# Fast-Mixing Markov Chains without Gradients

Robert Kutri*, Robert Scheichl†

29th June 2026

Labels: *Markov Chain Monte Carlo, SGLMM, Gaussian Process, Surrogates*

Most approaches for accelerating Markov chain mixing either rely on incorporating expensive geometric information in the proposals, or reduce the per-step cost of sampling via surrogate densities. We propose a localisation principle that allows a surrogate-based Metropolis–Hastings proposal to exploit gradient-level geometric information about the target density, without evaluating either the target gradient or the surrogate gradient. The construction relies on regularisation and tempering of the proposal measure. We show that the expected proposal displacement coincides with the Langevin drift up to controlled error. The resulting framework, Delayed Acceptance with Regularisation and Tempering (DART), achieves an $\mathcal{O}(\kappa \max\{\kappa, d\})$ mixing time from warm start for strongly log-concave targets with condition number $\kappa$ in $d$ dimensions. This matches the known $\mathcal{O}(\kappa d)$ rate for MALA when $d \geq \kappa$, and scales as $\mathcal{O}(\kappa^2)$, independent of dimension, otherwise. This is, to our knowledge, the first mixing time guarantee for a surrogate-transition-based MCMC method. We demonstrate DART on a hierarchical spatial generalised linear mixed model. In this setting, the Dirichlet–Neumann averaging parametrisation, originally introduced for the efficient simulation of Gaussian processes, is repurposed to supply the surrogate, and its linear memory and log-linear arithmetic scaling in the number of observation sites carry over to inference.

*. Institute for Mathematics and Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, 69120 Heidelberg, Germany (robert.kutri@uni-heidelberg.de).
†. Institute for Mathematics and Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, 69120 Heidelberg, Germany (r.scheichl@uni-heidelberg.de).

# 1. Motivation

We consider the problem of sampling from a probability measure $\Pi$ with density

$$\pi(x) \propto e^{-f(x)}, \quad \text{for } x \in \mathbb{R}^d,$$

where the continuous function $f : \mathbb{R}^d \to \mathbb{R}$ arises from a computationally intensive model and cannot be expressed in closed form.

Problems of this kind arise throughout statistics and probabilistic machine learning. The function $f$ may represent the negative log-posterior of a Bayesian model obtained through a costly likelihood evaluation [Stu10; Con+16], or a potential arising from marginalising over high-dimensional latent states [DTM98; ADH10; AR09]. In each setting, pointwise evaluation of $f$ at scale is prohibitively expensive, making direct sampling from $\Pi$ intractable. In moderate-to-high dimension, or when $f$ induces complicated geometry, MCMC methods are the standard tool. They require only pointwise evaluations of $f$ or its gradient and yield asymptotically exact samples from $\Pi$.

Their practical efficiency depends on several interacting design choices. The key factors we consider are the mixing time of the chain, which quantifies the number of steps required to reach approximate stationarity, the per-step cost, and the parametrisation of the target density. A suitable parametrisation of $\pi$ can alter the geometry of the sampling target, affecting both mixing and per-step cost simultaneously.

Incorporating local geometric information about $f$ into the transition mechanism is one of the most effective ways to accelerate mixing. Methods based on Langevin dynamics [RT96b; Dal17; BH13] and Hamiltonian Monte Carlo [Dua+87; Nea+11; HG+14] exploit gradient information to construct proposals that respect the local curvature of $f$, and rigorous mixing time guarantees are available in several settings [DMS17; Ebe14; Dwi+19; Che+20]. These gains, however, come at a price. Each transition requires one or more evaluations of $\nabla f$, with methods such as stochastic-Newton MCMC [Mar+12] additionally requiring Hessian evaluations. Moreover, when the target is multimodal, gradient-driven dynamics tend to mix well within individual modes but struggle to traverse the low-probability regions separating them.

Turning to the second strategy, the most direct way to reduce per-step cost is to replace $f$ by a cheaper surrogate $g$ induced by a lower-fidelity model, such as a Gaussian process [Gra20; ST18], a neural operator [Bha+21; Lu+21; Rao+23], or a polynomial approximation [Wes+26]. Substituting $g$ for $f$, however, yields an invariant measure that is biased relative to $\Pi$. Asymptotic exactness can be retained by embedding surrogates more carefully into the MCMC procedure. The surrogate transition method (STM) [Liu01] runs an auxiliary chain targeting a surrogate measure $\tilde{\Pi}$ with density $\tilde{\pi} \propto e^{-g}$ and corrects the resulting proposals via a Metropolis–Hastings step against the true target density. Delayed acceptance (DA) MCMC [CF05; CFO11] uses $g$ to pre-screen proposals before evaluating the expensive model. Both approaches preserve $\Pi$ as the invariant measure of the overall chain.

The Multi-Level Delayed Acceptance (MLDA) method [Lyk+23] combines surrogate transitions and screening across multiple surrogate levels. Explicit non-asymptotic mixing time

bounds in terms of the problem parameters, of the kind available for MALA or HMC [Dwi+19; Che+20], are to our knowledge not available for surrogate-transition, or delayed-acceptance methods. The efficiency of these methods further depends critically on the availability of a surrogate hierarchy whose geometry does not differ too severely from that of the target, and this hierarchy is fixed a priori. Adaptive error modelling schemes [Lyk+20] partially address this by adjusting the surrogate online, but come without a priori guarantees.

## Our Contributions

This manuscript makes three contributions to the design and analysis of surrogate-driven MCMC methods.

The first is methodological. We introduce a localisation principle, achieved through a combination of regularisation and tempering of the surrogate density. This allows a surrogate density with sufficient gradient fidelity to carry the same geometric information as an exact gradient-based proposal, without requiring evaluations of either $\nabla f$ or $\nabla g$. When the surrogate $g$ is quadratic, the localised surrogate density is Gaussian and proposals can be drawn directly, reducing the algorithm to a Metropolis–Hastings variant with a specific, surrogate-informed proposal. When the surrogate is not quadratic, we draw approximate samples by running a short auxiliary Markov chain targeting the localised surrogate density. This extends the surrogate transition framework of [Liu01] and the delayed acceptance approaches of [CF05; Lyk+23] by allowing for localised, state-dependent proposals. We refer to the resulting method as Delayed Acceptance with Regularisation and Tempering (DART).

The second is theoretical. We prove that for strongly log-concave targets with condition number $\kappa$, the DART chain achieves a total-variation mixing time of $\mathcal{O}(\kappa \max\{\kappa, d\})$ from a warm start, provided the surrogate has sufficient gradient fidelity. This bound exhibits two regimes. When $d \geq \kappa$, it matches the $\mathcal{O}(\kappa d)$ rate established for MALA in [Dwi+19].[1] When $\kappa \geq d$, it becomes $\mathcal{O}(\kappa^2)$, with no explicit dependence on $d$. To our knowledge, it is the first explicit mixing time bound for a surrogate-transition-based MCMC method.

The third is of independent interest, but strong synergy with DART. The Dirichlet–Neumann averaging (DNA) construction of [KS26] was developed for the efficient simulation of Gaussian processes on regular grids. We repurpose it as a parametrisation of the latent field in Bayesian models with isotropic Gaussian priors, where inference at scale is typically bottlenecked by a factorisation of the prior and computation of its log-determinant, as well as the simulation of corresponding realisations. We demonstrate the combination on a hierarchical spatial generalised linear mixed model (SGLMM). Under this new parametrisation, and for any sampler, the per-iteration latent field operations collapse from quadratic to near-linear:

1. This MALA bound has subsequently been sharpened to the minimax-optimal $\tilde{\mathcal{O}}(\kappa\sqrt{d})$ [WSC22]. Since our argument closely parallels that of [Dwi+19], we expect the same refinements to transfer. We do not pursue this here.

| | latent field evaluation | factorisation | prior log-determinant |
|---|---|---|---|
| Naive (Cholesky) | $\mathcal{O}(d^2)$ | $\mathcal{O}(d^3)$ | $\mathcal{O}(d^3)$ |
| DNA | $\mathcal{O}(d \log d)$ | $\mathcal{O}(d)$ | $\mathcal{O}(d)$ |

The parametrisation separates the latent field into low- and high-frequency components to yield a natural surrogate structure for DART. Combining the DNA parametrisation and DART unlocks log-linear computational scaling and linear memory scaling for hierarchical SGLMM problems with respect to the number of observations.

The manuscript is organised as follows. Section 2 specifies the general mathematical setting, fixes notation, and recalls the necessary background on MCMC methods and mixing times. Section 3 introduces the DART framework, including the algorithmic construction, the main theoretical results, and practical aspects of implementation. Numerical experiments illustrating the theoretical findings are given in Sections 4. Section 5 introduces the novel Gaussian prior parametrisation and its synergy with DART. Finally, Section 6 contains the proofs of the main results.

## 2. Preliminaries

### 2.1. Notation

We write Pr for the probability measure on a common underlying space carrying all random variables, and $\mathbb{E}$ for the corresponding expectation. All measures on $\mathbb{R}^d$ are defined on the Borel $\sigma$-algebra $\mathfrak{B}(\mathbb{R}^d)$. Throughout, *measurable* means Borel measurable. Absolute continuity and densities are understood with respect to the Lebesgue measure, and we leave this implicit.

We write $\mathfrak{P}(\mathbb{R}^d)$ for the space of Borel probability measures on $\mathbb{R}^d$, denote the Gaussian measure with mean $\bar{x}$ and covariance $C$ by $\mathcal{N}(\bar{x}, C)$, and the Dirac measure at $x$ by $\delta_x$. For $\nu \in \mathfrak{P}(\mathbb{R}^d)$ and $\psi \sim \nu$ we write $\mathbb{E}_\nu[\cdot]$ for expectation under $\nu$.

The total variation distance between $\mathcal{M}_1, \mathcal{M}_2 \in \mathfrak{P}(\mathbb{R}^d)$ is

$$\|\mathcal{M}_1 - \mathcal{M}_2\|_{\mathrm{TV}} := \sup_{A \in \mathfrak{B}(\mathbb{R}^d)} \big|\mathcal{M}_1(A) - \mathcal{M}_2(A)\big|.$$

With this normalisation, every $\mathcal{M} \in \mathfrak{P}(\mathbb{R}^d)$ satisfies the functional bound

$$\left| \int f \, \mathrm{d}\mathcal{M}_1 - \int f \, \mathrm{d}\mathcal{M}_2 \right| \leq \|\mathcal{M}_1 - \mathcal{M}_2\|_{\mathrm{TV}} \tag{2.1}$$

for any measurable $f : \mathbb{R}^d \to [0, 1]$.

We write $\mathbb{B}(x, R)$ for the closed Euclidean ball centred at $x$ with radius $R$, and $\|\cdot\|_2$ for the Euclidean norm. For non-negative quantities, $A \lesssim B$ means $A \leq cB$ for a universal constant $c$ that depends on none of the problem parameters, $A \gtrsim B$ means $B \lesssim A$, and $A \asymp B$ means

both $A \lesssim B$ and $B \lesssim A$. The same symbol may denote different constants in different expressions. We write $f \leq g$ for pointwise inequality between functions.

Given a density $\pi$ on $\mathbb{R}^d$, we write $\pi \propto e^{-f}$ to mean $\pi(x) = N^{-1} e^{-f(x)}$, where $N := \int_{\mathbb{R}^d} e^{-f(x)} \, \mathrm{d}x$ is the normalisation constant.

## 2.2. Markov Chains and Metropolis–Hastings

A Markov chain on $\mathbb{R}^d$ is a sequence of random variables $\Psi = (\psi_i)_{i \in \mathbb{N}}$ whose dynamics, is fully determined by a family of transition probability measures $\{\mathcal{T}_x\}_{x \in \mathbb{R}^d}$ satisfying

$$\Pr\left(\psi_i \in A \,\middle|\, \psi_{i-1} = x\right) = \mathcal{T}_x(A)$$

for any $A \in \mathfrak{B}(\mathbb{R}^d)$ and $i \in \mathbb{N}$. The initial state follows a probability measure $\psi_0 \sim \mu$, referred to as the *initial measure*. We assume throughout that $x \mapsto \mathcal{T}_x(A)$ is measurable. The associated *transition operator* is then

$$T : \mathcal{P}(\mathbb{R}^d) \to \mathcal{P}(\mathbb{R}^d), \qquad \big(T\nu\big)(A) := \int_{\mathbb{R}^d} \mathcal{T}_x(A) \, \mathrm{d}\nu(x), \quad A \in \mathfrak{B}(\mathbb{R}^d),$$

so that $\psi_i \sim T^i \mu$ for any $i \in \mathbb{N}$. A probability measure $\nu$ is *invariant* under $T$ if $T\nu = \nu$. We refer to $T$ itself as the (Markov) chain, when the context is clear. The goal of MCMC is to construct a chain for which

$$\| T^i \mu - \Pi \|_{\mathrm{TV}} \to 0, \quad \text{as} \quad i \to \infty, \tag{2.2}$$

so that for large $i$, the states of the chain can be used to perform inference with respect to $\Pi$. The Metropolis–Hastings (MH) algorithm [Met+53; Has70] provides a general construction achieving this. Given a family of absolutely continuous proposal measures $\{\mathcal{P}_x\}_{x \in \mathbb{R}^d}$ with densities $p_x$, a proposal $z \sim \mathcal{P}_x$ is accepted as the next state with probability

$$\alpha(x, z) := \min\left\{1, \frac{\pi(z)\, p_z(x)}{\pi(x)\, p_x(z)}\right\}, \tag{2.3}$$

and otherwise the chain remains at $x$. Provided that $\pi$ and the $p_x$ are positive a.e., the resulting chain has invariant measure $\Pi$ and satisfies (2.2) for $\Pi$-almost every initial state [RT96a; MT08]. Two standard instantiations are the Metropolised random walk (MRW), which uses Gaussian proposals $\mathcal{P}_x = \mathcal{N}(x, h^2 \mathbb{I})$, and the Metropolis-adjusted Langevin algorithm (MALA) [RT96b], for a step size parameter $h > 0$, which shifts proposals toward regions of higher density via $\mathcal{P}_x = \mathcal{N}\big(x + \frac{h^2}{2} \nabla \log \pi(x),\, h^2 \mathbb{I}\big)$.

## 2.3. Mixing Times and Conductance

While constructing a convergent Markov chain in the sense of (2.2) is possible for a broad class of target measures, quantifying the speed of this convergence is considerably more delicate. In particular, it depends sensitively on the geometry of the density $\pi$. We measure

the speed of convergence by the number of steps required for a prescribed total-variation tolerance.

**Definition 2.1** – *Mixing Time* –

Let $\delta > 0$. For a fixed initial measure $\mu$ on $(\mathbb{R}^d, \mathfrak{B}(\mathbb{R}^d))$, the *$\delta$-mixing time* $t_\delta(\mu)$ of a Markov chain with transition operator $T$ is defined as

$$t_\delta(\mu) := \inf\{n \in \mathbb{N} : \|T^n\mu - \Pi\|_{\mathrm{TV}} < \delta\}.$$

Mixing can be slow in general. For multi-modal target densities, local proposals struggle to cross low-probability regions separating modes, leading to poor global exploration. Even for unimodal targets, convergence can be arbitrarily slow if the chain is initialised from an unfavourable position [Ban+23]. This motivates restricting attention to classes of initial measures that represent beneficial initial configurations.

**Definition 2.2** – *Warmness* –

Let $\nu$ be a probability measure on $(\mathbb{R}^d, \mathfrak{B}(\mathbb{R}^d))$ with continuous density $\varphi$. For $\beta \geq 1$, we say that $\nu$ is *$\beta$-warm* with respect to $\Pi$ if

$$\sup_{x \in \mathbb{R}^d} \frac{\varphi(x)}{\pi(x)} \leq \beta.$$

We say that $\nu$ is *warm* with respect to $\Pi$ if it is $\beta$-warm for some finite $\beta$.

A complementary notion is the *conductance* $\Phi$ of a chain $T$ with invariant measure $\Pi$, defined by

$$\Phi(T) := \inf_{\Pi(A) \in (0,1/2)} \frac{\int_A \mathcal{T}_x(A^c)\,\pi(x)\,\mathrm{d}x}{\Pi(A)}.$$

Conductance measures the minimal probability flux leaving a set relative to its target mass and thus quantifies the presence of 'bottleneck regions' in the state space. Cheeger constants in differential geometry and graph theory are closely related notions.

To discount sets of negligible target mass, one introduces the $s$-conductance $\Phi_s(T)$, defined for $s \in (0, \frac{1}{2})$ by

$$\Phi_s(T) := \inf_{\Pi(A) \in (s,\, 1/2)} \frac{\int_A \mathcal{T}_x(A^c)\,\pi(x)\,\mathrm{d}x}{\Pi(A) - s}. \tag{2.4}$$

Warmness interacts favourably with $s$-conductance. Indeed, for any measurable $A \subset \mathbb{R}^d$ with $\Pi(A) < s$ and $s \in (0, \frac{1}{2})$, a $\beta$-warm measure $\nu$ satisfies

$$\nu(A) = \int_A \varphi(x)\,\mathrm{d}x \leq \beta \int_A \pi(x)\,\mathrm{d}x = \beta\Pi(A) \leq \beta s, \tag{2.5}$$

and therefore $|\nu(A) - \Pi(A)| \leq \beta s$.

As a consequence, Theorem 1.4 and Corollary 1.6 in [LS93] imply that for a $\beta$-warm initial measure $\mu$,

$$\|T^i\mu - \Pi\|_{\mathrm{TV}} \leq \beta s + \beta\left(1 - \tfrac{1}{2}\Phi_s^2(T)\right)^i, \qquad i \in \mathbb{N}. \tag{2.6}$$

Thus, lower bounds on the $s$-conductance (2.4) combined with a suitable choice of $s$, yield explicit non-asymptotic bounds on mixing times. This strategy has been successfully employed in, for example, [LS93; Lov99; Dwi+19; Che+20], and forms the basis of the analysis in Section 6.

## 3. Localised Surrogate Transitions

Let $g : \mathbb{R}^d \to \mathbb{R}$ be a computationally cheaper approximation to the target potential $f$. We construct a state-dependent proposal measure $\Pi_x$ by two modifications of the surrogate density $\tilde{\pi} \propto e^{-g}$. First, similar to a restricted Gaussian oracle [TP18; LST21] for the surrogate, we *localise* around the current state $x \in \mathbb{R}^d$ by adding a quadratic penalty scaled by $\gamma > 0$, confining proposals to a neighbourhood of $x$ where short steps are more likely to be accepted. Second, in the spirit of simulated annealing [KGV83; MP92; GT95], we *temper* the surrogate contribution by a factor $\theta \in (0, 1]$, making the resulting density easier to sample from. Together, these modifications yield the potential

$$V_x : \mathbb{R}^d \to \mathbb{R}, \qquad y \mapsto \theta\, g(y) + \frac{\gamma}{2}\|y - x\|_2^2,$$

and we define the corresponding proposal measure $\Pi_x$ with density

$$\pi_x(y) \propto e^{-V_x(y)}, \qquad y \in \mathbb{R}^d.$$

Our core claim for the construction above is the following: a draw $z \sim \Pi_x$ has the same expected drift towards high-density regions of $\pi$ as a Langevin proposal, despite never evaluating a gradient.

To illustrate when and why this might be true, define the mean displacement $\bar{u}_x := \mathbb{E}_{u\sim\Pi_x} u$ of the proposal for a fixed state $x \in \mathbb{R}^d$ and assume for the moment that $\nabla g = \nabla f$. If $\pi_x$ has (say) sub-Gaussian tails, then $\mathbb{E}_{u\sim\Pi_x}\nabla V_x(u) = 0$ by integration by parts. Expanding $\nabla V_x(u) = \theta\,\nabla g(u) + \gamma(u - x)$ and rearranging yields

$$\bar{u}_x = x - \frac{\theta}{\gamma}\,\mathbb{E}_{u\sim\Pi_x}\nabla g(u) = x - \frac{\theta}{\gamma}\,\mathbb{E}_{u\sim\Pi_x}\nabla f(u). \tag{3.1}$$

The proposal mean is shifted from $x$ along the $\Pi_x$-averaged target gradient, with $\theta/\gamma$ acting as an implicit step size. A MALA proposal with step size $h = \sqrt{2\theta/\gamma}$ produces the same drift with the pointwise gradient $\nabla f(x)$ in place of the average gradient $\mathbb{E}_{u\sim\Pi_x}\nabla f(u)$.

Two technical hurdles stand in the way of the heuristic (3.1) reliably producing gradient-informed proposals. No reasonably cheap surrogate can be expected to satisfy $\nabla g = \nabla f$, so the first hurdle lies in quantifying when the surrogate density carries sufficient gradient information. Secondly, for the average gradient to be a useful proxy for the pointwise target gradient, the concentration of $\Pi_x$, controlled chiefly through increasing $\gamma$, needs

to be strong enough. We address both issues for strongly log-concave target densities in Theorem 3.4, where the gradient fidelity condition (3.5) and the required level of localisation (3.6) are made explicit.

In DART, we therefore propose drawing $z \sim \Pi_x$ and accepting it via an MH step against the target density $\pi$. Denote by $N_x := \int_{\mathbb{R}^d} e^{-V_x(y)}\,\mathrm{d}y$ the normalisation constant of $\pi_x$. Since the densities $\pi_x$ and $\pi_z$ are normalised by different constants when $x \neq z$, these constants enter the standard MH ratio (2.3) explicitly. Owing to the symmetry of the quadratic penalty, we have

$$V_x(z) - V_z(x) = \theta\big(g(z) - g(x)\big),$$

and expanding the proposal density ratio gives the acceptance probability

$$\alpha(x, z) = \min\left\{1,\ \frac{\pi(z)}{\pi(x)}\,\frac{N_x}{N_z}\,e^{\theta\,(g(z)-g(x))}\right\}. \tag{3.2}$$

When $z \sim \Pi_x$, the resulting Markov chain on $\mathbb{R}^d$ is reversible with respect to $\Pi$ and admits $\Pi$ as its invariant measure.

Drawing $z \sim \Pi_x$ exactly is feasible only in special cases, the quadratic-surrogate case discussed in Section 3.2 being the main one. In general, we approximate samples from $\Pi_x$ by running a Markov chain, which we refer to as the *root chain*, for $n$ steps. That is, we replace the exact proposal $\mathcal{P}_x = \Pi_x$ by

$$\mathcal{P}_x = Q_x^n\,\mu_x \approx \Pi_x,$$

where $Q_x$ denotes the transition operator of a reversible Markov chain with invariant measure $\Pi_x$ and $\mu_x$ is an initial measure.

Because the density of $Q_x^n\mu_x$ is in general intractable, we cannot form its exact Metropolis ratio and instead retain (3.2), the ratio for exact proposals from $\Pi_x$. The implemented chain is then not in detailed balance with $\Pi$, and its bias is governed by the root-chain mixing via $\|\Pi_x - Q_x^n\mu_x\|_{\mathrm{TV}}$. Both localisation and tempering reduce this error directly, the quadratic penalty making the localised surrogate well-conditioned and tempering flattening it further. Condition 3.3 fixes the root-chain length $n$ this requires, which Theorem 3.4 subsequently folds into the mixing-time bound.

This use of a state-dependent root chain is the sense in which DART extends the surrogate transition method (STM) [Liu01], delayed acceptance (DA) [CF05], and the Multi-Level Delayed Acceptance (MLDA) method [Lyk+23], all of which propose from a fixed, global surrogate. Two-level MLDA is recovered as $\gamma \to 0$ and $\theta = 1$, a regime favourable only when the surrogate is at once faithful and cheap. Viewed from the proximal-sampling side, the root chain realises a restricted Gaussian oracle on the surrogate. The Metropolis step corrects it exactly, and our analysis (Section 6) shows the error left by running it for finite time can be made as small as desired at only logarithmic extra cost.

## 3.1. A Mixing Time Guarantee for DART

The main theoretical result of this manuscript is an explicit upper bound on the mixing time of a Markov chain using DART from a warm start, under the assumption of a strongly log-concave target density. The bound applies to the general setting where proposals are generated by a root chain of finite length targeting the localised surrogate measure $\Pi_x$. Corollary 3.5 below shows that the idealised and implemented chains coincide when proposals can be drawn from $\Pi_x$ directly, as is the case for quadratic surrogates, and the bound then holds in a stronger, monotone form.

Log-concave targets arise naturally in Gaussian models, logistic regression, and other members of the exponential family of response functions. More broadly, posterior measures arising from non-linear statistical inverse problems admit accurate log-concave approximations in Wasserstein distance under suitable conditions [Nic23, Theorem 5.1.3]. We summarise the assumptions on the target geometry in the following condition.

**Condition 3.1.** The measure $\Pi$ is absolutely continuous with respect to the Lebesgue measure with density $\pi \propto e^{-f}$, where

1. The potential $f$ is continuously differentiable and has a Lipschitz continuous gradient with constant $L > 0$. That is, for all $x, y \in \mathbb{R}^d$,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2.$$

2. The potential $f$ is $\lambda$-strongly convex for some $\lambda > 0$, meaning that for all $x, y \in \mathbb{R}^d$,

$$\langle \nabla f(x) - \nabla f(y),\, x - y \rangle \geq \lambda \|x - y\|_2^2.$$

Strong convexity guarantees concentration of $\Pi$ around a unique mode $x^\star := \arg\min_{\mathbb{R}^d} f$ (see e.g. [DM19]). Lipschitz continuity of the gradient provides quantitative control over how $f$ varies along a proposal move. We quantify the geometric complexity of a target density satisfying Condition 3.1 through its *condition number*

$$\kappa := \frac{L}{\lambda}. \tag{3.3}$$

The following definition summarises the surrogate construction for a target measure satisfying Condition 3.1.

**Definition 3.2** – *Strongly Log-Concave Surrogate Density* –

Let $g : \mathbb{R}^d \to \mathbb{R}$ be $m$-strongly convex and $M$-smooth with unique minimiser $\hat{x} := \arg\min_{\mathbb{R}^d} g$. For any $x \in \mathbb{R}^d$ and parameters $\theta \in (0, 1]$, $\gamma > 0$, define the potential

$$V_x(y) := \theta\, g(y) + \frac{\gamma}{2} \|x - y\|_2^2. \tag{3.4}$$

Let $\Pi_x$ denote the probability measure with density $\pi_x$ proportional to $e^{-V_x}$. We refer to $\pi_x$ as a *strongly log-concave surrogate density.*

The analysis requires two kinds of agreement between $g$ and $f$. The surrogate's mode $\hat{x}$ must lie in the same high-probability region of $\Pi$ as the target's mode $x^\star$, formalised in Theorem 3.4 by the inclusion $\hat{x} \in \mathcal{K}$. Within this region of high probability, the gradient $\nabla g$ must approximate $\nabla f$ sufficiently well. To ensure local regularity of $\nabla g$ as $\|x - y\|_2 \to 0$, we further require that $g$ is itself $m$-strongly convex with an $M$-Lipschitz gradient. Compatibility between these two requirements is ensured by $\lambda \le m \le M \le L$.

The final condition characterises the efficiency of the root chain used by the algorithm. We assume the root chain mixes in polynomial time in the dimension and condition number, with exponents $\omega, \tilde{\omega} > 0$. This is satisfied, for instance, by MRW, MALA, and HMC targeting strongly log-concave densities (see, for example, [Dwi+19; Che+20]).

**Condition 3.3.** Let $\{Q_x\}_{x\in\mathbb{R}^d}$ be a family of transition operators where each $Q_x$ has invariant measure $\Pi_x$ with potential $V_x$ that is strongly convex with Lipschitz continuous gradient. Denote by $\tilde{\kappa} := \sup_{x\in\mathbb{R}^d} \kappa(\Pi_x) \ge 1$ the largest condition number (3.3) among the $\Pi_x$, assumed finite. There exist $\omega, \tilde{\omega} > 0$ such that for every $x \in \mathbb{R}^d$, every $\beta_x$-warm initial measure $\mu_x$ with respect to $\Pi_x$, and every $\varepsilon \in (0, 1)$,

$$\|Q_x^n \mu_x - \Pi_x\|_{\mathrm{TV}} \le \varepsilon \quad \text{provided that} \quad n \gtrsim d^\omega \tilde{\kappa}^{\tilde{\omega}} \log \frac{2\beta_x}{\varepsilon}.$$

Under the parameter choices of Theorem 3.4, the surrogate condition number $\tilde{\kappa}$ is bounded by a universal constant (see (6.30)), and Lemma 6.6 shows that the warmness requirement is met by suitably concentrated Gaussian initial measures, with $\log \beta_x$ controlled uniformly over the region the chain explores.

### Theorem 3.4

*Let $f : \mathbb{R}^d \to \mathbb{R}$ satisfy Condition 3.1 and denote the mode $x^\star := \arg\min_{\mathbb{R}^d} f$. Let $g : \mathbb{R}^d \to \mathbb{R}$ be an $m$-strongly convex surrogate for $f$ with $M$-Lipschitz gradient, satisfying $\lambda \le m \le M \le L$, and denote $\hat{x} = \arg\min_{\mathbb{R}^d} g$. Let $\Pi$ be the target measure with density $\pi \propto e^{-f}$, let $\mu$ be a $\beta$-warm start w.r.t. $\Pi$, and fix an error tolerance $\delta \in (0, 1)$ satisfying $\log \frac{2\beta}{\delta} \le d$. Define the ball $\mathcal{K} := \mathbb{B}\big(x^\star, 3\sqrt{d/\lambda}\big)$, and assume $\hat{x} \in \mathcal{K}$. Suppose the following hold:*

*(i)* Gradient fidelity.

$$\sup_{x\in\mathcal{K}} \|\nabla f(x) - \nabla g(x)\|_2^2 \lesssim L \max\{1,\ \kappa/d\}. \tag{3.5}$$

*(ii)* Localisation. *The DART parameters satisfy*

$$\theta = \frac{1}{2} \quad \textit{and} \quad \gamma \asymp L \max\{\kappa,\ d\}. \tag{3.6}$$

*(iii)* Root chain mixing. *The root chains* $\{Q_x\}_{x\in\mathbb{R}^d}$ *satisfy Condition 3.3, each initialised at the Gaussian measure* $\mu_x = \mathcal{N}\big(x,\, (2(\gamma+\theta M))^{-1}\mathbb{I}\big)$ *of Lemma 6.6, and the root chain lengths satisfy*

$$n \gtrsim d^{\omega}\Big(d + \log\tfrac{\kappa}{\delta}\Big). \tag{3.7}$$

*Then the* $\delta$*-mixing time of the DART chain satisfies*

$$t_\delta(\mu) \lesssim \kappa \max\{\kappa,\ d\}\, \log\frac{2\beta}{\delta}. \tag{3.8}$$

When proposals are drawn from $\Pi_x$ exactly, the perturbation term $\|\Pi_x - Q_x^n\mu_x\|_{\mathrm{TV}}$ vanishes, the chain satisfies detailed balance with respect to $\Pi$, and the mixing time bound follows from conditions (i) and (ii) alone.

**Corollary 3.5.** *Under the hypotheses of Theorem 3.4, suppose that conditions* (i) *and* (ii) *hold, and that proposals are drawn exactly from* $\Pi_x$*. Then Condition 3.3 and condition* (iii) *are not required, and the mixing time bound* (3.8) *holds.*

**Remark 3.6.** The condition $\log(2\beta/\delta) \le d$ is cosmetic. Removing it introduces a dependence on $r(\delta/(4\beta), d)$ (see [Dwi+19, Theorems 1 and 2] and (6.24)), which increase only very slowly with decreasing error tolerance. As stated, Theorem 3.4 restricts the mixing times to error tolerances $\delta \gtrsim e^{-d}$. Furthermore, our results can be extended to weakly log-concave and perturbed log-concave targets using the reduction employed in [Dwi+19], ensuring algebraic mixing time bounds for those targets too. Since the arguments translate verbatim, we omit these computations here.

**Remark 3.7.** The mixing time bound (3.8) reveals two distinct regimes. If $\kappa \ge d$, then $t_\delta(\mu) \in \mathcal{O}(\kappa^2)$, with no explicit dependence on the dimension $d$. If $d \ge \kappa$, the bound becomes $t_\delta(\mu) \in \mathcal{O}(\kappa d)$, recovering the mixing time bound for MALA in [Dwi+19, Theorem 1]. As shown in Lemma 6.8, these regimes emerge from the localisation having to control two competing effects. The $\mathcal{O}(\kappa^2)$ scaling stems from the stronger localisation required to control the displacement of the surrogate mode from the current state in the presence of steep target gradients, requiring $\gamma \gtrsim L\kappa$. The $\mathcal{O}(\kappa d)$ regime arises from the impact of the natural fluctuations of $z \sim \Pi_x$ on the acceptance probability and can be controlled by $\gamma \gtrsim Ld$.

**Remark 3.8.** Using Markov chains to draw proposals $z \sim \mathcal{P}_x = Q_x^n\mu_x \approx \Pi_x$ is a natural choice and allows the proposal quality to be quantified in terms of the chain length $n$. The proof in Section 6, however, requires only $\|\mathcal{P}_x - \Pi_x\|_{\mathrm{TV}} \le \delta/(2N_\delta)$ on $\mathcal{K}_N$, as seen in (6.29), regardless of how the proposal measure is constructed. Any family of samplers whose single-draw law satisfies this bound yields the same mixing time bound as Theorem 3.4, with the acceptance step (3.2) unchanged. Exact sampling, as in Corollary 3.5, is one such instance. The inner sampler can therefore be chosen on purely computational grounds, provided the required total variation accuracy can be established, in which case the mixing analysis carries over without modification.

## 3.2. Practical Aspects

Theorem 3.4 is in a strongly regularised regime. This is deliberate, as the resulting strong localisation confines proposals to a neighbourhood of the current state, where gradient fidelity is the dominant concern. In practice, substantially smaller values of $\gamma$ are effective, for two reasons. First, the target's own strong convexity already provides a degree of localisation that the analysis neglects. In the proof, the localisation term is taken to dominate the curvature of $f$ (cf. Lemmas 6.6 and 6.8), but in practice the two curvatures compound, and a smaller $\gamma$ suffices. Second, the fidelity condition (3.5) only controls first-order information. If the surrogate also captures curvature information of the target, the localisation would only need to compensate for the remaining curvature mismatch. The theorem should therefore be read as a rigorous illustration of the principle that localised surrogate transitions can extract gradient-level geometric information from the surrogate, rather than as a prescription of optimal algorithmic parameters.

The acceptance probability (3.2) involves the ratio $N_x/N_z$ of normalisation constants. When $g$ is quadratic, say $g(y) = \frac{1}{2}(y-\hat{x})^\top A\,(y-\hat{x})$ for a positive-definite matrix $A$, the surrogate density $\pi_x$ is Gaussian with precision $\theta A + \gamma\mathbb{I}$. The normalisation constant $N_x$ then follows from the determinant of this precision, and the ratio $N_x/N_z$ is available in closed form. This covers, for instance, Laplace approximations, where $A$ is the Hessian at the MAP, and the high-frequency components of the DNA surrogate in Section 5, where $A$ is a diagonal matrix of prior precision eigenvalues.

When $g$ is not quadratic, the ratio must be estimated. The difference of potentials

$$V_z(v) - V_x(v) = \gamma\,\langle v - \tfrac{x+z}{2},\, x - z\rangle$$

is linear in $v$, since the quadratic terms cancel. This allows the reciprocal ratio to be expressed as

$$\frac{N_z}{N_x} = \mathbb{E}_{v\sim\Pi_x}\; e^{-\gamma\,\langle v-\frac{x+z}{2},\,x-z\rangle}. \tag{3.9}$$

The root chain already generates samples approximately distributed according to $\Pi_x$, so these can be repurposed to estimate (3.9) at no additional model cost. Given $\tilde{n}$ effectively independent states $\{v_i\}_{i=1}^{\tilde{n}}$ from the root chain, the estimator

$$\hat{R}_{\mathrm{IS}}^{-1}(x,z) \;:=\; \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} e^{-\gamma\,\langle v_i-\frac{x+z}{2},\,x-z\rangle} \tag{3.10}$$

is unbiased for $N_z/N_x$, requires no additional target evaluations, and is exact when $x = z$. Taking the reciprocal $\hat{R}_{\mathrm{IS}}(x,z) := 1/\hat{R}_{\mathrm{IS}}^{-1}(x,z)$ introduces an upward bias of order $\mathcal{O}(1/\tilde{n})$ by Jensen's inequality.

Equation (3.10) is the empirical mean of $e^{-\ell}$, with $\ell(v) := \gamma\,\langle v - \frac{x+z}{2},\, x - z\rangle$. A second estimator follows from approximating $\Pi_x$ by the Gaussian derived from the chain's empirical first two cumulants. This too costs no further evaluations and gives

$$\log \hat{R}_{\mathrm{G}}^{-1}(x,z) \;:=\; -\bar{\ell} + \tfrac{1}{2}s_\ell^2, \tag{3.11}$$

with the sample mean $\bar{\ell}$ and variance $s_\ell^2$. It estimates the log-ratio directly, avoiding the exponentiation and the reciprocal step that produced the Jensen bias of (3.10).

The two estimators trade consistency against efficiency. Estimator (3.11) is exact whenever $\Pi_x$ is Gaussian, recovering the closed-form quadratic case, and returns zero at $x = z$. For skewed, non-Gaussian $\Pi_x$ it retains a bias that does not vanish as $\tilde{n} \to \infty$. The importance-sampling estimator (3.10) is instead consistent, at higher variance and with less stable exponential weighting. The two estimators are compared empirically in Section 4.

When the surrogate $g$ is itself a Bayesian posterior with a known Gaussian prior of precision $C^{-1}$, the full Gaussian component of $V_x$ has precision $\theta\, C^{-1} + \gamma\mathbb{I}$, providing a natural pCN reference measure [Cot+13]. Tempering only the likelihood component of $g$ in such instances, while leaving the prior at full strength, preserves the prior geometry and ensures that the localisation remains solely responsible for confining proposals. This is employed in Sections 4 and 5.

In practice, we initialise the root chain at the current outer state, setting $\mu_x = \delta_x$. Although $\delta_x$ is not a warm start for $\Pi_x$, the transition operator $Q_x$ yields an absolutely continuous measure $Q_x\delta_x$ after one step, which, provided the localisation is sufficient, concentrates adequately to serve as an approximately warm start. We also implement a *crank-up* procedure: an auxiliary chain targeting $\widetilde{\pi} \propto e^{-g}$ is run for a fixed number of steps, and its terminal state is used as the initial state of the outer chain, giving $\mu \approx \widetilde{\Pi}$.

# 4. Empirical Behaviour

The experiments in Sections 4 and 5 rely on `styne`, a standalone Python library available at `https://github.com/rkutri/styne`. The scripts reproducing all experiments and figures in this paper reside under `reproducibility/dart` in release `v0.1.0`, browsable at `https://github.com/rkutri/styne/tree/v0.1.0`.

## 4.1. Stability of the Ratio Estimators

We isolate the accuracy of the estimation step on a two-component Gaussian mixture with unequal weights. The projected law of the root chain samples is skewed, so the experiment exercises $\hat{R}_{\mathrm{G}}$ outside the Gaussian regime in which it is exact. We set $\theta = 1$, since tempering would destroy the mixture form and with it the analytic ground truth. Localisation is therefore controlled by $\gamma$ alone.

We fix $x$ at the mode of the dominant component, draw independent realisations of $z \sim Q_x^n \mu_x$, where $Q_x$ is MALA on $\Pi_x$ initialised at $\mu_x = \delta_x$, and compute $\hat{R}_{\mathrm{IS}}(x, z)$ and $\hat{R}_{\mathrm{G}}(x, z)$ from the same trajectory.

Figure 1 reports the absolute error in $\log(N_z/N_x)$ as a function of root chain length $n$, across dimensions and localisation strengths. As $\gamma$ grows, $\Pi_x$ and $\Pi_z$ tighten and their overlap shrinks, so the ratio is harder to estimate, a difficulty that compounds with dimension. The two estimators agree while sampling is cheap. Once the log-weight variance grows, $\hat{R}_{\mathrm{IS}}$ degrades under large weights while $\hat{R}_{\mathrm{G}}$ does not. This is the trade-off of Section 3.2 made

visible, and it is why we use $\hat{R}_{\mathrm{G}}$ as the default estimator. While $\hat{R}_{\mathrm{G}}$ is more reliable in this experiment, it is expected to be more vulnerable to surrogate geometries that differ more strongly from its Gaussian approximation.

How much residual estimation noise affects DART's overall mixing depends on the relative sizes of all terms in (3.2), not on the ratio error alone. The subsequent experiments confirm that, once $\gamma$ is chosen appropriately, DART improves over its non-localised counterparts.

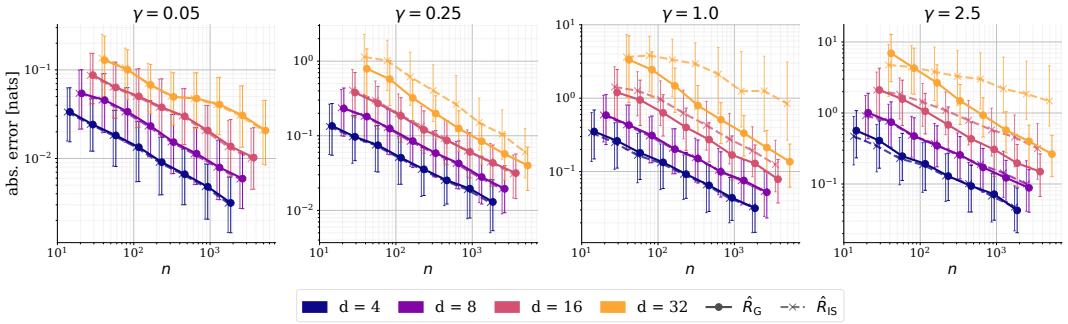


**Figure 1:** Median absolute error in $\log(N_z/N_x)$, with interquartile range over 10 000 independent runs, for $\hat{R}_{\mathrm{IS}}$ (dashed) and $\hat{R}_{\mathrm{G}}$ (solid) as a function of root chain length $n$. The mixture has weights $(0.7, 0.3)$, means $\pm a$ with $a = 0.5$, common covariance $\sigma^2\mathbb{I}$ with $\sigma = 1$, and $\theta = 1$. The root chain is MALA on $\Pi_x$ initialised at $\mu_x = \delta_x$, burn-in fraction 0.3, no thinning.

## 4.2. The Quadratic Surrogate Regime

We now test the full DART chain in the regime of Corollary 3.5, where the surrogate $g$ is quadratic and proposals can be drawn from $\Pi_x$ directly. Following [Dal17; Dwi+19], we consider a Bayesian logistic regression example. Given $n_{\mathrm{obs}}$ observations $(c_i, y_i) \in \mathbb{R}^d \times \{0, 1\}$, the probability that $y_i = 1$ is modelled as $\sigma(\langle x, c_i\rangle)$, with $\sigma(t) = (1 + e^{-t})^{-1}$. Covariates $c_i$ have i.i.d. Rademacher entries and unit Euclidean norm and responses are drawn from a fixed data-generating parameter $x^\star$. The prior is $\mathcal{N}(0, (\alpha\Sigma_C)^{-1})$ with $\Sigma_C = \frac{1}{n_{\mathrm{obs}}} C^\top C$ and parameter $\alpha$, whose value can be chosen freely.

After preconditioning by $\Sigma_C^{-1/2}$, the prior becomes $\mathcal{N}(0, \alpha^{-1}\mathbb{I}_d)$, the preconditioned design matrix satisfies $\tilde{C}^\top \tilde{C} = n_{\mathrm{obs}}\mathbb{I}_d$, and the negative log-posterior is

$$f(x) = -Y^\top \tilde{C}\, x + \sum_{i=1}^{n_{\mathrm{obs}}} \log\left(1 + e^{\langle x, \tilde{c}_i\rangle}\right) + \frac{\alpha}{2}\|x\|_2^2,$$

with gradient Lipschitz constant $L = \frac{1}{4} n_{\mathrm{obs}} + \alpha$, strong convexity constant $\lambda = \alpha$, and condition number $\kappa = L/\lambda$, independent of the spectrum of the original design matrix. With $n_{\mathrm{obs}}$ and $\alpha$ fixed, both $L$ and $\kappa$ are the same at every dimension.

As surrogate for $f$ we use the Laplace approximation at the MAP $\hat{x} = \arg\min_{\mathbb{R}^d} f$,

$$g(y) = \tfrac{1}{2}\langle y - \hat{x},\ H\,(y - \hat{x})\rangle,$$

where $H$ is the Hessian of $f$ at $\hat{x}$. Since $g$ is quadratic, the localised surrogate density is Gaussian with precision $P = \theta H + \gamma \mathbb{I}_d$ and mean $P^{-1}(\theta H \hat{x} + \gamma x)$. Both $P$ and the normalisation ratio $N_x/N_z$ are available in closed form, so no root chain or ratio estimation is needed.

We compare DART against MALA, MLDA and MRW in the preconditioned space. MALA and MRW step sizes are tuned to standard target acceptance rates and MLDA shares DART's surrogate, but canonically, without regularisation or tempering. The implicit step size of DART is governed by $\gamma$ and is swept over $\gamma/L \in [10^{-2}, 10]$. The value at the peak of the 'Efficiency' panel of Figure 2 is used for DART. To assess convergence, we track the running mean diagnostic

$$e_k = \frac{1}{d}\|\bar{x}_k - x_{\text{ref}}\|_1, \qquad \bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i,$$

where $x_{\text{ref}} = \Sigma_C^{1/2} x^\star$ is the data-generating parameter in the preconditioned coordinates. Since every chain targets the same posterior, $e_k$ decays to a common floor, the distance between the posterior mean and $x_{\text{ref}}$. All chains share common initialisations drawn from the prior. To compare stationary-phase efficiency, we report effective sample size (ESS) per iteration along the slowest direction of the target, the eigenvector of $H$ with the smallest eigenvalue. We project $M = 8$ replicate chains, each started from an independent prior draw, onto this direction and combine them with a multi-chain ESS estimator that pools the within-chain autocorrelation with the between-chain variance (c.f. [GR92]).

Figure 2 summarises the results. MLDA employs the same surrogate as DART but lacks localisation. As its sub-chain equilibrates, the proposals approach independent draws from the Laplace approximation. The left panel demonstrates that this measure lacks the fidelity required to serve the full model. The success of this same surrogate under DART isolates the effect of the localisation $\gamma$ in keeping each proposal local. Finally, MRW converges slowly but reliably.

DART converges faster than MALA, without evaluating the target gradient at any iteration. In line with the discussion in Section 3.2, when in a beneficial localisation regime (see middle panel in Figure 2), curvature information captured by the Laplace approximation further improves DART over MALA, which only has access to gradient information. The middle panel of Figure 2 illustrates the theoretical finding that the optimal localisation strength $\gamma$ is determined by the gradient-Lipschitz constant $L$, the largest curvature of the target. For this experiment, optimal ESS is achieved consistently for $\gamma \approx 0.2\,L$. The rightmost panel indicates the mechanism behind this: increasing $\gamma$ shrinks the proposal scale and raises acceptance monotonically. Below the peak, proposals overshoot and are rejected. Above it, proposals are accepted but take steps too small to decorrelate efficiently.

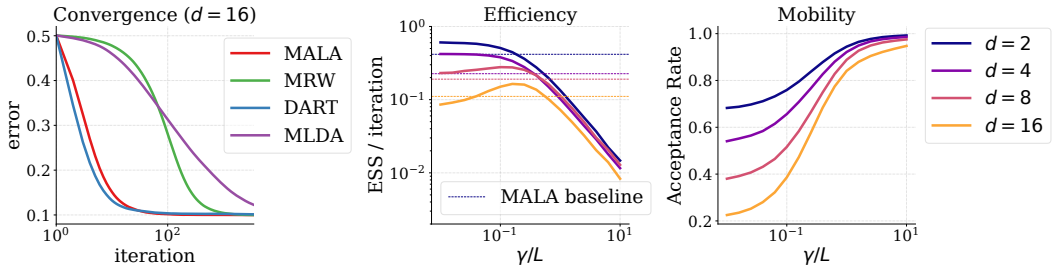


**Figure 2:** Bayesian logistic regression with a Laplace surrogate ($n_{\text{obs}} = 240$, $\alpha = 3$, giving $\kappa = 21$; $\theta = 1/2$). (a) Running mean error $e_k$ at $d = 16$ for MALA, MRW, MLDA, and DART at $\gamma = 0.2\,L$, averaged over $n_{\text{est}} = 500$ independent runs from prior initialisations. (b) ESS per iteration of DART along the slowest target direction, as a function of $\gamma/L$ for dimensions $d \in \{2, 4, 8, 16\}$, estimated from $M = 8$ replicate chains. Horizontal dashed lines show the same quantity for tuned MALA at matching dimensions. (c) Acceptance rate of DART as a function of $\gamma/L$. Target acceptance rates for tuning were 0.55 for MALA and 0.25 for MRW.

## 4.3. Multimodal Targets

Departing from the log-concave setting of Theorem 3.4, we investigate DART's behaviour on multimodal targets. Multimodality is a fundamental obstacle for MCMC methods that rely on local proposals, as the chain can become trapped in a single mode for extended periods. In DART, tempering the surrogate density reduces the effective depth of these barriers, making transitions that the outer chain would reject accessible to the surrogate chain. This is the same mechanism exploited by simulated annealing, but in DART it is built into the proposal step itself, requiring no separate tempered chains or swap moves. When a single level of tempering is insufficient to overcome the barrier, DART admits a recursive extension in the spirit of MLDA. Each level employs a smaller tempering parameter, producing a progressively flatter surrogate landscape. The resulting temperature ladder is traversed sequentially within a single outer proposal step.

We illustrate this mechanism on two bimodal Gaussian mixture targets with mode separation $\Delta$ and per-mode variance $\sigma^2$. To isolate the barrier-crossing effect from surrogate approximation error, we set $g = f$ throughout. For two-level DART, each outer step then requires $n + 1$ target evaluations: one for the acceptance step and $n$ for the surrogate chain. For three-level DART it is correspondingly even higher. The purpose of this experiment is to demonstrate the mixing mechanism, not to benchmark computational cost.

We compare MRW, MALA, two-level DART, and three-level DART. The DART variants use pCN, preconditioned against the quadratic penalty term, as the root chain. The regularisation $\gamma$ is chosen per target so that the localised surrogate density spans roughly the combined support of both modes. Figure 3 shows trace plots alongside the target density (grey) and, for the DART variants, the localised surrogate densities at each tempering level (dashed, coloured), evaluated at $x = 0$. Each trace is annotated with its integrated autocorrelation time estimate and its number of mode transitions.

On the moderate target ($\Delta = 4$, $\sigma^2 = 0.4$) MRW takes steps larger than the mode separation and tunnels across the barrier, transitioning freely. MALA crosses only a handful of times,

since its gradient drift pulls each proposal toward the nearest mode. The drift that speeds within-mode mixing impedes transitions between modes, and its autocorrelation time is large as a result. Both DART variants cross freely. On the harder target ($\Delta = 8$, $\sigma^2 = 0.2$), the modes are too far apart to tunnel at any feasible step size, so MRW fails alongside MALA, each confined to a single mode. Two-level DART still transitions but with visible periods of entrapment, while three-level DART mixes freely, showing that the flatter additional level is what carries the chain across the deeper barrier.

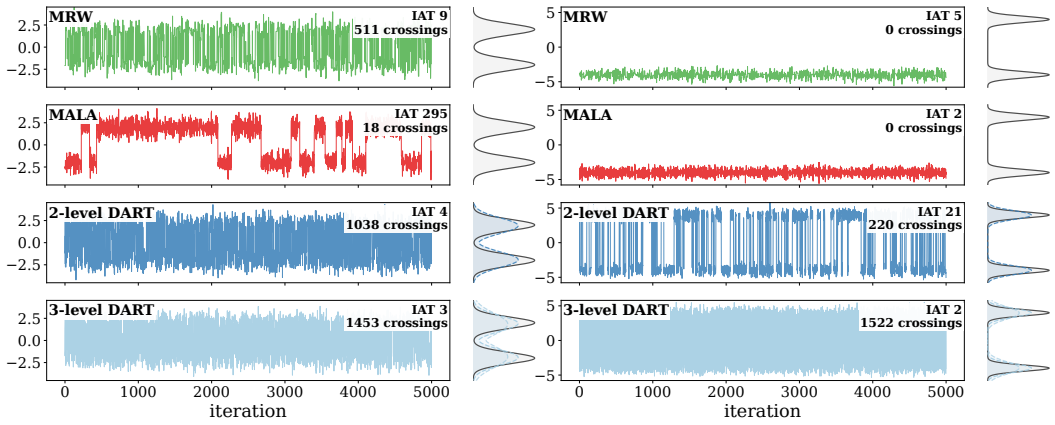


**Figure 3:** Trace plots (first 5 000 post-burn-in iterations) for MRW, MALA, two-level DART, and three-level DART on bimodal Gaussian mixture targets with $(\Delta,\ \sigma^2) = (4,\ 0.4)$ (left) and $(\Delta,\ \sigma^2) = (8,\ 0.2)$ (right). Side panels show the target density (grey) and, for the DART variants, the localised surrogate densities at each tempering level (dashed, coloured), evaluated at $x = 0$. Each trace is annotated with its IAT (estimated from 50 000 post-burn-in iterations) and mode-transition count. Acceptance tuned to MRW 0.3, MALA 0.5, and the pCN root 0.3. Two-level DART: $\theta = 0.5$, $n = 20$. Three-level DART: $\theta_1 = 0.25$, $\theta_2 = 0.5$, $n_1 = 20$, $n_2 = 10$. Regularisation: $\gamma = 0.05$ (moderate), $\gamma = 0.01$ (harder).

## 5. Spatial Generalised Linear Mixed Models

We now present the novel Gaussian process parametrisation based on Dirichlet-Neumann Averaging (DNA) [KS26] and show its synergy with DART on a hierarchical SGLMM problem. For background on SGLMMs and extensions to other observation models, see [DTM98; BC93].

### 5.1. Latent Field Inference

We consider a SGLMM with Poisson response. Let the spatial domain be $\Omega := (0,1)^D \subset \mathbb{R}^D$, with $N$ observation locations $\{s_i\}_{i=1}^N \subset \Omega$. At each location we observe the count $y_i \in \mathbb{N}_0$ of a spatially varying phenomenon. The underlying intensity is modelled by a latent, centred GP $u$ on $\Omega$. Conditional on this *latent field*, observations are assumed independent and

Poisson distributed:

$$y_i \mid u(s_i) \sim \text{Poisson}\big(e^{\eta_i}\big), \qquad \eta_i := \beta_0 + u(s_i), \qquad i = 1, \dots, N, \tag{5.1}$$

where $\beta_0 \in \mathbb{R}$ is a known mean log-intensity.

Covariates are omitted from (5.1) to streamline the exposition. We equip the latent field $u$ with a GP prior whose covariance function is Matérn with marginal variance $\sigma^2$, correlation length $\rho$, and fixed smoothness $\nu = 5/2$, and collect the free covariance parameters in $\psi = (\rho, \sigma^2)$. Together with the Poisson likelihood arising from (5.1), this prior defines a posterior density over $u$. As the Poisson likelihood is not conjugate to the GP prior, closed-form updates are unavailable and MCMC transitions are required. In Section 5.2 we extend this model by placing a hyperprior $p(\psi)$ on $\psi$ and targeting the full joint posterior over $u$ and $\psi$.

MCMC transitions require repeated evaluations of the latent field $u$ at the $N$ observation sites. A naive Cholesky parametrisation of the latent field values assembles and factors the $N \times N$ covariance matrix at cost $O(N^3)$, making each such evaluation prohibitively expensive even at moderate $N$.

### 5.1.1. Latent Field Parametrisation

The DNA framework of [KS26] represents the latent field $u$ as the scaled average of $2^D$ independent GPs, each satisfying a different combination of homogeneous Dirichlet and Neumann conditions on $\partial\Omega$. The boundary artifacts introduced by any single choice of boundary conditions cancel exactly under this averaging, yielding a genuinely isotropic field on $\Omega$. We refer to [KS26] for details in simulation and error analysis, and present mainly the novel elements pertaining to inference.

Boundary conditions are encoded in a vector $b \in \{0,1\}^D$, where $b_j = 0$ prescribes Neumann and $b_j = 1$ Dirichlet conditions on the $j$-th pair of opposing faces. For a spectral resolution $q \in \mathbb{N}$, define the index sets $I_b = I_{b_1} \times \cdots \times I_{b_D}$ with $I_0 = \{0, \dots, q\}$ and $I_1 = \{1, \dots, q\}$. Let $\hat{\varphi}_\psi$ denote the Fourier transform of the Matérn covariance function with parameters $\psi = (\rho, \sigma^2)$ (see [Gra+18, Eq. (2.22)]), and write $\omega_\nu = \omega_\nu(\psi) := \sqrt{\hat{\varphi}_\psi(\nu/2)}$ for the spectral weights at frequency $\nu$. The DNA parametrisation of $u$ is then given by

$$u(s,\, x) = 2^{-D/2} \sum_{b \in \{0,1\}^D} \sum_{\nu \in I_b} x_\nu^b \, \omega_\nu \, e_\nu^b(s), \qquad s \in \Omega, \tag{5.2}$$

where the basis functions $e_\nu^b$ are tensor products of cosine ($b_j = 0$) and sine ($b_j = 1$) modes, and the parameters $x = (x^b)_{b \in \{0,1\}^D} \in \mathbb{R}^d$, with total dimension $d = (2q+1)^D$, have prior $x \sim \mathcal{N}(0, I_d)$ (for details, see [KS26, Section 3]). Since each of the $2^D$ component fields carries its own coefficient vector but is evaluated on the same grid, $d$ exceeds the number of grid points $Q = (q+2)^D$ by a constant factor.

To evaluate the latent field at the $N$ observation sites, we introduce an equispaced grid $\mathcal{G}$ on $\Omega$ with $q+2$ vertices per dimension. On this grid, each inner sum in (5.2) is computed via the appropriate combination of discrete cosine and sine transforms, encoded in a matrix $W_b$.

A sparse bilinear interpolation matrix $\mathrm{I_F} \in \mathbb{R}^{N\times Q}$ maps the grid values to the observation sites (c.f. [Gra+15]), giving

$$\boldsymbol{u}(x) = 2^{-D/2}\ \mathrm{I_F}\left(\sum_{b\in\{0,1\}^D} W_b\,\Lambda_b\,x^b\right)\ \in\ \mathbb{R}^N,$$

where $\Lambda_b := \mathrm{diag}\big(\omega_\nu\big)_{\nu\in I_b}$ and $\boldsymbol{u}_i(x) := u(s_i, x)$. The total cost per evaluation is therefore $\mathcal{O}(Q\log Q + N)$.

### 5.1.2. Surrogate Construction

For fixed hyperparameters $\psi$, the posterior over the latent field in its DNA parametrisation is $\pi(x) \propto e^{-f(x)}$ with potential

$$f(x) = -\mathcal{L}(x) + \frac{1}{2}\|x\|_2^2, \qquad \mathcal{L}(x) := \sum_{i=1}^{N}\big(y_i\,\eta_i(x) - e^{\eta_i(x)}\big), \tag{5.3}$$

where $\mathcal{L}$ is the Poisson log-likelihood and $\eta_i(x) := \beta_0 + \boldsymbol{u}_i(x)$ is the linear predictor at site $s_i$. The gradient of $f$ with respect to each block $x^b$ is available at cost $\mathcal{O}(Q\log Q + N)$ via the adjoint of the interpolation and transform operators.

The structure of (5.2) provides a natural partition of the parameters. Fix a coarse resolution $q_\mathrm{C} < q$ and define coarse index sets $I_{b,\mathrm{C}}$ by replacing $q$ with $q_\mathrm{C}$. Each block decomposes as $x^b = (x^b_\mathrm{C},\ x^b_\mathrm{F})$, where $x^b_\mathrm{C}$ collects modes with indices in $I_{b,\mathrm{C}}$ and $x^b_\mathrm{F}$ collects the remainder. A coarse approximation to the log-likelihood is obtained by replacing the field evaluation in (5.3) with its coarse-grid analogue, using $W_{b,\mathrm{C}}$ and $\mathrm{I_C}$ in place of $W_b$ and $\mathrm{I_F}$, and acting only on $x_\mathrm{C}$. We denote this coarse log-likelihood by $\mathcal{L}_\mathrm{C}(z_\mathrm{C})$. Provided $q_\mathrm{C}$ is chosen large enough that the high-frequency modes carry negligible likelihood information, a natural choice is to let these modes contribute only via the prior. When at state $x = (x_\mathrm{C}, x_\mathrm{F})$, the localised surrogate potential used for DART is

$$V_x(z) = -\theta\,\mathcal{L}_\mathrm{C}(z_\mathrm{C})\ +\ \frac{\gamma}{2}\|z_\mathrm{C} - x_\mathrm{C}\|_2^2\ +\ \frac{1}{2}\|z\|_2^2, \tag{5.4}$$

where $\theta \in (0,1)$ tempers the coarse likelihood and the quadratic penalty localises only the coarse component. Since the parameters are independent under the prior, the prior density factorises across the partition, giving the $\|z_\mathrm{C}\|_2^2 + \|z_\mathrm{F}\|_2^2 = \|z\|_2^2$ decomposition in (5.4). A proposal $z = (z_\mathrm{C}, z_\mathrm{F})$ is drawn by running a root chain targeting the coarse component of $\Pi_x$, giving $z_\mathrm{C}$, and drawing $z_\mathrm{F}$ from a pCN step targeting $\mathcal{N}(0, I_{d_\mathrm{F}})$.

The acceptance probability (3.2) for this construction simplifies considerably. We track three cancellations in turn.

- *Quadratic penalty.* As in Section 3, the symmetry $V_x(z) - V_z(x) = -\theta(\mathcal{L}_\mathrm{C}(z_\mathrm{C}) - \mathcal{L}_\mathrm{C}(x_\mathrm{C})) + \frac{1}{2}(\|z\|_2^2 - \|x\|_2^2)$ eliminates the quadratic localisation term from the proposal density ratio; the prior difference that remains is cancelled below.

- *High-frequency block.* Since $x_{\mathrm{F}}$ is not localised in $V_x$, the dependence of the normalisation constant on the state enters only through the low-frequency component. We leverage $N_x = N_{x,\mathrm{C}} \cdot C_{\mathrm{F}}$, where $C_{\mathrm{F}}$ is the Gaussian integral over $z_{\mathrm{F}}$ and is independent of $x$. The ratio reduces to $N_x/N_z = N_{x,\mathrm{C}}/N_{z,\mathrm{C}}$, which is a problem in significantly lower dimension than then initial one.
- *Low-frequency prior.* Both $f$ and $V_x$ carry the prior term $\frac{1}{2}\|\cdot\|_2^2$. The contribution to $\pi(z)/\pi(x)$ is $\exp(-\frac{1}{2}(\|z\|_2^2 - \|x\|_2^2))$. The contribution to $p_z(x)/p_x(z)$ is its inverse, and the two cancel. This cancellation is structurally the same as in pCN [Cot+13], with the localisation parameter $\gamma$ playing no role in the preconditioned step.

Incorporating these three cancellations, the acceptance probability reduces to

$$\alpha(x, z) = \min\left\{1,\ \exp\Big(\mathcal{L}(z) - \mathcal{L}(x) + \theta\big(\mathcal{L}_{\mathrm{C}}(x_{\mathrm{C}}) - \mathcal{L}_{\mathrm{C}}(z_{\mathrm{C}})\big)\Big)\frac{N_{x,\mathrm{C}}}{N_{z,\mathrm{C}}}\right\}.$$

### 5.1.3. Experiment

The Hyperparameters are fixed at $\psi = (\rho, \sigma^2) = (0.1, 1.25)$ and we consider $D = 1$ spatial dimension. We begin in a regime where the coarse surrogate captures most likelihood information at the chosen correlation length. We compare DART against three alternatives: MALA on the posterior under a Cholesky parametrisation, MALA on the DNA parametrisation (5.3) of the full posterior, and MLDA on the DNA parametrisation using the same surrogate as DART but without regularisation or tempering. We deliberately do not include the INLA [RMC09] and stan [Car+17] HMC software routines in the comparison, as these occupy different points on the trade-off between cost, accuracy, and scalability.

Figure 4 presents three diagnostic views. The worst- and average-case IAT bars reveal a monotone hierarchy, from Cholesky, to DNA with MALA, to DNA with MLDA, to DNA with DART. The DNA parametrisation alone reduces the worst-case IAT over Cholesky; the surrogate hierarchy in MLDA buys a further factor over MALA on DNA, and regularisation and tempering bring DART to the lowest IAT. The best-case bars do not follow this order, nor is this expected: the best-mixing location is governed by the high-frequency blocks $x_{\mathrm{F}}$, which carry negligible likelihood information and so mix at essentially the prior rate for every DNA variant. The best-case summary therefore reflects prior mixing common to all three DNA methods and cannot separate them, but gives a clear advantage over the Cholesky parametrisation. The posterior mean and 95% credible band from DART (right panel) track the ground truth closely.

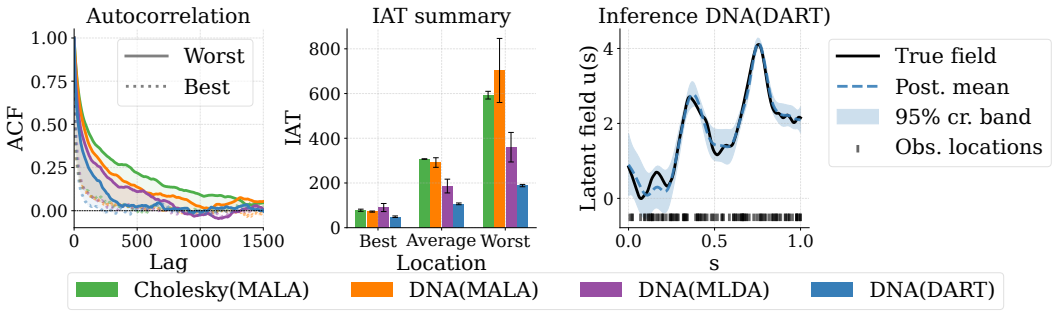


**Figure 4:** Good-surrogate regime. Ground truth is generated once at high resolution ($q_{\text{gt}} = 500$) and held fixed. $N = 100$ observation sites are drawn uniformly on $(0, 1)$ and Poisson counts are simulated from (5.1) with intercept $\beta_0 = 1.5$. DNA methods use fine resolution $q = 100$ and coarse resolution $q_C = 20$. DART and MLDA run $n = 25$ root chain steps per proposal, with DART using $\gamma = 0.25$ and $\theta = 0.75$. Each method is run for 150 000 iterations over 3 independent replications. The respective first 50 000 iterations are discarded as burn-in. Left: ACF at worst-case (solid) and best-case (dashed) spatial locations. Centre: IAT summary (best, average, worst); error bars show standard errors. Right: posterior mean and 95% credible band from DNA with DART, with ground truth (black) and observation locations (ticks).

## 5.2. Hierarchical Inference

We extend the SGLMM from Section 5.1 to joint inference over the latent field and the covariance hyperparameters $\psi = (\rho, \sigma)$, now in $D = 2$ spatial dimensions. We place a hyperprior on $\psi$ and use a Metropolis-within-Gibbs sampler alternating between the latent and hyperparameter blocks.

Placing a hyperprior $p(\psi)$ on the covariance parameters, the full joint posterior has density $\pi(x, \psi) \propto e^{-f(x,\psi)}$ with potential

$$f(x, \psi) = -\mathcal{L}(x, \psi) + \tfrac{1}{2}\|x\|_2^2 - \log p(\psi),$$

extending (5.3) to include the hyperprior. Here $\mathcal{L}(x, \psi)$ is the Poisson log-likelihood from (5.3), now written with its dependence on $\psi$ through the spectral weights $\omega_\nu(\psi)$ in (5.2) made explicit. The full conditionals for the two Gibbs blocks have potentials

$$f_0(x) := -\mathcal{L}(x, \psi) + \tfrac{1}{2}\|x\|_2^2, \quad \text{and} \quad f_1(\psi) := -\mathcal{L}(x, \psi) - \log p(\psi),$$

where $f_0$ coincides with (5.3) for fixed $\psi$. The coupling between the two Gibbs blocks is implicit, mediated through the spectral weights $\omega_\nu(\psi)$.

Under DNA, the per-sweep covariance update reduces to an $\mathcal{O}(d)$ spectral-weight recomputation. A Cholesky parametrisation would instead require a fresh $\mathcal{O}(N^3)$ factorisation each sweep, dominating the per-sweep cost at scale. The latent block targets the same posterior as Section 5.1.2, and the DART transition applies as in that section, with the surrogate re-evaluated at the current $\psi$. Since $f_1$ is only two-dimensional, we update $\psi$ on the log-scale via a random walk MH step with Robbins–Monro adaptation [RM51]. For the hyperprior we use a joint penalised complexity prior [Sim+17] on $(\rho, \sigma)$, penalising depar-

ture from a base model of zero variance and infinite correlation length at rates controlled by $\Pr(\rho < \rho_0) = \alpha_\rho$ and $\Pr(\sigma > \sigma_0) = \alpha_\sigma$.

We consider the SGLMM from (5.1) on the unit square $\Omega = (0, 1)^2$ with Matérn covariance and a constant trend $\beta_0 = 0.5$. We generate a smooth ($\nu = 5/2$, $\rho = 0.1$) and a rough ($\nu = 3/2$, $\rho = 0.05$) field as ground truth at high resolution ($q_{\text{gt}} = 500$). The aim of the experiment in this section is to demonstrate robust inference of these fields and the corresponding hyperparameters from the corresponding simulated Poisson count data. Observation locations are drawn from a Sobol sequence on $\Omega$, providing quasi-uniform coverage.

Figure 5 shows, for each smoothness, the ground-truth field, the DART posterior mean and the posterior standard deviation. The posterior mean recovers the truth in both cases. The posterior mean is slightly more regular than the truth. We attribute this to the penalised-complexity prior reverting towards the less complex, smoother model where the likelihood is uninformative. To guard against a single fortunate chain we run each demonstration at three independent seeds and pool the draws, and the recovered field and its uncertainty are stable across the three.

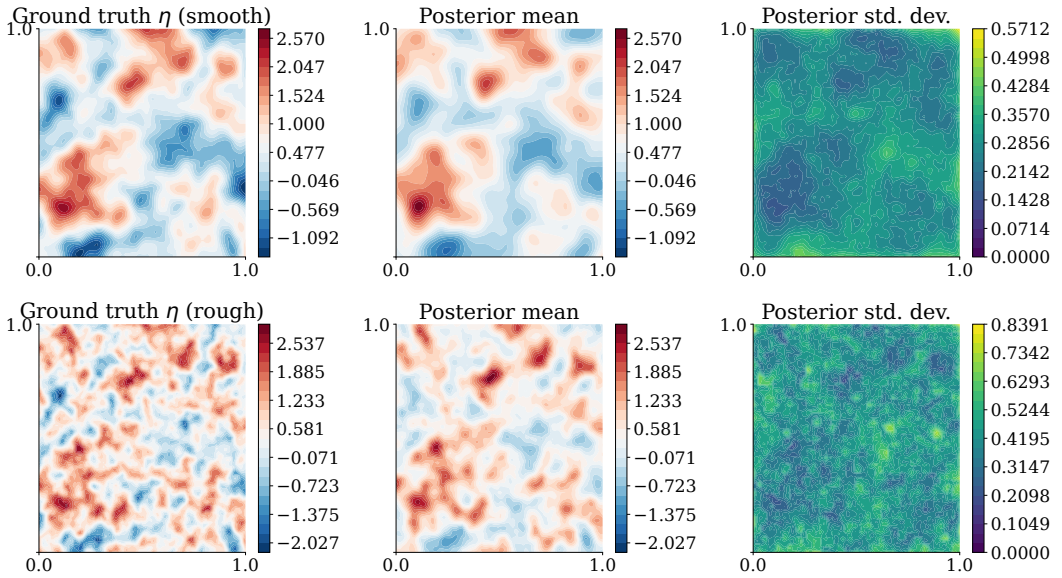


**Figure 5:** Hierarchical SGLMM, DART posterior fidelity at $N = 1024$. Rows: smooth field ($\nu = 5/2$, $\rho = 0.1$, $\sigma^2 = 3/4$) top, and rough field ($\nu = 3/2$, $\rho = 0.05$, $\sigma^2 = 3/4$) bottom. Columns: ground-truth field, DART posterior mean and posterior standard deviation. MCMC is run for 230 000 iterations, of which 30 000 are discarded as burn-in. Fine resolutions are $q = 64$ for both fields (16 641 parameters total) and the coarse resolutions are $q_{\text{C}} = 12$ (smooth, 625 parameters) and $q_{\text{C}} = 28$ (rough, 3 249 parameters). DART parameters: $\gamma = 5 \cdot 10^{-3}$, $\theta = 0.9$, $n = 6$ for both.

# 6. Proofs

The remainder of this manuscript assembles the proof of Theorem 3.4. The argument has two stages. In the first, we analyse an *idealised* chain $\hat{T}$ whose proposals are drawn exactly from the localised surrogate measures $\Pi_x$. This chain is $\Pi$-reversible, so its mixing time can be bounded by the $s$-conductance machinery. In the second stage, we transfer the bound to the *implemented* chain $T$, whose proposals come from a finite-length root chain. Section 6.3 combines all of these ingredients to conclude. Standard, cited, and technical auxiliary results are collected in Appendix A.

The first stage rests on the following result, due to [Dwi+19], which we state in notation adapted to our setting.

**Proposition 6.1.** *Let $t > 0$, $\varepsilon \in (0,1)$, and let $\nu$ be a probability measure on $\mathbb{R}^d$ with density $\varphi \propto e^{-g}$, where $g : \mathbb{R}^d \to \mathbb{R}$ is $\lambda$-strongly convex. Further, suppose $\{\mathcal{T}_x\}_{x\in\mathbb{R}^d}$ are the transition measures of a Markov chain with invariant measure $\nu$. If $\mathcal{K} \subset \mathbb{R}^d$ is a convex set such that for any $x, y \in \mathcal{K}$*

$$\|\mathcal{T}_x - \mathcal{T}_y\|_{\mathrm{TV}} \le \varepsilon \quad \text{whenever} \quad \|x - y\|_2 \le t,$$

*then for any measurable partition of $\mathbb{R}^d$ into $A_1$ and $A_2$,*

$$\int_{A_1} \mathcal{T}_z(A_2)\,\varphi(z)\,\mathrm{d}z \ge \frac{1-\varepsilon}{4} \min\left\{1,\ \frac{\log 2}{8}\,\nu^2(\mathcal{K})\,t\sqrt{\lambda}\right\} \min\Big\{\nu(A_1\cap\mathcal{K}),\ \nu(A_2\cap\mathcal{K})\Big\}. \quad (6.1)$$

If, in addition, $\nu$ assigns a significant fraction of its mass to $\mathcal{K}$, Proposition 6.1 yields an explicit conductance lower bound, which via (2.6) gives a mixing-time upper bound under a warm-start assumption.

**Lemma 6.2.** *Let $\nu \in \mathcal{P}(\mathbb{R}^d)$ have density $\varphi \propto e^{-g}$ with $g : \mathbb{R}^d \to \mathbb{R}$ $\lambda$-strongly convex, and let $\{\mathcal{T}_x\}_{x\in\mathbb{R}^d}$ be the transition measures of a $\nu$-reversible, lazy Markov chain with invariant measure $\nu$. Suppose there exist $t > 0$ and a convex set $\mathcal{K} \subset \mathbb{R}^d$ with*

$$\|\mathcal{T}_x - \mathcal{T}_y\|_{\mathrm{TV}} \le \tfrac{3}{4} \quad \text{whenever} \quad \|x - y\|_2 \le t, \qquad x, y \in \mathcal{K}. \quad (6.2)$$

*Let $\delta \in (0,1)$ and let $\mu$ be $\beta$-warm with respect to $\nu$, with $\beta \ge 1$. If*

$$\nu(\mathcal{K}) \ge 1 - \frac{\delta}{2\beta}, \quad (6.3)$$

*then*

$$\|T^n\mu - \nu\|_{\mathrm{TV}} \le \delta \qquad \text{for all} \qquad n \gtrsim \frac{1}{t^2\lambda} \log\frac{2\beta}{\delta}.$$

**Proof.** Set $s := \frac{\delta}{2\beta}$. Since $\beta \ge 1$ and $\delta < 1$, we have $s \in (0, \frac{1}{2})$, and the error budget in (2.6) is split evenly. With $\varepsilon = 3/4$, and taking the minimum in (6.1) to be attained by its second argument (the complementary case only strengthens the bound), (6.1) reduces to

$$\int_{A_1} \mathcal{T}_z(A_2)\,\varphi(z)\ \mathrm{d}z \ge \frac{\log 2}{128}\,t\sqrt{\lambda}\,\nu^2(\mathcal{K})\ \min\Big\{\nu(A_1\cap\mathcal{K}),\ \nu(A_2\cap\mathcal{K})\Big\}.$$

Now let $A$ be a measurable set with $s < \nu(A) \le \frac{1}{2}$, and set $A_1 = A$, $A_2 = A^{\mathrm{c}}$. By (6.3),

$$\nu(A \cap \mathcal{K}) \ge \nu(A) - \nu(\mathcal{K}^{\mathrm{c}}) \ge \nu(A) - s,$$

and analogously $\nu(A^{\mathrm{c}} \cap \mathcal{K}) \ge \nu(A^{\mathrm{c}}) - s$. Since $\nu(A) > s$ and $\nu(A^{\mathrm{c}}) \ge \nu(A) > s$, both lower bounds are positive, so

$$\min\Big\{ \nu(A \cap \mathcal{K}),\ \nu(A^{\mathrm{c}} \cap \mathcal{K}) \Big\} = \nu(A) - s.$$

Using $\log 2 \ge 1/2$, $\nu(\mathcal{K}) \ge 1 - s$ and $(1-s)^2 \ge \frac{1}{4}$, we arrive at the conductance lower bound

$$\Phi_s(T) \ := \ \inf_{\nu(A) \in (s,\, 1/2)} \frac{\int_A \mathcal{T}_x(A^{\mathrm{c}})\, \varphi(x)\ \mathrm{d}x}{\nu(A) - s} \ \ge \ \frac{t\sqrt{\lambda}}{1024}.$$

Inserting this into (2.6) and applying Bernoulli's inequality yields

$$\| T^n \mu - \nu \|_{\mathrm{TV}} \le \frac{\delta}{2} + \beta \big(1 - c\, t^2 \lambda\big)^n \le \frac{\delta}{2} + \beta\, e^{-c\, n t^2 \lambda},$$

where $c = 1/(2 \cdot 1024^2)$ is a universal constant. For $n \ge c^{-1}(t^2\lambda)^{-1} \log(2\beta/\delta)$ the second term is at most $\delta/2$, giving $\| T^n \mu - \nu \|_{\mathrm{TV}} \le \delta$ for all such $n$. The factor $c^{-1}$ is absorbed into the implicit constant of the statement.

□

The remaining ingredient is a convex set $\mathcal{K}$ satisfying (6.3). For strongly log-concave targets, this is supplied by concentration of $\nu$ around the minimiser of its potential [DM19; Dwi+19]. The radius of concentration is governed by

$$r : (0,\infty) \times \mathbb{N} \to \mathbb{R}, \qquad r(\varepsilon, d) \ := \left( 1 + 2\sqrt{\tfrac{\log \varepsilon^{-1}}{d}} + \tfrac{2 \log \varepsilon^{-1}}{d} \right)^{\frac{1}{2}},$$

and the concentration of measure is characterised as follows.

**Proposition 6.3.** *Let $\nu$ be a probability measure on $\mathbb{R}^d$ with density $\varphi \propto e^{-g}$, where $g : \mathbb{R}^d \to \mathbb{R}$ is $\lambda$-strongly convex, and set $\hat{x} := \arg\min_{\mathbb{R}^d} g$. Then for any $\varepsilon \in (0,1)$,*

$$\nu(\mathcal{K}(\varepsilon)) \ge 1 - \varepsilon, \qquad \textit{where} \qquad \mathcal{K}(\varepsilon) := \mathbb{B}\Big( \hat{x},\ r(\varepsilon, d) \sqrt{d/\lambda} \Big).$$

## 6.1. Properties of Localised Surrogate Measures

We first establish Lipschitz continuity of the map $x \mapsto \Pi_x$ in total variation, a key ingredient required by Proposition 6.1.

**Lemma 6.4.** *Let $\{\pi_x\}_{x \in \mathbb{R}^d}$ be a collection of strongly log-concave surrogate densities (Definition 3.2). Then*

$$\big\| \Pi_x - \Pi_y \big\|_{\mathrm{TV}} \le \frac{\gamma}{2\sqrt{\gamma + \theta m}} \, \|x - y\|_2.$$

**Proof.** Pinsker's inequality gives

$$\left\|\Pi_x - \Pi_y\right\|_{\mathrm{TV}}^2 \le \tfrac{1}{2}\mathrm{KL}\big(\Pi_x \,\|\, \Pi_y\big), \tag{6.4}$$

and the identity

$$\mathrm{KL}\big(\Pi_x \,\|\, \Pi_y\big) = \log \frac{N_y}{N_x} + \mathbb{E}_{z\sim\Pi_x}\left(V_y(z) - V_x(z)\right)$$

splits the divergence into a normalisation ratio and a potential discrepancy. As per (3.9), the ratio of normalisation constants can be rewritten as

$$\frac{N_y}{N_x} = \mathbb{E}_{z\sim\Pi_x}\, e^{-(V_y(z)-V_x(z))}. \tag{6.5}$$

Plugging (6.5) into (6.4),

$$\mathrm{KL}\big(\Pi_x \,\|\, \Pi_y\big) = \log \mathbb{E}_{z\sim\Pi_x}\, e^{-(V_y(z)-V_x(z))} + \mathbb{E}_{z\sim\Pi_x}\left(V_y(z) - V_x(z)\right). \tag{6.6}$$

By definition of the surrogate potentials (3.4),

$$V_y(z) - V_x(z) = \frac{\gamma}{2}\big(\|z-y\|_2^2 - \|z-x\|_2^2\big) = \gamma\,\langle x-y, z\rangle - c_{x,y}, \tag{6.7}$$

where $c_{x,y} := \frac{\gamma}{2}\big(\|x\|_2^2 - \|y\|_2^2\big)$. Inserting (6.7) into (6.6), the constants cancel, leaving

$$\mathrm{KL}\big(\Pi_x \,\|\, \Pi_y\big) = \log \mathbb{E}_{z\sim\Pi_x}\, e^{-\gamma\langle x-y,\, z\rangle} + \gamma\,\langle x-y,\ \bar{u}_x\rangle.$$

Writing $z = \bar{u}_x + (z-\bar{u}_x)$ inside the logarithm, the deterministic contribution $-\gamma\,\langle x-y,\ \bar{u}_x\rangle$ exits the expectation and cancels with the second term, giving

$$\mathrm{KL}\big(\Pi_x \,\|\, \Pi_y\big) = \log \mathbb{E}_{z\sim\Pi_x}\, e^{-\gamma\langle x-y,\, z-\bar{z}_x\rangle}. \tag{6.8}$$

Since $g$ is $m$-strongly convex, we may write $g = \frac{m}{2}\|\cdot\|_2^2 + \varphi$ with $\varphi$ convex. Expanding the surrogate potential and completing the square in $z$ shows that

$$\pi_x(z) \propto e^{-\theta\varphi(z)}\, e^{-\frac{\gamma+\theta m}{2}\|z-c_x\|_2^2},$$

where $c_x := \gamma x/(\gamma+\theta m)$. The second factor is proportional to the density of $\Gamma_x := \mathcal{N}\big(c_x,\, (\theta m+\gamma)^{-1}\mathbb{I}\big)$, and the first is log-concave. Hargé's inequality [Har04, Theorem 1.1] therefore gives, for any convex $\psi$,

$$\mathbb{E}_{z\sim\Pi_x}\,\psi(z-\bar{z}_x) \le \mathbb{E}_{w\sim\Gamma_x}\,\psi(w-c_x). \tag{6.9}$$

Applying (6.9) with the convex function $\psi:\ u\mapsto e^{-\gamma\langle x-y,\, u\rangle}$ and evaluating the resulting Gaussian moment generating function yields

$$\mathbb{E}_{z\sim\Pi_x}\, e^{-\gamma\langle x-y,\, z-\bar{z}_x\rangle} \le \mathbb{E}_{w\sim\mathcal{N}(0,\,(\theta m+\gamma)^{-1}\mathbb{I})}\, e^{-\gamma\langle x-y,\, w\rangle} = e^{\frac{\gamma^2}{2(\theta m+\gamma)}\|x-y\|_2^2}. \tag{6.10}$$

Inserting (6.10) into (6.8) and the result into (6.4), and taking the square root yields the claim.

□

Secondly, we show that the mode displacement of the localised surrogate measure is entirely controlled by the regularisation and tempering parameters. Denote the modes $a_x := \arg\min_{\mathbb{R}^d} V_x$.

**Lemma 6.5.** *Let $\{\pi_x\}_{x\in\mathbb{R}^d}$ be a collection of strongly log-concave surrogate densities (Definition 3.2). Then*

$$\|x - a_x\|_2 \le \frac{\theta M}{\gamma + \theta m} \|x - \hat{x}\|_2.$$

**Proof.** First-order optimality for $a_x$ gives $\nabla V_x(a_x) = \theta\nabla g(a_x) + \gamma(a_x - x) = 0$, which yields

$$\gamma(x - a_x) = \theta\,\nabla g(a_x). \tag{6.11}$$

Since $\nabla g(\hat{x}) = 0$ and $g$ is $M$-smooth,

$$\|\nabla g(a_x)\|_2 = \|\nabla g(a_x) - \nabla g(\hat{x})\|_2 \le M\|a_x - \hat{x}\|_2. \tag{6.12}$$

Combining (6.11) and (6.12),

$$\|x - a_x\|_2 \le \frac{\theta M}{\gamma} \|a_x - \hat{x}\|_2. \tag{6.13}$$

On the other hand, $\nabla V_x(\hat{x}) = \theta\nabla g(\hat{x}) + \gamma(\hat{x} - x) = \gamma(\hat{x} - x)$. Since $V_x$ is $(\gamma + \theta m)$-strongly convex with minimiser $a_x$,

$$\|a_x - \hat{x}\|_2 \le \frac{1}{\gamma + \theta m} \|\nabla V_x(\hat{x})\|_2 = \frac{\gamma}{\gamma + \theta m} \|x - \hat{x}\|_2. \tag{6.14}$$

Inserting (6.14) into (6.13) yields the claim.

□

Finally, we record a warmness estimate for a Gaussian measure centred at $x \in \mathbb{R}^d$, with respect to the corresponding surrogate measure $\Pi_x$. The mode displacement from Lemma 6.5 governs the resulting warmness constant.

**Lemma 6.6.** *Let $\{\pi_x\}_{x\in\mathbb{R}^d}$ be a collection of strongly log-concave surrogate densities (Definition 3.2), and write $\kappa_V := (\gamma + \theta M)/(\gamma + \theta m)$. Then for all $x \in \mathbb{R}^d$, the Gaussian measure $\mathcal{N}\big(x, (2(\gamma + \theta M))^{-1}\mathbb{I}\big)$ is $\beta$-warm with respect to $\Pi_x$, where*

$$\log\beta \le \frac{d}{2}\,\log(2\kappa_V) + \frac{\kappa_V\,\theta^2 M^2}{\gamma}\,\|x - \hat{x}\|_2^2. \tag{6.15}$$

**Proof.** The surrogate potential $V_x$ is $(\gamma + \theta m)$-strongly convex and $(\gamma + \theta M)$-smooth with minimiser $a_x$. In [Dwi+19, Section 3.2.1], the authors derive bounds on the density ratio for a Gaussian whose mean is displaced by $\varepsilon > 0$ from the target mode, where the target density has a $m_\star$-strongly convex and $L_\star$-smooth potential. Applying their result with $\varepsilon = \|x - a_x\|_2$, $m_\star = \gamma + \theta m$ and $L_\star = \gamma + \theta M$ yields

$$\log \beta \le \frac{d}{2} \log \frac{2(\gamma + \theta M)}{\gamma + \theta m} + (\gamma + \theta M) \|x - a_x\|_2^2.$$

The first term is already $\frac{d}{2} \log(2\kappa_V)$ by construction. For the second term, Lemma 6.5 bounds the mode displacement by $\|x - a_x\|_2 \le \frac{\theta M}{\gamma + \theta m} \|x - \hat{x}\|_2$, so that

$$(\gamma + \theta M) \|x - a_x\|_2^2 \le \frac{(\gamma + \theta M)\, \theta^2 M^2}{(\gamma + \theta m)^2} \|x - \hat{x}\|_2^2 = \frac{\kappa_V\, \theta^2 M^2}{\gamma + \theta m} \|x - \hat{x}\|_2^2 \le \frac{\kappa_V\, \theta^2 M^2}{\gamma} \|x - \hat{x}\|_2^2,$$

where the last step uses $\gamma + \theta m \ge \gamma$. This yields (6.15) and completes the proof.

□

## 6.2. Bounding the Acceptance Probability

The acceptance probability of DART is governed by the target potential difference $f(z) - f(x)$, the surrogate correction $\theta(g(x) - g(z))$, and the normalisation constant ratio $\log(N_z/N_x)$. We begin with a bound on the normalisation constant ratio. Bounds on the potential differences are folded into the proof of Lemma 6.8.

**Lemma 6.7.** *Let $\gamma > 0$ and let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be convex and continuously differentiable. For each $x \in \mathbb{R}^d$, define the potential*

$$V_x : \mathbb{R}^d \to \mathbb{R}, \quad y \mapsto \varphi(y) + \frac{\gamma}{2} \|y - x\|_2^2.$$

*Let $\Pi_x$ be the probability measure on $(\mathbb{R}^d, \mathfrak{B}(\mathbb{R}^d))$ with density $\pi_x \propto e^{-V_x}$ and normalisation constant $N_x$. Then, for all $x, y \in \mathbb{R}^d$,*

$$\log \frac{N_y}{N_x} \le \left\langle \mathbb{E}_{u \sim \Pi_x} \nabla \varphi(u),\ x - y \right\rangle.$$

**Proof.** As in (3.9), we express the ratio of normalisation constants as an expectation over $\Pi_x$:

$$\log \frac{N_y}{N_x} = \log \mathbb{E}_{u \sim \Pi_x}\, e^{V_x(u) - V_y(u)}.$$

Writing $u - y = (u - x) + (x - y)$ and expanding, the difference in potentials is

$$V_x(u) - V_y(u) = \frac{\gamma}{2}\left(\|u - x\|_2^2 - \|u - y\|_2^2\right) = -\gamma \langle x - y,\ u \rangle + \gamma \langle x - y,\ x \rangle - \frac{\gamma}{2} \|x - y\|_2^2.$$

The constant terms coincide with the negative logarithm of the Laplace transform of $\mathcal{N}(x, \gamma^{-1}\mathbb{I})$ evaluated at $\gamma(x-y)$, which allows us to write

$$\log \frac{N_y}{N_x} = \log \mathbb{E}_{u\sim\Pi_x}\, e^{-\gamma\langle x-y,u\rangle} - \log \mathbb{E}_{v\sim\mathcal{N}(x,\gamma^{-1}\mathbb{I})}\, e^{-\gamma\langle x-y,v\rangle}. \tag{6.16}$$

We now apply Hargé's inequality to bound the first term. Since $\varphi$ is convex, $\pi_x$ is a log-concave perturbation of the Gaussian $\mathcal{N}(x, \gamma^{-1}\mathbb{I})$, so [Har04, Theorem 1.1] gives

$$\mathbb{E}_{u\sim\Pi_x}\, e^{-\gamma\langle x-y,\, u-\bar{u}_x\rangle} \le \mathbb{E}_{v\sim\mathcal{N}(x,\gamma^{-1}\mathbb{I})}\, e^{-\gamma\langle x-y,\, v-x\rangle}. \tag{6.17}$$

Writing $u = \bar{u}_x + (u-\bar{u}_x)$ inside the first term of (6.16) and applying (6.17),

$$\begin{aligned}
\log \mathbb{E}_{u\sim\Pi_x}\, e^{-\gamma\langle x-y,u\rangle} &= -\gamma\langle x-y,\ \bar{u}_x\rangle + \log \mathbb{E}_{u\sim\Pi_x}\, e^{-\gamma\langle x-y,\, u-\bar{u}_x\rangle} \\
&\le -\gamma\langle x-y,\ \bar{u}_x\rangle + \log \mathbb{E}_{v\sim\mathcal{N}(x,\gamma^{-1}\mathbb{I})}\, e^{-\gamma\langle x-y,\, v-x\rangle} \\
&= \gamma\langle x-y,\ x-\bar{u}_x\rangle + \log \mathbb{E}_{v\sim\mathcal{N}(x,\gamma^{-1}\mathbb{I})}\, e^{-\gamma\langle x-y,\, v\rangle}.
\end{aligned} \tag{6.18}$$

The Laplace transform of the Gaussian density in (6.18) cancels with the second term of (6.16), leaving

$$\log \frac{N_y}{N_x} \le \gamma\langle x-\bar{u}_x,\ x-y\rangle. \tag{6.19}$$

The proof concludes by substituting the Stein identity (3.1) into (6.19).

□

With the normalisation constant ratio under control, we can bound the expected acceptance probability.

**Lemma 6.8.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be $\lambda$-strongly convex and $L$-smooth with minimiser $x^\star := \arg\min_{\mathbb{R}^d} f$, and define $\kappa := L/\lambda$. Consider an $m$-strongly convex and $M$-smooth surrogate $g : \mathbb{R}^d \to \mathbb{R}$ satisfying $\lambda \le m \le M \le L$, and let $\hat{x} := \arg\min_{\mathbb{R}^d} g$. Fix $\tau \in (0,1)$ and parameters $\theta \in (0,1)$, $\gamma > 0$, and let $\pi_x$ be the strongly log-concave surrogate density (Definition 3.2) using $g$ as the surrogate for $f$, with normalisation constant $N_x$. Define $\mathcal{K}_\tau := \mathbb{B}\big(x^\star,\ r(\tau,d)\sqrt{d/\lambda}\big)$ and assume that $g$ approximates $f$ in the sense that*

*(a1)* $\hat{x} \in \mathcal{K}_\tau$, *and*

*(a2)* $\sup_{x\in\mathcal{K}_\tau} \|\nabla g(x) - \nabla f(x)\|_2 \le \eta$ *for some* $\eta > 0$.

*Define the acceptance probability $\alpha(x,y) := \min\big\{1,\ e^{-A(x,y)}\big\}$, where*

$$A(x,y) := f(y) - f(x) + \theta\big(g(x) - g(y)\big) + \log \frac{N_y}{N_x}.$$

*Then, for any $x \in \mathcal{K}_\tau$ and $\varepsilon \in (0,1)$, we have $\mathbb{E}_{z\sim\Pi_x}\,\alpha(x,z) \geq 1-\varepsilon$, provided that the parameter choice satisfies*

$$\theta = \frac{1}{2} \quad \textit{and} \quad \gamma \geq c(\tau,d)\left(L\kappa + \frac{Ld}{\varepsilon} + \frac{\eta^2 d}{\varepsilon^2}\right), \tag{6.20}$$

*where $c(\tau,d) = 4(1+r(\tau,d))^2$.*

**Proof.** Fix $x \in \mathcal{K}_\tau$ and write $\sigma := (\gamma + \lambda/2)^{-1/2}$ and $A_+(x,z) := \max\{0,\, A(x,z)\}$. Since $\min\{1,\ e^{-t}\} \geq 1 - \max\{0,t\}$ for all $t \in \mathbb{R}$, it suffices to show $\mathbb{E}_{z\sim\Pi_x} A_+(x,z) \leq \varepsilon$. Setting $\theta = 1/2$, $L$-smoothness of $f$ gives $f(z) - f(x) \leq \langle -\nabla f(x),\, x - z\rangle + \frac{L}{2}\|x-z\|_2^2$, convexity of $g$ gives $\frac{1}{2}(g(x) - g(z)) \leq \frac{1}{2}\langle \nabla g(x),\, x - z\rangle$, and Lemma 6.7 applied with $\varphi = \frac{1}{2}g$ gives $\log(N_z/N_x) \leq \frac{1}{2}\langle \mathbb{E}_{u\sim\Pi_x}\nabla g(u),\ x - z\rangle$. Combining these three bounds yields

$$A(x,z) \leq \langle w,\, x - z\rangle + \frac{L}{2}\|x-z\|_2^2,$$

where, after regrouping the gradient terms,

$$w := \nabla g(x) - \nabla f(x) + \frac{1}{2}\,\mathbb{E}_{u\sim\Pi_x}\big(\nabla g(u) - \nabla g(x)\big).$$

By (a2), $M$-smoothness of $g$, and Jensen's inequality,

$$\|w\|_2 \leq \eta + \tfrac{M}{2}\sqrt{S}, \tag{6.21}$$

where $S := \mathbb{E}_{z\sim\Pi_x}\|z-x\|_2^2$. Cauchy–Schwarz followed by Jensen's inequality gives

$$\mathbb{E}_{z\sim\Pi_x} A_+(x,z) \leq \|w\|_2\sqrt{S} + \tfrac{L}{2}S \leq \eta\sqrt{S} + LS, \tag{6.22}$$

where the last step substitutes (6.21) and uses $M \leq L$. It remains to bound $S$ in terms of the algorithm parameters. The triangle inequality gives $\sqrt{S} \leq \|x - a_x\|_2 + \sqrt{\mathbb{E}\|z-a_x\|_2^2}$. Since both $x$ and $\hat{x}$ lie in $\mathcal{K}_\tau$ by (a1), the triangle inequality yields $\|x-\hat{x}\|_2 \leq \operatorname{diam}(\mathcal{K}_\tau) = 2r(\tau,d)\sqrt{d/\lambda}$. Combining with Lemma 6.5,

$$\|x - a_x\|_2 \leq \frac{\theta M}{\gamma + \theta m}\,\|x-\hat{x}\|_2 \leq \frac{L}{2}\,\sigma^2 \cdot 2r(\tau,d)\sqrt{d/\lambda} = r(\tau,d)L\sqrt{d/\lambda}\ \sigma^2,$$

where the second inequality uses $\theta M \leq L/2$, $M \leq L$, and $(\gamma + m/2)^{-1} \leq \sigma^2$ (the last since $m \geq \lambda$). From [DM19, Proposition 1(ii)], $\mathbb{E}\|z-a_x\|_2^2 \leq \sigma^2 d$, so

$$\sqrt{S} \leq r(\tau,d)L\sqrt{d/\lambda}\ \sigma^2 + \sigma\sqrt{d} = \sigma\sqrt{d}\Big(1 + r(\tau,d)\sqrt{L\kappa}\ \sigma\Big). \tag{6.23}$$

Write $c_s := 1 + r(\tau,d)$. We first impose $\gamma \geq L\kappa$. Since $\sigma^{-2} = \gamma + \lambda/2 \geq \gamma$, this gives $L\kappa\,\sigma^2 \leq 1$, i.e. $\sqrt{L\kappa}\,\sigma \leq 1$, so inequality (6.23) simplifies to $\sqrt{S} \leq c_s\,\sigma\sqrt{d}$. Substitution into (6.22) gives

$$\mathbb{E}_{z\sim\Pi_x} A_+(x,z) \leq c_s\,\eta\,\sigma\sqrt{d} + c_s^2\,L\,\sigma^2 d.$$

Bounding each of the two summands by $\varepsilon/2$, and again using $\sigma^{-2} = \gamma + \lambda/2 \geq \gamma$, yields

$$\gamma \geq 2c_s^2 Ld/\varepsilon \qquad \text{and} \qquad \gamma \geq 4c_s^2 \eta^2 d/\varepsilon^2.$$

Together with $\gamma \geq L\kappa$, all three conditions are implied by (6.20), since $c(\tau, d) \geq 1$, $c(\tau, d) \geq 2c_s^2$ and $c(\tau, d) \geq 4c_s^2$. Therefore $\mathbb{E}_{z\sim\Pi_x}\, \alpha(x, z) \geq 1 - \varepsilon$.

□

## 6.3. Proof of Theorem 3.4

Throughout, write $\mathcal{T}_x$ for the MH transition measure that proposes from $\mathcal{P}_x := Q_x^n \mu_x$ and accepts with probability $\alpha$ as in (3.2) and $\hat{\mathcal{T}}_x$ for the transition measure that proposes from $\Pi_x$ and accepts with this same probability. Since $\alpha$ is the exact MH ratio for proposals drawn from $\Pi_x$, the kernel $\hat{\mathcal{T}}_x$ satisfies detailed balance with respect to $\Pi$, so the chain is $\Pi$-reversible. The proof proceeds in three stages. We first bound the mixing time of $\hat{T}$, then transfer the bound to $T$ via the perturbation estimate of Lemma A.1, and finally, verify that condition (3.7) supplies the required root chain accuracy.

Define $\mathcal{K}_\tau := \mathbb{B}\big(x^\star,\, r(\delta/(4\beta), d)\sqrt{d/\lambda}\big)$. Under the constraint $\log \frac{2\beta}{\delta} \leq d$ we have $\log \frac{4\beta}{\delta}/d \leq 1 + \frac{\log 2}{d} \leq 1 + \log 2$, so

$$r(\delta/(4\beta), d)^2 \leq 1 + 2\sqrt{1 + \log 2} + 2(1 + \log 2) \leq 9, \tag{6.24}$$

giving $r(\delta/(4\beta), d) \leq 3$ and hence the inclusion $\mathcal{K}_\tau \subset \mathcal{K}$. Consequently, Proposition 6.3 implies

$$\Pi(\mathcal{K}) \geq \Pi(\mathcal{K}_\tau) \geq 1 - \frac{\delta}{4\beta}. \tag{6.25}$$

We next establish the TV-continuity of the measures $\hat{\mathcal{T}}_x$ on $\mathcal{K}$. Since $\hat{\mathcal{T}}_x$ corresponds to MH proposals from $\Pi_x$, the set $\{x\}$ is an atom of $\hat{\mathcal{T}}_x$ with mass $1 - \mathbb{E}_{z\sim\Pi_x}\, \alpha(x, z)$, while on $\mathbb{R}^d \setminus \{x\}$ the transition measure admits the density $\alpha(x, \cdot)\, \pi_x \leq \pi_x$. A direct computation gives

$$\|\hat{\mathcal{T}}_x - \Pi_x\|_{\mathrm{TV}} = 1 - \mathbb{E}_{z\sim\Pi_x}\, \alpha(x, z).$$

By the triangle inequality, for $y_1, y_2 \in \mathcal{K}$,

$$\|\hat{\mathcal{T}}_{y_1} - \hat{\mathcal{T}}_{y_2}\|_{\mathrm{TV}} \leq \underbrace{2\Big(1 - \inf_{x\in\mathcal{K}} \mathbb{E}_{z\sim\Pi_x}\, \alpha(x, z)\Big)}_{I} + \underbrace{\|\Pi_{y_1} - \Pi_{y_2}\|_{\mathrm{TV}}}_{II}.$$

It suffices to show that each term is at most $1/4$, so that $\|\hat{\mathcal{T}}_{y_1} - \hat{\mathcal{T}}_{y_2}\|_{\mathrm{TV}} \leq 1/2$.

For Term $I$, choose $\tau^\star \in (0, 1)$ so that $r(\tau^\star, d) = 3$, whence $\mathcal{K}_{\tau^\star} = \mathcal{K}$ and $c(\tau^\star, d) = 4(1 + 3)^2 = 64$. The gradient fidelity condition (3.5) supplies $\eta^2 \lesssim L \max\{1, \kappa/d\}$ in assumption (a2), and the localisation condition (3.6) gives $\theta = 1/2$ and $\gamma \gtrsim L \max\{\kappa, d\}$. Applying Lemma 6.8 with $\varepsilon = 1/8$, the requirement (6.20) reads $\gamma \geq 64\big(L\kappa + 8Ld + 64\,\eta^2 d\big)$, and since $\eta^2 d \lesssim L \max\{d, \kappa\}$, each summand is bounded by a multiple of $L \max\{\kappa, d\}$. Thus (3.6)

implies the requirement. We obtain

$$\inf_{x\in\mathcal{K}} \mathbb{E}_{z\sim\Pi_x}\, \alpha(x,z) \geq \frac{7}{8},$$

and therefore $I \leq 2(1-7/8) = 1/4$.

For Term $II$, Lemma 6.4 gives

$$\|\Pi_{y_1} - \Pi_{y_2}\|_{\mathrm{TV}} \leq \frac{\gamma}{2\sqrt{\gamma+\theta m}}\, \|y_1 - y_2\|_2.$$

Since $\gamma \gtrsim L \geq m = 2\theta m$, the Lipschitz constant satisfies $\frac{\gamma}{2\sqrt{\gamma+\theta m}} \lesssim \sqrt{\gamma}$, so $II \leq 1/4$ whenever

$$\|y_1 - y_2\|_2 \lesssim \gamma^{-1/2} =: t.$$

Collecting the two bounds establishes $\|\hat{\mathcal{T}}_{y_1} - \hat{\mathcal{T}}_{y_2}\|_{\mathrm{TV}} \leq 1/2$ whenever $\|y_1 - y_2\|_2 \leq t$ with $t \asymp \gamma^{-1/2}$, for all $y_1, y_2 \in \mathcal{K}$. A lazy version of $\hat{T}$ (c.f. Appendix A) satisfies continuity with constant $3/4$, which is the hypothesis (6.2) of Lemma 6.2. Since $\hat{T}$ is $\Pi$-reversible, lazy, and (6.25) supplies the mass requirement (6.3) at level $\delta/2$, Lemma 6.2 applies to $\hat{T}$ with tolerance $\delta/2$, giving

$$\|\hat{T}^N\mu - \Pi\|_{\mathrm{TV}} \leq \frac{\delta}{2} \quad \text{for all} \quad N \geq N_\delta := \left\lceil C\, t^{-2}\lambda^{-1} \log\tfrac{4\beta}{\delta} \right\rceil, \tag{6.26}$$

where $C > 0$ is the universal constant of Lemma 6.2. Since $t \asymp \gamma^{-1/2}$ and $\gamma \asymp L \max\{\kappa, d\}$,

$$N_\delta \asymp \frac{\gamma}{\lambda} \log\frac{4\beta}{\delta} \asymp \kappa \max\{\kappa, d\} \log\frac{2\beta}{\delta}. \tag{6.27}$$

Both chains, $\hat{T}$ and $T$, accept with the common probability $\alpha(x,\cdot)$. Hence, for each measurable $A$, $\mathcal{T}_x(A)$ is the integral of the $[0,1]$-valued function $z \mapsto \alpha(x,z)\,\delta_z(A) + (1-\alpha(x,z))\,\delta_x(A)$ against $\mathcal{P}_x$, and $\hat{\mathcal{T}}_x(A)$ the integral of the same function against $\Pi_x$. The functional bound (2.1) and a supremum over $A$ give

$$\|\mathcal{T}_x - \hat{\mathcal{T}}_x\|_{\mathrm{TV}} \leq \|\mathcal{P}_x - \Pi_x\|_{\mathrm{TV}}. \tag{6.28}$$

Define the enlarged region $\mathcal{K}_N := \mathbb{B}\big(x^\star,\, r(\delta/(4\beta N_\delta), d)\sqrt{d/\lambda}\big)$. Proposition 6.3 gives $\Pi({\mathcal{K}_N}^{\mathrm{c}}) \leq \delta/(4\beta N_\delta)$. Applying the perturbation estimate of Lemma A.1 to the lazy chains, whose per-step difference halves that of the kernels $\mathcal{T}_x, \hat{\mathcal{T}}_x$ (Appendix A), with $X = \mathcal{K}_N$, $\nu = \Pi$, and $N = N_\delta$, and substituting (6.28),

$$\|T^{N_\delta}\mu - \hat{T}^{N_\delta}\mu\|_{\mathrm{TV}} \leq N_\delta \sup_{x\in\mathcal{K}_N} \tfrac{1}{2}\|\mathcal{T}_x - \hat{\mathcal{T}}_x\|_{\mathrm{TV}} + \beta N_\delta\, \Pi({\mathcal{K}_N}^{\mathrm{c}}) \leq \frac{N_\delta}{2} \sup_{x\in\mathcal{K}_N} \|\mathcal{P}_x - \Pi_x\|_{\mathrm{TV}} + \frac{\delta}{4}. \tag{6.29}$$

It remains to make the first term at most $\delta/4$. Under (3.6), the condition number of each surrogate density $\pi_x$ is uniformly bounded:

$$\kappa_x := \frac{\gamma+\theta M}{\gamma+\theta m} \leq 1 + \frac{M}{2\gamma} \lesssim 1, \tag{6.30}$$

since $\gamma \gtrsim L \geq M$ and in particular $\tilde{\kappa} := \sup_x \kappa_x \lesssim 1$. For $x \in \mathcal{K}_N$, the triangle inequality through $x^\star$, using $x \in \mathcal{K}_N$ and the standing assumption $\hat{x} \in \mathcal{K}$, gives

$$\|x - \hat{x}\|_2^2 \leq \big(r(\delta/(4\beta N_\delta), d) + 3\big)^2 \frac{d}{\lambda} \lesssim \frac{d + \log(2N_\delta)}{\lambda},$$

where the last step uses $r(\delta/(4\beta N_\delta), d)^2 \lesssim 1 + \log(4\beta N_\delta/\delta)/d$ and $\log(4\beta/\delta) \leq 2d$ (valid for all $d \geq 1$ under $\log(2\beta/\delta) \leq d$). Lemma 6.6 then gives, with $\kappa_x \lesssim 1$ and $\theta^2 M^2/\gamma \leq L^2/(4\gamma) \lesssim L/\max\{\kappa, d\}$ from (3.6),

$$\log \beta_x \leq \frac{d}{2} \log(2\kappa_x) + \frac{\kappa_x\, \theta^2 M^2}{\gamma} \|x - \hat{x}\|_2^2 \lesssim d + \frac{L}{\max\{\kappa, d\}} \cdot \frac{d + \log(2N_\delta)}{\lambda}.$$

Since $L/(\lambda \max\{\kappa, d\}) = \kappa/\max\{\kappa, d\} \leq 1$, the second term is $\lesssim d + \log(2N_\delta)$, so

$$\sup_{x \in \mathcal{K}_N} \log \beta_x \lesssim d + \log(2N_\delta). \tag{6.31}$$

Now apply Condition 3.3 at each $x \in \mathcal{K}_N$, with the Gaussian initial measure $\mu_x$ of Lemma 6.6 and accuracy $\varepsilon_{\mathrm{root}} := \delta/(2N_\delta)$. A common root chain length $n$ is admissible provided

$$n \gtrsim d^\omega \tilde{\kappa}^{\tilde{\omega}} \sup_{x \in \mathcal{K}_N} \log \frac{2\beta_x}{\varepsilon_{\mathrm{root}}} \lesssim d^\omega \Big( d + \log(2N_\delta) + \log \tfrac{N_\delta}{\delta} \Big), \tag{6.32}$$

where we used $\tilde{\kappa} \lesssim 1$ and (6.31). By (6.27) and $\max\{\kappa, d\} \leq \kappa d$, $\log(2\beta/\delta) \leq d$, we have $N_\delta \lesssim \kappa^2 d^2$, so $\log(2N_\delta) \lesssim 1 + \log\kappa + \log d$ and $\log(N_\delta/\delta) \lesssim 1 + \log(\kappa/\delta) + \log d$. Since $\delta < 1$ gives $\log\kappa \leq \log(\kappa/\delta)$, and $1, \log d \leq d$, both terms are $\lesssim d + \log(\kappa/\delta)$. Hence the right-hand side of (6.32) is $\lesssim d^\omega(d + \log(\kappa/\delta))$, which is exactly (3.7). With this $n$, $\sup_{x \in \mathcal{K}_N} \|\mathcal{P}_x - \Pi_x\|_{\mathrm{TV}} \leq \delta/(2N_\delta)$, so the first term in (6.29) is at most $\frac{N_\delta}{2} \cdot \delta/(2N_\delta) = \delta/4$.

In summary, the first term in (6.29) is at most $\delta/4$, so together with its second term, $\|T^{N_\delta}\mu - \hat{T}^{N_\delta}\mu\|_{\mathrm{TV}} \leq \delta/2$. Combining with (6.26) at $N = N_\delta$,

$$\|T^{N_\delta}\mu - \Pi\|_{\mathrm{TV}} \leq \|T^{N_\delta}\mu - \hat{T}^{N_\delta}\mu\|_{\mathrm{TV}} + \|\hat{T}^{N_\delta}\mu - \Pi\|_{\mathrm{TV}} \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta.$$

Hence $t_\delta(\mu) \leq N_\delta \asymp \kappa \max\{\kappa, d\} \log(2\beta/\delta)$ by (6.27), completing the proof.


## Acknowledgements

This work is supported by the Deutsche Forschungsgemeinschaft (German Research Foundation) under Germany's Excellence Strategy EXC 2181/1 (the Heidelberg STRUCTURES Excellence Cluster)

R. Kutri would like to thank the Isaac Newton Institute for Mathematical Sciences for the support and hospitality during the programme 'Representing, calibrating & leveraging prediction uncertainty from statistics to machine learning' where work on this paper was undertaken. This work was supported by EPSRC grant no EP/R014604/1.

# References


[ADH10] Christophe Andrieu, Arnaud Doucet and Roman Holenstein. 'Particle markov chain monte carlo methods'. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 72.3 (2010), pp. 269–342.

[AR09] Christophe Andrieu and Gareth O Roberts. 'The pseudo-marginal approach for efficient Monte Carlo computations'. In: *The Annals of Statistics* 37.2 (2009).

[Ban+23] Afonso S Bandeira, Antoine Maillard, Richard Nickl and Sven Wang. 'On free energy barriers in Gaussian priors and failure of cold start MCMC for high-dimensional unimodal distributions'. In: *Philosophical Transactions of the Royal Society A* 381.2247 (2023), p. 20220150.

[BC93] Norman E Breslow and David G Clayton. 'Approximate inference in generalized linear mixed models'. In: *Journal of the American statistical Association* 88.421 (1993), pp. 9–25.

[BH13] Nawaf Bou-Rabee and Martin Hairer. 'Nonasymptotic mixing of the MALA algorithm'. In: *IMA Journal of Numerical Analysis* 33.1 (2013), pp. 80–110.

[Bha+21] Kaushik Bhattacharya, Bamdad Hosseini, Nikola B Kovachki and Andrew M Stuart. 'Model reduction and neural networks for parametric PDEs'. In: *The SMAI journal of computational mathematics* 7 (2021), pp. 121–157.

[Car+17] Bob Carpenter et al. 'Stan: A probabilistic programming language'. In: *Journal of statistical software* 76 (2017), pp. 1–32.

[CF05] J Andrés Christen and Colin Fox. 'Markov chain Monte Carlo using an approximation'. In: *Journal of Computational and Graphical statistics* 14.4 (2005), pp. 795–810.

[CFO11] Tiangang Cui, Colin Fox and MJ O'sullivan. 'Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis Hastings algorithm'. In: *Water Resources Research* 47.10 (2011).

[Che+20] Yuansi Chen, Raaz Dwivedi, Martin J Wainwright and Bin Yu. 'Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients'. In: *Journal of Machine Learning Research* 21.92 (2020), pp. 1–72.

[Con+16] Patrick R Conrad, Youssef M Marzouk, Natesh S Pillai and Aaron Smith. 'Accelerating asymptotically exact MCMC for computationally intensive models via local approximations'. In: *Journal of the American Statistical Association* 111.516 (2016), pp. 1591–1607.

[Cot+13] Simon L Cotter, Gareth O Roberts, Andrew M Stuart and David White. *MCMC methods for functions: modifying old algorithms to make them faster.* 2013.

[Dal17] Arnak S Dalalyan. 'Theoretical guarantees for approximate sampling from smooth and log-concave densities'. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 79.3 (2017), pp. 651–676.

[DM19] Alain Durmus and Eric Moulines. *High-dimensional Bayesian inference via the unadjusted Langevin algorithm.* 2019.

[DMS17] Alain Durmus, Eric Moulines and Eero Saksman. 'On the convergence of hamiltonian monte carlo'. In: *arXiv preprint arXiv:1705.00166* (2017).

[DTM98] Peter J Diggle, Jonathan A Tawn and Rana A Moyeed. 'Model-based geostatistics'. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 47.3 (1998), pp. 299–350.

[Dua+87] Simon Duane, Anthony D Kennedy, Brian J Pendleton and Duncan Roweth. 'Hybrid monte carlo'. In: *Physics letters B* 195.2 (1987), pp. 216–222.

[Dwi+19] Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright and Bin Yu. 'Log-Concave Sampling: Metropolis-Hastings Algorithms Are Fast'. In: *Journal of Machine Learning Research* 20.183 (2019), pp. 1–42.

[Ebe14] Andreas Eberle. 'Error bounds for Metropolis–Hastings algorithms applied to perturbations of Gaussian measures in high dimensions'. In: *The Annals of Applied Probability* 24.1 (2014).

[GR92] Andrew Gelman and Donald B Rubin. 'Inference from iterative simulation using multiple sequences'. In: *Statistical science* 7.4 (1992), pp. 457–472.

[Gra+15] Ivan G Graham et al. 'Quasi-Monte Carlo finite element methods for elliptic PDEs with lognormal random coefficients'. In: *Numerische Mathematik* 131.2 (2015), pp. 329–368.

[Gra+18] Ivan G Graham, Frances Y Kuo, Dirk Nuyens, Robert Scheichl and Ian H Sloan. 'Analysis of circulant embedding methods for sampling stationary random fields'. In: *SIAM Journal on Numerical Analysis* 56.3 (2018), pp. 1871–1895.

[Gra20] Robert B Gramacy. *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*. Chapman and Hall/CRC, 2020.

[GT95] Charles J Geyer and Elizabeth A Thompson. 'Annealing Markov chain Monte Carlo with applications to ancestral inference'. In: *Journal of the American Statistical Association* 90.431 (1995), pp. 909–920.

[Har04] Gilles Hargé. 'A convex/log-concave correlation inequality for Gaussian measure and an application to abstract Wiener spaces'. In: *Probability theory and related fields* 130 (2004), pp. 415–440.

[Has70] W Keith Hastings. 'Monte Carlo sampling methods using Markov chains and their applications'. In: *Biometrika* 57 (1970), pp. 97–109.

[HG+14] Matthew D Hoffman, Andrew Gelman et al. 'The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.' In: *J. Mach. Learn. Res.* 15.1 (2014), pp. 1593–1623.

[KGV83] Scott Kirkpatrick, C Daniel Gelatt Jr and Mario P Vecchi. 'Optimization by simulated annealing'. In: *science* 220.4598 (1983), pp. 671–680.

[KS26] Robert Kutri and Robert Scheichl. 'Dirichlet–Neumann Averaging: The DNA of Efficient Gaussian Process Simulation'. In: *SIAM/ASA Journal on Uncertainty Quantification* 14.2 (2026), pp. 313–340.

[Liu01] Jun S Liu. *Monte Carlo strategies in scientific computing*. Vol. 10. Springer, 2001.

[Lov99] László Lovász. 'Hit-and-run mixes fast'. In: *Mathematical programming* 86.3 (1999), pp. 443–461.

[LS93] László Lovász and Miklós Simonovits. 'Random walks in a convex body and an improved volume algorithm'. In: *Random structures & algorithms* 4.4 (1993), pp. 359–412.

[LST21] Yin Tat Lee, Ruoqi Shen and Kevin Tian. 'Structured logconcave sampling with a restricted Gaussian oracle'. In: *Conference on Learning Theory*. PMLR. 2021, pp. 2993–3050.

[Lu+21] Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang and George Em Karniadakis. 'Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators'. In: *Nature machine intelligence* 3.3 (2021), pp. 218–229.

[Lyk+20] Mikkel B Lykkegaard, Grigorios Mingas, Robert Scheichl, Colin Fox and Tim J Dodwell. 'Multilevel delayed acceptance MCMC with an adaptive error model in PyMC3'. In: *arXiv preprint arXiv:2012.05668* (2020).

[Lyk+23] Mikkel Bue Lykkegaard, Tim J Dodwell, Colin Fox, Grigorios Mingas and Robert Scheichl. 'Multilevel delayed acceptance MCMC'. In: *SIAM/ASA journal on uncertainty quantification* 11.1 (2023), pp. 1–30.

[Mar+12] James Martin, Lucas C Wilcox, Carsten Burstedde and Omar Ghattas. 'A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion'. In: *SIAM Journal on Scientific Computing* 34.3 (2012), A1460–A1487.

[Met+53] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller and Edward Teller. 'Equation of state calculations by fast computing machines'. In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.

[MP92] Enzo Marinari and Giorgio Parisi. 'Simulated tempering: a new Monte Carlo scheme'. In: *Europhysics letters* 19.6 (1992), p. 451.

[MT08] Sean P Meyn and Richard L Tweedie. *Markov Chains and stochastic stability*. Cambridge University Press, 2008.

[Nea+11] Radford M Neal et al. 'MCMC using Hamiltonian dynamics'. In: *Handbook of markov chain monte carlo* 2.11 (2011), p. 2.

[Nic23] Richard Nickl. *Bayesian non-linear statistical inverse problems*. EMS press Berlin, 2023.

[PS14] Natesh S Pillai and Aaron Smith. 'Ergodicity of approximate MCMC chains with applications to large data sets'. In: *arXiv preprint arXiv:1405.0182* (2014).

[Rao+23] Bogdan Raonic, Roberto Molinaro, Tobias Rohner, Siddhartha Mishra and Emmanuel de Bezenac. 'Convolutional neural operators'. In: *ICLR 2023 workshop on physics for machine learning*. 2023.

[RM51] Herbert Robbins and Sutton Monro. 'A Stochastic Approximation Method'. In: *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407.

[RMC09] Håvard Rue, Sara Martino and Nicolas Chopin. 'Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations'. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 71.2 (2009), pp. 319–392.

[RT96a] Gareth O Roberts and Richard L Tweedie. 'Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms'. In: *Biometrika* 83.1 (1996), pp. 95–110.

[RT96b] Gareth O. Roberts and Richard L. Tweedie. 'Exponential convergence of Langevin distributions and their discrete approximations'. In: *Bernoulli* 2.4 (1996), pp. 341–363.

[Sim+17] Daniel Simpson, Håvard Rue, Andrea Riebler, Thiago G. Martins and Sigrunn H. Sørbye. 'Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors'. In: *Statistical Science* 32.1 (2017), pp. 1–28.

[ST18] Andrew Stuart and Aretha Teckentrup. 'Posterior consistency for Gaussian process approximations of Bayesian posterior distributions'. In: *Mathematics of Computation* 87.310 (2018), pp. 721–753.

[Stu10] Andrew M Stuart. 'Inverse problems: a Bayesian perspective'. In: *Acta numerica* 19 (2010), pp. 451–559.

[TP18] Michalis K Titsias and Omiros Papaspiliopoulos. 'Auxiliary gradient-based sampling algorithms'. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80.4 (2018), pp. 749–767.

[Wes+26] Josephine Westermann, Benno Huber, Thomas O'Leary-Roseberry and Jakob Zech. 'Performance of neural and polynomial operator surrogates'. In: *arXiv preprint arXiv:2604.00689* (2026).

[WSC22] Keru Wu, Scott Schmidler and Yuansi Chen. 'Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling'. In: *Journal of Machine Learning Research* 23.270 (2022), pp. 1–63.

# A. Technical Auxiliary Results

Laziness is required by the conductance bound (2.6). The $\frac{1}{2}$-lazy version of a Markov chain with transition measures $\{\mathcal{T}_x\}_{x\in\mathbb{R}^d}$ has transition measures $\frac{1}{2}\delta_x + \frac{1}{2}\mathcal{T}_x$, which preserves the

invariant measure and reversibility while forcing a non-negative spectrum and removing periodicity. It rescales the constants of Section 6 in two elementary ways.

Comparing two lazy chains at the same state $x$, the shared atom $\frac{1}{2}\delta_x$ cancels, so

$$\big\|(\tfrac{1}{2}\delta_x + \tfrac{1}{2}\mathcal{T}_x) - (\tfrac{1}{2}\delta_x + \tfrac{1}{2}\hat{\mathcal{T}}_x)\big\|_{\mathrm{TV}} = \tfrac{1}{2}\|\mathcal{T}_x - \hat{\mathcal{T}}_x\|_{\mathrm{TV}}.$$

This is the halving used in (6.29). Comparing a single lazy chain at two states $y_1 \neq y_2$, the atoms do not cancel, and $\|\delta_{y_1} - \delta_{y_2}\|_{\mathrm{TV}} \leq 1$, so

$$\big\|(\tfrac{1}{2}\delta_{y_1} + \tfrac{1}{2}\hat{\mathcal{T}}_{y_1}) - (\tfrac{1}{2}\delta_{y_2} + \tfrac{1}{2}\hat{\mathcal{T}}_{y_2})\big\|_{\mathrm{TV}} \leq \tfrac{1}{2}\|\delta_{y_1} - \delta_{y_2}\|_{\mathrm{TV}} + \tfrac{1}{2}\|\hat{\mathcal{T}}_{y_1} - \hat{\mathcal{T}}_{y_2}\|_{\mathrm{TV}} \leq \tfrac{1}{2} + \tfrac{1}{2}\cdot\tfrac{1}{2} = \tfrac{3}{4}.$$

A continuity constant of $1/2$ for the underlying kernels therefore becomes $3/4$ for the lazy chain, as used before Lemma 6.2. See [LS93] for further detail.

The transfer from the idealised to the implemented chain rests on the following perturbation estimate, obtained by a standard telescoping argument with localisation to $X$ in the spirit of [PS14].

**Lemma A.1.** *Let $T$ and $\hat{T}$ be transition operators on $\mathcal{P}(\mathbb{R}^d)$ with transition measures $\{\mathcal{T}_x\}_{x\in\mathbb{R}^d}$ and $\{\hat{\mathcal{T}}_x\}_{x\in\mathbb{R}^d}$, let $\nu$ be invariant under $\hat{T}$, and let $\mu$ be $\beta$-warm with respect to $\nu$, with $\beta \geq 1$. Then for any measurable $X \subseteq \mathbb{R}^d$ and any $N \in \mathbb{N}$,*

$$\|T^N\mu - \hat{T}^N\mu\|_{\mathrm{TV}} \leq N \sup_{x\in X} \|\mathcal{T}_x - \hat{\mathcal{T}}_x\|_{\mathrm{TV}} + \beta N\, \nu(X^{\mathrm{c}}).$$

**Proof.** Writing $\nu_k := \hat{T}^k\mu$, define $\mu_k := T^{N-k}\nu_k$ for $k = 0, \dots, N$, so that $\mu_0 = T^N\mu$ and $\mu_N = \hat{T}^N\mu$. The triangle inequality gives

$$\|T^N\mu - \hat{T}^N\mu\|_{\mathrm{TV}} \leq \sum_{k=0}^{N-1} \|\mu_k - \mu_{k+1}\|_{\mathrm{TV}}. \tag{A.1}$$

Consecutive interpolants differ only in their $(k+1)$-st step, $\mu_k = T^{N-1-k}\big(T\nu_k\big)$ and $\mu_{k+1} = T^{N-1-k}\big(\hat{T}\nu_k\big)$. Both are images of $T\nu_k$ and $\hat{T}\nu_k$ under the common operator $T^{N-1-k}$, which is non-expansive in total variation by (2.1), so

$$\|\mu_k - \mu_{k+1}\|_{\mathrm{TV}} \leq \|T\nu_k - \hat{T}\nu_k\|_{\mathrm{TV}}.$$

For the single-step difference and any measurable $A$,

$$\big|(T\nu_k)(A) - (\hat{T}\nu_k)(A)\big| \leq \int_{\mathbb{R}^d} \big|\mathcal{T}_x(A) - \hat{\mathcal{T}}_x(A)\big|\, \mathrm{d}\nu_k(x) \leq \sup_{x\in X} \|\mathcal{T}_x - \hat{\mathcal{T}}_x\|_{\mathrm{TV}} + \nu_k(X^{\mathrm{c}}),$$

splitting the integral over $X$ and $X^{\mathrm{c}}$, bounding $\nu_k(X) \leq 1$ on the first part and $\big|\mathcal{T}_x(A) - \hat{\mathcal{T}}_x(A)\big| \leq 1$ on the second. Taking the supremum over $A$,

$$\|T\nu_k - \hat{T}\nu_k\|_{\mathrm{TV}} \leq \sup_{x\in X} \|\mathcal{T}_x - \hat{\mathcal{T}}_x\|_{\mathrm{TV}} + \nu_k(X^{\mathrm{c}}).$$

Finally, warmness is preserved under the $\nu$-invariant operator $\hat{T}$. With an analogous argument to (2.5), for every measurable $A$,

$$\nu_k(A) = \int \hat{\Upsilon}_y^k(A)\ \mathrm{d}\mu(y) \leq \beta \int \hat{\Upsilon}_y^k(A)\ \mathrm{d}\nu(y) = \beta\,(\hat{T}^k \nu)(A) = \beta\,\nu(A),$$

using $\hat{T}$-invariance in the last step, so $\nu_k(X^{\mathrm{c}}) \leq \beta\,\nu(X^{\mathrm{c}})$. Summing the $N$ terms in (A.1) yields the claim.

□