Computational inverse problems

Nuutti Hyvönen, Jenni Heino,
Juha-Pekka Puska

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

Tenth lecture, March 31, 2021.

Improper Gaussian priors

Motivation: Smoothness priors

Recall from the previous lecture that finding the *maximum a posteriori* (MAP) — or *conditional mean* (CM) — estimate for the linear inverse problem

$$Y = AX + E, \qquad Y, E \in \mathbb{R}^m,\ X \in \mathbb{R}^n,$$

where X and E are independent and Gaussian with zero mean,

$$X \sim \mathcal{N}(0, \Gamma), \qquad E \sim \mathcal{N}(0, \sigma^2 I),$$

is equivalent to minimizing the Tikhonov functional

$$T(x) \;=\; \|y - Ax\|^2 + \sigma^2 \|Rx\|^2,$$

where R satisfies $\Gamma^{-1} = R^{\mathrm{T}} R$. (The matrix R can be, e.g., the Cholesky factor of the positive definite and symmetric matrix Γ^{-1}.)

Let us then try to work our way in the opposite direction: Consider the corresponding classical linear inverse problem

$$Ax = y,$$

and let us solve it using Tikhonov regularization under the prior knowledge that $x \in \mathbb{R}^n$ represents point values of a smooth function.

We try to incorporate this extra information in the solution process by using a 'smoothness penalty term' for the Tikhonov functional:

$$T(x) = \|y - Ax\|^2 + \delta\|Lx\|^2,$$

where $L \in \mathbb{R}^{k\times n}$ is a discrete approximation of some suitable differential operator.

If you now compare the two Tikhonov functionals on the previous two slides, it seems natural that the Gaussian stochastic extension corresponding to the smoothness penalty approach would be

$$Y = AX + E,$$

with

$$X \sim \mathcal{N}(0, (L^{\mathrm{T}}L)^{-1}), \qquad E \sim \mathcal{N}(0, \sigma^2 I),$$

where $\sigma^2 = \delta$.

Unfortunately, there is a slight flaw in this logic: In order for the inverse $(L^{\mathrm{T}}L)^{-1}$ to exist — and to be positive definite — the matrix $L \in \mathbb{R}^{k \times n}$ needs to be injective, which is not always the case. (As an example, quite often $Lx = 0$ if all elements of x are the same.)

Due to this observation, we will next consider *improper densities* of the form:

$$\pi_{\mathrm{pr}}(x) \propto \exp\left(-\frac{1}{2}\|L(x-x_0)\|^2\right) = \exp\left(-\frac{1}{2}(x-x_0)^{\mathrm{T}}L^{\mathrm{T}}L(x-x_0)\right),$$

where $L \in \mathbb{R}^{k\times n}$ is a given, possible non-injective matrix.

We will show that the posterior density may be proper even if the prior is improper.

Proper posteriors corresponding to improper priors

When dealing with improper prior densities, the third theorem of the previous lecture is useless in the construction of the posterior: The prior covariance is used explicitly in the formula for the posterior covariance, but the natural candidate for the former, i.e., $(L^{\mathrm{T}}L)^{-1}$, does not typically exist.

However, recall that we also introduced alternative formulas for the posterior mean and covariance, namely

$$\Gamma_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}A)^{-1},$$

and

$$\bar{x} = \Gamma_{\mathrm{post}}(A^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}(y - e_0) + \Gamma_{\mathrm{pr}}^{-1}x_0).$$

These formulas look more promising as they involve only $\Gamma_{\mathrm{pr}}^{-1}$, not Γ_{pr}.

For simplicity let us only consider the zero mean case:

Theorem. *Consider the linear observation model* $Y = AX + E$, $A \in \mathbb{R}^{m\times n}$, *where* $X \in \mathbb{R}^n$ *and* $E \in \mathbb{R}^m$ *are mutually independent random variables, of which* E *is proper Gaussian,* $E \sim \mathcal{N}(0, \Gamma_{\mathrm{noise}})$. *Let* $L \in \mathbb{R}^{k\times n}$ *be a matrix such that* $\mathrm{Ker}(L) \cap \mathrm{Ker}(A) = \{0\}$. *Then the function*

$$x \mapsto \pi_{\mathrm{pr}}(x)\pi(y \,|\, x) \propto \exp\left(-\frac{1}{2}\left(\|Lx\|^2 + (y - Ax)^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}(y - Ax)\right)\right)$$

defines a Gaussian density over $\mathbb{R}^n$, *with the corresponding covariance and mean given by the formulas*

$$\Gamma_{\mathrm{post}} = (L^{\mathrm{T}}L + A^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}A)^{-1}, \qquad \bar{x} = \Gamma_{\mathrm{post}}A^{\mathrm{T}}\Gamma_{\mathrm{noise}}^{-1}y,$$

respectively.

Proof: Let us denote $G = L^{\mathrm{T}} L + A^{\mathrm{T}} \Gamma_{\mathrm{noise}}^{-1} A \in \mathbb{R}^{n \times n}$ and let $x \in \mathbb{R}^n$ be arbitrary. Because $\Gamma_{\mathrm{noise}}^{-1}$ is positive definite, we have

$$x^{\mathrm{T}} G x \;=\; \|Lx\|^2 + (Ax)^{\mathrm{T}} \Gamma_{\mathrm{noise}}^{-1} (Ax) \;\geq\; 0,$$

where the equality holds only if $x \in \mathrm{Ker}(L) \cap \mathrm{Ker}(A) = \{0\}$. In consequence, G is positive definite, meaning that $\Gamma_{\mathrm{post}} = G^{-1}$ is well-defined and also positive definite.

By completing the square with respect to x, the the quadratic functional in the exponent of the posterior density can be written as

$$\begin{aligned}\|Lx\|^2 + (y - Ax)^{\mathrm{T}} \Gamma_{\mathrm{noise}}^{-1} (y - Ax) &= x^{\mathrm{T}} G x - 2 x^{\mathrm{T}} A^{\mathrm{T}} \Gamma_{\mathrm{noise}}^{-1} y + y^{\mathrm{T}} \Gamma_{\mathrm{noise}}^{-1} y \\ &= (x - \bar{x})^{\mathrm{T}} G (x - \bar{x}) + c,\end{aligned}$$

where $c \in \mathbb{R}$ depends only on y, not on x, and

$$\bar{x} = G^{-1} A^{\mathrm{T}} \Gamma_{\mathrm{noise}}^{-1} y = \Gamma_{\mathrm{post}} A^{\mathrm{T}} \Gamma_{\mathrm{noise}}^{-1} y. \qquad \square$$

Exploring non-Gaussian densities

Why sampling is needed?

Remember that the CM estimate and the conditional covariance require solving integration problems involving the posterior density:

$$x_{\mathrm{CM}} = E\{x \mid y\} = \int_{\mathbb{R}^n} x\pi(x \,|\, y)dx$$

$$\mathrm{cov}(x \,|\, y) = \int_{\mathbb{R}^n} (x - x_{\mathrm{CM}})(x - x_{\mathrm{CM}})^{\mathrm{T}}\pi(x \,|\, y)dx.$$

In a non-Gaussian case, these integrals cannot typically be expressed in a closed form, and one must thus resort to numerical integration in $\mathbb{R}^n$.

Suppose that our aim is to estimate some quantity of the form

$$I = \int f(x)\pi(x)dx.$$

How about using quadrature rules? In principle, we could approximate

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

with some suitable weights $\{w_j\}$ and nodal points $\{x_j\}$. Unfortunately, if n is large, such computation is not feasible: For a quadrature rule with k discretization points per dimension, the total number of nodes is $N = k^n$. In addition, the implementation of a quadrature rule would require reliable information about the location of the ‘support’ of the probability density π.

Often it is more advisable to resort to sampling: Draw a large enough sample $\{x_j\}_{j=1}^N$ from the probability distribution corresponding to $\pi(x)$, and use these points to approximate the integral as

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

According to the Law of Large Numbers,

$$\lim_{N\to\infty} \frac{1}{N}\sum_{j=1}^{N} f(x_j) =: \lim_{N\to\infty} I_N = I$$

almost surely, i.e., the sample average converges almost surely to the expected value. Furthermore, the Central Limit Theorem states that

$$\mathrm{var}(I_N - I) \approx \frac{\mathrm{var}(f(X))}{N},$$

i.e., the discrepancy between I and I_N should go to zero like $1/\sqrt{N}$.

Markov Chain Monte Carlo

Random walk in $\mathbb{R}^n$

Random walk in $\mathbb{R}^n$ is a process of moving around by taking random steps. Elementary random walk:

1. Choose a starting point $x_0 \in \mathbb{R}^n$ and a 'step size' $\sigma > 0$. Set $k = 0$.
2. Draw a random vector $w_{k+1} \sim \mathcal{N}(0, I)$ and set $x_{k+1} = x_k + \sigma w_{k+1}$.
3. Set $k \leftarrow k + 1$ and return to step 2, unless your stopping criterion is satisfied.

The location of the random walk at time k is a realization of the random variable X_k, and we have an evolution model

$$X_{k+1} = X_k + \sigma W_{k+1}, \qquad W_{k+1} \sim \mathcal{N}(0, I).$$

The conditional density of X_{k+1}, given $X_k = x_k$, is

$$\pi(x_{k+1} \,|\, x_k) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|x_k - x_{k+1}\|^2\right) = q(x_k, x_{k+1}).$$

The function q is called the *transition kernel*. Since q does not depend on k, i.e., the step is always distributed in the same way, the kernel is called *time invariant*.

The process above defines a *chain* $\{X_k\}_{k=0}^{\infty}$ of random variables. This chain is a discrete time stochastic process. Note that

$$\pi(x_{k+1} \,|\, x_0, x_1, \ldots, x_k) = \pi(x_{k+1} \,|\, x_k),$$

i.e., the probability distribution of X_{k+1} depends on the past only through the preceding element X_k. A stochastic process with this property is called a *Markov chain*.

Example: Random walk in $\mathbb{R}^2$

A random walk model in $\mathbb{R}^2$:

$$X_{k+1} = X_k + \sigma W_{k+1}, \qquad W_{k+1} \sim \mathcal{N}(0, C), \quad C \in \mathbb{R}^{2\times 2}.$$

Since C is symmetric and positive definite, it has positive eigenvalues and allows an eigenvalue decomposition

$$C = UDU^{\mathrm{T}}.$$

Hence, the inverse of C can be written as

$$C^{-1} = UD^{-1}U^{\mathrm{T}} = (UD^{-1/2})\underbrace{(D^{-1/2}U^{\mathrm{T}})}_{=L},$$

which means that the transition Kernel can in turn be given as

$$q(x_k, x_{k+1}) = \pi(x_{k+1} \,|\, x_k) \propto \exp\left(-\frac{1}{2\sigma^2}\|L(x_k - x_{k+1})\|^2\right).$$

Consequently, the random walk model becomes

$$X_{k+1} = X_k + \sigma L^{-1}\tilde{W}_{k+1}, \qquad \tilde{W}_{k+1} \sim \mathcal{N}(0, I),$$

where we have used the fact that L is the whitening matrix of W_{k+1}.

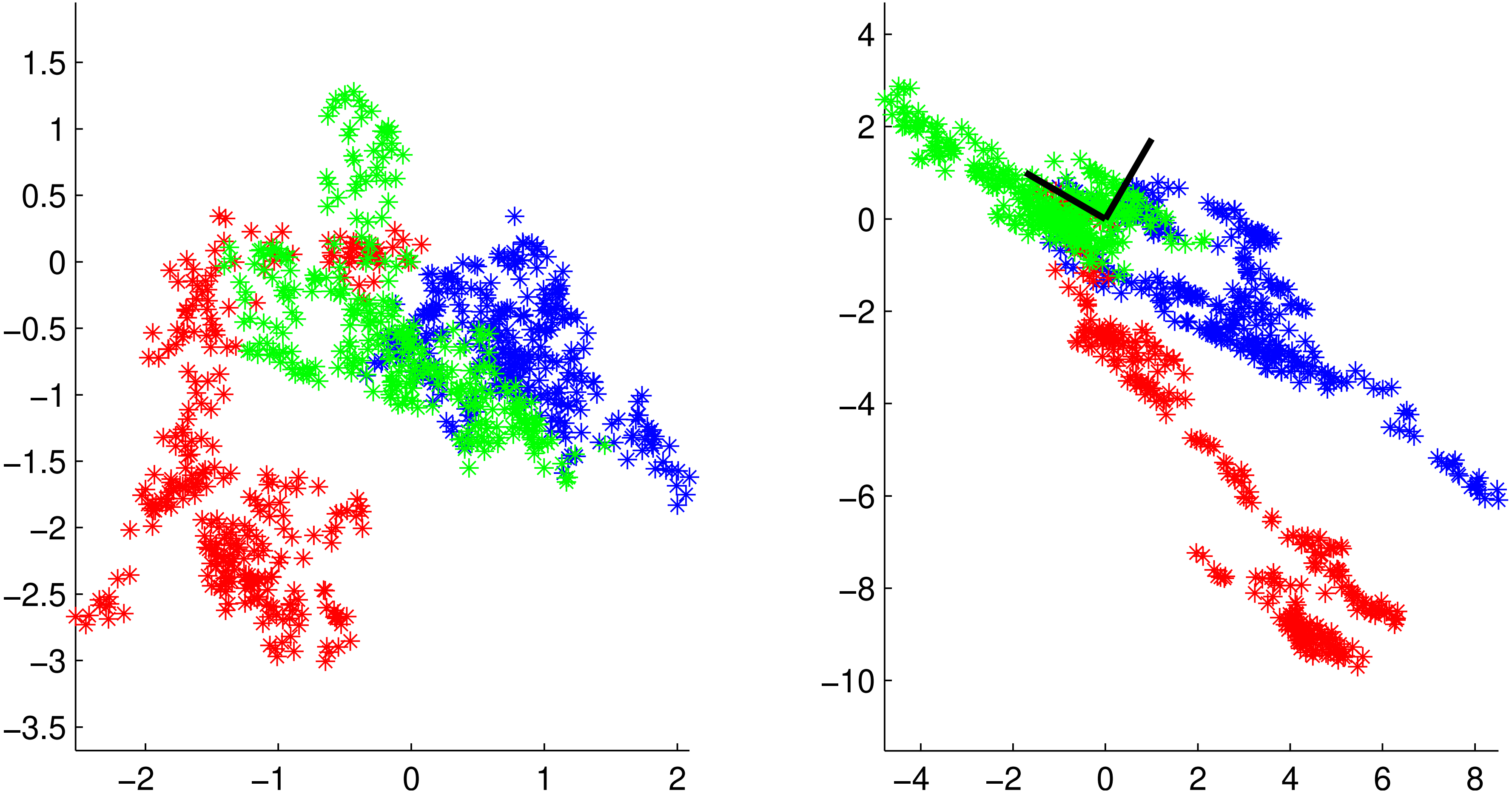
To demonstrate the effect of the covariance matrix, let

$$U = [u^{(1)}, u^{(2)}] = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \qquad \theta = \frac{\pi}{3},$$

and

$$D = \mathrm{diag}(s_1^2, s_2^2), \qquad s_1 = 1,\ s_2 = 4.$$

In the light of this random walk model, the random steps should on average have a component about four times larger in the direction of the second eigenvector e_2 than in the direction of the first eigenvector e_1.



On the left, three random walk realizations for $C = I$; on the right, three realizations for C given above. In both cases, $\sigma = 0.1$ and $x_0 = [0, 0]^{\mathrm{T}}$.

How about sampling from a given density $p(x)$?

Assume now that X is a random variable with a probability density $\pi(x) = p(x)$.

Consider an arbitrary transition kernel $q(x,y)$ that we use to generate a new random variable Y given $X = x$, that is,

$$\pi(y \,|\, x) = q(x,y).$$

The probability density of Y is found via marginalization,

$$\pi(y) = \int \pi(y \,|\, x)\pi(x)dx = \int q(x,y)p(x)dx.$$

If the probability density of Y is equal to the probability density of X, i.e.,

$$\int q(x,y)p(x)dx = p(y),$$

we say that p is an *invariant density* of the transition kernel q.

To summarize, if p is an invariant density of the transition kernel q and the random variable X obeys the density p, then the random variable Y defined via the conditional density $\pi(y \mid x) = q(x, y)$ is still distributed according to the density p. Loosely speaking, the transition defined by q does not affect the distribution of X.

This property of invariant densities and corresponding transition kernels can be put to use in sampling.

Theorem. *Let* $\{X_k\}_{k=0}^{\infty}$ *be a time invariant Markov chain with the transition kernel* q*, i.e.,*

$$\pi(x_{k+1} \,|\, x_k) \;=\; q(x_k, x_{k+1}).$$

Assume that p *is an invariant density of* q*, and that* q *satisfies some extra technical conditions (irreducibility and aperiodicity). Then, for all* $x_0 \in \mathbb{R}$ *and any Borel set* $B \subset \mathbb{R}^n$*, it holds that*

$$\lim_{N\to\infty} P\{X_N \in B \mid X_0 = x_0\} \;=\; \int_B p(x)dx.$$

Moreover, for any regular enough function f*,*

$$\lim_{N\to\infty} \frac{1}{N}\sum_{j=0}^{N} f(X_j) \;=\; \int_{\mathbb{R}^n} f(x)p(x)dx$$

almost surely.

Proof. Proof is omitted due to obvious reasons. □

Let us try to put the above theorem into practical use. Suppose that we want to sample some probability density p and happen to know that it is invariant with respect to some transition kernel q. Then, we can proceed as follows:

1. Select a starting point x_0 and set $k = 0$.
2. Draw x_{k+1} from $q(x_k, x_{k+1})$.
3. Set $k \leftarrow k + 1$ and return to step 2, unless your personal stopping criterion is satisfied.

According to the previous theorem, the sample $\{x_k\}_{k=0}^{N}$ should give a better and better representation of p as N increases.

Hence, we are facing an inverse problem: *Given a probability density p, we would like to find a kernel q such that p is its invariant density.*

Very popular technique for constructing such a transition kernel is the *Metropolis–Hastings* algorithm.

Metropolis–Hastings algorithm

Let us introduce a slightly more general Markov process: If you are currently at some $x \in \mathbb{R}^n$, either

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

Since R is a transition kernel, the mapping $y \mapsto R(x, y)$ defines a probability density, and thus

$$\int_{\mathbb{R}^n} R(x, y) dy = 1, \qquad \text{for all } x \in \mathbb{R}^n.$$

Denote by $\mathcal{A}$ the event of moving away from x and by $\neg\mathcal{A}$ the event of not moving, meaning that

$$P\{\mathcal{A}\} = 1 - r(x), \qquad P\{\neg\mathcal{A}\} = r(x).$$

What is the density of Y generated by this strategy, given $X = x$?

Let $B \subset \mathbb{R}^n$ be a Borel set and let us write

$$\begin{aligned} P\{Y \in B \mid X = x\} = P\{Y \in B \mid X = x, \mathcal{A}\}P\{\mathcal{A}\} \\ + P\{Y \in B \mid X = x, \neg\mathcal{A}\}P\{\neg\mathcal{A}\}. \end{aligned}$$

The probability of arriving in B if we happen to move:

$$P\{Y \in B \mid X = x, \mathcal{A}\} = \int_B R(x,y)dy.$$

Arriving in B without moving happens only if $x \in B$, i.e.,

$$P\{Y \in B \mid X = x, \neg\mathcal{A}\} = \chi_B(x) := \begin{cases} 1, & \text{if } x \in B, \\ 0, & \text{if } x \notin B. \end{cases}$$

To sum up, the probability of reaching B from x is

$$P\{Y \in B \mid X = x\} = (1 - r(x)) \int_B R(x,y)dy + r(x)\chi_B(x).$$

Finally, the probability of $Y \in B$ is found through marginalization:

$$\begin{aligned}
P\{Y \in B\} &= \int P\{Y \in B \mid X = x\} p(x)dx \\
&= \int p(x) \left(\int_B (1 - r(x))R(x,y)dy \right) dx + \int \chi_B(x)r(x)p(x)dx \\
&= \int_B \left(\int p(x)(1 - r(x))R(x,y)dx \right) dy + \int_B r(x)p(x)dx \\
&= \int_B \left(\int p(x)(1 - r(x))R(x,y)dx + r(y)p(y) \right) dy.
\end{aligned}$$

By definition

$$P\{Y \in B\} = \int_B \pi(y) dy,$$

and comparing this with the above formula, we see that the probability density of Y must be

$$\pi(y) = \int p(x)(1 - r(x))R(x, y)dx + r(y)p(y).$$

Our ultimate goal is to find a kernel R and a probability r such that $\pi(y) = p(y)$, that is,

$$p(y) = \int p(x)(1 - r(x))R(x, y)dx + r(y)p(y),$$

or, equivalently,

$$(1 - r(y))p(y) = \int p(x)(1 - r(x))R(x, y)dx.$$

Denote

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

and observe that, since R is a transition kernel,

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

The condition at the bottom of the previous slide can thus be written as

$$\int p(y)K(y,x)dx = \int p(x)K(x,y)dx,$$

which is called the *balance equation*. This condition is satisfied, in particular, if the integrands are equal, i.e.,

$$p(y)K(y,x) = p(x)K(x,y).$$

This condition is known as the *detailed balance equation*. The Metropolis–Hastings algorithm is simply a technique for finding a kernel K that satisfies the detailed version of the balance equation.

Start by selecting a *candidate generating kernel* $q(x, y)$, then define

$$\tilde{\alpha}(x, y) = \min \left\{ 1, \frac{p(y)q(y, x)}{p(x)q(x, y)} \right\},$$

and finally set

$$K(x, y) = \tilde{\alpha}(x, y)q(x, y).$$

A simple calculation shows that such K satisfies the detailed balance equation, i.e.,

$$p(y)\tilde{\alpha}(y, x)q(y, x) = p(x)\tilde{\alpha}(x, y)q(x, y).$$

To convince yourself, take note that for any $x, y \in \mathbb{R}^n$ either

$$\tilde{\alpha}(x, y) = \frac{p(y)q(y, x)}{p(x)q(x, y)} \qquad \text{and} \qquad \tilde{\alpha}(y, x) = 1,$$

or

$$\tilde{\alpha}(x, y) = 1 \qquad \text{and} \qquad \tilde{\alpha}(y, x) = \frac{p(x)q(x, y)}{p(y)q(y, x)}.$$

Metropolis–Hastings algorithm

The actual Metropolis–Hastings algorithm for drawing samples is as follows:

1. Choose $x_0 \in \mathbb{R}^n$. Set $k = 0$.
2. Given x_k, draw y using the transition kernel $q(x_k, y)$ of your choice.
3. Calculate the acceptance ratio,

$$\alpha(x_k, y) := \frac{p(y)q(y, x_k)}{p(x_k)q(x_k, y)}.$$

4. Flip the α-coin: Draw $t \sim \mathrm{Uniform}([0, 1])$. If $\alpha > t$, set $x_{k+1} = y$. Otherwise, stay put at x_k, i.e., set $x_{k+1} = x_k$.
5. Set $k \leftarrow k + 1$ and return to Step 2, unless your stopping criterion is satisfied.

The sample $\{x_k\}_{k=0}^{N}$ should represent p if N is large enough.