CORRECT LOCALLY, SWAP GLOBALLY: DISCRETE LANGEVIN REPLICA EXCHANGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradient-based discrete samplers enable efficient sampling from discrete distributions but struggle on multimodal targets, while parallel tempering complements their local exploitation with global exploration. Combining the two is nontrivial, however, as swapping cannot in general correct the discretization bias of unadjusted local dynamics. We characterize the small-step limits of discrete Langevin sampler as continuous-time Markov jump processes and identify an edge-balance condition ensuring reversibility w.r.t. the target distribution. Building on adjusted-score constructions, we develop A-DULA and AV-DULA, which rely only on discrete energy differences and require no continuous extension of the target. For the resulting limiting replica-exchange processes, we prove that swapping strictly reduces the asymptotic variance and increases the large-deviation rate governing deviations of the cold-chain empirical measure from its target. Under a multimodal decomposition, we further derive spectral-gap lower bounds that quantify the interplay among within-mode relaxation, hot-chain exploration, temperature overlap, and swap efficiency. Experiments on synthetic benchmarks, spin systems, combinatorial targets, and discrete energy-based models demonstrate improved mode exploration and sampling accuracy over the evaluated baselines.

1 INTRODUCTION

Sampling from high-dimensional discrete distributions is a fundamental problem in machine learning and statistical physics, with applications ranging from deep generative modeling (Mikolov et al., 2013; Devlin et al., 2019; Krizhevsky et al., 2012; Ronneberger et al., 2015) to computational biology (Binder, 1995; Zarfaty et al., 2022; Bollback, 2006; Macosko et al., 2015). For a target distribution of the form $\pi(x) \propto \exp U(x)$, the normalizing constant is typically intractable, making Markov chain Monte Carlo (MCMC) a standard computational approach (Robert & Casella, 1999; Neal, 2000). Locally balanced proposals construct informed transitions using local target ratios (Zanella, 2020), but explicitly evaluating coordinate-wise neighborhoods can become costly in high-dimensional spaces. Gradient-based discrete samplers (Grathwohl et al., 2021; Zhang et al., 2022; Rhodes & Gutmann, 2022; Sun et al., 2023a;b) instead exploit gradients of continuous extensions to guide proposal construction and improve sampling efficiency in discrete spaces.

Despite their efficiency, gradient-based discrete samplers can suffer from mode collapse due to an unfavorable exploration-exploitation trade-off. In continuous spaces, a standard remedy is to combine Langevin dynamics with parallel tempering (PT), or replica exchange, where multiple chains at different temperatures explore the landscape and swap states (Chen et al., 2019; Dong & Tong, 2021; Li et al., 2023; Ou et al., 2024). Such tempering is potentially even more valuable in discrete spaces, where discontinuities and large energy barriers can make transitions between modes difficult. Extending this idea to discrete Langevin sampler (DLS), however, raises a compatibility issue. Classical PT combines local kernels that preserve their respective tempered targets with a Metropolis swap criterion that preserves the corresponding joint distribution. For unadjusted discrete Langevin updates, however, the discretization bias is not generally guaranteed to vanish as the step size tends to zero, except in special cases. Consequently, coupling such biased local updates with classical Metropolis swaps does not, in general, guarantee convergence to the desired target distribution.

Recent approaches (Liang et al., 2025; Zheng et al., 2026) introduce history-dependent swap rules to address this mismatch. We show that the proposed detailed-balance characterization can fail even for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

a log-quadratic example (Appendix D.2). This suggests that PT for DLS better begin by establishing compatibility at the level of the local dynamics, rather than relying on the swap step to compensate for local discretization error. This leads to our central question: *How can DLS be coupled with PT while retaining controlled discretization error and provable acceleration?*

We address this by viewing DLS as small-step discretizations of continuous-time Markov jump processes (CTMJPs). This perspective leads to an edge-balance condition for reversibility w.r.t. the target distribution. Enforcing this condition yields adjusted scores based directly on discrete energy differences, giving rise to two parallel constructions: a one-stage discretization, **A-DULA**, and an auxiliary-variable proximal discretization, **AV-DULA**. Their Metropolis-adjusted counterparts, **A-DMALA** and **AV-DMALA**, further remove the finite-step bias. Our contributions are threefold:

1) We show that existing history-dependent swap rules (Liang et al., 2025; Zheng et al., 2026) can violate detailed balance even for log-quadratic targets (Appendix D.2). Viewing DLS as CTMJPs, we derive an edge-balance condition and extend adjusted-score (Gissler et al., 2026) to general discrete spaces, yielding A-DULA and AV-DULA compatible with standard Metropolis swaps.

2) We establish finite-step convergence bounds without the Dobrushin-type regularity conditions required by Gissler et al. (2026). For parallel tempering, we further derive multimodal spectral-gap bounds and prove strict improvements in cold-replica asymptotic variance and large-deviation rates.

3) Across sampling and learning benchmarks, with or without a differentiable continuous extension of the target energy, our methods improve mode exploration and sampling accuracy.

1.1 Related work

Gradient-based Discrete Sampling. Locally balanced proposals use local target ratios to construct informed transitions (Zanella, 2020), inspiring path-based, multiple-try, and any-scale variants (Power & Goldman, 2019; Sun et al., 2022; Gagnon et al., 2023; Sun et al., 2023a). Gradient-based methods reduce neighborhood-evaluation costs by approximating local energy differences through continuous extensions (Grathwohl et al., 2021; Zhang et al., 2022), with subsequent extensions incorporating discrete Wasserstein dynamics (Sun et al., 2023b), preconditioning (Rhodes & Gutmann, 2022), and momentum (Zhou & Tan, 2026; Barboni et al., 2026). Extension-free alternatives include finite-difference proposals (Xiang et al., 2023). Our local constructions generalize the discrete proximal sampler of Bach & Saremi (2025) and adjusted-score DLS of Gissler et al. (2026) from the binary hypercube to general discrete spaces. Our finite-step analysis also avoids the Dobrushin-type regularity condition required in the contraction theory of Gissler et al. (2026). We further combine these with classical PT and derive multimodal acceleration guarantees.

Sampling on Multimodal Distributions. Navigating complex multimodal landscapes is a long-standing challenge in statistical computation. Classical approaches incorporate temperature modulation to overcome energy barriers, including Simulated Annealing (Kirkpatrick et al., 1983), Simulated Tempering (Marinari & Parisi, 1992), and PT (Swendsen & Wang, 1986; Earl & Deem, 2005). In continuous spaces, combining PT with Langevin MCMC has been shown to improve convergence in non-convex settings through spectral-gap analysis (Chen et al., 2019; Dong & Tong, 2022; Deng et al., 2020). In discrete domains, however, multimodality poses a severe challenge. Recent approaches improve global exploration through adaptive step-size and balancing schedules (Pynadath et al., 2024), while other works incorporate PT ideas by introducing tailored swap rules based on historical energy estimates (Liang et al., 2025; Zheng et al., 2026). As shown in Section 2, modifying the swap mechanism alone does not generally ensure convergence to the target. This motivates correcting the local dynamics themselves. We therefore revisit DLS through the lens of CTMJPs and construct target-reversible dynamics compatible with the classical Metropolis swap criterion.

2 Motivation: Compatibility before Swap

Consider a finite product space $\mathcal{X} = \prod_{i=1}^{d} \mathcal{X}_i$, where $\mathcal{X}_i = \{0, \dots, S-1\}$, and the tempered target $\pi_\tau(x) \propto \exp\{U(x)/\tau\}$ for $\tau > 0$, with $\pi := \pi_1$. Given the Stein score $s(x) = \nabla \widetilde{U}(x)$ of a differentiable extension $\widetilde{U}$ of the discrete energy U, DLS samples each coordinate according to

$$q_{\eta,i}^{(\tau)}(y_i \mid x) \propto \exp\left\{ \frac{s_i(x)(y_i - x_i)}{2\tau} - \frac{(y_i - x_i)^2}{2\eta} \right\}, \qquad y_i \in \mathcal{X}_i. \tag{1}$$

This factorization enables parallel coordinate updates. Without Metropolis-Hastings (MH) correction, however, the stationary distribution generally differs from π_τ at finite step sizes. In the log-quadratic setting, this bias vanishes as $\eta \downarrow 0$ (Zhang et al., 2022, Theorem 5.1), although smaller steps make moves increasingly rare and slow down exploration (Liang et al., 2025, Theorem 5.2). Beyond this setting, smaller step sizes do not necessarily eliminate stationary bias.

Recent works (Liang et al., 2025; Zheng et al., 2026) incorporate PT into DLS by modifying the swap rule using historical energy information. Recall that the classical Metropolis swap

$$S\left(x^{(1)}, x^{(2)}\right) = \min\left\{1, \exp\left[\left(\frac{1}{\tau_2} - \frac{1}{\tau_1}\right)\left(U(x^{(1)}) - U(x^{(2)})\right)\right]\right\} \tag{2}$$

preserves the product of the two tempered target distributions. Thus, the fundamental mismatch arises from the local dynamics rather than the classical swap itself: outside special settings, unadjusted DLS lacks a general guarantee that its invariant distribution converges to the tempered target as $\eta \downarrow 0$. To compensate for this mismatch, the above works propose tailored swap rules:

$$\tilde{S}\left(x_{i+1}^{(1)}, x_{i+1}^{(2)} \middle| x_i^{(1)}, x_i^{(2)}\right) = \min\left\{1, \exp\left[\left(\frac{1}{\tau_2} - \frac{1}{\tau_1}\right)\left(U(x_{i+1}^{(1)}) + U(x_i^{(1)}) - U(x_{i+1}^{(2)}) - U(x_i^{(2)})\right)\right]\right\}. \tag{3}$$

Their analyses characterize these historical corrections as restoring detailed balance w.r.t. a step-size-dependent intermediate distribution that converges to the target in the log-quadratic setting (Liang et al., 2025; Zheng et al., 2026). Appendix D.2, however, gives a log-quadratic counterexample to this characterization and further shows that, in this example, modifying the swap rule alone cannot recover the intended joint target. We therefore take a different approach by redesigning the local dynamics rather than correcting them through the swap rule. We construct parallel updates with target-reversible small-step limits that integrate naturally with the classical Metropolis swap.

3 Discrete Langevin Samplers Revisited

We revisit DLS through its small-step continuous-time limit, which separates target reversibility of the limiting dynamics from finite-step discretization error. This characterization yields a simple edge-balance condition and motivates two discretizations.

3.1 Continuous-Time Limits and Edge Balance

Equip each coordinate domain $\mathcal{X}_i$ with a connected undirected graph, and let $\mathcal{N}_i(u)$ denote the neighbors of $u \in \mathcal{X}_i$. For $x \in \mathcal{X}$ and $a \in \mathcal{X}_i$, write $x^{i\to a}$ for the state obtained by replacing the ith coordinate of x with a, and define the tempered energy difference $\Delta_{i,\tau}(x,a) := \frac{U(x^{i\to a}) - U(x)}{\tau} = \log \frac{\pi_\tau(x^{i\to a})}{\pi_\tau(x)}$. Let $\ell_i : \mathcal{X}_i \times \mathcal{X}_i \to [0,\infty)$ be a symmetric move penalty satisfying, for some common $D > 0$, $\ell_i(u,u) = 0$, $\ell_i(u,v) = D$ for neighboring states, and $\ell_i(u,v) > D$ for distinct non-neighbors.[1] We use the effective step size $h_\eta := e^{-D/\eta}$. To quantify the contribution of non-neighboring moves, define

$$\gamma := \min\left\{1, \min_{\substack{1\le i\le d,\ u,v\in\mathcal{X}_i \\ v\notin\mathcal{N}_i(u)\cup\{u\}}} \left(\frac{\ell_i(u,v)}{D} - 1\right)\right\} \tag{4}$$

with $\gamma = 1$ if every pair of distinct states is neighboring. Since the state space is finite and $\ell_i(u,v) > D$ for every distinct non-neighbor, we have $\gamma > 0$. Consequently, every non-neighboring move satisfies $\ell_i(u,v) \ge (1+\gamma)D$, which implies that $e^{-\ell_i(u,v)/\eta} \le h_\eta^{1+\gamma}$. For each coordinate i, let $\sigma_i^{(\tau)} : \mathcal{X} \times \mathcal{X}_i \to \mathbb{R}$ be a score function, where $\sigma_i^{(\tau)}(x,a)$ assigns a score to the candidate value $a \in \mathcal{X}_i$ given the current state $x \in \mathcal{X}$. We set $\sigma_i^{(\tau)}(x, x_i) = 0$. Consider the factorized kernel

$$T_{\eta,\sigma}^{(\tau)}(y \mid x) = \prod_{i=1}^{d} \frac{\exp\left\{\frac{1}{2}\sigma_i^{(\tau)}(x, y_i) - \frac{\ell_i(x_i, y_i)}{\eta}\right\}}{\sum_{b\in\mathcal{X}_i} \exp\left\{\frac{1}{2}\sigma_i^{(\tau)}(x,b) - \frac{\ell_i(x_i,b)}{\eta}\right\}}, \qquad x, y \in \mathcal{X}. \tag{5}$$

[1] For example, on ordinal state spaces with nearest-neighbor graphs, the quadratic penalty $\ell_i(u,v) = (v-u)^2/2$ in equation 1 has edge penalty $D = 1/2$, while non-neighboring moves incur strictly larger penalties.

The original DLS in equation 1 is recovered by taking $\sigma_i^{(\tau)}(x,a) = s_i(x)(a-x_i)/\tau$ and $\ell_i(u,v) = (v-u)^2/2$. A single-coordinate move along a graph edge occurs with probability $O(h_\eta)$. Transitions involving two or more coordinate changes contribute $O(h_\eta^2)$, while single-coordinate moves to non-neighboring values contribute $O(h_\eta^{1+\gamma})$.

Theorem 1. *Let $Q_\sigma^{(\tau)}$ be the generator with off-diagonal jump rates, for $x \neq y$, $Q_\sigma^{(\tau)}(x,y) := \sum_{i=1}^d \sum_{a\in\mathcal{N}_i(x_i)} \exp\left\{\frac{1}{2}\sigma_i^{(\tau)}(x,a)\right\} \mathbf{1}_{\{y=x^{i\to a}\}}$, with $Q_\sigma^{(\tau)}(x,x) = -\sum_{y\neq x} Q_\sigma^{(\tau)}(x,y)$. Then, uniformly over $x,y \in \mathcal{X}$, $T_{\eta,\sigma}^{(\tau)}(y \mid x) = \mathbf{1}_{\{y=x\}} + h_\eta Q_\sigma^{(\tau)}(x,y) + \mathcal{O}(h_\eta^{1+\gamma})$. Moreover, $Q_\sigma^{(\tau)}$ is reversible w.r.t. π_τ if and only if*

$$\sigma_i^{(\tau)}(x,a) - \sigma_i^{(\tau)}(x^{i\to a}, x_i) = 2\Delta_{i,\tau}(x,a), \tag{6}$$

for every $x \in \mathcal{X}$, coordinate i, and $a \in \mathcal{N}_i(x_i)$.

We refer to equation 6 as the edge-balance condition. The expansion associates one iteration with a time interval h_η of the limiting jump process. For the original DLS score, this condition becomes $\big[s_i(x) + s_i(x^{i\to a})\big](a - x_i) = 2\big[U(x^{i\to a}) - U(x)\big]$. It holds whenever the continuous extension is quadratic along the ith coordinate direction, providing a continuous-time explanation for the vanishing-bias behavior of log-quadratic targets (Zhang et al., 2022), but need not hold for general energies. More generally, equation 6 reveals that target reversibility is governed by energy differences along graph edges, motivating constructions based directly on discrete energy differences.

3.2 Two Discretizations

One-Stage Discretization: A-DULA. To enforce the edge-balance condition, we adopt the Barker rate $q_{\mathrm{B}}(u) = \frac{e^u}{1+e^u}$, corresponding to the Barker balancing function in locally balanced MCMC (Zanella, 2020; Livingstone & Zanella, 2022). This yields the adjusted score

$$\sigma_{\mathrm{A},i}^{(\tau)}(x,a) = 2\log\left(\frac{e^{\Delta_{i,\tau}(x,a)}}{1+e^{\Delta_{i,\tau}(x,a)}}\right), \qquad \text{yielding} \qquad Q_{\mathrm{A},\tau}(x,x^{i\to a}) = \frac{e^{\Delta_{i,\tau}(x,a)}}{1+e^{\Delta_{i,\tau}(x,a)}}. \tag{7}$$

For self-transitions, set $\sigma_{\mathrm{A},i}^{(\tau)}(x,x_i) = 0$. Substituting into equation 5 defines *Adjusted Discrete Unadjusted Langevin Algorithm* (A-DULA), whose Barker rates are bounded and satisfy edge balance.

Auxiliary-Variable Construction: AV-DULA. A complementary route draws on the auxiliary-variable view of proximal and denoising-based sampling (Lee et al., 2021; Chen et al., 2022; Bach & Saremi, 2025), combining a target-agnostic perturbation with a target-informed correction. We adapt this idea to obtain factorized updates with target-reversible limiting dynamics. We introduce an auxiliary state $z \in \mathcal{X}$ via $A_\eta(z \mid x) = \prod_{i=1}^d A_{\eta,i}(z_i \mid x_i)$, $A_{\eta,i}(a \mid x_i) = \frac{e^{-\ell_i(x_i,a)/\eta}}{\sum_{b\in\mathcal{X}_i} e^{-\ell_i(x_i,b)/\eta}}$. The associated augmented distribution is $\widetilde{\pi}_{\eta,\tau}(x,z) = \pi_\tau(x)A_\eta(z \mid x)$. After drawing z, an exact conditional update would sample y with probability proportional to $\pi_\tau(y)A_\eta(z \mid y)$, which generally couples coordinates. We replace this conditional update with the factorized correction $B_{\eta,\sigma}^{(\tau)}(y \mid z) := T_{\eta,\sigma}^{(\tau)}(y \mid z)$, using the architecture in equation 5. One iteration first draws $z \sim A_\eta(\cdot \mid x)$ and then $y \sim B_{\eta,\sigma}^{(\tau)}(\cdot \mid z)$, giving the composite kernel $T_{\eta,\sigma,\mathrm{AV}}^{(\tau)}(y \mid x) = \sum_{z\in\mathcal{X}} A_\eta(z \mid x)B_{\eta,\sigma}^{(\tau)}(y \mid z)$.

Proposition 2. *The two-stage kernel satisfies $T_{\eta,\sigma,\mathrm{AV}}^{(\tau)} = I + h_\eta Q_{\sigma,\mathrm{AV}}^{(\tau)} + \mathcal{O}(h_\eta^{1+\gamma})$, where, for $x \neq y$, $Q_{\sigma,\mathrm{AV}}^{(\tau)}(x,y) = \sum_{i=1}^d \sum_{a\in\mathcal{N}_i(x_i)} \left(1 + e^{\sigma_i^{(\tau)}(x,a)/2}\right) \mathbf{1}_{\{y=x^{i\to a}\}}$. Moreover, $Q_{\sigma,\mathrm{AV}}^{(\tau)}$ is reversible w.r.t. π_τ if and only if, on every graph edge,*

$$\log\frac{1+e^{\sigma_i^{(\tau)}(x,a)/2}}{1+e^{\sigma_i^{(\tau)}(x^{i\to a},x_i)/2}} = \Delta_{i,\tau}(x,a). \tag{8}$$

In particular, $\sigma_{\mathrm{AV},i}^{(\tau)}(x,a) = 2\Delta_{i,\tau}(x,a)$ satisfies equation 8.

We call the resulting sampler *Auxiliary-Variable DULA* (AV-DULA). Both A/AV-DULA retain factorized coordinate-wise updates and use only discrete energy differences, without requiring a differentiable continuous extension. Under the common time scaling h_η, the AV limiting process has a

larger spectral gap and smaller asymptotic variance than A-DULA, although this need not translate into a finite-step advantage (Remark 13).

Theorem 3. *For $\mathrm{M} \in \{\mathrm{A}, \mathrm{AV}\}$, let $T^{(\tau)}_{\eta,\mathrm{M}}$ be the finite-step kernel, $Q_{\mathrm{M},\tau}$ its limiting generator, and $\lambda_{\mathrm{M},\tau}$ the corresponding spectral gap. Then there exists $C_{\mathrm{M},\tau} < \infty$ such that, for any initial distribution ν, any integer $n \geq 0$, and sufficiently small h_η,*

$$\left\| \nu \big(T^{(\tau)}_{\eta,\mathrm{M}}\big)^n - \pi_\tau \right\|_{\mathrm{TV}} \leq \frac{1}{2} e^{-\lambda_{\mathrm{M},\tau} n h_\eta} \sqrt{\chi^2(\nu \mid \pi_\tau)} + C_{\mathrm{M},\tau} h_\eta^\gamma.$$

The constant $C_{\mathrm{M},\tau}$ is independent of η, n, and ν, but may depend on the target, temperature, state space, neighborhood structure, and local dynamics.

Theorem 3 decomposes the total error into a mixing term associated with the limiting continuous-time process and a finite-step discretization term of order $\mathcal{O}(h_\eta^\gamma)$. We next study how replica exchange modifies the limiting dynamics and their convergence properties.

MH correction: A-DMALA and AV-DMALA. A-DULA and AV-DULA are reversible in the small-step limit, with finite-step bias of order $\mathcal{O}(h_\eta^\gamma)$, which can be removed by a MH correction. For A-DULA, given a proposal $y \sim T^{(\tau)}_{\eta,\mathrm{A}}(\cdot \mid x)$, the acceptance probability is $\alpha^{(\tau)}_{\mathrm{A}}(x,y) = \min\left\{1, \frac{\pi_\tau(y) T^{(\tau)}_{\eta,\mathrm{A}}(x|y)}{\pi_\tau(x) T^{(\tau)}_{\eta,\mathrm{A}}(y|x)}\right\}$. The resulting sampler, ***A-DMALA***, preserves π_τ exactly for any finite η while retaining the same factorized proposal structure. For AV-DULA, the proposal first draws $z \sim A_\eta(\cdot \mid x)$ and then $y \sim B^{(\tau)}_{\eta,\mathrm{AV}}(\cdot \mid z)$. Directly Metropolizing the marginal proposal would require evaluating $\sum_z A_\eta(z \mid x) B^{(\tau)}_{\eta,\mathrm{AV}}(y \mid z)$, which is generally intractable. We instead apply MH conditionally on the auxiliary state

$$\alpha^{(\tau)}_{\mathrm{AV},z}(x,y) = \min\left\{1, \frac{\pi_\tau(y) A_\eta(z \mid y) B^{(\tau)}_{\eta,\mathrm{AV}}(x \mid z)}{\pi_\tau(x) A_\eta(z \mid x) B^{(\tau)}_{\eta,\mathrm{AV}}(y \mid z)}\right\}.$$

For each fixed z, the conditional MH update satisfies $\pi_\tau(x) A_\eta(z \mid x) K_z(x,y) = \pi_\tau(y) A_\eta(z \mid y) K_z(y,x)$. Summing over z therefore shows that the marginal AV-DMALA kernel is reversible w.r.t. π_τ. Both A-DMALA and AV-DMALA thus remove finite-step bias while preserving the first-order limiting generators of their unadjusted counterparts (Proposition 11).

4 Parallel Tempering with Discrete Langevin Samplers

We now combine the local dynamics from Section 3 with PT. We derive the limiting replica-exchange process, control its finite-step error, and quantify the effect of swaps on asymptotic variance, the large-deviation rate, and the spectral gap. Throughout, $\mathrm{M} \in \{\mathrm{A}, \mathrm{AV}\}$ denotes A-DULA or AV-DULA, respectively.

4.1 Replica-Exchange Dynamics and Finite-Step Convergence

Let $1 = \tau_1 < \tau_2$ and consider the joint state $\mathbf{x} = (x^{(1)}, x^{(2)})$ with target $\Pi(\mathbf{x}) = \pi_{\tau_1}(x^{(1)}) \pi_{\tau_2}(x^{(2)})$. Given a swap intensity $a \geq 0$, define

$$\begin{aligned}\mathcal{L}_{\mathrm{M},a} f(\mathbf{x}) = &\sum_{y \neq x^{(1)}} Q_{\mathrm{M},\tau_1}(x^{(1)}, y)\big[f(y, x^{(2)}) - f(\mathbf{x})\big] + \sum_{y \neq x^{(2)}} Q_{\mathrm{M},\tau_2}(x^{(2)}, y)\big[f(x^{(1)}, y) - f(\mathbf{x})\big] \\ &+ a\, S\Big(x^{(1)}, x^{(2)}\Big)\big[f(x^{(2)}, x^{(1)}) - f(\mathbf{x})\big],\end{aligned}$$

where S is the classical Metropolis swap probability in equation 2. Since the local generators are irreducible and reversible w.r.t. their tempered targets, and the swap satisfies detailed balance w.r.t. Π, the generator $\mathcal{L}_{\mathrm{M},a}$ is irreducible and Π-reversible. Its finite-step counterpart $P_{\eta,a,\mathrm{M}}$ independently updates the two replicas and then swaps them with probability $a h_\eta S$, where $a h_\eta \leq 1$. The following proposition controls its deviation from Π uniformly in time.

Proposition 4. *For either local construction and any fixed swap intensity $a \geq 0$, there exist constants $C_\nu < \infty$ and $C_{a,\mathrm{M}} < \infty$ such that, for any initial distribution ν, any integer $n \geq 0$, and sufficiently small h_η,*

$$\left\|\nu P_{\eta,a,\mathrm{M}}^{n} - \Pi\right\|_{\mathrm{TV}} \leq C_\nu e^{-\lambda_{a,\mathrm{M}} n h_\eta} + C_{a,\mathrm{M}} h_\eta^{\gamma}, \tag{9}$$

where $\lambda_{a,\mathrm{M}}$ is the spectral gap of $\mathcal{L}_{\mathrm{M},a}$. The constants C_ν and $C_{a,\mathrm{M}}$ are independent of h_η and n, with C_ν depending on the initial distribution and $C_{a,\mathrm{M}}$ on the target, local dynamics, and a.

The bound combines exponential relaxation of the limiting replica-exchange process with a time-uniform finite-step discretization error. In particular, the invariant distribution of the finite-step sampler differs from Π by at most $\mathcal{O}(h_\eta^\gamma)$. Since the discretization constant may depend on a, faster limiting relaxation does not necessarily yield a smaller finite-step error.

4.2 Effects of Parallel Tempering

We first study the effect of swap without imposing a multimodal decomposition. The joint Dirichlet form satisfies $\mathcal{E}_{a,\mathrm{M}}(g) = \mathcal{E}_{0,\mathrm{M}}(g) + a\,\mathcal{E}_{\mathrm{sw}}(g)$ and $\mathcal{E}_{\mathrm{sw}}(g) \geq 0$, where $\mathcal{E}_{\mathrm{sw}}$ is the contribution of accepted swaps. Consequently, $\lambda_{a,\mathrm{M}} \geq \lambda_{0,\mathrm{M}} = \min\{\lambda_{\mathrm{M},\tau_1}, \lambda_{\mathrm{M},\tau_2}\}$, so swap cannot decrease the spectral gap of the limiting process[2]. This monotonicity concerns the joint dynamics and need not be strict. We next turn to cold-replica statistics, for which stronger conclusions are available. For an observable $f : \mathcal{X} \to \mathbb{R}$, define its asymptotic variance by $\mathrm{AVar}_{a,\mathrm{M}}(f) := \lim_{T\to\infty} \frac{1}{T} \mathrm{Var}_\Pi\left(\int_0^T f(X_t^{(1)})\,\mathrm{d}t\right)$. When $a = 0$, this reduces to the asymptotic variance of the uncoupled cold chain.

Theorem 5. *For every observable f, $\mathrm{AVar}_{a,\mathrm{M}}(f)$ is nonincreasing in $a \geq 0$. Moreover, for every $a > 0$ and every nonconstant f, $\mathrm{AVar}_{a,\mathrm{M}}(f) < \mathrm{AVar}_{0,\mathrm{M}}(f)$.*

Thus, swap strictly reduces the asymptotic variance of every nonconstant cold-replica observable in continuous time. We next move from individual observables to the empirical distribution of the cold replica. Define $\widehat{\mu}_T^{(1)}(x) := \frac{1}{T}\int_0^T \mathbf{1}_{\{X_t^{(1)}=x\}}\,\mathrm{d}t$ for $x \in \mathcal{X}$, which is the fraction of time the cold replica spends at state x. We study how swap affects the probability that this empirical distribution remains away from its target over a long time horizon.

Theorem 6. *The empirical measure $\widehat{\mu}_T^{(1)}$ satisfies a large deviation principle with speed T and rate function $J_{a,\mathrm{M}} : \mathcal{P}(\mathcal{X}) \to [0,\infty)$, where $\mathcal{P}(\mathcal{X})$ denotes the set of probability distributions on $\mathcal{X}$. For every $\mu \in \mathcal{P}(\mathcal{X})$, $J_{a,\mathrm{M}}(\mu)$ is nondecreasing in a. Moreover, for every $a > 0$, $J_{a,\mathrm{M}}(\mu) > J_{0,\mathrm{M}}(\mu)$ for $\mu \neq \pi_{\tau_1}$, where $J_{0,\mathrm{M}}$ is the rate function without swap.*

The rate function determines how rapidly persistent empirical-distribution deviations become unlikely. In particular, for any $\varepsilon > 0$,

$$\limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_\Pi\left(\|\widehat{\mu}_T^{(1)} - \pi_{\tau_1}\|_{\mathrm{TV}} \geq \varepsilon\right) \leq -\inf_{\|\mu-\pi_{\tau_1}\|_{\mathrm{TV}}\geq\varepsilon} J_{a,\mathrm{M}}(\mu).$$

For every $a > 0$, the infimum on the right is strictly larger than its value at $a = 0$ whenever the deviation set is nonempty (Corollary 16). Thus, swap increases the exponential decay rate of empirical-distribution deviations from the cold target, complementing the asymptotic-variance reduction for individual observables. The proofs are provided in Appendix D.4.5.

Multimodal Spectral Analysis. The preceding results do not require a multimodal decomposition. We now partition $\mathcal{X} = A_1 \cup \cdots \cup A_J$ into modes that mix internally but are difficult to traverse at the cold temperature. Following Woodard et al. (2009); Dong & Tong (2022), let λ_{in} denote the slowest within-mode gap, λ_{hot} the hot mode-projection gap, δ the minimum mode-wise overlap between temperatures, and $\overline{R}_{\mathrm{hot}}$ the maximal hot-chain exit rate. Precise definitions are deferred to Appendix D.4.6.

[2]The spectral gap is the infimum of the Dirichlet-form Rayleigh quotient. Since swap adds the nonnegative term $a\mathcal{E}_{\mathrm{sw}}$, the gap cannot decrease.

Proposition 7 (Informal). *Under the above decomposition, for every swap intensity $a > 0$,*

$$\lambda_{a,\mathrm{M}} \geq \max\left\{\lambda_{0,\mathrm{M}}, \frac{\delta^2}{13}\min\left\{\frac{\lambda_{\mathrm{in}}\lambda_{\mathrm{hot}}}{\lambda_{\mathrm{hot}} + \overline{R}_{\mathrm{hot}}}, \lambda_{\mathrm{hot}}, a\right\}\right\}.$$

Notably, this lower bound does not depend on the cold-chain transition rates between modes. Instead, inter-mode exploration can proceed through the hot chain and accepted swaps, while the bound is controlled by within-mode relaxation, hot-chain exploration, temperature overlap, and swap intensity. Whenever the displayed lower bound exceeds $\lambda_{0,\mathrm{M}}$, it certifies a larger spectral gap and hence a faster-decaying mixing term in equation 9.

5 Experiments

We evaluate four algorithms: (i) **PT-A-DULA** (one-stage unadjusted), (ii) **PT-AV-DULA** (auxiliary-variable unadjusted), (iii) **PT-A-DMALA** (one-stage Metropolis-adjusted), and (iv) **PT-AV-DMALA** (auxiliary-variable Metropolis-adjusted). The code is available at Anonymous GitHub.

Baselines. We categorize baselines based on whether they rely on a continuous extension of the discrete energy: **(1) Extension-based:** We evaluate discrete Langevin samplers (*DULA/DMALA* (Zhang et al., 2022)), step-size scheduled variants (*ACS/mh-ACS* (Pynadath et al., 2024)), discrete replica-exchange Langevin (*DREXEL/DREAM* (Zheng et al., 2026)), and continuous relaxation samplers (*R-MALA* (Grathwohl et al., 2021), *D-SVGD* (Han et al., 2020)). **(2) Extension-free:** We include Gibbs, the locally balanced proposal (*LB* (Zanella, 2020)), and discrete Newton proposals (*UNA/MANA* (Xiang et al., 2023)). We further equip UNA/MANA with PT, yielding PT-UNA/PT-MANA.

5.1 Synthetic Multimodal Experiments

Setup. We use discretized 16-Gaussian, 32-Gaussian, and 16-Student-t mixtures. Global accuracy and mode balance are measured by squared maximum mean discrepancy (MMD^2) (Gretton et al., 2012) and entropic mode coverage (EMC) (Blessing et al., 2024).

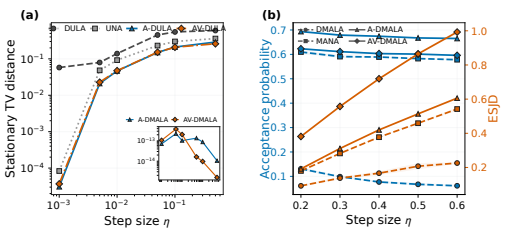


Figure 1: 16-Student-t diagnostics. (a) Stationary TV error (inset: MH-adjusted kernels). (b) Acceptance (blue) and ESJD (orange).

Local Accuracy. Figure 1 examines local behavior on the 16-Student-t target. As the step size decreases, A-DULA and AV-DULA become increasingly accurate, whereas DULA retains a visible bias floor; MH correction eliminates the remaining visible bias (panel a). A/AV-DMALA maintain high acceptance rates while achieving larger expected squared jumping distance (ESJD) than the competing methods (panel b).

Contribution of Swap. Adjusted single-chain samplers improve local accuracy but remain mode-imbalanced (Table 1). PT substantially improves global exploration: A-DULA's MMD^2 drops from 0.1250 to 0.0009 on 16-Gaussian and from 0.1165 to 0.0006 on 32-Gaussian. Across all three targets, the MH-adjusted PT methods achieve $\mathrm{MMD}^2 \leq 0.0009$ and $\mathrm{EMC} \geq 0.996$. We therefore use PT variants as our primary methods in the remaining benchmarks.

5.2 Sampling Tasks

Ising and Potts Models. We evaluate a 16×16 binary Ising model and a categorical Potts model, using long-run Swendsen–Wang samples as reference marginals. Following Zhang et al. (2022), we measure convergence by the RMSE between the estimated and reference marginals. We compare convergence over 50000 iterations and robustness across increasing coupling strengths J, with chains initialized away from the favored phase. Figure 2 shows that the PT methods achieve lower RMSE than the single-chain baselines. PT-AV-DMALA achieves the lowest final error on both targets. As coupling strengthens, several baselines deteriorate sharply, especially on Potts, whereas the MH-adjusted PT methods remain substantially more accurate, with PT-A-DMALA particularly competitive at the strongest couplings.

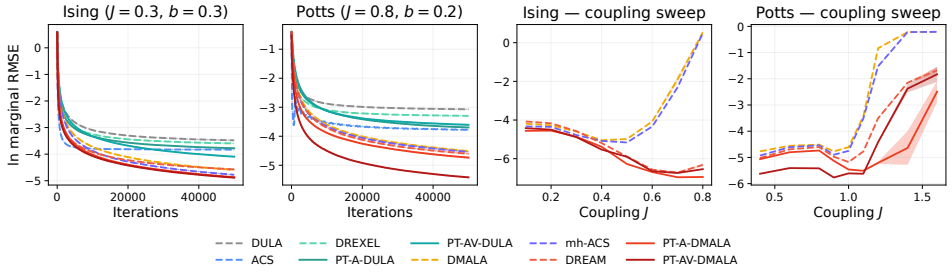


Figure 2: Iteration-only bias-aware log marginal-RMSE on 16×16 Ising and Potts targets.

Table 1: Synthetic multimodal benchmark results across 3 independent runs (mean $\pm$ std). **Bold** and underline indicate the best and second-best results within each group.

	16-Gaussian		**32-Gaussian**		**16-Student-t**	
Sampler	MMD^2 (↓)	EMC (↑)	MMD^2 (↓)	EMC (↑)	MMD^2 (↓)	EMC (↑)
Non-MH Methods						
DULA	0.6388 $\pm$0.0003	0.000 $\pm$0.000	0.5394 $\pm$0.1199	0.138 $\pm$0.138	0.3169 $\pm$0.0833	0.285 $\pm$0.109
A-DULA	0.1250 $\pm$0.0005	0.748 $\pm$0.001	0.1165 $\pm$0.0002	0.796 $\pm$0.000	0.1406 $\pm$0.0763	0.582 $\pm$0.130
AV-DULA	0.1295 $\pm$0.0007	0.740 $\pm$0.001	0.0850 $\pm$0.0355	0.858 $\pm$0.068	0.0724 $\pm$0.0377	0.710 $\pm$0.101
ACS	0.1120 $\pm$0.0546	0.740 $\pm$0.057	0.0803 $\pm$0.0797	0.812 $\pm$0.119	0.0171 $\pm$0.0038	0.920 $\pm$0.020
DREXEL	0.0723 $\pm$0.0059	0.838 $\pm$0.037	0.0363 $\pm$0.0304	0.899 $\pm$0.039	0.0011 $\pm$0.0001	0.995 $\pm$0.001
PT-A-DULA	**0.0009** $\pm$0.0004	**0.997** $\pm$0.001	**0.0006** $\pm$0.0001	**0.998** $\pm$0.000	0.0010 $\pm$0.0001	0.997 $\pm$0.000
PT-AV-DULA	0.0068 $\pm$0.0010	0.978 $\pm$0.002	0.0052 $\pm$0.0003	0.982 $\pm$0.001	**0.0006** $\pm$0.0000	**0.999** $\pm$0.000
MH-Adjusted Methods						
DMALA	0.1767 $\pm$0.0353	0.683 $\pm$0.051	0.1609 $\pm$0.0226	0.706 $\pm$0.046	0.1386 $\pm$0.0638	0.599 $\pm$0.102
A-DMALA	0.1251 $\pm$0.0003	0.750 $\pm$0.001	0.1134 $\pm$0.0001	0.800 $\pm$0.000	0.0932 $\pm$0.0383	0.645 $\pm$0.092
AV-DMALA	0.1248 $\pm$0.0002	0.750 $\pm$0.000	0.1134 $\pm$0.0003	0.800 $\pm$0.000	0.0420 $\pm$0.0152	0.822 $\pm$0.051
mh-ACS	0.0121 $\pm$0.0041	0.949 $\pm$0.019	0.0270 $\pm$0.0092	0.918 $\pm$0.016	0.0079 $\pm$0.0043	0.969 $\pm$0.017
DREAM	0.0136 $\pm$0.0063	0.955 $\pm$0.010	0.0249 $\pm$0.0139	0.943 $\pm$0.016	0.0038 $\pm$0.0021	0.984 $\pm$0.007
PT-A-DMALA	0.0007 $\pm$0.0005	**0.999** $\pm$0.001	**0.0001** $\pm$0.0001	**1.000** $\pm$0.000	0.0009 $\pm$0.0002	0.996 $\pm$0.001
PT-AV-DMALA	**0.0006** $\pm$0.0005	0.999 $\pm$0.001	0.0007 $\pm$0.0005	0.999 $\pm$0.001	**0.0006** $\pm$0.0004	**0.998** $\pm$0.002

Restricted Boltzmann Machines (RBMs). We evaluate MH-adjusted samplers on five pretrained 784–500 binary RBMs, using Block-Gibbs samples as reference. Sampling quality is measured by $\log \mathrm{MMD}^2$, with a disjoint reference split for tuning. As shown in Table 2, our methods perform best on four of the five datasets: PT-AV-DMALA leads on MNIST, PT-A-DMALA on eMNIST, Omniglot, and Caltech, while DREAM leads on kMNIST. [3]

Facility Location Problem. We consider a multimodal facility-location target $U(S) = \sum_{j=1}^{m} \max_{i \in S} c_{ij} - \lambda |S|$. Five separated customer groups with three nearly redundant facilities each yield 15 binary variables, 120 customers, and 243 dominant modes separated by configuration barriers. We set $\lambda = 12$ with a 3% utility perturbation and use exact enumeration for reference marginals. Table 3 shows that all proposed PT methods achieve RMSE below 0.003, substantially lower than the strongest extension-free baselines. PT-AV-DMALA attains the lowest mean RMSE, while PT-AV-DULA achieves the highest ESS/s[4] with slightly larger error.

Table 3: Performance on facility location problem.

Method	RMSE ↓	ESS/s ↑
Non-MH or continuous relaxation Methods		
R-MALA	0.2778 ± 0.0012	0.5123 ± 0.0604
D-SVGD	0.3346 ± 0.0003	0.5548 ± 0.2619
Gibbs	0.1965 ± 0.0048	0.2457 ± 0.2069
LB	0.1226 ± 0.0072	0.0197 ± 0.0042
UNA	0.0288 ± 0.0031	0.2712 ± 0.0768
PT-UNA	0.0129 ± 0.0006	0.2056 ± 0.0069
PT-A-DULA	0.0025 ± 0.0002	4.7666 ± 0.4080
PT-AV-DULA	$\mathbf{0.0021 \pm 0.0002}$	$\mathbf{6.2749 \pm 0.3413}$
MH-adjusted Methods		
MANA	0.0280 ± 0.0025	0.0829 ± 0.0009
PT-MANA	0.0114 ± 0.0025	0.0569 ± 0.0009
PT-A-DMALA	0.0020 ± 0.0006	1.9198 ± 0.1128
PT-AV-DMALA	$\mathbf{0.0018 \pm 0.0009}$	$\mathbf{3.6216 \pm 0.3699}$

[3] Appendix C.2 compares against PT-DMALA with classical MH: a proposed method outperforms in most settings, supporting the benefit of improved local dynamics with standard PT.

[4] ESS/s denotes effective sample size per second, accounting for autocorrelation and runtime.

Table 2: RBM sampling and learning results.

Task	Dataset	DMALA	mh-ACS	DREAM	PT-A-DMALA	PT-AV-DMALA
Sampling $\log \mathrm{MMD}^2$ (↓)	MNIST	-4.316 ± 0.215	-4.746 ± 0.174	-4.670 ± 0.226	-6.007 ± 0.515	$\mathbf{-6.194 \pm 0.034}$
	kMNIST	-5.666 ± 0.059	-5.681 ± 0.239	$\mathbf{-5.791 \pm 0.042}$	-5.764 ± 0.197	-5.674 ± 0.091
	eMNIST	-5.595 ± 0.195	-5.639 ± 0.118	-5.556 ± 0.397	$\mathbf{-5.983 \pm 0.268}$	-5.907 ± 0.292
	Omniglot	-6.393 ± 0.061	-6.432 ± 0.114	-6.221 ± 0.317	$\mathbf{-6.461 \pm 0.139}$	-6.432 ± 0.052
	Caltech	-5.712 ± 0.027	-5.727 ± 0.018	-5.729 ± 0.071	$\mathbf{-5.754 \pm 0.019}$	-5.730 ± 0.047
Learning NLL (↓)	MNIST	135.23 ± 6.17	136.78 ± 7.82	125.64 ± 2.88	$\mathbf{119.62 \pm 5.16}$	128.70 ± 1.95
	kMNIST	252.37 ± 3.18	254.28 ± 11.44	260.50 ± 27.55	241.20 ± 2.70	$\mathbf{238.74 \pm 8.25}$
	eMNIST	169.58 ± 19.22	174.78 ± 9.02	157.28 ± 5.36	$\mathbf{153.42 \pm 1.20}$	162.93 ± 6.53
	Omniglot	138.79 ± 7.06	134.60 ± 4.62	138.11 ± 5.62	$\mathbf{130.76 \pm 0.38}$	137.96 ± 4.39
	Caltech	250.01 ± 17.49	364.92 ± 58.40	246.40 ± 16.93	$\mathbf{245.56 \pm 12.46}$	253.67 ± 19.78

Table 4: EBM learning results on multi-planted MaxSAT.

	Easy: $\lambda = 6$			Medium: $\lambda = 10$			Hard: $\lambda = 14$		
Method	NLL	GradErr	ModeTV	NLL	GradErr	ModeTV	NLL	GradErr	ModeTV
Random initialization									
Gibbs	7.872±0.998	0.988±0.253	0.783±0.229	8.553±3.776	1.011±0.126	0.761±0.044	14.442±4.221	1.033±0.025	0.833±0.162
LB	8.660±1.233	0.984±0.103	0.803±0.039	11.662±1.381	0.928±0.055	0.797±0.009	14.438±0.401	0.978±0.052	0.877±0.010
UNA	10.619±1.643	1.100±0.166	0.876±0.099	10.297±1.283	1.082±0.198	0.802±0.156	13.655±2.218	1.125±0.044	0.893±0.024
MANA	6.303±1.490	1.090±0.147	0.743±0.161	7.586±3.326	1.035±0.138	0.731±0.247	6.444±2.119	1.040±0.210	0.675±0.306
PT-MANA	3.042±0.566	0.289±0.075	0.076±0.002	2.958±0.605	0.285±0.098	0.080±0.025	2.962±0.599	0.327±0.181	0.092±0.041
PT-A-DMALA	**3.038±0.570**	**0.235±0.032**	**0.068±0.007**	2.952±0.614	**0.218±0.009**	**0.063±0.011**	**2.950±0.612**	0.222±0.038	0.068±0.009
PT-AV-DMALA	3.046±0.567	0.252±0.061	0.075±0.019	**2.951±0.610**	0.248±0.102	0.068±0.026	2.953±0.614	**0.210±0.037**	**0.058±0.005**
Single-basin initialization									
Gibbs	11.081±7.065	1.420±0.256	0.863±0.107	13.112±4.434	1.246±0.210	0.889±0.128	16.421±1.638	1.587±0.106	0.998±0.001
LB	10.117±7.149	1.357±0.254	0.798±0.194	16.447±2.389	1.460±0.082	0.973±0.029	16.466±1.640	1.717±0.056	1.000±0.000
UNA	9.862±6.208	1.202±0.028	0.825±0.133	12.790±5.497	1.375±0.129	0.942±0.055	16.393±2.383	1.439±0.262	0.902±0.169
MANA	4.830±1.209	1.046±0.346	0.626±0.225	6.597±1.716	0.977±0.265	0.608±0.099	5.453±2.142	0.877±0.208	0.495±0.258
PT-MANA	3.041±0.567	0.292±0.087	0.083±0.020	2.961±0.600	0.319±0.130	0.094±0.035	2.955±0.608	0.330±0.148	0.094±0.027
PT-A-DMALA	**3.038±0.568**	**0.221±0.042**	**0.067±0.002**	2.956±0.616	**0.253±0.066**	0.079±0.014	**2.948±0.613**	**0.236±0.040**	**0.071±0.004**
PT-AV-DMALA	3.046±0.565	0.240±0.071	0.070±0.015	**2.950±0.611**	0.258±0.083	**0.062±0.013**	2.949±0.611	0.255±0.032	0.073±0.003

5.3 Learning Tasks

Learning RBMs. We train RBMs using persistent contrastive divergence (PCD) (Tieleman, 2008). Following Zhang et al. (2022), after marginalizing out the hidden units, the visible distribution takes the form $p_\theta(x) \propto \exp\{U_\theta(x)\}$, where $U_\theta(x) = \sum_i \mathrm{softplus}((Wx + a)_i) + b^\top x$ and $\theta = (W, a, b)$. At each parameter update, persistent particles are advanced using the corresponding sampler to approximate the model expectation in the likelihood gradient. We optimize with Adam and evaluate held-out negative log-likelihood (NLL) using annealed importance sampling (AIS). As shown in Table 2, our methods achieve the lowest mean NLL on all five datasets.

Learning Energy-Based Models (EBMs) on Multi-Planted MaxSAT. We study PCD learning on multimodal 3-SAT EBMs with R planted modes (Ahmadi et al., 2026), defined by $p_\theta(x) \propto \exp\{U_\theta(x)\}$ with $U_\theta(x) = \theta^\top(2x - \mathbf{1}) - \lambda \sum_{k=1}^{m} w_k \mathbb{1}_{\{C_k(x)=\mathrm{False}\}}$. We set $d = 18$, $R = 8$, and $m = 108$, and exactly enumerate all 2^{18} states to evaluate NLL, gradient error (GradErr), and mode coverage (ModeTV) under random and single-basin initialization. Table 4 reports representative penalties $\lambda \in \{6, 10, 14\}$. At larger λ, non-PT baselines incur substantial errors, with ModeTV approaching one under single-basin initialization. In contrast, PT-A-DMALA and PT-AV-DMALA remain robust, maintaining $\mathrm{GradErr} \leq 0.26$ and $\mathrm{ModeTV} \leq 0.08$. At $\lambda = 14$ under single-basin initialization, they attain $\mathrm{NLL}/\mathrm{ModeTV}$ values of $2.948/0.071$ and $2.949/0.073$, respectively.

6 Conclusion

In this paper, we address the compatibility between DLS and PT by enforcing local edge balance through discrete energy differences. The resulting A-DULA and AV-DULA retain parallel coordinate updates and admit target-reversible limiting dynamics, while their Metropolis-adjusted variants achieve exact finite-step invariance. Coupled with PT, these constructions enjoy finite-step convergence guarantees and provable improvements in cold-replica asymptotic variance, large-deviation rates, and multimodal spectral gaps. Promising directions include non-reversible local dynamics, non-reversible PT, and adaptive temperature selection.

AI USE STATEMENT

AI-based tools were used during manuscript preparation for limited purposes, including language editing, proofreading, and code implementation assistance. All mathematical derivations, theoretical results, algorithmic designs, and scientific conclusions were developed by the authors. Any AI-assisted code was manually reviewed, and all experimental implementations and empirical analyses were verified by the authors.

REFERENCES

Ali Ahmadi, Kiarash Banihashem, Iman Gholami, Mohammad Taghi Hajiaghayi, and Jan Olkowski. Quiet planting for k-sat, multiple solutions of arbitrary geometry. *arXiv preprint arXiv:2606.15979*, 2026.

Francis Bach and Saeed Saremi. Sampling binary data by denoising through score functions. *arXiv preprint arXiv:2502.00557*, 2025.

Raphaël Barboni, Sebastiano Grazzi, and Giacomo Zanella. Hamiltonian dynamics for sampling on discrete spaces. *arXiv preprint arXiv:2608.17961*, 2026.

Kurt Binder (ed.). *The Monte Carlo Method in Condensed Matter Physics*, volume 71 of *Topics in Applied Physics*. Springer, Berlin, Heidelberg, 2 edition, 1995. doi: 10.1007/3-540-60174-0.

Denis Blessing, Xiaogang Jia, Johannes Esslinger, Francisco Vargas, and Gerhard Neumann. Beyond elbos: A large-scale evaluation of variational methods for sampling. *arXiv preprint arXiv:2406.07423*, 2024.

Jonathan P. Bollback. SIMMAP: Stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics*, 7:88, 2006. doi: 10.1186/1471-2105-7-88.

Yi Chen, Jinglin Chen, Jing Dong, Jian Peng, and Zhaoran Wang. Accelerating nonconvex learning via replica exchange langevin diffusion. In *International Conference on Learning Representations*, 2019.

Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, pp. 2984–3014. PMLR, 2022.

Wei Deng, Qi Feng, Georgios Karagiannis, Guang Lin, and Faming Liang. Accelerating convergence of replica exchange stochastic gradient mcmc via variance reduction. *arXiv preprint arXiv:2010.01084*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

Jing Dong and Xin T Tong. Replica exchange for non-convex optimization. *Journal of Machine Learning Research*, 22(173):1–59, 2021.

Jing Dong and Xin T Tong. Spectral gap of replica exchange langevin diffusion on mixture distributions. *Stochastic Processes and their Applications*, 151:451–489, 2022.

David J Earl and Michael W Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.

Philippe Gagnon, Florian Maire, and Giacomo Zanella. Improving multiple-try metropolis with local balancing. *Journal of Machine Learning Research*, 24(248):1–59, 2023.

Armand Gissler, Saeed Saremi, and Francis Bach. Adjusted scores for discrete langevin algorithms. *arXiv preprint arXiv:2602.15587*, 2026.

Will Grathwohl, Kevin Swersky, Milad Hashemi, David Duvenaud, and Chris J Maddison. Oops i took a gradient: Scalable sampling for discrete distributions. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 3831–3841. PMLR, 2021.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Jun Han, Fan Ding, Xianglong Liu, Lorenzo Torresani, Jian Peng, and Qiang Liu. Stein variational inference for discrete distributions. In *International Conference on Artificial Intelligence and Statistics*, pp. 4563–4572. PMLR, 2020.

Scott Kirkpatrick, C Daniel Gelatt Jr, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

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

Guanxun Li, Guang Lin, Zecheng Zhang, and Quan Zhou. Fast replica exchange stochastic gradient langevin dynamics. *arXiv preprint arXiv:2301.01898*, 2023.

Luxu Liang, Yuhang Jia, and Feng Zhou. Enhancing gradient-based discrete sampling via parallel tempering. *arXiv preprint arXiv:2502.19240*, 2025.

Samuel Livingstone and Giacomo Zanella. The barker proposal: Combining robustness and efficiency in gradient-based mcmc. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):496–523, 2022.

Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

Enzo Marinari and Giorgio Parisi. Simulated tempering: a new monte carlo scheme. *Europhysics letters*, 19(6):451, 1992.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013.

Radford M Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

Na Ou, Zecheng Zhang, and Guang Lin. A replica exchange preconditioned crank-nicolson langevin dynamic mcmc method with multi-variance strategy for bayesian inverse problems. *Journal of Computational Physics*, 510:113067, 2024.

Samuel Power and Jacob Vorstrup Goldman. Accelerated sampling on discrete spaces with non-reversible markov processes. *arXiv preprint arXiv:1912.04681*, 2019.

Patrick Pynadath, Riddhiman Bhattacharya, Arun Hariharan, and Ruqi Zhang. Gradient-based discrete sampling with automatic cyclical scheduling. *Advances in Neural Information Processing Systems*, 37:46728–46763, 2024.

Benjamin Rhodes and Michael U. Gutmann. Enhanced gradient-based MCMC in discrete spaces. *Transactions on Machine Learning Research*, 2022.

Christian P Robert and George Casella. *Monte Carlo statistical methods*. Springer, 1999.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

Haoran Sun, Hanjun Dai, Wei Xia, and Arun Ramamurthy. Path auxiliary proposal for mcmc in discrete space. In *International Conference on Learning Representations*, 2022.

Haoran Sun, Bo Dai, Charles Sutton, Dale Schuurmans, and Hanjun Dai. Any-scale balanced samplers for discrete space. In *International Conference on Learning Representations*, 2023a.

Haoran Sun, Hanjun Dai, Bo Dai, Haomin Zhou, and Dale Schuurmans. Discrete langevin samplers via Wasserstein gradient flow. In *International Conference on Artificial Intelligence and Statistics*, pp. 6290–6313. PMLR, 2023b.

Robert H Swendsen and Jian-Sheng Wang. Replica monte carlo simulation of spin-glasses. *Physical review letters*, 57(21):2607, 1986.

Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pp. 1064–1071, 2008.

Dawn B. Woodard, Scott C. Schmidler, and Mark Huber. Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *The Annals of Applied Probability*, 19: 617–640, 2009.

Yue Xiang, Dongyao Zhu, Bowen Lei, Dongkuan Xu, and Ruqi Zhang. Efficient informed proposals for discrete distributions via newton's series approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 7288–7310. PMLR, 2023.

Giacomo Zanella. Informed proposals for local mcmc in discrete spaces. *Journal of the American Statistical Association*, 115(530):852–865, 2020.

Lior Zarfaty, Eli Barkai, and David A Kessler. Discrete sampling of extreme events modifies their statistics. *Physical Review Letters*, 129(9):094101, 2022.

Ruqi Zhang, Xingchao Liu, and Qiang Liu. A langevin-like sampler for discrete distributions. In *International Conference on Machine Learning*, pp. 26375–26396. PMLR, 2022.

Haoyang Zheng, Hengrong Du, Ruqi Zhang, and Guang Lin. Exploring non-convex discrete energy landscapes: An efficient langevin-like sampler with replica exchange. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 36775–36783, 2026.

Yuze Zhou and Zhiqiang Tan. Preconditioned discrete-hams: A second-order irreversible discrete sampler. *Statistics and Computing*, 36(4):159, 2026.

CONTENTS

A Algorithm **14**

B Discussion **14**

B.1 Computational complexity 14

B.2 Choice of swap intensity and temperature 17

C Experimental Details **17**

C.1 Synthetic Benchmarks 17

C.2 Sampling Tasks 18

C.3 Learning Tasks 19

D Proofs and Supporting Results **20**

D.1 Preliminaries 20

D.2 Proofs Omitted in Section 2 22

D.3 Proofs Omitted in Section 3 24

D.3.1 Continuous-Time Limits and Edge Balance 24

D.3.2 Finite-Step Convergence and Stationary Bias 26

D.3.3 MH Corrections 27

D.3.4 Comparison of the Limiting Processes 29

D.4 Proofs Omitted in Section 4 30

D.4.1 Replica-Exchange Generator 30

D.4.2 Finite-Step Convergence 31

D.4.3 Dirichlet Forms and Spectral Monotonicity 32

D.4.4 Cold-Replica Asymptotic Variance 32

D.4.5 Large Deviations Analysis 33

D.4.6 Multimodal Spectral-Gap Bounds 35

A ALGORITHM

In Sections 3 and 4, we formulated A-DULA, AV-DULA and their Metropolis-adjusted counterparts (A-DMALA, AV-DMALA) within a parallel tempering framework. In this appendix, we present their unified algorithmic implementation in Algorithm 1.

Canonical implementation: the FULL scheme. Under the factorized transition kernel equation 5, the exact proposal at any stage requires evaluating energy differences across all coordinate levels in $\mathcal{X}_i$. We refer to this scheme as the **FULL** implementation. Specifically, given an input state w for proposal generation (which represents the current configuration x in A-DULA/DMALA, or the intermediate perturbed state z in AV-DULA/DMALA), FULL computes the exact discrete energy difference $\Delta_{i,\tau}(w,b) = \frac{U(w^{i\to b})-U(w)}{\tau}$ for every candidate level $b \in \mathcal{X}_i \setminus \{w_i\}$. The corresponding local scores are then assigned according to the respective algorithm family:

$$\sigma^{(\tau)}_{\mathrm{A},i}(w,b) = 2\log\frac{e^{\Delta_{i,\tau}(w,b)}}{1+e^{\Delta_{i,\tau}(w,b)}}, \qquad \sigma^{(\tau)}_{\mathrm{AV},i}(w,b) = 2\Delta_{i,\tau}(w,b).$$

Self-transition scores are set to zero by definition. The coordinate-wise normalizers in equation 5 sum over the entire discrete domain $\mathcal{X}_i$, and coordinate proposals within each stage are drawn independently conditional on the input w. All primary results in Section 5 reported for the proposed A/AV samplers use the FULL implementation. Alternative lower-cost variants designed for ordinal spaces, specifically NBR (which truncates support to nearest graph neighbors) and APP (which linearly extrapolates neighbor differences), are detailed and ablated separately in Appendix B.

B DISCUSSION

B.1 COMPUTATIONAL COMPLEXITY

Evaluating a FULL score table requires $N_{\mathrm{full}} = \sum_i(|\mathcal{X}_i|-1)$ candidate queries of U. For binary and one-hot categorical domains, this matches the query complexity of finite-difference Newton-type proposals (Xiang et al., 2023), scaling as $O(d)$ and $O(dS)$, respectively. On ordinal grids, scalar finite differences can populate all Newton logits using $O(d)$ target queries, whereas FULL evaluates all $d(S-1)$ differences explicitly. This per-table query budget matches that of LB proposals over single-coordinate transitions (Zanella, 2020); crucially, however, our factorized construction leverages this information to propose simultaneous multi-coordinate updates. These counts refer to the construction of one score table rather than wall-clock cost. Structured energies, batching, and caching can substantially reduce the actual overhead. MH correction affects the two constructions differently: A-DMALA generally requires an additional score table anchored at the proposed state y to evaluate the reverse proposal, whereas AV-DMALA evaluates both correction probabilities using the same table anchored at z, and therefore requires no additional target-informed score table for the reverse move.

Lower-cost ordinal approximations. To reduce computational overhead on ordinal domains, we consider two lower-cost variants of the FULL scheme: NBR and APP. The NBR variant restricts each coordinate update to the current state and its adjacent graph neighbors, using exact energy differences while assigning zero probability to non-neighbor transitions and omitting them from the corresponding normalizers. For AV-DULA/DMALA, this support restriction is applied consistently to both the perturbation and correction stages. The APP variant instead retains the full support $\mathcal{X}_i$ but approximates nonlocal energy differences by extrapolating from immediate neighbors; in the AV construction, the perturbation kernel A_η remains unchanged and only the target-dependent differences in the correction stage are approximated. Specifically, given ordered levels $v_{i,0},\dots,v_{i,S-1}$ and a current coordinate state $w_i = v_{i,a}$, we compute the forward and backward finite differences to the immediate neighbors, $\delta_{i,+} := \Delta_{i,\tau}(w, v_{i,a+1})$ for $a < S-1$, and $\delta_{i,-} := \Delta_{i,\tau}(w, v_{i,a-1})$ for $a > 0$, and extrapolate them to arbitrary ordered levels $v_{i,b}$ via

$$\widetilde{\Delta}_{i,\tau}(w, v_{i,b}) = \begin{cases} (b-a)\delta_{i,+}, & b > a, \\ 0, & b = a, \\ (a-b)\delta_{i,-}, & b < a. \end{cases}$$

Algorithm 1 Two-Replica FULL A/AV-DULA and A/AV-DMALA with PT

Require: Target energy U; coordinate graphs and penalties ℓ_i with edge penalty D; temperatures $1 = \tau_1 < \tau_2$; step size $\eta > 0$; swap intensity $a \geq 0$ satisfying $ae^{-D/\eta} \leq 1$; algorithm family $\mathrm{M} \in \{\mathrm{A}, \mathrm{AV}\}$; MH flag $\mathrm{use_MH} \in \{\mathrm{true}, \mathrm{false}\}$; initial pair $(x_0^{(1)}, x_0^{(2)})$; total iterations N_{iter}.

Set $h \leftarrow e^{-D/\eta}$.
for $m = 1, \ldots, N_{\mathrm{iter}}$ **do**
  {**Stage 1: Local updates**}
  for all $k \in \{1, 2\}$ **in parallel do**
    Set $x \leftarrow x_{m-1}^{(k)}$ and $\alpha \leftarrow 1$.
    if $\mathrm{M} = \mathrm{A}$ **then**
      Evaluate $\sigma_{\mathrm{A},i}^{(\tau_k)}(x, b)$ for all $b \in \mathcal{X}_i \setminus \{x_i\}$ via equation 7, and set $\sigma_{\mathrm{A},i}^{(\tau_k)}(x, x_i) = 0$.
      Sample $y \sim T_{\eta,\mathrm{A}}^{(\tau_k)}(\cdot \mid x)$ using the full-support factorized kernel equation 5.
      if use_MH **then**
        Evaluate $\sigma_{\mathrm{A},i}^{(\tau_k)}(y, b)$ for all $b \in \mathcal{X}_i \setminus \{y_i\}$ and set $\sigma_{\mathrm{A},i}^{(\tau_k)}(y, y_i) = 0$ to form $T_{\eta,\mathrm{A}}^{(\tau_k)}(\cdot \mid y)$.
        Set

$$\alpha \leftarrow \min\left\{1,\, e^{[U(y)-U(x)]/\tau_k}\, \frac{T_{\eta,\mathrm{A}}^{(\tau_k)}(x \mid y)}{T_{\eta,\mathrm{A}}^{(\tau_k)}(y \mid x)}\right\}.$$

      end if
    else
      Sample $z \sim A_\eta(\cdot \mid x)$, where $A_{\eta,i}(b \mid x_i) \propto \exp\{-\ell_i(x_i, b)/\eta\}$ for all $b \in \mathcal{X}_i$.
      Evaluate $\sigma_{\mathrm{AV},i}^{(\tau_k)}(z, b) = 2\Delta_{i,\tau_k}(z, b)$ for all $b \in \mathcal{X}_i$ (with self-score zero).
      Sample $y \sim B_{\eta,\mathrm{AV}}^{(\tau_k)}(\cdot \mid z) \equiv T_{\eta,\sigma_{\mathrm{AV}}}^{(\tau_k)}(\cdot \mid z)$ using equation 5.
      if use_MH **then**
        Set

$$\alpha \leftarrow \min\left\{1,\, e^{[U(y)-U(x)]/\tau_k}\, \frac{A_\eta(z \mid y) B_{\eta,\mathrm{AV}}^{(\tau_k)}(x \mid z)}{A_\eta(z \mid x) B_{\eta,\mathrm{AV}}^{(\tau_k)}(y \mid z)}\right\}.$$

      end if
    end if
    Draw $u \sim \mathrm{Uniform}(0, 1)$ and set $x_m^{(k)} \leftarrow y$ if $u \leq \alpha$, otherwise $x_m^{(k)} \leftarrow x$.
  end for
  {**Stage 2: Swapping**}
  Compute the Metropolis swap acceptance $S(x_m^{(1)}, x_m^{(2)})$ via equation 2.
  Draw the swap-event indicator $b \sim \mathrm{Bernoulli}\left(ah\, S(x_m^{(1)}, x_m^{(2)})\right)$.
  if $b = 1$ **then**
    Swap $x_m^{(1)} \leftrightarrow x_m^{(2)}$.
  end if
end for
return Cold-replica trajectory $\{x_m^{(1)}\}_{m=1}^{N_{\mathrm{iter}}}$.

At the boundary levels, only the available one-sided difference is required.

These surrogate differences are substituted into the A/AV score definitions, with self-scores fixed to zero, while retaining the original coordinate penalties and all candidate states. Both variants require at most $2d$ exact energy differences per target-informed score table. However, APP still incurs $O(dS)$ logit-computation cost because it retains the full support, whereas NBR operates entirely in $O(d)$. These savings are specific to sparse coordinate graphs, such as chains or grids, and do not extend to binary or fully connected categorical domains, where every distinct state is adjacent. When MH correction is enabled, the acceptance probability is computed using the corresponding NBR or APP proposal kernel together with exact target energies; the PT exchange step is unchanged.

Theoretical consistency. Both variants retain exactly the same unnormalized edge scores as FULL; their finite-step transition probabilities differ because their normalizers and/or non-edge logits differ. On any fixed finite state space, APP assigns $O(h_\eta^{1+\gamma})$ weights to non-neighbor transitions, whereas NBR truncates them to zero. Consequently, invoking Appendix D.1 and the associated kernel expansions yields

$$T_{\eta,\mathrm{M}}^{(\tau),\mathrm{NBR}} = I + h_\eta Q_{\mathrm{M},\tau} + \mathcal{O}_\infty(h_\eta^2), \qquad T_{\eta,\mathrm{M}}^{(\tau),\mathrm{APP}} = I + h_\eta Q_{\mathrm{M},\tau} + \mathcal{O}_\infty(h_\eta^{1+\gamma}),$$

for $\mathrm{M} \in \{\mathrm{A}, \mathrm{AV}\}$. The limiting continuous-time generators and continuous-time PT guarantees thus remain unaltered. Furthermore, the MH-adjusted variants retain exact target stationarity and the same first-order limiting generators as their corresponding unadjusted kernels (Proposition 11). We emphasize that NBR and APP define distinct finite-step Markov kernels rather than mathematically equivalent implementations of FULL.

Experimental setup. We benchmark the unadjusted A/AV-DULA variants under the FULL, NBR, and APP formulations on two ordinal state spaces. For both benchmarks, we sweep a grid of step sizes for every sampler–implementation pair rather than fixing a single step size, thereby assessing the implementations across discretization scales. Results are averaged over three random seeds.

- 16-component Student-t mixture. We use the 16-component Student-t mixture on the uniform 512×512 grid over $[-2, 2]^2$ described in Appendix C.1, with 2 degrees of freedom and scale 0.025. Each coordinate is treated as an ordinal variable with domain $\mathcal{X}_i = \left\{-2 + \frac{4j}{511} : j = 0, \ldots, 511\right\}$, corresponding to 512 equally spaced ordered levels, with adjacent levels connected in the coordinate graph. All chains are initialized near the same mode and run for up to 10^7 iterations. Sampling error is measured by squared maximum mean discrepancy (MMD^2) to the exact target, averaged over six Gaussian RBF kernels with different bandwidths.
- 32-level grayscale image denoising. Each of the 16×16 pixels is treated as an ordinal variable with domain $\mathcal{X}_i = \left\{\frac{j}{31} : j = 0, \ldots, 31\right\}$, corresponding to 32 equally spaced grayscale levels in $[0, 1]$, with adjacent intensity levels connected in the coordinate graph. Given an observed noisy image y, the posterior distribution is

  $$\pi(x \mid y) \propto \exp\left[-\frac{\|x - y\|^2}{2(0.15)^2} - 6 \sum_{(i,j)\in E} (x_i - x_j)^2\right],$$

  where E denotes the non-periodic 4-nearest-neighbor spatial lattice over pixels. All chains are initialized from the zero image and run for 5×10^4 iterations. Long-run Gibbs trajectories serve as reference samples for computing full-state MMD^2.

Figure 3 compares the three implementations in terms of sampling accuracy and computational cost. On the Student-t mixture, APP exhibits per-iteration convergence competitive with FULL while demonstrating clear computational advantages in runtime, particularly under AV-DULA; by contrast, NBR converges substantially slower per iteration. In the denoising task, all three implementations achieve comparable terminal errors, with NBR displaying competitive wall-clock efficiency and APP offering no clear performance advantage. Note that the separately tuned step sizes for APP prevent attributing these differences solely to linear score extrapolation. Overall, these findings highlight task-dependent computational trade-offs rather than a uniform speedup.

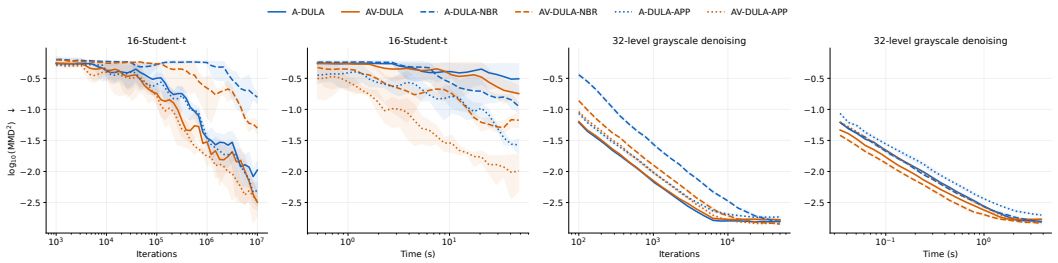


Figure 3: Computational ablation across ordinal implementations: FULL (solid), NBR (dashed), and APP (dotted). Curves report the $\log_{10}$ mean MMD^2 across three random seeds, with shaded regions denoting the seed-wise minimum and maximum. Time corresponds to total wall-clock runtime.

B.2 CHOICE OF SWAP INTENSITY AND TEMPERATURE

Choice of swap intensity. The continuous-time exchange rate a enters the discrete per-iteration swap probability as $ah_\eta S$, subject to the probability bound $ah_\eta \leq 1$. In the two-replica setting ($K = 2$), increasing a monotonically improves (or preserves) the continuous-time spectral gap and strictly reduces the asymptotic variance of every non-constant cold observable over identical continuous-time horizons (Lemma 15 and Theorem 5). We note that saturation of the theoretical multimodal lower bound does not imply that the true spectral gap has plateaued. Crucially, finite-step discretization bias does not necessarily decrease with a, as the underlying error constants scale with the exchange intensity. Because our convergence guarantees fix a as $h_\eta \to 0$ and do not hold uniformly for diverging rate schedules $a = a(h_\eta)$, the exchange intensity should be chosen jointly with the step size η and the targeted computational budget.

Choice of temperatures. We fix the cold temperature at $\tau_1 = 1$ and use $\tau_2 > 1$ to flatten the energy landscape of the hot replica. On any fixed graph edge, $\Delta_{i,\tau_2}(x, b) \to 0$ as $\tau_2 \to \infty$, so the A/AV jump rates become increasingly insensitive to the target energy and approach the graph-determined limits $1/2$ and 2, which explains why a hotter replica can facilitate cross-mode exploration.

However, faster exploration of the hot chain is only one part of the replica-exchange mechanism. States discovered at the hot temperature must also remain sufficiently compatible with the cold distribution to be transferred effectively through Metropolis swaps. This trade-off appears explicitly in our multimodal bound through the hot-chain mode-projection gap λ_{hot} and the cross-temperature overlap factor δ^2. Increasing τ_2 may improve hot-chain exploration while changing the overlap with the cold target, and our theory does not imply monotonicity of either quantity, nor of the resulting finite-step sampling performance. Temperature selection should balance hot-chain exploration against cross-temperature overlap, rather than simply make the auxiliary replica as hot as possible.

C EXPERIMENTAL DETAILS

C.1 SYNTHETIC BENCHMARKS

Targets. We discretize $[-2, 2]^2$ on a uniform 512×512 grid. The Gaussian benchmarks are equal-weight mixtures with isotropic standard deviation $\sigma = 0.02$ and centers

$$\mathcal{C}_{16}^{-} = \{-1.90\} \times \{-1.90, -1.50, -1.10, -0.70\} \cup \{-1.155151\} \times \{-1.84, -1.44, -1.04, -0.64\},$$
$$\mathcal{C}_{32}^{-} = \{-1.90\} \times \{-1.90, -1.70, \ldots, -0.50\} \cup \{-1.155151\} \times \{-1.84, -1.64, \ldots, -0.44\},$$

where $\mathcal{C}_m = \mathcal{C}_m^{-} \cup (-\mathcal{C}_m^{-})$. The third target is an equal-weight mixture of 16 isotropic bivariate Student-t components with $\nu = 2$, scale $s = 0.025$, and centers $\{-1.5, -0.5, 0.5, 1.5\}^2$.

Each chain runs for 10^6 transitions after a burn-in of 10^4. We report the mean $\pm$ standard deviation over the three evaluation seeds. All methods start from the same mode peak: mode 5 for 16-Gaussian and Student-t, and mode 10 for 32-Gaussian.

Hyperparameters. Each target-sampler pair is tuned separately. Replica-exchange methods use a common step size for both replicas and additionally tune the hot temperature and swap intensity, while ACS variants tune the cycle and shaped schedule. Candidates are screened for 12–20k steps and confirmed for 40–100k steps using multiscale MMD^2, mode-mass error, and core-category error. Selected settings are frozen before evaluation.

Metrics. For the empirical grid distribution $\widehat{p}$ and the exact target distribution p, let $\Delta = \widehat{p} - p$. We measure global sampling error using $\mathrm{MMD}_h^2 = \sum_{x,y} \Delta(x)\Delta(y) \exp\left(-\frac{\|x-y\|_2^2}{2h^2}\right)$, and report the average over bandwidths $h \in \{0.02, 0.04, 0.08, 0.16, 0.5, 1.0\}$. To quantify mode balance, let n_j denote the number of samples assigned to the nearest mixture center, $n = \sum_j n_j$, and $M \in \{16, 32\}$ the number of modes. We report the entropic mode coverage $\mathrm{EMC} = -\frac{1}{\log M} \sum_{j:n_j>0} \frac{n_j}{n} \log \frac{n_j}{n}$, where $\mathrm{EMC} = 1$ corresponds to perfectly balanced occupancy across all modes.

C.2 Sampling Tasks

Ising and Potts. We consider periodic 16×16 lattices with

$$\pi_{\mathrm{I}}(s) \propto \exp\left\{ J \sum_{(i,j)\in\mathcal{E}} s_i s_j + b \sum_i s_i \right\},$$

$$\pi_{\mathrm{P}}(x) \propto \exp\left\{ J \sum_{(i,j)\in\mathcal{E}} \mathbb{1}[x_i = x_j] + b \sum_i \mathbb{1}[x_i = 0] \right\},$$

where $\mathcal{E}$ denotes the nearest-neighbor edge set of the 16×16 toroidal lattice. Trajectory plots use $(J, b) = (0.3, 0.3)$ for Ising and $(0.8, 0.2)$ for Potts. With b fixed, coupling sweeps use $J \in \{0.1, 0.2, \ldots, 0.8\}$ for Ising and $J \in \{0.4, 0.6, 0.8, 0.9, 1.0, 1.1, 1.2, 1.4, 1.6\}$ for Potts. Reference one-site marginals are estimated from independently Swendsen–Wang ghost-cluster runs.

Each experiment uses 128 chains for 50000 updates. Cold chains are initialized in the phase opposite to the favored one, while hot replicas are initialized uniformly. We evaluate convergence using the natural logarithm of the RMSE between the running empirical one-site marginals and their reference values. For Ising models, the empirical marginals are the per-chain running spin means; for Potts models, they are the per-chain running category probabilities. Reference marginals are estimated from independent Swendsen–Wang ghost-cluster runs. Curves report the mean across seeds, with shaded regions indicating one standard deviation.

Restricted Boltzmann machines. We evaluate each method on pretrained 784–500 binary RBMs across MNIST, kMNIST, eMNIST, Omniglot, and Caltech, with the visible marginal given by

$$\pi(v) \propto \exp\left\{ b_v^\top v + \sum_{k=1}^{500} \mathrm{softplus}(W_k v + b_{h,k}) \right\}.$$

For each dataset, we generate a reference set of 1024 samples from 128 Block-Gibbs chains (10000 burn-in steps, thinned by 10), and split it into disjoint subsets of 500 states for tuning and evaluation. All samplers are run with 100 chains for 10000 updates. We report the performance of the cold chain using $\log \mathrm{MMD}^2$ under the kernel $k(v, v') = \exp\{-d_{\mathrm{H}}(v, v')/784\}$.

Comparison with Classical PT-DMALA across initializations. To isolate the contribution of the proposed local constructions within PT, we additionally compare against Classical PT-DMALA, which combines the original DMALA local kernel at each temperature with the standard Metropolis swap rule and is separately re-tuned on the same tuning split. Importantly, PT-A-DMALA and PT-AV-DMALA use the same classical form of PT; their distinction lies in replacing the original gradient-based DMALA proposal with the A/AV local constructions developed in Section 3. We also include DREAM as a recent alternative based on a history-dependent swap rule. We evaluate all methods under two cold-chain initializations: the highest-density reference state used in the main experiment and random initialization. Hot replicas are initialized uniformly over $\{0, 1\}^{784}$.

Initialization	Dataset	Classical PT-DMALA	DREAM	PT-A-DMALA	PT-AV-DMALA
Random	MNIST	-6.061 ± 0.106	$\underline{-6.146 \pm 0.103}$	$\mathbf{-6.362 \pm 0.151}$	-5.791 ± 0.268
	kMNIST	-5.644 ± 0.154	-5.651 ± 0.078	$\underline{-5.733 \pm 0.072}$	$\mathbf{-5.757 \pm 0.131}$
	eMNIST	-5.905 ± 0.168	$\underline{-5.935 \pm 0.152}$	$\mathbf{-5.977 \pm 0.188}$	-5.910 ± 0.150
	Omniglot	$\underline{-6.458 \pm 0.225}$	-6.357 ± 0.093	-6.435 ± 0.158	$\mathbf{-6.525 \pm 0.155}$
	Caltech	$\underline{-5.760 \pm 0.106}$	-5.729 ± 0.071	-5.754 ± 0.019	$\mathbf{-5.766 \pm 0.017}$
Highest-density	MNIST	-5.073 ± 0.398	-4.670 ± 0.226	$\underline{-6.007 \pm 0.515}$	$\mathbf{-6.194 \pm 0.034}$
	kMNIST	-5.737 ± 0.028	$\mathbf{-5.791 \pm 0.042}$	$\underline{-5.764 \pm 0.197}$	-5.674 ± 0.091
	eMNIST	$\underline{-5.912 \pm 0.140}$	-5.556 ± 0.397	$\mathbf{-5.983 \pm 0.268}$	-5.907 ± 0.292
	Omniglot	-6.411 ± 0.140	-6.221 ± 0.317	$\mathbf{-6.461 \pm 0.139}$	$\underline{-6.432 \pm 0.052}$
	Caltech	-5.707 ± 0.105	-5.729 ± 0.071	$\mathbf{-5.754 \pm 0.019}$	$\underline{-5.730 \pm 0.047}$

Table 5: RBM sampling under random and highest-density cold-chain initialization. Entries report the mean $\pm$ sample standard deviation of final $\log \mathrm{MMD}^2$ over three seeds. Bold indicates the best performance, and underline indicates the second best.

Table 5 shows that the advantage of the proposed methods persists across initialization schemes. In every dataset and initialization setting, at least one of PT-A-DMALA and PT-AV-DMALA improves upon Classical PT-DMALA, and a proposed method achieves the best final $\log \mathrm{MMD}^2$ in nine of the ten settings. The difference is particularly pronounced on MNIST under highest-density initialization: Classical PT-DMALA attains -5.073, compared with -6.007 for PT-A-DMALA and -6.194 for PT-AV-DMALA. Moreover, Classical PT-DMALA deteriorates markedly from its random-initialization value of -6.061, whereas the proposed constructions remain substantially more accurate. These results support the central design principle of correcting the local dynamics while retaining classical Metropolis swap: the observed gains do not require a specialized swap rule, but can instead be obtained by improving the local kernels coupled through standard PT.

Facility location. We construct five independent customer groups, each with three nearly redundant facilities and 24 customers, giving $d = 15$ and $m = 120$. Within-group utilities receive 3% perturbations, cross-group utilities are zero, and $\lambda = 12$, yielding $3^5 = 243$ dominant modes. Exact marginals are obtained by enumerating all 2^{15} states. Each run uses 128 chains for 15000 updates, with 3000 burn-in steps and thinning by 5. Chains start with facility 0 or 2 selected in every group; the two starts are averaged within each seed, and results are reported over three seeds. Marginal RMSE uses the 15 inclusion probabilities. ESS/s is the minimum Geyer ESS over energy, cardinality, and coordinate traces divided by wall time; for unadjusted methods, it is interpreted only as a movement diagnostic. Hyperparameters are selected on disjoint validation seeds.

C.3 Learning Tasks

RBM learning. We train 784–500 binary RBMs by PCD for 2000 Adam updates with 100 persistent particles. At each update, single-chain methods perform 80 local sampling rounds, while two-replica methods perform 40 rounds per replica. Hyperparameters are tuned on separate validation runs, and we report mean $\pm$ standard deviation of held-out NLL over three training seeds, estimated by AIS. All datasets use held-out evaluation, with standard binarization and dataset-specific train/test splits.

Multi-planted MaxSAT learning. We generate weighted 3-SAT targets with $d = 18$, $R = 8$ planted assignments, minimum pairwise Hamming distance $0.45d$, and $m = 6d = 108$ clauses. The teacher and student distributions are

$$p_b(x) \propto \exp\left\{ b^\top (2x - \mathbf{1}) - \lambda \sum_{k=1}^{108} w_k \mathbb{1}[C_k(x) = \mathrm{False}] \right\}, \qquad \lambda \in \{6, 8, 10, 12, 14\}.$$

For each formula, exact enumeration of all 2^{18} states yields the partition function, model moments, and basin masses. We draw 10000, 2000, and 5000 teacher samples for training, validation, and testing, respectively. The student bias b is trained by PCD for 2000 Adam updates with learning rate 10^{-3}, batch size 256, 256 persistent particles, and 20 sampler rounds per update. Under random initialization, both replicas are drawn uniformly; under single-basin initialization, both start from the same fixed planted assignment.

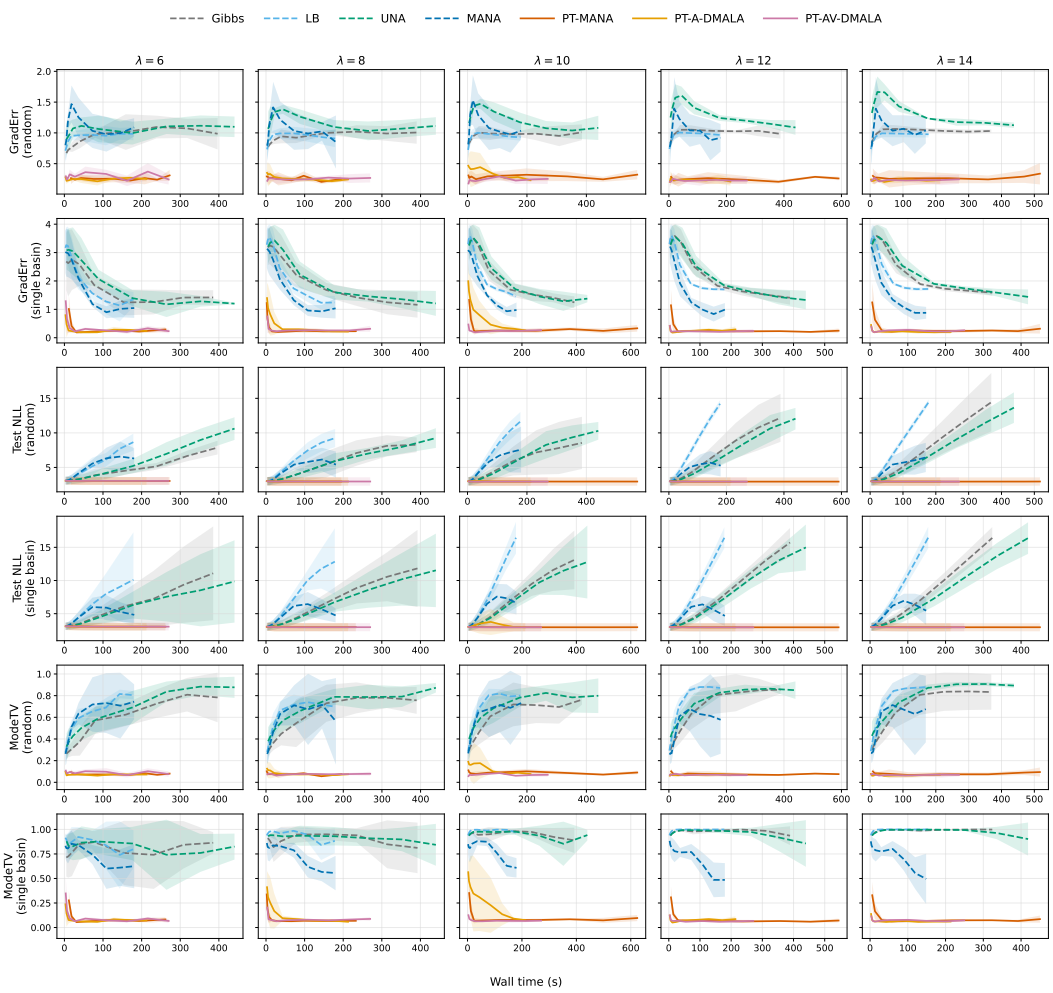


Figure 4: MaxSAT EBM learning curves versus wall-clock time across penalties $\lambda \in \{6, 8, 10, 12, 14\}$. Rows display negative-phase gradient error (rows 1 and 2), test NLL (rows 3 and 4), and ModeTV (rows 5 and 6), evaluated under random (odd rows) and single-basin (even rows) initializations. Curves and shaded regions report the mean and one standard deviation across three formulas with three seeds each.

Test NLL is computed using the exact partition function. GradErr measures the relative ℓ_2 error between the particle-based and exact model moments used in the likelihood gradient. ModeTV measures the discrepancy between the sampled and exact probability masses over the eight planted-solution basins. Figures 4 and 5 show the corresponding learning trajectories against wall-clock time and iterations. At larger penalties, several non-PT baselines incur substantial errors, whereas PT-A-DMALA and PT-AV-DMALA maintain low gradient error and ModeTV under both initialization schemes.

D Proofs and Supporting Results

D.1 Preliminaries

This section collects notation and conventions used throughout the proofs.

Markov processes and spectral gaps. Let Q denote the infinitesimal generator of an irreducible, continuous-time Markov chain on a finite state space $\mathcal{X}$, with transition semigroup $P_t = e^{tQ}$ and unique invariant distribution μ, satisfying $\mu Q = 0$ (or equivalently, $\mu P_t = \mu$ for all $t \geq 0$). The

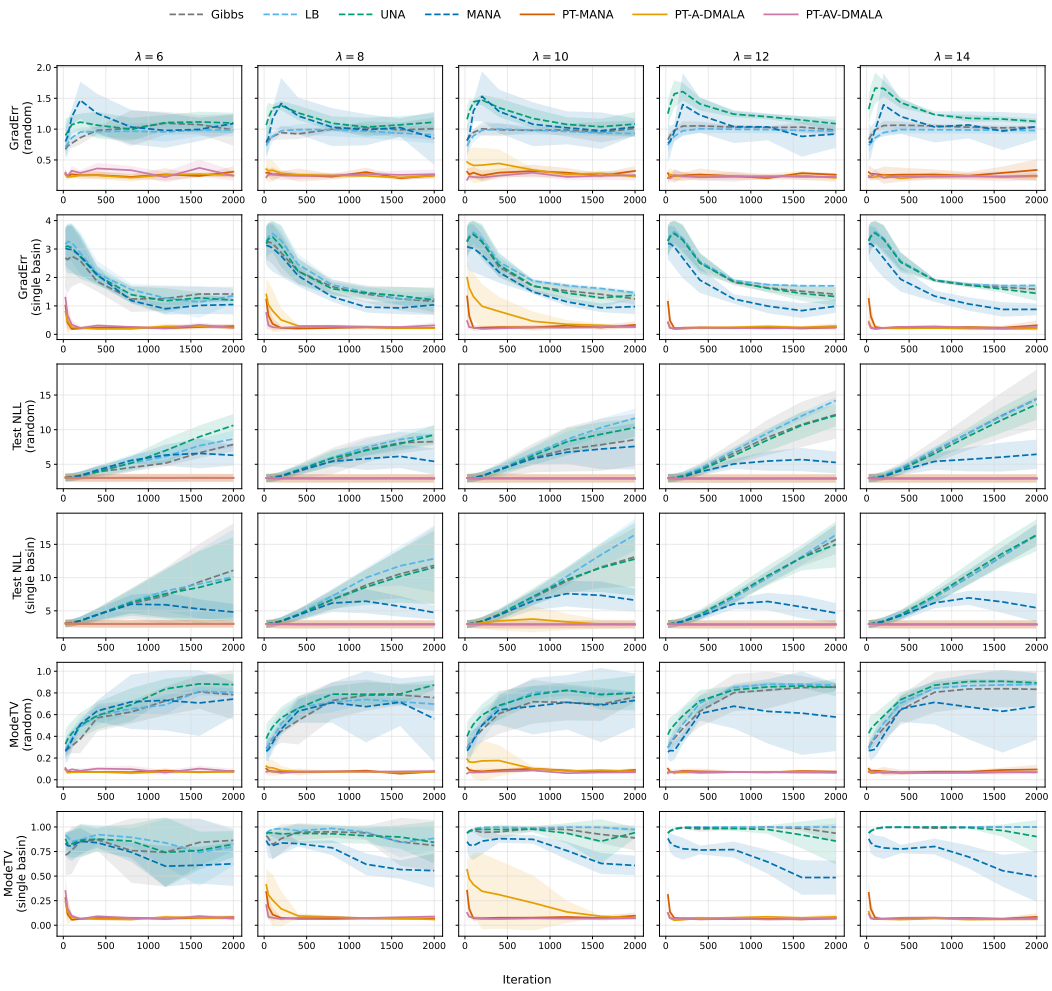


Figure 5: MaxSAT EBM trajectories plotted against PCD updates.

generator is reversible with respect to μ if it satisfies the detailed balance condition: $\mu(x)Q(x,y) = \mu(y)Q(y,x)$ for all $x, y \in \mathcal{X}$. For observables $f, g : \mathcal{X} \to \mathbb{R}$, we define

$$\mu(f) := \sum_{x\in\mathcal{X}} \mu(x)f(x), \qquad \langle f, g\rangle_\mu := \sum_{x\in\mathcal{X}} \mu(x)f(x)g(x), \qquad L_0^2(\mu) := \big\{f : \mu(f) = 0\big\},$$

and the variance $\mathrm{Var}_\mu(f) := \|f - \mu(f)\|^2_{L^2(\mu)}$. For a μ-reversible generator Q, we associate the Dirichlet form $\mathcal{E}_Q(f,g) := -\langle f, Qg\rangle_\mu$ and write $\mathcal{E}_Q(f) := \mathcal{E}_Q(f,f)$, which admits the explicit representation

$$\mathcal{E}_Q(f) = \frac{1}{2}\sum_{x,y\in\mathcal{X}} \mu(x)Q(x,y)\big[f(y) - f(x)\big]^2.$$

The spectral gap of Q is defined variationally as

$$\lambda(Q) := \inf_{\mathrm{Var}_\mu(f)>0} \frac{\mathcal{E}_Q(f)}{\mathrm{Var}_\mu(f)}.$$

Under irreducibility, μ is strictly positive on $\mathcal{X}$ and $\lambda(Q) > 0$. Consequently, the Poincaré inequality and semigroup L^2-decay hold:

$$\mathrm{Var}_\mu(f) \le \lambda(Q)^{-1}\mathcal{E}_Q(f), \qquad \|e^{tQ}f\|_{L^2(\mu)} \le e^{-\lambda(Q)t}\|f\|_{L^2(\mu)}, \quad \forall f \in L_0^2(\mu).$$

Norms and distances. For a signed row vector (measure) ξ and a matrix H, we define the standard norms

$$\|\xi\|_1 := \sum_x |\xi(x)|, \qquad \|H\|_\infty := \max_x \sum_y |H(x,y)|, \qquad \|\nu - \mu\|_{\mathrm{TV}} := \frac{1}{2}\|\nu - \mu\|_1.$$

For a strictly positive distribution μ, the chi-square divergence is defined by $\chi^2(\nu \mid \mu) := \sum_x [\nu(x) - \mu(x)]^2/\mu(x)$. The matrix infinity norm is submultiplicative, satisfying $\|HG\|_\infty \le \|H\|_\infty \|G\|_\infty$ and $\|\xi H\|_1 \le \|\xi\|_1 \|H\|_\infty$, with $\|P\|_\infty = 1$ for any transition probability matrix P.

We denote by $H_h = \mathcal{O}_\infty(h^r)$ if $\|H_h\|_\infty \le Ch^r$ for all sufficiently small $h > 0$. Unless stated otherwise, such constants C are independent of the step size h, but are not assumed to be uniform with respect to the dimension, state-space cardinality, target distribution, temperature, or underlying graph topology.

Poisson equation and asymptotic variance. Suppose Q is irreducible and μ-reversible. Then $-Q$, restricted to the zero-mean subspace $L_0^2(\mu)$, is self-adjoint, strictly positive definite, and invertible. For any observable f, letting $f_0 := f - \mu(f)$, the Poisson equation $-Qu = f_0$ has a unique centered solution $u = (-Q)^{-1} f_0 \in L_0^2(\mu)$. For the process $(X_t)_{t \ge 0}$ generated by Q with initial law $X_0 \sim \mu$, the asymptotic variance is given by

$$\mathrm{AVar}_Q(f) := \lim_{T \to \infty} \frac{1}{T} \mathrm{Var}_\mu \left(\int_0^T f(X_t)\, dt \right) = 2 \big\langle f_0, (-Q)^{-1} f_0 \big\rangle_\mu.$$

Throughout, all generator inverses are understood in the operator sense restricted to the mean-zero subspace $L_0^2(\mu)$.

Small-step conventions. For the discrete Langevin dynamics under consideration, the product state space is $\mathcal{X} = \prod_{i=1}^d \mathcal{X}_i$ (finite), where each coordinate graph is connected and undirected. We introduce the shorthand $h := h_\eta = e^{-D/\eta}$ and define the set of non-neighbor coordinates as $\mathcal{R}_i(u) := \mathcal{X}_i \setminus \big(\mathcal{N}_i(u) \cup \{u\}\big)$. We retain the exponent $\gamma \in (0, 1]$ specified in equation 4, which guarantees that $e^{-\ell_i(x_i,b)/\eta} \le h^{1+\gamma}$ for all $b \in \mathcal{R}_i(x_i)$.

Within a factorized proposal stage: (1) single-coordinate transitions along graph edges occur with probability $O(h)$; (2) single-coordinate transitions to non-neighboring states occur with probability $O(h^{1+\gamma})$; and (3) transitions modifying two or more coordinates simultaneously occur with probability $O(h^2)$. The score functions involved in the small-step analysis are finite on $\mathcal{X}$ and independent of the parameter η. Unless specified otherwise, all asymptotic statements refer to the limiting regime $h = h_\eta \downarrow 0$ while holding the target distribution, temperature, state space, coordinate graphs, score functions, and swap intensities fixed. Finally, $\mathrm{M} \in \{\mathrm{A}, \mathrm{AV}\}$ denotes the one-stage and auxiliary-variable schemes, respectively.

D.2 Proofs Omitted in Section 2

We show that replica exchange alone cannot debias unadjusted discrete local dynamics, establishing two key limitations. First, a log-quadratic counterexample demonstrates that the historical swap rule equation 3 violates detailed balance for both the target and intermediate invariant distributions. Second, we prove an impossibility theorem: no post-proposal rule restricted to retaining or swapping proposed states can preserve the product target.

A counterexample for the historical swap rule. Consider the two-state space $\mathcal{X} = \{0, 1\}$ equipped with the potential $U(x) = x$, corresponding to a log-quadratic potential with vanishing second-order differences. Set the step size $\eta = 1$ and temperatures $\tau_1 = 1$, $\tau_2 = 2$. For replica $k \in \{1, 2\}$, the unadjusted local proposal kernel is given by

$$q_k(y \mid x) = \frac{1}{Z_k(x)} \exp\left\{ -\frac{1}{2}(y - x)^2 + \frac{y - x}{2\tau_k} \right\}.$$

The corresponding normalizers are $Z_1(0) = 2$, $Z_1(1) = 1 + e^{-1}$, $Z_2(0) = 1 + e^{-1/4}$, and $Z_2(1) = 1 + e^{-3/4}$, yielding the explicit transition probabilities

$$q_1(0 \mid 0) = \frac{1}{2}, \quad q_1(0 \mid 1) = \frac{1}{1 + e}, \quad q_2(0 \mid 0) = \frac{1}{1 + e^{-1/4}}, \quad q_2(0 \mid 1) = \frac{1}{1 + e^{3/4}}.$$

The intended tempered target distribution is $\pi_{\tau_k}(x) \propto e^{x/\tau_k}$. By contrast, the finite-step local kernel q_k is reversible with respect to the intermediate law

$$\widetilde{\pi}_k(x) = C_k^{-1} e^{x/\tau_k} Z_k(x), \qquad C_k := \sum_{v \in \mathcal{X}} e^{v/\tau_k} Z_k(v),$$

which follows directly from the symmetry of the reversibility measure:

$$C_k \widetilde{\pi}_k(x) q_k(y \mid x) = \exp\left\{ -\frac{1}{2}(y-x)^2 + \frac{x+y}{2\tau_k} \right\} = C_k \widetilde{\pi}_k(y) q_k(x \mid y).$$

Let $\Pi := \pi_{\tau_1} \otimes \pi_{\tau_2}$ and $\widetilde{\Pi} := \widetilde{\pi}_1 \otimes \widetilde{\pi}_2$ denote the product targets. Under the historical swap rule equation 3, the overall transition probability—combining independent local proposals with the retain-or-swap decision—is given by

$$\begin{aligned} P(y_1, y_2 \mid x_1, x_2) = {} & [1 - \widetilde{S}(y_1, y_2 \mid x_1, x_2)]\, q_1(y_1 \mid x_1) q_2(y_2 \mid x_2) \\ & + \widetilde{S}(y_2, y_1 \mid x_1, x_2)\, q_1(y_2 \mid x_1) q_2(y_1 \mid x_2). \end{aligned}$$

Evaluating this rule at states $(0,1)$ and $(1,0)$ yields

$$\widetilde{S}(1,0 \mid 0,1) = \widetilde{S}(0,1 \mid 0,1) = \widetilde{S}(0,1 \mid 1,0) = 1, \qquad \widetilde{S}(1,0 \mid 1,0) = e^{-1},$$

which implies

$$P(1,0 \mid 0,1) = \frac{1}{2(1+e^{-3/4})}, \qquad P(0,1 \mid 1,0) = \frac{e^{-1}}{(1+e^{-1})(1+e^{-1/4})}.$$

Writing $\widetilde{C} = C_1 C_2$, the forward and backward probability fluxes under $\widetilde{\Pi}$ satisfy

$$\widetilde{C}\, \widetilde{\Pi}(0,1) P(1,0 \mid 0,1) = e^{1/2} \neq 1 = \widetilde{C}\, \widetilde{\Pi}(1,0) P(0,1 \mid 1,0).$$

Hence, detailed balance is violated with respect to the intermediate product distribution $\widetilde{\Pi}$.

Detailed balance also fails for the intended target Π. Setting $r = e^{1/4} > 1$ and $C = (1+e)(1+e^{1/2})$, we have

$$C\, \Pi(0,1) P(1,0 \mid 0,1) = \frac{r^5}{2(1+r^3)} > \frac{r^5}{(1+r^4)(1+r)} = C\, \Pi(1,0) P(0,1 \mid 1,0),$$

where the strict inequality follows from

$$(1+r^4)(1+r) - 2(1+r^3) = (r-1)(r^4 + 2r^3 + 1) > 0.$$

Therefore, the historical-swap kernel is reversible with respect to neither $\widetilde{\Pi}$ nor Π.

A structural limitation of swap-only correction. We next show that this pathology is not an artifact of the specific acceptance probability in equation 3, but represents an intrinsic structural limitation. Suppose local proposals (Y_1, Y_2) are generated independently according to $Y_1 \sim q_1(\cdot \mid X_1)$ and $Y_2 \sim q_2(\cdot \mid X_2)$. Subsequently, an arbitrary selection rule (which may be randomized, state-dependent, or history-dependent) selects the next state X_{new} from the candidate pair

$$\{(Y_1, Y_2),\ (Y_2, Y_1)\}.$$

For any symmetric observable F satisfying $F(x_1, x_2) = F(x_2, x_1)$, we have the pathwise identity

$$F(X_{\mathrm{new}}) = F(Y_1, Y_2),$$

which holds regardless of the specific exchange mechanism. Consequently, if this process admits Π as a stationary distribution, the stationarity condition requires that

$$\mathbb{E}_\Pi[F(X_1, X_2)] = \mathbb{E}_\Pi\big[\mathbb{E}[F(Y_1, Y_2) \mid X_1, X_2]\big] \tag{10}$$

for every symmetric test function F.

Now consider the symmetric indicator $F(x_1, x_2) = \mathbf{1}_{\{x_1 = x_2 = 0\}}$, and let $r = e^{1/4} > 1$. Under the target distribution Π, the marginal probabilities are

$$\pi_{\tau_1}(0) = \frac{1}{1+r^4}, \qquad \pi_{\tau_2}(0) = \frac{1}{1+r^2}.$$

On the other hand, the marginal distribution after one unadjusted local proposal step satisfies

$$(\pi_{\tau_1} q_1)(0) = \frac{1+3r^4}{2(1+r^4)^2}, \qquad (\pi_{\tau_2} q_2)(0) = \frac{r}{1+r^3}.$$

Crucially, both probabilities undergo a strict increase:

$$(\pi_{\tau_1} q_1)(0) - \pi_{\tau_1}(0) = \frac{r^4-1}{2(1+r^4)^2} > 0, \qquad (\pi_{\tau_2} q_2)(0) - \pi_{\tau_2}(0) = \frac{r-1}{(1+r^3)(1+r^2)} > 0.$$

By the conditional independence of the proposals given (X_1, X_2), we obtain

$$\mathbb{P}(Y_1 = Y_2 = 0) = (\pi_{\tau_1} q_1)(0)\,(\pi_{\tau_2} q_2)(0) > \pi_{\tau_1}(0)\,\pi_{\tau_2}(0) = \Pi(0,0),$$

which directly contradicts the stationarity identity equation 10.

We conclude that for these local kernels, no post-proposal mechanism whose action is confined to selecting between (Y_1, Y_2) and (Y_2, Y_1) can preserve Π as its stationary marginal, even if the decision depends on the current states or the historical trajectory. Thus, the asymptotic bias of discrete local dynamics cannot, in general, be eliminated through replica exchange alone.

Remark 8. *In the log-quadratic setting, unadjusted discrete Langevin sampling (DLS) with finite step size is reversible with respect to the modified density $\widetilde{\pi}_\eta^{(\tau)}(x) \propto \pi_\tau(x) Z_\eta^{(\tau)}(x)$, where*

$$Z_\eta^{(\tau)}(x) = \prod_{i=1}^{d} \sum_{b \in \mathcal{X}_i} \exp\left\{ \frac{s_i(x)(b-x_i)}{2\tau} - \frac{(b-x_i)^2}{2\eta} \right\}.$$

An idealized Metropolis exchange targeting $\widetilde{\pi}_\eta^{(\tau_1)} \otimes \widetilde{\pi}_\eta^{(\tau_2)}$ would therefore require the acceptance ratio

$$\frac{\widetilde{\pi}_\eta^{(\tau_1)}(x^{(2)})\,\widetilde{\pi}_\eta^{(\tau_2)}(x^{(1)})}{\widetilde{\pi}_\eta^{(\tau_1)}(x^{(1)})\,\widetilde{\pi}_\eta^{(\tau_2)}(x^{(2)})} = \underbrace{\frac{\pi_{\tau_1}(x^{(2)})\,\pi_{\tau_2}(x^{(1)})}{\pi_{\tau_1}(x^{(1)})\,\pi_{\tau_2}(x^{(2)})}}_{\textit{standard Metropolis ratio}} \times \underbrace{\frac{Z_\eta^{(\tau_1)}(x^{(2)})\,Z_\eta^{(\tau_2)}(x^{(1)})}{Z_\eta^{(\tau_1)}(x^{(1)})\,Z_\eta^{(\tau_2)}(x^{(2)})}}_{\textit{ratio of local normalizers}}.$$

The historical rule equation 3 omits this second ratio of local partition functions, which causes detailed balance to fail even in simple log-quadratic models. Furthermore, this product-form intermediate invariant measure is an exceptional artifact of the log-quadratic potential and does not generalize to arbitrary discrete targets.

D.3 Proofs Omitted in Section 3

We first establish small-step generator expansions for both the one-stage and auxiliary-variable kernels, and characterize the reversibility of their continuous-time limiting generators. Leveraging these expansions, we then establish finite-step error bounds, analyze their MH corrections, and compare the resulting limiting processes.

D.3.1 Continuous-Time Limits and Edge Balance

Throughout this subsection, we fix a temperature $\tau > 0$. We first establish the small-step asymptotic expansion of the one-stage transition kernel, which holds without requiring any edge-balance condition.

Lemma 9. *As $h \downarrow 0$, the one-stage kernel satisfies*

$$T_{\eta,\sigma}^{(\tau)} = I + hQ_\sigma^{(\tau)} + \mathcal{O}(h^{1+\gamma}),$$

where the remainder is bounded with respect to the matrix ℓ_∞-operator norm.

Proof. Fix $x \in \mathcal{X}$. For each coordinate i, the normalizing constant in equation 5 expands uniformly in x as

$$\sum_{b \in \mathcal{X}_i} \exp\left\{ \frac{1}{2}\sigma_i^{(\tau)}(x,b) - \frac{\ell_i(x_i,b)}{\eta} \right\} = 1 + h \sum_{b \in \mathcal{N}_i(x_i)} e^{\sigma_i^{(\tau)}(x,b)/2} + \mathcal{O}(h^{1+\gamma}).$$

Indeed, the diagonal term $b = x_i$ contributes 1, neighboring states $b \in \mathcal{N}_i(x_i)$ contribute at order $\mathcal{O}(h)$, and non-neighboring states contribute $\mathcal{O}(h^{1+\gamma})$. Expanding the reciprocal $(1+u)^{-1} = 1 - u + \mathcal{O}(u^2)$ then yields

$$T^{(\tau)}_{\eta,\sigma,i}(x_i \mid x) = 1 - h \sum_{b \in \mathcal{N}_i(x_i)} e^{\sigma_i^{(\tau)}(x,b)/2} + \mathcal{O}(h^{1+\gamma}), \quad T^{(\tau)}_{\eta,\sigma,i}(b \mid x) = h e^{\sigma_i^{(\tau)}(x,b)/2} + \mathcal{O}(h^{1+\gamma}),$$

for $b \in \mathcal{N}_i(x_i)$, while transitions to non-neighboring states are $\mathcal{O}(h^{1+\gamma})$. By the coordinate-wise factorization $T^{(\tau)}_{\eta,\sigma}(y \mid x) = \prod_i T^{(\tau)}_{\eta,\sigma,i}(y_i \mid x)$, we have

$$T^{(\tau)}_{\eta,\sigma}(y \mid x) = \begin{cases} 1 + hQ^{(\tau)}_{\sigma}(x,x) + \mathcal{O}(h^{1+\gamma}), & y = x, \\ hQ^{(\tau)}_{\sigma}(x,y) + \mathcal{O}(h^{1+\gamma}), & y = x^{i \to b},\ b \in \mathcal{N}_i(x_i), \\ \mathcal{O}(h^{1+\gamma}), & \text{otherwise.} \end{cases}$$

Any remaining transition requires either a non-neighboring move within a single coordinate (bounded by $\mathcal{O}(h^{1+\gamma})$) or coordinate changes in at least two coordinates (which contribute $\mathcal{O}(h^2) \subseteq \mathcal{O}(h^{1+\gamma})$ since $\gamma \leq 1$). Because $\mathcal{X}$ is finite, these bounds hold uniformly across all $x, y \in \mathcal{X}$, establishing the stated matrix ℓ_∞-operator norm bound. □

Proof of Theorem 1. The asymptotic expansion of the kernel is guaranteed by Lemma 9. For any graph edge $y = x^{i \to b}$, the transition rates are given by

$$Q^{(\tau)}_{\sigma}(x,y) = e^{\sigma_i^{(\tau)}(x,b)/2}, \qquad Q^{(\tau)}_{\sigma}(y,x) = e^{\sigma_i^{(\tau)}(y,x_i)/2}.$$

Therefore, detailed balance along this edge is equivalent to

$$\frac{\pi_\tau(x) Q^{(\tau)}_{\sigma}(x,y)}{\pi_\tau(y) Q^{(\tau)}_{\sigma}(y,x)} = \exp\left\{ \frac{\sigma_i^{(\tau)}(x,b) - \sigma_i^{(\tau)}(y,x_i)}{2} - \Delta_{i,\tau}(x,b) \right\} = 1,$$

which coincides with equation 6. Since transition rates vanish in both directions for any non-adjacent pair (x, y), this edge-wise balance condition is both necessary and sufficient for π_τ-reversibility. □

Specialization to A-DULA. Under the adjusted score defined in equation 7, every graph edge $y = x^{i \to b}$ satisfies

$$Q_{\mathrm{A},\tau}(x,y) = \frac{\pi_\tau(y)}{\pi_\tau(x) + \pi_\tau(y)}, \qquad \pi_\tau(x) Q_{\mathrm{A},\tau}(x,y) = \underbrace{\frac{\pi_\tau(x)\pi_\tau(y)}{\pi_\tau(x) + \pi_\tau(y)}}_{\text{symmetric edge flux}}.$$

Hence, $Q_{\mathrm{A},\tau}$ is π_τ-reversible. Furthermore, the strict positivity of the edge transition rates together with the connectedness of the coordinate graphs implies irreducibility.

We next turn to the auxiliary-variable (AV) construction.

Proof of Proposition 2. An AV update is a two-stage transition: it first samples an intermediate state $z \sim A_\eta(\cdot \mid x)$ and subsequently draws $y \sim T^{(\tau)}_{\eta,\sigma}(\cdot \mid z)$. For an edge transition $y = x^{i \to b}$, first-order contributions in h arise from exactly two disjoint paths:

(i) the perturbation stage stays at x and the correction stage transitions to y, contributing $h e^{\sigma_i^{(\tau)}(x,b)/2} + \mathcal{O}(h^{1+\gamma})$;

(ii) the perturbation stage transitions to y and the correction stage remains at y, contributing $h + \mathcal{O}(h^{1+\gamma})$.

Transitions involving off-diagonal moves in both stages are $\mathcal{O}(h^2)$, whereas transitions involving any non-neighboring move are $\mathcal{O}(h^{1+\gamma})$. Consequently,

$$T^{(\tau)}_{\eta,\sigma,\mathrm{AV}} = I + hQ^{(\tau)}_{\sigma,\mathrm{AV}} + \mathcal{O}(h^{1+\gamma}), \qquad Q^{(\tau)}_{\sigma,\mathrm{AV}}(x,y) = 1 + e^{\sigma_i^{(\tau)}(x,b)/2}$$

for every graph edge $y = x^{i\to b}$, with the remainder bounded in the matrix ℓ_∞-operator norm.

Detailed balance along this edge is equivalent to

$$\frac{1 + e^{\sigma_i^{(\tau)}(x,b)/2}}{1 + e^{\sigma_i^{(\tau)}(y,x_i)/2}} = \frac{\pi_\tau(y)}{\pi_\tau(x)} = e^{\Delta_{i,\tau}(x,b)},$$

which is precisely equation 8. Since non-edge transition rates vanish symmetrically, this condition is both necessary and sufficient.

In particular, for $\sigma^{(\tau)}_{\mathrm{AV},i}(x,b) = 2\Delta_{i,\tau}(x,b)$, the off-diagonal generator entries satisfy

$$Q_{\mathrm{AV},\tau}(x,y) = 1 + \frac{\pi_\tau(y)}{\pi_\tau(x)}, \qquad \pi_\tau(x)Q_{\mathrm{AV},\tau}(x,y) = \pi_\tau(x) + \pi_\tau(y) = \pi_\tau(y)Q_{\mathrm{AV},\tau}(y,x).$$

Thus, $Q_{\mathrm{AV},\tau}$ is π_τ-reversible. Strict positivity of the edge rates and connectedness of the coordinate graphs similarly guarantee irreducibility. □

Unified local truncation bound. Because the state space $\mathcal{X}$ is finite, the asymptotic expansions established above hold uniformly. Therefore, for each mechanism $\mathrm{M} \in \{\mathrm{A}, \mathrm{AV}\}$ and fixed $\tau > 0$, there exists a constant $c_{\mathrm{M},\tau} < \infty$, independent of h, such that for all sufficiently small $h > 0$,

$$\left\| T^{(\tau)}_{\eta,\mathrm{M}} - I - hQ_{\mathrm{M},\tau} \right\|_\infty \le c_{\mathrm{M},\tau} h^{1+\gamma}. \tag{11}$$

D.3.2 Finite-Step Convergence and Stationary Bias

We now translate the one-step generator approximation into an error bound that is uniform in the number of discrete steps. For any generator Q on a finite state space, we have $e^{hQ} - I - hQ = \int_0^h (h-s)Q^2 e^{sQ}\,\mathrm{d}s$. Since e^{sQ} is a Markov transition kernel, its induced ℓ_∞-norm satisfies $\|e^{sQ}\|_\infty = 1$. Consequently,

$$\left\| e^{hQ} - I - hQ \right\|_\infty \le \frac{1}{2}\|Q\|_\infty^2 h^2. \tag{12}$$

Lemma 10. *Let Q be an irreducible, reversible generator on a finite state space $\mathcal{X}$, with invariant distribution μ and spectral gap $\lambda > 0$. Suppose that a Markov kernel P_h satisfies, for all sufficiently small h,*

$$\left\| P_h - e^{hQ} \right\|_\infty \le c h^{1+\gamma}$$

for some constants $c < \infty$ and $\gamma \in (0, 1]$. Then, for every initial distribution ν and any integer $n \ge 0$,

$$\left\| \nu P_h^n - \nu e^{nhQ} \right\|_{\mathrm{TV}} \le \frac{c h^{1+\gamma}}{2\sqrt{\min_{x\in\mathcal{X}} \mu(x)}} \frac{1 - e^{-\lambda n h}}{1 - e^{-\lambda h}}. \tag{13}$$

If, in addition, $\lambda h \le 1$, then

$$\|\nu P_h^n - \mu\|_{\mathrm{TV}} \le \frac{1}{2} e^{-\lambda n h} \sqrt{\chi^2(\nu \mid \mu)} + \frac{c}{\lambda \sqrt{\min_{x\in\mathcal{X}} \mu(x)}} h^\gamma \left(1 - e^{-\lambda n h}\right). \tag{14}$$

Proof. Let ξ be a signed measure satisfying $\xi \mathbf{1} = 0$, and define the density $g := \xi/\mu$. Then $g \in L_0^2(\mu)$ and, by reversibility, $(\xi e^{tQ})/\mu = e^{tQ} g$. The spectral gap contraction in $L^2(\mu)$ together with Cauchy–Schwarz yields

$$\left\| \xi e^{tQ} \right\|_1 \le \frac{e^{-\lambda t}}{\sqrt{\min_{x\in\mathcal{X}} \mu(x)}} \|\xi\|_1. \tag{15}$$

Applying the same contraction to the signed measure $\xi = \nu - \mu$ gives

$$\left\| \nu e^{tQ} - \mu \right\|_{\mathrm{TV}} \le \frac{1}{2} e^{-\lambda t} \sqrt{\chi^2(\nu \mid \mu)}. \tag{16}$$

Now, set $\Delta_h := P_h - e^{hQ}$. Because both P_h and e^{hQ} are Markov kernels, we have $\Delta_h \mathbf{1} = 0$. Hence, $\nu P_h^j \Delta_h$ has zero total mass, and

$$\|\nu P_h^j \Delta_h\|_1 \le \|\nu P_h^j\|_1 \|\Delta_h\|_\infty \le \|\Delta_h\|_\infty \le c h^{1+\gamma}.$$

For $n \geq 1$, telescoping yields the representation

$$\nu(P_h^n - e^{nhQ}) = \sum_{j=0}^{n-1} \nu P_h^j \Delta_h e^{(n-1-j)hQ}.$$

Applying equation 15 term by term along with $\|\cdot\|_{\mathrm{TV}} = \frac{1}{2}\|\cdot\|_1$, we find

$$\left\|\nu(P_h^n - e^{nhQ})\right\|_{\mathrm{TV}} \leq \frac{ch^{1+\gamma}}{2\sqrt{\min_{x\in\mathcal{X}} \mu(x)}} \sum_{j=0}^{n-1} e^{-\lambda(n-1-j)h} = \frac{ch^{1+\gamma}}{2\sqrt{\min_{x\in\mathcal{X}} \mu(x)}} \frac{1-e^{-\lambda nh}}{1-e^{-\lambda h}},$$

which establishes equation 13. The geometrically decaying weights prevent linear accumulation of the local discretization error. When $\lambda h \leq 1$, the elementary inequality $1 - e^{-\lambda h} \geq \frac{1}{2}\lambda h$ holds. Combining equation 13 with equation 16 via the triangle inequality yields equation 14. The case $n = 0$ holds trivially. □

Proof of Theorem 3. Fix $\mathrm{M} \in \{\mathrm{A}, \mathrm{AV}\}$ and $\tau > 0$. By equation 11 and equation 12, together with the fact that $h^2 \leq h^{1+\gamma}$ for $h \in (0, 1]$ (as $\gamma \in (0, 1]$), we have for all sufficiently small h:

$$\begin{aligned}\left\|T_{\eta,\mathrm{M}}^{(\tau)} - e^{hQ_{\mathrm{M},\tau}}\right\|_\infty &\leq \left\|T_{\eta,\mathrm{M}}^{(\tau)} - I - hQ_{\mathrm{M},\tau}\right\|_\infty + \left\|e^{hQ_{\mathrm{M},\tau}} - I - hQ_{\mathrm{M},\tau}\right\|_\infty \\ &\leq \left(c_{\mathrm{M},\tau} + \frac{1}{2}\|Q_{\mathrm{M},\tau}\|_\infty^2\right) h^{1+\gamma}.\end{aligned}$$

For sufficiently small h, the condition $\lambda_{\mathrm{M},\tau} h \leq 1$ is satisfied. Applying Lemma 10 gives

$$\left\|\nu\big(T_{\eta,\mathrm{M}}^{(\tau)}\big)^n - \pi_\tau\right\|_{\mathrm{TV}} \leq \underbrace{\frac{1}{2} e^{-\lambda_{\mathrm{M},\tau} nh} \sqrt{\chi^2(\nu \mid \pi_\tau)}}_{\text{continuous-time mixing}} + \underbrace{\frac{c_{\mathrm{M},\tau} + \frac{1}{2}\|Q_{\mathrm{M},\tau}\|_\infty^2}{\lambda_{\mathrm{M},\tau}\sqrt{\min_{x\in\mathcal{X}} \pi_\tau(x)}} h^\gamma \left(1 - e^{-\lambda_{\mathrm{M},\tau} nh}\right)}_{\text{discretization error}}. \tag{17}$$

Thus, Theorem 3 holds with prefactor

$$C_{\mathrm{M},\tau} := \frac{c_{\mathrm{M},\tau} + \frac{1}{2}\|Q_{\mathrm{M},\tau}\|_\infty^2}{\lambda_{\mathrm{M},\tau}\sqrt{\min_{x\in\mathcal{X}} \pi_\tau(x)}}.$$

It remains to bound the stationary bias. For either construction, the transition kernel $T_{\eta,\mathrm{M}}^{(\tau)}$ possesses strictly positive diagonal entries (ensuring aperiodicity) and strictly positive transition probabilities along all coordinate edges (ensuring irreducibility, since the coordinate graph is connected). Therefore, it admits a unique invariant distribution $\widehat{\pi}_{\eta,\mathrm{M},\tau}$.

Taking $n \to \infty$ in equation 17, we obtain

$$\|\widehat{\pi}_{\eta,\mathrm{M},\tau} - \pi_\tau\|_{\mathrm{TV}} \leq C_{\mathrm{M},\tau} h^\gamma.$$

The prefactor $C_{\mathrm{M},\tau}$ is independent of the step size h, step index n, and initial distribution ν, though it may depend on the target distribution, temperature, state space geometry, and local dynamics. □

D.3.3 MH Corrections

The MH corrections ensure exact π_τ-reversibility at finite step sizes while preserving the limiting generators of the A and AV schemes as $h \to 0$.

A-DMALA. Let $T := T_{\eta,\sigma}^{(\tau)}$ denote the proposal kernel and K its standard MH-adjusted transition kernel. For any distinct states $x \neq y$, the accepted probability flux satisfies

$$\pi_\tau(x)K(x, y) = \min\{\pi_\tau(x)T(y \mid x),\, \pi_\tau(y)T(x \mid y)\} = \pi_\tau(y)K(y, x).$$

Consequently, K satisfies detailed balance with respect to π_τ, establishing its exact invariance.

AV-DMALA. For a fixed auxiliary state z, the conditional MH update targets the distribution proportional to $\pi_\tau(x)A_\eta(z \mid x)$. The corresponding accepted transition flux is given by

$$\begin{aligned}&\pi_\tau(x)A_\eta(z \mid x)B^{(\tau)}_{\eta,\mathrm{AV}}(y \mid z)\alpha^{(\tau)}_{\mathrm{AV},z}(x,y)\\&\quad= \min\Big\{\pi_\tau(x)A_\eta(z \mid x)B^{(\tau)}_{\eta,\mathrm{AV}}(y \mid z),\, \pi_\tau(y)A_\eta(z \mid y)B^{(\tau)}_{\eta,\mathrm{AV}}(x \mid z)\Big\},\end{aligned}$$

which is manifestly symmetric in x and y. Summing over all $z \in \mathcal{X}$ yields detailed balance for the marginal AV-DMALA kernel with respect to π_τ. Crucially, this conditional auxiliary-variable formulation circumvents the costly evaluation of the marginal proposal mixture $\sum_z A_\eta(z \mid x)B^{(\tau)}_{\eta,\mathrm{AV}}(y \mid z)$.

Proposition 11 (Continuous-time limits of Metropolis-adjusted kernels)**.** *Let $K^{(\tau)}_{\eta,\sigma}$ denote the standard Metropolis–Hastings (MH) correction of $T^{(\tau)}_{\eta,\sigma}$. Then, as $h \downarrow 0$,*

$$K^{(\tau)}_{\eta,\sigma} = I + hQ^{(\tau)}_{\sigma,\mathrm{MH}} + \mathcal{O}(h^{1+\gamma}),$$

where the remainder is bounded with respect to the matrix ℓ_∞-operator norm. The generator $Q^{(\tau)}_{\sigma,\mathrm{MH}}$ has zero rates on non-edges, diagonal entries satisfying $\sum_y Q^{(\tau)}_{\sigma,\mathrm{MH}}(x,y) = 0$, and for every graph edge $y = x^{i\to b}$,

$$Q^{(\tau)}_{\sigma,\mathrm{MH}}(x,y) = \min\left\{e^{\sigma_i^{(\tau)}(x,b)/2}, \frac{\pi_\tau(y)}{\pi_\tau(x)}e^{\sigma_i^{(\tau)}(y,x_i)/2}\right\}.$$

Moreover, $Q^{(\tau)}_{\sigma,\mathrm{MH}} = Q^{(\tau)}_{\sigma}$ if and only if the edge-balance condition equation 6 holds.

In particular, the expansions

$$K^{(\tau)}_{\eta,\mathrm{A}} = I + hQ_{\mathrm{A},\tau} + \mathcal{O}(h^{1+\gamma}) \qquad \textit{and} \qquad K^{(\tau)}_{\eta,\mathrm{AV}} = I + hQ_{\mathrm{AV},\tau} + \mathcal{O}(h^{1+\gamma})$$

hold in the matrix ℓ_∞-operator norm.

Proof. **One-stage MH adjustment.** For any graph edge $y = x^{i\to b}$, the accepted transition probability is

$$K^{(\tau)}_{\eta,\sigma}(x,y) = \min\left\{T^{(\tau)}_{\eta,\sigma}(y \mid x), \frac{\pi_\tau(y)}{\pi_\tau(x)}T^{(\tau)}_{\eta,\sigma}(x \mid y)\right\}.$$

By Lemma 9,

$$T^{(\tau)}_{\eta,\sigma}(y \mid x) = he^{\sigma_i^{(\tau)}(x,b)/2} + \mathcal{O}(h^{1+\gamma}), \qquad T^{(\tau)}_{\eta,\sigma}(x \mid y) = he^{\sigma_i^{(\tau)}(y,x_i)/2} + \mathcal{O}(h^{1+\gamma}).$$

Since the mapping $(u,v) \mapsto \min\{u,v\}$ is 1-Lipschitz, substituting these expansions gives

$$K^{(\tau)}_{\eta,\sigma}(x,y) = hQ^{(\tau)}_{\sigma,\mathrm{MH}}(x,y) + \mathcal{O}(h^{1+\gamma}).$$

The adjusted and unadjusted transition rates coincide on both orientations of every graph edge if and only if

$$e^{\sigma_i^{(\tau)}(x,b)/2} = \frac{\pi_\tau(y)}{\pi_\tau(x)}e^{\sigma_i^{(\tau)}(y,x_i)/2},$$

which is equivalent to equation 6. Because score A satisfies this balance condition, we immediately have

$$K^{(\tau)}_{\eta,\mathrm{A}}(x,y) = hQ_{\mathrm{A},\tau}(x,y) + \mathcal{O}(h^{1+\gamma})$$

for every graph edge.

AV-DMALA. For a graph edge $y = x^{i\to b}$, the Metropolis-adjusted marginal transition probability is

$$K^{(\tau)}_{\eta,\mathrm{AV}}(x,y) = \sum_{z\in\mathcal{X}} \min\left\{A_\eta(z \mid x)B^{(\tau)}_{\eta,\mathrm{AV}}(y \mid z), \frac{\pi_\tau(y)}{\pi_\tau(x)}A_\eta(z \mid y)B^{(\tau)}_{\eta,\mathrm{AV}}(x \mid z)\right\}.$$

At first order, only the intermediate states $z = x$ and $z = y$ yield non-vanishing contributions:

- For $z = x$, applying $e^{\sigma^{(\tau)}_{\mathrm{AV},i}(x,b)/2} = \frac{\pi_\tau(y)}{\pi_\tau(x)}$, both arguments inside the minimum evaluate to $h\frac{\pi_\tau(y)}{\pi_\tau(x)} + \mathcal{O}(h^{1+\gamma})$, contributing
$$h\frac{\pi_\tau(y)}{\pi_\tau(x)} + \mathcal{O}(h^{1+\gamma}).$$
- For $z = y$, using $e^{\sigma^{(\tau)}_{\mathrm{AV},i}(y,x_i)/2} = \frac{\pi_\tau(x)}{\pi_\tau(y)}$, both arguments evaluate to $h + \mathcal{O}(h^{1+\gamma})$, contributing
$$h + \mathcal{O}(h^{1+\gamma}).$$

Any path through $z \notin \{x, y\}$ either involves off-diagonal transitions in both steps (contributing $\mathcal{O}(h^2)$) or traverses a non-neighboring transition (contributing $\mathcal{O}(h^{1+\gamma})$). Since $\gamma \le 1$, the collective contribution from all $z \notin \{x, y\}$ is $\mathcal{O}(h^{1+\gamma})$. Summing these contributions yields

$$K^{(\tau)}_{\eta,\mathrm{AV}}(x, y) = h\left(1 + \frac{\pi_\tau(y)}{\pi_\tau(x)}\right) + \mathcal{O}(h^{1+\gamma}) = hQ_{\mathrm{AV},\tau}(x, y) + \mathcal{O}(h^{1+\gamma}).$$

Non-edge transitions and diagonal entries. For the one-stage MH adjustment, the accepted off-diagonal probability is bounded by the unadjusted proposal probability: $K^{(\tau)}_{\eta,\sigma}(x, y) \le T^{(\tau)}_{\eta,\sigma}(y \mid x) = \mathcal{O}(h^{1+\gamma})$ for any non-edge $y \neq x$. For AV-DMALA, the same bound follows by bounding the minimum by the unadjusted proposal path: $\sum_z A_\eta(z \mid x)B^{(\tau)}_{\eta,\mathrm{AV}}(y \mid z) = \mathcal{O}(h^{1+\gamma})$. Thus, non-edge transition rates remain $\mathcal{O}(h^{1+\gamma})$ in both schemes.

Finally, the diagonal errors satisfy

$$K(x, x) - \big(1 + hQ(x, x)\big) = -\sum_{y\neq x}\big(K(x, y) - hQ(x, y)\big).$$

Given that $\mathcal{X}$ is finite and the off-diagonal errors are uniformly $\mathcal{O}(h^{1+\gamma})$, the diagonal errors are also uniformly $\mathcal{O}(h^{1+\gamma})$. Since entrywise $\mathcal{O}(h^{1+\gamma})$ bounds on a finite state space imply the same order in the induced ℓ_∞-operator norm, which completes the proof. □

D.3.4 Comparison of the Limiting Processes

Proposition 12 (Comparison of Limiting Generators)**.** *For every edge $x \sim y$, the transition rates satisfy*

$$\frac{Q_{\mathrm{AV},\tau}(x, y)}{Q_{\mathrm{A},\tau}(x, y)} = \frac{\big(\pi_\tau(x) + \pi_\tau(y)\big)^2}{\pi_\tau(x)\pi_\tau(y)} = 2 + \frac{\pi_\tau(y)}{\pi_\tau(x)} + \frac{\pi_\tau(x)}{\pi_\tau(y)} \ge 4,$$

with equality if and only if $\pi_\tau(x) = \pi_\tau(y)$. Consequently, under the shared time scaling h_η, the spectral gaps and asymptotic variances satisfy

$$\lambda_{\mathrm{AV},\tau} \ge 4\lambda_{\mathrm{A},\tau}, \qquad \mathrm{AVar}_{Q_{\mathrm{AV},\tau}}(f) \le \frac{1}{4}\,\mathrm{AVar}_{Q_{\mathrm{A},\tau}}(f)$$

for every nonconstant observable $f\colon \mathcal{X} \to \mathbb{R}$.

Remark 13. *The factor-of-four advantage pertains strictly to the limiting continuous-time jump processes under the shared time scaling h_η. It does not immediately guarantee an identical ordering for the discrete-time chains or wall-clock runtime, where discretization errors, proposal rejection rates, and per-iteration computational costs come into play.*

Proof. By definition of the generators, for any edge $x \sim y$,

$$Q_{\mathrm{A},\tau}(x, y) = \frac{\pi_\tau(y)}{\pi_\tau(x) + \pi_\tau(y)}, \qquad Q_{\mathrm{AV},\tau}(x, y) = \frac{\pi_\tau(x) + \pi_\tau(y)}{\pi_\tau(x)}.$$

Taking their ratio gives

$$\frac{Q_{\mathrm{AV},\tau}(x, y)}{Q_{\mathrm{A},\tau}(x, y)} = \frac{\big(\pi_\tau(x) + \pi_\tau(y)\big)^2}{\pi_\tau(x)\pi_\tau(y)} \ge 4,$$

which follows from the inequality $(a+b)^2 \geq 4ab$, with equality if and only if $\pi_\tau(x) = \pi_\tau(y)$.

Since both generators are reversible with respect to π_τ and have zero rates outside the edge set, their Dirichlet forms satisfy, for any $g\colon \mathcal{X} \to \mathbb{R}$,

$$\begin{aligned}\mathcal{E}_{Q_{\mathrm{AV},\tau}}(g) &= \frac{1}{2}\sum_{x\sim y} \pi_\tau(x) Q_{\mathrm{AV},\tau}(x,y)\big(g(y)-g(x)\big)^2 \\ &\geq \frac{1}{2}\sum_{x\sim y} \pi_\tau(x)\big[4Q_{\mathrm{A},\tau}(x,y)\big]\big(g(y)-g(x)\big)^2 = 4\mathcal{E}_{Q_{\mathrm{A},\tau}}(g).\end{aligned}$$

By the variational characterization of the spectral gap on $L_0^2(\pi_\tau)$, we obtain

$$\lambda_{\mathrm{AV},\tau} = \inf_{\substack{g\in L_0^2(\pi_\tau)\\ g\neq 0}} \frac{\mathcal{E}_{Q_{\mathrm{AV},\tau}}(g)}{\|g\|^2_{L^2(\pi_\tau)}} \geq 4 \inf_{\substack{g\in L_0^2(\pi_\tau)\\ g\neq 0}} \frac{\mathcal{E}_{Q_{\mathrm{A},\tau}}(g)}{\|g\|^2_{L^2(\pi_\tau)}} = 4\lambda_{\mathrm{A},\tau}.$$

For the asymptotic variance, let $f_0 := f - \pi_\tau(f) \in L_0^2(\pi_\tau)$. For any positive-definite self-adjoint operator $-Q$ on $L_0^2(\pi_\tau)$, the Poisson equation yields the dual variational formulation:

$$\langle f_0, (-Q)^{-1} f_0\rangle_{\pi_\tau} = \sup_{g\in L_0^2(\pi_\tau)} \left\{2\langle f_0, g\rangle_{\pi_\tau} - \mathcal{E}_Q(g)\right\}.$$

Applying the Dirichlet form bound $\mathcal{E}_{Q_{\mathrm{AV},\tau}}(g) \geq 4\mathcal{E}_{Q_{\mathrm{A},\tau}}(g)$ and setting $g = u/4$, we obtain

$$\begin{aligned}\langle f_0, (-Q_{\mathrm{AV},\tau})^{-1} f_0\rangle_{\pi_\tau} &\leq \sup_{g\in L_0^2(\pi_\tau)} \left\{2\langle f_0, g\rangle_{\pi_\tau} - 4\mathcal{E}_{Q_{\mathrm{A},\tau}}(g)\right\} \\ &= \sup_{u\in L_0^2(\pi_\tau)} \left\{\frac{1}{2}\langle f_0, u\rangle_{\pi_\tau} - \frac{1}{4}\mathcal{E}_{Q_{\mathrm{A},\tau}}(u)\right\} = \frac{1}{4}\langle f_0, (-Q_{\mathrm{A},\tau})^{-1} f_0\rangle_{\pi_\tau}.\end{aligned}$$

Recalling the representation $\mathrm{AVar}_Q(f) = 2\langle f_0, (-Q)^{-1} f_0\rangle_{\pi_\tau}$, we conclude that

$$\mathrm{AVar}_{Q_{\mathrm{AV},\tau}}(f) \leq \frac{1}{4}\,\mathrm{AVar}_{Q_{\mathrm{A},\tau}}(f),$$

which completes the proof. □

D.4 Proofs Omitted in Section 4

We use the conventions of Appendix D.1 and the local results of Appendix D.3. Assume $|\mathcal{X}| \geq 2$, the singleton case being trivial. Throughout this section, fix $\mathrm{M} \in \{\mathrm{A}, \mathrm{AV}\}$ and write $\pi_k := \pi_{\tau_k}$ and $Q_k := Q_{\mathrm{M},\tau_k}$ for $k = 1, 2$, so that $\Pi(x,y) = \pi_1(x)\pi_2(y)$. The index k identifies the replica, not its temperature. All small-step estimates hold for fixed swap intensity $a \geq 0$.

D.4.1 Replica-Exchange Generator

For any bounded test function $g\colon \mathcal{X}^2 \to \mathbb{R}$, the replica-exchange generator with swap intensity $a \geq 0$ acts as

$$\begin{aligned}(\mathcal{L}_{\mathrm{M},a} g)(x,y) = &\sum_{u\neq x} Q_1(x,u)\big[g(u,y) - g(x,y)\big] + \sum_{v\neq y} Q_2(y,v)\big[g(x,v) - g(x,y)\big] \\ &+ aS(x,y)\big[g(y,x) - g(x,y)\big],\end{aligned}$$

where $S(x,y)$ is the standard Metropolis swap probability in equation 2. Its finite-step counterpart $P_{\eta,a,\mathrm{M}}$ first applies $T_{\eta,\mathrm{M}}^{(\tau_1)}$ and $T_{\eta,\mathrm{M}}^{(\tau_2)}$ independently to the two replicas and then performs a swap with probability ahS, where $ah \leq 1$.

Lemma 14 (Replica-exchange generator and one-step approximation)**.** *For every $a \geq 0$, the generator $\mathcal{L}_{\mathrm{M},a}$ is irreducible and reversible with respect to $\Pi = \pi_1 \otimes \pi_2$, and hence $\Pi\mathcal{L}_{\mathrm{M},a} = 0$. Moreover, for fixed $a \geq 0$, there exists $c_{a,\mathrm{M}} < \infty$, independent of h, such that for all sufficiently small h,*

$$\|P_{\eta,a,\mathrm{M}} - I - h\mathcal{L}_{\mathrm{M},a}\|_\infty \leq c_{a,\mathrm{M}} h^{1+\gamma}. \tag{18}$$

Proof. **Reversibility and irreducibility.** Since Q_k is reversible with respect to π_k for $k = 1, 2$, the independent local generator $\mathcal{L}_{\mathrm{M},0}$ is reversible with respect to Π. The swap flow satisfies

$$\Pi(x, y)S(x, y) = \min\{\pi_1(x)\pi_2(y), \pi_1(y)\pi_2(x)\} = \Pi(y, x)S(y, x). \tag{19}$$

Hence the swap generator $\mathcal{S}$, and therefore $\mathcal{L}_{\mathrm{M},a} = \mathcal{L}_{\mathrm{M},0} + a\mathcal{S}$, is Π-reversible, which implies $\Pi\mathcal{L}_{\mathrm{M},a} = 0$. Since the independent local dynamics are irreducible on $\mathcal{X}^2$, adding swap transitions preserves irreducibility.

One-step approximation. By equation 11, $T_{\eta,\mathrm{M}}^{(\tau_k)} = I + hQ_k + \mathcal{O}(h^{1+\gamma})$ for $k = 1, 2$, with remainders bounded in the matrix ℓ_∞-operator norm. Combining the two independent local updates gives

$$\|P_{\eta,0,\mathrm{M}} - I - h\mathcal{L}_{\mathrm{M},0}\|_\infty = \mathcal{O}(h^{1+\gamma}),$$

since their first-order interaction is $\mathcal{O}(h^2)$ and $\gamma \le 1$.

For sufficiently small h, $ah \le 1$ and the swap step has transition kernel $I + ah\mathcal{S}$. Thus

$$P_{\eta,a,\mathrm{M}} = P_{\eta,0,\mathrm{M}}(I + ah\mathcal{S}),$$

and

$$P_{\eta,a,\mathrm{M}} - I - h\mathcal{L}_{\mathrm{M},a} = \big(P_{\eta,0,\mathrm{M}} - I - h\mathcal{L}_{\mathrm{M},0}\big)(I + ah\mathcal{S}) + ah^2\mathcal{L}_{\mathrm{M},0}\mathcal{S}.$$

Since $I + ah\mathcal{S}$ is a Markov kernel, $\|I + ah\mathcal{S}\|_\infty = 1$, while $\|\mathcal{S}\|_\infty \le 2$ and $\|\mathcal{L}_{\mathrm{M},0}\|_\infty < \infty$. Hence the two terms on the right are respectively $\mathcal{O}(h^{1+\gamma})$ and $\mathcal{O}(h^2)$, with the latter absorbed into $\mathcal{O}(h^{1+\gamma})$ because $\gamma \le 1$. This proves equation 18. □

D.4.2 Finite-Step Convergence

Proof of Proposition 4. By Lemma 14 and the semigroup estimate equation 12, for fixed $a \ge 0$ and all sufficiently small h,

$$\begin{aligned}\big\|P_{\eta,a,\mathrm{M}} - e^{h\mathcal{L}_{\mathrm{M},a}}\big\|_\infty &\le \|P_{\eta,a,\mathrm{M}} - I - h\mathcal{L}_{\mathrm{M},a}\|_\infty + \big\|e^{h\mathcal{L}_{\mathrm{M},a}} - I - h\mathcal{L}_{\mathrm{M},a}\big\|_\infty \\ &\le \left(c_{a,\mathrm{M}} + \frac{1}{2}\|\mathcal{L}_{\mathrm{M},a}\|_\infty^2\right) h^{1+\gamma}.\end{aligned}$$

Since $\mathcal{L}_{\mathrm{M},a}$ is irreducible and Π-reversible, its spectral gap $\lambda_{a,\mathrm{M}} := \lambda(\mathcal{L}_{\mathrm{M},a})$ is strictly positive. For sufficiently small h, we also have $\lambda_{a,\mathrm{M}}h \le 1$. Applying Lemma 10 therefore gives, for every initial distribution ν and integer $n \ge 0$,

$$\big\|\nu P_{\eta,a,\mathrm{M}}^n - \Pi\big\|_{\mathrm{TV}} \le \frac{1}{2}e^{-\lambda_{a,\mathrm{M}}nh}\sqrt{\chi^2(\nu \mid \Pi)} + C_{a,\mathrm{M}}h^\gamma\left(1 - e^{-\lambda_{a,\mathrm{M}}nh}\right), \tag{20}$$

where

$$C_{a,\mathrm{M}} := \frac{c_{a,\mathrm{M}} + \frac{1}{2}\|\mathcal{L}_{\mathrm{M},a}\|_\infty^2}{\lambda_{a,\mathrm{M}}\sqrt{\min_{x,y}\Pi(x, y)}}.$$

Since $1 - e^{-\lambda_{a,\mathrm{M}}nh} \le 1$, this implies

$$\big\|\nu P_{\eta,a,\mathrm{M}}^n - \Pi\big\|_{\mathrm{TV}} \le C_\nu e^{-\lambda_{a,\mathrm{M}}nh} + C_{a,\mathrm{M}}h^\gamma, \qquad \text{with} \quad C_\nu := \frac{1}{2}\sqrt{\chi^2(\nu \mid \Pi)},$$

which proves the finite-step bound in Proposition 4.

For the stationary bias, fix a and take h sufficiently small that $ah < 1$. The two local kernels have positive self-transition probabilities and positive transition probabilities along every coordinate edge, while the swap step leaves the updated pair unchanged with probability at least $1 - ah > 0$. Hence $P_{\eta,a,\mathrm{M}}$ is irreducible and aperiodic on $\mathcal{X}^2$, and therefore has a unique invariant distribution $\widehat{\Pi}_{\eta,a,\mathrm{M}}$. Letting $n \to \infty$ in equation 20 gives

$$\left\|\widehat{\Pi}_{\eta,a,\mathrm{M}} - \Pi\right\|_{\mathrm{TV}} \le C_{a,\mathrm{M}}h^\gamma.$$

Thus the stationary bias of the finite-step replica-exchange chain is uniformly of order $\mathcal{O}(h^\gamma)$. □

D.4.3 Dirichlet Forms and Spectral Monotonicity

Define the unit-intensity swap contribution by $\mathcal{E}_{\mathrm{sw}}(g) := \frac{1}{2}\sum_{x,y\in\mathcal{X}} \Pi(x,y)S(x,y)\big[g(y,x) - g(x,y)\big]^2$. The swap intensity a is not included in $\mathcal{E}_{\mathrm{sw}}$.

Lemma 15. *For every $g \in L^2(\Pi)$,*

$$\mathcal{E}_{a,\mathrm{M}}(g) = \mathcal{E}_{0,\mathrm{M}}(g) + a\mathcal{E}_{\mathrm{sw}}(g), \tag{21}$$

where $\mathcal{E}_{0,\mathrm{M}}(g) = \sum_y \pi_2(y)\mathcal{E}_{Q_1}\big(g(\cdot,y)\big) + \sum_x \pi_1(x)\mathcal{E}_{Q_2}\big(g(x,\cdot)\big)$. Moreover,

$$0 \le \mathcal{E}_{\mathrm{sw}}(g) \le 2\|g\|^2_{L^2(\Pi)}, \tag{22}$$

and $\mathcal{E}_{\mathrm{sw}}(g) = 0$ if and only if $g(x,y) = g(y,x)$ for all $x, y \in \mathcal{X}$.

Consequently, $\lambda_{a,\mathrm{M}}$ is nondecreasing in a and

$$\lambda_{a,\mathrm{M}} \ge \lambda_{0,\mathrm{M}} = \min\{\lambda_{\mathrm{M},\tau_1}, \lambda_{\mathrm{M},\tau_2}\}. \tag{23}$$

Proof. **Dirichlet-form decomposition and swap bound.** Separating the two local components and the swap component in $\mathcal{L}_{\mathrm{M},a}$, and using the swap-flow symmetry equation 19, gives

$$\mathcal{E}_{a,\mathrm{M}}(g) = \mathcal{E}_{0,\mathrm{M}}(g) + \frac{a}{2}\sum_{x,y} \Pi(x,y)S(x,y)\big[g(y,x) - g(x,y)\big]^2,$$

which proves equation 21. Since $\Pi(x,y)S(x,y) > 0$, the swap contribution vanishes exactly when $g(x,y) = g(y,x)$ for every x, y.

Using $(u-v)^2 \le 2(u^2+v^2)$, equation 19, and $S(x,y) \le 1$,

$$\mathcal{E}_{\mathrm{sw}}(g) \le \sum_{x,y} \Pi(x,y)S(x,y)\big[g(y,x)^2 + g(x,y)^2\big] = 2\sum_{x,y} \Pi(x,y)S(x,y)g(x,y)^2 \le 2\|g\|^2_{L^2(\Pi)},$$

which proves equation 22.

Spectral monotonicity and the uncoupled gap. Since $\mathcal{E}_{\mathrm{sw}}(g) \ge 0$, equation 21 shows that $\mathcal{E}_{a,\mathrm{M}}(g)$ is nondecreasing in a for every g. All generators have the same invariant distribution Π, so the variational characterization of the spectral gap implies that $\lambda_{a,\mathrm{M}}$ is also nondecreasing.

For $a = 0$, variance tensorization under $\Pi = \pi_1 \otimes \pi_2$, followed by the Poincaré inequalities for Q_1 and Q_2, gives

$$\mathrm{Var}_\Pi(g) \le \sum_y \pi_2(y)\,\mathrm{Var}_{\pi_1}\big(g(\cdot,y)\big) + \sum_x \pi_1(x)\,\mathrm{Var}_{\pi_2}\big(g(x,\cdot)\big) \le \frac{\mathcal{E}_{0,\mathrm{M}}(g)}{\min\{\lambda_{\mathrm{M},\tau_1}, \lambda_{\mathrm{M},\tau_2}\}}.$$

Hence $\lambda_{0,\mathrm{M}} \ge \min\{\lambda_{\mathrm{M},\tau_1}, \lambda_{\mathrm{M},\tau_2}\}$. Because the state space is finite, each local spectral gap is attained; testing the joint Rayleigh quotient with a gap-achieving function of either coordinate alone gives the reverse inequality. Therefore $\lambda_{0,\mathrm{M}} = \min\{\lambda_{\mathrm{M},\tau_1}, \lambda_{\mathrm{M},\tau_2}\}$, which completes the proof. □

D.4.4 Cold-Replica Asymptotic Variance

For a cold observable f, define the centered observable $F(x,y) := f(x) - \pi_1(f)$, and let $G_a \in L^2_0(\Pi)$ be the unique solution to the Poisson equation $-\mathcal{L}_{\mathrm{M},a}G_a = F$.

Proof of Theorem 5. By the Poisson representation in Appendix D.1, $\mathrm{AVar}_{a,\mathrm{M}}(f) = 2\langle F, G_a\rangle_\Pi$. At $a = 0$, the two replicas decouple, yielding $G_0(x,y) = \big[(-Q_1)^{-1}\big(f - \pi_1(f)\big)\big](x)$, since the hot generator acts trivially on functions depending solely on the cold coordinate. Consequently, $\mathrm{AVar}_{0,\mathrm{M}}(f) = \mathrm{AVar}_{Q_1}(f)$.

We decompose the combined generator as $\mathcal{L}_{\mathrm{M},a} = \mathcal{L}_{\mathrm{M},0} + a\mathcal{S}$, where $\mathcal{S}$ denotes the unit-intensity swap generator. On $L^2_0(\Pi)$, the operator $-\mathcal{L}_{\mathrm{M},a}$ is positive definite and invertible. Differentiating its inverse with respect to a gives

$$\frac{\mathrm{d}}{\mathrm{d}a}(-\mathcal{L}_{\mathrm{M},a})^{-1} = (-\mathcal{L}_{\mathrm{M},a})^{-1}\mathcal{S}(-\mathcal{L}_{\mathrm{M},a})^{-1}.$$

It follows that

$$\frac{\mathrm{d}}{\mathrm{d}a}\,\mathrm{AVar}_{a,\mathrm{M}}(f) = 2\langle G_a, \mathcal{S}G_a\rangle_{\Pi} = -2\mathcal{E}_{\mathrm{sw}}(G_a) \le 0, \tag{24}$$

which establishes that $\mathrm{AVar}_{a,\mathrm{M}}(f)$ is nonincreasing in a.

It remains to establish strict monotonicity. Suppose that f is nonconstant and that the derivative in equation 24 vanishes at some $a \ge 0$ (interpreted as a right-derivative at $a = 0$). By Lemma 15, the condition $\mathcal{E}_{\mathrm{sw}}(G_a) = 0$ implies that G_a is swap-invariant, whence $\mathcal{S}G_a = 0$. The Poisson equation then collapses to

$$-\mathcal{L}_{\mathrm{M},0}G_a = F.$$

By the uniqueness of the centered Poisson solution, we must have $G_a = G_0$. Because G_0 depends exclusively on the cold coordinate, swap invariance forces G_0 to be spatially constant. This in turn yields $F = -\mathcal{L}_{\mathrm{M},0}G_0 = 0$, contradicting the assumption that f is nonconstant.

Therefore, the derivative in equation 24 is strictly negative for all $a > 0$ and strictly negative from the right at $a = 0$. Consequently, for any nonconstant observable f,

$$\mathrm{AVar}_{a,\mathrm{M}}(f) < \mathrm{AVar}_{0,\mathrm{M}}(f) \qquad \text{for all } a > 0.$$

□

D.4.5 Large Deviations Analysis

For any $\mu \in \mathcal{P}(\mathcal{X})$, we define the functional

$$\mathcal{K}(\mu) := \frac{1}{2}\sum_{x,y} \Pi(x,y)S(x,y)\left[\sqrt{\frac{\mu(y)}{\pi_1(y)}} - \sqrt{\frac{\mu(x)}{\pi_1(x)}}\right]^2.$$

Thus, $\mathcal{K}(\mu)$ represents the swap Dirichlet form evaluated at the joint observable $(x,y) \mapsto \sqrt{\mu(x)/\pi_1(x)}$, which depends solely on the cold coordinate.

Proof of Theorem 6. By the classical Donsker–Varadhan theorem on finite state spaces, the empirical measure of the joint stationary process satisfies a large deviation principle (LDP) with good rate function $\zeta \mapsto \mathcal{E}_{a,\mathrm{M}}\big(\sqrt{\zeta/\Pi}\big)$. Applying the contraction principle via the continuous marginal projection $\zeta \mapsto \zeta_1$, where $\zeta_1(x) := \sum_y \zeta(x,y)$, the cold marginal satisfies an LDP with rate function

$$J_{a,\mathrm{M}}(\mu) = \min_{\substack{\zeta\in\mathcal{P}(\mathcal{X}^2)\\ \zeta_1=\mu}} \mathcal{E}_{a,\mathrm{M}}\left(\sqrt{\frac{\zeta}{\Pi}}\right). \tag{25}$$

The infimum in equation 25 is attained as a minimum because the feasible set $\{\zeta \in \mathcal{P}(\mathcal{X}^2) : \zeta_1 = \mu\}$ is compact and the objective is lower semicontinuous. Furthermore, since $\mathcal{E}_{a,\mathrm{M}} = \mathcal{E}_{0,\mathrm{M}} + a\mathcal{E}_{\mathrm{sw}}$ with $\mathcal{E}_{\mathrm{sw}} \ge 0$, the rate $J_{a,\mathrm{M}}(\mu)$ is monotonically nondecreasing in $a \ge 0$.

To identify the uncoupled (no-swap) rate, fix any feasible ζ and define

$$g(x,y) := \sqrt{\frac{\zeta(x,y)}{\Pi(x,y)}}, \qquad \varphi(x) := \sqrt{\frac{\mu(x)}{\pi_1(x)}}.$$

The marginal constraint $\zeta_1 = \mu$ is equivalent to $\|g(x,\cdot)\|^2_{L^2(\pi_2)} = \sum_y \pi_2(y)g(x,y)^2 = \varphi(x)^2$. By the reverse triangle inequality in $L^2(\pi_2)$, we have

$$\|g(u,\cdot) - g(x,\cdot)\|_{L^2(\pi_2)} \ge \big|\|g(u,\cdot)\|_{L^2(\pi_2)} - \|g(x,\cdot)\|_{L^2(\pi_2)}\big| = |\varphi(u) - \varphi(x)|.$$

Consequently, the cold contribution to $\mathcal{E}_{0,\mathrm{M}}(g)$ is bounded below by $\mathcal{E}_{Q_1}(\varphi)$, while the hot contribution is non-negative. Equality is uniquely attained when $g(x,y) = \varphi(x)$ (i.e., $\zeta(x,y) = \mu(x)\pi_2(y)$). Therefore,

$$J_{0,\mathrm{M}}(\mu) = \mathcal{E}_{Q_1}\left(\sqrt{\frac{\mu}{\pi_1}}\right). \tag{26}$$

Next, we quantify the rate enhancement induced by swaps ($a > 0$). Since $g \geq 0$ and $\|g(x,\cdot)\|_{L^2(\pi_2)} = \varphi(x)$, the Cauchy–Schwarz inequality implies $0 \leq \pi_2\big(g(x,\cdot)\big) \leq \varphi(x)$, which yields the pointwise variance bound

$$\|g(x,\cdot) - \varphi(x)\|^2_{L^2(\pi_2)} = 2\varphi(x)\big(\varphi(x) - \pi_2(g(x,\cdot))\big) \leq 2\,\mathrm{Var}_{\pi_2}\big(g(x,\cdot)\big).$$

Integrating with respect to π_1, we obtain $\|g-\varphi\|^2_{L^2(\Pi)} \leq 2\sum_x \pi_1(x)\,\mathrm{Var}_{\pi_2}\big(g(x,\cdot)\big)$. Combining this with the lower bound on the cold Dirichlet energy and the hot chain's Poincaré inequality gives the coercivity estimate

$$\mathcal{E}_{0,\mathrm{M}}(g) \geq J_{0,\mathrm{M}}(\mu) + \frac{\lambda_{\mathrm{M},\tau_2}}{2}\|g-\varphi\|^2_{L^2(\Pi)}. \tag{27}$$

Notice that $\mathcal{K}(\mu) = \mathcal{E}_{\mathrm{sw}}(\varphi)$. By the triangle inequality for the seminorm $\sqrt{\mathcal{E}_{\mathrm{sw}}}$ combined with equation 22,

$$\sqrt{\mathcal{K}(\mu)} \leq \sqrt{\mathcal{E}_{\mathrm{sw}}(g)} + \sqrt{2}\,\|g-\varphi\|_{L^2(\Pi)}.$$

For any $a > 0$, applying the Cauchy–Schwarz inequality to the inner product of $\big(\frac{1}{\sqrt{a}}, \frac{2}{\sqrt{\lambda_{\mathrm{M},\tau_2}}}\big)$ and $\big(\sqrt{a\mathcal{E}_{\mathrm{sw}}(g)}, \sqrt{\frac{\lambda_{\mathrm{M},\tau_2}}{2}}\|g-\varphi\|_{L^2(\Pi)}\big)$ yields

$$\mathcal{K}(\mu) \leq \left(\frac{1}{a} + \frac{4}{\lambda_{\mathrm{M},\tau_2}}\right)\left[a\mathcal{E}_{\mathrm{sw}}(g) + \frac{\lambda_{\mathrm{M},\tau_2}}{2}\|g-\varphi\|^2_{L^2(\Pi)}\right].$$

Substituting this bound into equation 27 and taking the infimum over all feasible ζ leads to

$$J_{a,\mathrm{M}}(\mu) \geq J_{0,\mathrm{M}}(\mu) + \frac{a\lambda_{\mathrm{M},\tau_2}}{\lambda_{\mathrm{M},\tau_2} + 4a}\mathcal{K}(\mu), \tag{28}$$

where $\lambda_{\mathrm{M},\tau_2}$ denotes the spectral gap of the full hot chain. Under the irreducibility assumption, all swap-flow weights are strictly positive, implying that $\mathcal{K}(\mu) = 0$ if and only if $\sqrt{\mu/\pi_1}$ is constant, which (by the normalization $\sum_x \mu(x) = 1$) holds if and only if $\mu = \pi_1$. Consequently, equation 28 establishes strict acceleration:

$$J_{a,\mathrm{M}}(\mu) > J_{0,\mathrm{M}}(\mu) \qquad \text{for all } a > 0 \text{ and } \mu \neq \pi_1.$$

Finally, every feasible candidate g in equation 25 satisfies $\|g\|_{L^2(\Pi)} = 1$ due to $\sum_{x,y}\zeta(x,y) = 1$. By the Cauchy–Schwarz inequality and the marginal condition,

$$\Pi(g) = \sum_{x,y}\Pi(x,y)g(x,y) \leq \sum_x \sqrt{\pi_1(x)}\sqrt{\mu(x)}.$$

Applying the Poincaré inequality for the generator $\mathcal{L}_{\mathrm{M},a}$ gives

$$\mathcal{E}_{a,\mathrm{M}}(g) \geq \lambda_{a,\mathrm{M}}\,\mathrm{Var}_\Pi(g) = \lambda_{a,\mathrm{M}}\big(1-\Pi(g)^2\big) \geq \lambda_{a,\mathrm{M}}\left[1 - \left(\sum_x \sqrt{\mu(x)\pi_1(x)}\right)^2\right].$$

Taking the infimum over feasible ζ yields the global spectral gap bound

$$J_{a,\mathrm{M}}(\mu) \geq \lambda_{a,\mathrm{M}}\left[1 - \left(\sum_x \sqrt{\mu(x)\pi_1(x)}\right)^2\right]. \tag{29}$$

By continuity, these arguments remain valid when μ has zero-mass components. □

Corollary 16. *Let $C \subset \mathcal{P}(\mathcal{X})$ be a nonempty closed set with $\pi_1 \notin C$. Then, for every $a > 0$,*

$$\inf_{\mu\in C} J_{a,\mathrm{M}}(\mu) \geq \inf_{\mu\in C} J_{0,\mathrm{M}}(\mu) + \frac{a\lambda_{\mathrm{M},\tau_2}}{\lambda_{\mathrm{M},\tau_2}+4a}\min_{\mu\in C}\mathcal{K}(\mu) > \inf_{\mu\in C} J_{0,\mathrm{M}}(\mu).$$

Proof. The probability simplex $\mathcal{P}(\mathcal{X})$ is compact, and $\mathcal{K}$ is continuous with a unique minimum of zero attained at π_1. Since C is compact and $\pi_1 \notin C$, the extreme value theorem ensures that $\min_{\mu\in C}\mathcal{K}(\mu) > 0$. The claim then follows directly from equation 28. □

D.4.6 MULTIMODAL SPECTRAL-GAP BOUNDS

We now introduce a mode decomposition to establish a quantitative lower bound on the spectral gap. We emphasize that the preceding variance and large-deviation results do not rely on this partition structure.

Assumption 1 (Within-mode irreducibility)**.** *Let $\mathcal{X} = \bigsqcup_{i=1}^{J} A_i$ be a partition of $\mathcal{X}$ into $J \geq 2$ non-empty pairwise disjoint subsets. For each $k \in \{1, 2\}$ and each block A_i, let $Q_k^{(i)}$ denote the reflecting restriction of Q_k to A_i, obtained by retaining transitions within A_i and adjusting the diagonal entries to ensure zero row sums. We assume that every $Q_k^{(i)}$ with $|A_i| \geq 2$ is irreducible.*

Let $p_i := \pi_1(A_i)$ and $q_i := \pi_2(A_i)$ denote the mode weights. The projected generator of the hot chain onto the mode space is defined by

$$\overline{Q}_2(i,j) := \frac{1}{q_i} \sum_{x \in A_i} \sum_{y \in A_j} \pi_2(x) Q_2(x,y), \qquad i \neq j,$$

with diagonal entries chosen such that $\sum_j \overline{Q}_2(i,j) = 0$. By construction, $\overline{Q}_2$ is irreducible and reversible with respect to $q = (q_1, \ldots, q_J)$. We define the key spectral quantities

$$\lambda_{\text{in}} := \min_{k \in \{1,2\},\, i \in \{1,\ldots,J\}} \lambda(Q_k^{(i)}), \qquad \lambda_{\text{hot}} := \lambda(\overline{Q}_2), \qquad \overline{R}_{\text{hot}} := \max_{x \in \mathcal{X}} \sum_{y \neq x} Q_2(x,y),$$

under the convention that $\lambda(Q_k^{(i)}) = +\infty$ whenever $|A_i| = 1$. Finally, we quantify the mode-wise overlap between the two target measures via

$$\delta := \min_{1 \leq i \leq J} \frac{\sum_{x \in A_i} \min\{\pi_1(x), \pi_2(x)\}}{\max\{p_i, q_i\}} \in (0, 1].$$

Here, λ_{hot} denotes the spectral gap of the projected mode chain, whereas $\lambda_{\text{M},\tau_2}$ defined earlier denotes the gap of the full hot chain on $\mathcal{X}$.

Proposition 17 (Multimodal spectral-gap bound)**.** *Under Assumption 1, for every swap intensity $a > 0$,*

$$\lambda_{a,\text{M}} \geq \min\left\{ \frac{\delta^2 \lambda_{\text{in}} \lambda_{\text{hot}}}{(12 + \delta^2)\lambda_{\text{hot}} + 9\overline{R}_{\text{hot}}},\ \frac{\delta^2 \lambda_{\text{hot}}}{9},\ \frac{a\delta^2}{12} \right\}, \tag{30}$$

with the first term interpreted as $+\infty$ when all modes are singletons. In conjunction with equation 23, this yields the universal bound

$$\lambda_{a,\text{M}} \geq \max\left\{ \lambda_{0,\text{M}},\ \frac{\delta^2}{13} \min\left\{ \frac{\lambda_{\text{in}} \lambda_{\text{hot}}}{\lambda_{\text{hot}} + \overline{R}_{\text{hot}}},\ \lambda_{\text{hot}},\ a \right\} \right\}. \tag{31}$$

Proof. **Within-mode fluctuations.** For each mode pair (i,j), let $\bar{g}_{ij} := \mathbb{E}_\Pi[g \mid A_i \times A_j]$ and define the within-mode variance $W := \sum_{i,j} p_i q_j \operatorname{Var}_\Pi(g \mid A_i \times A_j)$. The law of total variance gives

$$\operatorname{Var}_\Pi(g) = W + \sum_{i,j} p_i q_j \left[\bar{g}_{ij} - \Pi(g)\right]^2. \tag{32}$$

Let $\mathcal{E}_{\text{in}}(g)$ be the Dirichlet energy of local transitions remaining within their respective modes. Since the reflecting product dynamics on each block $A_i \times A_j$ have a spectral gap of at least λ_{in}, the blockwise Poincaré inequalities yield

$$W \leq \lambda_{\text{in}}^{-1} \mathcal{E}_{\text{in}}(g), \tag{33}$$

with the standard convention $\lambda_{\text{in}}^{-1} = 0$ if all modes are singletons (in which case $W = 0$).

Between-mode fluctuations. The definition of δ ensures $p_i \leq \delta^{-1} q_i$ and $q_i \leq \delta^{-1} p_i$. By equation 19, the mode-swap capacity $s_{ij} := \sum_{x \in A_i, y \in A_j} \Pi(x) S(x,y)$ satisfies

$$s_{ij} \geq \sum_{x \in A_i, y \in A_j} \min\{\pi_1(x), \pi_2(x)\} \min\{\pi_1(y), \pi_2(y)\} \geq \delta^2 p_i q_j. \tag{34}$$

Next, we introduce the projected Dirichlet forms for the hot transitions and swap moves:

$$\overline{\mathcal{E}}_{\mathrm{hot}}(\bar{g}) := \frac{1}{2}\sum_i p_i \sum_{j\neq\ell} q_j \overline{Q}_2(j,\ell)(\bar{g}_{i\ell} - \bar{g}_{ij})^2, \qquad \overline{\mathcal{E}}_{\mathrm{swap}}(\bar{g}) := \frac{a}{2}\sum_{i,j} s_{ij}(\bar{g}_{ji} - \bar{g}_{ij})^2.$$

Conditioning on the cold-mode index, we decompose the inter-mode variance as $\sum_{i,j} p_i q_j[\bar{g}_{ij} - \Pi(g)]^2 = \sum_i p_i \operatorname{Var}_{j\sim q}(\bar{g}_{ij}) + \operatorname{Var}_{i\sim p}(\sum_j q_j \bar{g}_{ij})$. Writing $\sum_j q_j \bar{g}_{ij} = \sum_j q_j(\bar{g}_{ij} - \bar{g}_{ji}) + \sum_j q_j \bar{g}_{ji}$, applying variance subadditivity $\operatorname{Var}(u+v) \le 2\operatorname{Var}(u) + 2\operatorname{Var}(v)$, Jensen's inequality, and the comparison $p_i \le \delta^{-1} q_i$, we obtain $\operatorname{Var}_{i\sim p}(\sum_j q_j \bar{g}_{ij}) \le 2\sum_{i,j} p_i q_j(\bar{g}_{ij} - \bar{g}_{ji})^2 + 2\delta^{-2}\sum_i p_i \operatorname{Var}_{j\sim q}(\bar{g}_{ij})$. Combining this with $\delta \le 1$, the Poincaré inequality for $\overline{Q}_2$, and equation 34, we deduce

$$\sum_{i,j} p_i q_j[\bar{g}_{ij} - \Pi(g)]^2 \le \frac{3}{\delta^2 \lambda_{\mathrm{hot}}}\overline{\mathcal{E}}_{\mathrm{hot}}(\bar{g}) + \frac{4}{a\delta^2}\overline{\mathcal{E}}_{\mathrm{swap}}(\bar{g}). \tag{35}$$

Lifting to the original dynamics. Let $\mathcal{E}^{\mathrm{out}}_{\mathrm{hot}}(g)$ denote the Dirichlet contribution from hot transitions between distinct modes. For any $u \in A_i, x \in A_j$, and $y \in A_\ell$ with $j \neq \ell$, we decompose $\bar{g}_{i\ell} - \bar{g}_{ij} = [\bar{g}_{i\ell} - g(u,y)] + [g(u,y) - g(u,x)] + [g(u,x) - \bar{g}_{ij}]$. Expanding via $(r+s+t)^2 \le 3(r^2+s^2+t^2)$ and integrating against $\frac{1}{2}\pi_1(u)\pi_2(x)Q_2(x,y)$, the middle term produces $3\mathcal{E}^{\mathrm{out}}_{\mathrm{hot}}(g)$, while the two endpoint terms are together bounded by $3\overline{R}_{\mathrm{hot}}W$ via the reversibility of Q_2. Consequently,

$$\overline{\mathcal{E}}_{\mathrm{hot}}(\bar{g}) \le 3\mathcal{E}^{\mathrm{out}}_{\mathrm{hot}}(g) + 3\overline{R}_{\mathrm{hot}}W. \tag{36}$$

Similarly, writing $\bar{g}_{ji} - \bar{g}_{ij} = [\bar{g}_{ji} - g(y,x)] + [g(y,x) - g(x,y)] + [g(x,y) - \bar{g}_{ij}]$ and integrating against $\frac{a}{2}\Pi(x,y)S(x,y)$, the swap-flow symmetry and $S \le 1$ yield

$$\overline{\mathcal{E}}_{\mathrm{swap}}(\bar{g}) \le 3a\mathcal{E}_{\mathrm{sw}}(g) + 3aW. \tag{37}$$

Combining the estimates. Substituting equation 36 and equation 37 into equation 35, and using the total variance decomposition equation 32, we obtain

$$\begin{aligned}\operatorname{Var}_\Pi(g) &\le \left(1 + \frac{12}{\delta^2} + \frac{9\overline{R}_{\mathrm{hot}}}{\delta^2\lambda_{\mathrm{hot}}}\right)W + \frac{9}{\delta^2\lambda_{\mathrm{hot}}}\mathcal{E}^{\mathrm{out}}_{\mathrm{hot}}(g) + \frac{12}{\delta^2}\mathcal{E}_{\mathrm{sw}}(g) \\ &\le \max\left\{\frac{(12+\delta^2)\lambda_{\mathrm{hot}} + 9\overline{R}_{\mathrm{hot}}}{\delta^2\lambda_{\mathrm{in}}\lambda_{\mathrm{hot}}},\ \frac{9}{\delta^2\lambda_{\mathrm{hot}}},\ \frac{12}{a\delta^2}\right\}\mathcal{E}_{a,\mathrm{M}}(g),\end{aligned}$$

where the second inequality applies equation 33 and uses that $\mathcal{E}_{\mathrm{in}}(g)$, $\mathcal{E}^{\mathrm{out}}_{\mathrm{hot}}(g)$, and $a\mathcal{E}_{\mathrm{sw}}(g)$ are non-negative disjoint components of $\mathcal{E}_{a,\mathrm{M}}(g)$. Taking the infimum over the Rayleigh quotient establishes equation 30.

Finally, using $\delta \le 1$, we observe $(12+\delta^2)\lambda_{\mathrm{hot}} + 9\overline{R}_{\mathrm{hot}} \le 13(\lambda_{\mathrm{hot}} + \overline{R}_{\mathrm{hot}})$ and $\max\{9,12\} \le 13$. Hence, equation 30 simplifies to

$$\lambda_{a,\mathrm{M}} \ge \frac{\delta^2}{13}\min\left\{\frac{\lambda_{\mathrm{in}}\lambda_{\mathrm{hot}}}{\lambda_{\mathrm{hot}} + \overline{R}_{\mathrm{hot}}},\ \lambda_{\mathrm{hot}},\ a\right\}.$$

Together with equation 23, this completes the proof of equation 31 and Proposition 7. □