Probabilistic Artificial Intelligence

Solutions to Problem Set 3

October 28, 2018

1. Markov chains and detailed balance

Assume that you are given a Markov chain with finite state space Ω and transition matrix T, which is defined for all $x, y \in \Omega$ and $t \geq 0$ as $T(x, y) := P(X_{t+1} = y \mid X_t = x)$. Furthermore, let π be the stationary distribution of the chain.

(i) Show that, if for some t the current state X_t is distributed according to the stationary distribution and additionally the chain satisfies the detailed balance equations

$$\pi(x)T(x, y) = \pi(y)T(y, x), \text{ for all } x, y \in \Omega,$$

then the following holds for all $k \geq 0$ and $x_0, \ldots, x_k \in \Omega$:

$$P(X_t = x_0, \ldots, X_{t+k} = x_k) = P(X_t = x_k, \ldots, X_{t+k} = x_0).$$

(This is why a chain that satisfies detailed balance is called *reversible*.)

(ii) Show that, if T is a symmetric matrix, then the chain satisfies detailed balance, and the uniform distribution on Ω is stationary for that chain.

Solution

(i) We use the chain rule, as well as the detailed balance condition:

$$\begin{aligned}
&P(X_t = x_0, \ldots, X_{t+k} = x_k) \\
&= P(X_t = x_0)P(X_{t+1} = x_1 \mid X_t = x_0) \ldots P(X_{t+k} = x_k \mid X_{t+k-1} = x_{k-1}) && \text{ch. rule} \\
&= \pi(x_0)T(x_0, x_1) \ldots T(x_{k-1}, x_k) && X_t \sim \pi \\
&= T(x_1, x_0)\pi(x_1) \ldots T(x_{k-1}, x_k) && \text{detailed balance} \\
&= \ldots && \vdots \\
&= T(x_1, x_0) \ldots T(x_k, x_{k-1})\pi(x_k) && \text{detailed balance} \\
&= \pi(x_k)T(x_k, x_{k-1}) \ldots T(x_1, x_0) \\
&= P(X_t = x_k)P(X_{t+1} = x_{k-1} \mid X_t = x_k) \ldots P(X_{t+k} = x_0 \mid X_{t+k-1} = x_1) && X_t \sim \pi \\
&= P(X_t = x_k, \ldots, X_{t+k} = x_0). && \text{ch. rule}
\end{aligned}$$

(ii) By definition of a symmetric matrix, we have that $\pi(x)T(x, y) = \pi(x)T(y, x)$, for all $x, y \in \Omega$. Therefore, if $\pi(x) = \frac{1}{|\Omega|}$, for all $x \in \Omega$, then $\pi(x)T(x, y) = \pi(y)T(y, x)$, which means that detailed balance holds for the chain and the uniform distribution is stationary.

2. Convergence of the Metropolis-Hastings algorithm

We use Markov Chain Monte Carlo (MCMC) methods to sample from a target distribution $P(x) = Q(x)/Z$ using a proposal distribution $R(x|x')$, without computing the normalization constant Z. A famous MCMC method is the Metropolis-Hastings algorithm, given below:

Algorithm 1 Metropolis-Hastings

Input: Unnormalized target distribution $Q(x)$, proposal distribution $R(x|x')$
Initialize: x_1 arbitrary
For $t = 1, 2, \ldots, T$:

1. Sample proposal x from the proposal distribution $R(x|x_t)$.
2. Compute the *acceptance probability* $\alpha = \min\left(1, \frac{Q(x)R(x_t|x)}{Q(x_t)R(x|x_t)}\right)$.
3. With probability α, set $x_{t+1} = x$, else $x_{t+1} = x_t$.

The algorithm defines a Markov chain with transition kernel $T(x, x') = P(x_{t+1} = x'|x_t = x)$. In this exercise, we prove that the stationary distribution of this Markov chain is equal to the target distribution $P(x)$. *Remark:* While we show, that Metropolis-Hastings converges to the correct distribution, the proof doesn't tell us how fast it converges. In practice, we typically use samples only after a 'burn-in' period, which allows the chain to converge.

(i) Show that if an unnormalized distribution Q on Ω satisfies the detailed balance equations,

$$Q(x)T(x, y) = Q(y)T(y, x), \text{ for all } x, y \in \Omega, \tag{1}$$

then $\pi(x) = \frac{1}{Z}Q(x)$ is the stationary distribution of the Markov chain defined by the transition kernel $T(x, x')$.

(ii) Show that if Metropolis-Hastings transitions to a new state, i.e. $x_{t+1} \neq x_t$, then the transition probability $T(x_t, x_{t+1})$ can be written as

$$T(x_t, x_{t+1}) = \frac{1}{Q(x_t)} \min\left(Q(x_t)R(x_{t+1}|x_t), Q(x_{t+1})R(x_t|x_{t+1})\right). \tag{2}$$

Use this to show that the detailed balance equation for Q is satisfied if $x_{t+1} \neq x_t$.

(iii) Finally, show that if $x_t = x_{t+1}$, the detailed balance condition is trivially satisfied. *Remark:* You can still compute the transition probability $T(x_t, x_{t+1})$ for this case, but the result follows independent of the exact transition probability.

Solution

(i) If $Q(x)$ satisfies detailed balance, then so does the normalized distribution $\pi(x) = \frac{1}{Z}Q(x)$. Now, suppose that $P(X_t = x) = \pi(x)$. Then, for any $y \in \Omega$:

$$\begin{aligned}
P(X_{t+1} = y) &= \sum_{x\in\Omega} P(X_{t+1} = y, X_t = x) \\
&= \sum_{x\in\Omega} P(X_{t+1} = y | X_t = x) P(X_t = x) && \text{chain rule} \\
&= \sum_{x\in\Omega} T(x,y)\pi(x) && \text{def. of T \& } X_t \sim \pi \\
&= \sum_{x\in\Omega} \pi(y) T(y,x) && \text{detailed balance} \\
&= \sum_{x\in\Omega} \pi(y) P(X_{t+1} = x | X_t = y) && \text{def. of T} \\
&= \pi(y)\ . && \text{probs. sum to 1}
\end{aligned}$$

(ii) From algorithm 1, we get

$$\begin{aligned}
T(x_t, x_{t+1}) &= P(X_{t+1} = x_{t+1} | X_t = x_t) && \text{def. of T} \\
&= (\text{prob. to draw } x_{t+1} \text{ from } R \text{ given } X_t = x_t) \times (\text{prob. to accept } x_t) \\
&= \alpha R(x_{t+1}|x_t) && \text{def. of } \alpha \text{ and } R \\
&= \min\left(1, \frac{Q(x_{t+1})R(x_t|x_{t+1})}{Q(x_t)R(x_{t+1}|x_t)}\right) R(x_{t+1}|x_t) && \text{def. of } \alpha \\
&= \frac{1}{Q(x_t)} \min\left(Q(x_t)R(x_{t+1}|x_t), Q(x_{t+1})R(x_t|x_{t+1})\right)\ , && \times\tfrac{Q(x_t)}{Q(x_t)}
\end{aligned}$$

which proves (2). Exchanging x_t and x_{t+1} in (2) , we then get that:

$$\begin{aligned}
Q(x_{t+1})T(x_{t+1}, x_t) &= \min\left(Q(x_{t+1})R(x_t|x_{t+1}), Q(x_t)R(x_{t+1}|x_t)\right) \\
&= \min\left(Q(x_t)R(x_{t+1}|x_t), Q(x_{t+1})R(x_t|x_{t+1})\right) \\
&= T(x_t, x_{t+1})Q(x_t)\ .
\end{aligned}$$

which proves detailed balance for Q and T.

(iii) If $x_{t+1} = x_t$, equation $Q(x_{t+1})T(x_{t+1}, x_t) = T(x_t, x_{t+1})Q(x_t)$ trivially holds (for any Q and T). Combining question 2 & 3, we have proven that the latter equation holds for any $x_t, x_{t+1} \in \Omega$. Hence Q satisfies detailed balance, which, by question 1, shows that $P(x) = \frac{1}{Z}Q(x)$ is the stationary distribution of the Markov process defined by the Metropolis-Hastings algorithm.

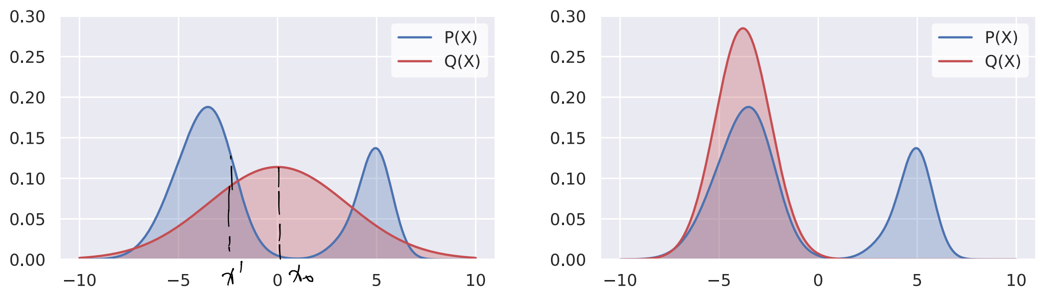
3. Forward and Reverse KL for Variational Inference

In variational inference we seek to find a distribution Q in a class of distributions $\mathcal{Q}$, that minimizes the KL-distance to a target distribution P, i.e.

$$Q \in \arg\min_{Q\in\mathcal{Q}} \mathrm{KL}(Q\|P)\ . \tag{3}$$

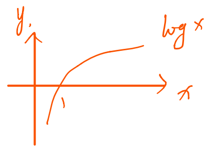
$Q(x) \to 0$ $\frac{P(x)}{Q(x)} \to 0$

The KL-distance (for finite support) is defined as $\mathrm{KL}(Q\|P) = \sum_x Q(x)\log(Q(x)/P(x))$. The KL-distance is not symmetric, so in general, $\mathrm{KL}(Q\|P) \neq \mathrm{KL}(P\|Q)$. We refer to $\mathrm{KL}(P\|Q)$ as *forward KL* and $\mathrm{KL}(Q\|P)$ as *reverse KL*. It is possible to use both, forward and reverse KL for variational inference, but the approximations will be different. In the plots below, we fit the true distribution $P(x)$ with a Gaussian $Q(x)$, using either forward or reverse KL. Explain which KL was used in either case!



Solution

$P(x)\log P(x) - P(x)\log Q(x)$



- Left: Forward-KL, i.e. $\mathrm{KL}(P\|Q)$. Justification:

  When $Q(x)$ tends to 0 while $P(x)$ stays bounded away from it $(P(x)/Q(x) \to 0)$, the integrand of the forward-KL, $P(x)\log\frac{P(x)}{Q(x)}$, diverges to $+\infty$. Hence the minimizer Q^* of $\mathrm{KL}(P\|Q)$ cannot have large regions where $P(x)/Q(x) \ll 1$, which excludes the right plot. Said differently, the reverse KL is visibly much smaller on the left plot (where Q roughly covers both modes of P) than on the right plot (where Q only covers one of the two modes of P).

  被积函数.

- Right: Reverse-KL, i.e. $\mathrm{KL}(Q\|P)$. Justification:

  Conversely, when $Q(x)$ tends to 0 while $P(x)$ stays bounded away from 0 $(Q(x)/P(x) \to 0)$, the integrand of the reverse-KL, $Q(x)\log\frac{Q(x)}{P(x)}$ tends to 0. Hence, the distribution $Q^\star$ that minimizes the reverse-KL might not cover all modes of P.