Scalable Metropolis–Hastings for Exact Bayesian Inference with Large Datasets: Supplementary Material

Robert Cornish [1] **Paul Vanetti** [1] **Alexandre Bouchard-Côté** [2] **George Deligiannidis** [1 3] **Arnaud Doucet** [1 3]

A. Guide to Notation

$a \wedge b$	$\min\{a, b\}$
$a \vee b$	$\max\{a, b\}$
$B(x, K)$	Euclidean ball centered at x of radius K
$\mathbb{I}_A$	Indicator function of the set A
$\partial_j F(x)$	j-th partial derivative of F at x, i.e. $\partial F(x)/\partial x_j$
$\nabla F(x)$	Gradient of F at x
$\nabla^2 F(x)$	Hessian of F at x
$\|\cdot\|$	The ℓ^2 norm
$\|\cdot\|_1$	The ℓ^1 norm
$\|\cdot\|_\infty$	The supremum norm
$\|\cdot\|_{\text{op}}$	The operator norm with respect to $\|\cdot\|$ on the domain and range
I_d	Identity matrix
$A \prec B$	$B - A$ is symmetric positive-definite
$A \preceq B$	$B - A$ is symmetric nonnegative-definite
$a(x) \asymp b(x)$ as $x \to x_0$	$\lim_{x \to x_0} a(x)/b(x) = 1$
$a(x) = O(b(x))$ as $x \to x_0$	$\limsup_{x \to x_0} \lvert a(x)/b(x)\rvert < \infty$
$a(x) = \Theta(b(x))$ as $x \to x_0$	$a(x) = O(b(x))$ as $x \to x_0$ and $\liminf_{x \to x_0} \lvert a(x)/b(x)\rvert > 0$. (Note that similar notation is used for our state space Θ, but the meaning will always be clear from context.)
$x \ll y$	(Informal) x is much smaller than y
$x \approx y$	(Informal) x is approximately equal to y
Leb	The Lebesgue measure
a.s.	Almost surely
i.i.d.	Independent and identically distributed
$X_n \xrightarrow{\mathbb{P}} X$	X_n converges to X in $\mathbb{P}$-probability
$X_n = O_{\mathbb{P}}(a_n)$	X_n/a_n is $\mathbb{P}$-tight, i.e. for all $\epsilon > 0$ there exists $c > 0$ such that $\mathbb{P}(\lvert X_n/a_n\rvert < c) > 1 - \epsilon$ for all n
$X_n = o_{\mathbb{P}}(a_n)$	$X_n/a_n \xrightarrow{\mathbb{P}} 0$
$\mathbb{E}[X]$	Expectation of a random variable X
$\mathbb{E}[X; A]$	$\mathbb{E}[X\mathbb{I}_A]$
L^p	The space of random variables X such that $\mathbb{E}[\lvert X\rvert^p] < \infty$
$L^p(\mu)$	The space of real-valued test functions f such that $f(X) \in L^p$ where $X \sim \mu$

We also use multi-index notation to express higher-order derivatives succinctly. Specifically, for $\beta = (\beta_1, \cdots, \beta_d) \in \mathbb{Z}^d_{\geq 0}$

[1]University of Oxford, Oxford, United Kingdom [2]University of British Columbia, Vancouver, Canada [3]The Alan Turing Institute, London, United Kingdom. Correspondence to: Rob Cornish <rcornish@robots.ox.ac.uk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

and $\theta = (\theta_1, \cdots, \theta_d) \in \Theta$, we define

$$|\beta| := \sum_{i=1}^{d} \beta_i \qquad \beta! := \prod_{i=1}^{d} \beta_i! \qquad \theta^\beta := \prod_{i=1}^{d} \theta_i^{\beta_i} \qquad \partial^\beta := \frac{\partial^{|\beta|}}{\partial^{\beta_1} \cdots \partial^{\beta_d}}.$$

B. Factorised Metropolis–Hastings

Note that the definition (2) of $\alpha_{\mathrm{FMH}}(\theta, \theta')$ technically does not apply when $\pi(\theta)q(\theta, \theta') = 0$. For concreteness, like Hastings (1970), we therefore define explicitly

$$\alpha_{\mathrm{FMH}}(\theta, \theta') := \begin{cases} \prod_{i=1}^{m} 1 \wedge \frac{\pi_i(\theta')q_i(\theta', \theta)}{\pi_i(\theta)q_i(\theta, \theta')} & \text{if each } \pi_i(\theta)q_i(\theta, \theta') \neq 0 \\ 1 & \text{otherwise,} \end{cases}$$

and take $\alpha_{\mathrm{MH}}(\theta, \theta')$ to be the case when $m = 1$. We still take $\alpha_{\mathrm{TFMH}}(\theta, \theta')$ to be defined by (6). We first establish a useful preliminary Proposition.

Proposition B.1. *For all $\theta, \theta' \in \Theta$, $\alpha_{\mathrm{FMH}}(\theta, \theta') = \alpha_{\mathrm{MH}}(\theta, \theta')(\alpha_{\mathrm{FMH}}(\theta, \theta') \vee \alpha_{\mathrm{FMH}}(\theta', \theta))$.*

Proof. The cases where $\pi_i(\theta)q_i(\theta, \theta') = 0$ or $\pi_i(\theta')q_i(\theta', \theta) = 0$ for some i are immediate from the definition above. Otherwise, since $(1 \wedge c)^{-1} = 1 \vee c^{-1}$ for all $c > 0$,

$$\begin{aligned} \alpha_{\mathrm{MH}}(\theta, \theta')^{-1} &= \left(1 \wedge \prod_{i=1}^{m} \frac{\pi_i(\theta')q_i(\theta', \theta)}{\pi_i(\theta)q_i(\theta, \theta')}\right)^{-1} \\ &= 1 \vee \prod_{i=1}^{m} \frac{\pi_i(\theta)q_i(\theta, \theta')}{\pi_i(\theta')q_i(\theta', \theta)}, \end{aligned}$$

and hence

$$\begin{aligned} \frac{\alpha_{\mathrm{FMH}}(\theta, \theta')}{\alpha_{\mathrm{MH}}(\theta, \theta')} &= \alpha_{\mathrm{FMH}}(\theta, \theta') \vee \left(\alpha_{\mathrm{FMH}}(\theta, \theta') \prod_{i=1}^{m} \frac{\pi_i(\theta)q_i(\theta, \theta')}{\pi_i(\theta')q_i(\theta', \theta)}\right) \\ &= \alpha_{\mathrm{FMH}}(\theta, \theta') \vee \left(\prod_{i=1}^{m} (1 \wedge \frac{\pi_i(\theta')q_i(\theta', \theta)}{\pi_i(\theta)q_i(\theta, \theta')}) \prod_{i=1}^{m} \frac{\pi_i(\theta)q_i(\theta, \theta')}{\pi_i(\theta')q_i(\theta', \theta)}\right) \\ &= \alpha_{\mathrm{FMH}}(\theta, \theta') \vee \left(\prod_{i=1}^{m} 1 \wedge \frac{\pi_i(\theta)q_i(\theta, \theta')}{\pi_i(\theta')q_i(\theta', \theta)}\right) \\ &= \alpha_{\mathrm{FMH}}(\theta, \theta') \vee \alpha_{\mathrm{FMH}}(\theta', \theta) \end{aligned}$$

which gives the result. □

Corollary B.1. *For all $\theta, \theta' \in \Theta$, $\alpha_{\mathrm{FMH}}(\theta, \theta') \leq \alpha_{\mathrm{MH}}(\theta, \theta')$.*

B.1. Reversibility

To show reversibility for P_{FMH} and P_{TFMH}, we will use the standard result (see e.g. (Geyer, 1998, Lemma 3.4)) that a kernel of the form

$$P(\theta, A) = \left(1 - \int q(\theta, \theta')\alpha(\theta, \theta')d\theta'\right) \mathbb{I}_A(\theta) + \int_A q(\theta, \theta')\alpha(\theta, \theta')d\theta'$$

is reversible if $\pi(\theta)q(\theta, \theta')\alpha(\theta, \theta')$ is symmetric in θ and θ'. It is straightforward to show for instance that

$$\pi(\theta)q(\theta, \theta')\alpha_{\mathrm{MH}}(\theta, \theta') = \pi(\theta')q(\theta', \theta)\alpha_{\mathrm{MH}}(\theta', \theta), \tag{B.1}$$

which is immediate if either $\pi(\theta) = 0$ or $\pi(\theta') = 0$, and otherwise

$$\begin{aligned} \pi(\theta)q(\theta, \theta')\alpha_{\mathrm{MH}}(\theta, \theta') &= \pi(\theta)q(\theta, \theta') \left(1 \wedge \frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta)q(\theta, \theta')}\right) \\ &= \pi(\theta)q(\theta, \theta') \wedge \pi(\theta')q(\theta', \theta). \end{aligned}$$

We use this result to establish reversiblity of P_{FMH}. This result is standard but we include it here for completeness.

Proposition B.2. *P_{FMH} is π-reversible.*

Proof. By Proposition B.1

$$\pi(\theta)q(\theta,\theta')\alpha_{\mathrm{FMH}}(\theta,\theta') = \pi(\theta)q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta')(\alpha_{\mathrm{FMH}}(\theta,\theta') \vee \alpha_{\mathrm{FMH}}(\theta',\theta)),$$

which is symmetric in θ and θ' by (B.1). □

Proposition B.3. *If $\overline{\lambda}(\theta,\theta')$ is symmetric in θ and θ', then P_{TFMH} is π-reversible.*

Proof. Simply write

$$\alpha_{\mathrm{TFMH}}(\theta,\theta') = \mathbb{I}(\overline{\lambda}(\theta,\theta') < R)\alpha_{\mathrm{FMH}}(\theta,\theta') + \mathbb{I}(\overline{\lambda}(\theta,\theta') \geq R)\alpha_{\mathrm{MH}}(\theta,\theta').$$

The result then follows from the symmetry of the indicator functions, (B.1), and the proof of Proposition B.2. □

B.2. Ergodic Properties

We provide a brief background to the theory of φ-irreducible Markov Chains. See (Meyn & Tweedie, 2009) for a comprehensive treatment.

For a transition kernel P, we inductively define the transition kernel P^k for $k \geq 1$ by setting $P^1 := P$ and

$$P^k(\theta, A) := \int P(\theta, d\theta')P^{k-1}(\theta', A)d\theta'$$

for $k > 1$, where $\theta \in \Theta$ and $A \subseteq \Theta$ is measurable. Given a nontrivial measure φ on Θ, we say P is *φ-irreducible* if $\varphi(A) > 0$ implies $P^k(\theta, A) > 0$ for some $k \geq 1$. For φ-irreducible P, we define a *k-cycle* of P to be a partition $D_1, \cdots, D_k, N$ of Θ such that $\varphi(N) = 0$, and for all $1 \leq i \leq k$, if $\theta \in D_i$ then $P(\theta, D_{i+1}) = 1$. (Here $i+1$ is meant modulo k.) If there exists a k-cycle with $k > 1$, we say that P is *periodic*; otherwise it is *aperiodic*.

If P is φ-irreducible and aperiodic and has invariant distribution π, we say P is *geometrically ergodic* if there exists constants $\rho < 1$, $C < \infty$, and a π-a.s. finite function $V \geq 1$ such that

$$\|P^k(\theta, \cdot) - \pi\|_V \leq C\, V(\theta)\rho^k$$

for all $\theta \in \Theta$ and $k \geq 1$. Here $\|\cdot\|_V$ denotes the V-norm on signed measures defined by

$$\|\mu\|_V = \sup_{|f| \leq V} |\pi(f)|,$$

where $\pi(f) := \int f(\theta)\pi(d\theta)$. By (Roberts & Rosenthal, 1997, Proposition 2.1), this is equivalent to the apparently weaker condition that there exist some constant $\rho > 0$ and π-a.s. finite function M such that

$$\|P^k(\theta, \cdot) - \pi\|_{\mathrm{TV}} \leq M(\theta)\rho^k$$

for all $\theta \in \Theta$ and $k \geq 1$, where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation distance on signed measures.

Our interest in geometric ergodicity is largely due to the implications it has for the *asymptotic variance* of the ergodic averages produced by a transition kernel. Suppose $(\theta_k)_{k\geq 1}$ is a stationary Markov chain with transition kernel P having invariant distribution π. For $f \in L^2(\pi)$, the asymptotic variance for the ergodic averages of f is defined

$$\mathrm{var}(f, P) := \lim_{k\to\infty} \mathrm{Var}\left(\sqrt{k}(\frac{1}{k}\sum_{i=1}^{k} f(\theta_k) - \pi(f))\right) = \lim_{k\to\infty} \frac{1}{k}\mathrm{Var}(\sum_{i=1}^{k} f(\theta_k)).$$

We abuse notation a little and denote the variance of $f(\theta)$ where $\theta \sim \pi$ by $\mathrm{var}(f, \pi)$.

Of interest is also the (right) *spectral gap*, which for a π-reversible transition kernel P is defined

$$\mathrm{Gap}(P) := \inf_{f \in L^2(\pi):\pi(f)=0} \frac{\int\int \frac{1}{2}(f(\theta)-f(\theta'))^2\pi(d\theta)P(\theta,d\theta')}{\int f(\theta)^2\pi(d\theta)}.$$

Finally, it is convenient to define the MH rejection probability

$$r_{\mathrm{MH}}(\theta) := 1 - \int q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta')d\theta',$$

and similarly r_{FMH} and r_{TFMH} for FMH and TFMH.

Proposition B.4. *P_{TFMH} is φ-irreducible and aperiodic whenever P_{MH} is.*

Proof. We use throughout the easily verified facts $\alpha_{\mathrm{FMH}}(\theta,\theta') \leq \alpha_{\mathrm{TFMH}}(\theta,\theta') \leq \alpha_{\mathrm{MH}}(\theta,\theta')$ and $r_{\mathrm{FMH}}(\theta) \geq r_{\mathrm{TFMH}}(\theta) \geq r_{\mathrm{MH}}(\theta)$ for all $\theta,\theta' \in \Theta$. See Proposition B.1.

For φ-irreducibility, first note that if $\alpha_{\mathrm{MH}}(\theta,\theta') > 0$ then $\alpha_{\mathrm{TFMH}}(\theta,\theta') > 0$. This holds since if $\alpha_{\mathrm{TFMH}}(\theta,\theta') = 0$, then either $\alpha_{\mathrm{MH}}(\theta,\theta') = 0$ or $\alpha_{\mathrm{FMH}}(\theta,\theta') = 0$. In the latter case we must have some $\pi_i(\theta')q_i(\theta',\theta) = 0$, so that $\pi(\theta')q(\theta,\theta') = 0$, and hence again $\alpha_{\mathrm{MH}}(\theta,\theta') = 0$.

We now show by induction on $k \in \mathbb{Z}_{\geq 1}$ that for all $\theta \in \Theta$, $P^k_{\mathrm{MH}}(\theta,A) > 0$ implies $P^k_{\mathrm{TFMH}}(\theta,A) > 0$. For $k = 1$, suppose $P_{\mathrm{MH}}(\theta,A) > 0$. Then either $r_{\mathrm{MH}}(\theta)\mathbb{I}_A(\theta) > 0$ or $\int_A q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta')d\theta' > 0$. In the former case we we have

$$r_{\mathrm{TFMH}}(\theta)\mathbb{I}_A(\theta) \geq r_{\mathrm{MH}}(\theta)\mathbb{I}_A(\theta) > 0.$$

In the latter case the above considerations give

$$\begin{aligned}\mathrm{Leb}(\{\theta' \in A \mid q(\theta,\theta')\alpha_{\mathrm{TFMH}}(\theta,\theta') > 0\}) &= \mathrm{Leb}(\{\theta' \in A \mid q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta') > 0\}) \\ &> 0.\end{aligned}$$

Either way we have $P_{\mathrm{TFMH}}(\theta,A) > 0$.

Suppose now $P^{k-1}_{\mathrm{MH}}(\theta,A) > 0$ implies $P^{k-1}_{\mathrm{TFMH}}(\theta,A) > 0$. Then observe

$$P^k_{\mathrm{MH}}(\theta,A) = r_{\mathrm{MH}}(\theta)P^{k-1}_{\mathrm{MH}}(\theta,A) + \int q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta')P^{k-1}_{\mathrm{MH}}(\theta',A)d\theta'$$

and likewise *mutatis mutandis* for $P^k_{\mathrm{TFMH}}(\theta,A)$. Thus if $P^k_{\mathrm{MH}}(\theta,A) > 0$, one possibility is $r_{\mathrm{MH}}(\theta)P^{k-1}_{\mathrm{MH}}(\theta,A) > 0$, which implies $r_{\mathrm{TFMH}}(\theta) > 0$ and, by the induction hypothesis, $P^{k-1}_{\mathrm{TFMH}}(\theta,A) > 0$. The only other possibility is

$$\begin{aligned}\mathrm{Leb}(\{\theta' \in \Theta \mid q(\theta,\theta')\alpha_{\mathrm{TFMH}}(\theta,\theta')P^{k-1}_{\mathrm{TFMH}}(\theta',A) > 0\}) &= \mathrm{Leb}(\{\theta' \in \Theta \mid q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta')P^{k-1}_{\mathrm{MH}}(\theta',A) > 0\}) \\ &> 0,\end{aligned}$$

again by the induction hypothesis. Either way, as desired $P^k_{\mathrm{TFMH}}(\theta,A) > 0$. It now follows that P_{TFMH} is φ-irreducible when P_{MH} is.

Now suppose P_{MH} and hence P_{TFMH} is φ-irreducible. If P_{TFMH} is periodic, then there exists a k-cycle $D_1,\cdots,D_k,N$ for P_{TFMH} with $k > 1$. But now if $\theta \in D_i$, then $\mathbb{I}_{D_{i+1}}(\theta) = 0$ and so

$$\begin{aligned}P_{\mathrm{MH}}(\theta,D_{i+1}) &= \int_{D_{i+1}} q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta')d\theta' \\ &\geq \int_{D_{i+1}} q(\theta,\theta')\alpha_{\mathrm{TFMH}}(\theta,\theta')d\theta' \\ &= P_{\mathrm{TFMH}}(\theta,D_{i+1}) \\ &= 1.\end{aligned}$$

Thus the same partition is a k-cycle for P_{MH} which is therefore periodic. □

Theorem B.1. *If P_{MH} is φ-irreducible, aperiodic, and geometrically ergodic, then P_{TFMH} is too if*

$$\delta := \inf_{\overline{\lambda}(\theta,\theta')<R} \alpha_{\mathrm{FMH}}(\theta,\theta') \vee \alpha_{\mathrm{FMH}}(\theta',\theta) > 0.$$

In this case, $\mathrm{Gap}(P_{\mathrm{FMH}}) \geq \delta \mathrm{Gap}(P_{\mathrm{MH}})$*, and for* $f \in L^2(\pi)$

$$\mathrm{var}(f, P_{\mathrm{TFMH}}) \leq (\delta^{-1} - 1)\mathrm{var}(f, \pi) + \delta^{-1}\mathrm{var}(f, P_{\mathrm{MH}}).$$

Proof. Our proof of this result is similar to (Banterle et al., 2015, Proposition 1), but differs in its use of Proposition B.1 to express the relationship between MH and FMH exactly.

For $\theta \in \Theta$, let

$$\mathcal{R}(\theta) := \{\theta' \in \Theta \mid \overline{\lambda}(\theta,\theta') < R\}.$$

Whenever $\theta \in \Theta$ and $A \subseteq \Theta$ is measurable,

$$\begin{aligned}
P_{\mathrm{TFMH}}(\theta, A) &= r_{\mathrm{TFMH}}(\theta)\mathbb{I}_A(\theta) + \int_{\mathcal{R}(\theta)\cap A} q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta')(\alpha_{\mathrm{FMH}}(\theta,\theta') \vee \alpha_{\mathrm{FMH}}(\theta',\theta))d\theta' \\
&\quad + \int_{\mathcal{R}(\theta)^c\cap A} q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta')d\theta' \\
&\geq r_{\mathrm{MH}}(\theta)\mathbb{I}_A(\theta) + \delta\int_{\mathcal{R}(\theta)\cap A} q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta')d\theta' + \int_{\mathcal{R}(\theta)^c\cap A} q(\theta,\theta')\alpha_{\mathrm{MH}}(\theta,\theta')d\theta' \\
&\geq \delta P_{\mathrm{MH}}(\theta, A).
\end{aligned}$$

The last line follows since certainly $\delta \leq 1$.

Suppose $\delta > 0$. If P_{MH} is geometrically ergodic, then (Jones et al., 2014, Theorem 1) entails that P_{TFMH} is geometrically ergodic also. The remaining claims follow directly from (Andrieu et al., 2018, Lemma 32). □

C. Fast Simulation of Bernoulli Random Variables

For sake of completeness, we provide here the proof of validity of Algorithm 1. It combines the Fukui-Todo procedure (Fukui & Todo, 2009) with a thinning argument.

Proposition C.1. *If*

$N \sim \mathrm{Poisson}\left(\overline{\lambda}(\theta,\theta')\right)$

$X_1, \cdots, X_N \overset{\mathrm{iid}}{\sim} \mathrm{Categorical}((\overline{\lambda}_i(\theta,\theta')/\overline{\lambda}(\theta,\theta'))_{1\leq i\leq m})$

$B_j \sim \mathrm{Bernoulli}(\lambda_{X_j}(\theta,\theta')/\overline{\lambda}_{X_j}(\theta,\theta'))$ *independently for* $1 \leq j \leq N$

then $\mathbb{P}(B=0) = \alpha_{\mathrm{FMH}}(\theta,\theta')$ *where* $B = \sum_{j=1}^N B_j$ *(and* $B = 0$ *if* $N = 0$*).*

Proof. Letting

$$\lambda(\theta,\theta') := \sum_{i=1}^m \lambda_i(\theta,\theta'),$$

our goal is to show that $\mathbb{P}(B=0) = \exp(-\lambda(\theta,\theta'))$. For brevity we omit all dependences on θ and θ' in the following.

Observe the random variables B_j's are i.i.d. with

$$\mathbb{P}(B_j = 0) = \sum_{i=1}^m \underbrace{\mathbb{P}(X_j = i)}_{=\overline{\lambda}_i/\overline{\lambda}} \underbrace{\mathbb{P}(B_j = 0 | X_j = i)}_{=1-\lambda_i/\overline{\lambda}_i} = \frac{\overline{\lambda} - \lambda}{\overline{\lambda}}.$$

Thus

$$\begin{aligned}\mathbb{P}(B=0) &= \sum_{\ell=0}^{\infty} \underbrace{\mathbb{P}(N=\ell)}_{=\exp(-\overline{\lambda})\overline{\lambda}^{\ell}/\ell!} \mathbb{P}(B_1=0)^{\ell} \\ &= \exp(-\overline{\lambda}) \sum_{\ell=0}^{\infty} \frac{(\overline{\lambda}-\lambda)^{\ell}}{\ell!} \\ &= \exp(-\lambda)\end{aligned}$$

as desired. □

D. Upper Bounds

We refer the reader to Section A for an explanation of multi-index notation β.

Proposition D.1. *If each U_i is $(k+1)$-times continuously differentiable with*

$$\overline{U}_{k+1,i} \geq \sup_{\substack{\theta\in\Theta \\ |\beta|=k+1}} |\partial^{\beta} U_i(\theta)|,$$

then

$$-\log\left(1 \wedge \frac{\pi_i(\theta')\widehat{\pi}_{k,i}(\theta)}{\pi_i(\theta)\widehat{\pi}_{k,i}(\theta')}\right) \leq (\|\theta-\widehat{\theta}\|_1^{k+1} + \|\theta'-\widehat{\theta}\|_1^{k+1}) \frac{\overline{U}_{k+1,i}}{(k+1)!}.$$

Proof. We have

$$\begin{aligned}-\log\left(1 \wedge \frac{\pi_i(\theta')\widehat{\pi}_{k,i}(\theta)}{\pi_i(\theta)\widehat{\pi}_{k,i}(\theta')}\right) &= 0 \vee (U_i(\theta') - \widehat{U}_{k,i}(\theta') - U_i(\theta) + \widehat{U}_{k,i}(\theta)) \\ &\leq |U_i(\theta') - \widehat{U}_{k,i}(\theta')| + |U_i(\theta) - \widehat{U}_{k,i}(\theta)|.\end{aligned}$$

Notice that $U_i(\theta) - \widehat{U}_{k,i}(\theta)$ is just the remainder of a Taylor expansion. As such, for each θ, Taylor's remainder theorem gives for some $\widetilde{\theta} \in \Theta$

$$\begin{aligned}|U_i(\theta) - \widehat{U}_{k,i}(\theta)| &= \left|\frac{1}{(k+1)!} \sum_{|\beta|=k+1} \partial^{\beta} U_i(\widetilde{\theta})(\theta-\widehat{\theta})^{\beta}\right| \\ &\leq \frac{\overline{U}_{k+1,i}}{(k+1)!} \sum_{|\beta|=k+1} \frac{|(\theta-\widehat{\theta})^{\beta}|}{\beta!} \\ &\leq \frac{\overline{U}_{k+1,i}}{(k+1)!} \|\theta-\widehat{\theta}\|_1^{k+1}.\end{aligned}$$

The result now follows. □

E. Reversible Proposals

E.1. General Conditions for Reversibility

We can handle both the first and second-order cases with the following Proposition.

Proposition E.1. *Suppose*

$$q(\theta,\theta') = \mathrm{Normal}(\theta' \mid A\theta + b, C)$$

and

$$-\log\hat{\pi}(\theta) = \frac{1}{2}\theta^{\top} D\theta + e^{\top}\theta + \mathrm{const}$$

where $A, C, D \in \mathbb{R}^{d\times d}$ with $C \succ 0$, and $b, e \in \mathbb{R}^d$. Then q is $\hat{\pi}$-reversible if and only if the following conditions hold:

$$\begin{aligned} A^\top C^{-1} &= C^{-1}A && \text{(E.1)}\\ A^2 &= I_d - CD && \text{(E.2)}\\ (A^\top + I_d)b &= -Ce, && \text{(E.3)} \end{aligned}$$

where $I_d \in \mathbb{R}^{d\times d}$ is the identity matrix.

Proof. Let

$$F(\theta, \theta') := -\log\hat{\pi}(\theta) - \log q(\theta, \theta').$$

Note that q is $\hat{\pi}$-reversible precisely when F is symmetric in its arguments. Since F is a polynomial of the form

$$F(\theta, \theta') = \frac{1}{2}\theta^\top J\theta + \frac{1}{2}\theta'^\top K\theta' + \theta^\top L\theta' + m^\top\theta + n^\top\theta' + \text{const}, \tag{E.4}$$

where $J, K, L \in \mathbb{R}^{d\times d}$ and $m, n \in \Theta$, then by equating coefficients it follows that $F(\theta, \theta') = F(\theta', \theta)$ precisely when

$$\begin{aligned} J &= K && \text{(E.5)}\\ L &= L^\top && \text{(E.6)}\\ m &= n. && \text{(E.7)} \end{aligned}$$

Now, we can expand

$$\begin{aligned} -\log q(\theta, \theta') &= \frac{1}{2}(\theta' - A\theta - b)^\top C^{-1}(\theta' - A\theta - b) + \text{const}\\ &= \frac{1}{2}\theta'^\top C^{-1}\theta' - (A\theta + b)^\top C^{-1}\theta' + \frac{1}{2}(A\theta + b)^\top C^{-1}(A\theta + b) + \text{const}\\ &= \frac{1}{2}\theta^\top A^\top C^{-1}A\theta + \frac{1}{2}\theta'^\top C^{-1}\theta' - \theta^\top A^\top C^{-1}\theta' + b^\top C^{-1}A\theta - b^\top C^{-1}\theta' + \frac{1}{2}b^\top C^{-1}b\\ &\quad +\text{const} \end{aligned}$$

Since $-\log q(\theta, \theta')$ must be the only source of terms in (E.4) containing both θ and θ', we see immediately that

$$L = -A^\top C^{-1},$$

and thus from (E.6) we have $-A^\top C^{-1} = -(C^{-1})^\top A$. Since $C \succ 0$, C^{-1} is symmetric and this condition becomes (E.1). Next we see that

$$\begin{aligned} J &= A^\top C^{-1}A + D\\ K &= C^{-1}, \end{aligned}$$

and from (E.5) and (E.1) we require $C^{-1}A^2 + D = C^{-1}$, or equivalently (E.2). Finally, since

$$\begin{aligned} m &= A^\top C^{-1}b + e\\ n &= -C^{-1}b, \end{aligned}$$

we require from (E.7) that $A^\top C^{-1}b + e = -C^{-1}b$, which combined with (E.1) gives (E.3).

Since (E.5), (E.6), and (E.7) are necessary and sufficient for symmetry of F, we see that (E.1), (E.2), and (E.3) are necessary and sufficient for reversibility also. □

We now specialise this to the first and second-order cases.

E.2. First-Order Case

When $k = 1$ we have

$$-\log \widehat{\pi}(\theta) = \widehat{U}_1(\theta) = U(\widehat{\theta}) + \nabla U(\widehat{\theta})^\top (\theta - \widehat{\theta}),$$

so that

$$\begin{aligned} D &= 0 \\ e &= \nabla U(\widehat{\theta}), \end{aligned}$$

and conditions (E.1), (E.2), and (E.3) become

$$\begin{aligned} A^\top C^{-1} &= C^{-1} A \\ A^2 &= I_d \\ (A^\top + I_d) b &= -C \nabla U(\widehat{\theta}). \end{aligned}$$

E.3. Second-Order Case

When $k = 2$,

$$-\log \widehat{\pi}(\theta) = \widehat{U}_2(\theta) = U(\widehat{\theta}) + \nabla U(\widehat{\theta})^\top (\theta - \widehat{\theta}) + \frac{1}{2}(\theta - \widehat{\theta})^\top \nabla^2 U(\widehat{\theta})(\theta - \widehat{\theta}).$$

In this case

$$\begin{aligned} D &= \nabla^2 U(\widehat{\theta}) \\ e &= \nabla U(\widehat{\theta}) - \nabla^2 U(\widehat{\theta})^\top \widehat{\theta}, \end{aligned}$$

so conditions (E.1), (E.2), and (E.3) become

$$\begin{aligned} A^\top C^{-1} &= C^{-1} A \\ A^2 &= I_d - C \nabla^2 U(\widehat{\theta}) \\ (A^\top + I_d) b &= C(\nabla^2 U(\widehat{\theta})^\top \widehat{\theta} - \nabla U(\widehat{\theta})). \end{aligned}$$

A common setting has $\nabla^2 U(\widehat{\theta}) \succ 0$, $A = A^\top$, and $A + I_d$ invertible. In this case the latter two conditions become

$$\begin{aligned} C &= (I_d - A^2)[\nabla^2 U(\widehat{\theta})]^{-1} \\ b &= (I_d - A)(\widehat{\theta} - [\nabla^2 U(\widehat{\theta})]^{-1} \nabla U(\widehat{\theta})). \end{aligned}$$

E.4. Decreasing Norm Property

Under usual circumstances for both first and second-order approximations, when $\|\theta\|$ is large, a $\widehat{\pi}$-reversible q will propose $\theta' \sim q(\theta, \cdot)$ with smaller norm than θ. This is made precise in the following Proposition:

Proposition E.2. *Suppose*

$$q(\theta, \theta') = \mathrm{Normal}(\theta' \mid A\theta + b, C)$$

and

$$-\log \hat{\pi}(\theta) = \frac{1}{2}\theta^\top D \theta + e^\top \theta + \mathrm{const},$$

where $A = A^\top$ is symmetric, $C \succ 0$, and $D \succeq 0$. If q is $\hat{\pi}$-reversible, then $\|A\|_{\mathrm{op}} \leq 1$. If $D \succ 0$ is strict, then $\|A\|_{\mathrm{op}} < 1$ is strict too. In this case, if $\theta' \sim q(\theta, \cdot)$, then $\|\theta\| - \|\theta'\| \to \infty$ in probability as $\|\theta\| \to \infty$.

Proof. By (E.2), we must have $CD = I_d - A^2$. Since $A = A^\top$, this entails $CD = (CD)^\top = DC$ and hence $CD \succeq 0$ since $D, C \succeq 0$. Thus $-CD \preceq 0$ and

$$A^2 = I_d - CD \preceq I_d.$$

Therefore each eigenvalue σ of A must have $|\sigma| \leq 1$, since σ^2 is an eigenvalue of A^2. But A is diagonalisable since it is symmetric, and hence $\|A\|_{\mathrm{op}} \leq 1$.

If $D \succ 0$ is strict, then the above matrix inequalities become strict also, and it follows that each $|\sigma| < 1$ and hence $\|A\|_{\mathrm{op}} < 1$. In this case, suppose $\theta' \sim q(\theta, \cdot)$, and fix $K > 0$ arbitrarily. Let $\epsilon > 0$, and choose $L > 0$ large enough that

$$\mathbb{P}(\theta' \in B(A\theta + b, L)) > 1 - \epsilon.$$

As $\|\theta\| \to \infty$,

$$\|\theta\| - \|A\theta + b\| \geq \|\theta\|(1 - \|A\|_{\mathrm{op}}) + \|b\| \to \infty$$

since $1 - \|A\|_{\mathrm{op}} > 0$, so if $\theta' \in B(A\theta + b, L)$, then $\|\theta\| - \|\theta'\| \to \infty$ also. Thus

$$\mathbb{P}(\|\theta\| - \|\theta'\| > K) > 1 - \epsilon$$

for all $\|\theta\|$ large enough. Taking $\epsilon \to 0$ gives the result. □

In practice the assumption $D \succeq 0$ makes sense, since $\widehat{\theta}$ is chosen near a minimum of U and since D is the Hessian of $\widehat{U}_k \approx U$ for $k = 1, 2$. Likewise, all sensible proposals (certainly including pCN) that we have found are such that A is symmetric, though we acknowledge the possibility that it may be desirable to violate this in some cases.

F. Performance Gains

Lemma F.1. *Suppose that $0 \leq X_n \in L^p$ and $\mathcal{F}_n$ is some σ-algebra for every $n \in \mathbb{Z}_{\geq 1}$. If $\mathbb{E}[X_n^p|\mathcal{F}_n] = O_{\mathbb{P}}(a_n)$, then $\mathbb{E}[X_n^\ell|\mathcal{F}_n] = O_{\mathbb{P}}(a_n^{\ell/p})$ for all $1 \leq \ell \leq p$. If moreover $0 \leq Y_n \in L^p$ gives $\mathbb{E}[Y_n^p|\mathcal{F}_n] = O_{\mathbb{P}}(a_n)$, then $\mathbb{E}[(X_n + Y_n)^p|\mathcal{F}_n] = O_{\mathbb{P}}(a_n)$.*

Proof. The first part is just Jensen's inequality:

$$\mathbb{E}[X_n^\ell|\mathcal{F}_n] \leq \mathbb{E}[X_n^p|\mathcal{F}_n]^{\ell/p} = O_{P_0}(a_n)^{\ell/p} = O_{P_0}(a_n^{\ell/p}).$$

The second part follows from the C_p-inequality, which gives

$$\mathbb{E}[(X_n + Y_n)^p|\mathcal{F}_n] \leq 2^{p-1}\left(\mathbb{E}[X_n^p|\mathcal{F}_n] + \mathbb{E}[Y_n^p|\mathcal{F}_n]\right) = 2^{p-1}(O_{\mathbb{P}}(a_n) + O_{\mathbb{P}}(a_n)) = O_{\mathbb{P}}(a_n).$$

□

Theorem F.1. *Suppose each U_i is $(k+1)$-times continuously differentiable, each $\overline{U}_{k+1,i} \in L^{k+2}$, and $\mathbb{E}[\sum_{i=1}^{m^{(n)}} \overline{U}_{k+1,i}|Y_{1:n}] = O_{P_0}(n)$. Likewise, assume each of $\|\theta^{(n)} - \theta_{\mathrm{MAP}}^{(n)}\|$, $\|\theta^{(n)} - \theta'^{(n)}\|$, and $\|\widehat{\theta}^{(n)} - \theta_{\mathrm{MAP}}^{(n)}\|$ is in L^{k+2}, and each of $\mathbb{E}[\|\theta^{(n)} - \theta_{\mathrm{MAP}}^{(n)}\|^{k+1}|Y_{1:n}]$, $\mathbb{E}[\|\theta^{(n)} - \theta'^{(n)}\|^{k+1}|Y_{1:n}]$, and $\|\widehat{\theta}^{(n)} - \theta_{\mathrm{MAP}}^{(n)}\|^{k+1}$ is $O_{P_0}(n^{-(k+1)/2})$ as $n \to \infty$. Then $\overline{\lambda}$ defined by (13) satisfies*

$$\mathbb{E}[\overline{\lambda}(\theta^{(n)}, \theta'^{(n)})|Y_{1:n}] = O_{P_0}(n^{(1-k)/2}).$$

Proof. Write

$$\overline{\lambda}(\theta^{(n)}, \theta'^{(n)}) = \varphi(\theta^{(n)}, \theta'^{(n)}) \sum_{i=1}^{m^{(n)}} \psi_i.$$

with φ and ψ defined by (13) also. Observe that

$$\begin{aligned}
\varphi(\theta^{(n)}, \theta'^{(n)}) &= \|\theta^{(n)} - \widehat{\theta}^{(n)}\|_1^{k+1} + \|\theta'^{(n)} - \widehat{\theta}^{(n)}\|_1^{k+1} \\
&\leq (\|\theta^{(n)} - \widehat{\theta}^{(n)}\|_1 + \|\theta'^{(n)} - \widehat{\theta}^{(n)}\|_1)^{k+1} \\
&\leq (\|\theta^{(n)} - \theta_{\mathrm{MAP}}^{(n)}\|_1 + \|\theta_{\mathrm{MAP}}^{(n)} - \widehat{\theta}^{(n)}\|_1 + \|\theta'^{(n)} - \theta^{(n)}\|_1 + \|\theta^{(n)} - \theta_{\mathrm{MAP}}^{(n)}\|_1 + \|\theta_{\mathrm{MAP}}^{(n)} - \widehat{\theta}^{(n)}\|_1)^{k+1} \\
&\leq c(\underbrace{\|\theta'^{(n)} - \theta^{(n)}\| + \|\theta^{(n)} - \theta_{\mathrm{MAP}}^{(n)}\| + \|\theta_{\mathrm{MAP}}^{(n)} - \widehat{\theta}^{(n)}\|}_{\in L^{k+2}})^{k+1}
\end{aligned}$$

for some $c > 0$, by the triangle inequality and norm equivalence. We thus have $\varphi(\theta^{(n)}, \theta'^{(n)}) \in L^{(k+2)/(k+1)}$ and

$$\mathbb{E}[\varphi(\theta^{(n)}, \theta'^{(n)})|Y_{1:n}] = O_{P_0}(n^{-(k+1)/2}).$$

Likewise,

$$\sum_{i=1}^{m^{(n)}} \psi_i = \frac{1}{(k+1)!} \sum_{i=1}^{m^{(n)}} \overline{U}_{k+1,i} \in L^{k+2}.$$

Together this gives $\overline{\lambda}(\theta^{(n)}, \theta'^{(n)}) \in L^1$ by Hölder's inequality. Since in our setup $(\theta^{(n)}, \theta'^{(n)})$ is conditionally independent of all other randomness given $Y_{1:n}$, we thus have

$$\mathbb{E}[\overline{\lambda}(\theta^{(n)}, \theta'^{(n)})|Y_{1:n}] = \mathbb{E}[\varphi(\theta^{(n)}, \theta'^{(n)})|Y_{1:n}]\mathbb{E}[\sum_{i=1}^{m^{(n)}} \psi_i|Y_{1:n}] = O_{P_0}(n^{(1-k)/2}). \tag{F.1}$$

□

Note that in the preceding result we could use weaker integrability assumptions on $\|\theta^{(n)} - \theta^{(n)}_{\mathrm{MAP}}\|$, $\|\theta^{(n)} - \theta'^{(n)}\|$, and $\|\widehat{\theta}^{(n)} - \theta^{(n)}_{\mathrm{MAP}}\|$ by using a stronger integrability assumption on $\overline{U}_{k+1,i}$. Most generally, for any $\epsilon \geq 0$ we could require each

$$\begin{aligned} \overline{U}_{k+1,i} &\in L^{(k+1+\epsilon)/\epsilon} \\ \|\theta^{(n)} - \theta^{(n)}_{\mathrm{MAP}}\|, \|\theta^{(n)} - \theta'^{(n)}\|, \|\widehat{\theta}^{(n)} - \theta^{(n)}_{\mathrm{MAP}}\| &\in L^{k+1+\epsilon}. \end{aligned}$$

The case $\epsilon = 0$ would mean $\overline{U}_{k+1,i} \in L^\infty$.

Lemma F.2. *Suppose each U_i is twice continuously differentiable, each $\overline{U}_{2,i} \in L^3$, and $\sum_{i=1}^{m^{(n)}} \overline{U}_{2,i} = O_{P_0}(n)$. If $\|\widehat{\theta}^{(n)} - \theta^{(n)}_{\mathrm{MAP}}\| = O_{P_0}(1/\sqrt{n})$, then $\|\nabla U^{(n)}(\widehat{\theta}^{(n)})\|$ is in $L^{3/2}$ and $O_{P_0}(\sqrt{n})$.*

Proof. By norm equivalence the Hessian satisfies

$$\|\nabla^2 U^{(n)}(\theta)\|_{\mathrm{op}} \leq c\|\nabla^2 U^{(n)}(\theta)\|_1 \leq c \sum_{i=1}^{m^{(n)}} \overline{U}_{2,i}$$

for some $c > 0$ (where $\|\cdot\|_1$ is understood to be applied as if $\nabla^2 U^{(n)}(\theta)$ were a vector), which means $\nabla U^{(n)}$ is $(c\sum_{i=1}^{m^{(n)}} \overline{U}_{2,i})$-Lipschitz. Thus

$$\begin{aligned} \|\nabla U^{(n)}(\widehat{\theta}^{(n)})\| &= \|\nabla U^{(n)}(\widehat{\theta}^{(n)}) - \nabla U^{(n)}(\theta^{(n)}_{\mathrm{MAP}})\| \\ &\leq \underbrace{c(\sum_{i=1}^{m^{(n)}} \overline{U}_{2,i})}_{\in L^3} \underbrace{\|\widehat{\theta}^{(n)} - \theta^{(n)}_{\mathrm{MAP}}\|}_{\in L^{k+2} \subseteq L^3} \end{aligned}$$

since $k \geq 1$. By Cauchy-Schwarz we have therefore $\|\nabla U^{(n)}(\widehat{\theta}^{(n)})\| \in L^{3/2}$.

Similarly, since $\widehat{\theta}^{(n)}$ and $\theta^{(n)}_{\mathrm{MAP}}$ are functions of $Y_{1:n}$,

$$\begin{aligned} \|\nabla U^{(n)}(\widehat{\theta}^{(n)})\| &= \mathbb{E}[\|\nabla U^{(n)}(\widehat{\theta}^{(n)}) - \nabla U^{(n)}(\theta^{(n)}_{\mathrm{MAP}})\||Y_{1:n}] \\ &\leq \mathbb{E}[c(\sum_{i=1}^{m^{(n)}} \overline{U}_{2,i})\|\widehat{\theta}^{(n)} - \theta^{(n)}_{\mathrm{MAP}}\||Y_{1:n}] \\ &= \underbrace{c\,\mathbb{E}[\sum_{i=1}^{m^{(n)}} \overline{U}_{2,i}|Y_{1:n}]}_{=O_{P_0}(n)} \underbrace{\|\widehat{\theta}^{(n)} - \theta^{(n)}_{\mathrm{MAP}}\|}_{=O_{P_0}(1/\sqrt{n})} \\ &= O_{P_0}(\sqrt{n}). \end{aligned}$$

□

Theorem F.2. *Suppose the assumptions of Theorem 3.1 hold, and additionally that for $2 \leq \ell \leq k$, each $\overline{U}_{\ell,i} \in L^{\ell+1}$, and $\mathbb{E}[\sum_{i=1}^{m^{(n)}} \overline{U}_{\ell,i} | Y_{1:n}] = O_{P_0}(n)$. Then*

$$-\log(1 \wedge \frac{\widehat{\pi}_k^{(n)}(\theta'^{(n)})}{\widehat{\pi}_k^{(n)}(\theta^{(n)})}) = O_{P_0}(1)$$

for all $k \geq 1$.

Proof. It is useful to denote

$$\begin{aligned} U^{(n)}(\theta) &:= \sum_{i=1}^{m^{(n)}} U_i(\theta) \\ \widehat{U}_k^{(n)}(\theta) &:= \sum_{i=1}^{m^{(n)}} \widehat{U}_{k,i}(\theta) = -\log(\widehat{\pi}^{(n)}(\theta)). \end{aligned}$$

Observe that

$$0 \leq -\log(1 \wedge \frac{\widehat{\pi}^{(n)}(\theta'^{(n)})}{\widehat{\pi}^{(n)}(\theta^{(n)})}) \leq |\widehat{U}_k^{(n)}(\theta'^{(n)}) - \widehat{U}_k^{(n)}(\theta^{(n)})|. \tag{F.2}$$

Now,

$$\widehat{U}_k^{(n)}(\theta'^{(n)}) - \widehat{U}_k^{(n)}(\theta^{(n)}) = \langle \nabla U^{(n)}(\widehat{\theta}^{(n)}), \theta'^{(n)} - \theta^{(n)} \rangle + \sum_{2 \leq |\beta| \leq k} \frac{\partial^\beta U^{(n)}(\widehat{\theta}^{(n)})}{\beta!} ((\theta'^{(n)} - \widehat{\theta}^{(n)})^\beta - (\theta^{(n)} - \widehat{\theta}^{(n)})^\beta). \tag{F.3}$$

For the first term here, Cauchy-Schwarz gives

$$\begin{aligned} \mathbb{E}[|\langle \nabla U^{(n)}(\widehat{\theta}^{(n)}), \theta'^{(n)} - \theta^{(n)} \rangle| | Y_{1:n}] &\leq \mathbb{E}[\underbrace{\|\nabla U^{(n)}(\widehat{\theta}^{(n)})\|}_{\in L^{3/2}} \underbrace{\|\theta'^{(n)} - \theta^{(n)}\|}_{\in L^{k+2} \subseteq L^3} | Y_{1:n}] \\ &= \underbrace{\|\nabla U^{(n)}(\widehat{\theta}^{(n)})\|}_{= O_{P_0}(\sqrt{n})} \underbrace{\mathbb{E}[\|\theta'^{(n)} - \theta^{(n)}\| | Y_{1:n}]}_{= O_{P_0}(1/\sqrt{n})} \\ &= O_{P_0}(1). \end{aligned}$$

Integrability follows from Lemma F.2 and Hölder's inequality, and the asymptotic statements from conditional independence, Lemma F.2, and Lemma F.1. For the summation in (F.3), note that

$$|\partial^\beta U^{(n)}(\widehat{\theta}^{(n)})| \leq \sum_{i=1}^{m^{(n)}} |\partial^\beta U_i(\widehat{\theta}^{(n)})| \leq \sum_{i=1}^{m^{(n)}} \overline{U}_{|\beta|,i},$$

and that for some $c > 0$,

$$\begin{aligned} |(\theta'^{(n)} - \widehat{\theta}^{(n)})^\beta - (\theta^{(n)} - \widehat{\theta}^{(n)})^\beta| &\leq \|\theta'^{(n)} - \widehat{\theta}^{(n)}\|_\infty^{|\beta|} + \|\theta^{(n)} - \widehat{\theta}^{(n)}\|_\infty^{|\beta|} \\ &\leq c\|\theta'^{(n)} - \widehat{\theta}^{(n)}\|^{|\beta|} + c\|\theta^{(n)} - \widehat{\theta}^{(n)}\|^{|\beta|} \end{aligned}$$

by norm equivalence. Thus, conditional on $Y_{1:n}$, the absolute value of the summation in (F.3) is bounded above by

$$\begin{aligned} &\sum_{2 \leq |\beta| \leq k} \frac{1}{\beta!} \mathbb{E}[(\underbrace{\sum_{i=1}^{m^{(n)}} \overline{U}_{|\beta|,i}}_{\in L^{|\beta|+1}})(c \underbrace{\|\theta'^{(n)} - \widehat{\theta}^{(n)}\|^{|\beta|}}_{\in L^{(|\beta|+1)/|\beta|}} + c \underbrace{\|\theta^{(n)} - \widehat{\theta}^{(n)}\|^{|\beta|}}_{\in L^{(|\beta|+1)/|\beta|}}) | Y_{1:n}] \\ = &\sum_{2 \leq |\beta| \leq k} \frac{c}{\beta!} \underbrace{\mathbb{E}[\sum_{i=1}^{m^{(n)}} \overline{U}_{|\beta|,i} | Y_{1:n}]}_{= O_{P_0}(n)} (\underbrace{\mathbb{E}[\|\theta'^{(n)} - \widehat{\theta}^{(n)}\|^{|\beta|} | Y_{1:n}]}_{= O_{P_0}(n^{-|\beta|/2})} + \underbrace{\mathbb{E}[\|\theta^{(n)} - \widehat{\theta}^{(n)}\|^{|\beta|} | Y_{1:n}]}_{= O_{P_0}(n^{-|\beta|/2})}) \\ = &\ O_{P_0}(1). \end{aligned}$$

Again, integrability follows from Hölder's inequality. The second line holds since $\widehat{\theta}^{(n)} \equiv \widehat{\theta}^{(n)}(Y_{1:n})$ and since $(\theta^{(n)}, \theta'^{(n)})$ is conditionally independent of all other randomness given $Y_{1:n}$. Finally, the asymptotics follow from the law of large numbers and Lemma F.1 (noting that each $|\beta| \geq 2$).

Inspection of (F.3) now shows that (F.3) is $O_{P_0}(1)$ as required. □

F.1. Sufficient Conditions

We are interested in sufficient conditions that guarantee the convergence rate assumptions in Theorem 3.1 will hold. For simplicity we assume throughout that the likelihood of a data point $p(y|\theta)$ admits a density w.r.t. Lebesgue measure and that P_0 also admits a Lebesgue density denoted $p_0(y)$.

F.1.1. Concentration Around the Mode

We first consider the assumption

$$\mathbb{E}[\|\theta^{(n)} - \theta^{(n)}_{\mathrm{MAP}}\|^{k+1}|Y_{1:n}] = O_{P_0}(n^{-(k+1)/2}).$$

Intuitively, this says that the distance of $\theta^{(n)}$ from the mode is $O(1/\sqrt{n})$, and hence connects directly with standard concentration results on Bayesian posteriors. To establish this rigorously, it is enough to show that for some $\theta^* \in \Theta$ both

$$\mathbb{E}[\|\theta^{(n)} - \theta^*\|^{k+1}|Y_{1:n}] = O_{P_0}(n^{-(k+1)/2}) \tag{F.4}$$

$$\mathbb{E}[\|\theta^{(n)}_{\mathrm{MAP}} - \theta^*\|^{k+1}|Y_{1:n}] = O_{P_0}(n^{-(k+1)/2}), \tag{F.5}$$

which entails the result by Lemma F.1 and the triangle inequality. Note that $\theta^{(n)}_{\mathrm{MAP}} \equiv \theta^{(n)}_{\mathrm{MAP}}(Y_{1:n})$ is deterministic function of the data, so that (F.5) may be written more simply as

$$\sqrt{n}(\theta^{(n)}_{\mathrm{MAP}} - \theta^*) = O_{P_0}(1). \tag{F.6}$$

We give sufficient conditions for (F.4) and (F.6) now.

By Proposition F.1 below, (F.4) holds as soon as we show that

$$\mathbb{E}[\|\sqrt{n}(\theta^{(n)} - \theta^*)\|^{k+1}\mathbb{I}(\|\sqrt{n}(\theta^{(n)} - \theta^*)\| > M_n)|Y_{1:n}] \xrightarrow{P_0} 0, \qquad \text{for all } M_n \to \infty. \tag{F.7}$$

This condition is a consequence of standard assumptions used to prove the Bernstein-von Mises theorem (BvM): in particular, it is (van der Vaart, 1998, (10.9)) when the model is well-specified (i.e. $p_0 = p(y|\theta_0)$ for some $\theta_0 \in \theta$), and (Kleijn & van der Vaart, 2012, (2.16)) in the misspecified case. In both cases

$$\theta^* = \arg\min_{\theta \in \Theta} D_{\mathrm{KL}}(p_0(y) \parallel p(y|\theta)),$$

where $D_{\mathrm{KL}}(\cdot \parallel \cdot)$ denotes the Kullback-Leibler divergence. The key assumption required for (F.7) is then the existence of certain test sequences $\phi_n \equiv \phi_n(Y_{1:n})$ with $0 \leq \phi_n \leq 1$ such that, whenever $\epsilon > 0$, both

$$\int \phi_n(y_{1:n}) \prod_{i=1}^n p_0(y_i) dy_{1:n} \to 0 \qquad \text{and} \qquad \sup_{\|\theta - \theta^*\| \geq \epsilon} \int (1 - \phi_n(y_{1:n})) \prod_{i=1}^n \frac{p(y_i|\theta)}{p(y_i|\theta^*)} p_0(y_i) dy_{1:n} \to 0, \tag{F.8}$$

Note that in the well-specified case these conditions say that ϕ_n is uniformly consistent for testing the hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \|\theta - \theta_0\| \geq \epsilon$. Since ϕ_n may have arbitrary form, this requirement does not seem arduous. Sufficient conditions are given by (van der Vaart, 1998, Lemma 10.4, Lemma 10.6) for the well-specified case, and (Kleijn & van der Vaart, 2012, Theorem 3.2) for the misspecified case.

In addition to (F.8), we require in both the well-specified and misspecified cases that the prior $p(\theta)$ be continuous and positive at θ^* and satisfy

$$\int \|\theta\|^{k+1} p(\theta) d\theta < \infty.$$

There are additionally some mild smoothness and regularity conditions imposed on the likelihood, which are naturally stronger in the misspecified case than in the well-specified one. In the well-specified case we require $p(y|\theta)$ is differentiable

in quadratic mean at θ^* (van der Vaart, 1998, (7.1)). In the misspecified case the conditions are more complicated. We omit repeating these for brevity and instead refer the reader to the statements of Lemma 2.1 and Theorem 3.1 in (Kleijn & van der Vaart, 2012).

Lemma F.3. *Suppose a sequence of random variables X_n is $O_{\mathbb{P}}(M_n)$ for every sequence $M_n \to \infty$. Then $X_n = O_{\mathbb{P}}(1)$.*

Proof. Suppose $X_n \neq O_{\mathbb{P}}(1)$. Then, for some $\epsilon > 0$, for every $c > 0$ we have $\mathbb{P}(|X_n| > c) \geq \epsilon$ for infinitely many X_n. This allows us to choose a subsequence X_{n_k} such that $\mathbb{P}(|X_{n_k}| > k) \geq \epsilon$ for each $k \in \mathbb{Z}_{\geq 1}$. Let

$$M_n := \begin{cases} k & \text{if } n = n_k \text{ for some (necessarily unique) } k \\ n & \text{otherwise.} \end{cases}$$

Then $M_n \to \infty$ but $\mathbb{P}(|X_n| > M_n) \geq \epsilon$ occurs for infinitely many n and hence $X_n \neq O_{\mathbb{P}}(M_n)$. □

Proposition F.1. *Suppose that for some $\theta^* \in \Theta$ and $\ell \geq 0$,*

$$\mathbb{E}[\|\sqrt{n}(\theta^{(n)} - \theta^*)\|^\ell \mathbb{I}(\|\sqrt{n}(\theta^{(n)} - \theta^*)\| > M_n)|Y_{1:n}] \xrightarrow{P_0} 0$$

whenever $M_n \to \infty$. Then

$$\mathbb{E}[\|\theta^{(n)} - \theta^*\|^\ell |Y_{1:n}] = O_{P_0}(n^{-\ell/2}).$$

Proof. For $M_n \to \infty$, our assumption lets us write

$$\begin{aligned} n^{\ell/2}\mathbb{E}[\|\theta^{(n)} - \theta^*\|^\ell |Y_{1:n}] &= \mathbb{E}[\|\sqrt{n}(\theta^{(n)} - \theta^*)\|^\ell \mathbb{I}(\|\sqrt{n}(\theta^{(n)} - \theta^*)\| \leq M_n)|Y_{1:n}] + o_{P_0}(1) \\ &\leq M_n^\ell + o_{P_0}(1) \\ &= O_{P_0}(M_n^\ell). \end{aligned}$$

Since M_n was arbitrary, Lemma F.3 entails the left-hand side is $O_{P_0}(1)$, so that

$$\mathbb{E}[\|\theta^{(n)} - \theta^*\|^\ell |Y_{1:n}] = O_{P_0}(n^{-\ell/2}).$$

□

It remains to give conditions for (F.6). Our discussion here is fairly standard. Recall that for $\theta^{(n)}_{\mathrm{MLE}}$ the maximum likelihood estimator,

$$\sqrt{n}(\theta^{(n)}_{\mathrm{MLE}} - \theta^*) = O_{P_0}(1)$$

often holds under under mild smoothness assumptions. We show here that effectively those same assumptions are also sufficient to guarantee a similar result for $\theta^{(n)}_{\mathrm{MAP}}$.

In the following we define

$$\mathcal{L}_n(\theta) := \frac{1}{n}\sum_{i=1}^n \log p(Y_i|\theta).$$

Note that by definition

$$\theta^{(n)}_{\mathrm{MLE}} = \sup_{\theta \in \Theta} \mathcal{L}_n(\theta).$$

Our first result here shows that if both the MAP and the MLE are consistent and the prior is well-behaved, then the MAP is a *near maximiser* of $\mathcal{L}_n$ in the sense that (F.9). Combined with mild smoothness assumptions on the likelihood, (F.9) is a standard condition used to show results such as (F.6). See for instance (van der Vaart, 1998, Theorem 5.23) for a detailed statement.

Proposition F.2. *Suppose for some $\theta^* \in \Theta$ that $\theta^{(n)}_{\mathrm{MAP}}, \theta^{(n)}_{\mathrm{MLE}} \xrightarrow{P_0} \theta^*$ and that the prior $p(\theta)$ is continuous and positive at θ^*, then*

$$\mathcal{L}_n(\theta^{(n)}_{\mathrm{MAP}}) \geq \mathcal{L}_n(\theta^{(n)}_{\mathrm{MLE}}) - o_{P_0}(1/n). \tag{F.9}$$

Proof. Observe that by definition of the MAP,

$$\mathcal{L}_n(\theta^{(n)}_{\mathrm{MLE}}) + \frac{1}{n}\log p(\theta^{(n)}_{\mathrm{MLE}}) \leq \mathcal{L}_n(\theta^{(n)}_{\mathrm{MAP}}) + \frac{1}{n}\log p(\theta^{(n)}_{\mathrm{MAP}}).$$

We can rewrite this inequality as

$$\mathcal{L}_n(\theta^{(n)}_{\mathrm{MAP}}) \geq \mathcal{L}_n(\theta^{(n)}_{\mathrm{MLE}}) + \frac{1}{n}\log \frac{p(\theta^{(n)}_{\mathrm{MLE}})}{p(\theta^{(n)}_{\mathrm{MAP}})}.$$

The second term on the right-hand side is $o_{P_0}(1/n)$, since our assumption on the prior gives

$$\frac{p(\theta^{(n)}_{\mathrm{MLE}})}{p(\theta^{(n)}_{\mathrm{MAP}})} \xrightarrow{P_0} 1.$$

□

We next consider how to show that the MAP is indeed consistent, as the vast majority of such results in this area only consider the MLE. However, assuming the prior is not pathological, arguments for the consistency of the MLE ought to apply also for the MAP, since the MAP optimises the objective function

$$\mathcal{L}_n(\theta) + \frac{1}{n}\log p(\theta),$$

which is asymptotically equivalent to $\mathcal{L}_n(\theta)$ as $n \to \infty$ whenever $p(\theta) > 0$. By way of example, we show that (van der Vaart, 1998, Theorem 5.7), which can be used to show the consistency of the MLE, also applies to the MAP. For this, we assume that

$$\int |\log p(y|\theta^*)| p_0(y) dy < \infty, \tag{F.10}$$

and define

$$\mathcal{L}(\theta) := \int \log p(y|\theta) p_0(y) dy.$$

Proposition F.3. *Suppose that* (F.10) *holds, that*

$$\sup_{\theta \in \Theta} |\mathcal{L}_n(\theta) - \mathcal{L}(\theta)| \xrightarrow{P_0} 0,$$

and that for some $\epsilon > 0$ *and* $\theta^* \in \Theta$

$$\sup_{\|\theta - \theta^*\| \geq \epsilon} \mathcal{L}(\theta) < \mathcal{L}(\theta^*). \tag{F.11}$$

Further, suppose the prior $p(\theta)$ *is continuous and positive at* θ^**, and that* $\sup_{\theta \in \Theta} p(\theta) < \infty$*. Then both* $\theta^{(n)}_{\mathrm{MLE}}, \theta^{(n)}_{\mathrm{MAP}} \xrightarrow{P_0} \theta^*$.

Proof. For each $\theta \in \Theta$ we have $\mathcal{L}_n(\theta) \xrightarrow{P_0} \mathcal{L}(\theta)$ as $n \to \infty$ by the law of large numbers, and thus $\theta^{(n)}_{\mathrm{MLE}} \xrightarrow{P_0} \theta^*$ by (van der Vaart, 1998, Theorem 5.7). Since $p(\theta)$ is continuous and positive at θ^*, this yields that

$$P_0(p(\theta^{(n)}_{\mathrm{MLE}}) > c) \to 1 \tag{F.12}$$

for some $c > 0$, as well as

$$\frac{1}{n}\log p(\theta^{(n)}_{\mathrm{MLE}}) = O_{P_0}(1/n).$$

Now, by maximality

$$\mathcal{L}_n(\theta^{(n)}_{\mathrm{MLE}}) + \frac{1}{n}\log p(\theta^{(n)}_{\mathrm{MLE}}) \leq \mathcal{L}_n(\theta^{(n)}_{\mathrm{MAP}}) + \frac{1}{n}\log p(\theta^{(n)}_{\mathrm{MAP}}) \leq \mathcal{L}_n(\theta^{(n)}_{\mathrm{MLE}}) + \frac{1}{n}\log p(\theta^{(n)}_{\mathrm{MAP}}).$$

Observe that it implies that $p(\theta^{(n)}_{\mathrm{MLE}}) \leq p(\theta^{(n)}_{\mathrm{MAP}})$. Together with (F.12) and our boundedness assumption on the prior, this gives

$$\frac{1}{n}\log p(\theta^{(n)}_{\mathrm{MAP}}) = O_{P_0}(1/n).$$

We can thus write

$$\mathcal{L}_n(\theta^{(n)}_{\mathrm{MAP}}) \geq \mathcal{L}_n(\theta^{(n)}_{\mathrm{MLE}}) + O_{P_0}(1/n).$$

The result now follows from (van der Vaart, 1998, Theorem 5.7). □

Observe that by negating (F.11) and adding the constant $\int p_0(y) \log p_0(y) dy$ to both sides, we see it is equivalent to the perhaps more intuitive condition

$$\inf_{\|\theta-\theta^*\|\geq\epsilon} D_{\mathrm{KL}}(p_0(y) \parallel p(y|\theta)) > D_{\mathrm{KL}}(p_0(y) \parallel p(y|\theta^*)).$$

F.1.2. Scaling of the Proposal

We now consider the assumption

$$\mathbb{E}[\|\theta^{(n)} - \theta'^{(n)}\|^{k+1}|Y_{1:n}] = O_{P_0}(n^{-(k+1)/2}). \tag{F.13}$$

Intuitively this holds if we scale our proposal like $1/\sqrt{n}$. We consider here proposals based on a noise distribution $\xi^{(n)} \overset{\text{iid}}{\sim} \mathrm{Normal}(0, I_d)$, but generalisations are possible. We immediately obtain (F.13) for instance with the scaled random walk proposal (15), for which

$$\theta'^{(n)} = \theta^{(n)} + \frac{\sigma}{\sqrt{n}}\xi^{(n)}.$$

Similarly, the $\widehat{\pi}_1$-reversible proposal defined by (17) has

$$\theta'^{(n)} = \theta^{(n)} - \frac{1}{2n}\nabla U^{(n)}(\widehat{\theta}^{(n)}) + \frac{\sigma}{\sqrt{n}}\xi^{(n)},$$

with $\xi^{(n)} \overset{\text{iid}}{\sim} \mathrm{Normal}(0, I_d)$. If the conditions of Lemma F.2 hold, then the second term is $O_{P_0}(1/\sqrt{n})$ and (F.13) follows.

More generally we can consider trying to match the covariance of our noise to the covariance of our target. Intuitively, under usual circumstances, $[\nabla^2 U^{(n)}(\widehat{\theta}^{(n)})]^{-1}$ is approximately proportional to the inverse observed Fisher information at θ^*, and hence preconditioning $\xi^{(n)}$ by $S^{(n)}$ such that

$$S^{(n)}{S^{(n)}}^\top = [\nabla^2 U^{(n)}(\widehat{\theta}^{(n)})]^{-1}$$

matches our proposal to the characteristics of the target. Such an $S^{(n)}$ can be computed for instance via a Cholesky decomposition.

Under usual circumstances this achieves a correctly scaled proposal. In particular, if

$$\widehat{\theta}^{(n)} \quad \overset{P_0}{\to} \quad \theta^* \tag{F.14}$$

$$\frac{1}{n}\partial_j\partial_k U^{(n)}(\theta^*) \quad \overset{P_0}{\to} \quad \mathcal{I}_{j,k} \tag{F.15}$$

for some constants $\mathcal{I}_{j,k}$, then Proposition F.4 below entails $\|S^{(n)}\|_{\mathrm{op}} = O_{P_0}(1/\sqrt{n})$. Thus (F.13) holds for the preconditioned random walk proposal (16) for which

$$\theta'^{(n)} = \theta^{(n)} + S^{(n)}\xi^{(n)},$$

since

$$\|S^{(n)}\xi^{(n)}\| \leq \|S^{(n)}\|_{\mathrm{op}}\|\xi^{(n)}\| = O_{P_0}(1/\sqrt{n}). \tag{F.16}$$

The same is also true a pCN proposal. In this case

$$\theta'^{(n)} - \theta^{(n)} = (\sqrt{\rho} - 1)(\theta^{(n)} - \widehat{\theta}^{(n)}) + (\sqrt{\rho} - 1)([\nabla^2 U^{(n)}(\widehat{\theta}^{(n)})]^{-1}\nabla U^{(n)}(\widehat{\theta}^{(n)})) + \sqrt{1-\rho}S^{(n)}\xi^{(n)}.$$

Note that here the first term satisfies

$$E[\|\theta^{(n)} - \widehat{\theta}^{(n)}\|^3|Y_{1:n}] = O_{P_0}(n^{-3/2}),$$

while the remaining two terms are $O_{P_0}(1/\sqrt{n})$ by Lemma F.2 and (F.16). This gives F.13 by Lemma F.1.

Condition (F.14) holds for instance under the assumptions of Theorem 3.1 and provided concentration around θ^* of the kind described in Section F.1.1 occurs. Condition (F.15) will also often hold in practice. For instance, if

$$U^{(n)}(\theta) = -\log p(\theta) - \sum_{i=1}^{n} \log p(Y_i|\theta),$$

and if the prior is positive at θ^*, then for all $1 \leq j, k \leq d$ the law of large numbers gives

$$\begin{aligned} \frac{1}{n}\partial_j\partial_k U^{(n)}(\theta^*) &= -\frac{1}{n}\partial_j\partial_k \log p(\theta^*) - \frac{1}{n}\sum_{i=1}^{n} \partial_j\partial_k \log p(Y_i|\theta^*) \\ &\xrightarrow{P_0} -\int \partial_j\partial_k \log p(y|\theta^*)p_0(y)dy \end{aligned}$$

when the derivatives and the integral exists. More generally our model may be specified conditional on i.i.d. covariates X_i so that

$$U^{(n)}(\theta) = -\log p(\theta) - \sum_{i=1}^{n} \log p(Y_i|\theta, X_i) + \log p(X_i),$$

in which case the same argument still applies. (Note that here abuse notation by considering our data $Y_i \equiv (X_i, Y_i)$, where the right-hand Y_i are response variables.)

Proposition F.4. *Suppose for some $\theta^* \in \Theta$ we have $\widehat{\theta}^{(n)} \xrightarrow{P_0} \theta^*$ and*

$$\frac{1}{n}\partial_j\partial_k U^{(n)}(\theta^*) \xrightarrow{P_0} \mathcal{I}_{jk}$$

for all $1 \leq j, k \leq d$. Suppose moreover that each $\overline{U}_{3,i} \in L^1$ and each $\nabla^2 U^{(n)}(\widehat{\theta}^{(n)}) \succ 0$. If $[\nabla^2 U^{(n)}(\widehat{\theta}^{(n)})]^{-1} = S^{(n)}S^{(n)\top}$ for some $S^{(n)} \in \mathbb{R}^{d\times d}$, then

$$\|S^{(n)}\|_{\mathrm{op}} = O_{P_0}(1/\sqrt{n}).$$

Proof. Suppose $|\beta| = 2$. Note that since for each i and θ

$$\|\nabla\partial^\beta U_i(\theta)\| \leq c\|\nabla\partial^\beta U_i(\theta)\|_1 = c\sum_{j=1}^{d} |\partial_j\partial^\beta U_i(\theta)| \leq cd\overline{U}_{3,i},$$

for some $c > 0$ by norm equivalence, it follows that $\partial^\beta U_i$ is $cd\overline{U}_{3,i}$-Lipschitz. Consequently for each θ

$$\left|\frac{1}{n}\partial^\beta U^{(n)}(\theta) - \frac{1}{n}\partial^\beta U^{(n)}(\theta^*)\right| \leq \frac{1}{n}\sum_{i=1}^{m^{(n)}} |\partial^\beta U_i(\theta) - \partial^\beta U_i(\theta^*)| \leq \frac{1}{n}(\sum_{i=1}^{m^{(n)}} \overline{U}_{3,i})cd\|\theta - \theta^*\|.$$

Thus given $K, \eta > 0$

$$\begin{aligned} \mathbb{P}\left(\sup_{\|\theta-\theta^*\|<K} \left|\frac{1}{n}\partial^\beta U^{(n)}(\theta) - \frac{1}{n}\partial^\beta U^{(n)}(\theta^*)\right| > \eta\right) &\leq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{m^{(n)}} \overline{U}_{3,i} > \eta c^{-1}d^{-1}K^{-1}\right), \\ &\leq \frac{\mathbb{E}[\overline{U}_{3,i}]}{\eta c^{-1}d^{-1}K^{-1}}, \end{aligned}$$

by Markov's inequality. It is clear that given any $\eta > 0$ the right-hand side can be made arbitrarily small by taking $K \to 0$, which yields $n^{-1}\partial^\beta U^{(n)}(\theta)$ is stochastic equicontinuous at θ^*, and consequently that

$$\frac{1}{n}\partial^\beta U^{(n)}(\widehat{\theta}^{(n)}) - \frac{1}{n}\partial^\beta U^{(n)}(\theta^*) \xrightarrow{P_0} 0,$$

see (Pollard, 2012, page 139).

Define the matrix $\mathcal{I} \in \mathbb{R}^{d\times d}$ by the constants $\mathcal{I}_{jk}$. We thus have

$$\frac{1}{n}\nabla^2 U^{(n)}(\widehat{\theta}^{(n)}) \xrightarrow{P_0} \mathcal{I}$$

since it converges element-wise. Thus by the continuous mapping theorem

$$n\|\nabla^2 U^{(n)}(\widehat{\theta}^{(n)})^{-1}\|_{\mathrm{op}} \xrightarrow{P_0} \|\mathcal{I}^{-1}\|_{\mathrm{op}},$$

from which it follows that

$$\|[\nabla^2 U^{(n)}(\widehat{\theta}^{(n)})]^{-1}\|_{\mathrm{op}} = O_{P_0}(1/n).$$

It is a standard result from linear algebra that

$$\|[\nabla^2 U^{(n)}(\widehat{\theta}^{(n)})]^{-1}\|_{\mathrm{op}} = \|S^{(n)}\|^2_{\mathrm{op}},$$

which gives the result. □

G. Applications

We give here the results of applying our method to a logistic regression and a robust linear regression example. In both cases we write our covariates as x_i and responses as y_i, and our target is the posterior

$$\pi(\theta) = p(\theta|x_{1:n}, y_{1:n}) \propto p(\theta)\prod_{i=1}^{n} p(y_i|\theta, x_i).$$

G.1. Logistic Regression

In this case we have $x_i \in \mathbb{R}^d$, $y_i \in \{0, 1\}$, and

$$p(y_i|\theta, x_i) = \mathrm{Bernoulli}(y_i|\frac{1}{1+\exp(-\theta^\top x_i)}).$$

For simplicity we assume a flat prior $p(\theta) \equiv 1$, which allows factorising π like (8) with $m = n$ and $\widetilde{\pi}_i(\theta) = p(y_i|\theta, x_i)$. It is then easy to show that

$$U_i(\theta) = -\log\widetilde{\pi}_i(\theta) = \log(1+\exp(\theta^T x_i)) - y_i\theta^\top x_i.$$

We require upper bounds $\overline{U}_{k+1,i}$ of the form (12) for these terms. For this we let $\sigma(z) = 1/(1+\exp(-z))$ and note the identity $\sigma'(z) = \sigma(z)(1-\sigma(z))$, which entails

$$\partial_j\sigma(\theta^\top x_i) = -x_{ij}(\sigma(\theta^\top x_i) - \sigma(\theta^\top x_i)^2).$$

We then have

$$\begin{aligned}
\partial_j U_i(\theta) &= x_{ij}(\sigma(\theta^\top x_i) - y_i)\\
\partial_k\partial_j U_i(\theta) &= x_{ij}x_{ik}(\sigma(\theta^\top x_i) - \sigma(\theta^\top x_i)^2)\\
\partial_\ell\partial_k\partial_j U_i(\theta) &= x_{ij}x_{ik}(x_{i\ell}(\sigma(\theta^\top x_i) - \sigma(\theta^\top x_i)^2) - 2\sigma(\theta^\top x_i)x_{i\ell}(\sigma(\theta^\top x_i) - \sigma(\theta^\top x_i)^2))\\
&= x_{ij}x_{ik}x_{i\ell}(\sigma(\theta^\top x_i) - \sigma(\theta^\top x_i)^2)(1 - 2\sigma(\theta^\top x_i)).
\end{aligned}$$

It is possible to show that (whether $y_i = 0$ or $y_1 = 1$)

$$\begin{aligned}
\sup_{t\in\mathbb{R}}|\sigma(t) - y_i| &= 1\\
\sup_{t\in\mathbb{R}}|\sigma(t) - \sigma(t)^2| &= \frac{1}{4}\\
\sup_{t\in\mathbb{R}}|(\sigma(t) - \sigma(t)^2)(1 - 2\sigma(t))| &= \frac{1}{6\sqrt{3}}.
\end{aligned}$$

Thus setting

$$\begin{aligned}\overline{U}_{1,i} &:= \max_{1\le j\le d}|x_{ij}| \\ \overline{U}_{2,i} &:= \frac{1}{4}\max_{1\le j\le d}|x_{ij}|^2 \\ \overline{U}_{3,i} &:= \frac{1}{6\sqrt{3}}\max_{1\le j\le d}|x_{ij}|^3\end{aligned}$$

satisfies (12).

In Figure 1 we compare the histogram of the samples of the first coordinate θ_1 to the marginal of the Gaussian approximation. This is done for $n = 2048$, the smallest data size for which we saw a significant ESS improvement of SMH-2 over MH, and for larger n showing the convergence of the Gaussian approximation and the posterior.

In Figure 2 we demonstrate the performance of the algorithm where θ is of dimension 20. The results are qualitatively similar to the 10-dimensional case in that SMH-2 eventually performs better than MH as the number of data increases. However for the 20-dimensional model SMH-2 yields superior performance to MH around the point at which n exceeds 32768, whereas in the 10-dimensional model this happens for n exceeding 2048.

G.2. Robust Linear Regression

Here $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. We use a flat prior $p(\theta) \equiv 1$, and the likelihood is given by

$$p(y_i|\theta, x_i) \quad = \quad \text{Student}(y_i - \theta^\top x_i \mid \nu).$$

Here $\text{Student}(\nu)$ denotes the Student-t distribution with ν degrees of freedom that the user will specify. This gives

$$U_i(\theta) = \frac{\nu+1}{2}\log\left(1 + \frac{(y_i - \theta^\top x_i)^2}{\nu}\right).$$

To derive bounds necessary for (12), let $\phi_i(\theta) := y_i - \theta^T x_i$ and note that $\partial_j \phi_i(\theta) = -x_{ij}$. Then we have

$$\begin{aligned}U_i(\theta) &= \frac{\nu+1}{2}\log\left(1 + \frac{\phi_i(\theta)^2}{\nu}\right) \\ \partial_j U_i(\theta) &= -(\nu+1)x_{ij}\frac{\phi_i(\theta)}{\nu + \phi_i(\theta)^2} \\ \partial_k\partial_j U_i(\theta) &= -(\nu+1)x_{ij}\frac{-x_{ik}(\nu+\phi_i(\theta)^2) + 2x_{ik}\phi_i(\theta)^2}{\nu+\phi_i(\theta)} \\ &= (\nu+1)x_{ij}x_{ik}\frac{\nu - \phi_i(\theta)^2}{(\nu+\phi_i(\theta)^2)^2} \\ \partial_\ell\partial_k\partial_j U_i(\theta) &= (\nu+1)x_{ij}x_{ik}\frac{2x_{i\ell}\phi_i(\theta)(\nu+\phi_i(\theta)^2)^2 + 4x_{i\ell}(\nu - \phi_i(\theta)^2)(\nu+\phi_i(\theta)^2)\phi_i(\theta)}{(\nu+\phi_i(\theta)^2)^4} \\ &= -2(\nu+1)x_{ij}x_{ik}x_{i\ell}\frac{\phi_i(\theta)(\phi_i(\theta)^2 - 3\nu)}{(\nu+\phi_i(\theta)^2)^3}\end{aligned}$$

In general,

$$\begin{aligned}\sup_{t\in\mathbb{R}}\left|\frac{t}{\nu+t^2}\right| &= \frac{1}{2\sqrt{\nu}} \\ \sup_{t\in\mathbb{R}}\left|\frac{\nu - t^2}{(\nu+t^2)^2}\right| &= \frac{1}{\nu} \\ \sup_{t\in\mathbb{R}}\left|\frac{t(t^2-3\nu)}{(\nu+t^2)^3}\right| &= \frac{3+2\sqrt{2}}{8\nu^{3/2}},\end{aligned}$$

so setting

$$\begin{aligned}\overline{U}_{1,i} &:= \frac{\nu+1}{2\sqrt{\nu}} \max_{1\le j\le d} |x_{ij}| \\ \overline{U}_{2,i} &:= \frac{\nu+1}{\nu} \max_{1\le j\le d} |x_{ij}|^2 \\ \overline{U}_{3,i} &:= \frac{(\nu+1)(3+2\sqrt{2})}{4\nu^{3/2}} \max_{1\le j\le d} |x_{ij}|^3\end{aligned}$$

satisfies (12).

In Figure 3 we show effective sample size (ESS) per second for the robust linear regression model; this experiment mimics the conditions of Figure 2 in the main text, where we used a logistic regression model. The performance for this model is qualitatively similar to that for logistic regression. Figures 4 and 5 show the ESS and acceptance rate for pCN proposals as ρ is varied. These mimic Figures 3 and 4 in the main text. For these experiments we use synthetic data, taking an $n \times 10$ matrix X with elements drawn independently from a standard normal distribution, and simulate $y_i = \sum_j X_{ij} + \epsilon$ where ϵ itself is drawn from a standard normal distribution. We choose as the model parameter $\nu = 4.0$.

References

Andrieu, C., Lee, A., and Vihola, M. Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli*, 24(2):842–872, 2018.

Banterle, M., Grazian, C., Lee, A., and Robert, C. P. Accelerating Metropolis–Hastings algorithms by delayed acceptance. *arXiv preprint arXiv:1503.00996*, 2015.

Fukui, K. and Todo, S. Order-n cluster Monte Carlo method for spin systems with long-range interactions. *Journal of Computational Physics*, 228(7):2629–2642, 2009.

Geyer, C. J. Markov chain Monte Carlo lecture notes. 1998. URL http://www.stat.umn.edu/geyer/f05/8931/n1998.pdf.

Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. URL http://dx.doi.org/.

Jones, G. L., Roberts, G. O., and Rosenthal, J. S. Convergence of conditional Metropolis–Hastings samplers. *Advances in Applied Probability*, 46(2):422445, 2014. doi: 10.1239/aap/1401369701.

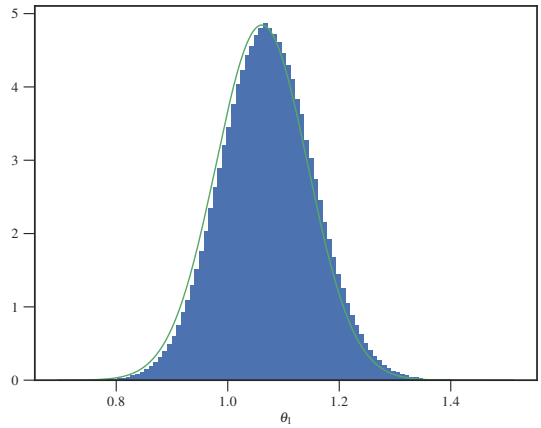
Kleijn, B. and van der Vaart, A. The Bernstein-Von-Mises theorem under misspecification. *Electron. J. Statist.*, 6:354–381, 2012. doi: 10.1214/12-EJS675. URL https://doi.org/10.1214/12-EJS675.

Meyn, S. and Tweedie, R. L. *Markov Chains and Stochastic Stability*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009. ISBN 0521731828, 9780521731829.

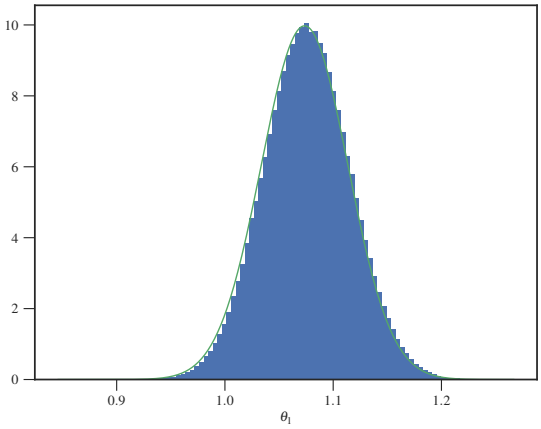
Pollard, D. *Convergence of Stochastic Processes*. Springer Series in Statistics. Springer New York, 2012. ISBN 9781461252542. URL https://books.google.co.uk/books?id=g5DbBwAAQBAJ.

Roberts, G. and Rosenthal, J. Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, 2:13–25, 1997. doi: 10.1214/ECP.v2-981. URL https://doi.org/10.1214/ECP.v2-981.

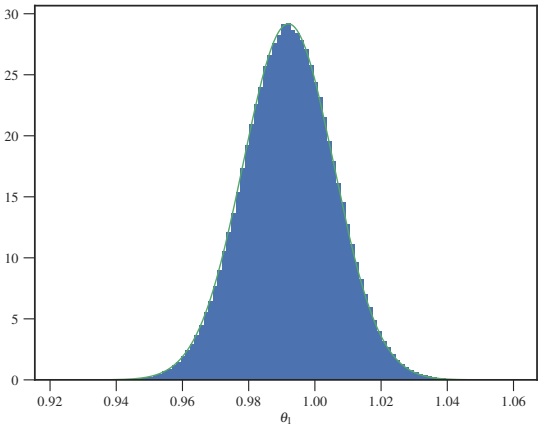
van der Vaart, A. W. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.



(a) $n = 2048$



(b) $n = 8192$



(c) $n = 65536$

Figure 1. Histogram of samples of first regression coefficient (θ_1) versus marginal of Gaussian approximation (green lines).

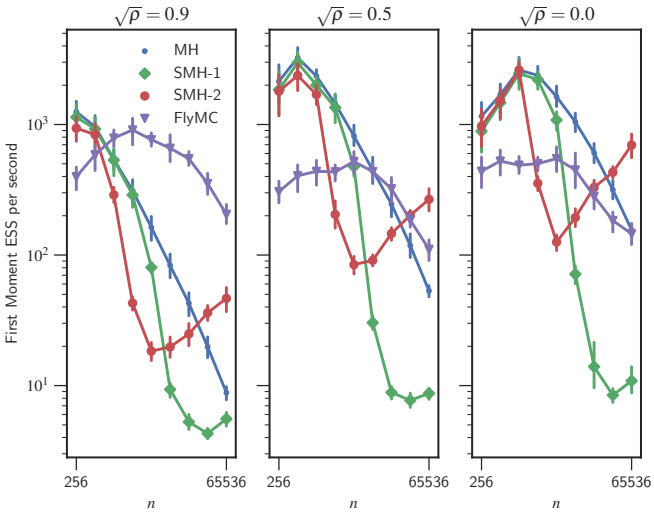


Figure 2. ESS of first regression coefficient for a logistic regression model of dimension 20.

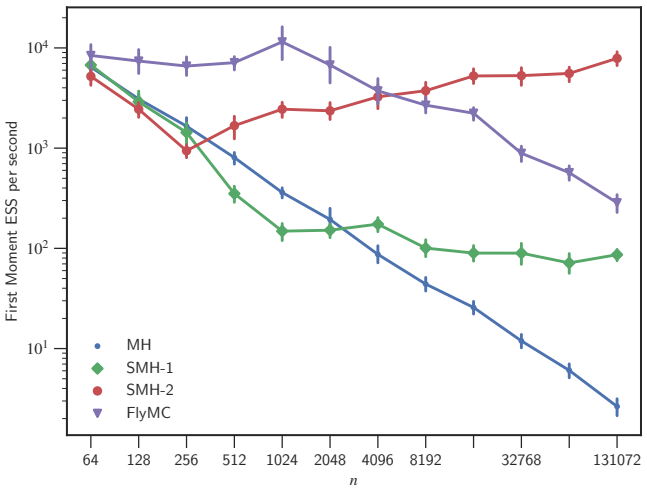


Figure 3. ESS for first regression coefficient of a robust linear regression posterior, scaled by execution time (higher is better).

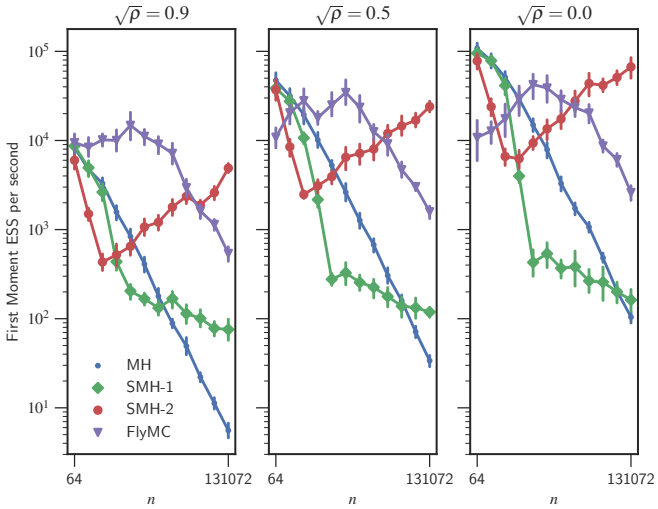


Figure 4. Effect of ρ on ESS for first regression coefficient of the robust linear regression model, scaled by execution time (higher is better).

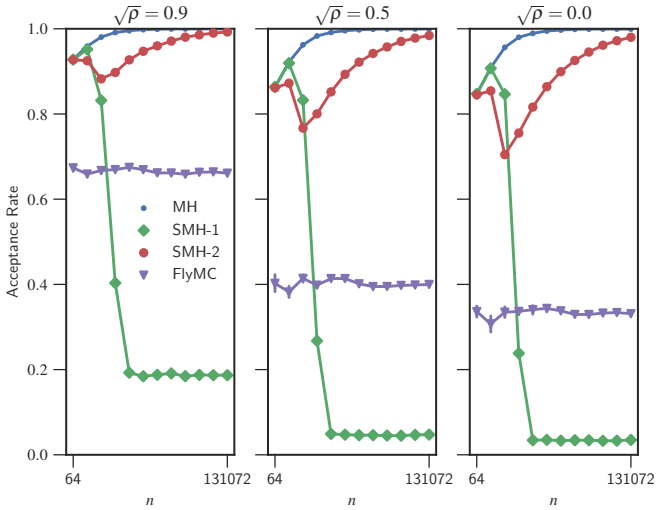


Figure 5. Acceptance rates for pCN proposals for the robust linear regression model.