A Local polynomial regression

Here we provide additional detail about the polynomial regression scheme described in Section 2.2. We consider the quadratic case, as the linear case is a simple restriction thereof. For each component f_j of $\mathbf{f}$, the quadratic regressor is of the form

$$\tilde{f}_j(\hat{\theta}) := a_j + b_j^T \hat{\theta} + \frac{1}{2}\hat{\theta}^T H_j \hat{\theta},$$

where $a_j \in \mathbb{R}$ is a constant term, $b_j \in \mathbb{R}^d$ is a linear term, and $H_j \in \mathbb{R}^{d\times d}$ is a symmetric Hessian matrix. Note that a_j, b_j, and H_j collectively contain $M = (d+2)(d+1)/2$ independent entries for each j. The coordinates $\hat{\theta} \in \mathbb{R}^d$ are obtained by shifting and scaling the original parameters θ as follows. Recall that the local regression scheme uses N samples $\{\theta^1, \ldots, \theta^N\}$ drawn from the ball of radius R centered on the point of interest θ, along with the corresponding model evaluations $y_j^i = f_j(\theta^i)$.[11] We assume that the components of θ have already been scaled so that they are of comparable magnitudes, then define $\hat{\theta}^i = (\theta^i - \theta)/R$, so that the transformed samples are centered at zero and have maximum radius one. Writing the error bounds as in (1) requires this rescaling along with the $1/2$ in the form of the regressor above (Conn et al., 2009).

Next, construct the diagonal weight matrix $W = \mathrm{diag}(w^1, \ldots, w^N)$ using the sample weights in (2), where we have $R = 1$ because of the rescaling. Then compute the N-by-M basis matrix Φ:

$$\Phi = \begin{pmatrix} 1 & \hat{\theta}_1^1 & \cdots & \hat{\theta}_d^1 & \frac{1}{2}\left(\hat{\theta}_1^1\right)^2 & \cdots & \frac{1}{2}\left(\hat{\theta}_d^1\right)^2 & \hat{\theta}_1^1\hat{\theta}_2^1 & \cdots & \hat{\theta}_{d-1}^1\hat{\theta}_d^1 \\ \vdots & & & & & & & & & \vdots \\ 1 & \hat{\theta}_1^N & \cdots & \hat{\theta}_d^N & \frac{1}{2}\left(\hat{\theta}_1^N\right)^2 & \cdots & \frac{1}{2}\left(\hat{\theta}_d^N\right)^2 & \hat{\theta}_1^N\hat{\theta}_2^N & \cdots & \hat{\theta}_{d-1}^N\hat{\theta}_d^N \end{pmatrix}$$

where we ensure that $N > M$. Finally, solve the n least squares problems,

$$\Phi^T W \Phi Z = \Phi^T W Y, \tag{8}$$

where each column of the N-by-n matrix Y contains the samples $\left(y_j^1, \ldots, y_j^N\right)^T$, $j = 1, \ldots, n$. Each column z_j of $Z \in \mathbb{R}^{M\times n}$ contains the desired regression coefficients for output j,

$$z_j^T = \begin{pmatrix} a_j & b_j^T & (H_j)_{1,1} & \cdots & (H_j)_{d,d} & (H_j)_{1,2} & \cdots & (H_j)_{d-1,d} \end{pmatrix}. \tag{9}$$

[11] To avoid any ambiguities, this appendix departs from the rest of the narrative by using a superscript to index samples and a subscript to index coordinates.

The least squares problem may be solved in a numerically stable fashion using a QR factorization of $W\Phi Z$, which may be computed once and reused for all n least squares problems. The cross-validation fit omitting sample i simply removes row i from both sides of (8). These least squares problems can be solved efficiently with a low-rank update of the QR factorization of the full least squares problem, rather than recomputing the QR factors from scratch (Hammarling and Lucas, 2008).

B Detailed theoretical results and proofs of theorems

B.1 Auxiliary notation

We now define some useful auxillary objects. For a fixed finite set $\mathcal{S} \subset \Theta$, we consider the stochastic process defined by Algorithm 4 with $\mathcal{S}_1 = \mathcal{S}$ and lines 11–20 and 22 removed. This process is essentially the original algorithm with all approximations based on a single set of points $\mathcal{S}$ and no refinements. Since there are no refinements, this process is in fact a Metropolis-Hastings Markov chain, and we write $K_{\mathcal{S}}$ for its transition kernel. For all measurable sets $U \subset \Theta$, this kernel can be written as $K_{\mathcal{S}}(x, U) = r_{\mathcal{S}}(x)\delta_x(U) + (1 - r_{\mathcal{S}}(x)) \int_{y \in U} p_{\mathcal{S}}(x, y)dy$ for some $0 \leq r_{\mathcal{S}}(x) \leq 1$ and density $p_{\mathcal{S}}(x, y)$. We denote by $\alpha_{\mathcal{S}}(x, y)$ the acceptance probability of $K_{\mathcal{S}}$.

We introduce another important piece of notation before giving our results. Let $\{Z_t\}_{t\in\mathbb{N}}$ be a (generally non-Markovian) stochastic process on some state space Ω. We say that a sequence of (generally random, dependent) kernels $\{Q_t\}_{t\in\mathbb{N}}$ is *adapted* to $\{Z_t\}_{t\in\mathbb{N}}$ if there exists an auxillary process $\{A_t\}_{t\in\mathbb{N}}$ so that:

- $\{(Z_t, A_t)\}_{t\in\mathbb{N}}$ is a Markov chain,
- Q_t is $\sigma(A_t)$-measurable, and
- $\mathbb{P}[Z_{t+1} \in \cdot | Z_t, A_t] = Q_t(Z_t, \cdot)$.

Let $\{X_t, \mathcal{S}_t\}_{t\in\mathbb{N}}$ be a sequence evolving according to the stochastic process defined by Algorithm 4 and define the following associated sequence of kernels:

$$\tilde{K}_t(x, A) \equiv \mathbb{P}[X_{t+1} \in A | \{X_s\}_{1\leq s<t}, X_t = x, \{\mathcal{S}_s\}_{1\leq s\leq t}].$$

The sequence of kernels $\{\tilde{K}_t\}_{t\in\mathbb{N}}$ is adapted to $\{X_t\}_{t\in\mathbb{N}}$, with $\{\mathcal{S}\}_{t\in\mathbb{N}}$ as the auxillary process. For any fixed t, one can sample from $\tilde{K}_t(x, \cdot)$ by first drawing a proposal y from $L(x, \cdot)$ and then accepting

with probability

$$\tilde{\alpha}_t(x,y) \equiv c_1\alpha_{\mathcal{S}_t}(x,y) + c_2\alpha_{\mathcal{S}_t\cup\{(x,f(x))\}}(x,y) + c_3\alpha_{\mathcal{S}_t\cup\{(y,f(y))\}}(x,y), \tag{10}$$

where c_1, c_2, c_3 are some positive constants that depend on x, y, β_t and γ_t and satisfy the identity $c_1 + c_2 + c_3 = 1$.

B.2 Book-keeping result

The following result will be used repeatedly in our ergodicity arguments.

Theorem B.1 (Approximate Ergodicity of Adaptive Chains)**.** *Fix a kernel K with stationary distribution π on state space $\mathcal{X}$ and let $\{Y_t\}_{t\geq 0}$ evolve according to K. Assume*

$$\|K^t(x,\cdot) - \pi\|_{\mathrm{TV}} \leq C_x(1-\alpha)^t \tag{11}$$

for some $0 < \alpha \leq 1$, $\{C_x\}_{x\in\mathcal{X}}$ and all $t \in \mathbb{N}$.

Let $\{K_t\}_{t\in\mathbb{N}}$ be a sequence of kernels adapted to some stochastic process $\{X_t\}_{t\in\mathbb{N}}$, with auxillary process $\{A_t\}_{t\in\mathbb{N}}$. Also fix a Lyapunov function V and constants $0 < a, \delta, \epsilon < 1$, $0 \leq b < \infty$ and $0 \leq B < \frac{2b}{a\epsilon}$. Assume that there exists a non-random time $\mathcal{T} = \mathcal{T}_{\epsilon,\delta}$ and a $\sigma\left(\{(X_s, A_s)\}_{s\in\mathbb{N}}^{\mathcal{T}}\right)$-measurable event $\mathcal{F}$ so that $\mathbb{P}[\mathcal{F}] > 1 - \epsilon$,

$$\mathbb{E}[V(X_{\mathcal{T}})\mathbf{1}_{\mathcal{F}}] < \infty, \tag{12}$$

$$\sup_{t>\mathcal{T}} \sup_{x\,:\,V(x)<B} \|K_t(x,\cdot) - K(x,\cdot)\|_{\mathrm{TV}} < \delta + \mathbf{1}_{\mathcal{F}^c}, \tag{13}$$

and the following inequalities are satisfied for all $t > \mathcal{T}$:

$$\begin{aligned}\mathbb{E}[V(X_{t+1})\mathbf{1}_{\mathcal{F}}|X_t = x, A_t] &\leq (1-a)V(x) + b \\ \mathbb{E}[V(Y_{t+1})|Y_t = y] &\leq (1-a)V(y) + b.\end{aligned} \tag{14}$$

Then

$$\limsup_{T\to\infty} \|\mathcal{L}(X_T) - \pi\|_{\mathrm{TV}} \leq 3\epsilon + \delta\frac{\log\left(\frac{e\delta}{\mathcal{C}\log(1-\alpha)}\right)}{\log(1-\alpha)} + \frac{4b}{aB}\lceil\frac{\log\left(\frac{\delta}{\mathcal{C}\log(1-\alpha)}\right)}{\log(1-\alpha)} + 1\rceil,$$

where $\mathcal{C} = \mathcal{C}(\epsilon) \equiv \sup\{C_x : V(x) \leq \frac{2b}{\epsilon a}\}$.

Proof. Assume WLOG that $\mathcal{T} = 0$, fix $\gamma > 0$ and fix $\frac{\log \frac{b}{a(\max(\mathbb{E}[V(X_0)\mathbf{1}_{\mathcal{F}}], \pi(V))+1)}}{\log(1-a)} \leq S < T$. Let $\{Y_t\}_{t \geq S}$, $\{Z_t\}_{t \geq S}$ be Markov chains evolving according to the kernel K and starting at time S, with $Y_S = X_S$ and Z_S distributed according to π . By inequality (11), it is possible to couple $\{Y_t\}_{S \leq t \leq T}$, $\{Z_t\}_{S \leq t \leq T}$ so that

$$\mathbb{P}[Y_T \neq Z_T | X_S] \leq C_{X_S}(1-\alpha)^{T-S} + \gamma. \tag{15}$$

By inequality (13) and a union bound over $S \leq t < T$, it is possible to couple $\{X_t\}_{S \leq t \leq T}$, $\{Y_t\}_{S \leq t \leq T}$ so that

$$\mathbb{P}[X_T \neq Y_T] \leq \delta(T-S) + \mathbb{P}[\mathcal{F}^c] + \mathbb{P}[\max_{S \leq t \leq T}(\max(V(X_t), V(Y_t))) > B] + \gamma. \tag{16}$$

By inequalities (12) and (14),

$$\mathbb{E}[V(X_S)\mathbf{1}_{\mathcal{F}} | X_0, A_0] \leq \mathbb{E}[V(X_0)\mathbf{1}_{\mathcal{F}}](1-a)^S + \frac{b}{a} \leq \frac{2b}{a},$$

and so by Markov's inequality,

$$\mathbb{P}\big[\{V(X_S) > \frac{2b}{a\epsilon}\} \cap \mathcal{F}\big] \leq \epsilon. \tag{17}$$

By the same calculations,

$$\mathbb{P}[\{\max_{S \leq t \leq T}(\max(V(X_t), V(Y_t))) > B\} \cap \mathcal{F}] \leq (T-S+1)\frac{4b}{aB}. \tag{18}$$

Couple $\{Y_t\}_{S \leq t \leq T}$ to $\{X_t\}_{S \leq t \leq T}$ so as to satisfy inequality (16), and then couple $\{Z_t\}_{S \leq t \leq T}$ to $\{Y_t\}_{S \leq t \leq T}$ so as to satisfy inequality (15). It is possible to combine these two couplings of pairs of processes into a coupling of all three processes by the standard 'gluing lemma' (see *e.g.*, Chapter 1 of Villani (2009)). Combining inequalities (15), (16), (17), and (18), we have

$$\begin{aligned}
\|\mathcal{L}(X_T) - \pi\|_{\mathrm{TV}} &\leq \mathbb{P}[X_T \neq Y_T] + \mathbb{P}[Y_T \neq Z_T] \\
&\leq \mathbb{P}[X_T \neq Y_T] + \mathbb{E}[\mathbf{1}_{Y_T \neq Z_T}\mathbf{1}_{V(X_S) > B}\mathbf{1}_{\mathcal{F}}] + \mathbb{E}[\mathbf{1}_{Y_T \neq Z_T}\mathbf{1}_{V(X_S) \leq B}] + \mathbb{P}[\mathcal{F}^c] \\
&\leq \delta(T-S) + 3\epsilon + (T-S+1)\frac{4b}{aB} + 2\gamma + \mathcal{C}(1-\alpha)^{T-S}.
\end{aligned}$$

Approximately optimizing over $S < T$ by choosing $S' = T - \lceil \frac{\log\left(\frac{\delta}{\mathcal{C}\log(1-\alpha)}\right)}{\log(1-\alpha)} \rceil$ for T large, we conclude

$$\begin{aligned}\limsup_{T\to\infty} \|\mathcal{L}(X_T) - \pi\|_{\mathrm{TV}} &\leq \limsup_{T\to\infty} \left(\delta(T - S') + 3\epsilon + (T - S + 1)\frac{4b}{aB} + 2\gamma + \mathcal{C}(1-\alpha)^{T-S'}\right) \\ &\leq 3\epsilon + 2\gamma + \delta\frac{\log\left(\frac{\delta}{\mathcal{C}\log(1-\alpha)}\right)}{\log(1-\alpha)} + \frac{\delta}{\log(1-\alpha)} + \frac{4b}{aB}\lceil\frac{\log\left(\frac{\delta}{\mathcal{C}\log(1-\alpha)}\right)}{\log(1-\alpha)} + 1\rceil.\end{aligned}$$

Since this holds for all $\gamma > 0$, the proof is finished. □

Remark B.2. *In the adaptive MCMC literature, similar results are often stated in terms of a diminishing adaptation condition (this roughly corresponds to inequality (13)) and a containment condition (this roughly corresponds to inequalities (12) and (14)). These phrases were introduced in Roberts and Rosenthal (2007), and there is now a large literature with many sophisticated variants; see, e.g., Fort et al. (2012) for related results that also give LLNs and CLTs under similar conditions. We included our result because its proof is very short, and because checking these simple conditions is easier than checking the more general conditions in the existing literature.*

B.3 Good sets and monotonicity

We give some notation that will be used in the proofs of Theorems 3.4 and 3.3. Fix $0 \leq c, r, R \leq \infty$. For $0 < \ell < \infty$ and $x \in \mathbb{R}^d$, denote by $\mathcal{B}_\ell(x)$ the ball of radius ℓ around x. Say that a finite set $\mathcal{S} \subset \Theta \subset \mathbb{R}^d$ is (c, r, R)-*good* with respect to a set $\mathcal{A} \subset \Theta$ if it satisfies:

1. $\sup_{x\in\mathcal{A}, \|x\|\leq r} \min_{y\in\mathcal{S}} \|x - y\| \leq c$.
2. For all $x \in \mathcal{A}$ with $\|x\| > R$, we have that $|\mathcal{S} \cap \mathcal{B}_{\frac{1}{2}\|x\|}(x)| \geq N$.

We say that it is (c, r, R)-*good* if it is (c, r, R)-*good* with respect to Θ itself. The first condition will imply that the approximation $p_{\mathcal{S}}(x)$ is quite good for x close to the origin. The second condition gives an extremely weak notion of 'locality'; it implies the points we use to construct a 'local' polynomial approximation around x do not remain near the origin when $\|x\|$ itself is very far from the origin. We observe that our definition is monotone in various parameters:

- If $\mathcal{S}$ is (c, r, R)-good, then it is also (c', r', R')-good for all $c' \geq c$, $r' \leq r$ and $R' \geq R$.
- If $\mathcal{S}$ is (c, r, R)-good, then $\mathcal{S} \cup \mathcal{S}'$ is also (c, r, R)-good for any finite set $\mathcal{S}' \subset \Theta$.
- If $\mathcal{S}$ is $(\infty, 0, R)$-good and (c, r, ∞)-good, it is also (c, r, R)-good.

Our arguments will involve showing that, for any finite (c, r, R), the sets $\{\mathcal{S}_t\}_{t\geq 0}$ are eventually (c, r, R)-good.

B.4 Proof of Theorem 3.4, ergodicity in the compact case

In this section we give the proof of Theorem 3.4. Note that some statements are made in slightly greater generality than necessary, as they will be reused in the proof of Theorem 3.3.

Lemma B.3 (Convergence of Kernels)**.** *Let the assumptions stated in the statement of Theorem 3.4 hold. For all $\delta > 0$, there exists a stopping time $\tau = \tau(\delta)$ with respect to $\{\mathcal{S}_t\}_{t\in\mathbb{N}}$* [12] *so that*

$$\sup_{t>\tau}\sup_{x\in\Theta} \|K_\infty(x,\cdot) - \tilde{K}_t(x,\cdot)\|_{\mathrm{TV}} < \delta \tag{19}$$

and so that $\mathbb{P}[\tau < \infty] = 1$.

Proof. Fix $R \in \mathbb{R}$ so that $\Theta \subset \mathcal{B}_R(0)$. By results in (Conn et al., 2009),[13] for any $\lambda, \alpha > 0$, there exists a constant $c = c(\alpha, \lambda) > 0$ so that $\sup_{\theta\in\Theta} |p_{\mathcal{S}}(\theta) - p(\theta|\mathbf{d})| < \alpha$ if $\mathcal{S}$ is λ-poised and (c, R, R)-good. Set $c = c(\delta, \lambda)$ and define $\tau = \inf\{t : \mathcal{S}_t \text{ is } (c, R, R) - \text{good}\}$. By definition, this is a stopping time with respect to $\{\mathcal{S}_t\}_{t\in\mathbb{N}}$ that satisfies inequality (19); we now check that $\mathbb{P}[\tau < \infty] = 1$.

By the assumption that $\ell(x, y)$ is bounded away from 0, there exist $\epsilon > 0$ and measures μ, $\{r_x\}_{x\in\Theta}$ so that

$$L(x,\cdot) = \epsilon\mu(\cdot) + (1-\epsilon)r_x(\cdot). \tag{20}$$

Let $\{A_i\}_{i\in\mathbb{N}}$ and $\{B_i\}_{i\in\mathbb{N}}$ be two sequences of i.i.d. Bernoulli random variables, with success probabilities ϵ and β respectively. Let $\tau_0 = \inf\{t : X_t \in \Theta\}$ and define inductively $\tau_{i+1} = \inf\{t > \tau_i + 1 : X_t \in \Theta\}$. By equality (20), it is possible to couple the sequences $\{X_t\}_{t\in\mathbb{N}}, \{A_i\}_{i\in\mathbb{N}}$ so that

$$\begin{aligned}\mathbb{P}[L_{\tau_i} \in \cdot|\tau_i, X_{\tau_i}, A_i = 1] &= \mu(\cdot) \\ \mathbb{P}[L_{\tau_i} \in \cdot|\tau_i, X_{\tau_i}, A_i = 0] &= r_{X_{\tau_i}}(\cdot).\end{aligned} \tag{21}$$

We can further couple $\{B_i\}_{i\in\mathbb{N}}$ to these sequences by using B_i for the random variable in step 12 of Algorithm 4 at time τ_i. That is, when running Algorithm 4, we would run the subroutine

[12] Throughout the note, for any stochastic process $\{Z_t\}_{t\geq 0}$, we use the phrase "τ is a stopping time with respect to $\{Z_t\}_{t\geq 0}$" as shorthand for "τ is a stopping time with respect to the filtration $\mathcal{F}_t$ given by $\mathcal{F}_t = \sigma(\{Z_s\}_{0\leq s\leq t})$."

[13] The required result is a combination of Theorems 3.14 and 3.16, as discussed in the text after the proof of Theorem 3.16 of (Conn et al., 2009).

RefineNear in step 13 of the algorithm at time $t = \tau_i$ if $B_i = 1$, and we would not run that subroutine in that step at that time if $B_i = 0$. Define $I = \{i \in \mathbb{N} : A_i = B_i = 1\}$. Under this coupling of $\{A_i\}_{i\in\mathbb{N}}, \{B_i\}_{i\in\mathbb{N}}$, and $\{X_t\}_{t\in\mathbb{N}}$,

$$\{L_{\tau_i}\}_{i\in I,\, \tau_i<t} \subset \mathcal{S}_t.$$

Furthermore, $\{L_{\tau_i}\}_{i\in I,\, i\leq N}$ is an i.i.d sequence of N draws from μ and $\mathbb{P}[\tau_i < \infty] = 1$ for all i. Let $\mathcal{E}_j$ be the event that $\{L_{\tau_i}\}_{i\leq j}$ is (c, R, R)-good. We have $\tau \leq \tau_{\inf\{j\,:\,\mathcal{E}_j \text{ holds}\}}$. By independence of the sequence $\{L_{\tau_i}\}_{i\in\mathbb{N}}$, we obtain

$$\mathbb{P}[\tau < \infty] \geq \liminf_{j\to\infty} \mathbb{P}[\mathcal{E}_j] = 1.$$

This completes the proof of the Lemma. □

Remark B.4. *We mention briefly that this lemma can also be used to obtain a quantitative bound on the asymptotic rate of convergence of the bias of our algorithm.*

Observe that τ as defined in the proof of Lemma B.3 is stochastically dominated by an exponential distribution with mean $O(-dc^{-d}\log(c))$ as long as both $\ell(x,\cdot)$ and $p(\cdot|\mathbf{d})$ are bounded below. This gives a rather poor bound on the amount of time it takes for inequality (29) to hold. Inequality (29), together with standard 'perturbation' bounds relating the distance between transition kernels and the distance between their stationary distributions, imply a quantitative bound on the asymptotic rate of convergence of the bias of our algorithm. An example of such a perturbation bound may be found by applying Theorem 1 of (Korattikara et al., 2013), which does not in fact rely on time-homogeneity, to a subsequence of the stochastic process generated by our algorithm. Unfortunately, the resulting bound is rather poor, and does not seem to reflect our algorithm's actual performance.

We now prove Theorem 3.4:

Proof. It is sufficient to show that, for all $\epsilon, \delta > 0$ sufficiently small, the conditions of Theorem B.1 can be satisfied. We now set the constants and functions associated with Theorem B.1; we begin by choosing $C_x \equiv V(x) \equiv b = a = 1$, setting $\alpha = \frac{\inf_{x,y\in\Theta} \ell(x,y) \inf_{\theta\in\Theta} p(\theta|d)}{\sup_{\theta\in\Theta} p(\theta|d)}$, and setting $B = \infty$.

By the minorization condition, inequality (11) is satisfied for this value of α; by the assumption that $\ell(x,y), p(\theta|d)$ are bounded away from 0 and infinity, we also have $\alpha > 0$. Next, for all $\delta > 0$, Lemma B.3 implies that implies that $\sup_x \|K(x,\cdot) - \tilde{K}(x,\cdot)\|_{\mathrm{TV}} < \delta$ for all times t greater than

some a.s. finite random time $\tau = \tau(\delta)$ that is a stopping time with respect to $\{\mathcal{S}_t\}_{t\in\mathbb{N}}$. Choosing $\mathcal{T} = \mathcal{T}_{\epsilon,\delta}$ to be the smallest integer so that $\mathbb{P}[\tau(\delta) > \mathcal{T}] \leq 1-\epsilon$ and setting $\mathcal{F} = \{\tau \leq \mathcal{T}\}$, this means that inequality (13) is satisfied. Inequalities (14) and (12) are trivially satisfied given our choice of V, a, b. Applying Theorem B.1 with this choice of $V, \alpha, a, b, \mathcal{T}$, we have for all $\epsilon, \delta > 0$ that

$$\limsup_{T\to\infty} \|\mathcal{L}(X_T) - \pi\|_{\mathrm{TV}} \leq 3\epsilon + \delta \frac{\log\left(\frac{e\delta}{\mathcal{C}\log(1-\alpha)}\right)}{\log(1-\alpha)}.$$

Letting δ go to 0 and then ϵ go to 0 completes the proof. □

B.5 Proof of Theorem 3.3, ergodicity in the non-compact case

In this section, we prove Theorem 3.3. The argument is similar to that of Theorem 3.4, but we must show the following to ensure that the sampler does not behave too badly when it is far from the posterior mode:

1. $\mathcal{S}_t$ is $(\infty, 0, R)$-good after some almost-surely finite random time τ; see Lemma B.6.
2. The kernel $\tilde{K}_t$ satisfies a drift condition if $\mathcal{S}_t$ is $(\infty, 0, R)$-good; see Lemmas B.8 and B.9.
3. This drift condition implies that the chain X_t spends most of its time in a compact subset of Θ; see Lemma B.10.

Remark B.5. *The Gaussian envelope condition (see Assumption 3.1) is used only to show the second step in the above proof strategy, which in turn is used to satisfy condition (14) of Theorem B.1. It can be replaced by any assumption on the target density for which $\mathcal{S}$ being $(\infty, 0, R)$-good for some $R < \infty$ implies that $\tilde{K}_{\mathcal{S}}$ satisfies a drift condition of the form given by inequality (14).*

We begin by showing, roughly, that for any $R > 0$, $\mathcal{S}_t$ is eventually $(\infty, 0, R)$-good:

Lemma B.6 (Approximations At Infinity Ignore Compact Sets)**.** *Fix any $\mathcal{X} > 0$ and any $k \geq 2$ and define*

$$\tau_{\mathcal{X}}^{(k)} = \sup\left\{t \,:\, \|L_t\| > k\mathcal{X},\ \|L_t\| - R_t < \mathcal{X}\right\}.$$

Then

$$\mathbb{P}[\{\text{There exists } k < \infty,\ \textit{s.t. } \tau_{\mathcal{X}}^{(k)} < \infty\}] = 1.$$

Proof. Fix $N \in \mathbb{N}$, $\delta > 0$ and $0 < r_1 < r_2 < \infty$. For $0 < \ell < \infty$, denote by $\partial\mathcal{B}_\ell(0)$ the sphere of radius ℓ. Fix a finite covering $\{P_i\}$ of $\partial\mathcal{B}_{\frac{r_1+r_2}{2}}(0)$ with the property that, for any $x \in \partial\mathcal{B}_{\frac{r_1+r_2}{2}}(0)$, there exists at least one i so that $P_i \subset \mathcal{B}_\delta(x)$. For $k \in \mathbb{N}$, define a thickening of P_i by:

$$\mathcal{P}_i^{(k)} = \left\{ x : \frac{r_1+r_2}{2}\frac{x}{\|x\|} \in P_i,\ \frac{r_1+r_2}{2} + (k-1)\frac{r_2-r_1}{2} \leq \|x\| \leq \frac{r_1+r_2}{2} + k\frac{r_2-r_1}{2} \right\}.$$

We will show that, almost surely, for every thickening $\mathcal{P}_i^{(k)}$ of an element P_i of the cover, either $|\mathcal{P}_i^{(k)} \cap \mathcal{S}_t|$ is eventually greater than N or $|\mathcal{P}_i^{(k)} \cap \{L_t\}_{t\in\mathbb{N}}|$ is finite. Note that it is trivial that either $|\mathcal{P}_i^{(k)} \cap \{L_t\}_{t\in\mathbb{N}}|$ is eventually greater than N or $|\mathcal{P}_i^{(k)} \cap \{L_t\}_{t\in\mathbb{N}}|$ is finite; the goal is to check that if $\{L_t\}_{t\in\mathbb{N}}$ visits $\mathcal{P}_i$ infinitely often, $|\mathcal{P}_i^{(k)} \cap \mathcal{S}_t|$ must eventually be greater than N.

To see this, we introduce a representation of the random variables used in step 12 of Algorithm 4. Recall that in this step, L_t is added to $\mathcal{S}_t$ with probability β, independently of the rest of the history of the walk. We will split up the sequence B_t of Bernoulli(β) random variables according to the covering as follows: for each element $\mathcal{P}_i^{(k)}$ of the covering, let $\{B_t^{(i,k)}\}_{t\in\mathbb{N}}$ be an i.i.d. sequence of Bernoulli random variables with success probability β. At the mth time L_t is in $\mathcal{P}_i^{(k)}$, we use $B_m^{(i,k)}$ as the indicator function in step 12 of Algorithm 4. This does not affect the distribution of the steps that the algorithm takes.

By the Borel-Cantelli lemma, we have for each i, k that $\mathbb{P}[B_t^{(i,k)} = 1, \text{infinitely often}] = 1$. If $B_t^{(i,k)} = 1$ infinitely often, then $|\mathcal{P}_i^{(k)} \cap \{L_t\}_{t\in\mathbb{N}}| = \infty$ implies that for all $M < \infty$, we have $|\mathcal{P}_i^{(k)} \cap \mathcal{S}_t| > M$ eventually. Let $\mathcal{C}_{i,k}$ be the event that $|\mathcal{P}_i^{(k)} \cap \mathcal{S}_t| > N$ eventually and let $\mathcal{D}_{i,k}$ be the event that $|\mathcal{P}_i^{(k)} \cap \{L_t\}_{t\in\mathbb{N}}| = \infty$. Then this argument implies that

$$\mathbb{P}[\mathcal{C}_{i,k} | \mathcal{D}_{i,k}] = 1.$$

Since there are only countably many sets $\mathcal{P}_i^{(k)}$, we have

$$\mathbb{P}[\cap_{i,k} \left(\mathcal{C}_{i,k} \cup \mathcal{D}_{i,k}^c\right)] = 1. \tag{22}$$

Thus, conditioned on the almost sure event $\cap_{i,k} \left(\mathcal{C}_{i,k} \cup \mathcal{D}_{i,k}^c\right)$, all sets $\mathcal{P}_i^{(k)}$ that L_t visits infinitely often will also contribute points to $\mathcal{S}_t$ infinitely often.

Let $k(i) = \min\{k : |\mathcal{P}_i^{(k)} \cap \{L_t\}_{t\in\mathbb{N}}| = \infty\}$ when that set is non-empty, and set $k(i) = \infty$

otherwise. Let $I = \{i : k(i) < \infty\}$. Finally, set

$$\tau_{r_1,r_2} = \inf\{t : \forall i \in I, |\mathcal{P}_i^{(k(i))} \cap \mathcal{S}_t| \geq N\}. \tag{23}$$

Since $|I|$ is finite, we have shown that, for all $N, \delta > 0$ and $0 < r_2 < r_1 < \infty$, $\mathbb{P}[\tau_{r_1,r_2} < \infty] = 1$. Finally, we observe that for all $\delta = \delta(\mathcal{X}, d)$ sufficiently small, all $N \geq N_{\text{def}}$ and all $k \geq \max_{i\in I} k(i)$,

$$\tau_{\mathcal{X}}^{(k)} \leq \tau_{\frac{2}{3}\mathcal{X},\frac{4}{3}\mathcal{X}}. \tag{24}$$

This completes the proof. □

Remark B.7. *We will eventually see that, in the notation of the proof of Lemma B.6, $k(i) = 1$ for all i.*

Next, we show that the approximation $p_{\mathcal{S}_t}(x)$ of the posterior used at time t is close to $p_\infty(X_t)$ when $\mathcal{S}_t$ is $(\infty, 0, R)$-good and $\|X_t\|$ is sufficiently large:

Lemma B.8 (Approximation at Infinity)**.** *For all $\epsilon > 0$ and $k \geq 2$, there exists a constant $\mathcal{X} = \mathcal{X}(\epsilon) > 0$ so that, if $R_t < (\|L_t\| - (k-1)\mathcal{X})\mathbf{1}_{\|L_t\|>k\mathcal{X}}$ and the set $\{q_t^{(1)}, \ldots, q_t^{(N)}\}$ is λ-poised, then*

$$|\log(p_{\mathcal{S}_t}(L_t)) - \log(p_\infty(L_t))| < \epsilon + \lambda(N+1)G.$$

Proof. Fix $\epsilon > 0$. By (6) in Assumption 3.1, there exists some $\mathcal{X} = \mathcal{X}(\epsilon)$ so that $\|x\| > \mathcal{X}$ implies

$$|\log(p(x|\mathbf{d})) - \log(p_\infty(x))| < G + \frac{\epsilon}{(N+1)\lambda}. \tag{25}$$

We fix this constant $\mathcal{X}$ in the remainder of the proof.

Denote by $\{f_i\}_{i=1}^{N+1}$ the Lagrange polynomials associated with the set $\{q_t^{(1)}, \ldots, q_t^{(N)}\}$. By Lemma 3.5 of (Conn et al., 2009),

$$\begin{aligned}
|\log(p_{\mathcal{S}_t}(L_t)) - \log(p_\infty(L_t))| &= |\sum_i f_i(L_t)\log(p(q_t^{(i)}|\mathbf{d})) - \log(p_\infty(L_t))| \\
&\leq |\sum_i \log(p_\infty(q_t^{(i)}))f_i(L_t) - \log(p_\infty(L_t))| \\
&\qquad + \sum_i |\log(p(q_t^{(i)}|\mathbf{d})) - \log(p_\infty(q_t^{(i)}))|\,|f_i(L_t)| \\
&\leq 0 + (N+1)\lambda \sup_i |\log(p(q_t^{(i)}|\mathbf{d})) - \log(p_\infty(q_t^{(i)}))|
\end{aligned}$$

where the last line follows from the definition of Lagrange polynomials and Definition 4.7 of (Conn et al., 2009). Under the assumption $\|q_t^{(i)}\| > \mathcal{X}$ for $\|L_t\| - R_t > (k-1)\mathcal{X}$, the conclusion follows from inequality (25). □

For $\epsilon > 0$, define $V_\epsilon(x) = V(x)^{\frac{1}{1+\epsilon}}$, where V is defined in Equation (7). Denote by $\alpha_\infty(x,y)$ the acceptance function of a Metropolis-Hastings chain with proposal kernel L and target distribution p_∞, and recall that $\tilde{\alpha}_t(x,y)$ as given in Equation (10) is the acceptance function for $\tilde{K}_t$. We show that $\tilde{K}_t$ inherits a drift condition from K_∞:

Lemma B.9 (Drift Condition)**.** *For $0 < \delta < \frac{1}{10}$ and $\mathcal{Y}, \mathcal{T} < \infty$, let $\mathcal{F}$ be the event that*

$$|\tilde{\alpha}_t(X_t, L_t) - \alpha_\infty(X_t, L_t)| < \delta + 2\mathbf{1}_{|X_t|<\mathcal{Y}} + 2\mathbf{1}_{|L_t|<\mathcal{Y}} \tag{26}$$

for all $t > \mathcal{T}$. Then, for $\epsilon = \epsilon_0$ as given in item 1 of Assumption 3.2, and all $\delta < \delta_0(\epsilon, a, b, V) < \frac{1}{10}$ sufficiently small and $\mathcal{Y}$ sufficiently large, X_t satisfies a drift condition of the form:

$$\mathbb{E}[V_\epsilon(X_{t+1})\mathbf{1}_{\mathcal{F}}|X_t, \mathcal{S}_t] \leq a_1 V_\epsilon(X_t) + b_1 \tag{27}$$

for some $0 \leq a_1 < 1$, $0 \leq b_1 < \infty$ and for all $t > \mathcal{T}$.

Proof. Assume WLOG that $\mathcal{T} = 0$. Let Z_t be a Metropolis-Hastings Markov chain with proposal kernel L and target distribution p_∞. By Jensen's inequality and Assumption 3.2

$$\mathbb{E}[V_\epsilon(Z_{t+1})|Z_t = x] \leq a_\epsilon V_\epsilon(x) + b_\epsilon$$

for some $0 < a_\epsilon < 1$ and some $0 \leq b_\epsilon < \infty$.

Assume $X_t = x$ and fix δ so that $\delta < \delta_0$ and $(1+3\delta)a_\epsilon < a_\epsilon + \frac{1}{2}(1-\alpha_\epsilon)$. Then

$$\begin{aligned}
\mathbb{E}[V_\epsilon(X_{t+1})\mathbf{1}_{\mathcal{F}}|X_t = x, \mathcal{S}_t] &\leq \int_{y\in\mathbb{R}^d} \left(\tilde{\alpha}_t(x,y)V_\epsilon(y) + (1-\tilde{\alpha}_t(x,y))V_\epsilon(x)\right)\ell(x,y)dy \\
&\leq \int_{\mathbb{R}^d\setminus[-\mathcal{Y},\mathcal{Y}]^d} \left(e^{2\delta}\alpha_\infty(x,y)V_\epsilon(y) + \left(1-e^{-2\delta}\alpha_\infty(x,y)\right)V_\epsilon(x)\right)\ell(x,y)dy \\
&\qquad + \int_{y\in[-\mathcal{Y},\mathcal{Y}]^d} \left(V_\epsilon(x) + \sup_{\|z\|\leq\mathcal{Y}} V_\epsilon(z)\right)\ell(x,y)dy \\
&\leq (1+3\delta)\int_{\mathbb{R}^d} \left(\alpha_\infty(x,y)V_\epsilon(y) + (1-\alpha_\infty(x,y))V_\epsilon(x)\right)\ell(x,y)dy \\
&\qquad + \left(V_\epsilon(x) + \sup_{\|z\|\leq\mathcal{Y}} V_\epsilon(z)\right) L\left(x, [-\mathcal{Y},\mathcal{Y}]^d\right)
\end{aligned}$$

$$\leq (1+3\delta)a_\epsilon V_\epsilon(x) + (1+3\delta)b_\epsilon + \left(V_\epsilon(x) + \sup_{\|z\|\leq \mathcal{Y}} V_\epsilon(z)\right) L\left(x, [-\mathcal{Y}, \mathcal{Y}]^d\right).$$

Since $\delta < \frac{1}{10}$ and $(1+3\delta)a_\epsilon < a_\epsilon + \frac{1}{2}(1-\alpha_\epsilon)$, we have

$$\mathbb{E}[V_\epsilon(X_{t+1})\mathbf{1}_{\mathcal{F}}|X_t = x, \mathcal{S}_t] \leq (a_\epsilon + \frac{1}{2}(1-\alpha_\epsilon))V(x) + (1+3\delta)b_\epsilon + \left(V_\epsilon(x) + \sup_{\|z\|\leq \mathcal{Y}} V_\epsilon(z)\right) L\left(x, [-\mathcal{Y}, \mathcal{Y}]^d\right).$$

Since $V_\epsilon(x)L\left(x, [-\mathcal{Y}, \mathcal{Y}]^d\right)$ is uniformly bounded in x for all fixed $\mathcal{Y}$ by item 2 of Assumption 3.2, the claim follows with

$$\begin{aligned} a_1 &= a_\epsilon + \frac{1}{2}(1-\alpha_\epsilon) < 1, \\ b_1 &= 2b_\epsilon + \sup_x V_\epsilon(x) L\left(x, [-\mathcal{Y}, \mathcal{Y}]^d\right) + \sup_{\|z\|\leq \mathcal{Y}} V_\epsilon(z), \end{aligned}$$

finishing the proof. □

We use these bounds to show that some compact set is returned to infinitely often:

Lemma B.10 (Infinitely Many Returns)**.** *For $G < G(L, p_\infty, \lambda, N)$ sufficiently small, there exists a compact set $\mathcal{A}$ that satisfies $\mathbb{P}[\sum_{t\in\mathbb{N}} \mathbf{1}_{X_t\in\mathcal{A}} = \infty] = 1$.*

Proof. Combining Lemmas B.6, B.8 and B.9, there exists some number $\mathcal{X} > 0$ and almost surely finite random time $\tau_{\mathcal{X}}$ so that X_t satisfies a drift condition of the form

$$\mathbb{E}[V(X_{t+1})\mathbf{1}_{t>\tau_{\mathcal{X}}}|X_t = x, \mathcal{S}_t] \leq aV(x) + b$$

for some function V and constants $0 \leq a < 1$, $b < \infty$. The existence of a recurrent compact set follows immediately from this drift condition and Lemma 4 of (Rosenthal, 1995). □

This allows us to slightly strengthen Lemma B.9:

Lemma B.11. *All times $\tau_{\mathcal{X},2\mathcal{X}}$ of the form given in Equation (23) satisfy $\mathbb{P}[\tau_{\mathcal{X},2\mathcal{X}} < \infty] = 1$ and are stopping times with respect to $\{\mathcal{S}_t\}$. Furthermore, for $G < G(L, p_\infty, \lambda, N)$ sufficiently small, there exists a random time τ of the form given in Equation (23) so that*

$$\mathbb{E}[V_\epsilon(X_{t+1})\mathbf{1}_{\tau<t}|X_t, \mathcal{S}_t] \leq a_1 V_\epsilon(X_t) + b_1 \tag{28}$$

for some $0 \leq a_1 < 1$, $0 \leq b_1 < \infty$.

Proof. By inequality (24), there exists a random time $\tau \equiv \tau_{\mathcal{X},2\mathcal{X}}$ of the form (23) that is at least as large as the random time $\tau_{\mathcal{X}}$ constructed in the proof of Lemma B.10 and that satisfies $\mathbb{P}[\tau < \infty] = 1$. As shown in Lemma B.10, an inequality of the form (28) holds for $\tau_{\mathcal{X}}$, and so the same inequality must also hold with $\tau_{\mathcal{X}}$ replaced by the larger time $\tau \geq \tau_{\mathcal{X}}$.

The only detail to check is that all random times $\tau_{\mathcal{X},2\mathcal{X}}$ of the form (23) are stopping times with respect to $\{\mathcal{S}_t\}_{t\in\mathbb{N}}$. Let $\{P_i\}$ be the partition associated with $\tau_{\mathcal{X},2\mathcal{X}}$, as constructed in Lemma B.6. By Lemma B.10 and part 3 of Assumption 3.2, we have $\mathbb{P}[|\{L_t\}_{t\in\mathbb{N}} \cap \mathcal{P}_i^{(1)}| = \infty] = 1$ for all i. Thus, in the notation of Lemma B.6, $I^c = \emptyset$ and $k(i) = 1$ for all $i \in I$. Thus, we have shown that $\tau_{\mathcal{X},2\mathcal{X}} = \inf\{t : \forall i, |\mathcal{P}_i^{(1)} \cap \mathcal{S}_t| \geq N\}$, which is clearly a stopping time with respect to $\{\mathcal{S}_t\}_{t\in\mathbb{N}}$, and the proof is finished. □

We now finish our proof of Theorem 3.3 analogously to our proof of Theorem 3.4.

The following bound is almost identical to Lemma B.3, but now proved under the Gaussian envelope assumption for the target density.

Lemma B.12 (Convergence of Kernels)**.** *Let the assumptions stated in the statement of Theorem 3.3 hold and fix a compact set $\mathcal{A} \subset \Theta$. For all $\delta > 0$, there exists a stopping time $\tau = \tau(\delta)$ with respect to $\{\mathcal{S}_t\}_{t\in\mathbb{N}}$ so that*

$$\sup_{t>\tau} \sup_{x\in\mathcal{A}} \|K_\infty(x,\cdot) - \tilde{K}_t(x,\cdot)\|_{\mathrm{TV}} < \delta \tag{29}$$

and so that $\mathbb{P}[\tau < \infty] = 1$.

Proof. Fix a constant $0 < R < \infty$ so that $\mathcal{A} \subset \mathcal{B}_R(0)$. By results in (Conn et al., 2009), for any $\lambda, \alpha > 0$, there exists a constant $c = c(\alpha, \lambda) > 0$ so that $\sup_{\theta\in\mathcal{A}} |p_{\mathcal{S}}(\theta) - p(\theta|\mathbf{d})| < \alpha$ if $\mathcal{S}$ is λ-poised and (c, R, R)-good. Set $c = c(\epsilon, \lambda)$ and define $\tau' = \inf\{t : \mathcal{S}_t \text{ is } (c, R, R) - \text{good}\}$. By definition, τ' is a stopping time with respect to $\{\mathcal{S}_t\}_{t\in\mathbb{N}}$ that satisfies inequality (29). We now check that $\mathbb{P}[\tau' < \infty] = 1$. By the assumption that $\ell(x, y)$ is bounded away from 0, there exist $\epsilon > 0$ and measures μ, $\{r_x\}_{x\in\Theta}$ so that

$$L(x,\cdot) = \epsilon\mu(\cdot) + (1-\epsilon)r_x(\cdot). \tag{30}$$

Let $\{A_i\}_{i\in\mathbb{N}}$ and $\{B_i\}_{i\in\mathbb{N}}$ be two sequences of i.i.d. Bernoulli random variables, with success probabilities ϵ and β respectively. Let $\tau_0 = \inf\{t : X_t \in \mathcal{A}\}$ and define inductively $\tau_{i+1} = \inf\{t >$

$\tau_i + 1 : X_t \in \mathcal{A}\}$. By equality (30), it is possible to couple the sequences $\{X_t\}_{t\in\mathbb{N}}, \{A_i\}_{i\in\mathbb{N}}$ so that

$$\begin{aligned} \mathbb{P}[L_{\tau_i} \in \cdot | \tau_i, X_{\tau_i}, A_i = 1] &= \mu(\cdot) \\ \mathbb{P}[L_{\tau_i} \in \cdot | \tau_i, X_{\tau_i}, A_i = 0] &= r_{X_{\tau_i}}(\cdot). \end{aligned} \tag{31}$$

We can further couple $\{B_i\}_{i\in\mathbb{N}}$ to these sequences by using B_i for the random variable in step 12 of Algorithm 4 at time τ_i. That is, when running Algorithm 4, we would run the subroutine RefineNear in step 13 of the algorithm at time $t = \tau_i$ if $B_i = 1$, and we would not run that subroutine in that step at that time if $B_i = 0$. Define $I = \{i \in \mathbb{N} : A_i = B_i = 1\}$. Under this coupling of $\{A_i\}_{i\in\mathbb{N}}, \{B_i\}_{i\in\mathbb{N}}$, and $\{X_t\}_{t\in\mathbb{N}}$,

$$\{L_{\tau_i}\}_{i\in I,\, \tau_i < t} \subset \mathcal{S}_t.$$

Furthermore, $\{L_{\tau_i}\}_{i\in I,\, i\leq N}$ is an i.i.d sequence of N draws from μ, and by Lemma B.10, $\mathbb{P}[\tau_i < \infty] = 1$ for all i. Let $\mathcal{E}_j$ be the event that $\{L_{\tau_i}\}_{i\leq j}$ is (c, R, R)-good. We have $\tau' \leq \inf\{\tau_j : \mathcal{E}_j \text{ holds}\}$. By independence of the sequence $\{L_{\tau_i}\}_{i\in\mathbb{N}}$, we obtain

$$\mathbb{P}[\tau' < \infty] \geq \liminf_{j\to\infty} \mathbb{P}[\mathcal{E}_j] = 1.$$

This argument shows that, for any compact set $\mathcal{A}$, there exists a stopping time τ' with respect to $\{\mathcal{S}_t\}_{t\in\mathbb{N}}$ so that $\mathbb{P}[\tau' < \infty] = 1$ and so that

$$\sup_{t>\tau'} \sup_{x\in\mathcal{A}} \|\tilde{K}_t(x, \cdot) - K_\infty(x, \cdot)\|_{\mathrm{TV}} < \delta. \tag{32}$$

This completes the proof of the Lemma.

□

We are finally ready to prove Theorem 3.3:

Proof of Theorem 3.3. As with the proof of Theorem 3.4, it is sufficient to show that, for all $\epsilon, \delta, G > 0$ sufficiently small and all $B \gg \epsilon^{-1}$ sufficiently large, the conditions of Theorem B.1 can be satisfied for some time $\mathcal{T} = \mathcal{T}_{\epsilon,\delta}$ with the same drift function V and constants α, a, b.

By Assumption 3.2 and Theorem 12 of Rosenthal (1995), inequality (11) holds for some $\alpha > 0$ and $\{C_x\}_{x\in\Theta}$. For any fixed $0 < B < \infty$ and all $0 < G, \delta$ sufficiently small, Lemma B.12 implies

that there exists some almost surely finite stopping time $\tau_1 = \tau_1(\delta)$ so that inequality (13) holds for the set $\mathcal{F}_1 = \{\tau_1 > t\}$. Lemma B.11 implies that, for all $G > 0$ sufficiently small, there exists some almost surely finite stopping time τ_2 so that inequality (14) holds for the set $\mathcal{F}_2 = \{\tau_2 > t\}$. Choose $\mathcal{T}$ to be the smallest integer so that $\mathbb{P}[\max(\tau_1, \tau_2) > \mathcal{T}] < \epsilon$ and set $\mathcal{F} = \{\min(\tau_1, \tau_2) > \mathcal{T}\}$. We then have that inequalities (13) and (14) are satisfied. Finally, inequality (12) holds by part 2 of Assumption 3.2. We have shown that there exist fixed values of $\mathcal{C}$ and α so that the conditions of Theorem B.1 hold for all $\epsilon, \delta > 0$ sufficiently small. We conclude that, for all $\epsilon, \delta > 0$ sufficiently small,

$$\limsup_{T\to\infty} \|\mathcal{L}(X_T) - \pi\|_{\mathrm{TV}} \leq 3\epsilon + \delta \frac{\log\left(\frac{e\delta}{\mathcal{C}\log(1-\alpha)}\right)}{\log(1-\alpha)} + \frac{4b}{aB}\left\lceil \frac{\log\left(\frac{\delta}{\mathcal{C}\log(1-\alpha)}\right)}{\log(1-\alpha)} + 1\right\rceil.$$

Letting B go to infinity, then δ go to 0 and finally ϵ go to 0 completes the proof. □

B.6 Alternative assumptions

In this section, we briefly give other sufficient conditions for ergodicity. We do not give detailed proofs but highlight the instances at which our current arguments should be modified.

The central difficulty in proving convergence of our algorithm is that, in general, the local polynomial fits we use may be very poor when R_t is large. This difficulty manifests in the fact that, for most target distributions, making the set $\mathcal{S}$ a (c, r, R)-good set does not guarantee that $\tilde{K}_{\mathcal{S}}$ inherits a drift condition of the form (14) from K_∞, for any value of c, r, R. Indeed, no property that is monotone in the set $\mathcal{S}$ can guarantee that $\tilde{K}_{\mathcal{S}}$ satisfies a drift condition. In a forthcoming project focused on theoretical issues, we plan to show convergence based on drift conditions that only hold 'on average' and over long time intervals. There are several other situations under which it is possible to guarantee the eventual existence of a drift condition, and thus ergodicity:

1. Fix a function $\delta_0 : \Theta \to \mathbb{R}^+$ and add the step "If $R_t > \delta_0(\theta^+)$, $\mathcal{S} \leftarrow \{(\theta^+, f(\theta^+)\} \cup \mathcal{S})$" between steps 7 and 8 of Algorithm 4. If $\lim_{r\to\infty} \sup_{\|x\|\geq r} \delta_0(x) = 0$ and

$$\lim_{r\to\infty} \sup_{\|x\|\geq r} \max(\|p'(\theta|\mathbf{d})\|, \|p''(\theta|\mathbf{d})\|) = 0,$$

then the main condition of Lemma B.9, inequality (26) (with α_∞ replaced by the acceptance function of K), holds by a combination of Theorems 3.14 and 3.16 of (Conn et al., 2009). If $p(\theta|\mathbf{d})$ has sub-Gaussian tails, the proof of Lemma B.9 can then continue largely as written

if we replace $p_\infty(x)$ with $p(x|\mathbf{d})$ wherever it appears. Since the Gaussian envelope condition is only used to prove that the condition in Lemma B.9 holds, Theorem 3.3 holds with the Gaussian envelope condition replaced by these requirements.

2. Similar results sometimes hold if we only require that $\delta_0(x) \equiv \delta_0$ be a sufficiently small constant. Theorem 1 of Ferré et al. (2013), combined with Theorems 3.14 and 3.16 of (Conn et al., 2009), can be used to obtain weaker sufficient conditions under which the condition in Lemma B.9 holds.

3. If $d = 1$, $N_{\text{def}} = 2$, and the approximations in Algorithm 4 are made using linear rather than quadratic models, we state without proof that a drift condition at infinity proved in Lemma B.9 can be verified directly. For $d \geq 2$, more work needs to be done.

4. Finally, we discuss analogous results that hold for other forms of local approximation, such as Gaussian processes. When the target distribution is compact, we expect Theorem 3.4 to hold as stated whenever local approximations to a function based on (c, R, R)-good sets converge to the true function value as c goes to 0. In our proof of Theorem 3.4, we cite (Conn et al., 2009) for this fact. The proof of Theorem 3.4 will hold as stated for other local approximations if all references to (Conn et al., 2009) are replaced by references to appropriate analogous results. Such results typically hold for reasonably constructed local approximation strategies (Cleveland and Loader, 1996; Atkeson et al., 1997).

   When the target distribution is not compact, modifying our arguments can be more difficult, though we expect similar conclusions to often hold.

B.7 Examples for parameter choices

Example B.13 (Decay Rate for β)**.** *We note that if β_t decays too quickly, our sampler may not converge, even if $\gamma_t \to 0$ at any rate. Consider the proposal distribution L that draws i.i.d. uniform samples from $[0,1]^d$ and let $\lambda(\cdot)$ denote the Lebesgue measure. Consider a target distribution of the form $p(\theta|\mathbf{d}) \propto \mathbf{1}_{\theta \in G}$ for set G with Lebesgue measure $0 < \lambda(G) < 1$. If $\sum_t \beta_t < \infty$, then by Bo+rel-Cantelli, the probability $p = p(\{\beta_t\}_{t \in \mathbb{N}})$ that no points are added to $\mathcal{S}$ except during the initial choice of reference points or failed cross-validation checks is strictly greater than 0. With probability $\lambda(G)^k > 0$, the first k reference points are all in G. But if both these events happen, all cross-validation checks are passed for any $\gamma > 0$, and so the walk never converges; it samples from the measure λ forever.*

Example B.14 (Decay Rate for γ)**.** *We note that we have not used the assumption that $\gamma < \infty$ anywhere. As pointed out in Example B.13, in a way this is justified—we can certainly find sequences $\{\beta_t\}_{t\in\mathbb{N}}$ and walks that are not ergodic for any sequence $\gamma_t > 0$ converging to zero at any rate.*

In the other direction, there exist examples for which having any reasonable fixed value of γ gives convergence, even with $\beta = 0$. We point out that this depends on the initially selected points; one could be unlucky and choose points with log-likelihoods that happen to lie exactly on some quadratic that does not match the true distribution. Consider a target density $\pi(x) \propto 1 + C\mathbf{1}_{x>\frac{1}{2}}$ on $[0, 1]$ with independent proposal moves from the uniform measure on $[0, 1]$. To simplify the discussion, we assume that our approximation of the density at each point is linear and based exactly on the three nearest sampled points. Denote by $\mathcal{S}_t$ the points which have been evaluated by time t, and let $\mathcal{S}_0 = \{\frac{1}{8}, \frac{2}{8}, \frac{3}{8}, \frac{5}{8}, \frac{6}{8}, \frac{7}{8}\}$. Write $x_1, \ldots, x_{m(t)} = \mathcal{S}_t \cap [0, \frac{1}{2}]$ and $x_{m(t)+1}, \ldots, x_{n(t)} = \mathcal{S}_t \cap [\frac{1}{2}, 1]$. It is easy to check that

$$\|\mathcal{L}(X_{t+1}) - \pi\|_{\mathrm{TV}} \leq x_{m(t)+3} - x_{m(t)-2}. \tag{33}$$

It is also easy to see that with probability one, for any $\gamma < \frac{1}{2}$, there will always be a subinterval of $[x_{m(t)-2}, x_{m(t)+3}]$ with strictly positive measure for which a cross-validation check will fail. Combining this with inequality (33) implies that the algorithm will converge in this situation, even with $\beta = 0$. Furthermore, in this situation choosing $\beta \equiv 0$ results in a set $\mathcal{S}_t$ that grows extremely slowly in t, without substantially increasing bias.

C Genetic toggle switch inference problem

Here we provide additional details about the setup of the genetic toggle switch inference problem from Section 4.2. This genetic circuit has a bistable response to the concentration of an input chemical, [IPTG]. Figure 13 illustrates these high and low responses, where the vertical axis corresponds to the expression level of a particular gene. (Gardner et al., 2000) proposed the following differential-algebraic model for the switch:

$$\begin{aligned} \frac{du}{dt} &= \frac{\alpha_1}{1 + v^\beta} - u, \\ \frac{dv}{dt} &= \frac{\alpha_2}{1 + w^\gamma} - v, \\ w &= \frac{u}{(1 + [\mathrm{IPTG}]/K)^\eta}. \end{aligned} \tag{34}$$

The model contains six unknown parameters $Z_\theta = \{\alpha_1, \alpha_2, \beta, \gamma, K, \eta\} \in \mathbb{R}^6$, while the data correspond to observations of the steady-state values $v(t = \infty)$ for six different input concentrations of [IPTG], averaged over several trials each. As in (Marzouk and Xiu, 2009), the parameters are centered and scaled around their nominal values so that they can be endowed with uniform priors over the hypercube $[-1, 1]^6$. Specifically, the six parameters of interest are normalized around their nominal values to have the form

$$Z_i = \bar{\theta}_i(1 + \zeta_i \theta_i), \ i = 1, \ldots, 6,$$

so that each θ_i has prior Uniform$[-1, 1]$. The values of $\bar{\theta}_i$ and ζ_i are given in Table 1. The data are observed at six different values of [IPTG]; the first corresponds to the "low" state of the switch while the rest are in the "high" state. Multiple experimental observations are averaged without affecting the posterior by correspondingly lowering the noise; hence, the data comprise one observation of v/v_{ref} at each concentration, where $v_{\text{ref}} = 15.5990$. The data are modeled as having independent Gaussian errors, *i.e.*, as draws from $\mathcal{N}(d_i, \sigma_i^2)$, where the high- and low-state observations have different standard deviations, specified in Table 2. The forward model may be computed by integrating the ODE system (35), or more simply by iterating until a fixed point for v is found.

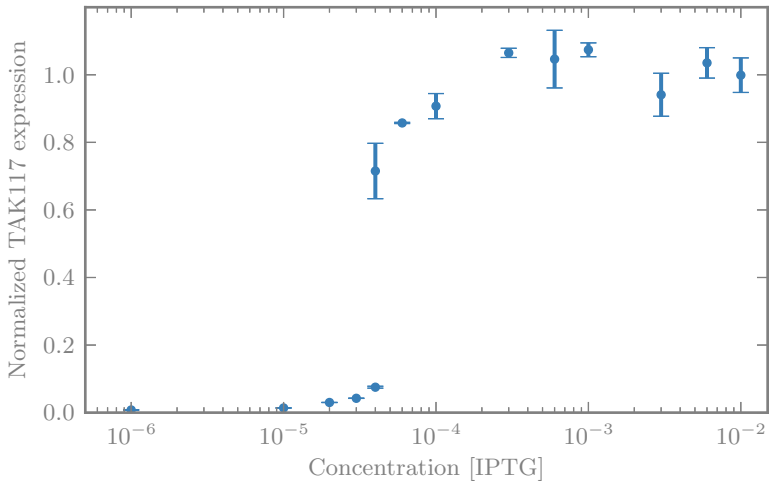

Figure 13: Response of the pTAK117 genetic toggle switch to the input concentration of IPTG (Gardner et al., 2000). The plot shows the mean and standard deviation of the experimentally-observed gene expression levels over a range of input concentrations. Expression levels are normalized by the mean response at the largest IPTG concentration.

Table 1: Normalization of the parameters in the genetic toggle switch example.

	α_1	α_2	β	γ	K	η
$\bar{\theta}_i$	156.25	15.6	2.5	1	2.0015	2.9618e-5
ζ_i	0.20	0.15	0.15	0.15	0.30	0.2

Table 2: Data and obervation error variances for the likelihood of the genetic toggle switch example.

[IPTG]	156.25	15.6	2.5	1	2.0015	2.9618e-5
d_i	0.00798491	1.07691684	1.05514201	0.95429837	1.02147051	1.0
σ_i	4.0e-5	0.005	0.005	0.005	0.005	0.005

D Elliptic PDE inverse problem

Here we provide details about the elliptic PDE inference problem. The forward model is given by the solution of an elliptic PDE in two spatial dimensions

$$\nabla_{\mathbf{s}} \cdot (k(\mathbf{s}, \theta) \nabla_{\mathbf{s}} u(\mathbf{s}, \theta)) = 0, \tag{35}$$

where $\mathbf{s} = (s_1, s_2) \in [0, 1]^2$ is the spatial coordinate. The boundary conditions are

$$\begin{aligned} u(\mathbf{s}, \theta)|_{s_2=0} &= s_1, \\ u(\mathbf{s}, \theta)|_{s_2=1} &= 1 - s_1, \\ \left.\frac{\partial u(\mathbf{s}, \theta)}{\partial s_1}\right|_{s_1=0} &= 0, \\ \left.\frac{\partial u(\mathbf{s}, \theta)}{\partial s_1}\right|_{s_1=1} &= 0. \end{aligned}$$

This PDE serves as a simple model of steady-state flow in aquifers and other subsurface systems; k can represent the permeability of a porous medium while u represents the hydraulic head. Our numerical solution of (35) uses the standard continuous Galerkin finite element method with bilinear basis functions on a uniform 30-by-30 quadrilateral mesh.

The log-diffusivity field $\log k(\mathbf{s})$ is endowed with a Gaussian process prior, with mean zero and an isotropic squared-exponential covariance kernel:

$$C(\mathbf{s}_1, \mathbf{s}_2) = \sigma^2 \exp\left(-\frac{\|\mathbf{s}_1 - \mathbf{s}_2\|^2}{2\ell^2}\right),$$

for which we choose variance $\sigma^2 = 1$ and a length scale $\ell = 0.2$. This prior allows the field to be easily parameterized with a Karhunen-Loève (K-L) expansion (Adler, 1981):

$$k(\mathbf{s}, \theta) \approx \exp\left(\sum_{i=1}^{d} \theta_i \sqrt{\lambda_i} k_i(\mathbf{s})\right),$$

where λ_i and $k_i(\mathbf{s})$ are the eigenvalues and eigenfunctions, respectively, of the integral operator on $[0,1]^2$ defined by the kernel C, and the parameters θ_i are endowed with independent standard normal priors, $\theta_i \sim \mathcal{N}(0,1)$. These parameters then become the targets of inference. In particular, we truncate the Karhunen-Loève expansion at $d = 6$ modes and condition the corresponding mode weights $(\theta_1, \ldots, \theta_6)$ on data. Data arise from observations of the solution field on a uniform 11×11 grid covering the unit square. The observational errors are taken to be additive and Gaussian:

$$d_j = u(\mathbf{s}_j, \theta) + \epsilon_j,$$

with $\epsilon_j \sim \mathcal{N}(0, 0.1^2)$.