A SPLITTING METHOD FOR SDE TERMINAL-LAW ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In many settings involving stochastic differential equations (SDEs), including diffusion based generative AI, our aim is to accurately generate samples from a distribution. Typically, this is done by generating i.i.d. samples of the SDE. Given a fixed simulation budget, a reasonable way to gain efficiency may be to instead generate a tree of paths through appropriately split partial paths. This suggests improved performance, but one worries about the injected dependence. In this paper, we study this issue comprehensively. We develop a general method to construct the splitting design and analyze its performance for a broad class of error measures, giving insight into the elegant underlying structure in the problem. For the Kolmogorov-Smirnov (KS) distance and maximum mean discrepancy (MMD), we identify the limiting errors of the associated empirical distributions as the simulation budget increases to infinity. We characterize a splitting strategy motivated by a corresponding asymptotic optimization problem. Overall, we observe a 12-26% improvement in mean KS over i.i.d. samples in many settings. For CIFAR-10 and FFHQ experiments, our method reduces the MMD by 9-15%.

1 INTRODUCTION

Many applications require samples from a terminal distribution of a stochastic differential equation (SDE). In diffusion-based generative AI, e.g., in DDPMs (Ho et al., 2020) and in EDMs (Karras et al., 2022), a trained model is used to run an SDE, and samples from its terminal distribution are of interest. These samples are often used to estimate quantities of interest to make decisions. For example, GenCast is a diffusion model that generates samples of weather forecasts, which are used to estimate performance measures such as the likelihood of extreme weather and uncertainty in wind power (Price et al., 2025). SDEs are also standard for modeling the evolution of financial quantities such as asset prices and interest rates. Here, the underlying SDE parameters are typically learned from data and one may be interested in ascertaining the probability distribution at the terminal time. Recently, diffusion models have also been used to generate financial time series (Huang et al., 2024).

In practice, one may be interested in a variety of functionals of the SDE terminal distribution, such as expectations of functions of the terminal output or tail risk measures. However, in this paper, we focus on the fundamental question of accurately estimating the underlying distribution given a fixed simulation budget. This is often done using i.i.d. sampling. We ask whether a well-optimized branching process that splits paths at intermediate times can estimate the terminal distribution better than i.i.d. sampling with the same computational effort.

Our theoretical results reveal the elegant dependence structure, which is completely captured by the covariance of the SDE across time segments. This leads to an allocation problem where the amount of splitting in the branching process is chosen to minimize the limiting error. We solve a relaxed optimization problem, and approximate the solution with a finite mixture of trees with a provable approximation guarantee. This is implemented using a two-phase procedure that first learns the variance contributions and then spends the remaining budget according to the learned allocation. The limit theory applies to any error that is a continuous function of the empirical process under the uniform norm over a class of test functions on which the process satisfies a functional central limit theorem. In this paper, we develop it for the Kolmogorov-Smirnov (KS) distance and the maximum mean discrepancy (MMD). Experiments spanning SDEs and diffusion models show consistent improvements over independent full-path simulation.

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 paper is structured as follows. Section 2 reviews the closest work. Section 3 develops theory for a finite-budget and fixed-threshold setting and identifies the optimal allocation in that setting. Section 4 develops limit theorems by modeling the empirical process. Section 5 studies the Kolmogorov-Smirnov error and the corresponding relaxed allocation problem. Section 6 converts the relaxed allocation into a mixture of exact dyadic trees. Finally, Section 7 presents the experiments.

2 Related Work

The idea of splitting to reduce Monte Carlo error is not new. In classical multilevel splitting, paths are copied when they reach intermediate levels leading to a rare event. Glasserman et al. (1999) study this through branching-process arguments and show how the number of copies affects both variance and computational cost. Their analysis suggests keeping the expected particle population stable across levels. However, this line of work is mainly concerned with estimating a rare-event probability.

Related allocation questions also appear outside rare-event simulation. Melas (1993) studies branching estimators for functionals of Markov chains and allows the branching rate to depend on the current state. More simulation effort is assigned where the future contribution to the estimator is more uncertain, while total cost is controlled. Splitting has also been combined with Markov chain Monte Carlo (MCMC) methods. Botev & Kroese (2012) apply splitting to static problems, such as estimating normalizing constants. Cérou & Guyader (2016) study an adaptive version that, at each stage, keeps only the particles with the highest scores, resamples from them, and moves the copies by MCMC. However, both papers target a single rare-event probability or integral for a static random vector, while we estimate the terminal CDF of a process that evolves in time.

Distributional estimation is considered more directly by Anderson & Ehlert (2022). They estimate the terminal probability mass function of a stochastic reaction network by simulating paths to one intermediate time and then generating several independent continuations from each partial path. They choose the split time and the number of continuations by balancing distributional error against simulation cost. However, their construction uses a discrete state space and a single split.

For discretized SDEs, Giles & Haji-Ali (2024) use repeated path branching for a digital payoff within multilevel Monte Carlo. Their analysis studies the dependence created by shared histories and gives rules for placing repeated binary split times near maturity. The target is one payoff correction, and the main design variable is the timing of the splits rather than a general allocation of particle counts.

3 Discretization and Splitting Estimator

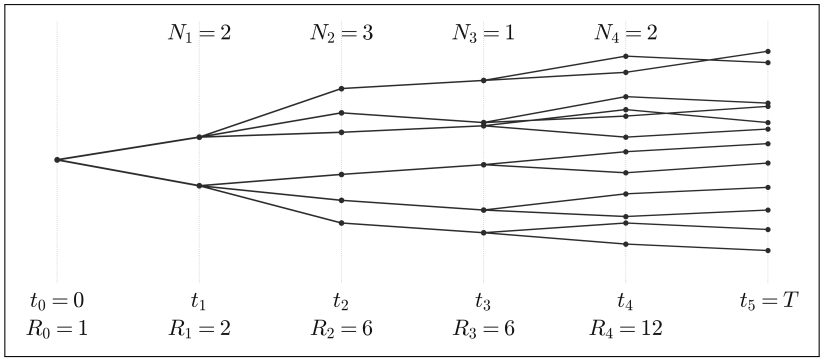

Figure 1: Illustration of splitting - at time $t_{i,B}$, each particle is split into $N_{i,B}$ copies.

Let $X = (X_t)_{0 \le t \le T}$ be a solution to

$$dX_t = b(t, X_t)\, dt + \sigma(t, X_t)\, dW_t, \qquad X_0 \sim \mu_0, \qquad X_0 \perp W.$$

We aim to estimate its terminal distribution function $F(x) = \mathbb{P}(X_T \le x)$ for all $x \in \mathbb{R}^d$. For each budget B, choose a time partition and the corresponding partition size

$$0 = t_{0,B} < t_{1,B} < \cdots < t_{K_B,B} = T, \qquad h_B = \max_{0 \le i < K_B} (t_{i+1,B} - t_{i,B}).$$

We assume throughout that $h_B \to 0$ as $B \to \infty$. Let $X_0^B, X_1^B, \ldots, X_{K_B}^B$ be a Markov-chain discretization of the SDE along this partition, with $X_0^B \sim \mu_0$. Its terminal distribution function is $F_B(x) = \mathbb{P}\left(X_{K_B}^B \le x\right)$. There are K_B transitions, where transition i propagates particles from $t_{i,B}$ to $t_{i+1,B}$. At $t_{0,B}$, there are $N_{0,B}$ independent root particles. Split i occurs after particles arrive at $t_{i,B}$ and before transition i. Each arriving particle is copied $N_{i,B} \ge 1$ times, and the copies are propagated independently through transition i. An illustration is shown in Figure 1. Define

$$R_{0,B} = 1, \qquad R_{i,B} = \prod_{j=1}^{i} N_{j,B}, \quad 1 \le i \le K_B - 1.$$

At the terminal time, there are $R_{K_B-1,B}$ descendants per root. If $Z_{B,j}^{(\ell)}$ denotes terminal descendant ℓ from root j, the splitting estimator is

$$\widehat{F}_B(x) = \frac{1}{N_{0,B} R_{K_B-1,B}} \sum_{j=1}^{N_{0,B}} \sum_{\ell=1}^{R_{K_B-1,B}} \mathbf{1}\left(Z_{B,j}^{(\ell)} \le x\right), \qquad x \in \mathbb{R}^d.$$

This estimator is unbiased for $F_B(x)$. For simplicity, we assume that each transition has the same unit cost. However, this can be relaxed, and our results extend to settings where each transition has a different cost; see A.17. In the unit cost setting, a fully used budget satisfies $B = N_{0,B} \sum_{i=0}^{K_B-1} R_{i,B}$.

To analyse the variance of this estimator, observe that particles that share a root remain dependent due to their common history. Once they split, their transitions are conditionally independent. The next result shows how this structure can be understood by the covariance of the empirical CDF. For $i = 0, \ldots, K_B$, define $u_{i,x}^B$ to be the conditional expectation of $\mathbf{1}(X_{K_B}^B \le x)$ as

$$u_{i,x}^B(z) = \mathbb{E}[\mathbf{1}(X_{K_B}^B \le x) \mid X_i^B = z].$$

Lemma 1 (Splitting covariance)**.** For every $x, y \in \mathbb{R}^d$,

$$\mathrm{Cov}(\widehat{F}_B(x), \widehat{F}_B(y)) = \frac{1}{N_{0,B}} \left[\Gamma_{0,B}(x,y) + \sum_{i=0}^{K_B-1} \frac{v_{i,B}(x,y)}{R_{i,B}}\right],$$

where

$$\Gamma_{0,B}(x,y) = \mathrm{Cov}\left(u_{0,x}^B(X_0^B), u_{0,y}^B(X_0^B)\right),$$
$$v_{i,B}(x,y) = \mathbb{E}\left[\mathrm{Cov}\left(u_{i+1,x}^B(X_{i+1}^B), u_{i+1,y}^B(X_{i+1}^B) \mid X_i^B\right)\right].$$

The proof is given in A.2. This separates variance of the estimator into the variance of each step of the process. Here, $v_{i,B}$ is the covariance while crossing transition i, and it is divided by the population propagated through that transition. Increasing the population before one transition therefore reduces variance created there and later. To get insight into the optimal allocation, we first consider the simpler problem of minimizing the variance at one threshold x for a fixed budget B. Write

$$\widetilde{v}_k = v_{k,B}(x,x) + \mathbf{1}_{\{k=0\}}\Gamma_{0,B}(x,x), \qquad n_k = N_{0,B} R_{k,B}.$$

Here, n_k is the number of particles propagated through transition k. We consider a relaxation where the n_k can be positive reals. The fixed-budget variance minimization problem is then

$$\min_{n \in \mathbb{R}_{>0}^{K_B}} \sum_{k=0}^{K_B-1} \frac{\widetilde{v}_k}{n_k} \quad \text{s.t.} \quad \sum_{k=0}^{K_B-1} n_k = B, \qquad n_0 \le n_1 \le \cdots \le n_{K_B-1}.$$

The monotonicity constraint is handled by using Pool-Adjacent-Violators Algorithm (PAVA) (Robertson et al., 1988, Chapter 1). See A.1 for a detailed discussion and construction of PAVA.

Lemma 2 (Variance Optimization with Monotone Allocation)**.** Fix $x \in \mathbb{R}^d$, and assume $\widetilde{v}_k > 0$ for all k. The unique minimizer is

$$n_k^* = B \frac{\sqrt{\widehat{v}_k}}{\sum_{j=0}^{K_B-1} \sqrt{\widehat{v}_j}}, \qquad \widehat{v} = \mathrm{PAVA}(\widetilde{v}).$$

The proof is given in A.3. The key idea here is that n_k follow a Neyman allocation with respect to the variances, after accounting for the monotonicity constraint. This solves the problem for one threshold, whereas a terminal-law estimator has to account for all thresholds, and a single allocation must control the entire CDF together. We develop theory for this in the next section.

4 ASYMPTOTIC ANALYSIS

In this section, we analyse the asymptotic error distribution of the splitting estimator. Our first goal is to find the continuous-time counterpart of the covariance formula in Lemma 1, which under suitable assumptions, is the limit of v_i. For $x \in \mathbb{R}^d$, write $a_x(z) = \mathbf{1}(z \le x)$ and the cumulative splitting function $r_B(t) = R_{i,B}$ for $t \in [t_{i,B}, t_{i+1,B})$. Let the natural filtration of the SDE be denoted as

$$\mathscr{F}_t = \sigma(X_0) \vee \sigma(W_s : 0 \le s \le t), \qquad 0 \le t \le T.$$

Assumption 1 (Strong terminal convergence)**.** The chain X^B is driven by X_0 and W, X^B and X are Markov with respect to $(\mathscr{F}_{t_{i,B}})_i$ and $(\mathscr{F}_t)_t$, and $\|X^B_{K_B} - X_T\|_{L^2} \to 0$.

Assumption 1 ensures consistency of the discretization scheme and assumes the Markov property for X. It is satisfied by the Euler-Maruyama scheme under global Lipschitz and linear-growth conditions on the drift and diffusion coefficients; see Kloeden & Platen (1992, Theorem 10.2.2).

Assumption 2 (Splitting-function limit)**.** The splitting functions satisfy $r_B(0+) = 1$, and each r_B is nondecreasing. Moreover, $r_B \to r$ in $L^1([0,T])$ with $r(0+) = 1$.

Since the splitting function is a design choice, many irregular and feasible splitting functions can be constructed. Assumption 2 is a regularity condition which restricts the class of functions permitted.

Assumption 3 (Continuous terminal CDF)**.** The terminal CDF $F(x) = \mathbb{P}(X_T \le x)$ is continuous.

We begin by identifying the continuous-time limit of the covariance of the splitting estimator. As we saw in the discrete case, $u^B_{i,x}$ was central to our arguments, and the continuous-time analog of it is

$$u_x(t,z) = \mathbb{E}\left[a_x(X_T) \mid X_t = z\right], \qquad u_x(t, X_t) = \mathbb{E}\left[a_x(X_T) \mid \mathscr{F}_t\right].$$

Since $a_x(X_T)$ is bounded, $u_x(t, X_t)$ is a bounded square-integrable martingale. Applying martingale representation conditionally on X_0, there is a predictable process H_x such that

$$u_x(t, X_t) = u_x(0, X_0) + \int_0^t H_x(\tau)^\top dW_\tau, \qquad 0 \le t \le T.$$

The following lemma shows that the entire covariance structure of the SDE is captured by the predictable process H_x and identifies the continuous-time analog of the one-step covariances $v_{i,B}$.

Lemma 3. For $x, y \in \mathbb{R}^d$, define, for almost every $t \in [0,T]$,

$$g_{x,y}(t) := \mathbb{E}\left[H_x(t)^\top H_y(t)\right].$$

Fix B and $0 \le i \le K_B - 1$, and set $s = t_{i,B}$ and $t = t_{i+1,B}$. Then the true SDE's covariance contribution over the interval $[s,t]$ is equal to the integral of $g_{x,y}$ over that interval, i.e.,

$$v^{\mathrm{SDE}}_{i,B}(x,y) := \mathbb{E}\left[\mathrm{Cov}(u_x(t,X_t), u_y(t,X_t) \mid \mathscr{F}_s)\right] = \int_s^t g_{x,y}(\tau)\, d\tau.$$

Moreover, $g_{x,y} \in L^1([0,T])$.

Proof is given in A.5. Here, $g_{x,y}(t)$ is interpreted as the covariance density at time t for the SDE, and represents the rate at which covariance is being changed at time t. As in the discrete case, let

$$\Gamma_0(x,y) := \mathrm{Cov}(u_x(0,X_0), u_y(0,X_0))$$

be the initial variance. The martingale representation separates the two sources of randomness, giving

$$\mathrm{Cov}(a_x(X_T), a_y(X_T)) = \Gamma_0(x,y) + \int_0^T g_{x,y}(t)\, dt.$$

Remark 1. If $u_x \in C^{1,2}$, then by Itô's formula, we have $H_x(t) = \sigma(t, X_t)^\top \nabla u_x(t, X_t)$.

Having identified the covariance density, we now show that the discrete one-step covariances converge to it. Using the splitting function defined above, define the scaled covariance kernel

$$\Gamma_B(x,y) := N_{0,B}\, \mathrm{Cov}(\widehat{F}_B(x), \widehat{F}_B(y)) = \Gamma_{0,B}(x,y) + \sum\nolimits_{i=0}^{K_B-1} \frac{v_{i,B}(x,y)}{R_{i,B}}.$$

Proposition 1. Suppose Assumptions 1 and 3 hold. Then, for every $x, y \in \mathbb{R}^d$,

$$\Gamma_{0,B}(x,y) \to \Gamma_0(x,y), \quad \text{and} \quad \sum_{i=0}^{K_B-1} \left| v_{i,B}(x,y) - v_{i,B}^{\mathrm{SDE}}(x,y) \right| \to 0.$$

The proof is given in A.6. This shows that each $v_{i,B}$ and the initial variance converge to their continuous-time analogs. However, this does not establish a covariance limit for the splitting estimator.

Proposition 2 (Covariance limit)**.** Suppose Assumptions 1, 2, and 3 hold. Then, for every $x, y \in \mathbb{R}^d$,

$$\Gamma_B(x,y) \to \Gamma(x,y) = \Gamma_0(x,y) + \int_0^T \frac{g_{x,y}(t)}{r(t)}\, dt.$$

Moreover, Γ is positive semidefinite and

$$d_\Gamma(x,y)^2 := \Gamma(x,x) + \Gamma(y,y) - 2\Gamma(x,y) \le \|a_x(X_T) - a_y(X_T)\|_{L^2}^2 .$$

The proof is given in A.7. This is the continuous-time analog of Lemma 1 and shows that the covariance of the splitting estimator converges to a natural limit.

We now show that the splitting estimator obeys a central limit theorem. In particular, we have a triangular-array empirical process, where the entire simulation depends on B. Let $\mathcal{A} = \{a_x : x \in \mathbb{R}^d\}$ be the lower-orthant class. Define $\rho(a_x, a_y) = \|a_x(X_T) - a_y(X_T)\|_{L^2}$ for $a_x, a_y \in \mathcal{A}$. For notational convenience, let $m_B = R_{K_B-1,B}$ and let

$$E_B = (\mathbb{R}^d)^{m_B}, \qquad Z_B = \left(Z_B^{(1)}, \ldots, Z_B^{(m_B)} \right) \in E_B,$$

where, E_B is the sample space of the terminal descendants from one root, and Z_B is the corresponding random vector. For $z \in E_B$, define

$$f_{B,x}(z) = \frac{1}{m_B} \sum_{\ell=1}^{m_B} a_x(z^{(\ell)}), \qquad \mathcal{F}_B = \left\{ f_{B,x} : x \in \mathbb{R}^d \right\}.$$

For each fixed B, the empirical process is naturally computed on $\mathcal{F}_B$ over the sample space E_B. Although E_B and $\mathcal{F}_B$ change with B, we label the coordinate corresponding to $f_{B,x}$ by a_x and regard the process as indexed by the fixed class $\mathcal{A}$. We are now ready to state the central limit theorem.

Theorem 1 (Splitting CLT)**.** Under Assumptions 1, 2, and 3, for $a_x \in \mathcal{A}$, define

$$\mathbb{G}_B(a_x) = \sqrt{N_{0,B}} \left(\widehat{F}_B(x) - F_B(x) \right).$$

If $N_{0,B} \to \infty$, then

$$\mathbb{G}_B \Rightarrow \mathbb{G} \quad \text{in } \ell^\infty(\mathcal{A}), \qquad \mathrm{Cov}(\mathbb{G}(a_x), \mathbb{G}(a_y)) = \Gamma(x,y).$$

Here, $\mathbb{G}$ is a centered Gaussian process indexed by $\mathcal{A}$ with covariance Γ defined in Proposition 2.

The proof is given in A.9. This shows us that the covariance of the Gaussian process is dependent on the splitting function r, and hence it completely specifies the limiting distribution.

The CLT above describes only the simulation noise around F_B. To obtain the error around F, we need to account for the bias due to discretization. We follow the approach of Duffie & Glynn (1995) to balance the bias and variance, which are modeled as functions of the budget B. One tree costs $C_B = B/N_{0,B} = \sum_{i=0}^{K_B-1} R_{i,B}$. On an asymptotically uniform time partition, Assumption 2 gives

$$h_B C_B \to \int_0^T r(t)\, dt =: \gamma,$$

where γ is the limiting scaled cost of one tree. Hence $N_{0,B} \sim B h_B/\gamma$; see A.10 for the details. For the discretization error $F_B - F$, we assume the following bound to provide a rate for the total error.

Assumption 4 (Weak discretization error)**.** There is $p > 0$ such that $\|F_B - F\|_\infty = O(h_B^p)$.

For the Euler scheme, Bally & Talay (1996) prove weak error of order one for a broad class of test functions, including indicator functions, under regularity conditions, motivating the case $p = 1$.

Corollary 1. Suppose the assumptions of Theorem 1 and Assumption 4 hold. Assume that the time partition is asymptotically uniform. Put $\lambda_B = h_B B^{1/(1+2p)}$. If $\lambda_B \to \lambda \in (0, \infty)$, then

$$B^{p/(1+2p)}(\widehat{F}_B - F_B) \Rightarrow (\gamma/\lambda)^{1/2}\,\mathbb{G} \qquad \text{in } \ell^\infty(\mathcal{A}).$$

Moreover,

$$B^{p/(1+2p)}\|\widehat{F}_B - F\|_\infty$$

is bounded in probability.

The proof is given in A.10. Assumption 4 controls only the magnitude of $F_B - F$, and only gives a rate for the total error rather than a weak limit. For the remainder of the paper, we consider the non-degenerate regime $\lambda_B \to \lambda \in (0, \infty)$, which happens when $h_B \propto B^{-1/(1+2p)}$.

Remark 2. We retain the unit-cost and asymptotically uniform partition requirements here for simplicity; a nonuniform discretization is handled by assigning different costs to different transitions, and A.17 shows that the corollary then holds with γ replaced by a cost-weighted γ_ω.

5 KS Statistic and Minimax Allocation

The Kolmogorov-Smirnov (KS) distance from the target law is

$$\mathrm{KS}(\widehat{F}_B, F) = \sup_{x\in\mathbb{R}^d} |\widehat{F}_B(x) - F(x)|.$$

Since $F_B - F$ is independent of the splitting function r, the allocation can affect only $\widehat{F}_B - F_B$. We therefore use the limiting KS loss of the centered simulation error as the allocation criterion. By the continuous mapping theorem,

$$B^{p/(1+2p)}\,\mathrm{KS}(\widehat{F}_B, F_B) \Rightarrow (\gamma/\lambda)^{1/2}\sup_{x\in\mathbb{R}^d}|\mathbb{G}(a_x)|.$$

We are interested in minimizing this quantity. However, it is generally intractable. Fortunately, we can upper and lower bound its expected value by bounds that differ up to a logarithmic factor.

Proposition 3 (Expected Limiting Centered KS Bounds)**.** Let $\Gamma^* = \sup_{x\in\mathbb{R}^d}\sqrt{\Gamma(x,x)}$. If $\Gamma^* > 0$, then there exists a constant $C < \infty$, depending only on d, such that

$$\sqrt{2/\pi}\,(\gamma/\lambda)^{1/2}\,\Gamma^* \le (\gamma/\lambda)^{1/2}\,\mathbb{E}\left[\sup_{x\in\mathbb{R}^d}|\mathbb{G}(a_x)|\right] \le C\,(\gamma/\lambda)^{1/2}\,\Gamma^*\sqrt{1-\log(\Gamma^*)}.$$

The proof is given in A.11. This proposition shows that for a fixed λ, the bias term does not depend on the splitting design, and we can limit our attention to the variance term. Notice that the bounds agree up to a logarithmic factor. Dropping this factor gives the surrogate $\gamma(r)^{1/2}\Gamma^*$. Since the surrogate is nonnegative, minimizing its square gives the proxy objective

$$\inf_{r\in\mathcal{R}}\gamma(r)\sup_{x\in\mathbb{R}^d}\left[\Gamma_0(x,x) + \int_0^T \frac{g_{x,x}(t)}{r(t)}\,dt\right],$$

where $\mathcal{R}$ is the set of nondecreasing functions $r \in L^1([0,T])$ with $r(0+) = 1$. This problem has an elegant interpretation: we are minimizing the worst-case variance of the Gaussian process over all thresholds x by choosing the splitting function r. To solve this problem, let $\mathcal{X} = [-\infty, \infty]^d$ have the product topology. Since $\mathbb{R}^d$ is dense in $\mathcal{X}$, compactification does not change the supremum.

Inspired by Lemma 2, we first describe the continuous-time analog of PAVA. For any integrable function $f : [0,T] \to \mathbb{R}$, let $C[f]$ be the greatest convex minorant of $t \mapsto \int_0^t f(s)\,ds$ on $[0,T]$. Then, the continuous-time PAVA of f is defined as the right derivative of $C[f]$, that is $\mathrm{PAVA}(f) = C[f]'_+$.

Now, note that the threshold that maximizes the variance depends on the allocation, and hence they need to be controlled jointly. This can be thought of as a game between the designer and an adversary, where the adversary chooses weights on the thresholds to maximize the variance. We therefore introduce mixtures over thresholds. For $\pi \in \mathcal{P}(\mathcal{X})$, define

$$\Gamma_0^\pi = \int_{\mathcal{X}} \Gamma_0(x,x)\,\pi(dx), \qquad g^\pi = \int_{\mathcal{X}} g_{x,x}\,\pi(dx).$$

These integrals are well defined by Lemma 6. The PAVA construction above can be extended to include an atom Γ_0^π at $t = 0$, as done in A.1. Applying this to the pair (Γ_0^π, g^π), we get $h_\pi = \mathrm{PAVA}(\Gamma_0^\pi, g^\pi)$.

We solve this problem by replacing the supremum over x with a maximization over probability distributions π on $\mathcal{X}$, which leaves its value unchanged. For $r \in \mathcal{R}$, set $n = r/\gamma(r)$. The objective is convex in n and affine in π. Under a suitable topology, $\mathcal{P}(\mathcal{X})$ is compact. We then use Sion's minimax theorem and the PAVA construction above to conclude. A.13 gives the details.

Theorem 2 (Minimax Allocation)**.** Suppose Assumption 3 holds. Then

$$\inf_{r\in\mathcal{R}} \gamma(r) \sup_{x\in\mathbb{R}^d} \left[\Gamma_0(x,x) + \int_0^T \frac{g_{x,x}(t)}{r(t)}\,dt\right] = \max_{\pi\in\mathcal{P}(\mathcal{X})} \left(\int_0^T \sqrt{h_\pi(t)}\,dt\right)^2.$$

The maximum on the right is attained. For any maximizer π^*, the infimum on the left is attained if and only if $h_{\pi^*}(0+) > 0$. In that case, $r^*(t) = \sqrt{h_{\pi^*}(t)/h_{\pi^*}(0+)}$ defines a minimizer.

The proof is given in A.13. Notably, the optimal allocation still follows Neyman allocation, but now with respect to the PAVA projection of a mixture of covariance densities.

Since simulation is performed with a finite budget, we also provide a finite-budget counterpart to Theorem 2, which uses the same argument for the minimax version of Lemma 2. Put $m = K_B$, and let $\widetilde{v}_i(x)$ and n_i be as in Lemma 2, now with the dependence on x. Define the feasible allocation set $\mathcal{N}$ and, to account for the coupling across thresholds, the closed convex hull $\mathcal{V}$ of the variances:

$$\mathcal{N} = \left\{n \in \mathbb{R}^m_{>0} \text{ nondecreasing} : \sum\nolimits_i n_i = B\right\}, \quad \mathcal{V} = \overline{\mathrm{conv}}\left\{\left(\widetilde{v}_i(x)\right)_{i<m} : x \in \mathbb{R}^d\right\} \subset \mathbb{R}^m_+.$$

Corollary 2 (Finite-Budget Minimax Allocation)**.** Assume that there exists $x_0 \in \mathbb{R}^d$ such that $\widetilde{v}_i(x_0) > 0$ for every i. For $a \in \mathcal{V}$, let $\widehat{a} = \mathrm{PAVA}(a)$. Then the minimum below is attained, and

$$\min\nolimits_{n\in\mathcal{N}} \sup\nolimits_{x\in\mathbb{R}^d} \sum\nolimits_{i=0}^{m-1} \frac{\widetilde{v}_i(x)}{n_i} = \frac{1}{B} \max\nolimits_{a\in\mathcal{V}} \left(\sum\nolimits_{i=0}^{m-1} \sqrt{\widehat{a}_i}\right)^2.$$

Moreover, if a^* maximizes the right-hand side and $h^* = \mathrm{PAVA}(a^*)$, then $h_i^* > 0$ for every i, and every minimizing allocation satisfies $n_i^* \propto \sqrt{h_i^*}$. The proof is given in A.14.

6 Rounding

For a fixed B, let n^* be an optimal allocation in Corollary 2. This need not define a valid tree because it can have noninteger entries. Hence, we approximate it with a mixture of trees. Set $m = K_B$ and let $r_i^* = n_i^*/n_0^*$ be the normalized allocation. Define the relaxed normalized variance

$$V^* = \left(\sum\nolimits_{i=0}^{m-1} r_i^*\right) \sup\nolimits_{x\in\mathbb{R}^d} \sum\nolimits_{i=0}^{m-1} \frac{\widetilde{v}_i(x)}{r_i^*}.$$

Given r^*, we first construct a finite set of dyadic trees following the construction of Roundy (1989, Section 2). Let $U \in [0, 1)$. For each value of U, associate a tree with profile

$$R_i(U) = 2^{\lfloor \log_2 r_i^* + U \rfloor}, \qquad 0 \le i < m.$$

Since r^* is nondecreasing, the exponents are also nondecreasing, so every ratio $R_i(U)/R_{i-1}(U)$ is a nonnegative power of two and every $R(U)$ defines a valid tree. As U varies over $[0, 1)$, each coordinate can change at most once. Hence, at most m distinct trees occur, which are our candidates.

We now construct a mixture of these trees. Let $R^{(1)}, \ldots, R^{(S)}$ be the distinct profiles, where $S \le m$. Suppose we generate M_s independent trees of type s to construct the empirical CDF $\widehat{F}_{B,s}$, and assign weights λ_s to each type with $\sum_s \lambda_s = 1$. We define the mixture CDF to be

$$F_B^\lambda = \sum\nolimits_{s=1}^{S} \lambda_s \widehat{F}_{B,s}.$$

Since each type is unbiased, the mixture is unbiased. From Lemma 1, we have for every $x, y \in \mathbb{R}^d$,

$$\mathrm{Cov}\left(F_B^\lambda(x), F_B^\lambda(y)\right) = \sum\nolimits_{s=1}^{S} \frac{\lambda_s^2}{M_s} \left[\Gamma_{0,B}(x,y) + \sum\nolimits_{i=0}^{m-1} \frac{v_{i,B}(x,y)}{R_i^{(s)}}\right].$$

Here, λ_s is interpreted as the budget share assigned to type s. If C_s is the cost of generating one tree of type s, we have $M_s = \lambda_s B / C_s$. We choose these weights to minimize the worst-case variance and show that the resulting mixture approximates the relaxed optimum up to a small constant factor.

Proposition 4 (Dyadic Mixture Approximation)**.** If the candidate types are generated from r^*, then

$$V^* \le \min_{\substack{\lambda_s \ge 0 \\ \sum_s \lambda_s = 1}} \sup_{x \in \mathbb{R}^d} \sum_{s=1}^{S} \lambda_s C_s \sum_{i=0}^{m-1} \frac{\widetilde{v}_i(x)}{R_i^{(s)}} \le \frac{2}{e \log 2} V^*.$$

The proof is given in A.15. We also show that the asymptotic analysis extends to this mixture in A.16.

7 Numerical Experiments

The optimal allocation is determined by the variance contributions $v_i(x, x)$ and $\Gamma_{0,B}(x, x)$. In practice they are not known. Hence, we propose the following two-phase procedure.

Two-Phase Algorithm. We choose a small set of split times $0 = t_0 < t_1 < \cdots < t_{L-1} < t_L = T$. We only allow splitting at these times and particles follow the original dynamics inside a segment. Phase 1 uses a budget $B_1 \propto B^\alpha$ with $\alpha < 1$ to simulate paths without splitting, and uses them to estimate the variance contributions of each segment. We then solve the allocation problem with these estimates and construct the corresponding mixture of trees. Phase 2 uses the remaining budget to run the selected trees. The final estimate combines the Phase 1 terminal samples with the weighted empirical distribution of the Phase 2 trees. For KS, we fit small MLPs to the Phase 1 paths to estimate $u_x(t, X_t) = \mathbb{P}(X_T \le x \mid X_t)$, and replace the supremum over thresholds in Corollary 2 by a maximum over a finite set. See B.1 for the procedure for KS and B.2 for the procedure for MMD.

Experimental Setup. We study five examples: (i) a two-dimensional Ornstein-Uhlenbeck (OU) process, (ii) two-dimensional overdamped Langevin dynamics, (iii) a stochastic EDM model trained on a two-dimensional Gaussian mixture, (iv) the pretrained `google/ddpm-cifar10-32` model, and (v) the pretrained `asparius/ldm-ffhq-256` latent diffusion model for FFHQ. For every example, we consider schedules with 9, 19 and 39 split times. See D for the complete setup.

At each budget, we compare learned splitting with independent full-path simulation and the Uniform-c and Learned-c baselines, where we split by a factor of c each time. Refer to D.3 for both baseline procedures. Due to compute constraints, we only test CIFAR-10 and FFHQ on Learned-c. For the two-dimensional examples, we compute the exact KS distance of the empirical CDF to a fixed reference CDF. Computing the KS distance exactly is not practical for images, so for CIFAR-10 and FFHQ we use the MMD with a random Fourier feature kernel which is also a continuous function of the empirical process under the uniform norm, now over a finite class of features. Appendix A.18 develops the theory for MMD and its allocation problem.

Results. Figure 2 shows the percentage reduction in the mean error metric relative to i.i.d. simulation, budget by budget. Figure 3 shows the learned cumulative allocation profiles at the largest tested budget for each model. Refer to F for complete numerical results.

Splitting lowers mean error at every tested budget and schedule across all five examples. Reductions reach 26% on the two-dimensional examples, 15% on CIFAR-10, and 12% on FFHQ. The baseline where c is chosen before running performs well. However, it requires a choice of c that is not known in advance, and some choices of c perform worse than i.i.d. simulation. The baseline where c is learned using a two-phase procedure also beats i.i.d. simulation in every case, and is close to the learned allocation in most cases.

A key observation is that the learned allocation profiles show a non-trivial splitting behavior. For example, for the EDM model, the learned allocation only splits near the midpoint. This agrees with the intuition that most denoising, and hence higher variance, occurs in the middle of the trajectory.

To demonstrate the effectiveness of the two-phase procedure in small dimensions, we compare the learned allocation with an oracle for the OU process, where the variance contributions of the OU process can be computed analytically. Figure 4 shows that the learned allocation closely matches the oracle allocation. At $B = 5 \times 10^6$, we find that our learned allocations are able to achieve performance close to the oracle across the three schedules. See E for the construction and results.

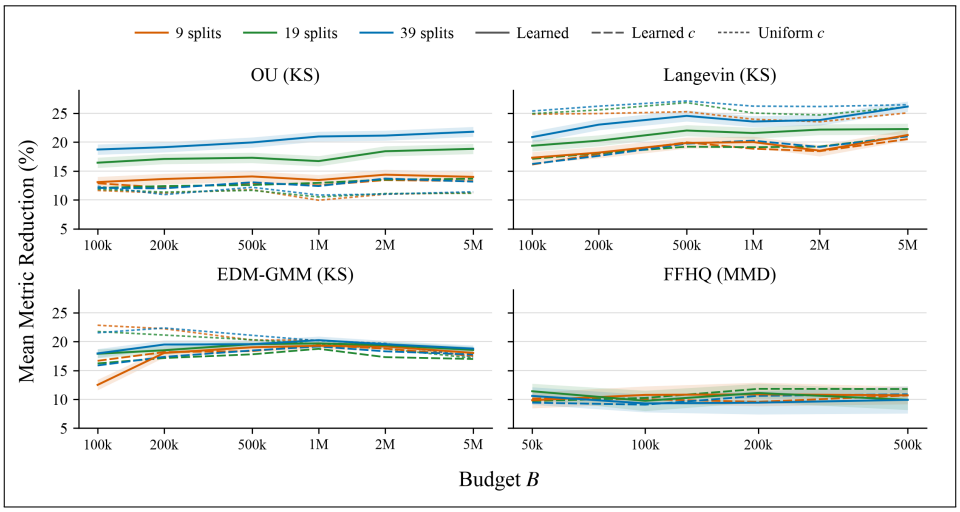

Figure 2: Percentage reduction in the metric relative to i.i.d. simulation for the learned allocation, Uniform-c, and Learned-c.

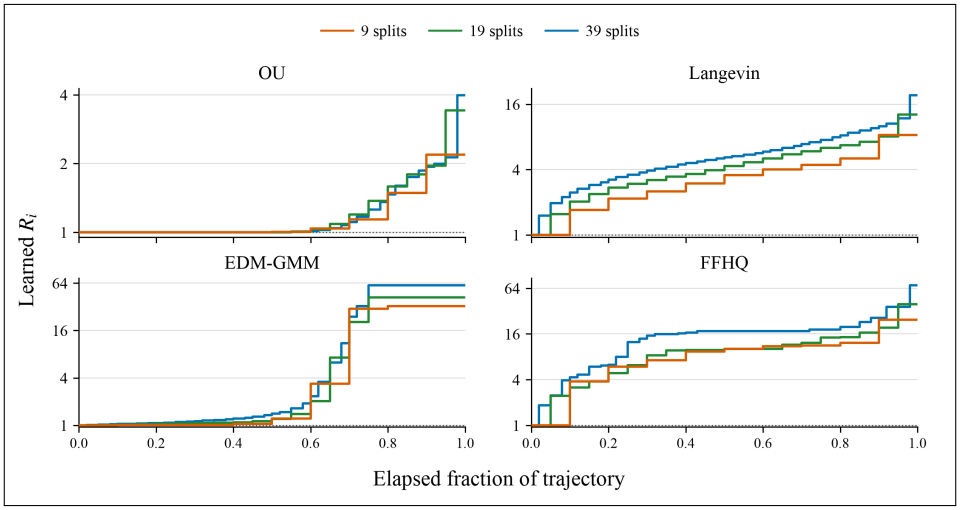

Figure 3: Allocation profiles R_i for the learned allocation at each model's largest tested budget.

8 Conclusion

In this paper, we propose a splitting estimator to estimate the terminal distribution of stochastic differential equations under a fixed simulation budget. We develop a general method that can be applied to a broad class of error measures. Focusing on the Kolmogorov-Smirnov distance and the maximum mean discrepancy, we do analysis to identify the optimal allocation. We also provide a practical two-phase procedure to implement the estimator, which is shown to be effective in extensive numerical experiments.

Our approach has several limitations. For the Kolmogorov-Smirnov distance, the upper and lower bounds differ by a logarithmic factor, and the design criterion ignores this gap. Moreover, some implementation details, such as the finite query set and estimated variance values, remain outside the theory, which does not account for these errors. Another important extension is to consider conditional diffusion models, where the target distribution is conditioned on some input, and the goal is to either (i) learn a splitting allocation that is optimal for all inputs or (ii) learn a mapping from inputs to splitting allocations that is optimal for each input. We leave these extensions to future work.

REPRODUCIBILITY STATEMENT

All theoretical claims are stated with their assumptions in the main text, and complete proofs are given in Appendix A. Appendices B.1 and B.2 specify the two-phase procedures for KS and MMD end to end, including the variance estimator of Phase 1, the optimization problem we solve, the dyadic rounding of Phase 2, and the final algorithm. Appendix D gives the five benchmark models, their simulation settings, and the Uniform-c and Learned-c baselines, and Appendix C defines the reference samples and the KS and MMD metrics used throughout. Appendix E describes the finite-query minimax benchmark against which the learned allocation is compared, and Appendix F reports every measurement behind the figures, with confidence-interval half-widths and repetition counts. Code to reproduce every experiment is available at https://anonymous.4open.science/r/splitting-7D9C/.

AI USE STATEMENT

We used generative AI tools to implement the experiments. Much of the experimental code was written with AI assistance, including the scripts that compute the reported statistics and produce the figures and tables. We did not use these tools for theoretical modeling, mathematical formulation, proofs, hypotheses, methodology, or the interpretation of results. Synthetic data generation, translation, and data cleaning do not apply to this work. We also used AI tools for literature search, but every cited work was found and read by the authors. Finally, we used them to polish the writing (grammar and consistency of notation) and to help with LaTeX formatting. All text was written by the authors. The authors reviewed all AI-assisted code line by line. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

David F. Anderson and Kurt W. Ehlert. Conditional Monte Carlo for reaction networks. *SIAM Journal on Scientific Computing*, 44(2):A993–A1019, 2022. doi: 10.1137/21M144267X.

Vlad Bally and Denis Talay. The law of the euler scheme for stochastic differential equations: I. convergence rate of the distribution function. *Probability Theory and Related Fields*, 104:43–60, 1996.

Zdravko I. Botev and Dirk P. Kroese. Efficient Monte Carlo simulation via the generalized splitting method. *Statistics and Computing*, 22(1):1–16, 2012. doi: 10.1007/s11222-010-9201-4.

Frédéric Cérou and Arnaud Guyader. Fluctuation analysis of adaptive multilevel splitting. *The Annals of Applied Probability*, 26(6):3319–3380, 2016. doi: 10.1214/16-AAP1177.

Darrell Duffie and Peter Glynn. Efficient monte carlo simulation of security prices. *The Annals of Applied Probability*, 5(4):897–905, 1995.

Michael B. Giles and Abdul-Lateef Haji-Ali. Multilevel path branching for digital options. *The Annals of Applied Probability*, 34(5):4836–4862, 2024. doi: 10.1214/24-AAP2083.

Paul Glasserman, Philip Heidelberger, Perwez Shahabuddin, and Tim Zajic. Multilevel splitting for estimating rare event probabilities. *Operations Research*, 47(4):585–600, 1999. doi: 10.1287/opre.47.4.585.

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://www.jmlr.org/papers/v13/gretton12a.html.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Hongbin Huang, Minghua Chen, and Xiao Qiao. Generative learning for financial time series with irregular and scale-invariant patterns. In *International Conference on Learning Representations*, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, 1992.

V. B. Melas. Optimal simulation design by branching technique. In Werner G. Müller, Henry P. Wynn, and Anatoly A. Zhigljavsky (eds.), *Model-Oriented Data Analysis*, pp. 113–126. Physica-Verlag, Heidelberg, 1993. URL https://pure.iiasa.ac.at/id/eprint/3705/.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637:84–90, 2025. doi: 10.1038/s41586-024-08252-9.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

Tim Robertson, Farroll T. Wright, and Richard L. Dykstra. *Order Restricted Statistical Inference*. Wiley, Chichester, 1988. ISBN 9780471917878.

Robin Roundy. Rounding off to powers of two in continuous relaxations of capacitated lot sizing problems. *Management Science*, 35(12):1433–1442, 1989. doi: 10.1287/mnsc.35.12.1433.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

A. W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer International Publishing, Cham, 2023.

Ji Zhao and Deyu Meng. FastMMD: Ensemble of circular discrepancy for efficient two-sample test. *Neural Computation*, 27(6):1345–1372, 2015.

Appendix Contents

A Proofs and Auxiliary Results **12**
A.1 PAVA and convex minorants . . . 12
A.2 Proof of Lemma 1 . . . 15
A.3 Proof of Lemma 2 . . . 15
A.4 Splitting-function convergence . . . 16
A.5 Proof of Lemma 3 . . . 17
A.6 Proof of Proposition 1 . . . 17
A.7 Proof of Proposition 2 . . . 19
A.8 Finite Donsker Entropy . . . 19
A.9 Proof of Theorem 1 . . . 20
A.10 Proof of Corollary 1 . . . 22
A.11 Proof of Proposition 3 . . . 22
A.12 Continuity in the Threshold . . . 23
A.13 Proof of Theorem 2 . . . 24
A.14 Proof of Corollary 2 . . . 26
A.15 Proof of Proposition 4 . . . 26
A.16 Asymptotic Limit of the Implemented Mixture . . . 28
A.17 Cost-weighted extensions . . . 30
A.18 Extension to MMD . . . 32

B Two-Phase Algorithms **34**
B.1 KS . . . 34
B.2 MMD . . . 37

C Metrics and Evaluations **38**

D Benchmark Models and Simulation Settings **40**
D.1 Models . . . 40
D.2 Simulation settings . . . 40
D.3 Uniform and Learned c Baselines . . . 41

E OU Finite-Query Minimax Benchmark **42**
E.1 Oracle allocation . . . 42
E.2 Simulated KS Evaluation . . . 42

F Complete Numerical Results **43**

A Proofs and Auxiliary Results

A.1 PAVA and convex minorants

This subsection defines the discrete and continuous PAVA constructions used in Lemma 2 and Theorem 2. In both settings, PAVA replaces a nonnegative variance profile by a nondecreasing

effective profile. We first describe the discrete construction and then define its continuous-time analogue. For $z = (z_0, \dots, z_{m-1}) \in [0, \infty)^m$, define the equal-weight nondecreasing PAVA fit by

$$\widehat{z} = \mathrm{PAVA}(z) := \arg \min_{\theta_0 \le \cdots \le \theta_{m-1}} \sum_{k=0}^{m-1} (z_k - \theta_k)^2.$$

The fit can be computed by starting with the singleton blocks $\{0\}, \dots, \{m-1\}$. For a consecutive block I, write

$$\overline{z}_I = \frac{1}{|I|} \sum_{k \in I} z_k.$$

Whenever two adjacent blocks I and J satisfy $\overline{z}_I > \overline{z}_J$, merge them and assign the combined block its average. Continue until the block averages are nondecreasing. If $I_1, \dots, I_L$ are the final PAVA blocks, then

$$\widehat{z}_k = \overline{z}_{I_\ell}, \qquad k \in I_\ell, \quad 1 \le \ell \le L.$$

The resulting fit is unique and does not depend on the order in which violations are pooled. This is the standard equal-weight PAVA construction (Robertson et al., 1988, Chapter 1). The same fit is obtained from a convex minorant. Define

$$S_0 = 0, \qquad S_j = \sum_{i=0}^{j-1} z_i, \quad 1 \le j \le m,$$

and let C_z be the greatest convex minorant of the polygonal line through

$$(0, S_0), (1, S_1), \dots, (m, S_m).$$

Then

$$\widehat{z}_i = C_z(i+1) - C_z(i), \qquad 0 \le i < m.$$

Thus the final PAVA blocks are the intervals on which the convex minorant has constant slope. Moreover, $C_z(m) = S_m$, and hence

$$\sum_{i=0}^{m-1} \widehat{z}_i = \sum_{i=0}^{m-1} z_i.$$

To see how this construction enters the discrete allocation problem, fix x and B, and put $m = K_B$. Define

$$\widetilde{v}_k = v_{k,B}(x, x) + \mathbf{1}_{\{k=0\}} \Gamma_{0,B}(x, x), \qquad n_k = N_{0,B} R_{k,B}.$$

Since $R_{0,B} = 1$, both $\Gamma_{0,B}(x, x)$ and $v_{0,B}(x, x)$ are divided by the initial population $n_0 = N_{0,B}$. This is why the initial-state variance is included in the first coordinate of the variance-contribution profile. The variance is

$$\sum_{k=0}^{m-1} \frac{\widetilde{v}_k}{n_k},$$

while the relaxed particle counts must satisfy

$$0 < n_0 \le n_1 \le \cdots \le n_{m-1}.$$

Write

$$\widetilde{v} = (\widetilde{v}_0, \dots, \widetilde{v}_{m-1}).$$

Under the positivity condition in Lemma 2, let

$$\widehat{v} = \mathrm{PAVA}(\widetilde{v}).$$

That lemma shows that the optimal relaxed allocation is

$$n_k^* = B \frac{\sqrt{\widehat{v}_k}}{\sum_{j=0}^{m-1} \sqrt{\widehat{v}_j}}, \qquad 0 \le k < m.$$

Thus PAVA pools adjacent variance contributions that conflict with the monotonicity constraint, after which the allocation follows the usual square-root rule.

We next define the continuous analogue, including an atom at time zero. Let $a_0 \geq 0$, and let $a \in L^1([0,T])$ be nonnegative almost everywhere. Define the cumulative variance profile $A[a_0, a]$ by

$$A[a_0, a](0) = 0, \qquad A[a_0, a](t) = a_0 + \int_0^t a(s)\, ds, \quad 0 < t \leq T.$$

The jump at zero records the initial variance atom. Let $C[a_0, a]$ be the greatest continuous convex function C on $[0, T]$ satisfying

$$C(0) = 0, \qquad C(t) \leq A[a_0, a](t), \quad 0 < t \leq T.$$

We define the continuous PAVA profile by

$$h(t) = \mathrm{PAVA}(a_0, a)(t) := \frac{d^+}{dt} C[a_0, a](t), \qquad 0 \leq t < T.$$

The function h is nonnegative, nondecreasing, right-continuous, and integrable. The greatest convex minorant meets the cumulative profile at the terminal time, so

$$C[a_0, a](T) = A[a_0, a](T) = a_0 + \int_0^T a(t)\, dt.$$

Consequently,

$$C[a_0, a](t) = \int_0^t h(s)\, ds, \qquad \int_0^T h(t)\, dt = a_0 + \int_0^T a(t)\, dt.$$

The gap

$$G(t) = A[a_0, a](t) - C[a_0, a](t)$$

is nonnegative, satisfies $G(T) = 0$, and the convex minorant is affine on every connected component of $\{t : G(t) > 0\}$. Therefore h is constant on every interval over which the continuous construction pools variance. For any bounded, nonnegative, nonincreasing, right-continuous function q, integration by parts gives

$$a_0 q(0+) + \int_0^T a(t) q(t)\, dt = \int_0^T h(t) q(t)\, dt - \int_{(0,T)} G(t)\, dq(t) \geq \int_0^T h(t) q(t)\, dt.$$

This is the continuous counterpart of the PAVA block inequality used in Lemma 2. The discrete and continuous constructions agree on a fixed grid. Let $\Delta = T/m$ and define

$$d_0 = a_0 + \int_0^\Delta a(s)\, ds, \qquad d_i = \int_{i\Delta}^{(i+1)\Delta} a(s)\, ds, \quad 1 \leq i < m.$$

Set

$$D_0 = 0, \qquad D_j = \sum_{i=0}^{j-1} d_i, \quad 1 \leq j \leq m.$$

Then $D_j = A[a_0, a](j\Delta)$ for $j \geq 1$. The slopes of the greatest convex minorant through

$$(0, D_0), (\Delta, D_1), \ldots, (m\Delta, D_m)$$

are

$$\frac{\mathrm{PAVA}(d)_i}{\Delta}, \qquad 0 \leq i < m.$$

Thus the initial atom enters the first grid cell and may be pooled with later intervals whenever monotonicity requires it. In Theorem 2, the inputs are

$$a_0 = \Gamma_0^\pi, \qquad a = g^\pi.$$

The notation used there is therefore

$$h_\pi = \mathrm{PAVA}(\Gamma_0^\pi, g^\pi).$$

The discrete construction in this subsection is equal-weight. The cost-weighted version PAVA_c, in which cumulative cost replaces the index on the horizontal axis and block averages use transition costs as weights, is defined in Appendix A.17.

A.2 Proof of Lemma 1

Proof. Let $\mathscr{F}_{-1} = \{\emptyset, \Omega\}$. For $0 \le i < K_B$, let $\mathscr{F}_i$ be the σ-algebra generated by the particle system immediately before transition i. Thus, $\mathscr{F}_0$ contains the initial root particles. For $i \ge 1$, $\mathscr{F}_i$ includes split i. Let $\mathscr{F}_{K_B}$ contain the terminal particle system. For $x \in \mathbb{R}^d$, define

$$U_i(x) = \mathbb{E}[\widehat{F}_B(x) \mid \mathscr{F}_i], \qquad i = -1, 0, \dots, K_B.$$

Then $U_i(x)$ is a martingale with $U_{K_B}(x) = \widehat{F}_B(x)$ and $U_{-1}(x) = F_B(x)$.

$$\Delta_i^x = U_{i+1}(x) - U_i(x), \qquad i = -1, 0, \dots, K_B - 1.$$

The martingale differences are orthogonal, hence

$$\widehat{F}_B(x) - F_B(x) = \sum_{i=-1}^{K_B-1} \Delta_i^x, \qquad \mathrm{Cov}(\widehat{F}_B(x), \widehat{F}_B(y)) = \sum_{i=-1}^{K_B-1} \mathbb{E}[\Delta_i^x \Delta_i^y].$$

Independence of the root initial states gives

$$\mathbb{E}[\Delta_{-1}^x \Delta_{-1}^y] = \frac{1}{N_{0,B}} \Gamma_{0,B}(x, y)$$

Moreover, we also have

$$\mathbb{E}[\Delta_i^x \Delta_i^y \mid \mathscr{F}_i] = \mathrm{Cov}(U_{i+1}(x), U_{i+1}(y) \mid \mathscr{F}_i), \quad 0 \le i \le K_B - 1.$$

For $0 \le i < K_B$, let I_i be the set of particle labels immediately before transition i. Then $|I_i| = N_{0,B} R_{i,B}$. For $\alpha \in I_i$, let X_i^α be the state of particle α. Conditional on $\mathscr{F}_i$, transition i independently produces X_{i+1}^α from each X_i^α. When $i < K_B - 1$, split $i+1$ repeats each arriving state $N_{i+1,B}$ times. Repeating every state the same number of times does not change the particle average. Therefore,

$$U_{i+1}(x) = \frac{1}{N_{0,B} R_{i,B}} \sum_{\alpha \in I_i} u_{i+1,x}^B \left(X_{i+1}^\alpha\right).$$

Since the X_{i+1}^α are conditionally independent given $\mathscr{F}_i$,

$$\begin{aligned} &\mathrm{Cov}(U_{i+1}(x), U_{i+1}(y) \mid \mathscr{F}_i) \\ &\quad = \frac{1}{(N_{0,B} R_{i,B})^2} \sum_{\alpha \in I_i} \mathrm{Cov}\big(u_{i+1,x}^B(X_{i+1}^\alpha), u_{i+1,y}^B(X_{i+1}^\alpha) \mid \mathscr{F}_i\big). \end{aligned}$$

The covariance in each summand depends on $\mathscr{F}_i$ only through X_i^α. Taking expectations, and using that each X_i^α has the same law, we obtain

$$\mathbb{E}\left[\mathrm{Cov}(U_{i+1}(x), U_{i+1}(y) \mid \mathscr{F}_i)\right] = \frac{v_{i,B}(x, y)}{N_{0,B} R_{i,B}}.$$

Therefore

$$\mathrm{Cov}(\widehat{F}_B(x), \widehat{F}_B(y)) = \frac{1}{N_{0,B}} \left[\Gamma_{0,B}(x, y) + \sum_{i=0}^{K_B-1} \frac{v_{i,B}(x, y)}{R_{i,B}}\right].$$

□

A.3 Proof of Lemma 2

Proof. Put $m = K_B$. Since $\widetilde{v}_k > 0$, the objective

$$n \mapsto \sum_{k=0}^{m-1} \frac{\widetilde{v}_k}{n_k}$$

is strictly convex and the feasible set is convex and hence, if a minimizer exists, it is unique. Let $\widehat{v} = \mathrm{PAVA}(\widetilde{v})$, and let $I_1, \dots, I_L$ be the final PAVA blocks. For a block $I = [a, b]$, $\widehat{v}$ is constant on I, with value

$$\widehat{v}_k = \bar{v}_I = \frac{1}{|I|} \sum_{j \in I} \widetilde{v}_j, \qquad k \in I.$$

By the standard min-max characterization of isotonic regression (Robertson et al., 1988, Chapter 1), if $I = [a, b]$ is a final block, then every initial segment $[a, \ell]$, $\ell < b$, has average at least the average of I. Thus,

$$\sum_{k=a}^{\ell} (\widetilde{v}_k - \widehat{v}_k) \geq 0 \quad \text{for } a \leq \ell < b, \qquad \sum_{k=a}^{b} (\widetilde{v}_k - \widehat{v}_k) = 0.$$

Now let n be any feasible monotone allocation. Since $n_a \leq n_{a+1} \leq \cdots \leq n_b$, we have

$$\frac{1}{n_a} \geq \frac{1}{n_{a+1}} \geq \cdots \geq \frac{1}{n_b}.$$

For a final PAVA block $I = [a, b]$, observe that

$$\begin{aligned}
\sum_{k=a}^{b} \frac{\widetilde{v}_k - \widehat{v}_k}{n_k} &= \sum_{k=a}^{b} (\widetilde{v}_k - \widehat{v}_k) \left(\frac{1}{n_k} - \frac{1}{n_b} \right) \\
&= \sum_{k=a}^{b-1} (\widetilde{v}_k - \widehat{v}_k) \sum_{\ell=k}^{b-1} \left(\frac{1}{n_\ell} - \frac{1}{n_{\ell+1}} \right) \\
&= \sum_{\ell=a}^{b-1} \left(\sum_{k=a}^{\ell} (\widetilde{v}_k - \widehat{v}_k) \right) \left(\frac{1}{n_\ell} - \frac{1}{n_{\ell+1}} \right) \geq 0.
\end{aligned}$$

The last inequality follows because both factors are nonnegative. Summing over all PAVA blocks yields

$$\sum_{k=0}^{m-1} \frac{\widehat{v}_k}{n_k} \leq \sum_{k=0}^{m-1} \frac{\widetilde{v}_k}{n_k}.$$

Thus, for every feasible monotone allocation n, the objective is bounded below by the same problem with $\widetilde{v}$ replaced by its PAVA projection $\widehat{v}$. We now minimize this lower bound and then verify that the resulting allocation makes the bound sharp. By Cauchy-Schwarz,

$$\sum_{k=0}^{m-1} \sqrt{\widehat{v}_k} = \sum_{k=0}^{m-1} \frac{\sqrt{\widehat{v}_k}}{\sqrt{n_k}} \sqrt{n_k} \leq \left(\sum_{k=0}^{m-1} \frac{\widehat{v}_k}{n_k} \right)^{1/2} \left(\sum_{k=0}^{m-1} n_k \right)^{1/2}.$$

Since every feasible allocation satisfies $\sum_k n_k = B$, we have the lower bound

$$\sum_{k=0}^{m-1} \frac{\widetilde{v}_k}{n_k} \geq \frac{\left(\sum_{k=0}^{m-1} \sqrt{\widehat{v}_k} \right)^2}{B}.$$

Equality holds in the Cauchy-Schwarz step if and only if $n_k \propto \sqrt{\widehat{v}_k}$. The constant is determined by the budget constraint. Denote this allocation by n^*. Since $\widehat{v}$ is nondecreasing, n^* is nondecreasing. Equality also holds in the PAVA-block inequality because n^* is constant on each final PAVA block. Therefore n^* attains the lower bound and is optimal. □

A.4 Splitting-function convergence

Lemma 4. Under Assumption 2, r has a nondecreasing representative on $[0, T)$ satisfying $r(0+) = 1$, $1 \leq r < \infty$ a.e., and

$$\frac{1}{r_B} \longrightarrow \frac{1}{r} \qquad \text{in } L^1([0, T]).$$

Proof. Since each r_B is nondecreasing and $r_B(0+) = 1$, we have $r_B \geq 1$. The $L^1([0, T])$-convergence gives a subsequence converging to r almost everywhere, so $r \geq 1$ almost everywhere. Moreover, $r < \infty$ almost everywhere because $r \in L^1([0, T])$. The nondecreasing representative satisfying $r(0+) = 1$ is specified in Assumption 2.

Finally,

$$\left| \frac{1}{r_B} - \frac{1}{r} \right| = \frac{|r_B - r|}{r_B r} \leq |r_B - r| \qquad \text{a.e.},$$

so the claimed L^1-convergence follows. □

A.5 Proof of Lemma 3

Proof. Since $u_x(t, X_t)$ and $u_y(t, X_t)$ are martingales, we write the conditional covariance as

$$\begin{aligned} v_{i,B}^{\mathrm{SDE}}(x,y) &= \mathbb{E}\left[\mathrm{Cov}\big(u_x(t,X_t), u_y(t,X_t) \mid \mathscr{F}_s\big)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(u_x(t,X_t) - u_x(s,X_s)\right)\left(u_y(t,X_t) - u_y(s,X_s)\right) \mid \mathscr{F}_s\right]\right] \\ &= \mathbb{E}\left[\left(u_x(t,X_t) - u_x(s,X_s)\right)\left(u_y(t,X_t) - u_y(s,X_s)\right)\right] \\ &= \mathbb{E}\left[u_x(t,X_t)u_y(t,X_t)\right] - \mathbb{E}\left[u_y(s,X_s)\mathbb{E}\left[u_x(t,X_t) \mid \mathcal{F}_s\right]\right] \\ &\quad - \mathbb{E}\left[u_x(s,X_s)\mathbb{E}\left[u_y(t,X_t) \mid \mathcal{F}_s\right]\right] + \mathbb{E}\left[u_x(s,X_s)u_y(s,X_s)\right] \\ &= \mathbb{E}\left[u_x(t,X_t)u_y(t,X_t) - u_x(s,X_s)u_y(s,X_s)\right]. \end{aligned}$$

Using the martingale representations,

$$du_x(t,X_t) = H_x(t)^\top dW_t, \qquad du_y(t,X_t) = H_y(t)^\top dW_t,$$

we get

$$d\left\langle u_x(\cdot,X_\cdot), u_y(\cdot,X_\cdot)\right\rangle_t = H_x(t)^\top H_y(t)\, dt.$$

Since $u_x(t,X_t)u_y(t,X_t) - \langle u_x(\cdot,X_\cdot), u_y(\cdot,X_\cdot)\rangle_t$ is a martingale, we have

$$v_{i,B}^{\mathrm{SDE}}(x,y) = \mathbb{E}\left[\int_s^t H_x(\tau)^\top H_y(\tau)\, d\tau\right] = \int_s^t g_{x,y}(\tau)\, d\tau.$$

It remains to establish integrability. By Itô's isometry,

$$\begin{aligned} \mathbb{E}\left[\int_0^T \|H_x(\tau)\|_2^2\, d\tau\right] &= \mathbb{E}\left[\left(\int_0^T H_x(\tau)^\top dW_\tau\right)^2\right] \\ &= \mathbb{E}\left[\left(u_x(T,X_T) - u_x(0,X_0)\right)^2\right] \\ &= \mathbb{E}\left[\mathrm{Var}(a_x(X_T) \mid X_0)\right] \\ &\le \mathrm{Var}(a_x(X_T)) \le \frac{1}{4}. \end{aligned}$$

Therefore, by Cauchy-Schwarz,

$$\begin{aligned} \int_0^T |g_{x,y}(\tau)|\, d\tau &\le \mathbb{E}\left[\int_0^T \|H_x(\tau)\|_2 \|H_y(\tau)\|_2\, d\tau\right] \\ &\le \mathbb{E}\left[\int_0^T \|H_x(\tau)\|_2^2\, d\tau\right]^{1/2} \mathbb{E}\left[\int_0^T \|H_y(\tau)\|_2^2\, d\tau\right]^{1/2} \\ &\le \frac{1}{4}. \end{aligned}$$

□

A.6 Proof of Proposition 1

Proof. Since $X_{K_B}^B \to X_T$ in L^2, we also have $X_{K_B}^B \to X_T$ in probability. Continuity of F implies that, for any $x \in \mathbb{R}^d$,

$$\mathbb{P}\left(X_T \in \partial(-\infty, x]\right) = 0.$$

Hence

$$a_x(X_{K_B}^B) \xrightarrow{\mathbb{P}} a_x(X_T).$$

Since these random variables are bounded by 1, the convergence also holds in L^2. Recall that

$$u_{i,x}^B(z) = \mathbb{E}[a_x(X_{K_B}^B) \mid X_i^B = z], \qquad u_x(t,z) = \mathbb{E}[a_x(X_T) \mid X_t = z].$$

For the remaining part of the proof, we suppress the arguments and write

$$u_{i,x}^B := u_{i,x}^B(X_i^B), \qquad u_{t,x} := u_x(t, X_t).$$

We first prove $\Gamma_{0,B}(x,y) \to \Gamma_0(x,y)$. For any $x \in \mathbb{R}^d$,

$$u^B_{0,x} - u_{0,x} = \mathbb{E}\left[a_x(X^B_{K_B}) - a_x(X_T) \mid \mathscr{F}_0\right].$$

By the L^2-contraction property of conditional expectation,

$$\left\|u^B_{0,x} - u_{0,x}\right\|_{L^2} \le \left\|a_x(X^B_{K_B}) - a_x(X_T)\right\|_{L^2} \to 0, \qquad \forall x \in \mathbb{R}^d.$$

Thus, for all $x, y \in \mathbb{R}^d$, $u^B_{0,x} \to u_{0,x}$ and $u^B_{0,y} \to u_{0,y}$ in L^2. Then, by triangle inequality and then Cauchy-Schwarz, we get

$$\begin{aligned}
&\left|\mathrm{Cov}(u^B_{0,x}, u^B_{0,y}) - \mathrm{Cov}(u_{0,x}, u_{0,y})\right| \\
&\quad = \left|\mathrm{Cov}(u^B_{0,x} - u_{0,x}, u^B_{0,y}) + \mathrm{Cov}(u_{0,x}, u^B_{0,y} - u_{0,y})\right| \\
&\quad \le \left|\mathrm{Cov}(u^B_{0,x} - u_{0,x}, u^B_{0,y})\right| + \left|\mathrm{Cov}(u_{0,x}, u^B_{0,y} - u_{0,y})\right| \\
&\quad \le \|u^B_{0,x} - u_{0,x}\|_{L^2}\|u^B_{0,y}\|_{L^2} + \|u_{0,x}\|_{L^2}\|u^B_{0,y} - u_{0,y}\|_{L^2} \\
&\quad \le \|u^B_{0,x} - u_{0,x}\|_{L^2} + \|u^B_{0,y} - u_{0,y}\|_{L^2} \to 0.
\end{aligned}$$

Hence,

$$\Gamma_{0,B}(x,y) = \mathrm{Cov}(u^B_{0,x}, u^B_{0,y}) \to \mathrm{Cov}(u_{0,x}, u_{0,y}) = \Gamma_0(x,y).$$

It remains to prove the one-step covariance convergence. Write

$$\Delta^{B,z}_i := u^B_{i+1,z} - u^B_{i,z}, \qquad \Delta^z_i := u_{t_{i+1,B},z} - u_{t_{i,B},z}.$$

Then

$$v_{i,B}(x,y) = \mathbb{E}\left[\Delta^{B,x}_i \Delta^{B,y}_i\right], \qquad v^{\mathrm{SDE}}_{i,B}(x,y) = \mathbb{E}\left[\Delta^x_i \Delta^y_i\right].$$

For every $z \in \mathbb{R}^d$, $u^B_{i,z}$, $u_{t_{i,B},z}$, and $u^B_{i,z} - u_{t_{i,B},z}$ are square-integrable martingales. In particular, their increments are centered. Hence, using $ab - cd = (a-c)b + c(b-d)$, the triangle inequality, and Cauchy-Schwarz,

$$\begin{aligned}
&\sum_{i=0}^{K_B-1} \left|v_{i,B}(x,y) - v^{\mathrm{SDE}}_{i,B}(x,y)\right| \\
&\quad = \sum_{i=0}^{K_B-1} \left|\mathbb{E}\left[\Delta^{B,x}_i \Delta^{B,y}_i - \Delta^x_i \Delta^y_i\right]\right| \\
&\quad \le \sum_{i=0}^{K_B-1} \left|\mathbb{E}\left[(\Delta^{B,x}_i - \Delta^x_i)\Delta^{B,y}_i\right]\right| + \sum_{i=0}^{K_B-1} \left|\mathbb{E}\left[\Delta^x_i(\Delta^{B,y}_i - \Delta^y_i)\right]\right| \\
&\quad \le \left(\sum_{i=0}^{K_B-1} \mathbb{E}\left[\left|\Delta^{B,x}_i - \Delta^x_i\right|^2\right]\right)^{1/2} \left(\sum_{i=0}^{K_B-1} \mathbb{E}\left[\left|\Delta^{B,y}_i\right|^2\right]\right)^{1/2} \\
&\qquad + \left(\sum_{i=0}^{K_B-1} \mathbb{E}\left[|\Delta^x_i|^2\right]\right)^{1/2} \left(\sum_{i=0}^{K_B-1} \mathbb{E}\left[\left|\Delta^{B,y}_i - \Delta^y_i\right|^2\right]\right)^{1/2}.
\end{aligned}$$

Therefore, by orthogonality of martingale increments, for all $x \in \mathbb{R}^d$,

$$\sum_{i=0}^{K_B-1} \mathbb{E}\left[\left|\Delta^{B,x}_i - \Delta^x_i\right|^2\right] \le \mathbb{E}\left[\left|u^B_{K_B,x} - u_{T,x}\right|^2\right] = \mathbb{E}\left[\left|a_x(X^B_{K_B}) - a_x(X_T)\right|^2\right] \to 0,$$

and,

$$\sum_{i=0}^{K_B-1} \mathbb{E}\left[\left|\Delta^{B,x}_i\right|^2\right] \le \mathbb{E}\left[\left|u^B_{K_B,x}\right|^2\right] \le 1, \qquad \sum_{i=0}^{K_B-1} \mathbb{E}\left[|\Delta^x_i|^2\right] \le \mathbb{E}\left[|u_{T,x}|^2\right] \le 1.$$

Substituting them into the preceding Cauchy-Schwarz bound gives

$$\sum_{i=0}^{K_B-1} \left|v_{i,B}(x,y) - v^{\mathrm{SDE}}_{i,B}(x,y)\right| \to 0.$$

□

A.7 Proof of Proposition 2

Proof. By Lemma 1 and the definition of Γ_B,

$$\Gamma_B(x,y) = \Gamma_{0,B}(x,y) + \sum_{i=0}^{K_B-1} \frac{v_{i,B}(x,y)}{R_{i,B}}.$$

By the second statement of Proposition 1, we have

$$\Gamma_B(x,y) = \Gamma_{0,B}(x,y) + \sum_{i=0}^{K_B-1} \frac{v_{i,B}^{\mathrm{SDE}}(x,y)}{R_{i,B}} + o(1).$$

By Lemma 3,

$$v_{i,B}^{\mathrm{SDE}}(x,y) = \int_{t_{i,B}}^{t_{i+1,B}} g_{x,y}(t)\,dt,$$

Therefore,

$$\Gamma_B(x,y) = \Gamma_{0,B}(x,y) + \int_0^T \frac{g_{x,y}(t)}{r_B(t)}\,dt + o(1).$$

By Lemma 4, $1 \le r < \infty$ almost everywhere and $\|1/r_B - 1/r\|_{L^1([0,T])} \to 0$. Fix $M > 0$. Since $|1/r_B - 1/r| \le 2$,

$$\int_0^T \left|\frac{1}{r_B(t)} - \frac{1}{r(t)}\right| |g_{x,y}(t)|\,dt \le M \left\|\frac{1}{r_B} - \frac{1}{r}\right\|_{L^1([0,T])} + 2\int_0^T |g_{x,y}(t)|\mathbf{1}(|g_{x,y}(t)| > M)\,dt.$$

Letting $B \to \infty$ and then $M \to \infty$, and using $g_{x,y} \in L^1([0,T])$, gives

$$\int_0^T \frac{g_{x,y}(t)}{r_B(t)}\,dt \to \int_0^T \frac{g_{x,y}(t)}{r(t)}\,dt.$$

The first statement of Proposition 1 then proves $\Gamma_B(x,y) \to \Gamma(x,y)$. Since each Γ_B is a covariance kernel, its pointwise limit Γ is positive semidefinite. It remains to prove the bound on d_Γ. By definition,

$$\begin{aligned} d_\Gamma(x,y)^2 &= \Gamma(x,x) + \Gamma(y,y) - 2\Gamma(x,y) \\ &= \mathrm{Var}(u_x(0,X_0) - u_y(0,X_0)) + \int_0^T \frac{1}{r(t)}[g_{x,x}(t) + g_{y,y}(t) - 2g_{x,y}(t)]\,dt. \end{aligned}$$

Since

$$g_{x,x}(t) + g_{y,y}(t) - 2g_{x,y}(t) = \mathbb{E}\left[\|H_x(t) - H_y(t)\|_2^2\right],$$

and $1/r \le 1$, we get

$$\begin{aligned} d_\Gamma(x,y)^2 &\le \mathrm{Var}(u_x(0,X_0) - u_y(0,X_0)) + \mathbb{E}\left[\int_0^T \|H_x(t) - H_y(t)\|_2^2\,dt\right] \\ &= \mathrm{Var}(a_x(X_T) - a_y(X_T)) \\ &\le \mathbb{E}\left[(a_x(X_T) - a_y(X_T))^2\right] = \|a_x(X_T) - a_y(X_T)\|_{L^2}^2. \end{aligned}$$

□

A.8 Finite Donsker Entropy

Lemma 5 (Finite Donsker Entropy)**.** There are constants $A < \infty$, $v < \infty$, $A' < \infty$, and $v' < \infty$, depending only on d, such that, for $0 < \varepsilon < 1$,

$$\sup_{\mu} N(\varepsilon, \mathcal{A}, L^2(\mu)) \le A\varepsilon^{-v}, \qquad \sup_{B,\nu} N(\varepsilon, \mathcal{F}_B, L^2(\nu)) \le A\varepsilon^{-v}.$$

If

$$\mathcal{H}_{B,\delta} = \{f_{B,x} - f_{B,y} : x, y \in \mathbb{R}^d,\ \rho(a_x, a_y) < \delta\},$$

then

$$\sup_{B,\delta,\nu} N(\varepsilon, \mathcal{H}_{B,\delta}, L^2(\nu)) \le A' \varepsilon^{-v'}.$$

Consequently, $(\mathcal{A}, \rho)$ is totally bounded and the uniform entropy integrals for $\mathcal{A}$ and $\mathcal{H}_{B,\delta}$ with respect to $L^2(\mu)$ and $L^2(\nu)$, respectively, are finite for all measures μ and ν, i.e.,

$$\sup_{\mu} \int_0^1 \sqrt{\log N(\varepsilon, \mathcal{A}, L^2(\mu))}\, d\varepsilon < \infty, \qquad \sup_{B,\delta,\nu} \int_0^1 \sqrt{\log N(\varepsilon, \mathcal{H}_{B,\delta}, L^2(\nu))}\, d\varepsilon < \infty.$$

Proof. The class $\mathcal{A}$ has VC dimension d, so by van der Vaart & Wellner (2023, Theorem 2.6.4) there are constants $A < \infty$ and $v < \infty$, depending only on d, such that

$$N(\varepsilon, \mathcal{A}, L^2(\mu)) \le A\varepsilon^{-v}$$

for every probability measure μ on $\mathbb{R}^d$. To transfer the same bound to $\mathcal{F}_B$, let ν be any probability measure on $(\mathbb{R}^d)^{m_B}$, and let $\nu_1, \dots, \nu_{m_B}$ be its coordinate marginals. Set

$$\overline{\nu} = \frac{1}{m_B} \sum_{\ell=1}^{m_B} \nu_\ell.$$

Let $\mathcal{C}$ be an ε-cover of $\mathcal{A}$ in $L^2(\overline{\nu})$. For $g \in \mathcal{C}$, define

$$\widetilde{g}(z) = \frac{1}{m_B} \sum_{\ell=1}^{m_B} g(z^{(\ell)}).$$

If g is within ε of a_x in $L^2(\overline{\nu})$, then Jensen's inequality gives

$$\begin{aligned} \|f_{B,x} - \widetilde{g}\|^2_{L^2(\nu)} &= \int \left[\frac{1}{m_B} \sum_{\ell=1}^{m_B} \{a_x(z^{(\ell)}) - g(z^{(\ell)})\} \right]^2 d\nu(z) \\ &\le \frac{1}{m_B} \sum_{\ell=1}^{m_B} \int \{a_x(z^{(\ell)}) - g(z^{(\ell)})\}^2 d\nu(z) \\ &= \|a_x - g\|^2_{L^2(\overline{\nu})} \le \varepsilon^2. \end{aligned}$$

Thus the same covering number bound transfers from $\mathcal{A}$ to $\mathcal{F}_B$. Now fix $\delta > 0$. Let $\mathcal{C}_B$ be an $\varepsilon/2$-cover of $\mathcal{F}_B$ in $L^2(\nu)$. For any $f_{B,x} - f_{B,y} \in \mathcal{H}_{B,\delta}$, choose $g_x, g_y \in \mathcal{C}_B$ such that

$$\|f_{B,x} - g_x\|_{L^2(\nu)} \le \varepsilon/2, \qquad \|f_{B,y} - g_y\|_{L^2(\nu)} \le \varepsilon/2.$$

Then the triangle inequality gives

$$\|(f_{B,x} - f_{B,y}) - (g_x - g_y)\|_{L^2(\nu)} \le \varepsilon.$$

Hence

$$N(\varepsilon, \mathcal{H}_{B,\delta}, L^2(\nu)) \le N(\varepsilon/2, \mathcal{F}_B, L^2(\nu))^2 \le A^2 2^{2v} \varepsilon^{-2v}.$$

Taking $A' = A^2 2^{2v}$ and $v' = 2v$ proves the claim. To get the finite entropy bound, use the uniform bound on the covering numbers and observe that the integral is finite. □

A.9 Proof of Theorem 1

Proof. Let $P_B = \mathcal{L}(Z_B)$ and $n_B = N_{0,B}$. For each B, $Z_{B,1}, \dots, Z_{B,n_B}$ are i.i.d. with law P_B, while E_B, P_B, and $\mathcal{F}_B$ may depend on B. We regard the row-wise empirical process as indexed by $\mathcal{A}$, with the coordinate labeled by a_x given by

$$\mathbb{G}_B(a_x) = \frac{1}{\sqrt{n_B}} \sum_{j=1}^{n_B} \left(f_{B,x}(Z_{B,j}) - \mathbb{E}\left[f_{B,x}(Z_B) \right] \right).$$

Indeed, $n_B^{-1} \sum_{j=1}^{n_B} f_{B,x}(Z_{B,j}) = \widehat{F}_B(x)$ and $\mathbb{E}[f_{B,x}(Z_B)] = F_B(x)$. To show weak convergence in $\ell^\infty(\mathcal{A})$, it suffices to prove finite-dimensional convergence and asymptotic ρ-equicontinuity on

$\mathcal{A}$, with a tight Gaussian limit. We first show that all finite-dimensional distributions converge. Fix $a_{x_1}, \ldots, a_{x_J} \in \mathcal{A}$. The summands are centered, i.i.d. within each row, and uniformly bounded. Their covariance matrix is $(\Gamma_B(x_\ell, x_m))_{\ell,m}$, which converges to $(\Gamma(x_\ell, x_m))_{\ell,m}$ by Proposition 2. By the multivariate Lindeberg-Feller theorem, the finite-dimensional distributions converge.

We now construct the limiting object as a tight Borel element of $\ell^\infty(\mathcal{A})$. Proposition 2 shows that Γ is positive semidefinite, so Kolmogorov's extension theorem gives a centered Gaussian process $\overline{\mathbb{G}}$ with covariance Γ. Define the canonical semimetric

$$d_\Gamma(a_x, a_y)^2 = \mathrm{Var}(\overline{\mathbb{G}}(a_x) - \overline{\mathbb{G}}(a_y)) = \Gamma(x,x) + \Gamma(y,y) - 2\Gamma(x,y).$$

Proposition 2 gives $d_\Gamma \le \rho$. By Lemma 5, $(\mathcal{A}, \rho)$ is totally bounded and has finite entropy integral, and therefore

$$\int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{A}, d_\Gamma)}\, d\varepsilon < \infty.$$

By Dudley's entropy theorem (van der Vaart & Wellner, 2023, Corollary 2.2.9), there is a separable version $\mathbb{G}$ of this Gaussian process with bounded and uniformly d_Γ-continuous paths. Since $d_\Gamma \le \rho$, these paths are also uniformly ρ-continuous. Hence, this version is a tight Borel element of $\ell^\infty(\mathcal{A})$. It remains to show asymptotic equicontinuity of $\mathbb{G}_B$, that is, for every $\eta > 0$,

$$\lim_{\delta \downarrow 0} \limsup_{B \to \infty} \mathbb{P}^* \left(\sup_{\rho(a_x, a_y) < \delta} |\mathbb{G}_B(a_x) - \mathbb{G}_B(a_y)| > \eta \right) = 0.$$

By Jensen's inequality and the fact that every terminal descendant has marginal law $X^B_{K_B}$,

$$\mathbb{E}[(f_{B,x}(Z_B) - f_{B,y}(Z_B))^2] \le \mathbb{E}[(a_x(X^B_{K_B}) - a_y(X^B_{K_B}))^2].$$

Then, observe that

$$\mathbb{E}[(a_x(X^B_{K_B}) - a_y(X^B_{K_B}))^2] = F_B(x) + F_B(y) - 2F_B(x \wedge y),$$
$$\rho(a_x, a_y)^2 = F(x) + F(y) - 2F(x \wedge y).$$

Therefore,

$$\mathbb{E}[(f_{B,x}(Z_B) - f_{B,y}(Z_B))^2] \le \rho(a_x, a_y)^2 + 4\|F_B - F\|_\infty.$$

For $\delta > 0$, let $\mathcal{H}_{B,\delta}$ be the class defined in Lemma 5. Put

$$\alpha^2_{B,\delta} = \sup_{h \in \mathcal{H}_{B,\delta}} \mathbb{E}[h(Z_B)^2] = \sup_{\rho(a_x, a_y) < \delta} \mathbb{E}[(f_{B,x}(Z_B) - f_{B,y}(Z_B))^2].$$

By Assumptions 1 and 3, $F_B(x) \to F(x)$ for every $x \in \mathbb{R}^d$. The multivariate Pólya theorem (van der Vaart, 1998, Lemma 2.11) therefore gives $\|F_B - F\|_\infty \to 0$. Hence

$$\lim_{\delta \downarrow 0} \limsup_{B \to \infty} \alpha_{B,\delta} = 0.$$

Fix $t \in (0,1)$. Choose $\delta_t > 0$ such that

$$\limsup_{B \to \infty} \alpha_{B,\delta_t} \le t/2.$$

Then, for every $\delta \le \delta_t$, we have $\alpha_{B,\delta} < t$ for all sufficiently large B. Since $|h| \le 1$ for $h \in \mathcal{H}_{B,\delta}$, the class $\mathcal{H}_{B,\delta}$ has envelope $H \equiv 1$, and van der Vaart & Wellner (2023, Theorem 2.14.2) applied to the B-th row at variance level t gives a constant C, independent of B and δ, with

$$\limsup_{B \to \infty} \mathbb{E}^* \left[\sup_{\rho(a_x, a_y) < \delta} |\mathbb{G}_B(a_x) - \mathbb{G}_B(a_y)| \right] \le \limsup_{B \to \infty} CI(t) \left(1 + \frac{I(t)}{t^2 \sqrt{N_{0,B}}} \right) = CI(t),$$

where

$$I(t) = \sup_{B,\delta,\nu} \int_0^t \sqrt{1 + \log N(\varepsilon, \mathcal{H}_{B,\delta}, L^2(\nu))}\, d\varepsilon,$$

the last equality holding because $I(t) < \infty$ and $N_{0,B} \to \infty$. By Lemma 5, $I(t) \to 0$ as $t \downarrow 0$. Therefore

$$\lim_{\delta \downarrow 0} \limsup_{B \to \infty} \mathbb{E}^* \left[\sup_{\rho(a_x, a_y) < \delta} |\mathbb{G}_B(a_x) - \mathbb{G}_B(a_y)| \right] = 0.$$

Then, by Markov's inequality, $\mathbb{G}_B$ is asymptotically ρ-equicontinuous on $\mathcal{A}$. Hence,

$$\mathbb{G}_B \Rightarrow \mathbb{G} \qquad \text{in } \ell^\infty(\mathcal{A}).$$

□

A.10 Proof of Corollary 1

We first record the cost accounting summarized in Section 4. For a given B, the number of root trees $N_{0,B}$ is determined by the cost of one root tree,

$$C_B = B/N_{0,B} = \sum_{i=0}^{K_B-1} R_{i,B}.$$

The time partition is asymptotically uniform, in the sense that

$$\max_{0\le i<K_B} \left| \frac{t_{i+1,B} - t_{i,B}}{h_B} - 1 \right| \longrightarrow 0.$$

By Assumption 2,

$$h_B C_B = \sum_{i=0}^{K_B-1} h_B R_{i,B} \longrightarrow \int_0^T r(t)\,dt = \gamma,$$

which is finite by the same assumption and is interpreted as the limiting cost of one root tree. Hence

$$\frac{B}{N_{0,B}} = \frac{\gamma}{h_B} + o(h_B^{-1}).$$

The functional CLT concerns $\widehat{F}_B - F_B$. To recover the error relative to F, we write

$$\widehat{F}_B - F = (\widehat{F}_B - F_B) + (F_B - F),$$

and treat the two terms separately.

Proof. Since $C_B = B/N_{0,B}$,

$$B^{p/(1+2p)} N_{0,B}^{-1/2} = \left(\frac{h_B C_B}{\lambda_B} \right)^{1/2} \longrightarrow \left(\frac{\gamma}{\lambda} \right)^{1/2}.$$

Theorem 1 and Slutsky's theorem therefore give

$$B^{p/(1+2p)} (\widehat{F}_B - F_B) \Rightarrow \left(\frac{\gamma}{\lambda} \right)^{1/2} \mathbb{G} \qquad \text{in } \ell^\infty(\mathcal{A}).$$

Moreover, Assumption 4 gives

$$\begin{aligned} B^{p/(1+2p)} \|F_B - F\|_\infty &= O\Big(B^{p/(1+2p)} h_B^p \Big) \\ &= O(\lambda_B^p) = O(1). \end{aligned}$$

Weak convergence in $\ell^\infty(\mathcal{A})$ implies that

$$B^{p/(1+2p)} \|\widehat{F}_B - F_B\|_\infty$$

is bounded in probability. The preceding bias bound and the triangle inequality applied to

$$\widehat{F}_B - F = (\widehat{F}_B - F_B) + (F_B - F)$$

complete the proof. □

A.11 Proof of Proposition 3

Proof. It suffices to bound

$$\mathbb{E}\left[\sup_{x\in\mathbb{R}^d} |\mathbb{G}(a_x)| \right],$$

since $(\gamma/\lambda)^{1/2}$ is deterministic. By Proposition 2, $0 \le \Gamma^* \le 1/2$ and $d_\Gamma \le \rho$. It follows that the d_Γ-diameter of $\mathbb{R}^d$ is at most $2\Gamma^*$. By van der Vaart & Wellner (2023, Corollary 2.2.9), there is a universal constant $C_0 < \infty$ such that

$$\mathbb{E}\left[\sup_{y\in\mathbb{R}^d} |\mathbb{G}(a_y)| \right] \le \mathbb{E}[|\mathbb{G}(a_x)|] + C_0 \int_0^{2\Gamma^*} \sqrt{\log N(\varepsilon, \mathbb{R}^d, d_\Gamma)}\, d\varepsilon, \quad \forall x \in \mathbb{R}^d.$$

The first term in the upper bound can be taken to zero by taking x to infinity in all coordinates. For the second term, the covering number is bounded from Lemma 5 involving a constant that only depends on d. Therefore, for constants $C, C' < \infty$ depending only on d,

$$\mathbb{E}\left[\sup_{x\in\mathbb{R}^d} |\mathbb{G}(a_x)|\right] \le C\int_0^{2\Gamma^*} \sqrt{\log A + v\log(1/\varepsilon)}\, d\varepsilon \le C'\Gamma^*\sqrt{1-\log(\Gamma^*)}.$$

The last inequality follows by the change of variables $\varepsilon = e^{-u}$, with constants absorbed into C'. For the lower bound,

$$\mathbb{E}\left[\sup_{x\in\mathbb{R}^d} |\mathbb{G}(a_x)|\right] \ge \sup_{x\in\mathbb{R}^d} \mathbb{E}[|\mathbb{G}(a_x)|] = \sqrt{\frac{2}{\pi}}\Gamma^*.$$

Restoring the factor $(\gamma/\lambda)^{1/2}$ yields the stated inequalities. □

A.12 Continuity in the Threshold

Lemma 6 (Continuity in the threshold)**.** Let $\mathcal{X} = [-\infty, \infty]^d$ have the product topology. For $x \in \mathcal{X}$, define

$$a_x(z) = \mathbf{1}(z \le x), \qquad u_x(t, X_t) = \mathbb{E}[a_x(X_T) \mid \mathscr{F}_t].$$

using the extended coordinate order. Write

$$u_x(t, X_t) = u_x(0, X_0) + \int_0^t H_x(s)^\top dW_s.$$

and set

$$\Gamma_0(x, x) = \mathrm{Var}(u_x(0, X_0)), \qquad g_{x,x}(t) = \mathbb{E}[\|H_x(t)\|_2^2].$$

Under Assumption 3,

$$x \longmapsto \Gamma_0(x, x)$$

is continuous on $\mathcal{X}$, and

$$x \longmapsto g_{x,x}$$

is continuous from $\mathcal{X}$ into $L^1([0, T])$.

Proof. Continuity of F implies that every marginal of X_T is atomless. Let $x_m \to x$ in $\mathcal{X}$. The atomless-marginal property gives

$$a_{x_m}(X_T) \longrightarrow a_x(X_T) \qquad \text{almost surely.}$$

This also covers coordinates equal to $+\infty$ or $-\infty$. Since the indicators are bounded,

$$\delta_m := \|a_{x_m}(X_T) - a_x(X_T)\|_{L^2} \longrightarrow 0.$$

Conditional expectation is an L^2-contraction. Hence

$$\|u_{x_m}(0, X_0) - u_x(0, X_0)\|_{L^2} \le \delta_m.$$

It follows that

$$\Gamma_0(x_m, x_m) \longrightarrow \Gamma_0(x, x).$$

Subtracting the martingale representations and applying Itô's isometry gives

$$\mathbb{E}\int_0^T \|H_{x_m}(t) - H_x(t)\|_2^2\, dt = \mathbb{E}\left[\mathrm{Var}\left(a_{x_m}(X_T) - a_x(X_T) \mid X_0\right)\right] \le \delta_m^2.$$

The same isometry gives, for every $z \in \mathcal{X}$,

$$\left(\mathbb{E}\int_0^T \|H_z(t)\|_2^2\, dt\right)^{1/2} \le \frac{1}{2}.$$

Therefore, by Cauchy-Schwarz,

$$\begin{aligned}\|g_{x_m,x_m} - g_{x,x}\|_{L^1} &\le \left(\mathbb{E}\int_0^T \|H_{x_m}(t) - H_x(t)\|_2^2\,dt\right)^{1/2} \\ &\quad\times\left[\left(\mathbb{E}\int_0^T \|H_{x_m}(t)\|_2^2\,dt\right)^{1/2} + \left(\mathbb{E}\int_0^T \|H_x(t)\|_2^2\,dt\right)^{1/2}\right] \\ &\le \delta_m \longrightarrow 0.\end{aligned}$$

□

A.13 Proof of Theorem 2

Proof. Let

$$\mathcal{N}_\infty = \left\{ n : [0,T) \to (0,\infty) : \begin{array}{l} n \text{ is nondecreasing and right-continuous,} \\ \displaystyle\int_0^T n(t)\,dt = 1, \quad n(0+) > 0 \end{array} \right\}.$$

For $r \in \mathcal{R}$, set $n = r/\gamma(r)$. Conversely, $r = n/n(0+)$ and $\gamma(r) = 1/n(0+)$. For $n \in \mathcal{N}_\infty$ and $\pi \in \mathcal{P}(\mathcal{X})$, define

$$L(n,\pi) = \frac{\Gamma_0^\pi}{n(0+)} + \int_0^T \frac{g^\pi(t)}{n(t)}\,dt.$$

For fixed $n \in \mathcal{N}_\infty$, let

$$\ell_n(x) = \frac{\Gamma_0(x,x)}{n(0+)} + \int_0^T \frac{g_{x,x}(t)}{n(t)}\,dt.$$

Since $n(t) \ge n(0+) > 0$, Lemma 6 shows that ℓ_n is continuous on $\mathcal{X}$. Moreover,

$$L(n,\pi) = \int_{\mathcal{X}} \ell_n(x)\,\pi(dx).$$

It follows that

$$\sup_{\pi\in\mathcal{P}(\mathcal{X})} L(n,\pi) = \max_{x\in\mathcal{X}} \ell_n(x) = \sup_{x\in\mathbb{R}^d} \ell_n(x).$$

The criterion therefore becomes

$$\inf_{n\in\mathcal{N}_\infty} \sup_{\pi\in\mathcal{P}(\mathcal{X})} L(n,\pi).$$

Fix π, write $h = h_\pi$, and let

$$H(\pi) = \left(\int_0^T \sqrt{h(t)}\,dt\right)^2.$$

Let $q = 1/n$, let $C = C[\Gamma_0^\pi, g^\pi]$, and set

$$\Delta(t) = \Gamma_0^\pi + \int_0^t g^\pi(s)\,ds - C(t).$$

The greatest convex minorant meets its upper bound at T. Since $\Delta \ge 0$ and q is nonincreasing, integration by parts and Cauchy-Schwarz give

$$\begin{aligned}L(n,\pi) &= \int_0^T \frac{h(t)}{n(t)}\,dt - \int_{(0,T)} \Delta(t)\,dq(t) \\ &\ge \int_0^T \frac{h(t)}{n(t)}\,dt \\ &\ge H(\pi).\end{aligned}$$

For $\varepsilon > 0$, define

$$n_\varepsilon(t) = \frac{\sqrt{h(t)+\varepsilon}}{\displaystyle\int_0^T \sqrt{h(s)+\varepsilon}\, ds}, \qquad q_\varepsilon = \frac{1}{n_\varepsilon}.$$

Since h is nondecreasing, $n_\varepsilon \in \mathcal{N}_\infty$. The function C is affine on every interval where $\Delta > 0$. Thus h and q_ε are constant on those intervals. Elsewhere $\Delta = 0$. Hence

$$\int_{(0,T)} \Delta(t)\, dq_\varepsilon(t) = 0.$$

It follows that

$$L(n_\varepsilon, \pi) = \left(\int_0^T \sqrt{h(t)+\varepsilon}\, dt\right) \int_0^T \frac{h(t)}{\sqrt{h(t)+\varepsilon}}\, dt \longrightarrow H(\pi).$$

Therefore

$$\inf_{n\in\mathcal{N}_\infty} L(n,\pi) = H(\pi).$$

When $H(\pi) > 0$, equality in Cauchy-Schwarz forces n to be proportional to $\sqrt{h}$. For this choice, q is constant wherever $\Delta > 0$, so the first inequality is also an equality. It belongs to $\mathcal{N}_\infty$ exactly when $h(0+) > 0$. In that case, the unique minimizer for this mixture is

$$n_\pi(t) = \frac{\sqrt{h_\pi(t)}}{\displaystyle\int_0^T \sqrt{h_\pi(s)}\, ds}.$$

Give $\mathcal{P}(\mathcal{X})$ the weak topology. Regard $\mathcal{N}_\infty$ as a convex subset of $\mathbb{R} \times L^1([0,T])$ through the map

$$n \longmapsto (n(0+), n).$$

For fixed n, the representation above shows that $L(n,\cdot)$ is continuous and affine. We next check continuity in n. Suppose

$$(n_k(0+), n_k) \longrightarrow (n(0+), n) \quad \text{in } \mathbb{R} \times L^1([0,T]).$$

Choose $c > 0$ so that, for all large k,

$$n_k(t) \geq c, \qquad n(t) \geq c$$

for almost every t. Then

$$\left\| \frac{1}{n_k} - \frac{1}{n} \right\|_{L^1} \leq \frac{1}{c^2} \|n_k - n\|_{L^1} \longrightarrow 0.$$

Since $g^\pi \geq 0$, for $A > 0$,

$$\int_0^T g^\pi(t) \left| \frac{1}{n_k(t)} - \frac{1}{n(t)} \right| dt \leq A \left\| \frac{1}{n_k} - \frac{1}{n} \right\|_{L^1} + \frac{2}{c} \int_0^T g^\pi(t) \mathbf{1}(g^\pi(t) > A)\, dt.$$

Letting $k \to \infty$ and then $A \to \infty$ proves convergence of the integral term in $L(n_k, \pi)$. The first term also converges because $n_k(0+) \to n(0+) > 0$. Thus

$$L(n_k, \pi) \longrightarrow L(n, \pi).$$

Thus $L(\cdot, \pi)$ is continuous. It is convex because $\Gamma_0^\pi \geq 0$, $g^\pi \geq 0$, and $z \mapsto 1/z$ is convex. The set $\mathcal{P}(\mathcal{X})$ is compact and convex. The set $\mathcal{N}_\infty$ is convex. Sion's theorem therefore gives

$$\inf_{n\in\mathcal{N}_\infty} \sup_{\pi\in\mathcal{P}(\mathcal{X})} L(n,\pi) = \sup_{\pi\in\mathcal{P}(\mathcal{X})} \inf_{n\in\mathcal{N}_\infty} L(n,\pi) = \sup_{\pi\in\mathcal{P}(\mathcal{X})} H(\pi).$$

Because $H(\pi) = \inf_n L(n,\pi)$, it is upper semicontinuous. Compactness of $\mathcal{P}(\mathcal{X})$ therefore gives a maximizer. Denote the maximum by V. Choose x with $0 < F(x) < 1$. The greatest convex minorant meets the cumulative variance at T, so the covariance decomposition gives

$$\int_0^T h_{\delta_x}(t)\, dt = \Gamma_0(x,x) + \int_0^T g_{x,x}(t)\, dt = F(x)(1-F(x)) > 0.$$

Hence $V \geq H(\delta_x) > 0$. Fix a maximizer π^* and write $h^* = h_{\pi^*}$. Suppose first that $h^*(0+) > 0$. The fixed-mixture result gives

$$n^*(t) = \frac{\sqrt{h^*(t)}}{\sqrt{V}}.$$

Choose $(n_k) \subset \mathcal{N}_\infty$ such that

$$\sup_\pi L(n_k, \pi) \longrightarrow V.$$

Since $H(\pi^*) = V$,

$$V \leq \int_0^T \frac{h^*(t)}{n_k(t)}\, dt \leq L(n_k, \pi^*) \leq \sup_\pi L(n_k, \pi) \longrightarrow V.$$

Since both n_k and n^* integrate to one, Cauchy-Schwarz gives

$$\|n_k - n^*\|_1^2 \leq \int_0^T \frac{(n_k(t) - n^*(t))^2}{n_k(t)}\, dt = \frac{1}{V}\int_0^T \frac{h^*(t)}{n_k(t)}\, dt - 1 \longrightarrow 0.$$

Thus $n_k \to n^*$ in $L^1([0,T])$. After passing to an almost-everywhere convergent subsequence, monotonicity and right-continuity give

$$\limsup_k n_k(0+) \leq n^*(0+).$$

Fatou's lemma now gives, for every π,

$$L(n^*, \pi) \leq \liminf_k L(n_k, \pi) \leq V.$$

Since $L(n^*, \pi^*) = V$, we have

$$\sup_\pi L(n^*, \pi) = V.$$

Thus n^* is optimal. Returning to r gives the formula in the theorem. If instead $h^*(0+) = 0$, a minimizer $\overline{n}$ would satisfy

$$V = \inf_n L(n, \pi^*) \leq L(\overline{n}, \pi^*) \leq \sup_\pi L(\overline{n}, \pi) = V.$$

It would therefore attain the fixed-mixture infimum, which is impossible when $h^*(0+) = 0$. Hence the infimum is not attained. □

A.14 Proof of Corollary 2

Proof. Set $L(n, a) = \sum_i a_i/n_i$. The minimax identity follows from the Sion argument in the proof of Theorem 2 and the fixed-profile calculation in the proof of Lemma 2, with sums in place of integrals. For zero coordinates, apply the calculation to $a + \varepsilon\mathbf{1}$ and let $\varepsilon \downarrow 0$. Compactness of $\mathcal{V}$ and continuity of PAVA give the maximum.

Let $c = \min_i \widetilde{v}_i(x_0) > 0$. Since $\sup_x L(n, \widetilde{V}(x)) \geq c\sum_i n_i^{-1}$, every minimizing sequence stays away from zero coordinates. Compactness of $\overline{\mathcal{N}}$ then gives the minimum. For any maximizer a^* and minimizer n^*, the minimax identity gives

$$\inf_n L(n, a^*) = L(n^*, a^*) = \max_a L(n^*, a).$$

This value is positive by the preceding bound. Thus n^* attains the fixed-profile infimum. The equality case in the proof of Lemma 2 forces every coordinate of $\mathrm{PAVA}(a^*)$ to be positive and gives the stated formula. □

A.15 Proof of Proposition 4

Proof. Let

$$\mathcal{R}_m = \{r \in \mathbb{R}^m_{>0} : 1 = r_0 \leq \cdots \leq r_{m-1}\}.$$

Every $r \in \mathcal{R}_m$ corresponds to the feasible allocation $n_i = Br_i / \sum_j r_j$. Corollary 2 and the definition of r^* therefore give

$$V^* = \inf_{r \in \mathcal{R}_m} \left(\sum_i r_i \right) \sup_{x \in \mathbb{R}^d} \sum_i \frac{\widetilde{v}_i(x)}{r_i}.$$

We first prove the lower bound. Fix any feasible weights λ, and define

$$z_i^{(s)} = \frac{R_i^{(s)}}{C_s}, \qquad \overline{z}_i = \sum_s \lambda_s z_i^{(s)}.$$

Since $C_s = \sum_i R_i^{(s)}$, we have $\sum_i z_i^{(s)} = 1$ and $\sum_i \overline{z}_i = 1$. Each $z^{(s)}$ is positive and nondecreasing. Hence, $\overline{z}$ is also positive and nondecreasing. Convexity of $z \mapsto 1/z$ gives

$$\sum_i \frac{\widetilde{v}_i(x)}{\overline{z}_i} \le \sum_s \lambda_s C_s \sum_i \frac{\widetilde{v}_i(x)}{R_i^{(s)}}.$$

Set

$$r_i = \frac{\overline{z}_i}{\overline{z}_0}.$$

Then $r \in \mathcal{R}_m$. Since $\sum_i \overline{z}_i = 1$,

$$\left(\sum_i r_i \right) \left(\sum_i \frac{\widetilde{v}_i(x)}{r_i} \right) = \sum_i \frac{\widetilde{v}_i(x)}{\overline{z}_i}.$$

Therefore,

$$V^* \le \left(\sum_i r_i \right) \sup_{x \in \mathbb{R}^d} \sum_i \frac{\widetilde{v}_i(x)}{r_i} \le \sup_{x \in \mathbb{R}^d} \sum_s \lambda_s C_s \sum_i \frac{\widetilde{v}_i(x)}{R_i^{(s)}}.$$

This holds for every feasible λ. Taking the minimum over λ proves the lower bound.

We now prove the upper bound. Let $U \sim \mathrm{Unif}[0,1)$. For each candidate profile $R^{(s)}$, define

$$I_s = \{u \in [0,1) : R(u) = R^{(s)}\}, \qquad w_s = \mathbb{P}(U \in I_s).$$

The sets I_s partition $[0,1)$, and w_s is their length. Thus $w_s \ge 0$ and $\sum_s w_s = 1$, so w is a feasible choice of mixture weights. Fix i, j and write

$$\log_2 \left(\frac{r_i^*}{r_j^*} \right) = k + f, \qquad k \in \mathbb{Z}, \quad 0 \le f < 1.$$

The fractional part of $\log_2 r_j^* + U$ is uniform on $[0,1)$. Hence, the rounded exponent difference equals k with probability $1 - f$ and $k+1$ with probability f. Therefore,

$$\mathbb{E}_U \left[\frac{R_i(U)}{R_j(U)} \right] = (1-f)2^k + f2^{k+1} = \frac{r_i^*}{r_j^*}(1+f)2^{-f}.$$

The last factor satisfies

$$\max_{0 \le f < 1} (1+f)2^{-f} = \frac{2}{e \log 2}.$$

For every $x \in \mathbb{R}^d$,

$$\sum_s w_s C_s \sum_j \frac{\widetilde{v}_j(x)}{R_j^{(s)}} = \mathbb{E}_U \left[\left(\sum_i R_i(U) \right) \left(\sum_j \frac{\widetilde{v}_j(x)}{R_j(U)} \right) \right].$$

The ratio bound therefore gives

$$\sum_s w_s C_s \sum_j \frac{\widetilde{v}_j(x)}{R_j^{(s)}} \le \frac{2}{e \log 2} \left(\sum_i r_i^* \right) \left(\sum_j \frac{\widetilde{v}_j(x)}{r_j^*} \right).$$

Taking the supremum over x and using the definition of V^* gives

$$\min_{\substack{\lambda_s \ge 0 \\ \sum_s \lambda_s = 1}} \sup_{x \in \mathbb{R}^d} \sum_s \lambda_s C_s \sum_j \frac{\widetilde{v}_j(x)}{R_j^{(s)}} \le \sup_{x \in \mathbb{R}^d} \sum_s w_s C_s \sum_j \frac{\widetilde{v}_j(x)}{R_j^{(s)}} \le \frac{2}{e \log 2} V^*.$$

This proves the upper bound. □

A.16 Asymptotic Limit of the Implemented Mixture

Write $C_{s,B} = \sum_{i=0}^{K_B-1} R_{i,B}^{(s)}$. For each B, let $\lambda_B = (\lambda_{1,B}, \dots, \lambda_{S_B,B})$ be an optimizer of the problem in Proposition 4. The optimizer need not be unique, and the number of candidate types S_B can grow with B. Instead, we analyse the normalized mixture cost, which is defined as

$$\overline{\gamma}_B = h_B \sum_{s=1}^{S_B} \lambda_{s,B} C_{s,B},$$

and, for $t \in [t_{i,B}, t_{i+1,B})$, define the effective splitting function

$$\overline{r}_B(t) = \frac{\sum_{s=1}^{S_B} \lambda_{s,B} C_{s,B}}{\sum_{s=1}^{S_B} \lambda_{s,B} C_{s,B} / R_{i,B}^{(s)}}.$$

The function $\overline{r}_B$ is nondecreasing and satisfies $\overline{r}_B(0+) = 1$. To obtain integer root counts, set

$$M_{s,B} = \left\lfloor \frac{B\lambda_{s,B}}{C_{s,B}} \right\rfloor$$

For type s, let $\widehat{F}_{B,s}$ be the empirical CDF formed from its $M_{s,B}$ root. The mixture estimator is then

$$\widehat{F}_B^{\mathrm{mix}} = \frac{\sum_{s=1}^{S_B} M_{s,B} C_{s,B} \widehat{F}_{B,s}}{\sum_{s=1}^{S_B} M_{s,B} C_{s,B}}.$$

The candidate profiles and weights are fixed before simulation, and all roots are simulated independently within and across candidate types.

Proposition 5 (Mixture Functional Limit)**.** Suppose Assumptions 1 and 3 hold, the time partitions are asymptotically uniform, and for some $\overline{\gamma} \in (0, \infty)$ and $\overline{r} \in \mathcal{R}$,

$$Bh_B^2 \longrightarrow \infty, \qquad \overline{\gamma}_B \longrightarrow \overline{\gamma}, \qquad \overline{r}_B \longrightarrow \overline{r} \quad \text{in } L^1([0,T]).$$

Then

$$\sqrt{Bh_B}\left(\widehat{F}_B^{\mathrm{mix}} - F_B\right) \Rightarrow \sqrt{\overline{\gamma}}\,\mathbb{G} \qquad \text{in } \ell^\infty(\mathcal{A}),$$

where $\mathbb{G}$ is a centered Gaussian process with covariance

$$\mathrm{Cov}(\mathbb{G}(a_x), \mathbb{G}(a_y)) = \Gamma_0(x,y) + \int_0^T \frac{g_{x,y}(t)}{\overline{r}(t)}\, dt.$$

Moreover,

$$\sqrt{Bh_B}\,\mathrm{KS}\left(\widehat{F}_B^{\mathrm{mix}}, F_B\right) \Rightarrow \sqrt{\overline{\gamma}} \sup_{x \in \mathbb{R}^d} |\mathbb{G}(a_x)|\,.$$

The condition $Bh_B^2 \to \infty$ makes the budget lost to flooring negligible. Under the nondegenerate scaling $h_B \propto B^{-1/(1+2p)}$, this holds whenever $p > 1/2$, including the usual Euler case $p = 1$.

The individual optimized weights need not converge, and a candidate type need not receive a growing number of roots. The earlier covariance, functional CLT, and KS limits centered at F_B carry over with (γ, r) replaced by $(\overline{\gamma}, \overline{r})$. Under Assumption 4, the bias-variance analysis carries over as well. Hence, the relaxed allocation can be approximated by a mixture of valid dyadic trees while retaining the same form of the asymptotic limit theory.

Proof. For any two candidate types s and t, logarithmic rounding gives $R_{i,B}^{(s)}/R_{i,B}^{(t)} \in [1/2, 2]$, and hence $C_{s,B} \le 2C_{t,B}$. Therefore,

$$\max_s C_{s,B} \le \frac{2\overline{\gamma}_B}{h_B} = O(h_B^{-1}), \qquad \sum_s C_{s,B} = O(h_B^{-2}) = o(B),$$

where the first bound follows by averaging $C_{s,B} \le 2C_{t,B}$ with respect to $\lambda_{t,B}$, and the last uses $S_B \le K_B = O(h_B^{-1})$ and $Bh_B^2 \to \infty$. Let

$$D_B = \sum_s M_{s,B} C_{s,B}, \qquad w_{s,B} = \frac{M_{s,B} C_{s,B}}{D_B},$$

be the realized budget and mixture weights after flooring. Then

$$0 \le B - D_B < \sum_s C_{s,B} = o(B), \qquad \frac{D_B}{B} \longrightarrow 1, \qquad \|w_B - \lambda_B\|_1 \le \frac{2(B - D_B)}{D_B} \longrightarrow 0.$$

In particular,

$$h_B \sum_s |w_{s,B} - \lambda_{s,B}| C_{s,B} \le \left(\max_s h_B C_{s,B}\right) \|w_B - \lambda_B\|_1 \longrightarrow 0.$$

Let $Z_{s,B,j}$ be the vector of terminal descendants from root j of type s, and let $f_{s,B,x}(Z_{s,B,j})$ be their empirical CDF at x. Then

$$\begin{aligned}\mathbb{G}_B^{\mathrm{mix}}(a_x) &:= \sqrt{Bh_B}\left(\widehat{F}_B^{\mathrm{mix}}(x) - F_B(x)\right) \\ &= \sum\nolimits_{s,j} a_{s,B}\left[f_{s,B,x}(Z_{s,B,j}) - F_B(x)\right], \qquad a_{s,B} = \frac{C_{s,B}\sqrt{Bh_B}}{D_B}.\end{aligned}$$

Moreover,

$$\max_s a_{s,B} = O((Bh_B)^{-1/2}) \longrightarrow 0, \qquad Q_B := \sum_s M_{s,B} a_{s,B}^2 = \frac{B}{D_B} h_B \sum_s w_{s,B} C_{s,B} \longrightarrow \overline{\gamma},$$

where the last limit follows from the preceding bound and the definition of $\overline{\gamma}_B$. Thus, no individual root dominates, while the total squared coefficient converges to $\overline{\gamma}$. We next identify the limiting covariance. For one root of type s, write

$$\Gamma_B^{(s)}(x,y) = \Gamma_{0,B}(x,y) + \sum_i \frac{v_{i,B}(x,y)}{R_{i,B}^{(s)}}.$$

Independence of the roots and the definition of $\overline{r}_B$ give

$$\begin{aligned}\mathrm{Cov}\big(\mathbb{G}_B^{\mathrm{mix}}(a_x), \mathbb{G}_B^{\mathrm{mix}}(a_y)\big) &= \frac{B}{D_B} h_B \sum_s w_{s,B} C_{s,B} \Gamma_B^{(s)}(x,y), \\ h_B \sum_s \lambda_{s,B} C_{s,B} \Gamma_B^{(s)}(x,y) &= \overline{\gamma}_B \left[\Gamma_{0,B}(x,y) + \sum_i \frac{v_{i,B}(x,y)}{\overline{r}_{i,B}}\right].\end{aligned}$$

Since $|\Gamma_B^{(s)}(x,y)| \le 1/4$, the preceding weight bound allows w_B to be replaced by λ_B at a cost of $o(1)$. Proposition 1 and the proof of Proposition 2, with $\overline{r}_B$ in place of r_B, therefore give

$$\mathrm{Cov}\big(\mathbb{G}_B^{\mathrm{mix}}(a_x), \mathbb{G}_B^{\mathrm{mix}}(a_y)\big) \longrightarrow \overline{\gamma}\left[\Gamma_0(x,y) + \int_0^T \frac{g_{x,y}(t)}{\overline{r}(t)}\, dt\right].$$

Since the centered root averages are bounded by one, for every $\eta > 0$,

$$\sum_{s,j} \mathbb{E}\Big[\|a_{s,B}(f_{s,B,\cdot} - F_B)\|_{\mathcal{A}}^2 \mathbf{1}\big\{\|a_{s,B}(f_{s,B,\cdot} - F_B)\|_{\mathcal{A}} > \eta\big\}\Big] \le Q_B \mathbf{1}\{\max_s a_{s,B} > \eta\} \longrightarrow 0.$$

Thus, the envelope Lindeberg condition holds, and the covariance limit and the multivariate Lindeberg-Feller theorem give the finite-dimensional limits.

To upgrade this to process convergence, it remains to prove asymptotic ρ-equicontinuity. For every root type, Jensen's inequality gives

$$\begin{aligned}\mathbb{E}\Big[\big(f_{s,B,x}(Z_{s,B,j}) - f_{s,B,y}(Z_{s,B,j})\big)^2\Big] &\le \mathbb{E}\Big[\big(a_x(X_{K_B}^B) - a_y(X_{K_B}^B)\big)^2\Big] \\ &\le \rho(a_x,a_y)^2 + 4\|F_B - F\|_\infty.\end{aligned}$$

The proof of Theorem 1 shows that $\|F_B - F\|_\infty \to 0$. Hence, for every $\delta_B \downarrow 0$,

$$\sup_{\rho(a_x,a_y)<\delta_B} \sum_{s,j} a_{s,B}^2 \mathbb{E}\Big[\big(f_{s,B,x}(Z_{s,B,j}) - f_{s,B,y}(Z_{s,B,j})\big)^2\Big] \le Q_B\left(\delta_B^2 + 4\|F_B - F\|_\infty\right) \longrightarrow 0.$$

For the entropy condition, define

$$d_B(a_x, a_y)^2 = \sum_{s,j} a_{s,B}^2 \left(f_{s,B,x}(Z_{s,B,j}) - f_{s,B,y}(Z_{s,B,j})\right)^2.$$

Let ν_B be the probability measure

$$\nu_B = \frac{1}{Q_B} \sum_{s,j} \frac{a_{s,B}^2}{R^{(s)}_{K_B-1,B}} \sum_{\ell=1}^{R^{(s)}_{K_B-1,B}} \delta_{Z^{(\ell)}_{s,B,j}}.$$

Jensen's inequality gives

$$d_B(a_x, a_y) \le \sqrt{Q_B}\, \|a_x - a_y\|_{L^2(\nu_B)}.$$

The first entropy bound in Lemma 5 therefore gives, uniformly over realizations,

$$N(\varepsilon, \mathcal{A}, d_B) \le A \left(1 + \frac{\sqrt{Q_B}}{\varepsilon}\right)^v.$$

Consequently, for every $\eta_B \downarrow 0$,

$$\int_0^{\eta_B} \sqrt{\log N(\varepsilon, \mathcal{A}, d_B)}\, d\varepsilon \lesssim \eta_B \sqrt{\log(1 + \eta_B^{-1})} \longrightarrow 0.$$

The class $\mathcal{A}$ is pointwise measurable and $(\mathcal{A}, \rho)$ is totally bounded by Lemma 5. The envelope, increment, and entropy bounds above therefore imply asymptotic ρ-equicontinuity by the independent-array equicontinuity theorem (van der Vaart & Wellner, 2023, Theorem 2.11.1).

The limiting covariance kernel is positive semidefinite, and Proposition 2 bounds its canonical semimetric by $\sqrt{\overline{\gamma}}\rho$. Lemma 5 and Dudley's entropy theorem therefore give a tight Gaussian process with the covariance in the statement. Together with the finite-dimensional convergence and asymptotic equicontinuity,

$$\mathbb{G}_B^{\mathrm{mix}} \Rightarrow \sqrt{\overline{\gamma}}\,\mathbb{G} \qquad \text{in } \ell^\infty(\mathcal{A}).$$

The KS limit follows from the continuous mapping theorem. □

A.17 Cost-weighted extensions

The main text assigns unit cost to every transition and uses an asymptotically uniform time partition. Neither restriction is essential, and neither holds in our experiments: the Heun sampler of Appendix D spends two network evaluations on every step but the last, and the diffusion samplers use graded time grids. This subsection removes both. The budget is spent on cost rather than on elapsed time, so we measure time in cost. Each result below is then the corresponding unit-cost result in this variable.

Let $c_{i,B} > 0$ be the cost of propagating one particle through transition i. One root tree costs

$$C_B = \sum_{i=0}^{K_B-1} c_{i,B} R_{i,B}, \qquad N_{0,B} = B/C_B.$$

For $t \in [t_{i,B}, t_{i+1,B})$, define the cost density

$$\omega_B(t) = \frac{h_B c_{i,B}}{t_{i+1,B} - t_{i,B}}.$$

Here ω_B is interpreted as computational cost per unit of physical time, measured on the scale h_B. Summing over transitions gives the exact identity

$$h_B C_B = \int_0^T \omega_B(t) r_B(t)\, dt.$$

Thus the normalized cost of a root tree is the area under the splitting function when time is weighted by cost. Unit costs on an asymptotically uniform partition give $\omega_B \to 1$ uniformly, and the right side then converges to γ, recovering the normalization of Appendix A.10.

Only one hypothesis is needed. It asks that the cost densities have a limit, and that this limit pairs with the splitting functions in the obvious way.

Assumption 5. There is a measurable $\omega : [0,T] \to (0,\infty)$, bounded above and bounded away from zero, such that

$$\int_0^T \omega_B(t) r_B(t)\, dt \longrightarrow \gamma_\omega(r) := \int_0^T \omega(t) r(t)\, dt$$

whenever $r_B \to r$ as in Assumption 2.

Uniform convergence $\omega_B \to \omega$ is sufficient, since the r_B are bounded in $L^1([0,T])$. Here $\gamma_\omega(r)$ is interpreted as the limiting cost of one root tree. It reduces to $\gamma(r)$ when $\omega \equiv 1$, and takes its place throughout.

Retain the remaining hypotheses of Corollary 1, and replace unit costs and the asymptotically uniform partition by Assumption 5. If $\lambda_B \to \lambda \in (0,\infty)$, then

$$B^{p/(1+2p)}(\widehat{F}_B - F_B) \Rightarrow \left(\frac{\gamma_\omega(r)}{\lambda}\right)^{1/2} \mathbb{G} \qquad \text{in } \ell^\infty(\mathcal{A}),$$

and $B^{p/(1+2p)}\|\widehat{F}_B - F\|_\infty$ is bounded in probability. The costs and the partition enter Corollary 1 only through the limit of $h_B C_B$, which the identity above and Assumption 5 now supply. Theorem 1 uses neither, so the rest of that proof is unchanged.

Cost therefore enters the limit through $\gamma_\omega(r)$ and nowhere else. Proposition 3 applies with $\gamma(r)$ replaced by $\gamma_\omega(r)$, and dropping the same logarithmic factor gives the surrogate $\gamma_\omega(r)^{1/2}\Gamma^*$. Minimizing its square gives the cost-weighted proxy objective

$$\inf_{r\in\mathcal{R}} \gamma_\omega(r) \sup_{x\in\mathbb{R}^d} \left[\Gamma_0(x,x) + \int_0^T \frac{g_{x,x}(t)}{r(t)}\, dt\right].$$

We now measure time in cost. Define the cumulative cost and the total cost

$$\Omega(t) = \int_0^t \omega(u)\, du, \qquad \Omega_T = \Omega(T).$$

Since $\omega > 0$, the map Ω is strictly increasing, so $s = \Omega(t)$ is a change of variables from $[0,T]$ onto $[0,\Omega_T]$. Writing $\overline{r}(s) = r(\Omega^{-1}(s))$ gives

$$\gamma_\omega(r) = \int_0^{\Omega_T} \overline{r}(s)\, ds, \qquad \int_0^T \frac{g_{x,x}(t)}{r(t)}\, dt = \int_0^{\Omega_T} \frac{\overline{g}_{x,x}(s)}{\overline{r}(s)}\, ds, \qquad \overline{g}_{x,x}(s) = \frac{g_{x,x}(\Omega^{-1}(s))}{\omega(\Omega^{-1}(s))}.$$

Since Ω is increasing, $\overline{r}$ is again nondecreasing with $\overline{r}(0+) = 1$, so $r \mapsto \overline{r}$ matches $\mathcal{R}$ with the same class on $[0,\Omega_T]$. The cost-weighted objective is thus the unit-cost objective in the variable s. Accordingly, for $\pi \in \mathcal{P}(\mathcal{X})$, set

$$\overline{g}^\pi(s) = \frac{g^\pi(\Omega^{-1}(s))}{\omega(\Omega^{-1}(s))}, \qquad \overline{h}_\pi = \mathrm{PAVA}(\Gamma_0^\pi, \overline{g}^\pi), \qquad 0 \le s < \Omega_T.$$

Under Assumption 3,

$$\inf_{r\in\mathcal{R}} \gamma_\omega(r) \sup_{x\in\mathbb{R}^d} \left[\Gamma_0(x,x) + \int_0^T \frac{g_{x,x}(t)}{r(t)}\, dt\right] = \max_{\pi\in\mathcal{P}(\mathcal{X})} \left(\int_0^{\Omega_T} \sqrt{\overline{h}_\pi(s)}\, ds\right)^2.$$

For any maximizer π^*, the infimum is attained if and only if $\overline{h}_{\pi^*}(0+) > 0$, in which case

$$r^*(t) = \sqrt{\frac{\overline{h}_{\pi^*}(\Omega(t))}{\overline{h}_{\pi^*}(0+)}}, \qquad 0 \le t < T,$$

defines a minimizer. Both claims are Theorem 2 applied on $[0,\Omega_T]$ with inputs $(\Gamma_0^\pi, \overline{g}^\pi)$, read back through Ω.

Geometrically, $\overline{h}_\pi$ is obtained by plotting cumulative variance against cumulative cost, with Γ_0^π as an initial atom at cost zero, and taking the slopes of the greatest convex minorant. In the main text every

transition costs the same, so cumulative cost is the index and this is the equal-weight construction of Appendix A.1. Here the horizontal axis is stretched by ω. Equivalently, in physical time PAVA acts on $g^\pi(t)/\omega(t)$ with weights $\omega(t)\,dt$.

The finite-budget construction is the same, with finitely many points, and defines the cost-weighted fit referred to in Appendix A.1. Fix B, put $m = K_B$, and write $c_i = c_{i,B}$. For $z \in [0,\infty)^m$, let $\mathrm{PAVA}_c(z)_i$ be the slope of the greatest convex minorant of the polygonal line through

$$\left(\sum_{j<k} c_j,\ \sum_{j<k} z_j\right), \qquad 0 \le k \le m.$$

As in Appendix A.1, these slopes are constant on the final blocks, so that

$$\mathrm{PAVA}_c(z)_i = \frac{\sum_{j\in I} z_j}{\sum_{j\in I} c_j}, \qquad i \in I.$$

Cumulative cost has replaced the index on the horizontal axis, and the block average has become a cost-weighted average. For $c \equiv 1$ this is the equal-weight fit PAVA.

Replace the feasible set $\mathcal{N}$ of Corollary 2 by

$$\mathcal{N}_c = \left\{ n \in \mathbb{R}^m_{>0} : n_0 \le \cdots \le n_{m-1},\ \sum_{i=0}^{m-1} c_i n_i = B \right\},$$

and retain $\widetilde{v}_i$ and $\mathcal{V}$ as defined there.

Under the assumptions of Corollary 2, the minimum below is attained, and

$$\min_{n\in\mathcal{N}_c} \sup_{x\in\mathbb{R}^d} \sum_{i=0}^{m-1} \frac{\widetilde{v}_i(x)}{n_i} = \frac{1}{B} \max_{a\in\mathcal{V}} \left(\sum_{i=0}^{m-1} c_i \sqrt{\mathrm{PAVA}_c(a)_i} \right)^2.$$

If a^* is a maximizer and $h^* = \mathrm{PAVA}_c(a^*)$, then $h^*_i > 0$ for every i, and every minimizing allocation satisfies

$$n^*_i = B \frac{\sqrt{h^*_i}}{\sum_{j=0}^{m-1} c_j \sqrt{h^*_j}}, \qquad 0 \le i < m.$$

The argument of Corollary 2 goes through with two substitutions. The PAVA block inequality becomes $\sum_i a_i q_i \ge \sum_i c_i h_i q_i$ for nonincreasing q, which is again summation by parts against a minorant gap that is nonnegative and vanishes at both ends. Cauchy-Schwarz is then applied as

$$\left(\sum_{i=0}^{m-1} \frac{c_i h_i}{n_i} \right) \left(\sum_{i=0}^{m-1} c_i n_i \right) \ge \left(\sum_{i=0}^{m-1} c_i \sqrt{h_i} \right)^2,$$

with equality when $n \propto \sqrt{h}$, which is nondecreasing and hence feasible. The Sion argument and the positivity of h^* are unchanged.

This is the allocation problem solved in the experiments. Appendix B.1.2 gives the Frank-Wolfe scheme used to compute a maximizing a^*, and Appendix E uses it for the oracle benchmark.

A.18 Extension to MMD

Let k be a bounded kernel on $\mathbb{R}^d$ with reproducing kernel Hilbert space $\mathcal{H}$. The maximum mean discrepancy between two distributions P and P' is

$$\mathrm{MMD}(P, P') = \sup_{\|f\|_{\mathcal{H}} \le 1} |Pf - P'f|.$$

For a Gaussian kernel, it is zero only when $P = P'$ (Gretton et al., 2012). Computing it between two samples requires the kernel at every pair of points. Random Fourier features (Rahimi & Recht, 2007) avoid this. For the Gaussian kernel $k(z, z') = \exp(-\|z - z'\|_2^2/(2\eta^2))$, Bochner's theorem gives

$k(z, z') = \mathbb{E}[\cos(\omega^\top(z - z'))]$ with $\omega \sim N(0, \eta^{-2}I)$. Draw frequencies $\omega_1, \dots, \omega_M$ and define $\phi : \mathbb{R}^d \to \mathbb{R}^{d_\phi}$, with $d_\phi = 2M$, by

$$\phi(z) = \frac{1}{\sqrt{M}} \left(\cos \omega_1^\top z, \dots, \cos \omega_M^\top z,\ \sin \omega_1^\top z, \dots, \sin \omega_M^\top z\right).$$

Then $\phi(z)^\top \phi(z') = M^{-1} \sum_j \cos(\omega_j^\top (z - z')) \approx k(z, z')$, and $\|\phi(z)\|_2 = 1$ for every z. Writing $P\phi = (P\phi_1, \dots, P\phi_{d_\phi})$, the MMD becomes

$$\mathrm{MMD}(P, P') \approx \|P\phi - P'\phi\|_2,$$

which costs time linear in the sample sizes. In our experiments, $M = 1{,}024$, the kernel is an equal mixture of Gaussian kernels at eight bandwidths, and ϕ is applied to an image after eight-bit quantization and standardization (Appendix C).

The error therefore depends on the splitting estimator only through the d_ϕ means $\widehat{P}_B\phi_k$, and it is a continuous function of them. The steps used for KS carry over. We prove a central limit theorem for the vector $\widehat{P}_B\phi$, bound the expected norm of its limit, and solve the resulting allocation problem.

Notation. Let $\widehat{P}_B\phi$ be the average of ϕ over the terminal particles of the splitting estimator, and let $P_B\phi = \mathbb{E}[\phi(X^B_{K_B})]$. The vector analogue of $u^B_{i,x}$ is

$$\psi_i(z) = \mathbb{E}\big[\phi(X^B_{K_B}) \mid X^B_i = z\big] \in \mathbb{R}^{d_\phi}, \qquad 0 \le i \le K_B.$$

Define the $d_\phi \times d_\phi$ matrices

$$\Gamma^\phi_{0,B} = \mathrm{Cov}\big(\psi_0(X^B_0)\big), \qquad v^\phi_{i,B} = \mathbb{E}\big[\mathrm{Cov}\big(\psi_{i+1}(X^B_{i+1}) \mid X^B_i\big)\big].$$

Let Γ_ϕ be the matrix whose (k, l) entry is the limit Γ of Proposition 2, with a_x and a_y replaced by ϕ_k and ϕ_l.

Because of the quantization step, ϕ is not continuous. It is discontinuous only where a pixel value lies on a rounding threshold. Let D_ϕ be this set of discontinuity points. We need $\mathbb{P}(X_T \in D_\phi) = 0$, which holds when every pixel of X_T has a continuous distribution. For FFHQ, ϕ is applied to the decoded image, and the condition refers to its pixels.

Lemma 7. For every B,

$$\mathrm{Cov}\Big(\widehat{P}_B\phi\Big) = \frac{1}{N_{0,B}} \left[\Gamma^\phi_{0,B} + \sum_{i=0}^{K_B - 1} \frac{v^\phi_{i,B}}{R_{i,B}}\right].$$

If, in addition, Assumptions 1 and 2 hold and $\mathbb{P}(X_T \in D_\phi) = 0$, then

$$N_{0,B}\, \mathrm{Cov}\Big(\widehat{P}_B\phi\Big) \longrightarrow \Gamma_\phi.$$

Proof. For the covariance formula, apply Lemma 1 to each pair of coordinates ϕ_k, ϕ_l. Its proof uses only that a_x is bounded, and each ϕ_k is bounded.

For the limit, the proofs of Propositions 1 and 2 use two facts about a_x. The first is $|a_x| \le 1$, and $|\phi_k| \le 1$ as well. The second is $a_x(X^B_{K_B}) \to a_x(X_T)$ in L^2. The same holds for ϕ_k. Assumption 1 gives $X^B_{K_B} \to X_T$ in probability. Since $\mathbb{P}(X_T \in D_\phi) = 0$ and ϕ_k is bounded,

$$\phi_k(X^B_{K_B}) \xrightarrow{\mathbb{P}} \phi_k(X_T) \qquad \text{and hence} \qquad \phi_k(X^B_{K_B}) \xrightarrow{L^2} \phi_k(X_T).$$

The first limit is the continuous mapping theorem, and the second follows from boundedness. The rest of both proofs is unchanged. □

The vector $\widehat{P}_B\phi$ is the average of $N_{0,B}$ i.i.d. root-tree averages, which are bounded and have mean $P_B\phi$. Hence, if $N_{0,B} \to \infty$, the Lindeberg-Feller theorem gives

$$\sqrt{N_{0,B}}\left(\widehat{P}_B\phi - P_B\phi\right) \Rightarrow \mathbb{Z} \sim N(0, \Gamma_\phi).$$

Since $\|h\|_2 \le \sqrt{d_\phi}\, \|h\|_\infty$ for $h \in \mathbb{R}^{d_\phi}$, the map $h \mapsto \|h\|_2$ is continuous, and the continuous mapping theorem gives

$$\sqrt{N_{0,B}}\left\|\widehat{P}_B\phi - P_B\phi\right\|_2 \Rightarrow \|\mathbb{Z}\|_2.$$

Lemma 8. Let $\mathbb{Z} \sim N(0, \Gamma)$ in $\mathbb{R}^{d_\phi}$. Then

$$\sqrt{2/\pi}\,\sqrt{\operatorname{tr}\Gamma} \le \mathbb{E}\|\mathbb{Z}\|_2 \le \sqrt{\operatorname{tr}\Gamma}.$$

Proof. For the upper bound, Jensen's inequality gives

$$\mathbb{E}\|\mathbb{Z}\|_2 \le \left(\mathbb{E}\|\mathbb{Z}\|_2^2\right)^{1/2} = \sqrt{\operatorname{tr}\Gamma}.$$

For the lower bound, let $\lambda_1, \dots, \lambda_{d_\phi}$ be the eigenvalues of Γ with orthonormal eigenvectors $e_1, \dots, e_{d_\phi}$. Write $\mathbb{Z} = \sum_k \sqrt{\lambda_k}\,\xi_k e_k$, where $\xi_1, \dots, \xi_{d_\phi}$ are i.i.d. standard normal. By the Cauchy-Schwarz inequality,

$$\sum_k \lambda_k |\xi_k| = \sum_k \sqrt{\lambda_k} \cdot \sqrt{\lambda_k}\,|\xi_k| \le \sqrt{\operatorname{tr}\Gamma}\,\|\mathbb{Z}\|_2.$$

Taking expectations and using $\mathbb{E}|\xi_k| = \sqrt{2/\pi}$ gives $\sqrt{2/\pi}\ \operatorname{tr}\Gamma \le \sqrt{\operatorname{tr}\Gamma}\,\mathbb{E}\|\mathbb{Z}\|_2$. □

Allocation. Lemma 8 is the analogue of Proposition 3. Its two sides differ only by the constant $\sqrt{\pi/2}$, so we choose the allocation to minimize the trace of the covariance. Let $n_i = N_{0,B} R_{i,B}$ be the number of particles propagated through transition i. Taking the trace of the covariance formula in Lemma 7 gives

$$\operatorname{tr}\operatorname{Cov}\left(\widehat{P}_B\phi\right) = \sum_{i=0}^{K_B-1} \frac{w_i}{n_i}, \qquad w_i = \operatorname{tr} v_{i,B}^{\phi} + \mathbf{1}_{\{i=0\}} \operatorname{tr}\Gamma_{0,B}^{\phi}.$$

This has the same form as the variance at a single threshold in Section 3, with w_i in place of $\widetilde{v}_i$. The minimizer over monotone allocations is therefore given by Lemma 2. With segment costs c_i, Appendix A.17 applies with $\mathcal{V} = \{w\}$, and

$$n_i^* = B\frac{\sqrt{h_i}}{\sum_j c_j \sqrt{h_j}}, \qquad h = \operatorname{PAVA}_c(w).$$

The dyadic mixture of Proposition 4 then uses a single tree type, and the factor $2/(e \log 2)$ still holds. The weights w_i can be written in terms of second moments, which is how Phase 1 estimates them in Appendix B.2. Let $Q_i = \mathbb{E}\|\psi_i(X_i^B)\|_2^2$. Since $\psi_i(X_i^B)$ is a martingale, $\operatorname{tr} v_{i,B}^{\phi} = Q_{i+1} - Q_i$ and $\operatorname{tr}\Gamma_{0,B}^{\phi} = Q_0 - \|P_B\phi\|_2^2$. Hence

$$w_0 = Q_1 - \|P_B\phi\|_2^2, \qquad w_i = Q_{i+1} - Q_i, \quad 1 \le i < K_B,$$

and $Q_{K_B} = \mathbb{E}\|\phi(X_{K_B}^B)\|_2^2 = 1$.

B Two-Phase Algorithms

This appendix gives the two-phase procedures used in the experiments. Both use Phase 1 to estimate the variance added by each path segment and Phase 2 to run a mixture of dyadic trees. They differ in what Phase 1 estimates and in how the allocation problem is solved.

B.1 KS

This procedure is used for the OU, overdamped Langevin, and EDM models. Phase 1 uses complete paths to estimate the variance added by each path segment. We then solve a cost-weighted allocation problem. Phase 2 converts the relaxed allocation into a finite mixture of exact dyadic trees. Appendix B.2 gives the corresponding procedure for the MMD.

B.1.1 Phase 1: Simulation, Query Generation, and Estimation

The allowed split times

$$0 = t_0 < t_1 < \cdots < t_{L-1} < t_L = T$$

divide each path into L segments. Let $c_i > 0$ be the cost of propagating one particle through segment i. One complete path costs

$$C_{\text{full}} = \sum_{i=0}^{L-1} c_i.$$

Phase 1 receives budget B_1. It simulates

$$J = \left\lfloor \frac{B_1}{C_{\text{full}}} \right\rfloor$$

independent complete paths without splitting. Each path is recorded at $t_0, \ldots, t_L$. The remaining Phase 2 budget is

$$B_2 = B - JC_{\text{full}}.$$

Define the Phase 1 empirical CDF and its realized budget share by

$$\widehat{F}_1(x) = \frac{1}{J} \sum_{j=1}^{J} \mathbf{1}\Big(X_T^{(j)} \le x\Big), \qquad \beta_{\text{P}} = \frac{JC_{\text{full}}}{B}.$$

We retain these terminal samples for the final estimate, where β_{P} is the fraction of the total budget actually spent on the complete Phase 1 paths. Given these paths, we want to estimate the quantities $v_i(x)$ for all $x \in \mathbb{R}^d$. We replace this with a finite-query approximation using Q queries. Further, to avoid sparse queries, we restrict the number of active coordinates in each query to be at most $d_{\max}$. We use $d_{\max} = 2$. Each repetition draws its own set of queries.

For query q, draw

$$k_q \sim \text{Unif}\{1, \ldots, d_{\max}\}.$$

Select k_q coordinates uniformly without replacement. Set

$$\mu_q = 0.05 + \left(q - \tfrac{1}{2}\right) \frac{0.90}{Q}, \qquad 1 \le q \le Q.$$

These levels are evenly spaced between 0.05 and 0.95. For each query, draw

$$(s_{q,1}, \ldots, s_{q,k_q}) \sim \text{Dirichlet}(1, \ldots, 1).$$

For each selected coordinate a, set $x_{q,a}$ to its Phase 1 empirical quantile at level $\mu_q^{s_{q,a}}$. Set every unselected coordinate to $+\infty$. Those coordinates then impose no restriction on the event.

Given these queries, we now estimate the variance contributions. For each query q, let x_q be the corresponding threshold vector. Define

$$Y_q^{(j)} = \mathbf{1}\Big(X_T^{(j)} \le x_q\Big), \qquad \widehat{F}_q = \frac{1}{J} \sum_{j=1}^{J} Y_q^{(j)}.$$

Let F_B denote the terminal CDF on the discretization used at budget B. For each query, define

$$u_q(t_i, z) = \mathbb{P}(X_T \le x_q \mid X_{t_i} = z)$$

and

$$Q_i(x_q) = \mathbb{E}\big[u_q(t_i, X_{t_i})^2\big].$$

We estimate $u_q(t_i, \cdot)$ by regressing $Y_q^{(j)}$ on the path state at t_i and the query x_q. We assign the pilot paths cyclically to K folds. For each fold, we train one MLP on the other folds and predict the held-out paths. The networks have hidden widths 128 and 64, SiLU activations, and a sigmoid output.

One MLP is shared across all recorded times and queries within a fold. Its input consists of the active-coordinate margins, time, the fraction of active coordinates, and the logit of the query mass. Each input is standardized before fitting. The label is $Y_q^{(j)}$.

We train for five epochs with squared-error loss. Let $p_{iq}^{(j)}$ be the held-out prediction for path j, time t_i, and query q. For a fixed prediction rule p,

$$\mathbb{E}\big[2p(X_{t_i})Y_q - p(X_{t_i})^2\big] = Q_i(x_q) - \mathbb{E}\Big[\big(p(X_{t_i}) - u_q(t_i, X_{t_i})\big)^2\Big].$$

This identity motivates

$$\widehat{Q}_i^{\mathrm{raw}}(x_q) = \frac{1}{J}\sum_{j=1}^{J}\left[2p_{iq}^{(j)}Y_q^{(j)} - \left(p_{iq}^{(j)}\right)^2\right], \qquad 0 \le i < L.$$

We set

$$\widehat{Q}_L^{\mathrm{raw}}(x_q) = \widehat{F}_q.$$

We note that our estimator is not unbiased. Each $p_{iq}^{(j)}$ comes from an MLP trained without path j, but the query thresholds and $\widehat{F}_q$ still use the full Phase 1 sample. Now, see that the true second moments satisfy

$$F_B(x_q)^2 \le Q_0(x_q) \le \cdots \le Q_L(x_q) = F_B(x_q).$$

Sampling error can violate this order. We first clip the raw estimates to $[\widehat{F}_q^2, \widehat{F}_q]$. We then apply nondecreasing PAVA to the full sequence while holding the terminal value fixed at $\widehat{F}_q$. Write $\widehat{Q}_i(x_q)$ for the projected values. The root population controls both the initial-state variance and the variance added in the first segment. The estimated contributions used by the optimizer are

$$\widehat{v}_0(x_q) = \widehat{Q}_1(x_q) - \widehat{F}_q^2$$

and

$$\widehat{v}_i(x_q) = \widehat{Q}_{i+1}(x_q) - \widehat{Q}_i(x_q), \qquad 1 \le i < L.$$

B.1.2 Optimization

Let $n_i > 0$ be the relaxed total number of particles propagated through segment i. For query q, define

$$\ell_q(n) = \sum_{i=0}^{L-1}\frac{\widehat{v}_i(x_q)}{n_i}.$$

We solve

$$\min_n \max_{1\le q\le Q} \ell_q(n)$$

subject to

$$\sum_{i=0}^{L-1} c_i n_i = B_2, \qquad 0 < n_0 \le \cdots \le n_{L-1}.$$

We solve the query-mixture dual with Frank-Wolfe. Let $\pi_q \ge 0$ be the current query weights, with

$$\sum_{q=1}^{Q}\pi_q = 1.$$

Form

$$p_i = \sum_{q=1}^{Q}\pi_q\widehat{v}_i(x_q).$$

Apply the cost-weighted fit of Appendix A.17 to obtain $h = \mathrm{PAVA}_c(p)$. The resulting relaxed allocation is

$$n_i = B_2\frac{\sqrt{h_i}}{\sum_{r=0}^{L-1} c_r\sqrt{h_r}}, \qquad 0 \le i < L.$$

At each Frank-Wolfe step, we recompute the allocation and choose

$$q^\star \in \arg\max_{1\le q\le Q} \ell_q(n).$$

A line search chooses $\eta \in [0,1]$. We then update

$$\pi \leftarrow (1-\eta)\pi + \eta e_{q^\star}.$$

Here, $e_{q^\star}$ is the unit vector for query $q^\star$. The final Frank-Wolfe iterate gives the relaxed particle counts $n_0, \ldots, n_{L-1}$.

B.1.3 FINAL ALGORITHM

Algorithm 1 summarizes the complete procedure. Phase 1 learns the allocation and supplies a full-path component. Phase 2 uses fresh paths to construct the splitting component.

Algorithm 1 Final two-phase algorithm

Require: Budget B, pilot budget B_1, split times, segment costs $c_0, \dots, c_{L-1}$, and queries $x_1, \dots, x_Q$

1: Use budget B_1 to estimate the variance contributions and retain $\widehat{F}_1$ as in Appendix B.1.1
2: Use the remaining budget B_2 to solve the relaxed allocation problem in Appendix B.1.2
3: Normalize the solution by setting $r_i = n_i/n_0$
4: Construct the candidate dyadic profiles $R^{(1)}, \dots, R^{(S)}$ from r as in Section 6
5: Solve the mixture problem in Proposition 4 to obtain λ
6: For each type s, set $C_s = \sum_i c_i R_i^{(s)}$, set $M_s = \lfloor \lambda_s B_2 / C_s \rfloor$, and simulate M_s fresh trees
7: Set $B_2' = \sum_{s=1}^{S} M_s C_s$ and, for $x \in \mathbb{R}^d$,

$$\widehat{F}_2(x) = \sum_{s=1}^{S} \sum_{j=1}^{M_s} \sum_{\ell=1}^{R_{L-1}^{(s)}} \frac{C_s}{B_2' R_{L-1}^{(s)}} \mathbf{1}\Big(Z_{s,j}^{(\ell)} \le x\Big)$$

8: Set $\beta_{\mathrm{P}} = JC_{\mathrm{full}}/B$
9: **return** $\widehat{F}_B^{\mathrm{reuse}} = \beta_{\mathrm{P}} \widehat{F}_1 + (1 - \beta_{\mathrm{P}}) \widehat{F}_2$

Each Phase 1 terminal sample receives weight β_{P}/J, and each Phase 2 leaf of type s receives weight

$$\frac{(1 - \beta_{\mathrm{P}}) C_s}{B_2' R_{L-1}^{(s)}}.$$

Conditional on Phase 1,

$$\mathbb{E}\Big[\widehat{F}_B^{\mathrm{reuse}}(x) \mid \text{Phase 1}\Big] = \beta_{\mathrm{P}} \widehat{F}_1(x) + (1 - \beta_{\mathrm{P}}) F_B(x),$$

so the reused estimator is generally not conditionally unbiased. It is exactly unconditionally unbiased because $\mathbb{E}[\widehat{F}_1(x)] = F_B(x)$. Moreover, $B_1 = o(B)$ implies $\beta_{\mathrm{P}} \to 0$, so the conditional discrepancy introduced by reuse vanishes asymptotically.

B.2 MMD

This procedure is used for CIFAR-10 and FFHQ. It follows Appendix B.1 with three changes. Phase 1 estimates the single profile w of Appendix A.18 from pairs of paths that share a history, without queries or regression. The allocation then has a closed form. The dyadic mixture uses a single tree type. For FFHQ the chain runs in the latent space and the features are computed on decoded images.

Design map. The reported MMD uses the feature map ϕ of Appendix C. Phase 1 uses a second map ϕ', built in the same way but with independently drawn frequencies, so the allocation is chosen without access to the features on which it is evaluated. Below, Q_i and w_i refer to ϕ'.

Pilot. Each pilot path consists of a complete path and one branch. The branch starts from the state of the complete path at an anchor time t_a and runs independently to T, so the path costs $C_{\mathrm{full}} + \sum_{i \ge a} c_i$. Given the state at t_a, the two terminal features are independent with mean ψ_a evaluated at that state, so their inner product has mean Q_a. The A anchor levels are the nearest integers to A evenly spaced points in $[0, L-1]$, and paths are assigned to them cyclically. We fit as many paths as possible within $B_1 = \lfloor 10 B^{0.66} \rfloor$, and take the largest A for which every anchor receives at least five paths. Let J be the number of pilot paths, $B_1' \le B_1$ their total cost, and $B_2 = B - B_1'$. In our experiments, the pilot uses between 40 and 87 paths and between 11% and 25% of the budget.

Estimation. Let Y_j and Y'_j be the features ϕ' of the terminal states of the complete path and the branch of pilot path j, and let $\mathcal{J}_a$ be the set of paths anchored at a. We estimate

$$\widehat{Q}_a = \frac{1}{|\mathcal{J}_a|} \sum_{j \in \mathcal{J}_a} Y_j^\top Y'_j, \qquad \widehat{S} = \max\left\{0,\ \frac{\left\|\sum_{j=1}^J Y_j\right\|_2^2 - \sum_{j=1}^J \|Y_j\|_2^2}{J(J-1)}\right\},$$

where $\widehat{S}$ estimates $\|P_B\phi'\|_2^2$. Values between anchors are obtained by linear interpolation, and we set $\widehat{Q}_L = 1$. The interpolation error is not covered by the theory. As in Appendix B.1.1, we clip to $[\widehat{S}, 1]$ and apply nondecreasing PAVA with the terminal value fixed. The estimate $\widehat{w}$ is then given by the formulas for w in Appendix A.18, with $\widehat{Q}$ and $\widehat{S}$ in place of Q and $\|P_B\phi'\|_2^2$.

Root floor. Entries of $\widehat{w}$ are raised to at least $10^{-12}\sum_i \widehat{w}_i$. For the learned allocation, if the design affords fewer than 32 Phase 2 roots, we raise this floor by bisection until it affords 32, since the limit theory requires $N_{0,B} \to \infty$.

Allocation. The learned allocation is the minimizer of Appendix A.18 with budget B_2 and $\widehat{w}$ in place of w. For Learned-c, in the notation of Appendix D.3,

$$\widehat{c} \in \operatorname*{argmin}_{1 \le c \le c_{\max}} \left(\sum_{i=0}^{L-1} \kappa_i c^i\right)\left(\sum_{i=0}^{L-1} \widehat{w}_i c^{-i}\right),$$

over real c, where $c_{\max}$ is the largest c for which $r(c)$ affords 32 roots with budget B_2. The dyadic mixture problem has a single row with entries $C_s \sum_i \widehat{w}_i / R_i^{(s)}$, so all of B_2 goes to the candidate with the smallest entry.

Estimator. Both terminal states of every pilot path have law P_B, and all $2J$ are retained. With $\beta_{\mathrm{P}} = B'_1/B$, each pilot state receives weight $\beta_{\mathrm{P}}/(2J)$, and the Phase 2 leaves share the weight $1-\beta_{\mathrm{P}}$ equally. The estimator of $P_B f$ is unconditionally unbiased for every bounded f. Algorithm 2 summarizes the procedure.

Algorithm 2 Two-phase algorithm for MMD

Require: Budget B, pilot budget B_1, split times, segment costs $c_0, \dots, c_{L-1}$, and design map ϕ'

1: Choose the anchors and simulate the J pilot paths, at cost B'_1
2: Estimate $\widehat{w}$ from the sibling inner products, and apply the root floor
3: Set $B_2 = B - B'_1$ and $n_i = B_2\sqrt{h_i}/\sum_r c_r\sqrt{h_r}$ with $h = \mathrm{PAVA}_c(\widehat{w})$
4: Normalize by setting $r_i = n_i/n_0$, and construct the dyadic candidates $R^{(1)}, \dots, R^{(S)}$ from r
5: Select the candidate s minimizing $C_s \sum_i \widehat{w}_i / R_i^{(s)}$, and simulate $M = \lfloor B_2/C_s \rfloor$ fresh trees of that type
6: Set $\beta_{\mathrm{P}} = B'_1/B$
7: **return** $\widehat{P}_G$, which gives weight $\beta_{\mathrm{P}}/(2J)$ to each pilot terminal state and weight $(1-\beta_{\mathrm{P}})/(MR_{L-1}^{(s)})$ to each Phase 2 leaf

C Metrics and Evaluations

Appendix D defines the models and simulation settings. To evaluate each model, we compute a metric between the estimated CDF and a reference CDF. For each model, the reference CDF is generated once and fixed, and the same sample is used for every method, budget, split schedule, and repetition. For the OU-process and overdamped Langevin models, the reference CDF is the empirical distribution of a large number of samples obtained by simulating a finely discretized SDE. For the EDM model, the reference CDF is the empirical distribution of a large number of samples drawn directly from the Gaussian mixture. For CIFAR-10 and FFHQ, the reference CDF is the empirical distribution of a large number of samples generated by the full DDPM sampler, which for FFHQ runs in latent space and is decoded to images. Table 1 gives the source and size of each sample.

Table 1: Reference samples used to evaluate the generated samples.

Model	Reference source	Samples
OU process	Euler-Maruyama with 20,000 steps	2,500,000
Overdamped Langevin	Euler-Maruyama with 20,000 steps	2,500,000
EDM	Direct draws from the Gaussian mixture	5,000,000
CIFAR-10 DDPM	Full 1,000-step DDPM sampler	20,000
FFHQ LDM	Full 1,000-step latent DDPM sampler, decoded to 256×256 images	20,000

For any function f on the terminal state space, define the generated and reference empirical operators by

$$\widehat{P}_G f = \frac{\beta_{\mathrm{P}}}{J} \sum_{j=1}^{J} f\left(X_T^{(j)}\right) + (1-\beta_{\mathrm{P}}) \sum_{s=1}^{S} \sum_{j=1}^{M_s} \sum_{\ell=1}^{R_{L-1}^{(s)}} \frac{C_s}{B_2' R_{L-1}^{(s)}} f\left(Z_{s,j}^{(\ell)}\right), \qquad \widehat{P}_R f = \frac{1}{|R|} \sum_{z \in R} f(z).$$

Thus every Phase 1 output receives weight β_{P}/J, every Phase 2 leaf of type s receives weight $(1-\beta_{\mathrm{P}})C_s/(B_2' R_{L-1}^{(s)})$, and the fixed reference sample remains uniformly weighted. This is the estimator of Appendix B.1. For the MMD procedure, the Phase 1 term averages over the $2J$ pilot terminal states instead, as in Appendix B.2. The OU, overdamped Langevin, and EDM examples are two-dimensional. For $a_x(z) = \mathbf{1}(z \le x)$, compare the generated output G with the fixed reference sample R using the two-sample lower-orthant KS distance

$$D_{\mathrm{KS}}(G, R) = \sup_{x \in \mathbb{R}^2} \left| \widehat{P}_G a_x - \widehat{P}_R a_x \right|.$$

Both empirical CDFs change only at observed coordinate values. We evaluate that grid and compute the weighted supremum exactly.

The same grid calculation is not practical for CIFAR-10, where each image has 3,072 coordinates, or for FFHQ, where each decoded image has 196,608. Hence, we use the maximum mean discrepancy (Gretton et al., 2012). We compute it with the random Fourier feature approximation of Zhao & Meng (2015), using a fixed map drawn once and reused for every comparison. Phase 1 of the MMD procedure uses a second, independently drawn map (Appendix B.2).

Images are quantized to eight bits. Each pixel coordinate is standardized using the mean and standard deviation of the fixed reference sample. Let η be the median Euclidean distance over sampled pairs of standardized reference images. We use the eight scales

$$\eta\, 2^{-4}, \ldots, \eta\, 2^{3}.$$

We draw 1,024 orthogonal random frequencies and divide them equally across these scales. Each frequency contributes one sine coordinate and one cosine coordinate. The coordinates are scaled so that squared feature distances are averaged over the frequencies.

Let ϕ denote the resulting feature map. For generated output G and reference sample R, define

$$\widehat{\mathrm{MMD}}_{\phi}(G, R) = \left\| \widehat{P}_G \phi - \widehat{P}_R \phi \right\|_2.$$

Each learned-splitting repetition reruns Phase 1, optimization, dyadic-mixture construction, and Phase 2, and reuses its Phase 1 terminal samples with weight β_{P}. All methods have the same total budget B. We use 2,500 repetitions for each two-dimensional setting and 50 repetitions for each CIFAR-10 and FFHQ setting. We report the mean empirical-reference error.

For a splitting method with mean error $\overline{D}_{\mathrm{split}}$ and an independent-path baseline with mean error $\overline{D}_{\mathrm{ind}}$, the reported percentage reduction is

$$100 \left(1 - \frac{\overline{D}_{\mathrm{split}}}{\overline{D}_{\mathrm{ind}}} \right).$$

D Benchmark Models and Simulation Settings

D.1 Models

Table 2: Benchmark models. The EDM denoiser is trained and sampled in standardized coordinates, so its prior is stated in that scale.

OU Process

$$dX_{t,j} = \big[1.35(-0.2 - X_{t,j}) + 0.25(\bar{X}_t - X_{t,j})\big]dt + 0.65\, dW_{t,j},$$
$$\bar{X}_t = (X_{t,1} + X_{t,2})/2$$
$$j = 1, 2, \qquad X_{0,j} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1), \qquad T = 1$$

Overdamped Langevin

$$U(x_1, x_2) = (x_1^2 - 1)^2 + (x_2^2 - 1)^2 + \tfrac{1}{2}(x_2 - x_1)^2,$$
$$dX_t = -\nabla U(X_t)\, dt + \sqrt{2}\, dW_t$$
$$j = 1, 2, \qquad X_{0,j} \overset{\text{iid}}{\sim} \mathcal{N}(-1, 0.05), \qquad T = 2$$

EDM

$$\frac{dX_\sigma}{d\sigma} = \frac{X_\sigma - D_\theta(X_\sigma, \sigma)}{\sigma} = -\sigma s_\theta(X_\sigma, \sigma),$$
$$s_\theta(x, \sigma) = \frac{D_\theta(x, \sigma) - x}{\sigma^2}$$
$$X_{\sigma_{\max}} \sim \mathcal{N}(0, \sigma_{\max}^2 I_2), \qquad \sigma_{\max} = 80, \qquad \sigma_{\min} = 0.002$$
Neural denoiser: D_θ

CIFAR-10 DDPM

`google/ddpm-cifar10-32`

Bundled DDPM scheduler, with step counts in Table 3

FFHQ LDM

`asparius/ldm-ffhq-256`, a latent DDPM paired with a KL autoencoder

DDPM scheduler rebuilt from the shipped betas, with step counts in Table 3. Sampling runs on $4 \times 32 \times 32$ latents, which are decoded to 256×256 images

The EDM block of Table 2 shows the probability-flow ODE used by the sampler. Following Karras et al. (2022), we trained the neural denoiser D_θ on samples from the equally weighted Gaussian mixture

$$\frac{1}{2}\mathcal{N}\left(\begin{bmatrix}-1\\1\end{bmatrix}, \begin{bmatrix}0.20 & 0.05\\0.05 & 0.30\end{bmatrix}\right) + \frac{1}{2}\mathcal{N}\left(\begin{bmatrix}1\\-1\end{bmatrix}, \begin{bmatrix}0.30 & -0.08\\-0.08 & 0.15\end{bmatrix}\right).$$

We use the stochastic EDM sampler with time-step exponent 3, in the parameterization of Karras et al. (2022). At each step, stochastic churn is applied before a second-order Heun update of the probability-flow ODE. The churn parameters are $(S_{\text{churn}}, S_{\min}, S_{\max}, S_{\text{noise}}) = (40, 0, 80, 1)$. Direct samples from the mixture are used only for the reference distribution.

D.2 Simulation settings

The nine-split schedule splits every 10% from 10% through 90%. The nineteen-split schedule splits every 5% from 5% through 95%. The thirty-nine-split schedule splits every 2.5% from 2.5% through 97.5%. For the KS procedure of Appendix B.1, we use $K = 5$ folds and $Q = 1{,}024$ queries. We take $B_1 = \lfloor 5B^{0.66} \rfloor$ for OU and overdamped Langevin and $B_1 = \lfloor 10B^{0.66} \rfloor$ for EDM. CIFAR-10 and FFHQ use the MMD procedure of Appendix B.2 with $B_1 = \lfloor 10B^{0.66} \rfloor$. All methods use the same time grid across budgets, with the number of steps per budget given in Table 3.

OU and overdamped Langevin are solved with Euler-Maruyama, EDM uses the second-order Heun solver of Karras et al. (2022), and DDPM uses the DDPM sampler. Table 3 gives the step counts and EDM network-evaluation counts.

Table 3: Numerical schedules by budget. Euler steps apply to OU and overdamped Langevin. DDPM steps apply to both CIFAR-10 and FFHQ, which are run only through $B = 500$k. EDM NFE is the number of network evaluations per sample.

B	Euler steps	EDM steps	EDM NFE	DDPM steps
50k	-	-	-	200
100k	130	40	79	252
200k	160	46	91	317
500k	220	55	109	430
1M	280	63	125	-
2M	350	73	145	-
5M	475	87	173	-

The step and network-evaluation counts in Table 3 determine the segment costs c_i. Cost is measured in numerical steps for the OU, overdamped Langevin, and DDPM samplers, and in network evaluations for EDM. The final EDM step omits the corrector and costs one evaluation, so K solver steps use $2K - 1$ network evaluations.

D.3 Uniform and Learned c Baselines

Both baselines restrict the normalized relaxed profile to

$$r_i(c) = c^i, \qquad 0 \le i < L.$$

To distinguish the split factor c from the segment costs c_i above, write those costs as κ_i in this subsection. Uniform-c uses no pilot phase and fixes c before simulation. We construct the exact dyadic candidate trees by the logarithmic rounding in Section 6. Since this baseline has no variance estimates, we assign each candidate the length of the set of shifts $u \in [0, 1)$ that produces it, floor the resulting root counts, and use the realized budget shares in the estimator. For OU, overdamped Langevin, and EDM, we use $c \in \{1.1, 1.15, 1.25, 1.5\}$ with nine splits, $c \in \{1.1, 1.15, 1.2\}$ with nineteen splits, and $c \in \{1.05, 1.1\}$ with thirty-nine splits.

Learned-c uses the same Phase 1 estimates $\widehat{v}_i(x_q)$, pilot budget, and remaining budget B_2 as the learned allocation, but restricts the optimization to the one-parameter family above. It selects

$$\widehat{c} \in \operatorname*{argmin}_{c \ge 1} \left(\sum_{i=0}^{L-1} \kappa_i c^i \right) \max_{1 \le q \le Q} \left(\sum_{i=0}^{L-1} \widehat{v}_i(x_q) c^{-i} \right).$$

After constructing the dyadic candidates $R^{(1)}, \dots, R^{(S)}$ from $r(\widehat{c})$, define

$$C_s = \sum_{i=0}^{L-1} \kappa_i R_i^{(s)}, \qquad E_{sq} = C_s \sum_{i=0}^{L-1} \frac{\widehat{v}_i(x_q)}{R_i^{(s)}}.$$

The Phase 2 budget shares solve the linear program

$$\begin{aligned} \min_{\lambda, t} \quad & t \\ \text{subject to} \quad & \sum_{s=1}^{S} \lambda_s E_{sq} \le t, \qquad 1 \le q \le Q, \\ & \sum_{s=1}^{S} \lambda_s = 1, \qquad \lambda_s \ge 0. \end{aligned}$$

We initially assign $M_s = \lfloor \lambda_s B_2 / C_s \rfloor$ roots. Candidate types receiving no roots are removed and the program is re-solved. As in Algorithm 1, the Phase 1 terminal samples are retained with their realized budget share and the Phase 2 weights are scaled by the remaining share. For CIFAR-10 and FFHQ, Learned-c uses the trace profile $\widehat{w}$ in place of the query values, as described in Appendix B.2.

E OU FINITE-QUERY MINIMAX BENCHMARK

For the linear-Gaussian OU process, the Euler chain gives the variance contributions without Phase 1 estimation. We use them to solve a finite-query version of the minimax problem in Corollary 2. We call the result the oracle allocation.

E.1 ORACLE ALLOCATION

For a lower-orthant event $A_x = \{X_K \le x\}$, define

$$M_k(x) = \mathbb{P}(A_x \mid X_k), \qquad Q_x(k) = \mathbb{E}[M_k(x)^2].$$

The quantity $Q_x(k)$ is the probability that two terminal descendants both lie in A_x when they share the state at step k and use independent noise afterward. Because the process is linear and Gaussian, we evaluate $F(x)$ and $Q_x(k)$ from the Euler-chain covariances and numerical Gaussian CDFs. If $k_1, \dots, k_{L-1}$ are the split steps, their successive differences give the segmentwise variance vector

$$\begin{aligned} v_0(x) &= Q_x(k_1) - F(x)^2, \\ v_i(x) &= Q_x(k_{i+1}) - Q_x(k_i), \quad 1 \le i < L-1, \\ v_{L-1}(x) &= F(x) - Q_x(k_{L-1}). \end{aligned}$$

We evaluate these vectors on a fixed set of 10,200 lower-orthant queries. A Frank-Wolfe solver optimizes over their convex hull using the cost-weighted finite minimax allocation of Appendix A.17. Thus the allocation comes from the finite minimax problem rather than one representative query. Normalize the relaxed oracle counts by their root count and generate the corresponding exact tree candidates. Since the exact variance table is available, solve the cost-weighted tree-mixture linear program with that table. Floor the root counts and use realized budget-share weights.

E.2 SIMULATED KS EVALUATION

At $B = 5 \times 10^6$, the Euler chain has $K = 475$ steps. We compute the oracle allocation once for each split schedule. Figure 4 compares this allocation with the learned mean allocation over 2,500 complete experiment repetitions. Here R_i is the number of particles per initial root after split i. A value of one means that no branching is needed.

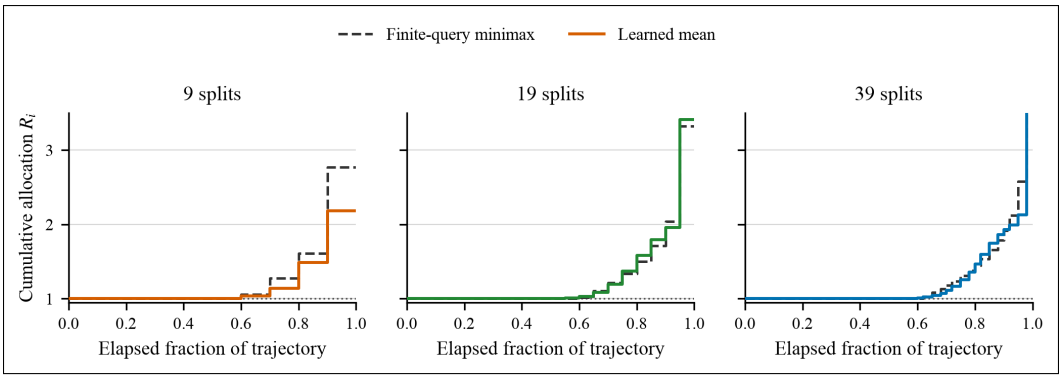

Figure 4: Finite-query minimax and learned mean cumulative OU allocations for schedules with 9, 19, and 39 splits of the 475-step Euler process at $B = 5 \times 10^6$. Learned values are means over 2,500 runs.

We fix the oracle factors and run the same budgeted splitting simulator used for the learned allocation. For each of 10,000 independent repetitions, we compute the exact two-sample lower-orthant KS distance to the fixed reference sample described in Appendix C. The distance for one repetition is computable, but its expectation under the branching scheme is not available in closed form. Table 4 therefore reports the simulated reduction in mean KS relative to the shared independent-path baseline.

Table 4: Reduction in measured mean OU KS at $B = 5 \times 10^6$ for the finite-query minimax and learned allocations relative to the shared independent-path baseline. Half-widths of two-sided 90% normal confidence intervals are in parentheses. Oracle means use 10,000 repetitions. Learned and independent-path means use 2,500. The final column is the learned reduction as a percentage of the minimax-allocation reduction.

Split times	Oracle reduction	Learned reduction	Oracle captured
9	13.9 (0.7)	14.0 (0.9)	100.7%
19	22.8 (0.6)	18.8 (0.9)	82.6%
39	22.4 (0.6)	21.8 (0.9)	97.2%

F Complete Numerical Results

Figure 5 reports the absolute error behind the percentage reductions in the main text. It also shows how the mean error changes with the computational budget.

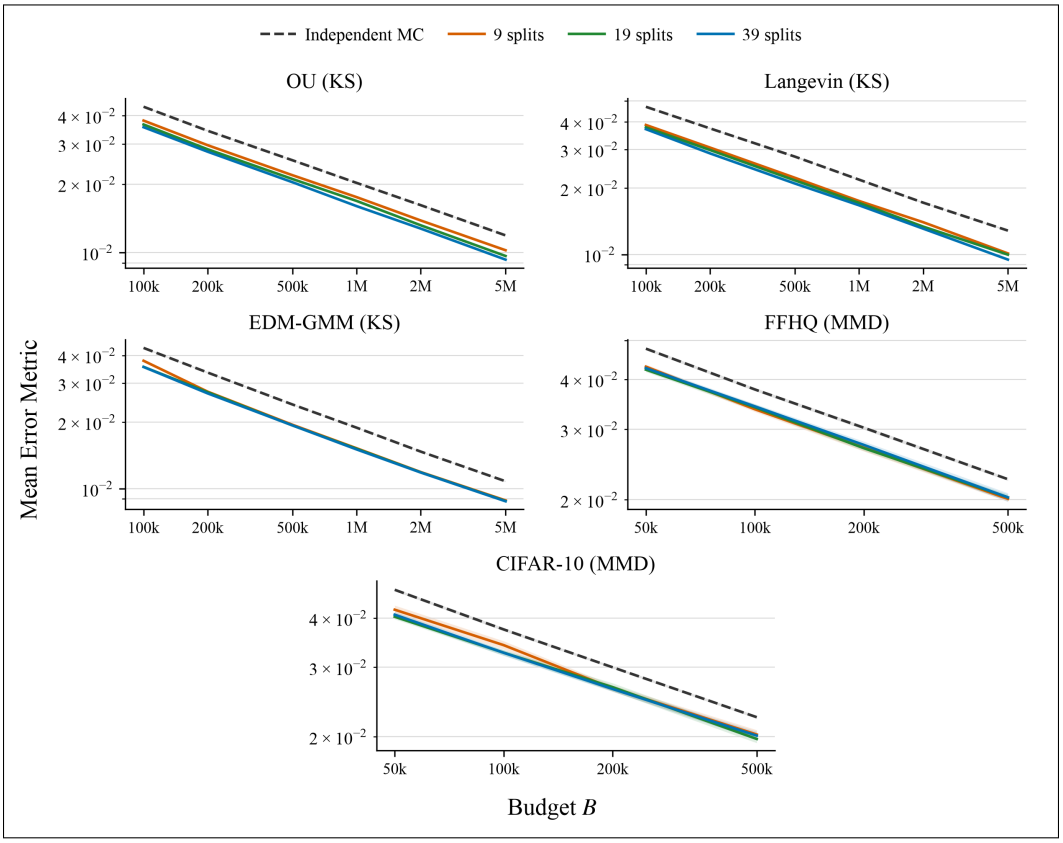

Figure 5: Absolute mean error versus budget for independent full-path simulation and the learned allocation. The first three panels show KS over 2,500 repetitions, and the FFHQ and CIFAR-10 panels show MMD over 50 repetitions. Both axes are logarithmic.

Tables 5-9 report the percentage reduction in mean error relative to independent full-path simulation. Each cell gives the estimated reduction, with the half-width of its two-sided 90% normal confidence interval in parentheses, computed by the delta method for the ratio of two independent sample means. In each budget column, the method with the lowest mean error is bold.

Table 5: OU process: percentage reduction in mean KS at each budget and split schedule, for the learned allocation, Uniform-c, and Learned-c. Half-widths of two-sided 90% normal confidence intervals are in parentheses.

Allocation	100k	200k	500k	1M	2M	5M
9 split points						
Learned	13.1 (0.9)	13.6 (0.9)	14.1 (0.9)	13.4 (0.9)	14.4 (0.9)	14.0 (0.9)
Learned c	12.9 (0.9)	12.1 (0.9)	12.9 (0.9)	12.5 (0.9)	13.4 (0.9)	13.3 (0.9)
Uniform ($c = 1.1$)	9.5 (1.0)	8.6 (1.0)	8.7 (1.0)	8.6 (0.9)	8.4 (1.0)	8.8 (0.9)
Uniform ($c = 1.15$)	11.5 (1.0)	12.0 (1.0)	10.2 (1.0)	9.8 (1.0)	10.8 (0.9)	10.5 (1.0)
Uniform ($c = 1.25$)	11.6 (1.0)	11.2 (1.0)	11.8 (1.0)	9.9 (1.0)	11.0 (1.0)	11.2 (1.0)
Uniform ($c = 1.5$)	-12.1 (1.5)	-12.6 (1.5)	-11.4 (1.5)	-13.8 (1.5)	-11.3 (1.5)	-11.0 (1.5)
19 split points						
Learned	16.4 (0.9)	17.1 (0.9)	17.3 (0.9)	16.7 (0.9)	18.4 (0.9)	18.8 (0.9)
Learned c	12.1 (0.9)	12.4 (0.9)	12.6 (0.9)	13.0 (0.9)	13.5 (0.9)	13.7 (0.9)
Uniform ($c = 1.1$)	11.8 (1.0)	11.3 (1.0)	11.7 (1.0)	10.5 (1.0)	11.1 (1.0)	11.2 (1.0)
Uniform ($c = 1.15$)	7.3 (1.1)	8.3 (1.1)	6.9 (1.2)	6.6 (1.1)	7.6 (1.1)	6.6 (1.1)
Uniform ($c = 1.2$)	-4.0 (1.4)	-5.9 (1.4)	-5.4 (1.4)	-4.9 (1.4)	-5.0 (1.4)	-5.2 (1.4)
39 split points						
Learned	**18.7 (0.9)**	**19.1 (0.9)**	**20.0 (0.9)**	**21.0 (0.8)**	**21.1 (0.9)**	**21.8 (0.9)**
Learned c	12.1 (0.9)	12.0 (0.9)	13.1 (0.9)	12.4 (0.9)	13.7 (0.9)	13.2 (0.9)
Uniform ($c = 1.05$)	12.3 (1.0)	10.9 (1.0)	12.2 (1.0)	10.8 (1.0)	11.0 (1.0)	11.4 (1.0)
Uniform ($c = 1.1$)	-9.9 (1.5)	-8.9 (1.5)	-8.8 (1.5)	-9.7 (1.5)	-9.5 (1.5)	-9.5 (1.5)

Table 6: Overdamped Langevin: percentage reduction in mean KS at each budget and split schedule, for the learned allocation, Uniform-c, and Learned-c. Half-widths of two-sided 90% normal confidence intervals are in parentheses.

Allocation	100k	200k	500k	1M	2M	5M
9 split points						
Learned	17.3 (1.0)	18.2 (1.0)	19.9 (1.0)	20.0 (0.9)	18.5 (1.0)	21.2 (0.9)
Learned c	16.1 (1.0)	18.0 (1.0)	20.0 (1.0)	18.9 (1.0)	18.4 (1.0)	20.5 (0.9)
Uniform ($c = 1.1$)	10.6 (1.0)	10.9 (1.0)	11.5 (1.0)	10.9 (1.0)	11.0 (1.0)	12.1 (1.0)
Uniform ($c = 1.15$)	13.9 (1.0)	14.2 (1.0)	16.2 (1.0)	15.5 (1.0)	14.7 (1.0)	15.0 (1.0)
Uniform ($c = 1.25$)	19.6 (1.0)	19.5 (1.0)	20.7 (1.0)	20.8 (0.9)	20.7 (0.9)	20.7 (1.0)
Uniform ($c = 1.5$)	24.9 (1.0)	25.0 (1.0)	25.3 (1.0)	24.0 (1.0)	23.5 (1.0)	25.1 (1.0)
19 split points						
Learned	19.4 (1.0)	20.3 (0.9)	22.0 (1.0)	21.6 (1.0)	22.2 (0.9)	22.3 (0.9)
Learned c	17.1 (1.0)	18.1 (1.0)	19.2 (1.0)	19.1 (1.0)	19.2 (1.0)	21.0 (0.9)
Uniform ($c = 1.1$)	18.2 (1.0)	18.7 (1.0)	19.7 (1.0)	18.5 (1.0)	19.0 (1.0)	20.1 (0.9)
Uniform ($c = 1.15$)	24.0 (0.9)	23.0 (0.9)	24.5 (0.9)	23.7 (0.9)	23.5 (0.9)	24.0 (0.9)
Uniform ($c = 1.2$)	25.0 (0.9)	25.6 (0.9)	26.9 (0.9)	25.0 (0.9)	24.7 (1.0)	26.1 (0.9)
39 split points						
Learned	20.9 (1.0)	23.0 (0.9)	24.6 (1.0)	23.6 (0.9)	23.8 (0.9)	26.2 (0.9)
Learned c	16.2 (1.0)	17.6 (1.0)	19.7 (1.0)	20.3 (1.0)	19.2 (1.0)	21.0 (0.9)
Uniform ($c = 1.05$)	19.3 (1.0)	19.7 (0.9)	20.5 (1.0)	20.3 (1.0)	19.6 (0.9)	21.1 (0.9)
Uniform ($c = 1.1$)	**25.4 (1.0)**	**26.2 (0.9)**	**27.1 (1.0)**	**26.2 (1.0)**	**26.2 (0.9)**	**26.5 (0.9)**

Table 7: EDM Gaussian mixture: percentage reduction in mean KS at each budget and split schedule, for the learned allocation, Uniform-c, and Learned-c. Half-widths of two-sided 90% normal confidence intervals are in parentheses.

Allocation	100k	200k	500k	1M	2M	5M
9 split points						
Learned	12.5 (0.9)	18.0 (0.8)	19.0 (0.7)	19.3 (0.7)	19.2 (0.6)	18.0 (0.6)
Learned c	16.7 (0.8)	18.2 (0.8)	18.4 (0.7)	19.4 (0.7)	18.8 (0.6)	17.5 (0.6)
Uniform ($c = 1.1$)	8.3 (0.9)	8.5 (0.9)	8.7 (0.8)	8.0 (0.8)	7.2 (0.7)	6.6 (0.7)
Uniform ($c = 1.15$)	12.0 (0.8)	12.1 (0.8)	10.8 (0.8)	11.2 (0.8)	10.4 (0.7)	9.4 (0.7)
Uniform ($c = 1.25$)	16.9 (0.8)	16.9 (0.8)	15.6 (0.7)	15.7 (0.7)	14.5 (0.7)	13.3 (0.6)
Uniform ($c = 1.5$)	**22.8 (0.7)**	22.2 (0.7)	20.3 (0.7)	**20.3 (0.7)**	19.4 (0.6)	17.5 (0.6)
19 split points						
Learned	17.9 (0.8)	18.5 (0.7)	19.6 (0.7)	19.7 (0.7)	19.4 (0.6)	18.6 (0.6)
Learned c	16.2 (0.8)	17.2 (0.8)	17.8 (0.7)	18.8 (0.7)	17.3 (0.6)	17.0 (0.6)
Uniform ($c = 1.1$)	14.6 (0.8)	15.5 (0.8)	14.4 (0.7)	13.8 (0.7)	13.3 (0.7)	11.8 (0.7)
Uniform ($c = 1.15$)	19.0 (0.8)	18.4 (0.8)	18.1 (0.7)	17.6 (0.7)	16.7 (0.7)	15.0 (0.6)
Uniform ($c = 1.2$)	21.8 (0.7)	21.1 (0.7)	20.3 (0.7)	19.6 (0.7)	18.8 (0.6)	17.1 (0.6)
39 split points						
Learned	18.0 (0.8)	19.5 (0.7)	19.6 (0.7)	20.3 (0.7)	19.6 (0.6)	**18.8 (0.6)**
Learned c	15.9 (0.8)	17.4 (0.8)	18.5 (0.7)	19.2 (0.7)	18.3 (0.6)	17.8 (0.6)
Uniform ($c = 1.05$)	15.1 (0.8)	15.4 (0.8)	14.8 (0.7)	14.1 (0.7)	13.6 (0.7)	12.4 (0.6)
Uniform ($c = 1.1$)	21.5 (0.7)	**22.4 (0.7)**	**21.1 (0.7)**	20.2 (0.7)	**19.7 (0.6)**	17.3 (0.6)

Table 8: CIFAR-10 DDPM: percentage reduction in mean MMD at each budget and split schedule, for the learned allocation and Learned-c. Half-widths of two-sided 90% normal confidence intervals are in parentheses.

Allocation	50k	100k	200k	500k
9 split points				
Learned	11.0 (2.3)	8.8 (2.3)	11.6 (1.7)	9.6 (2.5)
Learned c	12.5 (1.4)	12.7 (1.3)	11.3 (1.3)	**12.9 (1.4)**
19 split points				
Learned	**14.6 (1.4)**	12.7 (1.6)	11.0 (2.2)	11.9 (2.2)
Learned c	11.6 (1.6)	11.7 (1.5)	11.6 (1.5)	12.5 (2.1)
39 split points				
Learned	13.5 (1.6)	**12.8 (1.7)**	**11.9 (1.5)**	10.2 (2.3)
Learned c	11.4 (1.5)	12.1 (1.6)	11.4 (1.5)	10.6 (2.0)

Table 9: FFHQ LDM: percentage reduction in mean MMD at each budget and split schedule, for the learned allocation and Learned-c. Half-widths of two-sided 90% normal confidence intervals are in parentheses.

Allocation	50k	100k	200k	500k
9 split points				
Learned	9.8 (1.4)	**10.8 (1.5)**	10.9 (2.0)	10.7 (1.5)
Learned c	10.2 (1.4)	10.0 (1.5)	9.6 (1.7)	10.7 (1.7)
19 split points				
Learned	**11.4 (1.3)**	9.8 (1.8)	11.1 (1.6)	9.9 (1.7)
Learned c	9.7 (1.4)	10.2 (1.5)	**11.8 (1.4)**	**11.8 (1.5)**
39 split points				
Learned	10.6 (1.5)	9.3 (1.5)	9.4 (2.0)	9.9 (2.4)
Learned c	9.4 (1.6)	9.1 (1.4)	10.6 (1.4)	10.9 (1.3)