Analysis of the Gibbs sampler for a model related to James-Stein estimators

by

Jeffrey S. Rosenthal*

Department of Statistics
University of Toronto
Toronto, Ontario
Canada M5S 1A1

Phone: (416) 978-4594. Internet: jeff@utstat.toronto.edu

(Appeared in *Statistics and Computing* **6** (1996), 269–275.)

Summary. We analyze a hierarchical Bayes model which is related to the usual empirical Bayes formulation of James-Stein estimators. We consider running a Gibbs sampler on this model. Using previous results about convergence rates of Markov chains, we provide rigorous, numerical, reasonable bounds on the running time of the Gibbs sampler, for a suitable range of prior distributions. We apply these results to baseball data from Efron and Morris (1975). For a different range of prior distributions, we prove that the Gibbs sampler will fail to converge, and use this information to prove that in this case the associated posterior distribution is non-normalizable.

Acknowledgements. I am very grateful to Jun Liu for suggesting this project, and to Neal Madras for suggesting the use of the Submartingale Convergence Theorem herein. I thank Kate Cowles and Richard Tweedie for helpful conversations, and thank the referees for useful comments.

1. Introduction.

Markov chain Monte Carlo techniques, including the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), data augmentation (Tanner and Wong, 1986), and the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990) have become very popular in recent years as a way of generating a sample from a complicated probability distribution (such as the posterior distribution in a Bayesian inference problem). A fundamental issue regarding such techniques is their convergence properties, specifically whether or not the algorithm will converge to the correct distribution, and if so how quickly. In addition to the many general convergence results (e.g. Tierney, 1991) and convergence diagnostics (e.g. Roberts, 1992; Mykland, Tierney, and Yu, 1992) which have been developed, a number of papers have attempted to prove rigorous bounds on convergence rates for these algorithms (Jerrum and Sinclair, 1989; Amit and Grenander, 1991; Frieze, Kannan, and Polson, 1994; Meyn and Tweedie, 1994; Lund and Tweedie, 1993; Mengersen and Tweedie, 1993; Frigessi *et al.*, 1993; Rosenthal, 1993, 1995a, 1995b). However, most of the results are of a quite specific and limited nature, and the general question of convergence rates for these algorithms remains problematic and largely unsolved.

In this paper we investigate the convergence properties of the Gibbs sampler as applied to a particular hierarchical Bayes model. The model is related to James-Stein estimators (James and Stein, 1961; Efron and Morris, 1973, 1975; Morris, 1983). Briefly, James-Stein estimators may be defined as the mean of a certain empirical Bayes posterior distribution (as discussed in the next section). We consider the problem of using the Gibbs sampler as a way of sampling from a richer posterior distribution, as suggested by Jun Liu (personal communication). Such a technique would eliminate the need to estimate a certain parameter empirically and to provide a "guess" at another one, and would give additional information about the distribution of the parameters involved.

We consider, in particular, the convergence properties of this Gibbs sampler. For a certain range of prior distributions, we establish (Section 3) rigorous, numerical, reasonable rates of convergence. The bounds are obtained using the methods of Rosenthal (1995b). We thus rigorously bound the running time for this Gibbs sampler to converge to the posterior distribution, within a specified accuracy (as measured by total variation distance). We provide a general formula for this bound, which is of reasonable size, in terms of the prior distribution and the data. This Gibbs sampler is perhaps the most complicated example to date for which reasonable quantitative convergence rates have been obtained. We apply our bounds to the numerical baseball data of Efron and Morris (1975) and Morris (1983), based on batting averages of baseball players, and show that approximately 140 iterations are sufficient to achieve convergence in this case.

For a different range of prior distributions, we use the Submartingale Convergence Theorem to prove (Section 4) that this Gibbs sampler will in fact *not* converge to the desired posterior distribution (in *any* amount of time). On the other hand, standard theory indicates that, if the model were well-defined, this Gibbs sampler in fact *should* converge to this distribution. This apparent contradiction thus proves that the posterior distribution for this model, with this range of (improper) prior distributions, is not well defined, i.e. is itself improper. We have thus used the Gibbs sampler as a theoretical tool, to establish properties of the model itself. This suggests that analyses of Markov chain Monte Carlo algorithms may have important uses as analytical tools, in addition to their benefit in facilitating sampling. Furthermore, the example provides a cautionary note to the effect that naive use of the Gibbs sampler may lead to incorrect results (such as claiming convergence when there is actually nothing to converge to).

The precise model and Gibbs sampler studied are defined in Section 2. The development of rates of convergence is done in Section 3. The argument about improper posterior distributions is given in Section 4. Finally, for ease of reading, the more computational proofs are relegated to an Appendix.

Our emphasis throughout is on the theoretical analysis of this Gibbs sampler (with the hope that similar analyses will be applied to other models), rather than on specific implications for James-Stein estimators or for the particular statistical model at hand.

2. The model.

The empirical Bayes formulation of James-Stein estimators may be defined as follows. For $1 \leq i \leq K$, we observe data Y_i, where $Y_i \,|\, \theta_i \sim N(\theta_i, V)$ and are conditionally independent. Here θ_i are unknown parameters to be estimated, and $V > 0$ is assumed to be known (or estimated directly from the data). Furthermore $\theta_i \,|\, \mu, A \sim N(\mu, A)$ and are conditionally independent.

The standard James-Stein estimator for θ_i can be obtained (cf. Efron and Morris, 1975) as the posterior mean $\mathbf{E}(\theta_i \,|\, Y_i)$, where μ is taken to be an "initial guess" at the θ's, and where $(1 + A^2)^{-1}$ is replaced by its (unbiased) estimate $(K-2)/\sum(Y_i - \mu)^2$.

In this paper we follow the suggestion of Jun Liu (personal communication) to regard μ and A as further parameters to be estimated. The Bayesian approach then involves putting priors on μ and A, thus defining a posterior distribution

$$\pi(\cdot) \;=\; \mathcal{L}(A, \mu, \theta_1, \ldots, \theta_K \,|\, Y_1, \ldots, Y_K)\,.$$

To be specific, we use a flat prior for μ, and use a conjugate prior of the form $IG(a,b)$ for A (where $IG(a,b)$ is the inverse gamma distribution with density proportional to $e^{-b/x}x^{-(a+1)}$). We shall see that the chosen values of a and b can greatly affect the properties of $\pi(\cdot)$.

The remainder of this paper is thus concerned with the problem of sampling from the distribution $\pi(\cdot)$ defined above, with

$$Y_i \,|\, \theta_i \;\sim\; N(\theta_i, V) \qquad (1 \le i \le K)$$

$$\theta_i \,|\, \mu, A \;\sim\; N(\mu, A) \qquad (1 \le i \le K)$$

$$\mu \;\sim\; \text{flat prior on } \mathbf{R}$$

$$A \;\sim\; IG(a,b)\,.$$

To accomplish this sampling, we use a Gibbs sampler on $(A, \mu, \theta_1, \ldots, \theta_K)$. After choosing initial values $A^{(0)}, \mu^{(0)}, \theta_i^{(0)}$ from some initial distribution, we follow the suggestion of Jun Liu (personal communication) by letting the Gibbs sampler update these variables repeatedly (for iterations $k = 1, 2, 3, \ldots$) by the (easily-computed) conditional distributions

$$A^{(k)} \;\sim\; \mathcal{L}(A \,|\, \theta_i = \theta_i^{(k-1)}, Y_i) \;=\; IG\left(a + \frac{K-1}{2},\, b + \frac{1}{2}\sum(\theta_i^{(k-1)} - \overline{\theta}^{(k-1)})^2\right)\,;$$

$$\mu^{(k)} \;\sim\; \mathcal{L}(\mu \,|\, A = A^{(k)}, \theta_i = \theta_i^{(k-1)}, Y_i) \;=\; N(\overline{\theta}^{(k-1)}, A^{(k)}/K)\,;$$

$$\theta_i^{(k)} \;\sim\; \mathcal{L}(\theta_i \,|\, A = A^{(k)}, \mu = \mu^{(k)}, Y_i) \;=\; N\left(\frac{\mu^{(k)}V + Y_i A^{(k)}}{V + A^{(k)}}, \frac{A^{(k)}V}{V + A^{(k)}}\right)\,;$$

where $\overline{\theta}^{(k)} = \frac{1}{K}\sum \theta_i^{(k)}$. [From the point of view of the Gibbs sampler, this corresponds to treating (A, μ) as a single variable, and jointly updating $(A^{(k)}, \mu^{(k)}) \sim \mathcal{L}(A, \mu \,|\, \theta_i = \theta_i^{(k-1)}, Y_i)$.]

This definition specifies the distribution of the random variables $A^{(k)}, \mu^{(k)}, \theta_i^{(k)}$, for $k = 1, 2, 3, \ldots$. The Gibbs sampler is said to *converge* (in total variation distance $\|\cdot\|$) to the distribution $\pi(\cdot)$, if

$$\lim_{k\to\infty} \|\mathcal{L}(x^{(k)}) \;-\; \pi(\cdot)\| \;:=\; \lim_{k\to\infty} \sup_{S \subseteq \mathbf{R}^{K+2}} |P(x^{(k)} \in S) - \pi(S)| \;=\; 0\,,$$

where we have written $x^{(k)}$ as shorthand for $(A^{(k)}, \mu^{(k)}, \theta_1^{(k)}, \ldots, \theta_K^{(k)})$. If the Gibbs sampler does converge, then quantitative bounds on $\|\mathcal{L}(x^{(k)}) - \pi(\cdot)\|$ are important because they can determine how long the algorithm should be run (i.e. how large k needs to be) before $x^{(k)}$ can be regarded approximately as a sample from $\pi(\cdot)$.

We note that this Gibbs sampler is similar in form to the Gibbs sampler for variance components models, as suggested in Gelfand and Smith (1990), and partially analyzed (though not numerically) in Rosenthal (1995a). It is thus a realistic example of an applied use of the Gibbs sampler. However, like most uses of the Gibbs sampler, its convergence properties are not at all clear.

We analyze this Gibbs sampler in two different ways, corresponding to different ranges of values of a and b above. In Section 3, for certain ranges of a and b, we use results from Rosenthal (1995b) to get quantitative bounds on $\|\mathcal{L}(x^{(k)}) - \pi(\cdot)\|$ which converge to 0, and thus provide bounds on the required running times. In Section 4, we show that for certain other ranges of a and b, this Gibbs sampler will not converge to $\pi(\cdot)$ at all. We use this to prove that, for these values of a and b, the above model is improper.

3. Rates of convergence.

To get bounds on the rate of convergence of this Gibbs sampler, we recall a result from Rosenthal (1995b, Theorem 12).

Proposition 1. *Let $P(x, \cdot)$ be the transition probabilities for a Markov chain $X_0, X_1, X_2, \ldots$ on a state space $\mathcal{X}$, with stationary distribution $\pi(\cdot)$. Suppose there exist $\epsilon > 0$, $0 < \lambda < 1$, $0 < \Lambda < \infty$, $d > \frac{2\Lambda}{1-\lambda}$, a non-negative function $f : \mathcal{X} \to \mathbf{R}$, and a probability measure $Q(\cdot)$ on $\mathcal{X}$, such that*

$$\mathbf{E}\left(f(X_1) \,|\, X_0 = x\right) \;\leq\; \lambda f(x) + \Lambda\,, \qquad x \in \mathcal{X} \tag{1}$$

and

$$P(x, \cdot) \;\geq\; \epsilon\, Q(\cdot)\,, \qquad x \in f_d \tag{2}$$

where $f_d = \{x \in \mathcal{X} \,|\, f(x) \leq d\}$, and where $P(x, \cdot) \;\geq\; \epsilon\, Q(\cdot)$ means $P(x, S) \;\geq\; \epsilon\, Q(S)$ for every measurable $S \subseteq \mathcal{X}$. Then for any $0 < r < 1$, we have

$$\|\mathcal{L}(X_k) - \pi(\cdot)\| \;\leq\; (1-\epsilon)^{rk} \;+\; \left(\alpha^{-(1-r)}\gamma^r\right)^k \left(1 + \frac{\Lambda}{1-\lambda} + \mathbf{E}\left(f(X_0)\right)\right),$$

where

$$\alpha^{-1} \;=\; \frac{1 + 2\Lambda + \lambda d}{1+d} \;<\; 1\,; \qquad \gamma \;=\; 1 + 2(\lambda d + \Lambda)\,.$$

Equation (1) above is called a *drift condition*, while equation (2) above is called a *minorization condition*. These two conditions are surprisingly difficult to verify for complicated Markov chains, and although the above proposition appears to be quite general, it has been applied to date only in a few very specific examples.

To apply this result to the Gibbs sampler at hand, we need to choose a function $f(x)$. Intuitively, we need this function to have the dual properties that (1) if it is very large at one iteration, it tends to get smaller at the next, and (2) all values of x for which $f(x)$ is small have similar transition probabilities for the next iteration. We thus choose the function

$$f(x) \;=\; f(A, \mu, \theta_1, \ldots, \theta_K) \;=\; \sum_{i=1}^{K} (\theta_i - \overline{Y})^2 \;=\; K(\overline{\theta} - \overline{Y})^2 + \sum_{i=1}^{K} (\theta_i - \overline{\theta})^2 \,,$$

where $\overline{Y} = \frac{1}{K} \sum_i Y_i$. Intuitively, $f(x)$ is small if the values of θ_i are close to the average $\overline{Y}$ of the data values.

For this function, we have the following two key computational lemmas (proved in the Appendix).

Lemma 2. *Assume that $b > 0$ and $a > -\frac{K-1}{2}$. Then*

$$\mathbf{E}\left(f(x^{(k)}) \,|\, x^{(k-1)} = x \right) \;\le\; \lambda\, f(x) \;+\; \Lambda \,,$$

where

$$\begin{aligned} \lambda \;&=\; \mathbf{E}\left(1 + \frac{W}{V}\right)^{-2} \qquad \textit{with } W \sim IG\left(a + \frac{K-1}{2}, b\right) \\ &=\; \int_0^\infty \frac{b^{a+\frac{K-1}{2}} e^{-b/w}}{\Gamma(a + \frac{K-1}{2}) w^{a+\frac{K+1}{2}}} \left(1 + \frac{w}{V}\right)^{-2} dw \end{aligned}$$

and $\Lambda = \Delta + (K + \frac{1}{4})V$, with $\Delta = \sum (Y_i - \overline{Y})^2$.

Lemma 3. *Assume $b > 0$ and $a > -\frac{K-1}{2}$. Then there exists a probability measure $Q(\cdot)$ such that*

$$P(x, \cdot) \;\ge\; \epsilon\, Q(\cdot), \qquad x \in f_d$$

where

$$\epsilon \;=\; 2 \int_0^\infty dA \,\min\left[IG\left(a + \frac{K-1}{2}, b; A\right),\; IG\left(a + \frac{K-1}{2}, b + \frac{d}{2}; A\right)\right] \int_0^\infty d\mu\, N\left(\sqrt{\frac{d}{K}}, \frac{A}{K}; \mu\right),$$

with $IG(a, b; t) = \frac{b^a e^{-b/t}}{\Gamma(a) t^{a+1}}$ and $N(m, v; t) = \frac{1}{\sqrt{2\pi v}} e^{-(t-m)^2/2v}$ the density functions for the inverse gamma and normal distributions, respectively.

Putting the above three results together, we obtain

Theorem 4. *Assume* $b > 0$ *and* $a > -\frac{K-1}{2}$*. Then for any* $d > \frac{2\Lambda}{1-\lambda}$*, and any* $0 < r < 1$*, the Gibbs sampler defined in Section 2 satisfies*

$$\|\mathcal{L}(x^{(k)}) - \pi(\cdot)\| \leq (1-\epsilon)^{rk} + \left(\alpha^{-(1-r)}\gamma^r\right)^k \left(1 + \frac{\Lambda}{1-\lambda} + \mathbf{E}\left(f(x^{(0)})\right)\right),$$

where $\epsilon, \lambda, \Lambda, \alpha, \gamma$ *are as defined above.*

We apply this to the baseball data in Efron and Morris (1975) and Morris (1983). From the (modified) data of Table 1 of Morris (1983), we compute that $K = 18$, that $V = 0.00434$, and that $\Delta = 0.0822$, so that $\Lambda = 0.161$. We assign arbitrarily the prior values $a = -1$ (the same as for a flat prior) and $b = 2$ (since it has to be positive). We then compute, using numerical integration, that $\lambda = 0.000289$, Choosing $d = 1$, we then compute (using numerical double integration) that $\epsilon = 0.0656$.

From the formulas in Proposition 1, we compute that $\alpha^{-1} = 0.662$ and $\gamma = 1.32$. Hence, choosing $r = 0.5$ in an effort to "balance" the two terms involved, and noting that $\frac{\Lambda}{1-\lambda} < 0.17$, we conclude that

$$\|\mathcal{L}(x^{(k)}) - \pi(\cdot)\| \leq (0.967)^k + (0.935)^k \left(1.17 + \mathbf{E}\left(\sum(\theta_i^{(0)} - \overline{Y})^2\right)\right).$$

For example, if (say) we begin with $\theta_i^{(0)} = \overline{Y}$ for all i, and run the Gibbs sampler for $k = 140$ iterations, this bound is less than 0.009. We have thus proved that, in this case, 140 iterations suffice to achieve approximate convergence.

Remark. In principle, the method of proving rates of convergence used here could be used in any Markov chain Monte Carlo situation. However, the computations necessary to establish the drift and minorization conditions can be quite difficult in more complicated examples. It is to be hoped that, with time, these methods can be applied to more and more complicated Markov chains.

4. Improper posterior distributions.

In this section we prove that for a certain range of values of a, b, Δ, and V, the Gibbs sampler defined in Section 2 will in fact *not* converge to $\pi(\cdot)$ at all. We then use this to prove that for this range of values, the posterior distribution $\pi(\cdot)$ is in fact improper.

The key computation is the following.

Lemma 5. *Assume that $b \leq 0$, that $a \geq 1$, and that $\Delta \leq V$. Then for any $t > 0$,*

$$\mathbf{E}\left(A^{(k)} \mid A^{(k-1)} = t, x^{(k-2)}, x^{(k-3)}, \ldots\right) \;\leq\; t - \frac{t^3}{(V+t)^2} \;<\; t\,.$$

This lemma says that, under these hypotheses, the process $\{-A^{(k)}\}_{k=0}^{\infty}$ is a *submartingale*. Since the $A^{(k)}$ are non-negative, it follows from the Submartingale Convergence Theorem (see e.g. Billingsley, 1986, Theorem 35.4) that $A^{(k)}$ converges *almost surely*. That is, the values of $A^{(k)}$ actually converge to a fixed random variable (as opposed to converging in distribution). Since the correction term $\frac{t^3}{(V+t)^2}$ in the above lemma only goes to 0 as $t \to 0$, it must be that $A^{(k)}$ actually converges to 0 almost surely, so that $\mathcal{L}(A^{(k)})$ becomes more and more concentrated around 0.

On the other hand, if $\nu(\cdot)$ is any absolutely-continuous probability distribution on $\mathbf{R}^{K+2}$, then $\nu\{A = 0\} = 0$. Hence, it cannot be that $\mathcal{L}(x^{(k)})$ converges in distribution to $\nu(\cdot)$. We have thus proved

Proposition 6. *If $b \leq 0$, $a \geq 1$, and $\Delta \leq V$, and $\nu(\cdot)$ is any absolutely-continuous distribution, then the Gibbs sampler does not converge to $\nu(\cdot)$ from any initial distribution. In fact, $\lim_{k\to\infty} \|\mathcal{L}(x^{(k)}) - \nu(\cdot)\| = 1$.*

On the other hand, the density of $\pi(\cdot)$ is everywhere-positive. It follows from standard theory (see e.g. Tierney, 1991, Theorem 1; for related convergence-rates results, see Schervish and Carlin, 1992; Liu, Wong, and Kong, 1991a, 1991b; Baxter and Rosenthal, 1995) that, if $\pi(\cdot)$ were proper (i.e. normalizable), we would necessarily have $\mathcal{L}(x^{(k)})$ converging to $\pi(\cdot)$ in distribution, from almost every initial point. Since such $\pi(\cdot)$ would be absolutely continuous, the above proposition implies that this is impossible. This proves

Theorem 7. *If $b \leq 0$, $a \geq 1$, and $\Delta \leq V$, then the posterior distribution $\pi(\cdot)$ of the model defined in Section 2 is improper, i.e. the integral of the density function $(*)$ is infinite.*

Remarks.

(*i*) It is to be admitted that the choice of priors implied by $a \geq 1$ is not as natural as a flat prior (with $a = -1$ and $b = 0$). However, this theorem is interesting in that it indicates how rigorous theoretical analysis of a Gibbs sampler, associated with a given model, can imply information about the model itself. Furthermore, priors with $a \geq 1$ may sometimes occur as marginals of higher-dimensional priors, such as those for estimating covariance matrices (cf. Eaton and Sudderth, 1993).

(*ii*) For this particular model, it is possible to verify the directly (through integration) that the posterior distribution is improper. However, this might not be clear initially, so it is useful to see how it can be inferred from the associated Gibbs sampler. Furthermore, the analysis presented here may generalize to other situations in which the direct integration is too complicated.

(*iii*) The contrasting results of Section 3 and Section 4 indicate the importance of the prior distribution through the values of a and b, and in particular whether or not $b > 0$. Evidence of similar importance of related prior values for variance components models can be found in Gelfand et al. (1990, Figure 1).

5. Appendix: Proofs of computational lemmas.

The following lemma is easily verified, and we shall use it freely in the computations which follow.

Lemma 8. *Let $Z_1, \ldots, Z_n$ be independent random variables, and set $\overline{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$. Then*

$$\mathbf{E}\left(\sum_{i=1}^n (Z_i - \overline{Z})^2\right) \;=\; \left(\frac{n-1}{n}\right)\sum_{i=1}^n \mathbf{Var}(Z_i) \;+\; \sum_{i=1}^n \left((\mathbf{E}Z_i) - (\mathbf{E}\overline{Z})\right)^2 .$$

We now proceed to the proofs of the computational lemmas.

Proof of Lemma 2. We compute the conditional expected value in three stages, as we "peel away" those variables on which we are conditioning. We begin by noting that, using Lemma 8, we have

$$\mathbf{E}\left(\sum (\theta_i^{(k)} - \overline{\theta}^{(k)})^2 \mid A^{(k)}, \mu^{(k)}, x^{(k-1)}\right)$$

$$= \left[\left(\frac{K-1}{K}\right) K \left(\frac{A^{(k)}V}{V + A^{(k)}}\right) \;+\; \sum_{i=1}^K \left(\frac{\mu^{(k)}V + Y_i A^{(k)}}{V + A^{(k)}} \;-\; \frac{\mu^{(k)}V + \overline{Y} A^{(k)}}{V + A^{(k)}}\right)^2\right]$$

$$= \left[(K-1)\left(\frac{A^{(k)}V}{V+A^{(k)}}\right) \;+\; \left(\frac{A^{(k)}}{V+A^{(k)}}\right)^2 \Delta\right].$$

Also

$$\mathbf{E}\left(K(\overline{\theta}-\overline{Y})^2 \,|\, A^{(k)}, \mu^{(k)}, x^{(k-1)}\right)$$

$$= K\left[\mathbf{Var}(\overline{\theta} \,|\, A^{(k)}, \mu^{(k)}, x^{(k-1)}) \;+\; \left(\mathbf{E}(\overline{\theta}-\overline{Y} \,|\, A^{(k)}, \mu^{(k)}, x^{(k-1)})\right)^2\right]$$

$$= K\left[\frac{1}{K}\;\frac{A^{(k)}V}{V+A^{(k)}} \;+\; (\mu^{(k)}-\overline{Y})^2\left(\frac{V}{V+A^{(k)}}\right)^2\right].$$

We next take the expected value over $\mu^{(k)}$. We have that

$$\mathbf{E}\left((\mu^{(k)}-\overline{Y})^2 \,|\, A^{(k)}, x^{(k-1)}\right)$$

$$= \mathbf{Var}\left((\mu^{(k)}-\overline{Y}) \,|\, A^{(k)}, x^{(k-1)}\right) \;+\; \left(\mathbf{E}\left((\mu^{(k)}-\overline{Y}) \,|\, A^{(k)}, x^{(k-1)}\right)\right)^2$$

$$= \left(\overline{\theta}^{(k-1)}-\overline{Y}\right)^2 \;+\; A^{(k)}/K\,.$$

Now, recalling that $f(x) \;=\; K(\overline{\theta}-\overline{Y})^2 + \sum(\theta_i-\overline{\theta})^2$, and putting all of this together, we obtain

$$\mathbf{E}\left(f(x^{(k)}) \,|\, A^{(k)}, x^{(k-1)}\right) \;=\; K\left(\frac{A^{(k)}V}{V+A^{(k)}}\right) \;+\; \left(\frac{A^{(k)}}{V+A^{(k)}}\right)^2 \Delta$$

$$+\; \left(K(\overline{\theta}^{(k-1)}-\overline{Y})^2 + A^{(k)}\right)\left(\frac{V}{V+A^{(k)}}\right)^2.$$

Our final step will involve taking expectation with respect to $A^{(k)}$. Before doing so, we simplify the above (exact) formula using inequalities. By inspection we have $\frac{A^{(k)}}{V+A^{(k)}} \leq 1$ and $K(\overline{\theta}^{(k-1)}-\overline{Y})^2 \leq f(x^{(k-1)})$. Also by calculus we verify that $\frac{a}{(1+a)^2} \leq 1/4$ for any $a \geq 0$, so that $\frac{V^2 A^{(k)}}{\left(V+A^{(k)}\right)^2} \leq V/4$, for any value of $A^{(k)}$. (It may be possible to use more sophisticated bounds.) We thus obtain

$$\mathbf{E}\left(f(x^{(k)}) \,|\, A^{(k)}, x^{(k-1)}\right) \;\leq\; \left(1+\frac{A^{(k)}}{V}\right)^{-2} f(x^{(k-1)}) \;+\; \left(K+\frac{1}{4}\right)V \;+\; \Delta\,.$$

Note that this expression is strictly *decreasing* as a function of $A^{(k)}$.

To complete the calculation, we need to take expected value over $A^{(k)}$. Now, $\mathcal{L}\left(A^{(k)} \,|\, x^{(k-1)}\right) = IG\left(a+\frac{K-1}{2},\, b+\frac{1}{2}\sum(\theta_i^{(k-1)}-\overline{\theta}^{(k-1)})^2\right)$. Hence if $W \sim IG(a+\frac{K-1}{2},\, b)$, then $A^{(k)}$ is stochastically larger than W, and hence

$$\mathbf{E}\left(f(x^{(k)}) \,|\, x^{(k-1)}\right) \;\leq\; \mathbf{E}\left(\left(1+\frac{W}{V}\right)^{-2}\right) f(x^{(k-1)}) \;+\; \left(K+\frac{1}{4}\right)V \;+\; \Delta\,,$$

as desired. ∎

Proof of Lemma 3. Our proof is similar to Lemmas 3 and 4 of Rosenthal (1995b). Recall that for $x \in f_d$, we have in particular that $\sum(\theta_i - \overline{\theta})^2 \leq d$ and that $K(\overline{\theta} - \overline{Y})^2 \leq d$. We define the (non-normalized) measure $Q'(\cdot)$ on $\mathcal{X}$ inductively, in terms of Lebesgue measure, by

$$\begin{aligned} Q'(dA) &= \left(\inf_{0 \leq r \leq d} IG\left(a + \frac{K-1}{2}, b + \frac{r}{2}; A\right)\right) dA\,; \\ Q'(d\mu \,|\, A) &= \left(\inf_{K(s-\overline{Y})^2 \leq d} N\left(s, \frac{A}{K}; \mu\right)\right) d\mu\,; \\ Q'(d\theta_i \,|\, \mu, A) &= N\left(\frac{\mu V + Y_i A}{V + A}, \frac{AV}{V+A}\right), \end{aligned}$$

for $1 \leq i \leq K$, with the θ_i conditionally independent.

Intuitively, $Q'(\cdot)$ has been defined to mimic the transition probabilities $P(x, \cdot)$ (as defined by the conditional distributions in Section 2), but with appropriate infimums over values of $x \in f_d$. (Of course, once A and μ have been chosen, the distributions of the θ_i are completely determined, so no further infima are necessary.) This construction ensures that

$$P(x, \cdot) \geq Q'(\cdot), \qquad x \in f_d\,.$$

Hence we can take $Q(\cdot) = \frac{Q'(\cdot)}{Q'(\mathcal{X})}$ and $\epsilon = Q'(\mathcal{X})$, to get that

$$P(x, \cdot) \geq \epsilon\, Q(\cdot), \qquad x \in f_d\,,$$

with $Q(\cdot)$ a probability measure on $\mathcal{X}$.

On the other hand,

$$\begin{aligned} Q'(\cdot) &= \int_0^\infty Q'(dA) \int_{-\infty}^{\infty} Q'(d\mu \,|\, A) \prod_{i=1}^{K} \int_{-\infty}^{\infty} Q'(d\theta_i \,|\, \mu, A) \\ &= \int_0^\infty \left(\inf_{0 \leq r \leq d} IG\left(a + \frac{K-1}{2}, b + \frac{r}{2}; A\right)\right) dA \int_{-\infty}^{\infty} \left(\inf_{K(s-\overline{Y})^2 \leq d} N\left(s, \frac{A}{K}; \mu\right)\right) d\mu\,. \end{aligned}$$

Now, for fixed a and t, the function $IG(a, b; t)$ is unimodal as a function of b. It follows that

$$\inf_{0 \leq r \leq d} IG\left(a + \frac{K-1}{2}, b + \frac{r}{2}; A\right) = \min\left[IG\left(a + \frac{K-1}{2}, b; A\right),\ IG\left(a + \frac{K-1}{2}, b + \frac{d}{2}; A\right)\right].$$

Also, by replacing μ by $\mu + \overline{Y}$ in the inner integral, and then considering separately the cases $\mu < 0$ and $\mu > 0$, it is easily seen that

$$\int_{-\infty}^{\infty} \left(\inf_{K(s-\overline{Y})^2 \le d} N\left(s, \frac{A}{K}; \mu\right)\right) d\mu \;=\; 2\int_{0}^{\infty} N\left(-\sqrt{\frac{d}{K}}, \frac{A}{K}; \mu\right)\, d\mu\,.$$

The result follows. ∎

Proof of Lemma 5. We again compute the expected value in stages. We note first that, since the mean of $IG(\alpha, \beta)$ is $\frac{\beta}{\alpha-1}$ for $\alpha > 1$, we have

$$\mathbf{E}\left(A^{(k)} \,|\, x^{(k-1)}, x^{(k-2)}, \ldots\right) \;=\; \frac{b + \frac{1}{2}\sum(\theta_i^{(k-1)} - \overline{\theta}^{(k-1)})^2}{a + \frac{K-3}{2}}\,.$$

Hence, taking expectation over $\theta_i^{(k-1)}$ and using Lemma 8, we get that

$$\begin{aligned}
&\mathbf{E}\left(A^{(k)} \,|\, \mu^{(k-1)}, A^{(k-1)}, x^{(k-2)}, \ldots\right) \\
&= \left(\frac{1}{a + \frac{K-3}{2}}\right)\left(b \;+\; \frac{1}{2}\left[\left(\frac{K-1}{K}\right) K \left(\frac{A^{(k-1)}V}{V + A^{(k-1)}}\right)\right.\right. \\
&\left.\left. + \sum_{i=1}^{K}\left(\frac{\mu^{(k-1)}V + Y_i A^{(k-1)}}{V + A^{(k-1)}} \;-\; \frac{\mu^{(k-1)}V + \overline{Y}A^{(k-1)}}{V + A^{(k-1)}}\right)^2\right]\right) \\
&= \left(\frac{1}{a + \frac{K-3}{2}}\right)\left(b \;+\; \frac{1}{2}\left[(K-1)\left(\frac{A^{(k-1)}V}{V + A^{(k-1)}}\right) \;+\; \left(\frac{A^{(k-1)}}{V + A^{(k-1)}}\right)^2 \Delta\right]\right).
\end{aligned}$$

Since $\mu^{(k-1)}$ does not appear in this last expression, it follows that $\mathbf{E}\left(A^{(k)} \,|\, A^{(k-1)}, x^{(k-2)}, \ldots\right)$ equals this same expression.

Now, if $a \ge 1$, $b \le 0$, $\Delta \le V$, and $A^{(k-1)} = t > 0$, we see that

$$\begin{aligned}
\mathbf{E}\left(A^{(k)} \,|\, A^{(k-1)} = t, x^{(k-2)}, \ldots\right) \;&\le\; \left(\frac{tV}{V+t}\right) + \left(\frac{t}{V+t}\right)^2 V \\
= \frac{tV(V+t) + t^2 V}{(V+t)^2} \;&=\; \frac{t(V+t)^2 - t^3}{(V+t)^2} \;=\; t \;-\; \frac{t^3}{(V+t)^2}\,,
\end{aligned}$$

as desired. ∎

REFERENCES

Y. Amit and U. Grenander (1991), Comparing sweep strategies for stochastic relaxation. J. Multivariate Analysis **37**, No. **2**, 197-222.

J.R. Baxter and J.S. Rosenthal (1995), Rates of convergence for everywhere-positive Markov chains. Stat. Prob. Lett. **22**, 333-338.

P. Billingsley (1986). Probability and Measure, 2nd ed. Wiley & Sons, New York.

M.C. Eaton and W.D. Sudderth (1993), Prediction in a multivariate normal setting: coherence and incoherence. Tech. Rep., Dept. of Statistics, University of Minnesota.

B. Efron and C. Morris (1973), Stein's estimation rule and its competitors – An empirical Bayes approach. J. Amer. Stat. Assoc., Vol. **68**, No. **341**, 117- 130.

B. Efron and C. Morris (1975), Data analysis using Stein's estimator and its generalizations. J. Amer. Stat. Assoc., Vol. **70**, No. **350**, 311-319.

A. Frieze, R. Kannan, and N.G. Polson (1994), Sampling from log-concave distributions. Ann. Appl. Prob. **4**, 812-837.

A. Frigessi, C.R. Hwang, S.J. Sheu, and P. Di Stefano (1993), Convergence rates of the Gibbs sampler, the Metropolis algorithm, and other single-site updating dynamics. J. Roy. Stat. Soc. Ser. B **55**, 205-220.

A.E. Gelfand and A.F.M. Smith (1990), Sampling based approaches to calculating marginal densities. J. Amer. Stat. Assoc. **85**, 398-409.

A.E. Gelfand, S.E. Hills, A. Racine-Poon, and A.F.M. Smith (1990), Illustration of Bayesian inference in normal data models using Gibbs sampling. J. Amer. Stat. Soc. **85**, 972-985.

S. Geman and D. Geman (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on pattern analysis and machine intelligence **6**, 721-741.

W.K. Hastings (1970), Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97-109.

W. James and C. Stein (1961), Estimation with Quadratic Loss. Proceedings of the Fourth

Berkeley Symposium on Mathematical Statistics and Probability, Vol. **1**, University of California Press, Berkeley, 361-379.

M. Jerrum and A. Sinclair (1989), Approximating the permanent. SIAM J. Comput. **18**, 1149-1178.

J. Liu, W. Wong, and A. Kong (1991a), Correlation structure and the convergence of the Gibbs sampler, I. Tech Rep. **299**, Dept. of Statistics, University of Chicago. Biometrika, to appear.

J. Liu, W. Wong, and A. Kong (1991b), Correlation structure and the convergence of the Gibbs sampler, II: Applications to various scans. Tech Rep. **304**, Dept. of Statistics, University of Chicago. J. Royal Stat. Sci. **(B)**, to appear.

R.B. Lund and R.L. Tweedie (1993), Geometric convergence rates for stochastically ordered Markov chains. Tech. Rep., Dept. of Statistics, Colorado State University.

K.L. Mengersen and R.L. Tweedie (1993), Rates of convergence of the Hastings and Metropolis algorithms. Tech. Rep. **93/30**, Dept. of Statistics, Colorado State University.

N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953), Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087-1091.

S.P. Meyn and R.L. Tweedie (1994), Computable bounds for convergence rates of Markov chains. Ann. Appl. Prob. **4**, 981-1011.

C. Morris (1983), Parametric empirical Bayes confidence intervals. Scientific Inference, Data Analysis, and Robustness, 25-50.

P. Mykland, L. Tierney, and B. Yu (1992), Regeneration in Markov chain samplers. Tech. Rep. **585**, School of Statistics, University of Minnesota.

G.O. Roberts (1992), Convergence diagnostics of the Gibbs sampler. In Bayesian Statistics **4** (J.M. Bernardo et al., eds.), 777-784. Oxford University Press.

J.S. Rosenthal (1993), Rates of convergence for Data Augmentation on finite sample spaces. Ann. Appl. Prob., Vol. **3**, No. **3**, 319-339.

J.S. Rosenthal (1995a), Rates of convergence for Gibbs sampler for variance components models. Ann. Stat., to appear.

J.S. Rosenthal (1995b), Minorization conditions and convergence rates for Markov chain Monte Carlo. J. Amer. Stat. Assoc., to appear.

M.J. Schervish and B.P. Carlin (1992), On the convergence of successive substitution sampling, J. Comp. Graph. Stat. **1**, 111–127.

M.A. Tanner and W.H. Wong (1987), The calculation of posterior distributions by data augmentation (with discussion). J. Amer. Stat. Assoc. **82**, 528-550.

L. Tierney (1991), Markov chains for exploring posterior distributions. Tech. Rep. **560**, School of Statistics, University of Minnesota. Ann. Stat., to appear.