Markov Chain Monte Carlo (MCMC)

Dependent Sampling

Suppose we wish to sample from a density \( \pi \), and we can evaluate \( \pi \) as a function but have no means to directly generate a sample.

- Rejection sampling can be used if a trial density \( f \) can be found where \( \pi / f \) has a reasonable bound. When the dimension is high this is difficult or impossible.
- Rejection sampling, the inversion method, and the density transform method all produce independent realizations from \( \pi \). If these methods are inefficient or difficult to implement, we can drop independence as a goal, aiming instead to generate a dependent sequence

\[
X_1, X_2, \ldots
\]

such that the marginal distribution of each \( X_i \) is \( \pi \).
- Going a step further, we may allow the marginal distribution \( \pi_k(X_k) \) to be different from \( \pi \), but converge to \( \pi \) in some sense: \( \pi_k \to \pi \).
- By relaxing the goal of the problem, it becomes possible to overcome the key problem of rejection sampling – that it does not adapt as information about \( \pi \) is built up through repeated evaluations of \( \pi \) at the sampled points.
- A practical framework for constructing dependent sequences satisfying these goals is provided by discrete time Markov chains.

Review of Markov Chains

- A discrete time Markov chain is a sequence of random variables \( X_0, X_1, \ldots \) such that

\[
P(X_{n+1} \in A | X_0, \ldots, X_n) = P(X_{n+1} \in A | X_n) \equiv P(X_n, A).
\]

The function \( P \) is called the transition kernel.
- A probability distribution \( \pi \) is called the invariant distribution for \( P \) if

\[
\pi(A) = \int P(x, A)\pi(x)dx,
\]

for all measurable sets \( A \), in which case we may write \( \pi P = \pi \). More generally this relationship may hold for a non-finite measure \( \pi \), in which case \( \pi \) is called the invariant measure for \( P \).
• Suppose that $P$ has a density, denoted $q(x, y)$. This means that

$$P(x, A) = \int_A q(x, y)dy.$$ 

A sufficient condition for $\pi$ to be the invariant distribution for $P$ is detailed balance:

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

The proof is a simple application of Fubini’s theorem.

• The distribution $\pi$ is the equilibrium distribution for $X_0, X_1, \ldots$ if

$$P^n(X_0, A) \equiv P(X_n \in A|X_0)$$

satisfies

$$\lim_{n \to \infty} P^n(x, A) = \pi(A),$$

for all measurable sets $A$ and for $\pi$ almost-every $x$.

• Key definitions:

  – Irreducible: The Markov chain given by transition kernel $P$ with invariant distribution $\pi$ is irreducible if for each measurable set $A$ with $\pi(A) > 0$, for each $x$ there exists $n$ such that $P^n(x, A) > 0$.

  – Recurrent: The Markov chain given by transition kernel $P$ with invariant distribution $\pi$ is recurrent if for every measurable set $B$ with $\pi(B) > 0$,

    $$P(X_1, X_2, \ldots \in B \text{ i.o.}|X_0) > 0 \quad \forall X_0$$

    and

    $$P(X_1, X_2, \ldots \in B \text{ i.o.}|X_0) = 1$$

for $\pi$ almost-every $X_0$.

  – The Markov chain given by transition kernel $P$ with invariant distribution $\pi$ is Harris recurrent if

    $$P(X_1, X_2, \ldots \in B \text{ i.o.}|X_0) = 1 \quad \forall X_0.$$

• Fact: A Markov transition kernel $P$ that is $\pi$-irreducible and has an invariant measure is recurrent. A recurrent kernel $P$ with a finite invariant measure is positive recurrent, otherwise it is null recurrent.
• **Periodic:** A Markov transition kernel $P$ is periodic if there exists $n \geq 2$ and a sequence of nonempty, disjoint measurable sets $A_1, A_2, \ldots, A_n$ such that $\forall x \in A_j$ ($j < n$), $P(x, A_{j+1}) = 1$, $\forall x \in A_j$, $P(x, A_1) = 1$ $\forall x \in A_n$.

• **Total Variation norm:** For any measure $\lambda$, the total variation (TV) norm is

\[
\|\lambda\| = \sup_A \lambda(A) - \inf_A \lambda(A),
\]

where the supremum and infimum are taken over all $\lambda$–measurable sets. The distance between two probability measures $\pi_1$ and $\pi_2$ can be assessed using $\|\pi_1 - \pi_2\|$.  

• **Theorem:** Let $P$ be a Markov transition kernel with invariant distribution $\pi$, and suppose that $P$ is $\pi$ irreducible. Then $P$ is positive recurrent, and $\pi$ is the unique invariant distribution of $P$. If $P$ is aperiodic, then for $\pi$ almost-every $x$

\[
|P^n(x, A) - \pi(A)| \to 0
\]

for all measurable sets $A$ ($P^n$ converges to $\pi$ in the TV norm). If $P$ is Harris recurrent, then the convergence occurs for all $x$.

• In words, we can start at essentially any $X$, run the chain for a long time, and the final draw has a distribution that is approximately $\pi$. The “long time” is called the burn in, and it is generally impossible to know how long this needs to be in order to get a good approximation.

• We don’t need independence for the most common application of Monte Carlo studies: estimating $E\pi f$: the mean of a function $f$ with respect to the invariant distribution $\pi$. If chain is *uniformly ergodic*  

\[
\sup_x \|P^n(x, \cdot) - \pi\| \leq Mr^n
\]

for some $M \geq 0$ and $r < 1$, there is a central limit theorem: let $\bar{f}_n = \sum_{j=1}^n f(X_j)/n$ denote the sample mean. Then $\sqrt{n}(\bar{f}_n - E\pi f)$ converges weakly to a normal distribution with mean 0.

**Methods for generating chains with a given equilibrium distribution**

• We have a density $\pi$ that we can evaluate but not sample from. How do we construct a markov chain $P$ with invariant distribution $\pi$?

• **Gibbs sampling:** Suppose the random variable $X = (X_1, \ldots, X_d)$ can be partitioned into blocks, and we can sample from each of the conditional distributions $P(X_i|\{X_j, j \neq i\})$. These transitions are invariant with respect to $P(X_1, \ldots, X_d)$. Consider the case $X = (X_1, X_2)$, and for a set $A$, let $A_{X_1} = \{X_2|(X_1, X_2) \in A\}$. Under the proposal distribution $P(X_2|X_1)$:
\[
\int P((X_1, X_2), A) \pi(X_1, X_2) dX_1 dX_2 = \int \int_{A_{X_1}} \pi(y|X_1) dX_1 \pi(X_2|X_1) dX_1 dX_2 \\
= \int \int_{A_{X_1}} \pi(y|X_1) \pi(X_1) dX_1 dX_2 \\
= \pi(A).
\]

The Gibbs chain will in general only be irreducible and aperiodic if we combine all conditional distributions in random order.

- The Metropolis (or Metropolis-Hastings) algorithm. Suppose we have a transition density \( q(x, y) \) \( (P(X_{n+1} \in A|X_n = x) = \int_A q(x, y) dy) \). Define the Metropolis-Hastings ratio by

\[
\alpha(x, y) = \min\left\{ \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1 \right\} \quad \pi(x)q(x, y) > 0 \\
= 1 \quad \pi(x)q(x, y) = 0.
\]

The kernel \( q(x, y) \) will roughly play the role that the trial density plays in rejection sampling. Suppose we are at \( X_n \). Sample \( Y \) from \( q(X_n, y) \). With probability \( \alpha(X_n, Y) \) set \( X_{n+1} = Y \), otherwise set \( X_{n+1} = X_n \). The transition kernel can be represented as

\[
p(x, y) = q(x, y)\alpha(x, y) + r(x)\delta_x
\]

where \( \delta_x \) is the point mass at \( x \), and

\[
r(x) = 1 - \int q(x, y)\alpha(x, y) dy
\]

is the marginal probability of remaining at \( x \).

The resulting chain satisfies detailed balance, and hence has invariant distribution \( \pi \).

Only density evaluations are required, and unknown normalizing constants cancel.

**General Metropolis-Hastings Schemes**

**Gibbs sampling** If \( q(x, y) = \pi(y|x) \) then \( \alpha(x, y) \equiv 1 \), so we accept every move. This shows that Gibbs sampling is a special case of the Metropolis-Hastings algorithm.

**Random Walks** Suppose \( q(x, y) \) is a random walk: \( q(x, y) = q^*(y-x) \) for some distribution \( q^* \). The Metropolis-Hastings ratio is \( \min(\pi(y)q^*(x-y)/\pi(x)q^*(y-x), 1) \) (or 1 if \( \pi(x)q^*(y-x) = 0 \)). If \( q^* \) is symmetric, then the Metropolis-Hastings ratio reduces to \( \min(\pi(y)/\pi(x), 1) \) (or 1 if \( \pi(x) = 0 \)).
An important practical issue with random walk chains is the scaling of the random walk density $q^\ast$. If the variance of $q^\ast$ is too large, most steps will move into the tails of the distribution and will be rejected. On the other hand if the variance is too small, the autocorrelation in the sequence of accepted draws will be high and convergence to the target distribution will be slow. Roberts and Rosenthal (2001) have shown that an acceptance rate of around 0.23 is optimal when the target density is the product of identical marginal density functions.

**Independence chains** If $q(x, y) = q(y)$, so that the proposals are an iid sequence, the Metropolis-Hastings ratio is $\min(\pi(y)q(x)/\pi(x)q(y), 1)$ (or 1 if $\pi(x)q(y) = 0$).

This is similar to rejection sampling, however the acceptance probabilities are scaled by $q(x)/q(y)$ rather than by $1/c$, and points are repeated rather than rejected if the proposal fails. Two interesting special cases are:

1. $q(x) \propto 1$: if the proposal distribution is uniform, it is even more similar to rejection sampling, since the Metropolis-Hastings ratio becomes $\min(\pi(y)/\pi(x), 1)$.
2. If $q(y) = \pi(y)$, then the Metropolis-Hastings ratio is identically 1, and we are sampling directly from the target density.

**Rejection sampling chains** Suppose we want to use rejection sampling from a trial density $f$ to produce realizations from $\pi$. We can use the Metropolis-Hastings algorithm to rescue rejection sampling when we cannot find a good bound for $\pi/f$. We conjecture, but are not certain, that $\pi/f < c$. If we carry out rejection sampling we are actually drawing from the density proportional to $\tilde{f} = \min(\pi, cf)$.

Suppose we use an independence chain generated by $\tilde{f}$ to drive the Metropolis-Hastings algorithm. Thus we have $q(x, y) \propto \min(\pi(y), cf(y))$. Let $S$ denote the subset of the sample space where the bound $\pi/f < c$ holds. We have four possibilities:

1. $x, y \in S \Rightarrow \alpha = 1$.
2. $x \in S$, $y \notin S \Rightarrow \alpha = 1$.
3. $x \notin S$, $y \in S \Rightarrow \alpha < 1$.
4. $x \notin S$, $y \notin S$. In this case nothing can be said about $\alpha$ in general.

Since $S^c$ is the set of points that will be undersampled using the rejection sampling procedure, it is reasonable that moves from $S$ to $S^c$ are always accepted, while moves from $S^c$ to $S$ always have a nonzero probability of being rejected. The chain dwells in $S^c$ to build up mass that is missing from $\tilde{f}$.

**Examples:**

- **Normal mean and variance:** Suppose we observe $Y_1, \ldots, Y_n$ iid from the $N(\mu, \sigma^2)$ distribution. If we take flat (improper) priors on $\mu$ and $\sigma^2$, the posterior distribution is
\[ P(\mu, \sigma^2|\{Y\}) \propto (\sigma^2)^{-n/2} \exp \left( -\frac{1}{2\sigma^2} \sum_i (Y_i - \mu)^2 \right). \]

As a function of \( \mu \) this is proportional to

\[ \exp\left(-\frac{1}{2\sigma^2}(n\mu^2 - 2\mu Y)\right) \propto \exp\left(-\frac{n}{2\sigma^2}(\mu - \bar{Y})^2\right) \]

Thus

\[ P(\mu|\sigma^2, \{Y\}) = N(\bar{Y}, \sigma^2/n). \]

To get the conditional distribution of \( \sigma^2 \) given \( \{Y\} \) and \( \mu \), do the change of variables \( \eta = 1/\sigma^2 \), so

\[ P(\eta|\mu, \{Y\}) \propto \eta^{n/2-2} \exp(-\eta \sum_i (Y_i - \mu)^2/2). \]

This is a \( \Gamma(n/2 - 1, 2/\sum_i (Y_i - \mu)^2) \) distribution.

So the Gibbs sampling algorithm alternates between sampling \( \mu \) from a normal distribution and \( \eta \) from a gamma distribution.

- **Grouped binary data:** Suppose we observe \( m \) groups of Bernoulli trials, where the \( i \)th group consists of \( n_i \) iid trials. Let \( n_i \) be the number of trials in group \( i \), and let \( p_i \) be the success probability. Suppose the \( p_i \) are viewed as random effects from a symmetric Beta distribution, which has density proportional to

\[ p^{\alpha-1}(1-p)^{\alpha-1}. \]

Let \( k_i \) be the number of successes observed in the \( i \)th group, so

\[ P(\{k_i\}|\{p_i\})P(\{p_i\}|\alpha) \propto \prod_i p_i^{k_i+\alpha-1}(1-p_i)^{n_i-k_i+\alpha-1}. \]

To be safe, we will take a nearly-flat proper exponential prior on \( \alpha \) with mean \( \theta \) (set to an arbitrary large value). Gibbs sampling operates using the following distributions

\[ P(\alpha|\{p_i\}, \{k_i\}) \]
\[ P(\{p_i\}|\alpha, \{k_i\}) \]

The specific updates are

\[ \alpha \sim \text{exponential} \left( 1/(\sum_i p_i - \sum_i (1-p_i) + 1/\theta) \right) \]
\( p_i \sim \text{Beta}(k_i + \alpha, n_i - k_i + \alpha) \).

- **Grouped continuous data with right-skewed group means:** Suppose we observe data \( Y_{ij} \) for \( m \) groups with \( n_i \) replicates in group \( i \), so \( i = 1, \ldots, m \) and \( j = 1, \ldots, n_i \). Let \( \mu_i \) be the population mean for group \( i \), and suppose we take the \( \mu_i \) to be random effects from an exponential distribution with mean \( \lambda \). We will also take \( Y_{i1}, \ldots, Y_{in_i} | \mu_i \) to be independent and Gaussian with mean \( \mu_i \) and variance \( \sigma^2 \). If we form the improper posterior based on flat priors for \( \lambda \) and \( \sigma^2 \), then integrate out the random effects \( \mu_i \), we get

\[
P(\lambda, \sigma^2 | \{Y\}) \propto \prod_i (\sigma^2)^{-n_i/2} \lambda^{-1} \int \exp \left( -\frac{1}{2\sigma^2} (Y_i - \mu_i n_i)'(Y_i - \mu_i n_i) \right) \exp(-\mu/\lambda) d\mu
\]

\[
= (\sigma^2)^{-n/2+m/2} \lambda^{-m} \prod_i \exp \left( -\sigma^{-2}(S_i - M_i^2/n_i)/2 - M_i \lambda^{-1}/n_i + \sigma^2 \lambda^{-2}/2n_i \right),
\]

where \( Y_i = (Y_{i1}, \ldots, Y_{in_i})' \), \( M_i = \sum_j Y_{ij} \), and \( S_i = \sum_j Y_{ij}^2 \).

This does not have a finite integral, therefore a sampling algorithm applied to this formulation of the problem will not give meaningful results. For simplicity, we first suppose that \( \eta = 1/\lambda \) and \( \sigma^2 \) have a joint prior density proportional to

\[
\exp \left( -\frac{1}{2} \sum_i \frac{1}{n_i} \sigma^2 \eta^2 \right)
\]

on \((0, \infty) \times (0, \infty)\). In this case, the posterior on \( \eta, \sigma^2 \) is proportional to

\[
(\sigma^2)^{-n/2+m/2} \eta^{m-2} \exp \left( -\sigma^{-2} \sum_i (S_i - M_i^2/n_i)/2 - \eta \sum_i M_i/n_i \right).
\]

Now if we set \( \theta = 1/\sigma^2 \), we get the posterior expressed in terms of \( \eta \) and \( \theta \)

\[
\theta^{(n-m)/2-2} \eta^{m-2} \exp \left( -\theta \sum_i (S_i - M_i^2/n_i)/2 - \eta \sum_i M_i/n_i \right).
\]

Letting \( V_i = \sum_i S_i/n_i - M_i^2/n_i^2 \), approximately \( \text{var}(Y_i)/n_i \), and \( \bar{M}_i = M_i/n_i \), the sample mean of \( Y_i \), we get

\[
\theta^{(n-m)/2-2} \eta^{m-2} \exp \left( -\theta \sum_i n_i V_i/2 - \eta \sum_i \bar{M}_i \right).
\]
Thus \( \eta \) and \( \theta \) have independent Gamma posteriors

\[
\eta \sim \Gamma(m - 1, 1/\sum_i M_i)
\]

\[
\theta \sim \Gamma((n. - m)/2 - 1, 2/\sum_i n_i V_i).
\]

If we wanted to use a different prior, say with \( \lambda \) and \( \sigma^2 \) independent and exponential with means \( \beta \lambda \) and \( \beta \sigma^2 \), we could use the Gibbs sampler.

- **Cauchy regression**: Suppose we take the linear regression model
  \( Y_i = \alpha + \beta'X_i + \epsilon_i \)
  where \( Y_i \in \mathbb{R} \) and \( X_i \in \mathbb{R}^d \) are observed, and the \( \epsilon_i \) have central Cauchy distributions, with density \( \pi^{-1}/(1 + \epsilon_i^2) \). The joint distribution of the observed data is:

\[
\pi(Y_1, \ldots, Y_n|X_1, \ldots, X_n) = \prod_i \frac{\pi^{-1}}{1 + (Y_i - \alpha - \beta'X_i)^2}.
\]

Suppose we consider the Bayesian model with improper prior on \( \theta = (\alpha, \beta) \). We use a Metropolis-Hastings algorithm with a random walk distribution for \( q \): \( q(\theta', \theta) = \phi(\theta - \theta') \), where \( \phi \) is the bivariate normal distribution with mean 0. Due to the symmetry of the random walk, the Metropolis ratio \( \alpha \) simplifies to \( \alpha(\theta', \theta) = \pi(\theta')/\pi(\theta) \).

- **Rejection sampling using the wrong bound**.
  Suppose we want to sample from the density

\[
\pi(x) = \mathcal{I}(-1 \leq x \leq 1)|x|^{-1/2}/4.
\]

The density is unbounded at \( x = 0 \), so we can not use rejection sampling with a bounded trial density. Suppose we generate draws using the rejection sampling algorithm with a uniform trial density on \((-1, 1)\) and \( c = 6 \). These draws will have density

\[
q(x) \propto \mathcal{I}(-1 \leq x \leq 1) \cdot \min(|x|^{-1/2}/4, 3).
\]

Now we use these draws to drive a Metropolis algorithm. On \( S = \{ x \mid |x| \geq 1/144 \} \), the bound \( \pi/f < 6 \) is satisfied (where \( f \) is the uniform trial density). The Metropolis-Hastings ratio is given by the following.

1. \( x, y \in S \Rightarrow \alpha = 1 \).
2. \( x \in S, y \notin S \Rightarrow \alpha = 1 \).
3. \( x \notin S, y \in S \Rightarrow \alpha = 12\sqrt{|x|} < 1 \).
4. \( x \notin S, y \notin S \Rightarrow \alpha = \pi(y)/\pi(x) = \sqrt{|x/y|} \).
Figure 1: Sample paths for $\alpha$ and $\beta$ for the Cauchy regression model, using random walk input chains with Gaussian steps having mean zero and standard deviation $\sigma = .01$ (top), $\sigma = 3$ (middle), and $\sigma = 1$ (bottom).
• **Gaussian mixture model.** Suppose that $Y_1, \ldots, Y_n \in \mathbb{R}^d$ are independent and observed, $\delta_1, \ldots, \delta_n \in \{1, 2, \ldots, K\}$ are iid and unobserved, and $P(Y_j|\delta_j = k) = \phi(Y_j - \mu_k)$, where $\phi(\cdot)$ is the standard multivariate normal density. The parameters in the model are $\mu_1, \ldots, \mu_K$, and $\pi_1, \ldots, \pi_K$, where $\pi_k = P(\delta_j = k)$. We can take flat priors on these parameters, and estimate them through their posterior distribution, which can be simulated using Gibbs sampling. Here is a full set of conditional distributions, where $\mu_* = [\mu_1, \ldots, \mu_n]$, and similarly for $Y_*$ and $\delta_*$.  

$$P(\mu_*|\pi_*, Y_*, \delta_*) = P(\mu_*|Y_*, \delta_*) \propto P(Y_*|\mu_*, \delta_*)$$  

$$= \prod_j \phi(Y_j - \mu_{\delta_j})$$

This splits into a product of functions of $\mu_1, \ldots, \mu_K$, implying that the $\mu_k$ are independent given $\pi_*, Y_*$ and $\delta_*$. The factor that involves $\mu_k$ can be expressed

$$\exp\left(-\frac{1}{2}(-2\mu_k \sum_j Y_j I(\delta_j = k) + \mu_k^2 \sum_j I(\delta_j = k))\right),$$

which is a multivariate normal density with mean $\sum_j Y_j I(\delta_j = k)/\sum_j I(\delta_j = k)$, and variance $1/\sum I(\delta_j = k)$ (with zero correlation).

$$P(\delta_*|\pi_*, Y_*, \mu_*) \propto P(Y_*|\delta_*, \mu_*) P(\delta_*|\pi_*)$$  

$$= \prod_j \phi(Y_j - \mu_{\delta_j})\pi_{\delta_j}$$

This splits into a product of functions of $\delta_1, \ldots, \delta_n$, implying that the $\delta_j$ are independent given $\pi_*, Y_*$ and $\mu_*$. To sample $\delta_j$, first compute $q_k = \phi(Y_j - \mu_k)\pi_k$ for $k = 1, \ldots, K$, then rescale the $q_k$ so they sum to 1, and generate a multinomial draw according to the probabilities that result.

$$P(\pi_*|Y_*, \mu_*, \delta_*) \propto P(\delta_*|\pi_*)$$  

$$= \prod_k \pi_k^{n_k}$$

Let $n_k = \sum_j I(\delta_j = k)$, so the posterior density of $\pi_*$ is proportional to $\prod_k \pi_k^{n_k}$. This is a Dirichlet density, from which samples can be generated by simulating $Z_k \sim \Gamma(n_k + 1, 1)$, and setting $\pi_k = Z_k/\sum_\ell Z_\ell$.  

10
Figure 2: Gibbs sample paths for the probability parameters $\pi_1, \pi_2, \pi_3$ in the Gaussian mixture model with $k = 3$ components, $d = 10$ $Y$ dimensions, and $n = 200$ observations. The horizontal lines are the true values.

- **Spatial processes.** Suppose we have a random 0/1 array $Y_{ij} \in \{0, 1\}$, and we specify the conditional probabilities as

$$P(Y_{ij} = 1| Y_{i+1,j}, Y_{i-1,j}, Y_{i,j+1}, Y_{i,j-1}) = \epsilon + \frac{\tau - \epsilon}{4} (Y_{i+1,j} + Y_{i-1,j} + Y_{i,j+1} + Y_{i,j-1}).$$

where $0 \leq \epsilon \leq \tau \leq 1$. It is a fact that this family of conditional distributions is associated with a unique joint distribution on the array. The conditional probability of observing a 1 given the neighbors ranges from $\epsilon$ to $\tau$, and increases as more neighbors have 1’s. Larger values of $\epsilon - \tau$ produce greater spatial autocorrelation (the tendency of 1’s to cluster spatially in the array). We can simulate from this distribution using the Gibbs sampler.