MARKOV CHAINS AND CONVERGENCE CONCEPTS

Markov chains are among the simplest stochastic processes, just one step beyond *iid* sequences of random variables. Traditionally they've been used in modelling a variety of physical phenomena, but recently interest has grown enormously due to their applicability in facilitating Bayesian computation. These lecture notes and lectures are intended to introduce the elements of markov chain theory, emphasizing the convergence to stationary distributions and the resulting simulation-based numerical methods— Gibbs sampling and, more generally, Markov Chain Monte Carlo (MC^2).

Discrete-Time Finite-State Markov Chains

Let $\mathcal{S}$ be a finite or countable set of elements $\mathcal{S} = \{s_i\}$ and let $\{X_n\}$ be a sequence of $\mathcal{S}$-valued random variables (all defined on the same probablity space $(\Omega, \mathcal{F}, \mathsf{P})$). We call X_n a *Markov chain* if for each pair of integers $m \le n$ the probability distribution of X_n, given all the variables X_i for $i \le m$, depends only on X_m. In this case the joint distribution of all the X_i depends only on the initial distribution

$$\pi_i = \mathsf{P}[X_0 = s_i]$$

and on the *transition matrix*

$$P_{ij}^{(n)} = \mathsf{P}[X_{n+1} = s_j \mid X_n = s_i]$$

We can always arrange[1] for $P_{ij}^{(n)}$ to be the same matrix P for every n, in which case we omit the superscript from the notation and call X_n a *homogeneous* Markov chain with transition matrix P. By successive conditioning the joint probability distribution of $X_0, ..., X_n$ is given by

$$\mathsf{P}[X_0 = s_{i_0}, X_1 = s_{i_1}, ..., X_n = s_{i_n}] = \pi_{i_0} P_{i_0 i_1} \cdots P_{i_{n-1} i_n}$$

and, upon summing over $i_0, ..., i_{n-1}$ (i.e., after applying the "law of total probability"), we can see that the marginal probability distribution for X_n is just

$$\begin{aligned} \mathsf{P}[X_n = s_j] &= \sum_{i_0} ... \sum_{i_{n-1}} \pi_{i_0} P_{i_0 i_1} \cdots P_{i_{n-1} j} \\ &= [\pi P^n]_j, \end{aligned}$$

where π is the row-vector with components π_i and where P^n is the n^{th} power of the matrix P_{ij}.

[1] Try to figure out how! It has something to do with changing the state space $\mathcal{S}$....

Limiting Distributions

What happens to this probability distribution $\mathsf{P}[X_n = s_j] = [\pi P^n]_j$ as $n \to \infty$? To learn the answer, let's look at powers of matrices. For now let's consider only the finite-state case... so $\mathcal{S}$ will have only finitely many (say, p) points $s_1 ... s_p$. Now P is a $p \times p$ matrix, with p (not necessarily distinct) eigenvalues λ_i given by the roots of the polynomial $f(\lambda) = \det[P - \lambda I]$. If they *are* distinct (and often even if they aren't) we can number them in order of decreasing absolute value $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_p|$ and can find p linearly independent eigenvectors $u^{(k)}$ (some might have to have complex numbers for some of the components, but that won't cause any trouble for us) satisfying the eigenvector equation

$$P\,u^{(k)} = u^{(k)}\,\lambda_k \qquad \text{i.e.} \qquad \sum_{j \leq p} P_{ij}[u^{(k)}]_j = [u^{(k)}]_i\,\lambda_k$$

for each $1 \leq k \leq p$. If we denote by U the $p \times p$ matrix whose k^{th} column is $u^{(k)}$, and by Λ the diagonal $p \times p$ matrix whose k^{th} diagonal element is λ_k, we can write the p eigenvector equations all at once in the form

$$P\,U = U\,\Lambda.$$

But if the $u^{(k)}$ are *linearly independent*, then U is invertible and we can multiply both sides of this equation on the right by U^{-1} to find the so-called *spectral representation* of P,

$$P = U\,\Lambda U^{-1}.$$

One reason this representation is important is that it simplifies *powers* of P.... when we write $P^n = P\,P \cdots P = [U\,\Lambda U^{-1}][U\,\Lambda U^{-1}] \cdots [U\,\Lambda U^{-1}]$, terms of the form $[U^{-1}\,U]$ cancel out leaving only

$$P^n = U\,\Lambda^n U^{-1}.$$

Note that the n^{th} power Λ^n is again a diagonal matrix, whose k^{th} entry is the n^{th} power $(\lambda_k)^n$ of λ_k. Now taking *limits* of n^{th} powers P^n is easy— any eigenvalue with $|\lambda_k| < 1$ vanishes $(\lambda_k)^n \to 0$, any with $\lambda_k = 1$ remains fixed $(\lambda_k)^n \to 1$, and any with $|\lambda_k| > 1$ blows up $(\lambda_k)^n \to \infty$. If each $|\lambda_k| < 1$ except for $\lambda_1 = 1$, Λ^n converges to the matrix Λ^∞ with $\Lambda_{11} = 1$ and every other $\Lambda_{ij} = 0$, and P^n converges to $U\,\Lambda^\infty U^{-1}$, a projection onto the span of $u^{(1)}$. If there exist any eigenvalues with $\lambda_k \neq 1$ and $|\lambda_k| \geq 1$, then P^n will not converge.

It turns out that there are *never* any eigenvalues $|\lambda_k| > 1$, so $P^n = U\,\Lambda^n U^{-1}$ never blows up. To see why, remember first that each row of P is a (conditional) probability vector $\mathsf{P}[X_n = s_j \mid X_{n-1} = s_i]$, so $P_{ij} \geq 0$ and $\sum_j P_{ij} = 1$. Let λ_k be any of the

eigenvalues, and let i_k be the index of the largest absolute value $U_{i_k k}$ of the k^{th} column of U (remember this column is just the eigenvector $u^{(k)}$ for λ_k). Then[2]

$$\begin{aligned}|U_{i_k k}\,\lambda_k| &= \left|\sum_{j\le p} P_{i_k j}\,U_{jk}\right| \\ &\le \sum_{j\le p} P_{i_k j}|U_{jk}| \\ &\le \sum_{j\le p} P_{i_k j}|U_{i_k k}| \\ &= |U_{i_k k}|,\end{aligned}$$

so $|\lambda_k| \le 1$ for every k.

The *columns* of the matrix U are *right* eigenvectors, together satisfying $P\,U = U\Lambda$, and it is trivial to find a right eigenvector with eigenvalue $\lambda_1 = 1$: since $\sum_j \mathsf{P}_{ij} \equiv 1$, we can use (any scalar multiple of) $u^{(1)} = [1, 1, ..., 1]'$. There must also be a matrix V whose *rows* are the *left* eigenvectors satsifying $V\,P = \Lambda\,V$, and some left eigenvector $v^{(1)}$ with $\lambda_1 = 1$— in fact, a quick look at the spectral representation shows that $V = U^{-1}$ does the trick. In most cases there is only one left- and one right- eigenvector with $|\lambda_k| = 1$ (this is actually part of "Frobenius's Theorem"; there are some conditons that must be satisfied, to rule out rotations like $P = \left[\begin{smallmatrix}0 & 1\\ 1 & 0\end{smallmatrix}\right]$).

A left eigenvector $v^{(1)}$ with eigenvalue $\lambda_1 = 1$ is special. If the elements of $v^{(1)}$ are all real and of the same sign, we can define a probability vector π^* by $\pi^*_i \equiv v^{(1)}_i / \sum_j v^{(1)}_j$ that is also a unit eigenvector satisfying $\pi^*\,P = \pi^*$, and can start a Markov chain with initial distribution $\mathsf{P}[X_0 = s_i] = \pi^*_i$. Then we can see that $\mathsf{P}[X_n = s_i] = [\pi^*\,P^n]_i = \pi^*_i$, so π^* is the distribution for *every* X_n. It's called a *Stationary Distribution* for the chain.

If Frobenius's conditions are met, so that there is only one eigenvalue with $|\lambda| = 1$, then all the rest must be strictly smaller... hence smaller than

$$\max\left[|\lambda_k|\big|\lambda_k \ne 1\right] = |\lambda_2|.$$

But now we can see that starting at *any*[3] initial distribution π, the distribution of X_n must converge to π^*, and at rate $|\lambda_2|$, in the sense that

$$\begin{aligned}\mathsf{P}[X_n = s_i] &= [\pi\,P^n]_i \\ &= \pi^*_i + \sum_{k=2}^{p} [\pi\,U]_k (\lambda_k)^n V_{ki} \\ &= \pi^*_i + \mathcal{O}(|\lambda_2|^n)\end{aligned}$$

[2] Give an explanation for each of these four equalities and inequalities.

[3] Well, *almost* any... can you see what condition must be satisfied for this to work?

Computations

Frequently we need to find π^* for some given transition matrix P or, for MC^2, to construct a P having some given stationary distribution π^*; let's see how to do it. Given P, π^* is a normalized row vector π satisfying $\pi P = \pi$, i.e., $[P' - I]\pi' = 0$. Obviously if such a π exists then the matrix $[P' - I]$ cannot be of full rank; we can take advantage of that to force our solution to be properly normalized (i.e. satisfy $\sum_i \pi_i = 1$) by replacing any row (say, the first one) of $[P' - I]$ with $[1, 1, ..., 1]$ (call the resulting matrix B) and replace the right-hand-side with the vector $[1, 0, ..., 0]'$ (call it b) to solve:

$$B\pi' = b$$

This linear system can be solved by any standard method (Gaussian elimination, QR factorization, Householder's method, etc.) as implemented in any of the standard software packages (S-Plus, LAPACK, Mathematica, Maple, etc.), giving a simple method of calculating the stationary distribution $\pi^* = \pi$ for P.

Examples

A Random Walk with Reflecting Boundary Conditions

Consider a reflecting simple random walk on the set $\mathcal{S} = \{s_1, s_2, s_3\}$, starting at $X_0 = s_2$. Thus the initial distribution and transition matrix are

$$\pi = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \qquad P = \begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{pmatrix}$$

and the equation above for the stationary distribution is given by $B\,\pi' = b$ with the matrix $B_{ij} = 1$ if $i = 1$, $[P' - I]_{ij}$ if $i > 1$ and vector $b_i = 1$ if $i = 1$, 0 if $i > 1$, i.e.

$$B\,\pi' \equiv \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 0 & 1/2 & -1 \end{pmatrix} \pi' \qquad = \qquad \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \equiv b$$

The solution is easily seen to be $\pi = [1/4, 1/2, 1/4]$, and π is indeed a stationary distribution for this chain.

The eigenvalues of P are the roots of $f(\lambda) = \det[P - \lambda I] = -\lambda^3 + \lambda$, i.e. $\lambda_1 = 1$, $\lambda_2 = -1$, and $\lambda_3 = 0$; in particular the rate of convergence is $|\lambda_2| = 1$, i.e. the distribution does *not* converge— this chain is periodic (with period 2), with X_n is almost surely equal to s_2 for even n and almost surely unequal to s_2 for odd n, while the stationary distribution would have $\mathsf{P}[X_n = s_2] = \pi^*_2 = 1/2$ for every n.

A Random Walk with Periodic Boundary Conditions

The *periodic* simple random walk starting at $s_2 = 2$ has initial distribution and transition matrix given by

$$\pi = (\,0 \quad 1 \quad 0\,) \qquad\qquad P = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{pmatrix},$$

with stationary distribution $\pi^* = \pi$ given by the solution to the equation

$$B\,\pi' \equiv \begin{pmatrix} 1 & 1 & 1 \\ 1/2 & -1 & 1/2 \\ 1/2 & 1/2 & -1 \end{pmatrix} \pi' \qquad = \qquad \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \equiv b,$$

or $\pi = [1/3, 1/3, 1/3]$; again $\pi^* = \pi$ is stationary, but this time the eigenvalues are the roots of $f(\lambda) = -\lambda^3 + 3/4\lambda + 1/4 = -(\lambda - 1)(\lambda + 1/2)^2$, so $\lambda_2 = 1/2$ and the (aperiodic) chain converges in distribution to the uniform distribution at rate $(1/2)^n$.

Designing a Markov Chain

Now suppose we *start* with a desired stationary distribution (proportional to) π— how can we find a Markov chain $\{X_n\}$ with some transition matrix P with $\pi^* = \pi$, i.e., satisfying $\pi P = \pi$? This would allow us to compute approximations to expectations $\mathsf{E}[g] = \sum_j \pi_j^* g_j$ by looking at ergodic (time-) averages

$$\mathsf{E}[g] \approx \frac{1}{N-k} \sum_{n=k+1}^{N} g(X_n)$$

(of course you're supposed to be anticipating an extension from finite state space $\mathcal{S}$ to an uncountably infinite one, where the "stationary distribution" will become a posterior distribution; more about that below). One solution was given by Hastings. Algorithmically we proceed as follows:

I. 1. Choose some transition matrix Q that is *transitive*, i.e. for every i and j satisfies $[Q^n]_{ij} > 0$ for some n, so that it is possible to go from each state $s_i \in \mathcal{S}$ to every other s_j;

II. 2. Define a function α_{ij} for $Q_{ij} > 0$ by

$$\alpha_{ij} = \min\left(1, \frac{\pi_j Q_{ji}}{\pi_i Q_{ij}}\right),$$

and for $Q_{ij} = 0$ by $\alpha_{ij} = 1$; note that $\pi_i \alpha_{ij} Q_{ij} = \pi_j \alpha_{ji} Q_{ji}$ for every i, j.

III. 3. Starting from $X_n = s_i$, draw j from Q_{ij}; with probability α_{ij} "accept" the step and set $X_{n+1} = s_j$, and otherwise stay at $X_{n+1} = s_i$.

This process is a Markov chain with transition matrix P given for off-diagonal $i \neq j$ by $P_{ij} = Q_{ij}\,\alpha_{ij} = \min(Q_{ij}, \frac{\pi_j}{\pi_i} Q_{ji})$ if $Q_{ij} > 0$, by $P_{ij} = 0$ if $Q_{ij} = 0$, and for $i = j$ by $P_{ii} = Q_{ii} + \sum_{j \neq i}(1 - \alpha_{ij}) Q_{ij} = 1 - \sum_{j \neq i} P_{ij}$. Let's consider an example.

Example

Suppose we begin with the "reflecting" random walk and want to achieve a uniform stationary distribution; using Hastings' algorithm we set:

$$\begin{aligned}
\alpha_{12} &= \min\left(1, \tfrac{\pi_2 Q_{21}}{\pi_1 Q_{12}}\right) = \min\left(1, \tfrac{1/3 \cdot 1/2}{1/3 \cdot 1}\right) = 1/2 \\
\alpha_{21} &= \min\left(1, \tfrac{\pi_1 Q_{12}}{\pi_2 Q_{21}}\right) = \min\left(1, \tfrac{1/3 \cdot 1}{1/3 \cdot 1/2}\right) = 1 \\
\alpha_{23} &= \min\left(1, \tfrac{\pi_3 Q_{32}}{\pi_2 Q_{23}}\right) = \min\left(1, \tfrac{1/3 \cdot 1}{1/3 \cdot 1/2}\right) = 1 \\
\alpha_{32} &= \min\left(1, \tfrac{\pi_2 Q_{23}}{\pi_3 Q_{32}}\right) = \min\left(1, \tfrac{1/3 \cdot 1/2}{1/3 \cdot 1}\right) = 1/2
\end{aligned}$$

and, with $P_{ij} \equiv Q_{ij}\,\alpha_{ij}$ for $i \neq j$,

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix},$$

with stationary distribution given as usual by the solution to

$$B\,\pi' \equiv \begin{pmatrix} 1 & 1 & 1 \\ 1/2 & -1 & 1/2 \\ 0 & 1/2 & -1/2 \end{pmatrix} \pi' \qquad = \qquad \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \equiv b,$$

i.e. $\pi^* = \pi = [1/3, 1/3, 1/3]$. The eigenvalues of P are the roots of $f(\lambda) = \det[P - \lambda I] = -\lambda^3 + \lambda^2 + \lambda/4 - 1/4$, i.e. $\lambda_1 = 1$, $\lambda_2 = 1/2$, $\lambda_3 = -1/2$; again the rate of convergence is $(1/2)^n$. Note that the Hastings chains are *always* aperiodic, since some $\alpha_{ij} < 1$ and hence some $Q_{ii} > 0$.

An Alarming Example

Now look at a chain on $\mathcal{S} = \{s_1, s_2, s_3\}$ that only rarely visits s_3, and stays there a long time when it does... say, one with transition matrix

$$P = \begin{pmatrix} .50 & .50 & .00 \\ .50 & .49 & .01 \\ .00 & .01 & .99 \end{pmatrix};$$

then the eigenvalues are $\lambda_1 = 1$, $\lambda_2 = .985$, and $\lambda_3 = -.005$. As expected from the description, convergence to equilibrium is slow since $|\lambda_2|$ is close to one. The same phenomenon can happen (in more subtle ways) in practical applications of MCMC, as in the case of mixtures of Dirichlet processes. The second eigenvalue of the transition matrix determines the rate of convergence, but is seldom amenable to direct computation.

Finding λ_2

In general it's easy to find (or at least estimate) the *largest* eigenvalue of a matrix, but harder to find any of the others. The idea is that "almost" any vector w chosen haphazardly will have non-zero inner product with all the eigenvectors, and

$$A^n\, w = [U\Lambda^n V]w = \sum_{k=1}^{p} u^{(k)}(\lambda_k)^n(v^{(k)}\cdot w) = u^{(1)}(\lambda_1)^n(v^{(1)}\cdot w) + \mathcal{O}(\lambda_1)^n;$$

thus if we just repeatedly set $w_{n+1} = A\,\frac{w_n}{|A\,w_n|}$, the result should converge to $w_n \to u^{(1)}$, the eigenvector for the largest eigenvalue. We *know* the eigenvalue and right eigenvector for the *largest* eigenvector of the the transition matrix P— it's just $u^{(1)} \equiv [1, 1, \cdots, 1]'$. The trick is finding the *second* largest eigenvalue $|\lambda_2|$.

But that's easy, too, once we know the left (row) eigenvector $v^{(1)}$ for $\lambda_1 = 1$— look at the matrix $B = [A - u^{(1)} \otimes v^{(1)}]$ given by $B\,x = A\,x - (v^{(1)}\cdot x)\,u^{(1)}$; the largest eigenvalue of B is the second-largest of A, so we can simultaneously estimate $v^{(1)}$ and λ_2. Can you see how to implement this?

Extensions: Dynamic Transition Kernels

Note that we started with a chosen stationary distribution π and a completely arbitrary transitive Q and, through Hastings, produced a Markov chain with π for a stationary distribution. If we had used a different Q we would have found a different P and a different chain, but still with the same stationary distribution. In fact we can use *several different* transition matrices, and can choose which one to use *dynamically* as we generate the chain, without disturbing π's special status as the invariant measure for the chain. This observation can be used to accelerate the convergence of MC^2 methods by interjecting an occasional "big jump" (say, after some number of rejected smaller jumps) intended to get one out of "traps," i.e. local maxima of the posterior density function.[4]

Extensions: Continuous State Spaces

In real-life applications of MCMC (and Gibbs in particular) the state space is seldom finite or discrete; in Bayesian applications, $\mathcal{S}$ icludes the parameter space Θ as well as any unobserved ("latent") variables and any new parameters introduced to facilitate the application (for example, if we represent a t as a scale-mixture of normals with an inverse Gamma distribution on the scale parameter, then $\mathcal{S}$ must also include the scale parameter).

The notation changes a bit: instead of a transition *matrix* P_{ij} we have a conditional probability distribution $p(x, dy)$, a measure on $\mathcal{S}$ for each fixed $x \in \mathcal{S}$. It's unusual for that to have a density and be representable as $p(x, dy) = p(x, y)\,dy$, since most

[4] Can you see how to apply this idea to the "alarming example"?

MCMC algorithms either allow X_n to stay put (e.g. Metropolis/Hastings), in which case $p(x, dy)$ includes point mass components, or restrict X_{n+1} to a lower dimensional set (e.g. Gibbs). Whether or not it has a density, this transition induces an operator $\mathcal{P}$ operating on certain functions $f : \mathcal{S} \to \mathbb{R}$ by

$$\mathcal{P}[f](x) = \int_{\mathcal{S}} f(y)\, p(x, dy)$$

which obviously fixes $f(x) \equiv 1$. Mathematically we can construct spaces of such functions $f(x)$ and search for eigen*functions* f_k and eigenvalues λ_k such that $\mathcal{P}[f_k](x) = \lambda_k f_k(x)$; again convergence to the equilibrium distribution will prodeed at rate $|\lambda_2|$. *Functional analysis* is the mathematical study of such operators and function-spaces, including their spectral analysis. In some special cases it's not much more complicated than in $\mathbb{R}^n$ (for example, when $p(x, dy)$ has a density $p(x, y)$ satisfying $\iint p^2(x, y)\, dx\, dy < \infty$), while in other cases some new things can happen... like "spectral measures" instead of discrete sets of eigenvalues. The principles are still the same, the Hastings/Metropolis algorithm still works in exactly the same way, and the convergence issues remain.

Example

Suppose we want to find a Markov chain taking values in $\mathcal{S} = [0, 1]$ with equilibrium distribution proportional to $\pi(x) = x^2(1 - x)$ (perhaps this is an unnormalized posterior distribution). Starting with a transition measure $q(x, dy) = dy$ (i.e., take a uniform draw from $\mathcal{S}$ independent of the current location) the Hastings algorithm gives

$$\alpha(x, y) = \min\left(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right) = \min\left(1, \frac{y^2(1 - y)}{x^2(1 - x)}\right)$$

and the algorithm says at each stage to draw a random $Y \sim U[0, 1]$ and "accept" it and set $X_{n+1} = Y$ with probability $\alpha(X_n, Y)$, and otherwise to leave $X_{n+1} = X_n$. The resulting transition probability measure is

$$p(x, dy) = c_x \delta_x(dy) + \min\left(1, \frac{y^2(1 - y)}{x^2(1 - x)}\right) dy$$

where c_x is determined by the requirement $\int p(x, dy) = 1$. Can you see a way to estimate the second-largest-eigenvalue $|\lambda_2|$?

Example: Gaussian Gibbs

Suppose we want to generate a sample from the posterior distribution of $\theta \in \mathbb{R}^2$, upon observing multivariate normal random variable $X \sim N[\theta, \Sigma_x]$ with uncertain mean θ and known covariance Σ_x, and that we begin with either the conjugate prior distribution $\theta \sim N[\mu_\pi, \Sigma_\pi]$ or the improper noninformative prior $\theta \sim U[\mathbb{R}^p]$. The posterior distribution is $\theta|X \sim N[\mu, \Sigma]$ with $\Sigma = \left(\Sigma_x^{-1} + \Sigma_\pi^{-1}\right)^{-1}$ and $\mu = \Sigma\left(\Sigma_x^{-1}X + \Sigma_\pi^{-1}\mu_\pi\right)$, for the conjugate prior, or $\Sigma = \Sigma_x$ and $\mu = X$, for the noninformative one; in either case the components of θ have a joint normal posterior distribution, with one-dimensional conditional distributions $\theta_i|\theta_{\bar{i}} \sim N[\mu_{i|\bar{i}}, \sigma^2_{i|\bar{i}}]$, where

$$\begin{aligned}\mu_{i|\bar{i}} &= \mu_i + (\theta_{\bar{i}} - \mu_{\bar{i}})(\Sigma_{\bar{i}\bar{i}})^{-1}\Sigma_{\bar{i}i} \\ \sigma^2_{i|\bar{i}} &= \Sigma_{ii} - \Sigma_{i\bar{i}}(\Sigma_{\bar{i}\bar{i}})^{-1}\Sigma_{\bar{i}i}\end{aligned}$$

In the bivariate case where

$$\mu = \begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix} \qquad \Sigma = \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2\end{pmatrix}$$

this reduces to

$$\begin{aligned}\theta_i|\theta_j &\sim N[\mu_{i|j}, \sigma^2_{i|j}], \text{ where} \\ \mu_{1|2} &= \mu_1 + (\theta_2 - \mu_2)(\sigma_2^{-2})(\rho\sigma_2\sigma_1) \\ &= \mu_1 + (\theta_2 - \mu_2)\frac{\rho\sigma_1}{\sigma_2} \\ \sigma^2_{1|2} &= \sigma_1^2 - (\rho\sigma_1\sigma_2)(\sigma_2^{-2})(\rho\sigma_2\sigma_1) \\ &= \sigma_1^2(1-\rho^2) \\ \mu_{2|1} &= \mu_2 + (\theta_1 - \mu_1)\frac{\rho\sigma_2}{\sigma_1} \\ \sigma^2_{2|1} &= \sigma_2^2(1-\rho^2)\end{aligned}$$

Given an infinite sequence Z_i of *iid* $N[0,1]$ random variables and an initial value (or distribution) for $\theta_2^{(0)}$, we can implement the Gibbs Sampling algorithm by constructing a sequence $\theta^{(n)}$ of variables by the recursion relations (for $n \geq 1$)

$$\begin{aligned}\theta_1^{(n)} &= \mu_1 + (\theta_2^{(n-1)} - \mu_2)\frac{\rho\sigma_1}{\sigma_2} + \sigma_1\sqrt{1-\rho^2}\ Z_{2n-1} \\ \theta_2^{(n)} &= \mu_2 + (\theta_1^{(n)} - \mu_1)\frac{\rho\sigma_2}{\sigma_1} \quad + \sigma_2\sqrt{1-\rho^2}\ Z_{2n}.\end{aligned}$$

Upon subtracting μ_i and substituting, we have

$$\begin{aligned}
(\theta_2^{(n)} - \mu_2) &= (\theta_1^{(n)} - \mu_1)\frac{\rho\sigma_2}{\sigma_1} + \sigma_2\sqrt{1-\rho^2}\ Z_{2n} \\
&= \left((\theta_2^{(n-1)} - \mu_2)\frac{\rho\sigma_1}{\sigma_2} + \sigma_1\sqrt{1-\rho^2}\ Z_{2n-1}\right)\frac{\rho\sigma_2}{\sigma_1} + \sigma_2\sqrt{1-\rho^2}\ Z_{2n} \\
&= (\theta_2^{(n-1)} - \mu_2)\rho^2 + \sigma_2\sqrt{1-\rho^2}\ (Z_{2n} + \rho Z_{2n-1}) \\
&= (\theta_2^{(n-2)} - \mu_2)\rho^4 + \sigma_2\sqrt{1-\rho^2}\ (Z_{2n} + \rho Z_{2n-1} + \rho^2 Z_{2n-2} + \rho^3 Z_{2n-3}) \\
&= (\theta_2^{(0)} - \mu_2)\rho^{2n} + \sigma_2\sqrt{1-\rho^2}\ \sum_{k=1}^{2n} Z_k\rho^{2n-k}, \text{ so} \\
\theta_2^{(n)} &\sim N[\mu_2 + (\theta_2^{(0)} - \mu_2)\rho^{2n},\ \sigma_2^2(1-\rho^{2n})].
\end{aligned}$$

Thus both the mean and variance of $\theta_i^{(n)}$ converge geometrically at rate ρ^2, as $n \to \infty$, to their limiting values $\theta_i^{(\infty)} \sim N[\mu_i, \sigma_i^2]$.

Example: Overparametrized Model

Now suppose we model $X \sim N[\theta_1 + \theta_2, \sigma^2]$ for fixed variance σ^2 but unknown vector θ (θ_2 might represent the bias in some measuring instrument, for example, while θ_1 and X represent the true and measured quantities, respectively). The log likelihood function

$$\ell(\theta_1, \theta_2) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(X - \theta_1 - \theta_2)^2}{2\sigma^2}$$

is constant along lines of the form $\theta_1 + \theta_2 = c$ so there is no unique maximum likelihood estimate $\hat\theta = (\hat\theta_1, \hat\theta_2)$, but the model is perfectly amenable to a Bayesian analysis with a proper prior distribution. What if an *improper* prior is used— for example, the "noninformative" uniform prior $\pi(d\theta_1,\ d\theta_2) = d\theta_1\, d\theta_2$?

Although the *joint* distribution of θ_1 and θ_2 is improper, the *conditional* distributions are well-defined, proper, and trivial to calculate: $\theta_i|\theta_j \sim N[X - \theta_j, \sigma^2]$ for $i \neq j$. A naïve investigator could implement the Gibbs procedure just as before, beginning with an *iid* sequence $Z_i \sim N[0,1]$ and an initial value (or distribution) for $\theta_2^{(0)}$, by constructing a sequence $\theta^{(n)}$ of variables by the recursion relations (for $n \geq 1$)

$$\begin{aligned}
\theta_1^{(n)} &= X - \theta_2^{(n-1)} + \sigma\ Z_{2n-1} \\
\theta_2^{(n)} &= X - \theta_1^{(n)} \quad + \sigma\ Z_{2n}.
\end{aligned}$$

Now when we substitute we find

$$\begin{aligned}
\theta_2^{(n)} &= X - \left(X - \theta_2^{(n-1)} + \sigma\ Z_{2n-1}\right) + \sigma\ Z_{2n} \\
&= \theta_2^{(n-1)} + \sigma\ \left(Z_{2n} + Z_{2n-1}\right) \\
&= \theta_2^{(0)} + \sigma\ \sum_{k=1}^{2n} Z_k \\
&\sim\ N[\theta_2^{(0)}, 2n\sigma^2]
\end{aligned}$$

The sequences $\theta_i^{(n)}$ do *not* converge in distribution now— in fact, they are simple Gaussian random walks and so wander off to infinity as $n \to \infty$, despite the fact that the Gibbs procedure is straightforward to implement and reports no errors.