Multi-Marginal Couplings for Metropolis–Hastings

Buu Phan
University of Toronto
truong.phan@mail.utoronto.ca

Gergely Flamich
Imperial College London
g.flamich@imperial.ac.uk

Ashish Khisti
University of Toronto
akhisti@ece.utoronto.ca

Shahab Asoodeh
McMaster University
asoodeh@mcmaster.ca

Abstract

Convergence diagnosis for Markov chain Monte Carlo is a matter of fundamental importance in computational statistics: it determines the resources allocated to a particular sampling problem and influences the practitioner's view of the quality of estimates obtained from a Markov chain. Motivated by this, we contribute to the emerging class of coupling-based convergence diagnostic algorithms. Concretely, we study coupling multiple Metropolis–Hastings chains using multi-marginal coupling. We introduce a natural objective for this setting and establish lower and upper bounds by drawing connections to list-level distribution coupling and distributed pairwise-matching problems. This analysis ultimately leads to a shared-randomness Poisson Monte Carlo construction for coupling multiple Markov chains. In this process, we avoid a key dimension-dependent bottleneck in the runtime complexity of classical Poisson Monte Carlo by developing an adaptive rule for updating the point process, yielding significant gains in high-dimensional settings. Experiments on grand couplings of Markov chains show that our methods improve coalescence rates across dimensions, reducing meeting times by up to 50% compared with existing baselines.

1 Introduction

Markov chain Monte Carlo (MCMC) is a central methodology for sampling from complex probability distributions when exact simulation techniques are computationally intractable. It proceeds by constructing a Markov chain whose stationary distribution matches a target distribution of interest, allowing expectations under the target law to be approximated from the chain trajectory. Because of its broad applicability in Bayesian inference, statistical physics, and machine learning, MCMC has become one of the most widely used tools for high-dimensional probabilistic computation.

A fundamental challenge in MCMC is convergence diagnosis: determining whether the chain has run long enough for its distribution to be sufficiently close to stationarity. Classical approaches often rely on running multiple independent chains with different initializations and comparing their behavior, as in the Gelman–Rubin diagnostic and its extensions [3, 10]. Coupling provides another principled alternative by jointly evolving chains from different initializations and using their meeting behavior as a diagnostic signal, where a meeting occurs when two or more chains coalesce in the same state [2, 4]. As the effectiveness of such diagnostics depends on how quickly coalescence occurs, several studies have examined the design of efficient couplings, with a primary focus on two chains [30]. By comparison, the multi-chain coupling setting has received relatively limited attention, despite the existence of convergence diagnostics based on grand couplings, which jointly evolve chains from multiple initial states until they coalesce [16].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Coupling multiple chains is not as simple as pairwise coupling. A natural first strategy is to construct a multi-chain coupling using pairwise coupling as a building block. For example, we could have a "server-client" topology, in which we couple each chain to a distinguished reference chain; alternatively, we could randomly choose coupling partners at each transition [30, 2]. Such approaches, however, can be highly suboptimal: they may spend effort attempting to couple chains whose states are far apart, and thus whose coalescence is unlikely. A more effective strategy should preferentially match chains that are close under the transition kernel, but identifying such pairs is challenging in high-dimensional settings. In this work, we develop adaptive coupling strategies in which the joint construction itself determines which chains should meet by connecting the multi-marginal setting to list-level coupling and distributed matching. This strategy avoids committing to fixed partners, promotes coalescence among compatible chains, and leads to a kernel-coupling construction with effective meeting behavior and favorable runtime scaling in high-dimensional multi-chain settings. In particular, our contributions are:

1. In Section 2, we study multi-chain coupling from a multi-marginal perspective and analyze theoretical properties of the resulting meeting behavior through list-level coupling (membership inclusion), and distributed matching formulations.
2. In Section 3, we introduce an efficient Poisson-matching scheme for coupling multiple Metropolis–Hastings kernels. The scheme uses a flexible adaptive proposal update that yields runtime scaling primarily with the number of chains rather than the dimension.
3. In Section 4, we show experimentally that our method outperforms natural grand-coupling baselines, substantially reducing coalescence times in high-dimensional settings when using many chains.

2 Algorithms for Multi-Marginals Coupling

We first recall the pairwise case. For probability measures P and Q, the maximum agreement probability over all couplings is $\Pr(X^P = X^Q) = 1 - d_{\mathrm{TV}}(P, Q)$, where $d_{\mathrm{TV}}(P, Q)$ denotes the total variation distance. A coupling attaining this bound is called a maximal coupling [19]; its achievability is reviewed in Appendix A.1. For more than two marginals, however, there is no canonical notion of agreement. To explore this setting, consider C random variables $X^1, \dots, X^C$ with marginals $X^i \sim P^i$. Natural objectives could be to maximize the probability of full coalescence or the probability of at least one pairwise match, as studied in the multi-marginal coupling setting of Angel and Spinka [1]:

$$\Pr(X^1 = X^2 = \cdots = X^C) \qquad \text{or} \qquad \Pr(\exists\, i \neq j : X^i = X^j).$$

For coupling multiple Markov chains, coalescence is expected to occur progressively over iterations. Therefore, full coalescence is often too stringent, especially as C grows. In contrast, the existence of a single pairwise match does not distinguish between a coupling that merges two variables and one that merges many. These observations motivate the following measure of partial coalescence.

Problem setup. To quantify partial coalescence, we consider the optimization problem

$$\mathbf{G}^* = \inf_{\gamma \in \Gamma(P^1, \dots, P^C)} \mathbb{E}_\gamma[G], \qquad G \triangleq \big|\{\text{unique values among } X^1, \dots, X^C\}\big|, \tag{1}$$

where $\Gamma(P^1, \dots, P^C)$ is the set of all couplings of $P^1, \dots, P^C$ and G is the number of distinct realized values among the C variables. It directly measures progressive coalescence: $G{=}C$ corresponds to complete separation, $G{=}1$ to full coalescence, and intermediate values to partial merges.

The structure of an optimal coupling for equation 1, and how to construct one, are not immediately clear. A natural first attempt is to extend maximal coupling by selecting one variable as an anchor and coupling all other variables to it. However, this construction is asymmetric and designed to maximize agreement with the chosen anchor, rather than directly minimize the number of distinct realized values. This section will show that this objective admits informative lower and upper bounds.

Theorem 2.1. *Let $P^1, \dots, P^C$ be probability measures, and let σ be a permutation of $\{1, \dots, C\}$. For each $k \in \{1, \dots, C-1\}$, define the mixture distribution*

$$\bar{\nu}^{\sigma(k)} := \frac{1}{C-k} \sum_{j=k+1}^{C} P^{\sigma(j)}.$$

Recall that the Hockey-Stick divergence [27] of order $m \geq 1$ is defined by

$$E_m(P\|Q) \coloneqq \int \big(P(x) - mQ(x)\big)_+ \,\mathrm{d}x,$$

where $(a)_+ \coloneqq \max\{a, 0\}$. Then the optimal value $\mathbf{G}^*$ of equation 1 satisfies the following bounds.

Lower bound.

$$\mathbf{G}^* \geq 1 + \max_{\sigma} \sum_{k=1}^{C-1} E_{C-k}\Big(P^{\sigma(k)} \,\Big\|\, \bar{\nu}^{\sigma(k)}\Big). \tag{2}$$

Upper bounds.

$$\mathbf{G}^* \leq C - \max_{\sigma} \sum_{r=1}^{C-1} \int \left[\int \max\left\{\frac{P^{\sigma(r)}(v)}{P^{\sigma(r)}(y)}, \frac{P^{\sigma(r+1)}(v)}{P^{\sigma(r+1)}(y)}\right\} \mathrm{d}v\right]^{-1} \mathrm{d}y \tag{3}$$

$$\leq 1 + \min_{\sigma} \sum_{r=1}^{C-1} \frac{2\, d_{\mathrm{TV}}\big(P^{\sigma(r)}, P^{\sigma(r+1)}\big)}{1 + d_{\mathrm{TV}}\big(P^{\sigma(r)}, P^{\sigma(r+1)}\big)}. \tag{4}$$

Both bounds are tight at the two extreme configurations. If $X^1 = \cdots = X^C$ almost surely, then $d_{\mathrm{TV}}(P^i, P^j) = 0$ for all $i \neq j$ and $G = 1$, so both bounds hold with equality. Conversely, if $X^i \neq X^j$ almost surely for all $i \neq j$, then $G = C$, and both bounds are again tight. Beyond these extremes, the lower bound is tight when $C = 2$, in which case the problem reduces to the maximal coupling of two variables, and in certain other special cases; see Example 1 and Appendix C. The total variation and hockey-stick divergences appearing in these bounds are used only for theoretical analysis and are not required to implement the coupling algorithms. Numerical estimates of the bounds can be obtained via Monte Carlo simulation without explicitly computing the divergences, or through variational methods [22]. We now prove Theorem 2.1 and develop coupling constructions motivated by the structure of these bounds.

2.1 Lower Bound with List-Level Coupling

To establish the lower bound in Theorem 2.1, we first rewrite the expected cluster count in terms of list-level inclusion probabilities. Fixing any ordering $\sigma(1), \ldots, \sigma(C)$ of the indices, we show in Appendix B that, for any coupling of $X^1, \ldots, X^C$,

$$G = C - \sum_{k=1}^{C-1} \mathbf{1}\Big(X^{\sigma(k)} \in \{X^{\sigma(k+1)}, \ldots, X^{\sigma(C)}\}\Big) \tag{5}$$

Intuitively, scanning the ordering backward, the inclusion event means the current value has already appeared in the suffix, so it does not create a new cluster. Furthermore, this idea yields a natural coupling procedure in the multi-marginal setting, where the matching partner need not be fixed in advance. In particular, under a given ordering, each variable may coalesce with any later variable that realizes the same value. Thus, we get the following lower bound on the optimal objective $\mathbf{G}^\star$:

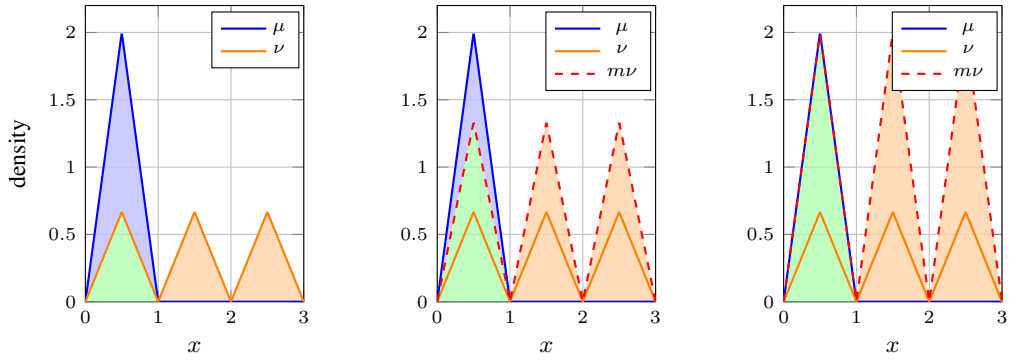
$$\mathbf{G}^* = \inf_{\gamma \in \Gamma} \mathbb{E}_\gamma[G] \geq C - \sum_{k=1}^{C-1} \sup_{\gamma \in \Gamma} \Pr\Big(X^{\sigma(k)} \in \{X^{\sigma(k+1)}, \ldots, X^{\sigma(C)}\}\Big), \tag{6}$$

where $\Gamma \triangleq \Gamma(P^1, \cdots, P^C)$ and the inequality in (6) relaxes the problem by optimizing each list-level inclusion (or membership) probability separately; note that a different coupling may attain the optimum for each term. The following maximal list-level coupling result upper bounds each separated inclusion probability. Together with equation 6, it directly yields the lower bound in Theorem 2.1.

Theorem 2.2 (List-Level Coupling)**.** *Let $X \sim \mu$ and $Y^1, \ldots, Y^m$ have marginals $\nu^1, \ldots, \nu^m$, all absolutely continuous w.r.t. a common dominating measure λ. Then we have:*

$$\sup_{\gamma \in \Gamma(\mu, \nu^1, \ldots, \nu^m)} \Pr_{\gamma}\big(X \in \{Y^1, \ldots, Y^m\}\big) = 1 - E_m(\mu\|\bar{\nu}),$$

where $\bar{\nu} \coloneqq m^{-1} \sum_{j=1}^m \nu^j$ is the mixture barycenter of $\nu^1, \ldots, \nu^m$.



(a) $m = 1$ (b) $m = 2$ (c) $m = 3$

Figure 1: Illustration of Theorem 2.2 for $m = 1, 2, 3$. Green regions denote the overlap $\ell(x) = \min\{\mu(x), m\nu(x)\}$, blue regions denote the missing mass $\mu_{\text{miss}} = \mu - \ell$, and the orange region denotes the residual $m\nu - \ell$. For $m = 3$, the overlap coincides with μ, so the missing mass vanishes.

Details and proof of this theorem are given in Appendix B.2. When $m = 1$, the result reduces to the classical maximal coupling identity. To the best of our knowledge, this provides the first list-level coupling analog of maximal coupling. Consider the example in Figure 1, where the measures ν^i are identical. The matching probability for $m = 1$ is exactly the mass of the green overlap region. As m increases, the scaled measure $m\nu$ covers a larger portion of μ, thereby increasing the overlap

$$\ell(x) = \min\{\mu(x), m\nu(x)\}$$

and decreasing the missing mass μ_{miss}. For $m = 3$, Theorem 2.2 implies that the optimal list-inclusion probability is 1: there exists a coupling for which X is always contained in the list (Y_1, Y_2, Y_3). Such a coupling can be obtained by sampling (Y_1, Y_2, Y_3) as a uniformly random permutation of the three possible outcomes. Each Y_j then has marginal distribution ν, while the list contains every outcome exactly once and therefore necessarily contains the value drawn from μ. A useful visual interpretation is obtained by scaling ν by the list size $m = 3$, which completely covers the single mode distribution $\mu = (1, 0, 0)$. Equivalently, the missing mass vanishes, so the list has sufficient mass on the support of μ to guarantee a match.

The supremum is attained by Algorithm 1, together with the residual-sampling subroutine in Algorithm 2. The construction samples $X \sim \mu$, then on the coupling event randomly chooses which Y^j coalesces with X, and samples the remaining Y's from their residual distributions. Specifically, the residual density of R^j is

$$r^j(x) = \frac{\nu^j(x)}{1 - \alpha^j}\left(1 - \min\left\{1, \frac{\mu(x)}{m\bar{\nu}(x)}\right\}\right), \qquad \alpha^j = \int \nu^j(x) \min\left\{1, \frac{\mu(x)}{m\bar{\nu}(x)}\right\} \lambda(\mathrm{d}x), \quad (7)$$

with the convention that the integrand is 0 whenever $\bar{\nu}(x)=0$.[1] Note that the sampling steps highlighted in blue in Algorithm 1 need not be independent. Hence, when direct residual sampling is available, such as when the distributions involved are discrete, there is a natural greedy strategy: after each list-level match is formed, remove the matched variables and recursively apply Algorithm 1 to the remaining unmatched list. For general cases, see Remark 1 below Algorithm 2. Although the greedy strategy does not attain the lower bound in general, it can be optimal in certain settings; see Example 1. Empirically, we find that it outperforms fixed-reference baselines, see Section 4.

Example 1. Suppose P^1, P^2, P^3, P^4 decompose into two separated overlap pairs, P^1 with P^2 and P^3 with P^4, with no cross-overlap between the two pairs, then the recursive list-level coupling attains maximal pairwise matching within both pairs, thereby achieving optimal performance. The lower bound in Theorem 2.1 is also achieved for a similar reason.

2.2 Upper Bound via Poisson Matching and Fast Coupling Construction

Although list-level coupling provides useful theoretical guarantees and suggests a natural adaptive strategy, its implementation for general continuous measures remains challenging. We now introduce

[1]With a slight abuse of notation, we denote their densities with the same letter when no confusion arises.

Algorithm 1: Maximal list coupling

Input: densities $\mu, \nu^1, \dots, \nu^m$; residual samplers $R^1, \dots, R^m$
Output: $(X, Y^1, \dots, Y^m)$
$\bar{\nu}(x) \leftarrow m^{-1} \sum_{j=1}^m \nu^j(x)$
$X, U \leftarrow \texttt{Sample}(\mu), \texttt{Sample}(\mathrm{Unif}(0,1))$
if $U \le \min\{1, m\bar{\nu}(X)/\mu(X)\}$ **then**
  Sample $I \in [m]$ with $\Pr(I = j \mid X) = \nu^j(X)/(m\bar{\nu}(X))$
  $Y^I \leftarrow X$
  $Y^j \leftarrow \texttt{Sample}(R^j)$ for $j \in [m] \setminus \{I\}$
end
else
  $Y^j \leftarrow \texttt{Sample}(R^j)$ for all $j \in [m]$
end
return $(X, Y^1, \dots, Y^m)$, matched index I

Algorithm 2: Residual sampler for R^j

Input: densities $\nu^j, \mu, \bar{\nu}$
Output: $Y^j \sim R^j$
repeat
  $Z, U \leftarrow \texttt{Sample}(\nu^j), \texttt{Sample}(\mathrm{Unif}(0,1))$
  $a(Z) \leftarrow \left(1 - \frac{\mu(Z)}{m\bar{\nu}(Z)}\right)_+$
until $U \le a(Z)$
return Z

Remark 1. In general, it is nontrivial to recursively apply the blue step together with maximal list coupling when residual sampling is required, e.g. when all measures are Gaussians. Without the normalizing constant $1 - \alpha^j$ of r^j in equation 7, the barycenter likelihood cannot be computed explicitly.

an alternative construction based on a shared marked Poisson process that preserves the idea of Algorithm 1: the matching partners are determined adaptively. Related constructions were also considered by Angel and Spinka [1], but computationally, their use was limited to special discrete settings, since direct implementation becomes prohibitively expensive in high dimensions. Here, we introduce a new technique that turns this Poisson-process construction into a practical coupling algorithm for continuous high-dimensional MCMC. The induced pairwise coupling probabilities provide the key control needed to establish the upper bound in Theorem 2.1.

Poisson Monte Carlo (PMC) . We begin by introducing PMC, a sampling method for drawing samples from a target distribution P using a proposal distribution μ. Assume $P \ll \mu$[2], and let $\Omega = \{(S_i, X_i)\}_{i=1}^\infty$ be a marked Poisson process, where $\{S_i\}_{i=1}^\infty$ are the arrival times of a unit-rate Poisson process on $\mathbb{R}_+$ and the marks X_i are i.i.d. samples from μ. Define the random index

$$J = \arg\min_i \; S_i \left(\frac{\mathrm{d}P}{\mathrm{d}\mu}(X_i) \right)^{-1}, \qquad \text{where } \frac{\mathrm{d}P}{\mathrm{d}\mu} \text{ is the Radon–Nikodym derivative.} \tag{8}$$

Then, setting $X^P = X_J$, we have $X^P \sim P$ (marginalize over Ω). The procedure to solve the above optimization problem is presented in Appendix A.3, following [21, 28, 8]. To couple $X^1, \dots, X^C$ simultaneously, we let all sampler parties use a common Poisson process Ω, which induces the pairwise matching guarantees, known as the Poisson Matching Lemma (PML) [20].

Lemma 2.3 (Poisson Matching Lemma [20])**.** *Let $\Omega = \{(S_i, X_i)\}_{i=1}^\infty$ be a marked Poisson process on $\mathbb{R}_+ \times \mathcal{X}$ with intensity measure $\mathrm{d}s \otimes \mu$, and P and Q be two probability measures such that $P, Q \ll \mu$. Let P and Q be coupled via the Poisson Monte Carlo procedure described above, producing samples $X^P \sim P$ and $X^Q \sim Q$. Then, we have:*

$$\Pr(X^P = X^Q | X^P = x) \ge \left(1 + \frac{\mathrm{d}P}{\mathrm{d}Q}(x)\right)^{-1}, \qquad \Pr(X^P = X^Q) \ge \frac{1 - d_{\mathrm{TV}}(P,Q)}{1 + d_{\mathrm{TV}}(P,Q)}, \tag{9}$$

where the gap in matching probability compared with maximal coupling is small in practice [5, 20].

A detailed proof that the PML construction achieves the upper bound in Theorem 2.1 is provided in Appendix D. Here, $\mathrm{d}P/\mathrm{d}Q$ is interpreted as the extended density ratio $(\mathrm{d}P/\mathrm{d}\mu)/(\mathrm{d}Q/\mathrm{d}\mu)$, taking the value $+\infty$ whenever the denominator vanishes and the numerator is positive. We adopt Poisson Monte Carlo rather than classical rejection sampling because it has been shown to achieve higher matching probabilities, both theoretically and empirically [23]. We next analyze the computational complexity of PML and introduce a new technique to accelerate the coupling procedure.

[2] In the context of PMC, this distribution μ serves as a proposal measure and should not be confused with the reference distribution used in the list-level coupling of Algorithm 1.

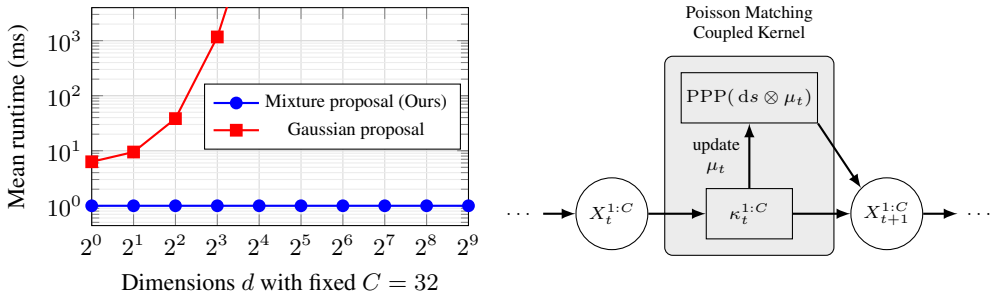


Figure 2: Poisson point process (PPP) coupling. (Left) Runtime comparison across dimensions between our constructed proposal and a naive strategy. (Right) Schematic illustration of how our method applies to multi-chain Markov coupling through adaptive updates of the shared PPP.

Runtime Complexity Problem. One caveat of the PMC approach is that its runtime can become prohibitively expensive as the dimension of X increases. In particular, for a target distribution P and proposal distribution μ then the expected computational cost typically scales as $\mathcal{O}(\operatorname{ess\,sup}_x\{(\mathrm{d}P/\mathrm{d}\mu)(x)\})$. Unfortunately, this can grow *exponentially* in the dimension even under a mild mismatch. For example, let P and μ be d-dimensional product measures, with $P = \bigotimes_{i=1}^d P_i$ and $\mu = \bigotimes_{i=1}^d \mu_i$. In the i.i.d. case where $P_i = P_1$ and $\mu_i = \mu_1$ for all i

$$\frac{\mathrm{d}P}{\mathrm{d}\mu}(x_{1:d}) = \prod_{i=1}^d \frac{\mathrm{d}P_1}{\mathrm{d}\mu_1}(x_i) \quad \Rightarrow \quad \operatorname*{ess\,sup}_{x_{1:d}} \frac{\mathrm{d}P}{\mathrm{d}\mu}(x_{1:d}) = \left(\operatorname*{ess\,sup}_x \frac{\mathrm{d}P_1}{\mathrm{d}\mu_1}(x)\right)^d = M^d,$$

where $M = \operatorname{ess\,sup}_x(\mathrm{d}P_1/\mathrm{d}\mu_1)(x) > 1$ is the one-dimensional mismatch factor . This leads to an M^d/C per-chain sample complexity in the grand-coupling setting. Thus, in the high-dimensional regimes typical of MCMC, a poorly matched proposal can lead to prohibitively large runtimes.

Adaptive Proposal for Fast Coupling. To improve the runtime complexity, our starting point is the common assumption that sampling from each target distribution can be performed in $\mathcal{O}(1)$ time. This assumption is particularly natural in the MCMC setting, where sampling from each transition kernel is typically cheap. Still, this does not automatically yield an efficient coupling: if the common measure μ is chosen to be one of the target distributions, say P, then sampling from another distribution Q may remain prohibitively expensive, since $\operatorname{ess\,sup}_x\{(\mathrm{d}Q/\mathrm{d}P)(x)\}$ can be large. Interestingly, this dimensional dependence can be replaced by dependence on the number of coupling proposals, which is often more manageable in practice, by setting μ to the barycenter of the target measures.

Formally, consider C probability measures $P^1, \ldots, P^C$ such that sampling from each P^i can be done in $\mathcal{O}(1)$ time. Then, we choose the proposal to be the barycenter measure μ as their uniform mixture:

$$\mu \triangleq \frac{1}{C}\sum_{j=1}^C P^j. \tag{10}$$

The following lemma shows that this choice yields favorable runtime complexity.

Lemma 2.4 (Mixture Barycenter Proposal)**.** *Let $P^1, \ldots, P^C$ be probability measures on $\mathcal{X}$, and the marked Poisson process defined previously with the common barycenter measure μ above. Then:*

$$P^i \ll \mu \qquad \textit{and} \qquad \operatorname*{ess\,sup}_{x\in\mathcal{X}} \left(\frac{\mathrm{d}P^i}{\mathrm{d}\mu}\right)(x) \le C, \qquad i = 1, \ldots, C. \tag{11}$$

As a result, sampling from each measure has worst-case complexity $\mathcal{O}(C)$, and since they all share the same marked Poisson process, the expected number of samples per probability measure is $\mathcal{O}(1)$.

The proof is in Appendix E. While evaluating $\mathcal{O}(C)$ shared candidates per marginal incurs $\mathcal{O}(C^2)$ total computational work, the construction is highly parallelizable. In particular, density evaluations can be batched into a $C \times C$ score tensor and computed in parallel on a GPU, followed by row-wise minimization, substantially reducing wall-clock runtime. Figure 2 (Left) shows the computational efficiency of this approach for $C = 32$ across dimensions $d \in \{2^0, \ldots, 2^9\}$. The target distributions are Gaussian with distinct means and identity covariance. We compare against a single-Gaussian

proposal centered at the average target mean with covariance CI_d, ensuring broad coverage. Both methods are evaluated using parallel workers under the same computational setting. For $C = 32$, the score tensor contains 1024 entries, which can be efficiently processed on the NVIDIA A40 GPU with negligible additional wall-clock overhead. Consequently, our runtime remains nearly constant with dimension, while the Gaussian-proposal baseline grows exponentially.

3 Coupling Multiple Metropolis-Hastings Chains

While list-level coupling provides useful guarantees and suggests a natural adaptive strategy, it is not trivial to implement for general continuous distributions. We therefore focus on adaptive Poisson matching constructions as tractable couplings for Markov chain Monte Carlo methods that use the Metropolis-Hastings (MH) algorithm. We now review the MH algorithm and introduce the setup.

Metropolis-Hastings (MH) Algorithm. Given a target distribution π, and a proposal kernel K with Lebesgue density $k(\cdot \mid x)$, so that $K(\mathrm{d}y \mid x) = k(y \mid x)\mathrm{d}y$, the Metropolis–Hastings algorithm proceeds from a current state x by proposing $y \sim k(\cdot \mid x)$ and accepting the move with probability

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)k(x \mid y)}{\pi(x)k(y \mid x)}\right\}.$$

If the proposal is accepted, the next state is y; otherwise, the chain remains at x. This construction defines a Markov chain with a stationary distribution π. Also see Algorithm 5, Appendix A.2.

Remark 2. In the case of MH, we note that the resulting Markov transition kernel κ can be written as $\kappa(\mathrm{d}y \mid x) = \alpha(x, y)\, k(y \mid x)\, \mathrm{d}y + \delta_x(\mathrm{d}y)\, r(x)$, where $r(x) = 1 - \int \alpha(x, y)\, k(y \mid x)\, \mathrm{d}y$ is the rejection probability at state x, and δ_x denotes the Dirac measure at x. Importantly, evaluating κ in closed form is nontrivial, since computing $r(x)$ is generally intractable.

The central goal in MCMC convergence analysis is to assess how close the chain's law at iteration t, denoted by π_t, is to the target distribution π. To this end, coupling provides a principled way to study convergence by jointly evolving two or more Markov chains. Following [17], we consider a *grand coupling* of C chains, all initialized from a prior distribution π_0 and evolved using a coupled kernel.

Definition 1 (Grand coupling)**.** Let κ be a Markov kernel on a measurable space $\mathcal{X}$. A *grand coupling* of κ for C chains is a Markov kernel $\bar{\kappa}$ on the product space $\mathcal{X}^C$ such that, if

$$(X_{t+1}^{(1)}, \ldots, X_{t+1}^{(C)}) \sim \bar{\kappa}(\cdot \mid (x_t^{(1)}, \ldots, x_t^{(C)})),$$

then $X_{t+1}^{(i)} \sim \kappa(\cdot \mid x_t^{(i)}),$ for $i = 1, \ldots, C.$ and the associated grand meeting time is

$$\tau \triangleq \inf\{t \geq 0 : X_t^{(1)} = \cdots = X_t^{(C)}\}.$$

We further assume faithfulness: once two or more chains meet at some time $\tilde{\tau}$, they are updated identically thereafter and remain equal for all $t \geq \tilde{\tau}$. Following [2, 16], faster coalescence is desirable because the tail probability $\Pr(\tau > t)$ often controls convergence bounds, so reducing this term leads to tighter bounds. See Appendix J for further details.

3.1 Poisson Matching Coupled Kernel

We introduce the *joint coupled kernel* method, which couples directly at the level of the full Markov transition kernel κ. As κ is typically unavailable in closed form for scoring in equation 8, we lift the construction to the augmented proposal–acceptance space. For each chain i, define the lifted measure

$$P_i'(\mathrm{d}y, \mathrm{d}u) = k(y \mid x_t^{(i)})\, \mathrm{d}y\, \mathrm{Ber}\big(\alpha(x_t^{(i)}, y)\big)(\mathrm{d}u), \tag{12}$$

where $Y_t^{(i)} \sim k(\cdot \mid x_t^{(i)})$, and $U_t^{(i)} \mid \{Y_t^{(i)} = y\} \sim \mathrm{Ber}(\alpha(x_t^{(i)}, y))$. The chains then update their respective next state by

$$x_{t+1}^{(i)} = \mathbf{1}\{U_t^{(i)} = 1\} Y_t^{(i)} + \mathbf{1}\{U_t^{(i)} = 0\} x_t^{(i)}. \tag{13}$$

This lifted representation avoids treating the rejection component as an explicit atom inside the PML construction; instead, rejection is encoded by the auxiliary variable $U_t^{(i)}$, while the pushforward

Algorithm 3: Metropolis–Hastings Coupled Kernel with Poisson Matching

Input : Current states $x_t^{(1)}, \dots, x_t^{(C)}$

Output : Updated states $x_{t+1}^{(1)}, \dots, x_{t+1}^{(C)}$

Set lifted target measures: $P_i'(\mathrm{d}y, \mathrm{d}u) = k(y \mid x_t^{(i)})\,\mathrm{d}y\,\mathrm{Ber}\big(\alpha(x_t^{(i)}, y)\big)(\mathrm{d}u), \quad i = 1, \dots, C$

Proposal measure: $\mu(\mathrm{d}y, \mathrm{d}u) = \dfrac{1}{C}\displaystyle\sum_{i=1}^{C} P_i'(\mathrm{d}y, \mathrm{d}u)$

Shared Poisson process: Generate $\Omega = \{(S_j, Y_j, U_j)\}_{j \geq 1}$ with reference measure $\mathrm{d}s \otimes \mu$

for $i = 1, \dots, C$ **do**

PML selection: $J^{(i)} = \arg\min_{j \geq 1} \left\{ S_j \left(\dfrac{\mathrm{d}P_i'}{\mathrm{d}\mu}(Y_j, U_j) \right)^{-1} \right\}$

MH update: $x_{t+1}^{(i)} \leftarrow U_{J^{(i)}} Y_{J^{(i)}} + \big(1 - U_{J^{(i)}}\big) x_t^{(i)}$

end

through equation 13 recovers the original MH transition kernel. How does this augmented-space strategy compare with the non-augmented one, assuming full knowledge about the point mass measure? Surprisingly, Lemma 3.1 shows that the augmented construction yields a better matching probability than the direct coupling scheme based on κ; see Appendix I for the proof.

Lemma 3.1. *Consider two chains at states x_t^1 and x_t^2 respectively. At $t+1$, let $\mathbb{P}_{\mathrm{aug}}(x_t^1, x_t^2)$ denote the meeting probability under the product-augmented construction, and let $\mathbb{P}_{\mathrm{direct}}(x_t^1, x_t^2)$ denote the meeting probability under the original direct construction. Then*

$$\mathbb{P}_{\mathrm{aug}}(x_t^1, x_t^2) \geq \mathbb{P}_{\mathrm{direct}}(x_t^1, x_t^2).$$

To construct a shared source of common randomness, we use a marked Poisson point process $\Omega = \{(Y_j, U_j, S_j)\}_{j \geq 1}$,[3] whose marks (Y_j, U_j) are drawn i.i.d. from the mixture measure[4]

$$\mu(\mathrm{d}y, \mathrm{d}u) \triangleq \frac{1}{C} \sum_{i=1}^{C} k(y \mid x_t^{(i)})\,\mathrm{d}y\,\mathrm{Ber}\big(\alpha(x_t^{(i)}, y)\big)(\mathrm{d}u).$$

We apply the PMC selection step using P_i' as the target measure and μ as the common proposal. The procedure is summarized in Algorithm 3. In Appendix H, we outline the *two-stage coupled kernel* , which separately couples proposals and MH accept/reject decisions. Our joint construction achieves stronger matching-probability guarantees, see Theorem H.1, consistent with our empirical findings.

4 Experiments

Multi-Marginal Coupling. We begin with the finite-state setup from Section 2, shown in the left panel of Figure 3. We evaluate the proposed multi-marginal coupling methods on randomly generated sparse discrete distributions over a state space with $|\mathcal{X}| = 60$, where each distribution is supported on 5 states, which induces an irregular overlap structure across pairs of measures. We estimate $\mathbb{E}[G]$ for $C \in \{2^1, \dots, 2^5\}$ over 2×10^5 runs, ensuring that multiple measures share at least one state. Our baselines are two maximal pairwise-coupling schemes. The first, called "random anchor", selects a reference measure at random in each trial and couples all other measures to it. The second, random sequence, applies pairwise coupling along a randomly generated sequence of pairs. For $C = 2$, the pairwise baselines and the list-level coupling recover the optimal coupling performance. For larger C, our schemes consistently achieve substantially lower $\mathbb{E}[G]$ than the pairwise baselines. The list-level coupling outperforms Poisson matching for $C \in \{2, 4, 8, 16\}$, likely because the sparse overlap structure resembles the optimality-achieving setting in Example 1. In such settings, list-level coupling can exploit local overlap groups, whereas Poisson matching does not generally maximize all

[3] We use j to index the Poisson points and t to denote the Markov chain time index.

[4] For simple implementation, one may also use $\mu(\mathrm{d}y, \mathrm{d}u) \triangleq \mathrm{Ber}\big(1/2\big)(\mathrm{d}u) \times \frac{1}{C} \sum_{i=1}^{C} k(y \mid x_t^{(i)})\,\mathrm{d}y$, with upper-bound ratio and expected runtime complexity $2C$.



Method	Number of Measures C				
	2	4	8	16	32
Poisson Matching	1.83	3.61	7.10	11.97	**18.99**
List	**1.71**	**3.44**	**6.82**	**11.56**	19.28
Random Anchor	**1.71**	3.62	7.37	13.56	23.61
Random Sequence	**1.71**	3.63	7.38	13.59	23.82
Lower Bound	1.71	3.40	6.60	10.28	14.60

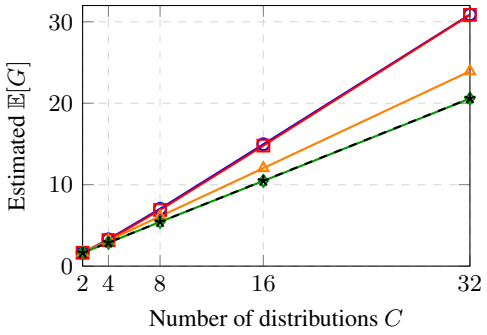


Figure 3: Estimated $\mathbb{E}[G]$ across different coupling strategies across number of measures C. Left: random discrete measures. Right: exponential target measures (scale$=1$ but different locations).

pairwise matches simultaneously. As C grows, the combined supports cover more of the state space, making this local-overlap advantage less pronounced; in this regime, Poisson matching becomes more effective and can outperform the list-level coupling.

The right panel of Figure 3 depicts a continuous setting where the recursive list-level step is analytically computable, using the shifted exponential family $P_i = i + \mathrm{Exp}(1)$, $i = 0, \ldots, C-1$. The list-level coupling again outperforms the pairwise baselines. Interestingly, Poisson matching exactly matches the lower bound; it turns out that theoretically, Poisson matching achieves an optimal $\mathbf{G}^*$ in this particular setting. The proof is given in Appendix C, where we show that $\mathbf{G}^* = C - (C-1)e^{-1}$.

Coupling Metropolis–Hastings chains. We evaluate our proposed scheme on two representative random-walk Metropolis–Hastings (RWMH) kernels in Figure 4. The first row considers a Gaussian target $\pi = \mathcal{N}(0, \mathrm{I}_d)$, with chains initialized from $\pi_0 = \mathcal{N}(\mathbf{1}_d, 16\mathrm{I}_d)$. The second row considers a heavy-tailed Cauchy target sampled using a Student-t random-walk proposal with 2 degrees of freedom, where the chains start with $\pi_0 = \mathcal{N}(0, \mathrm{I}_d)$. The proposal scale is set to σ=$2.4/\sqrt{d}$, following the classical high-dimensional optimal-scaling heuristic for random-walk Metropolis [11]. We compare our proposed scheme with grand-coupling extensions of the pairwise methods from [17] and [30], namely Algorithms 1 and 3 of [30] for general-purpose kernels. Since these methods were originally developed for coupling two chains, we adapt them to the C-chain setting by selecting one chain as a reference and coupling each of the remaining $C-1$ chains to it via maximal coupling; we refer to this as star coupling, which can be viewed as the two-step and one-step counterparts of our approach, respectively. For these baselines, we find that using a fixed anchor, or reference chain, across iterations works best. We measure the average meeting time as a function of the number of chains C with fixed d, and as a function of dimension d with fixed C. Across both targets, our method achieves smaller meeting times than the pairwise star-coupling baselines. We attribute the improvement to the adaptive construction of the shared point process, which couples all chains jointly rather than forcing every match through a fixed reference chain.

Additional experiments. See Appendix J for further discussion of other kernel and additional details on grand couplings for convergence diagnostics. There, we present an alternative bound different from that of Johnson [16], targetting the case where rejection coefficient between the two measures is not available. We also discuss applications to the weight harmonization [4] diagnostic.

5 Related Work

Multi-marginal coupling. Our construction builds on the Poisson matching lemma [20], which studies matching from shared Poisson randomness primarily for information-theoretic achievability using Poisson Monte Carlo (PMC), also known as A^* sampling [21, 28, 9, 8]. This framework has recently gained traction in compression [9, 7, 26], alongside related schemes based on different sampling procedures [23, 5, 24]. Several recent works study list-level coupling questions motivated by multi-draft speculative decoding [25, 18]; however, prior to our work, optimal achievability bounds for the general multi-marginal formulation were not known. For couplings of multiple measures, to the best of our knowledge, the closest existing work is that of Angel and Spinka [1], who study simultaneous pairwise disagreement guarantees for discrete distributions. In contrast, we study multi-

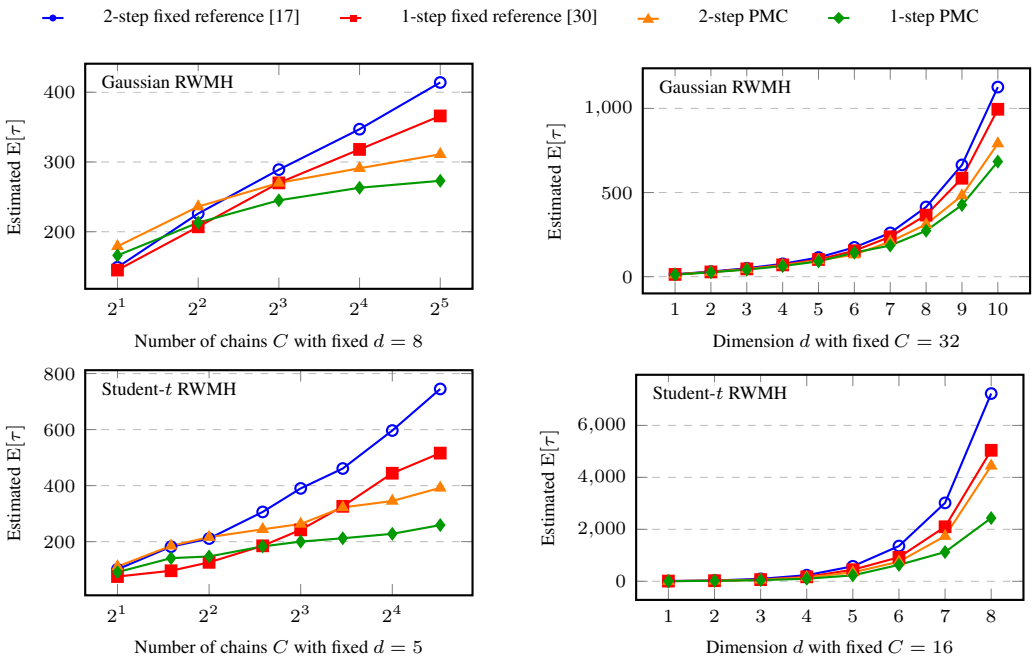


Figure 4: Comparison of estimated meeting times across Gaussian and Student-t targets. Top row: Gaussian target; bottom row: Student-t target. Left column: meeting time versus the number of coupled chains; right column: meeting time versus dimension. Results are averaged over 10000 runs.

marginal coupling through a cluster-count objective, relate it to list-level and pairwise mismatch, and develop efficient constructions that extend naturally to continuous measures.

Markov chain coupling. Couplings are widely used to study the convergence of Markov chains and to design MCMC convergence diagnostics [16, 2, 15, 13]. Existing approaches include contraction-based couplings, such as reflection couplings [6], and exact-meeting constructions based on maximal couplings of Metropolis–Hastings kernels [30] or Gibbs kernels [29]. We focus on multi-chain couplings that directly target list-level and pairwise matching events through a multi-marginal formulation, and can be viewed as extending maximal-coupling ideas for Metropolis–Hastings kernels to the multi-chain setting.

6 Conclusion

We studied coupling constructions for Markov chains from a multi-marginal perspective, with the goal of encouraging matching behavior beyond the classical pairwise maximal-coupling setting. Our construction leads to an optimization formulation for multi-marginal coupling, for which we provide a general bounds as well as optimal solutions in several specific settings. These solutions are attained using either list-level coupling or Poisson matching constructions. From a theoretical perspective, it would be interesting to develop more general-purpose solutions to this optimization problem and to design efficient algorithms for computing or approximating them. Such developments could, in turn, lead to stronger and more broadly applicable coupling constructions for Markov chains.

Related to convergence analysis, a promising direction for future work is to build and derive sharper convergence diagnostics based on multiple coupled chains, potentially enabling more informative assessments of mixing and convergence than standard two-chain methods. Another important application of coupled Markov chains is perfect simulation, such as coupling from the past [14]. These methods can be viewed as an extreme form of convergence diagnosis: they couple chains from all possible initial states and check whether they have coalesced, yielding exact samples when applicable. In contrast, our methods track coalescence among only finitely many chains as a practical diagnostic. A natural future direction is to explore whether Poisson coupling techniques can be used to construct update functions for perfect simulation as well.

References

[1] Omer Angel and Yinon Spinka. Pairwise optimal coupling of multiple random variables. *arXiv preprint arXiv:1903.00632*, 2019.

[2] Niloy Biswas, Pierre E Jacob, and Paul Vanetti. Estimating convergence of markov chains with l-lag couplings. *Advances in neural information processing systems*, 32, 2019.

[3] Stephen P Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455, 1998.

[4] Adrien Corenflos and Hai-Dang Dau. A coupling-based approach to f-divergences diagnostics for markov chain monte carlo. *arXiv preprint arXiv:2510.07559*, 2025.

[5] Majid Daliri, Christopher Musco, and Ananda Theertha Suresh. Coupling without communication and drafter-invariant speculative decoding. In *2025 IEEE International Symposium on Information Theory (ISIT)*, pages 1–6. IEEE, 2025.

[6] Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probability theory and related fields*, 166(3):851–886, 2016.

[7] Gergely Flamich. Greedy poisson rejection sampling. *Advances in Neural Information Processing Systems*, 36:37089–37127, 2023.

[8] Gergely Flamich. *Data Compression with Relative Entropy Coding*. PhD thesis, Apollo - University of Cambridge Repository, 2024. URL `https://www.repository.cam.ac.uk/handle/1810/385303`. PhD Thesis.

[9] Gergely Flamich, Stratis Markou, and José Miguel Hernández-Lobato. Fast relative entropy coding with a* coding. *arXiv preprint arXiv:2201.12857*, 2022.

[10] Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

[11] Andrew Gelman, Walter R Gilks, and Gareth O Roberts. Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1):110–120, 1997.

[12] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(2): 123–214, 2011.

[13] Geoffrey R Grimmett and Mark Holmes. Coalescence in markov chains. *arXiv preprint arXiv:2510.13572*, 2025.

[14] Mark L Huber. Perfect simulation. *Monographs on Statistics and Applied Probability*, 148:148, 2016.

[15] Pierre E Jacob, John O'leary, and Yves F Atchadé. Unbiased markov chain monte carlo methods with couplings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82 (3):543–600, 2020.

[16] Valen E Johnson. Studying convergence of markov chain monte carlo algorithms using coupled sample paths. *Journal of the American Statistical Association*, 91(433):154–166, 1996.

[17] Valen E Johnson. A coupling-regeneration scheme for diagnosing convergence in markov chain monte carlo algorithms. *Journal of the American Statistical Association*, 93(441):238–248, 1998.

[18] Ashish Khisti, M Reza Ebrahimi, Hassan Dbouk, Arash Behboodi, Roland Memisevic, and Christos Louizos. Multi-draft speculative sampling: Canonical decomposition and theoretical limits. *arXiv preprint arXiv:2410.18234*, 2024.

[19] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[20] Cheuk Ting Li and Venkat Anantharam. A unified framework for one-shot achievability via the poisson matching lemma. *IEEE Transactions on Information Theory*, 67(5):2624–2651, 2021.

[21] Chris J Maddison, Daniel Tarlow, and Tom Minka. A* sampling. *Advances in Neural Information Processing Systems*, 27:3086–3094, 2014.

[22] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[23] Buu Phan and Ashish J Khisti. Channel simulation and distributed compression with ensemble rejection sampling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[24] Buu Phan, Ashish Khisti, and Christos Louizos. Importance matching lemma for lossy compression with side information. In *International Conference on Artificial Intelligence and Statistics*, pages 1387–1395. PMLR, 2024.

[25] Joseph Rowan, Buu Phan, and Ashish Khisti. List-level distribution coupling with applications to speculative decoding and lossy compression. *arXiv preprint arXiv:2506.05632*, 2025.

[26] Joseph Rowan, Buu Phan, and Ashish Khisti. One-shot broadcast joint source-channel coding with codebook diversity. *arXiv preprint arXiv:2601.10648*, 2026.

[27] Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

[28] Lucas Theis and Noureldin Y Ahmed. Algorithms for the communication of samples. In *International Conference on Machine Learning*, pages 21308–21328. PMLR, 2022.

[29] Brian L Trippe, Tin D Nguyen, and Tamara Broderick. Optimal transport couplings of gibbs samplers on partitions for unbiased estimation. *arXiv preprint arXiv:2104.04514*, 2021.

[30] Guanyang Wang, John O'Leary, and Pierre Jacob. Maximal couplings of the metropolis-hastings algorithm. In *International Conference on Artificial Intelligence and Statistics*, pages 1225–1233. PMLR, 2021.

A Auxiliary Algorithms

A.1 Maximal Coupling

Achievability. Let P and Q be probability measures on a measurable space $(\mathcal{X}, \mathcal{X})$. Their meet measure is defined by

$$\gamma := P \wedge Q,$$

that is, γ is the largest measure dominated by both P and Q. Equivalently, when P and Q have densities p and q with respect to the same base measure, the common component has density

$$\gamma(x) = \min\{p(x), q(x)\},$$

with the abuse of notation, we denote the density with the same letter when no confusion arises. The total mass of the common component satisfies

$$\gamma(\mathcal{X}) = 1 - d_{\mathrm{TV}}(P, Q),$$

where

$$d_{\mathrm{TV}}(P, Q) = \sup_{A \in \mathcal{X}} |P(A) - Q(A)|.$$

A maximal coupling can be constructed by first identifying this common component. Namely,

(i) with probability $\gamma(\mathcal{X})$, sample $X^P = X^Q \sim \gamma/\gamma(\mathcal{X})$,

(ii) otherwise, sample $X^P \sim \dfrac{P - \gamma}{d_{\mathrm{TV}}(P, Q)}$ and $X^Q \sim \dfrac{Q - \gamma}{d_{\mathrm{TV}}(P, Q)}$.

By construction, $X^P \sim P$ and $X^Q \sim Q$. Moreover,

$$\Pr(X^P = X^Q) = \gamma(\mathcal{X}) = 1 - d_{\mathrm{TV}}(P, Q),$$

which is the largest possible agreement probability over all couplings of P and Q. We note that this construction requires sampling not only from the common component γ, but also from the residual measures $P - \gamma$ and $Q - \gamma$. In practice, when P and Q admit densities p and q with respect to the same base measure, the following rejection-based implementation provides a general maximal coupling algorithm.

A.2 Metropolis Hasting Algorithms

We provide the Metropolis-Hasting algorithm for target distribution $\pi(\cdot)$ using the proposal kernel $K(\cdot|x)$ in Algorithm 5.

A.3 Poisson Monte Carlo Algorithm

Algorithm 6 gives the procedure used to solve the optimization problem over the countably infinite set of Poisson points in equation 8, following the version of Theis and Ahmed [28].

B List-Level Theorems

Notation Notice. Unlike Poisson matching, the list-level coupling does not select atoms by Poisson indices. For this reason, we switch to use subscripts to denote distribution indices for convenience.

B.1 List-Level Coupling Representation

We first relate G to our list-level objective, fix an ordering of the variables and define, for $k = 1, \ldots, C-1$,

$$A_k := \{X_k \in \{X_{k+1}, \ldots, X_C\}\}.$$

The event A_k means that X_k does not introduce a new cluster relative to the suffix $(X_{k+1}, \ldots, X_C)$. Accordingly, if we define

$$G_k \triangleq |\{\text{unique values among } X_k, \ldots, X_C\}|,$$

Algorithm 4: Maximal coupling of P and Q

Input:
Probability measures P and Q with densities p and q w.r.t. a common dominating measure
Output:
A coupled pair (X^P, X^Q) with marginals P and Q
`// Step 1: Sample from` P`.`
Sample $X \sim P$ and $W \sim \mathrm{Unif}(0,1)$
if $W \le \min\left\{1, \dfrac{q(X)}{p(X)}\right\}$ **:**
  $X^P \leftarrow X$
  $X^Q \leftarrow X$
  return (X^P, X^Q)
end
else
  $X^P \leftarrow X$
  `// Step 2: Sample from the residual part of` Q`.`
  repeat
    Sample $Y \sim Q$ and $W^* \sim \mathrm{Unif}(0,1)$
  until $W^* > \min\left\{1, \dfrac{p(Y)}{q(Y)}\right\}$
  $X^Q \leftarrow Y$
  return (X^P, X^Q)
end

Algorithm 5: One step of the Metropolis–Hastings algorithm

Input: current state $x \in \mathcal{X}$, target density π, proposal kernel $K(\mathrm{d}y|x)$ with density $k(y \mid x)$
Output: next state $x' \in \mathcal{X}$
`// Step 1: Propose a candidate move.`
Sample $y \sim K(\cdot|x)$
`// Step 2: Compute the MH acceptance probability.`

$$\alpha(x,y) \leftarrow \min\left(1, \frac{\pi(y)k(x \mid y)}{\pi(x)k(y \mid x)}\right).$$

Sample $u \sim \mathrm{Unif}(0,1)$
`// Step 3: Accept or reject the proposal.`
if $u \le \alpha(x,y)$ **:**
  $x' \leftarrow y$
end
else
  $x' \leftarrow x$
end
return x'

then

$$G_k = G_{k+1} + \mathbf{1}\{X_k \notin \{X_{k+1}, \dots, X_C\}\} = G_{k+1} + 1 - \mathbf{1}(A_k).$$

Since $G_C = 1$, summing over k yields the pathwise identity

$$G = C - \sum_{k=1}^{C-1} \mathbf{1}(A_k),$$

and therefore

$$\mathbb{E}[G] = C - \sum_{k=1}^{C-1} \Pr(X_k \in \{X_{k+1}, \dots, X_C\}).$$

Algorithm 6: PFR sampling with proposal measure μ, following Theis and Ahmed [28]

Input : Target measure P, proposal measure μ, and constant $w_{\min} > 0$ such that

$$w_{\min} \le \inf_{x \in \mathcal{X}} \frac{\mathrm{d}\mu}{\mathrm{d}P}(x).$$

Output : Sample $X \sim P$

Generate the atoms

$$\Omega = \{(X_j, S_j)\}_{j \ge 1}$$

of a marked Poisson process on $\mathcal{X} \times \mathbb{R}_+$ with intensity $\mu(\mathrm{d}x) \otimes \mathrm{d}s$

Set $j \leftarrow 1$, $s^\star \leftarrow \infty$, and $j^\star \leftarrow 1$

while true **do**

$$\widetilde{S}_j \leftarrow S_j \frac{\mathrm{d}\mu}{\mathrm{d}P}(X_j)$$

  if $\widetilde{S}_j < s^\star$ **then**

    $s^\star \leftarrow \widetilde{S}_j$

    $j^\star \leftarrow j$

  end

  if $s^\star \le S_j w_{\min}$ **then**

    return $X_{j^\star}$

  end

  $j \leftarrow j + 1$

end

B.2 Proof of Theorem 2.2

We restate the theorem for convenience below.

Theorem (List-Level Coupling)**.** Let $\mu, \nu_1, \dots, \nu_m$ be probability measures on a common measurable space, all absolutely continuous with respect to a common dominating measure λ. With the abuse of notation, we denote their densities with the same letter when no confusion arises.

$$\mu(x) = \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x), \qquad \nu_i(x) = \frac{\mathrm{d}\nu_i}{\mathrm{d}\lambda}(x).$$

Define the barycenter

$$\bar{\nu}(x) \coloneqq \frac{1}{m} \sum_{i=1}^{m} \nu_i(x).$$

Then there exists a coupling such that

$$\Pr\big(X \notin \{Y_1, \dots, Y_m\}\big) = E_m(\mu \| \bar{\nu}),$$

which is optimal. Moreover, this coupling can be implemented without computing $E_m(\mu \| \bar{\nu})$.

Proof. Set

$$\ell(x) \coloneqq \min\{\mu(x), m\bar{\nu}(x)\}.$$

Then

$$1 - E_m(\mu \| \bar{\nu}) = \int \ell(x)\, d\lambda(x),$$

since $\ell(x) = \mu(x) - (\mu(x) - m\bar{\nu}(x))_+$. For each $j \in [m]$, define

$$h_j(x) = \begin{cases} \nu_j(x), & \mu(x) > m\bar{\nu}(x), \\ \mu(x) \dfrac{\nu_j(x)}{m\bar{\nu}(x)}, & \mu(x) \le m\bar{\nu}(x),\ m\bar{\nu}(x) > 0, \\ 0, & \bar{\nu}(x) = 0. \end{cases}$$

Then, we have $\sum_{j=1}^{m} h_j(x) = \ell(x)$ for every x. Indeed, if $\mu(x) > m\bar{\nu}(x)$, then

$$\sum_{j=1}^{m} h_j(x) = \sum_{j=1}^{m} \nu_j(x) = m\bar{\nu}(x) = \ell(x),$$

whereas if $\mu(x) \le m\bar{\nu}(x)$ and $m\bar{\nu}(x) > 0$, then

$$\sum_{j=1}^{m} h_j(x) = \mu(x)\frac{\sum_{j=1}^{m} \nu_j(x)}{m\bar{\nu}(x)} = \mu(x) = \ell(x).$$

When $\bar{\nu}(x) = 0$, both sides are 0. Let

$$\alpha_j \coloneqq \int h_j(x)\, d\lambda(x).$$

Since $0 \le h_j \le \nu_j$ and $\int \nu_j\, d\lambda = 1$, we have $0 \le \alpha_j \le 1$. Define the residual density

$$r_j(x) \coloneqq \begin{cases} \dfrac{\nu_j(x) - h_j(x)}{1 - \alpha_j}, & \alpha_j < 1, \\ 0, & \alpha_j = 1. \end{cases}$$

Let R_j denote the corresponding probability measure when $\alpha_j < 1$; if $\alpha_j = 1$, then $\nu_j = h_j$ a.e. and the residual part is never needed. Observe that

$$\nu_j(x) - h_j(x) = \nu_j(x)\Big(1 - \frac{\mu(x)}{m\bar{\nu}(x)}\Big)_+,$$

with the convention that the right-hand side is 0 when $\bar{\nu}(x) = 0$. Indeed, if $\mu(x) > m\bar{\nu}(x)$ then $h_j(x) = \nu_j(x)$, so both sides are 0; if $\mu(x) \le m\bar{\nu}(x)$ and $m\bar{\nu}(x) > 0$, then

$$\nu_j(x) - h_j(x) = \nu_j(x) - \mu(x)\frac{\nu_j(x)}{m\bar{\nu}(x)} = \nu_j(x)\Big(1 - \frac{\mu(x)}{m\bar{\nu}(x)}\Big).$$

Now construct the coupling as follows. Sample $X \sim \mu$ and independently $U \sim \mathrm{Unif}(0,1)$. Define the hit event

$$H \coloneqq \left\{ U \le \min\Big(1, \frac{m\bar{\nu}(X)}{\mu(X)}\Big) \right\}.$$

On H, choose an index $I \in \{1, \dots, m\}$ conditionally on X according to

$$\Pr(I = i \mid X) = \frac{\nu_i(X)}{m\bar{\nu}(X)}.$$

This is well-defined on H, because H implies $m\bar{\nu}(X) > 0$ whenever $\mu(X) > 0$. Then set $Y_I = X$. For the remaining coordinates $j \ne I$, generate random variables with marginals R_j. On H^c, generate all coordinates Y_j with marginals R_j. Notice that Y_j drawn from the residual laws need not be independent; any joint coupling with the specified marginals is allowed. Independence is merely a convenient implementation choice.

We now verify the properties of this construction.

Membership probability. Since $X \sim \mu$, we can write

$$\Pr(H) = \int \mu(x) \min\Big(1, \frac{m\bar{\nu}(x)}{\mu(x)}\Big)\, \mathrm{d}\lambda(x) = \int \ell(x)\, \mathrm{d}\lambda(x) = 1 - E_m(\mu \| \bar{\nu}).$$

On the event H, one coordinate is set equal to X, namely $Y_I = X$, hence $\Pr\big(X \in \{Y_1, \dots, Y_m\}\big) \ge \Pr(H)$. Conversely, on H^c we have

$$\mu(X) > m\bar{\nu}(X).$$

For any j, if $\mu(x) > m\bar{\nu}(x)$ then $h_j(x) = \nu_j(x)$, so $\nu_j(x) - h_j(x) = 0$, and hence $r_j(x) = 0$ whenever $\mu(x) > m\bar{\nu}(x)$. Therefore each residual law R_j is supported on $\{x : \mu(x) \le m\bar{\nu}(x)\}$, so on H^c none of the residual draws can equal X. Thus

$$\Pr\big(X \in \{Y_1, \dots, Y_m\}\big) = \Pr(H) = 1 - E_m(\mu \| \bar{\nu}),$$

Correct marginals. Fix $i \in [m]$. We need to show that $Y_i \sim \nu_i$. To do that, note that for measurable A,

$$\Pr(X \in A,\ H,\ I = i) = \int_A \mu(x) \min\Big(1, \frac{m\bar{\nu}(x)}{\mu(x)}\Big) \frac{\nu_i(x)}{m\bar{\nu}(x)}\ \mathrm{d}\lambda(x).$$

The integrand simplifies to $h_i(x)$, because:

- if $\mu(x) \le m\bar{\nu}(x)$ and $m\bar{\nu}(x) > 0$, then

$$\mu(x)\frac{\nu_i(x)}{m\bar{\nu}(x)} = \mu(x)\frac{\nu_i(x)}{m\bar{\nu}(x)} = h_i(x);$$

- if $\mu(x) > m\bar{\nu}(x)$, then

$$\mu(x)\frac{m\bar{\nu}(x)}{\mu(x)}\frac{\nu_i(x)}{m\bar{\nu}(x)} = \nu_i(x) = h_i(x).$$

Hence

$$\Pr(X \in A,\ H,\ I = i) = \int_A h_i(x)\ \mathrm{d}\lambda(x).$$

Since $Y_i = X$ on the event $\{H, I = i\}$, this event contributes exactly the measure with density h_i to the law of Y_i.

On the complementary event $\{H, I \neq i\} \cup H^c$, the coordinate Y_i is drawn from the residual law R_i. The probability of this complementary event is

$$1 - \Pr(H, I = i) = 1 - \alpha_i,$$

because $\Pr(H, I = i) = \int h_i(x)\, d\lambda(x) = \alpha_i$. Therefore the contribution of this complementary event to the law of Y_i is

$$(1 - \alpha_i)R_i,$$

whose density is $(1 - \alpha_i)r_i(x) = \nu_i(x) - h_i(x)$.

Combining the two contributions, the law of Y_i has density $h_i(x) + \nu_i(x) - h_i(x) = \nu_i(x)$. Thus $Y_i \sim \nu_i$ for every i.

Optimality. We shall show that $\mathbb{E}_m(\mu\|\bar{\nu})$ is in fact the optimal value for the list membership probability. To that goal, define for any measurable A

$$\Gamma(A) \coloneqq \Pr\big(X \in A,\ X \in \{Y_1, \ldots, Y_m\}\big).$$

Then Γ is a finite measure and clearly have $\Gamma(A) \le \Pr(X \in A) = \mu(A)$. Also, we can write

$$\begin{aligned}
\Gamma(A) &= \Pr\Big(X \in A,\ \bigcup_{i=1}^{m}\{X = Y_i\}\Big) \\
&\le \sum_{i=1}^{m} \Pr(X \in A,\ X = Y_i) \\
&\le \sum_{i=1}^{m} \Pr(Y_i \in A) \\
&= \sum_{i=1}^{m} \nu_i(A) = m\bar{\nu}(A).
\end{aligned}$$

Hence, Γ is dominated by both μ and $m\bar{\nu}$, so its density satisfies $\frac{\mathrm{d}\Gamma}{\mathrm{d}\lambda}(x) \le \min\{\mu(x), m\bar{\nu}(x)\}$ a.e., implying that

$$\Pr\big(X \in \{Y_1, \ldots, Y_m\}\big) = \Gamma(\mathcal{X}) = \int \frac{\mathrm{d}\Gamma}{\mathrm{d}\lambda}(x)\ \mathrm{d}\lambda \le \int \min\{\mu(x), m\bar{\nu}(x)\}\ \mathrm{d}\lambda(x).$$

Equivalently,

$$\Pr\big(X \notin \{Y_1, \ldots, Y_m\}\big) \ge 1 - \int \min\{\mu(x), m\bar{\nu}(x)\}\ \mathrm{d}\lambda(x) = E_m(\mu\|\bar{\nu}).$$

□

C Optimality for Shifted-Exponential Setting

Let $P^0 = \mathrm{Exp}(1)$ and, for $i = 1, \dots, m$, let $P^i = i + \mathrm{Exp}(1)$, where all exponentials have scale 1. Thus

$$p_0(x) = e^{-x}\mathbf{1}\{x \geq 0\}, \qquad p_i(x) = e^{-(x-i)}\mathbf{1}\{x \geq i\}, \qquad i = 1, \dots, m.$$

Poisson matching is optimal. Let μ be any common dominating measure for $P^0, \dots, P^m$, and write

$$f_i(x) = \frac{dP^i}{d\mu}(x), \qquad i = 0, \dots, m.$$

Let

$$\Pi = \{(T_n, W_n)\}_{n \geq 1}$$

be the shared Poisson process on $\mathbb{R}_+ \times \mathbb{R}$ with intensity $dt\,\mu(dw)$. Under the Poisson matching rule, the output corresponding to P^i is

$$X^i = W_{I_i}, \qquad I_i = \arg\min_{n \geq 1} \frac{T_n}{f_i(W_n)}.$$

The key observation is that, on $[1, \infty)$,

$$P^1(dx) = e\,P^0(dx),$$

and therefore

$$f_1(x) = e f_0(x), \qquad x \geq 1,$$

μ-almost everywhere. Also, $f_1(x) = 0$ on $(-\infty, 1)$.

Let (T_*, X^0) be the atom selected by P^0, and define

$$R_* = \frac{T_*}{f_0(X^0)}.$$

By definition of I_0,

$$\frac{T_n}{f_0(W_n)} \geq R_* \qquad \text{for all } n \geq 1. \tag{14}$$

If $X^0 < 1$, then X^0 is outside the support of every P^i, $i \geq 1$. Hence no X^i, $i \geq 1$, can equal X^0, and therefore

$$X^0 \notin \{X^1, \dots, X^m\}.$$

If $X^0 \geq 1$, then $f_1(X^0) = e f_0(X^0)$, so the P^1-score of the same atom is

$$\frac{T_*}{f_1(X^0)} = \frac{T_*}{e f_0(X^0)} = \frac{R_*}{e}.$$

For any other atom with $W_n \geq 1$, we have

$$\frac{T_n}{f_1(W_n)} = \frac{T_n}{e f_0(W_n)} \geq \frac{R_*}{e},$$

where the last inequality follows from equation 14. Atoms with $W_n < 1$ have $f_1(W_n) = 0$, so they cannot be selected by P^1. Therefore P^1 selects the same atom as P^0, and hence

$$X^1 = X^0.$$

Consequently,

$$X^0 \in \{X^1, \dots, X^m\} \quad \Longleftrightarrow \quad X^0 \geq 1.$$

Since $X^0 \sim P^0 = \mathrm{Exp}(0)$,

$$\Pr(X^0 \geq 1) = e^{-1}.$$

Thus, under the shared Poisson matching construction,

$$\Pr\big(X^0 \notin \{X^1, \dots, X^m\}\big) = \Pr(X^0 < 1) = 1 - e^{-1}.$$

or

$$\Pr\big(X^0 \in \{X^1, \dots, X^m\}\big) = e^{-1}.$$

Hockey-stick lower bound. It turns out that the quantity above is also the desired optimal Hockey-Stick divergence.

$$\bar{P} \coloneqq \frac{1}{m}\sum_{i=1}^{m} P^i$$

be the barycenter of $P^1, \ldots, P^m$. The order-m hockey-stick divergence between P^0 and $\bar{P}$ is

$$E_m(P^0 \| \bar{P}) = \int \left(p_0(x) - m\bar{p}(x)\right)_+ dx = \int \left(p_0(x) - \sum_{i=1}^{m} p_i(x)\right)_+ dx.$$

Using

$$p_0(x) = e^{-x}\mathbf{1}\{x \geq 0\}, \qquad p_i(x) = e^{-(x-i)}\mathbf{1}\{x \geq i\},$$

we have

$$p_0(x) - \sum_{i=1}^{m} p_i(x) = e^{-x}\mathbf{1}\{x \geq 0\} - \sum_{i=1}^{m} e^{-(x-i)}\mathbf{1}\{x \geq i\}.$$

For $0 \leq x < 1$, none of $p_1, \ldots, p_m$ are active, so

$$p_0(x) - \sum_{i=1}^{m} p_i(x) = e^{-x} > 0.$$

For $x \geq 1$, the first shifted exponential is active and satisfies

$$p_1(x) = e^{-(x-1)} = e\, p_0(x) > p_0(x).$$

Hence

$$p_0(x) - \sum_{i=1}^{m} p_i(x) < 0, \qquad x \geq 1.$$

Therefore the positive part is supported only on $[0, 1)$, and

$$E_m(P^0 \| \bar{P}) = \int_0^1 e^{-x}\, dx = 1 - e^{-1}.$$

Thus the optimal list-level inclusion probability is

$$1 - E_m(P^0 \| \bar{P}) = e^{-1}.$$

This concludes that, Poisson matching achieves the optimal inclusion probability.

Optimal value of $\mathbf{G}^*$. Fix the left-to-right ordering $P^0, P^1, \ldots, P^m$, and let $C = m + 1$. Recall that

$$\mathbf{G}^* = C - \sup_{\gamma \in \Gamma} \sum_{k=0}^{m-1} \Pr_{\gamma}\left(X^k \in \{X^{k+1}, \ldots, X^m\}\right),$$

where Γ denotes the set of all couplings of $P^0, \ldots, P^m$.

For the shifted-exponential family, the same argument as above applies to every $k = 0, \ldots, m-1$. Indeed, P^{k+1} is just a unit shift of P^k, and on the overlap region $[k+1, \infty)$ its density is e times that of P^k. Therefore, under the shared Poisson matching construction,

$$\Pr\left(X^k \in \{X^{k+1}, \ldots, X^m\}\right) = e^{-1}.$$

The corresponding order-$(m-k)$ hockey-stick bound has the same value, so Poisson matching attains the optimal list-level probability for every term in the left-to-right decomposition. Hence

$$\sup_{\gamma \in \Gamma} \sum_{k=0}^{m-1} \Pr_{\gamma}\left(X^k \in \{X^{k+1}, \ldots, X^m\}\right) = me^{-1}.$$

Since $C = m + 1$, we obtain

$$\mathbf{G}^* = C - (C-1)e^{-1}.$$

Thus the lower bound using this left-to-right ordering is also achievable.

D Proof for the Upperbound of Theorem 2.1

The upperbound of interest can be restated as follow.

Theorem. Let $X^1, \dots, X^C$ be the random variables defined previously, and let

$$G := \big|\{\text{distinct values among } X^1, \dots, X^C\}\big|, \qquad D := \sum_{1 \le i < j \le C} \mathbf{1}\{X^i \neq X^j\}.$$

Then

$$\mathbb{E}[G] \le 1 + \min_{\sigma \in S_C} \sum_{r=1}^{C-1} \Pr\Big(X^{\sigma(r)} \neq X^{\sigma(r+1)}\Big), \tag{15}$$

where S_C denotes the set of all permutations of $\{1, \dots, C\}$. Furthermore, there exists a coupling scheme such that

$$\mathbb{E}[G] \le C - \max_{\sigma \in S_C} \sum_{r=1}^{C-1} \int \frac{\mathrm{d}y}{\int \max\left\{\dfrac{P^{\sigma(r)}(v)}{P^{\sigma(r)}(y)}, \dfrac{P^{\sigma(r+1)}(v)}{P^{\sigma(r+1)}(y)}\right\} \mathrm{d}v} \tag{16}$$

Proof. For any permutation σ of $\{1, \dots, C\}$, the number of distinct values is at most the number of consecutive runs in the ordered sequence $X^{\sigma(1)}, \dots, X^{\sigma(C)}$. Hence, pathwise,

$$G \le 1 + \sum_{r=1}^{C-1} \mathbf{1}\left\{X^{\sigma(r)} \neq X^{\sigma(r+1)}\right\}.$$

This holds for any permutation σ and thus taking the expectation yields the first bound. For the second bound, we use the Poisson matching probability:

$$\Pr\left(X^i \neq X^j\right) = 1 - \int \frac{dy}{\int \max\left\{\frac{P^i(v)}{P^i(y)}, \frac{P^j(v)}{P^j(y)}\right\} dv},$$

which completes the proof. □

E Proof of Lemma 2.4

Lemma (Restatement of Lemma 2.4)**.** Let $P^1, \dots, P^C$ be probability measures on $\mathcal{X}$, and the marked Poisson process defined previously with the common barycenter measure μ above. Then:

$$P^i \ll \mu \qquad \text{and} \qquad \operatorname*{ess\,sup}_{x \in \mathcal{X}} \left(\frac{\mathrm{d}P^i}{\mathrm{d}\mu}\right)(x) \le C, \qquad i = 1, \dots, C. \tag{17}$$

As a result, sampling from each measure has worst-case complexity $O(C)$, and since they all share the same marked Poisson process, the expected number of samples per probability measure is $O(1)$.

Proof. Fix $i \in \{1, \dots, C\}$. By construction,

$$\mu(A) = \frac{1}{C} \sum_{j=1}^{C} P^j(A) \ge \frac{1}{C} P^i(A) \tag{18}$$

for every measurable set $A \subseteq \mathcal{X}$. Therefore, if $\mu(A) = 0$, then necessarily $P^i(A) = 0$, which shows that $P^i \ll \mu$.

Since P^i is absolutely continuous with respect to μ, the Radon–Nikodym derivative $\mathrm{d}P^i/\mathrm{d}\mu$ is well-defined. Moreover, the measure inequality

$$P^i(A) \le C\, \mu(A) \tag{19}$$

holds for every measurable set A. This implies

$$\frac{\mathrm{d}P^i}{\mathrm{d}\mu}(x) \leq C \tag{20}$$

for μ-almost every x. Hence,

$$\operatorname*{ess\,sup}_{x\in\mathcal{X}} \frac{\mathrm{d}P^i}{\mathrm{d}\mu}(x) \leq C, \tag{21}$$

which is also the expected runtime complexity of PMC, concluding the proof. □

F Two-Step Metropolis Hasting Coupling Algorithms

We provide the 2-step MH coupling algorithms for comparision in Algorithm 7.

Algorithm 7: Two-stage coupled kernel for Metropolis–Hastings chains

Input : Current states $x_t^{(1)}, \ldots, x_t^{(C)}$
Output : Updated states $x_{t+1}^{(1)}, \ldots, x_{t+1}^{(C)}$
Set $K_i(\mathrm{d}y) \triangleq k(y \mid x_t^{(i)})\,\mathrm{d}y$, $i = 1, \ldots, C$, and

$$\mu(\mathrm{d}y) = \frac{1}{C}\sum_{i=1}^{C} K_i(\mathrm{d}y).$$

`// Step 1: PML proposal coupling`
Generate $\Omega = \{(S_j, Y_j,)\}_{j\geq 1}$ with intensity measure $\mathrm{d}s \otimes \mu$
for $i = 1, \ldots, C$ **do**

$$J^{(i)} = \arg\min_{j\geq 1} \left\{ S_j \left(\frac{\mathrm{d}K_i}{\mathrm{d}\mu}(Y_j) \right)^{-1} \right\}$$

$$Y_t^{(i)} \leftarrow Y_{J^{(i)}}$$

end
`// Step 2: Shared accept-reject`
$U_t \sim \mathrm{Unif}(0, 1)$
for $i = 1, \ldots, C$ **do**
  if $U_t \leq \alpha(x_t^{(i)}, Y_t^{(i)})$ **then**
    $x_{t+1}^{(i)} \leftarrow Y_t^{(i)}$
  else
    $x_{t+1}^{(i)} \leftarrow x_t^{(i)}$
  end
end

G Two-Stage Coupled Kernel

This construction is given in Algorithm 7, Appendix F, where in the first step, we couple the proposal kernels $K(\,\mathrm{d}y\cdot \mid x_t^{(i)}) \triangleq k(y|x)\,\mathrm{d}y$ using the adaptive Poisson matching method developed in Section 2.2. In the second step, we couple the accept/reject decisions using a shared uniform random variable which achieves optimal pairwise matching probability for Bernouli random variables.

Pairwise Matching Probability. For any two distinct chains $x_t^{(a)} \neq x_t^{(b)}$, assume that the proposal kernel admits a density $k(x, y)$ with respect to Lebesgue measure. Then the event

$$\{x_{t+1}^{(a)} = x_{t+1}^{(b)}\}$$

can occur with nonzero probability only if both chains obtain and accept the same proposed value. Indeed, the alternatives

$$(x_{t+1}^{(a)}, x_{t+1}^{(b)}) = (x_t^{(a)}, x_t^{(b)}), \qquad (x_{t+1}^{(a)}, x_{t+1}^{(b)}) = (Y_t^{(a)}, x_t^{(b)}), \qquad (x_{t+1}^{(a)}, x_{t+1}^{(b)}) = (x_t^{(a)}, Y_t^{(b)})$$

cannot yield equality except on null events, such as $Y_t^{(a)} = x_t^{(b)}$ or $Y_t^{(b)} = x_t^{(a)}$, which have probability zero by continuity. Hence, writing

$$\mathbb{P}_{\text{2-step}} \coloneqq \Pr\Big(x_{t+1}^{(a)} = x_{t+1}^{(b)}\Big),$$

we have

$$\mathbb{P}_{\text{2-step}} \geq \mathbb{P}_{\text{2,bound}},$$

where

$$\mathbb{P}_{\text{2,bound}} \coloneqq \int \min\Big\{\alpha(x_t^{(a)}, y), \alpha(x_t^{(b)}, y)\Big\} \frac{k(y|x_t^{(a)})k(y|x_t^{(b)})}{k(y|x_t^{(a)}) + k(y|x_t^{(b)})}\,\mathrm{d}y. \tag{22}$$

See Appendix H.1 for comparison to the joint method. To show the above bound, consider the following steps:

$$\begin{aligned}
\mathbb{P}_{\text{2-step}} &= \Pr\Big(x_{t+1}^{(a)} = x_{t+1}^{(b)}\Big) && (23)\\
&= \Pr\Big(U_t \leq \alpha(x_t^{(a)}, Y_t^{(a)}),\ U_t \leq \alpha(x_t^{(b)}, Y_t^{(b)}),\ Y_t^{(a)} = Y_t^{(b)}\Big) && (24)\\
&= \int_{y \notin \{x_t^{(a)}, x_t^{(b)}\}} \min\Big\{\alpha(x_t^{(a)}, y), \alpha(x_t^{(b)}, y)\Big\}\ \Pr\Big(Y_t^{(a)} = Y_t^{(b)} \in \mathrm{d}y\Big) && (25)\\
&\geq \int_{y \notin \{x_t^{(a)}, x_t^{(b)}\}} \min\Big\{\alpha(x_t^{(a)}, y), \alpha(x_t^{(b)}, y)\Big\} \frac{k(y|x_t^{(a)})k(y|x_t^{(b)})}{k(y|x_t^{(a)}) + k(y|x_t^{(b)})}\,\mathrm{d}y && (26)\\
&\triangleq \mathbb{P}_{\text{2,bound}}. && (27)
\end{aligned}$$

Here, $\Pr\Big(Y_t^{(a)} = Y_t^{(b)} \in \mathrm{d}y\Big)$ denotes the diagonal measure induced by the PML coupling. The inequality is due to Poisson Matching Lemma.

H Matching Probability of Joint Coupled Kernel

Pairwise Matching Probability. For two chains indexed by a and b, the resulting matching probability $\mathbb{P}_{\text{direct}} \triangleq \Pr\Big(x_{t+1}^{(a)} = x_{t+1}^{(b)}\Big)$ admits the following lower bound $\mathbb{P}_{\text{1,bound}}$:

$$\mathbb{P}_{\text{1,bound}} = \int \frac{\alpha(x_t^{(a)}, y)k(y \mid x_t^{(a)})\,\alpha(x_t^{(b)}, y)k(y \mid x_t^{(b)})}{\alpha(x_t^{(a)}, y)k(y \mid x_t^{(a)}) + \alpha(x_t^{(b)}, y)k(y \mid x_t^{(b)})}\,\mathrm{d}y. \tag{28}$$

In particular,

$$\begin{aligned}
\mathbb{P}_{\text{direct}} &\triangleq \Pr\Big(x_{t+1}^{(a)} = x_{t+1}^{(b)}\Big) && (29)\\
&= \Pr\Big(Y_t^{(a)} = Y_t^{(b)},\ U_t^{(a)} = U_t^{(b)} = 1\Big) && (30)\\
&= \int \Pr\Big(Y_t^{(a)} = Y_t^{(b)} = y,\ U_t^{(a)} = U_t^{(b)} = 1\Big)\,\mathrm{d}y && (31)\\
&= \int \Pr\Big(Y_t^{(a)} = Y_t^{(b)},\ U_t^{(a)} = U_t^{(b)} \mid Y_t^{(a)} = y, U_t^{(a)} = 1\Big)\ \alpha(x_t^{(a)}, y)k(y \mid x_t^{(a)})\,\mathrm{d}y. && (32)
\end{aligned}$$

Applying PML, we have the lower bound

$$\begin{aligned}
\mathbb{P}_{\text{direct}} &\geq \int \frac{\alpha(x_t^{(a)}, y)k(y \mid x_t^{(a)})\,\alpha(x_t^{(b)}, y)k(y \mid x_t^{(b)})}{\alpha(x_t^{(a)}, y)k(y \mid x_t^{(a)}) + \alpha(x_t^{(b)}, y)k(y \mid x_t^{(b)})}\,\mathrm{d}y && (33)\\
&\triangleq \mathbb{P}_{\text{1,bound}}. && (34)
\end{aligned}$$

H.1 Compare Matching Probability Bounds

Theorem H.1. *Let $\mathbb{P}_{1,\mathrm{bound}}$ denote the matching probability bound obtained using PMC applied directly to the Metropolis–Hastings transition kernel, and let $\mathbb{P}_{2,\mathrm{bound}}$ denote the matching probability induced by the two-step construction described above. Then we have $\mathbb{P}_{1,\mathrm{bound}} \geq \mathbb{P}_{2,\mathrm{bound}}$.*

Proof. It is sufficient to prove that for every y,

$$\frac{\alpha(x^A,y)k(y \mid x^A)\,\alpha(x^B,y)k(y \mid x^B)}{\alpha(x^A,y)k(y \mid x^A)+\alpha(x^B,y)k(y \mid x^B)} \geq \min\{\alpha(x^A,y),\alpha(x^B,y)\}\frac{k(y \mid x^A)k(y \mid x^B)}{k(y \mid x^A)+k(y \mid x^B)}.$$

Using the shorthand

$$\alpha^A=\alpha(x^A,y), \qquad \alpha^B=\alpha(x^B,y), \qquad k^A=k(y \mid x^A), \qquad k^B=k(y \mid x^B),$$

and assuming without loss of generality that $\alpha^A \leq \alpha^B$, the inequality becomes

$$\frac{\alpha^A k^A\,\alpha^B k^B}{\alpha^A k^A+\alpha^B k^B} \geq \alpha^A\,\frac{k^A k^B}{k^A+k^B}.$$

If $\alpha^A k^A k^B = 0$, the inequality is trivial. Otherwise, canceling the positive factor $\alpha^A k^A k^B$ yields

$$\frac{\alpha^B}{\alpha^A k^A+\alpha^B k^B} \geq \frac{1}{k^A+k^B}.$$

Since all terms are nonnegative, cross-multiplication gives

$$\alpha^B(k^A+k^B) \geq \alpha^A k^A+\alpha^B k^B \iff \alpha^B k^A \geq \alpha^A k^A,$$

which holds because $k^A \geq 0$ and $\alpha^B \geq \alpha^A$. This completes the proof. □

I Proof of Lemma 3.1

Lemma I.1. *Consider two chains at states x_t^1 and x_t^2 respectively. At $t+1$, let $\mathbb{P}_{\mathrm{aug}}(x_t^1,x_t^2)$ denote the meeting probability under the product-augmented construction, and let $\mathbb{P}_{\mathrm{direct}}(x_t^1,x_t^2)$ denote the meeting probability under the original direct construction. Then*

$$\mathbb{P}_{\mathrm{aug}}(x_t^1,x_t^2) \geq \mathbb{P}_{\mathrm{direct}}(x_t^1,x_t^2).$$

Proof. We will show later that, for every u in the continuous accepted support, one can write the following conditional matching probability for the direct matching strategy:

$$\Pr_{\mathrm{direct}}(X_{t+1}^2=x \mid X_{t+1}^1=x, X_t^1=x_t^1, X_t^2=x_t^2)=\frac{1}{1+\beta(x)} \tag{35}$$

For the augmented strategy:

$$\Pr_{\mathrm{aug}}(X_{t+1}^2=x \mid X_{t+1}^1=x, X_t^1=x_t^1, X_t^2=x_t^2) \tag{36}$$

$$=\Pr_{\mathrm{aug}}(Y_{t+1}^2=x, U_{t+1}^2=1, \mid Y_{t+1}^1=x, U_{t+1}^1=1, X_t^1=x_t^1, X_t^2=x_t^2) \tag{37}$$

$$=\frac{1}{1+\beta'(1,x)} \tag{38}$$

where the exact form of β, β' will be provided later. Furthermore, we have

$$\beta'(1,x) \leq \beta(x).$$

Therefore,

$$\frac{1}{1+\beta'(1,x)} \geq \frac{1}{1+\beta(x)}. \tag{39}$$

So we have, for any x:

$$\Pr_{\mathrm{aug}}(X_{t+1}^2=x \mid X_{t+1}^1=x, X_t^1=x_t^1, X_t^2=x_t^2) \geq \Pr_{\mathrm{direct}}(X_{t+1}^2=x \mid X_{t+1}^1=x, X_t^1=x_t^1, X_t^2=x_t^2)$$

and since $P(X_{t+1}^1=x \mid X_t^1=x_t^1, X_t^2=x_t^2)$ is the same these two chains, integrate w.r.t. this gives the desired inequality. □

I.1 Proof of conditional matching probabilities

We prove the result by directly comparing the two matching probabilities. For notational simplicity, we use a different notation in this proof. With a slight abuse of notation, we preserve the subscripts a and b for the two chains while omitting the explicit dependence on time t, i.e. let $x_t^1 = x_a$ and $x_t^2 = x_b$ denote their current states.

Assume $x_a \neq x_b$, and let λ be a common dominating measure for the proposal parts of the two kernels. For the direct kernel between two chains at state x_a and x_b, define

$$\begin{aligned}
P(\mathrm{d}y) &\triangleq \kappa(\mathrm{d}y \mid x_a) = \alpha(x_a, y)K(\mathrm{d}y \mid x_a) + r(x_a)\delta_{x_a}(\mathrm{d}y), \text{ (chain } a \text{ kernel)} \\
Q(\mathrm{d}y) &\triangleq \kappa(\mathrm{d}y \mid x_b) = \alpha(x_b, y)K(\mathrm{d}y \mid x_b) + r(x_b)\delta_{x_b}(\mathrm{d}y), \text{ (chain } b \text{ kernel)} \\
P_a &\triangleq r(x_a), \qquad Q_b \triangleq r(x_b), \\
p(y)\lambda(\mathrm{d}y) &\triangleq \alpha(x_a, y)K(\mathrm{d}y \mid x_a), \qquad q(y)\lambda(\mathrm{d}y) \triangleq \alpha(x_b, y)K(\mathrm{d}y \mid x_b), \\
a(y)\lambda(\mathrm{d}y) &\triangleq \big(1 - \alpha(x_a, y)\big)K(\mathrm{d}y \mid x_a), \qquad b(y)\lambda(\mathrm{d}y) \triangleq \big(1 - \alpha(x_b, y)\big)K(\mathrm{d}y \mid x_b).
\end{aligned}$$

Thus

$$P(\mathrm{d}y) = P_a\delta_{x_a}(\mathrm{d}y) + p(y)\lambda(\mathrm{d}y), \qquad Q(\mathrm{d}y) = Q_b\delta_{x_b}(\mathrm{d}y) + q(y)\lambda(\mathrm{d}y).$$

Direct Matching. Under the notation above, the matching probability at kernel level is equivalent to that with P and Q. In particular, recall the two probability measures

$$P(\mathrm{d}y) = P_a\,\delta_{x_a}(\mathrm{d}y) + p(y)\lambda(\mathrm{d}y), \qquad Q(\mathrm{d}y) = Q_b\,\delta_{x_b}(\mathrm{d}y) + q(y)\lambda(\mathrm{d}y),$$

where $P_a = P(\{x_a\})$, $Q_b = Q(\{x_b\})$, and p, q are the densities of the non-atomic parts with respect to λ. Hence

$$\int p(v)\lambda(\mathrm{d}v) = 1 - P_a, \qquad \int q(v)\lambda(\mathrm{d}v) = 1 - Q_b.$$

Let $X_P \sim P$ and $X_Q \sim Q$, then we have:

$$\Pr(X_Q = u | X_P = u) = \Pr(X_{t+1}^2 = u | X_{t+1}^1 = u, X_t^2 = x_b, X_t^1 = x_a) \tag{40}$$

For direct matching with the Poisson Matching Lemma (PML), we choose a proposal measure μ satisfying $P, Q \ll \mu$. In particular, take

$$\mu(\mathrm{d}y) = \eta_a\,\delta_{x_a}(\mathrm{d}y) + \eta_b\,\delta_{x_b}(\mathrm{d}y) + t(y)\lambda(\mathrm{d}y),$$

where $\eta_a, \eta_b > 0$, and $t(y) > 0$ on the support of $p + q$. Define

$$f = \frac{\mathrm{d}P}{\mathrm{d}\mu}, \qquad g = \frac{\mathrm{d}Q}{\mathrm{d}\mu}.$$

Then

$$f(x_a) = \frac{P_a}{\eta_a}, \qquad f(x_b) = 0, \qquad f(y) = \frac{p(y)}{t(y)},$$

and

$$g(x_a) = 0, \qquad g(x_b) = \frac{Q_b}{\eta_b}, \qquad g(y) = \frac{q(y)}{t(y)}.$$

For u in the continuous support with $p(u), q(u) > 0$, define

$$\beta(u) \triangleq f(u) \int \left(\frac{g(v)}{g(u)} - \frac{f(v)}{f(u)}\right)_+ \mu(\mathrm{d}v).$$

This is the quantity appearing in Eq. (16) of [20], and the conditional meeting probability is

$$\Pr(\widetilde{U}_Q = \widetilde{U}_P \mid \widetilde{U}_P = u) = \frac{1}{1 + \beta(u)},$$

where $\widetilde{U}_P$ and $\widetilde{U}_Q$ are the values selected according to the PML from the common marked point process $\Pi = \{(T_i, U_i)\}_{i=1}^{\infty}$, with $U_i \sim \mu$. In our case, those are X_P and X_Q respectively.

We now compute $\beta(u)$. Splitting the integral into the atoms x_a, x_b and the non-atomic part gives

$$\beta(u) = Q_b \frac{p(u)}{q(u)} + p(u) \int \left(\frac{q(v)}{q(u)} - \frac{p(v)}{p(u)} \right)_+ \lambda(\mathrm{d}v).$$

Indeed, at $v = x_a$, we have $g(x_a) = 0$, so the positive part vanishes. At $v = x_b$, we have $f(x_b) = 0$ and $g(x_b) = Q_b/\eta_b$, hence the atomic contribution is

$$f(u)\,\eta_b\,\frac{g(x_b)}{g(u)} = \frac{p(u)}{t(u)}\,\eta_b\,\frac{Q_b/\eta_b}{q(u)/t(u)} = Q_b \frac{p(u)}{q(u)}.$$

For the non-atomic part,

$$f(v) = \frac{p(v)}{t(v)}, \qquad g(v) = \frac{q(v)}{t(v)}.$$

Therefore

$$\begin{aligned} & f(u) \int \left(\frac{g(v)}{g(u)} - \frac{f(v)}{f(u)} \right)_+ t(v)\lambda(\mathrm{d}v) \\ &= \frac{p(u)}{t(u)} \int \left[\frac{t(u)}{t(v)} \left(\frac{q(v)}{q(u)} - \frac{p(v)}{p(u)} \right) \right]_+ t(v)\lambda(\mathrm{d}v) \\ &= p(u) \int \left(\frac{q(v)}{q(u)} - \frac{p(v)}{p(u)} \right)_+ \lambda(\mathrm{d}v). \end{aligned}$$

Adding the atomic and non-atomic contributions yields the displayed formula for $\beta(u)$.

Augmented Matching. Now consider the augmented space $\{0,1\} \times \mathcal{X}$, where the first coordinate records whether the proposal is accepted. Define

$$\begin{aligned} P'(\mathrm{d}j, \mathrm{d}y) &\triangleq K(\mathrm{d}y \mid x_a)\,\mathrm{Ber}\big(\alpha(x_a, y)\big)(\mathrm{d}j), \\ Q'(\mathrm{d}j, \mathrm{d}y) &\triangleq K(\mathrm{d}y \mid x_b)\,\mathrm{Ber}\big(\alpha(x_b, y)\big)(\mathrm{d}j). \end{aligned}$$

Equivalently, using the notation introduced above,

$$\begin{aligned} P'(\{1\}, \mathrm{d}y) &= p(y)\lambda(\mathrm{d}y), & Q'(\{1\}, \mathrm{d}y) &= q(y)\lambda(\mathrm{d}y), \\ P'(\{0\}, \mathrm{d}y) &= a(y)\lambda(\mathrm{d}y), & Q'(\{0\}, \mathrm{d}y) &= b(y)\lambda(\mathrm{d}y). \end{aligned}$$

Compared to the kernel-level notation, our event of interest is:

$$\begin{aligned} &\Pr(Y_Q = u, U_Q = 1 | Y_P = u, U_P = 1) && (41) \\ &\quad = \Pr(Y^2_{t+1} = u, U^2_{t+1} = 1 | Y^1_{t+1} = u, U^1_{t+1} = 1, X^2_t = x_b, X^1_t = x_a) && (42) \end{aligned}$$

Let μ' be a dominating reference measure such that $P', Q' \ll \mu'$, and define

$$f'(j, y) = \frac{\mathrm{d}P'}{\mathrm{d}\mu'}(j, y), \qquad g'(j, y) = \frac{\mathrm{d}Q'}{\mathrm{d}\mu'}(j, y).$$

We are interested in the accepted-proposal matching event, which corresponds to the branch $j = 1$. Applying Eq. (16) of [20], for $p(u), q(u) > 0$, define

$$\beta'(1, u) \triangleq f'(1, u) \int \left(\frac{g'(j, v)}{g'(1, u)} - \frac{f'(j, v)}{f'(1, u)} \right)_+ \mu'(\mathrm{d}j, \mathrm{d}v).$$

The corresponding conditional meeting probability is

$$\Pr(\widetilde{U}_{Q'} = \widetilde{U}_{P'} \mid \widetilde{U}_{P'} = (1, u)) = \frac{1}{1 + \beta'(1, u)}.$$

Splitting the integral over the two branches $j = 0$ and $j = 1$, write

$$\beta'(1, u) = I_0 + I_1,$$

where

$$I_0 \triangleq f'(1,u)\int\left(\frac{g'(0,v)}{g'(1,u)}-\frac{f'(0,v)}{f'(1,u)}\right)_+\mu'(\{0\},\mathrm{d}v),$$
$$I_1 \triangleq f'(1,u)\int\left(\frac{g'(1,v)}{g'(1,u)}-\frac{f'(1,v)}{f'(1,u)}\right)_+\mu'(\{1\},\mathrm{d}v).$$

First consider I_1. By the definition of the Radon–Nikodym derivatives,

$$f'(1,v)\mu'(\{1\},\mathrm{d}v)=p(v)\lambda(\mathrm{d}v),\qquad g'(1,v)\mu'(\{1\},\mathrm{d}v)=q(v)\lambda(\mathrm{d}v).$$

Moreover, for $p(u),q(u)>0$,

$$\frac{f'(1,u)}{g'(1,u)}=\frac{p(u)}{q(u)}.$$

Therefore

$$\begin{aligned}I_1&=f'(1,u)\int\left(\frac{g'(1,v)}{g'(1,u)}-\frac{f'(1,v)}{f'(1,u)}\right)_+\mu'(\{1\},\mathrm{d}v)\\&=\int\left(\frac{f'(1,u)}{g'(1,u)}g'(1,v)-f'(1,v)\right)_+\mu'(\{1\},\mathrm{d}v)\\&=\int\left(\frac{p(u)}{q(u)}q(v)-p(v)\right)_+\lambda(\mathrm{d}v)\\&=p(u)\int\left(\frac{q(v)}{q(u)}-\frac{p(v)}{p(u)}\right)_+\lambda(\mathrm{d}v).\end{aligned}$$

Now consider I_0. Similarly,

$$f'(0,v)\mu'(\{0\},\mathrm{d}v)=a(v)\lambda(\mathrm{d}v),\qquad g'(0,v)\mu'(\{0\},\mathrm{d}v)=b(v)\lambda(\mathrm{d}v).$$

Thus

$$\begin{aligned}I_0&=f'(1,u)\int\left(\frac{g'(0,v)}{g'(1,u)}-\frac{f'(0,v)}{f'(1,u)}\right)_+\mu'(\{0\},\mathrm{d}v)\\&=\int\left(\frac{f'(1,u)}{g'(1,u)}g'(0,v)-f'(0,v)\right)_+\mu'(\{0\},\mathrm{d}v)\\&=\int\left(\frac{p(u)}{q(u)}b(v)-a(v)\right)_+\lambda(\mathrm{d}v)\\&=p(u)\int\left(\frac{b(v)}{q(u)}-\frac{a(v)}{p(u)}\right)_+\lambda(\mathrm{d}v).\end{aligned}$$

Using $(x-y)_+\le x$ for $x,y\ge 0$, we obtain

$$\begin{aligned}I_0&\le p(u)\int\frac{b(v)}{q(u)}\lambda(\mathrm{d}v)\\&=\frac{p(u)}{q(u)}\int b(v)\lambda(\mathrm{d}v)\\&=Q_b\frac{p(u)}{q(u)}.\end{aligned}$$

Therefore

$$\begin{aligned}\beta'(1,u)&=I_0+I_1\\&\le Q_b\frac{p(u)}{q(u)}+p(u)\int\left(\frac{q(v)}{q(u)}-\frac{p(v)}{p(u)}\right)_+\lambda(\mathrm{d}v)\\&=\beta(u),\end{aligned}$$

which completes the proof.

J Additional Experiments

Riemann manifold Metropolis-adjusted Langevin Beyond random-walk Metropolis–Hastings, we also study a geometry-dependent proposal that adapts to local structure of the target. In particular, we consider a Riemannian manifold MALA-type proposal [12] on a two-dimensional banana distribution, which is defined by the density

$$\pi(x_1, x_2) \propto \exp\left\{-\frac{x_1^2}{2\sigma_1^2} - \frac{\left(x_2 + b(x_1^2 - \sigma_1^2)\right)^2}{2\sigma_2^2}\right\}, \qquad x = (x_1, x_2) \in \mathbb{R}^2,$$

where $b > 0$ controls the curvature of the banana shape. In our experiment, we set $b = 0.05$.

The Riemannian proposal at the current state x is

$$y \sim \mathcal{N}\left(x + \frac{\sigma^2}{2} G(x)^{-1} \nabla \log \pi(x),\, \sigma^2 G(x)^{-1}\right),$$

where $G(x) = -\nabla^2 \log \pi(x)$ is the local metric tensor. This proposal adapts to the local curvature of the target and is followed by the standard Metropolis–Hastings accept–reject step. In our experiment, we set $\sigma = 0.4$ and initialize the chains from $\pi_0 = \mathrm{Unif}([-2, 2]^d)$. As shown in Figure 5, we observe the same trend as in Section 4: Poisson matching outperforms the pairwise baselines in terms of average meeting time. Furthermore, the gap between the two-step and one-step methods is smaller than in the random-walk experiments. We attribute this to the higher acceptance rate in this regime, which reduces the difference between coupling proposals before and after the Metropolis–Hastings accept–reject step. Overall, this experiment suggests that our method can be adapted beyond random-walk proposals to geometry-dependent settings.

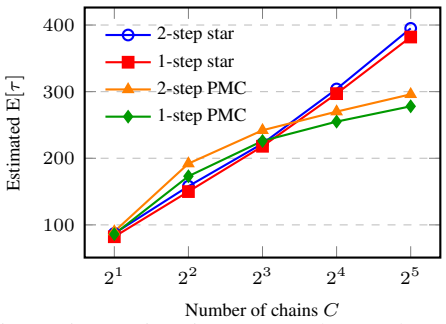


Figure 5: Meeting time versus the number of coupled chains for Riemann MMALA.

J.1 Convergence Diagnostic

We now provide an application of our method for problems in convergence diagnostic.

J.1.1 Grand-Coupling Diagnostic

This line of work follows the discussion in Section 3 in the main text. First, we restate the classical results by Johnson [16].

Lemma J.1 (Johnson [16])**.** *Let κ be a Markov kernel with invariant distribution π, and let $X_0^{(1)}, \ldots, X_0^{(C)} \overset{\text{iid}}{\sim} \pi_0$. Define*

$$\omega := \sup\{a \in [0, 1] : \pi_0 \geq a\pi\}.$$

For a grand coupling of the C chains, let τ be the first time at which all chains coalesce. Then

$$\|\pi_t - \pi\|_{\mathrm{TV}} \leq \frac{\Pr(\tau > t)}{1 - (1 - \omega)^C},$$

regardless of the coupling strategy.

We observe that for sufficiently large C, the denominator $1 - (1 - \omega)^C$ is close to 1. In this regime, the meeting-time tail $\Pr(\tau > t)$ becomes a direct proxy for the distributional bias $\|\pi_t - \pi\|_{\mathrm{TV}}$. This also highlights why faster meeting times are desirable: the bound is controlled by the tail probability $\Pr(\tau > t)$.

However, this diagnostic depends on knowledge of ω. In many settings, ω is not easy to determined and may be zero, making the bound vacuous. For example, if π_0 is Gaussian while π is heavy-tailed, then π_0 cannot dominate any positive multiple of π uniformly over the state space. We now provide an alternative bound using the idea of list-level coupling.

Table 1: Comparison between the additive bias term $1-\alpha_C$ from our list-level bound and the Johnson denominator $1-(1-\omega)^C$ for $\pi=\mathcal{N}(0,I_d)$ and $\pi_0=\mathcal{N}(\mathbf{1}_d,16I_d)$.

$d=1$			$d=2$			$d=3$		
C	$1-\alpha_C$	Johnson	C	$1-\alpha_C$	Johnson	C	$1-\alpha_C$	Johnson
2	0.5673	0.4251	2	0.7468	0.1135	2	0.8538	0.0281
8	0.2627	0.8908	8	0.4748	0.3824	8	0.6661	0.1077
16	0.1526	0.9881	16	0.3289	0.6186	16	0.5292	0.2037
32	0.0846	0.9999	32	0.2039	0.8546	32	0.3934	0.3660
64	0.0439	1.0000	64	0.1149	0.9788	64	0.2619	0.5980
128	0.0242	1.0000	128	0.0653	0.9996	128	0.1549	0.8384

Lemma J.2. *In the setting stated above, for any faithful grand coupling of the C chains,*

$$\|\pi_t-\pi\|_{\mathrm{TV}}\le 1-\alpha_C+\Pr(\tau>t),$$

where α_C is an achievable probability of the initial list-level inclusion event

$$\{Y_0\in\{X_0^{(1)},\dots,X_0^{(C)}\}\},\qquad Y_0\sim\pi,\quad X_0^{(i)}\sim\pi_0.$$

Using list-level Poisson matching [20, 26], one may take

$$\alpha_C\ge\mathbb{E}_{X\sim\pi_0}\left[\frac{C\pi(X)}{C\pi_0(X)+\pi(X)}\right].$$

The proof for this Lemma is at the end of this part. We note that the lower bound on α_C can be estimated by Monte Carlo, since it is an expectation under $X\sim\pi_0$. Moreover, this quantity is well-defined and bounded between 0 and 1. While this form is convenient in a wider range of scenarios, it introduces an additive term in the upper bound; improving this term, for instance using ideas from Biswas et al. [2], is an interesting direction for future work. We now provide some examples of convergence diagnostic analysis using the RWMH examples in Section 4.

Student-t RWMH. This setting considers a heavy-tailed Cauchy target sampled using a Student-t random-walk proposal with 2 degrees of freedom, where the chains are initialized from $\pi_0=\mathcal{N}(0,\mathrm{I}_d)$. Note that in this case, Lemma J.1 is not applicable because the relevant rejection coefficient is unbounded, as discussed above. We provide the estimated bias in Figure 6 (left), where we test dimension $d=1$. Overall, the results suggest that using a sufficiently large number of chains can substantially improve the diagnostic bound in this heavy-tailed setting; among the tested values, $C=16$ gives the tightest bound by reducing the additive bias term associated with α_C in Lemma 6 (left). We note, however, that α_C need not approach 1 even for large C; for example, if π_0 is uniform on a bounded set while π has mass outside this support, then the list-level inclusion probability is bounded away from 1. Thus this bias may persist, but Lemma J.2 still provides a useful heuristic for understanding the behavior of the coupled chains, and additional chain-level statistics may be used to further reduce or calibrate it. For example, one can switch to another diagnostic method when the bound becomes saturated. We illustrate this idea in the next experiment.

Gaussian RWMH. We recall the Gaussian target setting $\pi=\mathcal{N}(0,\mathrm{I}_d)$, with chains initialized from $\pi_0=\mathcal{N}(\mathbf{1}_d,16\mathrm{I}_d)$. This setting is more tractable and allows us to illustrate how one can combine the bounds from Lemma J.1 and Lemma J.2. First, Table 1 shows how the additive bias term $1-\alpha_C$ in Lemma J.2 and Johnson's denominator $1-(1-\omega)^C$ vary with the number of chains C. As expected in this example, $1-\alpha_C$ decreases toward 0 as C increases, while Johnson's denominator increases toward 1.

Figure 6 (right) reports the resulting bounds for $d=3$ and $C\in\{32,64\}$. In this case, $C=64$ yields a tighter bound for both approaches. Furthermore, the bound from Lemma J.2 is tighter than the one from Lemma J.1 for $t<50$, where the additive bias term is less restrictive than Johnson's multiplicative denominator. For larger t, Johnson's bound can become tighter as the meeting-time tail decreases. Thus, in this scenario, it is natural to take the minimum of the two bounds rather than relying on a single diagnostic metric.

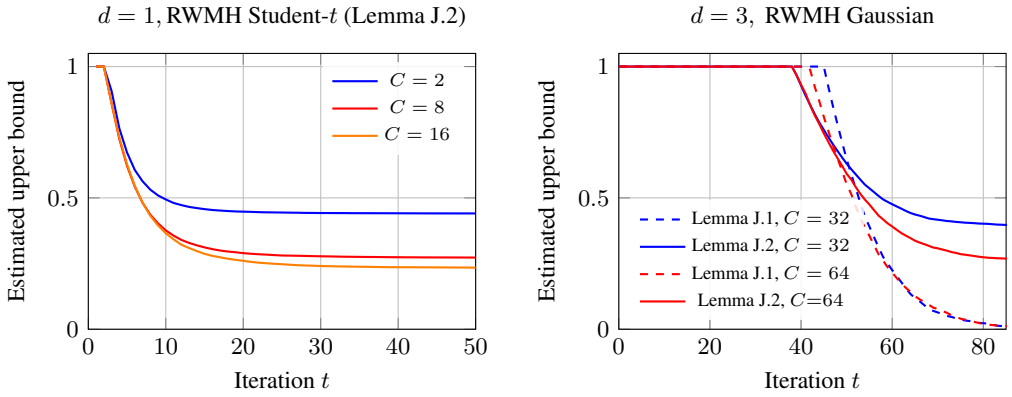


Figure 6: (Left) Estimated upper bounds for $|\pi_t - \pi|_{\mathrm{TV}}$ using Lemma J.2 (Right) Estimated upper bounds $|\pi_t - \pi|_{\mathrm{TV}}$ comparing Johnson's bound [16] and the list-level bound, i.e. Lemma J.2, in the Gaussian target setting with $d = 3$.

Proof of Lemma J.2 We introduce the proof of Lemma J.2 as follows.

Proof. Our goal is to generate $X_0^{(1)}, \ldots, X_0^{(C)} \sim \pi_0$ independently, together with an auxiliary stationary chain $Y_0 \sim \pi$, such that

$$\Pr\Big(Y_0 \in \{X_0^{(1)}, \ldots, X_0^{(C)}\}\Big) \geq \alpha_C.$$

To do this, we draw a Poisson point process

$$\Pi = \{(X_i, S_i)\}_{i \geq 1},$$

where the marks X_i have base distribution π_0 and S_i denotes the arrival time. We take the first C marked points as the initial chains, $X_0^{(1)}, \ldots, X_0^{(C)}$, and select Y_0 using the Poisson Monte Carlo procedure described in Section 2.2. By the list-level Poisson matching lemma [20], this construction attains an inclusion probability α_C satisfying the lower bound stated above. Other Poisson-based coupling constructions, such as Rowan et al. [26], may also be used.

We define the auxiliary chain Y_t in the following way. At time t, first update the observed list $X_t^{1:C}$ to $X_{t+1}^{1:C}$ using the original grand coupling. Then:

- if $Y_t = X_t^{(i)}$ for some i, choose one such index $i^\star$ and set $Y_{t+1} = X_{t+1}^{(i^\star)}$;
- if $Y_t \notin \{X_t^{(1)}, \ldots, X_t^{(C)}\}$, draw $Y_{t+1} \sim \kappa(\cdot|Y_t)$ independently of the observed update.

Then, under this construction, we have $Y_{t+1} \sim \kappa(\cdot|Y_t)$ and thus Y_t is at stationary. We now prove the event implication. On

$$A = \{Y_0 \in \{X_0^{(1)}, \ldots, X_0^{(C)}\}\},$$

choose an index I such that

$$Y_0 = X_0^{(I)}.$$

We claim that

$$Y_s = X_s^{(I)} \qquad \text{for all } s \geq 0.$$

This holds at $s = 0$. If $Y_s = X_s^{(I)}$, then Y_s belongs to the current list. The auxiliary update chooses some index $i^\star$ with $Y_s = X_s^{(i^\star)}$, hence $X_s^{(i^\star)} = X_s^{(I)}$. By faithfulness of the observed grand coupling,

$$X_{s+1}^{(i^\star)} = X_{s+1}^{(I)}.$$

Since $Y_{s+1} = X_{s+1}^{(i^\star)}$, we get $Y_{s+1} = X_{s+1}^{(I)}$.

Therefore $Y_s = X_s^{(I)}$ for all s. If also $\tau \le t$, then all observed chains have coalesced by time t. Thus for any fixed j,

$$X_t^{(j)} = X_t^{(I)} = Y_t.$$

Therefore,

$$A \cap \{\tau \le t\} \subseteq \{X_t^{(j)} = Y_t\},$$

or equivalently,

$$\{X_t^{(j)} \neq Y_t\} \subseteq A^c \cup \{\tau > t\}.$$

Since $X_t^{(j)}$ has law π_t and $Y_t \sim \pi$, the coupling inequality gives

$$\|\pi_t - \pi\|_{\mathrm{TV}} \le \Pr(X_t^{(j)} \neq Y_t).$$

Combining this with the event inclusion gives

$$\|\pi_t - \pi\|_{\mathrm{TV}} \le \Pr(A^c \cup \{\tau > t\}).$$

Finally,

$$\Pr(A^c \cup \{\tau > t\}) \le \Pr(A^c) + \Pr(\tau > t) = 1 - \alpha_C + \Pr(\tau > t).$$

□

J.1.2 Weight-Harmonization

We show that coupling multiple chains can potentially be used to extend and improve the weight-harmonization framework for convergence diagnostics [4]. Weight harmonization starts from N parallel MCMC chains, each associated with an importance weight $W_0^{(i)}$ that reflects the discrepancy between the initial distribution $\hat{\pi}_0$ and the target distribution π. The basic intuition is that, as the empirical distribution $\hat{\pi}_t$ approaches π, the empirical distribution of the weights $W_t^{(i)}$ should become closer to uniform. Since $\hat{\pi}_t$ is itself unknown, the main challenge is to design an interactive update rule that propagates these weights consistently over time.

In [4], this is achieved through pairwise couplings together with partner exchanges across chains. Unlike grand coupling, where the goal is for all chains to coalesce into a single state, weight harmonization uses meetings between chains only as a mechanism for updating and propagating weights, after which the chains are randomly reassigned to new partners. Here, we show that, instead of coupling chains only in pairs, one can couple multiple chains jointly to improve diagnostic performance. The motivation is that, under pairwise couplings, two selected chains may be far apart, making coalescence unlikely and thus slowing down the weight updates. By allowing several chains to interact simultaneously, the joint coupling construction creates more opportunities for partial coalescence among nearby chains. This can lead to faster weight propagation and a more effective convergence diagnostic. We show the procedure in Algorithm 8 where replacing group size $m = 2$ recovers the proposed algorithm by Corenflos and Dau [4]. We note that the proposed procedure naturally satisfies the criteria for convergence diagnostics, i.e. Proposition 2 and Theorem 2 in [4].

Experiment. We follow the setup in [4], where the target distribution is $\pi = \mathcal{N}(0, \mathrm{I}_d)$. We set $d = 20$ and use $C = 10000$ chains for convergence diagnostics. The initial distribution is $\pi_0 = \mathcal{N}(10\,\mathbf{1}_d, 5\mathrm{I}_d)$, and we consider the autoregressive kernel

$$\kappa(\mathrm{d}y \mid x) = \mathcal{N}(y; \rho x, (1-\rho^2)\mathrm{I}_d)\,\mathrm{d}y,$$

since in this case the marginal law at iteration t admits a closed-form expression,

$$\hat{\pi}_t = \mathcal{N}\big(\rho^t 10\,\mathbf{1}_d,\ \big(1 + 4\rho^{2t}\big)\mathrm{I}_d\big)\,.$$

This allows us to evaluate the coupling schemes in a setting where the transient distribution is available exactly, and hence to directly assess the tightness of the bound. As the discrepancy measure, we use the squared Hellinger distance.

The performance of the different estimation schemes is shown in Figure 7. We observe that our approach consistently outperforms the baseline. We attribute this improvement to the order-invariant coupling property of the Poisson Matching framework, which avoids the asymmetry of reference-based constructions and enables more effective matching across chains.

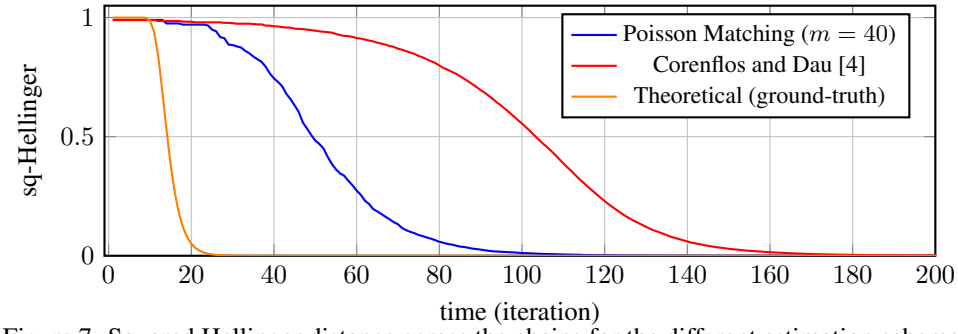


Figure 7: Squared Hellinger distance across the chains for the different estimation schemes.

Algorithm 8: Weight harmonization with groupwise grand coupling

Input : Chain states $\{X_t^{(i)}\}_{i=1}^N$, weights $\{W_t^{(i)}\}_{i=1}^N$, group size m, and a grand-coupling transition rule

Output : Updated chain states $\{X_{t+1}^{(i)}\}_{i=1}^N$, updated weights $\{W_{t+1}^{(i)}\}_{i=1}^N$

Partition the chains into groups

$$\mathcal{G}_t = \{G_t^{(1)}, \dots, G_t^{(L)}\}, \qquad |G_t^{(\ell)}| = m \text{ for all } \ell \text{ where } L = N/m$$

$\mathcal{A}_t \leftarrow \emptyset$ `// indices of groups with at least one coalescence`

for $\ell = 1, \dots, L$ **do**

  Let $G_t^{(\ell)} = \{i_1, \dots, i_m\}$

  Sample the joint transition: $(X_{t+1}^{(i_1)}, \dots, X_{t+1}^{(i_m)}) \leftarrow \texttt{GrandCouple}(X_t^{(i_1)}, \dots, X_t^{(i_m)})$

  Find all coalesced clusters: $\mathcal{C}_\ell \leftarrow \texttt{FindClusters}(X_{t+1}^{(i_1)}, \dots, X_{t+1}^{(i_m)})$

  foreach *cluster* $C \in \mathcal{C}_\ell$ *such that* $|C| \geq 2$ **do**

    $\bar{W} \leftarrow \texttt{Average}\big(\{W_t^{(j)} : j \in C\}\big)$

    foreach $j \in C$ **do**

      $W_{t+1}^{(j)} \leftarrow \bar{W}$

    end

  end

  foreach $j \in G_t^{(\ell)}$ *not belonging to any cluster* $C \in \mathcal{C}_\ell$ *with* $|C| \geq 2$ **do**

    $W_{t+1}^{(j)} \leftarrow W_t^{(j)}$

  end

  if *there exists a cluster* $C \in \mathcal{C}_\ell$ *with* $|C| \geq 2$ **then**

    $\mathcal{A}_t \leftarrow \mathcal{A}_t \cup \{\ell\}$

  end

end

if $|\mathcal{A}_t| \geq 2$ **then**

  Reshuffle the groups indexed by $\mathcal{A}_t$ to form the next grouping $\mathcal{G}_{t+1}$

end

else

  $\mathcal{G}_{t+1} \leftarrow \mathcal{G}_t$

end

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We state the contributions in the introduction and provide details proof in the Appendix

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Conlcusion Section

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All proofs are included and verified through experiments.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer:[Yes]

   Justification: Details are presented in the Experiment Section + Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code will be released after paper results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Parameters are described in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Every plot is averaged over 10000 runs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experiments are run on a A4500 GPU.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper is about coupling and Markov chains, we do not see any potential ethics problem.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper is about coupling and Markov chains, we do not see any negative associated societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer:[NA]

    Justification: This work studies MCMC and is not not related to the mentioned problem.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: No assets involved.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing experiments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No such research is done in this paper.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: No LLM usage was involved.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.