Notes on the KL-divergence between a Markov chain and its equilibrium distribution

Iain Murray and Ruslan Salakhutdinov

June 24, 2008

Abstract

After drawing a sample from a distribution, further correlated samples can be obtained by simulating a Markov chain that leaves the target distribution stationary. Often drawing even one sample from a distribution of interest is intractable, so the Markov chain is initialized arbitrarily. This note considers the marginal distribution over the Markov chain's position at each time step. We show that this marginal *never* moves further away from the chain's stationary distribution, as measured by KL-divergence either way around. This is a known result (Cover and Thomas, 1991). The presentation here is for review purposes only.

1 Introduction

Markov chain Monte Carlo (MCMC) is a method for drawing correlated samples from a target distribution of interest, $\pi(x)$. Usually, a Markov chain with unique equilibrium distribution equal to $\pi(x)$ is simulated, generating a sequence of states $x_1, x_2, \dots x_T$. If the chain was initialized at the equilibrium distribution: $x_0 \sim \pi$, then the marginal distribution over each state the chain visits is correct: $p(x_t) = \pi(x_t)$.

Often drawing even one sample from a distribution of interest is intractable, so the Markov chain is initialized arbitrarily. Then the distribution over the position at each iteration, q_t, is biased away from the target equilibrium distribution. Most users pick Markov chains where this bias disappears over time: $\lim_{t\to\infty} q_t(x) = \pi(x)$.

This note considers the marginal distribution over a Markov chain's position at each time step. We see that this marginal *never* moves further away from the target equilibrium distribution, as measured by KL-divergence either way around. This has relevance to learning algorithms that use only one or a few Markov chain steps per iteration, which are sometimes motivated by improvements in KL.

2 Preliminaries

We will consider a single Markov chain transition drawn from a distribution T. That is, if the marginal distribution over the chain's position at time t is $q_t(x)$, the distribution at the next time step is:

$$q_{t+1}(x') = \sum_x T(x' \!\leftarrow\! x)\, q_t(x). \tag{1}$$

In this note, our only requirement for T is that given a sample from $\pi(x)$, the marginal distribution over the next state in the chain is also the target distribution of interest π:

$$\pi(x') = \sum_x T(x' \!\leftarrow\! x)\, \pi(x) \quad \text{for all } x'. \tag{2}$$

That is, it leaves the target distribution stationary.

Given any transition operator T satisfying the stationary condition, (2), we can attempt to construct a *reverse operator* $\widetilde{T}$ defined by

$$\widetilde{T}(x \!\leftarrow\! x') \propto T(x' \!\leftarrow\! x)\, \pi(x) = \frac{T(x' \!\leftarrow\! x)\, \pi(x)}{\sum_x T(x' \!\leftarrow\! x)\, \pi(x)} = \frac{T(x' \!\leftarrow\! x)\, \pi(x)}{\pi(x')}. \tag{3}$$

Operators satisfying detailed balance are their own reverse operator.

Technical detail: we have not defined the reverse operator for moves starting from states with zero target probability: $\pi(x')=0$. This could be an issue; some Markov chains define $T(x' \leftarrow x)$ between pairs of states that both have zero target probability, i.e. $\pi(x)=\pi(x')=0$, and may be initialized in such zero-probability states. In what follows we can avoid needing to specify the undefined reverse moves.

We will find it helpful to use p^+ to denote a positive function derived from a distribution p, but with zeros removed:

$$p^+(x) = \begin{cases} p(x) & p(x) > 0 \\ \epsilon & p(x) = 0, \end{cases} \tag{4}$$

where ϵ is an arbitrary positive value.

3 The results

The KL-divergence measured under the target distribution never gets worse:

$$\mathrm{KL}[\pi \,\|\, q_{t+1}] = \sum_{x':\pi(x')>0} \pi(x') \,\log \frac{\pi(x')}{q_{t+1}(x')}$$

If ever $q_{t+1}(x')=0$ when $\pi(x') > 0$ then the KL is defined to be infinite.

$$= \sum_{x':\pi(x')>0} \pi(x') \,\log \frac{\pi(x')}{\sum_x T(x' \leftarrow x)\, q_t(x)} \qquad \text{(substituting (1))}$$

If $q_{t+1}(x')=0$ or a restricted sum $\sum_{x:\pi(x)>0} T(x' \leftarrow x) q_t(x)=0$, then equations (1) and (2) imply that $q_t(x) = 0$ for some x where $\pi(x) > 0$. Therefore the previous divergence, $\mathrm{KL}[\pi \,\|\, q_t]$, was infinite and $\mathrm{KL}[\pi \,\|\, q_{t+1}]$ can be no worse. We now consider the remaining cases:

$$\leq -\sum_{x':\pi(x')>0} \pi(x') \,\log \sum_{x:\pi(x)>0} T(x' \leftarrow x) \frac{q_t(x)}{\pi(x')}$$

$$= -\sum_{x':\pi(x')>0} \pi(x') \,\log \sum_{x:\pi(x)>0} \widetilde{T}(x \leftarrow x') \frac{q_t(x)}{\pi(x)} \qquad \text{(substituting (3), minus sign flips log)}$$

When $\pi(x')>0$ and $\pi(x)=0$, $\widetilde{T}(x \leftarrow x')$ is defined and zero:

$$= -\sum_{x':\pi(x')>0} \pi(x') \,\log \sum_{x} \widetilde{T}(x \leftarrow x') \frac{q_t(x)}{\pi^+(x)}$$

$$\leq -\sum_{x':\pi(x')>0} \pi(x') \sum_{x} \widetilde{T}(x \leftarrow x') \,\log \frac{q_t(x)}{\pi^+(x)} \qquad \text{(Jensen's inequality, average under } \widetilde{T}\text{)}$$

The minus sign flips the log, and we note that (3) implies $\sum_{x'} \widetilde{T}(x \leftarrow x')\, \pi(x') = \pi(x)$:

$$= \sum_{x} \pi(x) \,\log \frac{\pi^+(x)}{q_t(x)} = \sum_{x:\pi(x)>0} \pi(x) \,\log \frac{\pi(x)}{q_t(x)}$$

$\mathrm{KL}[\pi \,\|\, q_{t+1}] \leq \mathrm{KL}[\pi \,\|\, q_t]$.

The divergence measured the other way around also never increases:

$$\mathrm{KL}[q_{t+1} \,\|\, \pi\,] = \sum_{x':q_{t+1}(x')>0} q_{t+1}(x')\,\log\frac{q_{t+1}(x')}{\pi(x')}$$

If $\pi(x') = 0$ for some x' where $q_{t+1}(x') > 0$ then the KL is defined to be infinite. From equations (1) and (2) we see that there must be some x for which $q_t(x) > 0$ and $\pi(x) = 0$. Therefore, the previous divergence, $\mathrm{KL}[q_t \,\|\, \pi\,]$, was also infinite. From now on we assume that the KL is finite.

$$\begin{aligned}
&= \sum_{x'} q_{t+1}(x')\,\log\frac{q^{+}_{t+1}(x')}{\pi^{+}(x')}\\
&= \sum_{x'}\sum_{x:q_t(x)>0} T(x' \!\leftarrow\! x)\,q_t(x)\,\log\frac{q^{+}_{t+1}(x')}{\pi^{+}(x')} \qquad \text{(from (1))}\\
&\le \sum_{x:q_t(x)>0} q_t(x)\,\log\sum_{x'}\frac{T(x' \!\leftarrow\! x)}{\pi^{+}(x')}\,q^{+}_{t+1}(x') \qquad \text{(Jensen's, average under } T\text{)}
\end{aligned}$$

If there is an x where $\pi(x) = 0$ and $q_t(x) > 0$ then $\mathrm{KL}[q_t \,\|\, \pi\,]$ was infinite and the new KL can be no worse. We now assume $\pi(x) > 0$ whenever $q_t(x) > 0$. Then, in the expression, $T(x' \leftarrow x) > 0$ only if $\pi(x') > 0$. Also, note that $T(x' \leftarrow x) > 0$ implies $q_{t+1}(x') > 0$ for $q_t(x) > 0$.

$$\begin{aligned}
&= \sum_{x:q_t(x)>0} q_t(x)\,\log\sum_{x':\pi(x')>0}\frac{\widetilde{T}(x \!\leftarrow\! x')}{\pi(x)}\,q_{t+1}(x') \qquad \text{(substituted (3))}\\
&= \sum_{x:q_t(x)>0} q_t(x)\,\log\frac{q_t(x)}{\pi(x)}\,\frac{\sum_{x':\pi(x')>0}\widetilde{T}(x \!\leftarrow\! x')\,q_{t+1}(x')}{q_t(x)}\\
&\le \mathrm{KL}[q_t \,\|\, \pi\,] + \log\sum_{x:q_t(x)>0}\;\sum_{x':\pi(x')>0}\widetilde{T}(x \!\leftarrow\! x')\,q_{t+1}(x') \qquad \text{(Jensen's again. Second term is} \le 0.\text{)}
\end{aligned}$$

$$\mathrm{KL}[q_{t+1} \,\|\, \pi\,] \le \mathrm{KL}[q_t \,\|\, \pi\,]\,.$$

4 Discussion

Reducing the KL divergences to be near zero will require more than one step of an ergodic Markov chain in general.

Note that other divergence measures between distributions, such as $\max(|\pi(x) - q(x)|)$, *can* increase after a single Markov chain step. It is pleasantly surprising that neither KL ever transiently increases, and that the only technical condition is (2).

In fact both of these results drop out from a more general treatment result given in Cover and Thomas (1991, section 2.9). There has since been a generalization to the whole family of so-called alpha divergence functions, which include the KL-divergences (Friedman et al., 2007).

Acknowledgements

Thanks to Geoffrey Hinton for stating the result for Gibbs sampling; Ilya Sutskever for pointing us to Cover and Thomas; Ryan Adams for suggesting looking for generalizations to alpha divergences. Iain Murray's research is supported by the government of Canada.

References

T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, Inc., 1991. ISBN 0-471-06259-6.

C. Friedman, J. Huang, and S. Sandow. A utility-based approach to some information measures. *Entropy*, 9:1–26, 2007.