Theory for CDA

JEJ

This version: January 15, 2017

Here we describe some basic theoretical properties of CDA Gibbs and M-H. We show that CDA Gibbs is ergodic, and has lower autocorrelation at stationarity than the original Gibbs sampler from which it is derived. We also show that CDA M-H is ergodic.

Consider a data augmentation Gibbs sampler with generic update rule

$$z \mid y, \theta \sim g(z \mid \theta, y)$$
$$\theta \mid y, z \sim f(\theta \mid z, y)$$

The lag-1 autocorrelation of a function $f : \Theta \to \mathbb{R}$ at stationarity can be expressed as the Bayesian fraction of missing information [CITES]

$$\gamma_f = 1 - \frac{\mathbb{E}[\mathrm{var}(f(\theta) \mid z)]}{\mathrm{var}(f(\theta))},$$

where the integrals are with respect to the invariant measure Π. Let

$$L_2(\Pi) = \left\{ f : \Theta \to \mathbb{R}, \int_{\theta \in \Theta} \{f(\theta)\}^2 \Pi(d\theta) < \infty \right\}$$

be the set of real-valued functions square-integrable with respect to the invariant measure. The *maximal autocorrelation*

$$\gamma = \sup_{f \in L^2(\Pi)} \gamma_f = 1 - \inf_{f \in L^2(\Pi)} \frac{\mathbb{E}[\mathrm{var}(f(\theta) \mid z)]}{\mathrm{var}(f(\theta))}$$

is equal to the geometric convergence rate of the data augmentation Gibbs sampler [CITES].

[CITES] describe methods for improving mixing in data augmentation Gibbs by reparametrization. For example, noncentered parametrizations of the model, which are reparametrizations that result in $z \perp \theta$ *a priori*, can often significantly improve mixing when the corresponding centered parametrization mixes poorly. However, reparametrization does not change the invariant measure, a constraint that can limit the effectiveness of the strategy purely from the point of view of generating fast mixing Markov chains.

In contrast, CDA does not require that the invariant measure be preserved. It introduces additional parameters r, b, resulting in a new family of likelihoods $L_{r,b}(\theta; y)$, with corresponding posterior $\Pi_{r,b}(\theta; y) \propto L_{r,b}(\theta; y)\Pi_0(\theta)$. It is easy to show that the resulting algorithm is ergodic.

REMARK **0.1** (ergodicity). *For fixed r, b, CDA Gibbs is ergodic with invariant measure $\Pi_{r,b}(z, \theta)$. Moreover, a Metropolis-Hastings algorithm with proposal kernel $q(\theta'; \theta)$ equal to the θ-marginal CDA Gibbs transition kernel $K_{r,b}(\theta'; \theta)$ with fixed r, b is ergodic with invariant measure $\Pi(\theta; y)$.*

Proof. For any r, b, the conditionals $\Pi_{r,b}(z \mid \theta)$ and $\Pi_{r,b}(\theta \mid z)$ are well-defined for all $z \in \mathcal{Z}, \theta \in \Theta$, and therefore the Gibbs transition kernel $K_{r,b}(\theta', z'; \theta, z)$ and corresponding marginal kernel

$K_{r,b}(\theta';\theta)$ are well-defined. Moreover, for any $(z,\theta) \in \mathcal{Z} \times \Theta$, we have $\mathbb{P}[(\theta', z') \in A \mid (\theta, z)] > 0$ whenever $\Pi_{r,b}(A) > 0$. Thus $K_{r,b}$ is aperiodic and $\Pi_{r,b}$-irreducible.

Let

$$\Pi_{r,b}(\theta) = \int \Pi_{r,b}(\theta, z)dz$$

be the θ-marginal of the invariant measure for any r, b. $q(\theta';\theta) = K_{r,b}(\theta';\theta)$ is aperiodic and $\Pi_{r,b}(\theta)$-irreducible. Thus, it is also $\Pi(\theta)$-irreducible so long as $\Pi(\theta) \gg \Pi_{r,b}(\theta)$. Since $\Pi(\theta), \Pi_{r,b}(\theta)$ are supported on the same subset $\Theta \subset \mathbb{R}^p$, $\Pi_{r,b}(\theta)$-irreducibility implies $\Pi(\theta)$ irreducibility. Moreover, $q(\theta';\theta) > 0$ everywhere on Θ. Thus, by Theorem 3 of [CITE Roberts 1994], CDA M-H is Π-irreducible and aperiodic. ♢

JEJ: some of the statements above should really become assumptions about CDA gibbs, for example that $\mathbb{P}[(\theta', z') \in A \mid (\theta, z)] > 0$ whenever $\Pi_{r,b}(A) > 0$.

Having established ergodicity of both CDA Gibbs and CDA M-H, we now provide a semi-rigorous argument for why our approach to tuning r and b results in both rapid convergence and closeness of $\Pi_{r,b}$ to Π. In general, it is possible to choose r to set

$$\mathbb{E}_{\Pi_{r,b}}[\mathrm{var}(\theta \mid z)] = \mathrm{var}_{\Pi_{r,b}}(\theta)$$

for *any value of* b, although an analytic expression for the correct value of r may not be available. JEJ: we should have more justification for this – is it always true when r is a scale parameter and b a location parameter? What do we need for this to be true?. By tuning r during the adaptation phase to make the lag-1 autocorrelation for the identity function small, we can numerically approximate the correct value of r.

This is obviously much weaker than minimizing the autocorrelation for worst-case functions. However, for the sake of exposition, we will proceed on the assumption that (1) we can make the lag-1 autocorrelation for the identity function zero by appropriately tuning r and (2) this is sufficient to obtain a Gibbs transition kernel that generates nearly *independent* samples. This makes the rationale for tuning b to increase the Metropolis acceptance probability much clearer. First, we note the form of the Metropolis acceptance ratios

REMARK **0.2**. *The CDA M-H acceptance ratio is given by*

$$\frac{L(\theta';y)\Pi_0(\theta')q(\theta;\theta')}{L(\theta;y)\Pi_0(\theta)q(\theta';\theta)} = \frac{L(\theta';y)L_{r,b}(\theta;y)}{L(\theta;y)L_{r,b}(\theta';y)} \tag{1}$$

Proof. Since $q(\theta;\theta') = K_{r,b}(\theta;\theta')$ is the θ marginal of a Gibbs transition kernel, and Gibbs is reversible on its margins, we have

$$q(\theta;\theta')\Pi_{r,b}(\theta') = q(\theta';\theta)\Pi_{r,b}(\theta),$$

and so

$$\begin{aligned}\frac{L(\theta';y)\Pi_0(\theta')q(\theta;\theta')}{L(\theta;y)\Pi_0(\theta)q(\theta';\theta)} &= \frac{L(\theta';y)\Pi_0(\theta')L_{r,b}(\theta;y)\Pi_0(\theta)}{L(\theta;y)\Pi_0(\theta)L_{r,b}(\theta';y)\Pi_0(\theta')} \\ &= \frac{L(\theta';y)L_{r,b}(\theta;y)}{L(\theta;y)L_{r,b}(\theta';y)}.\end{aligned}$$

♢

The expression in (1) will be near 1 at stationarity if

$$\int \log\left(\frac{L(\theta';y)L_{r,b}(\theta;y)}{L(\theta;y)L_{r,b}(\theta';y)}\right) K_{r,b}(\theta';\theta)\Pi(\theta)d\theta \approx 0.$$

Now, suppose that a Markov chain evolving according to $K_{r,b}$ is rapidly mixing, so that for starting measures satisfying a condition like

$$\sup_A \frac{\nu(A)}{\Pi_{r,b}(A)} < M$$

for M not too large we have

$$\mathrm{KL}\left(\Pi_{r,b} \parallel \int K_{r,b}(\theta';\theta)\nu(d\theta)\right) \quad \text{small.}$$

Then the symmetric KL is

$$\begin{aligned}
\mathrm{KL}\left(\Pi_{r,b} \parallel \Pi\right) + \mathrm{KL}\left(\Pi \parallel \Pi_{r,b}\right) &= \int \Pi_{r,b}(d\theta)\log\frac{\Pi_{r,b}(\theta)}{\Pi(\theta)} + \int \Pi(d\theta)\log\frac{\Pi(\theta)}{\Pi_{r,b}(\theta)} \\
&= \int \Pi_{r,b}(d\theta)\log\frac{c_{r,b}L_{r,b}(\theta)\Pi_0(\theta)}{cL(\theta)\Pi_0(\theta)} + \int \Pi(d\theta)\log\frac{cL(\theta)\Pi_0(\theta)}{c_{r,b}L_{r,b}(\theta)\Pi_0(\theta)} \\
&\approx \int K_{r,b}(\theta';\theta)\Pi(d\theta)\log\frac{L_{r,b}(\theta')}{L(\theta')} + \int \Pi(d\theta)\log\frac{L(\theta)}{L_{r,b}(\theta)} \\
&= \mathbb{E}\left[\frac{L_{r,b}(\theta')L(\theta)}{L_{r,b}(\theta)L(\theta')}\right],
\end{aligned}$$

for $\theta \sim \Pi$ and $\theta' \mid \theta \sim K_{r,b}(\theta';\theta)$, so that tuning b to make the M-H acceptance ratio larger will tend to make the symmetric KL between $\Pi_{r,b}$ and Π small. This justifies the approach of using the acceptance ratio to tune b. As the acceptance ratio approaches 1, CDA M-H and CDA Gibbs coincide, and the CDA Gibbs invariant measure is identically Π, but the corresponding Gibbs sampler converges rapidly.