ONE-STEP SAMPLING VIA MCMC DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Markov chain Monte Carlo (MCMC) enables sampling from unnormalized distributions, but obtaining well-mixed samples often requires many sequential transitions. We study MCMC distillation, which amortizes these iterative refinements into an implicit neural sampler that generates samples in a single forward pass. Starting from the current generator, we apply short MCMC refinements and train the generator to match the resulting distribution using sample-based discrepancies, without requiring target samples or evaluating the generator density. We establish a unified weak convergence framework based on a finite-window error decomposition that separates reference mixing, accumulated distillation error, and transition bias. We investigate sliced Wasserstein distance, MMD with Riesz and Matérn kernel, and Sinkhorn divergence as potential sample-based divergences in this theoretical framework. For unadjusted Langevin algorithm (ULA) teachers, we derive explicit sufficient conditions for weak convergence to the target and conditional rates for expectation errors. Experiments on synthetic and high-dimensional Bayesian inference tasks achieves comparable results to ULA teachers, demonstrating the effectiveness of repeated distillation across different sample-based discrepancies.

1 INTRODUCTION

Markov chain Monte Carlo (MCMC) provides a general approach to sampling from distributions known only up to a normalizing constant (Metropolis et al., 1953; Duane et al., 1987; Neal, 2011), but obtaining well-mixed samples can require many sequential transitions. This cost motivates learning a reusable sampler that generates new samples directly. Variational inference pursues this goal by optimizing a parameterized distribution, with its expressiveness and training objective constrained by the available density evaluations or bounds (Jordan et al., 1999; Wainwright & Jordan, 2008; Blei et al., 2017; Yin & Zhou, 2018; Titsias & Ruiz, 2019; Yu & Zhang, 2023). An implicit neural generator offers a flexible alternative: it maps a simple random input Z to $G_\theta(Z)$ in one forward pass, defining a distribution μ_θ whose density need not be tractable. The challenge is to train this generator when the target density is unnormalized and samples from the target are unavailable.

Building on amortized MCMC (Li et al., 2017), we study *MCMC distillation*: initialize a short teacher chain from the current generator, then fit the generator to the chain's output distribution. Writing T_n for the teacher transition and D for a fitting discrepancy, the population update takes the form

$$\theta_{n+1} \approx \arg\min_{\theta\in\Theta} D(\mu_\theta, (T_n)_{\#}\mu_{\theta_n}), \tag{1}$$

where the teacher distribution is held fixed during optimization. Sample-based discrepancies make this update accessible without evaluating the generator density or drawing target samples; access to the target is provided through the teacher transition, for example through its score in a Langevin update. After training, sampling requires only a generator forward pass.

The amortization principle connects our approach to several related methods. Li et al. (2017) formulate the update as a projected fixed-point iteration and discuss KL, adversarial, and energy-matching objectives. MCMC teaching also appears in cooperative learning (Xie et al., 2018), where a generator learns from short MCMC refinements while an energy model is learned jointly. Our setting instead treats the target as fixed. Concurrent work by Pauline et al. (2026) also investigates one-step implicit sampling through short-chain MCMC refinement. These connections motivate our focus on the accuracy required of successive fits and the propagation of their errors.

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

The central question is whether repeated approximate distillation preserves enough of the teacher's progress to approach the target. Teacher mixing alone does not answer this question. In general, the transition may not contract the discrepancy used for fitting, and restricting each update to a neural family introduces an additional error. With $\mu_n = \mu_{\theta_n}$, this error is measured by the residual $\varepsilon_n = D(\mu_{n+1}, (T_n)_{\#}\mu_n)$. Numerical teacher transitions, such as the unadjusted Langevin algorithm (ULA), can introduce further bias because they need not preserve the target exactly. We address these effects through expectation errors of test functions. A finite-window decomposition separates reference mixing, accumulated distillation error, and transition bias.

Recent one-step generative models provide a complementary perspective. Drifting Models (Deng et al., 2026) construct training targets from interactions between generated and target samples. W-Flow (Han et al., 2026) uses Wasserstein gradient flows and analyzes the finite-time approximation of their particle dynamics. MMD gradient flows similarly use a discrepancy to define distributional evolution (Arbel et al., 2019). In our analysis, prescribed Markov transitions determine the teacher evolution, while a sample-based discrepancy controls how accurately the generator fits it. Our results concern the long-time behavior of these repeatedly fitted laws, conditional on their population approximation errors.

We summarize our contributions as follows.

- **A convergence framework for approximate distillation.** We combine the finite-window decomposition with discrepancy-dependent test-function bounds to give sufficient conditions for weak convergence. The framework separates fitting accuracy from reference mixing and transition consistency, and accommodates both target-invariant and biased teacher transitions under the corresponding assumptions.
- **Concrete discrepancy and Langevin guarantees.** We verify the required test-function bounds for Wasserstein and sliced Wasserstein distances, and for Riesz and Matérn MMDs. For ULA teachers, we establish weak convergence under explicit regularity, mixing, and student-stability conditions when step sizes h_n vanish with $\sum_n h_n = \infty$ and $\varepsilon_n = o(h_n)$. We also derive conditional rates for expectation errors and extend the analysis to Sinkhorn fitting through a Wasserstein comparison with vanishing entropic regularization.
- **Sample-based implementations and empirical evaluation.** We implement distillation with kernel and sliced-Wasserstein matching and examine how the fitting criterion affects learned distributions on two-dimensional toy targets, a logistic regression posterior and a high dimensional conditioned-diffusion posterior with efficient sampling and high accuracy.

2 BACKGROUND

Bayesian inference. Given observations y and an unknown parameter ϕ, Bayesian inference targets the posterior distribution

$$\pi(\phi \mid y) = \frac{p(y \mid \phi)p(\phi)}{p(y)}, \qquad p(y) = \int p(y \mid \phi)p(\phi)\, d\phi. \tag{2}$$

The evidence $p(y)$, also called the marginal likelihood, is the normalizing constant and is often intractable, even when the unnormalized posterior $p(y \mid \phi)p(\phi)$ can be evaluated. The central task in Bayesian computation is posterior approximation which supports parameter estimation, prediction, and uncertainty quantification.

MCMC. A common approach for Bayesian inference is Markov chain Monte Carlo (MCMC). Let π be a target distribution on $\mathbb{R}^d$ with density $\pi(x) = Z_\pi^{-1} \exp\{-U(x)\}$, where U is the energy and Z_π is the generally unknown normalizing constant. For example, π may be the posterior distribution in equation 2. A Markov transition kernel K is invariant with respect to π if

$$\pi(dx') = \int_{\mathbb{R}^d} K(x, dx')\, \pi(dx). \tag{3}$$

A sufficient condition for invariance is detailed balance,

$$\pi(dx)\, K(x, dx') = \pi(dx')\, K(x', dx). \tag{4}$$

When K admits a transition density k, this condition becomes $\pi(x)k(x,x') = \pi(x')k(x',x)$. MCMC constructs a π-invariant kernel and generates a chain $X_{n+1} \mid X_n \sim K(X_n, \cdot)$. Under suitable ergodicity conditions, the law of X_n converges to π as $n \to \infty$. Although multiple chains can run in parallel, transitions within each chain are sequential, and slowly mixing chains may require many transitions to approximate the target.

3 METHOD

To amortize the sampling cost, we consider a neural sampler G_θ with output distribution $\mu_\theta = (G_\theta)_\# \lambda$, where λ is a fixed input-noise distribution. Our goal is to learn θ so that μ_θ approximates π. Once trained, the sampler generates independent samples $G_\theta(\epsilon)$ using independent inputs $\epsilon \sim \lambda$, with a single forward pass per sample. More specifically, we alternate between evolving the current sampler distribution through a Markov transition and fitting the resulting distribution with the generator. At round n, let $\mu_n = \mu_{\theta_n}$ and let T_n be a Markov kernel representing one or several sampling steps for π. Write $\nu_n = (T_n)_\# \mu_n$ for the resulting teacher distribution, defined by $\nu_n(A) = \int T_n(x, A)\, \mu_n(dx)$. The generator is updated by approximately solving

$$\theta_{n+1} \approx \arg\min_\theta D(\mu_\theta, \mathrm{sg}[\nu_n])\,, \tag{5}$$

where D is a discrepancy between distributions and sg denotes stop-gradient: the teacher distribution is held fixed while the student is optimized. The updated student then initializes the teacher at the next round.

For example, a teacher based on the unadjusted Langevin algorithm (ULA) applies the transition

$$Y = X - h_n \nabla U(X) + \sqrt{2h_n}\,\xi, \qquad X \sim \mu_n, \quad \xi \sim \mathcal{N}(0, I_d), \quad \xi \perp X, \tag{6}$$

where $h_n > 0$ is the step size. This produces teacher samples directly from the current generator, without running a chain to equilibrium. Unlike an exactly π-invariant transition, ULA generally introduces discretization bias.

In practice, we draw $x_1, \ldots, x_B \overset{\text{i.i.d.}}{\sim} \mu_n$ and independently simulate $y_i \mid x_i \sim T_n(x_i, \cdot)$ for $i = 1, \ldots, B$. The resulting teacher samples approximate ν_n and are held fixed during the student update. We optimize an empirical version of equation 5 using samples from μ_θ and these teacher samples, with gradients propagated only through the student samples. Algorithm 1 summarizes the training procedure.

The choice of D determines how the student is fitted to the teacher. Li et al. (2017) discuss KL minimization, adversarial divergence estimation, and energy matching. Their direct KL-based update requires evaluating the student density, whereas adversarial estimation requires training an auxiliary discriminator. Here, we use discrepancies computable from empirical samples without either requirement, including maximum mean discrepancy (MMD), sliced Wasserstein distance, and Sinkhorn divergence.

Maximum Mean Discrepancy For a positive definite kernel k, we minimize the empirical squared maximum mean discrepancy (MMD) (Gretton et al., 2012):

$$\widehat{D}^2_{n,k}(\theta) = \frac{1}{B^2} \sum_{i,j=1}^{B} \left[k(x_i, x_j) + k(y_i, y_j) - 2k(x_i, \bar{y}_j)\right]. \tag{7}$$

We consider Matérn kernels and the anchored Riesz kernel $k_\alpha(x,y) = (|x|^\alpha + |y|^\alpha - |x-y|^\alpha)/2$, $0 < \alpha < 2$, whose MMD is an energy discrepancy (Sejdinovic et al., 2013). The loss uses pairwise kernel evaluations, with fixed kernel parameters during each update. Its teacher–teacher term is constant in θ. Equation equation 7 is a V-statistic, and the two batches share their initial samples.

Sliced Wasserstein distance. Let $u_1, \ldots, u_J$ be independent uniform directions on $\mathbb{S}^{d-1}$, and let $x_{\ell,(i)}(\theta)$ and $\bar{y}_{\ell,(i)}$ denote the sorted projections of the two batches onto u_ℓ. For $p \geq 1$, the empirical sliced transport objective is

$$\widehat{D}^p_{n,\mathrm{SW}_p}(\theta) = \frac{1}{JB} \sum_{\ell=1}^{J} \sum_{i=1}^{B} |x_{\ell,(i)}(\theta) - \bar{y}_{\ell,(i)}|^p. \tag{8}$$

Algorithm 1 Incremental MCMC distillation for one-step neural sampling

Require: Generator G_θ, initial parameters θ_0, base law λ;
teacher kernels $\{T_n\}_{n=0}^{N-1}$ targeting $\pi \propto e^{-U}$;
batch size B, number of rounds N, sample-based loss $\mathcal{L}$.

for $n = 0, \dots, N-1$ **do**
  Draw $z_1, \dots, z_B \overset{\text{iid}}{\sim} \lambda$.
  for $i = 1, \dots, B$ **do**
    $x_i^{\text{teach}} \leftarrow \text{stopgrad}[G_{\theta_n}(z_i)]$.
    Draw $y_i \sim T_n(x_i^{\text{teach}}, \cdot)$ independently
    $\bar{y}_i \leftarrow \text{sg}[y_i]$.
  end for
  $\widehat{\nu}_n \leftarrow B^{-1} \sum_{i=1}^B \delta_{\bar{y}_i}$, $\widehat{\mu}_\theta \leftarrow B^{-1} \sum_{i=1}^B \delta_{G_\theta(z_i)}$.
  $\widehat{D}_n(\theta) \leftarrow \widehat{D}(\widehat{\mu}_\theta, \widehat{\nu}_n)$ ▷ D is a sample-based metric, with estimator $\widehat{D}$
  $\theta_{n+1} \leftarrow \text{OptimizerStep}(\theta_n, \nabla_\theta \widehat{\mathcal{L}}_n(\theta_n))$.
end for
return G_{θ_N}; generate $X = G_{\theta_N}(Z)$ with fresh $Z \sim \lambda$.

This is a Monte Carlo approximation to the angular average of empirical one-dimensional transport costs (Nadjahi et al., 2019). The population distance satisfies $\text{SW}_1 \le \text{SW}_p$.

Sinkhorn divergence. We also consider $S_{\tau,p}(\widehat{\mu}_\theta, \widehat{\nu}_n)$ with ground cost $|x-y|^p$ and entropy coefficient $\tau > 0$ (Feydy et al., 2019). The debiased objective is $S_{\tau,p}(\mu,\nu) = \text{OT}_{\tau,p}(\mu,\nu) - \frac{1}{2}\text{OT}_{\tau,p}(\mu,\mu) - \frac{1}{2}\text{OT}_{\tau,p}(\nu,\nu)$, where $\text{OT}_{\tau,p}$ minimizes transport cost plus $\tau\text{KL}(\zeta \| \mu \otimes \nu)$ over couplings ζ. The student self-cost is included in the gradient. Appendix A derives a Wasserstein bound that accounts for the regularization error.

4 Weak Convergence of MCMC distillation

We now study when MCMC distillation converges to the target distribution despite errors in fitting the teacher at each round. Write $\varepsilon_n = D(\mu_{n+1}, (T_n)_\# \mu_n)$ for the round-n fitting error. As noted by Li et al. (2017), amortized MCMC shares a projected fixed-point structure with reinforcement-learning methods such as Q-learning and temporal-difference learning (Watkins & Dayan, 1992; Tsitsiklis & Van Roy, 1997). Motivated by this connection, we develop a unified analysis based on test functions. The framework covers W_p and SW_p for $1 \le p < \infty$, as well as Riesz MMD and Matérn MMD, under the corresponding assumptions stated below.

Our convergence criterion is *weak convergence*: $\mu_n \xrightarrow{w} \pi$ means that $\mathbb{E}_{X\sim\mu_n} f(X) \to \mathbb{E}_{X\sim\pi} f(X)$ for every bounded continuous function $f \in C_b(\mathbb{R}^d)$. Thus, expectations under the learned distributions converge to their target values for every bounded continuous observable. Since both μ_n and π are probability measures, it suffices to establish this convergence for $f \in C_c^\infty(\mathbb{R}^d)$, the smooth compactly supported functions (Billingsley, 1999; Dudley, 2002). This reduction provides the basis for our test-function analysis. Weak convergence also admits a metric characterization: for any fixed Matérn kernel with positive smoothness and length-scale parameters,

$$\mu_n \xrightarrow{w} \pi \quad \Longleftrightarrow \quad \text{MMD}_{\text{Mat}}(\mu_n, \pi) \longrightarrow 0$$

(Sriperumbudur et al., 2011; Simon-Gabriel et al., 2023). For a ULA teacher, we further derive convergence rates for expectation errors under the additional assumptions stated below.

4.1 Decomposition of expected error on test functions

For a finite signed measure μ and a bounded measurable function f, write $\mu(f) := \int f \, d\mu$. For a Markov kernel T, define $Tf(x) := \int f(y)\, T(x, dy)$, $(T_\# \mu)(f) = \mu(Tf)$. Let $\mathcal{M}_n = \{(G_\theta)_\# \lambda : \theta \in \Theta_n\}$ denote the generator family used at round n, where λ is the fixed input-noise distribution. Fitting the teacher law $(T_n)_\# \mu_n$ returns $\mu_{n+1} \in \mathcal{M}_n$, with T_n denoting the implemented Markov transition.

To separate fitting error from bias in the teacher transition, introduce reference kernels $\widetilde{T}_n$ satisfying $(\widetilde{T}_n)_{\#}\pi = \pi$. For example, when T_n is a ULA step of size h_n, $\widetilde{T}_n$ can be the exact Langevin transition over time h_n. The implemented kernel T_n need not preserve π; when it does, we take $\widetilde{T}_n = T_n$. We index rounds and step sizes from zero and define the accumulated time by $s_0 = 0,\ s_n := \sum_{j=0}^{n-1} h_j,\ n \geq 1$. For the convergence results, we assume $0 < h_n \leq \bar{h} < \infty$ and $s_n \to \infty$. We impose the following dual bound on the training discrepancy D.

Assumption 4.1. There exists a seminorm $\mathcal{N}_D : C_b(\mathbb{R}^d) \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$ such that

$$|\mu(f) - \nu(f)| \leq D(\mu, \nu)\mathcal{N}_D(f), \quad f \in C_b(\mathbb{R}^d), \tag{9}$$

for all probability measures in the domain of D. Seminorm also requires $\mathcal{N}_D(f+g) \leq \mathcal{N}_D(f) + \mathcal{N}_D(g), \mathcal{N}(cf) = |c|\mathcal{N}_D(f)$.

Remark This bound is the standard dual control underlying integral probability metrics (IPMs) (Müller, 1997). Indeed, for $D(\mu,\nu) = \sup_{\mathcal{N}_D(g)\leq 1} |(\mu - \nu)(g)|$, equation 9 follows by rescaling the test function. It separates the discrepancy between measures from the regularity of the test. Constants have zero seminorm because their expectations agree under any two probability measures. Kantorovich–Rubinstein duality gives $\mathcal{N}_D = \mathrm{Lip}$ for W_1 (Villani, 2009); for MMD, the bound follows from Cauchy–Schwarz in its RKHS, with $\mathcal{N}_D(f) = \inf_{c: f-c \in \mathcal{H}_k} \|f - c\|_{\mathcal{H}_k}$ (Gretton et al., 2012). We only require an upper bound of this form, not that every training distance be represented exactly as an IPM. Appendix A.1 gives the detailed bounds. For the Langevin verification, Appendix A.4 uses higher-order Sobolev norms as auxiliary majorants to establish propagation of the controlling seminorms. In particular, the distillation residual satisfies $|\eta_n(g)| \leq \varepsilon_n \mathcal{N}_D(g)$. Sinkhorn fitting (Feydy et al., 2019) can also be included through a comparison with Wasserstein distance as its entropy regularisation vanishes; see Appendix A.7.

4.2 Finite-window decomposition and convergence

The analysis rests on an exact decomposition of the expectation error. Set $K_{k,n} = \widetilde{T}_k \circ \widetilde{T}_{k+1} \circ \cdots \widetilde{T}_{n-1}$ for $k < n$ and $K_{n,n} = I$. Define the distillation residual $\eta_n = \mu_{n+1} - T_{n\#}\mu_n$. One step update gives $(\mu_{n+1} - \pi)f = [\widetilde{T_{n\#}}(\mu_n - \pi)]f + \eta_n(f) + \big(T_{n\#}\mu_n - \widetilde{T_{n\#}}\mu_n\big)(f)$. For $0 \leq k < n$ and $f \in C_c^\infty(\mathbb{R}^d)$, we can decompose the error into three parts:

$$(\mu_n - \pi)f = \underbrace{[(K_{k,n})_{\#}\mu_k - \pi]f}_{\text{Term 1}} + \underbrace{\sum_{j=k}^{n-1} \eta_j(K_{j+1,n}f)}_{\text{Term 2}} + \underbrace{\sum_{j=k}^{n-1} \mu_j((T_j - \widetilde{T}_j)K_{j+1,n}f)}_{\text{Term 3}}. \tag{10}$$

The three terms represent reference mixing, accumulated fitting error, and transition bias, respectively. The last term vanishes for an exact teacher. This identity allows the choice of training discrepancy to be separated from the mixing and discretization properties of the teacher. Its proof is given in Appendix A.2.

For $A > 0$ and $s_n \geq A$, define $k = k(n, A) := \max\{0 \leq j \leq n : s_j \leq s_n - A\}$. Then $A \leq s_n - s_k < A + h_k \leq A + \bar{h}$. The following assumption provide a control for reference mixing, which is also proposed in former MCMC theoretical analysis literature, like (Benaïm et al., 2017).

Assumption 4.2 (Uniform Ergodicity)**.** For every $f \in C_c^{+\infty}(\mathbb{R}^d)$, there exists constants $C_f, \lambda_f > 0$ only depending on f, such that

$$\big|[(K_{j,n})_{\#}\mu_j - \pi]f\big| \leq C_f e^{-\lambda_f(s_n - s_j)}, \quad 0 \leq j \leq n. \tag{11}$$

We assume an exponential growth bound for the propagated seminorm, in line with standard semigroup and evolution-family estimates (Engel & Nagel, 2000; Buşe et al., 2017). For $D = W_1$, this reduces to a Lipschitz-seminorm bound for the Langevin semigroup (Bakry et al., 2014).

Assumption 4.3. There exist constants $M_D > 0$ and $\omega_D \geq 0$ such that for every $f \in C_c^\infty(\mathbb{R}^d)$, $\mathcal{N}_D(f) < \infty$, the propagated test $K_{j,n}f$ belongs to $C_b(\mathbb{R}^d)$ and

$$\mathcal{N}_D(K_{j,n}f) \leq M_D e^{\omega_D(s_n - s_j)}\mathcal{N}_D(f), \quad 0 \leq j \leq n \tag{12}$$

We also assume the implemented transition kernel T_j to be locally consistent with the reference transition $\widetilde{T}_j$ on the physical time scale. More precisely, its one-step weak error should be $o(h_j)$. Conditions of this type are standard in the numerical analysis of SDEs and in perturbation analyses of numerical SDE literature (Talay & Tubaro, 1990; Kloeden & Platen, 1992; Milstein & Tretyakov, 2004; Shardlow & Stuart, 2000).

Assumption 4.4 (Local consistency)**.** For every $f \in C_c^\infty(\mathbb{R}^d)$, every finite $A > 0$, and every sequence $(n_j)_{j\geq 0}$ satisfying $n_j > j$, $s_{n_j} - s_{j+1} \leq A$, we have

$$\lim_{j\to+\infty} \frac{1}{h_j} \left| \mathbb{E}_{\mu_j} (T_j - \widetilde{T}_j) K_{j+1,n_j} f \right| \longrightarrow 0. \tag{13}$$

Based on the settings above, we obtain the general convergence result for MCMC distillation, with a detailed proof in *Proof: Appendix A.2.*

Theorem 4.5 (Weak convergence of MCMC distillation)**.** *Suppose Assumptions 4.1–4.4 hold. If $\varepsilon_n = o(h_n)$, then $\mu_n \xrightarrow{w} \pi$.*

4.3 Verification for Langevin kernels

We specialize the preceding analysis to a ULA teacher for the target $\pi(dx) = Z_U^{-1} e^{-U(x)}\, dx$, where $0 < Z_U < \infty$. Let $(P_t)_{t\geq 0}$ denote the transition semigroup of the overdamped Langevin diffusion $dX_t = -\nabla U(X_t)\, dt + \sqrt{2}\, dB_t$. The implemented ULA kernel and its target-invariant reference are

$$T_n g(x) = \mathbb{E}\, g\Big(x - h_n \nabla U(x) + \sqrt{2h_n} Z\Big), \quad Z \sim \mathcal{N}(0, I_d), \quad \widetilde{T}_n = P_{h_n}, \tag{14}$$

By the semigroup property, the reference propagator reduces to $K_{j,n} = P_{s_n - s_j}$. We use the log-Sobolev convention $\mathrm{Ent}_\pi(v^2) \leq (2/\rho)\pi(|\nabla v|^2)$, with $\rho > 0$. The next theorem verifies the three assumptions and gives an explicit bound on the resulting expectation error.

Theorem 4.6 (Langevin distillation)**.** *Suppose $U \in C^2(\mathbb{R}^d)$ and ∇U is globally L-Lipschitz, and let $h_n \to 0$. Let D be W_p, SW_p ($1 \leq p < \infty$), Riesz MMD with $0 < \alpha < 2$, or Matérn MMD with smoothness $\nu > 0$, with fixed parameters and well-defined student–teacher discrepancies. For every $f \in C_c^\infty(\mathbb{R}^d)$ and $t \geq 0$:*

1. *If $H_{t_0} := \sup_j \mathrm{KL}((P_{t_0})_{\#}\mu_j \| \pi) < \infty$ for some $t_0 \geq 0$ and π satisfies LSI with constant ρ, then Assumption 4.2 holds with rate $\lambda_D = \rho$ and $C_f = e^{\rho t_0} \|f\|_\infty \max\{2, \sqrt{2H_{t_0}}\}$.*
2. *If U is sufficiently smooth (See Appendix A.4), the controlling seminorms D satisfy Assumption 4.3 with $M_D > 0$ and $\omega_D \geq 0$.*
3. *If $\sup_{j\geq 0} \mathbb{E}_{\mu_j} \|\nabla U\| = G < \infty$, then Assumption 4.4 holds with*

$$\mathbb{E}_{\mu_j} \left| (T_j - P_{h_j}) P_t f \right| \leq L e^{L(t+h_j)} \mathrm{Lip}(f) \left[\frac{G}{2} h_j^2 + \frac{2\sqrt{2d}}{3} h_j^{3/2} \right]. \tag{15}$$

Under these conditions, for $0 < A \leq s_n$ and $k = k(n, A)$,

$$|(\mu_n - \pi) f| = O_f \left(e^{-\rho A} + e^{\omega_D A} \sum_{j=k}^{n-1} \varepsilon_j + e^{LA} \sum_{j=k}^{n-1} h_j^{3/2} \right), \tag{16}$$

with constants independent of n and A. Thus $\varepsilon_n = o(h_n)$ imply $\mu_n \xrightarrow{w} \pi$.

The three bounds are proved in Appendices A.3–A.5. The decomposition bound also admits discussions on a polynomial convergence rate, with additional assumptions on stepsize schedule and residual control. We leave it to A.6.

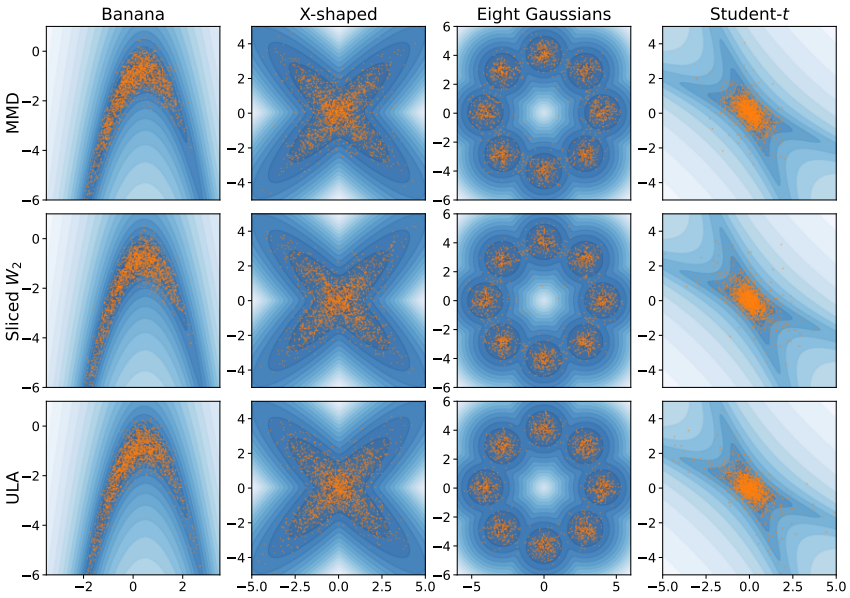

Figure 1: Samples from the three methods on the two-dimensional targets.

5 EXPERIMENTS

We study four two-dimensional targets, a 22-dimensional logistic-regression posterior, and a 100-dimensional conditioned-diffusion posterior to demonstrate our proposed methods empirically (Cheng et al., 2024; Pielok et al., 2025). The distilled sampler is fitted using either Matérn MMD with a Laplacian kernel or sliced Wasserstein-2. The unadjusted Langevin algorithm (ULA) provides a direct-sampling baseline, as well as serving as the MCMC teacher. The two-dimensional runs use 10,000 outer distillation iterations and the other runs use 100,000 distillation iterations. Each student iteration uses five ULA steps as teaacher; step sizes depend on the target (Table 4). The ULA baseline uses 50,000 transitions on two-dimensional targets and 500,000 otherwise, matching the transition count of the students. Results are means $\pm$ sample standard deviations across five runs. Target definitions, training configurations and evaluation protocols are given in Appendix B.

5.1 TWO-DIMENSIONAL TARGETS

We evaluate on targets including banana, X-shaped and eight-Gaussian mixtures, and transformed Student-t, using estimations of KL and W2 distance as metrics. For Student-t we use a sliced-W_2 diagnostic truncated to coordinate range $|x_i| < 6$ as global W2 distance is not well defined.

We present the visualized results per method in Figure 1, plotted against contours with 2000 samples each. We can observe from the visualizations that, all methods are able to capture the distribution successfully. The quantitative metrics are present in Table 1. Sliced W_2 gives the lowest reported W_2 on banana, eight Gaussians, and Student-t, while MMD gives the lowest on X-shaped. The lowest KL estimates are split between sliced W_2 and MMD.

Table 1: Final metrics for the two-dimensional targets.

	KL		W_2	
Target	MMD	Sliced W_2	MMD	Sliced W_2
Banana	0.035 ± 0.008	$\mathbf{0.022} \pm 0.007$	0.463 ± 0.071	$\mathbf{0.211} \pm 0.020$
X-shaped	$\mathbf{0.053} \pm 0.073$	0.216 ± 0.183	$\mathbf{0.108} \pm 0.058$	0.248 ± 0.177
Eight Gaussians	0.074 ± 0.017	$\mathbf{0.067} \pm 0.006$	0.295 ± 0.182	$\mathbf{0.233} \pm 0.106$
Student-t	$\mathbf{0.211} \pm 0.008$	0.222 ± 0.016	0.156 ± 0.002	$\mathbf{0.132} \pm 0.008$

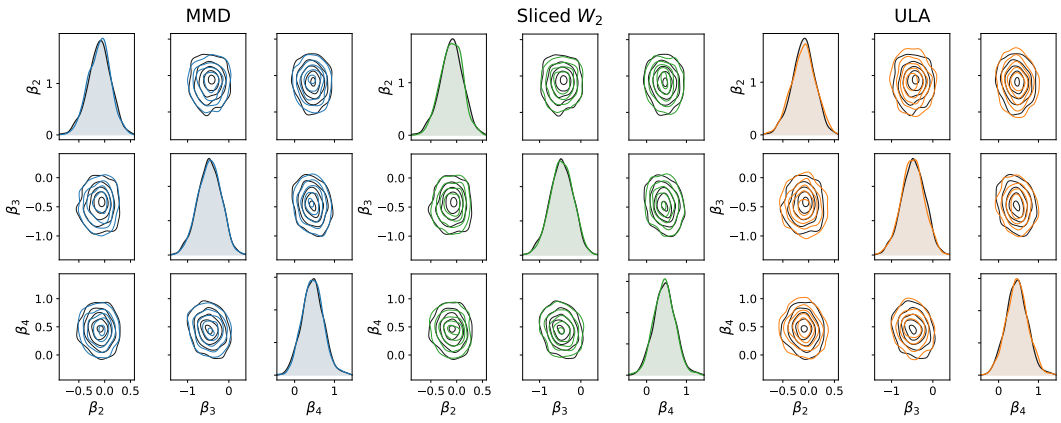

Figure 2: Waveform posterior marginals (diagonal) and pairwise density contours (off diagonal) for coefficients β_2, β_3, and β_4. Black denotes the independent reference ULA trace; colored curves denote each compared method.

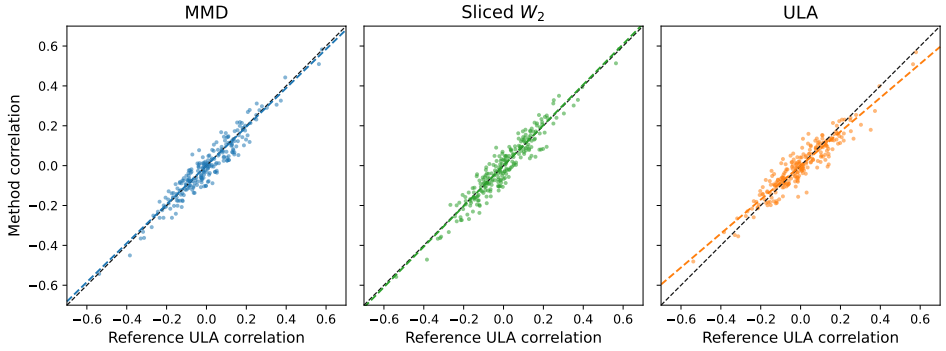

Figure 3: Method versus reference correlations for all 231 distinct pairs of coefficients. The black dashed line is equality; colored dashed lines are least-squares fits constrained through the origin.

5.2 Logistic-regression posterior

We next evaluate on the Bayesian logistic regression setting which emits a 22-dimensional posterior distribution (Cheng et al., 2024; Pielok et al., 2025). We use 400 labeled observations from the WAVEFORM dataset[1], each with 21 explanatory variables, to infer the coefficients of a binary classifier. With an intercept-augmented covariate $\bar{x}_i = (1, x_i^\top)^\top$, the model assigns probability $[1 + \exp(-\beta^\top \bar{x}_i)]^{-1}$ to label $y_i = 1$. The prior of β is defined to be the zero meaned gaussian $\mathcal{N}(0, \alpha^{-1} I)$ where $\alpha = 0.01$. We defer the additional details to Appendix B.4.

Figure 2 compares one-dimensional marginal densities and pairwise contours for three non-intercept coefficients. We compare these colored curves against the curves emitted by 1000 reference samples. Here we perform ULA with a smaller step size of 10^{-4} and 200,000 steps to obtain a set of 1000 reference samples. It can be seen that the students' distribution shape closely follow that of the reference distribution. We also study the pairwise correlation coefficients defined as $\rho = \left\{ \rho_{i,j} := \frac{\mathrm{cov}(\beta_i, \beta_j)}{\sqrt{\mathrm{cov}(\beta_i, \beta_i)\mathrm{cov}(\beta_j, \beta_j)}} \right\}_{1 \le i < j \le 22}$ and compare the efficient emitted by the learned posterior against those emitted by the reference samples. Figure 3 compares all 231 distinct pairwise coefficient correlations to the reference. It can be seen that the MMD and Sliced W2 students perform better than the large step size $5 * 10^{-3}$ teacher.

[1] https://archive.ics.uci.edu/ml/machine-learning-databases/waveform

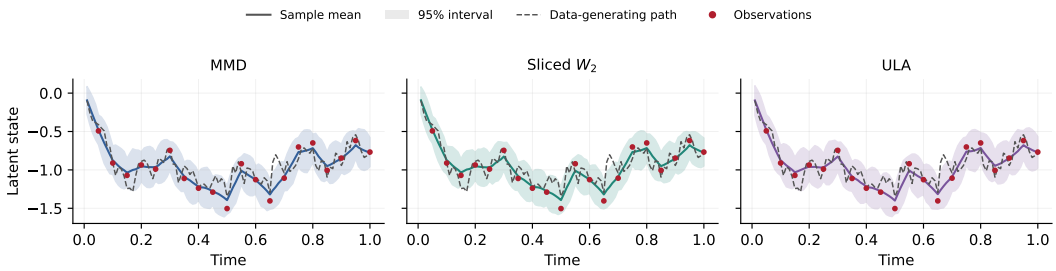

Figure 4: Sample paths and per-coordinate 95% sample intervals for the conditioned diffusion target.

5.3 Conditioned diffusion posterior

The conditioned-diffusion scenario concerns a 100-dimensional latent path observed at 20 time points in $[0, 1]$. The data-generating double-well SDE is

$$\mathrm{d}u_t = 10u_t(1 - u_t^2)\,\mathrm{d}t + \mathrm{d}W_t, \qquad u_0 = 0. \tag{17}$$

We defer the target definition details to Appendix B.5. Figure 4 shows the sample mean and central per-coordinate 95% sample interval from ULA and the two distilled samplers. Both distilled samplers have similar pointwise summaries to ULA in the displayed runs.

Table 2: Final per-coordinate ELM for the conditioned diffusion posterior.

Method	ELM
MMD	74.862 ± 0.042
Sliced W_2	74.973 ± 0.094
ULA	74.703 ± 0.004

We further evaluate coordinatewise expected log marginal likelihood (ELM) on 1,000 fixed reference paths against the 100,000 sample-induced posterior distribution, which is attained by performing per-coordinate Gaussian KDEs to those samples (Pielok et al., 2025). The reference paths were generated by ULA with step size 10^{-4} and 200,000 transitions. ULA candidates come from specialized large-batch 100,000-particle runs, whereas the student samples come from the checkpoint directly, a direct display of the scalability of our methods by distilling MCMC. It can be seen in Table 2 that the marginal likelihoods of distilled samplers slightly outperform ULA teacher, which we contribute to the fact that distillation is able to absorb and smooth the noise in the ULA teacher.

6 Conclusion

We studied MCMC distillation as a way to amortize short Markov-chain updates into an implicit neural sampler. The resulting procedure needs only generator samples and applications of the teacher transition, and can therefore be trained with sliced Wasserstein, kernel, or Sinkhorn matching objectives without evaluating the generator density. Theoretically, we establish a weak convergence analysis framework. For ULA teachers, we verify the required properties under explicit sufficient conditions, obtaining conditional expectation-error bounds for various discrepancies. Empirically, the learned samplers capture the target geometry across two-dimensional toy examples, logistic regression posterior and high dimensional conditioned diffusion posterior, demonstrating that the distillation is effective to transfer the information of MCMC into the neural sampler. A limitation of our current analysis is that the step-size schedules required for theoretical convergence differ from those used in practice. Bridging this gap and establishing convergence guarantees under more practical step-size schedules are left for future work.

AI use statement

We used generative AI tools to assist in drafting and editing the manuscript; sourcing relevant literature; and exploring portions of the proofs. All literature references were checked against the cited

sources.We do not use generative AI for the distillation idea and the proof framework. The authors reviewed all AI-assisted material and take responsibility for the final content of this work.

REFERENCES

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1906.04370.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2014. doi: 10.1007/978-3-319-00227-9. URL https://link.springer.com/book/10.1007/978-3-319-00227-9.

Michel Benaïm, Florian Bouguet, and Bertrand Cloez. Ergodicity of inhomogeneous Markov chains through asymptotic pseudotrajectories. *The Annals of Applied Probability*, 27(5):3004–3049, 2017. doi: 10.1214/17-AAP1275. URL https://doi.org/10.1214/17-AAP1275.

Patrick Billingsley. *Convergence of Probability Measures*. Wiley, 2nd edition, 1999. doi: 10.1002/9780470316962.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL https://www.cs.columbia.edu/~blei/papers/BleiKucukelbirMcAuliffe2017.pdf.

Nicolas Bonnotte. *Unidimensional and Evolution Methods for Optimal Transportation*. PhD thesis, Université Paris-Sud and Scuola Normale Superiore, 2013. URL https://www.normalesup.org/~bonnotte/thesis/thesis-pdf.pdf.

Constantin Buşe, Donal O'Regan, and Olivia Saierli. An inequality concerning the growth bound of an evolution family and the norm of a convolution operator. *Semigroup Forum*, 94(3):618–631, 2017. doi: 10.1007/s00233-016-9822-9.

Ziheng Cheng, Longlin Yu, Tianyu Xie, Shiyue Zhang, and Cheng Zhang. Kernel semi-implicit variational inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8248–8269, 2024. URL https://proceedings.mlr.press/v235/cheng24l.html.

Arnak Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 678–689, 2017. URL https://proceedings.mlr.press/v65/dalalyan17a.html.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting, 2026. URL https://arxiv.org/abs/2602.04770v2. Version 2.

Simon Duane, A. D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987. doi: 10.1016/0370-2693(87)91197-X.

Richard M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2nd edition, 2002. doi: 10.1017/CBO9780511755347.

Weinan E, Tiejun Li, and Eric Vanden-Eijnden. *Applied Stochastic Analysis*, volume 199 of *Graduate Studies in Mathematics*. American Mathematical Society, 2019. ISBN 978-1-4704-4933-9. doi: 10.1090/gsm/199.

Klaus-Jochen Engel and Rainer Nagel. *One-Parameter Semigroups for Linear Evolution Equations*, volume 194 of *Graduate Texts in Mathematics*. Springer, 2000. doi: 10.1007/b97696.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2681–2690, 2019. URL https://proceedings.mlr.press/v89/feydy19a.html.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

Jiaqi Han, Puheng Li, Qiushan Guo, Renyuan Xu, Stefano Ermon, and Emmanuel J. Candès. One-step generative modeling via wasserstein gradient flows, 2026. URL https://arxiv.org/abs/2605.11755v2. Version 2; convergence theorem in Appendix B, Theorem B.4.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999. doi: 10.1023/A:1007665907178.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences, 2018. URL https://arxiv.org/abs/1807.02582.

Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, 1992. doi: 10.1007/978-3-662-12616-5.

Hiroshi Kunita. *Stochastic Flows and Stochastic Differential Equations*, volume 24 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1990a. ISBN 9780521350501.

Hiroshi Kunita. *Stochastic Flows and Stochastic Differential Equations*, volume 24 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 1990b. URL https://assets.cambridge.org/97805213/50501/sample/9780521350501ws.pdf.

Xuechen Li, Denny Wu, Lester Mackey, and Murat A. Erdogdu. Stochastic Runge–Kutta accelerates Langevin Monte Carlo and beyond. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper/8990-stochastic-runge-kutta-accelerates-langevin-monte-carlo-and-beyond.pdf.

Yingzhen Li, Richard E. Turner, and Qiang Liu. Approximate inference with amortised MCMC, 2017. URL https://arxiv.org/abs/1702.08343v2. Version 2.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. doi: 10.1063/1.1699114.

Grigori N. Milstein and Michael V. Tretyakov. *Stochastic Numerics for Mathematical Physics*. Springer, 2004. doi: 10.1007/978-3-662-10063-9.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997. doi: 10.2307/1428011. URL https://doi.org/10.2307/1428011.

Kimia Nadjahi, Alain Durmus, Umut Simsekli, and Roland Badeau. Asymptotic guarantees for learning generative models with the sliced-Wasserstein distance. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1906.04516.

Radford M. Neal. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng (eds.), *Handbook of Markov Chain Monte Carlo*, pp. 113–162. Chapman and Hall/CRC, 2011. doi: 10.1201/b10905-6. URL https://www.cs.toronto.edu/~radford/ham-mcmc.abstract.html.

Vincent Pauline, Kirill Tamogashev, Arran Carter, Sanghyeok Choi, Stefan Bauer, and Esmeralda S. Whitammer. Amortised inference through one-step implicit sampling. In *ICML Workshop on Structured Probabilistic Inference and Generative Modeling*, 2026. URL https://mcml.ai/publications/ptc%2B26/.

Tobias Pielok, Bernd Bischl, and David Rügamer. Revisiting unbiased implicit variational inference. *arXiv preprint arXiv:2506.03839*, 2025.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, 2013. doi: 10.1214/13-AOS1140. URL https://arxiv.org/abs/1207.6076.

Tony Shardlow and Andrew M. Stuart. A perturbation theory for ergodic Markov chains and application to numerical approximations. *SIAM Journal on Numerical Analysis*, 37(4):1120–1137, 2000. URL https://people.bath.ac.uk/tjs42/assets/pubs/R7.pdf.

Carl-Johann Simon-Gabriel, Alessandro Barp, Bernhard Schölkopf, and Lester Mackey. Metrizing weak convergence with maximum mean discrepancies. *Journal of Machine Learning Research*, 24(184):1–20, 2023. URL https://jmlr.org/papers/v24/21-0599.html.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(70): 2389–2410, 2011. URL https://jmlr.org/papers/v12/sriperumbudur11a.html.

Zoltán Szabó. Information theoretical estimators toolbox. *Journal of Machine Learning Research*, 15:283–287, 2014.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. doi: 10.1016/j.jspi.2013.03.018. URL https://www.sciencedirect.com/science/article/pii/S0378375813000633.

Denis Talay and Luciano Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Analysis and Applications*, 8(4):483–509, 1990. doi: 10.1080/07362999008809220.

Michalis K. Titsias and Francisco Ruiz. Unbiased implicit variational inference. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 167–176, 2019. URL https://proceedings.mlr.press/v89/titsias19a.html.

John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997. doi: 10.1109/9.580874. URL https://web.mit.edu/~jnt/www/Papers/J063-97-bvr-td.pdf.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1903.08568.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009. doi: 10.1007/978-3-540-71050-9. URL https://doi.org/10.1007/978-3-540-71050-9.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8:279–292, 1992. doi: 10.1007/BF00992698. URL https://doi.org/10.1007/BF00992698.

Jianwen Xie, Yang Lu, Ruiqi Gao, and Ying Nian Wu. Cooperative learning of energy-based model and latent variable model via MCMC teaching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11834. URL https://ojs.aaai.org/index.php/AAAI/article/view/11834.

Mingzhang Yin and Mingyuan Zhou. Semi-implicit variational inference. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5660–5669, 2018. URL https://proceedings.mlr.press/v80/yin18b.html.

Longlin Yu and Cheng Zhang. Semi-implicit variational inference via score matching. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sd90a2ytrt.

A THEORETICAL DETAILS AND PROOFS

In this section, we provide necessary proofs and illustrations for the propositions.

A.1 CONTROLLING SEMINORMS FOR DISTRIBUTION DISCREPANCIES

For a signed residual $\eta = \mu - \nu$ of mass zero, we seek a function seminorm $\mathcal{N}_D$ such that

$$|\eta(g)| \le D(\mu,\nu)\mathcal{N}_D(g). \tag{18}$$

It is homogeneous and subadditive. One intrinsic choice is $\sup_{\mu,\nu} |(\mu-\nu)g|/D(\mu,\nu)$ over admissible pairs with positive discrepancy. However, we do not always need the exact dual choice; a larger, explicitly computable controlling seminorm is sufficient.

Table 3: Valid controlling seminorms. The precise domains and normalisations are given below

Population discrepancy D	A choice of $\mathcal{N}_D(g)$
W_1	$\mathrm{Lip}(g)$
Matérn MMD	$C_{d,\nu}\,\|g\|_{H^{\nu+d/2}}$
Riesz MMD, $0<\alpha<2$	$C_{d,\alpha}\big(\int |\widehat{g}(\xi)|^2|\xi|^{d+\alpha}\,d\xi\big)^{1/2}$
SW_1	$C_d \sup_{\theta\in\mathbb{S}^{d-1}} \int_{\mathbb{R}} |r|^d|\widehat{g}(r\theta)|\,dr$

Wasserstein distance. For laws with finite first moments, Kantorovich duality gives $|\eta(g)| \le W_1(\mu,\nu)\,\mathrm{Lip}(g)$.

Matérn kernels. Let $m_\mu = \int k(x,\cdot)\,d\mu(x)$ be a well-defined RKHS mean embedding and $D(\mu,\nu) = \|m_\mu - m_\nu\|_{\mathcal{H}_k}$. For $g - c \in \mathcal{H}_k$, the reproducing property gives

$$|\eta(g)| = |\langle m_\mu - m_\nu, g - c\rangle_{\mathcal{H}_k}| \le D(\mu,\nu)\|g-c\|_{\mathcal{H}_k}. \tag{19}$$

Taking the infimum over c gives Table 3; the infimum of an empty admissible set is $+\infty$. For a translation-invariant kernel with spectral density w, using $\widehat{g}(\xi) = \int e^{-i\xi\cdot x} g(x)\,dx$, the RKHS norm is proportional to $(\int |\widehat{g}|^2/w)^{1/2}$. For Matérn smoothness $\nu > 0$, $w(\xi)$ is proportional to $(\lambda^2 + |\xi|^2)^{-\nu-d/2}$, giving the Sobolev norm in the table (Kanagawa et al., 2018).

Riesz MMD. For $0<\alpha<2$, use the positive definite anchored kernel $k_\alpha(x,y) = (|x|^\alpha + |y|^\alpha - |x-y|^\alpha)/2$. For laws with finite αth moments, its squared MMD is

$$D_\alpha^2(\mu,\nu) = \mathbb{E}|X-Y|^\alpha - \tfrac{1}{2}\mathbb{E}|X-X'|^\alpha - \tfrac{1}{2}\mathbb{E}|Y-Y'|^\alpha = c_{d,\alpha}\int \frac{|\widehat{\eta}(\xi)|^2}{|\xi|^{d+\alpha}}\,d\xi, \tag{20}$$

where the samples are independent with their indicated laws and $c_{d,\alpha} > 0$. This is the energy-distance correspondence (Sejdinovic et al., 2013). The characteristic-function representation is given by (Székely & Rizzo, 2013), up to the factor of two in our kernel normalization. For Schwartz g, Fourier inversion and weighted Cauchy–Schwarz give

$$|\eta(g)| \le (2\pi)^{-d}\left(\int |\widehat{g}(\xi)|^2|\xi|^{d+\alpha}d\xi\right)^{1/2}\left(\int |\widehat{\eta}(\xi)|^2|\xi|^{-d-\alpha}d\xi\right)^{1/2}. \tag{21}$$

This proves the stated seminorm with its normalisation-dependent constant.

Sliced Wasserstein. Let σ be uniform probability measure on $\mathbb{S}^{d-1}$, $\mu_\theta = (x \mapsto \theta^\top x)_{\#}\mu$, and $\mathrm{SW}_1(\mu,\nu) = \int W_1(\mu_\theta,\nu_\theta)d\sigma(\theta)$ (Bonnotte, 2013; Nadjahi et al., 2019). Assume finite first moments. Any measurable representation $g(x) = c + \int g_\theta(\theta^\top x)d\sigma(\theta)$ with uniformly bounded Lipschitz constants and normalised $g_\theta(0) = 0$ gives

$$|\eta(g)| \le \left(\operatorname*{ess\,sup}_\theta \mathrm{Lip}(g_\theta)\right)\mathrm{SW}_1(\mu,\nu). \tag{22}$$

To construct these ridge functions, use Fourier inversion in polar coordinates and the slice identity $\widehat{\mu_\theta}(r) = \widehat{\mu}(r\theta)$. (Bonnotte, 2013). With our Fourier normalization, a real Schwartz test g satisfies

$$g(x) = C_d \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} e^{ir\theta^\top x} |r|^{d-1} \widehat{g}(r\theta)\, dr\, d\sigma(\theta), \quad C_d = \frac{|\mathbb{S}^{d-1}|}{2(2\pi)^d}.$$

The factor $1/2$ accounts for using all signed radii. Thus set

$$g_\theta(u) = C_d \int_{\mathbb{R}} (e^{iru} - 1)|r|^{d-1} \widehat{g}(r\theta)\, dr, \quad g(x) = g(0) + \int g_\theta(\theta^\top x)\, d\sigma(\theta). \tag{23}$$

Conjugate symmetry makes g_θ real, and $g_\theta(0) = 0$. Differentiating in u bounds its Lipschitz constant by $C_d \int |r|^d |\widehat{g}(r\theta)| dr$ and proves the table's sufficient seminorm. It is finite for C_c^∞ tests. The same derivation applies when the Fourier integrals remain absolutely convergent, including the propagated tests controlled in Appendix A.4.

Higher-order Wasserstein discrepancies. For $1 \le p < \infty$, monotonicity of transport costs gives $W_1(\mu, \nu) \le W_p(\mu, \nu)$ whenever the pth moments are finite. With the conventional definition $\mathrm{SW}_p(\mu, \nu) = (\int W_p^p(\mu_\theta, \nu_\theta)\, d\sigma(\theta))^{1/p}$, the same monotonicity and Hölder's inequality give

$$\mathrm{SW}_1(\mu, \nu) \le \int W_p(\mu_\theta, \nu_\theta)\, d\sigma(\theta) \le \mathrm{SW}_p(\mu, \nu).$$

Thus the W_1 and SW_1 controlling seminorms also apply to W_p and SW_p, respectively, with no change of constants.

A.2 Proof of Theorem 4.5

Proof. For $k < n$, subtract consecutive quantities $\mu_{j+1}(K_{j+1,n} f)$ and $\mu_j(K_{j,n} f)$, using $K_{j,n} = \widetilde{T}_j K_{j+1,n}$ and $\mu_{j+1} = (T_j)_\# \mu_j + \eta_j$. Summation gives

$$\mu_n(f) - \mu_k(K_{k,n} f) = \sum_{j=k}^{n-1} \eta_j(K_{j+1,n} f) + \sum_{j=k}^{n-1} \mu_j((T_j - \widetilde{T}_j) K_{j+1,n} f).$$

Since $(K_{k,n})_\# \pi = \pi$, this proves equation 10.

We first express local consistency uniformly over the remaining propagation time. For each fixed $f \in C_c^\infty$ and $A > 0$, set

$$\beta_j(f, A) := \sup_{\substack{n > j \\ s_n - s_{j+1} \le A}} \frac{|\mu_j((T_j - \widetilde{T}_j) K_{j+1,n} f)|}{h_j}. \tag{24}$$

The index set is nonempty because it contains $n = j + 1$, and the supremum is finite because Markov kernels contract the supremum norm. Assumption 4.4 implies $\beta_j(f, A) \to 0$. Indeed, otherwise there would be a subsequence of indices and admissible terminal times for which the displayed quotient is bounded below by a positive constant. Choosing those terminal times on the subsequence and $n_j = j + 1$ elsewhere would contradict the assumption.

Fix $A > 0$, take $s_n \ge A$, and write $k = k(n, A)$. Every remaining propagation time in the decomposition is at most $A + \bar{h}$. The assumptions give

$$\begin{aligned} |(\mu_n - \pi) f| &\le C_f e^{-\lambda_f A} + M_D e^{\omega_D(A+\bar{h})} \mathcal{N}_D(f) \sum_{j=k}^{n-1} \varepsilon_j + \sum_{j=k}^{n-1} h_j \beta_j(f, A + \bar{h}) \\ &\le C_f e^{-\lambda_f A} + (A + \bar{h}) \left[M_D e^{\omega_D(A+\bar{h})} \mathcal{N}_D(f) \sup_{j \ge k} \frac{\varepsilon_j}{h_j} + \sup_{j \ge k} \beta_j(f, A + \bar{h}) \right]. \end{aligned} \tag{25}$$

Here $\sum_{j=k}^{n-1} h_j = s_n - s_k \le A + \bar{h}$. For fixed A, $k(n, A) \to \infty$ as $s_n \to \infty$. Both tail suprema therefore vanish, and $\limsup_n |(\mu_n - \pi) f| \le C_f e^{-\lambda_f A}$. Letting $A \to \infty$ proves convergence for every $f \in C_c^\infty$.

For completeness, these tests determine weak convergence to the probability measure π. Uniform approximation first extends convergence to $C_0(\mathbb{R}^d)$, since $|(\mu_n - \pi)g| \le 2\|g\|_\infty$. Given $\delta > 0$, choose a compactly supported cutoff $0 \le \chi \le 1$ with $\pi(\chi) > 1 - \delta$. Then $\mu_n(\chi) > 1 - 2\delta$ for all sufficiently large n. For $g \in C_b$, the test $g\chi$ belongs to C_0, while $|(\mu_n - \pi)(g(1-\chi))| \le 3\delta\|g\|_\infty$ eventually. Letting $\delta \downarrow 0$ yields $\mu_n \xrightarrow{w} \pi$; see also Billingsley (1999); Dudley (2002). □

The proof first fixes A and lets $n \to \infty$, then lets $A \to \infty$. For quantitative rates, a growing window can instead balance reference mixing against the accumulated errors.

A.3 Verification of reference mixing

We prove the first part of Theorem 4.6 under its LSI and uniform smoothed-entropy conditions.

Proof of the LSI mixing estimate. Write $\nu_{j,t} = (P_t)_\# \mu_j$. Entropy dissipation and the log-Sobolev inequality give

$$\frac{d}{dt}\mathrm{KL}(\nu_{j,t}\|\pi) = -I(\nu_{j,t}\|\pi) \le -2\rho\,\mathrm{KL}(\nu_{j,t}\|\pi),$$

where I denotes relative Fisher information. Consequently,

$$\mathrm{KL}(\nu_{j,t}\|\pi) \le e^{-2\rho(t-t_0)} H_{t_0}, \quad t \ge t_0. \tag{26}$$

These identities can first be applied to regular densities and then extended by semigroup regularisation and lower semicontinuity of entropy (Bakry et al., 2014; Vempala & Wibisono, 2019). Pinsker's inequality implies

$$|\nu_{j,t}(f) - \pi(f)| \le \|f\|_\infty \sqrt{2\mathrm{KL}(\nu_{j,t}\|\pi)} \le \|f\|_\infty \sqrt{2H_{t_0}}\, e^{-\rho(t-t_0)}.$$

For $0 \le t < t_0$, the bound $2\|f\|_\infty$ suffices. Combining the two ranges gives $|\mu_j(P_t f) - \pi(f)| \le C_f e^{-\rho t}$ with the stated C_f, uniformly in j. Substituting $t = s_n - s_j$ verifies Assumption 4.2. □

A.4 Verification of seminorm propagation

Appendix A.1 gives Fourier- and RKHS-based controlling seminorms for the discrepancies. To verify their propagation under the Langevin semigroup, we use stronger integer-order Sobolev norms as auxiliary majorants. For an integer $q \ge 0$ and $r \in \{1, 2\}$, set For SW_p, Riesz MMD, and Matérn MMD, the propagation argument has two steps. First, we prove that the Langevin semigroup grows at most exponentially in suitable integer-order Sobolev norms. Second, we show that these Sobolev norms dominate the Fourier- or RKHS-based controlling seminorms from Appendix A.1. We may therefore use the Sobolev norm itself as the controlling seminorm $\mathcal{N}_D$ in Assumption 4.3. This assumption only requires a seminorm that controls the discrepancy; it does not require the smallest or most intrinsic one. For an integer $q \ge 0$ and $r \in \{1, 2\}$, define

$$\|g\|_{W^{q,r}} := \sum_{|\beta| \le q} \|D^\beta g\|_{L^r},$$

where β ranges over multi-indices. These derivative norms are proof tools: below we show that, at a sufficiently high order, they dominate the Fourier or RKHS controlling seminorm and are propagated by the semigroup. They do not change the training discrepancy D. tools for verifying the required propagation assumption; they do not change the training discrepancy D.

The derivative estimates for the stochastic flow and its inverse are standard consequences of the regularity theory of stochastic flows; see, e.g., Kunita Kunita (1990a). Under our assumptions on the derivatives of U, the flow map Φ_t is a global C^q-diffeomorphism, and for $1 \le r \le q$,

$$\sup_x \|D^r \Phi_t(x)\| \le C_q e^{c_q t}, \qquad \sup_y |\det D\Phi_t^{-1}(y)| \le e^{dLt}.$$

Combining these estimates with the chain rule and a change of variables yields a propagation bound for the auxiliary Sobolev norms. We verify Assumption 4.3 for the Langevin semigroup in Theorem 4.6. For $D = W_p$, no smoothness beyond the globally L-Lipschitz gradient is needed. For the other discrepancies, choose an integer q such that $q > d + 1$ for $D = \mathrm{SW}_p$, $q \ge (d + \alpha)/2$

for Riesz MMD, and $q \geq \nu + d/2$ for Matérn MMD. In these cases, assume that $U \in C^{q+1}(\mathbb{R}^d)$ and its derivatives of orders $2, \ldots, q+1$ are globally bounded. These are sufficient smoothness conditions for the Sobolev controlling seminorms specified below; their dual bounds follow from Appendix A.1.

Wasserstein discrepancies. For $D = W_p$, synchronous coupling and Gronwall's inequality give $|X_t^x - X_t^y| \leq e^{Lt}|x-y|$. Since

$$\mathcal{N}_D(P_t f) = \mathrm{Lip}(P_t f) \leq e^{Lt}\,\mathrm{Lip}(f) = e^{Lt}\mathcal{N}_D(f), \tag{27}$$

and $W_1 \leq W_p$, this proves the required propagation bound for every $p \geq 1$.

Derivative bounds for the Langevin flow. Fix one of the remaining discrepancies and let q be as above. On each Brownian path, subtracting $\sqrt{2}B_t$ from the diffusion reduces the equation to an ODE with a time-dependent vector field. The assumed bounds on the derivatives of U ensure that its solution map $\Phi_t(x) = X_t^x$ is a global C^q diffeomorphism, as in the stochastic-flow framework of Kunita (1990b). We record the derivative and Jacobian estimates used below.

Set $A_r(t) = \sup_x \|D^r \Phi_t(x)\|$. Differentiating the flow equation gives $A_1(t) \leq e^{Lt}$. For $2 \leq r \leq q$, the chain rule yields

$$A_r(t) \leq \int_0^t \left[L A_r(u) + C_r \sum_{\ell=2}^{r} \sum_{\substack{r_1+\cdots+r_\ell=r \\ r_i \geq 1}} \prod_{i=1}^{\ell} A_{r_i}(u) \right] du. \tag{28}$$

Each $r_i < r$, so induction and Gronwall's inequality give $A_r(t) \leq C_q e^{c_q t}$, with deterministic constants independent of the Brownian path. Liouville's formula gives the Jacobian bound

$$\det D\Phi_t(x) = \exp\left[-\int_0^t \Delta U(\Phi_u(x))\, du \right], \quad |\det D\Phi_t^{-1}(y)| \leq e^{dLt}. \tag{29}$$

Propagation in Sobolev norms. For $r \in \{1, 2\}$, the chain rule, Minkowski's inequality, and the pathwise change of variables imply

$$\begin{aligned} \|D^\gamma P_t f\|_{L^r} &\leq C_q e^{c_q t} \sum_{|\beta| \leq |\gamma|} \mathbb{E}\|D^\beta f \circ \Phi_t\|_{L^r} \\ &\leq C_q e^{(c_q + dL/r)t} \sum_{|\beta| \leq |\gamma|} \|D^\beta f\|_{L^r}, \quad |\gamma| \leq q. \end{aligned} \tag{30}$$

For $|\gamma| = 0$ the same bound follows without differentiating. The pathwise derivative bounds justify differentiation under the expectation for $f \in C_c^\infty$. Summing over γ yields

$$\|P_t f\|_{W^{q,r}} \leq M_{q,r} e^{\omega_{q,r} t} \|f\|_{W^{q,r}}, \quad r \in \{1, 2\}, \quad t \geq 0. \tag{31}$$

Equivalent integer Sobolev norms only change $M_{q,2}$, so the same statement holds for $H^q = W^{q,2}$. The constants are independent of the Brownian path and of the student laws.

Using Sobolev norms to control $\mathcal{N}_D$. We now complete the second step by checking that the selected Sobolev norms dominate the Fourier or RKHS bounds in Appendix A.1. The Fourier estimates there, with integration by parts for SW_p, allow us to use $\mathcal{N}_D(g) = C_D\|g\|_{W^{q,1}}$ for SW_p when $q > d+1$, and $\mathcal{N}_D(g) = C_D\|g\|_{W^{q,2}}$ for Riesz MMD when $q \geq (d+\alpha)/2$. For Matérn MMD, take $\mathcal{N}_D(g) = C_D\|g\|_{W^{q,2}}$ with $q \geq \nu + d/2$; this Sobolev norm dominates the $H_{\nu+d/2}$ seminorm in Appendix A.1. These are valid controlling seminorms since each dominates the corresponding Fourier or RKHS seminorm used to prove the dual inequality there. Applying equation 31 with $r = 1$ for SW_p and $r = 2$ for Riesz and Matérn MMD therefore gives constants $M_D > 0$ and $\omega_D \geq 0$ such that $\mathcal{N}_D(P_t f) \leq M_D e^{\omega_D t}\mathcal{N}_D(f)$. $\mathcal{N}_D(P_t f) \leq M_D e^{\omega_D t}\mathcal{N}_D(f)$. This is the propagation bound required by Assumption 4.3 for these three discrepancies. The propagated test belongs to C_b by the Feller property and the supremum-norm contraction of P_t. Since $K_{j,n} = P_{s_n - s_j}$, substituting $t = s_n - s_j$ proves Assumption 4.3.

The constants are independent of the student distributions. They depend on D through the required Sobolev order and the normalisation of $\mathcal{N}_D$, and on the derivative bounds of U. No contractivity in the original Fourier or RKHS norm is asserted or needed.

A.5 VERIFICATION OF LOCAL CONSISTENCY AND COMPLETION OF THEOREM 4.6

The local estimate uses only the Lipschitz gradient and its first moment under the current student law.

Proof. From the same initial point x, couple the exact diffusion and one ULA step using the same Brownian motion:

$$X_u^x = x - \int_0^u \nabla U(X_v^x)\, dv + \sqrt{2}B_u, \quad Y_h^x = x - h\nabla U(x) + \sqrt{2}B_h.$$

The Lipschitz property and $\mathbb{E}|B_u| \le \sqrt{du}$ imply

$$\mathbb{E}|X_u^x - x| \le L \int_0^u \mathbb{E}|X_v^x - x|\, dv + u|\nabla U(x)| + \sqrt{2du}.$$

The last two terms are nondecreasing in u, so Gronwall's inequality gives $\mathbb{E}|X_u^x - x| \le e^{Lu}(u|\nabla U(x)| + \sqrt{2du})$. The Brownian increments cancel in $X_h^x - Y_h^x$, and therefore

$$\mathbb{E}|X_h^x - Y_h^x| \le L \int_0^h \mathbb{E}|X_u^x - x|\, du \le Le^{Lh}\left[\frac{h^2}{2}|\nabla U(x)| + \frac{2\sqrt{2d}}{3}h^{3/2}\right]. \tag{32}$$

For a Lipschitz test g, the same coupling bounds $|(T_h - P_h)g(x)|$ by $\mathrm{Lip}(g)\mathbb{E}|X_h^x - Y_h^x|$. Set $g = P_t f$, use equation 27, and integrate the pointwise absolute value against μ_j. This gives

$$\mathbb{E}_{\mu_j}|(T_j - P_{h_j})P_t f| \le Le^{L(t+h_j)}\,\mathrm{Lip}(f)\left[\frac{h_j^2}{2}\mathbb{E}_{\mu_j}|\nabla U| + \frac{2\sqrt{2d}}{3}h_j^{3/2}\right]. \tag{33}$$

Replacing the moment by G proves equation 15. This coupling argument is consistent with the local-deviation analyses used for Langevin discretisations (Dalalyan, 2017; Li et al., 2019).

For any admissible sequence (n_j) in Assumption 4.4, take $t = s_{n_j} - s_{j+1} \le A$. Then

$$\frac{|\mu_j((T_j - P_{h_j})P_t f)|}{h_j} \le Le^{L(A+\bar{h})}\,\mathrm{Lip}(f)\left[\frac{G}{2}h_j + \frac{2\sqrt{2d}}{3}\sqrt{h_j}\right] \longrightarrow 0 \tag{34}$$

whenever $h_j \to 0$, verifying local consistency.

To conclude the theorem, fix $k = k(n, A)$ and use equation 10 with $K_{j,n} = P_{s_n - s_j}$. Set $t_{j,n} = s_n - s_{j+1}$ and $B_f = L\,\mathrm{Lip}(f)(G\sqrt{h}/2 + 2\sqrt{2d}/3)$. Since $h_j^2 \le \sqrt{h}\, h_j^{3/2}$, the estimates above give

$$|(\mu_n - \pi)f| \le C_f e^{-\rho(s_n - s_k)} + M_D\mathcal{N}_D(f)\sum_{j=k}^{n-1} e^{\omega_D t_{j,n}}\varepsilon_j + B_f\sum_{j=k}^{n-1} e^{L(t_{j,n}+h_j)}h_j^{3/2}. \tag{35}$$

Use $A \le s_n - s_k \le A + \bar{h}$ to obtain equation 16. For a fixed window length,

$$\sum_{j=k}^{n-1}\varepsilon_j \le (A + \bar{h})\sup_{j\ge k}\frac{\varepsilon_j}{h_j}, \quad \sum_{j=k}^{n-1} h_j^{3/2} \le (A + \bar{h})\sup_{j\ge k}\sqrt{h_j}.$$

Both quantities tend to zero as $n \to \infty$ under the stated schedules. Letting $A \to \infty$ establishes weak convergence, equivalently by Theorem 4.5. For an exact Langevin teacher, $(T_j - P_{h_j})P_t f = 0$, so the second schedule condition $h_j \to 0$ is unnecessary. □

The order $h_j^{3/2}$ is sufficient for local consistency after division by h_j. Under additional regularity and integrability conditions, we can obtain standard global first-order weak convergence rate(thus second order locally) for Euler-Maruyama scheme (E et al., 2019).

A.6 Discussions of convergence rates

Suppose reference mixing is bounded by $C_f e^{-\rho t}$ as established in Appendix A.3. Set $B_f := L\,\mathrm{Lip}(f)(G\sqrt{\bar{h}}/2 + 2\sqrt{2d}/3)$. The explicit estimate equation 35 gives

$$|(\mu_n - \pi)f| \le C_f e^{-\lambda A} + (A+\bar{h})M_D \mathcal{N}_D(f) e^{\omega_D(A+\bar{h})} \sup_{j \ge k(n,A)} \frac{\varepsilon_j}{h_j} + (A+\bar{h}) B_f e^{L(A+\bar{h})} \sup_{j \ge k(n,A)} \sqrt{h_j}. \tag{36}$$

This bound exhibits the balance between the decay of reference mixing and the accumulation of fitting and discretization errors. Put $\omega = \max\{\omega_D, L\}$.

Proposition A.1 (Rates under polynomial budgets)**.** *Under the conditions of Theorem 4.6, suppose $h_n = c(n+1)^{-a}$ with $c > 0$ and $0 < a < 1$, and $\varepsilon_n/h_n = O((n+1)^{-\gamma})$ with $\gamma > 0$. Then, for each $f \in C_c^\infty$, setting $\delta = \min\{\gamma, a/2\}$ gives*

$$|\mathbb{E}_{\mu_n} f(X) - \mathbb{E}_\pi f(X)| = O_f\left(\left[\frac{\log(n+1)}{(n+1)^\delta} \right]^{\rho/(\rho+\omega)} \right). \tag{37}$$

Proof. For sufficiently large n, choose

$$A_n = \frac{\delta \log(n+1) - \log\log(n+e)}{\rho + \omega}.$$

Since $s_n - s_{\lfloor n/2 \rfloor} \asymp n^{1-a}$ and $A_n = O(\log n)$, we have $k(n, A_n) \ge \lfloor n/2 \rfloor$ eventually. The two suprema in equation 36 are therefore $O(n^{-\gamma})$ and $O(n^{-a/2})$. Their combined contribution is $O_f((A_n+1)e^{\omega A_n} n^{-\delta})$, while the mixing term is $O_f(e^{-\rho A_n})$. The chosen window bounds both by $n^{-\delta\rho/(\rho+\omega)}(\log n)^{\rho/(\rho+\omega)}$. □

A.7 Sinkhorn fitting with vanishing regularisation

Sinkhorn divergences provide sample-based fitting objectives (Feydy et al., 2019), but their definition does not directly give the fixed controlling seminorm in equation 18. We instead compare them with Wasserstein distance and account for the entropy bias as an additional error. This subsection uses the explicit normalization

$$\begin{aligned} \mathrm{OT}_{\tau,p}(\mu,\nu) &= \inf_{\xi \in \Pi(\mu,\nu)} \left\{ \int \|x-y\|^p \, d\zeta + \tau \mathrm{KL}(\xi \| \mu \otimes \nu) \right\}, \\ S_{\tau,p}(\mu,\nu) &= \mathrm{OT}_{\tau,p}(\mu,\nu) - \tfrac{1}{2}\mathrm{OT}_{\tau,p}(\mu,\mu) - \tfrac{1}{2}\mathrm{OT}_{\tau,p}(\nu,\nu), \end{aligned} \tag{38}$$

where $p \ge 1$, $\tau > 0$, and $\Pi(\mu,\nu)$ is the set of couplings. We also establish convergence results with decaying τ_n.

Proposition A.2. *Let μ, ν have finite pth moments and put $M = \max\{(\mathbb{E}_\mu |X|^p)^{1/p}, (\mathbb{E}_\nu |X|^p)^{1/p}\}$. Then*

$$\begin{aligned} W_1(\mu,\nu) \le W_p(\mu,\nu) &\le \left[S_{\tau,p}(\mu,\nu) + B_{d,p}(\tau, M) \right]^{1/p}, \\ B_{d,p}(\tau, M) &:= d^{p/2}\tau + d\tau \log\left[2e \left(2 + \frac{M}{\tau^{1/p}} \right) \right] = \mathcal{O}(\tau \log \frac{1}{\tau}). \end{aligned} \tag{39}$$

Proof. Partition $\mathbb{R}^d$ into half-open cubes of side length $a > 0$, indexed by $z \in \mathbb{Z}^d$. For $X \sim \mu$, let $N = \lfloor X/a \rfloor \in \mathbb{R}^d$ be its cell index. Draw a cell with its μ-probability, then draw X, Y independently from μ conditioned on that cell. Denote the distribution by $\xi = \sum_z \frac{1}{p_z} \mu_z \otimes \mu_z$. Since $\mu \otimes \mu = \sum_{z,w} p_z p_w \mu_z \otimes \mu_w$, we have $\frac{\mathrm{d}\xi}{\mathrm{d}\mu\otimes\mu} = \frac{1}{p_z}$ on the relative cube z.

Under the distribution ξ we constructed, we can control

$$\int \|x-y\|^p \mathrm{d}\xi = \sum_z p_z \int_z \|x-y\|^p d\mu_z \otimes \mu_z \le \sum_z p_z d^{\frac{p}{2}} a^p = (a\sqrt{d})^p. \tag{40}$$

We also have

$$\mathrm{KL}(\xi \parallel \mu \otimes \mu) = \int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\xi}{\mathrm{d}(\mu \otimes \mu)}\right) \mathrm{d}(\mu \otimes \mu) = \sum_z p_z \log \frac{1}{p_z} = H(N) \tag{41}$$

Where $H(N)$ is the Shannon entropy of the discrete distribution of N. For each coordinate $N_i \in \mathbb{Z}$, we can add an independent noise $U \sim \mathrm{Unif}[-\frac{1}{2}, \frac{1}{2})$, so that $N_i + U$ is a continuous random variable on $\mathbb{R}$. Since the intervals corresponding to different values of N_i are disjoint, we have $h(N_i + U) = H(N_i)$. By the moment assumption, $\mathbb{E}|N_i| \le \mathbb{E}|X_i| + 1 \le \frac{M}{a} + 1$.

Given the first-order absolute moment, the Laplace distribution maximizes the differential entropy, which gives

$$H(N_i) = h(N_i + U) \le \log\left(2e\,\mathbb{E}|N_i + U|\right) \le \log\left(2e\left(\frac{M}{a} + 1 + \frac{1}{4}\right)\right) < \log\left[2e\left(2 + \frac{M}{a}\right)\right]. \tag{42}$$

Here, we used $\mathbb{E}|U| = \frac{1}{4}$ and the triangle inequality. Consequently, by the subadditivity of Shannon entropy,

$$H(N) \le \sum_{i=1}^{d} H(N_i) \le \sum_{i=1}^{d} \log\left[2e\left(1 + \mathbb{E}|N_i|\right)\right] \le d \log\left[2e\left(2 + \frac{M}{a}\right)\right] \tag{43}$$

where the last inequality follows from Jensen's inequality. According to the formulas above, we have

$$\mathrm{OT}_{\tau,p}(\mu, \mu) \le \int \|x - y\|^p \mathrm{d}\xi + \tau H(N) \le a^p d^{\frac{p}{2}} + d\tau \log\left[2e\left(2 + \frac{M}{a}\right)\right], \quad \forall a > 0. \tag{44}$$

Taking $a = \tau^{1/p}$, the constructed coupling therefore yields

$$\mathrm{OT}_{\tau,p}(\mu, \mu) \le d^{p/2}\tau + d\tau \log\left[2e\left(2 + \frac{M}{\tau^{1/p}}\right)\right] = B_{d,p}(\tau, M).$$

The same bound holds for ν. Finally, nonnegativity of the KL penalty implies

$$W_p^p(\mu, \nu) \le \mathrm{OT}_{\tau,p}(\mu, \nu) = S_{\tau,p}(\mu, \nu) + \tfrac{1}{2}\mathrm{OT}_{\tau,p}(\mu, \mu) + \tfrac{1}{2}\mathrm{OT}_{\tau,p}(\nu, \nu) \le S_{\tau,p}(\mu, \nu) + B_{d,p}(\tau, M).$$

Taking the pth root and using $W_1 \le W_p$ completes the proof. □

Back to our MCMC distillation setting. Assume $\sup_n \left[\mathbb{E}_{\mu_n}|X|^p, \mathbb{E}_{T_{n\#}\mu_n}|X|^p\right] < +\infty$. For $\tau_n \le \tau_1$, we can bound

$$W_1(\mu_{n+1}, T_{n\#}\mu_n) \le \left(\varepsilon_n^p + C\tau_n \log\frac{1}{\tau_n}\right)^{1/p} \tag{45}$$

where C depends on the dimension, p, and a uniform bound on the pairwise pth moments. Consequently, to ensure that the distillation residual is $o(h_n)$, it suffices to require

$$\varepsilon_n = o(h_n), \qquad \tau_n \log\frac{1}{\tau_n} = o(h_n^p).$$

Under these conditions, the W_1 branch of Theorem 4.6 guarantees that $\mu_n \xrightarrow{w} \pi$, provided its remaining assumptions hold.

B Experimental Details

B.1 Shared protocol

The learned sampler is instantiated with either MMD with a Laplace kernel whose bandwidth is fitted to the median of the current generator batch or sliced Wasserstein-2 with 128 random projections per training update. Both variants use 1,024 training samples per update.

For all learned variants, each outer iteration first makes five Langevin refinement steps and then updates the generator. On annealed targets, let

$$\beta_t = 0.1 + 0.9 \min\left\{\frac{t}{T_{\text{ann}}}, 1\right\}, \qquad \delta_t = \frac{\delta_0}{\beta_t}, \tag{46}$$

where T_{ann} is the annealing interval and δ_0 is the base step size in Table 4. If z is a current sample, one refinement step is

$$z^+ = z + \frac{\delta_t}{2}\beta_t \nabla_z \log \widetilde{p}(z) + \sqrt{\delta_t}\,\xi, \qquad \xi \sim \mathcal{N}(0, I), \tag{47}$$

where $\widetilde{p}$ may be unnormalized. ULA applies this transition with $\beta_t = 1$ and $\delta_t = \delta_0$ to a persistent particle population initialized independently from $\mathcal{N}(0, I)$.

The two-dimensional generator is a conditional diagonal Gaussian. A three-hidden-layer ELU network of width 256 maps 32-dimensional standard Gaussian noise to its mean and log standard deviation; log standard deviations are bounded below by -3. The logistic regression generator, we use a two-hidden-layer ReLU network of width 100 to map a 10-dimensional standard gaussian noise to the 22-dimensional conditional mean. For the diffusion posterior, a two-hidden-layer ReLU network of width 128 maps 100-dimensional uniform noise to the conditional mean. It then adds independent Gaussian noise in each path coordinate, with a learned, coordinate-specific standard deviation shared across noise inputs. Each standard deviation is initialized to approximately 0.13 and then learned along with the network weights.

All learned variants use Adam with $(\beta_1, \beta_2) = (0.9, 0.999)$, gradient clipping at norm 10, and a stepwise learning-rate multiplier of 0.75. Table 4 gives the two-dimensional and diffusion schedules. Among those targets, Student-t uses no annealing; the others anneal over the first half of training. ULA uses the listed transition step throughout.

For the two-dimensional and diffusion targets, each method is run with five fixed random seeds. Each plotted panel uses 2,000 draws. Evaluation occurs at ten equally spaced checkpoints.

B.2 Two-dimensional target densities

All targets in this subsection are distributions on $\mathbb{R}^2$.

Banana. Let (U, V) be jointly Gaussian with zero mean and covariance

$$\Sigma_{\text{ban}} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}. \tag{48}$$

The target is the volume-preserving transformation

$$Z_1 = U, \qquad Z_2 = V - U^2 - 1, \tag{49}$$

or, equivalently,

$$p_{\text{ban}}(z) = \mathcal{N}\left(\begin{bmatrix} z_1 \\ z_1^2 + z_2 + 1 \end{bmatrix}; 0, \Sigma_{\text{ban}}\right). \tag{50}$$

This transformation produces a narrow, curved high-density region.

Table 4: Target-specific training settings. The decay interval is measured in outer iterations; every learned-method iteration contains five Langevin refinements.

Target	Iterations	Learning rate	Decay interval	Annealing	ULA step
Banana	10,000	10^{-4}	500	5,000	0.01
X-shaped	10,000	2×10^{-3}	500	5,000	0.10
Eight Gaussians	10,000	5×10^{-3}	500	5,000	0.05
Student-t	10,000	10^{-3}	500	disabled	0.10
Logistic Regression	100,000	10^{-3}	5,000	disabled	0.005
Diffusion posterior	100,000	10^{-3}	5,000	50,000	0.005

X-shaped Gaussian mixture. The target is

$$p_{\mathrm{X}}(z) = \frac{1}{2}\mathcal{N}(z; 0, \Sigma_+) + \frac{1}{2}\mathcal{N}(z; 0, \Sigma_-), \quad \Sigma_+ = \begin{pmatrix} 2 & 1.8 \\ 1.8 & 2 \end{pmatrix}, \quad \Sigma_- = \begin{pmatrix} 2 & -1.8 \\ -1.8 & 2 \end{pmatrix}. \tag{51}$$

The component correlations are ± 0.9, giving two narrow ridges that cross at the origin.

Eight-Gaussian mixture. Let $\theta_k = 2\pi k/8$ for $k = 0, \ldots, 7$ and $\mu_k = 4(\cos\theta_k, \sin\theta_k)$. The target is

$$p_{8\mathrm{G}}(z) = \frac{1}{8}\sum_{k=0}^{7} \mathcal{N}(z; \mu_k, 0.5^2 I_2). \tag{52}$$

The component separation makes missing modes and spurious bridges directly visible.

Transformed Student-t. Let U_1 and U_2 be independent univariate Student-t random variables with $\nu = 2$ degrees of freedom, and set $Z = AU$ for a fixed invertible linear transformation $A \in \mathbb{R}^{2\times 2}$ used in every run. Writing $u = A^{-1}z$, the change-of-variables formula gives the density

$$p_t(z) = |\det A|^{-1} \prod_{j=1}^{2} \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)} \left(1 + \frac{u_j^2}{\nu}\right)^{-(\nu+1)/2}, \qquad \nu = 2. \tag{53}$$

This mixing produces a nonelliptical distribution rather than a standard multivariate Student-t. Each base marginal has tails proportional to $|u_j|^{-3}$, so the target has infinite second moment and global Wasserstein-2 is unsuitable as an evaluation metric.

B.3 Two-dimensional evaluation

For targets with finite second moments, we report the sliced Wasserstein-2 diagnostic

$$\mathrm{SW}_2(q, p) = \left\{\mathbb{E}_{\theta\sim\mathrm{Unif}(\mathbb{S}^1)}\left[W_2^2(\theta^\top q, \theta^\top p)\right]\right\}^{1/2}, \tag{54}$$

approximated with 1,000 random projections and 10,000 samples from each distribution. For the Student-t target, we first retain samples from both distributions inside $[-a, a]^2$, then compute the same statistic using equal retained sample counts. We evaluate $a \in \{6, 8, 10\}$ and report $a = 6$ in the main text. Because this conditioning does not penalize mismatched retained mass, the retained fractions are a necessary companion diagnostic.

We estimate $\mathrm{KL}(q\|p)$ from 10,000 samples from each distribution using a k-nearest-neighbor estimator with $k = 5$ in the ITE toolbox Szabó (2014). For the Student-t diagnostic at $a = 6$, all samplers and an independent target reference retain over 99% of draws in the box. The conditional sliced-W_2 statistic still omits differences in tail mass outside the box.

B.4 Waveform logistic regression

The WAVEFORM data have 400 labeled training observations and 22 predictors, including an intercept. With binary labels $y_i \in \{0, 1\}$, design rows $x_i \in \mathbb{R}^{22}$, and coefficient vector $\beta \in \mathbb{R}^{22}$, the unnormalized posterior is

$$\log \widetilde{p}(\beta \mid X, y) = \sum_{i=1}^{400} \left[y_i x_i^\top \beta - \log\left(1 + \exp(x_i^\top \beta)\right)\right] - \frac{0.01}{2}\|\beta\|_2^2. \tag{55}$$

All refinement and ULA transitions evaluate the full-data posterior score.

B.5 Conditioned diffusion posterior

Let $\Delta t = 0.01$ and $u_0 = 0$. The 100-dimensional latent path prior used for inference is the Euler discretization

$$u_j \mid u_{j-1} \sim \mathcal{N}\left(u_{j-1} + 10\frac{u_{j-1} - u_{j-1}^3}{1 + u_{j-1}^2}\Delta t, \Delta t\right), \qquad j = 1, \ldots, 100. \tag{56}$$

Twenty observations are taken every five time points:

$$y_k \mid u_{5k} \sim \mathcal{N}(u_{5k}, 0.1^2), \qquad k = 1, \ldots, 20. \tag{57}$$

Thus the unnormalized posterior is

$$\log \widetilde{p}(u \mid y) = -\sum_{j=1}^{100} \frac{\left[u_j - u_{j-1} - 10\frac{u_{j-1}-u_{j-1}^3}{1+u_{j-1}^2}\Delta t\right]^2}{2\Delta t} \tag{58}$$

$$-\sum_{k=1}^{20} \frac{(u_{5k} - y_k)^2}{2(0.1)^2}. \tag{59}$$

The displayed latent path and observations are fixed by data-generation seed 2022. The data generator uses the untamed cubic drift $10u(1-u^2)$, whereas the inference density above uses the tamed drift. This difference applies to every method.

The posterior diagnostic is coordinatewise expected log marginal density (ELM). For fixed reference paths $r^{(i)}$ and Gaussian kernel density estimates $\widehat{q}_j$ fitted separately to each coordinate of candidate paths, we compute

$$\mathrm{ELM}(q) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{100} \log \widehat{q}_j\left(r_j^{(i)}\right). \tag{60}$$

The KDE bandwidth for coordinate j is $\max\{10^{-6}, \widehat{\sigma}_j n^{-1/5}\}$, with $\widehat{\sigma}_j$ the candidate sample standard deviation and n the candidate count. The fixed reference paths were generated with ULA using the smaller step size 10^{-4} and 200,000 transitions. The same 1,000 reference paths are used for each ELM evaluation. ELM assesses coordinatewise density and does not measure temporal dependence.