Bayesian Posterior Sampling via Stochastic Gradient Fisher Scoring
Supplementary Material

Abstract

This file contains the proof of a theorem stated in the main narrative.

In section 3.3 we state the following: “In the supplementary material we show that this online average converges to $I_1$ plus $O(1/N)$ corrections if we can assume that the samples are actually drawn from the posterior.” Here will will state this theorem more precisely and provide a proof.

**Theorem 1.** Consider a sampling algorithm which generates a sample $\theta_i$ from the posterior distribution of the model parameters $p(\theta|X_N)$ in each iteration $t$. In each iteration, we draw a random mini-batch of size $n$, $X_n^t = \{x_{t1},...,x_{tn}\}$ and compute the empirical covariance of the scores $V(\theta_i;X_n^t) = \frac{1}{n-1} \sum_{i=1}^{n} \{g(\theta_i;x_{ti}) - \bar{g}_n(\theta_i)\}\{g(\theta_i;x_{ti}) - \bar{g}_n(\theta_i)\}^T$. Let $V_T$ be the average of $V(\theta_i)$ across $T$ iterations. For large $N$, as $T \to \infty$, $V_T$ converges to the Fisher information $I(\theta_0) + O(\frac{1}{N})$ i.e.

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} V(\theta_i;X_n^t) = I(\theta_0) + O(\frac{1}{N}) \quad (1)$$

Proof. We will now provide an outline of the proof. First, note that using a little algebra, we can rewrite $V(\theta_i;X_n^t)$ as:

$$V(\theta_i;X_n^t) = \frac{1}{n} \sum_{i=1}^{n} g(\theta_i;x_{ti})g(\theta_i;x_{ti})^T - \frac{1}{n(n-1)} \sum_{i=1,j=1,i \neq j}^{n} g(\theta_i;x_{ti})g(\theta_i;x_{tj})^T \quad (2)$$

As $T \to \infty$, the averaging of $V(\theta_i,X_n^t)$ across different iterations is equivalent to taking an expectation of $V(\theta_i,X_n^t)$ w.r.t the posterior distribution of $\theta_i$. Also, when $N$ is very large, averaging across the different mini-batches $X_n^t$ drawn in different iterations is equivalent to taking an expectation w.r.t $p(x;\theta_0)$. Thus we have:

$$V_T = E_{\theta_i,X_n^t} [V(\theta_i,X_n^t)] = E_{\theta,x} \left[ g(\theta;x)g(\theta;x)^T \right] - E_{\theta,x,x'} \left[ g(\theta;x)g(\theta;x')^T \right] \quad (3)$$

According to the Bernstein-von Mises theorem, $\theta|X_N \sim N(\theta_0, \frac{I^{-1}(\theta_0)}{N})$ for large $N$. This means that we can consider $\theta \perp x_0$ and the expectations in Eqn. (3) can be taken w.r.t $\theta$ and $x$ independently. Also, since most of the posterior density is in a small region around $\theta_0$, we can consider a Taylor series expansion of $g(\theta;x)$ around $\theta_0$, (treating $\theta$ as a scalar for simplicity):

$$g(\theta;x) = \sum_{k=0}^{\infty} \frac{g^{(k)}(\theta_0;x)(\theta - \theta_0)^k}{k!} \quad (4)$$

We will now apply this series expansion to each of the terms in Eqn. (3). Thus we have:

$$E_{\theta,x} \left[ \{g(\theta;x)\}^2 \right] = \sum_{k=0}^{\infty} A_k(\theta_0)E_{\theta} \left[ (\theta - \theta_0)^k \right] \quad (5)$$

where,

$$A_k(\theta_0) = \sum_{m=0}^{k} \frac{E_x \left[ g^{(m)}(\theta_0;x)g^{(k-m)}(\theta_0;x) \right]}{m!(k-m)!} \quad (6)$$

Since $\theta$ has a Gaussian distribution, its odd central moments are zero, and the even central moments are given by:

$$E_{\theta} \left[ (\theta - \theta_0)^{2r} \right] = \frac{(2r)!}{2^r r!} \left[ \frac{I^{-1}(\theta_0)}{N} \right]^r \quad (7)$$

Thus, we have:

$$E_{\theta,x} \left[ \{g(\theta;x)\}^2 \right] = \sum_{r=0}^{\infty} A_{2r}(\theta_0) \frac{(2r)!}{2^r r!} \left[ \frac{I^{-1}(\theta_0)}{N} \right]^r \quad (8)$$
Similarly, we can show that:

\[
E_{\theta, x, x'} [g(\theta; x)g(\theta; x')] = \sum_{r=0}^{\infty} B_{2r}(\theta_0) \frac{(2r)!}{2^r r!} \left[ \frac{I^{-1}(\theta_0)}{N} \right]^r
\]

where,

\[
B_k(\theta_0) = \sum_{m=0}^{k} \frac{E_x \left[ g^{(m)}(\theta_0; x) \right] E_x \left[ g^{(k-m)}(\theta_0; x) \right]}{m!(k-m)!}
\]

From Eqns. (8) and (9), we have for large \(N\):

\[
\lim_{T \to \infty} V_T = \sum_{r=0}^{\infty} C_{2r}(\theta_0) \frac{(2r)!}{2^r r!} \left[ \frac{I^{-1}(\theta_0)}{N} \right]^r
\]

where,

\[
C_k(\theta_0) = A_k(\theta_0) - B_k(\theta_0)
\]

For example,

\[
C_0(\theta_0) = I(\theta_0)
\]

\[
C_2(\theta_0) = E_x \left[ g(\theta_0; x)g^{(2)}(\theta_0; x) + g^{(1)}(\theta_0; x)g^{(1)}(\theta_0; x) \right] - I^2(\theta_0)
\]

Thus for large \(N\),

\[
\lim_{T \to \infty} V_T = I(\theta_0) - \frac{I(\theta_0)}{N} + \frac{I^{-1}(\theta_0)}{N} E_x \left[ g(\theta_0; x)g^{(2)}(\theta_0; x) + g^{(1)}(\theta_0; x)g^{(1)}(\theta_0; x) \right] + \sum_{r=2}^{\infty} C_{2r}(\theta_0) \frac{(2r)!}{2^r r!} \left[ \frac{I^{-1}(\theta_0)}{N} \right]^r
\]

\[
= I(\theta_0) + O\left( \frac{1}{N} \right)
\]

\(\square\)