Rates of Convergence for Data Augmentation on Finite Sample Spaces

by

Jeffrey S. Rosenthal

School of Mathematics, University of Minnesota, Minneapolis, MN 55455, USA

(Appeared in *Annals of Applied Probability* **3** (1993), 819–839.)

Key phrases. Data Augmentation, Gibbs Sampler, Harris Recurrence, Convergence Rate.

AMS 1980 Subject Classifications. 60J10, 62F15.

Running Title. Convergence of Data Augmentation.

1. Introduction.

Tanner and Wong [TW] have defined an iterative process for obtaining closer and closer approximations to the (Bayes) posterior distribution of certain parameters given certain data. Their approach, which they call Data Augmentation, is closely related to the Gibbs Sampler algorithm as developed by Geman and Geman [GG]. It is used in the following situation. Suppose we observe data $\vec{Y}$, and wish to compute the posterior of a parameter $\vec{\theta}$ give $\vec{Y}$. Suppose further that there is some some other data $\vec{X}$ which is not observed, but such that the posterior of $\vec{\theta}$ given both $\vec{X}$ and $\vec{Y}$ is fairly simple. Furthermore, suppose the conditional distribution of $\vec{X}$ given $\vec{Y}$ and $\vec{\theta}$ is also simple. Under these conditions, the Data Augmentation algorithm provides a straightforward way to obtain better and better approximations of the true posterior of $\vec{\theta}$ given $\vec{Y}$. The idea is to *augment* the data $\vec{Y}$ with "simulated" values of the unknown $\vec{X}$.

Tanner and Wong study convergence properties of the Data Augmentation algorithm. Specifically, they show that under mild conditions, the iterative process will converge in total variation distance to the true posterior. However, they do not obtain a useful estimate for the rate of convergence.

In this paper, we examine this rate of convergence more carefully. We restrict our attention to the case where $\vec{X}$ and $\vec{Y}$ take values in a *finite* set. Thus, we imagine coin-tossing or the rolling of a finite die. Our set-up is that $\vec{X} = (X_1, \ldots, X_n)$ are n independent, unobserved results of a coin-toss or finite die. While we do not observe $\vec{X}$, we do observe $\vec{Y} = (Y_1, \ldots, Y_n)$. Here Y_i depends only on X_i, in a known way. (For example, if the X_i represent whether or not the i'th subject has a certain disease, the Y_i might be the observed results of an imperfect medical test.) We wish to compute the posterior for the distribution of the X_i, given only the "imperfect" data $\vec{Y}$. The idea is to augment the Y_i by "fake" values of X_i at each step, and then update our estimate for the posterior using these fake values of X_i. This provides an iterative procedure for obtaining successively better approximations to the desired posterior.

The main result of this paper states that under certain assumptions, such a process on a finite sample space will converge to the true posterior after $O(\log n)$ steps. Thus, the number of steps required to approach the true posterior does not grow too quickly with the amount of observed data. This suggests the feasibility of running this iterative process when given a large but finite amount of data. In [R], similar results are obtained for a more complicated model, namely the variance component models as discussed in [GS].

The plan of this paper is as follows. In Section 2 we review the definition of the Data Augmentation algorithm, and state the key lemma to be used in proving convergence results. In Section 3 we prove the convergence result for the case of coin-tossing (i.e. when X_i and Y_i only take values 0 and 1). In Section 4 we examine the general finite case (i.e. when X_i and Y_i take on an arbitrary finite number of values). Section 4 includes an analysis of a Dynamical System on the K-simplex that arises in the study of Data Augmentation in this case. We prove the convergence of this dynamical system under certain conditions, but the general question remains open.

2. Preliminaries.

To define the Data Augmentation algorithm as we shall study it, let $X_1, X_2, \ldots, X_n$ be iid random variables taking values in a set $\mathcal{X}$, with unknown distribution G. For $1 \leq i \leq n$, let Y_i be a random variable, also taking values in the set $\mathcal{X}$, which is a (known) random function of the corresponding X_i. Specifically, we assume there is a family of distributions H_x such that

$$\mathcal{L}(Y_i \mid X_1, \ldots, X_n, Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n) \;=\; \mathcal{L}(Y_i \mid X_i) \;=\; H_{X_i}\;.$$

We suppose that we observe $Y_1, Y_2, \ldots, Y_n$ but not $X_1, X_2, \ldots, X_n$, and we are interested in the posterior distribution μ of G, conditional on the observed values of the Y_i, and relative to some prior distribution ν. (Here μ and ν are probability distributions on the set $M_1(\mathcal{X})$ of all probability distributions on $\mathcal{X}$: $\mu, \nu \in M_1(M_1(\mathcal{X}))$.)

The Markov chain is defined as follows. Given a probability distribution $\theta_k \in M_1(\mathcal{X})$, choose $\theta_{k+1} \in M_1(\mathcal{X})$ by (a) choosing $x_1^{(k)}, \ldots, x_n^{(k)}$ independently according to θ_k conditional on the observations $Y_1, \ldots, Y_n$:

$$x_i^{(k)} \sim \theta_k(\,\cdot\,|Y_1, \ldots, Y_n)$$

and then (b) choosing θ_{k+1} from the posterior distribution of G conditional on the newly produced values $x_1^{(k)}, \ldots, x_n^{(k)}$ (and relative to the same prior distribution ν).

Formally, the transition probabilities are given by

$$K(\theta_k, d\theta_{k+1}) = \int_{\mathcal{X}^n} \theta_k(d\vec{x}|\vec{y})\ B_\nu(d\theta_{k+1}|\vec{x})$$

where $\vec{x}$ stands for the possible values $(x_1^{(k)}, \ldots, x_n^{(k)}) \in \mathcal{X}^n$, $\vec{y}$ stands for the observed data $(Y_1, \ldots, Y_n)$, and $B_\nu(\ \cdot\ |\vec{x})$ means the posterior distribution on $M_1(M_1(\mathcal{X}))$ conditional on the observations $\vec{x}$ and relative to the prior ν.

The following Proposition is from [TW]. (See also [GS] for a survey of the relevant literature.) We include a proof for completeness.

Proposition 1. *The above description defines a time-homogeneous Markov chain on $M_1(\mathcal{X})$ with stationary distribution given by μ, the posterior distribution of G conditional on the observed data $\vec{y} = (Y_1, \ldots, Y_n)$.*

Proof. The time-homogeneity is immediate since the prior distribution ν does not vary with time. The statement about stationarity follows from the computation

$$\begin{aligned} P(\theta_{k+1} \in S \mid \theta_k \sim \mu) &= \int_{M_1(\mathcal{X})} \mu(d\theta_k) \int_{\mathcal{X}^n} \theta_k(d\vec{x}|\vec{y})\ B_\nu(S|\vec{x}) \\ &= \int_{\mathcal{X}^n} \left(\int_{M_1(\mathcal{X})} \mu(d\theta_k)\ \theta_k(d\vec{x}|\vec{y}) \right) B_\nu(S|\vec{x}) \\ &= \int_{\mathcal{X}^n} P_\nu(d\vec{x}|\vec{y})\ B_\nu(S|\vec{x}) \\ &= \mu(S)\ . \end{aligned}$$

∎

Proposition 1 provides the motivation for the Data Augmentation algorithm. The algorithm provides a recipe for a Markov chain whose stationary distribution is the desired posterior distribution. Of particular interest is the question of convergence of the Markov chain to its stationary distribution. The main result of this paper is the following.

Let $\mathcal{X}$ be finite. Then, under certain assumptions about the data $\{Y_i\}$ and about the dependence of Y_i on X_i, and with a uniform prior ν, the Data Augmentation algorithm *converges in total variation distance after $O(\log n)$ steps, where n is the number of observed data.*

Remarks.

1. The convergence results in this paper are all stated using the $O(\cdot)$ notation, or in terms of unspecified constants. However, the proofs allow for a specific determination of the constants involved. Thus, a result such as the above actually states that given $\epsilon > 0$, there is a computable constant A_ϵ such that for any n, after $A_\epsilon \log n$ steps the total variation distance is less than ϵ.
2. We consider only uniform priors throughout most of this paper. However, the results go through quite generally; this is explored to some extent in the paper's final remark.
3. The posteriors considered in this paper are all for finite sample spaces, and can all be computed by other, non-iterative methods. Thus, the main thrust of the current paper is not that Data Augmentation should be used in these cases, but rather that the convergence results obtained here may provide some insight into using Data Augmentation and Gibbs Sampler in more complicated examples, such as those considered in [GS] and [GHRS]. We intend to consider some of those examples elsewhere [R]. Also, the methods used here may be applicable to many other Markov chain problems.

The main tool used in proving the above result will be the following "Upper Bound Lemma", inspired by the discussion on page 151 of [A]. It is closely related to the notions of Doeblin and Harris-recurrence (see [A], [AN], [AMN], [N], [Do]). In fact, a very similar (but less quantitative) result appears as Theorem 6.15 in [N]. But since this Lemma will be crucial to what follows, we include a complete proof.

Lemma 2. *Let $P(x,\cdot)$ be the transition probabilities for a time-homogeneous Markov chain on a general state space $\mathcal{X}$. Suppose that for some probability distribution $Q(\,\cdot\,)$ on $\mathcal{X}$, some positive integer k_0, and some $\epsilon > 0$,*

$$P^{k_0}(x,\cdot) \geq \epsilon\, Q(\,\cdot\,) \qquad \text{for all } x \in \mathcal{X}\ ,$$

where P^{k_0} represents the k_0-step transition probabilities. Then for any initial distribution

π_0, the distribution π_k of the Markov chain after k steps satisfies

$$\|\pi_k - \pi\| \leq (1-\epsilon)^{\lfloor k/k_0 \rfloor}$$

where $\| \cdot \|$ is total variation distance, π is any stationary distribution, and $\lfloor r \rfloor$ is the greatest integer not exceeding r. (In particular, the stationary distribution is unique.)

Proof. The proof shall be by a coupling argument. (For background on coupling, see, e.g., [P] or chapter 4E of [Di].) We let $\{X_k\}$ be the Markov chain beginning in the distribution π_0, and let $\{Y_k\}$ be the Markov chain beginning in the distribution π. We realize each Markov chain as follows. At time $k=0$, we choose the positions at time $k=k_0$ by (a) with probability ϵ letting them both go to a point $p \in \mathcal{X}$ chosen according to $Q(\cdot)$, and (b) with probability $1-\epsilon$ letting them move independently according to the distributions $\frac{1}{1-\epsilon}(P(X_0, \cdot) - \epsilon Q(\cdot))$ and $\frac{1}{1-\epsilon}(P(Y_0, \cdot) - \epsilon Q(\cdot))$, respectively. We then fill in the values $X_1, X_2, \ldots, X_{k_0-1}$ [resp. $Y_1, Y_2, \ldots, Y_{k_0-1}$] conditionally on X_0 and X_{k_0} [resp. Y_0 and Y_{k_0}]. Having done so, we similarly choose the values of $X_{k_0+1}, \ldots, X_{2k_0}$ and $Y_{k_0+1}, \ldots, Y_{2k_0}$. Continuing in this manner, we choose $\{X_k\}$ and $\{Y_k\}$ for all k.

It is easily checked that the above recipe realizes $\{X_k\}$ and $\{Y_k\}$ according to the transition probabilities $P(\cdot, \cdot)$. The coupling time T is the first time we choose option (a) above. This happens with probability ϵ every k_0 steps. Thus,

$$Prob(T > k) \leq (1-\epsilon)^{\lfloor k/k_0 \rfloor} \ .$$

The result now follows from the coupling inequality

$$\|\mathcal{L}(X_k) - \mathcal{L}(Y_k)\| = \|\pi_k - \pi\| \leq Prob(T > k) \ .$$

∎

Remarks.

1. It is in fact not necessary that the Markov chain be time-homogenous. The proof above works with very minor changes for a general Markov chain, provided we have $P^{t,t+k_0}(x, \cdot) > \epsilon Q(\cdot)$ for all $x \in \mathcal{X}$ *and* for all times t, and provided that $\pi(\cdot)$ is stationary for *each* $P^{t,t+1}(\cdot, \cdot)$.

2. Lemma 2 is similar in appearance to the Strong Stopping Times of Aldous and Diaconis (see [Di], Chapter 4A). However, in Lemma 2 the measure $Q(\cdot)$ is arbitrary, while in the case of Strong Stopping Times $Q(\cdot)$ is required to be a stationary distribution for the chain. This difference is significant since in many cases the stationary distribution is unknown or difficult to work with.
3. A generalization of Lemma 2, more suitable for unbounded spaces $\mathcal{X}$, is presented in [R].

3. The case of a two-element state space.

In this section we let $\mathcal{X} = \{0, 1\}$ have two elements only. (The extension to the case $\mathcal{X} = \{1, 2, \ldots, K\}$ is treated in the next Section.) Thus the random walk takes place on $M_1(\mathcal{X}) = [0, 1]$, the unit interval, with $\theta \in [0, 1]$ identified with the distribution on $\mathcal{X}$ giving mass θ to 1, and mass 1-θ to 0. The random variables $X_1, \ldots, X_n$ take values in $\{0, 1\}$ according to some unknown distribution $G \in [0, 1]$. For each i, Y_i is a random function of the value of X_i, and the observed data $Y_1, \ldots, Y_n$ are all in $\{0, 1\}$. We let the prior distribution ν be Lebesgue measure on $[0, 1]$, and we are interested in the posterior distribution μ of G given the observations $\{Y_i\}$.

We let p_{ab} $(a, b \in \{0, 1\})$ be the probability that $Y_i = b$ given that $X_i = a$. We set $p_{10} = s, p_{01} = t, p_{11} = 1 - s$, and $p_{00} = 1 - t$. We further let γ be the proportion of the data $\{Y_i\}$ which are 1:

$$\gamma = (\text{number of } i \text{ for which } Y_i = 1)/n \ .$$

(As an example, the Y_i might be the results of a medical test for a certain disease in n subjects. The X_i would indicate whether the i'th subject actually had the disease. In this case, γ would be the proportion of positive test results, while s and t would be the probabilities of false negatives and false positives, respectively.)

In this setting, the Markov chain $\{\theta_k\}$ (where $\theta_k \in [0, 1]$) may be described as follows. Set

$$\eta(\theta) = P(Y_r = 1) = (1 - s)\theta + t(1 - \theta) \ ,$$

and let

$$q_1(\theta) = P(X_r = 1 \mid Y_r = 1) = \frac{(1-s)\theta}{\eta(\theta)}$$

$$q_0(\theta) = P(X_r = 1 \mid Y_r = 0) = \frac{s\,\theta}{1-\eta(\theta)}\ .$$

Given θ_k, we choose $x_1^{(k)}, \ldots, x_n^{(k)} \in \{0,1\}$ where the probability that $x_r^{(k)} = 1$ is given by

$$P(x_r^{(k)} = 1) = \begin{cases} q_1(\theta), & Y_r = 1 \\ q_0(\theta), & Y_r = 0 \end{cases}$$

Then choose θ_{k+1} from the beta distribution $\beta(S_k + 1, n - S_k + 1)$ where $S_k = \sum\limits_{r=1}^{n} x_r^{(k)}$ is the number of $x_r^{(k)}$ which equal 1.

With this notation, our assumptions can be stated. We assume that s, t, and γ remain fixed as n increases. (The observant reader will object that γ must always be an integer multiple of $1/n$, and therefore cannot remain fixed for all n. However, this difficulty can be avoided by allowing γ to vary by an amount which is less than $1/n$. Such small changes will not affect the arguments which follow, and shall not be considered further.) We further assume that

$$0 < s < \frac{1}{2}; \qquad 0 < t < \frac{1}{2};$$

these assumptions merely state that X_r and Y_r are positively correlated.

Under the above assumptions, we shall prove

Theorem 3. *For the Data Augmentation process corresponding to $\mathcal{X} = \{0,1\}$, there exist positive numbers Λ and α (depending on s, t, and γ, but not depending on n) such that for any initial distribution π_0, the distribution π_k of the Markov chain after k steps satisfies*

$$\|\pi_k - \mu\| \le (1-\alpha)^{\lfloor k/\Lambda \log n \rfloor}$$

where $\|\ \cdot\ \|$ is total variation distance, μ is the posterior distribution given the observed data $Y_1, \ldots, Y_n$, and $\lfloor x \rfloor$ is the greatest integer not exceeding x.

Theorem 3 says that after $O(\log n)$ steps, the Markov chain is close in total variation distance to its stationary distribution μ.

Remark. In the case $\mathcal{X} = \{0, 1\}$, it is easy to see directly that the posterior distribution μ is absolutely continuous with respect to Lebesgue measure, with density proportional to $\eta(\theta)^{n\gamma}(1 - \eta(\theta))^{n(1-\gamma)}$, where $\eta(\theta) = (1 - s)\theta + t(1 - \theta)$ is the probability that $Y_r = 1$. Thus, μ has a peak (of width $O(1/\sqrt{n})$) near $\eta(\theta) = \gamma$, i.e. near $\theta = \frac{\gamma - t}{1-s-t}$. The quantity $\frac{\gamma-t}{1-s-t}$ will re- appear as the quantity F below.

To prove Theorem 3, we shall make use of Lemma 2. We must first examine the Markov chain in question more carefully. In particular, let us consider the distribution of θ_{k+1} given θ_k. Recall that given θ_k, we compute θ_{k+1} by flipping $n\gamma$ "$q_1(\theta_k)$-coins", and $n(1 - \gamma)$ "$q_0(\theta_k)$-coins", and then choosing θ_{k+1} from $\beta(S_k + 1, n - S_k + 1)$ where S_k is the number of "heads" we obtained in the n coin flips. Now, the distribution of S_k will be peaked within $O(1/\sqrt{n})$ of $n\gamma q_1(\theta_k) + n(1 - \gamma)q_0(\theta_k)$ with width of order $1/\sqrt{n}$. Then the distribution of θ_{k+1} will be peaked around $\frac{S_k+1}{n+2}$ with width again of order $1/\sqrt{n}$. We conclude that $L(\theta_{k+1} \mid \theta_k)$ will be peaked around $e(\theta_k)$ with width $O(1/\sqrt{n})$, where

$$\begin{aligned} e(\theta) &= \gamma q_1(\theta) + (1 - \gamma)q_0(\theta) \\ &= \gamma \frac{(1 - s)\theta}{(1 - s)\theta + t(1 - \theta)} + (1 - \gamma) \frac{s\theta}{1 - (1 - s)\theta - t(1 - \theta)} \ . \end{aligned}$$

This last observation gives us a picture of how things "ought to proceed". Aside from a small amount of "spreading", the values $\{\theta_k\}$ will follow the deterministic prescription

$$\theta_{k+1} = e(\theta_k) \ .$$

This suggest studying the "dynamical system" given by $\theta_{k+1} = e(\theta_k)$, and using this to infer information about our original Markov chain. We emphasize that the dynamical system is merely a useful approximation, and that its properties do not coincide with those of the Markov chain. On the other hand, we note that $e(\theta)$ does not depend on n, which simplifies the analysis.

The equation $\theta_{k+1} = e(\theta_k)$ is easily seen to have three fixed points $\theta_{k+1} = \theta_k$: when θ_k is 0, 1, or

$$F = F(s, t, \gamma) = \frac{\gamma - t}{1 - s - t} \ .$$

We shall assume for convenience (see Remark 2 at the end of this section) that $t < \gamma < 1-s$, i.e. that the proportion of 1's observed is not "exceptionally high" or "exceptionally low".

This assumption ensures that $0 < F < 1$, and that the fixed points 0 and 1 are unstable: if θ_k is "near" to 0 (say), then θ_{k+1} will tend to be a bit further away. The fixed point F, on the other hand, is stable: $\{\theta_k\}$ will tend to get closer and closer to F at an exponential rate.

We now return our attention to the Markov chain itself. The above analysis suggests that after $O(\log n)$ steps, θ_k for the Markov chain ought to be within, say, $1/\sqrt{n}$ of F. Then, since the binomial distributions above tend to "spread" things by $O(1/\sqrt{n})$, we expect that after one more step, θ_k will have a reasonable chance of going to any point within (say) $1/\sqrt{n}$ of F. Hence, if in Lemma 2 we make $Q(\,\cdot\,)$ roughly uniform on $[F-(1/\sqrt{n}), F+(1/\sqrt{n})]$, and set $k_0 = \Lambda \log n$ for some Λ, we should be able to choose ϵ independent of n, proving Theorem 3.

To make the above argument more precise, we need the following lemma. It says that after one step the Markov chain is at least a little bit away from 0 and 1, that $A \log n$ steps after that the Markov chain is far away from 0 and 1, and that $B_1 \log n + B_2$ steps after that the Markov chain is within about $1/\sqrt{n}$ of F, all with probabilities bounded below independently of n.

Lemma 4. *Let F be as above, and assume that $t < \gamma < 1-s$. Then there are constants $A, B_1, B_2, M_1, M_2, m_1, m_2$ and m_3, depending on s, t, and γ but all independent of n, such that for all sufficiently large n, if*

$$R_1 = [M_1/n, 1-(M_1/n)]\ ,$$

$$R_2 = [F/4, (F+3)/4]\ ,$$

$$\text{and} \qquad R_3 = [F-(M_2/\sqrt{n}), F+(M_2/\sqrt{n})]\ ,$$

then

(1) $Prob(\theta_1 \in R_1) \geq m_1 > 0$;

(2) $Prob(\theta_{T+A\log n} \in R_2 \mid \theta_T \in R_1) \geq m_2 > 0$;

(3) $Prob(\theta_{T+B_1\log n+B_2} \in R_3 \mid \theta_T \in R_2) \geq m_3 > 0$;

Proof. We let $f(\theta) = e(\theta) - \theta$. It is easily seen that $f(0) = f(F) = f(1) = 0$, that $f(\theta) > 0$ for $0 < \theta < F$, and that $f(\theta) < 0$ for $F < \theta < 1$. (For example, as $\theta \to 0$,

$\frac{e(\theta)}{\theta} \to \gamma \frac{1-s}{t} + (1-\gamma)\frac{s}{1-t}$, and this last expression is easily seen to be greater than 1 since $\gamma > t$.) Furthermore, f has non-zero derivative at each of 0, 1, and F. Thus we can define

$$C_1 = \min\left(\inf_{\theta<F/4} \frac{f(\theta)}{\theta}, \inf_{\theta<F/4} \frac{-f(\theta)}{1-\theta}\right) > 0 \ ;$$

$$C_2 = 1 + \frac{C_1}{2} > 1 \ ;$$

$$C_3 = \max\left(\sup_{0<\theta<1} \frac{q_1(\theta)}{\theta}, \sup_{0<\theta<1} \frac{1-q_0(\theta)}{1-\theta}, 1+\frac{3C_1}{4}\right) = \max\left(\frac{1-s}{t}, \frac{1-t}{s}, 1+\frac{3C_1}{4}\right) \ ;$$

$$M_1 = \frac{96C_3}{(C_1)^2(1-\frac{1}{C_2})} \ ;$$

We state these definitions here to emphasize their independence of n. With these definitions, we proceed to the proofs.

For (1), we note that $Prob(\theta_1 \in R_1)$ is smallest when $\theta_0 = 0$ (or equivalently when $\theta_0 = 1$). If $\theta_0 = 0$, then θ_1 is chosen from $\beta(n+1, 1)$, so

$$\begin{aligned} Prob(\theta_1 \in R_1 \mid \theta_0 = 0) &= (n+1) \int_{M_1/n}^{1-(M_1/n)} \theta^n d\theta \\ &= (1-(M_1/n))^{n+1} - (M_1/n)^{n+1} \\ &\geq e^{-2M_1} \text{ (say)}, \end{aligned}$$

for n sufficiently large, proving (1) with $m_1 = e^{-2M_1} > 0$.

For (2), we set $T = 0$ for simplicity, and we set $A_1 = \frac{\log(F/4)}{\log C_2}$, $A_2 = \frac{\log((F+3)/4)}{\log C_2}$. Then $Prob(\theta_{A_1} < F/4)$ is largest (for $\theta_0 \in R_1$) when $\theta_0 = M_1/n$. Now,

$$\begin{aligned} Prob(\theta_{A_1} < F/4 \ | \theta_0 = M_1/n) &\leq \sum_{k=1}^{A_1} Prob(\theta_{k+1} < (M_1/n)(C_2)^{k+1} \ | \theta_k \geq (M_1/n)(C_2)^k) \\ &\leq \sum_{k=1}^{A_1} Prob(\theta_{k+1} < (M_1/n)(C_2)^{k+1} \ | \theta_k = (M_1/n)(C_2)^k) \ . \end{aligned}$$

Also

$$Prob\left(\theta_{k+1} < (M_1/n)(C_2)^{k+1} \mid \theta_k = (M_1/n)(C_2)^k\right) \ \leq \ Prob_1 + Prob_2$$

where $Prob_1$ is the probability that starting from $\theta_k = (M_1/n)(C_2)^k$, the "binomial part" of the Markov chain mechanism gets us a proportion S_k/n of 1's less than $(M_1/n)(C_2)^k(1+$

$\frac{3C_1}{4}$), and where $Prob_2$ is the probability that starting from $S_k/n = (M_1/n)(C_2)^k(1+\frac{3C_1}{4})$, the "beta part" of the Markov chain mechanism results in a value of θ_{k+1} which is less than $(M_1/n)(C_2)^{k+1}$.

Now, starting from $\theta_k = (M_1/n)(C_2)^k$, S_k/n is a random variable with mean $\geq (M_1/n)C_2^k(1+C_1)$ and variance equal to

$$\begin{aligned}
&(\gamma q_1(\theta_k)(1-q_1(\theta_k)) + (1-\gamma)q_0(\theta_k)(1-q_0(\theta_k)))/n \\
&\qquad \leq \max(q_0(\theta_k), q_1(\theta_k))/n \\
&\qquad = q_1(\theta_k)/n \\
&\qquad \leq C_3\theta_k/n \\
&\qquad = (M_1/n^2)C_3C_2^k \ .
\end{aligned}$$

Thus, by Chebychev's inequality,

$$Prob_1 \leq \frac{(M_1/n^2)C_3C_2^k}{\big((M_1/n)C_2^k(C_1/4)\big)^2} = \frac{16C_3}{M_1C_1^2C_2^k} \leq \frac{1}{6}(1-\frac{1}{C_2})(C_2)^{-k} \ .$$

Similarly, starting from $S_k/n = (M_1/n)(C_2)^k(1+\frac{3C_1}{4})$, the result of the "beta part" is a random variable $\beta(S_k+1, n-S_k+1)$ with mean $(M_1/n)C_2^k(1+\frac{3C_1}{4})$ and variance

$$\frac{(S_k+1)(n-S_k+1)}{(n+2)^2(n+3)} \leq (S_k/n^2) \leq (M_1/n^2)(C_2)^k(1+\frac{3C_1}{4}) \leq (M_1/n^2)C_3(C_2)^k \ ,$$

so that also

$$Prob_2 \leq \frac{1}{6}(1-\frac{1}{C_2})(C_2)^{-k} \ .$$

Thus

$$Prob(\theta_{k+1} < (M_1/n)(C_2)^{k+1} \mid \theta_k = (M_1/n)(C_2)^k) \leq \frac{1}{3}(1-\frac{1}{C_2})(C_2)^{-k} \ .$$

Hence,

$$Prob(\theta_{A_1} < F/4 \mid \theta_0 = M_1/n) \leq \sum_{k=0}^{A_1} \frac{1}{3}(1-\frac{1}{C_2})(C_2)^{-k} < 1/3 \ .$$

Similarly, $Prob(\theta_{A_2} > (F+3)/4)$ is largest when $\theta_0 = 1-(M_1/n)$. A computation very similar to the above then shows that

$$Prob\big(\theta_{A_2} > (F+3)/4 \mid \theta_0 = 1-(M_1/n)\big) < 1/3 \ .$$

Now, it is easily checked that once θ_k is in R_2, its chances of leaving R_2 on any one step are $O(e^{-n})$. Hence, if we set $A = \max(A_1, A_2)$, then

$$Prob(\theta_A \notin R_2) \le 1/3 + 1/3 + O(A\, e^{-n}) = 2/3 + O(e^{-n} \log n) \le 3/4 \text{ (say)},$$

for n sufficiently large, so that

$$Prob(\theta_A \in R_2) \ge 1/4 \ ,$$

proving (2).

The computation for (3) is similar but easier. We again set $T = 0$, and we set

$$C_5 = \min\left(\inf_{F/4<\theta<F} \frac{f(\theta)}{F-\theta}, \inf_{F<\theta<(F+3)/4} \frac{-f(\theta)}{\theta - F} \right) > 0 \ ;$$

$$C_6 = 1 + \frac{C_5}{2} > 1 \ ;$$

$$C_7 = \max\left(\frac{3F}{4}, \frac{3(1-F)}{4} \right) \ ;$$

$$B_1 = \frac{1}{2\log C_6} \ ; \ B_2 = \frac{\log\left(\frac{1}{8} C_5^2 C_7^2 (1 - \frac{1}{(C_6)^2})\right)}{2\log C_6} \ .$$

We wish to compute the probability that $|F - \theta_k| \le C_7 (C_6)^{-k}$ for $0 \le k \le B$, where $B = B_1 \log n + B_2$. For $k = 0$ it follows from the assumption that $\theta_0 \in R_2$. As above,

$$\begin{aligned} &Prob\big(|F - \theta_{k+1}| \le C_7 (C_6)^{-k-1} \bigm| |F - \theta_k| \le C_7(C_6)^{-k}\big) \\ &\quad \le Prob\big(|F - \theta_{k+1}| \le C_7 (C_6)^{-k-1} \bigm| |F - \theta_k| = C_7(C_6)^{-k}\big) \\ &\quad \le Prob_1 + Prob_2 \ , \end{aligned}$$

where $Prob_1$ and $Prob_2$ are the probabilities that the "binomial part" and the "beta part", respectively, are more than $C_7(C_6)^{-k-1}(C_5/4)$ away from their means. Now, it is easily checked that the variances of the "binomial part" and the "beta part" are each bounded by $\frac{1}{4n}$. Thus by Chebychev's inequality

$$Prob_1, \ Prob_2 \ \le \ \frac{\frac{1}{4n}}{\left(C_7(C_6)^{-k-1}(C_5/4)\right)^2} \ = \ \frac{4}{C_5^2 C_7^2 n} C_6^{2k} \ .$$

Hence,

$$
\begin{aligned}
&Prob\big(|F-\theta_B| \le C_7 (C_6)^{-B} \;\big|\; \theta_0 \in R_2\big) \\
&\quad \ge 1 \;-\; \sum_{k=0}^{B} Prob\big(|F-\theta_{k+1}| \le C_7 (C_6)^{-k-1} \;\big|\; |F-\theta_k| \le C_7 (C_6)^{-k}\big) \\
&\quad \ge 1 \;-\; \sum_{k=0}^{B} (Prob_1 + Prob_2) \\
&\quad \ge 1 \;-\; \sum_{k=0}^{B} 2\,\frac{4}{C_5^2 C_7^2 n} C_6^{2k} \\
&\quad \ge 1 \;-\; (C_6)^{2B} \frac{8}{C_5^2 C_7^2 n} \sum_{k=0}^{\infty} (C_6)^{-2k} \\
&\quad = 1 \;-\; \frac{8 (C_6)^{2B}}{C_5^2 C_7^2 n (1 - \frac{1}{(C_6)^2})} \\
&\quad = \frac{1}{2} \qquad \text{(by construction of } B) \; .
\end{aligned}
$$

Thus with probability $\ge \frac{1}{2}$,

$$
|F - \theta_{B_1 \log n + B_2}| \le C_7 (C_6)^{-B_1 \log n - B_2} = C_7 \sqrt{\frac{1}{8} C_5^2 C_7^2 (1 - \frac{1}{C_6^2})} \sqrt{n} \; .
$$

This completes the proof of (3), with $m_3 = \frac{1}{2}$ and $M_2 = C_7 \sqrt{\frac{1}{8} C_5^2 C_7^2 (1 - \frac{1}{C_6^2})}$. ∎

Lemma 4 shows that after $(A + B_1) \log n + B_2 + 1$ steps, the Markov chain will be in R_3 (so that $|\theta_k - F| \le M_2/\sqrt{n}$) with probability at least $m_1 m_2 m_3 > 0$.

Let us now consider S_{k+1}, the result of the "binomial part" of the Markov chain on the next step. Given $\theta_k \in R_3$, we note that S_{k+1} will be binomially distributed, with S_{k+1}/n having mean also inside R_3 (because F is attractive), and having variance within $O(1/n\sqrt{n})$ of C_8/n (where $C_8 = \gamma q_1(F)(1 - q_1(F)) + (1 - \gamma) q_0(F)(1 - q_0(F))$). It follows from the Central Limit Theorem that for sufficiently large n, if i is an integer within $O(\sqrt{n})$ of Fn, then the probability that $S_{k+1} = i$ will be at least $m_4 = e^{-2(M_2+1)^2/C_8}$ (say). In other words, S_{k+1}/n will have $O(1/\sqrt{n})$ spread around the set R_3 and therefore about the point F.

Once S_k is chosen, recall that $\mathcal{L}(\theta_{k+1} \mid S_{k+1}) \;=\; \beta(i + 1, n - i + 1)$.

Now set

$$Q(\,\cdot\,) = \frac{1}{2\sqrt{n}} \sum_{i=Fn-\sqrt{n}}^{Fn+\sqrt{n}} \beta(i+1, n-i+1)\ ,$$

a linear combination of these beta distributions with means near F. It follows from the above that for sufficiently large n,

$$P(\theta,\,\cdot\,) \geq m_4 Q(\,\cdot\,) \qquad \text{for all } \theta \in R_3\ .$$

In other words, once the Markov chain is in R_3 it will tend to "spread out" over all of the interval $[F - (1/\sqrt{n}), F + (1/\sqrt{n})]$ in one more step.

Combining the above reasoning with Lemma 4, we see that we can use Lemma 2 with $k_0 = (A + B_1)\log n + B_2 + 2$, and with $\epsilon = m_1 m_2 m_3 m_4$, to complete the proof of Theorem 3 (with $\Lambda = A + B_1 + \max(B_2, 0) + 2$, and with $\alpha = \epsilon$).

Remarks.

1. We note that the result of Theorem 3 is "tight" in the sense that it really does take $O(\log n)$ steps to approach stationarity in total variation distance. Indeed, let the Markov chain begin in some initial state $\theta_0 \neq F$, say $\theta_0 < F$. Set

$$C_9 = \inf_{0<\theta<F} \frac{F - e(\theta)}{F - \theta}\ .$$

Thus C_9 is a measure (up to $O(1/\sqrt{n})$ errors) of the smallest fraction by which θ_k likely gets closer to F in a single step. Note that $C_9 > 0$ since as $\theta \to F$ this ratio approaches the derivative of the function $f(\theta)$ at F, i.e. $\frac{(1-s)t}{\gamma} + \frac{s(1-t)}{1-\gamma} - 1$ which is positive since $t < \gamma < 1 - s$. We now set $C_{10} = C_9/2$, a ratio strictly smaller (for sufficiently large n) than the smallest fraction by which θ_k likely gets closer to F. Specifically, the probability that θ_k will get closer to F by a ratio smaller than this is exponentially small as a function of n.

We now set $\Gamma = -\frac{1}{4}\log C_{10} > 0$. Then if $k = \Gamma \log n$, then except for events of exponentially small probability, we will have

$$|F - \theta_k| \geq |F - \theta_0|(C_{10})^k = |F - \theta_0| n^{-\frac{1}{4}}\ .$$

But by the remark after Therem 3, the stationary distribution μ is exponentially peaked near F for large n, with width of order $\frac{1}{\sqrt{n}}$. This shows that $L(\theta_k)$ is essentially disjoint from μ for large n, so for such n we will have

$$\|\mathcal{L}(\theta_k) - \mu\| \approx 1 \ .$$

2. The assumption that $t < \gamma < 1-s$, despite its "reasonableness", is not at all necessary. Indeed, if $\gamma \leq t$, we simply replace F by 0 in the above proof, while if $\gamma \geq 1-s$ we simply replace F by 1. The entire proof goes through with only minor modifications. The main differences are that now instead of getting close to a point in the middle of the interval [0,1], the Markov chain will get close to one of the endpoints; also, the "errors" in setting $E(\theta_{k+1} \mid \theta_k = \theta)$ equal to $e(\theta)$ are now $O(1/n)$, instead of $O(1/\sqrt{n})$, once we get close to 0 or 1 (so that $Q(\cdot)$ should now be taken to be roughly uniform on an interval of length about $1/n$ instead of $1/\sqrt{n}$).

4. The case of a general finite state space.

We now turn our attention to the case of general finite $\mathcal{X}$. We set $\mathcal{X} = \{1, 2, \ldots, K\}$, where $K = |\mathcal{X}|$ is regarded as fixed. We set $p_{ab} = P(Y_r = b \mid X_r = a)$ for $1 \leq a, b \leq K$, and we set

$$\gamma_a = (\text{number of } i \text{ for which } Y_i = a)/n \ .$$

We write $\vec{\gamma}$ for $(\gamma_1, \ldots, \gamma_K)$.

The Markov chain takes place on the $(K-1)$-dimensional simplex

$$S_{K-1} = \{\vec{\theta} = (\theta_1, \ldots, \theta_K) \mid \theta_i \geq 0, \ \sum_{i=1}^{K} \theta_i = 1\} \ .$$

The procedure is as follows. Set

$$\eta_b(\vec{\theta}) = P(Y_r = b | X_r \sim \vec{\theta}) = \sum_{a=1}^{K} p_{ab}\theta_a \ , \qquad 1 \leq b \leq K \ ,$$

and set

$$q_{ab}(\vec{\theta}) = P(X_r = a | Y_r = b) = \frac{p_{ab}\theta_a}{\eta_b(\vec{\theta})}, \qquad 1 \leq a, b \leq K \ .$$

Given $\vec{\theta}_k = (\theta_{k,1}, \ldots, \theta_{k,K})$, choose $x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)} \in \{1, \ldots, K\}$ where

$$P(x_r^{(k)} = a) = q_{ab}(\vec{\theta}) \qquad (\text{where } b = Y_r) \ .$$

Then choose $\vec{\theta}_{k+1}$ from the Dirichlet distribution $\mathcal{D}(S_{k,1}+1, S_{k,2}+1, \ldots, S_{k,K}+1)$, where $S_{k,a}$ is the number of r with $x_r^{(k)} = a$.

As in the case $\mathcal{X} = \{0,1\}$, we assume that p_{ab} and γ_a do not vary with n. Under these assumptions we prove that, at least with certain restrictions on p_{ab} and on γ_a, the Data Augmentation algorithm converges (in total variation distance) in $O(\log n)$ steps.

As in the case $\mathcal{X} = \{0,1\}$, we begin the analysis by noting that

$$E(\theta_{k+1,a} \mid \vec{\theta}_k = \vec{\theta}) = e_a(\vec{\theta}) + O(1/\sqrt{n}) \text{ errors}) \ ,$$

where

$$e_a(\vec{\theta}) = \sum_{b=1}^{K} \gamma_b q_{ab}(\vec{\theta}) = \theta_a \sum_{b=1}^{K} \gamma_b \frac{p_{ab}}{\eta_b(\vec{\theta})} \ .$$

Hence, up to $O(1/\sqrt{n})$, the values of $\vec{\theta}_k$ should follow the deterministic prescription

$$(*) \qquad \theta_{k+1,a} = e_a(\vec{\theta}_k) \ .$$

This situation is very similar to the case $\mathcal{X} = \{0,1\}$: the Markov chain follows a dynamical system except for random errors of about $1/\sqrt{n}$. The main difference here is that the dynamical system takes place on the $(K-1)$-dimensional simplex S_{K-1} instead of simply on the interval $[0,1]$. This makes the dynamical system $(*)$ more difficult to analyze, and prevents a complete solution. The following Theorem reduces the study of the Markov chain to the study of the related dynamical system, and we subsequently obtain results about the dynamical system under certain more restrictive assumptions.

Theorem 5. *Suppose that, for given values of $\{p_{ab}\}$ and $\{\gamma_a\}$, the dynamical system given by $(*)$ has the following property: there is a point $\vec{f}$ on the simplex S_{K-1} such that if the dynamical system is started at any point on the simplex S_{K-1} except for a finite number of "exceptional" points, it will converge to $\vec{f}$ exponentially quickly. Then the Data Augmentation algorithm for $\mathcal{X} = \{1, \ldots, K\}$ corresponding to those values of p_{ab} and γ_a will converge in total variation distance to the true posterior in $O(\log n)$ steps, where n is the number of observed data.*

Remarks.

1. Here "exponentially quickly" convergence means that there is a constant A such that for any $\epsilon > 0$, if we start at least ϵ away from all exceptional points (in, say, the L^∞ norm), then after $A \log(1/\epsilon)$ steps we will be within ϵ of $\vec{f}$. Equivalently, if we are close to $\vec{f}$ [resp. to an exceptional point], we can get twice as close [resp. twice as far away] in a constant number of steps.
2. The "exceptional points" here correspond to the points 0 and 1 in the case $\mathcal{X} = \{0, 1\}$; they are possible unstable fixed points of the dynamical system. For example, the K extreme points of the simplex S_{K-1} are all seen to be fixed points of $(*)$, for any p_{ab} and γ_a (though whether or not they are stable does depend on p_{ab} and γ_a). Note that we cannot simply throw away the boundary of the simplex, because that boundary may contain a stable fixed point in addition to various exceptional points.
3. While Theorem 4 does not definitively settle the question of whether the Data Augmentation algorithm will converge in $O(\log n)$ steps, it does reduce the study of a Markov chain to the (simpler) study of an associated dynamical system.
4. It appears (for example from computer simulations) that provided p_{aa} is not too small, the hypothesis of Theorem 4 always holds, i.e. that the dynamical system $(*)$ always converges exponentially quickly to a unique stable fixed point. (Note, however, that this unique stable fixed point may have some coordinates equal to zero in some cases.) However, we are unable to prove this in general; see Propositions 6 and 9 for some partial results. In a particular case (i.e. for particular values of p_{ab} and γ_a), it shouldn't be difficult to check the convergence properties of $(*)$.

Proof of Theorem 5. The proof is of a similar flavour to that of Lemma 4. However, this Theorem is easier because we *assume* the dynamical system has certain convergence properties which had to be proved in Lemma 4.

Here, as there, the key idea is that the Markov chain approximately follows a deterministic prescription which takes it exponentially quickly to a particular fixed point. As in Lemma 4 part (1), after one step the Data Augmentation Markov chain will be about $1/n$ away from the exceptional points, with probability bounded away from 0. This follows from the $O(1/n)$ standard deviation "spreading" of the Dirichlet, just as in Lemma 4.

Then, similar to Lemma 4 parts (2) and (3), the exponential convergence of the dynamical system takes over. For sufficiently large n, with high probability the Markov chain will get close to $\vec{f}$ at a fixed exponential rate chosen to be slightly slower than the rate of the dynamical system. By the assumption of exponential convergence, we see that after $C \log n$ steps (for a constant C independent of n), the Markov chain will be within, say, $1/\sqrt{n}$ of $\vec{f}$, with probability bounded below independently of n. This follows just as in Lemma 4, from noting that as $n \to \infty$, the dynamical system $(*)$ becomes a better and better approximation, with higher and higher probability, to the Markov chain itself. Hence, the probability that the Markov chain *fails* to converge to $\vec{f}$ at a rate slightly *slower* than the dynamical system rate becomes exponentially small.

We finish the proof of Theorem 5 in much the same way we finished the proof of Theorem 3. Once the Markov chain is within $1/\sqrt{n}$ of $\vec{f}$, then after one more step, it will tend (by the spreading of the multinomial) to "spread out" over an area on the simplex with sides about $1/\sqrt{n}$ long. Hence, we can apply Lemma 2 with $Q(\cdot)$ chosen to be a uniform linear combination of Dirichlet distributions with means within $1/\sqrt{n}$ of $\vec{f}$. Setting $k_0 = C \log n$, we can choose ϵ independent of n, to get the desired result. ∎

Theorem 5 suggests that we further analyze the dynamical system given by $(*)$. This appears difficult in general. While there is a huge literature on dynamical systems (see [De], [PdM], and references therein), including the promising theory of Liapounov functions (see [De] p. 176) for showing convergence to fixed points, we are unable to adapt this literature to our present purposes. Instead, we here take a direct approach, and show exponential convergence of our dynamical system in two special cases only. The first, Proposition 6, is a "highly symmetric" case which is very special and whose proof is omitted to save space. The second, Proposition 9, holds for a range of parameters in which the $\vec{Y}$ have high enough probability of being equal to the $\vec{X}$.

Proposition 6. *Suppose $\gamma_a = 1/K$ for each a, and that for some $d < 1/K$, we have $p_{ab} = d$ for each $a \neq b$. Then the dynamical system given by $(*)$ has a unique stable fixed point $\vec{f}$ given by $f_a = 1/K$ for each a. Furthermore, the system converges exponentially* *fast (in the sense of Theorem 5) to $\vec{f}$.*

Combining Proposition 6 with Theorem 5, we immediately obtain

Corollary 7. *Under the hypothesis of Proposition 6, the Data Augmentation algorithm will converge in $O(\log n)$ steps.*

To further analyze the dynamical system given by $(*)$, it is necessary to determine, in somewhat general situations, where the hoped-for stable fixed point $\vec{f}$ might be. To this end, we observe that if $\vec{\theta}$ is such that $\eta_b(\vec{\theta}) = \gamma_b$ for each b, then $e_a(\vec{\theta}) = \theta_a$ for each a, so $\vec{\theta}$ is a fixed point. (This fixed point corresponds to the point F in the case $\mathcal{X} = \{0, 1\}$, and with $t < \gamma < 1 - s$.) Now, it will not always be the case that such an element $\vec{\theta} \in S_{K-1}$ exists. However, if the p_{aa} are sufficiently large, and the γ_a are sufficiently "balanced", then there will be such an element $\vec{\theta}$ as the following Lemma shows.

Lemma 8. *Let $d = \max_a p_{aa}$, and assume $p_{aa} > \frac{1}{2}$. Further, let*

$$y = \max_b \sum_{a \neq b} p_{ab} \, ; \qquad z = \min_b (p_{bb} - \sum_{a \neq b} p_{ab}) \, ; \qquad s = \max_{a,b} \, \frac{\gamma_a}{\gamma_b} \, ;$$

and assume that $s < z/y$ (and in particular that $z > 0$). Then there is a unique point $\vec{f} = (f_1, \ldots, f_K)$ on the simplex S_{K-1} such that $\eta_b(\vec{f}) = \gamma_b$ for each b.

Proof. We write $[p]$ for the matrix with entries p_{ab}. It is easily checked that since $[p]$ is stochastic, and $p_{aa} > \frac{1}{2}$, $[p]$ has no kernel and is therefore invertible. Denote its inverse by $[p]^{-1}$. Set $\vec{f} = [p]^{-1}\vec{\gamma}$, where $\vec{\gamma} = (\gamma_1, \ldots, \gamma_K)$. Then $\vec{\eta}(\vec{f}) = [p]\vec{f} = \vec{\gamma}$ as required. Also $\vec{f}$ is unique by the invertibility of $[p]$. Hence, we need only verify that $\vec{f} \in S_{K-1}$. To this end, we observe that it is easily checked (by working in a basis contained in S_{K-1}) that $[p]^{-1}$ preserves the property of a vector's coordinates summing to 1. Hence since $\sum \gamma_a = 1$, we have $\sum f_a = 1$. We need therefore only verify that $f_a \geq 0$ for each a.

Suppose, to the contrary, that $f_a < 0$ for some a. We shall obtain a contradiction to the statement that $\sum_a p_{ab} f_a = \gamma_b$ for each b. Let i be such that f_i is smallest (and negative), and let I be such that f_I is largest. Let $m = -f_i > 0$, and let $M = f_I > 0$. Clearly $M \geq m$, for if $m > M$ then

$$\gamma_i = \sum_a p_{ai} f_a \leq -m p_{ii} + M \sum_{a \neq i} p_{ai} < -mz < 0 \ ,$$

which is impossible. We then have

$$\gamma_i = \sum_a p_{ai} f_a \le -p_{ii} m + \sum_{a \ne i} p_{ai} M \le yM \ .$$

Also

$$\begin{aligned}\gamma_I &= \sum_a p_{aI} f_a \ge p_{II} M - \sum_{a \ne I} p_{aI} m \\ &\ge \sum_a p_{aI} f_a \ge p_{II} M - \sum_{a \ne I} p_{aI} M \ge zM \ .\end{aligned}$$

Hence

$$s \ge \frac{\gamma_I}{\gamma_i} \ge \frac{z}{y} \ ,$$

contradicting the hypothesis. ∎

Lemma 8 guarantees the existence of a fixed point $\vec{f} \in S_{K-1}$. Under slightly stronger hypothesis, we can actually show that $\vec{\theta}_k$ approaches $\vec{f}$ exponentially quickly.

Proposition 9. *Let d, y, z, and s be as in Lemma 8. For each a, let $p_{*a} = \max_{a' \ne a} p_{a'a}$, and let*

$$r = \min_a \left(1 - \frac{p_{*a}}{\gamma_a p_{aa}}\right); \qquad x = \max_b \sum_a p_{ab} \ .$$

Assume that $d > \frac{1}{2}$, that $s < z/y$, that $r > y/d$, and that

$$(rd - y)z > sx(1 - d) \ .$$

Then the dynamical system $(*)$ converges exponentially quickly.

Remark. Intuitively, p_{aa} is close to 1 for each a, and p_{ab} is small for $a \ne b$. Hence, d is close to 1, y is small, z is close to 1, and r is somewhat close to 1. Also, the parameters are "balanced" so that s and x are not too much greater than 1.

Proof. By Lemma 7 there is a point $\vec{f} \in S_{K-1}$ with $\vec{\eta}(\vec{f}) = \vec{\gamma}$. We shall show that $\vec{\theta}$ approaches $\vec{f}$ exponentially quickly. To that end, we fix $\epsilon > 0$. We assume that initially $\theta_a > \epsilon$ for all a. We let $\vec{\theta}$ progress according to $(*)$.

For technical reasons, we begin by replacing r by a slightly smaller r', so that the hypotheses of the Proposition still hold. We break the proof up into three claims.

Claim 1. After $O(\log(1/\epsilon))$ steps, $\theta_a \geq r'\gamma_a$ for all a.

Indeed, if $0 < \theta_a < r'\gamma_a$ for some a, then

$$\eta_a(\vec{\theta}) = \sum_{a'} p_{a'a}\theta_{a'} \leq p_{aa}\theta_a + p_{*a}(1-\theta_a)$$

$$\leq p_{aa}r'\gamma_a + p_{*a} < p_{aa}\gamma_a \ ,$$

by the definition of r. Hence

$$\frac{e_a(\vec{\theta})}{\theta_a} = \sum_{a'} p_{a'a}\frac{\gamma_{a'}}{\eta_{a'}(\vec{\theta})} \geq p_{aa}\frac{\gamma_a}{\eta_a(\vec{\theta})} > 1 \ .$$

Furthermore, since we replaced r by the smaller r', $\frac{e_a(\vec{\theta})}{\theta_a}$ is actually bounded away from 1. Hence, θ_a will increase at an exponential rate (similar to Lemma 4 (2)) until it is at least $r'\gamma_a$. Finally, since $e_a(\vec{\theta})$ is "monotonic in θ_a" in an appropriate sense, it follows that once $\theta_a \geq r'\gamma_a$, it will remain at least $r'\gamma_a$ thereafter. Claim 1 follows.

We now replace r' by a still smaller r'', such that the hypotheses of the Proposition remain true.

Claim 2. Once Claim 1 is true, then after a constant number of steps $\theta_a \leq R\gamma_a$ for each a, where $R = \frac{r''d}{r''d-y}$.

The proof is similar to that for Claim 1. By Claim 1 we have $\eta_b(\vec{\theta}) \geq p_{bb}\theta_b \geq p_{bb}\gamma_b r'$. Then if $\theta_a > R\gamma_a$ for some a, then

$$\frac{e_a(\vec{\theta})}{\theta_a} \leq p_{aa}\frac{\gamma_a}{\eta_a(\vec{\theta})} \ + \ \left(\sum_{a'\neq a} p_{a'a}\right)\max_b \frac{\gamma_b}{\eta_b(\vec{\theta})}$$

$$\leq p_{aa}\frac{\gamma_a}{p_{aa}\theta_a} \ + \ y\ \max_b \frac{\gamma_b}{p_{bb}\gamma_b r'} < (1/R)y\,\frac{1}{dr'} < 1 \ ,$$

by the definition of R. Furthermore, since we replaced r' by the smaller r'', $\frac{e_a(\vec{\theta})}{\theta_a}$ is bounded away from 1. Also by Claim 1, θ_a is bounded away from 0. Hence θ_a will decrease independently of ϵ until it is less than $R\gamma_a$. Finally, as with Claim 1, once Claim 2 is true it remains true by "monotonicity".

Claim 3. Once Claims 1 and 2 are true, then after $O(\log(1/\epsilon))$ steps we will have

$$\max_a |\theta_a - f_a| < \epsilon \ .$$

Indeed, by Claims 1 and 2, we have that $r' \le \frac{\theta_a}{\gamma_a} \le R$, for each a. This implies that

$$\max_{a,b} \frac{\theta_a}{\theta_b} \le \frac{sR}{r'} \ .$$

Hence,

$$(**) \qquad \max_{a,b} \ \frac{\eta_a(\vec{\theta}) \sum_{a'} p_{a'b}}{\eta_b(\vec{\theta})} \le \frac{sRx}{r'} \ ,$$

by the definition of $\eta_a(\vec{\theta})$.

Now, suppose $|\theta_a - f_a|$ takes its maximum at $a = i$. Assume $|\theta_i - f_i| > 0$. (If $\theta_a = f_a$ for all a, then there is nothing to be proved.) For definiteness suppose $\theta_i > f_i$ (the case $\theta_i < f_i$ is entirely similar). Let $D = \theta_i - f_i > 0$. Then

$$\begin{aligned} \frac{e_i(\vec{\theta})}{\theta_i} - 1 &= \sum_b p_{ib} \left(\frac{\gamma_b}{\eta_b(\vec{\theta})} - 1 \right) = \sum_b p_{ib} \frac{\gamma_b - \eta_b(\vec{\theta})}{\eta_b(\vec{\theta})} \\ &= p_{ii} \frac{\gamma_i - \eta_i(\vec{\theta})}{\eta_i(\vec{\theta})} \ + \ \sum_{b \ne i} p_{ib} \frac{\gamma_b - \eta_b(\vec{\theta})}{\eta_b(\vec{\theta})} \ . \end{aligned}$$

Now,

$$\begin{aligned} \gamma_i - \eta_i(\vec{\theta}) &= \sum_a p_{ai}(f_a - \theta_a) \\ &\le -p_{ai} D + \sum_{a \ne i} p_{ai} D \le -zD \ . \end{aligned}$$

Also $\gamma_b - \eta_b(\vec{\theta}) \le D \sum_a p_{ab}$ for $b \ne i$. Using $(**)$, we obtain that

$$\begin{aligned} \frac{e_i(\vec{\theta})}{\theta_i} - 1 &\le p_{ii} \frac{-zD}{\eta_i(\vec{\theta})} + \left(\sum_{b \ne i} p_{ib} \right) \frac{D \frac{sRx}{r'}}{\eta_a(\vec{\theta})} \\ &\le \frac{D}{\eta_a(\vec{\theta})} \left(-dz + (1-d) \frac{sRx}{r'} \right) \ . \end{aligned}$$

This last expression is strictly negative by the hypothesis and the definition of R. Hence, the value of $\theta_i - f_i$ will decrease. Furthermore, an identical proof to the above shows that *each* $|\theta_a - f_a|$ will be less, on the next step, then the bound on $\theta_i - f_i$ proved above. Hence, $\max_a |\theta_a - f_a|$ will decrease exponentially quickly. This proves Claim 3, and hence establishes the Proposition. ∎

Combining Proposition 9 with Theorem 5, we immediately obtain

Corollary 10. *Under the hypothesis of Proposition 9, the Data Augmentation algorithm will converge in $O(\log n)$ steps.*

We conclude with a remark about priors other than the uniform prior.

Remark. *Other priors.* The results in this paper have all been stated in terms of using a *uniform* prior for the Data Augmentation algorithm. However, the proofs actually work much more generally. In particular, they work for any prior (independent of n) which is bounded above and below by a positive constant times a conjugate (i.e. beta or Dirichlet) prior.

To see this, consider the $\mathcal{X} = \{0, 1\}$ case, and suppose first that we have a $\beta(a_1, a_2)$ prior (with a_1, a_2 independent of n). This affects the Data Augmentation as follows. The law of θ_{k+1} given S_{k+1} will now be $\beta(S_{k+1} + a_1, n - S_{k+1} + a_2)$ instead of of $\beta(S_{k+1} + 1, n - S_{k+1} + 1)$. Hence the mean will be $\frac{S_{k+1}+a_1}{n-S_{k+1}+a_2}$ instead of $\frac{S_{k+1}+1}{n-S_{k+1}+1}$, and the variance will be similarly affected. However, all that was needed in the proof of Theorem 3 (and Lemma 4) was that this law would be peaked (exponentially as a function of n) within $O(1/\sqrt{n})$ of S_{k+1}/n, with width $O(1/\sqrt{n})$. By inspection, this property is preserved, so the proof of Theorem 3 goes through essentially without change. Identical comments apply to a $\mathcal{D}(a_1, a_2, \ldots, a_K)$ prior in Theorem 5.

Now suppose instead that the prior has density $z(x)$ satisfying $m\beta(a_1, a_2; x) \leq z(x) \leq M\beta(b_1, b_2; x)$ for some $m, M > 0$ (and with the prior again independent of n). Then, the law of θ_{k+1} given S_{k+1} will have density $\beta(S_{k+1} + 1, n - S_{k+1} + 1; x)\ z(x)$ which may be rather complicated. On the other hand, the density at any point x will be between $m\beta(S_{k+1} + a_1, n - S_{k+1} + a_2; x)$ and $M\beta(S_{k+1} + b_1, n - S_{k+1} + b_2; x)$. Since $m, M > 0$ are independent of n, for sufficiently large n we see that this density will still be peaked within $O(1/\sqrt{n})$ of S_{k+1}/n, and will still have width $O(1/\sqrt{n})$. Thus, once again the proof of Theorem 3 goes through essentially without change. And, once again, similar comments apply to Theorem 5, using a prior bounded above and below by Dirichet distributions.

Acknowledgements. I am very grateful to Persi Diaconis, my Ph.D. advisor at Harvard University, for suggesting this problem and for many helpful discussions. I thank Peter

Ney, James M^c^Kernan, and the referee for helpful comments. This work was partially supported by the Sloan Foundation and by NSERC of Canada.

REFERENCES

[A] S. Asmussen (1987), *Applied Probability and Queues,* John Wiley & Sons, New York.

[AMN] K.B. Athreya, D. McDonald, and P. Ney (1978), *Limit theorems for semi-Markov processes and renewel theory for Markov chains,* Ann. Prob. **6**, 788-797.

[AN] K.B. Athreya and P. Ney (1978), *A new approach to the limit theory of recurrent Markov chains,* Trans. Amer. Math. Soc. **245**, 493-501.

[De] R.L. Devaney (1989), *Chaotic Dynamical Systems,* Addison-Wesley, New York.

[Di] P. Diaconis (1988), *Group Representations in Probability and Statistics,* IMS Lecture Series volume **11**, Institute of Mathematical Statistics, Hayward, California.

[Do] J.L. Doob (1953), *Stochastic Processes,* Wiley, New York.

[GG] S. Geman and D. Geman (1984), *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images,* IEEE Trans. Pattern Anal. Machine Intel. **6**, 721-741.

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

[GHRS] 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.

[N] E. Nummelin (1984), *General irreducible Markov chains and non-negative operators,* Cambridge University Press.

[P] J.W. Pitman (1976), *On coupling of Markov chains,* Z. Wahrscheinlichkeitstheorie verw. Gebiete **35,** 315-322.

[PdM] J. Palis, Jr. and W. de Melo, *Geometric Theory of Dynamical Systems,* Springer-Verlag, New York.

[R] J.S. Rosenthal (1991), *Rates of convergence for Gibbs sampling for variance component models,* Tech. Rep., Dept. of Mathematics, Harvard University.

[TW] M. Tanner and W. Wong (1987), *The calculation of posterior distributions by data augmentation* (with discussion), J. Amer. Stat. Soc. **81**, 528-550.