A Appendix: convergence analysis

Here we prove that, provided we start with a well-defined Markov chain, exact asynchronous Gibbs sampling will converge to the correct target distribution. Note first that asynchronous versions of valid MCMC algorithms for 1 and 2-dimensional target distributions can be proven to always converge, because the random variables representing states of the algorithm can always be re-ordered to recover the Markov property – see Terenin and Xing (2017) for details.

Our strategy has two parts. First, we define a serialized parallel MCMC algorithm that formalizes the way in which workers draw samples and communicate with one another under the assumption that communication is instantaneous, using ideas inspired by the coupling of chains in *parallel tempering* (Swendsen and Wang, 1986). Then, we note that MCMC methods belong to the class of *fixed-point* algorithms, and hence we can use a result from the asynchronous convergence of these algorithms, due to Baudet (1978) and Bertsekas (1983), to prove that the asynchronous version of the parallel algorithm with non-instantaneous communication converges as well. We begin by defining the MCMC algorithm that we wish to parallelize.

Definition 5 (Preliminaries)**.** *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. Let X be a Polish space, and let $\mathscr{X}$ be its Borel σ-algebra. Let $\mathcal{M}_s(X)$ be the Banach space of signed measures on X, equipped with the total variation norm $\|\cdot\|_{\text{TV}}$. Let $\mathcal{M}_1(X) \subset \mathcal{M}_s(X)$ be the space of probability measures over X. Let $\pi \in \mathcal{M}_1(X)$ be the target measure.*

Definition 6 (Underlying chain)**.** *Let $k \in \mathbb{N}$. Define a Markov chain $\xi : \Omega \times \mathbb{N} \to X$, $(\omega, k) \mapsto \xi^k(\omega)$. For all $B \in \mathscr{X}$ and all $k \in \mathbb{N}$, let the regular conditional probability measure $P : \mathscr{X} \times X \to \mathbb{R}$ defined by $P(B \mid \xi) = \mathbb{P}(\xi^{k+1} \in B \mid \xi^k = \xi)$ be its transition kernel. This is well-defined, as the latter expression does not depend on k by the Markov property and time-homogeneity. Note that by definition of a regular conditional probability measure, for all $B \in \mathscr{X}$ the map $\xi \mapsto P(B \mid \xi)$ is $(X, \mathscr{X})$-measurable, and for all $\xi \in X$ the map $B \mapsto P(B \mid \xi)$ is a probability measure. Define the Markov operator $P : \mathcal{M}_1(X) \to \mathcal{M}_1(X)$ by $(P(\mu))(B) = \int_X P(B \mid \xi) \mathrm{d}\mu(\xi)$. Assume that for all $\mu \in \mathcal{M}_1(X)$, we have that $\left\|P^k(\mu) - \pi\right\|_{\text{TV}} \to 0$ as $k \to \infty$. We say that ξ^k the* underlying chain*.*

Here, X is the parameter space for the given problem, μ is the initial measure, π is the target measure, k is the current iteration of the chain, and P is the Markov operator for the chain we wish to parallelize, which we assume converges to π in total variation. Our analysis will center on the relationships between the workers' Markov chains, and we now introduce the definitions needed to consider this formally.

Definition 7 (Instantaneous parallel chain)**.** *Let $m \in \mathbb{N}$. Let $\mathcal{X} = \bigtimes_{i=1}^m X$, equipped with its product σ-algebra. Let L be any index set, and let $\{\xi_{(l)}^k : l \in L\}$ be a set of underlying chains. Let $x : \Omega \times \mathbb{N} \to \mathcal{X}$, $(\omega, k) \mapsto x^k(\omega)$ be a Markov chain such that for any $x \in \mathcal{X}$ and any $i \in \{1, .., m\}$ there exists an $l \in L$ such that for any $B \in \mathscr{X}$ we have $\mathbb{P}(x_i^{k+1} \in B \mid x^k = x) = \mathbb{P}(\xi_{(l)}^{k+1} \in B \mid \xi_{(l)}^k = x_i)$. We say that x^k is the* instantaneous parallel chain*.*

Here, $\mathcal{X}$ is the state space for the entire compute cluster's computation. Since we have assumed temporarily that communication is instantaneous, this means that the entire cluster's computation is also a Markov chain. We assume that this much larger Markov chain is made up of individual components representing the worker nodes. We also assume that each worker node performs a Markov update based on two components, namely its previous state $x_i \in X$, and a choice of proposal distribution indexed by a parameter $l \in L$, whose value can depend on the state of other workers.

Example 8 (Instantaneous parallel Gibbs sampler)**.** *Take $X = \mathbb{R}^d$ and $\mathcal{X} = \mathbb{R}^{d \times m}$. For all $i \in \{1, .., m\}$, let $C_i \subseteq \{1, .., d\}$ such that $\bigcup_{i=1}^m C_i = \{1, .., d\}$. Assume that π admits an absolutely continuous density f with respect to the Lebesgue measure. Let $\mathbf{X}$ be a Markov chain defined on $\mathbb{R}^{d \times m}$ as follows.*

1. *Select an index $s \in \{1, .., m\}$ uniformly at random.*
2. *Select a coordinate $j \in C_s$ uniformly at random.*
3. *Randomly draw x'_{sj} from $f(x_{sj} \mid \boldsymbol{x}_{s,-j})$.*
4. *For all i, set x_{ij} at the next iteration to x'_{sj} with probability*

$$\alpha_i = \min\left\{1, \frac{f(x'_{sj}, \boldsymbol{x}_{i,-j})\, f(x_{ij} \mid \boldsymbol{x}_{s,-j})}{f(x_{ij}, \boldsymbol{x}_{i,-j})\, f(x'_{sj} \mid \boldsymbol{x}_{s,-j})}\right\} \tag{18}$$

and set it to x_{ij} otherwise.

This chain describes how exact asynchronous Gibbs sampling would behave under instantaneous communication, with no asynchronous delays and all messages sent and received. It selects a worker at random and proposes from that worker's full conditional at every worker. Note that $\alpha_{i'} = 1$, because on worker whose full conditional is selected, the proposal is exactly a Gibbs step and is hence always accepted.

It is easily seen that this is, in fact, an instantaneous parallel chain, because at every iteration it performs a Metropolis-Hastings transition with respect to some proposal distribution determined by the current state of another worker. Thus, we can take L to be the set of all such proposal distributions.

At this stage, we don't yet know anything about the stationarity properties of the instantaneous parallel chain, due to the expanded state space. Indeed, depending on how parameters are partitioned and the details of how workers communicate, which at this stage have been abstracted out of the problem, this chain can be reducible, making stationarity analysis non-trivial. We would therefore like to avoid speaking about the joint distribution of the chain altogether, and instead only study the marginal distributions at every worker. To do this, we introduce a notion of coupling.

Definition 9 (Marginally coupled Markov operator)**.** *Let x^k be an instantaneous parallel chain. Let $E = \bigtimes_{i=1}^m \mathcal{M}_1(X)$. For $\varepsilon, \varpi \in E$, define the metric $d(\varepsilon, \varpi) = \sum_{i=1}^m ||\varepsilon_i - \varpi_i||_{\mathrm{TV}}$. For all $i \in \{1, .., m\}$ and all $B, B_i \in \mathscr{X}$, define the map*

$$H_i : \mathscr{X} \times \mathcal{X} \to \mathbb{R} \qquad H_i(B \mid x_1, .., x_m) = \mathbb{P}(x_i^{k+1} \in B \mid x_1^k = x_1, .., x_m^k = x_m) \tag{19}$$

and the operator

$$H : E \to E \qquad (H_i(\varepsilon))(B_i) = \int_\Omega .. \int_\Omega H_i(B_i \mid x_1, .., x_m) \mathrm{d}\varepsilon_1(x_1) .. \mathrm{d}\varepsilon_m(x_m). \tag{20}$$

Call H the marginally coupled Markov operator *for the instantaneous parallel chain.*

Here, E is a space in which each $\varepsilon \in E$ represents the distributional state of the entire cluster. The marginally coupled Markov operator H – analogous to the underlying chain's Markov operator P – captures how the cluster transitions from one state to the next probabilistically, while only tracking marginal distributions rather than the full joint. This means that we are only analyzing whether each worker converges to the target distribution, and ignoring any dependence between workers. To continue, we need an assumption.

Assumption 10 (Simultaneous worker-wise contraction)**.** *Let $\mu \in \mathcal{M}_1(X)$. Consider the marginally coupled Markov transition kernel H_i. Recall that for any fixed set of values $x_{1:m}^{-i} = \{x_j : j \in \{1, .., m\}, j \neq i\}$, by Definition 7 there exists an $l \in L$ such that H_i is the Markov transition kernel of an underlying chain $\xi_{(l)}^k$. Let $P_{(l)}$ be the Markov operator of that chain. Assume that for all l and all i there exists a $\rho < 1$ such that*

$$\left\| P_{(l)}(\mu) - \pi \right\|_{\mathrm{TV}} \leq \rho \|\mu - \pi\|_{\mathrm{TV}}. \tag{21}$$

This is a condition on how quickly each worker's chain converges to the target posterior with respect to the behavior of the other workers in the cluster. It says that, regardless of what the other workers are doing, no worker can proceed at an arbitrarily slow rate of convergence. We use the term *simultaneous* to emphasize that uniformity is only required with respect to workers, rather than other quantities such as initial conditions of the chain, as is typical in uniform ergodicity and related conditions. It is through this assumption that properties of the communication scheme, such as how frequently workers transmit their messages to one another, enter the theory. Whether or not the assumption will hold for a given Gibbs sampler will depend on properties of the target distribution. Note that at this stage, communication is still instantaneous, and asynchronicity properties such as message delays do not yet enter the theory. These will be considered later.

Proposition 11 (Coupled convergence)**.** *Let $\Pi = \bigtimes_{i=1}^m \pi \in E$. For any instantaneous parallel chain, we have that $H(\Pi) = \Pi$. Furthermore for all $\varepsilon \in E$ and all $i \in \{1, .., m\}$, the function $||H_i^k(\varepsilon) - \pi||_{\mathrm{TV}}$ is non-increasing in k, and we have*

$$d(H^k(\varepsilon), \Pi) \to 0 \tag{22}$$

as $k \to \infty$.

Proof. By additivity of d, it suffices to show that all three claims hold for each H_i, so fix an arbitrary $i \in \{1, .., m\}$. We have for all $B \in \mathscr{X}$ that

$$(H_i(\Pi))(B) = \int_\Omega .. \int_\Omega H_i(B \mid x_1, .., x_m) \mathrm{d}\pi(x_1) .. \mathrm{d}\pi(x_m). \tag{23}$$

Since H_i is non-negative, we may use Tonelli's Theorem to switch the order of integration so that x_i is the inner-most component being integrated. We then have

$$\int_\Omega H_i(B \mid x_1, .., x_m) \mathrm{d}\pi(x_i) = \pi(B) \tag{24}$$

because for all $x_1, .., x_m$ except x_i, there exists an $l \in L$ and an underlying chain $\xi_{(l)}^k$ with Markov operator $P_{(l)}$ for which we have $H_i = P_{(l)}$. This gives the first claim. Next, we check that $||H_i^k(\varepsilon) - \pi||_{\mathrm{TV}}$ is non-increasing in k, as well as convergence. We have that

$$\begin{aligned}
d(H(\varepsilon), \Pi) &= \sum_{i=1}^m \left\| \int_\Omega .. \int_\Omega H_i(\cdot \mid x_1, .., x_m) \mathrm{d}\varepsilon_1(x_1) .. \mathrm{d}\varepsilon_m(x_m) - \pi \right\|_{\mathrm{TV}} && (25)\\
&= \sum_{i=1}^m \left\| \int_\Omega .. \int_\Omega H_i(\cdot \mid x_1, .., x_m) \mathrm{d}\varepsilon_i(x_i) \underbrace{\mathrm{d}\varepsilon_1(x_1) .. \mathrm{d}\varepsilon_m(x_m)}_{\text{except } \mathrm{d}\varepsilon_i(x_i)} - \int_\Omega .. \int_\Omega \pi \underbrace{\mathrm{d}\varepsilon_1(x_1) .. \mathrm{d}\varepsilon_m(x_m)}_{\text{except } \mathrm{d}\varepsilon_i(x_i)} \right\|_{\mathrm{TV}} && (26)\\
&= \sum_{i=1}^m \left\| \int_\Omega .. \int_\Omega P_{(l)}(\varepsilon_i) - \pi \underbrace{\mathrm{d}\varepsilon_1(x_1) .. \mathrm{d}\varepsilon_m(x_m)}_{\text{except } \mathrm{d}\varepsilon_i(x_i)} \right\|_{\mathrm{TV}} && (27)\\
&\leq \sum_{i=1}^m \int_\Omega .. \int_\Omega \left\| P_{(l)}(\varepsilon_i) - \pi \right\|_{\mathrm{TV}} \underbrace{\mathrm{d}\varepsilon_1(x_1) .. \mathrm{d}\varepsilon_m(x_m)}_{\text{except } \mathrm{d}\varepsilon_i(x_i)} && (28)\\
&< \sum_{i=1}^m \int_\Omega .. \int_\Omega \rho \|\varepsilon_i - \pi\|_{\mathrm{TV}} \underbrace{\mathrm{d}\varepsilon_1(x_1) .. \mathrm{d}\varepsilon_m(x_m)}_{\text{except } \mathrm{d}\varepsilon_i(x_i)} && (29)\\
&= \sum_{i=1}^m \rho \|\varepsilon_i - \pi\|_{\mathrm{TV}} \int_\Omega .. \int_\Omega \underbrace{\mathrm{d}\varepsilon_1(x_1) .. \mathrm{d}\varepsilon_m(x_m)}_{\text{except } \mathrm{d}\varepsilon_i(x_i)} && (30)\\
&= \rho \, d(\varepsilon, \Pi). && (31)
\end{aligned}$$

Here, the second line follows from Tonelli's Theorem since H_i is non-negative, and since the integral of each ε_j is equal to 1. The third line follows by linearity and the definition of $P_{(l)}$ in Assumption 10. The fourth line follows from definition of $||\cdot||_{\mathrm{TV}}$, because the supremum of an integral is less than the integral of the supremum. The fifth line follows from Assumption 10. The sixth line follows because each ε_i is a probability measure and thus integrates to one. The last line follows by definition. Since $\rho < 1$, convergence follows from the Banach fixed point theorem. □

The set of Markov kernels $\{P_{(l)}, l \in L\}$, whose properties underly the above analysis, can be viewed as an *adaptive MCMC* algorithm. From this perspective, the first part of our argument is similar to Proposition 1 of Roberts and Rosenthal (2007), and the second part is similar to their Theorem 5, where our Assumption 10 is similar to their condition (a).

We now move to the second stage of the proof. From here, we want to show that H converges asynchronously, i.e., convergence is still valid in the setting in which each worker does not necessarily know the precise current state of all other workers, and instead works with the latest state that it knows about. We begin by stating the Baudet (1978), Bertsekas (1983), and Frommer and Szyld (2000) model of distributed computation, within which we base our analysis.

Definition 12 (Asynchronous computation)**.** *Start with the following fixed-point computation problem.*

(P1) Let $E = \bigtimes_{i=1}^m E_i$ be a product space, where i indexes workers. We take $E_i = \mathcal{M}_1(X)$.

(P2) Let $H : E \to E$ be a function with components H_i.

(P3) Let $\Pi = H(\Pi)$ be a fixed point of H.

Now, define the following cluster computation model:

- *Let $k \in \mathbb{N}_0$ be the total number of iterations performed by all workers.*
- *Let $s_i(k) \in \mathbb{N}_0$ be the total number of iterations on component i by all workers.*
- *Let I_k be an index set containing the components updated at iteration k.*

Next, assume the following basic regularity conditions on the cluster:

(R1) No worker's state is based on future values: $s_i(k) \leq k-1$.

(R2) No worker stops permanently: $\lim_{k\to\infty} s_i(k) = \infty$.

(R3) No component stops being updated or communicated by workers: $|\{k \in \mathbb{N} : i \in I^k\}| = \infty$.

Finally, define ε^k component-wise via the following:

$$\varepsilon_i^k = \begin{cases} H_i\big(\varepsilon_1^{s_1(k)}, .., \varepsilon_m^{s_m(k)}\big) \text{ if } i \in I^k, \\ \varepsilon_i^{k-1} \text{ otherwise.} \end{cases} \tag{32}$$

Then ε^k is termed an asynchronous iteration*, and $\{\varepsilon^k : k \in \mathbb{N}_0\}$ is termed an* asynchronous computation*.*

Definition 12 is broad enough to encompass most asynchronous computations, and it is at this stage that properties such as message delay enter the theory. With this computational model in mind, the following general theorem gives a sufficient set of conditions under which the asynchronous iterates ε^k converge to the correct answer.

Result 13 (Convergence of asynchronous computations)**.** *Given a well-defined asynchronous computation as in Definition 12, assume the following conditions hold for all $k \in \mathbb{N}_0$:*

(C1) There are sets $E^k \subseteq E$ satisfying $E^k = \bigtimes_{i=1}^m E_i^k$ (box condition)*.*

(C2) For E^k in (C1), $H(E^k) \subseteq E^{k+1} \subseteq E^k$ (nested sets condition)*.*

(C3) There exists a Π such that $\varepsilon \in E^k \implies \varepsilon \to \Pi$ in some metric (synchronous convergence condition)*.*

Then $\varepsilon^k \to \Pi$ in the same metric.

Proof. Baudet (1978), Bertsekas (1983), and Frommer and Szyld (2000). □

For MCMC, the main challenge in using this result is that an arbitrary measure space is not a product space – to avoid this, we instead work with Definition 9. We now proceed to verify its conditions.

Lemma 14 (Box condition)**.** *Fix the initial distribution $\varepsilon \in E$. Define the following:*

$$E^k = \big\{\varpi \in E : ||\varpi_i - \pi||_{\mathrm{TV}} \leq ||H_i^k(\varepsilon) - \pi||_{\mathrm{TV}} \text{ for all } i \in \{1, .., m\}\big\}. \tag{33}$$

Then there exist sets E_i^k such that $E^k = \bigtimes_{i=1}^m E_i^k$.

Proof. Take $E_i^k = \{\mu \in \mathcal{M}_1(X) : ||\mu - \pi||_{\mathrm{TV}} \leq ||H_i^k(\varepsilon) - \pi||_{\mathrm{TV}}\}$. □

Lemma 15 (Nested sets condition)**.** *Let E^k be defined as in the previous lemma. Then $H(E^k) \subseteq E^{k+1} \subseteq E^k$.*

Proof. By Proposition 11, $||H_i^k(\varepsilon) - \pi||_{\mathrm{TV}}$ is non-increasing in k for each i, so $E^{k+1} \subseteq E^k$, and $E^{k+1} = H(E^k)$ by construction. □

Theorem 16 (Asynchronous convergence)**.** *Asynchronous Markov chains in the sense of Definition 7 and Definition 12 satisfying Assumption 10 converge to π on each worker in total variation.*

Proof. We verify that all of the conditions required in Result 13 hold.

(P1–P3) Take E, H, Π as in Definition 7.

(R1–R3) All satisfied by assumption.

(C1–C3) Satisfied by Lemma 14, Lemma 15, and Proposition 11.

The claim follows. □

B Appendix: details of Gibbs sampler and approximate analytic matrix inversion in the Gaussian process example

We propose the following scheme to sample from the posterior of $(\mu, \sigma^2, \tau^2, \boldsymbol{\theta})$. We update individual slices of $\boldsymbol{\theta}$, consisting of 500 elements, via Gibbs steps. To do this, we sample from full conditional distributions of the form $\boldsymbol{\theta}_{1:500} \mid \boldsymbol{\theta}_{501:n}, \mu, \sigma^2, \tau^2$ for arbitrary blocks of 500 adjacent indices – recall that ϕ is fixed. Thus we need to sample from conditional Gaussian distributions of portions of $\boldsymbol{\theta}$, given the rest of $\boldsymbol{\theta}$. To do this without ever constructing the large covariance matrix, which may be too big to store in memory, we need to be able to invert $\mathbf{K}$, add $\sigma^{-2}\,\mathbf{I}_n$, and invert back. The following scheme allows us to do this element-wise, with only one approximate inversion along the way, which can with further work likely be refined into an exact inversion.

Since we have made the simplifying assumption that our grid is evenly spaced, the normalized covariance matrix $\tau^{-2}\mathbf{K}$ is Toeplitz. Additionally, since our covariance function is exponential, the resulting covariance matrix is hyperbolic, and can be inverted element-wise analytically via a technique due to Dow (2003), with inverse that simplifies to

$$\tau^2\mathbf{K}^{-1} = \begin{bmatrix} d_0 & a & 0 & \dots & \dots & \dots & 0 \\ a & b & a & 0 & \ddots & \ddots & \vdots \\ 0 & a & b & \ddots & \ddots & \ddots & \vdots \\ \vdots & 0 & \ddots & \ddots & \ddots & 0 & \ddots \\ \vdots & \ddots & \ddots & \ddots & b & a & 0 \\ \vdots & \ddots & \ddots & 0 & a & b & a \\ 0 & \dots & \dots & \dots & 0 & a & d_0 \end{bmatrix} \qquad \begin{cases} b = -\coth(-\phi\rho) \\ a = \dfrac{\operatorname{csch}(-\phi\rho)}{2} \\ d_0 = \dfrac{e^{-\phi\rho(2N-3)}\operatorname{csch}(-\phi\rho) + 1 - \coth(-\phi\rho)}{2 - 2e^{-\phi\rho(2N-3)}} \\ \rho = \text{grid spacing size} = 0.06 \\ N = \text{dimension of } \mathbf{K}. \end{cases} \tag{34}$$

Note that this $\mathbf{K}^{-1}$ is tridiagonal with modified corner elements. While this technique limits the generality of our Gaussian process prior, more complicated ways of avoiding large matrix inversions are available with modern spatial priors such as nearest neighbor Gaussian processes (Banerjee et al., 2012). If we had not fixed ϕ, we would have needed to compute a large matrix expression involving $\mathbf{K}^{-1}$ in its entirety for every sample of τ^2 and μ. Here, this is tractable, but we opted to avoid it for simplicity.

After we add σ^{-2} to the diagonal, the resulting covariance matrix is still tridiagonal with modified corner elements. We do not know how to invert this matrix analytically, but we do know how to invert the general tridiagonal Toeplitz matrix without modified corner elements, via a technique due to Hu and O'Connell (1996). We approximate the tridiagonal form by assuming that $d_0 = b$ in Equation (34) – this works well except at the points where the partition slices of $\boldsymbol{\theta}$ join, where a small amount of error is introduced.

Finally, to find the mean vector, we need to multiply the covariance matrix defined by Equation (34) by a term that includes the full data. This multiplication can be carried out to arbitrary precision by simply taking a slice in the center of the matrix in a neighborhood around the full conditional of interest, avoiding use of the full data. This idea also underlies *covariance tapering* (Furrer et al., 2006) and *composite likelihood methods* for spatial problems (Stein et al., 2004). After all of these steps, we can sample any slice of $\boldsymbol{\theta}$ full conditionally via the standard Schur complement formula, since the full conditional of a Gaussian is Gaussian.

C Appendix: trace plots for hierarchical mixed-effects model of Section 3.2

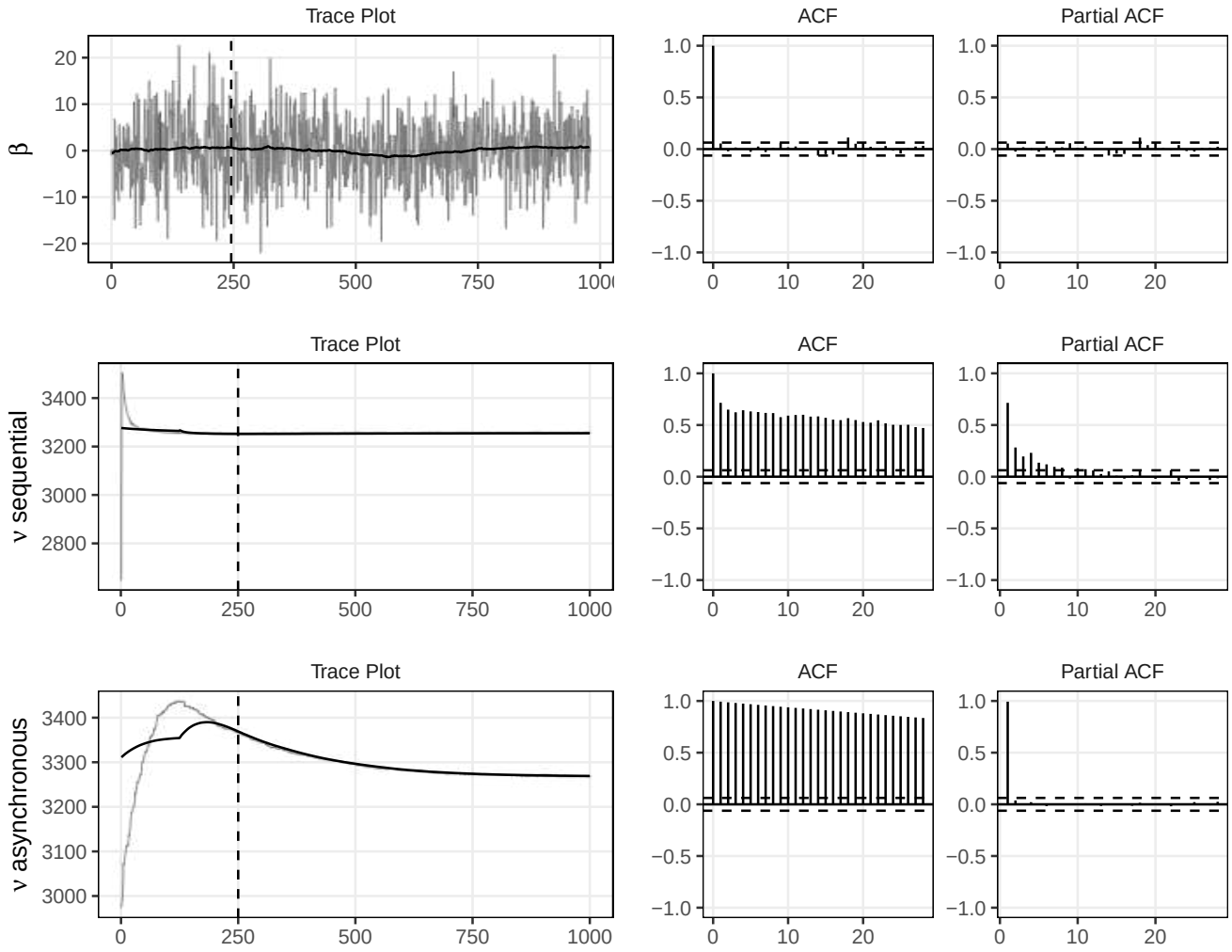

Figure 6: Trace plots and autocorrelation plots for an unspecified $\boldsymbol{\beta}_i$ component for the asynchronous Gibbs sampler, and of ν for the asynchronous and sequential-scan Gibbs samplers in Section 3.2.