Markov Chain Monte Carlo

Justin Domke `people.cs.umass.edu/domke/`

August 2, 2017

1 Setup

Given a *distribution* $p(x) : \mathbb{R}^n \to \mathbb{R}$ and a *function* $f(x) : \mathbb{R}^n \to \mathbb{R}$, the goal is to estimate the mean

$$\mu = \mathop{\mathbb{E}}_{p(x)}[f(x)] = \int_x p(x) f(x) dx.$$

You can query $p(x)$ and $f(x)$ at individual points, but otherwise, these are "black boxes". You are not able to access the inner structure of p or f.

2 Detailed Balance

Definition: A Markov chain with transition probabilities $M(x \to x')$ is said to satisfy *detailed balance* with respect to a distribution $p(x)$ if for all x and x',

$$p(x)M(x \to x') = p(x')M(x' \to x).$$

Definition: Informally, a Markov chain M has a *stationary distribution* p if, if you pick a point x from the $p(x)$ and run a single step of the chain, the resulting sample is still distributed like p. Formally, p is a stationary distribution of M if for all x', $\sum_x p(x)M(x \to x') = p(x')$.
Claim: If a Markov chain $M(x \to x')$ satisfies detailed balance with respect to some distribution $p(x)$ then $p(x)$ is a stationary distribution of that chain.
Proof: Suppose that M satisfies detailed balance with respect to p. Then,

$$\sum_x p(x)M(x \to x') = \sum_x p(x')M(x' \to x) = p(x') \sum_x M(x' \to x) = p(x').$$

3 Barker's Algorithm

- Pick x^1 randomly
- For $t = 1, 2, ..., T-1$
  - Generate $\eta \sim \mathcal{N}(0, I)$ and $r \sim U(0, 1)$.
  - Propose $x' \leftarrow x + \sqrt{\epsilon}\eta$
  - If $r < p(x') / \left(p(x) + p(x')\right)$, then $x^{t+1} \leftarrow x'$ ("accept"). Else, $x^{t+1} \leftarrow x^t$ ("reject").
- Return $\hat{\mu} = \sum_{t=1}^T f(x^t)$.

Claim: Barker's algorithm satisfies detailed balance with respect to p.
Proof: First, what is the probability that this algorithm transitions from x to x' in any given iteration? To do this, you must first "propose" x' by drawing the appropriate η, and then "accept" it by drawing the appropriate r. This leads to

$$M(x \to x') = \underbrace{\mathcal{N}(x' - x | 0, \epsilon I)}_{\text{probability of right } \eta} \underbrace{\frac{p(x')}{p(x) + p(x')}}_{\text{probability of right } r}.$$

Once you've observed this, it's trivial to check that

$$p(x)M(x \to x') = \mathcal{N}(x' - x|0, \epsilon I)\frac{p(x)p(x')}{p(x) + p(x')} = p(x')M(x' \to x).$$

This uses the fact that by symmetry $\mathcal{N}(x' - x|0, \epsilon I) = \mathcal{N}(x - x'|0, \epsilon I)$.
Comments:

- You can actually use any noise distribution rather than a Gaussian distribution. But it must be symmetric!
- The Metropolis method is a variant where you change $r < p(x')/\left(p(x) + p(x')\right)$ to $r < p(x')/p(x)$.
  - This "accepts" moves up to 50% more often. (Consider the case where $p(x) = p(x')$).
  - You should **always** use Metropolis instead of Barker.
- The Metropolis-Hastings methods is another variant where you can use arbitrary noise instead of Gaussian noise.

4 Langevin Dynamics

Langevin dynamics use a different setting. Rather than a black box that outputs $p(x)$, suppose we have one that also returns $\log p(x)$ and the gradient $\nabla \log p(x)$.

- Pick x^1 randomly
- For $t = 1, 2, ..., T - 1$
  - Generate $\eta \sim \mathcal{N}(0, I)$ and $r \sim U(0, 1)$.
  - $x' \leftarrow x + \frac{\epsilon}{2}\nabla \log p(x) + \sqrt{\epsilon}\eta$.
  - If $r < \boxed{\text{complicated}(x, x')}$, then $x^{t+1} \leftarrow x'$ ("accept"). Else, $x^{t+1} \leftarrow x^t$ ("reject").
- Return $\hat{\mu} = \sum_{t=1}^{T} f(x^t)$.

$$\boxed{\text{complicated}(x, x')} = \frac{p(x')}{p(x)} \exp\left(\frac{\epsilon}{8}\left\|\nabla \log p(x)\right\|^2 - \frac{\epsilon}{8}\left\|\nabla \log p(x')\right\|^2 + \frac{1}{2}(x - x') \cdot \left(\nabla \log p(x) + \nabla \log p(x')\right)\right)$$

Langevin dynamics are a special case of an algorithm known as Hamiltonian Monte Carlo that takes multiple gradients steps in each iteration (to try to increase mixing).

4.1 Why this is interesting

Firstly, this is interesting because it uses the gradient information. Experimentally, this often seems to make mixing faster.

A second reason is that this can form a "stochastic" MCMC algorithm. Often to evaluate $\log p(x)$ for a single x is very expensive. Commonly, x is some latent variable and $p(x)$ reflects how well it fits to your data. Then, to evaluate $p(x)$ once requires a full pass over your dataset— very expensive if you have 2TB of data!

Observations:

- In the limit of a small step-size ϵ, $\boxed{\text{complicated}(x, x')} \to 1$. That is, the algorithm always accepts.
- Often, one can easily get an *unbiased estimate* of $\nabla \log p(x)$ in just constant time by looking at a *single datum*.
- In the limit of a small step-size ϵ, if you have an error in the estimate of $\frac{\epsilon}{2}\nabla \log p(x)$, this will be dwarfed by the noise $\sqrt{\epsilon}\eta$.

Thus, for small ϵ, we can hopefully:

- Forget about computing $p(x)$ (we'd always accept anyway)
- Get away with only using a cheap estimate of $\nabla \log p(x)$.

4.2 Stochastic Langevin

- Pick x^1 randomly
- For $t = 1, 2, ..., T-1$
  - $g \leftarrow$ estimate of $\nabla \log p(x)$
  - Generate $\eta \sim \mathcal{N}(0, I)$
  - $x^{t+1} \leftarrow x^t + \sqrt{\epsilon}\eta$.
- Return $\hat{\mu} = \sum_{t=1}^{T} f(x^t)$

4.3 Proof that the acceptance rate becomes one in the limit of small ϵ

We claim that for small ϵ the acceptance rate becomes one. This section sacrifices some rigor for the sake of clarity (though it's easy to turn each "$\approx$" statement into a limit statement.)
Lemma: For small ϵ, $\log p(x) - \log p(x') \approx \frac{1}{2}(x - x') \cdot (\nabla \log p(x) + \nabla \log p(x'))$.
Proof: Take the two linear approximations

$$\log p(x) \approx \log p(x') + (x - x') \cdot \nabla \log p(x')$$
$$\log p(x') \approx \log p(x) + (x' - x) \cdot \nabla \log p(x).$$

If we subtract the second from the first, we get

$$\log p(x) - \log p(x') \approx \log p(x') - \log p(x) + (x - x') \cdot (\nabla \log p(x') + \nabla \log p(x)),$$

which is equivalent to the result.
Claim: $\lim_{\epsilon \to 0} \boxed{\text{complicated}(x, x')} = 1$.
Proof: We can use the above Lemma to write that

$$\begin{aligned}\boxed{\text{complicated}(x, x')} &\approx \frac{p(x')}{p(x)} \exp\left(\frac{\epsilon}{8}\|\nabla \log p(x)\|^2 - \frac{\epsilon}{8}\|\nabla \log p(x')\|^2 + \log p(x) - \log p(x')\right) \\ &= \exp\left(\frac{\epsilon}{8}\|\nabla \log p(x)\|^2 - \frac{\epsilon}{8}\|\nabla \log p(x')\|^2\right),\end{aligned}$$

which obviously goes to one in the limit.

4.4 Proof that Langevin Dynamics satisfy detailed balance.

To see that this satisfies detailed balance, first identify the transition probabilities. The probability of getting from x to x' in one step is

$$M(x \to x') = \mathcal{N}\left(x' - x + \frac{\epsilon}{2}\nabla \log p(x), 0, \epsilon I\right) \min\left(1, \boxed{\text{complicated}(x, x')}\right).$$

While the probability of getting from x' to x is

$$M(x' \to x) = \mathcal{N}\left(x - x' + \frac{\epsilon}{2}\nabla \log p(x'), 0, \epsilon I\right) \min\left(1, \boxed{\text{complicated}(x', x)}\right).$$

Now, note that

$$\begin{aligned}\frac{1}{\boxed{\text{complicated}(x, x')}} &= 1 \Big/ \left(\frac{p(x')}{p(x)} \exp\left(\frac{\epsilon}{8}\|\nabla \log p(x)\|^2 - \frac{\epsilon}{8}\|\nabla \log p(x')\|^2 + \frac{1}{2}(x - x') \cdot (\nabla \log p(x) + \nabla \log p(x'))\right)\right) \\ &= \frac{p(x)}{p(x')} \exp\left(\frac{\epsilon}{8}\|\nabla \log p(x')\|^2 - \frac{\epsilon}{8}\|\nabla \log p(x)\|^2 + \frac{1}{2}(x' - x) \cdot (\nabla \log p(x) + \nabla \log p(x'))\right) \\ &= \boxed{\text{complicated}(x', x)}\end{aligned}$$

Therefore, notice that exactly one of $\boxed{\text{complicated}(x, x')}$ and $\boxed{\text{complicated}(x', x)}$ will be less than one. Suppose without loss of generality that $\boxed{\text{complicated}(x, x')} < 1$. Then (where C is the normalizer of a Gaussian), we can do a lot of manipulation to show that

$$
\begin{aligned}
& p(x)M(x \to x') \\
=& p(x)\mathcal{N}\left(x' - x - \frac{\epsilon}{2}\nabla \log p(x), 0, \epsilon I\right) \boxed{\text{complicated}(x, x')} \\
=& C\ p(x) \exp\left(-\frac{\left\|x' - x - \frac{\epsilon}{2}\nabla \log p(x)\right\|^2}{2\epsilon}\right) \boxed{\text{complicated}(x, x')} \\
=& C\ p(x) \exp\left(-\frac{1}{2\epsilon}\left\|x' - x\right\|^2 + \frac{1}{2}(x' - x) \cdot \nabla \log p(x) - \frac{\epsilon}{8}\left\|\nabla \log p(x)\right\|^2\right) \boxed{\text{complicated}(x, x')} \\
=& C\ p(x) \exp\left(-\frac{1}{2\epsilon}\left\|x' - x\right\|^2 + \frac{1}{2}(x' - x) \cdot \nabla \log p(x) - \frac{\epsilon}{8}\left\|\nabla \log p(x)\right\|^2\right) \\
& \times \frac{p(x')}{p(x)} \exp\left(\frac{\epsilon}{8}\left\|\nabla \log p(x)\right\|^2 - \frac{\epsilon}{8}\left\|\nabla \log p(x')\right\|^2 + \frac{1}{2}(x - x') \cdot (\nabla \log p(x) + \nabla \log p(x'))\right) \\
=& C\ p(x') \exp\left(-\frac{1}{2\epsilon}\left\|x' - x\right\|^2 + \frac{1}{2}(x - x') \cdot \nabla \log p(x') - \frac{\epsilon}{8}\left\|\nabla \log p(x')\right\|^2\right) \\
=& C\ p(x') \exp\left(-\frac{\left\|x' - x\right\|^2 - \epsilon(x - x') \cdot \nabla \log p(x') + \left\|\frac{\epsilon}{2}\nabla \log p(x')\right\|^2}{2\epsilon}\right) \\
&= C\ p(x') \exp\left(-\frac{\left\|x - x' - \frac{\epsilon}{2}\nabla \log p(x')\right\|^2}{2\epsilon}\right) \\
&= p(x')\mathcal{N}\left(x - x' - \frac{\epsilon}{2}\nabla \log p(x'), 0, I\right) \\
&= p(x')M(x' \to x).
\end{aligned}
$$

5 Recommended References

- David MacKay, "Information Theory, Inference, and Learning Algorithms", Chapter 29 (Monte Carlo Methods) and Chapter 30 (Efficient Monte Carlo Methods).
- Max Welling and Yee Whye Teh "Bayesian Learning via Stochastic Gradient Langevin Dynamics", ICML 2011
- Radford Neal, "MCMC using Hamiltonian dynamics", Handbook of Markov Chain Monte Carlo, CRC Press, 2011