** MALA discretises Langevin dynamics:

$$\mathrm{d}Y_\tau = -\frac{1}{2}\Sigma\nabla U(Y_\tau)\mathrm{d}\tau + \sqrt{\Sigma}\mathrm{d}W_\tau$$

  - Results in update:

# PROPOSALS AS MOMENTUM



- **Example**: RWM

$$p = \Sigma W, \qquad W \sim \mathcal{N}(0, I)$$

$$q' = q + \epsilon p \sim \mathcal{N}(q, \epsilon\Sigma)$$

- **Example:** MALA discretises Langevin dynamics:

$$\mathrm{d}Y_\tau = -\frac{1}{2}\Sigma\nabla U(Y_\tau)\mathrm{d}\tau + \sqrt{\Sigma}\mathrm{d}W_\tau$$

  - Results in update:

$$p = -\frac{\epsilon}{2}\Sigma\nabla U(q) + \sqrt{\Sigma}W, \qquad W \sim \mathcal{N}(0, I)$$

# PROPOSALS AS MOMENTUM



- **Example**: RWM

$$p = \Sigma W, \qquad W \sim \mathcal{N}(0, I)$$

$$q' = q + \epsilon p \sim \mathcal{N}(q, \epsilon\Sigma)$$

- **Example:** MALA discretises Langevin dynamics:

$$\mathrm{d}Y_\tau = -\frac{1}{2}\Sigma\nabla U(Y_\tau)\mathrm{d}\tau + \sqrt{\Sigma}\mathrm{d}W_\tau$$

  - Results in update:

$$p = -\frac{\epsilon}{2}\Sigma\nabla U(q) + \sqrt{\Sigma}W, \qquad W \sim \mathcal{N}(0, I)$$

$$q' = q + \epsilon p \sim \mathcal{N}\left(q - \frac{\epsilon^2}{2}\Sigma\nabla U(q), \epsilon^2\Sigma\right)$$

# PROPOSALS AS MOMENTUM



- **Example**: RWM

$$p = \Sigma W, \qquad W \sim \mathcal{N}(0, I)$$

$$q' = q + \epsilon p \sim \mathcal{N}(q, \epsilon\Sigma)$$

- **Example:** MALA discretises Langevin dynamics:

$$\mathrm{d}Y_\tau = -\frac{1}{2}\Sigma\nabla U(Y_\tau)\mathrm{d}\tau + \sqrt{\Sigma}\mathrm{d}W_\tau$$

  - Results in update:

$$p = -\frac{\epsilon}{2}\Sigma\nabla U(q) + \sqrt{\Sigma}W, \qquad W \sim \mathcal{N}(0, I)$$

$$q' = q + \epsilon p \sim \mathcal{N}\left(q - \frac{\epsilon^2}{2}\Sigma\nabla U(q), \epsilon^2\Sigma\right)$$

- MALA has momentum pushing particle locally towards a lower potential energy

# PROPOSALS AS MOMENTUM



- **Example**: RWM

$$p = \Sigma W, \qquad W \sim \mathcal{N}(0, I)$$

$$q' = q + \epsilon p \sim \mathcal{N}(q, \epsilon\Sigma)$$

- **Example:** MALA discretises Langevin dynamics:

$$\mathrm{d}Y_\tau = -\frac{1}{2}\Sigma\nabla U(Y_\tau)\mathrm{d}\tau + \sqrt{\Sigma}\mathrm{d}W_\tau$$

  - Results in update:

$$p = -\frac{\epsilon}{2}\Sigma\nabla U(q) + \sqrt{\Sigma}W, \qquad W \sim \mathcal{N}(0, I)$$

$$q' = q + \epsilon p \sim \mathcal{N}\left(q - \frac{\epsilon^2}{2}\Sigma\nabla U(q), \epsilon^2\Sigma\right)$$

- MALA has momentum pushing particle locally towards a lower potential energy
  - Momentum retains no memory of previous momentums

# PROPOSALS AS MOMENTUM



- **Example**: RWM

$$p = \Sigma W, \qquad W \sim \mathcal{N}(0, I)$$

$$q' = q + \epsilon p \sim \mathcal{N}(q, \epsilon\Sigma)$$

- **Example:** MALA discretises Langevin dynamics:

$$\mathrm{d}Y_\tau = -\frac{1}{2}\Sigma\nabla U(Y_\tau)\mathrm{d}\tau + \sqrt{\Sigma}\mathrm{d}W_\tau$$

  - Results in update:

$$p = -\frac{\epsilon}{2}\Sigma\nabla U(q) + \sqrt{\Sigma}W, \qquad W \sim \mathcal{N}(0, I)$$

$$q' = q + \epsilon p \sim \mathcal{N}\left(q - \frac{\epsilon^2}{2}\Sigma\nabla U(q), \epsilon^2\Sigma\right)$$

- MALA has momentum pushing particle locally towards a lower potential energy
  - Momentum retains no memory of previous momentums
- Can fix this if we track position and momentum. Let $z = (q, p) \in \mathbb{R}^{2d}$ be the phase space.

# HAMILTONIAN DYNAMICS

# HAMILTONIAN DYNAMICS



- Let $q \in \mathbb{R}^d$ and $p \in \mathbb{R}^d$ denote the posituon and momentum, $U : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable potential function, and $M \in \mathbb{R}^{d \times d}$ be a symmetric positive definite matrix.

$$\frac{\mathrm{d}q}{\mathrm{d}t} = M^{-1}p, \qquad \frac{\mathrm{d}p}{\mathrm{d}t} = -\nabla U(q)$$

# HAMILTONIAN DYNAMICS



- Let $q \in \mathbb{R}^d$ and $p \in \mathbb{R}^d$ denote the posituon and momentum, $U : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable potential function, and $M \in \mathbb{R}^{d \times d}$ be a symmetric positive definite matrix.

$$\frac{\mathrm{d}q}{\mathrm{d}t} = M^{-1}p, \qquad \frac{\mathrm{d}p}{\mathrm{d}t} = -\nabla U(q)$$

- In the case when, these equations are equivalent to Newtonian mechanics with a non-dissipative force $F(q) = -\nabla U(q)$ arising from the potential field $U$.

$$M\frac{\mathrm{d}^2q}{\mathrm{d}t^2} = \frac{\mathrm{d}p}{\mathrm{d}t} = -\nabla U(q) = F(q)$$

# HAMILTONIAN DYNAMICS



- Let $q \in \mathbb{R}^d$ and $p \in \mathbb{R}^d$ denote the posituon and momentum, $U : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable potential function, and $M \in \mathbb{R}^{d \times d}$ be a symmetric positive definite matrix.

$$\frac{\mathrm{d}q}{\mathrm{d}t} = M^{-1}p, \qquad \frac{\mathrm{d}p}{\mathrm{d}t} = -\nabla U(q)$$

- In the case when, these equations are equivalent to Newtonian mechanics with a non-dissipative force $F(q) = -\nabla U(q)$ arising from the potential field $U$.

$$M\frac{\mathrm{d}^2q}{\mathrm{d}t^2} = \frac{\mathrm{d}p}{\mathrm{d}t} = -\nabla U(q) = F(q)$$

- Recall that if $H(z) = U(q) + K(p)$ then it has gradient,

$$\nabla_q H = \nabla U \qquad \nabla_p H = \nabla K = M^{-1}p$$

# HAMILTONIAN DYNAMICS



- Let $q \in \mathbb{R}^d$ and $p \in \mathbb{R}^d$ denote the posituon and momentum, $U : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable potential function, and $M \in \mathbb{R}^{d \times d}$ be a symmetric positive definite matrix.

$$\frac{\mathrm{d}q}{\mathrm{d}t} = M^{-1}p, \qquad \frac{\mathrm{d}p}{\mathrm{d}t} = -\nabla U(q)$$

- In the case when, these equations are equivalent to Newtonian mechanics with a non-dissipative force $F(q) = -\nabla U(q)$ arising from the potential field $U$.

$$M\frac{\mathrm{d}^2q}{\mathrm{d}t^2} = \frac{\mathrm{d}p}{\mathrm{d}t} = -\nabla U(q) = F(q)$$

- Recall that if $H(z) = U(q) + K(p)$ then it has gradient,

$$\nabla_q H = \nabla U \qquad \nabla_p H = \nabla K = M^{-1}p$$

- Can express in terms of $z(t) = (q(t), p(t))$ and hamiltonian

$$\frac{\mathrm{d}z}{\mathrm{d}t} = J\nabla_z H(z), \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

# CONSERVATION OF ENERGY

# CONSERVATION OF ENERGY



- Define the flow of the Hamiltonian flow $z(t) = (q(t), p(t))$ satisfying

$$\frac{\mathrm{d}z}{\mathrm{d}t} = J\nabla_z H(z), \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

# CONSERVATION OF ENERGY



- Define the flow of the Hamiltonian flow $z(t) = (q(t), p(t))$ satisfying

$$\frac{\mathrm{d}z}{\mathrm{d}t} = J\nabla_z H(z), \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

- **Proposition:** Hamiltonian dynamics preserves the Hamiltonian:

$$H(z(t)) = H(z(0))$$

# CONSERVATION OF ENERGY



- Define the flow of the Hamiltonian flow $z(t) = (q(t), p(t))$ satisfying

$$\frac{\mathrm{d}z}{\mathrm{d}t} = J\nabla_z H(z), \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

- **Proposition:** Hamiltonian dynamics preserves the Hamiltonian:

$$H(z(t)) = H(z(0))$$

  - Changes in potential energy (position) are offset the change in kinetic energy (momentum)!

# CONSERVATION OF ENERGY



- Define the flow of the Hamiltonian flow $z(t) = (q(t), p(t))$ satisfying

$$\frac{\mathrm{d}z}{\mathrm{d}t} = J\nabla_z H(z), \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

- **Proposition:** Hamiltonian dynamics preserves the Hamiltonian:

$$H(z(t)) = H(z(0))$$

  - Changes in potential energy (position) are offset the change in kinetic energy (momentum)!

- **Proof:**

# CONSERVATION OF ENERGY



- Define the flow of the Hamiltonian flow $z(t) = (q(t), p(t))$ satisfying

$$\frac{\mathrm{d}z}{\mathrm{d}t} = J\nabla_z H(z), \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

- **Proposition:** Hamiltonian dynamics preserves the Hamiltonian:

$$H(z(t)) = H(z(0))$$

  - Changes in potential energy (position) are offset the change in kinetic energy (momentum)!

- **Proof:**

$$\frac{\mathrm{d}}{\mathrm{d}t} H(z(t)) = \frac{\mathrm{d}}{\mathrm{d}t} U(q(t)) + \frac{\mathrm{d}}{\mathrm{d}t} K(p(t))$$

# CONSERVATION OF ENERGY



- Define the flow of the Hamiltonian flow $z(t) = (q(t), p(t))$ satisfying

$$\frac{\mathrm{d}z}{\mathrm{d}t} = J\nabla_z H(z), \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

- **Proposition:** Hamiltonian dynamics preserves the Hamiltonian:

$$H(z(t)) = H(z(0))$$

  - Changes in potential energy (position) are offset the change in kinetic energy (momentum)!

- **Proof:**

$$\frac{\mathrm{d}}{\mathrm{d}t} H(z(t)) = \frac{\mathrm{d}}{\mathrm{d}t} U(q(t)) + \frac{\mathrm{d}}{\mathrm{d}t} K(p(t))$$

$$= \nabla U(q(t))^\top \dot{q}(t) + \nabla K(p(t))^\top \dot{p}(t)$$

# CONSERVATION OF ENERGY



- Define the flow of the Hamiltonian flow $z(t) = (q(t), p(t))$ satisfying

$$\frac{\mathrm{d}z}{\mathrm{d}t} = J\nabla_z H(z), \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

- **Proposition:** Hamiltonian dynamics preserves the Hamiltonian:

$$H(z(t)) = H(z(0))$$

  - Changes in potential energy (position) are offset the change in kinetic energy (momentum)!

- **Proof:**

$$\frac{\mathrm{d}}{\mathrm{d}t} H(z(t)) = \frac{\mathrm{d}}{\mathrm{d}t} U(q(t)) + \frac{\mathrm{d}}{\mathrm{d}t} K(p(t))$$

$$= \nabla U(q(t))^\top \dot{q}(t) + \nabla K(p(t))^\top \dot{p}(t)$$

$$= \nabla U(q(t))^\top \nabla K(p(t))$$

# CONSERVATION OF ENERGY



- Define the flow of the Hamiltonian flow $z(t) = (q(t), p(t))$ satisfying

$$\frac{\mathrm{d}z}{\mathrm{d}t} = J\nabla_z H(z), \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

- **Proposition:** Hamiltonian dynamics preserves the Hamiltonian:

$$H(z(t)) = H(z(0))$$

  - Changes in potential energy (position) are offset the change in kinetic energy (momentum)!

- **Proof:**

$$\frac{\mathrm{d}}{\mathrm{d}t}H(z(t)) = \frac{\mathrm{d}}{\mathrm{d}t}U(q(t)) + \frac{\mathrm{d}}{\mathrm{d}t}K(p(t))$$

$$= \nabla U(q(t))^\top \dot{q}(t) + \nabla K(p(t))^\top \dot{p}(t)$$

$$= \nabla U(q(t))^\top \nabla K(p(t)) \quad - \nabla K(q(t))^\top \nabla U(q(t))$$

# CONSERVATION OF ENERGY



- Define the flow of the Hamiltonian flow $z(t) = (q(t), p(t))$ satisfying

$$\frac{\mathrm{d}z}{\mathrm{d}t} = J\nabla_z H(z), \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

- **Proposition:** Hamiltonian dynamics preserves the Hamiltonian:

$$H(z(t)) = H(z(0))$$

  - Changes in potential energy (position) are offset the change in kinetic energy (momentum)!

- **Proof:**

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} H(z(t)) &= \frac{\mathrm{d}}{\mathrm{d}t} U(q(t)) + \frac{\mathrm{d}}{\mathrm{d}t} K(p(t)) \\ &= \nabla U(q(t))^\top \dot{q}(t) + \nabla K(p(t))^\top \dot{p}(t) \\ &= \nabla U(q(t))^\top \nabla K(p(t)) \quad - \nabla K(q(t))^\top \nabla U(q(t)) \\ &= 0 \end{aligned}$$

# EXAMPLE: SIMPLE HARMONIC OSCILLATOR

# EXAMPLE: SIMPLE HARMONIC OSCILLATOR



- Suppose $d = 1$

$$U(q) = \frac{1}{2}\omega^2 q^2 \qquad K(p) = \frac{1}{2}p^2$$

# EXAMPLE: SIMPLE HARMONIC OSCILLATOR



- Suppose $d = 1$

$$U(q) = \frac{1}{2}\omega^2 q^2 \qquad K(p) = \frac{1}{2}p^2$$

- Then $H(z) = U(q) + K(p)$ satisfies

$$H(z) = z^\top L z, \qquad L = \begin{pmatrix} \omega^2 & 0 \\ 0 & 1 \end{pmatrix}.$$

# EXAMPLE: SIMPLE HARMONIC OSCILLATOR



- Suppose $d = 1$

$$U(q) = \frac{1}{2}\omega^2 q^2 \qquad K(p) = \frac{1}{2}p^2$$

- Then $H(z) = U(q) + K(p)$ satisfies

$$H(z) = z^\top L z, \qquad L = \begin{pmatrix} \omega^2 & 0 \\ 0 & 1 \end{pmatrix}.$$

- Hamilton's equations reduce to

$$\dot{q}_t = p, \qquad \dot{p}_t = -\omega^2 q$$

# EXAMPLE: SIMPLE HARMONIC OSCILLATOR



- Suppose $d = 1$

$$U(q) = \frac{1}{2}\omega^2 q^2 \qquad K(p) = \frac{1}{2}p^2$$

- Then $H(z) = U(q) + K(p)$ satisfies

$$H(z) = z^\top L z, \qquad L = \begin{pmatrix} \omega^2 & 0 \\ 0 & 1 \end{pmatrix}.$$

- Hamilton's equations reduce to

$$\dot{q}_t = p, \qquad \dot{p}_t = -\omega^2 q$$

- Solutions have the form:

$$q(t) = \cos(\omega t + \phi), \qquad p(t) = \omega \sin(\omega t + \phi)$$

# EXAMPLE: SIMPLE HARMONIC OSCILLATOR



- Suppose $d = 1$

$$U(q) = \frac{1}{2}\omega^2 q^2 \qquad K(p) = \frac{1}{2}p^2$$

- Then $H(z) = U(q) + K(p)$ satisfies

$$H(z) = z^\top L z, \qquad L = \begin{pmatrix} \omega^2 & 0 \\ 0 & 1 \end{pmatrix}.$$

- Hamilton's equations reduce to

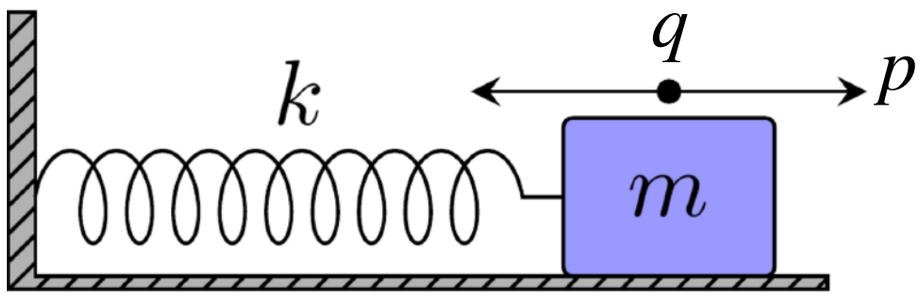
$$\dot{q}_t = p, \qquad \dot{p}_t = -\omega^2 q$$

- Solutions have the form:

$$q(t) = \cos(\omega t + \phi), \qquad p(t) = \omega \sin(\omega t + \phi)$$

- Corresponds to the motion of mass spring system:

$$m\ddot{q}_t = -kq$$



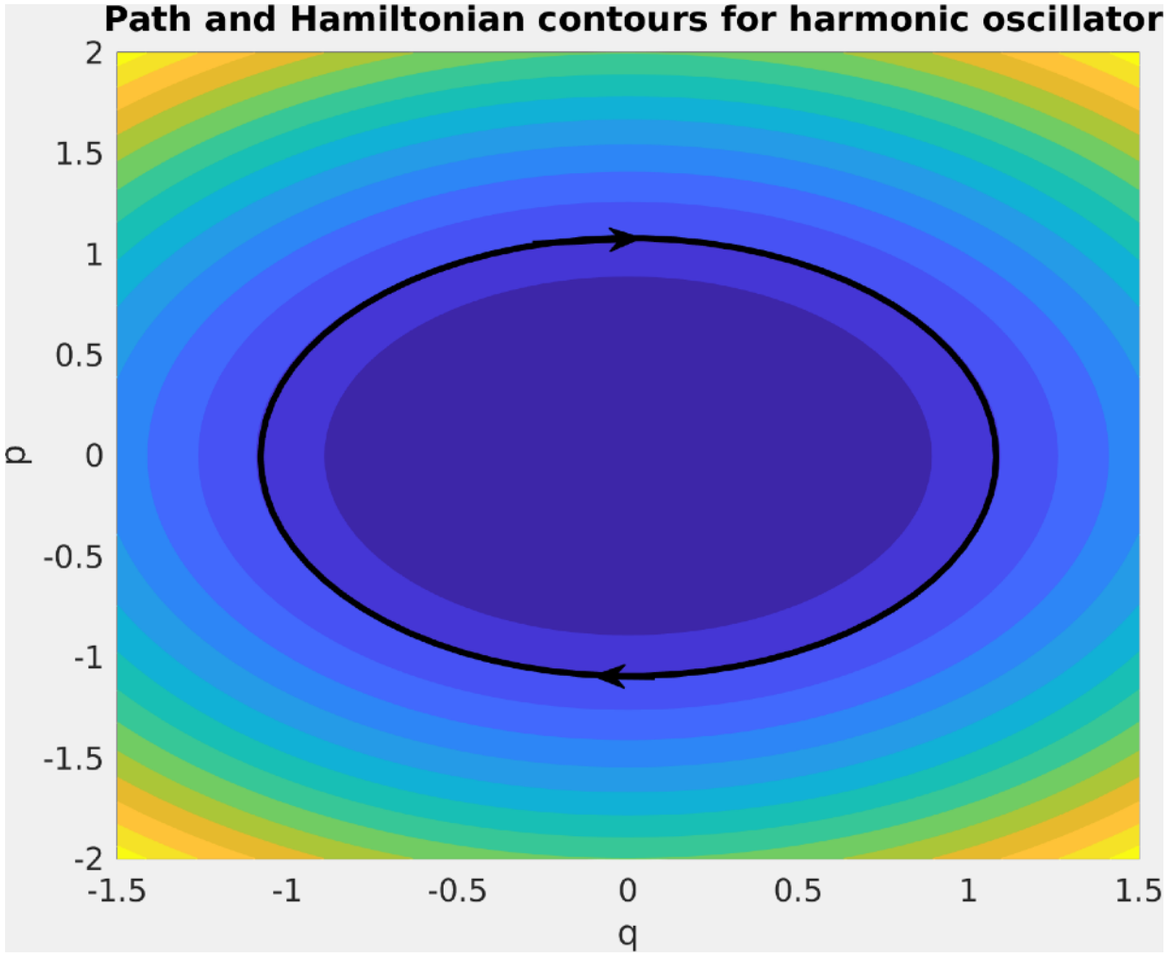

$$\omega^2 = \frac{k}{m}$$

# EXAMPLE: SIMPLE HARMONIC OSCILLATOR



- Note that the hamiltonian is conserved:

$$H(q(t), p(t)) = \frac{1}{2}\omega^2 q(t)^2 + \frac{1}{2}p(t)^2 = \omega^2$$



# PROPERTIES OF HAMILTONIAN MECHANICS

# PROPERTIES OF HAMILTONIAN MECHANICS



- Given $z(t) = (q(t), p(t))$ satisfying

$$\frac{\mathrm{d}z}{\mathrm{d}t} = J \nabla_z H(z), \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

# PROPERTIES OF HAMILTONIAN MECHANICS



- Given $z(t) = (q(t), p(t))$ satisfying

$$\frac{\mathrm{d}z}{\mathrm{d}t} = J\nabla_z H(z), \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

- Let $\Psi_t : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ denote the flow-map of the propagating $z(0)$ for time $t$

$$\Psi_t(z(0)) = z(t)$$

# PROPERTIES OF HAMILTONIAN MECHANICS



- Given $z(t) = (q(t), p(t))$ satisfying

$$\frac{\mathrm{d}z}{\mathrm{d}t} = J\nabla_z H(z), \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

- Let $\Psi_t : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ denote the flow-map of the propagating $z(0)$ for time $t$

$$\Psi_t(z(0)) = z(t)$$

- Let $N : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ denote the momentum flip

$$N(q, p) = (q, -p)$$

# PROPERTIES OF HAMILTONIAN MECHANICS



- Given $z(t) = (q(t), p(t))$ satisfying

$$\frac{\mathrm{d}z}{\mathrm{d}t} = J\nabla_z H(z), \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

- Let $\Psi_t : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ denote the flow-map of the propagating $z(0)$ for time $t$

$$\Psi_t(z(0)) = z(t)$$

- Let $N : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ denote the momentum flip

$$N(q, p) = (q, -p)$$

  - **Proposition:** The flow map is one-to-one and symplectic

$$\nabla\Psi_t^\top J^{-1} \nabla\Psi_t = J^{-1}$$

# PROPERTIES OF HAMILTONIAN MECHANICS



- Given $z(t) = (q(t), p(t))$ satisfying

$$\frac{\mathrm{d}z}{\mathrm{d}t} = J\nabla_z H(z), \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

- Let $\Psi_t : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ denote the flow-map of the propagating $z(0)$ for time $t$

$$\Psi_t(z(0)) = z(t)$$

- Let $N : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ denote the momentum flip

$$N(q, p) = (q, -p)$$

  - **Proposition:** The flow map is one-to-one and symplectic

$$\nabla\Psi_t^\top J^{-1} \nabla\Psi_t = J^{-1}$$

    - Consequentially it is volume preserving i.e. $\det(\nabla\Psi_t) = 1$

# PROPERTIES OF HAMILTONIAN MECHANICS



- Given $z(t) = (q(t), p(t))$ satisfying

$$\frac{\mathrm{d}z}{\mathrm{d}t} = J\nabla_z H(z), \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

- Let $\Psi_t : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ denote the flow-map of the propagating $z(0)$ for time $t$

$$\Psi_t(z(0)) = z(t)$$

- Let $N : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ denote the momentum flip

$$N(q, p) = (q, -p)$$

  - **Proposition:** The flow map is one-to-one and symplectic

$$\nabla\Psi_t^\top J^{-1} \nabla\Psi_t = J^{-1}$$

    - Consequentially it is volume preserving i.e. $\det(\nabla\Psi_t) = 1$

- **Proposition:** The time reversal solves Hamiltons equations and

$$z(t) = \Psi_t(z(0)), \quad \iff \quad z(0) = \Psi_t(N \circ z(t))$$

# PROPERTIES OF HAMILTONIAN MECHANICS

# PROPERTIES OF HAMILTONIAN MECHANICS



- **Proposition:** The target $\pi(z) \propto \exp(-H(z))$ is stationary with respect to the Hamiltonian flow.

$$z(0) \sim \pi, \quad \Longrightarrow \quad z(t) \sim \pi$$

# PROPERTIES OF HAMILTONIAN MECHANICS



- **Proposition:** The target $\pi(z) \propto \exp(-H(z))$ is stationary with respect to the Hamiltonian flow.

$$z(0) \sim \pi, \quad \Longrightarrow \quad z(t) \sim \pi$$

- **Proof:**

# PROPERTIES OF HAMILTONIAN MECHANICS



- **Proposition:** The target $\pi(z) \propto \exp(-H(z))$ is stationary with respect to the Hamiltonian flow.

$$z(0) \sim \pi, \quad \Longrightarrow \quad z(t) \sim \pi$$

- **Proof:**
  - Suppose $z_t \sim \pi_t$

# PROPERTIES OF HAMILTONIAN MECHANICS



- **Proposition:** The target $\pi(z) \propto \exp(-H(z))$ is stationary with respect to the Hamiltonian flow.

$$z(0) \sim \pi, \quad \Longrightarrow \quad z(t) \sim \pi$$

- **Proof:**
  - Suppose $z_t \sim \pi_t$
  - Since $z(t) = \Psi_t(z(0))$, and $\Psi_t$ is a one-to-one determinisitic map,

$$\pi_t(z(t)) = \pi(z(0)) \, |\det \nabla \Psi_t(z(0)) \, |$$

# PROPERTIES OF HAMILTONIAN MECHANICS



- **Proposition:** The target $\pi(z) \propto \exp(-H(z))$ is stationary with respect to the Hamiltonian flow.

$$z(0) \sim \pi, \quad \Longrightarrow \quad z(t) \sim \pi$$

- **Proof:**
  - Suppose $z_t \sim \pi_t$
  - Since $z(t) = \Psi_t(z(0))$, and $\Psi_t$ is a one-to-one determinisitic map,

$$\pi_t(z(t)) = \pi(z(0)) \, |\det \nabla \Psi_t(z(0)|$$

  - Since $\Psi_t$ is volume preserving the determinant is 1 and since the hamiltonian is preserved

$$\pi(z(0)) \propto \exp(H(z(0)) = \exp(H(z(t)) \propto \pi(z(t))$$

# PROPERTIES OF HAMILTONIAN MECHANICS



- **Proposition:** The target $\pi(z) \propto \exp(-H(z))$ is stationary with respect to the Hamiltonian flow.

$$z(0) \sim \pi, \quad \Longrightarrow \quad z(t) \sim \pi$$

- **Proof:**
  - Suppose $z_t \sim \pi_t$
  - Since $z(t) = \Psi_t(z(0))$, and $\Psi_t$ is a one-to-one determinisitic map,

$$\pi_t(z(t)) = \pi(z(0)) \, |\det \nabla \Psi_t(z(0)|$$

  - Since $\Psi_t$ is volume preserving the determinant is 1 and since the hamiltonian is preserved

$$\pi(z(0)) \propto \exp(H(z(0)) = \exp(H(z(t)) \propto \pi(z(t))$$

  - Hence for all $t$ we have

$$\pi_t(z(t)) = \pi(z(t))$$

# IRRIDUCIBILITY

# IRRIDUCIBILITY



- Defined the kernel $K_t(z, \mathrm{d}z')$ propogating $z$ according to Hamilton's equations for time $t$

$$K_t(z, \mathrm{d}z') = \delta_{\Psi_t(z)}(\mathrm{d}z')$$

# IRRIDUCIBILITY



- Defined the kernel $K_t(z, \mathrm{d}z')$ propogating $z$ according to Hamilton's equations for time $t$

$$K_t(z, \mathrm{d}z') = \delta_{\Psi_t(z)}(\mathrm{d}z')$$

  - Hamilton's equations themselves do not define an ergodic Markov chain, since they preserve the Hamiltonian.

# IRRIDUCIBILITY



- Defined the kernel $K_t(z, \mathrm{d}z')$ propogating $z$ according to Hamilton's equations for time $t$

$$K_t(z, \mathrm{d}z') = \delta_{\Psi_t(z)}(\mathrm{d}z')$$

  - Hamilton's equations themselves do not define an ergodic Markov chain, since they preserve the Hamiltonian.

- Let $K_R(z, \mathrm{d}z')$ be the Markov kernel that resample the momentum according to the Gaussian distribution with covariance matrix $M$

$$K_R(z, \mathrm{d}z') \propto \delta_q(\mathrm{d}q')\exp(-K(p'))\mathrm{d}p'$$

$$= \delta_q(\mathrm{d}q')\mathcal{N}(\mathrm{d}p'; 0, M)$$

# IRRIDUCIBILITY



- Defined the kernel $K_t(z, \mathrm{d}z')$ propogating $z$ according to Hamilton's equations for time $t$

$$K_t(z, \mathrm{d}z') = \delta_{\Psi_t(z)}(\mathrm{d}z')$$

  - Hamilton's equations themselves do not define an ergodic Markov chain, since they preserve the Hamiltonian.

- Let $K_R(z, \mathrm{d}z')$ be the Markov kernel that resample the momentum according to the Gaussian distribution with covariance matrix $M$

$$K_R(z, \mathrm{d}z') \propto \delta_q(\mathrm{d}q')\exp(-K(p'))\mathrm{d}p'$$

$$= \delta_q(\mathrm{d}q')\mathcal{N}(\mathrm{d}p'; 0, M)$$

- The combination $K_R K_t$ defines an irriducible and ergodic Markov kernel!
  - Hamiltonian dynamics propogates withina level set
  - Momentum refresh propogates between level sets

# HAMILTONIAN MONTE CARLO

# HAMILTONIAN MONTE CARLO



- **Theorem:** Let $\pi(q) \propto \exp(-U(q))$ and Let $T \sim \nu_T$ with positive density on an interval $[0,\tau]$ for $\tau > 0$. Let $K(q, \mathrm{d}q)$ be the Markov kernel for the position variables corresponding to:

    1. Sampling a momentum $p \sim \mathcal{N}(0,I)$ and $T \sim \nu_T$
    2. Running Hamiltonian dynamics initialised at $z(0) = (q, p)$ for time $T$
    3. Flip the momentum the momentum variable

# HAMILTONIAN MONTE CARLO



- **Theorem:** Let $\pi(q) \propto \exp(-U(q))$ and Let $T \sim \nu_T$ with positive density on an interval $[0,\tau]$ for $\tau > 0$. Let $K(q, \mathrm{d}q)$ be the Markov kernel for the position variables corresponding to:

    1. Sampling a momentum $p \sim \mathcal{N}(0,I)$ and $T \sim \nu_T$
    2. Running Hamiltonian dynamics initialised at $z(0) = (q, p)$ for time $T$
    3. Flip the momentum the momentum variable

- Suppose that $U$ is continuously differentiable on $\mathbb{R}^d$, and satisfies that

$$\inf_{q \in \mathbb{R}^d} U(q) > -\infty, \qquad \sup_q \|\nabla^2 U(q)\| \leq L$$

# HAMILTONIAN MONTE CARLO



- **Theorem:** Let $\pi(q) \propto \exp(-U(q))$ and Let $T \sim \nu_T$ with positive density on an interval $[0,\tau]$ for $\tau > 0$. Let $K(q, \mathrm{d}q)$ be the Markov kernel for the position variables corresponding to:

    1. Sampling a momentum $p \sim \mathcal{N}(0,I)$ and $T \sim \nu_T$
    2. Running Hamiltonian dynamics initialised at $z(0) = (q, p)$ for time $T$
    3. Flip the momentum the momentum variable

- Suppose that $U$ is continuously differentiable on $\mathbb{R}^d$, and satisfies that

$$\inf_{q \in \mathbb{R}^d} U(q) > -\infty, \qquad \sup_q \|\nabla^2 U(q)\| \leq L$$

- Then $K$ is strongly $\pi$-irreducible and $\pi$-reversible

# DISCRETIZATION

# DISCRETIZATION



- In general we cannot simulate from Hamilton's equations directly and have to discretize

$$\dot{q}(t) = M^{-1}p(t), \qquad \dot{p}(t) = -\nabla U(q(t))$$

# DISCRETIZATION



- In general we cannot simulate from Hamilton's equations directly and have to discretize

$$\dot{q}(t) = M^{-1}p(t), \qquad \dot{p}(t) = -\nabla U(q(t))$$

- **Euler scheme:**

$$p(t+\epsilon) = p(t) - \epsilon \nabla U(q(t))$$
$$q(t+\epsilon) = q(t) + \epsilon M^{-1}p(t)$$

# DISCRETIZATION



- In general we cannot simulate from Hamilton's equations directly and have to discretize

$$\dot{q}(t) = M^{-1}p(t), \qquad \dot{p}(t) = -\nabla U(q(t))$$

- **Euler scheme:**

$$p(t+\epsilon) = p(t) - \epsilon \nabla U(q(t))$$
$$q(t+\epsilon) = q(t) + \epsilon M^{-1}p(t)$$

- **Modified Euler scheme:**

$$p(t+\epsilon) = p(t) - \epsilon \nabla U(q(t))$$
$$q(t+\epsilon) = q(t) + \epsilon M^{-1}p(t+\epsilon)$$

# DISCRETIZATION



- In general we cannot simulate from Hamilton's equations directly and have to discretize

$$\dot{q}(t) = M^{-1}p(t), \qquad \dot{p}(t) = -\nabla U(q(t))$$

- **Euler scheme:**

$$p(t+\epsilon) = p(t) - \epsilon \nabla U(q(t))$$
$$q(t+\epsilon) = q(t) + \epsilon M^{-1}p(t)$$

- **Modified Euler scheme:**

$$p(t+\epsilon) = p(t) - \epsilon \nabla U(q(t))$$
$$q(t+\epsilon) = q(t) + \epsilon M^{-1}p(t+\epsilon)$$

- **Leapfrog scheme:**

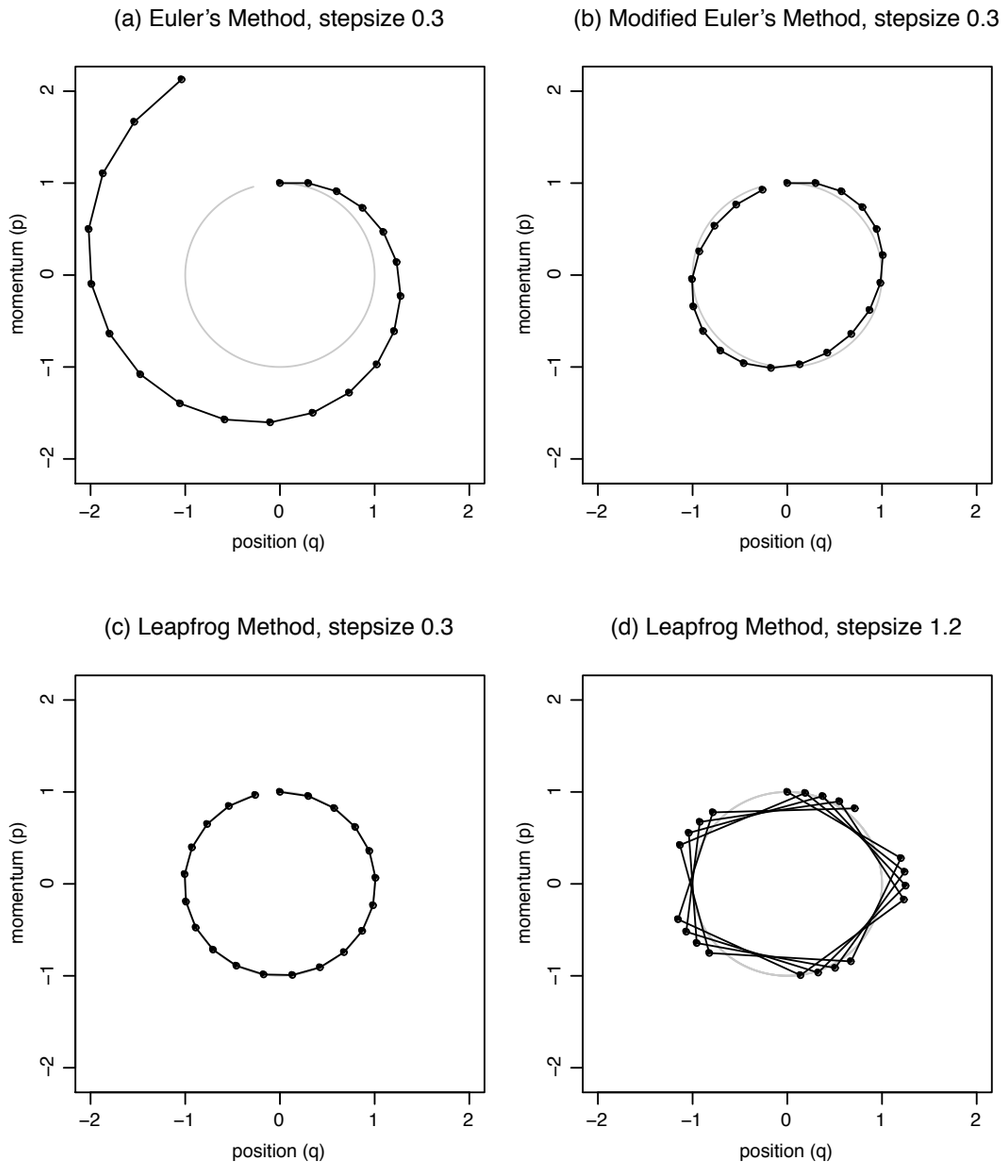
$$p(t+\epsilon/2) = p(t) - \frac{\epsilon}{2}\nabla U(q(t))$$
$$q(t+\epsilon) = q(t) + \epsilon M^{-1}p(t+\epsilon/2)$$
$$p(t+\epsilon) = p(t+\epsilon/2) - \frac{\epsilon}{2}\nabla U(q(t+\epsilon))$$

# DISCRETISATION



- Consider the simple harmonic oscillator $H(q,p) = \frac{1}{2}p^2 + \frac{1}{2}q^2$

- Approximation of Hamiltonian dynamics by 3 schemes. 20 steps in each case in black, the true trajectory in grey



# HAMILTONIAN MONTE CARLO

# HAMILTONIAN MONTE CARLO



- In practice we use the discretized dynamics with a step size $\epsilon$ for $L$ leap steps
  - Appromiates Hamiltonian dynamics for $T = \epsilon L$ time

# HAMILTONIAN MONTE CARLO



- In practice we use the discretized dynamics with a step size $\epsilon$ for $L$ leap steps
  - Appromiates Hamiltonian dynamics for $T = \epsilon L$ time

- This defines a proposal which we correct using Metropolis Hastings

# HAMILTONIAN MONTE CARLO



- In practice we use the discretized dynamics with a step size $\epsilon$ for $L$ leap steps
  - Appromiates Hamiltonian dynamics for $T = \epsilon L$ time

- This defines a proposal which we correct using Metropolis Hastings

- Given state $X_{t-1}$ defines the

# HAMILTONIAN MONTE CARLO



- In practice we use the discretized dynamics with a step size $\epsilon$ for $L$ leap steps
  - Appromiates Hamiltonian dynamics for $T = \epsilon L$ time

- This defines a proposal which we correct using Metropolis Hastings

- Given state $X_{t-1}$ defines the
  - Let $q = X_{t-1}$ and resample the $p \sim \mathcal{N}(0, I)$

# HAMILTONIAN MONTE CARLO



- In practice we use the discretized dynamics with a step size $\epsilon$ for $L$ leap steps
  - Appromiates Hamiltonian dynamics for $T = \epsilon L$ time

- This defines a proposal which we correct using Metropolis Hastings

- Given state $X_{t-1}$ defines the
  - Let $q = X_{t-1}$ and resample the $p \sim \mathcal{N}(0, I)$
  - Generate $(q', p')$ by running $L$ leapfrog steps with a step size $\epsilon$ initialised at $(q, p)$

# HAMILTONIAN MONTE CARLO



- In practice we use the discretized dynamics with a step size $\epsilon$ for $L$ leap steps
  - Appromiates Hamiltonian dynamics for $T = \epsilon L$ time

- This defines a proposal which we correct using Metropolis Hastings

- Given state $X_{t-1}$ defines the
  - Let $q = X_{t-1}$ and resample the $p \sim \mathcal{N}(0, I)$
  - Generate $(q', p')$ by running $L$ leapfrog steps with a step size $\epsilon$ initialised at $(q, p)$
  - Set $X_t = q'$ with probability $\alpha$

$$\alpha = 1 \wedge \exp(H(q, p) - H(q', p'))$$

# HAMILTONIAN MONTE CARLO



- In practice we use the discretized dynamics with a step size $\epsilon$ for $L$ leap steps
  - Appromiates Hamiltonian dynamics for $T = \epsilon L$ time

- This defines a proposal which we correct using Metropolis Hastings

- Given state $X_{t-1}$ defines the
  - Let $q = X_{t-1}$ and resample the $p \sim \mathcal{N}(0, I)$
  - Generate $(q', p')$ by running $L$ leapfrog steps with a step size $\epsilon$ initialised at $(q, p)$
  - Set $X_t = q'$ with probability $\alpha$

$$\alpha = 1 \wedge \exp(H(q, p) - H(q', p'))$$

- In general we often pick a random length or step size to avoid any periodicity

# HAMILTONIAN MONTE CARLO



- In practice we use the discretized dynamics with a step size $\epsilon$ for $L$ leap steps
  - Appromiates Hamiltonian dynamics for $T = \epsilon L$ time

- This defines a proposal which we correct using Metropolis Hastings

- Given state $X_{t-1}$ defines the
  - Let $q = X_{t-1}$ and resample the $p \sim \mathcal{N}(0, I)$
  - Generate $(q', p')$ by running $L$ leapfrog steps with a step size $\epsilon$ initialised at $(q, p)$
  - Set $X_t = q'$ with probability $\alpha$

$$\alpha = 1 \wedge \exp(H(q,p) - H(q',p'))$$

- In general we often pick a random length or step size to avoid any periodicity

- Demo:

  https://www.saifsyed.com/sampling-demo/app.html?algorithm=HMC

# HISTORY

# HISTORY



- Was first proposed in “Hybrid Monte Carlo” published in  Physical Review Letters B in 1987

# HISTORY



- Was first proposed in “Hybrid Monte Carlo” published in  Physical Review Letters B in 1987
  - Authors: Simon Duane, Anthony Kennedy, Brian Pendleton and Duncan Roweth

# HISTORY



- Was first proposed in “Hybrid Monte Carlo” published in  Physical Review Letters B in 1987
  - Authors: Simon Duane, Anthony Kennedy, Brian Pendleton and Duncan Roweth
  - Originates from the lattice feild theory community

# HISTORY



- Was first proposed in “Hybrid Monte Carlo” published in  Physical Review Letters B in 1987
  - Authors: Simon Duane, Anthony Kennedy, Brian Pendleton and Duncan Roweth
  - Originates from the lattice feild theory community

- The intuition is derived from physics, but the ideas extend beyond

# HISTORY



- Was first proposed in “Hybrid Monte Carlo” published in  Physical Review Letters B in 1987
  - Authors: Simon Duane, Anthony Kennedy, Brian Pendleton and Duncan Roweth
  - Originates from the lattice feild theory community

- The intuition is derived from physics, but the ideas extend beyond
  - Hamiltonians and energy are social constructs.

# HISTORY



- Was first proposed in “Hybrid Monte Carlo” published in  Physical Review Letters B in 1987
  - Authors: Simon Duane, Anthony Kennedy, Brian Pendleton and Duncan Roweth
  - Originates from the lattice feild theory community

- The intuition is derived from physics, but the ideas extend beyond
  - Hamiltonians and energy are social constructs.

- In 1996  Radford showed how the method could be used for a broader class of statistical problems, in particular Bayesian neural networks

# HISTORY



- Was first proposed in “Hybrid Monte Carlo” published in  Physical Review Letters B in 1987
  - Authors: Simon Duane, Anthony Kennedy, Brian Pendleton and Duncan Roweth
  - Originates from the lattice feild theory community

- The intuition is derived from physics, but the ideas extend beyond
  - Hamiltonians and energy are social constructs.

- In 1996  Radford showed how the method could be used for a broader class of statistical problems, in particular Bayesian neural networks

- Introduced to the statistics community as HMC in a review paper  “MCMC using Hamiltonian dynamics” in 2011 by Radford Neal

# HISTORY



- Was first proposed in “Hybrid Monte Carlo” published in  Physical Review Letters B in 1987
  - Authors: Simon Duane, Anthony Kennedy, Brian Pendleton and Duncan Roweth
  - Originates from the lattice feild theory community

- The intuition is derived from physics, but the ideas extend beyond
  - Hamiltonians and energy are social constructs.

- In 1996  Radford showed how the method could be used for a broader class of statistical problems, in particular Bayesian neural networks

- Introduced to the statistics community as HMC in a review paper  “MCMC using Hamiltonian dynamics” in 2011 by Radford Neal

- Popularised within the Bayesian statistics community due to popularity of Stan in 2017

# HISTORY



- Was first proposed in “Hybrid Monte Carlo” published in  Physical Review Letters B in 1987
  - Authors: Simon Duane, Anthony Kennedy, Brian Pendleton and Duncan Roweth
  - Originates from the lattice feild theory community

- The intuition is derived from physics, but the ideas extend beyond
  - Hamiltonians and energy are social constructs.

- In 1996  Radford showed how the method could be used for a broader class of statistical problems, in particular Bayesian neural networks

- Introduced to the statistics community as HMC in a review paper  “MCMC using Hamiltonian dynamics” in 2011 by Radford Neal

- Popularised within the Bayesian statistics community due to popularity of Stan in 2017

- Huge literature studying HMC and its variants

# HISTORY



- Was first proposed in “Hybrid Monte Carlo” published in  Physical Review Letters B in 1987
  - Authors: Simon Duane, Anthony Kennedy, Brian Pendleton and Duncan Roweth
  - Originates from the lattice feild theory community

- The intuition is derived from physics, but the ideas extend beyond
  - Hamiltonians and energy are social constructs.

- In 1996  Radford showed how the method could be used for a broader class of statistical problems, in particular Bayesian neural networks

- Introduced to the statistics community as HMC in a review paper  “MCMC using Hamiltonian dynamics” in 2011 by Radford Neal

- Popularised within the Bayesian statistics community due to popularity of Stan in 2017

- Huge literature studying HMC and its variants

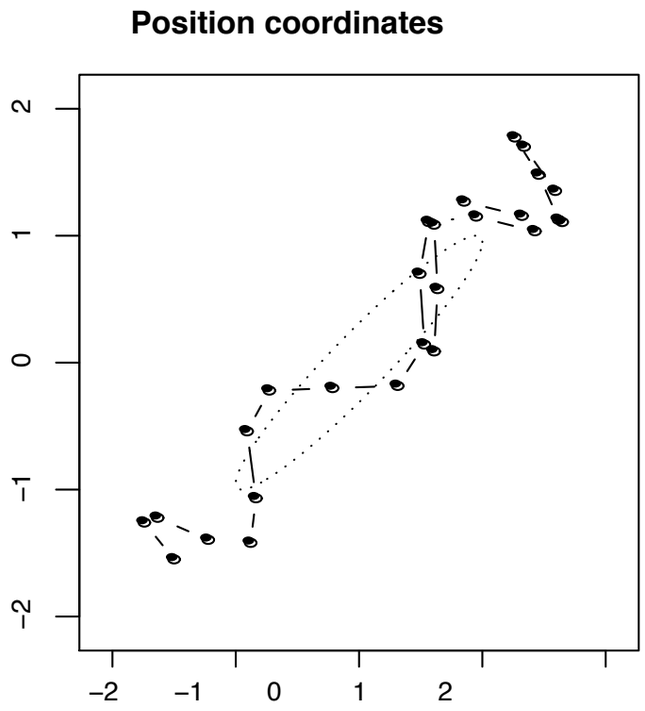
- HMC is a very good default sampler understanding the shape of a mode

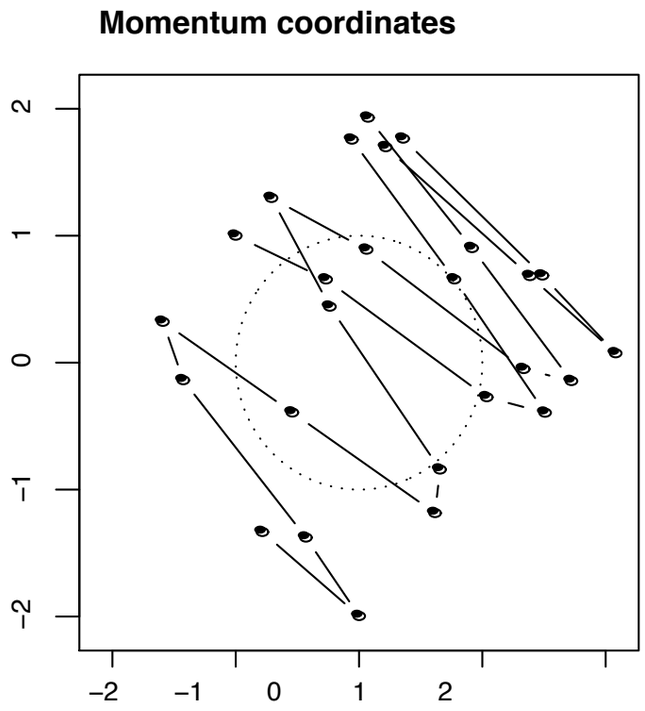
# EXAMPLE

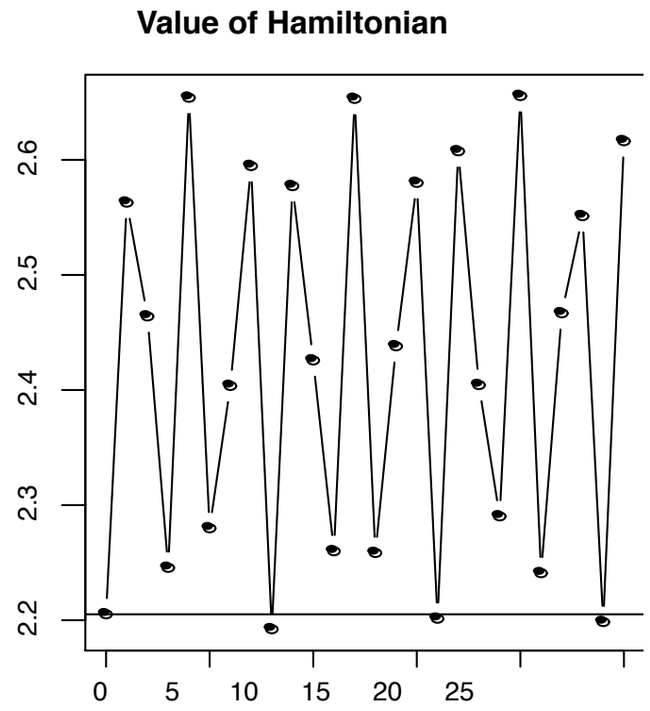


- Consider a 2D Gaussian distribution with covariance matrix $\Sigma = \begin{pmatrix} 1 & 0.98 \\ 0.98 & 1 \end{pmatrix}$

- We let $M = I$, and do $L = 25$ leapfrog steps per iteration using stepsize $\epsilon = 0.25$.



# EXAMPLE

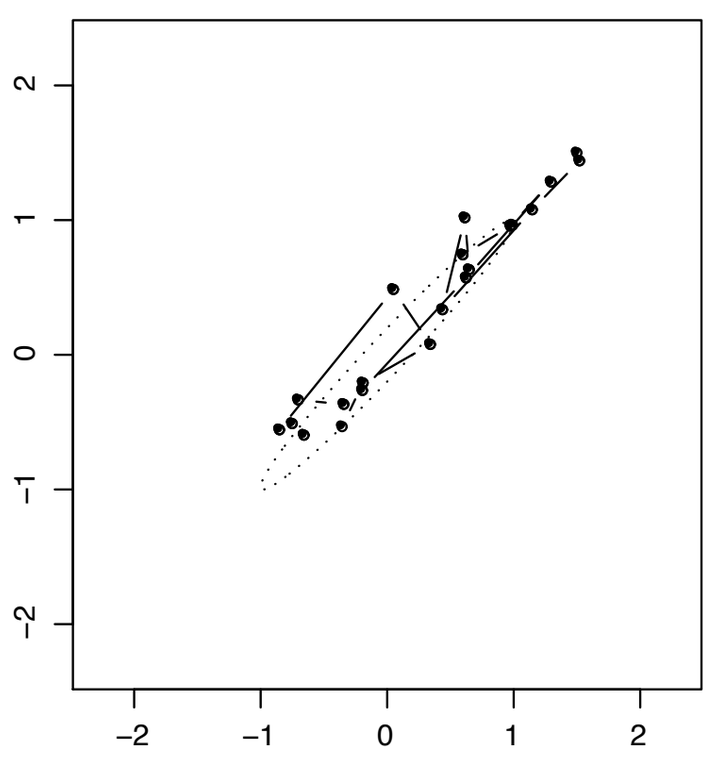


- Consider a 2D Gaussian distribution with covariance matrix $\Sigma = \begin{pmatrix} 1 & 0.98 \\ 0.98 & 1 \end{pmatrix}$

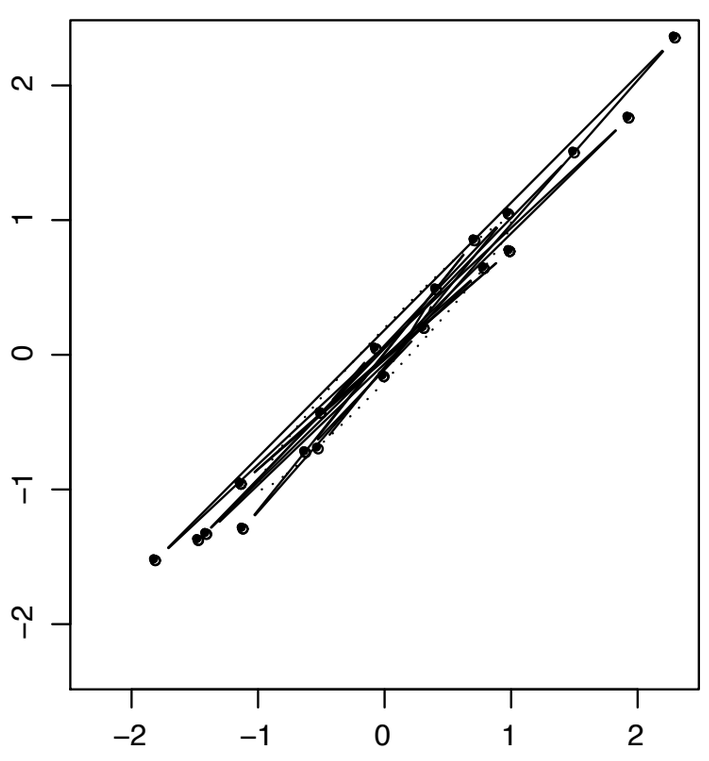
- We let $M = I$, and do $L = 25$ leapfrog steps per iteration using stepsize $\epsilon = 0.25$.

- Twenty iterations of the random walk Metropolis method (20 updates per iteration). HMC is making much larger moves and mixes faster than random walk Metropolis.

**Random–walk Metropolis**



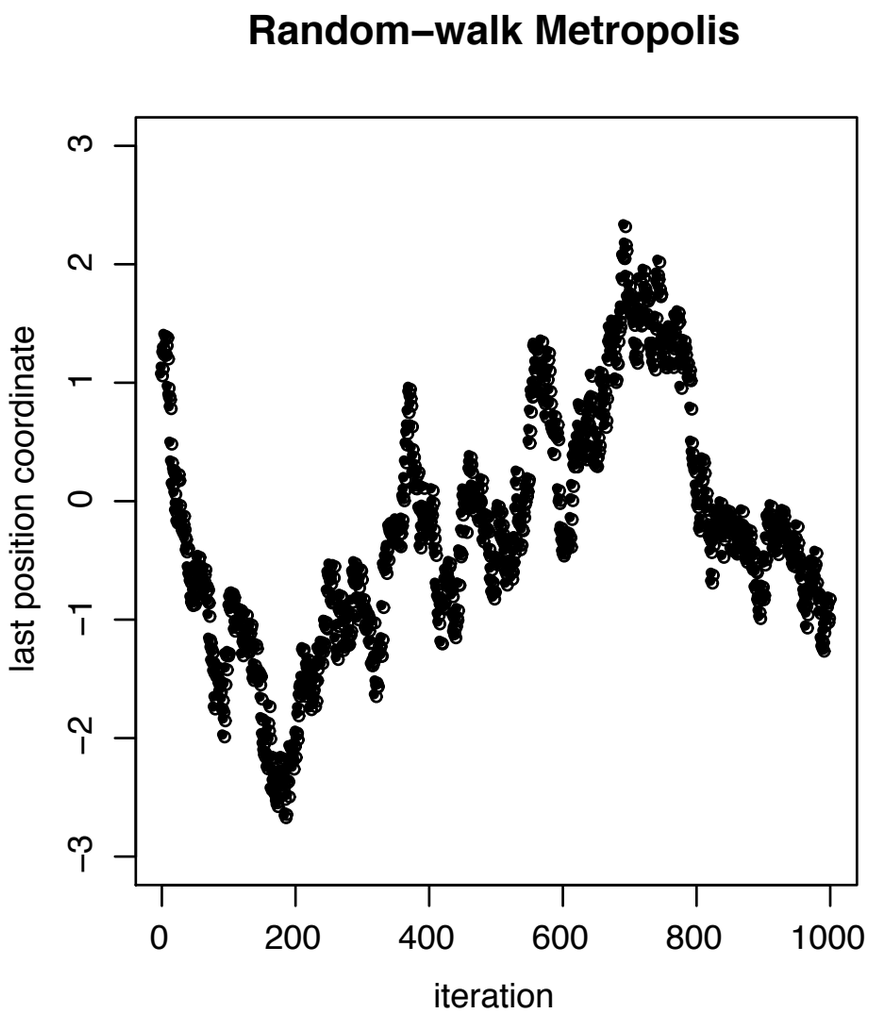

**Hamiltonian Monte Carlo**

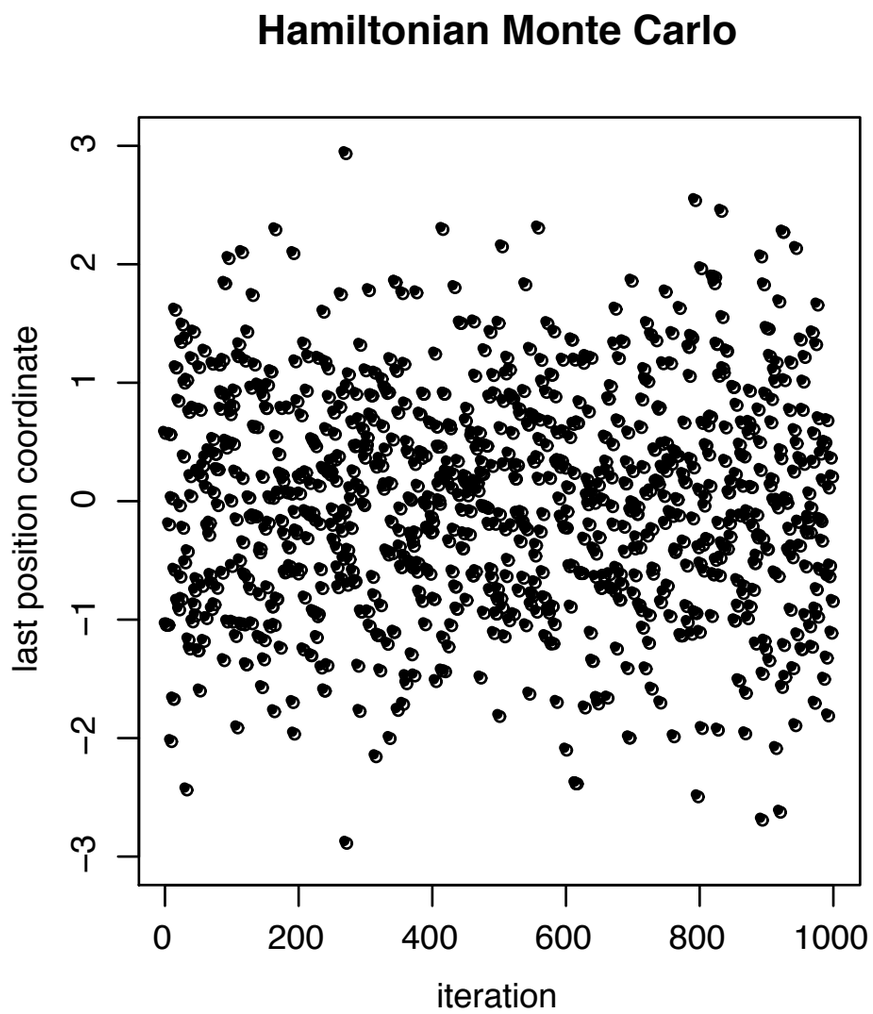


# EXAMPLE



- 100 dimensional multivariate Gaussian distribution with independent components of standard deviations $0.01,\ldots,1.00$, with $\epsilon = 0.1$ and $L = 150$ leapfrog steps. The figure shows the last component.



# CONCENTRATION OF MEASURE AND HMC

# CONCENTRATION OF MEASURE AND HMC



- HMC performs very well in high-dimensions

# CONCENTRATION OF MEASURE AND HMC

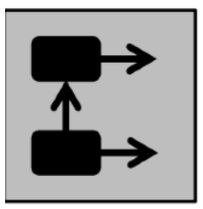


- HMC performs very well in high-dimensions
- In high dimensions, target distributions tend to be **non-identifiable**
  - i.e the target concentrates on a submanifold

# CONCENTRATION OF MEASURE AND HMC



- HMC performs very well in high-dimensions
- In high dimensions, target distributions tend to be **non-identifiable**
  - i.e the target concentrates on a submanifold
- **Example:** an ODE parameter inference problem from Leonhardt et al (2014)

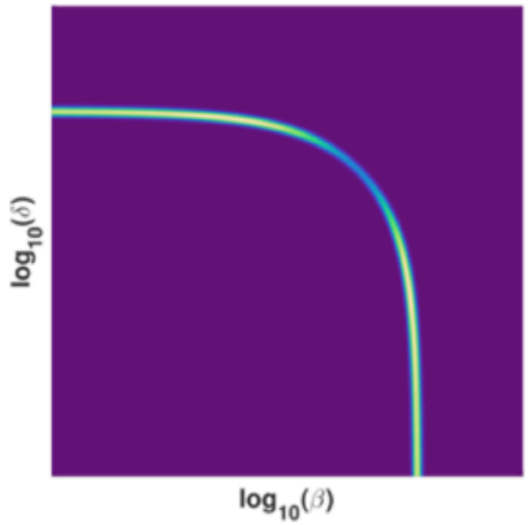


## (M1) mRNA Transfection

- Equally weighted modes; Parameters $\delta$ and $\beta$ exchangable
- Analytical solution available
- Only one of two states observable: $y_1(t) = G(t)$
- $\forall\, t < t_0 : m(t) = G(t) = 0$
- Two data sets: Experimental (a) and Artificial (b)

$$\dot{G} = k_{TL} m - \beta G, \qquad G(t_0) = 0$$

$$\dot{m} = -\delta m, \qquad m(t_0) = m_0$$



# CONCENTRATION OF MEASURE AND HMC

# CONCENTRATION OF MEASURE AND HMC



- In general, if the Hessian of the target potential satisfies that $\lambda I \preceq \nabla^2 U(x) \preceq LI$ for some $$ $0 < \lambda < L < \infty$, (strongly convex and smooth potential), and $H_{\min} := \inf_z H(z)$, then it is possible to show that

# CONCENTRATION OF MEASURE AND HMC



- In general, if the Hessian of the target potential satisfies that $\lambda I \preceq \nabla^2 U(x) \preceq LI$ for some $0 < \lambda < L < \infty$, (strongly convex and smooth potential), and $H_{\min} := \inf_z H(z)$, then it is possible to show that

$$\mathbb{P}\left( \left| \sqrt{H(z) - H_{\min}} - \mathbb{E}\sqrt{H(z) - H_{\min}} \right| \geq t \right) \leq C \exp\left( -\frac{t^2}{C} \right)$$

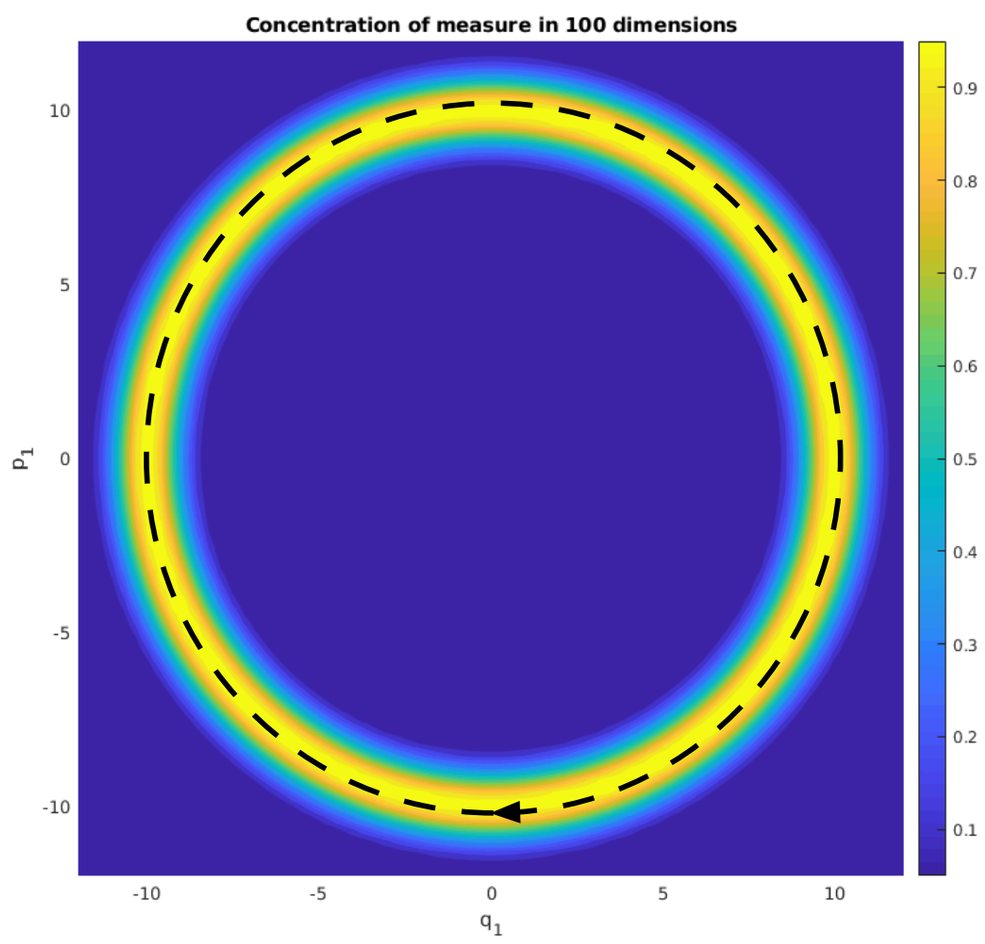
# CONCENTRATION OF MEASURE AND HMC



- In general, if the Hessian of the target potential satisfies that $\lambda I \preceq \nabla^2 U(x) \preceq LI$ for some $0 < \lambda < L < \infty$, (strongly convex and smooth potential), and $H_{\min} := \inf_z H(z)$, then it is possible to show that

$$\mathbb{P}\left( \left| \sqrt{H(z) - H_{\min}} - \mathbb{E}\sqrt{H(z) - H_{\min}} \right| \geq t \right) \leq C \exp\left( -\frac{t^2}{C} \right)$$

- Hamiltonian is close to constant high probability density area



# CONCENTRATION OF MEASURE AND HMC

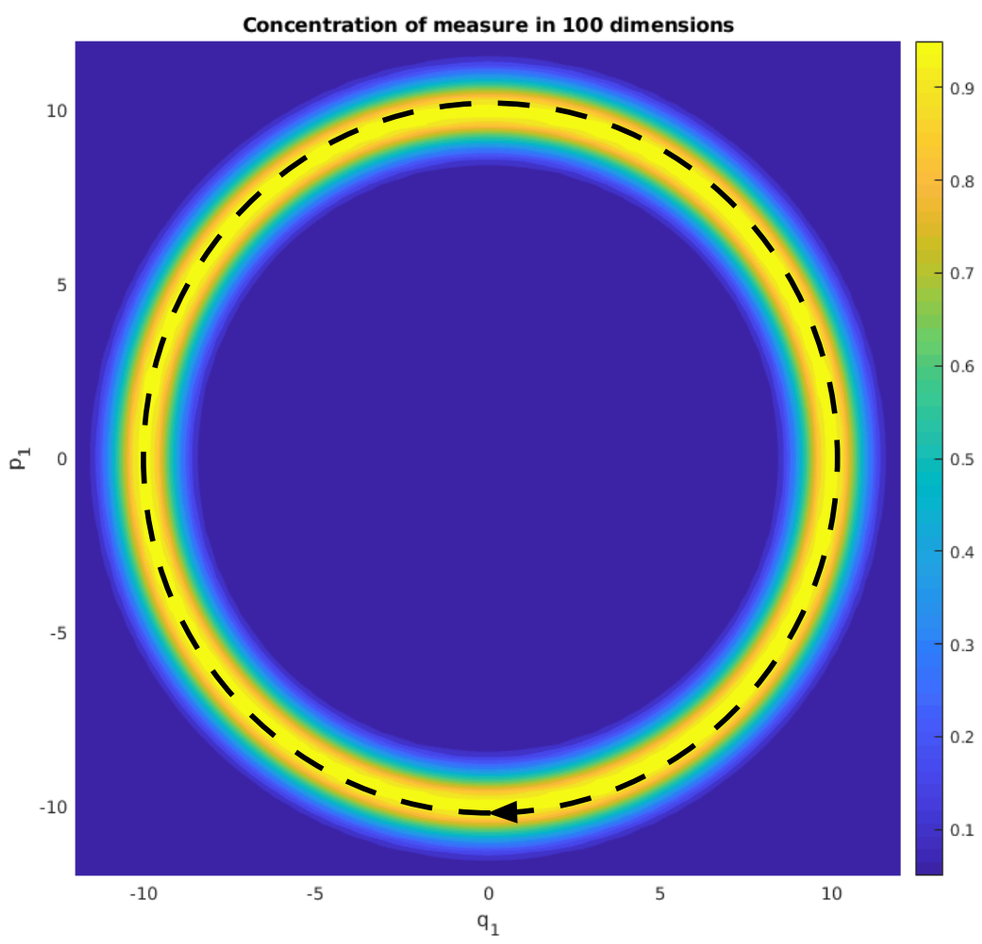


- In general, if the Hessian of the target potential satisfies that $\lambda I \preceq \nabla^2 U(x) \preceq LI$ for some $0 < \lambda < L < \infty$, (strongly convex and smooth potential), and $H_{\min} := \inf_z H(z)$, then it is possible to show that

$$\mathbb{P}\left( \left| \sqrt{H(z) - H_{\min}} - \mathbb{E}\sqrt{H(z) - H_{\min}} \right| \geq t \right) \leq C \exp\left( -\frac{t^2}{C} \right)$$

- Hamiltonian is close to constant high probability density area

- HMC is very efficient in exploring this potentially complicated modes automatically.



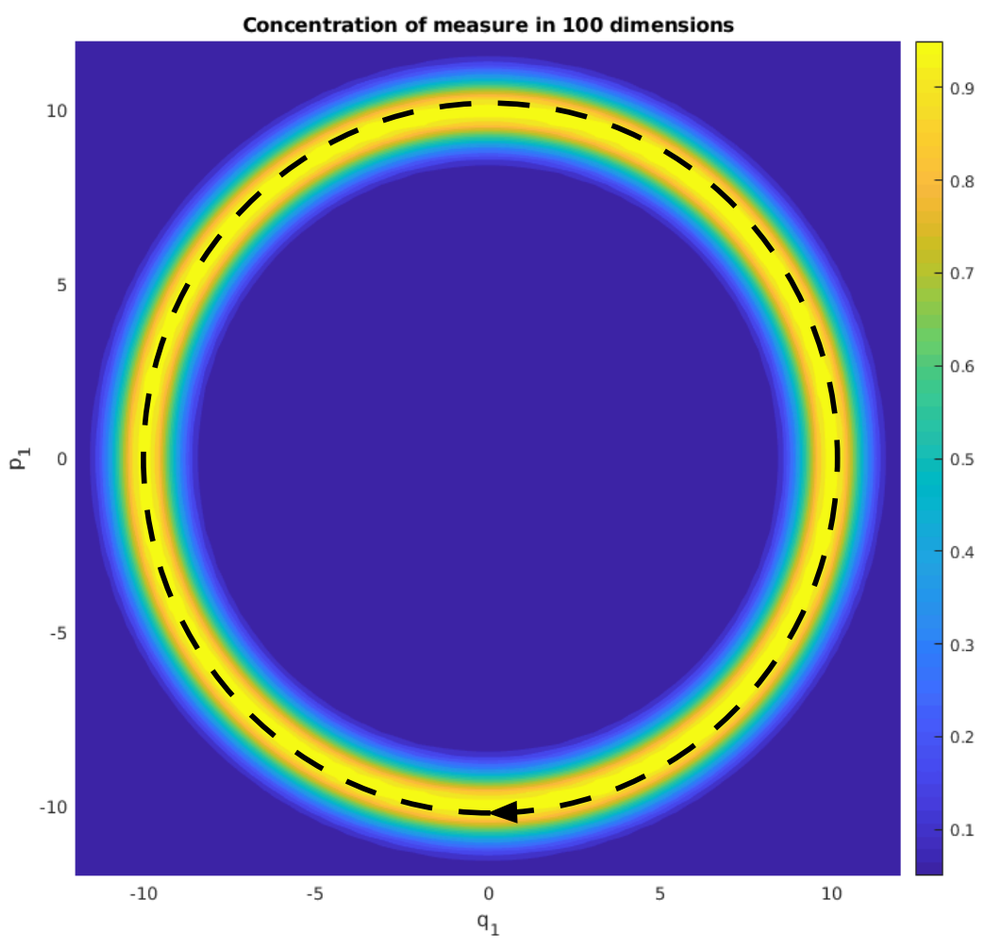
# CONCENTRATION OF MEASURE AND HMC



- In general, if the Hessian of the target potential satisfies that $\lambda I \preceq \nabla^2 U(x) \preceq LI$ for some $0 < \lambda < L < \infty$, (strongly convex and smooth potential), and $H_{\min} := \inf_z H(z)$, then it is possible to show that

$$\mathbb{P}\left( \left| \sqrt{H(z) - H_{\min}} - \mathbb{E}\sqrt{H(z) - H_{\min}} \right| \geq t \right) \leq C \exp\left( -\frac{t^2}{C} \right)$$

- Hamiltonian is close to constant high probability density area
- HMC is very efficient in exploring this potentially complicated modes automatically.
- Does not apply to multi-modal distributions



# SUMMARY OF LOCAL SAMPLERS

# SUMMARY OF LOCAL SAMPLERS



- **Example**: RWM $q' \sim \mathcal{N}(q, \epsilon\Sigma)$

# SUMMARY OF LOCAL SAMPLERS



- **Example**: RWM

$$q' \sim \mathcal{N}(q, \epsilon\Sigma)$$

- **Example:** MALA

$$q' \sim \mathcal{N}\left(q - \frac{\epsilon^2}{2}\Sigma\nabla U(q), \epsilon^2\Sigma\right)$$

# SUMMARY OF LOCAL SAMPLERS



- **Example**: RWM

$$q' \sim \mathcal{N}(q, \epsilon\Sigma)$$

- **Example:** MALA

$$q' \sim \mathcal{N}\left(q - \frac{\epsilon^2}{2}\Sigma\nabla U(q), \epsilon^2\Sigma\right)$$

- **HMC:**

$$q' = q_L$$

  - Where $q_0 = q$ and $p_0 \sim \mathcal{N}(0,I)$ and for $i = 1,\ldots,L$ do leap frog steps:

$$p_{i+1/2} = p_i - \frac{\epsilon}{2}\nabla \log U(q)$$

$$q_{i+1} = q_i + \epsilon M^{-1} p_{i+1/2}$$

$$p_{i+1} = p_{i+1/2} - \frac{\epsilon}{2}\nabla \log U(q_{i+1})$$

# SUMMARY OF LOCAL SAMPLERS



- **Example**: RWM

$$q' \sim \mathcal{N}(q, \epsilon\Sigma)$$

- **Example:** MALA

$$q' \sim \mathcal{N}\left(q - \frac{\epsilon^2}{2}\Sigma\nabla U(q), \epsilon^2\Sigma\right)$$

- **HMC:**

$$q' = q_L$$

  - Where $q_0 = q$ and $p_0 \sim \mathcal{N}(0,I)$ and for $i = 1,\ldots,L$ do leap frog steps:

$$p_{i+1/2} = p_i - \frac{\epsilon}{2}\nabla \log U(q)$$

$$q_{i+1} = q_i + \epsilon M^{-1} p_{i+1/2}$$

$$p_{i+1} = p_{i+1/2} - \frac{\epsilon}{2}\nabla \log U(q_{i+1})$$

- When $L = 1$ an HMC is equivalent to corresponds to iteration of MALA with $M = \Sigma^{-1}$

$$q' \sim \mathcal{N}\left(q - \frac{\epsilon^2}{2}M^{-1}\nabla \log U(q), \epsilon^2 M^{-1}\right)$$

# TUNING STEP SIZE

# TUNING STEP SIZE



- Step size $\epsilon$ is chosen to account for the scale of the target

# TUNING STEP SIZE



- Step size $\epsilon$ is chosen to account for the scale of the target
  - When distributionis wide, want to pick a big step size

# TUNING STEP SIZE



- Step size $\epsilon$ is chosen to account for the scale of the target
  - When distributionis wide, want to pick a big step size
  - When the distribution is narrow want to pick a small step size

# TUNING STEP SIZE



- Step size $\epsilon$ is chosen to account for the scale of the target
  - When distributionis wide, want to pick a big step size
  - When the distribution is narrow want to pick a small step size
- Small $\epsilon$ leads to high acceptance rates but small movement

# TUNING STEP SIZE



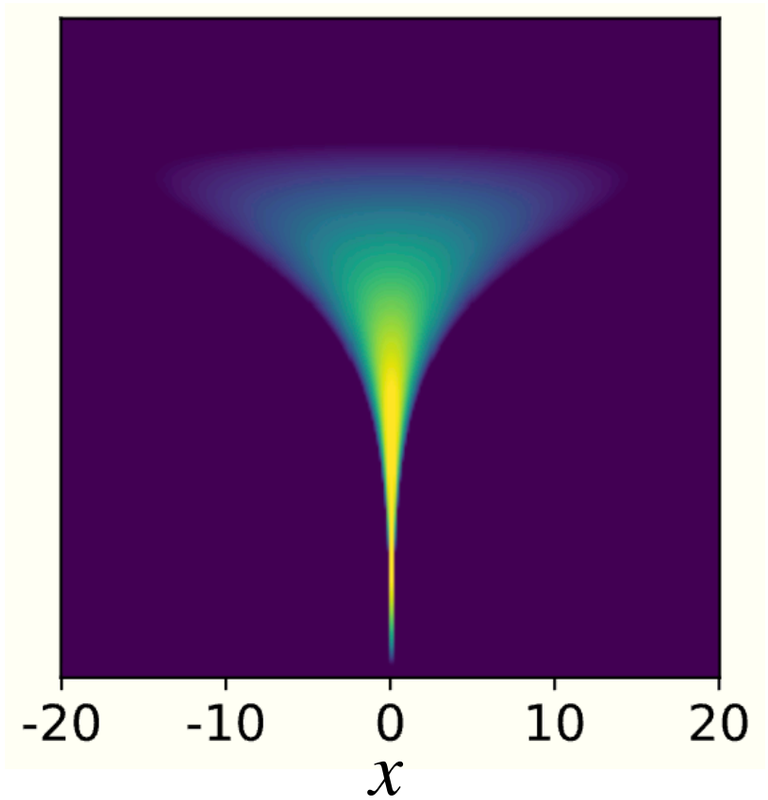
- Step size $\epsilon$ is chosen to account for the scale of the target
  - When distributionis wide, want to pick a big step size
  - When the distribution is narrow want to pick a small step size
- Small $\epsilon$ leads to high acceptance rates but small movement
- Large $\epsilon$ leads to low acceptance rates and propose samples in unlikely regions

# TUNING STEP SIZE



- Step size $\epsilon$ is chosen to account for the scale of the target
  - When distributionis wide, want to pick a big step size
  - When the distribution is narrow want to pick a small step size
- Small $\epsilon$ leads to high acceptance rates but small movement
- Large $\epsilon$ leads to low acceptance rates and propose samples in unlikely regions
- **Example:** Neil's funnel

$$Y \sim \mathcal{N}(0,3), \qquad X \sim \mathcal{N}(0,e^Y)$$

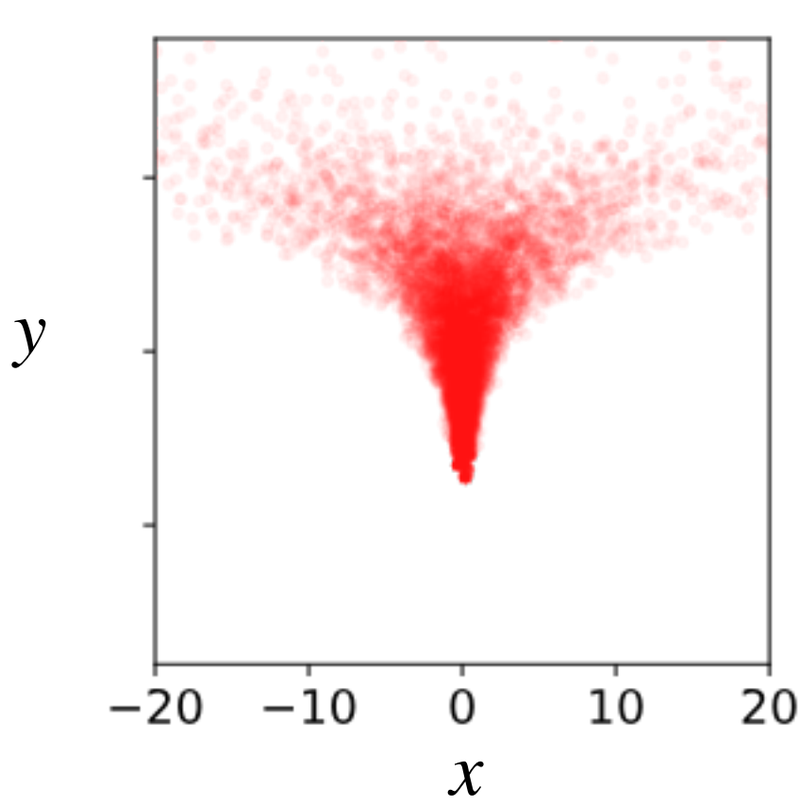


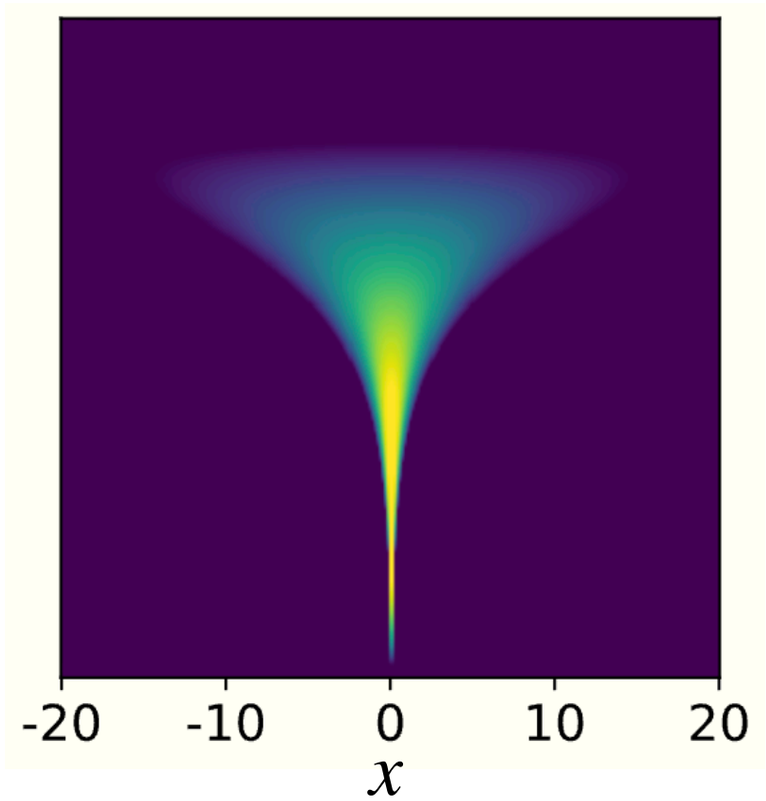
# TUNING STEP SIZE



- Step size $\epsilon$ is chosen to account for the scale of the target
  - When distributionis wide, want to pick a big step size
  - When the distribution is narrow want to pick a small step size
- Small $\epsilon$ leads to high acceptance rates but small movement
- Large $\epsilon$ leads to low acceptance rates and propose samples in unlikely regions
- **Example:** Neil's funnel

$$Y \sim \mathcal{N}(0,3), \qquad X \sim \mathcal{N}(0,e^{Y})$$



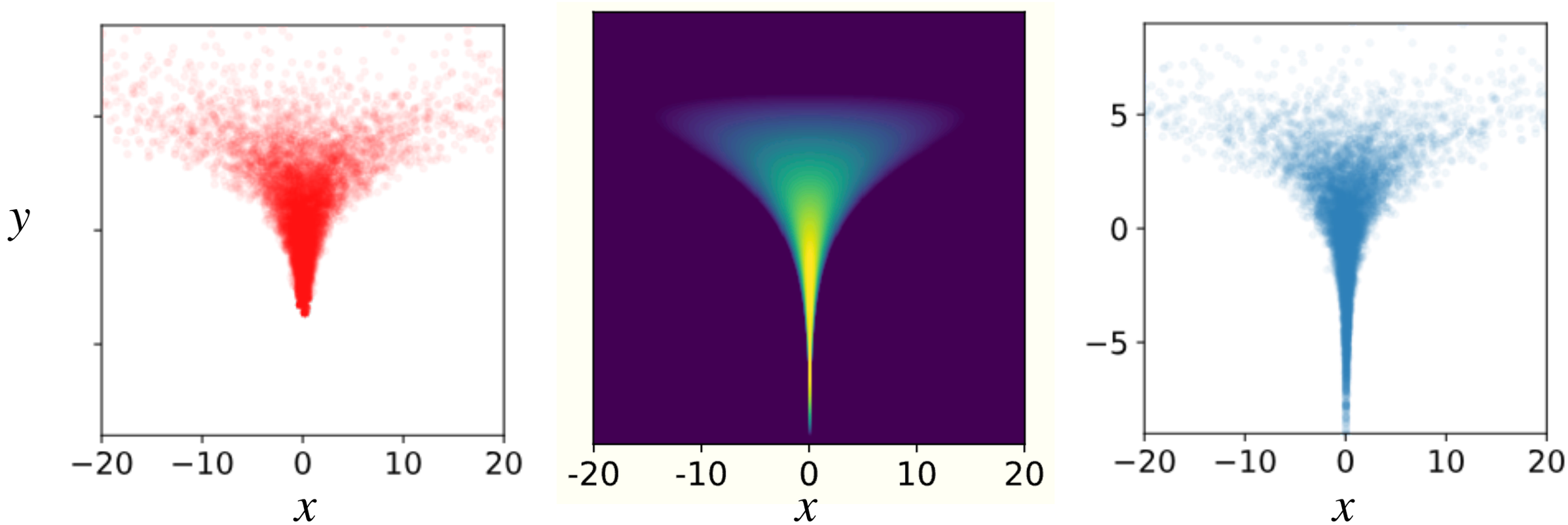



# TUNING STEP SIZE



- Step size $\epsilon$ is chosen to account for the scale of the target
  - When distributionis wide, want to pick a big step size
  - When the distribution is narrow want to pick a small step size
- Small $\epsilon$ leads to high acceptance rates but small movement
- Large $\epsilon$ leads to low acceptance rates and propose samples in unlikely regions
- **Example:** Neil's funnel

$$Y \sim \mathcal{N}(0,3), \qquad X \sim \mathcal{N}(0,e^Y)$$



# OPTIMAL SCALING

# OPTIMAL SCALING



- Maximize the expected square jumping distance (ESJD):

$$\mathbb{E}\left[\,||X_{t+1} - X_t||^2\right]$$

# OPTIMAL SCALING



- Maximize the expected square jumping distance (ESJD):

$$\mathbb{E}\left[||X_{t+1} - X_t||^2\right]$$

  - In general these optimal scaling results are assumed the target is an iide product in $\mathbb{X}^d$

$$\pi_d(x_{1:d}) = \prod_{i=1}^{d} \pi(x_i)$$

# OPTIMAL SCALING



- Maximize the expected square jumping distance (ESJD):

$$\mathbb{E}\left[||X_{t+1} - X_t||^2\right]$$

  - In general these optimal scaling results are assumed the target is an iide product in $\mathbb{X}^d$

$$\pi_d(x_{1:d}) = \prod_{i=1}^{d} \pi(x_i)$$

  - Take a limit as $d \to \infty$ and thus CLT turns the target in to a high dimension gaussian

# OPTIMAL SCALING



- Maximize the expected square jumping distance (ESJD):

$$\mathbb{E}\left[||X_{t+1} - X_t||^2\right]$$

  - In general these optimal scaling results are assumed the target is an iide product in $\mathbb{X}^d$

$$\pi_d(x_{1:d}) = \prod_{i=1}^{d} \pi(x_i)$$

  - Take a limit as $d \to \infty$ and thus CLT turns the target in to a high dimension gaussian
    - Analyse the average acceptance in terms of the ESJD and scaling with dimensions

# OPTIMAL SCALING



- Maximize the expected square jumping distance (ESJD):

$$\mathbb{E}\left[||X_{t+1} - X_t||^2\right]$$

  - In general these optimal scaling results are assumed the target is an iide product in $\mathbb{X}^d$

$$\pi_d(x_{1:d}) = \prod_{i=1}^{d} \pi(x_i)$$

  - Take a limit as $d \to \infty$ and thus CLT turns the target in to a high dimension gaussian
    - Analyse the average acceptance in terms of the ESJD and scaling with dimensions

- Stepsize and acceptance probability required to obtain stable ESJD:

# OPTIMAL SCALING



- Maximize the expected square jumping distance (ESJD):

$$\mathbb{E}\left[||X_{t+1} - X_t||^2\right]$$

  - In general these optimal scaling results are assumed the target is an iide product in $\mathbb{X}^d$

$$\pi_d(x_{1:d}) = \prod_{i=1}^{d} \pi(x_i)$$

  - Take a limit as $d \to \infty$ and thus CLT turns the target in to a high dimension gaussian
    - Analyse the average acceptance in terms of the ESJD and scaling with dimensions

- Stepsize and acceptance probability required to obtain stable ESJD:
  - **RWM** step size scales like $\epsilon = O(d^{-1})$ with "optimal" acceptance probability $\alpha = 0.23$

# OPTIMAL SCALING



- Maximize the expected square jumping distance (ESJD):

$$\mathbb{E}\left[||X_{t+1} - X_t||^2\right]$$

  - In general these optimal scaling results are assumed the target is an iide product in $\mathbb{X}^d$

$$\pi_d(x_{1:d}) = \prod_{i=1}^{d} \pi(x_i)$$

  - Take a limit as $d \to \infty$ and thus CLT turns the target in to a high dimension gaussian
    - Analyse the average acceptance in terms of the ESJD and scaling with dimensions

- Stepsize and acceptance probability required to obtain stable ESJD:
  - **RWM** step size scales like $\epsilon = O(d^{-1})$ with "optimal" acceptance probability $\alpha = 0.23$
  - **MALA** step size scales like $\epsilon = O(d^{-1/3})$ with "optimal" acceptance probability $\alpha = 0.57$

# OPTIMAL SCALING



- Maximize the expected square jumping distance (ESJD):

$$\mathbb{E}\left[||X_{t+1} - X_t||^2\right]$$

  - In general these optimal scaling results are assumed the target is an iide product in $\mathbb{X}^d$

$$\pi_d(x_{1:d}) = \prod_{i=1}^{d} \pi(x_i)$$

  - Take a limit as $d \to \infty$ and thus CLT turns the target in to a high dimension gaussian
    - Analyse the average acceptance in terms of the ESJD and scaling with dimensions

- Stepsize and acceptance probability required to obtain stable ESJD:
  - **RWM** step size scales like $\epsilon = O(d^{-1})$ with "optimal" acceptance probability $\alpha = 0.23$
  - **MALA** step size scales like $\epsilon = O(d^{-1/3})$ with "optimal" acceptance probability $\alpha = 0.57$
  - **HMC** step size scales like $\epsilon = O(d^{-1/4})$ with "optimal" acceptance probability $\alpha = 0.65$

# OPTIMAL SCALING



- Maximize the expected square jumping distance (ESJD):

$$\mathbb{E}\left[||X_{t+1} - X_t||^2\right]$$

  - In general these optimal scaling results are assumed the target is an iide product in $\mathbb{X}^d$

$$\pi_d(x_{1:d}) = \prod_{i=1}^{d} \pi(x_i)$$

  - Take a limit as $d \to \infty$ and thus CLT turns the target in to a high dimension gaussian
    - Analyse the average acceptance in terms of the ESJD and scaling with dimensions

- Stepsize and acceptance probability required to obtain stable ESJD:
  - **RWM** step size scales like $\epsilon = O(d^{-1})$ with "optimal" acceptance probability $\alpha = 0.23$
  - **MALA** step size scales like $\epsilon = O(d^{-1/3})$ with "optimal" acceptance probability $\alpha = 0.57$
  - **HMC** step size scales like $\epsilon = O(d^{-1/4})$ with "optimal" acceptance probability $\alpha = 0.65$

- This provides a powerful heuristic for uni-modal targets, but should be used with caution!

# OPTIMAL SCALING



- Maximize the expected square jumping distance (ESJD):

$$\mathbb{E}\left[||X_{t+1} - X_t||^2\right]$$

  - In general these optimal scaling results are assumed the target is an iide product in $\mathbb{X}^d$

$$\pi_d(x_{1:d}) = \prod_{i=1}^{d} \pi(x_i)$$

  - Take a limit as $d \to \infty$ and thus CLT turns the target in to a high dimension gaussian
    - Analyse the average acceptance in terms of the ESJD and scaling with dimensions

- Stepsize and acceptance probability required to obtain stable ESJD:
  - **RWM** step size scales like $\epsilon = O(d^{-1})$ with "optimal" acceptance probability $\alpha = 0.23$
  - **MALA** step size scales like $\epsilon = O(d^{-1/3})$ with "optimal" acceptance probability $\alpha = 0.57$
  - **HMC** step size scales like $\epsilon = O(d^{-1/4})$ with "optimal" acceptance probability $\alpha = 0.65$

- This provides a powerful heuristic for uni-modal targets, but should be used with caution!
  - In practice, anything between $\alpha \in (0,1)$ is good enough.

# TUNING MASS/COVARIANCE

# TUNING MASS/COVARIANCE

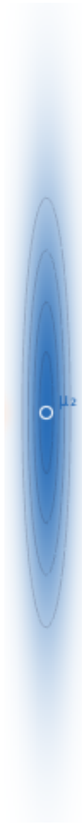


- The mass/covariance matrix $M = \Sigma^{-1}$ is a locally should match the covariance structure of mode

# TUNING MASS/COVARIANCE



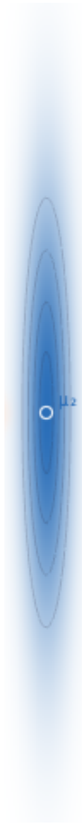
- The mass/covariance matrix $M = \Sigma^{-1}$ is a locally should match the covariance structure of mode



# TUNING MASS/COVARIANCE



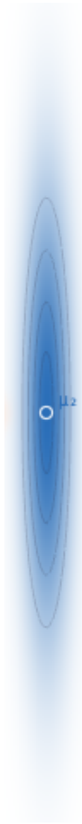
- The mass/covariance matrix $M = \Sigma^{-1}$ is a locally should match the covariance structure of mode
  - Move faster (low mass) in directions where mode stretches



# TUNING MASS/COVARIANCE



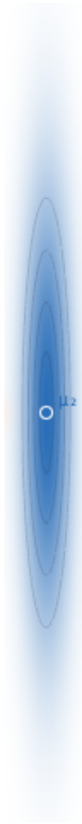
- The mass/covariance matrix $M = \Sigma^{-1}$ is a locally should match the covariance structure of mode
  - Move faster (low mass) in directions where mode stretches
  - Move slow (high mass) in direction where mode contracts



# TUNING MASS/COVARIANCE



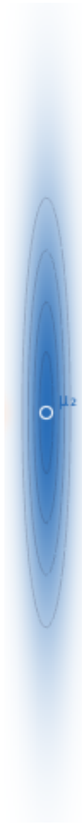
- The mass/covariance matrix $M = \Sigma^{-1}$ is a locally should match the covariance structure of mode
  - Move faster (low mass) in directions where mode stretches
  - Move slow (high mass) in direction where mode contracts
- Common to assume diagonal structure $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ and adaptively learn $\sigma_i^2$



# TUNING MASS/COVARIANCE



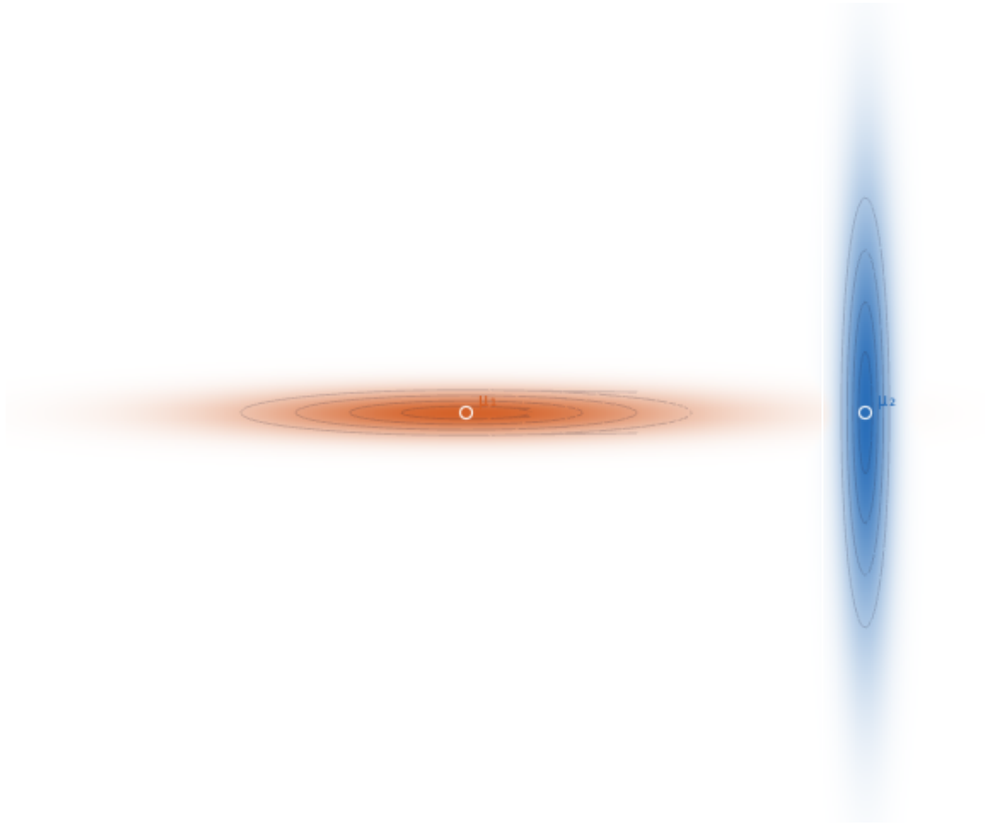
- The mass/covariance matrix $M = \Sigma^{-1}$ is a locally should match the covariance structure of mode
  - Move faster (low mass) in directions where mode stretches
  - Move slow (high mass) in direction where mode contracts
- Common to assume diagonal structure $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ and adaptively learn $\sigma_i^2$
- Hard to tune in multi-modal distributions with wildly different modes



# TUNING MASS/COVARIANCE



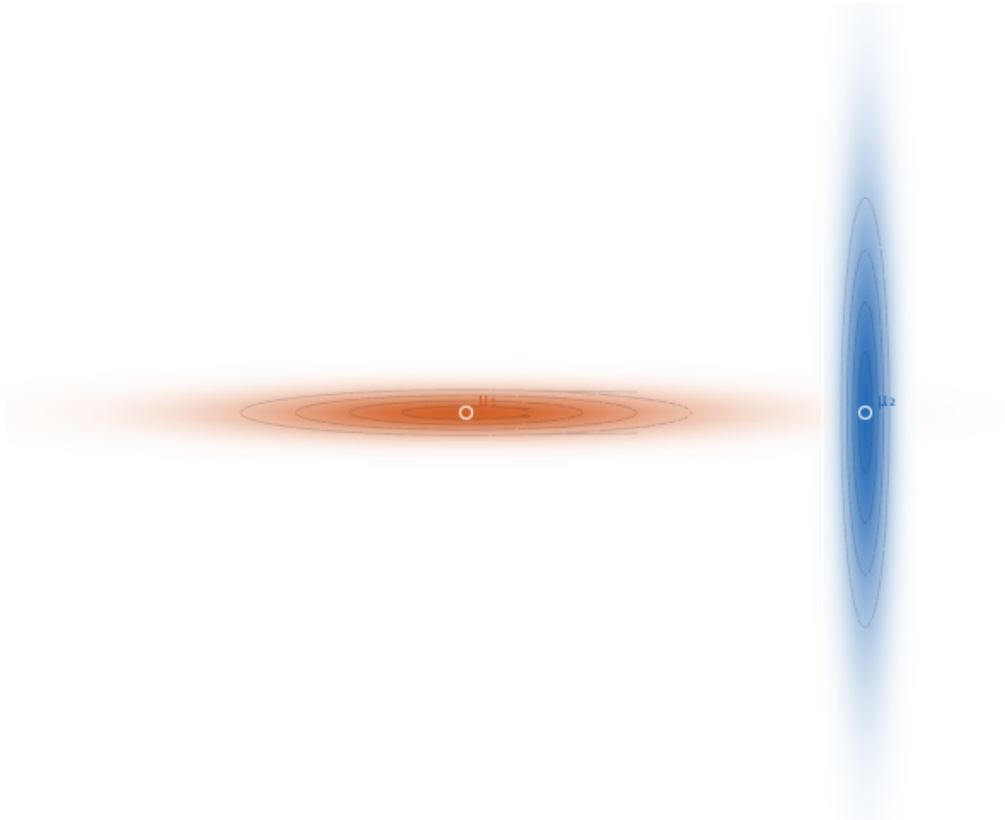
- The mass/covariance matrix $M = \Sigma^{-1}$ is a locally should match the covariance structure of mode
  - Move faster (low mass) in directions where mode stretches
  - Move slow (high mass) in direction where mode contracts
- Common to assume diagonal structure $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ and adaptively learn $\sigma_i^2$
- Hard to tune in multi-modal distributions with wildly different modes



# TUNING MASS/COVARIANCE



- The mass/covariance matrix $M = \Sigma^{-1}$ is a locally should match the covariance structure of mode
  - Move faster (low mass) in directions where mode stretches
  - Move slow (high mass) in direction where mode contracts
- Common to assume diagonal structure $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ and adaptively learn $\sigma_i^2$
- Hard to tune in multi-modal distributions with wildly different modes
  - The best matrix in one mode can be the worst matrix in another…



# TUNING TRAJECTORY LENGTH

# TUNING TRAJECTORY LENGTH

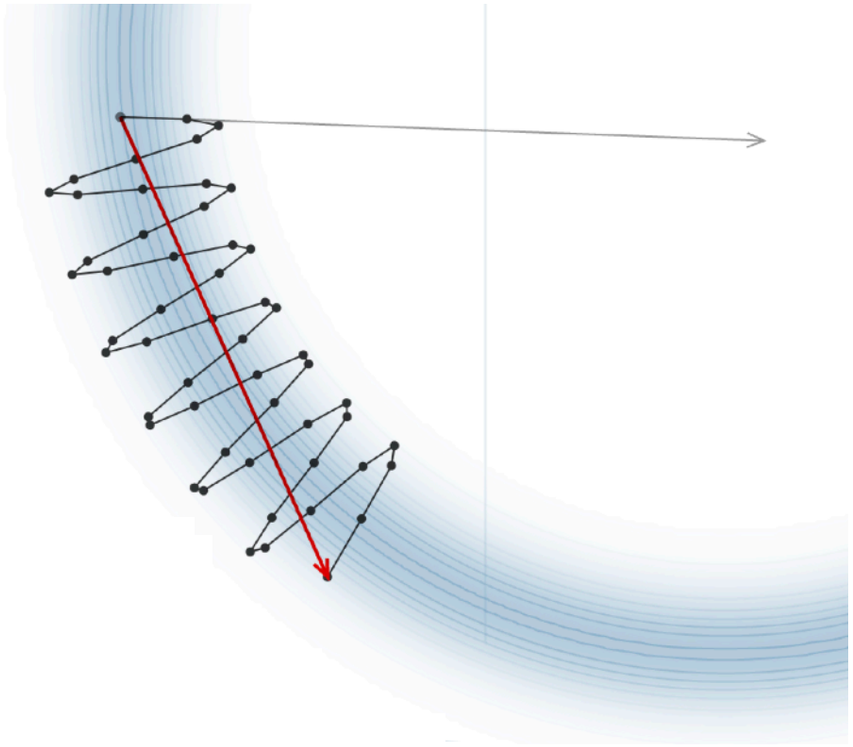


- For HMC additionaly tune integration time $L\epsilon$

# TUNING TRAJECTORY LENGTH



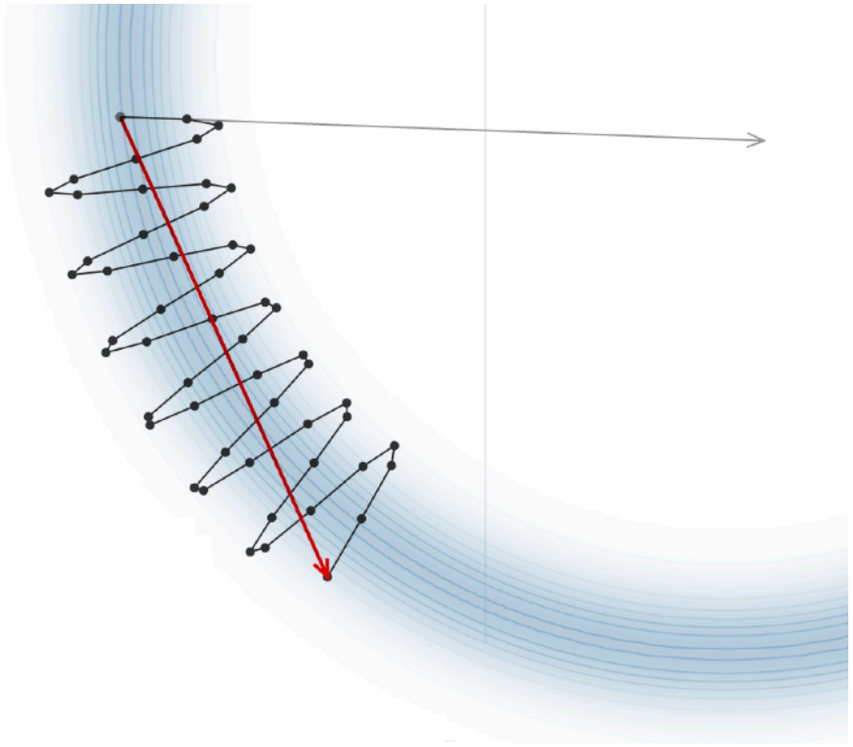
- For HMC additionaly tune integration time $L\epsilon$
- Concersvation of energy means kinetic energy turns into potential
  - Leads to high occilations

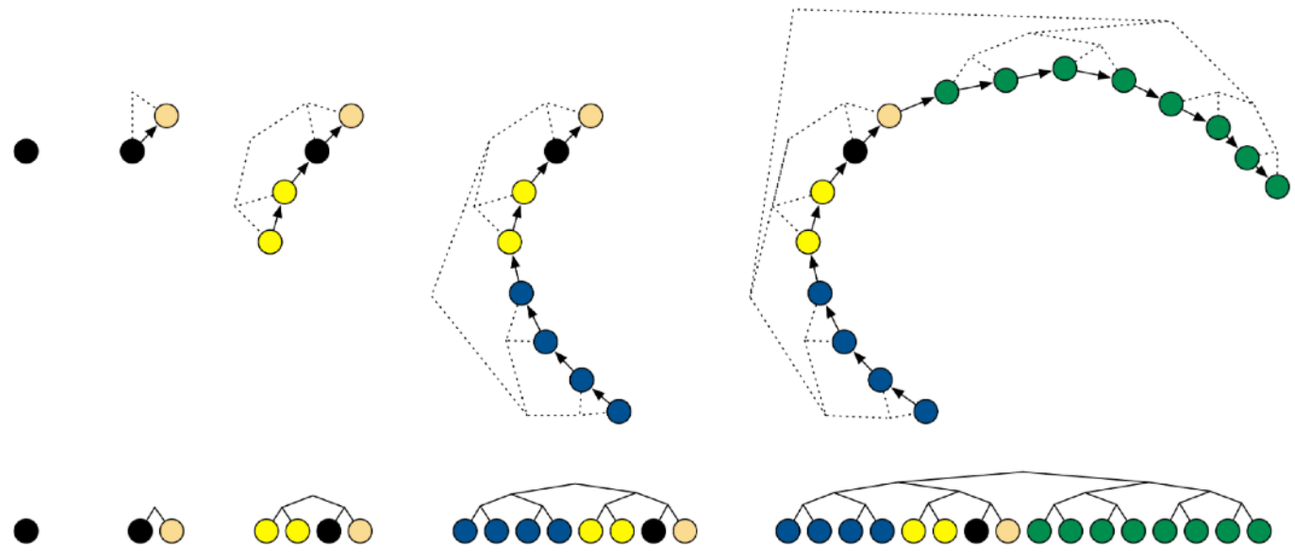


# TUNING TRAJECTORY LENGTH



- For HMC additionaly tune integration time $L\epsilon$
- Concersvation of energy means kinetic energy turns into potential
  - Leads to high occilations
- No U-Turn Samplers (NUTS) (Hoffman et al, 2011) avoid this issue by choosing a clever proposal
  - Admits random run-time for each iteration





# MULTI-MODALITY

# MULTI-MODALITY



- Most MCMC methods are traditionally designed to be **local**

# MULTI-MODALITY



- Most MCMC methods are traditionally designed to be **local**
- We often assume the target distribution is continuous in the topological features of $\mathbb{X}$

$$y \approx x \qquad \Longrightarrow \qquad \pi(y) \approx \pi(x)$$

# MULTI-MODALITY



- Most MCMC methods are traditionally designed to be **local**
- We often assume the target distribution is continuous in the topological features of $\mathbb{X}$

$$y \approx x \qquad \Longrightarrow \qquad \pi(y) \approx \pi(x)$$

- The proposal $Q$ appeals to the topology of the underlying statepace

# MULTI-MODALITY



- Most MCMC methods are traditionally designed to be **local**
- We often assume the target distribution is continuous in the topological features of $\mathbb{X}$

$$y \approx x \qquad \Longrightarrow \qquad \pi(y) \approx \pi(x)$$

- The proposal $Q$ appeals to the topology of the underlying statepace
  - Given $X$ proposal samples $Y \sim Q(X, \mathrm{d}y)$ within a neighbourhood of $X$

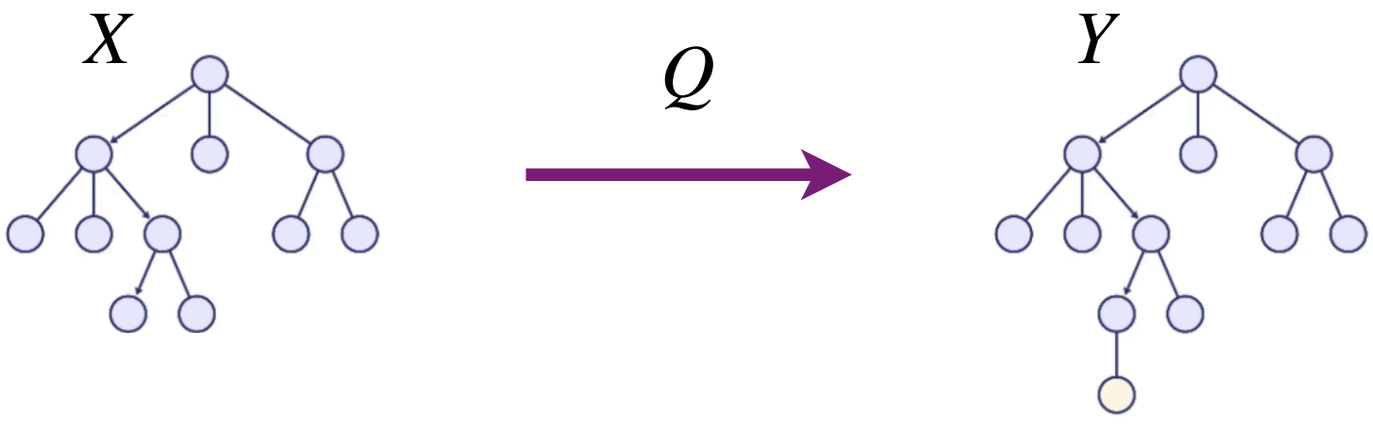
# MULTI-MODALITY



- Most MCMC methods are traditionally designed to be **local**
- We often assume the target distribution is continuous in the topological features of $\mathbb{X}$

$$y \approx x \qquad \Longrightarrow \qquad \pi(y) \approx \pi(x)$$

- The proposal $Q$ appeals to the topology of the underlying statepace
  - Given $X$ proposal samples $Y \sim Q(X, \mathrm{d}y)$ within a neighbourhood of $X$



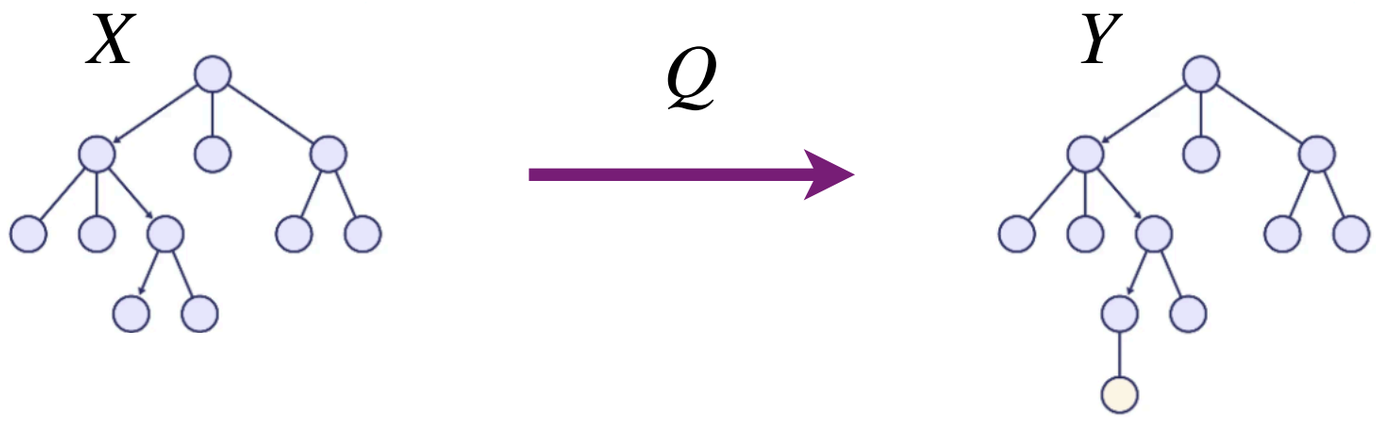
# MULTI-MODALITY



- Most MCMC methods are traditionally designed to be **local**
- We often assume the target distribution is continuous in the topological features of $\mathbb{X}$

$$y \approx x \qquad \Longrightarrow \qquad \pi(y) \approx \pi(x)$$

- The proposal $Q$ appeals to the topology of the underlying statepace
  - Given $X$ proposal samples $Y \sim Q(X, \mathrm{d}y)$ within a neighbourhood of $X$




- Often appeals to the dynamics driven by a differential equation such as Langevin or Hamiltionian

# PITFALLS OF LOCAL SAMPLERS

# PITFALLS OF LOCAL SAMPLERS



- Suppose $\pi_i$ are multiple distributions with support $\mathrm{supp}(\pi_i) = \mathbb{X}_i$ with disjoint support

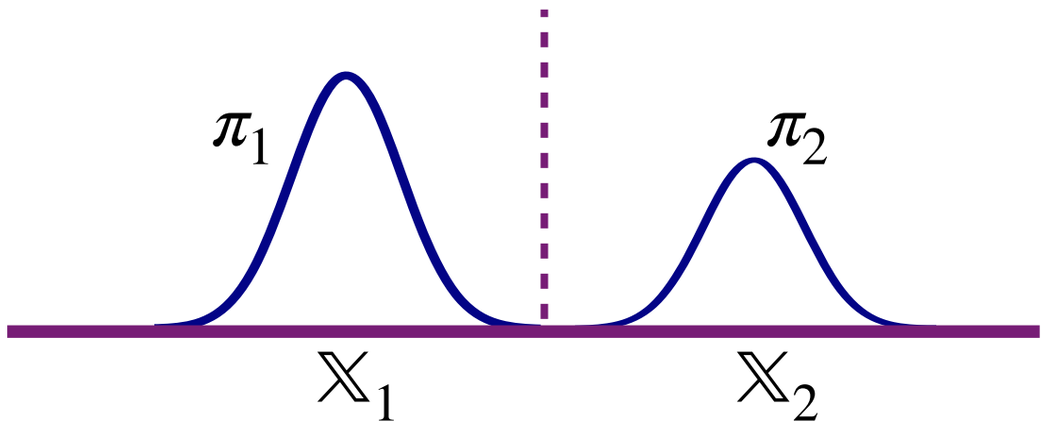
$$\pi(x) = \sum_i w_i \pi_i(x) \delta_{\mathbb{X}_i}(x)$$

# PITFALLS OF LOCAL SAMPLERS



- Suppose $\pi_i$ are multiple distributions with support $\mathrm{supp}(\pi_i) = \mathbb{X}_i$ with disjoint support

$$\pi(x) = \sum_i w_i \pi_i(x) \delta_{\mathbb{X}_i}(x)$$



# PITFALLS OF LOCAL SAMPLERS



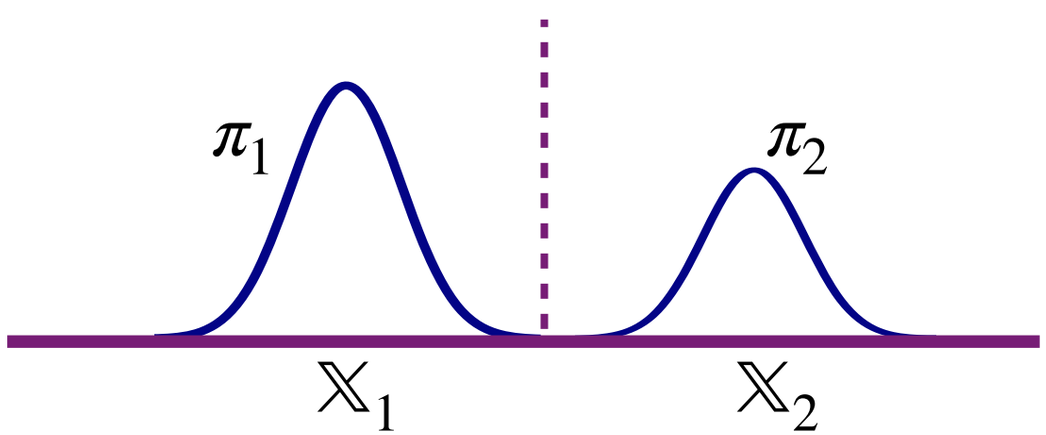
- Suppose $\pi_i$ are multiple distributions with support $\mathrm{supp}(\pi_i) = \mathbb{X}_i$ with disjoint support

$$\pi(x) = \sum_i w_i \pi_i(x) \delta_{\mathbb{X}_i}(x)$$

- Langevin dynamics and Hamiltonian dynamics are not irreducible!

$$\mathrm{d}Y_\tau = -\nabla \log \pi(Y_\tau)\mathrm{d}\tau + \sqrt{2}\mathrm{d}W_\tau$$

$$\mathrm{d}q_t = p_t \mathrm{d}t \qquad \mathrm{d}p_t = \log \pi(q_t)\mathrm{d}t$$



# PITFALLS OF LOCAL SAMPLERS



- Suppose $\pi_i$ are multiple distributions with support $\mathrm{supp}(\pi_i) = \mathbb{X}_i$ with disjoint support

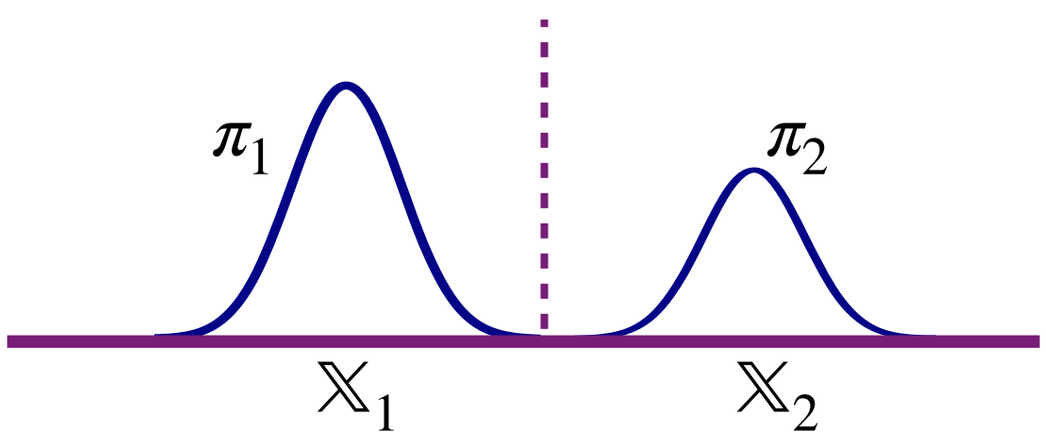
$$\pi(x) = \sum_i w_i \pi_i(x) \delta_{\mathbb{X}_i}(x)$$

- Langevin dynamics and Hamiltonian dynamics are not irreducible!

$$\mathrm{d}Y_\tau = -\nabla \log \pi(Y_\tau)\mathrm{d}\tau + \sqrt{2}\mathrm{d}W_\tau$$

$$\mathrm{d}q_t = p_t \mathrm{d}t \qquad \mathrm{d}p_t = \log \pi(q_t)\mathrm{d}t$$

- The spend zero time in a region where $\pi(x) = 0$



# PITFALLS OF LOCAL SAMPLERS



- Suppose $\pi_i$ are multiple distributions with support $\text{supp}(\pi_i) = \mathbb{X}_i$ with disjoint support

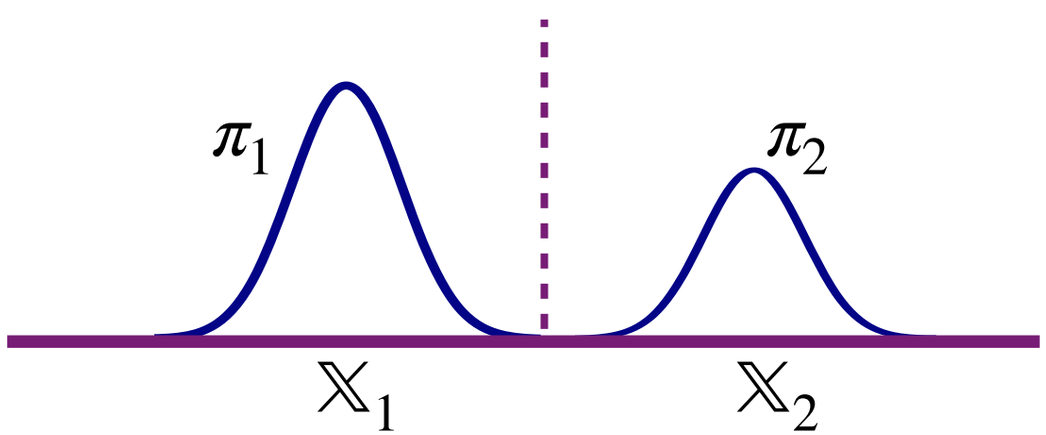
$$\pi(x) = \sum_i w_i \pi_i(x) \delta_{\mathbb{X}_i}(x)$$

- Langevin dynamics and Hamiltonian dynamics are not irreducible!

$$\mathrm{d}Y_\tau = -\nabla \log \pi(Y_\tau)\mathrm{d}\tau + \sqrt{2}\mathrm{d}W_\tau$$

$$\mathrm{d}q_t = p_t \mathrm{d}t \qquad \mathrm{d}p_t = \log \pi(q_t)\mathrm{d}t$$

- The spend zero time in a region where $\pi(x) = 0$
- Random initialisations don't help since attracted to the close modes not big big modes



# PITFALLS OF FIRST ORDER METHODS

# PITFALLS OF FIRST ORDER METHODS



- We we know each mode up to a normalising constant

$$\pi_i(x) = \frac{\gamma_i(x)}{Z_i}, \qquad Z_i = \int_{\mathbb{X}_i} \gamma_i(x) \mathrm{d}x$$

# PITFALLS OF FIRST ORDER METHODS



- We we know each mode up to a normalising constant

$$\pi_i(x) = \frac{\gamma_i(x)}{Z_i}, \qquad Z_i = \int_{\mathbb{X}_i} \gamma_i(x) \mathrm{d}x$$

- The normalising constant provides information about releative weighting of modes:

$$\pi(x) \propto \sum_i w_i \gamma_i(x) \delta_{\mathbb{X}_i}(x), \qquad Z = \sum_i w_i Z_i$$

# PITFALLS OF FIRST ORDER METHODS



- We we know each mode up to a normalising constant

$$\pi_i(x) = \frac{\gamma_i(x)}{Z_i}, \qquad Z_i = \int_{\mathbb{X}_i} \gamma_i(x) \mathrm{d}x$$

- The normalising constant provides information about releative weighting of modes:

$$\pi(x) \propto \sum_i w_i \gamma_i(x) \delta_{\mathbb{X}_i}(x), \qquad Z = \sum_i w_i Z_i$$

- The score is agnostic to the normalising constant and relative weight of the modes

$$\nabla \log \pi_i(x) = \nabla \log \gamma_i(x)$$

$$\nabla \log \pi(x) = \sum_i \nabla \log \pi_i(x) \delta_{\mathbb{X}_i}(x)$$

# PITFALLS OF FIRST ORDER METHODS



- We we know each mode up to a normalising constant

$$\pi_i(x) = \frac{\gamma_i(x)}{Z_i}, \qquad Z_i = \int_{\mathbb{X}_i} \gamma_i(x)\mathrm{d}x$$

- The normalising constant provides information about releative weighting of modes:

$$\pi(x) \propto \sum_i w_i \gamma_i(x)\delta_{\mathbb{X}_i}(x), \qquad Z = \sum_i w_i Z_i$$

- The score is agnostic to the normalising constant and relative weight of the modes

$$\nabla \log \pi_i(x) = \nabla \log \gamma_i(x)$$

$$\nabla \log \pi(x) = \sum_i \nabla \log \pi_i(x)\delta_{\mathbb{X}_i}(x)$$