Sampling

Anita Faul

Laboratory for Scientific Computing, University of Cambridge

Problem Description

- Let $\mathcal{D}$ be the given data.
- Let $\boldsymbol{\theta}_1$ be the vector of all parameters governing the assumed distribution of $\mathcal{D}$.
- Let $\boldsymbol{\theta}_2$ the vector of all hyperparameters governing the prior distributions for the parameters $\boldsymbol{\theta}_1$.
- The posterior distribution is $f(\boldsymbol{\theta}_1|\mathcal{D}, \boldsymbol{\theta}_2) = \dfrac{p(\mathcal{D}|\boldsymbol{\theta}_1) f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)}{\int p(\mathcal{D}|\boldsymbol{\theta}_1) f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_1}$.

Problem Description

$$\int p(\mathcal{D}|\boldsymbol{\theta}_1) f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_1 = \int p(\mathcal{D}, \boldsymbol{\theta}_1|\boldsymbol{\theta}_2) d\boldsymbol{\theta}_1 = f(\mathcal{D}|\boldsymbol{\theta}_2).$$

- Expectation of the function $p(\mathcal{D}|\boldsymbol{\theta}_1)$ with respect to the distribution described by the probability density function $f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)$.
- More generally, the expectation $\mathbb{E}[h] = \int h(x) f(x) dx$ with respect to the pdf $f(x)$ for some function $h(x)$ has to be calculated.
- Let $x_1, \ldots, x_N$ be independent samples. Then the expectation is approximated by

$$\mathbb{E}[h] \approx \frac{1}{N} \sum_{n=1}^{N} h(x_n).$$

Problem Description

- If some samples are dependent on each other, the sample size N has to be increased.
- If $h(x_n)$ is large in areas where $f(x_n)$ is small and vice versa, then the expectation is dominated by the few large values of h, even though the probability density is small in this region.

Inverse Transform Sampling

Inverse transform sampling, *inverse transformation*, *inversion sampling*, *inverse probability integral transform* and *Smirnov transform*.

- Algorithm generating uniformly distributed random numbers in $(0, 1)$ available.
- Cumulative probability density function is known and invertible, $F(x) = \int_{-\infty}^{x} f(t)dt$.
- Assumption: continuous and strictly monotonically increasing in (a, b) and 0 for $x \leq a$ and 1 for $x \geq b$.
- Let y be drawn from the uniform distribution over $(0, 1)$, then there exists a unique number in (a, b) such that $F(x) = y$, i.e. $x = F^{-1}(y)$.

Inverse Transform Sampling

$$p(x \leq \hat{x}) = p(F^{-1}(y) \leq \hat{x})$$

- $F^{-1}(y) \leq \hat{x}$ if and only if $y \leq F(\hat{x})$, since F and therefore F^{-1} strictly monotonically increasing.
- Since y is from the uniform distribution on $(0, 1)$, the probability of y being less than or equal to $F(\hat{x})$ is in fact $F(\hat{x})$ itself.

$$p(x \leq \hat{x}) = F(\hat{x}).$$

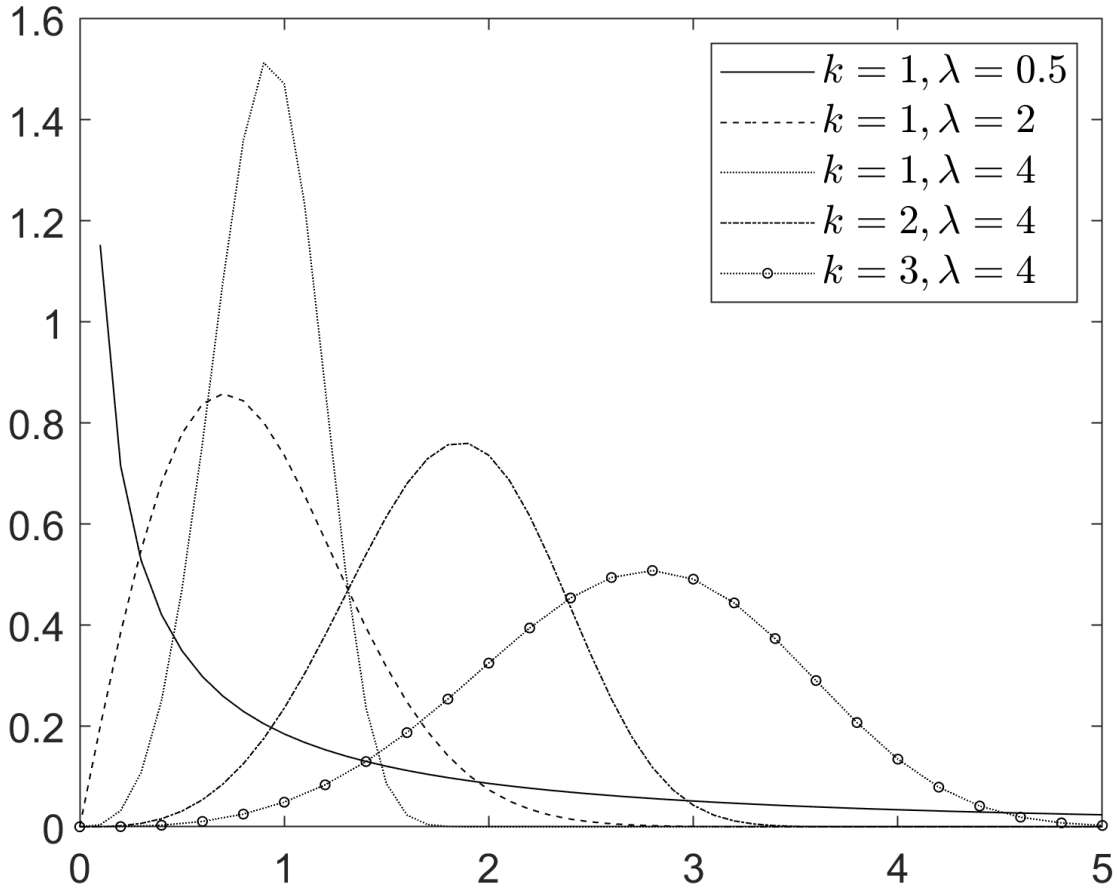
Weibull Distribution

Weibull distribution as example.

$$\text{Weibull}(x|\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \exp\left(-\left(\frac{x}{\lambda}\right)^{k}\right) & \text{for} \quad x \geq 0, \\ 0 & \text{for} \quad x < 0, \end{cases}$$

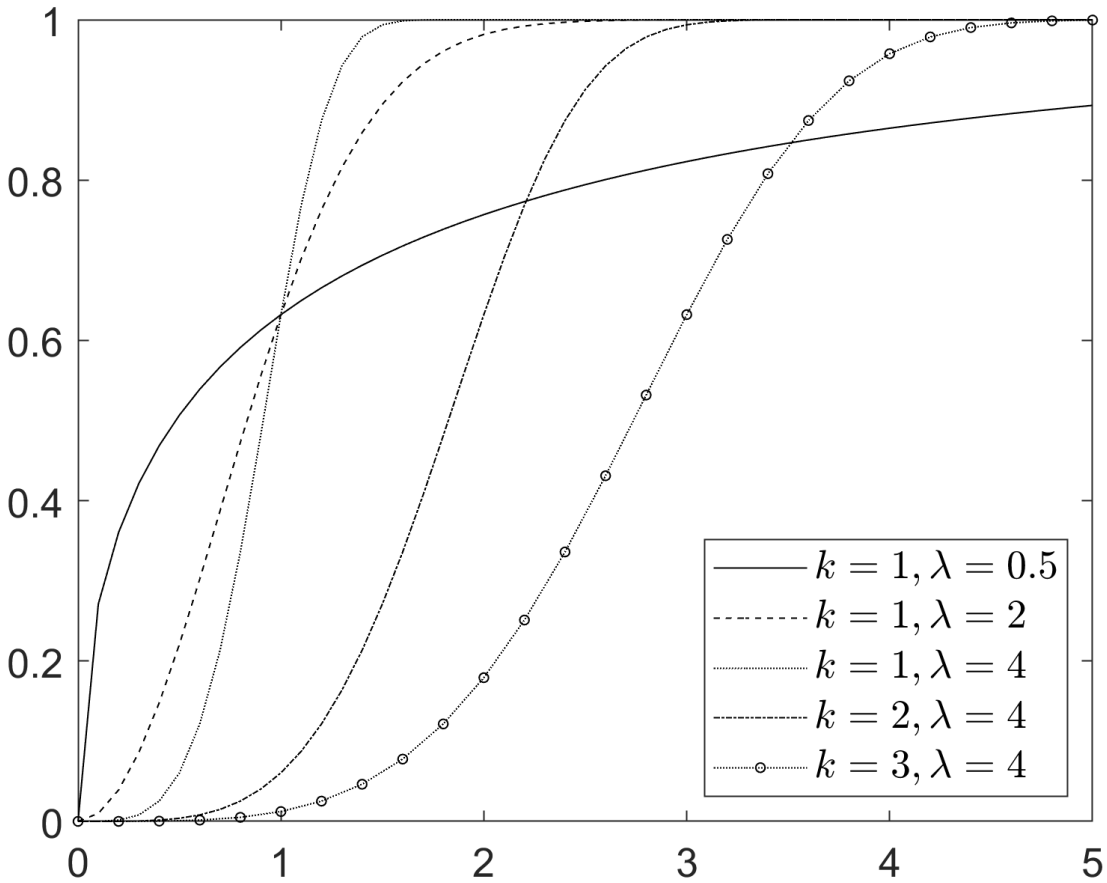
- *Scale parameter* $\lambda > 0$.
- *Shape parameter* $k > 0$.
- If $k > 1$, then the failure rate increases with time as parts are more likely to fail as time goes on.
- If $k = 1$, the failure rate is constant, the system is stable and there is no aging process.
- If $k < 1$, the failure rate decreases with time.
- Mean $\lambda\Gamma(1 + 1/k)$.
- Variance $\lambda^2[\Gamma(1 + 2/k) - (\Gamma(1 + 1/k))^2]$.

Weibull Distribution



The probability density function of the Weibull distribution for various choices of k and λ. For $k = 1$, it is the exponential probability distribution.

Weibull Distribution



The cumulative distribution function of the Weibull distribution for various choices of k and λ. For $k = 1$, it is the exponential probability distribution.

Weibull Distribution

$$
\begin{aligned}
F(x) &= \int_0^x \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} \exp\left(-\left(\frac{t}{\lambda}\right)^k\right) dt \\
&= \left[-\exp\left(-\left(\frac{t}{\lambda}\right)^k\right)\right]_0^x = 1 - \exp\left(-\left(\frac{x}{\lambda}\right)^k\right).
\end{aligned}
$$

Setting this equal to y and solving for x gives

$$x = \lambda\left[-\log(1-y)\right]^{1/k},$$

or equivalently for $z = 1 - y$

$$x = \lambda\left[-\log z\right]^{1/k}.$$

Box–Muller Transform

- Cumulative distribution function needs to be known and invertible.
- Not the case for the normal distribution.
- For x_1 and x_2 two standard normal random variables, let $x_1 = r\cos\theta$ and $x_2 = r\sin\theta$

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{x_1^2 + x_2^2}{2}\right) dydz = \int_0^{2\pi}\int_0^{\infty} r\exp\left(\frac{-r^2}{2}\right) drd\theta,$$

- θ follows the uniform distribution on the interval $(0, 2\pi)$.
- r has the probability density function $r\exp\left(-r^2/2\right)$ on $(0, \infty)$.

Box–Muller Transform

$$F(r) = \int_0^r t \exp\left(\frac{-t^2}{2}\right) dt = \left[-\exp\left(\frac{-t^2}{2}\right)\right]_0^r = 1 - \exp\left(\frac{-r^2}{2}\right).$$

- Let q be a random variable from the uniform distribution on $(0, 1)$.
- Setting $F(r) = q$ and solving for r, gives $r = \sqrt{-2\log(1-q)}$.
- Generate random standard normal variables x_1 and x_2 by drawing two variables y_1 and y_2 from the uniform distribution on $(0, 1)$ and letting

$$\begin{aligned} x_1 &= \sqrt{-2\log y_1}\cos(2\pi y_2), \\ x_2 &= \sqrt{-2\log y_1}\sin(2\pi y_2). \end{aligned}$$

- *Basic* form of the *Box–Muller transform*.

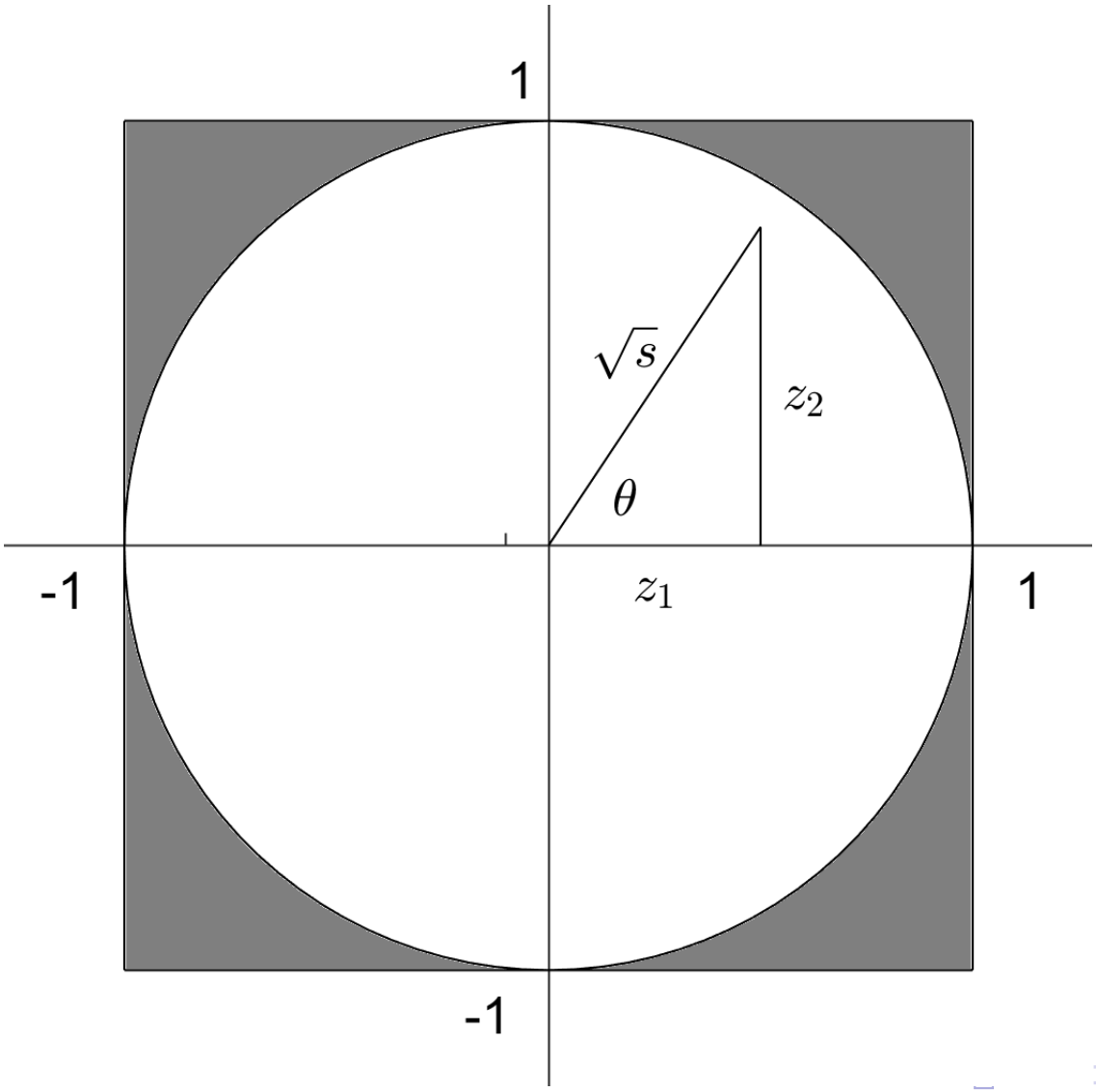
Polar Box–Muller Transform

Polar Box–Muller Transform

- Generate uniformly distributed random numbers $z_1, z_2 \in (-1, 1)$ by letting $z_i = 2y_i - 1$ for variables y_i uniformly distributed in $(0, 1)$, until $s = z_1^2 + z_2^2 < 1$.
- Probability $p(s \leq \hat{s})$ is the area of the circle with radius $\sqrt{\hat{s}}$, which is $\pi\hat{s}$, divided by the area of the unit circle which is π.
- $p(s \leq \hat{s}) = \hat{s}$, and s follows a uniform distribution on $(0, 1)$.

$$
\begin{aligned}
x_1 &= \sqrt{-2\log s}\,\overbrace{\frac{z_1}{\sqrt{s}}}^{\cos\theta} = z_1\sqrt{\frac{-2\log s}{s}}, \\
x_2 &= \sqrt{-2\log s}\,\underbrace{\frac{z_2}{\sqrt{s}}}_{\sin\theta} = z_2\sqrt{\frac{-2\log s}{s}}
\end{aligned}
$$

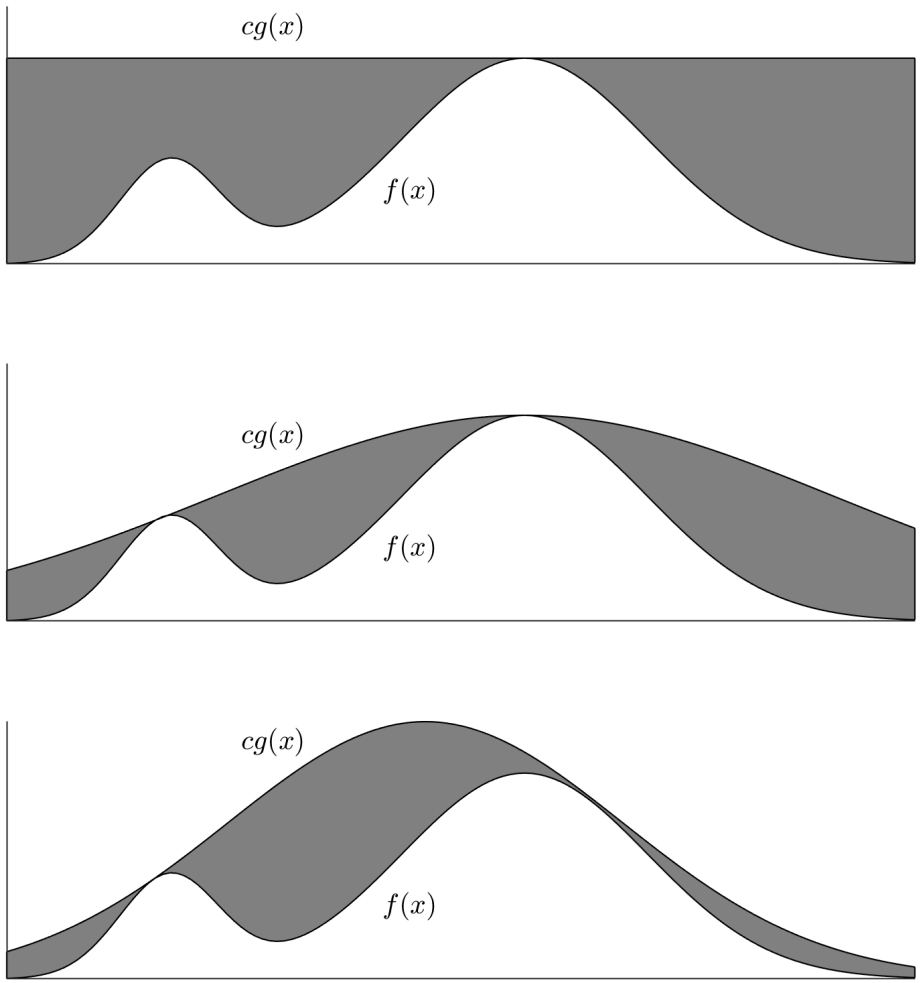
Polar Box–Muller Transform



Rejection Sampling

- $1 - \pi/4$ of samples in the square are rejected.
- More generally, the probability density function f is available and can be evaluated, but the cumulative distribution function F is not.
- Let $g(x)$ be a simpler probability density function such that $f(x) \leq cg(x)$ for all x for some finite constant $c > 1$.
- A method to draw samples from the associated cumulative distribution function G is available.
- $cg(x)$ is an *envelope* to $f(x)$.

Rejection Sampling



Rejection Sampling

Rejection sampling, a.k.a. *acceptance-rejection* method, has the steps:

1. Draw a random variable x following the distribution given by g;
2. Draw a random variable u from the uniform distribution over $(0, 1)$;
3. If $u \leq \frac{f(x)}{cg(x)}$, accept x as a sample from the distribution given by f. Otherwise return to 1.

Rejection Sampling

- Method samples uniformly points from the area under cg and discards those which fall in the shaded area between the curves of cg and f.
- The x-position of the retained points are samples from the distribution governed by f.
- Ratio of the areas under cg and f needs to be as close to 1, to reject as few samples as possible.

Rejection Sampling

- Samples follow the distribution defined by g conditioned on $u \leq \frac{f(x)}{cg(x)}$.
- The cumulative distribution function is

$$p\left(x \leq \hat{x} | u \leq \frac{f(x)}{cg(x)}\right) = \frac{p\left(u \leq \frac{f(x)}{cg(x)}, x \leq \hat{x}\right)}{p\left(u \leq \frac{f(x)}{cg(x)}\right)}.$$

- Denominator:

$$\begin{aligned} p\left(u \leq \tfrac{f(x)}{cg(x)}\right) &= \int_{-\infty}^{\infty} p\left(u \leq \frac{f(x)}{cg(x)} | x = \tilde{x}\right) p(x = \tilde{x}) d\tilde{x} \\ &= \int_{-\infty}^{\infty} \frac{f(\tilde{x})}{cg(\tilde{x})} g(\tilde{x}) d\tilde{x} = \frac{1}{c}. \end{aligned}$$

Rejection Sampling

Numerator:

$$\begin{aligned} p\left(u \leq \frac{f(x)}{cg(x)}, x \leq \hat{x}\right) &= \int_{-\infty}^{\hat{x}} p\left(u \leq \frac{f(x)}{cg(x)}, x = t\right) dt \\ &= \int_{-\infty}^{\hat{x}} p\left(u \leq \frac{f(x)}{cg(x)} | x = t\right) g(t) dt \\ &= \int_{-\infty}^{\hat{x}} \frac{f(t)}{cg(t)} g(t) dt = \frac{F(\hat{x})}{c}, \end{aligned}$$

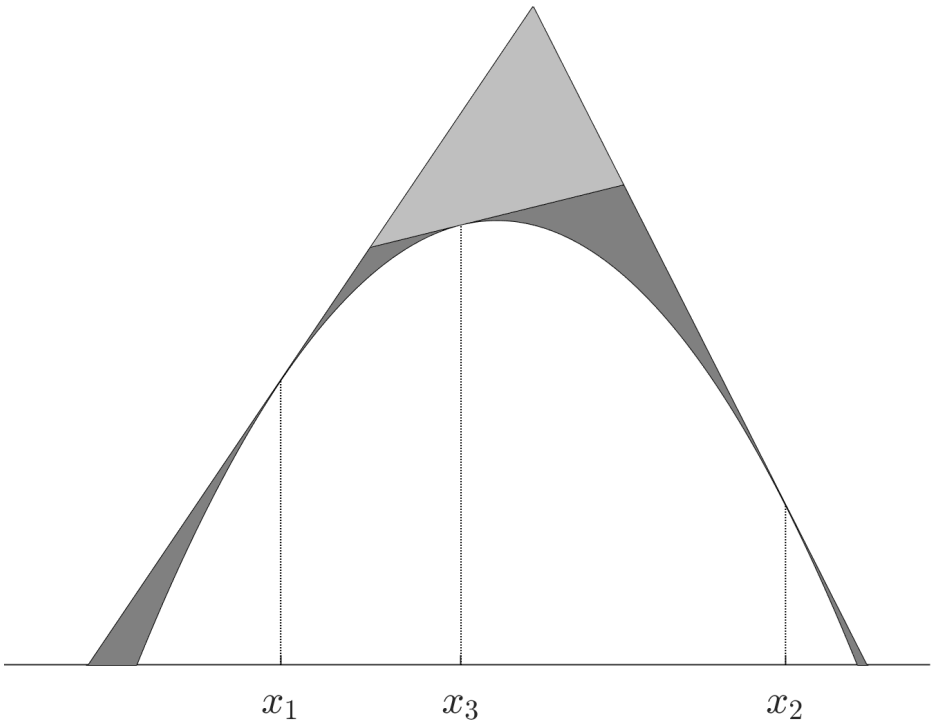
Therefore,

$$p\left(x \leq \hat{x} | u \leq \frac{f(x)}{cg(x)}\right) = F(\hat{x})$$

Rejection Sampling

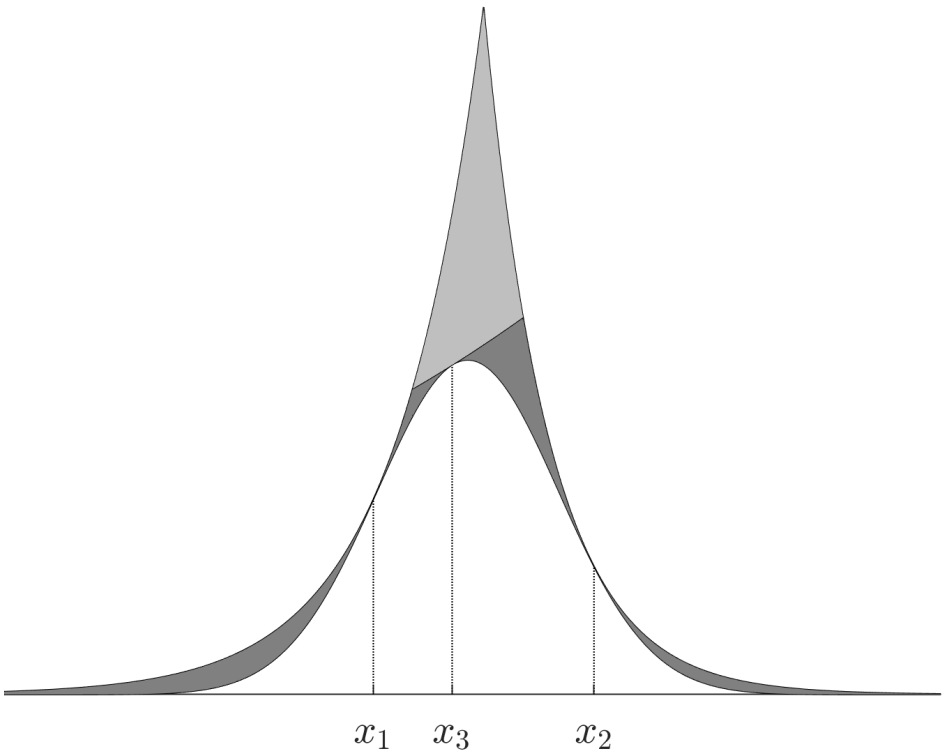
- How many attempts are necessary to draw a sample which we accept?
- Geometric distribution with $\mu = p\left(u \leq \frac{f(x)}{cg(x)}\right) = 1/c$.
- Expectation is $1/\mu = c$.
- Ratio of the areas under cg and f.
- If $f(x) = \frac{1}{c_f}\hat{f}(x)$, where $c_f = \int_{-\infty}^{\infty} \hat{f}(x)dx$ is the normalizing constant, we find $\hat{c}g(x)$ as envelope of $\hat{f}$.

Adaptive Rejection Sampling



- f *concave*.
- Tangent lines above the graph.
- Improve envelope with tangent at a rejected sample.

Adaptive Rejection Sampling



- f *log concave*, i.e. $\log(f)$ is concave.
- The piecewise linear envelop is transformed back by applying the exponential function.
- The envelope is a piecewise exponential function.

Importance Sampling

- Rejection sampling is unsuitable for high-dimensional problems due to the *curse of dimensionality*.
- Number of attempts necessary increases exponentially with the number of dimensions.
- *Importance sampling* concentrates on the regions of space considered important.

Importance Sampling

- Proposal distribution g which can be easily sampled.
- Samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are drawn from g.

$$\begin{aligned}\mathbb{E}_f[h] &= \int h(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \int h(\mathbf{x}) \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x}) d\mathbf{x} \\ &= \mathbb{E}_g[hf/g] \approx \frac{1}{N} \sum_{n=1}^{N} \frac{f(\mathbf{x}_n)}{g(\mathbf{x}_n)} h(\mathbf{x}_n),\end{aligned}$$

- $w_n = f(\mathbf{x}_n)/g(\mathbf{x}_n)$ are the *importance weights*.
- Correction to sampling from the wrong distribution.
- Where $f = g$, the correction factor is 1.
- If g is large, where f is small, the value of h needs to be reduced.
- If g is small, where f is large, the value of h needs to be magnified.

Importance Sampling

- When $f(\mathbf{x}) = \frac{1}{c_f}\hat{f}(\mathbf{x})$ and $g(\mathbf{x}) = \frac{1}{c_g}\hat{g}(\mathbf{x})$ with unknown normalizing constants $c_f = \int \hat{f}(\mathbf{x})d\mathbf{x}$ and $c_g = \int \hat{g}(\mathbf{x})d\mathbf{x}$,

$$\mathbb{E}_f[h] \approx \frac{c_g}{c_f}\frac{1}{N}\sum_{n=1}^{N}\frac{\hat{f}(\mathbf{x}_n)}{\hat{g}(\mathbf{x}_n)}h(\mathbf{x}_n).$$

- Estimating the ratio of normalizing constants as

$$\frac{c_f}{c_g} = \frac{1}{c_g}\int \hat{f}(\mathbf{x})d\mathbf{x} = \int \frac{\hat{f}(\mathbf{x})}{\hat{g}(\mathbf{x})}g(\mathbf{x})d\mathbf{x} = \mathbb{E}_g[\hat{f}/\hat{g}] \approx \frac{1}{N}\sum_{n=1}^{N}\frac{\hat{f}(\mathbf{x}_n)}{\hat{g}(\mathbf{x}_n)}.$$

- The proposal distribution should not be small, where f is large. The weighting can only make a correction, if an actual sample is drawn there.

Sampling-Importance-Resampling

Sampling-Importance-Resampling (SIR)

- Draw N samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$ from the proposal distribution g.
- Calculate importance weights $w_1, \ldots, w_N$.
- Sample from the set $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ according to the importance weights.
- A sample can feature several times in the final set.

Markov Chains

- A *Markov chain* is a series of random variables $\mathbf{x}_1, \ldots, \mathbf{x}_N$ generated one after the other.
- It is of *order* m, if

$$p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{x}_{n-2}, \ldots, \mathbf{x}_1) = p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{x}_{n-2}, \ldots, \mathbf{x}_{n-m})$$

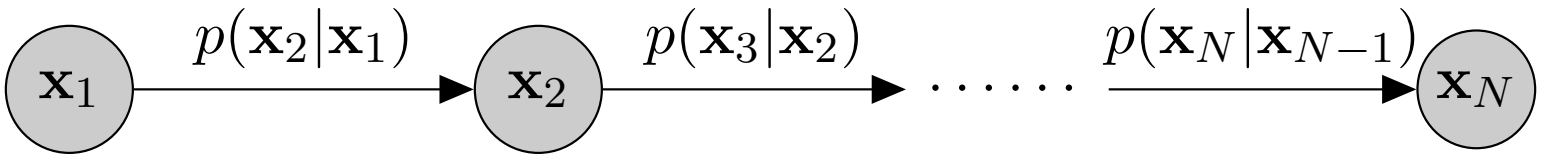
- First-order Markov chain satisfies

$$p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{x}_{n-2}, \ldots, \mathbf{x}_1) = p(\mathbf{x}_n|\mathbf{x}_{n-1}).$$

- Any Markov chain of order m can be transcribed into a first-order Markov chain by letting $\mathbf{y}_{n-m+1}$ be the tuple $(\mathbf{x}_n, \ldots, \mathbf{x}_{n-m+1})$, since then $p(\mathbf{y}_k|\mathbf{y}_{k-1}, \ldots, \mathbf{y}_1) = p(\mathbf{y}_k|\mathbf{y}_{k-1})$.

Markov Chains

- Probability distribution of the initial variable $\mathbf{x}_1$ is specified.
- *Transition probabilities*: $T_n(\mathbf{x}_{n-1}, \mathbf{x}_n) = p(\mathbf{x}_n|\mathbf{x}_{n-1})$.
- *Homogenous* if all transition probabilities are the same.



Markov Chains

- The *state space* are the possible values $\mathbf{x}_n$ can take.
- Discrete and countable, possibly finite state space.
- General, continuous state space.
- Marginal probability is

$$p(\mathbf{x}_n) = \sum_{\mathbf{x}_{n-1}} p(\mathbf{x}_n|\mathbf{x}_{n-1})p(\mathbf{x}_{n-1}) \quad \text{or} \quad p(\mathbf{x}_n) = \int p(\mathbf{x}_n|\mathbf{x}_{n-1})p(\mathbf{x}_{n-1})d\mathbf{x}_{n-1}.$$

- For a homogeneous Markov chain, the transition probabilities can be completely described by $T(\mathbf{x}, \hat{\mathbf{x}})$ for all states $\mathbf{x}$ and $\hat{\mathbf{x}}$.
- *Irreducible*, if the probability of reaching state $\mathbf{x}$ from state $\hat{\mathbf{x}}$ in a finite number of steps is non-zero for all states $\mathbf{x}$ and $\hat{\mathbf{x}}$.

Invariance

- Let f be a probability mass function or probability density function.
- It is *invariant* or *stationary* with respect to the Markov chain, if $\mathbf{x}_{n-1}$ follows the distribution, then so does $\mathbf{x}_n$ for all n.
- For example, in the degenerate case where

$$T(\mathbf{x}, \hat{\mathbf{x}}) = \begin{cases} 1 & \text{if } \hat{\mathbf{x}} = \mathbf{x}, \\ 0 & \text{otherwise}, \end{cases}$$

we have $\mathbf{x}_1 = \mathbf{x}_2 = \cdots = \mathbf{x}_N$ and any distribution is invariant.

Invariance

- For a homogeneous Markov chain, f is invariant, if

$$f(\hat{\mathbf{x}}) = \sum_{\mathbf{x}} T(\mathbf{x}, \hat{\mathbf{x}}) f(\mathbf{x}) \qquad \text{or} \qquad f(\hat{\mathbf{x}}) = \int T(\mathbf{x}, \hat{\mathbf{x}}) f(\mathbf{x}) d\mathbf{x}.$$

- The Markov chain is *reversible*, if the transition probabilities satisfy the property of *detailed balance* for all pairs of states $\mathbf{x}$ and $\hat{\mathbf{x}}$:

$$T(\mathbf{x}, \hat{\mathbf{x}}) f(\mathbf{x}) = T(\hat{\mathbf{x}}, \mathbf{x}) f(\hat{\mathbf{x}}).$$

- f is invariant, since

$$\sum_{\mathbf{x}} T(\mathbf{x}, \hat{\mathbf{x}}) f(\mathbf{x}) = \sum_{\mathbf{x}} T(\hat{\mathbf{x}}, \mathbf{x}) f(\hat{\mathbf{x}}) = f(\hat{\mathbf{x}}) \sum_{\mathbf{x}} p(\mathbf{x}|\hat{\mathbf{x}}) = f(\hat{\mathbf{x}}).$$

Base Transitions

- Transition probabilities can be constructed as linear combinations of base transition probabilities $B_1, \ldots, B_M$,

$$T(\mathbf{x}, \hat{\mathbf{x}}) = \sum_{m=1}^{M} b_m B_m(\mathbf{x}, \hat{\mathbf{x}}).$$

- $b_1, \ldots, b_M$ are known as *mixing coefficients*.
- $b_m \geq 0$ and $\sum_{m=1}^{M} b_m = 1$.
- If each of the base transitions satisfies detailed balance, then so does the linear combination.
- Often, the base transitions are chosen such that each only changes a subset of components in $\mathbf{x}$.

Finite State Space

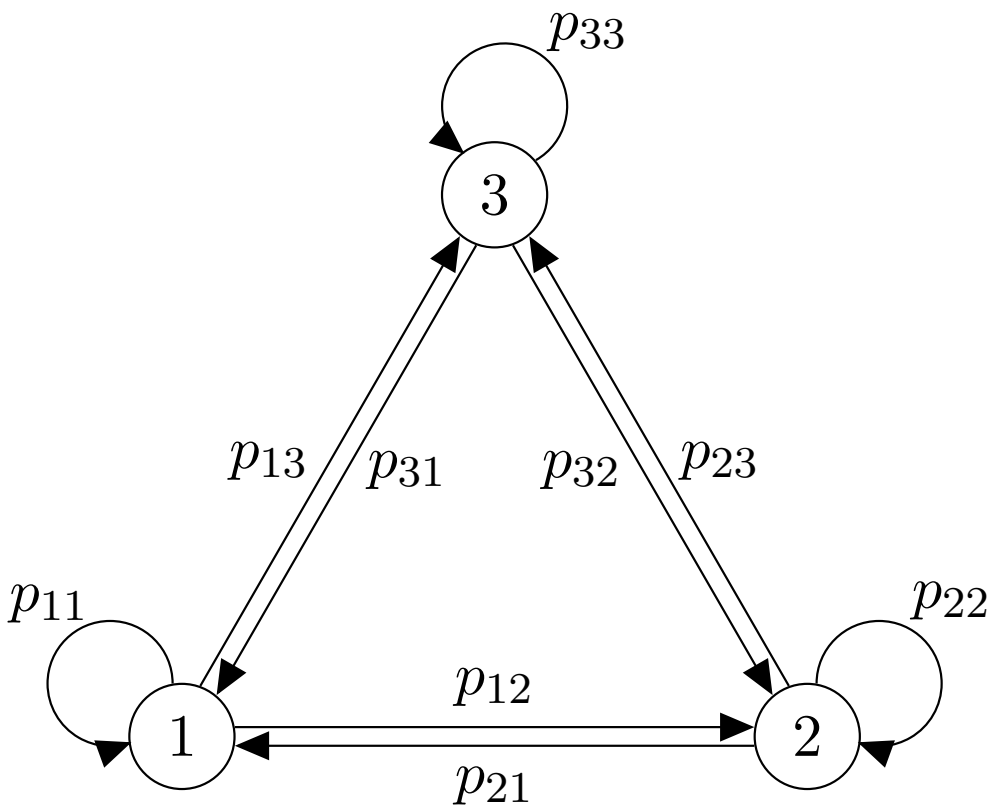
- A finite state space is represented by a 1-of-K vector.
- The transition probabilities $p_{kl} = p(x_{n+1,l} = 1 | x_{n,k} = 1)$, that is $\mathbf{x}_n$ in state k generates $\mathbf{x}_{n+1}$ in state l, are represented by the *transition matrix*

$$P = \begin{pmatrix} p_{11} & \cdots & p_{1K} \\ \vdots & \ddots & \vdots \\ p_{K1} & \cdots & p_{KK} \end{pmatrix}.$$

- Each row of P sums to 1. It is therefore a *right stochastic matrix*.
- In a *left stochastic matrix*, each column sums to 1.
- In a *doubly stochastic matrix* both columns and rows sum to 1, e.g. if P is symmetric, that is the probability of transitioning from state k to l is the same as from state l to k.

Finite State Space

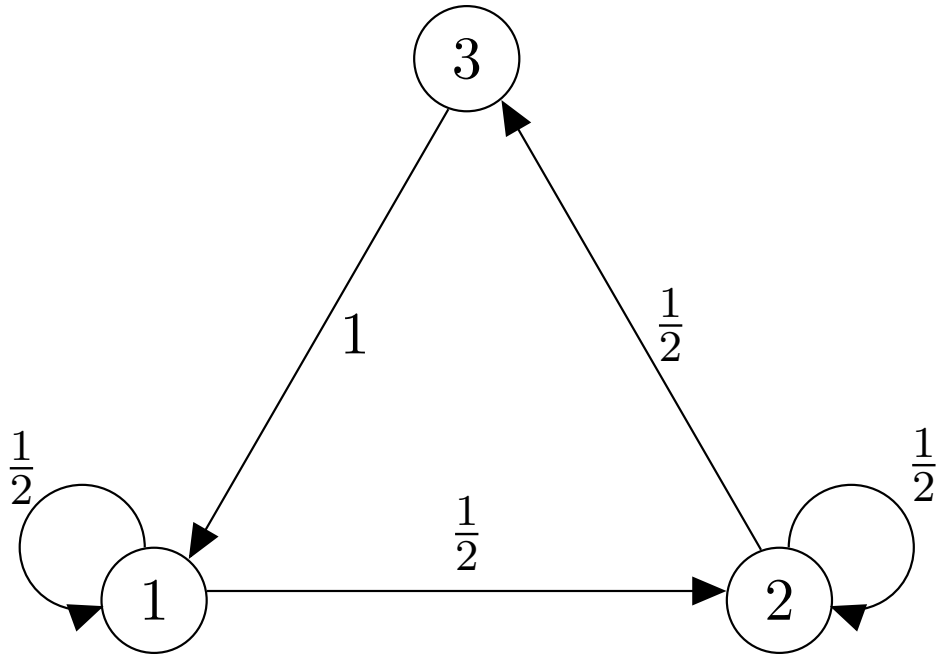
The transition matrix is depicted in a *state diagram*:



Finite State Space

- If the initial distribution $p(\mathbf{x}_1)$ is given by $\mathbf{p} = (p_1, \ldots, p_K)$, then the distribution of $\mathbf{x}_{n+1}$ is given by $\mathbf{p}P^n$.
- The (k, l) entry in P^m is the probability of transitioning from stage k to l in m steps.
- A state k has *period* m, if any return to state k occurs in multiples of m time steps.
- If $m = 1$, the state is called *aperiodic*, e.g. if the probability of transitioning to itself is non-zero.
- A Markov chain is *aperiodic*, if every state is aperiodic.
- A irreducible Markov chain only needs one aperiodic state, to be aperiodic.

Example



$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \end{pmatrix}.$$

Solving $\mathbf{f}P = \mathbf{f}$ gives the invariant distribution $\mathbf{f} = (2/5, 2/5, 1/5)$.

Example

- Constructing a Markov chain which is invariant for a given f, has more degrees of freedom.
- The property of detailed balance gives three equations,

$$\begin{aligned} p_{12}f_1 &= p_{21}f_2 \\ p_{13}f_1 &= p_{31}f_3 \\ p_{23}f_2 &= p_{32}f_3. \end{aligned}$$

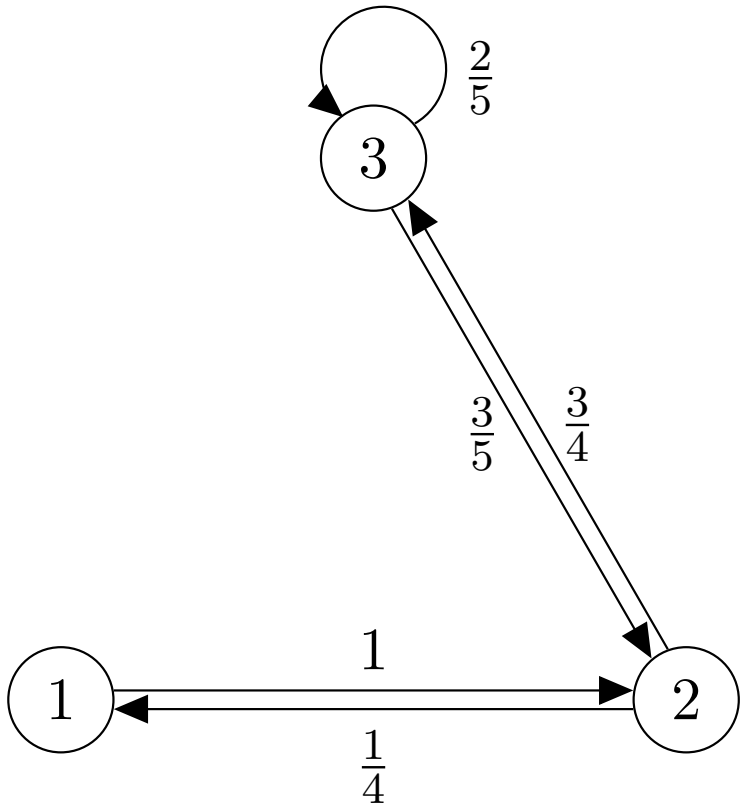
- The diagonal elements p_{11}, p_{22} and p_{33} can be determined, once the off-diagonal elements are chosen by using

$$\sum_{j=1}^{3} p_{ij} = 1.$$

Example

For $\mathbf{f} = (1/10, 2/5, 1/2)$:

$$P = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} \\ 0 & \frac{3}{5} & \frac{2}{5} \end{pmatrix}.$$



Example

$$P^{20} \approx \begin{pmatrix} 0.1006 & 0.3984 & 0.5011 \\ 0.0996 & 0.4012 & 0.4992 \\ 0.1002 & 0.3994 & 0.5004 \end{pmatrix}.$$

If P^n converges to

$$F = \begin{pmatrix} f_1 & f_2 & f_3 \\ f_1 & f_2 & f_3 \\ f_1 & f_2 & f_3 \end{pmatrix}$$

as n converges to infinity, then the distribution of $\mathbf{x}_{n+1}$ converges to

$$\mathbf{p}F = (f_1(p_1 + p_2 + p_3), f_2(p_1 + p_2 + p_3), f_3(p_1 + p_2 + p_3)) = \mathbf{f}.$$

Markov Chains

The *ergodic theorem* proves that if a finite state Markov chain is irreducible and aperiodic, the distribution of $\mathbf{x}_n$ converges to the the *equilibrium* which is the invariant distribution f irrespective of the initial distribution.

Example

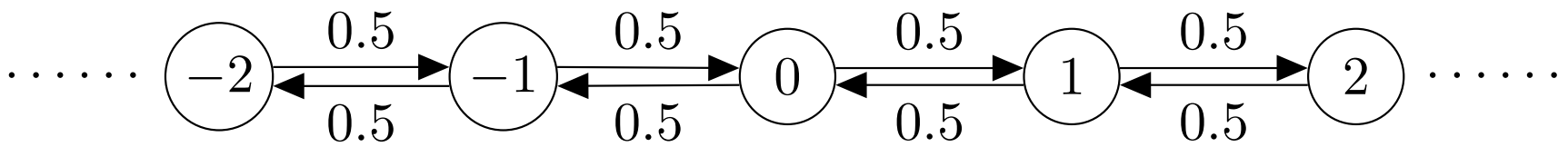
- Any matrix of the form

$$P = \begin{pmatrix} 1-(\alpha+\beta) & \alpha & \beta \\ \frac{\alpha}{4} & 1-\frac{\alpha+4\gamma}{4} & \gamma \\ \frac{\beta}{5} & \frac{4\gamma}{5} & 1-\frac{\beta+4\gamma}{5} \end{pmatrix}$$

  satisfies $\mathbf{f}P = \mathbf{f} = (1/10, 2/5, 1/2)$.
- If any two of α, β and γ are both zero, we have a 1 on the diagonal.
- It is impossible to leave that state, making it an *absorbing* state.
- If there is a non-zero probability of every state to reach that state, then the Markov chain is an *absorbing Markov chain*.
- Not irreducible.
- If at most one of α, β and γ is zero, the Markov chain is irreducible and aperiodic and can be used to generate approximate samples for the invariant distribution f.

State Space of Integers

- The *drunkards walk* is an example of a Markov chain on the state space of integers.
- The drunkard starts at the pub denoted by 0 and either steps forward $(x_{n+1} = x_n + 1)$ with probability $1/2$, or steps backward $(x_{n+1} = x_n - 1)$ also with probability $1/2$.
- The transition matrix would be infinite with zero on the diagonal and $1/2$ on the subdiagonal and superdiagonal.
- The state diagram is



State Space of Integers

- The expectation $\mathbb{E}[x_n]$ is 0, since the expectation of each individual step is that the drunkard stays put.
- $\mathbb{E}[x_n^2] = n$ implies that the distance traveled from the pub is of the order of $\sqrt{n}$.
- Ineffective in exploring the state space of the integers.
- If the random walk carries on indefinitely, it will reach each integer an infinite number of times.
- This is known as the *level-crossing phenomenon*, *recurrence* or *gambler's ruin*.

Metropolis Algorithm

- Let $f(\mathbf{x}) = \frac{1}{c_f}\hat{f}(\mathbf{x})$.
- Normalizing constant c_f not necessarily known.
- Proposal distribution $g(\mathbf{x}, \hat{\mathbf{x}})$ describes the probability of drawing $\hat{\mathbf{x}}$ when $\mathbf{x}$ is given.
- Known as *Metropolis* algorithm, if $g(\mathbf{x}, \hat{\mathbf{x}}) = g(\hat{\mathbf{x}}, \mathbf{x})$.
- Homogeneous, if $g(\mathbf{x}, \hat{\mathbf{x}})$ is independent of $\mathbf{x}$.

Metropolis Algorithm

- Candidate $\mathbf{x}^*$ is drawn from the proposal distribution and accepted with probability

$$\min\left(1, \frac{f(\mathbf{x}^*)}{f(\mathbf{x}_n)}\right) = \min\left(1, \frac{\hat{f}(\mathbf{x}^*)}{\hat{f}(\mathbf{x}_n)}\right).$$

- If the candidate sample is accepted, then $\mathbf{x}_{n+1} = \mathbf{x}^*$, otherwise $\mathbf{x}_{n+1} = \mathbf{x}_n$.
- Duplicating the sample is in contrast to rejection sampling.
- Counter acts as weight when, for example, the expectation is calculated.

Metropolis Algorithm

- If $\hat{f}$ is non-zero over the entire state space, there is always a non-zero probability of $\mathbf{x}_{n+1} = \mathbf{x}_n$ and the Markov chain is aperiodic.
- If $g(\mathbf{x}, \hat{\mathbf{x}})$ is non-zero over the entire state space, the Markov chain is irreducible.
- Successive samples are highly correlated, if $g(\mathbf{x}, \hat{\mathbf{x}})$ depends on $\mathbf{x}$.
- *Thinning* only takes every m^{th} element from the chain. For sufficiently large m, this approximates independence.
- *Burn-in* discards the first elements of a Markov chain
- Often $g(\mathbf{x}, \hat{\mathbf{x}})$ is the normal distribution with mean $\mathbf{x}$ and covariance matrix $\sigma^2\mathbf{I}$.

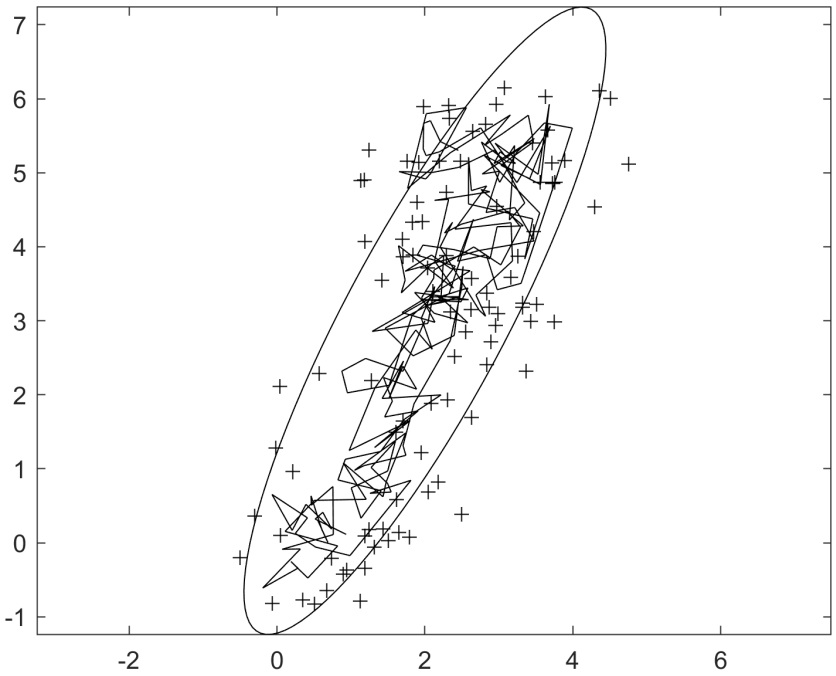
Example

- Normal target distribution with mean and variance

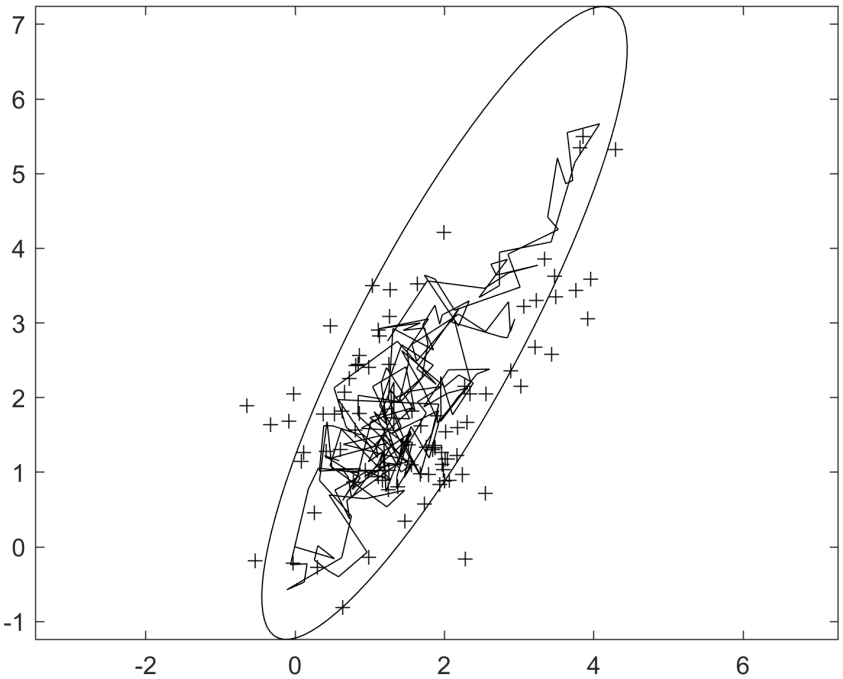
$$\boldsymbol{\mu} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 3/2 \\ 3/2 & 3 \end{pmatrix}.$$

- Eigenvalues of $\boldsymbol{\Sigma}$ are the smallest and largest variances, $\sigma^2_{\text{min}} \approx 0.2$ and $\sigma^2_{\text{max}} \approx 3.8$.

Example

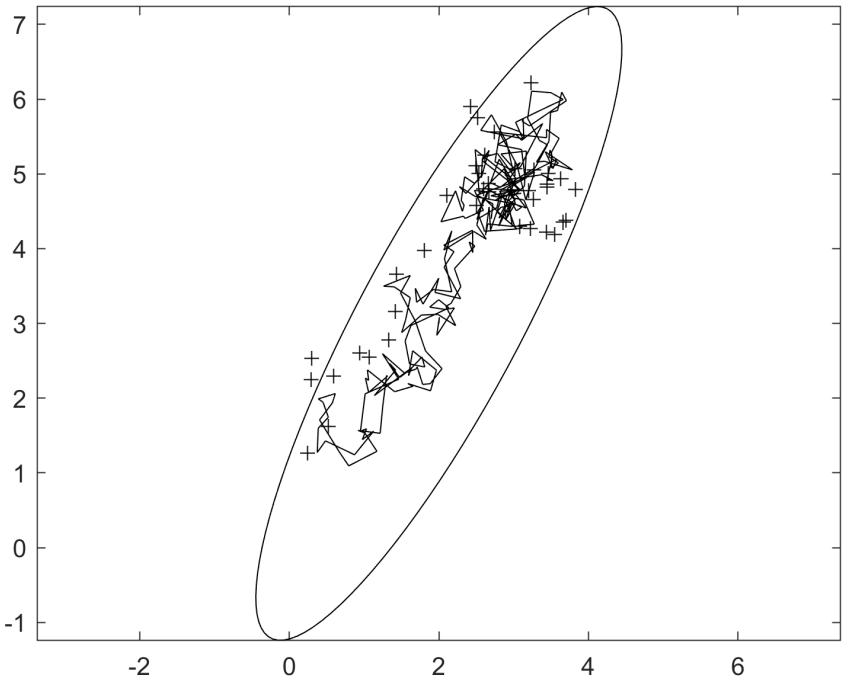


$\mathbf{x}_1 = \boldsymbol{\mu}, \sigma = \sigma_{\min}$.
About two thirds accepted.

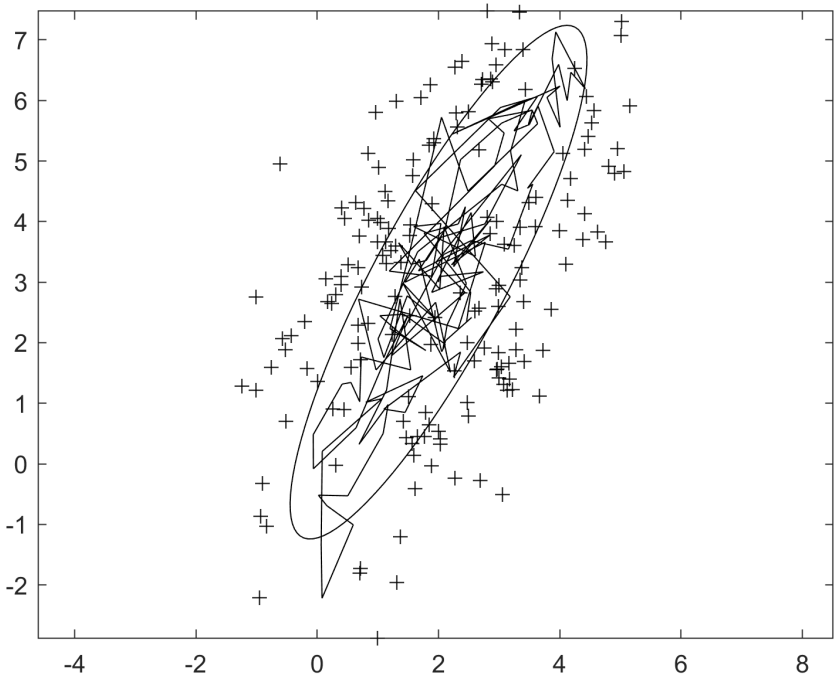


$\mathbf{x}_1 = (0,0)^T, \sigma = \sigma_{\min}$.
About two thirds accepted.

Example



$\mathbf{x}_1 = \boldsymbol{\mu}, \sigma = \sigma_{\min}/2.$
About 87% accepted.



$\mathbf{x}_1 = \boldsymbol{\mu}, \sigma = 2 * \sigma_{\min}.$
About 46% accepted.

Metropolis–Hastings Algorithm

- Proposal distribution not symmetric.
- Acceptance probability of candidate $\mathbf{x}^*$ is

$$\min\left(1, \frac{f(\mathbf{x}^*)g(\mathbf{x}^*, \mathbf{x}_n)}{f(\mathbf{x}_n)g(\mathbf{x}_n, \mathbf{x}^*)}\right) = \min\left(1, \frac{\hat{f}(\mathbf{x}^*)g(\mathbf{x}^*, \mathbf{x}_n)}{\hat{f}(\mathbf{x}_n)g(\mathbf{x}_n, \mathbf{x}^*)}\right).$$

- Transition probability from $\mathbf{x}$ to $\hat{\mathbf{x}}$ is given by

$$T(\mathbf{x}, \hat{\mathbf{x}}) = g(\mathbf{x}, \hat{\mathbf{x}}) \min\left(1, \frac{\hat{f}(\hat{\mathbf{x}})g(\hat{\mathbf{x}}, \mathbf{x})}{\hat{f}(\mathbf{x})g(\mathbf{x}, \hat{\mathbf{x}})}\right).$$

Metropolis–Hastings Algorithm

Detailed balance:

$$
\begin{aligned}
f(\mathbf{x})T(\mathbf{x},\hat{\mathbf{x}}) &= \frac{\hat{f}(\mathbf{x})}{c_f} g(\mathbf{x},\hat{\mathbf{x}}) \min\left(1, \frac{\hat{f}(\hat{\mathbf{x}})g(\hat{\mathbf{x}},\mathbf{x})}{\hat{f}(\mathbf{x})g(\mathbf{x},\hat{\mathbf{x}})}\right) \\
&= \frac{1}{c_f} \min\left(\hat{f}(\mathbf{x})g(\mathbf{x},\hat{\mathbf{x}}), \hat{f}(\hat{\mathbf{x}})g(\hat{\mathbf{x}},\mathbf{x})\right) \\
&= \frac{1}{c_f} \min\left(\hat{f}(\hat{\mathbf{x}})g(\hat{\mathbf{x}},\mathbf{x}), \hat{f}(\mathbf{x})g(\mathbf{x},\hat{\mathbf{x}})\right) \\
&= \frac{\hat{f}(\hat{\mathbf{x}})}{c_f} g(\hat{\mathbf{x}},\mathbf{x}) \min\left(1, \frac{\hat{f}(\mathbf{x})g(\mathbf{x},\hat{\mathbf{x}})}{\hat{f}(\hat{\mathbf{x}})g(\hat{\mathbf{x}},\mathbf{x})}\right) = f(\hat{\mathbf{x}})T(\hat{\mathbf{x}},\mathbf{x}).
\end{aligned}
$$

Gibbs Sampling

- *Gibbs sampling* is used, when it is easier to sample from the conditional distribution of the components of $\mathbf{x}$ than from the distribution of $\mathbf{x}$ itself.
- Let $\mathbf{x}_n = (x_{n,1}, \ldots, x_{n,D})^T$.
- Next element $\mathbf{x}_{n+1}$ is constructed in D steps by drawing $x_{n+1,d}$ sequentially from the conditional probability $p(x|x_{n+1,1}, \ldots, x_{n+1,d-1}, x_{n,d+1}, \ldots, x_{n,D})$ for $d = 1, \ldots, D$.
- Or the next component to be updated is chosen randomly.
- *Blocking* uses the conditional probabilities of sets of components.

Gibbs Sampling

- The marginal distribution of $\mathbf{x}_{-d} = (x_1, \ldots, x_{d-1}, x_{d+1}, \ldots, x_D)$ is invariant, since none of these components changes in the d^{th} step.
- The d^{th} component is sampled from the correct conditional distribution, and therefore invariant.
- The joint distribution is the product the marginal and conditional distribution by the product rule, and therefore invariant.
- If none of the conditional probabilities is zero anywhere in the state space, the ergodic theorem can be applied and the Markov chain generates samples from the desired distribution.

Gibbs Sampling

Gibbs sampling as Metropolis–Hastings:

- Let the current sample be $\mathbf{x}$. Note, this could be an element of the Markov chain or one of the intermediate steps.
- Candidate $\mathbf{x}^*$ differs from the previous sample $\mathbf{x}$ in only one component. Let this be the d^{th} component, $\mathbf{x}^*_{-d} = \mathbf{x}_{-d}$.
- Proposal distribution

$$g(\mathbf{x}, \mathbf{x}^*) = p(x^*_d | \mathbf{x}_{-d}).$$

Gibbs sampling as Metropolis–Hastings

- Acceptance probability

$$\begin{aligned}\frac{f(\mathbf{x}^*)g(\mathbf{x}^*,\mathbf{x})}{f(\mathbf{x})g(\mathbf{x},\mathbf{x}^*)} &= \frac{f(\mathbf{x}^*)p(x_d|\mathbf{x}^*_{-d})}{f(\mathbf{x})p(x^*_d|\mathbf{x}_{-d})} \\ &= \frac{f(x_1,\ldots,x_{d-1},x^*_d,x_{d+1},\ldots,x_D)p(x_d|\mathbf{x}_{-d})}{f(\mathbf{x})p(x^*_d|\mathbf{x}_{-d})}.\end{aligned}$$

- Product rules

$$\begin{aligned}f(\mathbf{x}) &= p(x_d|\mathbf{x}_{-d})p(\mathbf{x}_{-d}) \\ f(x_1,\ldots,x_{d-1},x^*_d,x_{d+1},\ldots,x_D) &= p(x^*_d|\mathbf{x}_{-d})p(\mathbf{x}_{-d})\end{aligned}$$

- Every candidate is accepted.

Gibbs Sampling