Supplement to “Scalable Hyperparameter Selection for Latent Dirichlet Allocation”

Wei Xia
Department of Statistics
University of Florida

Hani Doss
Department of Statistics
University of Florida

Abstract

This document provides proofs for some of the theoretical results in “Scalable Hyperparameter Selection for Latent Dirichlet Allocation” by Wei Xia and Hani Doss, and also the results of an additional experiment.

Throughout this document, equations, tables, etc. are labelled with the prefix “S”. We do this in order to avoid confusion with the equations, tables, etc. of the main paper.

1 Impropriety of the Posterior Induced by the Uniform Prior on the Hyperparameters

Here we show that if the prior on h is the uniform distribution on $(0, \infty)^{T+1}$, then the marginal posterior distribution of $(\boldsymbol{z}, h)$ is improper. The marginal posterior distribution of $(\boldsymbol{z}, h)$ is obtained by integrating out $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ from the full posterior, and may be written explicitly, up to a normalizing constant, as

$$\nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}(\boldsymbol{z}, h) \propto \left[\prod_{d=1}^{D}\left(\frac{\Gamma\big(\sum_{t=1}^{T}\alpha_t\big)}{\prod_{t=1}^{T}\Gamma(\alpha_t)}\frac{\prod_{t=1}^{T}\Gamma(n_{dt}+\alpha_t)}{\Gamma\big(n_d+\sum_{t=1}^{T}\alpha_t\big)}\right)\right]\left[\prod_{t=1}^{T}\left(\frac{\Gamma(V\eta)}{\Gamma(\eta)^V}\frac{\prod_{v=1}^{V}\Gamma(m_{\cdot tv}+\eta)}{\Gamma(m_{\cdot t\cdot}+V\eta)}\right)\right].$$

(See (2.2).) To show posterior impropriety, we need only show that there exists $\boldsymbol{z}$ such that $\int \nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}(\boldsymbol{z}, h)\, dh = \infty$, since this implies $\sum_{\boldsymbol{z}\in\mathcal{Z}} \int \nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}(\boldsymbol{z}, h)\, dh = \infty$, and in fact, it clearly suffices to show that

$$\int_0^\infty \prod_{t=1}^{T}\left(\frac{\Gamma(V\eta)}{\Gamma(\eta)^V}\frac{\prod_{v=1}^{V}\Gamma(m_{\cdot tv}+\eta)}{\Gamma(m_{\cdot t\cdot}+V\eta)}\right) d\eta = \infty. \tag{S-1.1}$$

We rewrite the integrand in (S-1.1) as

$$\prod_{t=1}^{T}\biggl(\frac{\Gamma(V\eta)}{\Gamma(\eta)^V}\frac{\prod_{v=1}^{V}\Gamma(m_{\cdot tv}+\eta)}{\Gamma(m_{\cdot t\cdot}+V\eta)}\biggr)=\biggl(\prod_{t=1}^{T}\frac{\Gamma(V\eta)}{\Gamma(m_{\cdot t\cdot}+V\eta)}\biggr)\biggl(\prod_{t=1}^{T}\prod_{v=1}^{V}\frac{\Gamma(m_{\cdot tv}+\eta)}{\Gamma(\eta)}\biggr). \tag{S-1.2}$$

We then have

$$\begin{aligned}\prod_{t=1}^{T}\frac{\Gamma(V\eta)}{\Gamma(m_{\cdot t\cdot}+V\eta)}&=\prod_{t=1}^{T}\frac{1}{V\eta(V\eta+1)\cdots(V\eta+m_{\cdot t\cdot}-1)}\\&=\prod_{t=1}^{T}\frac{1}{V^{m_{\cdot t\cdot}}\eta^{m_{\cdot t\cdot}}+o(\eta^{m_{\cdot t\cdot}})}=\frac{1}{V^N\eta^N+o(\eta^N)}\qquad\text{as }\eta\to\infty,\end{aligned}\tag{S-1.3}$$

and

$$\prod_{t=1}^{T}\prod_{v=1}^{V}\frac{\Gamma(m_{\cdot tv}+\eta)}{\Gamma(\eta)}=\prod_{t=1}^{T}\prod_{v=1}^{V}\eta(\eta+1)\cdots(\eta+m_{\cdot tv}-1)=\eta^N+o(\eta^N)\qquad\text{as }\eta\to\infty,\tag{S-1.4}$$

where $x=o(y)$ means $\lim_{y\to\infty}x/y=0$ and we recall that$N=\sum_t m_{\cdot t\cdot}=\sum_t\sum_v m_{\cdot tv}$. We now combine (S-1.2), (S-1.3), and (S-1.4) to get

$$\prod_{t=1}^{T}\biggl(\frac{\Gamma(V\eta)}{\Gamma(\eta)^V}\frac{\prod_{v=1}^{V}\Gamma(m_{\cdot tv}+\eta)}{\Gamma(m_{\cdot t\cdot}+V\eta)}\biggr)=\frac{\eta^N+o(\eta^N)}{V^N\eta^N+o(\eta^N)}\to\frac{1}{V^N}\qquad\text{as }\eta\to\infty,$$

which implies (S-1.1). We conclude that the marginal posterior distribution of $(\boldsymbol{z},h)$ is improper.

2 Brief Review of Hamiltonian Monte Carlo

Let X be a random variable on $\mathbb{R}^d$ with a differentiable density q, known up to a normalizing constant, i.e. $q(x)=p(x)/c$, where $p(x)$ is known but c is not, and suppose we wish to sample from q. In HMC, we introduce an auxiliary variable Y, of the same dimension as X, with the following properties: Y is independent of X, and $Y\sim\mathcal{N}(0,M)$, where M is a variance matrix, often taken to be a multiple of the identity matrix. The joint distribution of (X,Y) is therefore given by

$$p_{X,Y}(x,y)\propto\exp\bigl(\log(p(x))-y^\top M^{-1}y/2\bigr).$$

Define

$$H(x,y)=-\log(p_{X,Y}(x,y)),\quad U(x)=-\log(p(x)),\quad\text{and}\quad K(y)=y^\top M^{-1}y/2.\tag{S-2.1}$$

We then have $H(x,y)=U(x)+K(y)+c'$, where c' is a constant which will never play a role, and so may be ignored.

The function H is called the Hamiltonian, and in a physical interpretation represents the total energy of a system consisting of a puck sliding over a frictionless surface of varying heights. At position x, the height of the surface is $-\log(p(x))$, and the variable y represents the momentum of the puck. The *potential energy*, $U(x)$, is proportional to the height at position x. The *kinetic energy* is $(1/2)mv^\top v$, where v is the velocity vector and m is the mass of the puck. Since momentum y is given by $y = mv$, the kinetic energy of the puck with momentum y is $K(y) = (1/2)m^{-1}y^\top y$, which is $(1/2)y^\top M^{-1}y$ if $M = mI_{d\times d}$. The total energy consists of the potential energy $U(x)$ and the kinetic energy $K(y)$, i.e. $H(x, y) = U(x) + K(y)$. In this system the total energy is constant, while the potential and kinetic energy depend on the puck's position. For example, on a level part of the surface, the puck moves at a constant velocity. If it encounters a rising slope, the puck's momentum allows it to continue, with its kinetic energy decreasing and its potential energy increasing, until the kinetic energy is zero. Then it will slide back down with kinetic energy increasing and potential energy decreasing.

Define $x(t)$ to be the position of the puck at time t, and define $y(t)$ to be its momentum at time t. In Hamiltonian dynamics, the pair $(x(t), y(t))$ satisfies a set of differential equations, given by

$$\frac{dx_i}{dt} = \frac{\partial H}{\partial y_i}, \qquad i = 1, \ldots, d,$$
$$\frac{dy_i}{dt} = -\frac{\partial H}{\partial x_i}, \qquad i = 1, \ldots, d.$$

These determine how the pair $(x(t), y(t))$ evolves over time. With the specific form of the potential and kinetic energy functions given by (S-2.1), these equations become

$$\begin{aligned} \frac{dx_i}{dt} &= (M^{-1}y)_i, \qquad i = 1, \ldots, d, \\ \frac{dy_i}{dt} &= \frac{\partial \log(p(x))}{\partial x_i}, \qquad i = 1, \ldots, d. \end{aligned} \tag{S-2.2}$$

Given a starting state $(x(0), y(0))$, this system of equations has a unique solution, which is deterministic, i.e. non-random. Thus, given a starting state $(x(0), y(0))$, the state at time t, $(x(t), y(t))$, is known exactly, and this determines a map T_t from (x, y)-space to itself.

There are three important properties of the Hamiltonian which enable us to construct a Markov chain having invariant distribution $p_{X,Y}$ (for which the X-marginal is q).

P1 Hamiltonian dynamics is reversible, meaning that the maps T_t, $t > 0$ are invertible.

P2 The energy function is preserved, i.e. $H(T_t(x(s), y(s))) = H(x(s), y(s))$ for any t and s.

P3 Volume is preserved, i.e. if R is a region in (x, y)-space having volume V, then the image of

R under T_t also has volume V. This implies that the absolute value of the Jacobian of the map T_t is 1.

Remark 1 *Properties P1–P3 together imply that Hamiltonian dynamics preserves the joint density $p_{X,Y}$: if $(X_0, Y_0) \sim p_{X,Y}$, and we apply the map T_t to (X_0, Y_0), obtaining (X_t, Y_t), then $(X_t, Y_t) \sim p_{X,Y}$.*

We now return to the joint density $p_{X,Y}$. Consider the following sampling scheme. Suppose that the current state is (X, Y), and fix some $t > 0$.

1. We make a draw from the Y-marginal of $p_{X,Y}$, i.e. make a draw from the $\mathcal{N}(0, M)$ distribution. Denote the result by Y'.
2. Set $(X(0), Y(0)) = (X, Y')$.

   A Evolve $(X(0), Y(0))$ for time t according to Hamiltonian dynamics, i.e. using the solution to the Hamiltonian equations (S-2.2). This produces $(X(t), Y(t))$.

   B Negate the momentum variable $Y(t)$, repeat Step 2A above, and again negate the resulting momentum variable.

It is obvious that Step 1 preserves $p_{X,Y}$. Step 2 also preserves $p_{X,Y}$, as noted in Remark 1. In fact, because $p_{X,Y}(x(t), y(t)) = p_{X,Y}(x(0), y(0))$, Step 2 may be viewed as a Metropolis step for which the acceptance probability is always 1 (Step 2B ensures that the Metropolis proposal is symmetric, so there is no need for the Hastings correction in the acceptance probability). Therefore, the transition consisting of both steps also preserves $p_{X,Y}$. Consequently, this scheme can be used to form a Markov chain whose invariant distribution is $p_{X,Y}$ and, assuming we can establish ergodicity, this chain can be used to estimate the distribution $p_{X,Y}$ and its features.

Unfortunately, the scheme described above is rarely feasible, because the differential equations given by (S-2.2) cannot be solved explicitly. There are several methods for producing numerical approximations to the solution, all of which depend on the idea of discretizing time t, introducing some small step size ϵ. The most widely used scheme is the so-called *leapfrog* method. One iteration of this method maps the state $(x(t), y(t))$ to $(x(t + \epsilon), y(t + \epsilon))$, through the following steps:

$$\begin{aligned} y(t + \epsilon/2) &= y(t) + (\epsilon/2)\nabla_x \log(p(x)), \\ x(t + \epsilon) &= x(t) + \epsilon M^{-1} y(t + \epsilon/2), \\ y(t + \epsilon) &= y(t + \epsilon/2) + (\epsilon/2)\nabla_x \log(p(x(t + \epsilon))). \end{aligned}$$

The steps above are repeated for L iterations, after which the momentum variable is negated, the L iterations are repeated, and the momentum variable is negated again.

Suppose now that we carry out Steps 1 and 2, except that in Step 2, instead of using the exact solution to the Hamiltonian equations, we use the approximation given by the leapfrog algorithm. Step 2 will then not preserve the density $p_{X,Y}$. However, this problem can be dealt with by accepting or rejecting the proposed state with the Metropolis acceptance probability. Let (x^*, y^*) denote the state produced by the adjusted scheme. We accept (x^*, y^*) with probability

$$r = \min\bigl\{1, \exp\bigl(-H(x^*, y^*) + H(x, y)\bigr)\bigr\}, \tag{S-2.3}$$

where (x, y) denotes the state obtained after Step 1. The scheme consisting of Step 1 and the new Step 2 preserves $p_{X,Y}$. Note that if the leapfrog algorithm gives a solution which is close to the exact solution, then the Hamiltonian for the new position is close to the Hamiltonian for the old position, so that in (S-2.3), r is close to 1, which is a major advantage of the method. Generally speaking, HMC makes better use of what is known about the un-normalized density p (including the gradient of its logarithm) in comparison with random walk Metropolis algorithms. The result is that it can make proposals that are far from the current state, yet maintain a high acceptance rate, and thus can greatly outperform random walk Metropolis chains. For details and references, see Neal (2011).

In practice, the negation of the momentum variable is not needed, because for our particular choice of K, $K(-y) = K(y)$. One complete cycle of HMC implemented via leapfrog is given by Algorithm S-1. In the leapfrog algorithm described above, ϵ and L are parameters of the algorithm. Generally speaking, if ϵ is small, then the leapfrog approximation to the exact solution to the Hamiltonian equations is accurate, and so the acceptance probability is close to 1; but using a small ϵ requires more computation. Also, generally speaking, a large value for L results in bigger moves, and thus smaller correlation between successive moves; on the other hand, using a large L requires more computation. These two parameters need to be tuned. This can be done by trial and error, using preliminary runs; see Neal (2011) for guidelines. There is also some recent work on automatic tuning of these parameters; see Girolami and Calderhead (2011) and Hoffman and Gelman (2014).

Algorithm S-1: General Hamiltonian Monte Carlo

```
Data: Un-normalized density $p$
Result: A Markov chain $(X_1, Y_1), (X_2, Y_2), \dots$ for which the $X$-sequence has invariant
        distribution equal to the normalized $p$
Initialize $x^{(0)}, \epsilon, L, n$;
for $i = 1, \dots, n$ do
    generate $y^{(0)} \sim \mathcal{N}(0, M)$;
    set $x^{(i)} \leftarrow x^{(i-1)}$, $\tilde{x} \leftarrow x^{(i-1)}$, $\tilde{y} \leftarrow y^{(0)}$;
    for $j = 1, \dots, L$ do
        set $\tilde{y} \leftarrow \tilde{y} + \epsilon \nabla_x \log(p(\tilde{x}))/2$;
        set $\tilde{x} \leftarrow \tilde{x} + \epsilon M^{-1} \tilde{y}$;
        set $\tilde{y} \leftarrow \tilde{y} + \epsilon \nabla_x \log(p(\tilde{x}))/2$
    set $r = \exp\{\log(p(\tilde{x})) - \frac{1}{2}\tilde{y}^\top M^{-1} \tilde{y} - \log(p(x^{(i-1)})) + \frac{1}{2}(y^{(0)})^\top M^{-1} y^{(0)}\}$;
    with probability $\min\{1, r\}$ set $x^{(i)} \leftarrow \tilde{x}$;
```

3 Proof of Theorem 1

Uniform ergodicity is equivalent to the so-called Doeblin condition, which is that there exist a probability measure Π on $(\Lambda, \mathcal{B}_\Lambda)$, an integer m, and a constant $\epsilon > 0$ such that $P^m(\lambda, C) \geq \epsilon \Pi(C)$ for all $\lambda \in \Lambda$ and $C \in \mathcal{B}_\Lambda$; see Theorem 3 of Athreya et al. (1996). We will prove uniform ergodicity by establishing a Doeblin condition (with $m = 1$) stated in terms of the Markov transition density, i.e. $p(\lambda, \lambda') \geq \epsilon \pi(\lambda')$ for all $\lambda, \lambda' \in \Lambda$, where $p(\lambda, \cdot)$ and $\pi(\cdot)$ are densities on Λ (with respect to the natural measure on Λ, namely a product measure involving Lebesgue measure on the continuous components and counting measure on the discrete components).

Without loss of generality, we assume that the bounded hyper-rectangle has the form $\mathcal{H}_0 = [H_l, H_u]^{T+1}$, where $H_u > H_l > 0$. The Markov transition density $p(\cdot, \cdot)$ for the data augmentation chain may be decomposed as

$$\begin{aligned} p(\lambda, \lambda') = \nu_{\boldsymbol{z} \mid h}(\boldsymbol{z}' \mid h)\, \nu_{\boldsymbol{I} \mid (h, \boldsymbol{z})}(\boldsymbol{i}' \mid h, \boldsymbol{z}')\, \nu_{\boldsymbol{J} \mid (h, \boldsymbol{z})}(\boldsymbol{j}' \mid h, \boldsymbol{z}')\, \nu_{\boldsymbol{Q} \mid (h, \boldsymbol{z})}(\boldsymbol{q}' \mid h, \boldsymbol{z}') \\ \nu_{\boldsymbol{R} \mid (h, \boldsymbol{z})}(\boldsymbol{r}' \mid h, \boldsymbol{z}')\, \nu_{h \mid (\boldsymbol{z}, \boldsymbol{I}, \boldsymbol{Q}, \boldsymbol{J}, \boldsymbol{R})}(h' \mid \boldsymbol{z}', \boldsymbol{I}', \boldsymbol{Q}', \boldsymbol{J}', \boldsymbol{R}'), \end{aligned} \tag{S-3.1}$$

where we have used several conditional independence properties of the data augmentation chain. All the quantities on the right side of (S-3.1) depend on $\boldsymbol{w}$, but we have suppressed this dependence to lighten the notation. The right side of (S-3.1) is the product of six conditional densities. Our plan is to obtain lower bounds for the first five of these, having the following form:

$$\nu_{\boldsymbol{z}\,|\,h}(\boldsymbol{z}'\,|\,h) \geq \epsilon_1 \pi_1(\boldsymbol{z}'), \tag{S-3.2a}$$

$$\nu_{\boldsymbol{I}\,|\,(h,\boldsymbol{z})}(\boldsymbol{i}'\,|\,h,\boldsymbol{z}') \geq \epsilon_2 \pi_2(\boldsymbol{i}'\,|\,\boldsymbol{z}'), \tag{S-3.2b}$$

$$\nu_{\boldsymbol{J}\,|\,(h,\boldsymbol{z})}(\boldsymbol{j}'\,|\,h,\boldsymbol{z}') \geq \epsilon_3 \pi_3(\boldsymbol{j}'\,|\,\boldsymbol{z}'), \tag{S-3.2c}$$

$$\nu_{\boldsymbol{Q}\,|\,(h,\boldsymbol{z})}(\boldsymbol{q}'\,|\,h,\boldsymbol{z}') \geq \epsilon_4 \pi_4(\boldsymbol{q}'), \tag{S-3.2d}$$

$$\nu_{\boldsymbol{R}\,|\,(h,\boldsymbol{z})}(\boldsymbol{r}'\,|\,h,\boldsymbol{z}') \geq \epsilon_5 \pi_5(\boldsymbol{r}'). \tag{S-3.2e}$$

In (S-3.2), $\pi_1(\cdot)$, $\pi_4(\cdot)$, and $\pi_5(\cdot)$ are single distributions, but $\pi_2(\cdot\,|\,\boldsymbol{z}')$ and $\pi_3(\cdot\,|\,\boldsymbol{z}')$ are distributions which depend on $\boldsymbol{z}'$. Also, in (S-3.2), $\epsilon_1, \ldots, \epsilon_5 > 0$. The fact that $\pi_2(\cdot\,|\,\boldsymbol{z}')$ and $\pi_3(\cdot\,|\,\boldsymbol{z}')$ depend on $\boldsymbol{z}'$ does not create a problem, because if we take

$$\begin{aligned}\pi(\boldsymbol{z}', \boldsymbol{I}', \boldsymbol{Q}', \boldsymbol{J}', \boldsymbol{R}', h') = \pi_1(\boldsymbol{z}')\, \pi_2(\boldsymbol{i}'\,|\,\boldsymbol{z}')\, \pi_3(\boldsymbol{j}'\,|\,\boldsymbol{z}')\, \pi_4(\boldsymbol{q}')\, \pi_5(\boldsymbol{r}') \\ \nu_{h\,|\,(\boldsymbol{z},\boldsymbol{I},\boldsymbol{Q},\boldsymbol{J},\boldsymbol{R})}(h'\,|\,\boldsymbol{z}', \boldsymbol{I}', \boldsymbol{Q}', \boldsymbol{J}', \boldsymbol{R}'),\end{aligned} \tag{S-3.3}$$

then it is easy to see that π is a probability measure on Λ. It is a *single* distribution, i.e. it does not depend on $(\boldsymbol{z}, \boldsymbol{I}, \boldsymbol{Q}, \boldsymbol{J}, \boldsymbol{R}, h)$. Thus, taking $\epsilon = \epsilon_1 \cdots \epsilon_5$, from (S-3.1) and (S-3.2) we get $p(\lambda, \lambda') \geq \epsilon \pi(\boldsymbol{z}', \boldsymbol{I}', \boldsymbol{Q}', \boldsymbol{J}', \boldsymbol{R}', h')$, i.e. $p(\lambda, \lambda') \geq \epsilon \pi(\lambda')$. We now proceed with the details.

Proof of (S-3.2a) From (2.3), for every $d = 1, \ldots, D$ and $i = 1, \ldots, n_d$, we have

$$\begin{aligned}&\nu_{z_{di}\,|\,(\boldsymbol{z}_{(-di)},h,\boldsymbol{w})}(e_t) \\ &= \biggl(\frac{n_{dt(-di)} + \alpha_t}{n_d - 1 + \sum_{t'=1}^{T} \alpha_{t'}}\biggr) \biggl(\frac{m_{\cdot tv(-di)} + \eta}{m_{\cdot t\cdot(-di)} + V\eta}\biggr) \Biggl[\sum_{s=1}^{T} \biggl(\frac{n_{ds(-di)} + \alpha_s}{n_d - 1 + \sum_{s'=1}^{T} \alpha_{s'}}\biggr) \biggl(\frac{m_{\cdot sv(-di)} + \eta}{m_{\cdot s\cdot(-di)} + V\eta}\biggr)\Biggr]^{-1}.\end{aligned}$$

Since

$$\frac{n_{dt(-i)} + \alpha_t}{n_d - 1 + \sum_{t'=1}^{T} \alpha_{t'}} \leq 1 \qquad \text{and} \qquad \frac{m_{\cdot tv(-di)} + \eta}{m_{\cdot t\cdot(-di)} + V\eta} \leq 1,$$

it follows that

$$\begin{aligned}\nu_{z_{di}\,|\,(\boldsymbol{z}_{(-di)},h,\boldsymbol{w})}(e_t) &\geq \biggl(\frac{n_{dt(-di)} + \alpha_t}{n_d - 1 + \sum_{t'=1}^{T} \alpha_{t'}}\biggr) \biggl(\frac{m_{\cdot tv(-di)} + \eta}{m_{\cdot t\cdot(-di)} + V\eta}\biggr) \frac{1}{T} \\ &\geq \frac{H_l}{N - 1 + TH_u} \cdot \frac{H_l}{N + VH_u} \cdot \frac{1}{T}.\end{aligned}$$

Therefore,

$$\begin{aligned}\nu_{\boldsymbol{z}\,|\,h}(\boldsymbol{z}'\,|\,h) &\geq \prod_{d=1}^{D} \prod_{i=1}^{n_d} \biggl[\frac{H_l}{N - 1 + TH_u} \cdot \frac{H_l}{N + VH_u} \cdot \frac{1}{T}\biggr] = \biggl[\frac{H_l}{N - 1 + TH_u} \cdot \frac{H_l}{N + VH_u} \cdot \frac{1}{T}\biggr]^{N} \\ &= \epsilon_1 \pi_1(\boldsymbol{z}'),\end{aligned}$$

where π_1 denotes the uniform distribution on $\boldsymbol{z}$-space (which has cardinality T^N), and $\epsilon_1 = \big\{H_l^2/[(N-1+TH_u)(N+VH_u)]\big\}^N$.

Proof of (S-3.2b) *and* (S-3.2c) We first prove (S-3.2b). From (2.16) and (2.8), for every $d = 1, \dots, D$ and $t = 1, \dots, T$, we have

$$\nu_{I_{dt} \,|\, (h,\boldsymbol{z})}(i_{dt} \,|\, h, \boldsymbol{z}) = \begin{cases} \dfrac{\Gamma(\alpha_t)}{\Gamma(n_{dt}+\alpha_t)} S(n_{dt}, i_{dt}) \alpha_t^{i_{dt}} & \text{if } n_{dt} > 0, \\ \delta_0(i_{dt}) & \text{if } n_{dt} = 0. \end{cases}$$

Note that n_{dt} is a function of $\boldsymbol{z}$. If $n_{dt} > 0$, then because $S(n_{dt}, i_{dt}) \geq 1$, we have

$$\nu_{I_{dt} \,|\, (h,\boldsymbol{z})}(i_{dt} \,|\, h, \boldsymbol{z}) \geq \frac{\Gamma(\alpha_t)}{\Gamma(n_{dt}+\alpha_t)} \alpha_t^{n_{dt}} \delta_{n_{dt}}(i_{dt}) \geq M \frac{\Gamma(H_l)}{\Gamma(N+H_u)} \delta_{n_{dt}}(i_{dt}),$$

where $M = \min_{\alpha \in [H_l, H_u],\, i \in \{1,\dots,N\}} \alpha^i$. If $n_{dt} = 0$, then trivially

$$\nu_{I_{dt} \,|\, (h,\boldsymbol{z})}(i_{dt} \,|\, h, \boldsymbol{z}) = \delta_{n_{dt}}(i_{dt}).$$

To combine the two cases, we let $M^* = \min\{1, M\Gamma(H_l)/\Gamma(N+H_u)\}$ and write

$$\nu_{I_{dt} \,|\, (h,\boldsymbol{z})}(i_{dt} \,|\, h, \boldsymbol{z}) \geq M^* \delta_{dt}(i_{dt}).$$

Therefore,

$$\nu_{\boldsymbol{I} \,|\, (h,\boldsymbol{z})}(\boldsymbol{i} \,|\, h, \boldsymbol{z}) = \prod_{d=1}^{D} \prod_{t=1}^{T} \nu_{i_{dt} \,|\, (h,\boldsymbol{z})}(i_{dt} \,|\, h, \boldsymbol{z}) \geq \big[M^*\big]^{DT} \prod_{d=1}^{D} \prod_{t=1}^{T} \delta_{dt}(i_{dt}),$$

and this establishes (S-3.2b) with $\pi_2(\boldsymbol{i}' \,|\, \boldsymbol{z}') = \prod_{d=1}^{D} \prod_{t=1}^{T} \delta_{dt}(i'_{dt})$ and $\epsilon_2 = \big[M^*\big]^{DT}$. The proof of (S-3.2c) is very similar and is omitted.

Proof of (S-3.2d) *and* (S-3.2e) We begin with (S-3.2d). From (2.17), we have

$$\begin{aligned} \nu_{\boldsymbol{Q} \,|\, (h,\boldsymbol{z})}(\boldsymbol{q} \,|\, h, \boldsymbol{z}) &= \prod_{d=1}^{D} \left[\frac{\Gamma\big(\sum_{t=1}^{T} \alpha_t + n_d\big)}{\Gamma\big(\sum_{t=1}^{T} \alpha_t\big)\Gamma(n_d)} q_d^{\sum_{t=1}^{T} \alpha_t - 1} (1-q_d)^{n_d - 1} \right] \\ &\geq \prod_{d=1}^{D} \Big[q_d^{\sum_{t=1}^{T} \alpha_t - 1} (1-q_d)^{n_d - 1} \Big] \\ &\geq \prod_{d=1}^{D} \big[q_d^{TH_u - 1} (1-q_d)^{n_d - 1} \big] \\ &= \left[\prod_{d=1}^{D} \frac{\Gamma(TH_u)\Gamma(n_d)}{\Gamma(TH_u + n_d)} \right] \left[\prod_{d=1}^{D} \frac{\Gamma(TH_u + n_d)}{\Gamma(TH_u)\Gamma(n_d)} q_d^{TH_u - 1} (1-q_d)^{n_d - 1} \right] \\ &= \epsilon_4 \pi_4(\boldsymbol{q}), \end{aligned}$$

where $\epsilon_4 = \prod_{d=1}^{D} \bigl[\Gamma(TH_u)\Gamma(n_d)/\Gamma(TH_u + n_d)\bigr]$, and π_4 denotes the product of D beta densities with parameters TH_u and n_d. This gives (S-3.2d). The proof of (S-3.2e) is very similar and is omitted.

To wrap up, combining (S-3.2a)–(S-3.2e), we have shown that the Markov transition density $p(\cdot,\cdot)$ given by (S-3.1) satisfies a Doeblin condition with π given by (S-3.3) and $\epsilon = \epsilon_1 \cdots \epsilon_5$, and we conclude that the data augmentation chain is uniformly ergodic. □

4 Proof of Theorem 2

The proof of Theorem 2 is based on results in Doss and Park (2018), who use empirical process theory to obtain uniformity in the strong law of large numbers and the central limit theorem ((3.2) and (3.3) in our context). Let $\xi = (\boldsymbol{z}, \boldsymbol{I}, \boldsymbol{Q}, \boldsymbol{J}, \boldsymbol{R})$ (so ξ consists of all the components of λ except for h), and let $f_h(\xi) = \nu_{h \,|\, (\boldsymbol{w},\xi)}(h)/\nu_h(h)$, where $\nu_{h \,|\, (\boldsymbol{w},\xi)}(h)$ has the form of the quantity inside the braces in (3.4). Note that $\widehat{m}_n(h)$ is an average of the functions $f_h(\xi^{(k)})$, $k = 1, \dots, n$. Empirical process theory asserts uniformity in the strong law and the central limit theorem under very strong regularity conditions, of which the main ones, in our context, are that $E\bigl(\sup_h f_h(\xi)\bigr) < \infty$ and $E\bigl(\sup_h f_h^2(\xi)\bigr) < \infty$. Doss and Park (2018) consider a set of regularity conditions which, in our context, are A1–A4, and additionally the following.

B1 The Markov chain $\xi^{(1)}, \xi^{(2)} \dots$ is geometrically ergodic.

B2 For each $h \in \mathcal{H}$, there exists $\epsilon > 0$ such that $E\bigl(\|\nabla_h f_h(\xi)\|^{2+\epsilon}\bigr) < \infty$.

B3 For some $d \geq 1$, there exist $h_1, \dots, h_d \in \mathcal{H}$ and constants $c_1, \dots, c_d$, such that

$$\sup_h f_h(\xi) \leq \sum_{j=1}^{d} c_j f_{h_j}(\xi) \qquad \text{for all } \xi \in \Xi.$$

B4 For some $d \geq 1$, there exist $h_1, \dots, h_d \in \mathcal{H}$ and constants $c_1, \dots, c_d$, such that

$$\sup_h \|\nabla_h f_h(\xi)\|_\infty \leq \sum_{j=1}^{d} c_j \|\nabla_h f_{h_j}(\xi)\|_\infty \qquad \text{for all } \xi \in \Xi.$$

B5 For some $d \geq 1$, there exist $h_1, \dots, h_d \in \mathcal{H}$ and constants $c_1, \dots, c_d$, such that

$$\sup_h \|\nabla_h^2 f_h(\xi)\|_\infty \leq \sum_{j=1}^{d} c_j \|\nabla_h^2 f_{h_j}(\xi)\|_\infty \qquad \text{for all } \xi \in \Xi.$$

In B4 and B5, $\|\cdot\|_\infty$ is defined by $\|x\|_\infty = \max_i |x_i|$ and $\|x\|_\infty = \max_{ij} |x_{ij}|$, respectively. Doss and Park (2018) show (see their Theorem 4 and Remarks 1 and 5) that if Conditions A1–A4 and B1–B5 are satisfied, then statements (3.2) and (3.3) regarding uniformity in the strong law and central limit theorem are true, and the conclusions of our Theorem 2 hold. Therefore, to prove Theorem 2, we will establish Conditions B1–B5. Condition B1 is implied by Theorem 1, which asserts the stronger statement of uniform ergodicity of the data augmentation chain. We will first establish B3–B5 (with $d = 1$), and then deal with B2.

Proof of B3 We may write $f_h(\xi)$ in simplified form as

$$f_h(\xi) = C(\xi) \exp\Big\{ \tilde{J} \log(\eta) - V\tilde{R}\eta + \sum_{t=1}^{T} \big[\tilde{I}_t \log(\alpha_t) - \tilde{Q}\alpha_t \big] \Big\},$$

where $C(\xi)$ denotes the normalizing constant of $\nu_{h\,|\,\boldsymbol{w},\xi}$, $\tilde{I}_t = \sum_{d=1}^{D} I_{dt}$, $\tilde{Q} = -\sum_{d=1}^{D} \log(Q_d)$, $\tilde{J} = \sum_{t=1}^{T}\sum_{v=1}^{V} J_{tv}$, and $\tilde{R} = -\sum_{t=1}^{T} \log(R_t)$. For simplicity, we have omitted the normalizing constant of the prior ν_h, and this is without loss of generality because this constant depends only on the prespecified values a and b. Without loss of generality, we assume that the compact set $\mathcal{H}$ has the form $\mathcal{H} = [H_l, H_u]^{T+1}$, where $H_u > H_l > 0$. Note that $\tilde{Q}, \tilde{R} > 0$ and that $\tilde{I}_t, \tilde{J} \in \{1, \dots, N\}$ for $t = 1, \dots, T$ (recall that $N = \sum_{d=1}^{D} n_d$). For any $\xi \in \Xi$ we have

$$\begin{aligned}
f_h(\xi) &\le C(\xi) \exp\Big\{ \tilde{J} \log(H_u) - V\tilde{R}H_l + \textstyle\sum_{t=1}^{T} \big[\tilde{I}_t \log(H_u) - \tilde{Q}H_l \big] \Big\} \\
&= C(\xi) \exp\Big\{ \tilde{J} \log(H_l) - V\tilde{R}H_l + \textstyle\sum_{t=1}^{T} \big[\tilde{I}_t \log(H_l) - \tilde{Q}H_l \big] + \log(H_u/H_l)\big[\tilde{J} + \textstyle\sum_{t=1}^{T} \tilde{I}_t \big] \Big\} \\
&\le C(\xi) \exp\Big\{ \tilde{J} \log(H_l) - V\tilde{R}H_l + \textstyle\sum_{t=1}^{T} \big[\tilde{I}_t \log(H_l) - \tilde{Q}H_l \big] + \log(H_u/H_l)(T+1)N \Big\} \\
&= c_* f_{h_*}(\xi), \qquad \text{(S-4.1)}
\end{aligned}$$

where $c_* = (H_u/H_l)^{\exp[(T+1)N]}$ and $h_* = (H_l, \dots, H_l)$. This establishes B3 with $d = 1$.

Proof of B4 The partial derivative of f_h with respect to η is given by

$$\frac{\partial f_h(\xi)}{\partial \eta} = \big(\tilde{J}/\eta - V\tilde{R} \big) f_h(\xi), \qquad \text{(S-4.2)}$$

and the partial derivatives of f_h with respect to α_t, $t = 1, \dots, T$ are

$$\frac{\partial f_h(\xi)}{\partial \alpha_t} = \big(\tilde{I}_t/\alpha_t - \tilde{Q} \big) f_h(\xi).$$

We now consider the partial derivative with respect to η. For any $\xi \in \Xi$ we have

$$\begin{aligned}
|\partial f_h(\xi)/\partial\eta| &= \big|\big(\tilde{J}/\eta - V\tilde{R}\big) f_h(\xi)\big| \\
&\le \max\big\{\big|\big(\tilde{J}/H_l - V\tilde{R}\big) f_h(\xi)\big|,\ \big|\big(\tilde{J}/H_u - V\tilde{R}\big) f_h(\xi)\big|\big\} \\
&\le \big|\big(\tilde{J}/H_l - V\tilde{R}\big) f_h(\xi)\big| + \big|\big(\tilde{J}/H_u - V\tilde{R}\big) f_h(\xi)\big| \\
&= \left(1 + \left|\frac{\tilde{J}/H_u - V\tilde{R}}{\tilde{J}/H_l - V\tilde{R}}\right|\right) \cdot \big|\big(\tilde{J}/H_l - V\tilde{R}\big) f_h(\xi)\big|.
\end{aligned}$$

It is easy to see that for any fixed $\tilde{J} \in \{1, \dots, N\}$,

$$\left|\frac{\tilde{J}/H_u - V\tilde{R}}{\tilde{J}/H_l - V\tilde{R}}\right| \to \frac{H_l}{H_u} \quad \text{as } \tilde{R} \to 0 \qquad \text{and} \qquad \left|\frac{\tilde{J}/H_u - V\tilde{R}}{\tilde{J}/H_l - V\tilde{R}}\right| \to 1 \quad \text{as } \tilde{R} \to \infty. \tag{S-4.3}$$

Let $g_{\tilde{J}}(\tilde{R}) = \big|\big(\tilde{J}/H_u - V\tilde{R}\big)/\big(\tilde{J}/H_l - V\tilde{R}\big)\big|$. Since $g_{\tilde{J}}$ is continuous, (S-4.3) implies that $g_{\tilde{J}}$ is uniformly bounded on $(0, \infty)$. Letting M_{T+1} be a bound, and with c_* and h_* defined as in the proof of B3, for all $\xi \in \Xi$ we have

$$\begin{aligned}
|\partial f_h(\xi)/\partial\eta| &\le (1 + M_{T+1})\big|\big(\tilde{J}/H_l - V\tilde{R}\big) f_h(\xi)\big| \\
&\le (1 + M_{T+1})c_*\big|\big(\tilde{J}/H_l - V\tilde{R}\big) f_{h_*}(\xi)\big| \\
&= c_*(1 + M_{T+1})|\partial f_{h_*}(\xi)/\partial\eta|,
\end{aligned} \tag{S-4.4}$$

where the second inequality comes from (S-4.1), and the equality results from (S-4.2) after we recall that $h_* = (H_l, \dots, H_l)$.

Similarly, for any $t = 1, \dots, T$, there exists M_t such that

$$|\partial f_h(\xi)/\partial\alpha_t| \le c_*(1 + M_t)|\partial f_{h_*}(\xi)/\partial\alpha_t| \qquad \text{for all } \xi \in \Xi. \tag{S-4.5}$$

We conclude that

$$\|\nabla_h f_h(\xi)\|_\infty \le c_{1*}\|\nabla_h f_{h_*}(\xi)\|_\infty \qquad \text{for all } \xi \in \Xi,$$

where $c_{1*} = c_* \max_{t=1,\dots,T+1}\{1 + M_t\}$. This proves B4 with $d = 1$.

Proof of B5 To prove B5, we will find bounds on the absolute values of the second-order partial derivatives $\partial^2 f_h(\xi)/\partial\alpha_t\partial\eta$, $\partial^2 f_h(\xi)/\partial\alpha_t\partial\alpha_{t'}$ $(t \ne t')$, $\partial^2 f_h(\xi)/\partial\eta^2$, and $\partial^2 f_h(\xi)/\partial\alpha_t^2$, and we will proceed in this order. In what follows, c_* and M_t are defined as in the proofs of B3 and B4.

For $\partial^2 f_h(\xi)/\partial\alpha_t\partial\eta$, we first note that

$$\frac{\partial^2 f_h(\xi)}{\partial\alpha_t\partial\eta} = \left(\frac{\tilde{J}}{\eta} - V\tilde{R}\right)\frac{\partial f_h(\xi)}{\partial\alpha_t}. \tag{S-4.6}$$

Therefore, we have

$$\begin{aligned}
\left|\frac{\partial^2 f_h(\xi)}{\partial\alpha_t\partial\eta}\right| &= \left|\left(\frac{\tilde{J}}{\eta} - V\tilde{R}\right)\frac{\partial f_h(\xi)}{\partial\alpha_t}\right| \\
&\leq \left|\left(\frac{\tilde{J}}{H_l} - V\tilde{R}\right)\frac{\partial f_h(\xi)}{\partial\alpha_t}\right| + \left|\left(\frac{\tilde{J}}{H_u} - V\tilde{R}\right)\frac{\partial f_h(\xi)}{\partial\alpha_t}\right| \\
&= \left(1 + \left|\frac{\tilde{J}/H_u - V\tilde{R}}{\tilde{J}/H_l - V\tilde{R}}\right|\right)\cdot\left|\left(\frac{\tilde{J}}{H_l} - V\tilde{R}\right)\frac{\partial f_h(\xi)}{\partial\alpha_t}\right| \\
&\leq (1+M_{T+1})c_*(1+M_t)\left|\left(\frac{\tilde{J}}{H_l} - V\tilde{R}\right)\frac{\partial f_{h_*}(\xi)}{\partial\alpha_t}\right| \\
&= c_*(1+M_{T+1})(1+M_t)\left|\frac{\partial^2 f_{h_*}(\xi)}{\partial\alpha_t\partial\eta}\right|,
\end{aligned}$$

where the second inequality comes from (S-4.5) and the definition of M_{T+1}, and the last equality comes from (S-4.6) and the fact that $h_* = (H_l, \ldots, H_l)$.

Similarly, we can prove that for $t \neq t'$,

$$\left|\frac{\partial^2 f_h(\xi)}{\partial\alpha_t\partial\alpha_{t'}}\right| = \left|\left(\frac{\tilde{I}_{t'}}{\alpha_{t'}} - \tilde{Q}\right)\frac{\partial f_h(\xi)}{\partial\alpha_t}\right| \leq c_*(1+M_{t'})(1+M_t)\left|\frac{\partial^2 f_{h_*}(\xi)}{\partial\alpha_t\partial\alpha_{t'}}\right|.$$

For $\partial^2 f_h(\xi)/\partial\eta^2$, we have

$$\frac{\partial^2 f_h(\xi)}{\partial\eta^2} = -\frac{\tilde{J}}{\eta^2}f_h(\xi) + \left(\frac{\tilde{J}}{\eta} - V\tilde{R}\right)\frac{\partial f_h(\xi)}{\partial\eta} = \left(\frac{\tilde{J}\big(\tilde{J}-1\big)}{\eta^2} - \frac{2V\tilde{J}\tilde{R}}{\eta} + V^2\tilde{R}^2\right)f_h(\xi).$$

In the equation above, the coefficient of $f_h(\xi)$ may be viewed as a quadratic polynomial in $1/\eta$. The value of η minimizing it is $\eta' = \big(\tilde{J}-1\big)/\big(V\tilde{R}\big)$, and at that point the value of the coefficient is $-V^2\tilde{R}^2/\big(\tilde{J}-1\big)$. Therefore, the absolute value of the maximum must occur at H_u, H_l, or η'.

Consequently,

$$\begin{aligned}
\left|\frac{\partial^2 f_h(\xi)}{\partial\eta^2}\right| &\le \left|\left(\frac{\tilde J(\tilde J-1)}{H_l^2}-\frac{2V\tilde J\tilde R}{H_l}+V^2\tilde R^2\right)f_h(\xi)\right| + \left|\left(\frac{\tilde J(\tilde J-1)}{H_u^2}-\frac{2V\tilde J\tilde R}{H_u}+V^2\tilde R^2\right)f_h(\xi)\right| \\
&\qquad + \left|\frac{V^2\tilde R^2}{\tilde J-1}f_h(\xi)\right| \\
&= \left(1+\left|\frac{\tilde J(\tilde J-1)/H_u^2-2V\tilde J\tilde R/H_u+V^2\tilde R^2}{\tilde J(\tilde J-1)/H_l^2-2V\tilde J\tilde R/H_l+V^2\tilde R^2}\right|\right)\cdot\left|\left(\frac{\tilde J(\tilde J-1)}{H_l^2}-\frac{2V\tilde J\tilde R}{H_l}+V^2\tilde R^2\right)f_h(\xi)\right| \\
&\qquad + \left|\frac{V^2\tilde R^2/(\tilde J-1)}{\tilde J(\tilde J-1)/H_l^2-2V\tilde J\tilde R/H_l+V^2\tilde R^2}\right|\cdot\left|\left(\frac{\tilde J(\tilde J-1)}{H_l^2}-\frac{2V\tilde J\tilde R}{H_l}+V^2\tilde R^2\right)f_h(\xi)\right| \\
&\le (1+c'+c'')c_*\left|\left(\frac{\tilde J(\tilde J-1)}{H_l^2}-\frac{2V\tilde J\tilde R}{H_l}+V^2\tilde R^2\right)f_{h_*}(\xi)\right| \\
&= c_*(1+c'+c'')\left|\frac{\partial^2 f_{h_*}(\xi)}{\partial\eta^2}\right| \\
&= M'_{T+1}\left|\frac{\partial^2 f_{h_*}(\xi)}{\partial\eta^2}\right|,
\end{aligned}$$

where $M'_{T+1}=c_*(1+c'+c'')$. Here, c' and c'' satisfy

$$\left|\frac{\tilde J(\tilde J-1)/H_u^2-2V\tilde J\tilde R/H_u+V^2\tilde R^2}{\tilde J(\tilde J-1)/H_l^2-2V\tilde J\tilde R/H_l+V^2\tilde R^2}\right|\le c' \quad\text{and}\quad \left|\frac{V^2\tilde R^2/(\tilde J-1)}{\tilde J(\tilde J-1)/H_l^2-2V\tilde J\tilde R/H_l+V^2\tilde R^2}\right|\le c''$$

for all $\tilde J$ and $\tilde R$ in their range. The existence of c' and c'' is established through arguments similar to those used in the proof of B4.

Similarly, we can show that for $t=1,\dots,T$, there exists M'_t such that

$$\left|\frac{\partial^2 f_h(\xi)}{\partial\alpha_t^2}\right|\le M'_t\left|\frac{\partial^2 f_{h_*}(\xi)}{\partial\alpha_t^2}\right|.$$

Combining all the inequalities for the second-order partial derivatives, we conclude that $\|\nabla_h^2 f_h(\xi)\|_\infty \le c_{2*}\|\nabla_h^2 f_{h_*}(\xi)\|_\infty$ for all $\xi\in\Xi$, where

$$c_{2*}=\max\bigl\{c_*(1+M_t)(1+M_{t'}),\ M'_t,\ \text{for } t,t'=1,\dots,T+1 \text{ and } t\ne t'\bigr\}.$$

Proof of B2 Recall that $f_h(\xi)=\nu_{h\,|\,(\boldsymbol{w},\xi)}(h)/\nu_h(h)$ where $\nu_{h\,|\,(\boldsymbol{w},\xi)}(h)$ is the product of $T+1$ gammas (see the quantity inside the braces in (3.4)). We may express this function as $f_h(\xi)=$

$g_{T+1,\xi}(\eta)\prod_{t=1}^{T} g_{t,\xi}(\alpha_t)$, where

$$\begin{aligned} g_{T+1,\xi}(\eta) &= \frac{\Gamma(a)}{b^a}\frac{\big(b+V\tilde{R}\big)^{a+\tilde{J}}}{\Gamma\big(a+\tilde{J}\big)}\eta^{\tilde{J}}\exp\big(-V\tilde{R}\eta\big), \\ g_{t,\xi}(\alpha_t) &= \frac{\Gamma(a)}{b^a}\frac{\big(b+\tilde{Q}\big)^{a+\tilde{I}_t}}{\Gamma\big(a+\tilde{I}_t\big)}\alpha_t^{\tilde{I}_t}\exp\big(-\tilde{Q}\alpha_t\big), \qquad t=1,\dots,T. \end{aligned} \tag{S-4.7}$$

Our plan is to show that $\partial f_h(\xi)/\partial\eta$ and $\partial f_h(\xi)/\partial\alpha_t$ are uniformly bounded, and because the set of possible values of $\tilde{I}_t$ and $\tilde{J}$ is finite, it suffices to show that for fixed values of $\tilde{I}_t$ and $\tilde{J}$ these partial derivatives are uniformly bounded as $\tilde{Q}$ and $\tilde{R}$ vary. To find uniform bounds on $\partial f_h(\xi)/\partial\eta$ and $\partial f_h(\xi)/\partial\alpha_t$, we will make use of the inequalities for these partial derivatives in terms of the values of these partial derivatives at h_* which we obtained in the proof of B4. We have

$$g_{T+1,\xi}(H_l) = \frac{\Gamma(a)}{b^a}\frac{H_l^{\tilde{J}}}{\Gamma\big(a+\tilde{J}\big)}\frac{\big(b+V\tilde{R}\big)^{a+\tilde{J}}}{\exp\big(VH_l\tilde{R}\big)} \le M_{T+1}(\tilde{J}) < \infty \qquad \text{for all } \xi\in\Xi.$$

The existence of the finite upper bound $M_{T+1}(\tilde{J})$ is easy to establish and this is done as in the proof of B4. Similarly, for any $t=1,\dots,T$, there exists $M_t(\tilde{I}_t)<\infty$ such that

$$g_{t,\xi}(H_l) \le M_t(\tilde{I}_t) < \infty \qquad \text{for all } \xi\in\Xi. \tag{S-4.8}$$

We now consider the partial derivatives. We have

$$\begin{aligned} \left|\frac{\partial f_h(\xi)}{\partial\eta}\right| &\le c_*(1+M_{T+1})\left|\left(\frac{\tilde{J}}{H_l}-V\tilde{R}\right)f_{h_*}(\xi)\right| \\ &\le c_*(1+M_{T+1})\left|\frac{\Gamma(a)}{b^a}\frac{H_l^{\tilde{J}}}{\Gamma(a+\tilde{J})}\frac{\big(b+V\tilde{R}\big)^{a+\tilde{J}}\big(\tilde{J}/H_l-V\tilde{R}\big)}{\exp\big(VH_l\tilde{R}\big)}\right|\prod_{t=1}^{T}M_t(\tilde{I}_t) \\ &\le c_*(1+M_{T+1})M'_{T+1}(\tilde{J})\prod_{t=1}^{T}M_t(\tilde{I}_t), \end{aligned}$$

where the first inequality is from the second inequality in (S-4.4), the second inequality comes from (S-4.7) and (S-4.8), and the third inequality comes from the fact that there exists a finite $M'_{T+1}(\tilde{J})$ satisfying

$$\left|\frac{\Gamma(a)}{b^a}\frac{H_l^{\tilde{J}}}{\Gamma\big(a+\tilde{J}\big)}\frac{\big(b+V\tilde{R}\big)^{a+\tilde{J}}\big(\tilde{J}/H_l-V\tilde{R}\big)}{\exp\big(VH_l\tilde{R}\big)}\right| \le M'_{T+1}(\tilde{J}) \qquad \text{for all } \tilde{R}.$$

We conclude that there exists $\widetilde{M}_{T+1}>0$ such that $|\partial f_h(\xi)/\partial\eta| \le \widetilde{M}_{T+1}$ for all $\xi\in\Xi$. Similarly, we can prove that for any $t=1,\dots,T$, there exists $\widetilde{M}_t$ such that $|\partial f_h(\xi)/\partial\alpha_t| \le \widetilde{M}_t$ for all $\xi\in\Xi$. These last two statements immediately imply that B2 holds. □

5 Consistency of the Importance Weighted Marginal Density Estimator

Let $h^* \in \mathcal{H}$ be fixed, and let $\{\omega_{\boldsymbol{z}},\, \boldsymbol{z} \in \mathcal{Z}\}$ be a family of densities on $\mathcal{H}$. We do not require that this be the family of conditional densities corresponding to some joint distribution on $\mathcal{Z} \times \mathcal{H}$; we require only that for each $\boldsymbol{z} \in \mathcal{Z}$, $\omega_{\boldsymbol{z}}$ is a density on $\mathcal{H}$. The Markov chain based on HMC and the chain based on data augmentation both satisfy the conditions of Theorem 2 of Athreya et al. (1996). Consequently, for μ denoting the product of counting measure on $\mathcal{Z}$ and Lebesgue measure on $\mathcal{H}$, we have

$$\begin{aligned}
\hat{\nu}_{h\,|\,\boldsymbol{w}}(h^*) &\xrightarrow{\text{a.s.}} \int \omega_{\boldsymbol{z}}(h) \frac{\nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}(\boldsymbol{z}, h^*)}{\nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}(\boldsymbol{z}, h)} \nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}(\boldsymbol{z}, h)\, d\mu(\boldsymbol{z}, h) \\
&= \int \sum_{\boldsymbol{z}} \omega_{\boldsymbol{z}}(h) \frac{\nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}(\boldsymbol{z}, h^*)}{\nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}(\boldsymbol{z}, h)} \nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}(\boldsymbol{z}, h)\, dh \\
&= \sum_{\boldsymbol{z}} \biggl\{ \nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}(\boldsymbol{z}, h^*) \int \omega_{\boldsymbol{z}}(h)\, dh \biggr\} \\
&= \sum_{\boldsymbol{z}} \nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}(\boldsymbol{z}, h^*) \\
&= \nu_{h\,|\,\boldsymbol{w}}(h^*),
\end{aligned}$$

where the third equality is due to the fact that for each $\boldsymbol{z}$, $\omega_{\boldsymbol{z}}$ is a density on $\mathcal{H}$.

6 Proof of Theorem 3

The data augmentation algorithm gives a sequence $\lambda^{(1)}, \ldots, \lambda^{(n)}$, where the λ's have the general form $\lambda = (\boldsymbol{z}, \boldsymbol{I}, \boldsymbol{Q}, \boldsymbol{J}, \boldsymbol{R}, \tilde{h})$. Let ζ be the subvector of λ given by $\zeta = (\boldsymbol{z}, \tilde{h})$. Letting $f_h(\zeta) = \hat{\nu}_{h\,|\,\boldsymbol{w}}(h)/\hat{\nu}(h)$, where $\hat{\nu}_{h\,|\,\boldsymbol{w}}$ is now given by (3.5), we see that $\widehat{m}_n(h)$ is an average of the functions $f_h(\zeta^{(k)})$, $k = 1, \ldots, n$. As in the proof of Theorem 2, we need to check Conditions B2–B5 and, again, we will first establish B3–B5 (with $d = 1$), and then deal with B2.

Proof of B3 We first express $f_h(\zeta)$ as $f_h(\zeta) = C(\zeta) g_{\boldsymbol{z}}(h)$, where $C(\zeta) = \omega_{\boldsymbol{z}}(\tilde{h})/\nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}(\boldsymbol{z}, \tilde{h})$ and $g_{\boldsymbol{z}}(h) = \nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}(\boldsymbol{z}, h)/\nu_h(h)$, so that by (2.2), $g_{\boldsymbol{z}}(h)$ is given by

$$\begin{aligned}
g_{\boldsymbol{z}}(h) = &\left[\prod_{d=1}^{D} \frac{\Gamma\bigl(\sum_{t=1}^{T} \alpha_t\bigr)}{\Gamma\bigl(n_d + \sum_{t=1}^{T} \alpha_t\bigr)}\right] \left[\prod_{d=1}^{D} \prod_{t=1}^{T} \frac{\Gamma(n_{dt} + \alpha_t)}{\Gamma(\alpha_t)}\right] \left[\prod_{t=1}^{T} \frac{\Gamma(V\eta)}{\Gamma(m_{\cdot t \cdot} + V\eta)}\right] \\
&\times \left[\prod_{t=1}^{T} \prod_{v=1}^{V} \frac{\Gamma(m_{\cdot tv} + \eta)}{\Gamma(\eta)}\right].
\end{aligned} \tag{S-6.1}$$

We then consider the component-wise monotonicity in $h = (\alpha_1, \ldots, \alpha_T, \eta)$ of the functions inside the brackets. We have

$$\prod_{d=1}^{D} \frac{\Gamma\big(\sum_{t=1}^{T} \alpha_t\big)}{\Gamma\big(n_d + \sum_{t=1}^{T} \alpha_t\big)} = \prod_{d=1}^{D} \frac{1}{\big(\sum_{t=1}^{T} \alpha_t + n_d - 1\big) \cdots \big(\sum_{t=1}^{T} \alpha_t\big)},$$

and it is easy to see that for each $t = 1, \ldots, T$, this product is monotonically decreasing in α_t with all the other variables fixed. Also,

$$\prod_{d=1}^{D} \prod_{t=1}^{T} \frac{\Gamma(n_{dt} + \alpha_t)}{\Gamma(\alpha_t)} = \prod_{d=1}^{D} \prod_{t=1}^{T} (\alpha_t + n_{dt} - 1) \cdots \alpha_t,$$

and for each $t = 1, \ldots, T$, this product is monotonically increasing in α_t with all the other variables fixed. For the functions involving η the results are analogous. Recall that $\mathcal{H}$ is assumed compact, and without loss of generality has the form $\mathcal{H} = [H_l, H_u]^{T+1}$ with $0 < H_l < H_u$. Also, redefine h_* by $h_* = (H_u, \ldots, H_u)$. We then have

$$\begin{aligned} f_h(\zeta) \leq C(\zeta) &\left[\prod_{d=1}^{D} \frac{\Gamma\big(TH_l\big)}{\Gamma\big(n_d + TH_l\big)}\right] \left[\prod_{d=1}^{D} \prod_{t=1}^{T} \frac{\Gamma(n_{dt} + H_u)}{\Gamma(H_u)}\right] \left[\prod_{t=1}^{T} \frac{\Gamma(VH_l)}{\Gamma(m_{\cdot t \cdot} + VH_l)}\right] \\ &\left[\prod_{t=1}^{T} \prod_{v=1}^{V} \frac{\Gamma(m_{\cdot tv} + H_u)}{\Gamma(H_u)}\right] \end{aligned}$$

$$= f_{h_*}(\zeta) \left[\prod_{d=1}^{D} \frac{\Gamma(n_d + TH_u)}{\Gamma(n_d + TH_l)} \frac{\Gamma(TH_l)}{\Gamma(TH_u)}\right] \left[\prod_{t=1}^{T} \frac{\Gamma(m_{\cdot t \cdot} + VH_u)}{\Gamma(m_{\cdot t \cdot} + VH_l)} \frac{\Gamma(VH_l)}{\Gamma(VH_u)}\right] \tag{S-6.2a}$$

$$\leq f_{h_*}(\zeta) \left[\frac{\Gamma(N + TH_u)}{\Gamma(TH_l)}\right]^D \left[\frac{\Gamma(N + VH_u)}{\Gamma(VH_l)}\right]^T \tag{S-6.2b}$$

$$= c_* f_{h_*}(\zeta), \tag{S-6.2c}$$

where (S-6.2a) comes from the definition of h_* given above, (S-6.2b) results from the inequalities $0 \leq n_d \leq N$, $d = 1, \ldots, D$, and $0 \leq m_{\cdot t \cdot} \leq N$, $t = 1, \ldots, T$, and in (S-6.2c) c_* is defined by $c_* = \big[\Gamma(N + TH_u)/\Gamma(TH_l)\big]^D \big[\Gamma(N + VH_u)/\Gamma(VH_l)\big]^T$. This gives B3.

Proof of B4, B5, and B2 The hyperparameter h is given by $h = (\alpha_1, \ldots, \alpha_T, \eta)$, but for convenience, we will temporarily denote it by $h = (h_1, \ldots, h_{T+1})$. The function $g_{\boldsymbol{z}}(\cdot)$, defined in (S-6.1), is a product of ratios of polynomials in h, and over the compact domain of h, these ratios are bounded above and are non-zero. Consequently, for any $\boldsymbol{z} \in \mathcal{Z}$ and any $t = 1, \ldots, T+1$, $\partial g_{\boldsymbol{z}}(h)/\partial h_t$ is continuous in h. Therefore, there exists $h_{\boldsymbol{z}^{(t)}}$ that maximizes $|\partial g_{\boldsymbol{z}}(h)/\partial h_t|$. Because $\mathcal{Z}$ is finite, there exists $\boldsymbol{z}^{(*t)}$ and a corresponding $h_{\boldsymbol{z}^{(*t)}}$ such that $|\partial g_{\boldsymbol{z}^{(*t)}}(h_{\boldsymbol{z}^{(*t)}})/\partial h_t)| =$

$\max_{\boldsymbol{z}} |\partial g_{\boldsymbol{z}}(h_{\boldsymbol{z}^{(t)}})/\partial h_t|$. Define $M_u^{(t)} = |\partial g_{\boldsymbol{z}^{(*t)}}(h_{\boldsymbol{z}^{(*t)}})/\partial h_t|$. We then have

$$\left|\frac{\partial g_{\boldsymbol{z}}(h)}{\partial h_t}\right| \leq \left|\frac{\partial g_{\boldsymbol{z}}(h_{\boldsymbol{z}})}{\partial h_t}\right| \leq \frac{\partial g_{\boldsymbol{z}^{(t)}}(h_{(t)})}{\partial h_t} = M_u^{(t)} \qquad \text{for all } \boldsymbol{z} \in \mathcal{Z}, h \in \mathcal{H}.$$

Using a similar argument, we can show that there exists $M_l^{(t)} > 0$ such that

$$M_l^{(t)} \leq \left|\frac{\partial g_{\boldsymbol{z}}(h)}{\partial h_t}\right| \qquad \text{for all } \boldsymbol{z} \in \mathcal{Z}, h \in \mathcal{H}.$$

We now consider $f_h(\zeta)$. Choose an arbitrary $h_0 \in \mathcal{H}$. For any $t = 1, \ldots, T+1$, we have

$$\left|\frac{\partial f_h(\zeta)}{\partial h_t}\right| = \left|C(\zeta)\frac{\partial g_{\boldsymbol{z}}(h)}{\partial h_t}\right| \leq C(\zeta) M_u^{(t)} \leq C(\zeta)\frac{M_u^{(t)}}{M_l^{(t)}}\left|\frac{\partial g_{\boldsymbol{z}}(h_0)}{\partial h_t}\right| = \frac{M_u^{(t)}}{M_l^{(t)}}\left|\frac{\partial f_{h_0}(\zeta)}{\partial h_t}\right|. \tag{S-6.3}$$

We conclude that

$$\|\nabla_h f_h(\zeta)\|_\infty \leq M \|\nabla_h f_{h_0}(\zeta)\|_\infty,$$

where $M = \max\bigl\{M_u^{(t)}/M_l^{(t)},\, t = 1, \ldots, T+1\bigr\}$. This proves B4.

The proof of B5 is very similar and is omitted. Condition B2 is trivially satisfied because by (S-6.3), $\partial f_h(\zeta)/\partial h_t$ has a uniform upper bound. □

7 Feasibility for Large Corpora

Here we describe an experiment involving a large number of documents. The experiment has only one purpose, namely to demonstrate that our method can handle large corpora. It does not do a comparison with other methods, because a thorough comparison has already been done in Sections 4.2.2–4.2.4, and also because it is not possible to compare our method with ISST, since ISST does not easily handle corpora of the size we use.

In the experiment, we generated artificial corpora from LDA models with configuration parameters set as follows. All of them had $D = 10^5$, $T = 50$, $V = 3{,}000$, and $n_d = 200$ for each document in the corpus, and we used the nine values of h_{true} given in the first row of Table S-1. For each corpus, we applied our method, implemented via HMC, and formed $\hat{\hat{h}}$. Our goal was to demonstrate that $\hat{\hat{h}}$ does a good job of estimating $\hat{h} = \arg\max_h m(h)$. We have here the subtle point mentioned before, that $\hat{h}$ is unknown. So instead, we show that $\hat{\hat{h}}$ does a good job of estimating h_{true}, the rationale for this being that $\hat{h}$ is the maximum likelihood estimator of h_{true}, and when D is large, $\hat{h}$ is very close to h_{true}. Table S-1 gives the results. The second row shows that $\hat{\hat{h}}$ is very close to h_{true} in all cases, which shows that our method is working well. The third row gives the execution times (in hours) for our method. These are long, but they can be massively reduced if we use parallelization (see Section 5 of Xia and Doss (2018)). (Interestingly, applying our procedure to a random subsample of only 5% or 10% of the documents in each corpus produces essentially the same estimates, so the execution times can also be reduced in that way.)

In our experiments we used synthetic corpora, as opposed to real corpora, because for synthetic corpora we know the true hyperparameter, i.e. the hyperparameter value under which the corpus is generated, so there is no ambiguity in reporting and interpreting our results.

$(\alpha_{\text{true}}, \eta_{\text{true}})$	$(.2, .2)$	$(.2, .7)$	$(.7, .2)$	$(.4, .4)$	$(.7, .7)$	$(.5, .5)$	$(1.0, .5)$	$(1.0, 1.0)$	$(1.0, 1.5)$
$(\hat{\alpha}, \hat{\eta})$	$(.2, .2)$	$(.2, .7)$	$(.7, .2)$	$(.4, .4)$	$(.65, .66)$	$(.5, .51)$	$(.95, .46)$	$(.95, .96)$	$(.95, 1.46)$
Time (h)	55.9	53.7	54.1	53.8	53.6	50.27	54.6	54.8	53.9

Table S-1: Performance of the FBEB method on large synthetic corpora ($D = 10^5$) generated under nine different values of the hyperparameter. The method does an excellent job of estimating h_{true}, and hence also of estimating $\hat{h}$, in all nine cases.

References

Athreya, K. B., Doss, H. and Sethuraman, J. (1996). On the convergence of the Markov chain simulation method. *The Annals of Statistics* **24** 69–100.

Doss, H. and Park, Y. (2018). An MCMC approach to empirical Bayes inference and Bayesian sensitivity analysis via empirical processes. *The Annals of Statistics* **46** 1630–1663.

Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society,* Series B **73** 123–214.

Hoffman, M. D. and Gelman, A. (2014). The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15** 1593–1623.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo* (S. P. Brooks, A. Gelman, G. L. Jones and X.-L. Meng, eds.). CRC Press, Boca Raton, 113–162.

Xia, W. and Doss, H. (2018). Scalable hyperparameter selection for latent Dirichlet allocation. Tech. rep., Department of Statistics, University of Florida.