Advanced Simulation - Lecture 7

Patrick Rebeschini

February 5th, 2018

Outline

- Given a target $\pi(x) = \pi(x_1, x_2, ..., x_d)$, Gibbs sampling works by sampling from $\pi_{X_j|X_{-j}}(x_j|x_{-j})$ for $j = 1, ..., d$.
- Sampling exactly from one of these conditionals might be a hard problem itself.
- Even if it is possible, the Gibbs sampler might converge slowly if components are highly correlated.
- If the components are not highly correlated then Gibbs sampling performs well, even when $d \to \infty$, e.g. with an error increasing "only" polynomially with d.
- Metropolis–Hastings algorithm (1953, 1970) is a more general algorithm that can bypass these problems.
- Additionally Gibbs can be recovered as a special case.

Metropolis–Hastings algorithm

- Target distribution on $\mathbb{X} = \mathbb{R}^d$ of density $\pi(x)$.
- Proposal distribution: for any $x, x' \in \mathbb{X}$, we have $q(x' \mid x) \geq 0$ and $\int_{\mathbb{X}} q(x' \mid x)\, dx' = 1$.
- Starting with $X^{(1)}$, for $t = 2, 3, ...$

1. Sample $X^{\star} \sim q\left(\cdot \mid X^{(t-1)} \right)$.
2. Compute

$$\alpha\left(X^{\star} \mid X^{(t-1)} \right) = \min\left(1, \frac{\pi(X^{\star})\, q\left(X^{(t-1)} \middle| X^{\star} \right)}{\pi\left(X^{(t-1)}\right) q\left(X^{\star} \mid X^{(t-1)} \right)} \right).$$

3. Sample $U \sim \mathcal{U}_{[0,1]}$. If $U \leq \alpha\left(X^{\star} \mid X^{(t-1)} \right)$, set $X^{(t)} = X^{\star}$, otherwise set $X^{(t)} = X^{(t-1)}$.

Metropolis–Hastings algorithm

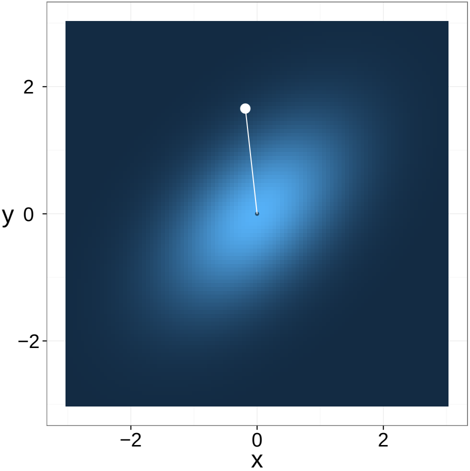

Figure: Metropolis–Hastings on a bivariate Gaussian target.

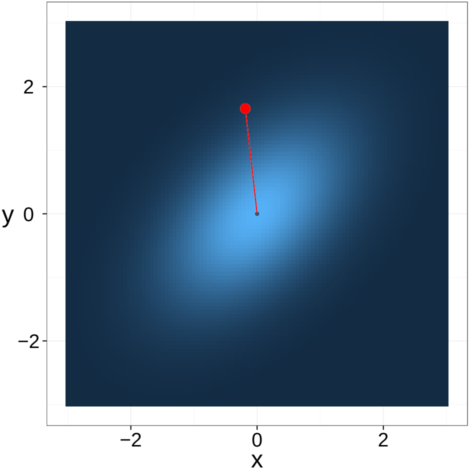

Figure: Metropolis–Hastings on a bivariate Gaussian target.

Metropolis–Hastings algorithm

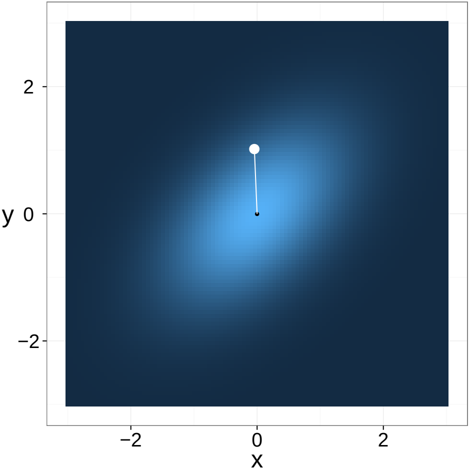

Figure: Metropolis–Hastings on a bivariate Gaussian target.

Metropolis–Hastings algorithm

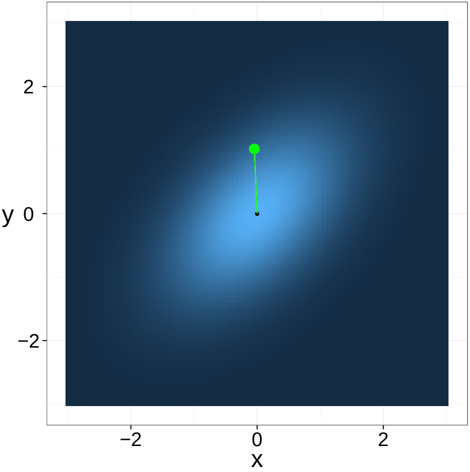

Figure: Metropolis–Hastings on a bivariate Gaussian target.

Metropolis–Hastings algorithm

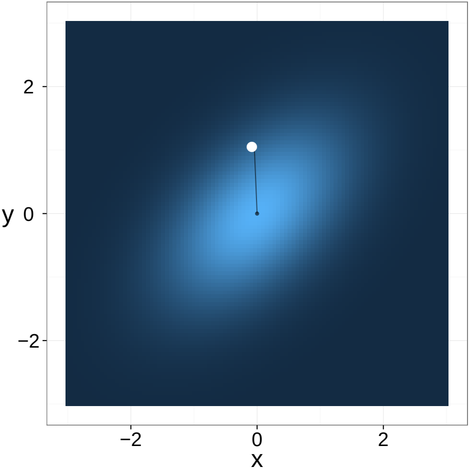

Figure: Metropolis–Hastings on a bivariate Gaussian target.

Metropolis–Hastings algorithm

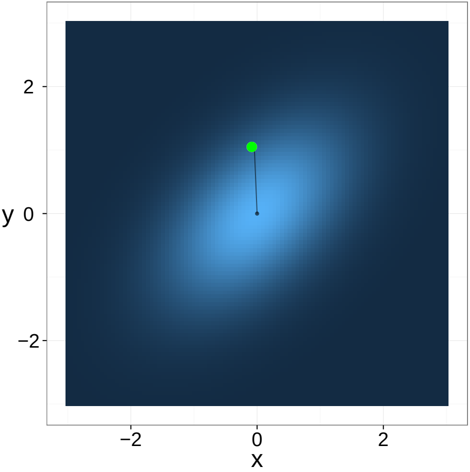

Figure: Metropolis–Hastings on a bivariate Gaussian target.

Metropolis–Hastings algorithm

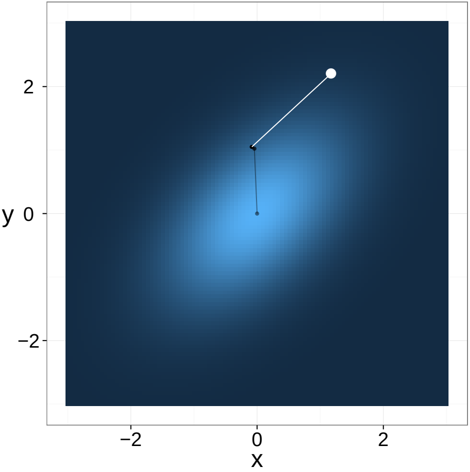

Figure: Metropolis–Hastings on a bivariate Gaussian target.

Metropolis–Hastings algorithm

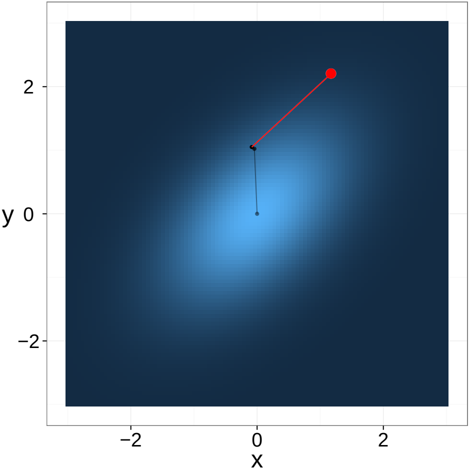

Figure: Metropolis–Hastings on a bivariate Gaussian target.

Metropolis–Hastings algorithm

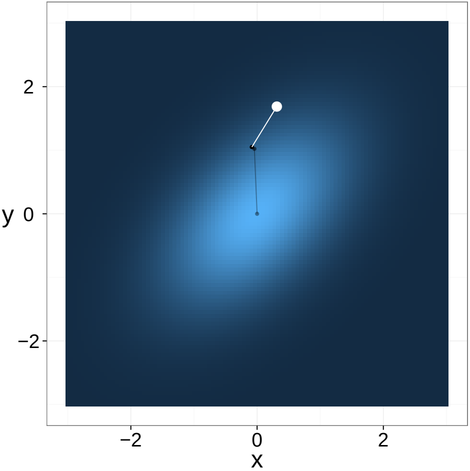

Figure: Metropolis–Hastings on a bivariate Gaussian target.

Metropolis–Hastings algorithm

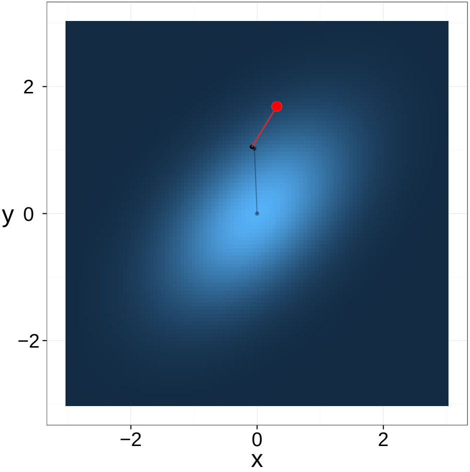

Figure: Metropolis–Hastings on a bivariate Gaussian target.

Metropolis–Hastings algorithm

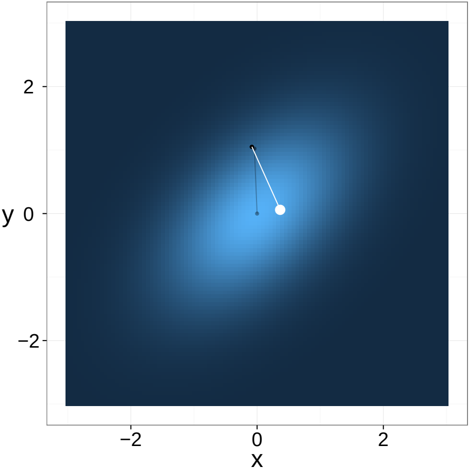

Figure: Metropolis–Hastings on a bivariate Gaussian target.

Metropolis–Hastings algorithm

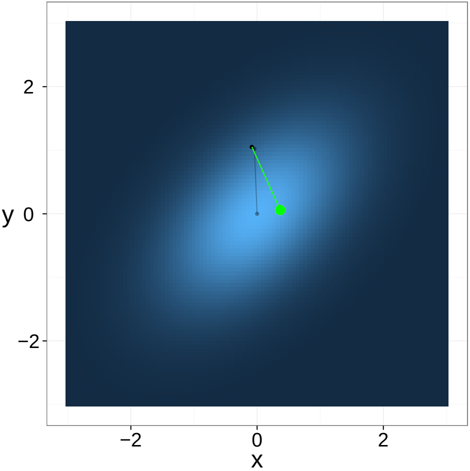

Figure: Metropolis–Hastings on a bivariate Gaussian target.

Metropolis–Hastings algorithm

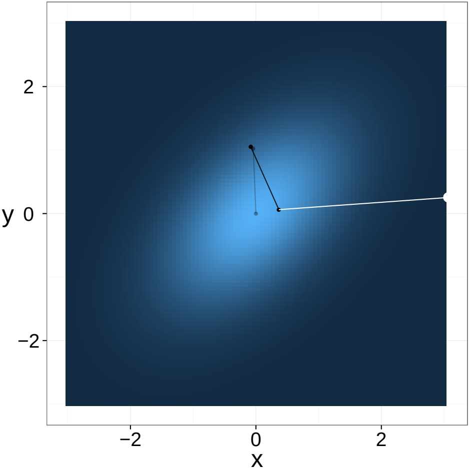

Figure: Metropolis–Hastings on a bivariate Gaussian target.

Metropolis–Hastings algorithm

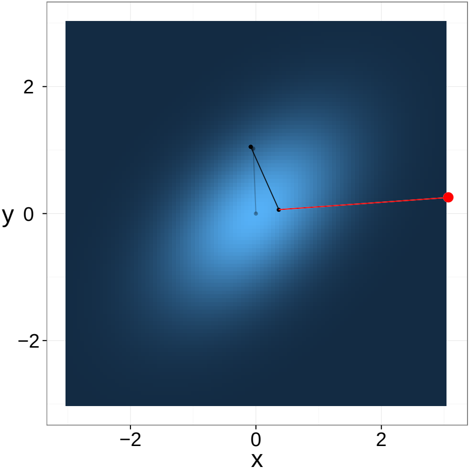

Figure: Metropolis–Hastings on a bivariate Gaussian target.

Metropolis–Hastings algorithm

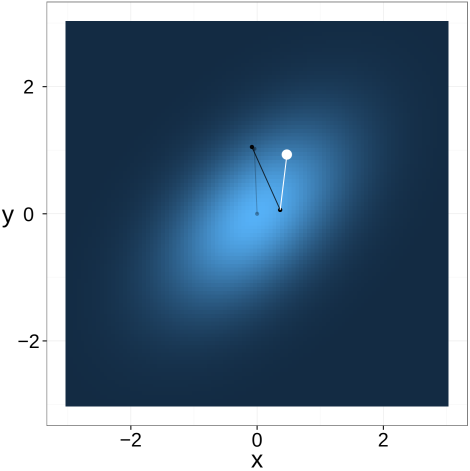

Figure: Metropolis–Hastings on a bivariate Gaussian target.

Metropolis–Hastings algorithm

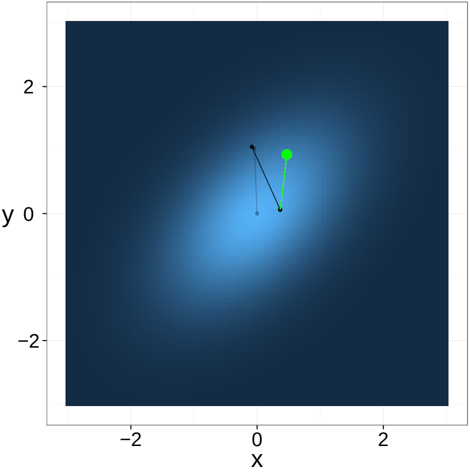

Figure: Metropolis–Hastings on a bivariate Gaussian target.

Metropolis–Hastings algorithm

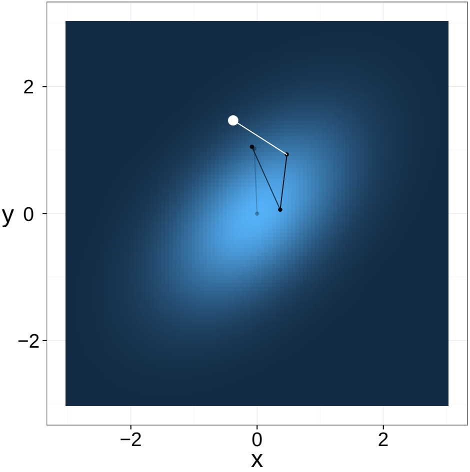

Figure: Metropolis–Hastings on a bivariate Gaussian target.

Metropolis–Hastings algorithm

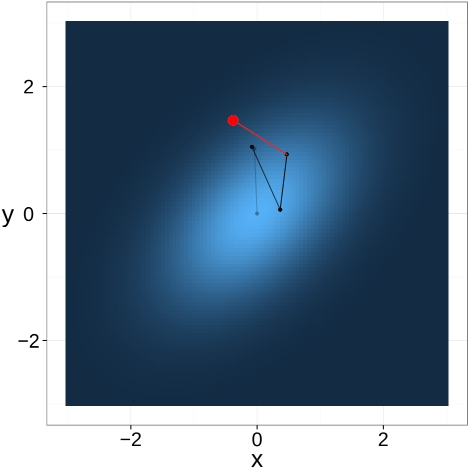

Figure: Metropolis–Hastings on a bivariate Gaussian target.

Metropolis–Hastings algorithm

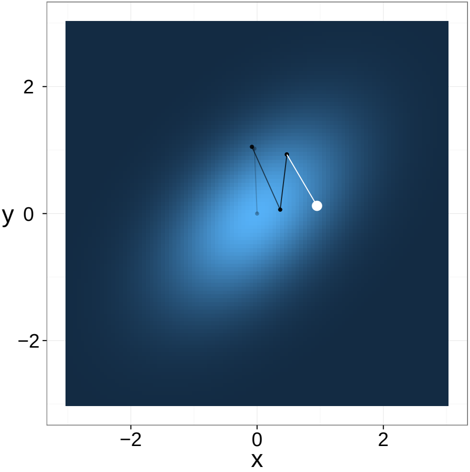

Figure: Metropolis–Hastings on a bivariate Gaussian target.

Metropolis–Hastings algorithm

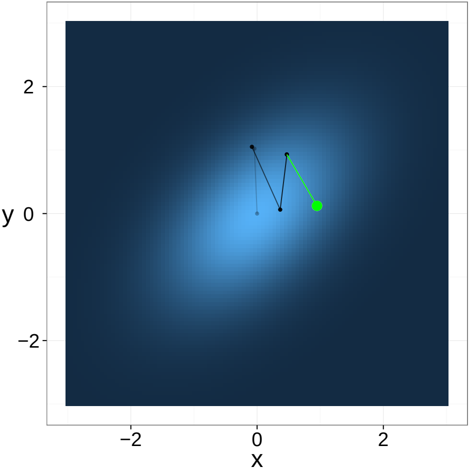

Figure: Metropolis–Hastings on a bivariate Gaussian target.

Metropolis–Hastings algorithm

- Metropolis–Hastings only requires point-wise evaluations of $\pi(x)$ up to a normalizing constant; indeed if $\widetilde{\pi}(x) \propto \pi(x)$ then

$$\frac{\pi(x^\star)\, q\left(x^{(t-1)} \middle| x^\star\right)}{\pi\left(x^{(t-1)}\right) q\left(x^\star \middle| x^{(t-1)}\right)} = \frac{\widetilde{\pi}(x^\star)\, q\left(x^{(t-1)} \middle| x^\star\right)}{\widetilde{\pi}\left(x^{(t-1)}\right) q\left(x^\star \middle| x^{(t-1)}\right)}.$$

- At each iteration t, a candidate is proposed. The probability of a candidate being accepted is given by

$$a\left(x^{(t-1)}\right) = \int_{\mathbb{X}} \alpha\left(x \middle| x^{(t-1)}\right) q\left(x \middle| x^{(t-1)}\right) dx$$

  in which case $X^{(t)} = X$, otherwise $X^{(t)} = X^{(t-1)}$.
- This algorithm clearly defines a Markov chain $\left(X^{(t)}\right)_{t \geq 1}$.

Transition Kernel and Reversibility

- **Lemma**. The transition kernel of the Metropolis–Hastings algorithm is given by

$$K(y \mid x) \equiv K(x, y) = \alpha(y \mid x) q(y \mid x) + (1 - a(x)) \delta_x(y)$$

where δ_x denotes the Dirac mass at x.

- *Proof.* We have

$$\begin{aligned} K(x, y) &= \int q(x^\star \mid x) \{\alpha(x^\star \mid x) \delta_{x^\star}(y) + (1 - \alpha(x^\star \mid x)) \delta_x(y)\} dx^\star \\ &= q(y \mid x) \alpha(y \mid x) + \left\{ \int q(x^\star \mid x)(1 - \alpha(x^\star \mid x)) dx^\star \right\} \delta_x(y) \\ &= q(y \mid x) \alpha(y \mid x) + \left\{ 1 - \int q(x^\star \mid x) \alpha(x^\star \mid x) dx^\star \right\} \delta_x(y) \\ &= q(y \mid x) \alpha(y \mid x) + \{1 - a(x)\} \delta_x(y). \end{aligned}$$

Reversibility

- **Proposition**. The Metropolis–Hastings kernel K is $\pi-$reversible and thus admit π as invariant distribution.

- *Proof.* For any $x, y \in \mathbb{X}$, with $x \neq y$

$$\begin{aligned}
\pi(x)K(x,y) &= \pi(x)q(y \mid x)\alpha(y \mid x) \\
&= \pi(x)q(y \mid x)\min\left(1, \frac{\pi(y)q(x \mid y)}{\pi(x)q(y \mid x)}\right) \\
&= \min\left(\pi(x)q(y \mid x), \pi(y)q(x \mid y)\right) \\
&= \pi(y)q(x \mid y)\min\left(\frac{\pi(x)q(y \mid x)}{\pi(y)q(x \mid y)}, 1\right) \\
&= \pi(y)K(y,x).
\end{aligned}$$

If $x = y$, then obviously $\pi(x)K(x,y) = \pi(y)K(y,x)$.

Reducibility and periodicity of Metropolis–Hastings

- Consider the target distribution

$$\pi\left(x\right)=\left(\mathcal{U}_{[0,1]}\left(x\right)+\mathcal{U}_{[2,3]}\left(x\right)\right)/2$$

  and the proposal distribution

$$q\left(x^{\star}\middle|\,x\right)=\mathcal{U}_{(x-\delta,x+\delta)}\left(x^{\star}\right).$$

- The MH chain is reducible if $\delta\leq 1$: the chain stays either in $[0,1]$ or $[2,3]$.

- Note that the MH chain is aperiodic if it always has a non-zero chance of staying where it is.

Law of Large Numbers

- The MH chain $\left(X^{(t)}\right)_{t\geq 1}$ is irreducible if $q\left(x^{\star}\middle|\, x\right) > 0$ for any $x, x^{\star} \in \mathrm{supp}(\pi)$: every state can be reached in a single step (strongly irreducible). Less strict conditions in (Roberts & Rosenthal, 2004).
- The MH chain is Harris recurrent if it is irreducible (see Tierney, 1994).
- **Theorem**. If the Markov chain generated by the Metropolis–Hastings sampler is $\pi-$irreducible, then we have for any integrable function $\varphi : \mathbb{X} \to \mathbb{R}$:

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} \varphi\left(X^{(i)}\right) = \int_{\mathbb{X}} \varphi\left(x\right) \pi\left(x\right) dx$$

for *every* starting value $X^{(1)}$.

Random Walk Metropolis–Hastings

- In the Metropolis–Hastings, pick $q(x^\star \mid x) = g(x^\star - x)$ with g being a *symmetric* distribution, thus

$$X^\star = X + \varepsilon, \quad \varepsilon \sim g;$$

  e.g. g is a zero-mean multivariate normal or t-student.

- Acceptance probability becomes

$$\alpha(x^\star \mid x) = \min\left(1, \frac{\pi(x^\star)}{\pi(x)}\right).$$

- We accept...
  - a move to a more probable state with probability 1;
  - a move to a less probable state with probability

$$\pi(x^\star)/\pi(x) < 1.$$

Independent Metropolis–Hastings

- If the proposal distribution $q(x^\star \mid x)$ does not depend on x, we call it an independent proposal.
- Acceptance probability becomes

$$\alpha(x^\star \mid x) = \min\left(1, \frac{\pi(x^\star)q(x)}{\pi(x)q(x^\star)}\right).$$

- For instance, multivariate normal or t-student distribution.
- If $\pi(x)/q(x) < M$ for all x and some $M < \infty$, then the chain is uniformly ergodic.
- It can be shown that the acceptance probability at stationarity is then at least $1/M$ (Lemma 7.9 of Robert & Casella).
- On the other hand, if such an M does not exist, the chain is not even geometrically ergodic!

Choosing a good proposal distribution

- Goal: to design a Markov chain with small correlation $\rho\left(X^{(t-1)}, X^{(t)}\right)$ between subsequent values (why?).
- Two sources of correlation:
  - between the current state $X^{(t-1)}$ and proposed value $X \sim q\left(\cdot \mid X^{(t-1)}\right)$,
  - correlation induced if $X^{(t)} = X^{(t-1)}$, if proposal is rejected.
- Trade-off: there is a compromise between
  - proposing large moves,
  - obtaining a decent acceptance probability.
- For multivariate distributions: covariance of proposal should reflect the covariance structure of the target.

Choice of proposal

- Target distribution, we want to sample from

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

- We use a random walk Metropolis—Hastings algorithm with

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

- What is the optimal choice of σ^2?
- We consider three choices: $\sigma^2 = 0.1^2, 1, 10^2$.

Metropolis–Hastings algorithm

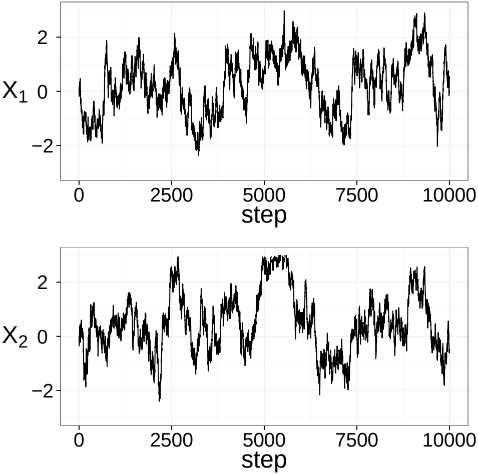

Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 0.1^2$, the acceptance rate is $\approx 94\%$.

Metropolis–Hastings algorithm

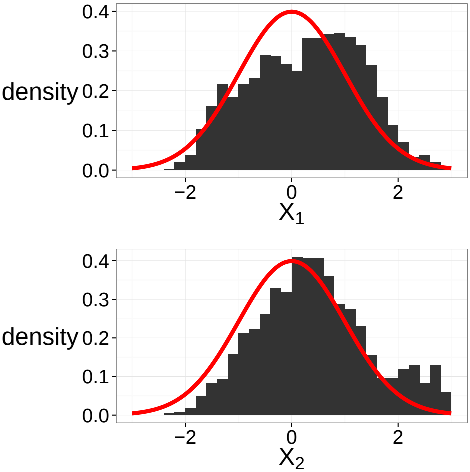

Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 0.1^2$, the acceptance rate is $\approx 94\%$.

Metropolis–Hastings algorithm

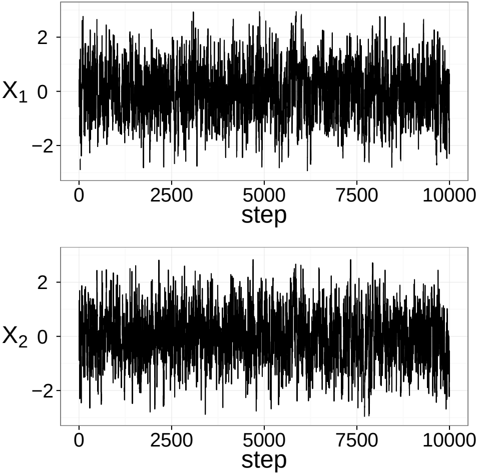

Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 1$, the acceptance rate is $\approx 52\%$.

Metropolis–Hastings algorithm

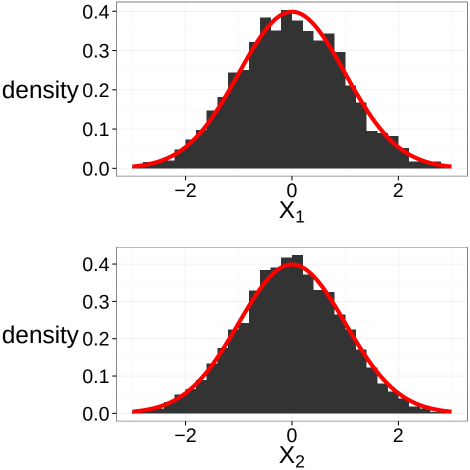

Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 1$, the acceptance rate is $\approx 52\%$.

Metropolis–Hastings algorithm

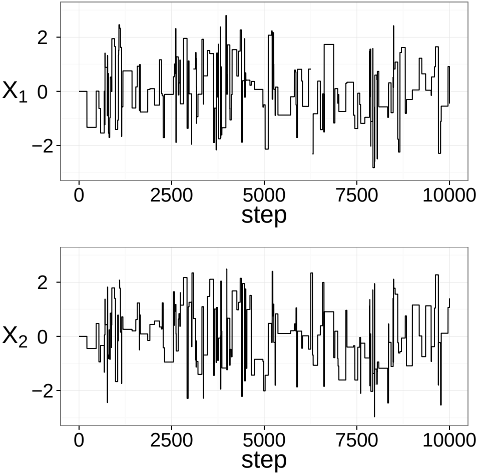

Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 10$, the acceptance rate is $\approx 1.5\%$.

Metropolis–Hastings algorithm

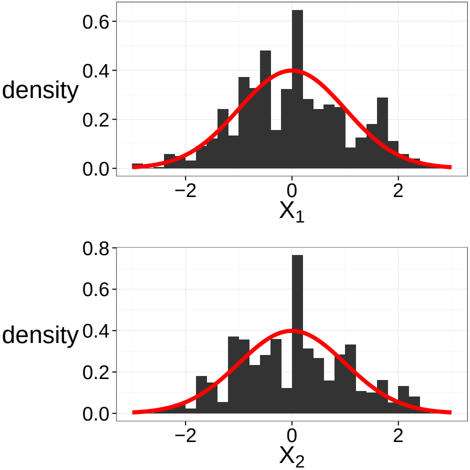

Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 10$, the acceptance rate is $\approx 1.5\%$.

Choice of proposal

- Aim at some intermediate acceptance ratio: 20%? 40%? Some hints come from the literature on "optimal scaling".
- Maximize the expected square jumping distance:

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

- In multivariate cases, try to mimick the covariance structure of the target distribution.

Cooking recipe: run the algorithm for T iterations, check some criterion, tune the proposal distribution accordingly, run the algorithm for T iterations again ...

"Constructing a chain that mixes well is somewhat of an art."
All of Statistics, L. Wasserman.

The adaptive MCMC approach

- One can make the transition kernel K adaptive, i.e. use K_t at iteration t and choose K_t using the past sample $(X_1, \ldots, X_{t-1})$.
- The Markov chain is not homogeneous anymore: the mathematical study of the algorithm is much more complicated.
- Adaptation can be counterproductive in some cases (see Atchadé & Rosenthal, 2005)!
- Adaptive Gibbs samplers also exist.