Unbiased Markov chain Monte Carlo with couplings

Pierre E. Jacob[*] John O'Leary[†] Yves F. Atchadé[‡]

Supplementary Materials

This document contains the proofs of the results of the article, numerical experiments with coupled Hamiltonian Monte Carlo on a Normal target of varying dimension, Gibbs sampling on baseball batting average data, Pólya-Gamma Gibbs sampling on the German credit data, and Bayesian Lasso Gibbs sampling on the diabetes data set.

S1 Proofs

S1.1 Proof of Proposition 3.1

We present the proof for $H_0(X,Y)$, a similar reasoning holds for $H_k(X,Y)$ and $H_{k:m}(X,Y)$. We follow the same arguments as in Glynn and Rhee [2014], Vihola [2017]. To study $H_0(X,Y) = h(X_0) + \sum_{t=1}^{\tau-1}(h(X_t) - h(Y_{t-1}))$, we introduce $\Delta_0 = h(X_0)$ and $\Delta_t = h(X_t) - h(Y_{t-1})$ for all $t \geq 1$, and define $H_0^n(X,Y) = \sum_{t=0}^{n} \Delta_t$. For simplicity we assume that $\Delta_t \in \mathbb{R}$, which corresponds to studying the component-wise behavior of $H_0(X,Y)$.

We have $\mathbb{E}[\tau] < \infty$ from Assumption 2.2, so that the computing time of $H_0(X,Y)$ has a finite expectation. Together with Assumption 2.3, this implies that $H_0^n(X,Y) \to H_0(X,Y)$ almost surely as $n \to \infty$. We now show that $H_0^n(X,Y)$ is a Cauchy sequence in L_2, the complete space of random variables with finite two moments, that is, $\sup_{n' \geq n} \mathbb{E}[(H_0^{n'}(X,Y) - H_0^n(X,Y))^2] \to 0$ as $n \to \infty$. For $n' \geq n$, we compute

$$\mathbb{E}[(H_0^{n'}(X,Y) - H_0^n(X,Y))^2] = \sum_{s=n+1}^{n'} \sum_{t=n+1}^{n'} \mathbb{E}[\Delta_s \Delta_t],$$

and consider each term $\mathbb{E}[\Delta_s \Delta_t]$ for $(s,t) \in \{n+1, \ldots, n'\}^2$. The Cauchy-Schwarz inequality implies $\mathbb{E}[\Delta_s \Delta_t] \leq \left(\mathbb{E}[\Delta_s^2] \cdot \mathbb{E}[\Delta_t^2]\right)^{1/2}$, and noting that $\mathbb{E}[\Delta_t^2] = \mathbb{E}[\Delta_t^2 \cdot \mathbb{1}(\tau > t)]$, we can apply Hölder's inequality with $p = 1 + \eta/2$, $q = (2+\eta)/\eta$ for any $\eta > 0$ to obtain

$$\mathbb{E}[\Delta_t^2 \cdot \mathbb{1}(\tau > t)] \leq \mathbb{E}[|\Delta_t|^{2+\eta}]^{1/(1+\eta/2)} \cdot \mathbb{E}[\mathbb{1}(\tau > t)]^{\eta/(2+\eta)} \leq \mathbb{E}[|\Delta_t|^{2+\eta}]^{1/(1+\eta/2)} \cdot \left(C\delta^t\right)^{\eta/(2+\eta)}.$$

Here we have used Assumption 2.2 to bound $\mathbb{E}[\mathbb{1}(\tau > t)]$. We can also use Assumption 2.1 together with Minkowski's inequality to bound $\mathbb{E}[|\Delta_t|^{2+\eta}]^{1/(1+\eta/2)}$ by a constant $\tilde{C}$, for all $t \geq 0$. Defining $\tilde{\delta} = \delta^{\eta/(2+\eta)} \in (0,1)$ then gives the bound $\mathbb{E}[\Delta_t^2] \leq \tilde{C}\tilde{\delta}^t$ for all $t \geq 0$. This implies $\mathbb{E}[(H_0^{n'}(X,Y) - H_0^n(X,Y))^2] \leq \bar{C}\tilde{\delta}^n$ for some other constant $\bar{C}$, and thus $(H_0^n(X,Y))$ is Cauchy in L_2. This proves that its limit $H_0(X,Y)$ has finite first and second moments. Assumption 2.1 implies that $\lim_{n\to\infty} \mathbb{E}[H_0^n(X,Y)] = \mathbb{E}_\pi[h(X)]$, by

[*]Department of Statistics, Harvard University, Cambridge, USA. Email: pjacob@fas.harvard.edu
[†]Department of Statistics, Harvard University, Cambridge, USA. Email: joleary@g.harvard.edu
[‡]Department of Mathematics & Statistics, Boston University, Boston, USA. Email: atchade@bu.edu

a telescopic sum argument, so we conclude that $\mathbb{E}[H_0(X,Y)] = \mathbb{E}_\pi[h(X)]$. We can also obtain an explicit representation of $\mathbb{E}[H_0(X,Y)^2]$ as the limit of $\mathbb{E}[H_0^n(X,Y)^2]$ when $n \to \infty$.

S1.2 Proof of Proposition 3.2

We adopt a similar strategy to that of Glynn and Rhee [2014], Section 4. For the $H_k(X,Y)$ case, the unbiased estimator of $F(s) = \mathbb{P}_\pi(X \le s)$ over R samples is of the form

$$\hat{F}^R(s) = \frac{1}{R}\sum_{r=1}^{R}\left(\mathbb{1}(X_k^{(r)} \le s) + \sum_{\ell=k+1}^{\tau^{(r)}-1}(\mathbb{1}(X_\ell^{(r)} \le s) - \mathbb{1}(Y_{\ell-1}^{(r)} \le s))\right).$$

Here $\tau^{(r)}$ denotes the meeting time of the r-th independent run. We want to show that as $R \to \infty$, $\sup_s |F(s) - \hat{F}^R(s)| \xrightarrow{a.s.} 0$. Define $G_{X,k}^R(s) = R^{-1}\sum_{r=1}^R \mathbb{1}(X_k^{(r)} \le s)$. For $\ell > k$, define

$$G_{X,\ell}^R(s) = \frac{1}{R}\sum_{r=1}^{R}\mathbb{1}(X_\ell^{(r)} \le s)\cdot\mathbb{1}(\ell \le \tau^{(r)}), \qquad G_{Y,\ell}^R(s) = \frac{1}{R}\sum_{r=1}^{R}\mathbb{1}(Y_{\ell-1}^{(r)} \le s)\cdot\mathbb{1}(\ell \le \tau^{(r)}).$$

Then $\hat{F}^R(s) = G_{X,k}^R(s) + \sum_{\ell=k+1}^{\infty}(G_{X,\ell}^R(s) - G_{Y,\ell}^R(s))$. By the standard Glivenko-Cantelli theorem, as $\mathbb{R} \to \infty$ we have

$$\sup_s \left|G_{X,k}^R(s) - \mathbb{P}(X_k \le s)\right| \xrightarrow{a.s.} 0, \qquad \sup_s \left|G_{X,\ell}^R(s) - \mathbb{E}[\mathbb{1}(X_\ell \le s)\cdot\mathbb{1}(\ell \le \tau)]\right| \xrightarrow{a.s.} 0,$$

$$\sup_s \left|G_{Y,\ell}^R(s) - \mathbb{E}[\mathbb{1}(Y_{\ell-1} \le s)\cdot\mathbb{1}(\ell \le \tau)]\right| \xrightarrow{a.s.} 0,$$

for each $\ell > k$. Next, we observe that, for all s, ℓ,

$$\mathbb{E}[(\mathbb{1}(X_\ell \le s) - \mathbb{1}(Y_{\ell-1} \le s))\cdot\mathbb{1}(\tau \ge \ell)] = \mathbb{P}(X_\ell \le s) - \mathbb{P}(X_{\ell-1} \le s).$$

This holds for a simple reason in our setting. For any $h(\cdot)$ and any ℓ,

$$\begin{aligned}\mathbb{E}[h(X_\ell) - h(Y_{\ell-1})] &= \mathbb{E}[(h(X_\ell) - h(Y_{\ell-1}))\mathbb{1}(\tau > \ell)] + \mathbb{E}[(h(X_\ell) - h(Y_{\ell-1}))\mathbb{1}(\tau \le \ell)] \\ &= \mathbb{E}[(h(X_\ell) - h(Y_{\ell-1}))\mathbb{1}(\tau > \ell)]\end{aligned}$$

since if $\tau \le \ell$ then $X_\ell = Y_{\ell-1}$ by Assumption 2.3. Applying this result with $h(\cdot) = \mathbb{1}(\cdot \le s)$ yields the desired statement.

The above implies that for any finite $i \ge k$ we have

$$\begin{aligned}&\left|G_{X,k}^R(s) - \mathbb{P}(X_k \le s) + \sum_{\ell=k+1}^{i}\left(G_{X,\ell}^R(s) - G_{Y,\ell}^R(s) - \mathbb{E}[(\mathbb{1}(X_\ell \le s) - \mathbb{1}(Y_{\ell-1} \le s))\cdot\mathbb{1}(\ell \le \tau)]\right)\right| \\ &= \left|G_{X,k}^R(s) - \mathbb{P}(X_k \le s) + \sum_{\ell=k+1}^{i}\left(G_{X,\ell}^R(s) - G_{Y,\ell}^R(s) - (\mathbb{P}(X_\ell \le s) - \mathbb{P}(X_{\ell-1} \le s))\right)\right| \\ &= \left|\left(G_{X,k}^R(s) + \sum_{\ell=k+1}^{i}(G_{X,\ell}^R(s) - G_{Y,\ell}^R(s))\right) - \mathbb{P}(X_i \le s)\right|.\end{aligned}$$

Hence

$$\sup_s \left|\left(G_{X,k}^R(s) + \sum_{\ell=k+1}^{i}(G_{X,\ell}^R(s) - G_{Y,\ell}^R(s))\right) - \mathbb{P}(X_i \le s)\right| \to 0.$$

We have assumed that $(X_t)_{t\geq 0}$ converges to π in total variation, which implies $\sup_s |\mathbb{P}(X_i \leq s) - F(s)| \to 0$ as $i \to \infty$. Also, we note that for all s,

$$\left| \sum_{\ell > i} G^R_{X,\ell}(s) - G^R_{Y,\ell}(s) \right| \leq \sum_{\ell > i} \frac{1}{R} \sum_{r=1}^{R} \mathbb{1}(\ell \leq \tau^{(r)}) \to \sum_{\ell > i} \mathbb{P}(\ell \leq \tau)$$

almost surely by the strong law of large numbers. Assumption 2.2 implies that this quantity goes to 0 as $i \to \infty$.

Combining these observations with the result obtained for finite i, we conclude that $\sup_s |\hat{F}^R(s) - F(s)| \to 0$ as $R \to \infty$, almost surely. The reasoning holds for the function $\hat{F}^R$ corresponding to $H_{k:m}(X, Y)$ instead of $H_k(X, Y)$; such a function is simply the average of a finite number of functions associated with $H_\ell(X, Y)$ for $\ell \in \{k, \ldots, m\}$.

S1.3 Proof of Proposition 3.3

Throughout the proof C denotes a generic finite constant whose actual value may change from one appearance to the next. We will use the usual Markov chain notation. In particular if $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a measurable function then $\bar{P}[f](x, y) := \int_{\mathcal{X}\times\mathcal{X}} \bar{P}((x, y), dz) f(z)$. Note that from the construction of $\bar{P}$, if f is depends only on x (resp. y), that is $f(x, y) = f(x)$ (resp. $f(x, y) = f(y)$), then $\bar{P}[f](x, y) = Pf(x)$ (resp. $\bar{P}[f](x, y) = Pf(y)$).

For $k \geq 1$, we consider the general problem of bounding $\mathbb{E}[S_k^2]$, where S_k is of the form

$$S_k = \mathbb{1}(\tau > k) \sum_{t=k}^{\tau-1} b_t \left(h(X_t) - h(Y_{t-1}) \right),$$

for some arbitrary bounded sequence $(b_t)_{t\geq 0}$. Fix an integer $N \geq k$, and set

$$S_k^{(N)} = \mathbb{1}(\tau > k) \sum_{t=k}^{N\wedge\tau-1} b_t \left(h(X_t) - h(Y_{t-1}) \right).$$

The same argument as in the proof of Proposition 3.1 can be applied here and shows that $S_k^{(N)}$ is a Cauchy sequence in L_2 that converges to S_k, as $N \to \infty$, so that

$$\mathbb{E}[S_k^2] = \lim_{N\to\infty} \mathbb{E}\left[(S_k^{(N)})^2 \right].$$

Since $|h|_{V^\beta} := \sup_x |h(x)|/V^\beta(x) < \infty$, and under Assumption 3.1, there exists a measurable function $g : \mathcal{X} \to \mathbb{R}$ such that $|g|_{V^\beta} < \infty$, and $g - Pg = h - \pi(h)$. To see this, first note that the drift condition (3.1) implies that for any $\beta \in (0, 1/2)$, we have $PV^\beta(x) \leq \lambda^\beta V^\beta(x) + b^\beta \mathbb{1}(x \in \mathcal{C})$, for all $x \in \mathcal{X}$. Indeed by Jensen's inequality $PV^\beta(x) \leq (PV(x))^\beta \leq (\lambda V(x) + b\mathbb{1}(x \in \mathcal{C}))^\beta \leq \lambda^\beta V^\beta(x) + b^\beta \mathbb{1}(x \in \mathcal{C})$, using the fact that for all $x, y \geq 0$ and $\alpha \in [0, 1]$, $(x + y)^\alpha \leq x^\alpha + y^\alpha$. The drift condition in V^β together with the fact that P is ϕ-irreducible and aperiodic implies that there exists $\rho_\beta \in (0, 1)$, $C_\beta < \infty$ such that for all $x \in \mathcal{X}$, $n \geq 0$,

$$\|P^n(x, \cdot) - \pi\|_{V^\beta} \leq C_\beta V^\beta(x) \rho_\beta^n, \tag{S1.1}$$

where for a function $W : \mathsf{X} \to [1, \infty)$, the W-norm between two probability measures μ, ν is defined as

$$\|\mu - \nu\|_W := \sup_{f \text{ meas.: } |f|_W \leq 1} |\mu(f) - \nu(f)|,$$

and $|f|_W := \sup_x |f(x)|/W(x)$. This result can be found in Theorem 15.0.1 of Meyn and Tweedie [2009]. It follows from (S1.1) that $\sum_{j\geq 0} |P^j(h-\pi(h))(x)| \leq |h|_{V^\beta} \sum_{j\geq 0} \|P^j(x,\cdot)-\pi\|_{V^\beta} \leq \frac{C_\beta |h|_{V^\beta} V^\beta(x)}{1-\rho_\beta} < \infty$. Hence the function

$$g(x) = \sum_{j\geq 0} P^j(h-\pi(h))(x), \quad x \in \mathcal{X},$$

is well-defined and measurable (as a limit of a sequence of measurable functions) and satisfies $|g(x)| \leq C_\beta |h|_{V^\beta} V^\beta(x)/(1-\rho_\beta)$. And since PV^β is finite everywhere, by Lebesgue's dominated convergence we deduce that Pg is finite everywhere as well and

$$Pg(x) = \int g(y)P(x,dy) = \sum_{j\geq 0} \int (P^j(h-\pi(h))(y))P(x,dy) = \sum_{j\geq 1} P^j(h-\pi(h))(x).$$

Hence $g - Pg = h - \pi(h)$, as claimed.

Hence, with $Z_t := (X_t, Y_{t-1})$, and $\bar{g}(x,y) = g(x) - g(y)$, we have

$$h(X_t) - h(Y_{t-1}) = \bar{g}(Z_t) - \bar{P}[\bar{g}](Z_t).$$

Using this and a telescoping sum argument, we write

$$\begin{aligned}
S_k^{(N)} &= \sum_{t=k}^{N-1} b_t \left(h(X_t) - h(Y_{t-1})\right) \mathbb{1}(\tau > t) \\
&= \sum_{t=k}^{N-1} b_t \left(\bar{g}(Z_t) - \bar{P}[\bar{g}](Z_t)\right) \mathbb{1}(\tau > t) \\
&= \sum_{t=k}^{N-1} b_t \left(\bar{g}(Z_{t+1}) - \bar{P}[\bar{g}](Z_t)\right) \mathbb{1}(\tau > t) \\
&\quad + \sum_{t=k}^{N-1} (b_t \bar{g}(Z_t)\mathbb{1}(\tau > t) - b_{t+1}\bar{g}(Z_{t+1})\mathbb{1}(\tau > t+1)) \\
&\quad + \sum_{t=k}^{N-1} (b_{t+1}\mathbb{1}(\tau > t+1) - b_t \mathbb{1}(\tau > t))\, \bar{g}(Z_{t+1}).
\end{aligned}$$

Since $\bar{g}(Z_{t+1}) = 0$ on $\{\tau = t+1\}$, the last term in the above display reduces to $\sum_{t=k}^{N-1}(b_{t+1} - b_t)\bar{g}(Z_{t+1})\mathbb{1}(\tau > t+1)$, and we obtain

$$\begin{aligned}
S_k^{(N)} = &\sum_{t=k}^{N-1} b_t \left(\bar{g}(Z_{t+1}) - \bar{P}[\bar{g}](Z_t)\right) \mathbb{1}(\tau > t) \\
&+ b_k \bar{g}(Z_k)\mathbb{1}(\tau > k) - b_N \bar{g}(Z_N)\mathbb{1}(\tau > N) + \sum_{t=k}^{N-1} (b_{t+1} - b_t)\bar{g}(Z_{t+1})\mathbb{1}(\tau > t+1). \quad \text{(S1.2)}
\end{aligned}$$

Let $\mathcal{F}_t$ denote the sigma-algebra generated by the variables $X_0, (X_1, Y_0), \ldots, (X_t, Y_{t-1})$. Note that $\{\tau > t\}$ belongs to $\mathcal{F}_t$. Hence

$$\begin{aligned}
\mathbb{E}\left[b_t \left(\bar{g}(Z_{t+1}) - \bar{P}[\bar{g}](Z_t)\right) \mathbb{1}(\tau > t) | \mathcal{F}_t\right] &= b_t \mathbb{1}(\tau > t) \mathbb{E}\left[\bar{g}(Z_{t+1}) - \bar{P}[\bar{g}](Z_t) | \mathcal{F}_t\right] \\
&= b_t \mathbb{1}(\tau > t)\left(\bar{P}[\bar{g}](Z_t) - \bar{P}[\bar{g}](Z_t)\right) = 0.
\end{aligned}$$

In other words, $\left\{\left(\sum_{t=k}^{j} b_t \left(\bar{g}(Z_{t+1}) - \bar{P}[\bar{g}](Z_t)\right) \mathbb{1}(\tau > t), \mathcal{F}_j\right),\ k \leq j \leq N-1\right\}$ is a martingale. The

orthogonality of the martingale increments gives

$$\mathbb{E}\left[\left(\sum_{t=k}^{N-1} b_t\left(\bar{g}(Z_{t+1})-\bar{P}[\bar{g}](Z_t)\right)\mathbb{1}(\tau>t)\right)^2\right]=\sum_{t=k}^{N-1} b_t^2\mathbb{E}\left[\left(\bar{g}(Z_{t+1})-\bar{P}[\bar{g}](Z_t)\right)^2\mathbb{1}(\tau>t)\right].$$

We use this together with (S1.2), the convexity of the squared norm, and Minkowski's inequality to conclude that

$$\begin{aligned}\mathbb{E}\left[(S_k^{(N)})^2\right]\le 4\sum_{t=k}^{N-1} b_t^2\mathbb{E}\left[\left(\bar{g}(Z_{t+1})-\bar{P}[\bar{g}](Z_t)\right)^2\mathbb{1}(\tau>t)\right]+4b_k^2\mathbb{E}\left[\bar{g}^2(Z_k)\mathbb{1}(\tau>k)\right]\\ +4b_N^2\mathbb{E}\left[\bar{g}^2(Z_N)\mathbb{1}(\tau>N)\right]+4\left[\sum_{t=k}^{N-1}|b_{t+1}-b_t|\mathbb{E}^{1/2}\left(\bar{g}^2(Z_{t+1})\mathbb{1}(\tau>t+1)\right)\right]^2. \quad \text{(S1.3)}\end{aligned}$$

Assumption 3.1 together with $\pi_0(V)<\infty$, implies that

$$\sup_{n\ge 0}\mathbb{E}[V(X_n)]\le C, \quad \text{(S1.4)}$$

for some finite constant C. Indeed, $\mathbb{E}[V(X_n)]=\int\pi_0(dx)P^nV(x)$, and a repeated application of the drift condition (3.1) implies that $P^nV(x)\le\lambda^nV(x)+\frac{b}{1-\lambda}$, for all $x\in\mathcal{X}$. For any $t\ge 0$, and for $1<p=\frac{1}{2\beta}$, and $\frac{1}{p}+\frac{1}{q}=1$, we have

$$\begin{aligned}\mathbb{E}\left[\left(V^\beta(X_t)+V^\beta(Y_{t-1})\right)^2\mathbb{1}(\tau>t)\right] &\le \mathbb{E}^{1/p}\left[\left(V^\beta(X_t)+V^\beta(Y_{t-1})\right)^{2p}\right]\mathbb{P}^{1/q}(\tau>t) && \text{(Hölder)}\\ &\le \left\{\mathbb{E}^{1/(2p)}\left[V^{2p\beta}(X_t)\right]+\mathbb{E}^{1/(2p)}\left[V^{2p\beta}(Y_{t-1})\right]\right\}^2 \\ &\qquad\times\mathbb{P}^{1/q}(\tau>t) && \text{(Minkowski)}\\ &\le C\mathbb{P}^{1/q}(\tau>t) && \text{(by (S1.4))}\\ &\le C\delta_\beta^t && \text{(by Assumption 2.2)},\end{aligned}$$

where $\delta_\beta=\delta^{1/q}$ with δ as in Assumption 2.2. Note that all the expectations on the right hand side of (S1.3) are of the form $\mathbb{E}\left[\bar{g}^2(Z_t)\mathbb{1}(\tau>t)\right]$, except for the term $\mathbb{E}\left[\left(\bar{g}(Z_{t+1})-\bar{P}[\bar{g}](Z_t)\right)^2\mathbb{1}(\tau>t)\right]$. However by the martingale difference property, $\mathbb{E}\left[\bar{P}[\bar{g}](Z_t)\left(\bar{g}(Z_{t+1})-\bar{P}[\bar{g}](Z_t)\right)\mathbb{1}(\tau>t)\right]=0$, so that for all $t\ge k$,

$$\begin{aligned}\mathbb{E}\left[\left(\bar{g}(Z_{t+1})-\bar{P}[\bar{g}](Z_t)\right)^2\mathbb{1}(\tau>t)\right] &= \mathbb{E}\left[\mathbb{1}(\tau>t)\bar{g}(Z_{t+1})^2\right]-\mathbb{E}\left[\bar{g}(Z_{t+1})\bar{P}[\bar{g}](Z_t)\mathbb{1}(\tau>t)\right]\\ &= \mathbb{E}\left[\mathbb{1}(\tau>t)\bar{g}(Z_{t+1})^2\right]-\mathbb{E}\left[\left(\bar{P}[\bar{g}](Z_t)\right)^2\mathbb{1}(\tau>t)\right] \text{ (by conditioning on } \mathcal{F}_t)\\ &\le \mathbb{E}\left[\mathbb{1}(\tau>t)\bar{g}(Z_{t+1})^2\right]\\ &\le |g|_{V^\beta}^2\mathbb{E}\left[\mathbb{1}(\tau>t)\left(V^\beta(X_{t+1})+V^\beta(Y_t)\right)^2\right]\\ &\le C\delta_\beta^t,\end{aligned}$$

using the same arguments as above. On the other hand, for all $t\ge k$, the term $\mathbb{E}\left[\bar{g}^2(Z_t)\mathbb{1}_{\{\tau>t\}}\right]$ satisfies

$$\mathbb{E}\left[\bar{g}^2(Z_t)\mathbb{1}(\tau>t)\right]\le|g|_{V^\beta}^2\mathbb{E}\left[\left(V^\beta(X_t)+V^\beta(Y_{t-1})\right)^2\mathbb{1}(\tau>t)\right]\le C\delta_\beta^t,$$

as seen above. In conclusion, all the expectations appearing in (S1.3) are upper bounded by some

constant times terms of the form δ_β^t. We conclude that

$$\mathbb{E}\left[(S_k^{(N)})^2\right] \leq C\left(b_k^2\delta_\beta^k + b_N^2\delta_\beta^N + \sum_{t=k}^{N-1} b_k^2\delta_\beta^t + \left[\sum_{t=k}^{N-1}|b_{t+1}-b_t|\delta_\beta^t\right]^2\right).$$

Letting $N \to \infty$, we conclude that

$$\mathbb{E}[S_k^2] \leq C\left(b_k^2\delta_\beta^k + \sum_{j\geq k} b_j^2\delta_\beta^j + \left[\sum_{j\geq k}|b_{j+1}-b_j|\delta_\beta^j\right]^2\right).$$

In the particular case of $\eta_{k:m}$, we have $\eta_{k:m} = \mathbb{1}(\tau > k)\sum_{t=k}^{\tau-1}\min\left(1, \frac{t+1-k}{m+1-k}\right)(h(X_{t+1}) - h(Y_t))$. Hence $b_k = 0$, $b_t = (t-k)/(m-k+1)$ if $k < t \leq m+1$, $b_t = 1$ if $t > m+1$. We then obtain the bound of Proposition 3.3.

S1.4 Proof of Proposition 3.4

Here Z_n is defined as (X_n, Y_{n-1}) for all $n \geq 1$. The assumption in (3.3), within the statement of Proposition 3.4, implies that for $(x,y) \in \mathcal{C}\times\mathcal{C}$, $\bar{P}$ can be written as a mixture

$$\bar{P}\left((x,y), dz\right) = \epsilon_{x,y}\nu_{x,y}(dz) + (1-\epsilon_{x,y})R\left((x,y), dz\right),$$

where $\epsilon_{x,y} \geq \epsilon$, $\nu_{x,y}(dz)$ is a restriction of $\bar{P}\left((x,y), dz\right)$ on $\mathcal{D}$ (that is for any measurable subset A of $\mathcal{D}$, $\bar{P}\left((x,y), A\right) = P\left((x,y), A\cap\mathcal{D}\right)/P\left((x,y), \mathcal{D}\right)$), and $R\left((x,y), dz\right)$ is the restriction of $\bar{P}\left((x,y), dz\right)$ on $(\mathcal{X}\times\mathcal{X})\setminus\mathcal{D}$. This means that whenever $(x,y) \in \mathcal{C}\times\mathcal{C}$ one can sample from $\bar{P}((x,y),\cdot)$ by drawing independently a Bernoulli random variable J, with probability of success $\epsilon_{x,y}$. Then if $J = 1$, we draw from $\nu_{x,y}$, if $J = 0$, we draw from $R\left((x,y),\cdot\right)$. From this decomposition, the proof of the proposition follows the same lines as in Douc et al. [2004], and we give the details only for completeness. We cannot directly invoke their result since their assumptions do not seem to apply to our setting.

Set $\bar{V}(x,y) = \frac{1}{2}(V(x) + V(y))$. First we show that the bivariate kernel satisfies a geometric drift towards $\mathcal{C}\times\mathcal{C}$. That is, there exists $\alpha \in (0,1)$ such that

$$\bar{P}\bar{V}(x,y) \leq \alpha\bar{V}(x,y), \quad (x,y) \notin \mathcal{C}\times\mathcal{C}. \tag{S1.5}$$

Indeed for $(x,y) \notin \mathcal{C}\times\mathcal{C}$, since $V \geq 1$, and $\mathcal{C} = \{V \leq L\}$, $\bar{V}(x,y) \geq (1+L)/2$. In other words, $\frac{1}{2} \leq \bar{V}(x,y)/(1+L)$. Therefore,

$$\bar{P}\bar{V}(x,y) = \frac{1}{2}\left(PV(x) + PV(y)\right) \leq \lambda\bar{V}(x,y) + \frac{b}{2} \leq \lambda\bar{V}(x,y) + \frac{b}{1+L}\bar{V}(x,y) \leq \alpha\bar{V}(x,y),$$

with $\alpha = \lambda + \frac{b}{1+L} < 1$. We set

$$B = \max\left(1, \frac{1}{\alpha}\sup_{(x,y)\in\mathcal{C}\times\mathcal{C}}\frac{\bar{P}[\bar{V}\mathbb{1}_{\mathcal{D}^c}](x,y)}{\bar{V}(x,y)}\right) \leq \frac{\lambda+b}{\alpha}.$$

In this section $\mathbb{1}_{\mathcal{S}}(\cdot)$ refers to the indicator function on the set $\mathcal{S}$. Let N_n denote the number of visits to $\mathcal{C}\times\mathcal{C}$ by time n. Then

$$\mathbb{P}(\tau > n) = \mathbb{P}(\tau > n,\ N_{n-1} \geq j) + \mathbb{P}(\tau > n,\ N_{n-1} < j).$$

The event $\{\tau > n,\ N_{n-1} \geq j\}$ implies that no success occurred within at least j independent Bernoulli random variables each with probability of success at least ϵ. Hence

$$\mathbb{P}(\tau > n,\ N_{n-1} \geq j) \leq (1-\epsilon)^j.$$

For the second term, we have (since $B \geq 1$, and the chains stay together after meeting via Assumption 2.3),

$$\mathbb{P}(\tau > n,\ N_{n-1} \leq j-1) \leq \mathbb{P}\left(Z_n \notin \mathcal{D}, B^{-N_{n-1}} \geq B^{-(j-1)}\right) = \mathbb{P}\left(\mathbb{1}_{\mathcal{D}^c}(Z_n) B^{-N_{n-1}} \geq B^{-(j-1)}\right).$$

Then use Markov's inequality to conclude that

$$\begin{aligned}\mathbb{P}(\tau > n,\ N_{n-1} \leq j-1) &\leq B^{j-1}\mathbb{E}\left[\mathbb{1}_{\mathcal{D}^c}(Z_n) B^{-N_{n-1}}\right] \\ &\leq B^{j-1}\mathbb{E}\left[\mathbb{1}_{\mathcal{D}^c}(Z_n) B^{-N_{n-1}} \bar{V}(Z_n)\right] = \alpha^n B^{j-1}\mathbb{E}[M_n],\end{aligned}$$

where $M_n = \mathbb{1}_{\mathcal{D}^c}(Z_n)\alpha^{-n}B^{-N_{n-1}}\bar{V}(Z_n)$ (set $N_0 = 0$ so that M_1 is well-defined). The result follows by noting that $\{M_n, \mathcal{F}_n\}$ is a super-martingale, where $\mathcal{F}_n = \sigma(Z_1, \ldots, Z_n)$, so that $\mathbb{E}[M_n] \leq \mathbb{E}[M_1] \leq \pi_0(V) + \pi_0(PV) \leq (1+\lambda)\pi_0(V) + b$, which implies that

$$\mathbb{P}(\tau > n) \leq (1-\epsilon)^j + \alpha^n B^{j-1}\left((1+\lambda)\pi_0(V) + b\right).$$

Since $\alpha < 1$, there exists an integer $k_0 \geq 1$ such that $\alpha B^{\frac{1}{k_0}} < 1$. In that case for $n \geq k_0$ one can take $j = \lceil n/k_0 \rceil$, to get

$$\mathbb{P}(\tau > n) \leq \left\{(1-\epsilon)^{\frac{1}{k_0}}\right\}^n + \left((1+\lambda)\pi_0(V) + b\right)\left\{\alpha B^{\frac{1}{k_0}}\right\}^n,$$

as claimed.

The argument that $\{M_n, \mathcal{F}_n\}$ is a super-martingale is as follows. We need to show that for all $n \geq 1$, $\mathbb{E}[M_{n+1}|\mathcal{F}_n] \leq M_n$. Note that $\mathbb{E}[M_{n+1}|\mathcal{F}_n] = 0 \leq M_n$ on $Z_n \in \mathcal{D}$. So it is enough to assume that $Z_n \notin \mathcal{D}$. Now, suppose also that $Z_n \notin \mathcal{C} \times \mathcal{C}$. Then $N_n = N_{n-1}$, and

$$\begin{aligned}\mathbb{E}[M_{n+1}|\mathcal{F}_n] &= \alpha^{-n-1}\mathbb{E}\left[B^{-N_{n-1}}\mathbb{1}_{\mathcal{D}^c}(Z_{n+1})\bar{V}(Z_{n+1})|\mathcal{F}_n\right], \\ &= \alpha^{-n-1}B^{-N_{n-1}}\mathbb{E}\left[\mathbb{1}_{\mathcal{D}^c}(Z_{n+1})\bar{V}(Z_{n+1})|Z_n\right], \\ &\leq \alpha^{-n-1}B^{-N_{n-1}}\mathbb{E}\left[\bar{V}(Z_{n+1})|Z_n\right], \\ &\leq \alpha^{-n}B^{-N_{n-1}}\bar{V}(Z_n), \\ &= M_n.\end{aligned}$$

Suppose now that $Z_n \in \mathcal{C} \times \mathcal{C}$. Then $N_n = N_{n-1} + 1$. Hence

$$\begin{aligned}\mathbb{E}[M_{n+1}|\mathcal{F}_n] &= \alpha^{-n-1}B^{-N_{n-1}-1}\mathbb{E}\left[\mathbb{1}_{\mathcal{D}^c}(Z_{n+1})\bar{V}(Z_{n+1})|\mathcal{F}_n\right], \\ &= \alpha^{-n}B^{-N_{n-1}}\bar{V}(Z_n)\frac{1}{\alpha B}\frac{\bar{P}[\mathbb{1}_{\mathcal{D}^c}\bar{V}](Z_n)}{\bar{V}(Z_n)}, \\ &\leq \alpha^{-n}B^{-N_{n-1}}\bar{V}(Z_n) = M_n.\end{aligned}$$

S2 Hamiltonian Monte Carlo on multivariate Normals

As Section 4 in the main document, we perform experiments with a d-dimensional Normal target distribution $\pi = \mathcal{N}(0, V)$, where V is the inverse of a matrix drawn from a Wishart distribution, with identity scale matrix and d degrees of freedom. We provide average meeting times obtained in varying dimensions, when using the coupled HMC algorithm described in Heng and Jacob [2019]. The latter article presents similar experiments but for a different choice of matrix V, thus we provide the present section for completeness.

Let us introduce a Markov kernel P as a mixture of two kernels, an MH kernel P_{MH} and an HMC kernel P_{HMC}. We first describe P_{MH} and a coupling of it. The kernel is an MH kernel with Normal random walk proposals, with a covariance matrix equal to 10^{-8} times the identity matrix. The coupled version of P_{MH} uses a maximal coupling of the proposals (as in Algorithm 2 of the main document).

The kernel P_{HMC} corresponds to an HMC algorithm, with mass matrix given by the inverse of the target variance V. This preconditioning mechanism is motivated by considerations similar to those in Girolami and Calderhead [2011]. In the present case of Normal distributions, this is particularly advantageous as it leads to a complete decoupling of the d components of the target; see Proposition 3.1 in Bou-Rabee and Sanz-Serna [2018]. We discretize Hamiltonian equations with a leap-frog integrator, using a stepsize of $\varepsilon = 0.1 \times d^{-1/4}$, and a number of steps of $L = 1 + \lfloor \varepsilon^{-1} \rfloor$, which corresponds to a trajectory length εL of approximately one. The coupling of such Hamiltonian kernels is done by using common random numbers for the momentum variables, i.e. a synchronous coupling [Bou-Rabee et al., 2018]. The kernels P_{MH} and P_{HMC}, and their coupled counterparts, are combined into mixtures P and $\bar{P}$, by assigning weights respectively of 0.05 and 0.95, i.e. an MH step is performed with probability 0.05.

We consider two types of initialization π_0: either the target distribution π, or a Normal distribution $\mathcal{N}(1_d, I_d)$, with 1_d a vector of ones and I_d the identity matrix. With the latter initialization, we observed a very low acceptance rate when using the stepsize ε given above. We did not observe any issue when the chains were started from π. We also did not observe such issues when V was replaced by the identity matrix. In principle, smaller stepsizes could be chosen, in order to increase the acceptance rate. However, this would result in more expensive iterations as L is defined as $1 + \lfloor \varepsilon^{-1} \rfloor$ above. Instead, we resort to the following heuristic strategy, which appears to solve the issue in the present example. We draw an initial position from $\mathcal{N}(1_d, I_d)$, then we perform 10 steps of "unadjusted HMC", with $\varepsilon = 0.1 \times d^{-1/4}$, and $L = 1 + \lfloor \varepsilon^{-1} \rfloor$ as described above. By "unadjusted HMC", we refer to a scheme where the final point of the Hamiltonian trajectory is accepted with probability one, i.e. no MH correction is applied. This initialization procedure is simply a redefinition of π_0, and thus would not jeopardize the validity of the proposed estimators.

The results under both initializations are shown in Figure 1, where we observe average meeting times that increase very slowly with the dimension of the target distribution. Thanks to the initialization strategy described above, we obtain similar meeting times when starting from the chains from π or from $\mathcal{N}(1_d, I_d)$.

S3 Baseball batting averages

We consider a classic Gibbs sampler discussed in the context of parallel computing in Rosenthal [2000]. In Rosenthal [1996], it was proved that the chain produced by this Gibbs sampler converges in total variation to within 1% of its stationary distribution after at most 140 iterations. This type of derivation is technically challenging and has been done only in specific cases. Using this result, Rosenthal [2000]

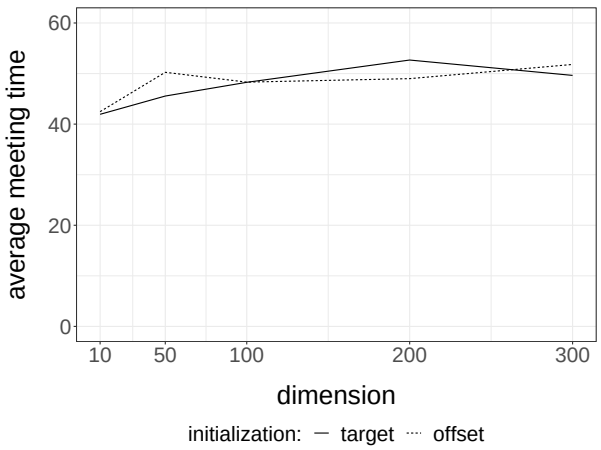


Figure 1: Scaling of the average meeting time of a coupled HMC algorithm with the dimension of the target $\mathcal{N}(0, V)$, where V is the inverse of a Wishart draw, as described in Section S2. The chains are either initialized from the target, or from a Normal $\mathcal{N}(1_d, I_d)$ followed by 10 steps of unadjusted HMC, as described in Section S2; this is referred to as "offset" in the legend.

recommends to run parallel chains with a burn-in of 140 iterations. We will compare this value with choices of k and m for the proposed unbiased estimators.

The data $(Z_n)_{n=1}^K$ are baseball players' batting averages taken from Table 1 of Morris [1983], where $K = 18$. In the model, each Z_n is assumed to follow $\mathcal{N}(\theta_n, V)$ where V is fixed to 0.00434. Then, θ_n is assumed to follow $\mathcal{N}(\mu, A)$, where μ is given a flat prior and A an Inverse Gamma (a, b), with $a = -1$ and $b = 2$. Here the Inverse Gamma (α, β) distribution has pdf $p(x; \alpha, \beta) = \Gamma(\alpha)^{-1}\beta^\alpha x^{-\alpha-1}\exp(-\beta/x)$. The Gibbs updates are as follows:

$$
\begin{aligned}
A|\text{rest} &\sim \text{Inverse Gamma}(a + (K-1)/2, b + \sum_{n=1}^K (\theta_n - \bar{\theta})^2/2), \text{ where } \bar{\theta} = K^{-1}\sum_{n=1}^K \theta_n, \\
\mu|\text{rest} &\sim \mathcal{N}\left(\bar{\theta}, A/K\right), \\
\theta_n|\text{rest} &\sim \mathcal{N}\left((V+A)^{-1}(\mu V + Z_n A), (V+A)^{-1}(AV)\right), \quad \text{for all } n \in \{1, \ldots, K\}.
\end{aligned}
$$

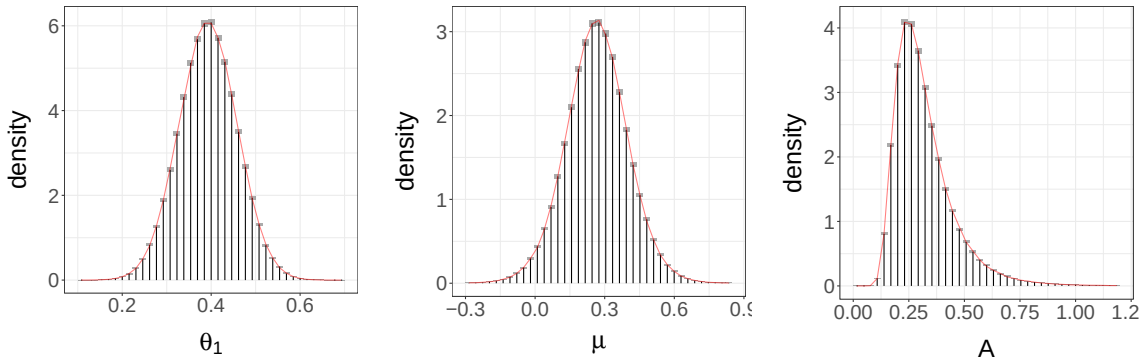
The initial values of the Gibbs sampler can be taken as $K^{-1}\sum_{n=1}^K Z_n$ for all θ_n [Rosenthal, 2000]. We couple this Gibbs sampler using maximal couplings of the conditional updates. The parameter space is 20-dimensional, but the chains meet as soon as the components (A, μ) meet, by construction. We consider the test function $h : (A, \mu, \theta_1, \ldots, \theta_K) \mapsto \theta_1$, that is, we are interested in the posterior expectation of θ_1, which represents the mean of the batting average of the first player.

We first run the coupled chains 1,000 times independently in parallel. All observed meeting times were less or equal to 4, so we consider k equal to 1, 3 and 5. Then, we consider m equal to k, $5k$ and $10k$. Over 10,000 independent experiments, we approximate the expected cost $\mathbb{E}[2(\tau-1) + \max(1, m-\tau+1)]$, the variance $\mathbb{V}[H_{k:m}(X, Y)]$, and we compute the inefficiency as the product of expected cost and variance. We then divide the inefficiency by the asymptotic variance of the MCMC estimator, denoted by V_∞, obtained from 5×10^5 iterations and a burn-in period of 10^3, and the CODA package [Plummer et al., 2006]. The results are shown in Table 1. We see that when k and m are large enough we can retrieve an inefficiency comparable to that of the underlying MCMC algorithm; for instance $k = 3$ and $m = 30$ yields an efficiency close to 1. The value less than 1 obtained for $k = 5$, $m = 50$ for the inefficiency ratio is likely due to Monte Carlo variability; we expect a value greater or equal to 1.

Since the efficiency of the underlying MCMC kernel is retrieved with $k = 3$, $m = 30$, for an associated

k	m	Cost	Variance	Inefficiency / V_∞
1	$1 \times k$	5.1494	0.0070	5.2152
1	$5 \times k$	8.0747	0.0013	1.5164
1	$10 \times k$	13.0747	0.0006	1.1838
3	$1 \times k$	6.0755	0.0061	5.3332
3	$5 \times k$	18.0747	0.0005	1.1990
3	$10 \times k$	33.0747	0.0002	1.0044
5	$1 \times k$	8.0747	0.0059	6.8677
5	$5 \times k$	28.0747	0.0003	1.1328
5	$10 \times k$	53.0747	0.0001	0.9816

Table 1: Cost, variance, and inefficiency divided by MCMC asymptotic variance V_∞ for various choices of k and m, for the test function $h : (A, \mu, \theta_1, \ldots, \theta_K) \mapsto \theta_1$, in the baseball batting averages example of Section S3.



(a) Histogram of θ_1. (b) Histogram of μ. (c) Histogram of A.

Figure 2: Gibbs sampling in the baseball batting averages example of Section S3. Histograms of marginal posterior distributions of θ_1, A and μ in 2a, 2b, 2c, produced using $R = 10,000$ estimators, $k = 3$, $m = 30$.

cost comparable to 33 steps of MCMC, we see that we can perform in parallel what would be equivalent to MCMC runs of 33 iterations. This is considerably less than the recommended burn-in of 140 derived in Rosenthal [1996], which is a strong indication that this recommended burn-in is conservative.

We plot histograms of θ_1, A and μ in Figures 2a, 2b and 2c, obtained for $R = 10,000$ estimators with $k = 3$ and $m = 30$. The overlaid red curves indicate the marginal target densities estimated from an MCMC run with 5×10^5 iterations and a burn-in of $1,000$ (which is unnecessarily conservative). These histograms confirm that accurate approximations of the posterior distribution are obtained with the proposed method.

S4 Pólya-Gamma Gibbs sampler for logistic regression

Next, we turn to a more modern MCMC setting and demonstrate that the proposed estimators can be constructed from the Pólya-Gamma Gibbs (PGG) sampler [Polson et al., 2013].

We consider the logistic regression of an outcome $Y = (Y_1, \ldots, Y_n) \in \{0, 1\}^n$ on covariates $X = (x_1, \ldots, x_n) \in \mathbb{R}^{n \times p}$. The model specifies $\mathbb{P}(Y_i = 1|\beta) = \text{expit}\left(x_i^T \beta\right)$, where $\text{expit} : z \mapsto 1/\left(1 + \exp(-z)\right)$, and the vector $\beta \in \mathbb{R}^p$ is the regression parameter. Realizations of $Y = (Y_1, \ldots, Y_n)$ are denoted by $y = (y_1, \ldots, y_n)$. The prior distribution on β is $\mathcal{N}(b, B)$, where we set the mean b to zero and the covariance B as diagonal, with non-zero entries equal to 10. The Pólya-Gamma Gibbs (PGG) sampler [Polson et al., 2013] is an MCMC algorithm targeting the posterior distribution, extended with n Pólya-Gamma variables $W = (W_1, \ldots, W_n)$. The Pólya-Gamma distribution with parameters $(1, c)$, denoted

by PG(1,c), has density defined for all $c \geq 0$ as

$$\forall x > 0 \quad \text{pg}(x; c) = \cosh\left(\frac{c}{2}\right) \exp\left(-\frac{c^2 x}{2}\right) \sum_{k=0}^{\infty} (-1)^k \frac{(2k+1)}{\sqrt{2\pi x^3}} \exp\left(-\frac{(2k+1)^2}{8x}\right).$$

Under the extended target distribution the variables $W = (W_1, \ldots, W_n)$ are independent of each other given β, and have the property that W_i follows $\text{PG}(1, |x_i^T \beta|)$ for all $1 \leq i \leq n$. The PGG sampler is a Gibbs sampler which alternates between the following updates:

$$\begin{aligned} W_i|\text{rest} &\sim \text{PG}(1, |x_i^T \beta|) \quad \text{for all } i \in \{1, \ldots, n\}, \\ \beta|\text{rest} &\sim \mathcal{N}(\Sigma(W)(X^T \tilde{y} + B^{-1} b), \Sigma(W)), \quad \text{with } \Sigma(W) = (X^T \text{diag}(W) X + B^{-1})^{-1}, \end{aligned}$$

where $\tilde{y} = (y_1 - 1/2, \ldots, y_n - 1/2)$. The resulting chain targets the posterior distribution and is uniformly ergodic [Choi and Hobert, 2013]. Here we initialize the algorithm with draws from the prior. We couple this chain using maximal couplings of the conditional updates. For the Pólya-Gamma updates, the probability density functions are intractable but the ratio of two density evaluations can be calculated using the identity

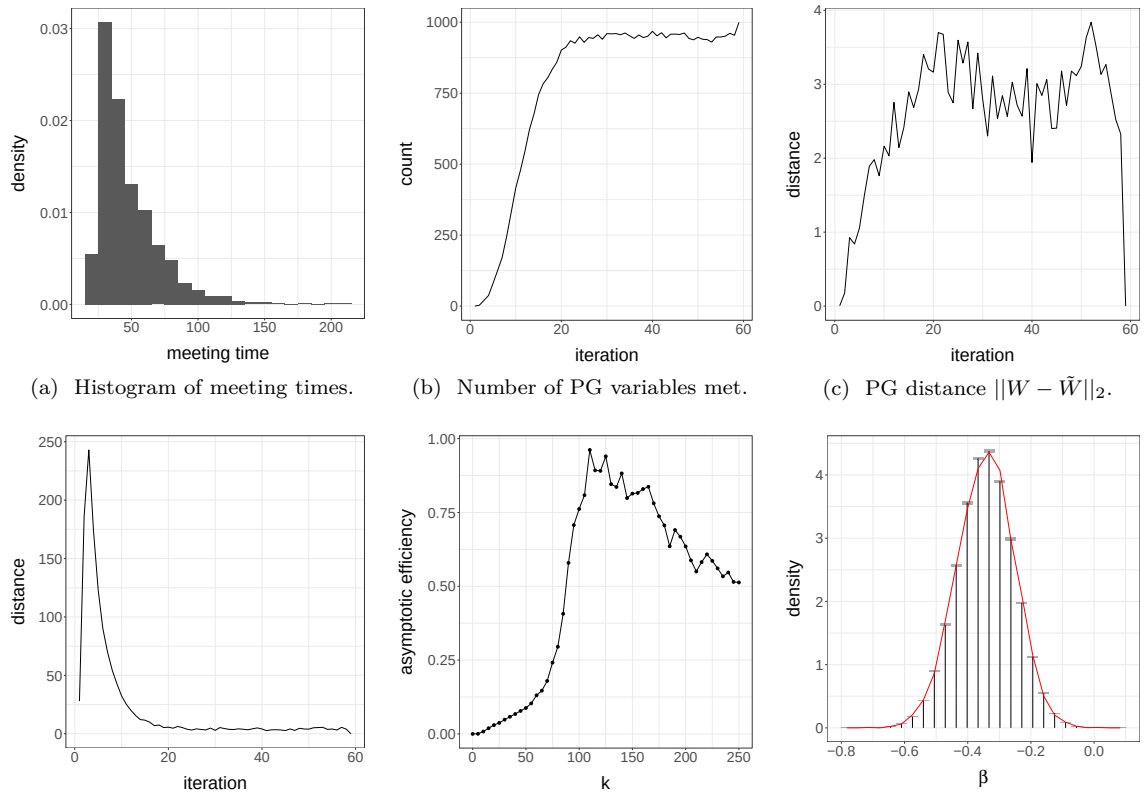
$$\forall x > 0 \quad \frac{\text{pg}(x; c_2)}{\text{pg}(x; c_1)} = \frac{\cosh(c_2/2)}{\cosh(c_1/2)} \exp\left(-\left(\frac{c_2^2}{2} - \frac{c_1^2}{2}\right) x\right),$$

which enables a fast implementation of the maximal coupling algorithm described in Algorithm 2 of the main document.

We apply the proposed method to the German credit data of [Lichman, 2013], a common example in binary regression and machine learning studies such as Polson et al. [2013], Huang et al. [2007], West [2000]. This dataset consists of $1,000$ loan application records, 700 of which were rated as creditworthy and 300 were rated as not creditworthy. Each record includes 20 additional variables including loan purpose, demographic information, bank account balances, marital, housing, employment status, and job type. Seven of these are quantitative and the rest are categorical. After translating categorical variables into indicators we obtain $p = 49$ regressors on $n = 1,000$ observations. Histograms of the meeting times for $R = 1,000$ coupled chains are shown in Figure 3a. These chains took between 18 and 164 steps to meet, with an average meeting time of 48 iterations.

The extended space of the Gibbs sampler is of dimension $n + p = 1,049$. However the two chains meet as soon as either all the n auxiliary PG variables or all the p regression coefficients meet. Since we use a maximal coupling of the update of the full vector of regression coefficients, either all or none of these meet at each iteration; MH-within-Gibbs strategies could be employed instead. As we show in Figure 3b, for one run of the coupled chains, the number of met PG variables rapidly increases to a plateau, at which point the chains are close enough to make coupling on the regression coefficients possible. Figure 3c shows the Euclidean distance between the PG variables of the two chains; it starts at zero as an artefact of the initial values of the PG variables being set to zero. Figure 3d shows the Euclidean distance between the regression coefficients. For this particular run, both PG variables and regression variables diverge at first, before converging as an increasing number of PG variables meet.

Finally, we consider the choice of k for the estimator $H_k(X, Y)$, and of m for the estimator $\bar{H}_{k:m}(X, Y)$. We consider the task of estimating the posterior mean of a particular regression coefficient corresponding to the installment payment as a percentage of disposable income. The efficiency of $H_k(X, Y)$, defined as one over the product of the variance times the cost, is shown in Figure 3e as a function of k. We see that choosing a large quantile of the distribution of τ, as shown in Figure 3a, would result in an efficiency close to its maximum. We choose $k = 110$, and, in line with the heuristics suggested in the



(a) Histogram of meeting times. (b) Number of PG variables met. (c) PG distance $||W - \tilde{W}||_2$.

(d) Coefficient distance $||\beta - \tilde{\beta}||_2$. (e) Efficiency of $H_k(X, Y)$ vs. k. (f) Histogram of β.

Figure 3: PGG sampling in the German credit example of Section S4. Histogram of meeting times in 3a. Trace of the number of Pólya-Gamma variables met in 3b, for one typical run of the coupled chains. Trace of the between-chain distance for Pólya-Gamma variables and regression coefficients in 3c-3d, for one typical run of the coupled chains. Efficiency of $H_k(X, Y)$ as a function of k in 3e, computed over $R = 1,000$ replicates. Histogram of the marginal posterior for the 'installment percent' regression coefficient in 3f, based on $R = 1,000$ estimators, $k = 110$ and $m = 1,100$.

main document, we take m to be $10k = 1,100$, that is, a large multiple of k. For the estimation of the posterior mean, we find that the above values of k and m yield an inefficiency about 7 times greater to that of the underlying Gibbs sampler, based on a long run; larger values of k and m would further reduce this inefficiency ratio.

With these tuning parameters, we produce a histogram of the posterior distribution for that coefficient in Figure 3f. We find agreement with a density estimated from a long MCMC run, depicted here by the overlaid curve in red. To summarize, in this example the proposed methodology effectively allows to run PGG chains in parallel, in chunks of approximately $1,000$ iterations, while bypassing the usual difficulties related to the choice of burn-in and the construction of confidence intervals.

In passing, in the particular case of the coupled PGG algorithm, we can directly show that the meeting time has an ε probability of occurring at each step t, for large enough t and some $\varepsilon > 0$ that does not depend on the starting points of the chains. Denote the β-component of the two chains by $(\beta_t)_{t\geq 0}$ and $(\tilde{\beta}_t)_{t\geq 0}$. From Choi and Hobert [2013], $(\beta_t)_{t\geq 0}$ and $(\tilde{\beta}_t)_{t\geq 0}$ are uniformly ergodic. Consider Fréchet's inequality $\mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1$ and the events $A = \{\beta_t \in S\}$ and $B = \{\tilde{\beta}_{t-1} \in S\}$, for some compact set S of $\mathbb{R}^p$ and some $t \geq 0$. We can define S such that $\{\beta_t \in S\}$ has probability at least $0.5 + \delta$, for some small $\delta > 0$, provided t is large enough, since

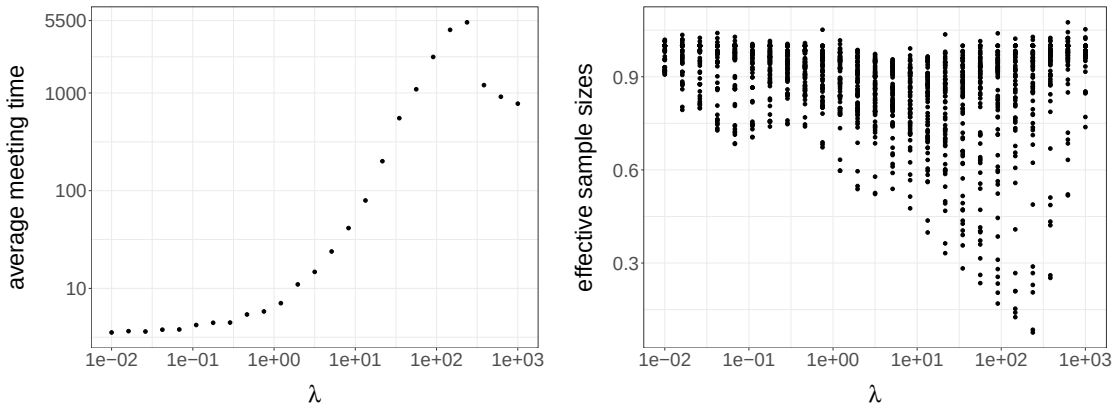
$$|\pi(S|y) - K^t(S|\beta)| \leq d_{\mathrm{TV}}(\pi(\cdot|y), K^t(\cdot|\beta)) < M\rho^t,$$

where β is any starting point, $K^t(\cdot|\beta)$ is the distribution of the t-th iterate of the chain starting from β, $M > 0$ and $\rho < 1$ are constants independent of β, and d_{TV} stands for the total variation distance. We take S to be a compact set of $\mathbb{R}^p$ such that $\pi(S|y) > 0.5 + 2\delta$. There is some t_0 such that $t \geq t_0$ implies $K^t(S|\beta) > 0.5 + \delta$; an identical reasoning can be done for the second chain. Using Fréchet's inequality, $\mathbb{P}(\beta_t \in S, \tilde{\beta}_{t-1} \in S) > 2\delta$. On these events, $|x_i^T \beta_t|^2 - |x_i^T \tilde{\beta}_{t-1}|^2$ are lower and upper-bounded for all $1 \leq i \leq n$, which results in a strictly non-zero probability of coupling each pair of Pólya-Gamma variables. This, in turns, leads to an $\varepsilon > 0$ probability of β_{t+1} meeting with $\tilde{\beta}_t$. Therefore, Assumption 2.2 on the meeting time in the main document is satisfied.

S5 Bayesian Lasso

We consider the setting of Bayesian inference in regression models. The Bayesian Lasso [Park and Casella, 2008] assigns a hierarchical prior on the parameters of a linear regression in such a way that the posterior mode corresponds to the Lasso estimator [Tibshirani, 1996, Efron et al., 2004]. The posterior distribution can be approximated by Gibbs sampling, independently for a range of regularization parameters $\lambda > 0$, which can then be selected by cross-validation or as described in Park and Casella [2008]; see also an alternative computational approach in Bornn et al. [2010]. On top of demonstrating the applicability of the proposed methodology in this setting, we illustrate the use of confidence intervals to guide the allocation of computational resources.

The model and associated Gibbs sampler are as follows. Consider an $n \times p$ matrix of standardized covariates X, and an n-vector of outcomes Y. The centered outcomes are denoted by $\tilde{Y}$, i.e. $\tilde{Y} = Y - \bar{Y}1_n$, where $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$ and 1_n is a n-vector of 1's. Introduce a regularization parameter $\lambda \geq 0$. The outcome Y is assumed to follow $\mathcal{N}(\mu 1_n + X\beta, \sigma^2 I_n)$, where μ is the intercept, and β is the p-vector of regression coefficients; I_n denotes a unit $n \times n$ diagonal matrix. The hierarchical prior specifies $\beta \sim \mathcal{N}(0_p, \sigma^2 D_\tau)$, where 0_p is a p-vector of 0's, and $D_\tau = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2)$. The prior on τ_j^2 is Exponential$(\lambda^2/2)$ for all $j \in \{1, \ldots, p\}$, and finally $\sigma^2 \sim$ Inverse Gamma(a, γ). A Gibbs sampler



(a) Average meeting time vs λ.

(b) Effective sample sizes vs λ.

Figure 4: Bayesian Lasso for the diabetes data as described in Section S5. Figure 4a shows the average of 100 independent meeting times against the regularization parameter λ. Figure 4b shows the effective sample size of each of the 64 components of β, from MCMC runs of length $50,000$ and the CODA package, as a function of λ.

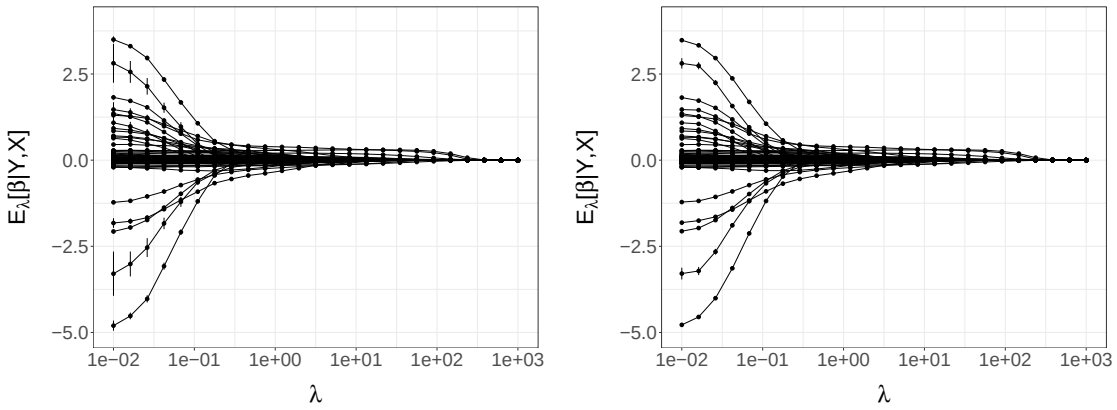
proposed in Park and Casella [2008] performs the following updates, after having integrated μ out:

$$\begin{aligned}
\beta|\text{rest} &\sim \mathcal{N}(A_\tau^{-1}X^T\tilde{Y}, \sigma^2 A_\tau^{-1}), \quad \text{where} \quad A_\tau = X^TX + D_\tau^{-1}, \\
\sigma^2|\text{rest} &\sim \text{Inverse Gamma}(a + \frac{n-1}{2} + \frac{p}{2}, \gamma + \left(\tilde{Y} - X\beta\right)^T \left(\tilde{Y} - X\beta\right)/2 + \beta^T D_\tau^{-1}\beta/2), \\
\tau_j^{-2}|\text{rest} &\sim \text{Inverse Gaussian}\left(\left(\lambda^2\sigma^2/\beta_j^2\right)^{1/2}, \lambda^2\right) \quad \text{for all } j \in \{1,\dots,p\}.
\end{aligned}$$

Here the Inverse Gaussian (μ, λ) distribution has pdf $p(x;\mu,\lambda) = (\lambda/2\pi x^3)^{1/2}\exp(-\lambda(x-\mu)^2/(2\mu^2 x))$. We initialize the sampler by setting $\beta_j = 0, \tau_j^2 = 1$ for all $j \in \{1,\dots,p\}$ and $\sigma^2 = 1$. We consider the diabetes dataset used in Efron et al. [2004], Park and Casella [2008], with $n = 442$ individuals and $p = 64$ covariates, as provided in the `lars` package accompanying Efron et al. [2004]. The parameter space is of dimension $2p + 1 = 129$. We set $a = 0$ and $\gamma = 0$ throughout. To couple the Gibbs sampler we use a maximal coupling of each conditional update, and focus on producing unbiased estimators of the posterior means $\mathbb{E}_\lambda[\beta|Y, X]$ given λ.

For a range of values of λ between 10^{-2} and 10^3, we run 100 coupled chains until they meet, and plot the empirical average of the meeting times as a function of λ in Figure 4a. First, we note that the average meeting times are very small for small values of λ. Next we observe a peak located around $\lambda = 10^2$, which implies a similar peak for the computational cost of the proposed estimators. This could be either a consequence of the mixing properties of the underlying MCMC, or a defect of the coupling strategy. To investigate this, we run the underlying Gibbs sampler for the same values of λ, for $50,000$ iterations, discard the first $5,000$ iterations, and compute the effective sample size using the CODA package [Plummer et al., 2006]. We do so for each of the 64 components of β, and plot the results in Figure 4b; each dot corresponds to the ESS of one of the components, for a particular value of λ. The effective sample size is divided by the number of iterations post burn-in, and thus we expect a number between 0 and 1. We see a drastic loss of efficiency around $\lambda = 10^2$, indicating that the Gibbs sampler of Park and Casella [2008] mixes poorly for these values of λ.

For each λ, we now choose k as the 99% quantile of the 100 meeting times previously obtained, and we choose $m = 10k$. We produce $R = 100$ unbiased estimators for each posterior mean $\mathbb{E}_\lambda[\beta|Y, X]$, and plot them against λ in Figure 5a, component-wise. On top of each dot lies a 95% confidence interval



(a) Posterior mean of β vs λ, with 100 estimators per λ. (b) Posterior mean of β vs λ, with $1,000$ more estimators for the 10 smallest values of λ.

Figure 5: Bayesian Lasso for the diabetes data with $n = 442$ and $p = 64$, described in Section S5. Figure 5a shows the paths of posterior means $\mathbb{E}_\lambda[\beta|Y, X]$ against the regularization parameter λ, obtained with $R = 100$ estimators, k chosen as one plus the 90% quantile of the distribution of meeting times, and $m = 10k$. Figure 5b shows the estimates obtained after having run $1,000$ more estimators for the first 10 values of λ.

represented by a vertical segment: these are too small to be noticeable except for the smallest values of λ.

In order to refine the posterior mean estimates, we can allocate computational resources based on the confidence intervals and the costs associated with each λ (there are 25 values of λ in total). In order to tighten the most visible confidence intervals, we produce $1,000$ more estimators for the 10 smallest values of λ, and obtain the refined estimates shown in Figure 5b. That is, these estimates are obtained from $1,100$ unbiased estimators for the 10 smallest values of λ, and for 100 for the other values of λ. This refinement procedure could be automatized, adaptively producing more estimators for values of λ where confidence intervals are wider.

References

L. Bornn, A. Doucet, and R. Gottardo. An efficient computational approach for prior sensitivity analysis and cross-validation. *Canadian Journal of Statistics*, 38(1):47–64, 2010. 13

N. Bou-Rabee and J. M. Sanz-Serna. Geometric integrators and the Hamiltonian Monte Carlo method. *Acta Numerica*, 27:113–206, 2018. 8

N. Bou-Rabee, A. Eberle, and R. Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. *arXiv preprint arXiv:1805.00452*, 2018. 8

H. M. Choi and J. P. Hobert. The Pólya–Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics*, 7:2054–2064, 2013. 11, 13

R. Douc, E. Moulines, and J. S. Rosenthal. Quantitative bounds on convergence of time-inhomogeneous markoc chains. *Annals of Applied Probability*, 14(4):1643–1665, 2004. 6

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004. 13, 14

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011. 8

P. W. Glynn and C.-H. Rhee. Exact estimation for Markov chain equilibrium expectations. *Journal of Applied Probability*, 51(A):377–389, 2014. 1, 2

J. Heng and P. E. Jacob. Unbiased Hamiltonian Monte Carlo with couplings. *Biometrika*, 106(2): 287–302, 2019. 8

C.-L. Huang, M.-C. Chen, and C.-J. Wang. Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications*, 33(4):847–856, 2007. 11

M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml. 11

S. Meyn and R. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, second edition, 2009. 4

C. N. Morris. Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55, 1983. 9

T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482): 681–686, 2008. 13, 14

M. Plummer, N. Best, K. Cowles, and K. Vines. CODA: convergence diagnosis and output analysis for MCMC. *R news*, 6(1):7–11, 2006. 9, 14

N. G. Polson, J. G. Scott, and J. Windle. Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013. 10, 11

J. S. Rosenthal. Analysis of the Gibbs sampler for a model related to James-Stein estimators. *Statistics and Computing*, 6(3):269–275, 1996. 8, 10

J. S. Rosenthal. Parallel computing and Monte Carlo algorithms. *Far east journal of theoretical statistics*, 4(2):207–236, 2000. 8, 9

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 13

M. Vihola. Unbiased estimators and multilevel Monte Carlo. *Operations Research*, 66(2):448–462, 2017. 1

D. West. Neural network credit scoring models. *Computers & Operations Research*, 27(11):1131–1152, 2000. 11