# CS281A/Stat241A Lecture 22

## *Monte Carlo Methods*

Peter Bartlett

# Key ideas of this lecture

- Sampling in Bayesian methods:
  - Predictive distribution
  - Example: Normal with conjugate prior
  - Posterior predictive model checking
- Example: Normal with semiconjugate prior
- Gibbs sampling
- Nonconjugate priors
- Metropolis algorithm
- Metropolis-Hastings algorithm
- Markov Chain Monte Carlo

# Sampling in Bayesian Methods

- Consider a model

$$p(x, \theta) = \underbrace{p(\theta)}_{\text{prior}} p(x|\theta) p(y|x, \theta).$$

- Given observations $(x_1, y_1), \ldots, (x_n, y_n)$ and $x$, we wish to estimate the distribution of $y$, perhaps to make a forecast $\hat{y}$ that minimizes the expected loss $\mathbb{E}L(\hat{y}, y)$.
- We condition on observed variables $((x_1, y_1), \ldots, (x_n, y_n), x)$ and marginalize out unknown variables ($\theta$) in the joint

$$p(x_1, y_1, \ldots, x_n, y_n, x, y, \theta) = p(\theta) p(x, y|\theta) \prod_{i=1}^{n} p(x_i, y_i|\theta).$$

# Sampling in Bayesian Methods

- This gives the *posterior predictive distribution*

$$p(y|x_1, y_1, \ldots, x_n, y_n, x) = \int p(y|\theta, D_n, x)p(\theta|D_n, x)d\theta$$
$$= \int p(y|\theta, x)p(\theta|D_n, x)d\theta,$$

  where $D_n = (x_1, y_1, \ldots, x_n, y_n)$ is the observed data.

- We can also consider the *prior predictive distribution*, which is the same quantity when nothing is observed:

$$p(y) = \int p(y|\theta)p(\theta)d\theta.$$

It is often useful, to evaluate if the prior reflects reasonable beliefs for the observations.

# Sampling from posterior

- We wish to sample from the posterior predictive distribution,

$$p(y|D_n) = \int p(y|\theta)p(\theta|D_n)d\theta.$$

- It might be straightforward to sample from

$$p(\theta|D_n) \propto p(D_n|\theta)p(\theta).$$

  For example, if we have a *conjugate prior* $p(\theta)$, it can be an easy calculation to obtain the posterior, and then we can sample from it.

- It is typically straightforward to sample from $p(y|\theta)$.

# Sampling from posterior

- In these cases, we can
  1. Sample $\theta^t \sim p(\theta|D_n)$,
  2. Sample $y^t \sim p(y|\theta^t)$.
- Then we have

$$(\theta^t, y^t) \sim p(y, \theta|D_n),$$

and hence we have samples from the posterior predictive distribution,

$$y^t \sim p(y|D_n).$$

# Example: Gaussian with Conjugate Prior

- Consider the model

$$y \sim \mathcal{N}(\mu, \sigma^2)$$
$$\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2/\kappa_0)$$
$$1/\sigma^2 \sim \text{gamma}(\nu_0/2, \nu_0\sigma_0^2/2),$$

where $x \sim \text{gamma}(a, b)$ for

$$p(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}.$$

# Example: Gaussian with Conjugate Prior

- This is a conjugate prior: posterior is

$$1/\sigma^2|y_1,\ldots,y_n \sim \text{gamma}(\nu_n/2, \nu_n\sigma^2/2)$$

$$\mu|y_1,\ldots,y_n,\sigma^2 \sim \mathcal{N}(\mu_n, \sigma^2/\kappa_n)$$

$$\begin{aligned}
\nu_n &= \nu_0 + n \\
\sigma_n^2 &= \left(\nu_0\sigma_0^2 + \sum_i (y_i - \bar{y})^2 + (\bar{y} - \mu_0)^2 \kappa_0 n/\kappa_n\right)/\nu_n \\
\kappa_n &= \kappa_0 + n \\
\mu_n &= (\kappa_0\mu_0 + n\bar{y})/\kappa_n
\end{aligned}$$

# Example: Gaussian with Conjugate Prior

- So we can easily compute the posterior distribution $p(\theta|D_n)$ (with $\theta = (\mu, \sigma^2)$), and then:
  1. Sample $\theta^t \sim p(\theta|D_n)$,
  2. Sample $y^t \sim p(y|\theta^t)$.

# Posterior Predictive Model Checking

- Another application of sampling:
- Suppose we have a model $p(\theta)$, $p(y|\theta)$.
- We see data $D_n = (y_1, \ldots, y_n)$.
- We obtain the predictive distribution $p(y|y_1, \ldots, y_n)$.
- Suppose that some important feature of the predictive distribution does not appear to be consistent with the empirical distribution. Is this due to sampling variability, or because of a model mismatch?

# Posterior Predictive Model Checking

1. Sample $\theta^t \sim p(\theta|D_n)$.
2. Sample $Y^t = (y_1^t, \ldots, y_n^t)$, with $y_i^t \sim p(y|\theta^t)$.
3. Calculate $f(Y^t)$, the statistic of interest.

We can use the Monte Carlo approximation to the distribution of $f(Y^t)$ to assess the model fit.

# Key ideas of this lecture

- Sampling in Bayesian methods:
  - Predictive distribution
  - Example: Normal with conjugate prior
  - Posterior predictive model checking
- Example: Normal with semiconjugate prior
- Gibbs sampling
- Nonconjugate priors
- Metropolis algorithm
- Metropolis-Hastings algorithm
- Markov Chain Monte Carlo

# Semiconjugate priors

- Consider again a normal distribution with conjugate prior:

$$
\begin{aligned}
y &\sim \mathcal{N}(\mu, \sigma^2) \\
\mu|\sigma^2 &\sim \mathcal{N}(\mu_0, \sigma^2/\kappa_0) \\
1/\sigma^2 &\sim \text{gamma}(\nu_0/2, \nu_0\sigma_0^2/2).
\end{aligned}
$$

  In order to ensure conjugacy, the $\mu$ and $\sigma^2$ distributions are not marginally independent.

# Semiconjugate priors

- If it's more appropriate to have decoupled prior distributions, we might consider a *semiconjugate* prior, that is, a prior that is a product of priors,

$$p(\theta) = p(\mu)p(\sigma^2),$$

  each of which, conditioned on the other parameters, is conjugate.

- For example, for the normal distribution, we could choose

$$\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$$
$$1/\sigma^2 \sim \text{gamma}(\nu_0/2, \nu_0\sigma_0^2/2).$$

# Semiconjugate priors

- The posterior of $1/\sigma^2$ is not a gamma distribution. However, the conditional distribution $p(1/\sigma^2|\mu, y_1, \ldots, y_n)$ is a gamma:

$$p(\sigma^2|\mu, y_1, \ldots, y_n) \propto p(y_1, \ldots, y_n|\mu, \sigma^2)p(\sigma^2),$$
$$p(\mu|\sigma^2, y_1, \ldots, y_n) \propto p(y_1, \ldots, y_n|\mu, \sigma^2)p(\mu).$$

- These distributions are called the *full conditional distributions* of $\sigma^2$ and of $\mu$: conditioned on everything else. In both cases, the distribution is the posterior of a conjugate, so it's easy to calculate.

# Semiconjugate priors: Gibbs sampling

$$p(\sigma^2|\mu, y_1, \ldots, y_n) \propto p(y_1, \ldots, y_n|\mu, \sigma^2)p(\sigma^2),$$
$$p(\mu|\sigma^2, y_1, \ldots, y_n) \propto p(y_1, \ldots, y_n|\mu, \sigma^2)p(\mu).$$

- Notice that, because the full conditional distributions are posteriors of conjugates, they are easy to calculate, and hence easy to sample from.
- Suppose we had $\sigma^{2(1)} \sim p(\sigma^2|y_1, \ldots, y_n)$.
- Then if we choose $\mu^{(1)} \sim p(\mu|\sigma^{2(1)}, y_1, \ldots, y_n)$, we would have

$$(\mu^{(1)}, \sigma^{2(1)}) \sim p(\mu, \sigma^2|y_1, \ldots, y_n).$$

# Semiconjugate priors

- In particular, $\mu^{(1)} \sim p(\mu^{(1)}|y_1, \ldots, y_n)$, so we can choose $\sigma^{2(2)} \sim p(\sigma^2|\mu^{(1)}, y_1, \ldots, y_n)$, so

$$(\mu^{(1)}, \sigma^{2(2)}) \sim p(\mu, \sigma^2|y_1, \ldots, y_n).$$

- etc...

# Gibbs sampling

For parameters $\theta = (\theta_1, \ldots, \theta_p)$:

1. Set some initial values $\theta^0$.
2. For $t = 1, \ldots, T$,
   - For $i = 1, \ldots, p$,
     - Sample $\theta_i^t \sim p(\theta_i | \theta_1^t, \ldots, \theta_{i-1}^t, \theta_{i+1}^{t-1}, \ldots, \theta_p^{t-1})$.

- Notice that the parameters $\theta^1, \theta^2, \ldots, \theta^T$ are dependent.
- They form a Markov chain: $\theta^t$ depends on $\theta^1, \ldots, \theta^{t-1}$ only through $\theta^{t-1}$.
- The posterior distribution is a stationary distribution of the Markov chain (by argument on previous slide).
- And the distribution of $\theta^T$ approaches the posterior.

# Key ideas of this lecture

- Sampling in Bayesian methods:
  - Predictive distribution
  - Example: Normal with conjugate prior
  - Posterior predictive model checking
- Example: Normal with semiconjugate prior
- Gibbs sampling
- Nonconjugate priors
- Metropolis algorithm
- Metropolis-Hastings algorithm
- Markov Chain Monte Carlo

# Nonconjugate priors

- Sometimes conjugate or semiconjugate priors are not available or are unsuitable.
- In such cases, it can be difficult to sample directly from $p(\theta|y)$ (or $p(\theta_i|\text{all else})$).
- Note that we do not need i.i.d. samples from $p(\theta|y)$, rather we need $\theta^1, \ldots, \theta^m$ so that the empirical distribution approximates $p(\theta|y)$.
- Equivalently, for any $\theta, \theta'$, we want

$$\frac{\mathbb{E}[\#\ \theta \text{ in sample}]}{\mathbb{E}[\#\ \theta' \text{ in sample}]} \approx \frac{p(\theta|y)}{p(\theta'|y)}.$$

# Nonconjugate priors

- Some intuition: suppose we already have $\theta^1, \ldots, \theta^t$, and another $\theta$. Should we add it (set $\theta^{t+1} = \theta$)?
- We could base the decision on how $p(\theta|y)$ compares to $p(\theta^t|y)$.
- Happily, we do not need to be able to compute $p(\theta|y)$:

$$\frac{p(\theta|y)}{p(\theta^t|y)} = \frac{p(y|\theta)p(\theta)p(y)}{p(y)p(y|\theta^t)p(\theta^t)} = \frac{p(y|\theta)p(\theta)}{p(y|\theta^t)p(\theta^t)}.$$

- If $p(\theta|y) > p(\theta^t|y)$, we set $\theta^{t+1} = \theta$.
- If $r = p(\theta|y)/p(\theta^t|y) < 1$, we expect to have $\theta$ appear $r$ times as often as $\theta^t$, so we accept $\theta$ (set $\theta^t = \theta$) with probability $r$.

# Metropolis Algorithm

Fix a *symmetric proposal distribution*, $q(\theta|\theta') = q(\theta'|\theta)$.
Given a sample $\theta^1, \ldots, \theta^t$,

1. Sample $\theta \sim q(\theta|\theta^t)$.
2. Calculate
$$r = \frac{p(\theta|y)}{p(\theta^t|y)} = \frac{p(y|\theta)p(\theta)}{p(y|\theta^t)p(\theta^t)}.$$
3. Set
$$\theta^{t+1} = \begin{cases} \theta & \text{with probability } \min(r, 1), \\ \theta^t & \text{with probability } 1 - \min(r, 1). \end{cases}$$

# Metropolis Algorithm

- Notice that $\theta^1, \ldots, \theta^t$ is a Markov chain.
- We'll see that its stationary distribution is $p(\theta|y)$.
- But we have to wait until the Markov chain mixes.

# Metropolis Algorithm

In practice:

1. Wait until some time $\tau$ (burn-in time) when it seems that the chain has reached the stationary distribution.
2. Gather more samples, $\theta^{\tau+1}, \ldots, \theta^{\tau+m}$.
3. Approximate the posterior $p(\theta|y)$ using the empirical distribution of $\{\theta^{\tau+1}, \ldots, \theta^{\tau+m}\}$.

# Metropolis Algorithm

- The higher the correlation between $\theta^t$ and $\theta^{t+1}$, the longer the burn-in period needs to be, and the worse the approximation of the posterior that we get from the empirical distribution of an $m$-sample: The effective sample size is much less than $m$.
- We can control the correlation using the proposal distribution.
- Typically, the best performance occurs for an intermediate value of a scale parameter of the proposal distribution: small variance gives high correlation; high variance gives samples far from the posterior mode, hence with low acceptance probability.

# Metropolis Algorithm

- Need to choose a proposal distribution that allows the Markov chain to move around the parameter space quickly, but not with such big steps that the proposals are frequently rejected.
- Rule of thumb: Tune proposal distribution so that the acceptance probability is between 20% and 50%.

# Key ideas of this lecture

- Sampling in Bayesian methods:
  - Predictive distribution
  - Example: Normal with conjugate prior
  - Posterior predictive model checking
- Example: Normal with semiconjugate prior
- Gibbs sampling
- Nonconjugate priors
- Metropolis algorithm
- Metropolis-Hastings algorithm
- Markov Chain Monte Carlo

# Metropolis-Hastings Algorithm

- Gibbs sampling and the Metropolis algorithm are special cases of *Metropolis-Hastings*.
- Suppose we wish to sample from $p(u, v)$.
- We come up with two proposal distributions, $q_u(u|u', v')$ and $q_v(v|u', v')$. Notice that they can depend on the other variable, and do not need to be symmetric as in the Metropolis algorithm.

# Metropolis-Hastings Algorithm

1. Update $u$ sample:
   (a) Sample $u \sim q_u(u|u^t, v^t)$.
   (b) Compute

$$r = \frac{p(u, v^t) q_u(u^t|u, v^t)}{p(u^t, v^t) q_u(u|u^t, v^t)}.$$

   (c) Set

$$u^{t+1} = \begin{cases} u & \text{with prob } \min(1, r), \\ u^t & \text{with prob } 1 - \min(1, r). \end{cases}$$

# Metropolis-Hastings Algorithm

2. Update $v$ sample:
   (a) Sample $v \sim q_v(v|u^{t+1}, v^t)$.
   (b) Compute

$$r = \frac{p(u^{t+1}, v) q_v(v^t | u^{t+1}, v)}{p(u^{t+1}, v^t) q_v(v | v^{t+1}, v^t)}.$$

   (c) Set

$$v^{t+1} = \begin{cases} v & \text{with prob } \min(1, r), \\ v^t & \text{with prob } 1 - \min(1, r). \end{cases}$$

# Metropolis-Hastings Algorithm

- Just like Metropolis algorithm, but the acceptance ratio

$$r = \frac{p(u, v^t) q_u(u^t | u, v^t)}{p(u^t, v^t) q_u(u | u^t, v^t)}$$

  contains an extra factor

$$\frac{q_u(u^t | u, v^t)}{q_u(u | u^t, v^t)}.$$

- If $u$ is much more likely to be reached from $u^t$ than vice versa, downweight the probability of acceptance (to avoid over-representing $u$).

# Metropolis versus M-H

- If $q_u$ is symmetric,

$$q_u(u|u', v) = q_u(u'|u, v),$$

then the correction factor

$$\frac{q_u(u^t|u, v^t)}{q_u(u|u^t, v^t)}$$

is 1, so acceptance probability is the same as the Metropolis algorithm.

- That is, Metropolis is a special case of Metropolis-Hastings.

# Metropolis versus Gibbs

- If $q_u$ is the full conditional distribution of $u$ given $v$,

$$q_u(u|u', v) = p(u|v),$$

so the acceptance ratio is

$$\begin{aligned} r &= \frac{p(u, v^t) q_u(u^t|u, v^t)}{p(u^t, v^t) q_u(u|u^t, v^t)} \\ &= \frac{p(u, v^t) p(u^t|v^t)}{p(u^t, v^t) p(u|v^t)} \\ &= \frac{p(u|v^t) p(v^t) p(u^t|v^t)}{p(u^t|v^t) p(v^t) p(u|v^t)} \\ &= 1. \end{aligned}$$

# Metropolis versus Gibbs

- That is, if we propose a value from the full conditional distribution, we accept it with probability 1.
- So Gibbs sampling is a special case of Metropolis-Hastings.

# Gibbs plus Metropolis plus MH

- Notice that we can choose different proposal distributions for different variables $u, v, \ldots$.
- Since Metropolis and Gibbs correspond to specific choices of the proposal distribution, we can easily combine these algorithms:
  - For some variables, full conditional distributions might be available (typically, because of a conjugate prior for that variable), and we can do Gibbs sampling.
  - For some variables, they are not available, but we can choose an alternative proposal distribution.

# Key ideas of this lecture

- Sampling in Bayesian methods:
  - Predictive distribution
  - Example: Normal with conjugate prior
  - Posterior predictive model checking
- Example: Normal with semiconjugate prior
- Gibbs sampling
- Nonconjugate priors
- Metropolis algorithm
- Metropolis-Hastings algorithm
- Markov Chain Monte Carlo

# Markov Chain Monte Carlo

- The transition probability matrix of a Markov chain determines the state evolution:

$$A_{ij} = \Pr(x_{t+1} = j | x_t = i).$$

- Recall that a distribution over states $p_t(x)' = (\Pr(x_t = 1), \ldots, \Pr(x_t = N))$ evolves as

$$p'_{t+1} = p'_t A.$$

- A *stationary distribution* $p$ on $\mathcal{X}$ satisfies $p'A = p'$.

# Markov Chain Monte Carlo

- An *ergodic* Markov chain is irreducible (no islands) and aperiodic. It always has a *unique* stationary distribution: for all $p_0$,

$$p_0' A^t \to p.$$

- An ergodic MC *mixes* exponentially: for some $C, \tau$ and stationary distribution $p$,

$$\|p_0' A^t - p\|_1 \leq C e^{-t/\tau}.$$

# Markov Chain Monte Carlo

- If $p$ satisfies the detailed balance equations

$$p_i A_{ij} = p_j A_{ji},$$

then $p$ is a stationary distribution, and the chain is called *reversible*:

$$\Pr(x_t = i, x_{t+1} = j) = \Pr(x_t = j, x_{t+1} = i).$$

# Metropolis-Hastings

- We'll show that the distribution $p$ and the transition probabilities $A$ of Metropolis-Hastings satisfy the detailed balance equations

$$p_i A_{ij} = p_j A_{ji},$$

  and hence $p$ is the stationary distribution.
- Thus, if the Markov chain is ergodic, it converges to $p$, and in particular

$$\sum_{t=\tau}^{T} f(x_t) \to \mathbb{E}_p f.$$

- For simplicity, we'll consider a single variable update.

# Metropolis-Hastings

MH proposal to move from $i$ to $j$ is accepted with probability

$$a(i,j) = \min\left(1, \frac{p(j)q(i|j)}{p(i)q(j|i)}\right).$$

Thus,

$$\begin{aligned} p_i A_{ij} &= p_i q(j|i) \min\left(1, \frac{p(j)q(i|j)}{p(i)q(j|i)}\right) \\ &= p_j q(i|j) \min\left(\frac{p_i q(j|i)}{p_j q(i|j)}, 1\right) \\ &= p_j A_{ji}. \end{aligned}$$

So $p$ is a stationary distribution.

# Metropolis-Hastings

- Will the distribution converge to the stationary distribution?
- When is the Markov chain ergodic?
- For instance, if $p$ is continuous on $\mathbb{R}^d$ and the proposal distribution $q(x'|x)$ is a Gaussian centered at the current value, that is enough.

# Metropolis-Hastings

- There are cases where we do not have an ergodic Markov chain. For example, consider Gibbs sampling in a directed graphical model

$$X_1 \rightarrow S \leftarrow X_2,$$

  where $X_1, X_2$ are outcomes of coin tosses and $S$ is the indicator for $X_1 = X_2$.

- In such cases, we can update blocks of variables at once: do exact inference for the block in a Gibbs sampling step.

# Metropolis-Hastings

- The mixing time of the Markov chain (the time constant in the exponential convergence to the stationary distribution) is of crucial importance in practice, but often we do not have useful upper bounds on the mixing time.

# Key ideas of this lecture

- Sampling in Bayesian methods:
  - Predictive distribution
  - Example: Normal with conjugate prior
  - Posterior predictive model checking
- Example: Normal with semiconjugate prior
- Gibbs sampling
- Nonconjugate priors
- Metropolis algorithm
- Metropolis-Hastings algorithm
- Markov Chain Monte Carlo