LECTURE 3

METROPOLIS-HASTINGS

Saifuddin Syed

RECALL

RECALL

- Suppose we have a distribution π that we can evaluate upto a normalising constant

$$\pi(x) = \frac{\gamma(x)}{Z}$$

RECALL

- Suppose we have a distribution π that we can evaluate upto a normalising constant

$$\pi(x) = \frac{\gamma(x)}{Z}$$

- We want to construct an π-invariant, Markov kernel K

RECALL

- Suppose we have a distribution π that we can evaluate upto a normalising constant

$$\pi(x) = \frac{\gamma(x)}{Z}$$

- We want to construct an π-invariant, Markov kernel K

- We build a Markov chain X_t initialised at μ

$$X_0 \sim \mu, \quad X_t \sim K(X_{t-1}, \mathrm{d}x_t)$$

RECALL

- Suppose we have a distribution π that we can evaluate upto a normalising constant

$$\pi(x) = \frac{\gamma(x)}{Z}$$

- We want to construct an π-invariant, Markov kernel K

- We build a Markov chain X_t initialised at μ

$$X_0 \sim \mu, \quad X_t \sim K(X_{t-1}, \mathrm{d}x_t)$$

- Assuming it K is π-invariant, irreducible, aperiod, Harris recurrent, etc…

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} f(X_t) = \pi[f]$$

RECALL

- Suppose we have a distribution π that we can evaluate upto a normalising constant

$$\pi(x) = \frac{\gamma(x)}{Z}$$

- We want to construct an π-invariant, Markov kernel K

- We build a Markov chain X_t initialised at μ

$$X_0 \sim \mu, \quad X_t \sim K(X_{t-1}, \mathrm{d}x_t)$$

- Assuming it K is π-invariant, irreducible, aperiod, Harris recurrent, etc…

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} f(X_t) = \pi[f]$$

- How do we actually construct such a chain in practice?

PROPOSAL KERNELS

PROPOSAL KERNELS

- Suppose we have Q is a Markov kernel such that:

PROPOSAL KERNELS

- Suppose we have Q is a Markov kernel such that:

    1. For all $x \in \mathbb{X}$ we can efficiently sample $Q(x, \mathrm{d}y)$

PROPOSAL KERNELS

- Suppose we have Q is a Markov kernel such that:

    1. For all $x \in \mathbb{X}$ we can efficiently sample $Q(x, \mathrm{d}y)$

    2. Suppose $\mu \ll \tilde{\mu}$ where

PROPOSAL KERNELS

- Suppose we have Q is a Markov kernel such that:

  1. For all $x \in \mathbb{X}$ we can efficiently sample $Q(x, \mathrm{d}y)$

  2. Suppose $\mu \ll \tilde{\mu}$ where

     - μ is the law of (X, Y) where $X \sim \pi$ and $Y \sim Q(X, \mathrm{d}y)$

$$\mu(\mathrm{d}x, \mathrm{d}y) = \pi \otimes Q(\mathrm{d}x, \mathrm{d}y) = \pi(\mathrm{d}x)Q(x, \mathrm{d}y)$$

PROPOSAL KERNELS

- Suppose we have Q is a Markov kernel such that:

  1. For all $x \in \mathbb{X}$ we can efficiently sample $Q(x, \mathrm{d}y)$

  2. Suppose $\mu \ll \tilde{\mu}$ where

     - μ is the law of (X, Y) where $X \sim \pi$ and $Y \sim Q(X, \mathrm{d}y)$

$$\mu(\mathrm{d}x, \mathrm{d}y) = \pi \otimes Q(\mathrm{d}x, \mathrm{d}y) = \pi(\mathrm{d}x)Q(x, \mathrm{d}y)$$

     - $\tilde{\mu}$ is the law of (Y, X) where $Y \sim \pi$ and $X \sim Q(Y, \mathrm{d}x)$

$$\tilde{\mu}(\mathrm{d}x, \mathrm{d}y) = \pi \otimes Q(\mathrm{d}y, \mathrm{d}x) = \pi(\mathrm{d}y)Q(y, \mathrm{d}x)$$

PROPOSAL KERNELS

- Suppose we have Q is a Markov kernel such that:

1. For all $x \in \mathbb{X}$ we can efficiently sample $Q(x, \mathrm{d}y)$

2. Suppose $\mu \ll \tilde{\mu}$ where

   - μ is the law of (X, Y) where $X \sim \pi$ and $Y \sim Q(X, \mathrm{d}y)$

$$\mu(\mathrm{d}x, \mathrm{d}y) = \pi \otimes Q(\mathrm{d}x, \mathrm{d}y) = \pi(\mathrm{d}x)Q(x, \mathrm{d}y)$$

   - $\tilde{\mu}$ is the law of (Y, X) where $Y \sim \pi$ and $X \sim Q(Y, \mathrm{d}x)$

$$\tilde{\mu}(\mathrm{d}x, \mathrm{d}y) = \pi \otimes Q(\mathrm{d}y, \mathrm{d}x) = \pi(\mathrm{d}y)Q(y, \mathrm{d}x)$$

3. We can evaluate the Radon-Nykodym derivative $A : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_+$

$$A(x, y) = \frac{\mathrm{d}\tilde{\mu}}{\mathrm{d}\mu}(x, y) := \frac{\pi \otimes Q(\mathrm{d}y, \mathrm{d}x)}{\pi \otimes Q(\mathrm{d}x, \mathrm{d}y)}$$

PROPOSAL KERNELS

- Suppose we have Q is a Markov kernel such that:

1. For all $x \in \mathbb{X}$ we can efficiently sample $Q(x, \mathrm{d}y)$

2. Suppose $\mu \ll \tilde{\mu}$ where

   - μ is the law of (X, Y) where $X \sim \pi$ and $Y \sim Q(X, \mathrm{d}y)$

$$\mu(\mathrm{d}x, \mathrm{d}y) = \pi \otimes Q(\mathrm{d}x, \mathrm{d}y) = \pi(\mathrm{d}x)Q(x, \mathrm{d}y)$$

   - $\tilde{\mu}$ is the law of (Y, X) where $Y \sim \pi$ and $X \sim Q(Y, \mathrm{d}x)$

$$\tilde{\mu}(\mathrm{d}x, \mathrm{d}y) = \pi \otimes Q(\mathrm{d}y, \mathrm{d}x) = \pi(\mathrm{d}y)Q(y, \mathrm{d}x)$$

3. We can evaluate the Radon-Nykodym derivative $A : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_+$

$$A(x, y) = \frac{\mathrm{d}\tilde{\mu}}{\mathrm{d}\mu}(x, y) := \frac{\pi \otimes Q(\mathrm{d}y, \mathrm{d}x)}{\pi \otimes Q(\mathrm{d}x, \mathrm{d}y)}$$

- Note that $A(x, y)$ does not depend on the normalising constant Z

METROPOLIS-HASTINGS

METROPOLIS-HASTINGS

- Initialise $X_0 \sim \mu$ where $\mu \in \mathscr{P}(\mathbb{X})$ is some intitial condition

METROPOLIS-HASTINGS

- Initialise $X_0 \sim \mu$ where $\mu \in \mathscr{P}(\mathbb{X})$ is some intitial condition
- Given X_{t-1} the Metroplis-Hastings algorithsm generates X_t as follows:

METROPOLIS-HASTINGS

- Initialise $X_0 \sim \mu$ where $\mu \in \mathscr{P}(\mathbb{X})$ is some intitial condition
- Given X_{t-1} the Metroplis-Hastings algorithsm generates X_t as follows:
  - Generate a proposal state Y_t using a proposal kernel $Q(x, \mathrm{d}y)$ that we can sample fro

$$Y_t \sim Q(X_{t-1}, \mathrm{d}y)$$

METROPOLIS-HASTINGS

- Initialise $X_0 \sim \mu$ where $\mu \in \mathscr{P}(\mathbb{X})$ is some intitial condition
- Given X_{t-1} the Metroplis-Hastings algorithsm generates X_t as follows:
  - Generate a proposal state Y_t using a proposal kernel $Q(x, \mathrm{d}y)$ that we can sample fro

$$Y_t \sim Q(X_{t-1}, \mathrm{d}y)$$

  - Set $X_t = Y_t$ with probability $\alpha(X_{t-1}, Y_t)$ and $X_t = X_{t-1}$ otherwise.

$$\alpha(x, y) = 1 \wedge A(x, y) = 1 \wedge \frac{\pi(\mathrm{d}y)Q(y, \mathrm{d}x)}{\pi(\mathrm{d}x)Q(x, \mathrm{d}y)}$$

METROPOLIS-HASTINGS

- Initialise $X_0 \sim \mu$ where $\mu \in \mathscr{P}(\mathbb{X})$ is some intitial condition
- Given X_{t-1} the Metroplis-Hastings algorithsm generates X_t as follows:
  - Generate a proposal state Y_t using a proposal kernel $Q(x, \mathrm{d}y)$ that we can sample fro

$$Y_t \sim Q(X_{t-1}, \mathrm{d}y)$$

  - Set $X_t = Y_t$ with probability $\alpha(X_{t-1}, Y_t)$ and $X_t = X_{t-1}$ otherwise.

$$\alpha(x, y) = 1 \wedge A(x, y) = 1 \wedge \frac{\pi(\mathrm{d}y)Q(y, \mathrm{d}x)}{\pi(\mathrm{d}x)Q(x, \mathrm{d}y)}$$

- **Algorithm:**

Initialise $X_0 \sim \eta$
For $t = 1, \ldots, T$:
1. Generate $Y_t \sim Q(X_{t-1}, \mathrm{d}y)$
2. Compute $\alpha(X_{t-1}, Y_t)$
3. Generate $U_t \sim \mathrm{Uniform}([0,1])$
   A. If $U_t \leq \alpha_t(X_{t-1}, Y_t)$ set $X_t = Y_t$
   B. Else return $X_t = X_{t-1}$

PROPOSAL DENSITY

PROPOSAL DENSITY

- Example: $Q(x, \mathrm{d}y) = q(x, y)\mathrm{d}y$ then

$$A(x, y) = \frac{\pi \otimes Q(\mathrm{d}y, \mathrm{d}x)}{\pi \otimes Q(\mathrm{d}x, \mathrm{d}y)} = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}$$

PROPOSAL DENSITY

- Example: $Q(x, \mathrm{d}y) = q(x, y)\mathrm{d}y$ then

$$A(x, y) = \frac{\pi \otimes Q(\mathrm{d}y, \mathrm{d}x)}{\pi \otimes Q(\mathrm{d}x, \mathrm{d}y)} = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}$$

  - The acceptance probability is independent of the normalising constnat Z

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}$$

PROPOSAL DENSITY

- Example: $Q(x, \mathrm{d}y) = q(x, y)\mathrm{d}y$ then

$$A(x, y) = \frac{\pi \otimes Q(\mathrm{d}y, \mathrm{d}x)}{\pi \otimes Q(\mathrm{d}x, \mathrm{d}y)} = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}$$

  - The acceptance probability is independent of the normalising constnat Z

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} = 1 \wedge \frac{\gamma(y)q(y, x)}{\gamma(x)q(x, y)}$$

PROPOSAL DENSITY

- Example: $Q(x, \mathrm{d}y) = q(x, y)\mathrm{d}y$ then

$$A(x, y) = \frac{\pi \otimes Q(\mathrm{d}y, \mathrm{d}x)}{\pi \otimes Q(\mathrm{d}x, \mathrm{d}y)} = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}$$

  - The acceptance probability is independent of the normalising constnat Z

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} = 1 \wedge \frac{\gamma(y)q(y, x)}{\gamma(x)q(x, y)}$$

- If the proposal is symmetric then $q(x, y) = q(y, x)$

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)}{\pi(x)}$$

PROPOSAL DENSITY

- Example: $Q(x, \mathrm{d}y) = q(x, y)\mathrm{d}y$ then

$$A(x, y) = \frac{\pi \otimes Q(\mathrm{d}y, \mathrm{d}x)}{\pi \otimes Q(\mathrm{d}x, \mathrm{d}y)} = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}$$

  - The acceptance probability is independent of the normalising constnat Z

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} = 1 \wedge \frac{\gamma(y)q(y, x)}{\gamma(x)q(x, y)}$$

- If the proposal is symmetric then $q(x, y) = q(y, x)$

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)}{\pi(x)}$$

  - If the proposed state is more likely then always accept

$$\pi(y) \geq \pi(x) \quad \Longrightarrow \quad \alpha(x, y) = 1$$

PROPOSAL DENSITY

- Example: $Q(x, \mathrm{d}y) = q(x, y)\mathrm{d}y$ then

$$A(x, y) = \frac{\pi \otimes Q(\mathrm{d}y, \mathrm{d}x)}{\pi \otimes Q(\mathrm{d}x, \mathrm{d}y)} = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}$$

  - The acceptance probability is independent of the normalising constnat Z

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} = 1 \wedge \frac{\gamma(y)q(y, x)}{\gamma(x)q(x, y)}$$

- If the proposal is symmetric then $q(x, y) = q(y, x)$

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)}{\pi(x)}$$

  - If the proposed state is more likely then always accept

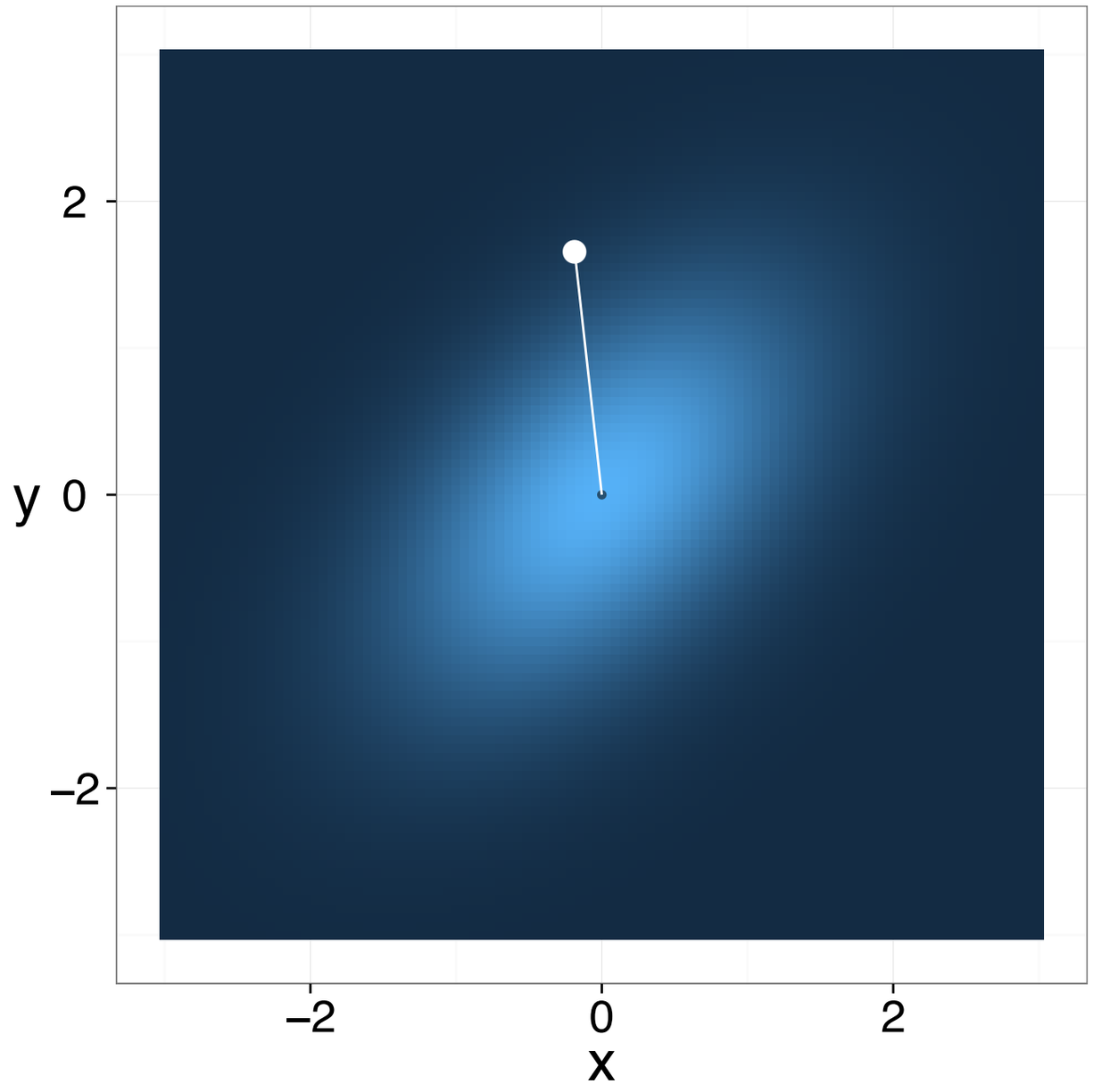
$$\pi(y) \geq \pi(x) \quad \Longrightarrow \quad \alpha(x, y) = 1$$

  - Other reject with proportional to the likelihood

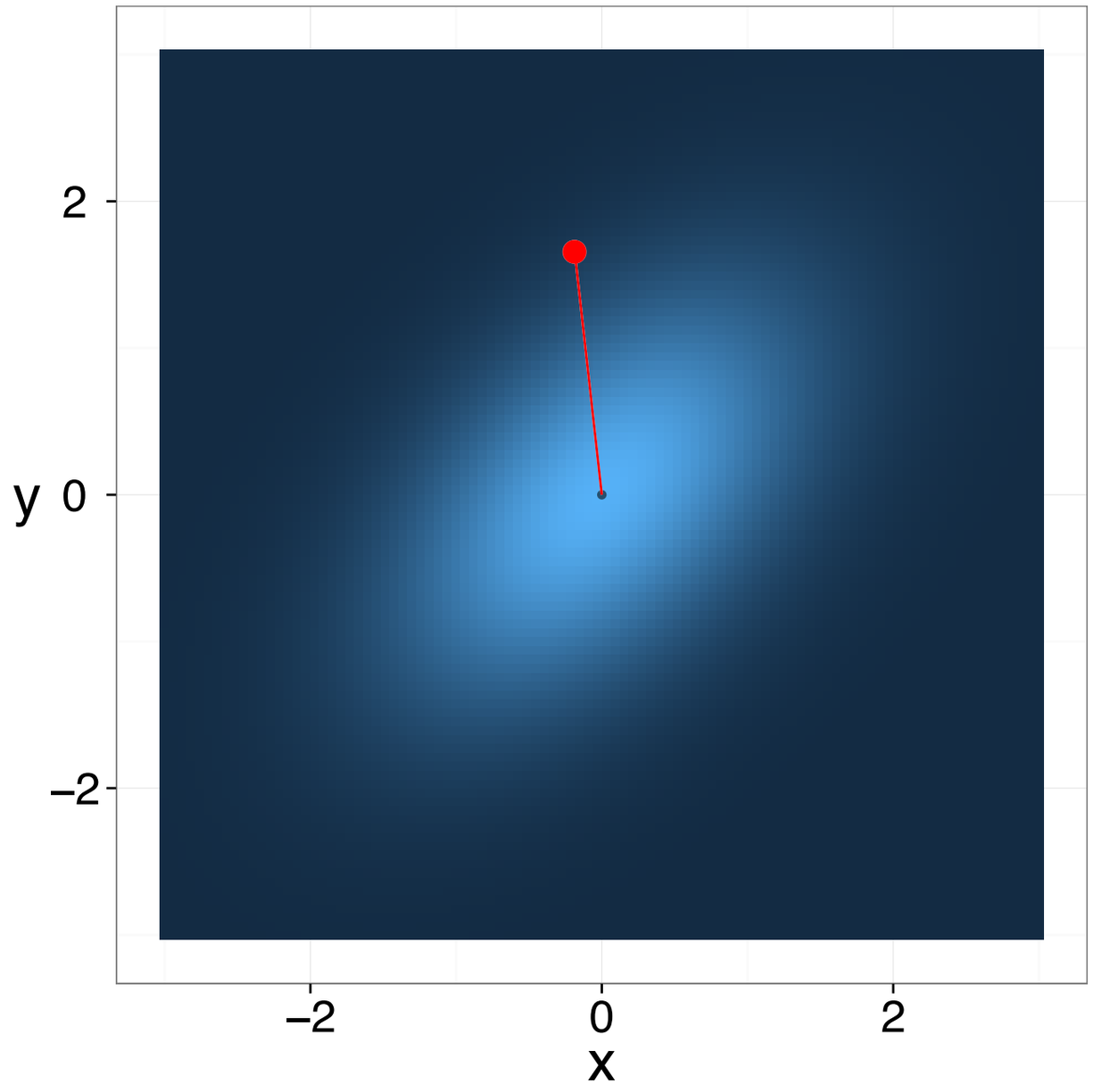
$$\pi(y) < \pi(x) \quad \Longrightarrow \quad \alpha(x, y) = \frac{\pi(y)}{\pi(x)} < 1$$

EXAMPLE: BIVARIATE GAUSSIAN

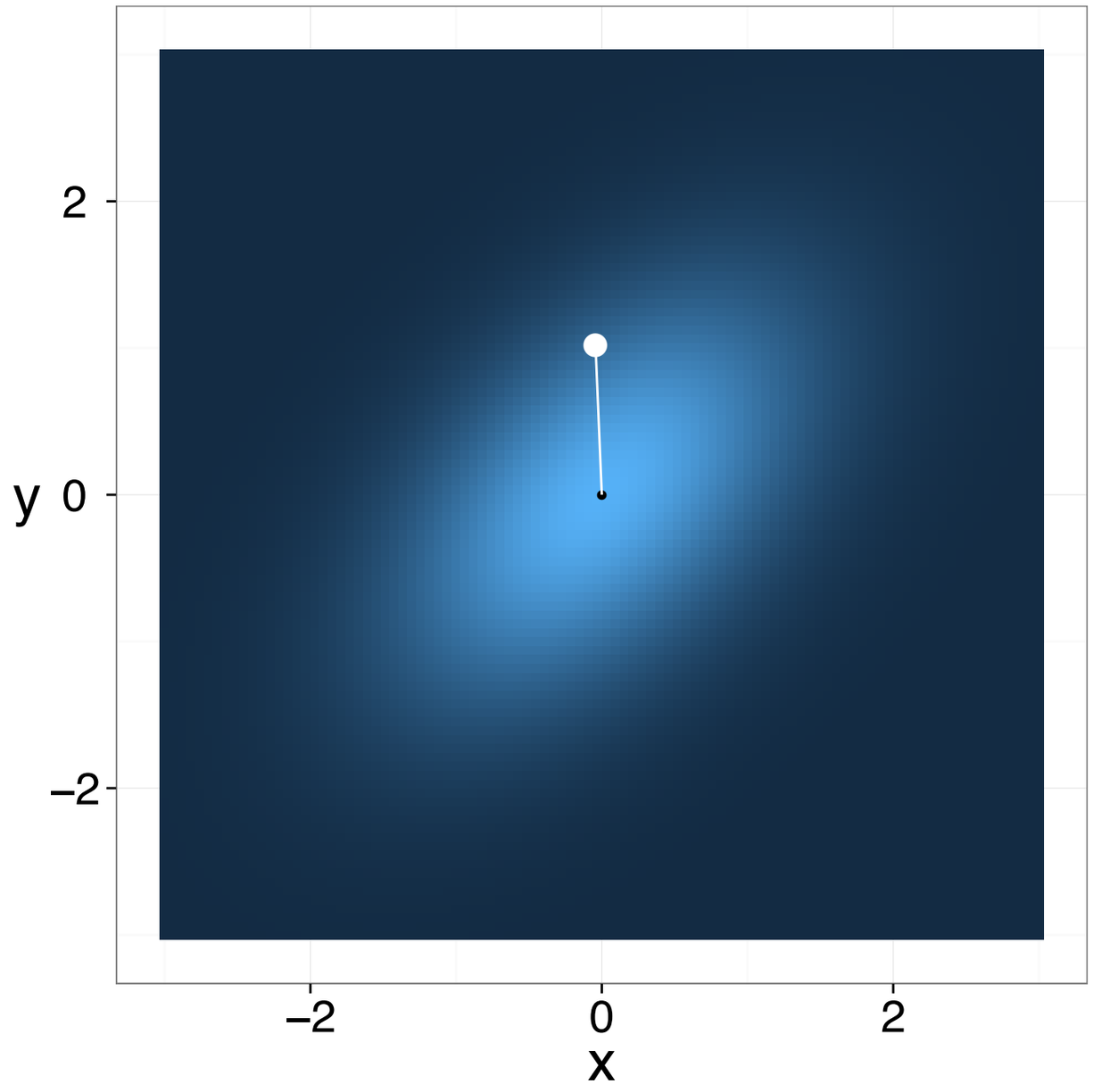
EXAMPLE: BIVARIATE GAUSSIAN



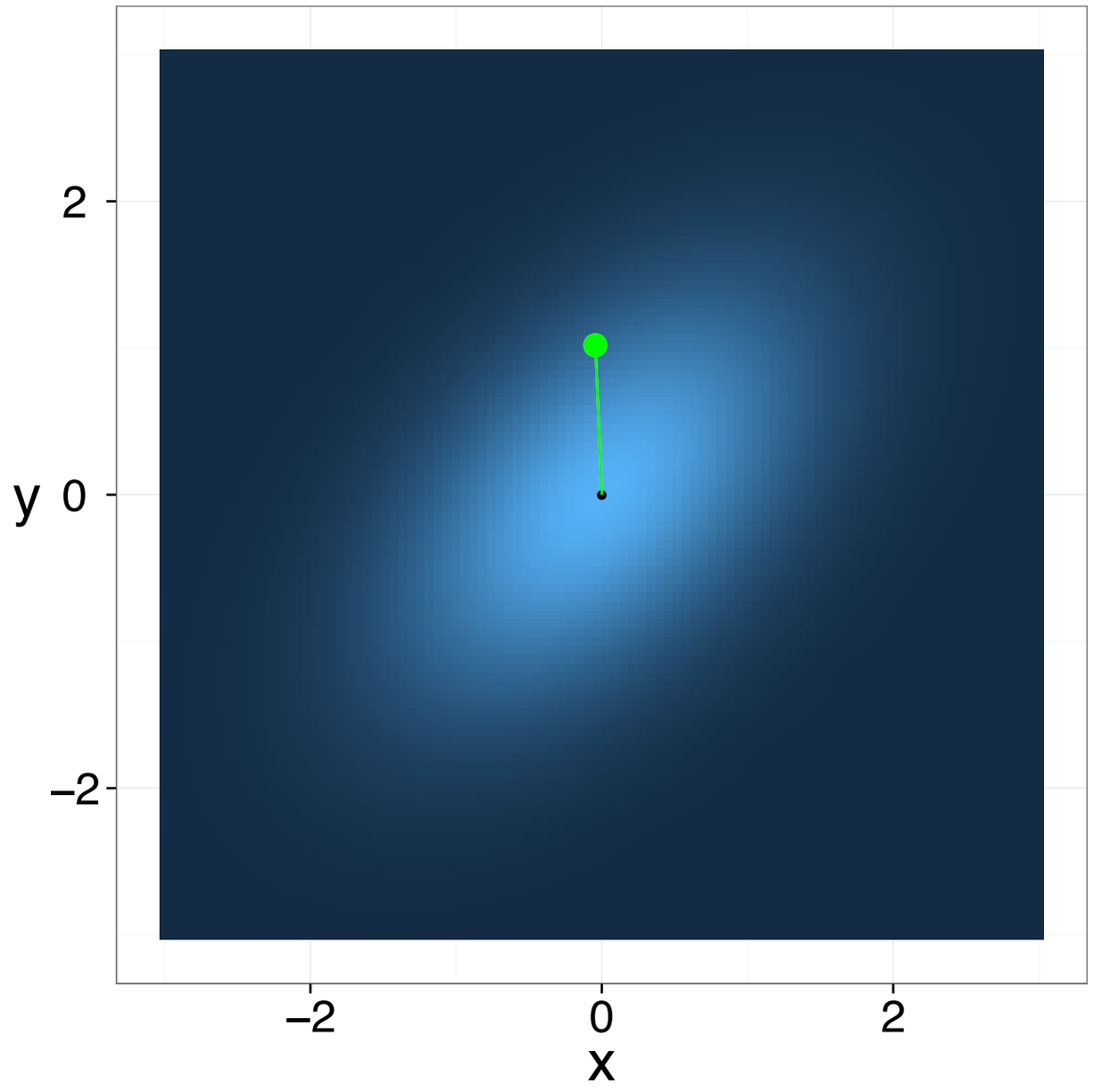
EXAMPLE: BIVARIATE GAUSSIAN



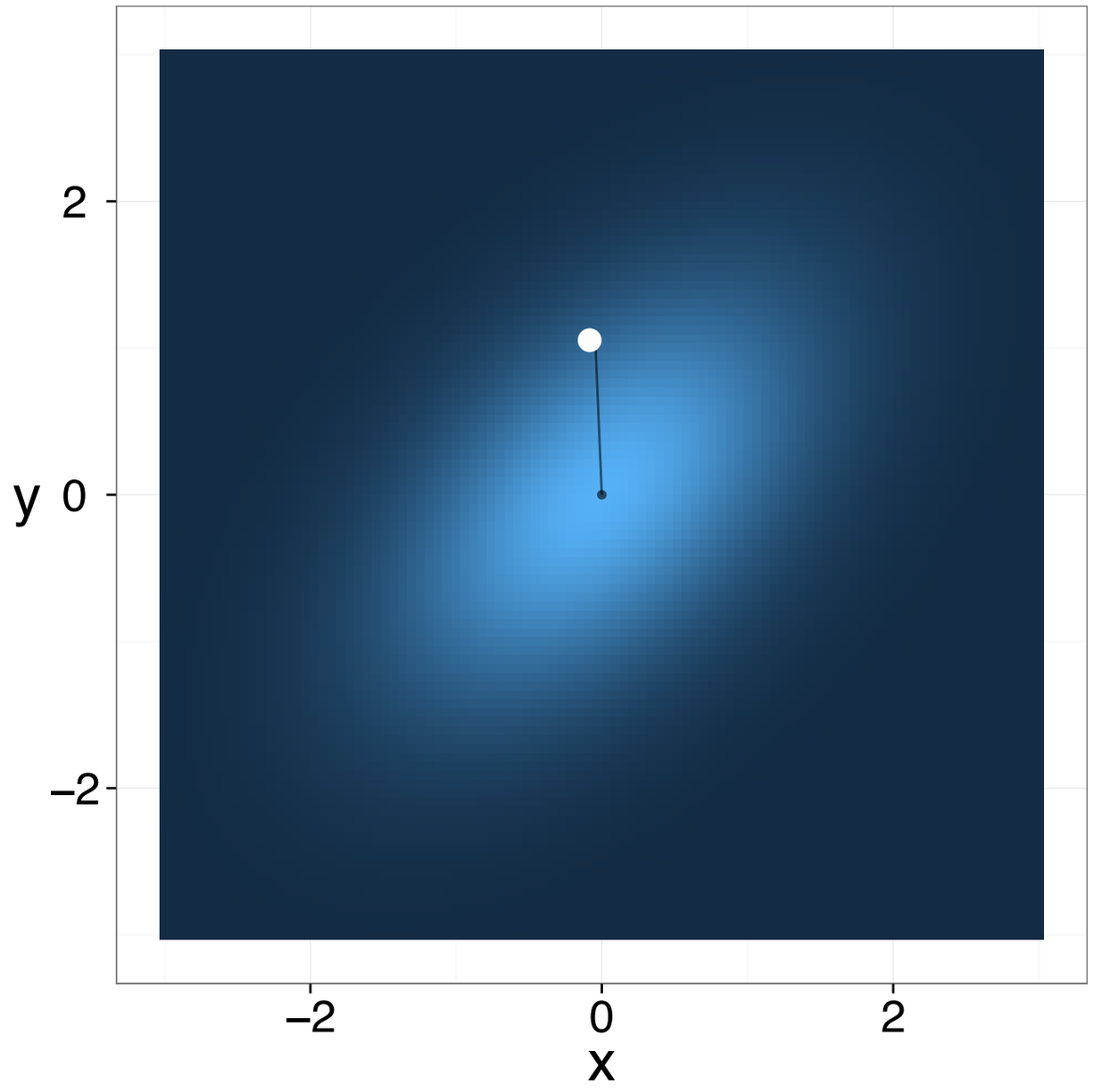
EXAMPLE: BIVARIATE GAUSSIAN



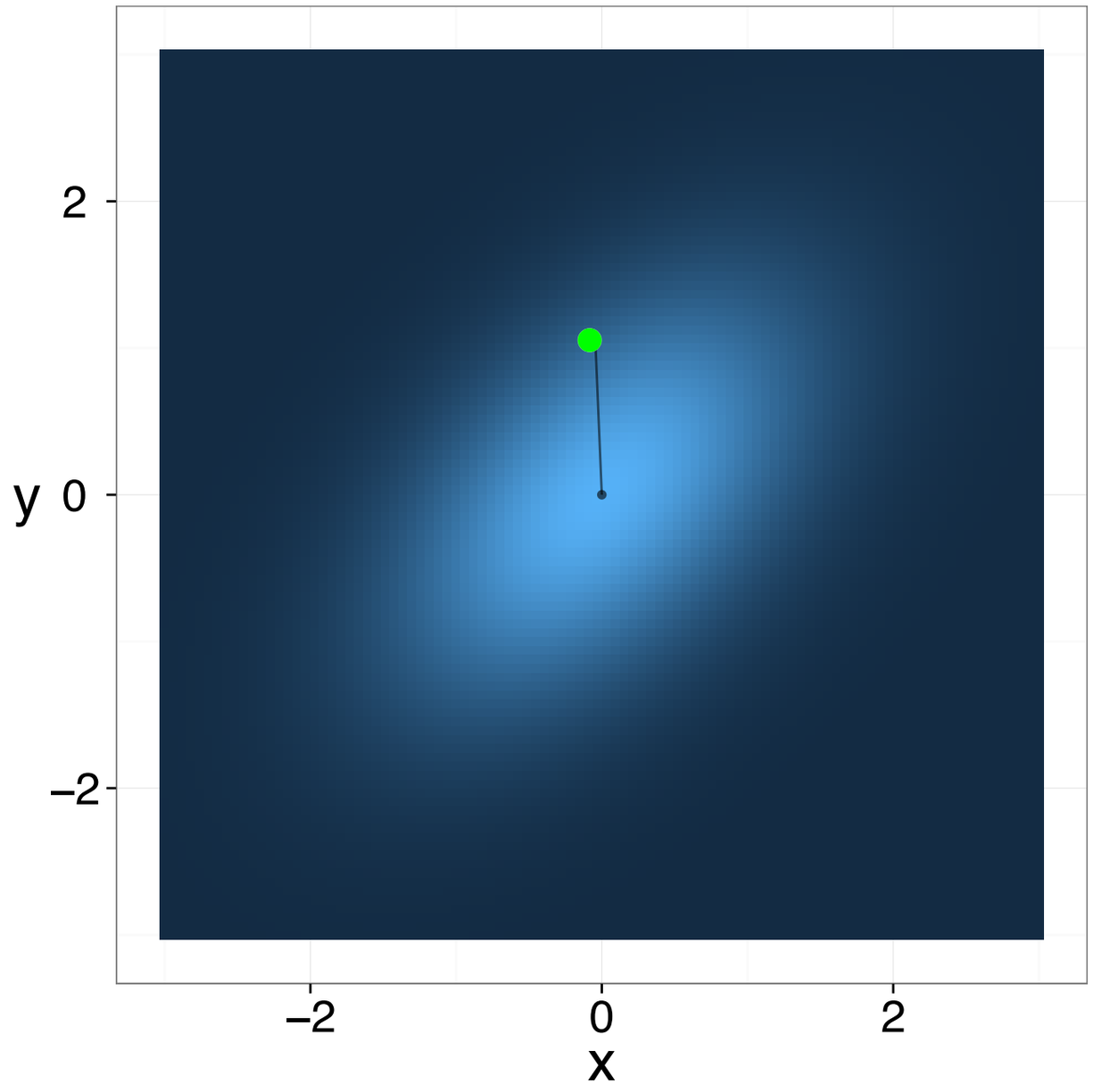
EXAMPLE: BIVARIATE GAUSSIAN



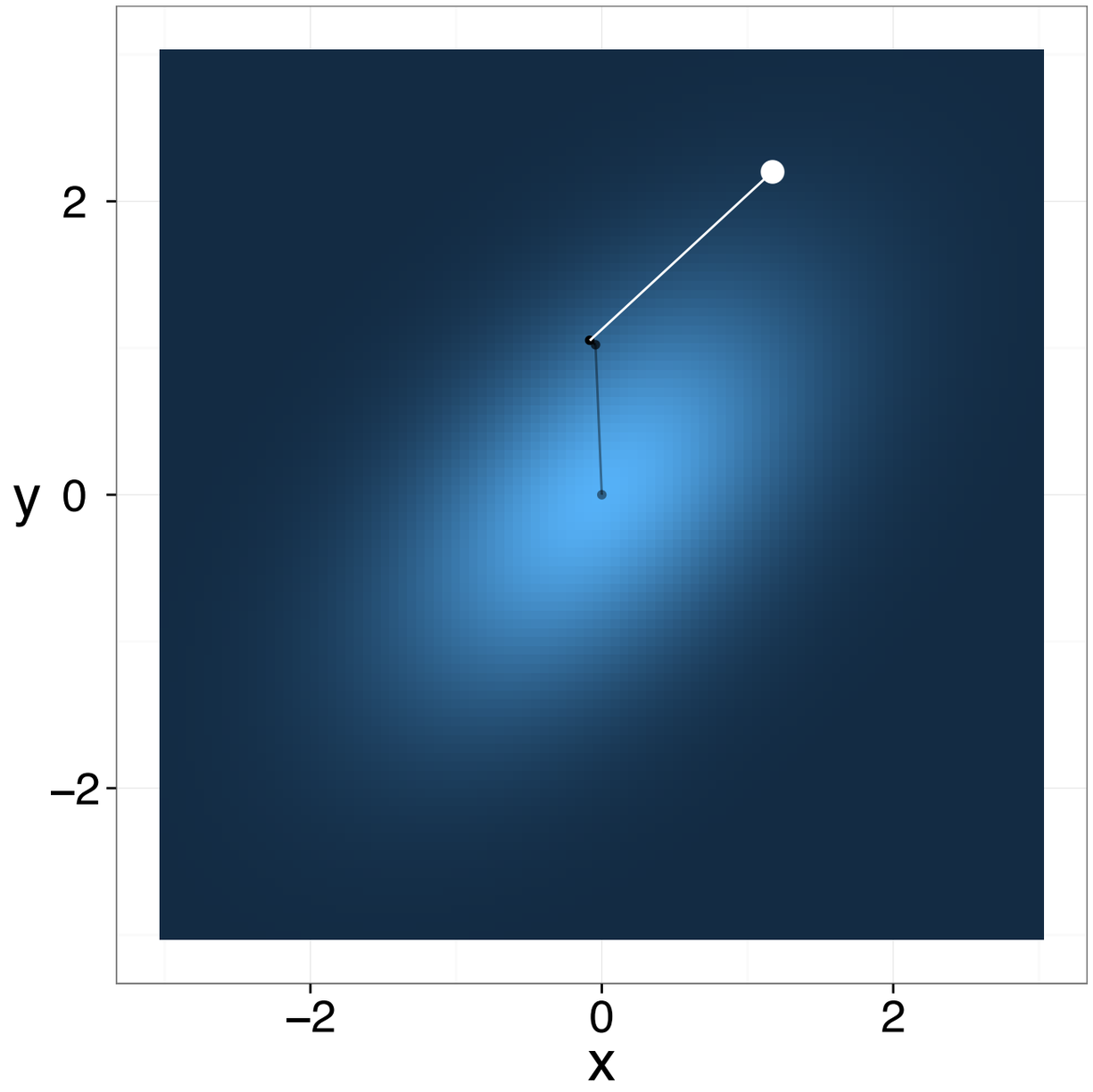
EXAMPLE: BIVARIATE GAUSSIAN



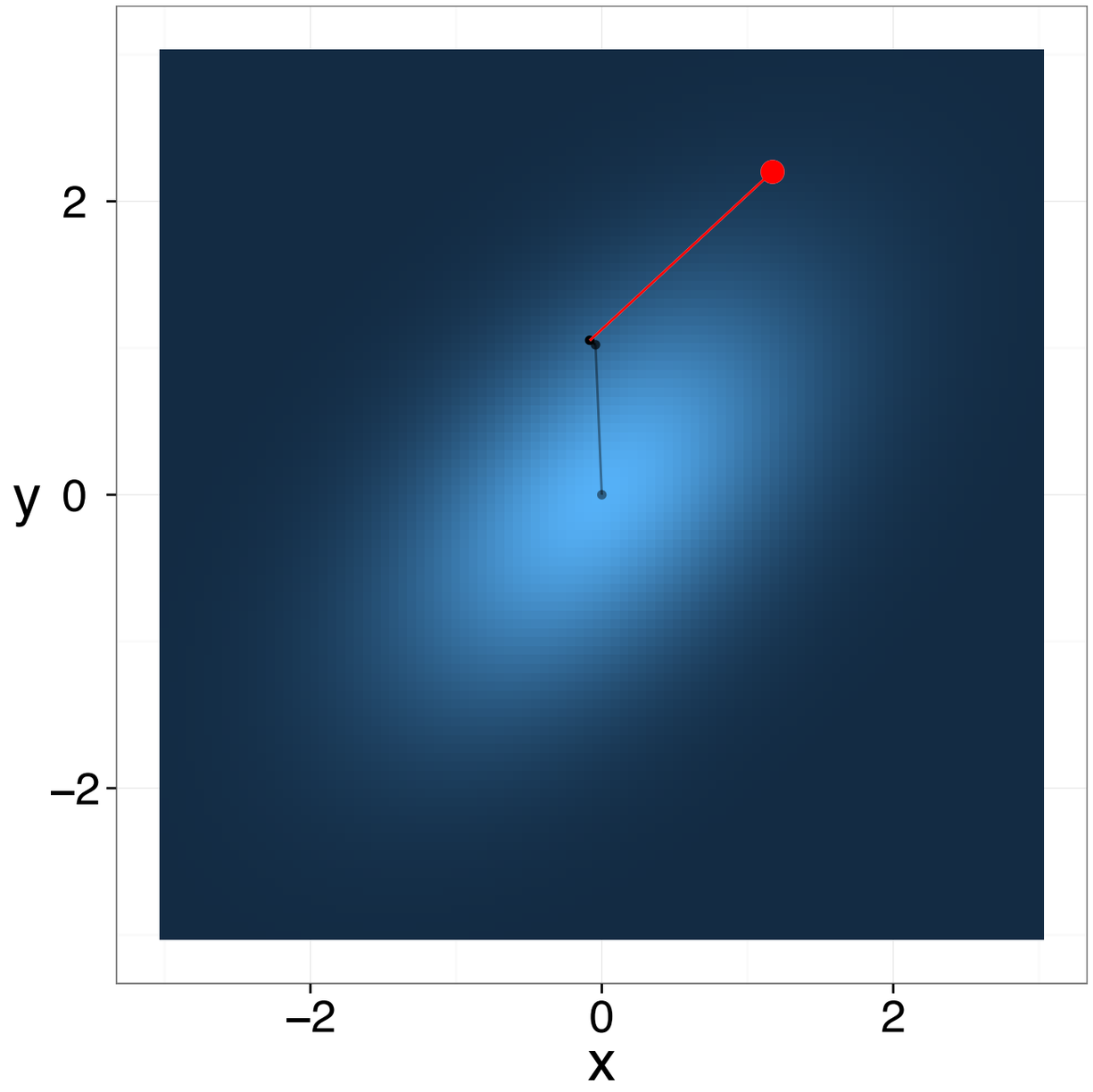
EXAMPLE: BIVARIATE GAUSSIAN



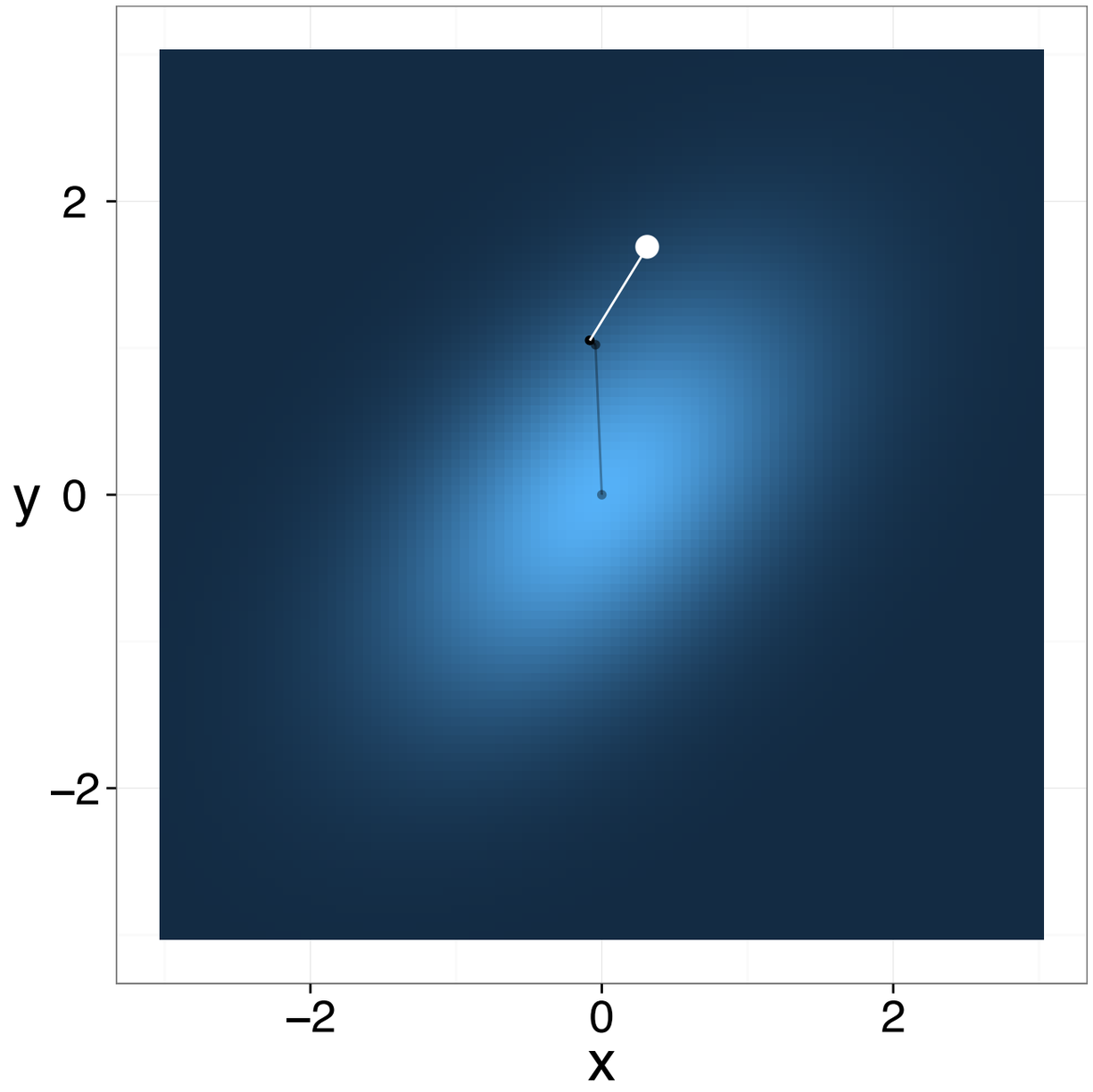
EXAMPLE: BIVARIATE GAUSSIAN



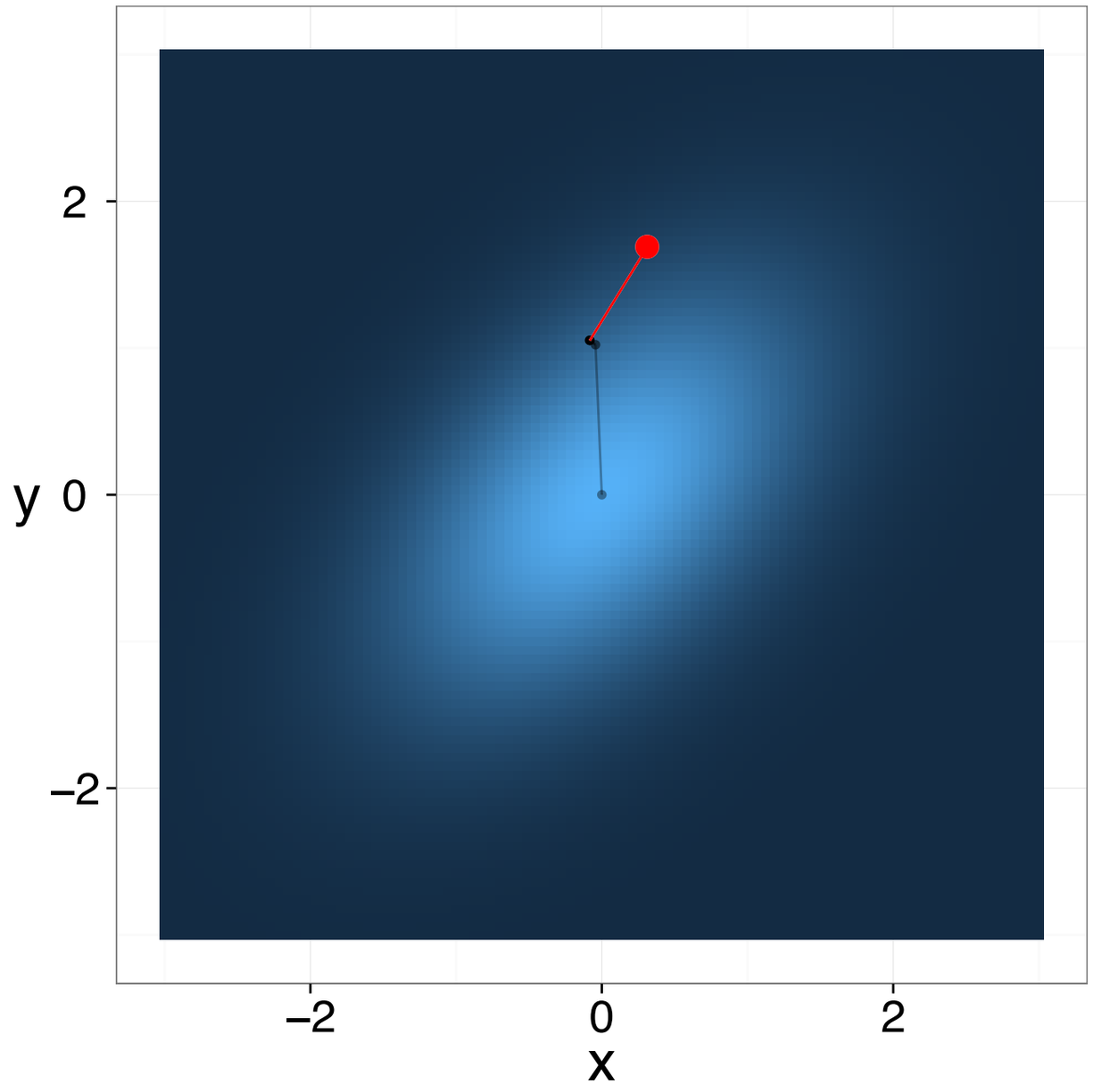
EXAMPLE: BIVARIATE GAUSSIAN



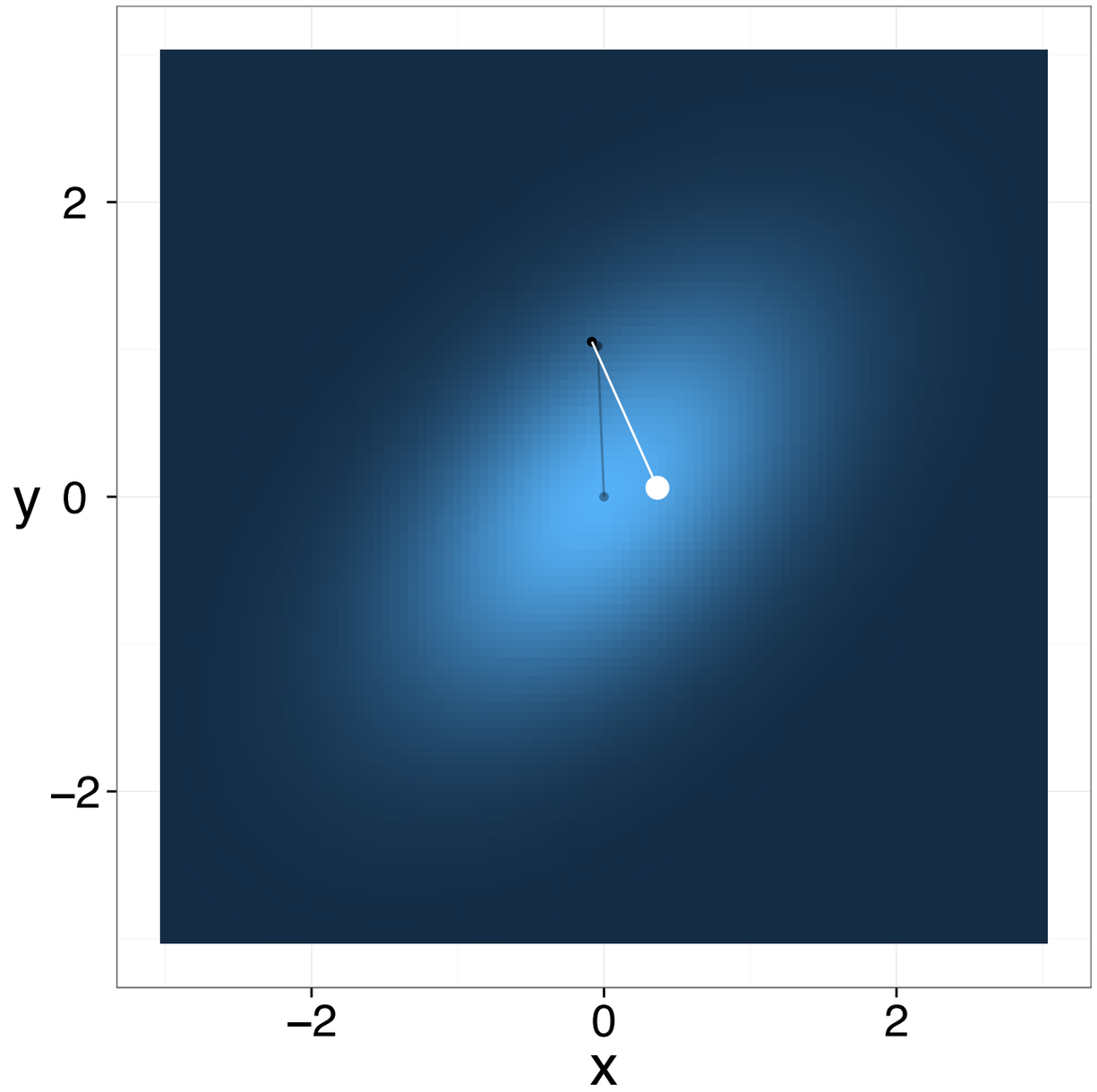
EXAMPLE: BIVARIATE GAUSSIAN



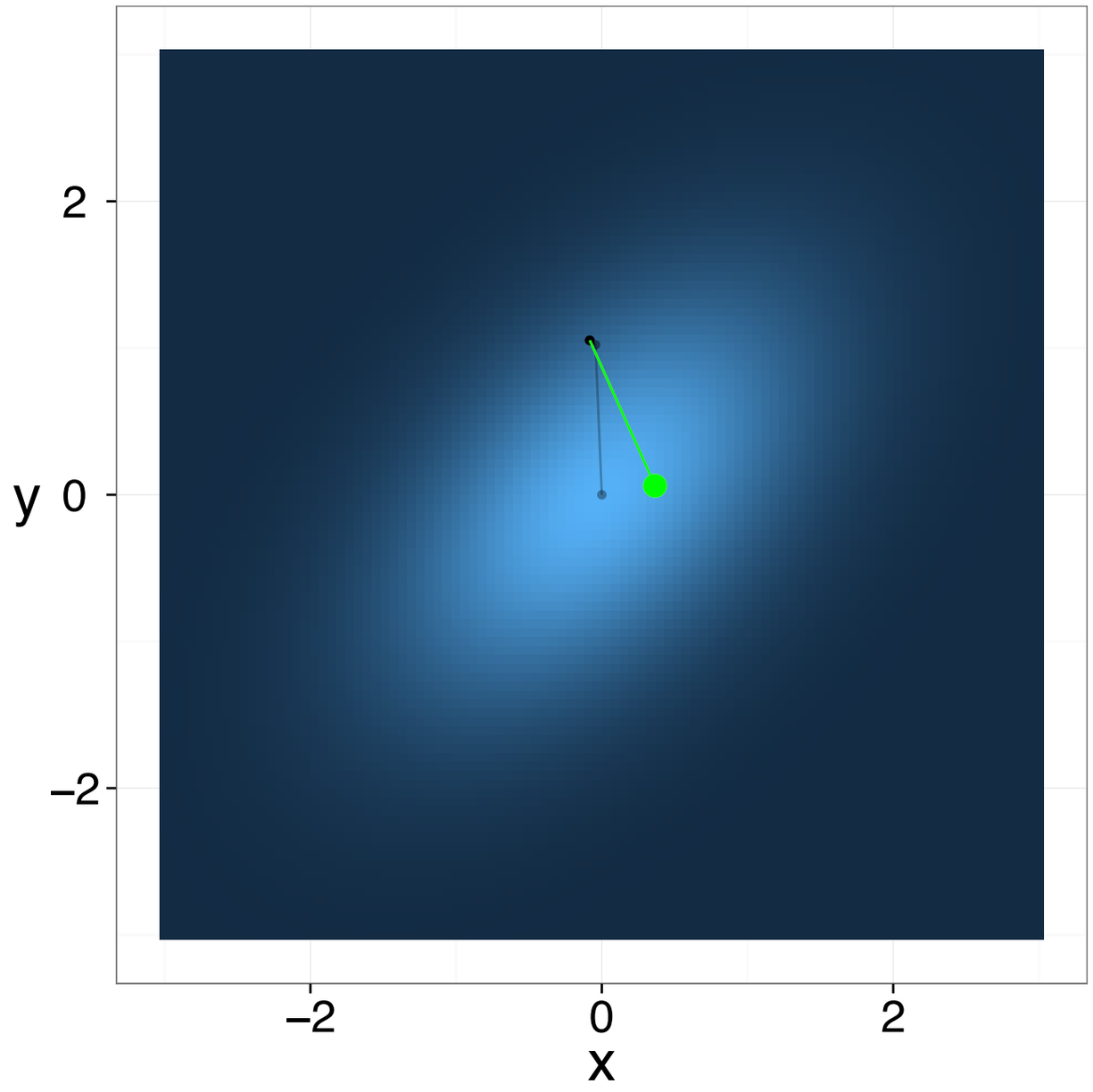
EXAMPLE: BIVARIATE GAUSSIAN



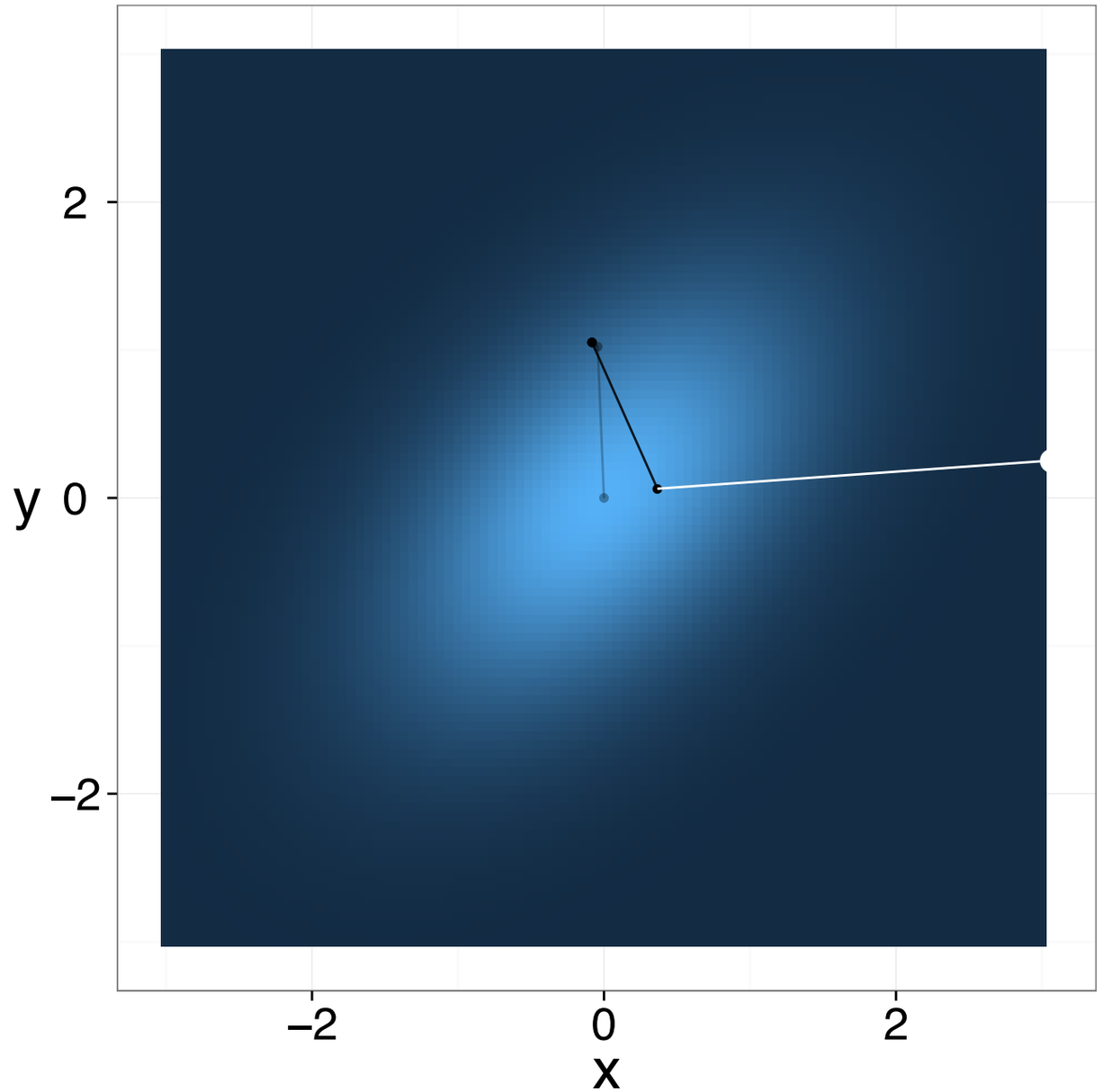
EXAMPLE: BIVARIATE GAUSSIAN



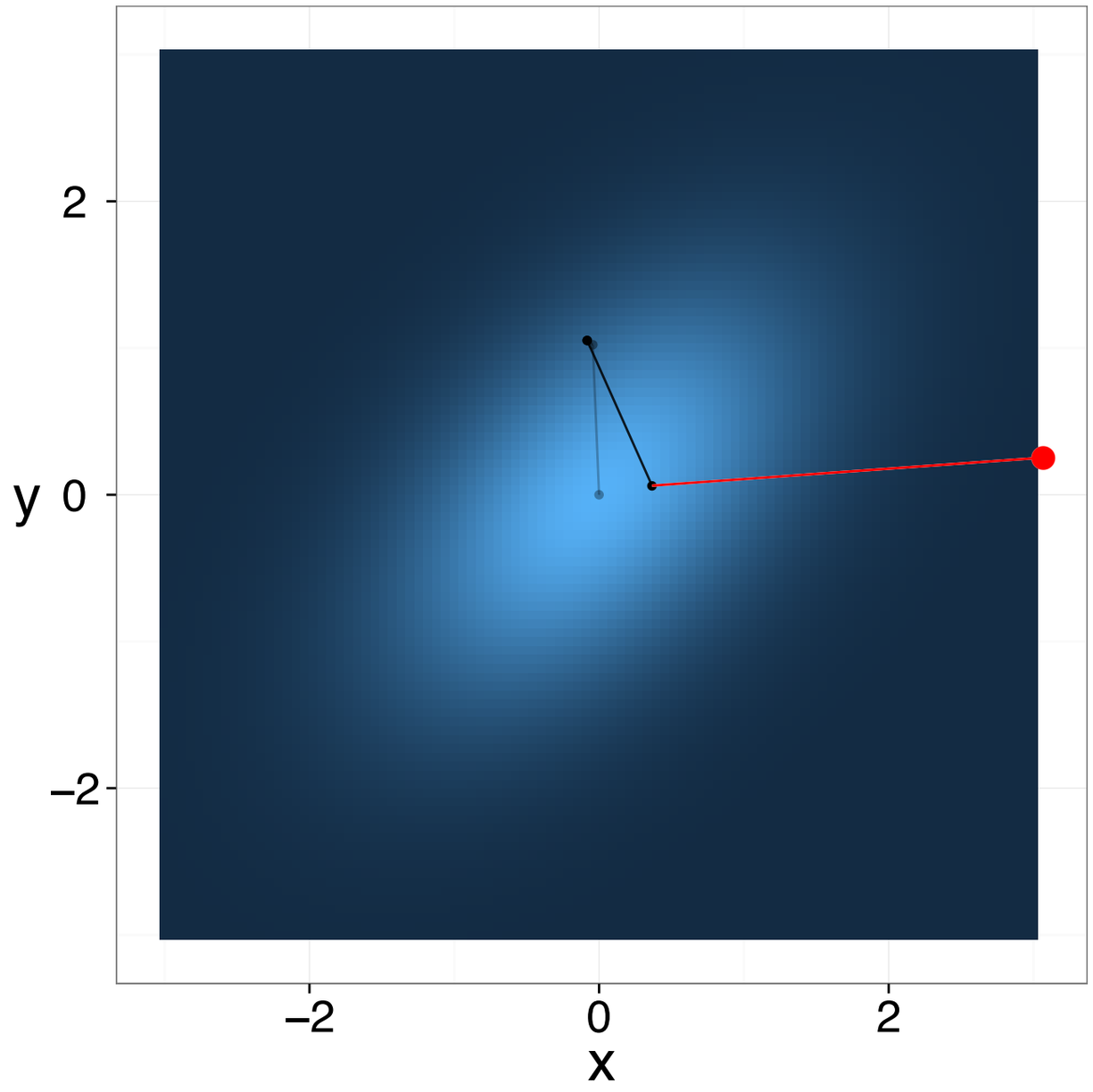
EXAMPLE: BIVARIATE GAUSSIAN



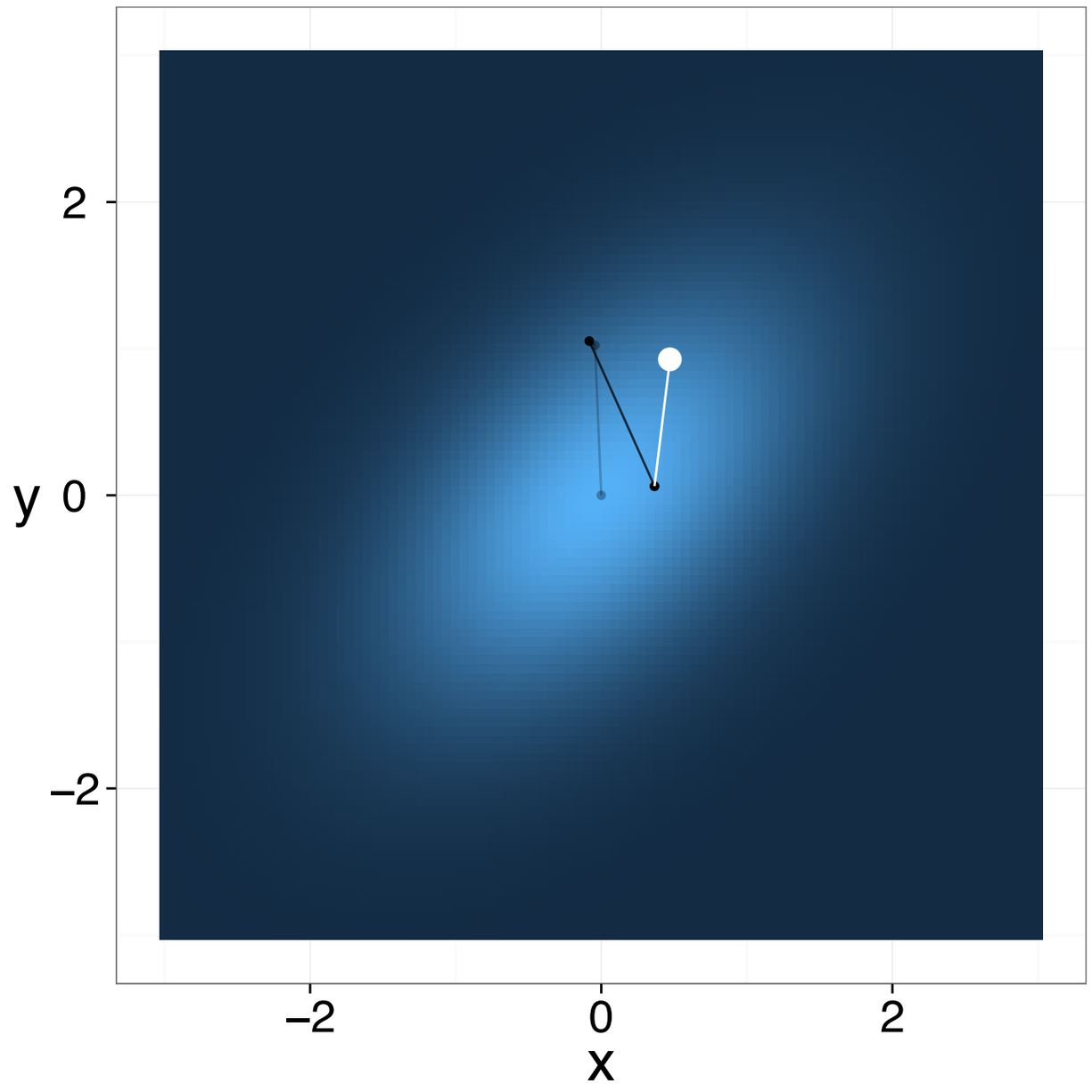
EXAMPLE: BIVARIATE GAUSSIAN



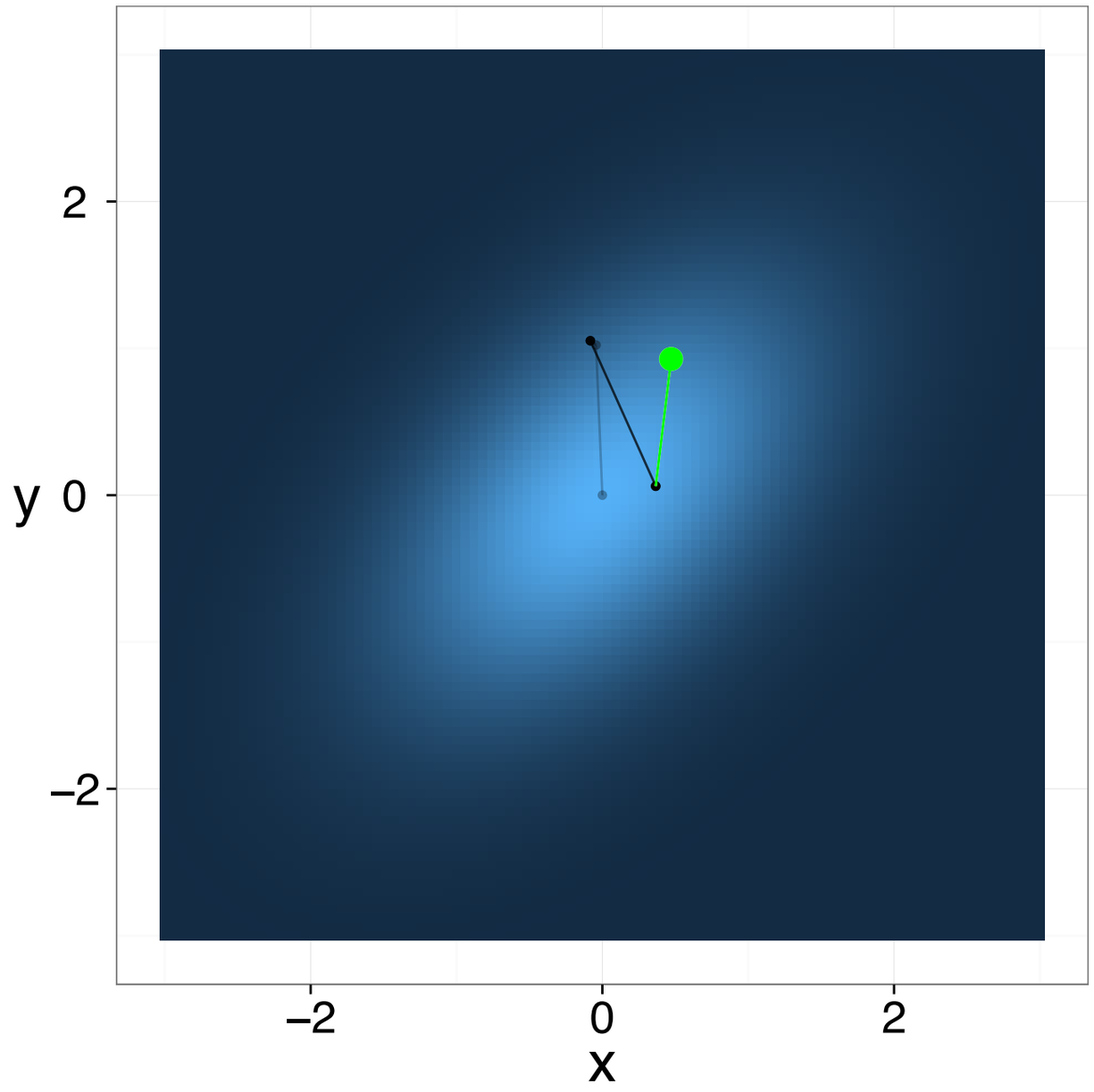
EXAMPLE: BIVARIATE GAUSSIAN



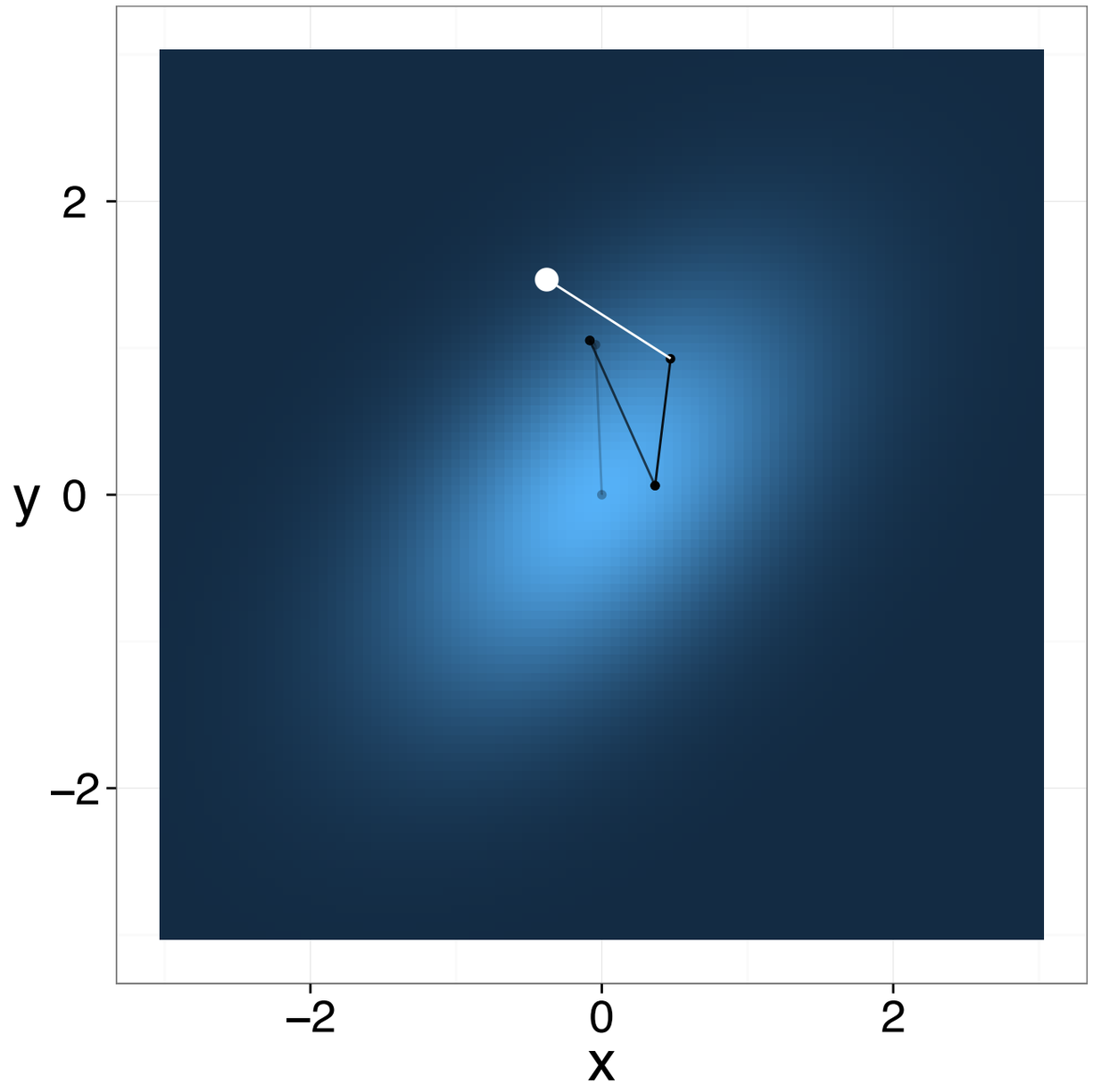
EXAMPLE: BIVARIATE GAUSSIAN



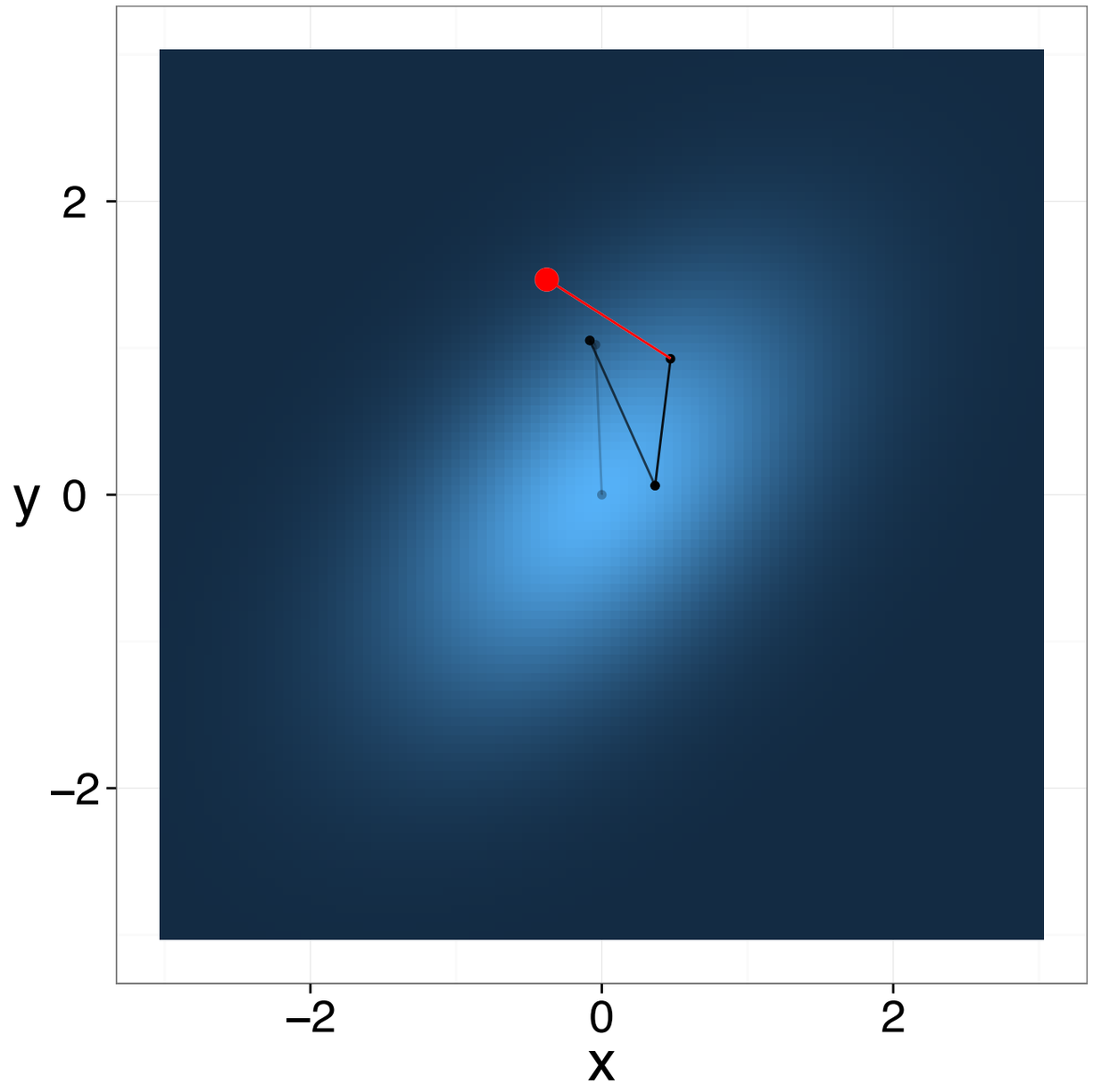
EXAMPLE: BIVARIATE GAUSSIAN



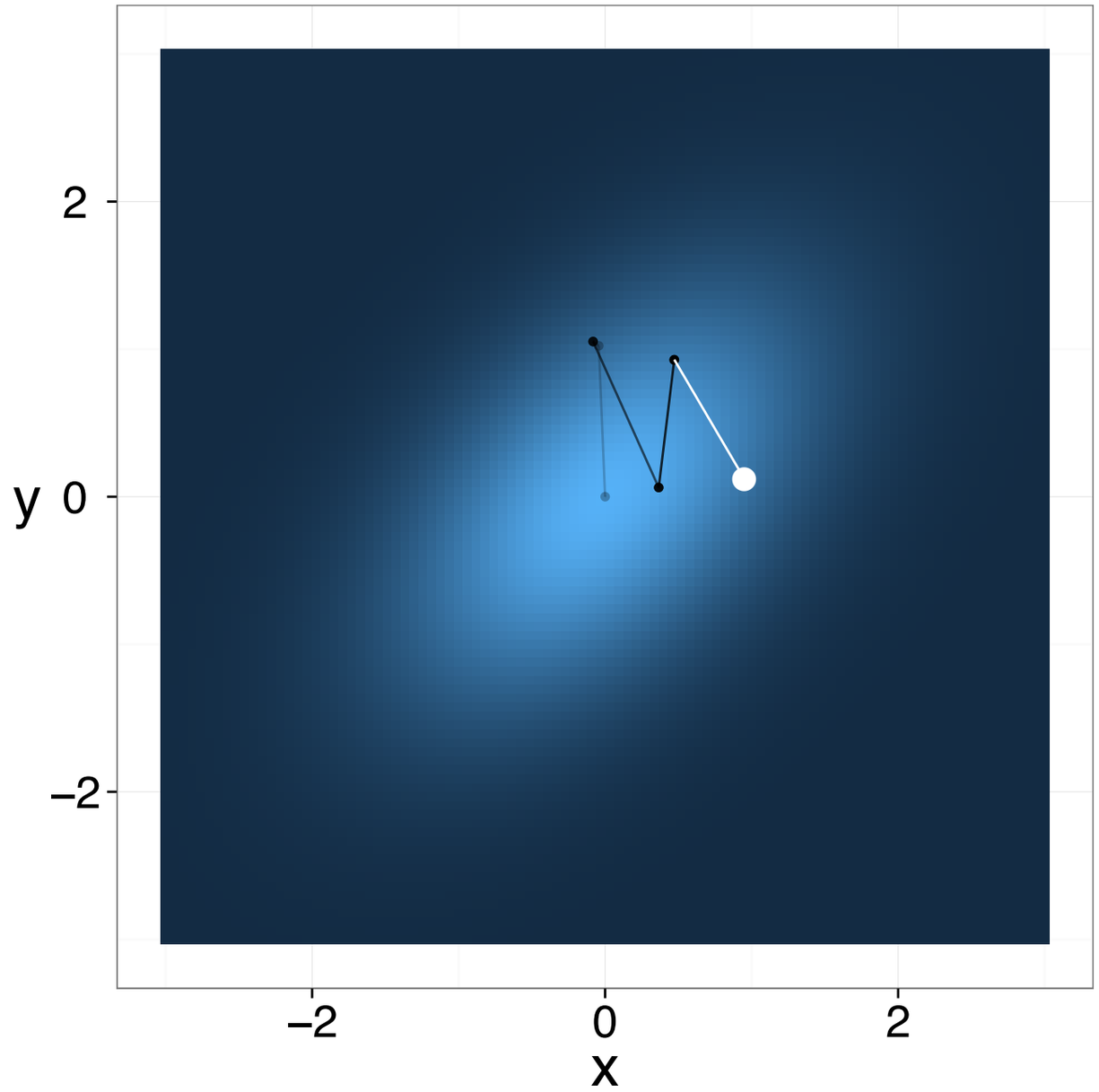
EXAMPLE: BIVARIATE GAUSSIAN



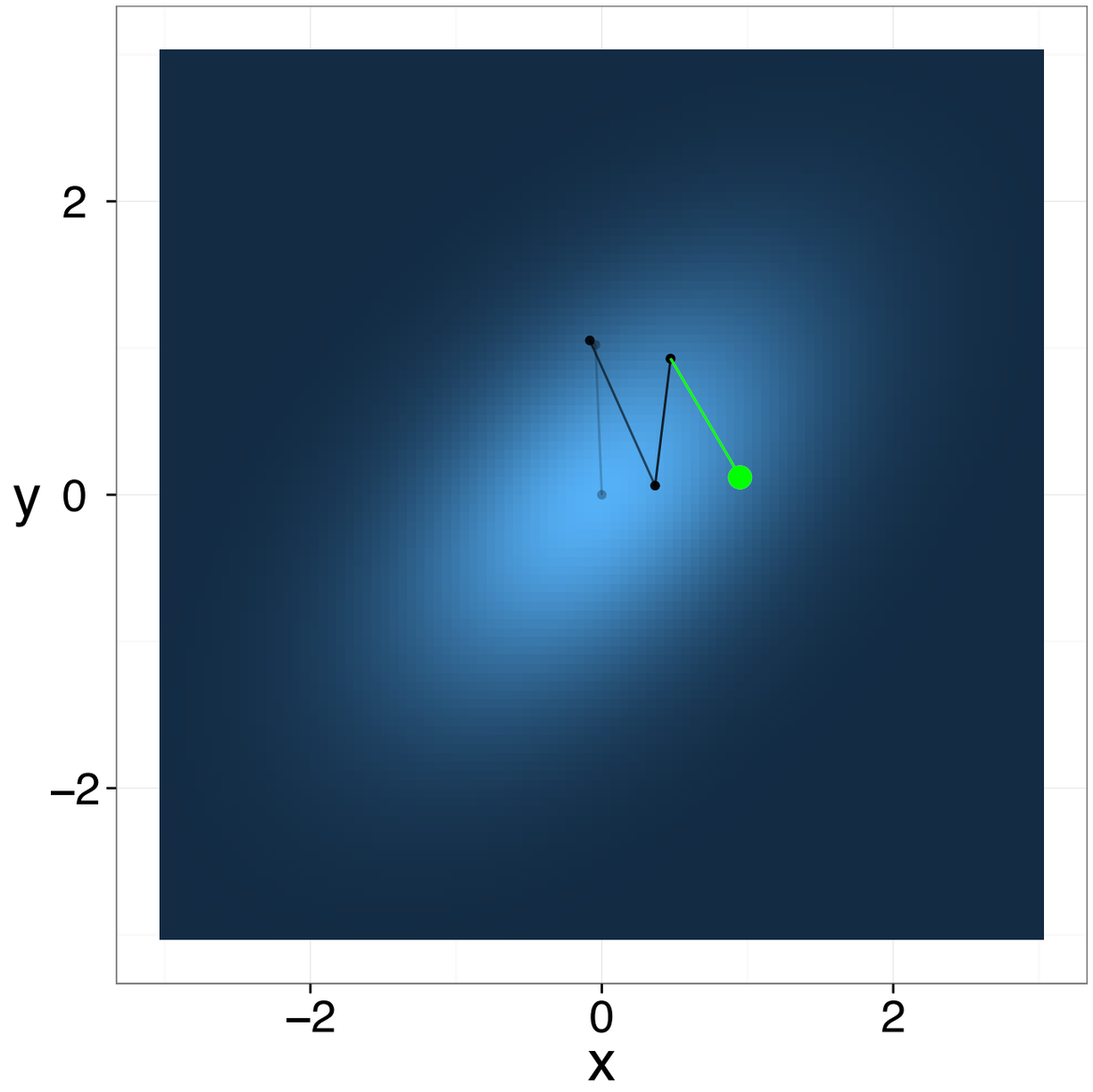
EXAMPLE: BIVARIATE GAUSSIAN



EXAMPLE: BIVARIATE GAUSSIAN



EXAMPLE: BIVARIATE GAUSSIAN



HISTORY OF METROPLIS

HISTORY OF METROPLIS

- **1953 —** The “Metroplis” algorithm was first introduced for symmetric proposals at Los Alamos

HISTORY OF METROPLIS

- **1953 —** The “Metroplis” algorithm was first introduced for symmetric proposals at Los Alamos
  - “Equation of State Calculations by Fast Computing Machines” published in the Journal of Chemical Physics

HISTORY OF METROPLIS

- **1953 —** The “Metroplis” algorithm was first introduced for symmetric proposals at Los Alamos
  - “Equation of State Calculations by Fast Computing Machines” published in the Journal of Chemical Physics
  - Developed by Arianna W. Rosenbluth, Marshall Rosenbluth, Augusta H. Teller and Edward Teller

HISTORY OF METROPLIS

- **1953 —** The “Metroplis” algorithm was first introduced for symmetric proposals at Los Alamos
  - “Equation of State Calculations by Fast Computing Machines” published in the Journal of Chemical Physics
  - Developed by Arianna W. Rosenbluth, Marshall Rosenbluth, Augusta H. Teller and Edward Teller
  - Debate over who did most of the work, but consencus is that Nicolas Metroplis played no role

HISTORY OF METROPLIS

- **1953 —** The “Metroplis” algorithm was first introduced for symmetric proposals at Los Alamos
  - “Equation of State Calculations by Fast Computing Machines” published in the Journal of Chemical Physics
  - Developed by Arianna W. Rosenbluth, Marshall Rosenbluth, Augusta H. Teller and Edward Teller
  - Debate over who did most of the work, but consencus is that Nicolas Metroplis played no role
- **1970 —** Keith W. Hastings generalised this to non-symmetric proposals at U of T

HISTORY OF METROPLIS

- **1953 —** The “Metroplis” algorithm was first introduced for symmetric proposals at Los Alamos
  - “Equation of State Calculations by Fast Computing Machines” published in the Journal of Chemical Physics
  - Developed by Arianna W. Rosenbluth, Marshall Rosenbluth, Augusta H. Teller and Edward Teller
  - Debate over who did most of the work, but consencus is that Nicolas Metroplis played no role

- **1970 —** Keith W. Hastings generalised this to non-symmetric proposals at U of T
  - “Monte Carlo sampling methods using Markov chains and their applications” published in Biometrika

HISTORY OF METROPLIS

- **1953 —** The “Metroplis” algorithm was first introduced for symmetric proposals at Los Alamos
  - “Equation of State Calculations by Fast Computing Machines” published in the Journal of Chemical Physics
  - Developed by Arianna W. Rosenbluth, Marshall Rosenbluth, Augusta H. Teller and Edward Teller
  - Debate over who did most of the work, but consencus is that Nicolas Metroplis played no role

- **1970 —** Keith W. Hastings generalised this to non-symmetric proposals at U of T
  - “Monte Carlo sampling methods using Markov chains and their applications” published in Biometrika

- One of the most important algorithms since the history of computing

HISTORY OF METROPLIS

- **1953 —** The “Metroplis” algorithm was first introduced for symmetric proposals at Los Alamos
  - “Equation of State Calculations by Fast Computing Machines” published in the Journal of Chemical Physics
  - Developed by Arianna W. Rosenbluth, Marshall Rosenbluth, Augusta H. Teller and Edward Teller
  - Debate over who did most of the work, but consencus is that Nicolas Metroplis played no role

- **1970 —** Keith W. Hastings generalised this to non-symmetric proposals at U of T
  - “Monte Carlo sampling methods using Markov chains and their applications” published in Biometrika

- One of the most important algorithms since the history of computing
  - Foundational to the entire feild computational statistics and physics

HISTORY OF METROPLIS

- **1953 —** The “Metroplis” algorithm was first introduced for symmetric proposals at Los Alamos
  - “Equation of State Calculations by Fast Computing Machines” published in the Journal of Chemical Physics
  - Developed by Arianna W. Rosenbluth, Marshall Rosenbluth, Augusta H. Teller and Edward Teller
  - Debate over who did most of the work, but consencus is that Nicolas Metroplis played no role

- **1970 —** Keith W. Hastings generalised this to non-symmetric proposals at U of T
  - “Monte Carlo sampling methods using Markov chains and their applications” published in Biometrika

- One of the most important algorithms since the history of computing
  - Foundational to the entire feild computational statistics and physics
  - Bayesian statistics would probably not be a field of study without this algorithm

KERNEL OF MH

KERNEL OF MH

- **Proposition:** The Metropolis-Hastening kernels

$$K(x, \mathrm{d}x') = \alpha(x, x')Q(x, \mathrm{d}x') + r(x)\delta_x(\mathrm{d}x')$$

KERNEL OF MH

- **Proposition:** The Metropolis-Hastening kernels

$$K(x, \mathrm{d}x') = \alpha(x, x')Q(x, \mathrm{d}x') + r(x)\delta_x(\mathrm{d}x')$$

  - Where $r(x) = 1 - \alpha(x)$ and $\alpha(x)$ the average acceptance probability:

$$\alpha(x) = \int_{\mathbb{X}} \alpha(x, y)Q(x, \mathrm{d}y)$$

KERNEL OF MH

- **Proposition:** The Metropolis-Hastening kernels

$$K(x, \mathrm{d}x') = \alpha(x, x')Q(x, \mathrm{d}x') + r(x)\delta_x(\mathrm{d}x')$$

  - Where $r(x) = 1 - \alpha(x)$ and $\alpha(x)$ the average acceptance probability:

$$\alpha(x) = \int_{\mathbb{X}} \alpha(x, y)Q(x, \mathrm{d}y)$$

- **Proof:**

KERNEL OF MH

- **Proposition:** The Metropolis-Hastening kernels

$$K(x, \mathrm{d}x') = \alpha(x, x')Q(x, \mathrm{d}x') + r(x)\delta_x(\mathrm{d}x')$$

  - Where $r(x) = 1 - \alpha(x)$ and $\alpha(x)$ the average acceptance probability:

$$\alpha(x) = \int_{\mathbb{X}} \alpha(x, y)Q(x, \mathrm{d}y)$$

- **Proof:**

$$K(x, \mathrm{d}x') = \int_{\mathbb{X}} Q(x, \mathrm{d}y)\Big[\alpha(x, y)\delta_y(\mathrm{d}x') + (1 - \alpha(x, y))\delta_x(\mathrm{d}x')\Big]$$

KERNEL OF MH

- **Proposition:** The Metropolis-Hastening kernels

$$K(x, \mathrm{d}x') = \alpha(x, x')Q(x, \mathrm{d}x') + r(x)\delta_x(\mathrm{d}x')$$

  - Where $r(x) = 1 - \alpha(x)$ and $\alpha(x)$ the average acceptance probability:

$$\alpha(x) = \int_{\mathbb{X}} \alpha(x, y)Q(x, \mathrm{d}y)$$

- **Proof:**

$$K(x, \mathrm{d}x') = \int_{\mathbb{X}} Q(x, \mathrm{d}y)\Big[\alpha(x, y)\delta_y(\mathrm{d}x') + (1 - \alpha(x, y))\delta_x(\mathrm{d}x')\Big]$$

$$= \int_{\mathbb{X}} \alpha(x, y)Q(x, \mathrm{d}y)\delta_y(\mathrm{d}x')$$

KERNEL OF MH

- **Proposition:** The Metropolis-Hastening kernels

$$K(x, \mathrm{d}x') = \alpha(x, x')Q(x, \mathrm{d}x') + r(x)\delta_x(\mathrm{d}x')$$

  - Where $r(x) = 1 - \alpha(x)$ and $\alpha(x)$ the average acceptance probability:

$$\alpha(x) = \int_{\mathbb{X}} \alpha(x, y)Q(x, \mathrm{d}y)$$

- **Proof:**

$$K(x, \mathrm{d}x') = \int_{\mathbb{X}} Q(x, \mathrm{d}y)\left[\alpha(x, y)\delta_y(\mathrm{d}x') + (1 - \alpha(x, y))\delta_x(\mathrm{d}x')\right]$$

$$= \int_{\mathbb{X}} \alpha(x, y)Q(x, \mathrm{d}y)\delta_y(\mathrm{d}x') \; + \left[\int_{\mathbb{X}} (1 - \alpha(x, y))Q(x, \mathrm{d}y)\right] \delta_x(\mathrm{d}x')$$

KERNEL OF MH

- **Proposition:** The Metropolis-Hastening kernels

$$K(x, \mathrm{d}x') = \alpha(x, x')Q(x, \mathrm{d}x') + r(x)\delta_x(\mathrm{d}x')$$

  - Where $r(x) = 1 - \alpha(x)$ and $\alpha(x)$ the average acceptance probability:

$$\alpha(x) = \int_{\mathbb{X}} \alpha(x, y)Q(x, \mathrm{d}y)$$

- **Proof:**

$$K(x, \mathrm{d}x') = \int_{\mathbb{X}} Q(x, \mathrm{d}y)\left[\alpha(x, y)\delta_y(\mathrm{d}x') + (1 - \alpha(x, y))\delta_x(\mathrm{d}x')\right]$$

$$= \int_{\mathbb{X}} \alpha(x, y)Q(x, \mathrm{d}y)\delta_y(\mathrm{d}x') + \left[\int_{\mathbb{X}} (1 - \alpha(x, y))Q(x, \mathrm{d}y)\right] \delta_x(\mathrm{d}x')$$

$$= \alpha(x, x')Q(x, \mathrm{d}x')$$

KERNEL OF MH

- **Proposition:** The Metropolis-Hastening kernels

$$K(x, \mathrm{d}x') = \alpha(x, x')Q(x, \mathrm{d}x') + r(x)\delta_x(\mathrm{d}x')$$

  - Where $r(x) = 1 - \alpha(x)$ and $\alpha(x)$ the average acceptance probability:

$$\alpha(x) = \int_{\mathbb{X}} \alpha(x, y)Q(x, \mathrm{d}y)$$

- **Proof:**

$$K(x, \mathrm{d}x') = \int_{\mathbb{X}} Q(x, \mathrm{d}y)\left[\alpha(x, y)\delta_y(\mathrm{d}x') + (1 - \alpha(x, y))\delta_x(\mathrm{d}x')\right]$$

$$= \int_{\mathbb{X}} \alpha(x, y)Q(x, \mathrm{d}y)\delta_y(\mathrm{d}x') + \left[\int_{\mathbb{X}} (1 - \alpha(x, y))Q(x, \mathrm{d}y)\right] \delta_x(\mathrm{d}x')$$

$$= \alpha(x, x')Q(x, \mathrm{d}x') + [1 - \alpha(x)]\delta_x(\mathrm{d}x')$$

REVERSIBILITY

REVERSIBILITY

- The MH kernel is π-reversible and is thus π-invariant

REVERSIBILITY

- The MH kernel is π-reversible and is thus π-invariant
- **Proof:** Recall that the kernel is

$$K(x, \mathrm{d}x') = \alpha(x, x')Q(x, \mathrm{d}x') + r(x)\delta_x(\mathrm{d}x')$$

REVERSIBILITY

- The MH kernel is π-reversible and is thus π-invariant
- **Proof:** Recall that the kernel is

$$K(x, \mathrm{d}x') = \alpha(x, x')Q(x, \mathrm{d}x') + r(x)\delta_x(\mathrm{d}x')$$

  - Therefore:

$$\pi(\mathrm{d}x)K(x, \mathrm{d}x') = \alpha(x, x')\pi(\mathrm{d}x)Q(x, \mathrm{d}x') + r(x)\pi(\mathrm{d}x)\delta_x(\mathrm{d}x')$$

REVERSIBILITY

- The MH kernel is π-reversible and is thus π-invariant
- **Proof:** Recall that the kernel is

$$K(x, \mathrm{d}x') = \alpha(x, x')Q(x, \mathrm{d}x') + r(x)\delta_x(\mathrm{d}x')$$

  - Therefore:

$$\pi(\mathrm{d}x)K(x, \mathrm{d}x') = \alpha(x, x')\pi(\mathrm{d}x)Q(x, \mathrm{d}x') + r(x)\pi(\mathrm{d}x)\delta_x(\mathrm{d}x')$$

  - The first term:

$$\alpha(x, x')\pi(\mathrm{d}x)Q(x, \mathrm{d}x') = \left[1 \wedge \frac{\pi(\mathrm{d}x')Q(x', \mathrm{d}x)}{\pi(\mathrm{d}x)Q(x, \mathrm{d}x')}\right] \pi(\mathrm{d}x)Q(x, \mathrm{d}x')$$

REVERSIBILITY

- The MH kernel is π-reversible and is thus π-invariant
- **Proof:** Recall that the kernel is

$$K(x, \mathrm{d}x') = \alpha(x, x')Q(x, \mathrm{d}x') + r(x)\delta_x(\mathrm{d}x')$$

  - Therefore:

$$\pi(\mathrm{d}x)K(x, \mathrm{d}x') = \alpha(x, x')\pi(\mathrm{d}x)Q(x, \mathrm{d}x') + r(x)\pi(\mathrm{d}x)\delta_x(\mathrm{d}x')$$

  - The first term:

$$\alpha(x, x')\pi(\mathrm{d}x)Q(x, \mathrm{d}x') = \left[1 \wedge \frac{\pi(\mathrm{d}x')Q(x', \mathrm{d}x)}{\pi(\mathrm{d}x)Q(x, \mathrm{d}x')}\right] \pi(\mathrm{d}x)Q(x, \mathrm{d}x')$$

$$= \pi(\mathrm{d}x)Q(x, \mathrm{d}x') \wedge \pi(\mathrm{d}x')Q(x', \mathrm{d}x)$$

REVERSIBILITY

- The MH kernel is π-reversible and is thus π-invariant
- **Proof:** Recall that the kernel is

$$K(x, \mathrm{d}x') = \alpha(x, x')Q(x, \mathrm{d}x') + r(x)\delta_x(\mathrm{d}x')$$

  - Therefore:

$$\pi(\mathrm{d}x)K(x, \mathrm{d}x') = \alpha(x, x')\pi(\mathrm{d}x)Q(x, \mathrm{d}x') + r(x)\pi(\mathrm{d}x)\delta_x(\mathrm{d}x')$$

  - The first term:

$$\begin{aligned}
\alpha(x, x')\pi(\mathrm{d}x)Q(x, \mathrm{d}x') &= \left[1 \wedge \frac{\pi(\mathrm{d}x')Q(x', \mathrm{d}x)}{\pi(\mathrm{d}x)Q(x, \mathrm{d}x')}\right] \pi(\mathrm{d}x)Q(x, \mathrm{d}x') \\
&= \pi(\mathrm{d}x)Q(x, \mathrm{d}x') \wedge \pi(\mathrm{d}x')Q(x', \mathrm{d}x) \\
&= \left[\frac{\pi(\mathrm{d}x)Q(x, \mathrm{d}x')}{\pi(\mathrm{d}x')Q(x', \mathrm{d}x)} \wedge 1\right] \pi(\mathrm{d}x')Q(x', \mathrm{d}x)
\end{aligned}$$

REVERSIBILITY

- The MH kernel is π-reversible and is thus π-invariant
- **Proof:** Recall that the kernel is

$$K(x, \mathrm{d}x') = \alpha(x, x')Q(x, \mathrm{d}x') + r(x)\delta_x(\mathrm{d}x')$$

  - Therefore:

$$\pi(\mathrm{d}x)K(x, \mathrm{d}x') = \alpha(x, x')\pi(\mathrm{d}x)Q(x, \mathrm{d}x') + r(x)\pi(\mathrm{d}x)\delta_x(\mathrm{d}x')$$

  - The first term:

$$\begin{aligned}
\alpha(x, x')\pi(\mathrm{d}x)Q(x, \mathrm{d}x') &= \left[1 \wedge \frac{\pi(\mathrm{d}x')Q(x', \mathrm{d}x)}{\pi(\mathrm{d}x)Q(x, \mathrm{d}x')}\right] \pi(\mathrm{d}x)Q(x, \mathrm{d}x') \\
&= \pi(\mathrm{d}x)Q(x, \mathrm{d}x') \wedge \pi(\mathrm{d}x')Q(x', \mathrm{d}x) \\
&= \left[\frac{\pi(\mathrm{d}x)Q(x, \mathrm{d}x')}{\pi(\mathrm{d}x')Q(x', \mathrm{d}x)} \wedge 1\right] \pi(\mathrm{d}x')Q(x', \mathrm{d}x) \\
&= \alpha(x', x)\pi(\mathrm{d}x')Q(x', \mathrm{d}x)
\end{aligned}$$

REVERSIBILITY

- The MH kernel is π-reversible and is thus π-invariant
- **Proof:** Recall that the kernel is

$$K(x, \mathrm{d}x') = \alpha(x, x')Q(x, \mathrm{d}x') + r(x)\delta_x(\mathrm{d}x')$$

  - Therefore:

$$\pi(\mathrm{d}x)K(x, \mathrm{d}x') = \alpha(x, x')\pi(\mathrm{d}x)Q(x, \mathrm{d}x') + r(x)\pi(\mathrm{d}x)\delta_x(\mathrm{d}x')$$

  - The first term:

$$\begin{aligned}
\alpha(x, x')\pi(\mathrm{d}x)Q(x, \mathrm{d}x') &= \left[1 \wedge \frac{\pi(\mathrm{d}x')Q(x', \mathrm{d}x)}{\pi(\mathrm{d}x)Q(x, \mathrm{d}x')}\right] \pi(\mathrm{d}x)Q(x, \mathrm{d}x') \\
&= \pi(\mathrm{d}x)Q(x, \mathrm{d}x') \wedge \pi(\mathrm{d}x')Q(x', \mathrm{d}x) \\
&= \left[\frac{\pi(\mathrm{d}x)Q(x, \mathrm{d}x')}{\pi(\mathrm{d}x')Q(x', \mathrm{d}x)} \wedge 1\right] \pi(\mathrm{d}x')Q(x', \mathrm{d}x) \\
&= \alpha(x', x)\pi(\mathrm{d}x')Q(x', \mathrm{d}x)
\end{aligned}$$

  - The second term:

$$r(x)\pi(\mathrm{d}x)\delta_x(\mathrm{d}x') = r(x')\pi(\mathrm{d}x')\delta_{x'}(\mathrm{d}x)$$

REDUCIBILITY AND PERIODICITY

REDUCIBILITY AND PERIODICITY

- The MH chain aperiodic in all reasonable circumstances

REDUCIBILITY AND PERIODICITY

- The MH chain aperiodic in all reasonable circumstances
  - The MH chain is π-periodic provided there $\alpha = 1$ and Q is periodic.

REDUCIBILITY AND PERIODICITY

- The MH chain aperiodic in all reasonable circumstances
  - The MH chain is π-periodic provided there $\alpha = 1$ and Q is periodic.
- It is not always guaranteed that MH is π-irreducible

REDUCIBILITY AND PERIODICITY

- The MH chain aperiodic in all reasonable circumstances
  - The MH chain is π-periodic provided there $\alpha = 1$ and Q is periodic.
- It is not always guaranteed that MH is π-irreducible
- **Example:** consider the following target:

$$\pi(x) = \frac{1}{2}\delta_{[0,1]}(x) + \frac{1}{2}\delta_{[2,3]}(x)$$

REDUCIBILITY AND PERIODICITY

- The MH chain aperiodic in all reasonable circumstances
  - The MH chain is π-periodic provided there $\alpha = 1$ and Q is periodic.
- It is not always guaranteed that MH is π-irreducible
- **Example:** consider the following target:

$$\pi(x) = \frac{1}{2}\delta_{[0,1]}(x) + \frac{1}{2}\delta_{[2,3]}(x)$$

- For $\delta > 0$ define the proposal $Q(x, \mathrm{d}y)$ that uniformly proposes a sample in $[x - \delta, x + \delta]$

$$Q(x, \mathrm{d}y) = \mathrm{Uniform}([x - \delta, x + \delta])$$

REDUCIBILITY AND PERIODICITY

- The MH chain aperiodic in all reasonable circumstances
  - The MH chain is π-periodic provided there $\alpha = 1$ and Q is periodic.
- It is not always guaranteed that MH is π-irreducible
- **Example:** consider the following target:

$$\pi(x) = \frac{1}{2}\delta_{[0,1]}(x) + \frac{1}{2}\delta_{[2,3]}(x)$$

- For $\delta > 0$ define the proposal $Q(x, \mathrm{d}y)$ that uniformly proposes a sample in $[x - \delta, x + \delta]$

$$Q(x, \mathrm{d}y) = \mathrm{Uniform}([x - \delta, x + \delta])$$

- Under what conditions is the MH kernel K_δ π-irreducible?

REDUCIBILITY AND PERIODICITY

- The MH chain aperiodic in all reasonable circumstances
  - The MH chain is π-periodic provided there $\alpha = 1$ and Q is periodic.
- It is not always guaranteed that MH is π-irreducible
- **Example:** consider the following target:

$$\pi(x) = \frac{1}{2}\delta_{[0,1]}(x) + \frac{1}{2}\delta_{[2,3]}(x)$$

- For $\delta > 0$ define the proposal $Q(x, \mathrm{d}y)$ that uniformly proposes a sample in $[x - \delta, x + \delta]$

$$Q(x, \mathrm{d}y) = \mathrm{Uniform}([x - \delta, x + \delta])$$

- Under what conditions is the MH kernel K_δ π-irreducible?
  - K_δ is π-irreducible if and only if $\delta > 1$

REDUCIBILITY AND PERIODICITY

- The MH chain aperiodic in all reasonable circumstances
  - The MH chain is π-periodic provided there $\alpha = 1$ and Q is periodic.
- It is not always guaranteed that MH is π-irreducible
- **Example:** consider the following target:

$$\pi(x) = \frac{1}{2}\delta_{[0,1]}(x) + \frac{1}{2}\delta_{[2,3]}(x)$$

- For $\delta > 0$ define the proposal $Q(x, \mathrm{d}y)$ that uniformly proposes a sample in $[x - \delta, x + \delta]$

$$Q(x, \mathrm{d}y) = \mathrm{Uniform}([x - \delta, x + \delta])$$

- Under what conditions is the MH kernel K_δ π-irreducible?
  - K_δ is π-irreducible if and only if $\delta > 1$
  - If $\delta \leq 1$ then the chain stays in $[0,1]$ or $[2,3]$ depending on the initial condition

REDUCIBILITY AND PERIODICITY

REDUCIBILITY AND PERIODICITY

- **Proposition:** If $Q\left(x, \mathrm{d}y\right) = q(x, y)\mathrm{d}y$ and $q(x, y) > 0$ for all $x, x' \in \mathrm{supp}(\pi)$, then the MH chain is π-irreducible in fact, every state can be reached in a single step (strongly irreducible).
  - See less strict condition in Roberts & Rosenthal (2004)

REDUCIBILITY AND PERIODICITY

- **Proposition:** If $Q\left(x, \mathrm{d}y\right) = q(x, y)\mathrm{d}y$ and $q(x, y) > 0$ for all $x, x' \in \mathrm{supp}(\pi)$, then the MH chain is π-irreducible in fact, every state can be reached in a single step (strongly irreducible).
  - See less strict condition in Roberts & Rosenthal (2004)
- **Proposition:** If the MH chain is π-irreducible then it is also Harris recurrent.
  - See Tierney (1994)

REDUCIBILITY AND PERIODICITY

- **Proposition:** If $Q\left(x, \mathrm{d}y\right) = q(x, y)\mathrm{d}y$ and $q(x, y) > 0$ for all $x, x' \in \mathrm{supp}(\pi)$, then the MH chain is π-irreducible in fact, every state can be reached in a single step (strongly irreducible).
  - See less strict condition in Roberts & Rosenthal (2004)
- **Proposition:** If the MH chain is π-irreducible then it is also Harris recurrent.
  - See Tierney (1994)

- **Proposition (LLN for MH):** If the MH chain X_t is π-irreducible, then for $f : \mathbb{X} \to \mathbb{R}$

$$\pi[f] = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} f(X_t)$$

  - Where the convergence is almost sure and holds for any initial condition $X_0 \sim \mu$

REDUCIBILITY AND PERIODICITY

- **Proposition:** If $Q\left(x, \mathrm{d}y\right) = q(x, y)\mathrm{d}y$ and $q(x, y) > 0$ for all $x, x' \in \mathrm{supp}(\pi)$, then the MH chain is π-irreducible in fact, every state can be reached in a single step (strongly irreducible).
  - See less strict condition in Roberts & Rosenthal (2004)
- **Proposition:** If the MH chain is π-irreducible then it is also Harris recurrent.
  - See Tierney (1994)
- **Proposition (LLN for MH):** If the MH chain X_t is π-irreducible, then for $f : \mathbb{X} \to \mathbb{R}$

$$\pi[f] = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} f(X_t)$$

  - Where the convergence is almost sure and holds for any initial condition $X_0 \sim \mu$
- In general, MH is not guaranteed to be geometrically ergodic or satisfy the CLT

INDEPENDENT METROPOLIS-HASTINGS (IMH)

INDEPENDENT METROPOLIS-HASTINGS (IMH)

- Suppose $\eta \in \mathscr{P}(\mathbb{X})$ with density $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$

INDEPENDENT METROPOLIS-HASTINGS (IMH)

- Suppose $\eta \in \mathscr{P}(\mathbb{X})$ with density $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$
  - 1. We can generate samples from η

INDEPENDENT METROPOLIS-HASTINGS (IMH)

- Suppose $\eta \in \mathscr{P}(\mathbb{X})$ with density $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$
  - 1. We can generate samples from η
  - 2. η is mututally absolutely continuous with respect to π

INDEPENDENT METROPOLIS-HASTINGS (IMH)

- Suppose $\eta \in \mathscr{P}(\mathbb{X})$ with density $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$
  - 1. We can generate samples from η
  - 2. η is mututally absolutely continuous with respect to π
  - 3. We can evaluate $w : \mathbb{X} \to \mathbb{R}_+$ the un-normalised likelihood ratio

$$w(x) = \frac{\gamma(x)}{\eta(x)} = Z\frac{\mathrm{d}\pi}{\mathrm{d}\eta}(x)$$

INDEPENDENT METROPOLIS-HASTINGS (IMH)

- Suppose $\eta \in \mathscr{P}(\mathbb{X})$ with density $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$
  - 1. We can generate samples from η
  - 2. η is mututally absolutely continuous with respect to π
  - 3. We can evaluate $w : \mathbb{X} \to \mathbb{R}_+$ the un-normalised likelihood ratio

$$w(x) = \frac{\gamma(x)}{\eta(x)} = Z\frac{\mathrm{d}\pi}{\mathrm{d}\eta}(x)$$

- Suppose $Q(x, \mathrm{d}x') = \eta(\mathrm{d}x')$ is the independent kernel independent of x

INDEPENDENT METROPOLIS-HASTINGS (IMH)

- Suppose $\eta \in \mathscr{P}(\mathbb{X})$ with density $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$
  - 1. We can generate samples from η
  - 2. η is mututally absolutely continuous with respect to π
  - 3. We can evaluate $w : \mathbb{X} \to \mathbb{R}_+$ the un-normalised likelihood ratio

$$w(x) = \frac{\gamma(x)}{\eta(x)} = Z\frac{\mathrm{d}\pi}{\mathrm{d}\eta}(x)$$

- Suppose $Q(x, \mathrm{d}x') = \eta(\mathrm{d}x')$ is the independent kernel independent of x
- We can evaluate the acceptance ratio in terms of w

$$A(x, y) = \frac{\pi(\mathrm{d}y)Q(y, \mathrm{d}x)}{\pi(\mathrm{d}x)Q(x, \mathrm{d}y)}$$

INDEPENDENT METROPOLIS-HASTINGS (IMH)

- Suppose $\eta \in \mathscr{P}(\mathbb{X})$ with density $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$
  - 1. We can generate samples from η
  - 2. η is mututally absolutely continuous with respect to π
  - 3. We can evaluate $w : \mathbb{X} \to \mathbb{R}_+$ the un-normalised likelihood ratio

$$w(x) = \frac{\gamma(x)}{\eta(x)} = Z\frac{\mathrm{d}\pi}{\mathrm{d}\eta}(x)$$

- Suppose $Q(x, \mathrm{d}x') = \eta(\mathrm{d}x')$ is the independent kernel independent of x
- We can evaluate the acceptance ratio in terms of w

$$A(x, y) = \frac{\pi(\mathrm{d}y)Q(y, \mathrm{d}x)}{\pi(\mathrm{d}x)Q(x, \mathrm{d}y)} = \frac{\pi(y)\eta(x)}{\pi(x)\eta(y)}$$

INDEPENDENT METROPOLIS-HASTINGS (IMH)

- Suppose $\eta \in \mathscr{P}(\mathbb{X})$ with density $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$
  - 1. We can generate samples from η
  - 2. η is mututally absolutely continuous with respect to π
  - 3. We can evaluate $w : \mathbb{X} \to \mathbb{R}_+$ the un-normalised likelihood ratio

$$w(x) = \frac{\gamma(x)}{\eta(x)} = Z\frac{\mathrm{d}\pi}{\mathrm{d}\eta}(x)$$

- Suppose $Q(x, \mathrm{d}x') = \eta(\mathrm{d}x')$ is the independent kernel independent of x
- We can evaluate the acceptance ratio in terms of w

$$A(x, y) = \frac{\pi(\mathrm{d}y)Q(y, \mathrm{d}x)}{\pi(\mathrm{d}x)Q(x, \mathrm{d}y)} = \frac{\pi(y)\eta(x)}{\pi(x)\eta(y)} = \frac{w(y)}{w(x)}$$

INDEPENDENT METROPOLIS-HASTINGS (IMH)

- Suppose $\eta \in \mathscr{P}(\mathbb{X})$ with density $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$
  - 1. We can generate samples from η
  - 2. η is mututally absolutely continuous with respect to π
  - 3. We can evaluate $w : \mathbb{X} \to \mathbb{R}_+$ the un-normalised likelihood ratio

$$w(x) = \frac{\gamma(x)}{\eta(x)} = Z\frac{\mathrm{d}\pi}{\mathrm{d}\eta}(x)$$

- Suppose $Q(x, \mathrm{d}x') = \eta(\mathrm{d}x')$ is the independent kernel independent of x
- We can evaluate the acceptance ratio in terms of w

$$A(x, y) = \frac{\pi(\mathrm{d}y)Q(y, \mathrm{d}x)}{\pi(\mathrm{d}x)Q(x, \mathrm{d}y)} = \frac{\pi(y)\eta(x)}{\pi(x)\eta(y)} = \frac{w(y)}{w(x)}$$

- Therefore Q is a valid kernel for MH kernel is irriducible and satisfies the LLN

INDEPENDENT METROPOLIS-HASTINGS (IMH)

INDEPENDENT METROPOLIS-HASTINGS (IMH)

- The choice of reference is critical!

INDEPENDENT METROPOLIS-HASTINGS (IMH)

- The choice of reference is critical!
- Recall the average acceptance probabiity:

INDEPENDENT METROPOLIS-HASTINGS (IMH)

- The choice of reference is critical!
- Recall the average acceptance probabiity:

$$\alpha(x) = \int_{\mathbb{X}} \alpha(x, x')Q(x, \mathrm{d}x') = \int \left[1 \wedge \frac{\pi(x')\eta(x)}{\pi(x)\eta(x')}\right] \eta(x')\mathrm{d}x'$$

INDEPENDENT METROPOLIS-HASTINGS (IMH)

- The choice of reference is critical!
- Recall the average acceptance probabiity:

$$\alpha(x) = \int_{\mathbb{X}} \alpha(x, x')Q(x, \mathrm{d}x') = \int \left[1 \wedge \frac{\pi(x')\eta(x)}{\pi(x)\eta(x')}\right] \eta(x')\mathrm{d}x'$$

  - At stationarity the avergage acceptance satisfies,

$$\pi[\alpha] = \int \alpha(x)\pi(x)\mathrm{d}x$$

INDEPENDENT METROPOLIS-HASTINGS (IMH)

- The choice of reference is critical!
- Recall the average acceptance probabiity:

$$\alpha(x) = \int_{\mathbb{X}} \alpha(x, x')Q(x, \mathrm{d}x') = \int \left[1 \wedge \frac{\pi(x')\eta(x)}{\pi(x)\eta(x')}\right] \eta(x')\mathrm{d}x'$$

  - At stationarity the avergage acceptance satisfies,

$$\pi[\alpha] = \int \alpha(x)\pi(x)\mathrm{d}x \quad = \int_{\mathbb{X}^2} \pi(x)\eta(x') \wedge \pi(x')\eta(x)\mathrm{d}x\mathrm{d}x'$$

INDEPENDENT METROPOLIS-HASTINGS (IMH)

- The choice of reference is critical!
- Recall the average acceptance probabiity:

$$\alpha(x) = \int_{\mathbb{X}} \alpha(x, x')Q(x, \mathrm{d}x') = \int \left[1 \wedge \frac{\pi(x')\eta(x)}{\pi(x)\eta(x')}\right] \eta(x')\mathrm{d}x'$$

  - At stationarity the avergage acceptance satisfies,

$$\pi[\alpha] = \int \alpha(x)\pi(x)\mathrm{d}x \quad = \int_{\mathbb{X}^2} \pi(x)\eta(x') \wedge \pi(x')\eta(x)\mathrm{d}x\mathrm{d}x' \quad = 1 - \|\eta \otimes \pi - \pi \otimes \eta\|_{\mathrm{TV}}$$

INDEPENDENT METROPOLIS-HASTINGS (IMH)

- The choice of reference is critical!
- Recall the average acceptance probabiity:

$$\alpha(x) = \int_{\mathbb{X}} \alpha(x, x')Q(x, \mathrm{d}x') = \int \left[1 \wedge \frac{\pi(x')\eta(x)}{\pi(x)\eta(x')}\right] \eta(x')\mathrm{d}x'$$

  - At stationarity the avergage acceptance satisfies,

$$\pi[\alpha] = \int \alpha(x)\pi(x)\mathrm{d}x \quad = \int_{\mathbb{X}^2} \pi(x)\eta(x') \wedge \pi(x')\eta(x)\mathrm{d}x\mathrm{d}x' \quad = 1 - \|\eta \otimes \pi - \pi \otimes \eta\|_{\mathrm{TV}}$$

  - The rejection rate is controlled by the overap between η and π

$$\|\eta \otimes \pi - \pi \otimes \eta\|_{\mathrm{TV}} \leq 2\|\eta - \pi\|_{\mathrm{TV}}$$

INDEPENDENT METROPOLIS-HASTINGS (IMH)

- The choice of reference is critical!
- Recall the average acceptance probabiity:

$$\alpha(x) = \int_{\mathbb{X}} \alpha(x, x')Q(x, \mathrm{d}x') = \int \left[1 \wedge \frac{\pi(x')\eta(x)}{\pi(x)\eta(x')}\right] \eta(x')\mathrm{d}x'$$

  - At stationarity the avergage acceptance satisfies,

$$\pi[\alpha] = \int \alpha(x)\pi(x)\mathrm{d}x \quad = \int_{\mathbb{X}^2} \pi(x)\eta(x') \wedge \pi(x')\eta(x)\mathrm{d}x\mathrm{d}x' \quad = 1 - \|\eta \otimes \pi - \pi \otimes \eta\|_{\mathrm{TV}}$$

    - The rejection rate is controlled by the overap between η and π

$$\|\eta \otimes \pi - \pi \otimes \eta\|_{\mathrm{TV}} \leq 2\|\eta - \pi\|_{\mathrm{TV}}$$

- The IMH kernel is uniformly ergodic if and only if $w(x) \leq M$ for some $M < \infty$

INDEPENDENT METROPOLIS-HASTINGS (IMH)

- The choice of reference is critical!
- Recall the average acceptance probabiity:

$$\alpha(x) = \int_{\mathbb{X}} \alpha(x, x')Q(x, \mathrm{d}x') = \int \left[1 \wedge \frac{\pi(x')\eta(x)}{\pi(x)\eta(x')}\right] \eta(x')\mathrm{d}x'$$

  - At stationarity the avergage acceptance satisfies,

$$\pi[\alpha] = \int \alpha(x)\pi(x)\mathrm{d}x \quad = \int_{\mathbb{X}^2} \pi(x)\eta(x') \wedge \pi(x')\eta(x)\mathrm{d}x\mathrm{d}x' \quad = 1 - \|\eta \otimes \pi - \pi \otimes \eta\|_{\mathrm{TV}}$$

    - The rejection rate is controlled by the overap between η and π

$$\|\eta \otimes \pi - \pi \otimes \eta\|_{\mathrm{TV}} \leq 2\|\eta - \pi\|_{\mathrm{TV}}$$

- The IMH kernel is uniformly ergodic if and only if $w(x) \leq M$ for some $M < \infty$
- The acceptance probability at stationarity is at least $1/M$
  - See Lemma 7.9 of Robert & Casella

INDEPENDENT METROPOLIS-HASTINGS (IMH)

- The choice of reference is critical!
- Recall the average acceptance probabiity:

$$\alpha(x) = \int_{\mathbb{X}} \alpha(x, x')Q(x, \mathrm{d}x') = \int \left[1 \wedge \frac{\pi(x')\eta(x)}{\pi(x)\eta(x')}\right] \eta(x')\mathrm{d}x'$$

  - At stationarity the avergage acceptance satisfies,

$$\pi[\alpha] = \int \alpha(x)\pi(x)\mathrm{d}x \quad = \int_{\mathbb{X}^2} \pi(x)\eta(x') \wedge \pi(x')\eta(x)\mathrm{d}x\mathrm{d}x' \quad = 1 - \|\eta \otimes \pi - \pi \otimes \eta\|_{\mathrm{TV}}$$

    - The rejection rate is controlled by the overap between η and π

$$\|\eta \otimes \pi - \pi \otimes \eta\|_{\mathrm{TV}} \leq 2\|\eta - \pi\|_{\mathrm{TV}}$$

- The IMH kernel is uniformly ergodic if and only if $w(x) \leq M$ for some $M < \infty$
- The acceptance probability at stationarity is at least $1/M$
  - See Lemma 7.9 of Robert & Casella
- IMH is not very useful on it's own but we will see it is a powerful building block

RANDOM WALK METROPOLIS (RWM)

RANDOM WALK METROPOLIS (RWM)

- Suppose we have $\mathbb{X} = \mathbb{R}^d$ is an open subset containing the support of $\boldsymbol{\pi}$

RANDOM WALK METROPOLIS (RWM)

- Suppose we have $\mathbb{X} = \mathbb{R}^d$ is an open subset containing the support of π
- Let $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$ be a symmetric probability distribution that we can sample $\epsilon \sim \eta$

RANDOM WALK METROPOLIS (RWM)

- Suppose we have $\mathbb{X} = \mathbb{R}^d$ is an open subset containing the support of π
- Let $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$ be a symmetric probability distribution that we can sample $\epsilon \sim \eta$
  - For example η is a mean-zero Gaussian or student-t distribution with covariance Σ

RANDOM WALK METROPOLIS (RWM)

- Suppose we have $\mathbb{X} = \mathbb{R}^d$ is an open subset containing the support of π
- Let $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$ be a symmetric probability distribution that we can sample $\epsilon \sim \eta$
  - For example η is a mean-zero Gaussian or student-t distribution with covariance Σ
- Define the random walk proposal $Q(x, \mathrm{d}y) = \eta(x - y)\mathrm{d}y = \mathcal{N}(\mathrm{d}y; x, \Sigma)$

RANDOM WALK METROPOLIS (RWM)

- Suppose we have $\mathbb{X} = \mathbb{R}^d$ is an open subset containing the support of π
- Let $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$ be a symmetric probability distribution that we can sample $\epsilon \sim \eta$
  - For example η is a mean-zero Gaussian or student-t distribution with covariance Σ
- Define the random walk proposal $Q(x, \mathrm{d}y) = \eta(x - y)\mathrm{d}y = \mathcal{N}(\mathrm{d}y; x, \Sigma)$
  - Given $x \in \mathbb{X}$ we can sample $Y \sim Q(x, \mathrm{d}y)$ since

$$X' = x + z, \quad z \sim \eta = N(0,\Sigma)$$

RANDOM WALK METROPOLIS (RWM)

- Suppose we have $\mathbb{X} = \mathbb{R}^d$ is an open subset containing the support of π
- Let $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$ be a symmetric probability distribution that we can sample $\epsilon \sim \eta$
  - For example η is a mean-zero Gaussian or student-t distribution with covariance Σ
- Define the random walk proposal $Q(x, \mathrm{d}y) = \eta(x - y)\mathrm{d}y = \mathcal{N}(\mathrm{d}y; x, \Sigma)$
  - Given $x \in \mathbb{X}$ we can sample $Y \sim Q(x, \mathrm{d}y)$ since

$$X' = x + z, \quad z \sim \eta = N(0, \Sigma)$$

- Recall that since Q is symmetric, we can evaluate the acceptance ratio

$$A(x, x') = \frac{\pi(x')}{\pi(x)} = \frac{\gamma(x')}{\gamma(x)}$$

RANDOM WALK METROPOLIS (RWM)

- Suppose we have $\mathbb{X} = \mathbb{R}^d$ is an open subset containing the support of π
- Let $\eta(\mathrm{d}x) = \eta(x)\mathrm{d}x$ be a symmetric probability distribution that we can sample $\epsilon \sim \eta$
  - For example η is a mean-zero Gaussian or student-t distribution with covariance Σ
- Define the random walk proposal $Q(x, \mathrm{d}y) = \eta(x - y)\mathrm{d}y = \mathcal{N}(\mathrm{d}y; x, \Sigma)$
  - Given $x \in \mathbb{X}$ we can sample $Y \sim Q(x, \mathrm{d}y)$ since

$$X' = x + z, \quad z \sim \eta = N(0, \Sigma)$$

  - Recall that since Q is symmetric, we can evaluate the acceptance ratio

$$A(x, x') = \frac{\pi(x')}{\pi(x)} = \frac{\gamma(x')}{\gamma(x)}$$

- See demo:

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

CHOOSING A GOOD PROPOSAL

CHOOSING A GOOD PROPOSAL

- The goal of the proposal is to design a Markov chain is small auto-correlations

CHOOSING A GOOD PROPOSAL

- The goal of the proposal is to design a Markov chain is small auto-correlations
- There are two sources of correllation:

CHOOSING A GOOD PROPOSAL

- The goal of the proposal is to design a Markov chain is small auto-correlations
- There are two sources of correllation:
  - Between the current state X_{t-1} and the proposed state $Y \sim Q(x, \mathrm{d}y)$,

CHOOSING A GOOD PROPOSAL

- The goal of the proposal is to design a Markov chain is small auto-correlations
- There are two sources of correllation:
  - Between the current state X_{t-1} and the proposed state $Y \sim Q(x, \mathrm{d}y)$,
  - Correlation induced if $X_t = X_{t-1}$ if the proposal is rejected.

CHOOSING A GOOD PROPOSAL

- The goal of the proposal is to design a Markov chain is small auto-correlations
- There are two sources of correllation:
  - Between the current state X_{t-1} and the proposed state $Y \sim Q(x, \mathrm{d}y)$,
  - Correlation induced if $X_t = X_{t-1}$ if the proposal is rejected.
- There is a compromise between

CHOOSING A GOOD PROPOSAL

- The goal of the proposal is to design a Markov chain is small auto-correlations
- There are two sources of correllation:
  - Between the current state X_{t-1} and the proposed state $Y \sim Q(x, \mathrm{d}y)$,
  - Correlation induced if $X_t = X_{t-1}$ if the proposal is rejected.
- There is a compromise between
  - proposing small conservative moves that lead to high acceptance

CHOOSING A GOOD PROPOSAL

- The goal of the proposal is to design a Markov chain is small auto-correlations
- There are two sources of correllation:
  - Between the current state X_{t-1} and the proposed state $Y \sim Q(x, \mathrm{d}y)$,
  - Correlation induced if $X_t = X_{t-1}$ if the proposal is rejected.
- There is a compromise between
  - proposing small conservative moves that lead to high acceptance
  - proposing large aggressive moves that lead to low acceptance

CHOOSING A GOOD PROPOSAL

- The goal of the proposal is to design a Markov chain is small auto-correlations
- There are two sources of correllation:
  - Between the current state X_{t-1} and the proposed state $Y \sim Q(x, \mathrm{d}y)$,
  - Correlation induced if $X_t = X_{t-1}$ if the proposal is rejected.
- There is a compromise between
  - proposing small conservative moves that lead to high acceptance
  - proposing large aggressive moves that lead to low acceptance
- For multivariate distributions: covariance of the proposal should reflect the covariance structure of the target.

CHOOSING A GOOD PROPOSAL

- The goal of the proposal is to design a Markov chain is small auto-correlations
- There are two sources of correllation:
  - Between the current state X_{t-1} and the proposed state $Y \sim Q(x, \mathrm{d}y)$,
  - Correlation induced if $X_t = X_{t-1}$ if the proposal is rejected.
- There is a compromise between
  - proposing small conservative moves that lead to high acceptance
  - proposing large aggressive moves that lead to low acceptance
- For multivariate distributions: covariance of the proposal should reflect the covariance structure of the target.
- What about multi-modal or multi-scale targets?

EXAMPLE: BIVARIATE GAUSSIAN

- Suppose the target distributions over $\mathbb{R}^2$

$$\pi(x) = \mathcal{N}\left(x; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$$

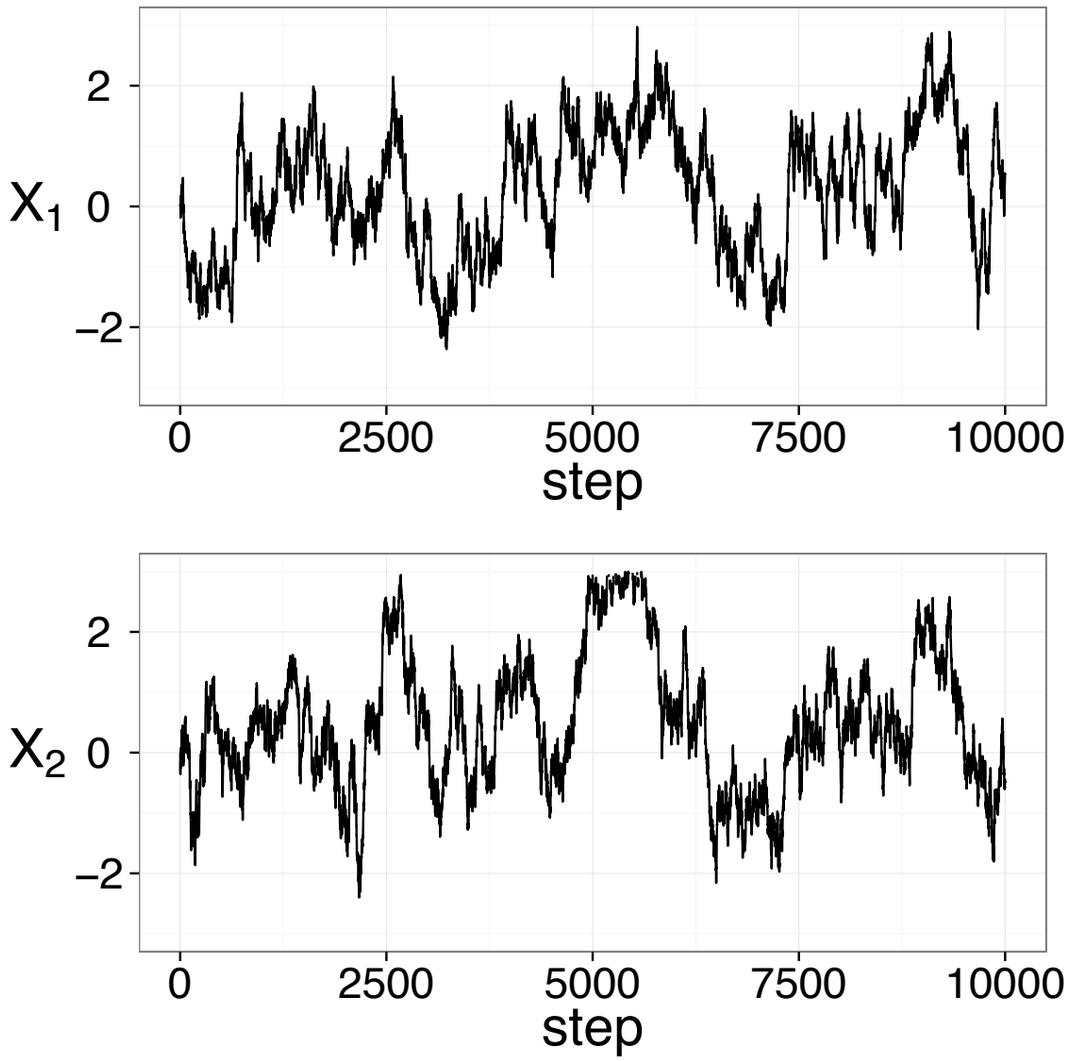
- Suppose we use a RWM proposal with covariance $\sigma^2 I$

$$\eta(\varepsilon) = \mathcal{N}\left(\varepsilon; 0, \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right).$$

- What is the optimal choice of σ?

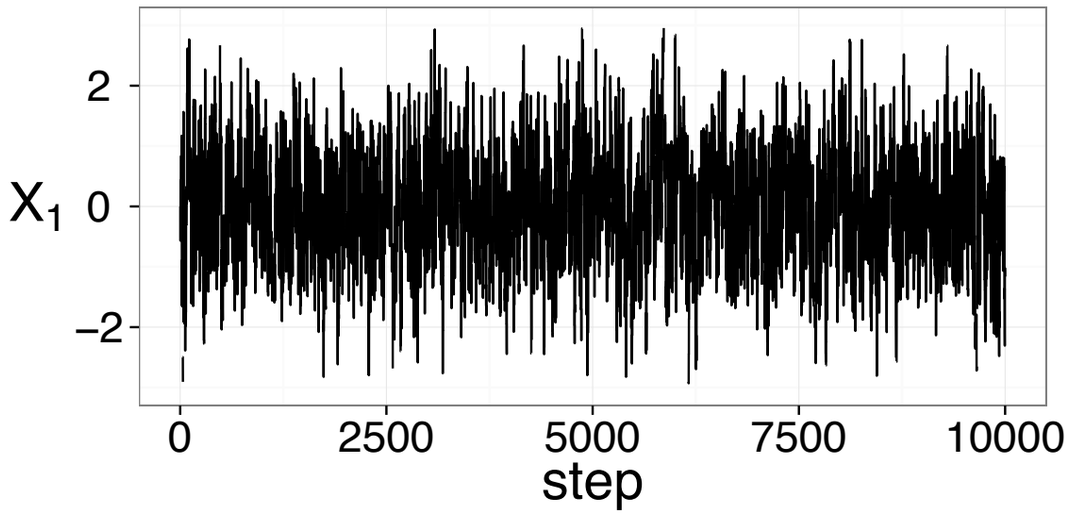
RANDOM WALK METROPOLIS

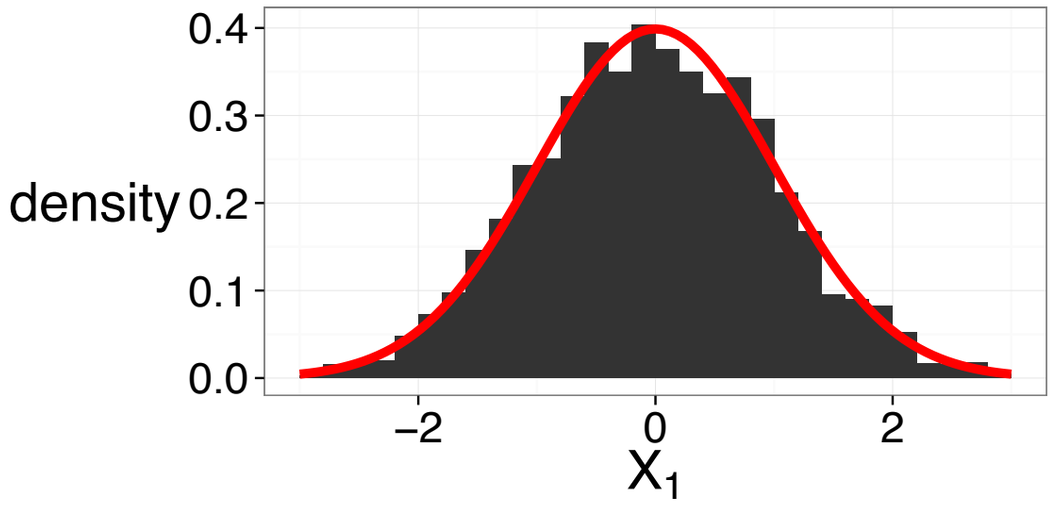
- RMW on a bivariate Gaussian target with $\sigma = 0.1$ and acceptance rate $\alpha \approx 0.94$

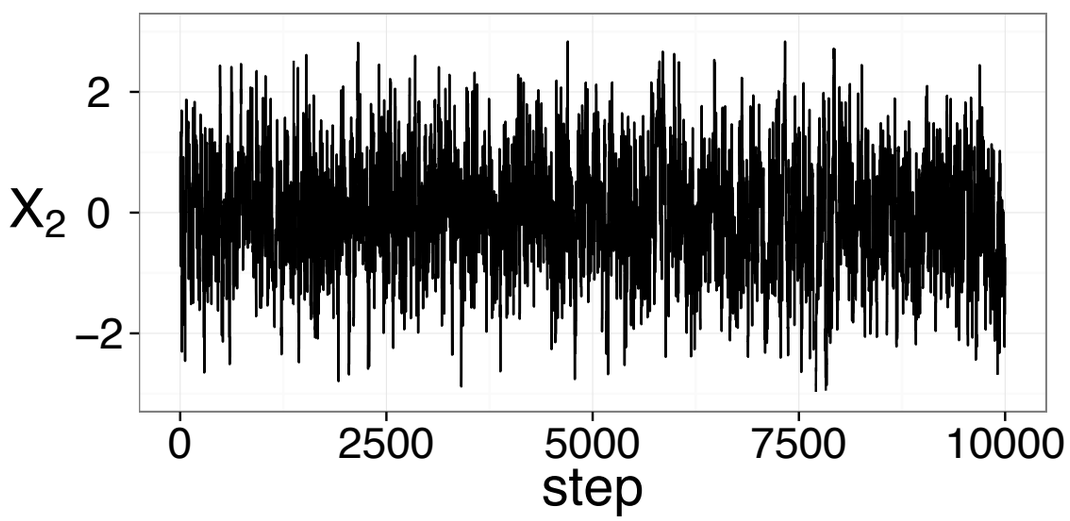


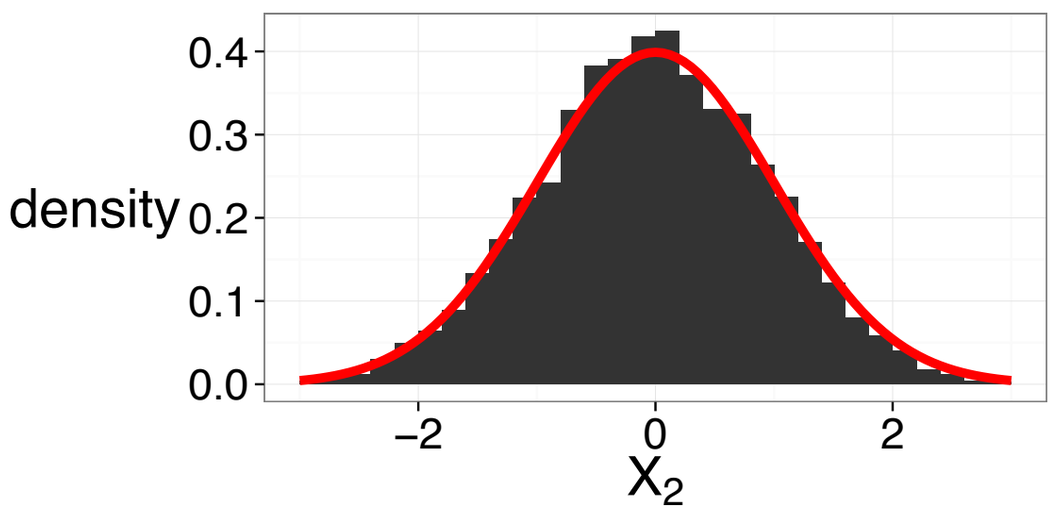
RANDOM WALK METROPOLIS

- RMW on a bivariate Gaussian target with $\sigma = 1$ and acceptance rate $\alpha \approx 0.52$



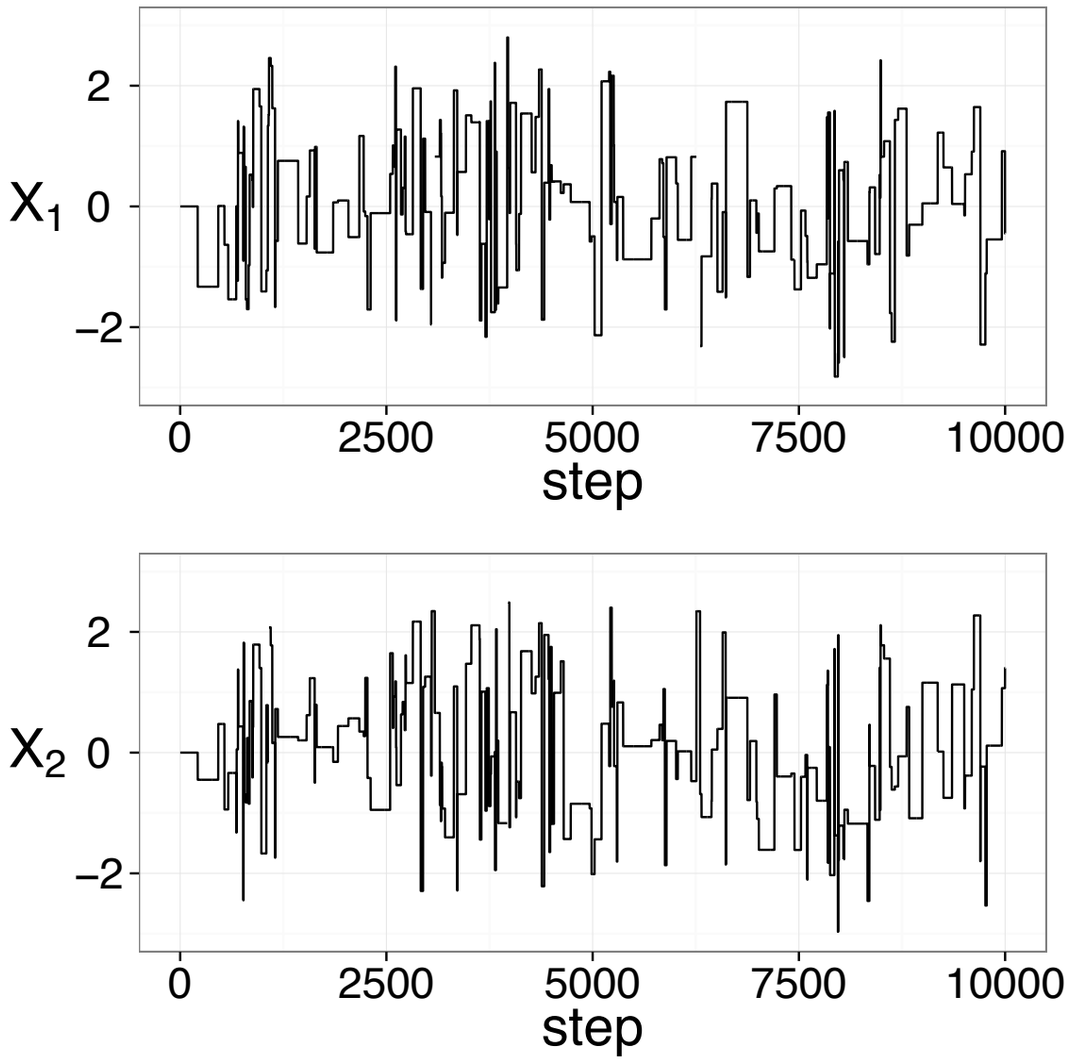


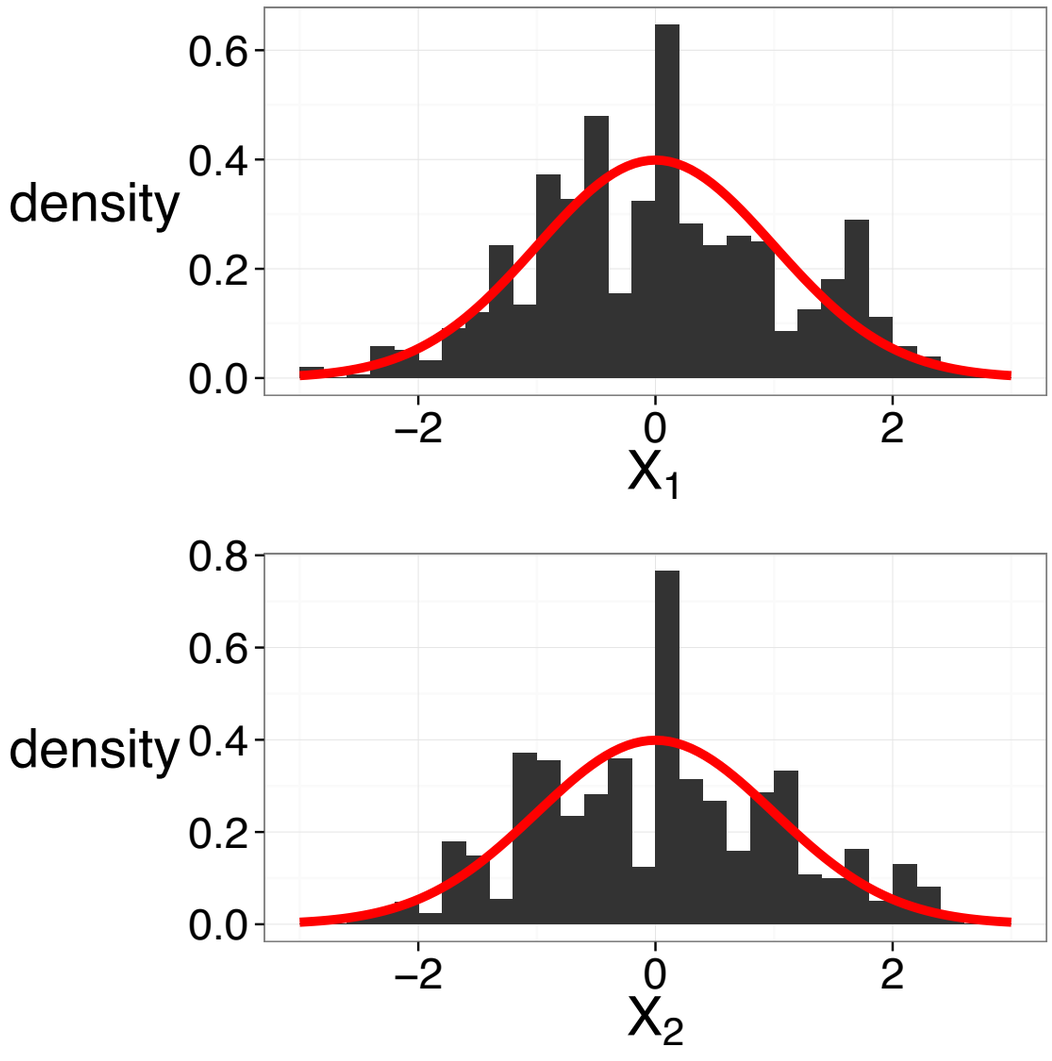




RANDOM WALK METROPOLIS

- RMW on a bivariate Gaussian target with $\sigma = 10$ and acceptance rate $\alpha \approx 0.015$





CONCENTRATION OF MEASURE

CONCENTRATION OF MEASURE

- Suppose that $z \sim N(0, I_d)$ is a d-dimensional standard normal random vector.

CONCENTRATION OF MEASURE

- Suppose that $z \sim N(0, I_d)$ is a d-dimensional standard normal random vector.
- For all $\delta \geq 0$ the Euclidean norm of $\|z\|$ satisfies

$$\mathbb{P}(|\|Z\| - \sqrt{d}\,| \geq \delta) \leq C \exp\left(-\frac{\delta^2}{C}\right)$$

  - Where C is an absolute constant independent of d

CONCENTRATION OF MEASURE

- Suppose that $z \sim N(0, I_d)$ is a d-dimensional standard normal random vector.
- For all $\delta \geq 0$ the Euclidean norm of $\|z\|$ satisfies

$$\mathbb{P}(|\|Z\| - \sqrt{d}| \geq \delta) \leq C \exp\left(-\frac{\delta^2}{C}\right)$$

  - Where C is an absolute constant independent of d

- This means that with high probability, $\|Z\| = \sqrt{d} + O(1)$ as $d \to \infty$

CONCENTRATION OF MEASURE

- Suppose that $z \sim N(0, I_d)$ is a d-dimensional standard normal random vector.
- For all $\delta \geq 0$ the Euclidean norm of $\|z\|$ satisfies

$$\mathbb{P}(|\|Z\| - \sqrt{d}| \geq \delta) \leq C \exp\left(-\frac{\delta^2}{C}\right)$$

  - Where C is an absolute constant independent of d

- This means that with high probability, $\|Z\| = \sqrt{d} + O(1)$ as $d \to \infty$
  - i.e. most of the probability is concentrated in a thin layer around the sphere of radius $\sqrt{d}$

CONCENTRATION OF MEASURE

- Suppose that $z \sim N(0, I_d)$ is a d-dimensional standard normal random vector.
- For all $\delta \geq 0$ the Euclidean norm of $\|z\|$ satisfies

$$\mathbb{P}(|\|Z\| - \sqrt{d}| \geq \delta) \leq C \exp\left(-\frac{\delta^2}{C}\right)$$

  - Where C is an absolute constant independent of d
- This means that with high probability, $\|Z\| = \sqrt{d} + O(1)$ as $d \to \infty$
  - i.e. most of the probability is concentrated in a thin layer around the sphere of radius $\sqrt{d}$
- More generally in high dimensions, mass tends to concentrate on a submanifold

CONCENTRATION OF MEASURE

- Suppose that $z \sim N(0, I_d)$ is a d-dimensional standard normal random vector.
- For all $\delta \geq 0$ the Euclidean norm of $\|z\|$ satisfies

$$\mathbb{P}(|\|Z\| - \sqrt{d}| \geq \delta) \leq C \exp\left(-\frac{\delta^2}{C}\right)$$

  - Where C is an absolute constant independent of d
- This means that with high probability, $\|Z\| = \sqrt{d} + O(1)$ as $d \to \infty$
  - i.e. most of the probability is concentrated in a thin layer around the sphere of radius $\sqrt{d}$
- More generally in high dimensions, mass tends to concentrate on a submanifold
- Most directions of proposals lead to rejection

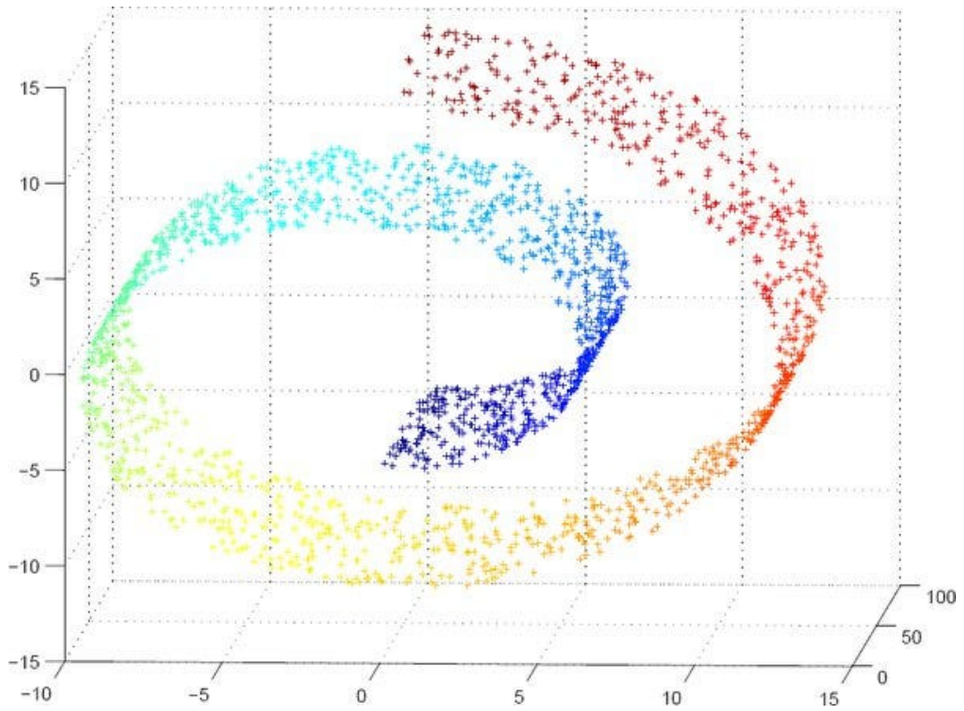
CONCENTRATION OF MEASURE

- Suppose that $z \sim N(0, I_d)$ is a d-dimensional standard normal random vector.
- For all $\delta \geq 0$ the Euclidean norm of $\|z\|$ satisfies

$$\mathbb{P}(|\|Z\| - \sqrt{d}| \geq \delta) \leq C \exp\left(-\frac{\delta^2}{C}\right)$$

  - Where C is an absolute constant independent of d
- This means that with high probability, $\|Z\| = \sqrt{d} + O(1)$ as $d \to \infty$
  - i.e. most of the probability is concentrated in a thin layer around the sphere of radius $\sqrt{d}$

- More generally in high dimensions, mass tends to concentrate on a submanifold
- Most directions of proposals lead to rejection



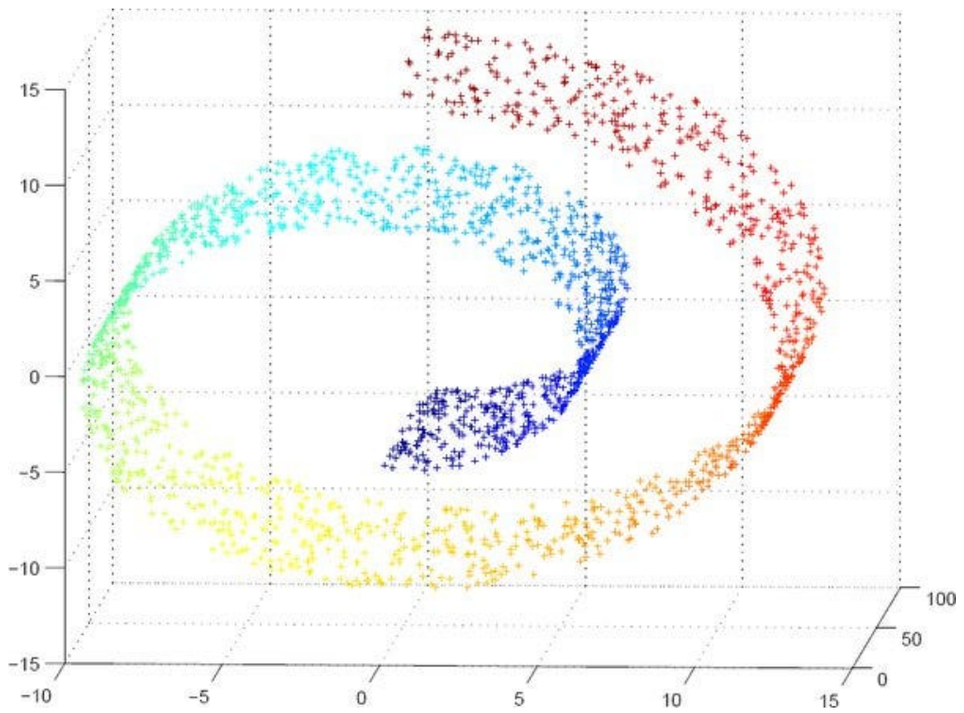
CONCENTRATION OF MEASURE

- Suppose that $z \sim N(0, I_d)$ is a d-dimensional standard normal random vector.
- For all $\delta \geq 0$ the Euclidean norm of $\|z\|$ satisfies

$$\mathbb{P}(|\|Z\| - \sqrt{d}\,| \geq \delta) \leq C \exp\left(-\frac{\delta^2}{C}\right)$$

  - Where C is an absolute constant independent of d
- This means that with high probability, $\|Z\| = \sqrt{d} + O(1)$ as $d \to \infty$
  - i.e. most of the probability is concentrated in a thin layer around the sphere of radius $\sqrt{d}$

- More generally in high dimensions, mass tends to concentrate on a submanifold
- Most directions of proposals lead to rejection
- Given x want to propose states $Y \sim Q(x, \mathrm{d}y)$ that follow the geometry of the manifold



FIRST ORDER PROPOSAL

FIRST ORDER PROPOSAL

- Recall the RWM proposal proposes Y , according to the noise $z \sim \eta$

$$Y = x + z$$

FIRST ORDER PROPOSAL

- Recall the RWM proposal proposes Y , according to the noise $z \sim \eta$

$$Y = x + z$$

  - Most directions are going to rejected proposals.

FIRST ORDER PROPOSAL

- Recall the RWM proposal proposes Y , according to the noise $z \sim \eta$

$$Y = x + z$$

  - Most directions are going to rejected proposals.
  - Only uses pointwise evaluation of the target density

FIRST ORDER PROPOSAL

- Recall the RWM proposal proposes Y , according to the noise $z \sim \eta$

$$Y = x + z$$

  - Most directions are going to rejected proposals.
  - Only uses pointwise evaluation of the target density

- Gien $\epsilon > 0$ can refine the proposal to propogate sample towards locally high probability regions:

$$Y = x + \epsilon \nabla \log \pi(x) + \sqrt{2\epsilon} z, \quad z \sim \eta$$

FIRST ORDER PROPOSAL

- Recall the RWM proposal proposes Y , according to the noise $z \sim \eta$

$$Y = x + z$$

  - Most directions are going to rejected proposals.
  - Only uses pointwise evaluation of the target density

- Gien $\epsilon > 0$ can refine the proposal to propogate sample towards locally high probability regions:

$$Y = x + \epsilon \nabla \log \pi(x) + \sqrt{2\epsilon} z, \quad z \sim \eta$$

  - Requires evaluatiing the (Stein) score $\nabla \log \pi(x)$ identifying the direction of high probability

FIRST ORDER PROPOSAL

- Recall the RWM proposal proposes Y , according to the noise $z \sim \eta$

$$Y = x + z$$

  - Most directions are going to rejected proposals.
  - Only uses pointwise evaluation of the target density

- Gien $\epsilon > 0$ can refine the proposal to propogate sample towards locally high probability regions:

$$Y = x + \epsilon \nabla \log \pi(x) + \sqrt{2\epsilon} z, \quad z \sim \eta$$

  - Requires evaluatiing the (Stein) score $\nabla \log \pi(x)$ identifying the direction of high probability
  - Analagous to noisy gradient descen

FIRST ORDER PROPOSAL

- Recall the RWM proposal proposes Y , according to the noise $z \sim \eta$

$$Y = x + z$$

  - Most directions are going to rejected proposals.
  - Only uses pointwise evaluation of the target density

- Gien $\epsilon > 0$ can refine the proposal to propogate sample towards locally high probability regions:

$$Y = x + \epsilon \nabla \log \pi(x) + \sqrt{2\epsilon} z, \quad z \sim \eta$$

  - Requires evaluatiing the (Stein) score $\nabla \log \pi(x)$ identifying the direction of high probability
  - Analagous to noisy gradient descen

- If $\eta \sim \mathcal{N}(0,\Sigma)$ and $\epsilon > 0$, then, $Y \sim Q(x, \mathrm{d}y)$ where

$$Q(x, \mathrm{d}y) = \mathcal{N}(x + \epsilon \nabla \log \pi(x), 2\epsilon\Sigma)$$

METROPOLIS ADJUSTED LANGEVIN ALGORITHM (MALA)

METROPOLIS ADJUSTED LANGEVIN ALGORITHM (MALA)

- If $\eta \sim \mathcal{N}(0,\Sigma)$ and $\epsilon > 0$, then, $Y \sim Q(x, \mathrm{d}y)$ where

$$Y = x + \epsilon \nabla \log \pi(x) + \sqrt{2\epsilon} z, \quad z \sim \eta$$

METROPOLIS ADJUSTED LANGEVIN ALGORITHM (MALA)

- If $\eta \sim \mathcal{N}(0,\Sigma)$ and $\epsilon > 0$, then, $Y \sim Q(x, \mathrm{d}y)$ where

$$Y = x + \epsilon \nabla \log \pi(x) + \sqrt{2\epsilon} z, \quad z \sim \eta$$

- The proposal Y is a discretisation of the overdamped Langevin SDE with step size ϵ

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

METROPOLIS ADJUSTED LANGEVIN ALGORITHM (MALA)

- If $\eta \sim \mathcal{N}(0,\Sigma)$ and $\epsilon > 0$, then, $Y \sim Q(x, \mathrm{d}y)$ where

$$Y = x + \epsilon \nabla \log \pi(x) + \sqrt{2\epsilon} z, \quad z \sim \eta$$

- The proposal Y is a discretisation of the overdamped Langevin SDE with step size ϵ

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

  - Which is stationary with respect to π, i.e. $X_0 \sim \pi$ then the $X_\tau \sim \pi$

METROPOLIS ADJUSTED LANGEVIN ALGORITHM (MALA)

- If $\eta \sim \mathcal{N}(0,\Sigma)$ and $\epsilon > 0$, then, $Y \sim Q(x, \mathrm{d}y)$ where

$$Y = x + \epsilon \nabla \log \pi(x) + \sqrt{2\epsilon} z, \quad z \sim \eta$$

- The proposal Y is a discretisation of the overdamped Langevin SDE with step size ϵ

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

  - Which is stationary with respect to π, i.e. $X_0 \sim \pi$ then the $X_\tau \sim \pi$

- The accept/reject steps acts correction to the discretization error of Langevin

METROPOLIS ADJUSTED LANGEVIN ALGORITHM (MALA)

- If $\eta \sim \mathcal{N}(0,\Sigma)$ and $\epsilon > 0$, then, $Y \sim Q(x, \mathrm{d}y)$ where

$$Y = x + \epsilon \nabla \log \pi(x) + \sqrt{2\epsilon} z, \quad z \sim \eta$$

- The proposal Y is a discretisation of the overdamped Langevin SDE with step size ϵ

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

  - Which is stationary with respect to π, i.e. $X_0 \sim \pi$ then the $X_\tau \sim \pi$

- The accept/reject steps acts correction to the discretization error of Langevin

- Without MH-correction is referred to the unadjusted Langevin algorithm (ULA) which is biased

METROPOLIS ADJUSTED LANGEVIN ALGORITHM (MALA)

- If $\eta \sim \mathcal{N}(0,\Sigma)$ and $\epsilon > 0$, then, $Y \sim Q(x, \mathrm{d}y)$ where

$$Y = x + \epsilon \nabla \log \pi(x) + \sqrt{2\epsilon} z, \quad z \sim \eta$$

- The proposal Y is a discretisation of the overdamped Langevin SDE with step size ϵ

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

  - Which is stationary with respect to π, i.e. $X_0 \sim \pi$ then the $X_\tau \sim \pi$

- The accept/reject steps acts correction to the discretization error of Langevin

- Without MH-correction is referred to the unadjusted Langevin algorithm (ULA) which is biased

- See demo:

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

MALA VS RMW

MALA VS RMW

- **RMW:** $Y = x + z, \qquad z \sim \mathcal{N}(0,\Sigma)$

MALA VS RMW

- **RMW:** $Y = x + z, \qquad z \sim \mathcal{N}(0, \Sigma)$
  - Can probe target pointwise

MALA VS RMW

- **RMW:** $Y = x + z, \qquad z \sim \mathcal{N}(0, \Sigma)$
  - Can probe target pointwise
  - Only requires tuning noise Σ

MALA VS RMW

- **RMW:** $Y = x + z, \qquad z \sim \mathcal{N}(0, \Sigma)$
  - Can probe target pointwise
  - Only requires tuning noise Σ
  - Behaves like a random walk in all directions

MALA VS RMW

- **RMW:** $Y = x + z, \qquad z \sim \mathcal{N}(0,\Sigma)$
  - Can probe target pointwise
  - Only requires tuning noise Σ
  - Behaves like a random walk in all directions
- **MALA:** $Y = x + \epsilon \nabla \log \pi(x) + \sqrt{2\epsilon} z, \qquad z \sim \mathcal{N}(0,\Sigma)$

MALA VS RMW

- **RMW:** $Y = x + z, \qquad z \sim \mathcal{N}(0, \Sigma)$
  - Can probe target pointwise
  - Only requires tuning noise Σ
  - Behaves like a random walk in all directions
- **MALA:** $Y = x + \epsilon \nabla \log \pi(x) + \sqrt{2\epsilon} z, \qquad z \sim \mathcal{N}(0, \Sigma)$
  - Gradient probes the local geometry of the target and identifies optimal directions to traverse

MALA VS RMW

- **RMW:** $Y = x + z, \qquad z \sim \mathcal{N}(0, \Sigma)$
  - Can probe target pointwise
  - Only requires tuning noise Σ
  - Behaves like a random walk in all directions

- **MALA:** $Y = x + \epsilon \nabla \log \pi(x) + \sqrt{2\epsilon} z, \qquad z \sim \mathcal{N}(0, \Sigma)$
  - Gradient probes the local geometry of the target and identifies optimal directions to traverse
  - Additionaly required tuning the step size ϵ

MALA VS RMW

- **RMW:** $Y = x + z, \qquad z \sim \mathcal{N}(0, \Sigma)$
  - Can probe target pointwise
  - Only requires tuning noise Σ
  - Behaves like a random walk in all directions
- **MALA:** $Y = x + \epsilon \nabla \log \pi(x) + \sqrt{2\epsilon} z, \qquad z \sim \mathcal{N}(0, \Sigma)$
  - Gradient probes the local geometry of the target and identifies optimal directions to traverse
  - Additionaly required tuning the step size ϵ
  - Still behaves like a random walk in the high-probability directions

MALA VS RMW

- **RMW:** $Y = x + z, \qquad z \sim \mathcal{N}(0, \Sigma)$
  - Can probe target pointwise
  - Only requires tuning noise Σ
  - Behaves like a random walk in all directions
- **MALA:** $Y = x + \epsilon \nabla \log \pi(x) + \sqrt{2\epsilon} z, \qquad z \sim \mathcal{N}(0, \Sigma)$
  - Gradient probes the local geometry of the target and identifies optimal directions to traverse
  - Additionaly required tuning the step size ϵ
  - Still behaves like a random walk in the high-probability directions
- Both MALA and RMW lack memory to maintain a persistent direction of travel

MALA VS RMW

- **RMW:** $Y = x + z, \qquad z \sim \mathcal{N}(0, \Sigma)$
  - Can probe target pointwise
  - Only requires tuning noise Σ
  - Behaves like a random walk in all directions
- **MALA:** $Y = x + \epsilon \nabla \log \pi(x) + \sqrt{2\epsilon} z, \qquad z \sim \mathcal{N}(0, \Sigma)$
  - Gradient probes the local geometry of the target and identifies optimal directions to traverse
  - Additionaly required tuning the step size ϵ
  - Still behaves like a random walk in the high-probability directions
- Both MALA and RMW lack memory to maintain a persistent direction of travel
- We will fix this by adding a momentum

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 $\Delta q_t = \epsilon p_t$
  - ϵ 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 $\Delta q_t = \epsilon p_t$
  - ϵ 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 $\Delta q_t = \epsilon p_t$
  - ϵ is the step size a momentum p_t, proposing direction of travel
  - p_t is the momentum indicating direction of travel

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

PROPOSALS AS MOMENTUM

PROPOSALS AS MOMENTUM

- **Example**: RWM

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

$$p \sim \mathcal{N}(0, M)$$

  - M is the covariance matrix

PROPOSALS AS MOMENTUM

- **Example**: RWM

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

$$p \sim \mathcal{N}(0, M)$$

  - M is the covariance matrix

- **Example:** MALA

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

$$p \sim \mathcal{N}(-\nabla U(q), M)$$

PROPOSALS AS MOMENTUM

- **Example**: RWM

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

$$p \sim \mathcal{N}(0, M)$$

  - M is the covariance matrix

- **Example:** MALA

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

$$p \sim \mathcal{N}(-\nabla U(q), M)$$

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

PROPOSALS AS MOMENTUM

- **Example**: RWM

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

$$p \sim \mathcal{N}(0, M)$$

  - M is the covariance matrix

- **Example:** MALA

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

$$p \sim \mathcal{N}(-\nabla U(q), M)$$

- MALA has momentum pushing particle locally towards a lower potential energy
  - It is hard to maintain momentum since the proposal is local

PROPOSALS AS MOMENTUM

- **Example**: RWM

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

$$p \sim \mathcal{N}(0, M)$$

  - M is the covariance matrix

- **Example:** MALA

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

$$p \sim \mathcal{N}(-\nabla U(q), M)$$

- MALA has momentum pushing particle locally towards a lower potential energy
  - It is hard to maintain momentum since the proposal is local
  - It forgets the momentum (proposal) from the last iteration

PROPOSALS AS MOMENTUM

- **Example**: RWM

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

$$p \sim \mathcal{N}(0, M)$$

  - M is the covariance matrix

- **Example:** MALA

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

$$p \sim \mathcal{N}(-\nabla U(q), M)$$

- MALA has momentum pushing particle locally towards a lower potential energy
  - It is hard to maintain momentum since the proposal is local
  - It forgets the momentum (proposal) from the last iteration
- Can fix this if we track position and momentum. Let $z = (q, p) \in \mathbb{R}^{2d}$ be the phase space.