Computational inverse problems

Nuutti Hyvönen, Jenni Heino,
Juha-Pekka Puska

nuutti.hyvonen@aalto.fi, juha-pekka.puska@aalto.fi

Eleventh lecture, April 5, 2021.

Does the above M–H algorithm really work? It is not quite obvious...

According to our construction, the Markov process introduced at the beginning, i.e. the one involving R and r, is with the choice

$$K(x,y) = (1 - r(x))R(x,y) = \tilde{\alpha}(x,y)q(x,y)$$

such that p is its invariant density. In the actual M–H algorithm, $q(x,y)$ is the employed proposal kernel, i.e. a probability density in its second variable, and $0 \leq \tilde{\alpha}(x,y) \leq 1$ is the acceptance probability *that depends on both the current location x and the proposed location y (unlike $r(x)$)*. If one is able to define $0 \leq r(x) \leq 1$ and a transition kernel $R(x,y)$ for any given $q(x,y)$ and $\tilde{\alpha}(x,y)$ so that the above identity is satisfied, our construction is legitimate.

It is easy to verify that this is achived by first setting

$$r(x) = 1 - \int \tilde{\alpha}(x,y) q(x,y) dy, \qquad x \in \mathbb{R}^n,$$

and then simply defining

$$R(x,y) = \frac{\tilde{\alpha}(x,y) q(x,y)}{1 - r(x)}, \qquad x, y \in \mathbb{R}^n.$$

Indeed, with these choices it is obvious that

$$(1 - r(x)) R(x,y) = \tilde{\alpha}(x,y) q(x,y) \quad \text{and} \quad \int R(x,y) dy = 1.$$

Moreover, clearly $r(x) \leq 1$ and also

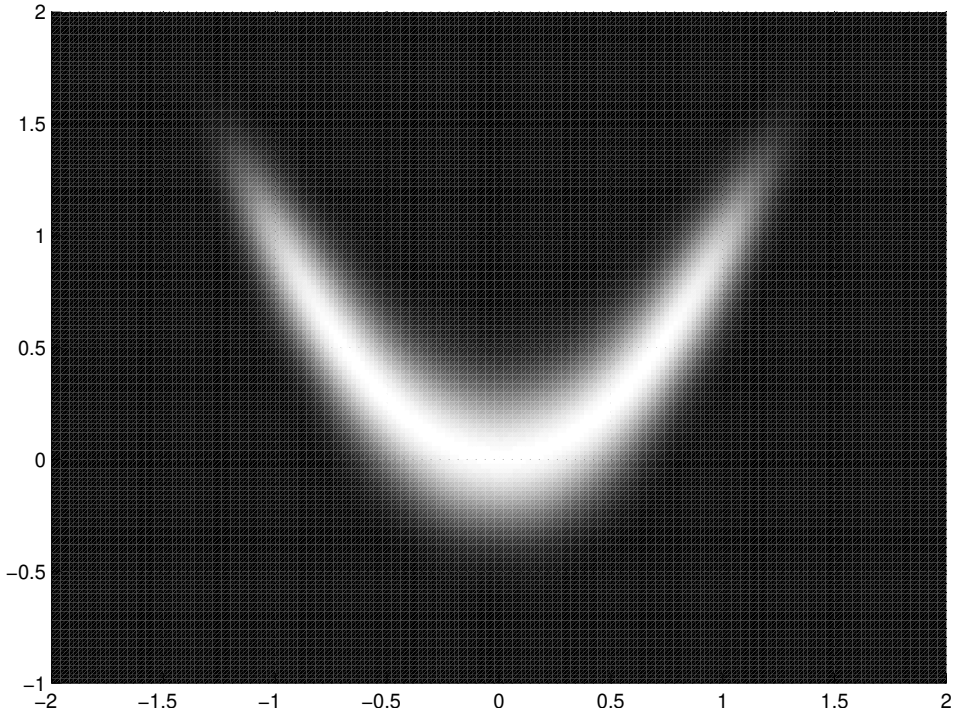
$$r(x) \geq 1 - \int q(x,y) dy = 0$$

for any $x \in \mathbb{R}^n$.

Example

Consider sampling in $\mathbb{R}^2$ from the density

$$\pi(x) \propto \exp\left(-10(x_1^2 - x_2)^2 - (x_2 - \frac{1}{4})^4\right).$$



We use white noise random walk proposal

$$q(x, y) = \frac{1}{\sqrt{2\pi\gamma^2}} \exp\left(-\frac{1}{2\gamma^2}\|x - y\|^2\right).$$

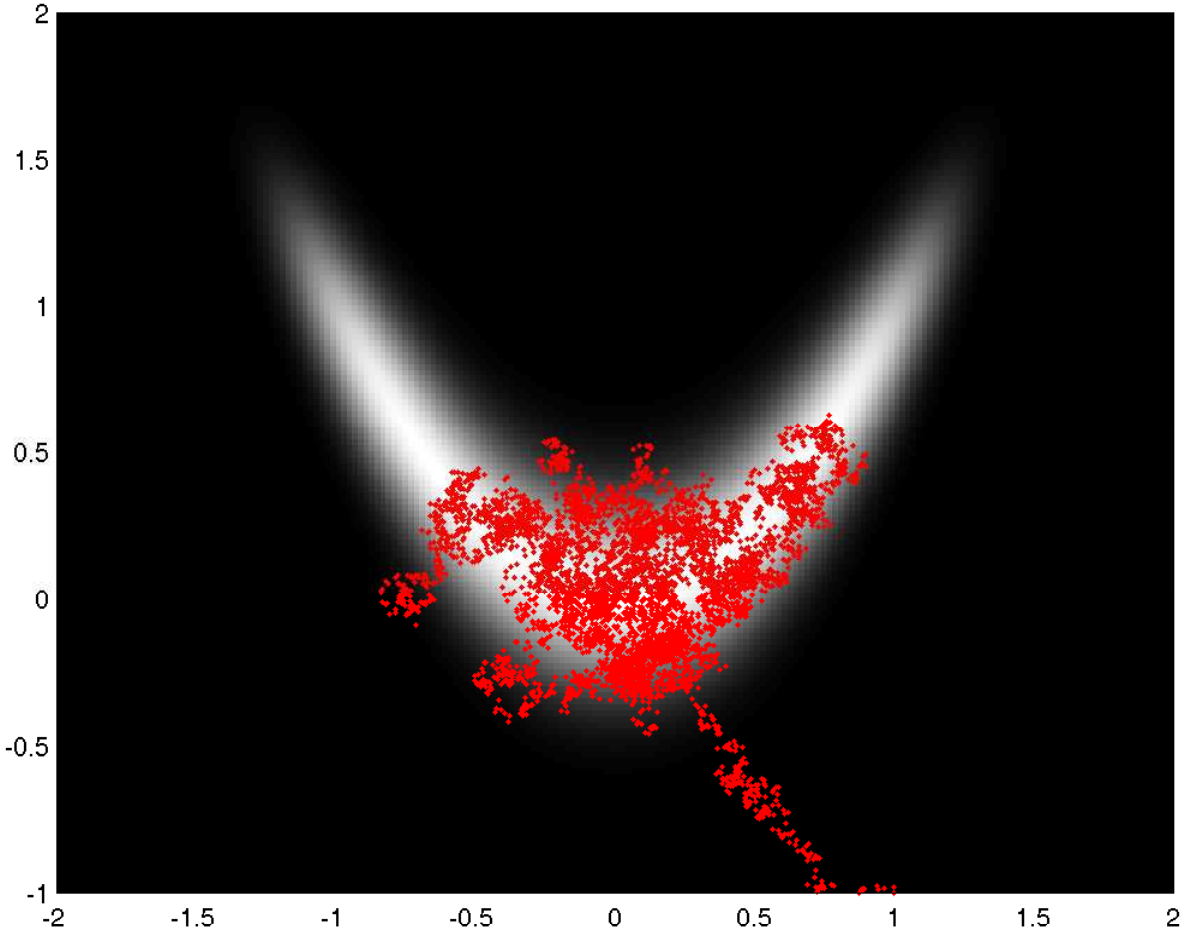
Note that now the transition kernel is symmetric, i.e.,

$$q(x, y) = q(y, x),$$

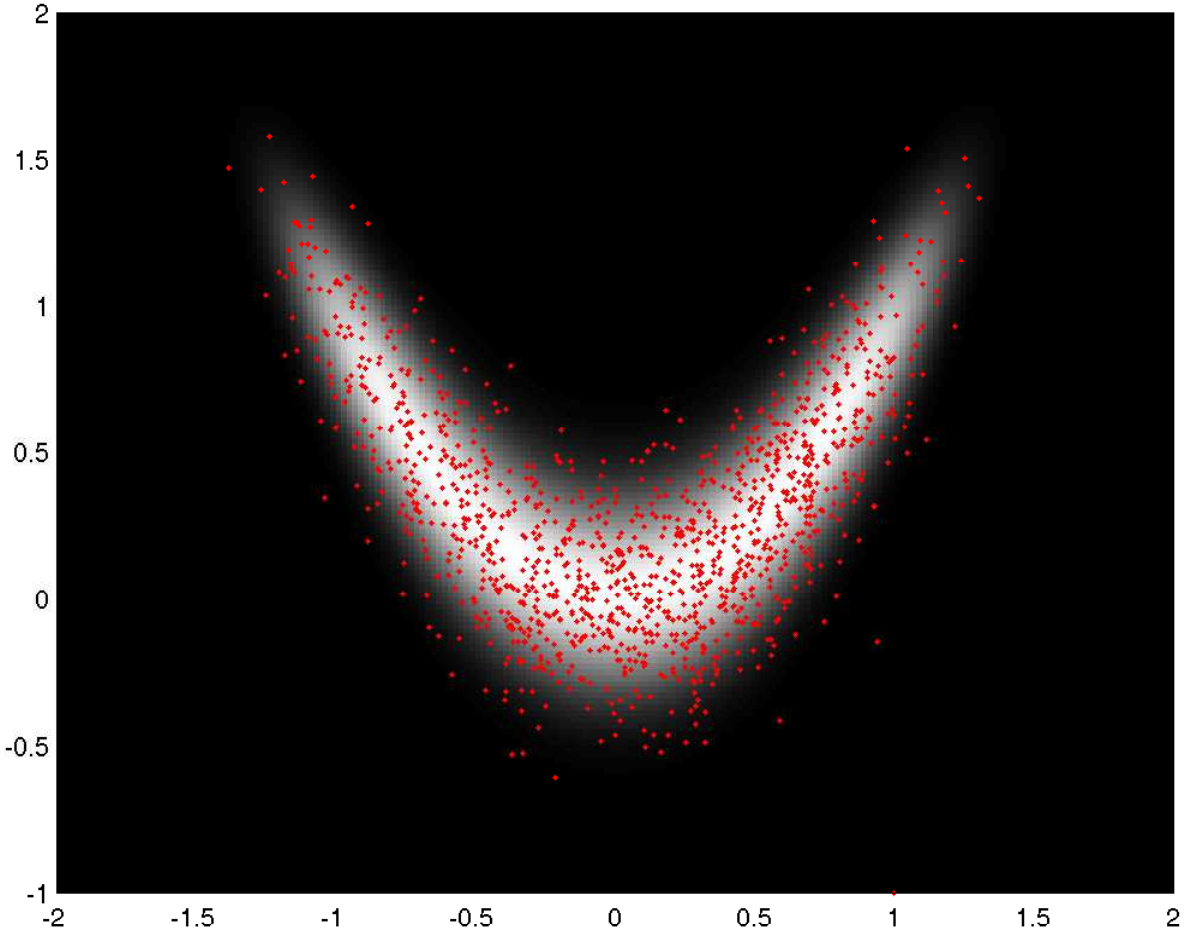
and hence

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

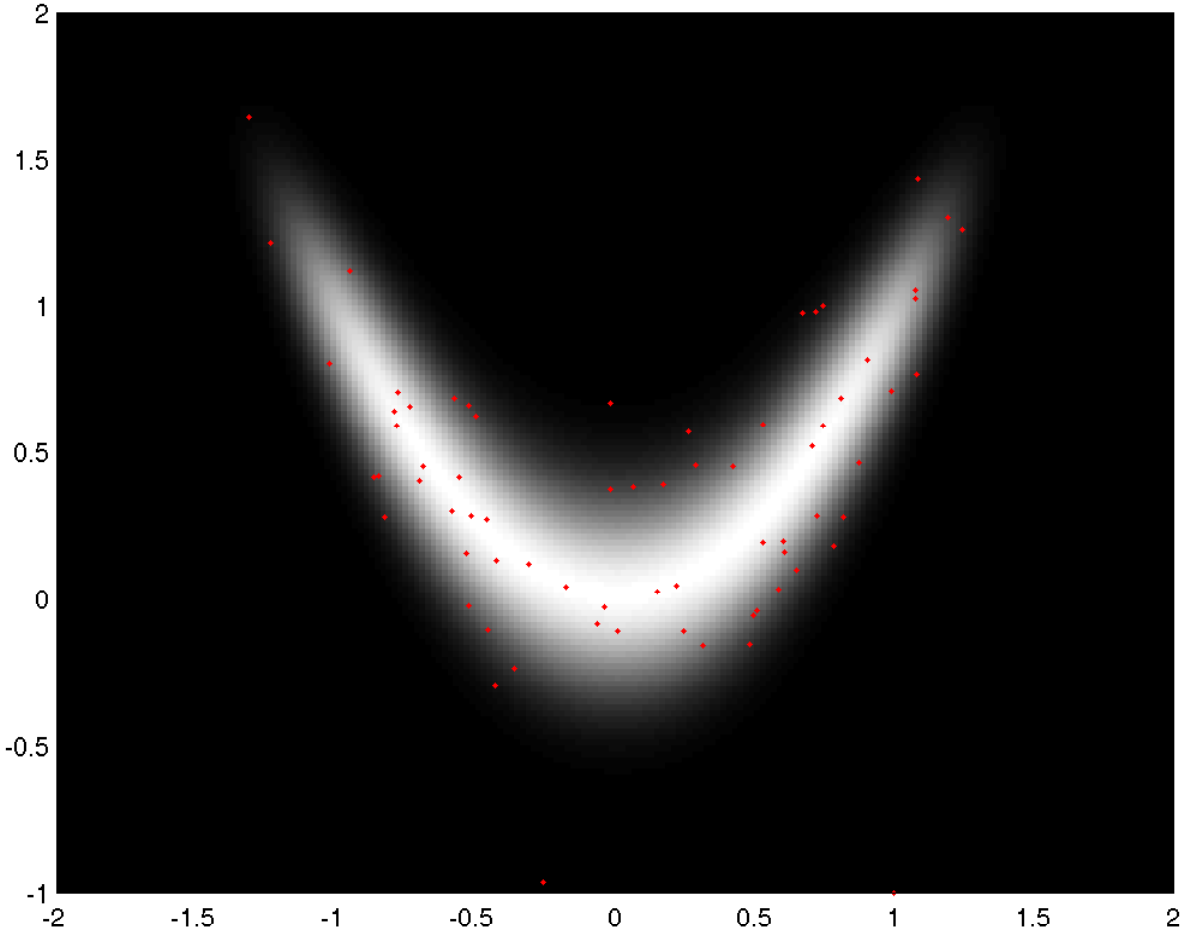
$\gamma = 0.02$; acceptance rate 95.6 %



$\gamma = 0.7$; acceptance rate 24.5 %



$\gamma = 4$; acceptance rate 1.4 %



Adapting the Metropolis-Hastings sampler

With the white noise random walk proposal density (used in the numerical example of the previous lecture), the sampler does not take into account the form of the posterior density.

However, the shape of the density *can* be taken into account when designing the proposal density in order to minimize the number of 'wasted proposals'. In high-dimensional setting, this becomes especially useful if the posterior density is highly *anisotropic*, i.e., if the posterior is stretched in some directions.

The proposal distribution can be updated while the sampling algorithm moves around the posterior density. This process is often called *adaptation*.

Gibbs sampler

Let us first consider some notational details:

- $I = \{1, 2, \ldots, n\}$ is the index set of $\mathbb{R}^n$.
- $I = \bigcup_{j=1}^{m} I_m$ is a partitioning of the index set into disjoint nonempty subsets.
- The number of elements in I_j is denoted by k_j; $k_1 + \cdots + k_m = n$.
- We partition $\mathbb{R}^n$ as $\mathbb{R}^n = \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_m}$, and correspondingly

$$x = [x_{I_1}; \ldots; x_{I_m}] \in \mathbb{R}^n, \qquad x_{I_j} \in \mathbb{R}^{k_j},$$

  where $x_i \in \mathbb{R}$ is a component of the vector x_{I_j} if and only if $i \in I_j$.

In practice, it often holds that $k_j = 1$ for all $j = 1, \ldots, m$, meaning that $m = n$ and x_{I_j} is just the jth component of the original vector x.

Transition kernel for the Gibbs sampler

Suppose that we are still aiming at sampling some given probability density $p : \mathbb{R}^n \to \mathbb{R}_+$, and recall the Markov process considered at the previous lecture: If you are currently at some $x \in \mathbb{R}^n$, either

1. stay put at x with the probability $r(x)$, $0 \leq r(x) \leq 1$, or
2. move away from x using a transition kernel $R(x, y)$ otherwise.

Recall also that we made the definition

$$K(x, y) \;=\; (1 - r(x))R(x, y).$$

For the Gibbs sampler, we choose $r(x) = 0$ for all $x \in \mathbb{R}^n$, i.e., moving is obligatory, and define

$$K(x, y) \;=\; R(x, y) \;=\; \prod_{i=1}^{m} p(y_{I_i} \,|\, y_{I_1}, \ldots, y_{I_{i-1}}, x_{I_{i+1}}, \ldots, x_{I_m}),$$

where the conditional densities are defined in the natural way based on p, i.e.,

$$p(y_{I_i} \,|\, y_{I_1}, \ldots, y_{I_{i-1}}, x_{I_{i+1}}, \ldots, x_{I_m}) = \frac{p(y_{I_1}, \ldots, y_{I_i}, x_{I_{i+1}}, \ldots, x_{I_m})}{\int_{\mathbb{R}^{k_i}} p(y_{I_1}, \ldots, y_{I_i}, x_{I_{i+1}}, \ldots x_{I_m}) dy_{I_i}}.$$

Such a transition kernel K does not, in general, satisfy the detailed balance equation, i.e.,

$$p(y)K(y,x) \neq p(x)K(x,y),$$

but it satisfies the (standard) balance equation,

$$\int_{\mathbb{R}^n} p(y)K(y,x)dx = \int_{\mathbb{R}^n} p(x)K(x,y)dx,$$

which is a sufficient condition for p being an invariant density of the above introduced Markov process. (See the slides of the previous lecture for the details.)

Proof: Consider first the left-hand side of the balance equation.

Due to the basic properties of probability densities, we have

$$\int_{\mathbb{R}^{k_i}} p(x_{I_i} \,|\, x_{I_1}, \ldots, x_{I_{i-1}}, y_{I_{i+1}}, \ldots, y_{I_m}) dx_{I_i} = 1$$

for all $i = 1, \ldots, m$. By integrating the kernel $K(y, x)$ over $\mathbb{R}^{k_m}$, we thus get

$$\int_{\mathbb{R}^{k_m}} K(y, x) dx_{I_m} = \int_{\mathbb{R}^{k_m}} \prod_{i=1}^{m} p(x_{I_i} \,|\, x_{I_1}, \ldots, x_{I_{i-1}}, y_{I_{i+1}}, \ldots, y_{I_m}) dx_{I_m}$$

$$= \prod_{i=1}^{m-1} p(x_{I_i} \,|\, x_{I_1}, \ldots, x_{I_{i-1}}, y_{I_{i+1}}, \ldots, y_{I_m}) \int_{\mathbb{R}^{k_m}} p(x_{I_m} \,|\, x_{I_1}, \ldots, x_{I_{m-1}}) dx_{I_m}$$

$$= \prod_{i=1}^{m-1} p(x_{I_i} \,|\, x_{I_1}, \ldots, x_{I_{i-1}}, y_{I_{i+1}}, \ldots, y_{I_m}).$$

Inductively, by always integrating with respect to the last block of x with respect to which we have not yet integrated, we easily obtain that altogether

$$\int_{\mathbb{R}^n} K(y,x)dx = 1,$$

which in turn implies that

$$\int_{\mathbb{R}^n} p(y)K(y,x)dx = p(y)\int_{\mathbb{R}^n} K(y,x)dx = p(y).$$

Next, we consider the right-hand side of the balance equation. Since $K(x,y)$ is independent of x_{I_1} and due to the definition of marginal probability densities, we have

$$\int_{\mathbb{R}^{k_1}} p(x)K(x,y)dx_{I_1} = K(x,y)\int_{\mathbb{R}^{k_1}} p(x)dx_{I_1} =: K(x,y)p(x_{I_2},\ldots,x_{I_m}).$$

By substituting the definition of K in the above formula, we see that

$$\begin{aligned}\int_{\mathbb{R}^{k_1}} p(x)K(x,y)dx_{I_1} &= K(x,y)p(x_{I_2},\ldots,x_{I_m}) \\ &= \left(\prod_{i=2}^{m} p(y_{I_i} \,|\, y_{I_1},\ldots,y_{I_{i-1}},x_{I_{i+1}},\ldots,x_{I_m})\right) \\ &\qquad \times p(y_{I_1} \,|\, x_{I_2},\ldots,x_{I_m})p(x_{I_2},\ldots,x_{I_m}) \\ &= \left(\prod_{i=2}^{m} p(y_{I_i} \,|\, y_{I_1},\ldots,y_{I_{i-1}},x_{I_{i+1}},\ldots,x_{I_m})\right) p(y_{I_1},x_{I_2},\ldots,x_{I_m}).\end{aligned}$$

Next, we integrate with respect to x_{I_2} over $\mathbb{R}^{k_2}$. By denoting

$$a_i \;=\; p(y_{I_i} \,|\, y_{I_1}, \ldots, y_{I_{i-1}}, x_{I_{i+1}}, \ldots, x_{I_m}), \qquad i = 2, \ldots, m,$$

we may write

$$\int_{\mathbb{R}^{k_2}} \int_{\mathbb{R}^{k_1}} p(x) K(x,y) dx_{I_1} dx_{I_2} = \int_{\mathbb{R}^{k_2}} \prod_{i=2}^{m} a_i \, p(y_{I_1}, x_{I_2}, \ldots, x_{I_m}) dx_{I_2}$$

$$= \prod_{i=3}^{m} a_i \, p(y_{I_2} \,|\, y_{I_1}, x_{I_3}, \ldots, x_{I_m}) \int_{\mathbb{R}^{k_2}} p(y_{I_1}, x_{I_2}, \ldots, x_{I_m}) dx_{I_2}$$

$$= \prod_{i=3}^{m} a_i \, p(y_{I_2} \,|\, y_{I_1}, x_{I_3}, \ldots, x_{I_m}) p(y_{I_1}, x_{I_3}, \ldots, x_{I_m})$$

$$= \prod_{i=3}^{m} a_i \, p(y_{I_1}, y_{I_2}, x_{I_3}, \ldots, x_{I_m}).$$

We can continue inductively integrating over the remaining blocks $x_{I_3}, \ldots, x_{I_m}$ in turns, which eventually results in

$$\int_{\mathbb{R}^n} p(x)K(x,y)dx \;=\; p(y_{I_1}, \ldots, y_{I_m}) \;=\; p(y),$$

and the proof is complete. □

Gibbs sampler algorithm

1. Choose the initial value $x^0 \in \mathbb{R}^n$ and set $k = 0$.
2. Draw the next sample as follows:
   (a) Set $x = x^k$ and $j = 1$.
   (b) Draw $y_{I_j} \in \mathbb{R}^{k_j}$ from the k_j-dimensional distribution $p(y_{I_j} \,|\, y_{I_1}, \ldots, y_{I_{j-1}}, x_{I_{j+1}}, \ldots, x_{I_m})$.
   (c) If $j = m$, set $y = [y_{I_1}; \ldots; y_m]$ and terminate the inner loop. Otherwise, set $j \leftarrow j + 1$ and return to step (b).
3. Set $x^{k+1} = y$, increase $k \leftarrow k + 1$ and return to step 2, unless the chosen stopping criterion is satisfied.

Single component Gibbs sampler algorithm

1. Choose the initial value $x^0 \in \mathbb{R}^n$ and set $k = 0$.
2. Draw the next sample as follows:
   (a) Set $x = x^k$ and $j = 1$.
   (b) Draw $t \in \mathbb{R}$ from the one-dimensional distribution

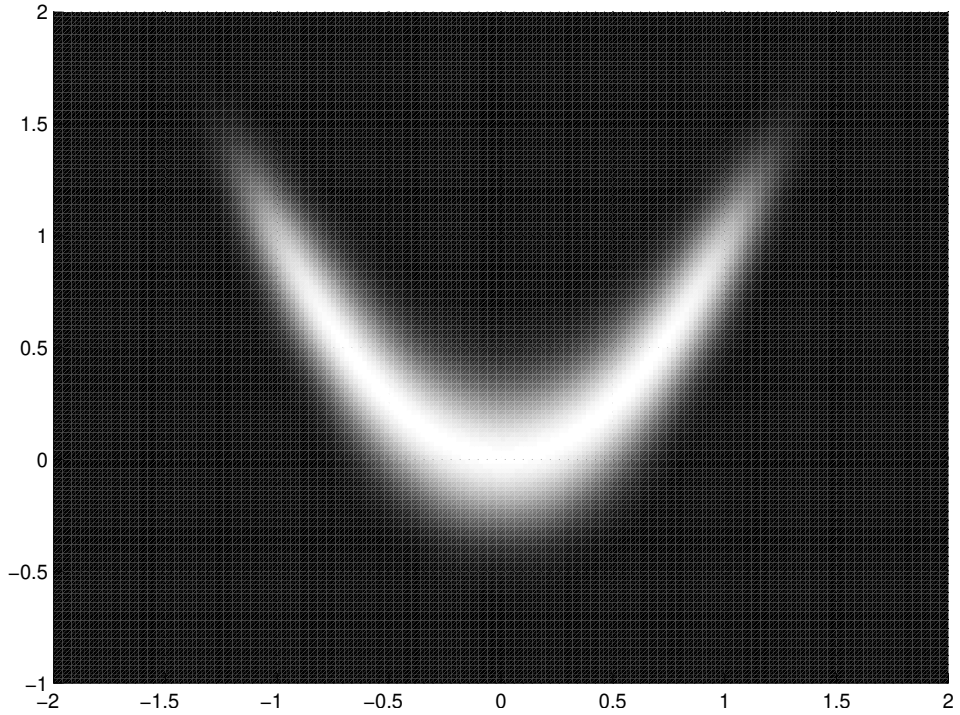
   $$p(t \,|\, y_1, \ldots, y_{j-1}, x_{j+1}, \ldots, x_n) \propto p(y_1, \ldots, y_{j-1}, t, x_{j+1}, \ldots, x_n)$$

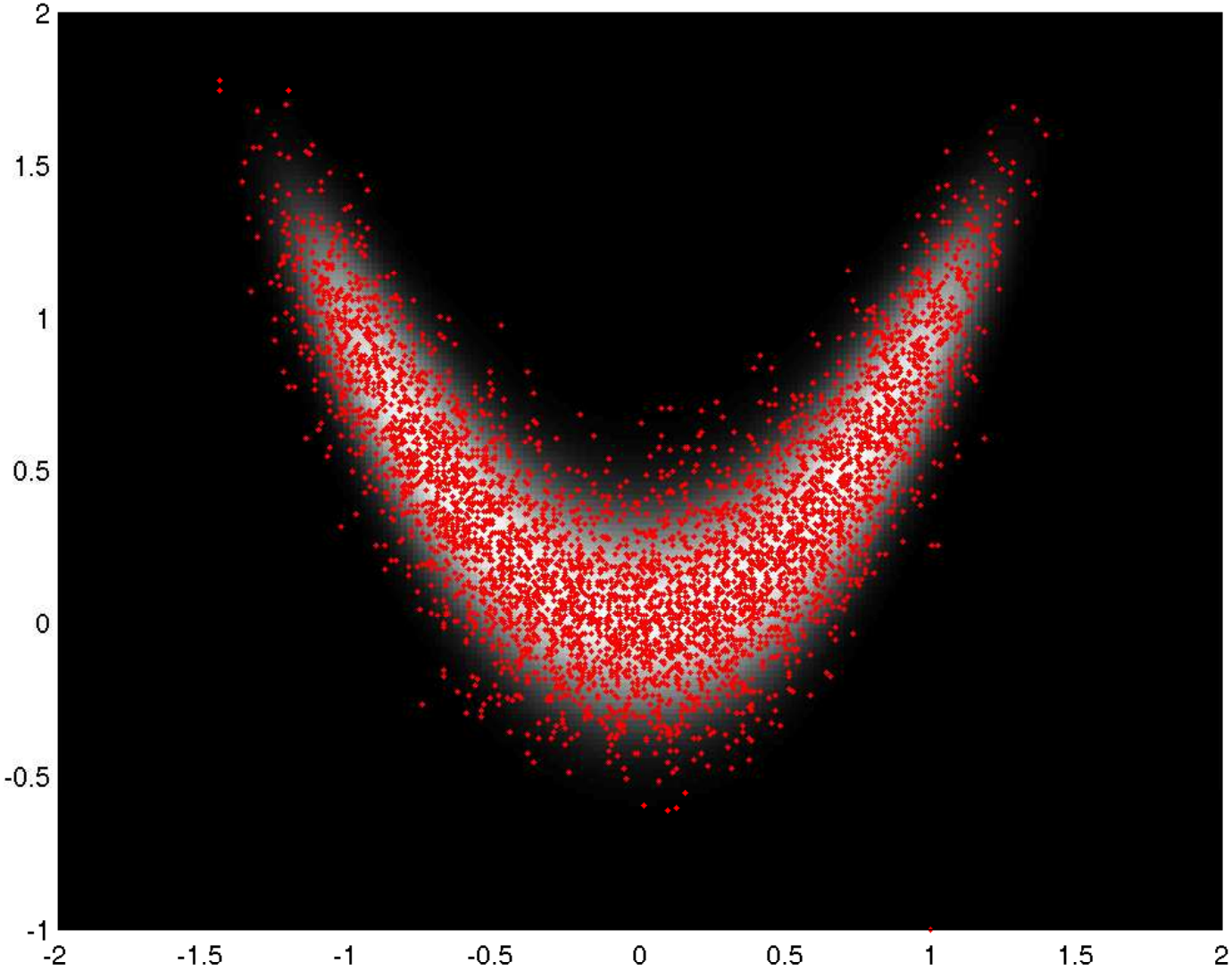
   and set $y_j = t$.
   (c) If $j = n$, set $y = [y_1, \ldots, y_n]^{\mathrm{T}}$ and terminate the inner loop. Otherwise, set $j \leftarrow j + 1$ and return to step (b).
3. Set $x^{k+1} = y$, increase $k \leftarrow k + 1$ and return to step 2, unless the chosen stopping criterion is satisfied.

Example

Consider again the density

$$\pi(x) \propto \exp\left(-10(x_1^2 - x_2)^2 - (x_2 - \frac{1}{4})^4\right), \quad x \in \mathbb{R}^2.$$




2
1.5
1
0.5
0
-0.5
-1
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

How to judge the quality of a sample?

Essential questions:

- What sampling strategy and/or proposal distribution works the best?
- Is the sample big enough?

Consider estimates of the form

$$\int f(x)\pi(x)dx = E\{f(X)\} \approx \frac{1}{N}\sum_{j=1}^{N} f(x_j),$$

and recall that the Central Limit Theorem gives some answers regarding the convergence.

Assume that the variables $Y_j = f(X_j) \in \mathbb{R}$ are mutually independent and identically distributed with $E\{Y_j\} = \bar{y}$ and $\mathrm{var}(Y_j) = \sigma^2$, and define

$$\tilde{Y}_N = \frac{1}{N}\sum_{j=1}^{N} Y_j \qquad \text{and} \qquad Z_N = \frac{\sqrt{N}(\tilde{Y}_N - \bar{y})}{\sigma}.$$

Then, $\tilde{Y}_N \to E\{Y\}$ almost surely (LLN). Moreover, Z_N is asymptotically (standard) normally distributed, that is,

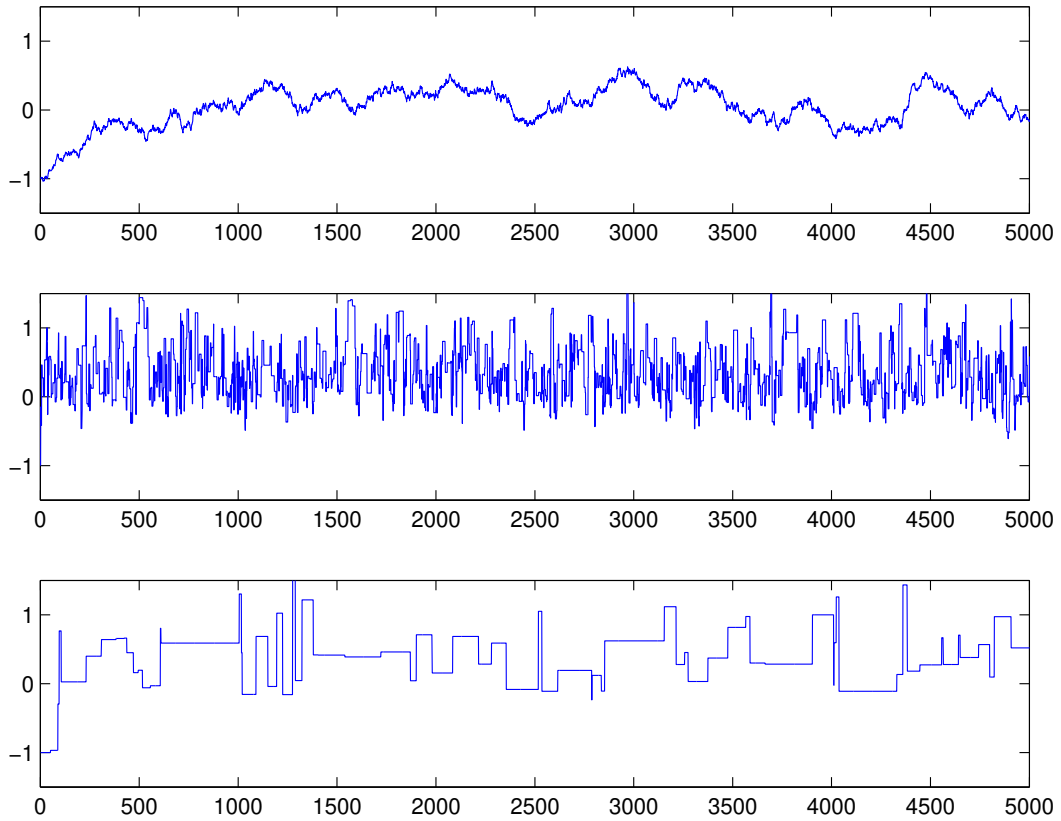
$$\lim_{N\to\infty} P\{Z_n \le z\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{1}{2}s^2\right) ds.$$

Loosely speaking, the above result says that the approximation error behaves as

$$\frac{1}{N}\sum_{j=1}^{N} f(x_j) - \int f(x)\pi(x)dx \;\approx\; \frac{\sigma}{\sqrt{N}}$$

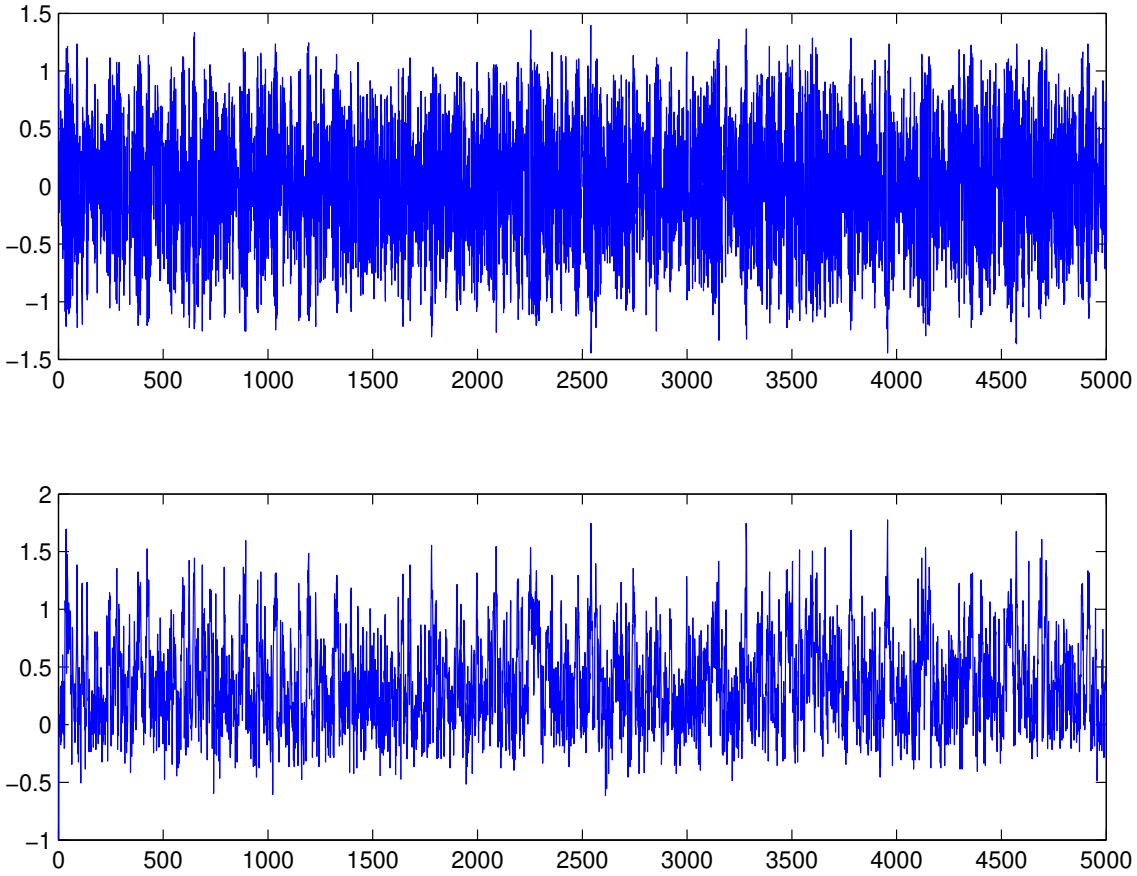
provided that the samples $\{x_j\}$ are *independent*.

Let us have another look at the sample histories corresponding to our standard example. First, the Metropolis–Hastings algorithm for the three choices of γ (the vertical component is plotted):



Clearly, consecutive elements are not independent.

Then, the Gibbs sampler (both components are plotted):

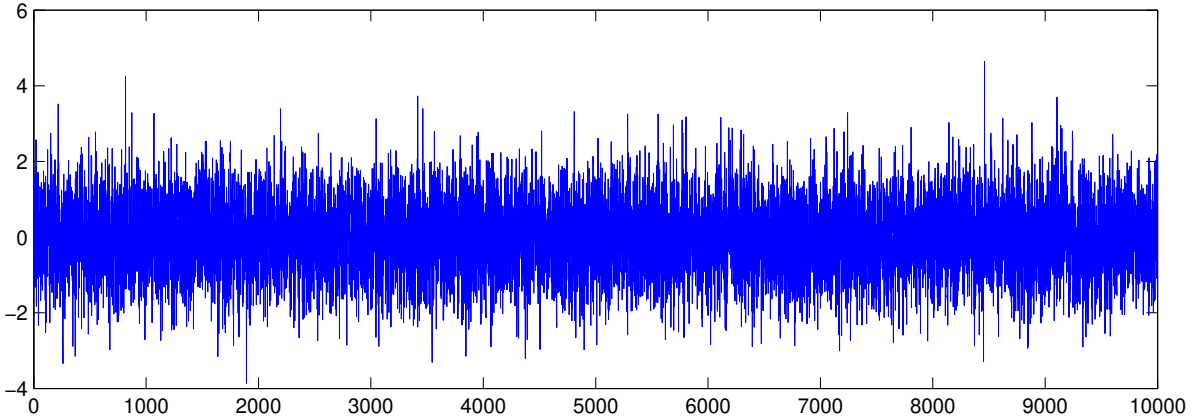


The results are somewhat better, but there is still some correlation between consecutive elements — especially for the vertical component.

If every kth sample point is independent, one might expect the discrepancy to behave as $1/\sqrt{N/k} = \sqrt{k/N}$ instead of $1/\sqrt{N}$. Consequently, one should try to choose the proposal distribution so that the *correlation length* is as small as possible.

Quick visual assessment: Take a look at the sample histories of individual components. How should they look like?

Consider a *white noise* signal, where the sample points are independent and the sample history looks like a "fuzzy worm". This is something one could aim at.



Autocovariance and correlation length

Denote by $f_{\mathrm{c}}(x_j) \in \mathbb{R}$, $j = 1, \ldots, N$, the centered sample points, i.e.,

$$f_{\mathrm{c}}(x_j) = f(x_j) - \frac{1}{N} \sum_{i=1}^{N} f(x_i), \qquad j = 1, \ldots, N.$$

Define the normalized autocovariance of the sample as

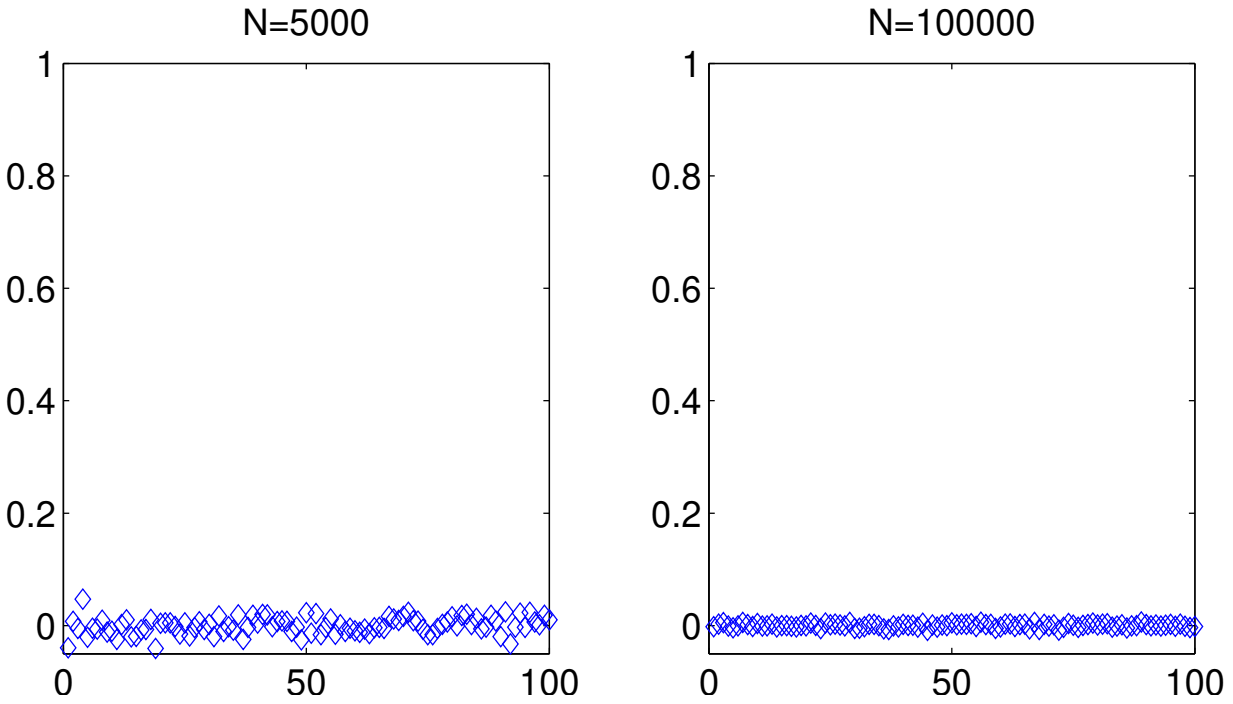
$$\gamma_k^2 = \frac{1}{\gamma_0^2 (N-k)} \sum_{j=1}^{N-k} f_{\mathrm{c}}(x_j) f_{\mathrm{c}}(x_{j+k}), \qquad k \geq 1,$$

where $\gamma_0^2 = \frac{1}{N} \sum_{j=1}^{N} f_{\mathrm{c}}(x_j)^2$ is the mean energy of the signal.

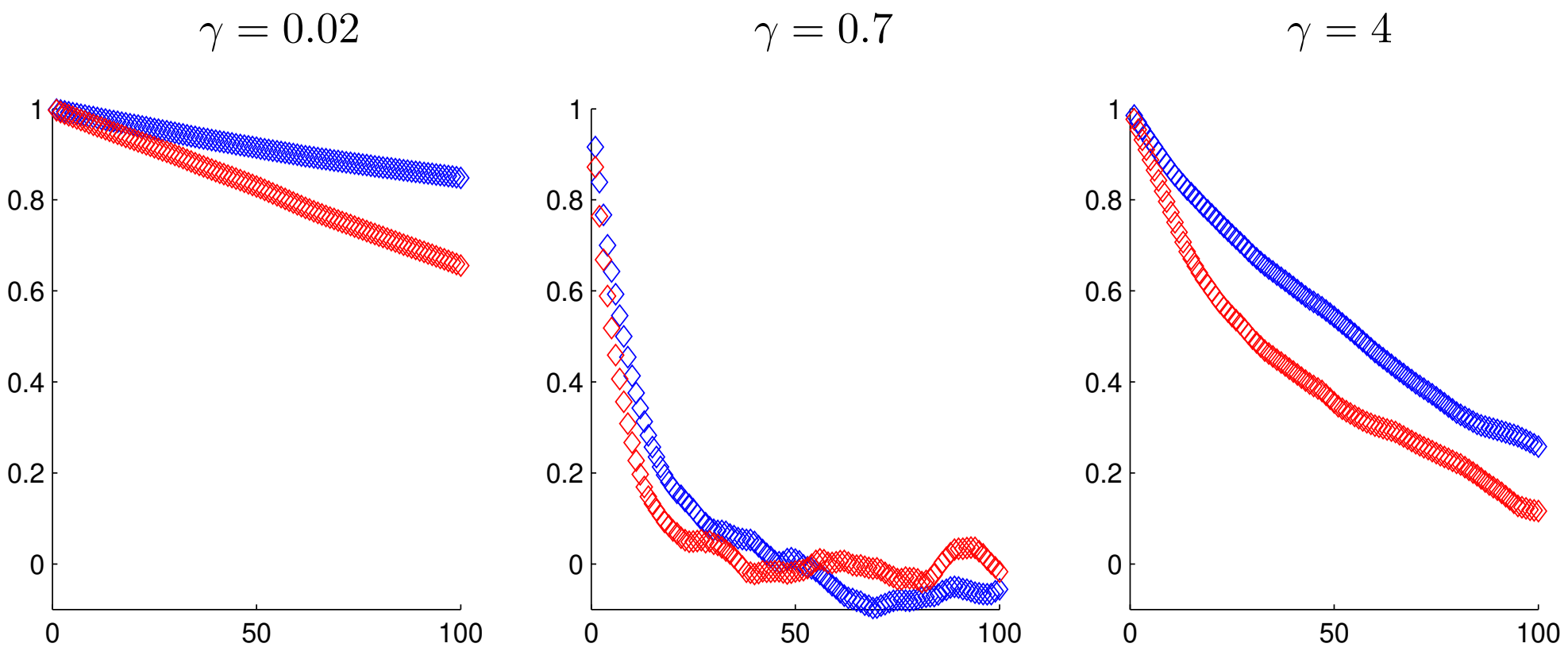
The correlation length of the sample $\{f(x_j)\}_{j=1}^{N}$ can be estimated based on the decay of the normalized autocovariance sequence of the sample.

For a white noise sample, $\gamma_k^2 \approx 0$ for any $k > 0$, where the estimate gets better as the sample, i.e., N, increases.

We test this hypothesis by drawing two white noise samples ($N = 5000$ and $N = 100000$) and plotting the function $k \mapsto \gamma_k^2$ in both cases.



Normalized autocovariance sequences for the MH example.



Normalized autocovariances for the Gibbs example; horizontal component in blue and vertical in red.