Safe Particle Selection for Diffusion Steering

Kianoosh Ashouritakllimi [1] **Iskander Azangulov** [1] **Leo Zhang** [1]

Abstract

We consider the problem particle selection for diffusion steering, where each current particle is branched, candidate children are evaluated by stochastic rollouts, and only a subset is propagated to the next denoising step. Standard Tweedie plug–in selection is computationally attractive, but for nonlinear rewards it can underestimate high–value ambiguous particles and prune them too early. We propose **Safe Particle Selection (SPS)**, which uses the Tweedie plug–in only as a proxy while making pruning decisions through anytime–valid confidence sequences built from e–processes. This yields finite–sample no–false-prune guarantees under adaptive rollout allocation, and the guarantee composes across the diffusion trajectory. Preliminary synthetic experiments show that our method better tracks the oracle, reduces cumulative regret, and retains more genuinely high–value particles than plug-in selection.

1. Introduction

Diffusion models generate samples through a sequence of denoising steps (Ho et al., 2020). In many applications, we want to steer this process toward samples with high terminal reward, such as an alignment score, classifier score, or physical objective (Singhal et al., 2025). A common approach is to maintain a population of particles, branch each particle at every denoising step, and keep only the most promising children for the next step. The central challenge is that the value of a child is not directly observed: it depends on the expected reward obtained after completing the rest of the diffusion trajectory.

Formally, after branching at time $t_{\ell+1}$, the ideal score for a child Y_a is its *look-ahead value*

$$\theta_a := \mathbb{E}\big[h(X_1) \mid X_{t_{\ell+1}} = Y_a\big] ,$$

where h is the terminal reward. Since θ_a is unknown, a common heuristic is to use Tweedie's formula (Efron, 2011) to estimate the denoised endpoint and then rank particles by the plug–in score

$$h\big(\mathbb{E}[X_1 \mid X_{t_{\ell+1}} = Y_a]\big) .$$

This can be misleading when h is nonlinear, because the reward of the posterior mean need not equal the posterior expected reward. Consequently, the Tweedie plug–in can discard particles whose continuations have high expected reward, especially early in denoising when uncertainty is large.

We propose **Safe Particle Selection (SPS)** for diffusion steering. Our method estimates each child value using stochastic rollouts, but eliminates a child only when an anytime–valid confidence sequence (Ramdas et al., 2023) certifies that it cannot belong to the top–N set. This gives a finite–sample no–false–prune guarantee: with high probability, no genuinely top–N child is discarded. The guarantee remains valid under adaptive rollout allocation and composes across the full diffusion trajectory.

We also use the Tweedie plug–in in a principled way. Rather than treating it as a reliable value estimate, we use it only as a *proxy* to guide betting and rollout allocation. Thus, the proxy can make selection faster when it is informative, but it does not affect the validity of the pruning certificate. Preliminary experiments on a synthetic diffusion task are promising: our approach more reliably tracks the oracle steering trajectory than the Tweedie plug–in baseline, avoiding premature elimination of high–value ambiguous particles and yielding lower end–to–end regret.

2. Background

2.1. Diffusion Models and Reward Steering

A diffusion model (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021) defines a Markov process on a time grid $0 = t_0 < \cdots < t_L = 1$ with transition kernels $K_\ell(\mathrm{d}x' \mid x)$ that map a tractable initial distribution to the data distribution at t_L. *Reward steering* (Singhal et al.,

[1]Department of Statistics, University of Oxford. Correspondence to: Kianoosh Ashouritaklimi <kianoosh.ashouritaklimi@stats.ox.ac.uk>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

2025) biases generation toward high values of a bounded[1] terminal score $h : \mathcal{X} \to [0,1]$ encoding the user objective (e.g. $h = r$ for a normalised reward, $h = \mathbf{1}\{r \geq u\}$ for tail events, or a tilted potential $h = \phi \circ r$.

2.2. Particle–Based Selection

Particle steering (Doucet et al., 2001; Skreta et al., 2025; Del Moral et al., 2006) carries N particles $\{X_i^\ell\}_{i=1}^N$ at time t_ℓ, branches each into K children $Y_{i,k}^\ell \sim K_\ell(\cdot \mid X_i^\ell)$, and selects N out of the $M = NK$ children to propagate to $t_{\ell+1}$. Flattening $a = (i,k)$, write the realised child pool as $Y_1, \ldots, Y_M$. The *look–ahead value* of child a is

$$\theta_a \;:=\; \mathbb{E}\big[h(X_1) \,\big|\, X_{t_{\ell+1}} = Y_a\big]\,, \tag{1}$$

and the oracle target is the top–N set $\mathcal{T}_N^\star := \arg\operatorname{top}_N\{\theta_1, \ldots, \theta_M\}$. Since θ_a is not available in closed form, it is estimated by stochastic rollouts: simulating a completion from Y_a to t_L yields $R_{a,j} = h(X_1^{a,j}) \in [0,1]$ with $\mathbb{E}[R_{a,j} \mid Y_{1:M}] = \theta_a$.

2.3. E-Processes and Confidence Sequences

A *confidence sequence* (CS) for θ is a sequence of intervals $(\widetilde{I}(n))_{n\geq 0}$ with time-uniform coverage

$$\mathbb{P}\Big(\forall n \geq 0 :\; \theta \in \widetilde{I}(n)\Big) \;\geq\; 1 - \alpha, \tag{2}$$

which remains valid at arbitrary, data-driven stopping times. CSs are obtained by inverting *e-processes* via Ville's inequality (Ville, 1939; Ramdas et al., 2023); betting–style martingales such as the hedged capital process of (Waudby-Smith & Ramdas, 2024) are variance–adaptive, parameter–free, and state of the art for bounded means. We build upon this betting strategy in §3.

3. Safe Particle Selection via Anytime–Valid Confidence Sequences

While the Tweedie plug–in can be a useful computational signal, but it cannot itself certify which particles are safe to keep or discard. We therefore combine stochastic look–ahead rollouts with anytime–valid confidence sequences. The plug–in guides betting and rollout allocation, while all pruning decisions are made only from valid confidence bounds.

3.1. Top–N Selection as Pure Exploration

Fix a diffusion step ℓ and condition on the realised child pool $Y_{1:M} = (Y_1, \ldots, Y_M)$, where $M = NK$. Each child has unknown look-ahead value

$$\theta_a := \mathbb{E}\big[h(X_1) \mid X_{t_{\ell+1}} = Y_a\big]\,, \qquad a = 1, \ldots, M,$$

and the oracle would retain the top–N set $\mathcal{T}_N^\star = \arg\operatorname{top}_N\{\theta_1, \ldots, \theta_M\}$. Conditional on $Y_{1:M}$, the values θ_a are fixed but unknown numbers in $[0,1]$. Pulling arm a means completing an independent diffusion rollout from Y_a and observing $R_{a,j} = h(X_1^{a,j}) \in [0,1]$ with mean θ_a. Thus, at each diffusion step, safe particle selection is a top-N pure-exploration bandit problem (Bubeck et al., 2013)

Because rollouts are expensive, the sampling rule should be adaptive. Let A_m be the arm sampled at inner round m, let R_m be the observed reward, and define

$$\mathcal{F}_m := \sigma(Y_{1:M}, A_1, R_1, \ldots, A_m, R_m)\,, \tag{3}$$

$$N_a(m) := \sum_{s=1}^{m} \mathbf{1}\{A_s = a\}. \tag{4}$$

We assume predictable allocation, unbiased rollouts, and i.i.d. conditional samples, as stated formally in Appendix E. For each arm we construct a per–arm confidence sequence $\widetilde{I}_a(n)$ for θ_a at level α_a. The global-clock interval is then

$$I_a(m) := \widetilde{I}_a(N_a(m)) = [L_a(m), U_a(m)].$$

By the standard time-change argument in Appendix D,

$$\mathbb{P}(\forall m \geq 0 : \theta_a \in I_a(m) \,|\, Y_{1:M}) \geq 1 - \alpha_a, \tag{5}$$

for any predictable rollout allocation.

3.2. Tweedie–Guided Confidence Sequences

For each child, Tweedie's formula gives the cheap proxy

$$\widehat{x}_{0,a} = \mathbb{E}\big[X_1 \mid X_{t_{\ell+1}} = Y_a\big]\,, \qquad p_a := h(\widehat{x}_{0,a}). \tag{6}$$

When h is nonlinear, p_a need not equal $\theta_a = \mathbb{E}[h(X_1) \mid X_{t_{\ell+1}} = Y_a]$, so it cannot be used as a pruning certificate. However, since p_a is measurable with respect to the realised child pool and does not depend on rollout outcomes, it can safely be used as side information to guide computation.

We use this side information inside a hedged betting e–process. Fix a candidate value $q \in (0,1)$ and predictable bets $\lambda_{a,n}^{\pm}(q)$ satisfying $\lambda_{a,n}^{+}(q) \in [0, 1/q]$ and $\lambda_{a,n}^{-}(q) \in [0, 1/(1-q)]$. Define the one-sided capital processes

$$K_{a,n}^{\pm}(q) := \prod_{i=1}^{n} \big(1 \pm \lambda_{a,i}^{\pm}(q)(R_{a,i} - q)\big)\,,$$

and combine them using a Tweedie–dependent hedge weight:

$$K_{a,n}(q) := w_a(q) K_{a,n}^{+}(q) + \big(1 - w_a(q)\big) K_{a,n}^{-}(q), \tag{7}$$

$$K_{a,0}(q) = 1, \tag{8}$$

[1] While unbounded terminal scores exist, we focus on the bounded case which is common in many reward steering applications (Singhal et al., 2025; Grünwald et al., 2024; Skreta et al., 2025)

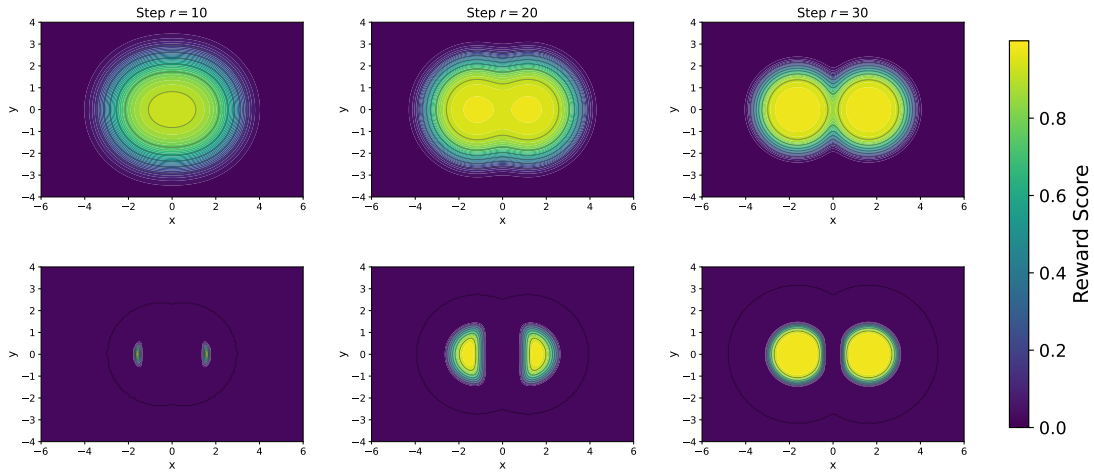

Figure 1. Comparison of the oracle posterior expected reward and the Tweedie plug-in reward across denoising steps. The top row shows the true look-ahead value $\theta_r(a) = \mathbb{E}[h(X^\star) \mid X_r = a]$, while the bottom row shows the plug-in score $h(\mathbb{E}[X^\star \mid X_r = a])$. Early in denoising, the plug-in score underestimates high-value ambiguous regions, illustrating how naive Tweedie-based selection can prematurely discard promising particles.

where

$$w_a(q) := \varepsilon + (1 - 2\varepsilon)\,\sigma(\gamma(p_a - q))\,,\ \varepsilon \in (0, \tfrac{1}{2}),\ \gamma \geq 0. \tag{9}$$

If $p_a > q$, more capital is placed on the upper–side bet; if $p_a < q$, more capital is placed on the lower–side bet. The parameters ε and γ control the minimum hedge in each direction and the sharpness of the proxy weighting. Setting $\gamma = 0$ recovers the symmetric hedged construction of (Waudby-Smith & Ramdas, 2024).

Lemma 1 (Validity of the Tweedie-guided hedge)**.** *Under Assumption 6, conditional on $Y_{1:M}$ and under $H_0 : \theta_a = q$, the process $(K_{a,n}(q))_{n\geq 0}$ defined in (8)–(9) is a non–negative $\mathcal{G}_{a,n}$-martingale with initial value one.*

The key point is that $w_a(q)$ depends only on $Y_{1:M}$ and q, not on the rollout sequence. Therefore the proxy can change how quickly the e–process grows, but it cannot invalidate the martingale property. By Ville's inequality, inverting the e–process over q gives the per–arm confidence sequence

$$\widetilde{I}_a(n) := \{q \in [0,1] : K_{a,n}(q) < 1/\alpha_a\}\,. \tag{10}$$

Combined with the time change above, this yields the global intervals $I_a(m) = [L_a(m), U_a(m)]$ satisfying (5). Details of the predictable plug–in bets, grid implementation, and proxy–regularised allocation score are given in §E.

3.3. Safe Elimination and Whole–Run Validity

At a fixed diffusion step, choose per–arm levels $\alpha_1, \ldots, \alpha_M$ with $\sum_{a=1}^{M} \alpha_a \leq \delta_\ell$. Define the simultaneous coverage event

$$\mathcal{E}_\ell := \bigcap_{a=1}^{M} \{\forall m \geq 0 : \theta_a \in I_a(m)\}\,.$$

By a union bound and (5),

$$\mathbb{P}(\mathcal{E}_\ell \mid Y_{1:M}) \geq 1 - \delta_\ell. \tag{11}$$

After m rollouts, let

$$S_m := \arg\operatorname{top}_N\{L_1(m), \ldots, L_M(m)\}, \tag{12}$$

$$\ell_m := \min_{a \in S_m} L_a(m), \tag{13}$$

$$u_m := \max_{b \notin S_m} U_b(m), \tag{14}$$

with deterministic tie–breaking. We eliminate a child $b \notin S_m$ only when

$$U_b(m) < \ell_m. \tag{15}$$

This is the main pruning rule used by our method.

Proposition 2 (Safe elimination)**.** *On $\mathcal{E}_\ell$, any child eliminated by (15) is not in the oracle top–N set $\mathcal{T}_N^\star$.*

Theorem 3 (Exact identification under separation)**.** *On $\mathcal{E}_\ell$, if $\ell_m > u_m$ at some round m, then $S_m = \mathcal{T}_N^\star$.*

Corollary 4 (No–false–prune guarantee)**.** *On $\mathcal{E}_\ell$, no child in $\mathcal{T}_N^\star$ is ever pruned by the rule (15).*

Thus the method can adaptively concentrate rollouts on ambiguous children without compromising validity. Across diffusion steps, choose stagewise error budgets $\delta_0, \ldots, \delta_{L-1}$ with $\sum_{\ell=0}^{L-1} \delta_\ell \leq \delta$ and set, for example, $\alpha_a = \delta_\ell / M$ at step ℓ.

Theorem 5 (Whole–run validity)**.** *Let $Y_{1:M}^{(\ell)}$ be the realised child pool at diffusion step ℓ, and let $\mathcal{H}_\ell$ be the full history before that step. If*

$$\mathbb{P}\Big(\mathcal{E}_\ell \mid \mathcal{H}_\ell, Y_{1:M}^{(\ell)}\Big) \geq 1 - \delta_\ell \quad \textit{a.s.}$$

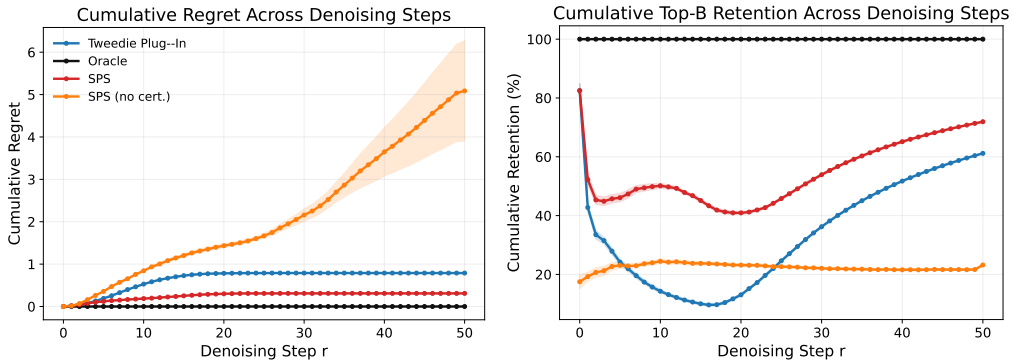

Figure 2. Cumulative regret and cumulative top–B retention across denoising steps. Safe Particle Selection (SPS) tracks the oracle more closely than the Tweedie plug-in baseline, achieving lower cumulative regret while retaining a larger fraction of the true top-B particles throughout the diffusion trajectory.

for every ℓ, and $\sum_{\ell=0}^{L-1} \delta_\ell \le \delta$, then

$$\mathbb{P}\left(\bigcap_{\ell=0}^{L-1} \mathcal{E}_\ell\right) \ge 1-\delta.$$

Consequently, with probability at least $1-\delta$, every pruning certificate, no-false-prune statement, and exact–identification statement holds simultaneously across the full diffusion trajectory. The Tweedie proxy may improve computational efficiency, but it never enters as evidence in the safety certificate.

4. Experiments

Synthetic diffusion task. We evaluate our approach on a 2D synthetic diffusion task. The clean endpoint $X^\star \in \mathbb{R}^2$ is drawn from a 3–component Gaussian mixture with two narrow high–reward modes and a broad background component. We use a variance–preserving noising schedule (Song et al., 2021) over $T = 50$ steps and a two–mode reward $h : \mathbb{R}^2 \to [0, 1]$ with bandwidth $\eta = 0.1$, peaked at the two narrow modes. In this model both the posterior expected reward $\theta_r(a) = \mathbb{E}[h(X^\star) \mid X_r = a]$ and the exact posterior predictive transition $K_r(a, \cdot)$ are available in closed form (§C), so the oracle steering trajectory is a tractable reference.

Particle steering. All methods start from the same paired initial population $\mathcal{P}_0 = \{Y_{0,i}\}_{i=1}^M$, $Y_{0,i} \sim \mathcal{N}(0, I_2)$. At each step $r < T$, a method selects a parent set of size B; each selected parent is branched into $L = M/B$ children using the exact posterior predictive transition, so population size is constant across denoising steps. We use $M = 50$, $B = 10$, $L = 5$, and average over 20 independent trials with paired seeds; full hyperparameters are in §C.

Baselines. The **oracle** baseline selects the B particles with largest analytic $\theta_r(Y_{r,i})$ at every step; it is unavailable in realistic problems and is used only to define regret. The **Tweedie plug–in** baseline computes the exact posterior mean $\widehat{x}_{0,r}(a) = \mathbb{E}[X^\star \mid X_r = a]$ and selects the B particles with the largest plug–in score $h(\widehat{x}_{0,r}(a))$. Our method, **Safe Particle Selections (SPS)** treats each particle as an arm with unknown mean $\theta_r(Y_{r,i})$, draws unbiased samples $R_{r,i}^{(s)} = h(X_{r,i}^{\star(s)})$ with $X_{r,i}^{\star(s)} \sim p(X^\star \mid X_r = Y_{r,i})$, and certifies selection by the anytime–valid CS $[L_{r,i}, U_{r,i}]$ of §3. We also run a **no–certificate** ablation that uses the same rollout samples and proxy–regularised scores but disables certified elimination.

Metric. We report cumulative end–to–end regret $\mathrm{CumRegret}_r^{(m)} = \sum_{s=0}^{r}(V_s^\star - V_s^{(m)})$, where $V_r^{(m)} = B^{-1}\sum_{i\in S_r^{(m)}} \theta_r(Y_{r,i}^{(m)})$ is the average analytic look–ahead value of method m's selected set and $V_r^\star$ is the same quantity for the oracle run on paired seeds. Lower regret means closer tracking of the oracle trajectory; the oracle has zero regret by definition. In addition to cumulative regret, we report the cumulative top–B retained particles: the number of particles among the true top–B set that remain available after each selection step. This metric measures whether a method preserves promising particles throughout the diffusion trajectory, rather than prematurely discarding candidates that may become high–value after further denoising.

Results. Figure 1 illustrates a failure mode of the naive Tweedie plug–in approach. Early in the diffusion process, the plug–in estimate can underestimate the reward of particles that would later become promising, causing the method to discard them without accounting for uncertainty.

Figure 2 reports cumulative regret and cumulative top–B retention across diffusion steps. Our method outperforms the baselines on both metrics: it achieves substantially lower regret while retaining more of the true top–B particles. This shows that safe elimination allows the method to track the oracle more closely than existing approaches by pruning

only when there is statistical evidence that a particle is unlikely to belong to the top set.

References

Bubeck, S., Wang, T., and Viswanathan, N. Multiple identifications in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 258–265, 2013.

Del Moral, P., Doucet, A., and Jasra, A. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B*, 68(3):411–436, 2006.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Doucet, A., de Freitas, N., and Gordon, N. (eds.). *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

Efron, B. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

Even-Dar, E., Mannor, S., and Mansour, Y. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.

Grünwald, P., de Heide, R., and Koolen, W. M. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 2024. doi: 10.1093/jrsssb/qkae011.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

Kaufmann, E., Cappé, O., and Garivier, A. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1): 1–42, 2016.

Kim, S., Kim, M., and Park, D. Test-time alignment of diffusion models without reward over-optimization. In *International Conference on Learning Representations*, 2025.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

Singhal, R., Horvitz, Z., Teehan, R., Ren, M., Yu, Z., McKeown, K., and Ranganath, R. A general framework for inference-time scaling and steering of diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55810–55827, 2025.

Skreta, M., Akhound-Sadegh, T., Ohanesian, V., Bondesan, R., Aspuru-Guzik, A., Doucet, A., Brekelmans, R., Tong, A., and Neklyudov, K. Feynman-kac correctors in diffusion: Annealing, guidance, and product of experts. *arXiv preprint arXiv:2503.02819*, 2025.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265. PMLR, 2015.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Ville, J. *Étude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

Waudby-Smith, I. and Ramdas, A. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B*, 86(1):1–27, 2024.

Appendix Contents

A Further Discussions **8**

B Related Work **9**

C Experimental Setup **10**

D Proofs **12**

D.1 Notation, Ties, and Rollout Model 12

D.2 Per–Arm Confidence Sequences and Adaptive Time Change 12

D.3 Tweedie–Guided Betting Confidence Sequences 13

D.4 Stagewise Simultaneous Coverage 14

D.5 Safe Elimination and Exact Identification 14

D.6 Validity of the Tweedie Proxy 15

D.7 Whole–Run Validity Across Diffusion Steps 15

D.8 Closed–Form Oracle Quantities in the Synthetic Experiment 16

E Confidence–Sequence Construction for Safe Particle Selection **17**

E.1 One Diffusion Step as a Bounded–Mean Bandit Problem 17

E.2 Tweedie Proxy and Hedge Weights 17

E.3 One–Sided Betting Capital Processes 17

E.4 Predictable Plug–In Bets Used in the Experiments 18

E.5 Numerical Inversion 18

E.6 Proxy–Regularised Allocation 19

E.7 Certified Pruning and Stopping 19

E.8 Stagewise and Whole–Run Error Budgets 20

A. Further Discussions

Safe Particle Selection (SPS) improves diffusion steering by replacing naive Tweedie plug–in selection with anytime–valid confidence sequences, preventing the premature pruning of high–value particles. Preliminary experiments confirm that SPS tracks the oracle more accurately and reduces end–to–end regret compared to the commonly used baseline.

Building on this framework, future research could pursue two main directions. First, scaling to high–dimensional modalities requires testing SPS on large–scale diffusion models (e.g., image or text generation) using complex, real–world terminal rewards. Second, improving rollout efficiency is critical due to the computational cost of stochastic look–ahead rollouts. Future work could explore safely truncating rollouts or integrating learned value–function critics without compromising the method's safety guarantees.

B. Related Work

Diffusion guidance and reward steering. Diffusion models generate samples through a gradual denoising process (Ho et al., 2020; Song et al., 2021). A large body of work steers this process by modifying the reverse dynamics using class information, classifier gradients, or classifier-free guidance (Dhariwal & Nichol, 2021; Ho & Salimans, 2022). More recent inference–time alignment methods use rewards without retraining, often through particle–based or Sequential Monte Carlo (SMC) schemes (Kim et al., 2025; Singhal et al., 2025; Skreta et al., 2025). Our work is complementary: rather than proposing a new reward–tilted sampler, we study the local particle–selection problem that arises during steering and provide statistical certificates that high–value particles are not prematurely discarded.

Particle methods and top–N identification. SMC methods propagate and resample interacting particles to approximate evolving target distributions (Doucet et al., 2001; Del Moral et al., 2006). In our setting, each diffusion step induces a finite population of candidate children, and selecting the best subset from stochastic rollout rewards is naturally a top–N pure–exploration problem. This connects to best–arm and multiple–arm identification in bandits, where confidence bounds are used to adaptively allocate samples and eliminate suboptimal arms (Even-Dar et al., 2006; Bubeck et al., 2013; Kaufmann et al., 2016). Unlike standard bandit identification, our arms are transient diffusion particles, and the selection guarantees must remain valid under adaptive rollout allocation and across the full denoising trajectory.

Anytime–valid inference and Tweedie proxies. Confidence sequences provide time–uniform uncertainty quantification and remain valid under data–dependent stopping rules (Ville, 1939; Howard et al., 2021). Recent betting–based constructions give sharp confidence sequences for bounded means via nonnegative martingales and e–processes (Ramdas et al., 2023; Waudby-Smith & Ramdas, 2024). We use these tools to certify safe elimination decisions. Tweedie's formula gives a posterior–mean reconstruction and is closely connected to empirical Bayes estimation (Efron, 2011); in diffusion models, this type of denoising estimate is often useful as a cheap proxy. However, for nonlinear rewards, $h(\mathbb{E}[X_0 \mid X_t])$ need not equal $\mathbb{E}[h(X_0) \mid X_t]$. Our method therefore uses the Tweedie plug–in only to guide computation, while all pruning decisions are based on anytime–valid confidence bounds.

C. Experimental Setup

Mixture and noising schedule. The clean endpoint $X^\star \in \mathbb{R}^2$ is drawn from a 3–component Gaussian mixture

$$p_\star(x) = \sum_{k=1}^{3} \pi_k \mathcal{N}(x; \mu_k, s_k^2 I_2), \tag{16}$$

with $(\pi_1, \pi_2, \pi_3) = (0.3, 0.3, 0.4)$, means $\mu_1 = (-2, 0)$, $\mu_2 = (2, 0)$, $\mu_3 = (0, 0)$, and standard deviations $(s_1, s_2, s_3) = (0.01, 0.01, 10)$. The first two components are narrow high–reward modes; the third is a broad background component that creates ambiguous particles during early denoising. For denoising steps $r = 0, \ldots, T$ with $T = 50$, we use the variance–preserving schedule (Song et al., 2021)

$$X_r = \alpha_r X^\star + \sigma_r \epsilon_r, \quad \epsilon_r \sim \mathcal{N}(0, I_2), \quad \alpha_r = \sin\left(\tfrac{\pi r}{2T}\right), \quad \sigma_r = \cos\left(\tfrac{\pi r}{2T}\right), \tag{17}$$

so $r = 0$ is pure noise and $r = T$ is clean. The reward has two modes at $\nu_1 = (-2, 0)$, $\nu_2 = (2, 0)$:

$$h(x) = \frac{1}{Z} \sum_{\ell=1}^{2} \exp\{-\|x - \nu_\ell\|_2^2/(2\eta^2)\}, \quad Z = 1 + \exp\{-\|\nu_1 - \nu_2\|_2^2/(2\eta^2)\}, \tag{18}$$

with bandwidth $\eta = 0.1$, normalised so the reward equals one at the two target modes.

Analytic oracle values and exact transitions. For a particle $a = X_r$, the posterior expected reward $\theta_r(a) = \mathbb{E}[h(X^\star) \mid X_r = a]$ is available in closed form. For $r < T$, let

$$\rho_{r,k}(a) \propto \pi_k \, \phi_2(a; \alpha_r \mu_k, v_{r,k} I_2), \quad v_{r,k} = \alpha_r^2 s_k^2 + \sigma_r^2 \tag{19}$$

denote the *posterior responsibility* of component k, and let

$$c_{r,k} = \left(s_k^{-2} + \alpha_r^2 \sigma_r^{-2}\right)^{-1}, \quad m_{r,k}(a) = c_{r,k}\left(s_k^{-2}\mu_k + \alpha_r \sigma_r^{-2} a\right) \tag{20}$$

be the posterior variance and mean of $X^\star$ conditional on component k. Then

$$\theta_r(a) = \frac{1}{Z} \sum_{k=1}^{3} \sum_{\ell=1}^{2} \rho_{r,k}(a) \frac{\eta^2}{c_{r,k} + \eta^2} \exp\left\{-\frac{\|m_{r,k}(a) - \nu_\ell\|_2^2}{2(c_{r,k} + \eta^2)}\right\}, \tag{21}$$

and at the clean step $\theta_T(a) = h(a)$. The Tweedie posterior mean used by the plug–in baseline is $\widehat{x}_{0,r}(a) = \sum_{k=1}^{3} \rho_{r,k}(a) m_{r,k}(a)$. Selected particles are propagated with the exact posterior predictive transition

$$K_r(a, \cdot) = \sum_{k=1}^{3} \rho_{r,k}(a) \, \mathcal{N}\big(\alpha_{r+1} m_{r,k}(a), \, (\alpha_{r+1}^2 c_{r,k} + \sigma_{r+1}^2) I_2\big), \tag{22}$$

which preserves the correct marginal at the next denoising level.

Hyperparameters. Below, we describe the hyperparameters used for our experiment:

- Population size $M = 50$, branch size $B = 10$, branching factor $L = M/B = 5$, $T = 50$ denoising steps, 20 independent trials.
- The CS hyperparameters are: total error budget $\delta = 0.1$, grid size $G = 200$, betting truncation $c = 1/2$, proxy strength $\tau = 5$, hedge parameters $\epsilon = 0.1$ and $\gamma = 10$. See also § E.4.
- The rollout–budget schedule $\{B_r^{\text{roll}}\}_{r=0}^{T-1}$ is fixed before the trial.

For our budget schedule, we use a deterministic uncertainty–annealed schedule. For each $r \in \{0, \ldots, T-1\}$, let

$$\alpha_r = \sin\left(\frac{\pi r}{2T}\right), \qquad \sigma_r = \cos\left(\frac{\pi r}{2T}\right), \qquad \omega_r = \alpha_r^2 \left(\sigma_r^2\right)^\kappa.$$

We normalize

$$\bar{\omega}_r = \frac{\omega_r}{\max_{s\in\{0,\dots,T-1\}} \omega_s},$$

set

$$e_r = \mathrm{round}\Big(\texttt{peak_extra_budget} \cdot \bar{\omega}_r\Big),$$

and define

$$B_r^{\mathrm{roll}} = \begin{cases} 0, & e_r < \texttt{min_extra_budget}, \\ M\, n_{\mathrm{init}} + e_r, & \text{otherwise}. \end{cases}$$

In our experiment, we use

$$n_{\mathrm{init}} = 10, \quad \texttt{peak_extra_budget} = 2000, \quad \kappa = 10, \quad \texttt{min_extra_budget} = 50.$$

This allocates most rollouts to early/intermediate steps, where X_r is informative but $p(X_0 \mid X_r)$ is still diffuse (so the plug–in bias can be large), and allocates little or no budget near $r = 0$ (pure noise) or near $r \approx T$ (where the posterior is concentrated, and the plug–in already reliable).

D. Proofs

Here, we provide the proofs for the safe particle selection part of our approach. The arguments are standard consequences of nonnegative martingales, Ville's inequality, and confidence–sequence inversion, but are written here in the notation of the paper; see (Ville, 1939), (Howard et al., 2021) for the general theory.

D.1. Notation, Ties, and Rollout Model

Fix one outer diffusion step. Conditional on the realised child pool $Y_{1:M} = (Y_1, \ldots, Y_M)$, the look–ahead value of child a is

$$\theta_a := \mathbb{E}[h(X_1) \mid X_{t_{\ell+1}} = Y_a] \in [0, 1].$$

We use a deterministic tie–breaking order on $[M]$. The oracle top–N set $T_N^\star$ is the set of the first N children after sorting by decreasing θ_a and then by this fixed tie–breaking order. When the inequalities below are strict, the tie–breaking convention is irrelevant.

For each child a, let $(R_{a,j})_{j\geq 1}$ denote the child–specific rollout rewards in their order of arrival. The global sampling rule selects a child A_m at inner round m and observes

$$R_m = R_{A_m, N_{A_m}(m)}, \qquad N_a(m) := \sum_{s=1}^{m} \mathbf{1}\{A_s = a\}.$$

The global and child–specific filtrations are

$$\mathcal{F}_m := \sigma(Y_{1:M}, A_1, R_1, \ldots, A_m, R_m), \qquad \mathcal{G}_{a,n} := \sigma(Y_{1:M}, R_{a,1}, \ldots, R_{a,n}).$$

Assumption 6 (Predictable, unbiased, fresh rollouts)**.** *Conditional on $Y_{1:M}$, the following hold.*

1. *A_m is $\mathcal{F}_{m-1}$–measurable for every $m \geq 1$.*
2. *For every child a and every $j \geq 1$, $R_{a,j} \in [0, 1]$ and $\mathbb{E}[R_{a,j} \mid Y_{1:M}] = \theta_a$.*
3. *For each fixed a, $(R_{a,j})_{j\geq 1}$ is conditionally i.i.d.; rollout sequences for distinct children are conditionally independent.*

A consequence of the above is that the realised reward at the global clock satisfies

$$\mathbb{E}[R_m \mid \mathcal{F}_{m-1}, A_m = a, Y_{1:M}] = \theta_a.$$

D.2. Per–Arm Confidence Sequences and Adaptive Time Change

Definition 7 (Per–arm confidence sequence)**.** A sequence of intervals $(\widetilde{I}_a(n))_{n\geq 0}$ is a level–α_a confidence sequence for θ_a on $(\mathcal{G}_{a,n})_{n\geq 0}$ if

$$\mathbb{P}\Big(\forall n \geq 0 : \ \theta_a \in \widetilde{I}_a(n) \ \Big| \ Y_{1:M}\Big) \geq 1 - \alpha_a.$$

The global–clock interval is

$$I_a(m) := \widetilde{I}_a(N_a(m)) = [L_a(m), U_a(m)].$$

Lemma 8 (Adaptive time change)**.** *Under Assumption 6, if $(\widetilde{I}_a(n))_{n\geq 0}$ is a level-α_a per–arm confidence sequence, then*

$$\mathbb{P}\Big(\forall m \geq 0 : \ \theta_a \in I_a(m) \ \Big| \ Y_{1:M}\Big) \geq 1 - \alpha_a.$$

The statement is valid for any predictable rollout allocation rule.

Proof. On the event $\{\forall n \geq 0 : \theta_a \in \widetilde{I}_a(n)\}$, every realised integer $N_a(m) \in \{0, 1, 2, \ldots\}$ also satisfies $\theta_a \in \widetilde{I}_a(N_a(m)) = I_a(m)$. Hence

$$\{\exists m \geq 0 : \theta_a \notin I_a(m)\} \subseteq \{\exists n \geq 0 : \theta_a \notin \widetilde{I}_a(n)\}.$$

Taking conditional probabilities gives the claim. The predictability and fresh–rollout assumptions ensure that the per–arm sequence used to construct $\widetilde{I}_a$ has the stated conditional distribution even though the global sampling rule is adaptive. □

D.3. Tweedie–Guided Betting Confidence Sequences

For a candidate mean $q \in (0,1)$ and child a, define predictable bets $\lambda^{+}_{a,i}(q)$ and $\lambda^{-}_{a,i}(q)$ satisfying

$$\lambda^{+}_{a,i}(q) \in [0, 1/q], \qquad \lambda^{-}_{a,i}(q) \in [0, 1/(1-q)],$$

where $\lambda^{\pm}_{a,i}(q)$ is $\mathcal{G}_{a,i-1}$–measurable. The one–sided capital processes are

$$K^{+}_{a,n}(q) := \prod_{i=1}^{n} \big(1 + \lambda^{+}_{a,i}(q)(R_{a,i} - q)\big), \qquad K^{-}_{a,n}(q) := \prod_{i=1}^{n} \big(1 - \lambda^{-}_{a,i}(q)(R_{a,i} - q)\big).$$

The constraints on $\lambda^{\pm}_{a,i}$ imply nonnegative multiplicative increments for every $R_{a,i} \in [0,1]$.

Let $p_a = h(\widehat{x}_{0,a})$ be the Tweedie plug–in proxy, where $\widehat{x}_{0,a} = \mathbb{E}[X_1 \mid X_{t_{\ell+1}} = Y_a]$. For fixed $\varepsilon \in (0, 1/2)$ and $\gamma \geq 0$, set

$$w_a(q) := \varepsilon + (1 - 2\varepsilon)\, \frac{1}{1 + \exp\{-\gamma(p_a - q)\}} \in [\varepsilon, 1 - \varepsilon]$$

and define the hedged capital

$$K_{a,n}(q) := w_a(q) K^{+}_{a,n}(q) + \big(1 - w_a(q)\big) K^{-}_{a,n}(q).$$

The proxy p_a is measurable with respect to $Y_{1:M}$ and is fixed before any rollout rewards from the current child pool are observed.

Lemma 9 (Validity of the Tweedie–guided hedge)**.** *Under Assumption 6, conditional on $Y_{1:M}$ and under $H_0 : \theta_a = q$, the process $(K_{a,n}(q))_{n \geq 0}$ is a nonnegative $\mathcal{G}_{a,n}$–martingale with $K_{a,0}(q) = 1$.*

Proof. Under $H_0 : \theta_a = q$,

$$\mathbb{E}[R_{a,i} - q \mid \mathcal{G}_{a,i-1}, Y_{1:M}] = 0.$$

Since $\lambda^{\pm}_{a,i}(q)$ are predictable,

$$\mathbb{E}\big[1 + \lambda^{+}_{a,i}(q)(R_{a,i} - q) \mid \mathcal{G}_{a,i-1}, Y_{1:M}\big] = 1,$$
$$\mathbb{E}\big[1 - \lambda^{-}_{a,i}(q)(R_{a,i} - q) \mid \mathcal{G}_{a,i-1}, Y_{1:M}\big] = 1.$$

Together with the nonnegativity of the factors, this proves that $(K^{+}_{a,n}(q))_{n \geq 0}$ and $(K^{-}_{a,n}(q))_{n \geq 0}$ are nonnegative martingales. The weight $w_a(q)$ is a deterministic function of $Y_{1:M}$ and q, so it is fixed under the conditional null. Therefore $K_{a,n}(q)$ is a fixed convex combination of two nonnegative martingales and is itself a nonnegative martingale starting at one. □

Proposition 10 (Betting confidence sequence)**.** *For $\alpha_a \in (0,1)$, define the continuous–inversion confidence set*

$$\widetilde{I}_a(n) := \mathrm{cl}\Big\{ q \in (0,1) : K_{a,n}(q) < 1/\alpha_a \Big\},$$

where cl *denotes closure in* $[0,1]$*. Then $(\widetilde{I}_a(n))_{n \geq 0}$ is a level–α_a confidence sequence for θ_a.*

Proof. Fix $q \in (0,1)$. By Lemma 9 and Ville's inequality,

$$\mathbb{P}_{\theta_a = q}\Big(\sup_{n \geq 0} K_{a,n}(q) \geq 1/\alpha_a \;\Big|\; Y_{1:M} \Big) \leq \alpha_a.$$

Taking $q = \theta_a$ shows that, with conditional probability at least $1 - \alpha_a$, $K_{a,n}(\theta_a) < 1/\alpha_a$ for every n; equivalently, $\theta_a \in \widetilde{I}_a(n)$ for every n. The closure convention includes endpoint means by the corresponding one–sided limits. □

Remark 11. Note that the set $\widetilde{I}_a(n)$ need not be connected. In practise, we use only its convex hull

$$\widetilde{A}_a(n) \;:=\; [\, L_a(n), U_a(n) \,], \qquad L_a(n) := \inf \widetilde{I}_a(n), \quad U_a(n) := \sup \widetilde{I}_a(n),$$

with the convention $\inf \emptyset = 1$, $\sup \emptyset = 0$. Since $\widetilde{I}_a(n) \subseteq \widetilde{A}_a(n)$, on the coverage event of Proposition 10 we have $\theta_a \in \widetilde{A}_a(n)$ for every n, so $\widetilde{A}_a(n)$ is itself a level-α_a confidence sequence for θ_a. All elimination statements in the paper use only the endpoints $L_a(m)$ and $U_b(m)$, so the safe-elimination and exact-identification results of §3.3 hold verbatim with $I_a(m)$ replaced by $\widetilde{A}_a(N_a(m))$.

D.4. Stagewise Simultaneous Coverage

At the fixed diffusion step, choose per–child error budgets $\alpha_1, \ldots, \alpha_M$ with $\sum_{a=1}^{M} \alpha_a \le \delta_\ell$ and define

$$E_\ell := \bigcap_{a=1}^{M} \{\forall m \ge 0 : \ \theta_a \in I_a(m)\}.$$

Lemma 12 (Simultaneous stagewise coverage)**.** *Under Assumption 6 and the betting construction above,*

$$\mathbb{P}(E_\ell \mid Y_{1:M}) \ge 1 - \delta_\ell.$$

The same bound holds conditionally on any outer history $\mathcal{H}_\ell$ before the current child pool is generated.

Proof. By Proposition 10 and Lemma 8, for each a,

$$\mathbb{P}\big(\exists m \ge 0 : \theta_a \notin I_a(m) \mid Y_{1:M}\big) \le \alpha_a.$$

A union bound over $a \in [M]$ gives

$$\mathbb{P}(E_\ell^c \mid Y_{1:M}) \le \sum_{a=1}^{M} \alpha_a \le \delta_\ell.$$

Conditioning additionally on $\mathcal{H}_\ell$ does not change the argument, because all statements are conditional on the realised child pool and on the history available before the inner rollouts begin. □

D.5. Safe Elimination and Exact Identification

At inner round m, let

$$S_m := \arg\operatorname{top}_N\{L_1(m), \ldots, L_M(m)\}$$

with the same deterministic tie–breaking convention as above, and define

$$\ell_m := \min_{a \in S_m} L_a(m), \qquad u_m := \max_{b \notin S_m} U_b(m).$$

A child $b \notin S_m$ is eliminated only if

$$U_b(m) < \ell_m.$$

Proposition 13 (Safe elimination)**.** *On E_ℓ, every child eliminated by the rule $U_b(m) < \ell_m$ is not in the oracle top–N set $T_N^\star$.*

Proof. Fix a round m and an eliminated child $b \notin S_m$. On E_ℓ, $\theta_b \le U_b(m)$. For every $a \in S_m$, again on E_ℓ, $\theta_a \ge L_a(m) \ge \ell_m$. The elimination condition gives

$$\theta_a \ge L_a(m) \ge \ell_m > U_b(m) \ge \theta_b, \qquad a \in S_m.$$

Thus there are N children whose look–ahead values are strictly larger than θ_b. Therefore b cannot be among the oracle top N children, no matter how ties are resolved. □

Theorem 14 (Exact identification under separation)**.** *On E_ℓ, if $\ell_m > u_m$ at some inner round m, then $S_m = T_N^\star$.*

Proof. For every $a \in S_m$ and every $b \notin S_m$, the event E_ℓ and the separation condition imply

$$\theta_a \ge L_a(m) \ge \min_{c \in S_m} L_c(m) = \ell_m > u_m = \max_{d \notin S_m} U_d(m) \ge U_b(m) \ge \theta_b.$$

Hence every child in S_m has value strictly larger than every child outside S_m. Since $|S_m| = N$, S_m is exactly the oracle top–N set. □

Corollary 15 (No false prune)**.** *On E_ℓ, no child in $T_N^\star$ is ever pruned by the safe–elimination rule during the diffusion step.*

Proof. Apply Proposition 13 at each inner round at which a pruning decision is made. The event E_ℓ is time–uniform, so the same coverage event holds simultaneously for all rounds. □

Corollary 16 (Stopping with N active children)**.** *Suppose the algorithm prunes only by the safe–elimination rule and stops when exactly N children remain active. On E_ℓ, the remaining active set is $T_N^\star$.*

Proof. By Corollary 15, all children in $T_N^\star$ remain active throughout the step. If only N active children remain and $|T_N^\star| = N$, the active set must equal $T_N^\star$. □

D.6. Validity of the Tweedie Proxy

The Tweedie plug–in proxy is used in two places: in the hedge weight $w_a(q)$ and in optional allocation scores. Neither use treats the proxy as rollout evidence.

Proposition 17 (Proxy–guided allocation preserves validity)**.** *Let p_a be any $\sigma(Y_{1:M})$–measurable proxy. Suppose the next rollout arm A_{m+1} is chosen as a measurable function of $(Y_{1:M}, A_1, R_1, \ldots, A_m, R_m)$ and the proxies $(p_1, \ldots, p_M)$. Then A_{m+1} is $\mathcal{F}_m$–measurable, and all coverage, safe–elimination, and exact–identification guarantees above remain valid.*

Proof. Because each p_a is $\sigma(Y_{1:M})$–measurable and $Y_{1:M} \subset \mathcal{F}_m$, any allocation rule based on the past data and the proxies is $\mathcal{F}_m$– measurable. Thus Assumption 6 is preserved. The martingale proof in Lemma 9 uses only conditional unbiasedness of the rollout rewards and predictability of the bets; it does not require the proxy to be accurate. Therefore the proxy can affect sampling efficiency and the rate at which capital grows, but not the validity of the confidence sequence or of any pruning certificate derived from it. □

D.7. Whole–Run Validity Across Diffusion Steps

Let E_ℓ denote the stagewise simultaneous coverage event at diffusion step ℓ, and let $\mathcal{H}_\ell$ be the full outer history before generating the child pool at that step. Let $Y_{1:M}^{(\ell)}$ be the realised child pool at step ℓ.

Theorem 18 (Whole–run validity)**.** *If, for every $\ell = 0, \ldots, L-1$,*

$$\mathbb{P}\big(E_\ell \mid \mathcal{H}_\ell, Y_{1:M}^{(\ell)}\big) \geq 1 - \delta_\ell \quad a.s.,$$

and $\sum_{\ell=0}^{L-1} \delta_\ell \leq \delta$, then

$$\mathbb{P}\Big(\bigcap_{\ell=0}^{L-1} E_\ell\Big) \geq 1 - \delta.$$

Consequently, with probability at least $1-\delta$, all stagewise pruning certificates, no–false–prune statements, and exact–identification statements hold simultaneously over the entire diffusion trajectory.

Proof. By the tower property,

$$\mathbb{P}(E_\ell^c) = \mathbb{E}\Big[\mathbb{P}\big(E_\ell^c \mid \mathcal{H}_\ell, Y_{1:M}^{(\ell)}\big)\Big] \leq \delta_\ell.$$

A union bound over diffusion steps yields

$$\mathbb{P}\Big(\bigcup_{\ell=0}^{L-1} E_\ell^c\Big) \leq \sum_{\ell=0}^{L-1} \mathbb{P}(E_\ell^c) \leq \sum_{\ell=0}^{L-1} \delta_\ell \leq \delta.$$

Taking complements proves the claim. No independence across diffusion steps is required. □

Remark 19 (Interpretation of whole–run validity)**.** The above result certifies the coverage events $\mathcal{E}_\ell$ for the child pools $Y_{1:M}^{(\ell)}$ that the algorithm *actually generates*. It is therefore a guarantee *along the realized trajectory*: every certified elimination is safe with respect to the look-ahead values θ_a defined relative to the realized $Y_{1:M}^{(\ell)}$.

D.8. Closed–Form Oracle Quantities in the Synthetic Experiment

This subsection derives the analytic quantities used only in the synthetic experiment. Let $Z \in \{1, 2, 3\}$ denote the mixture component, with $\mathbb{P}(Z = k) = \pi_k$ and

$$X^\star \mid Z = k \sim \mathcal{N}(\mu_k, s_k^2 I_2).$$

The noising model at step r is

$$X_r = \alpha_r X^\star + \sigma_r \varepsilon_r, \qquad \varepsilon_r \sim \mathcal{N}(0, I_2).$$

Proposition 20 (Posterior responsibilities, oracle values, and predictive kernel)**.** *For $a \in \mathbb{R}^2$ and $r < T$, define*

$$v_{r,k} := \alpha_r^2 s_k^2 + \sigma_r^2, \qquad \rho_{r,k}(a) := \frac{\pi_k \varphi_2(a; \alpha_r \mu_k, v_{r,k} I_2)}{\sum_{j=1}^3 \pi_j \varphi_2(a; \alpha_r \mu_j, v_{r,j} I_2)},$$

where $\varphi_2(\cdot; m, \Sigma)$ is the two–dimensional Gaussian density. Also define

$$c_{r,k} := \big(s_k^{-2} + \alpha_r^2 \sigma_r^{-2}\big)^{-1}, \qquad m_{r,k}(a) := c_{r,k}\big(s_k^{-2}\mu_k + \alpha_r \sigma_r^{-2} a\big).$$

Then

$$X^\star \mid (X_r = a, Z = k) \sim \mathcal{N}(m_{r,k}(a), c_{r,k} I_2), \qquad \mathbb{P}(Z = k \mid X_r = a) = \rho_{r,k}(a).$$

For the two–mode reward

$$h(x) = \frac{1}{Z_h} \sum_{j=1}^{2} \exp\{-\|x - \nu_j\|_2^2/(2\eta^2)\},$$

the oracle look–ahead value is

$$\theta_r(a) = \frac{1}{Z_h} \sum_{k=1}^{3} \rho_{r,k}(a) \sum_{j=1}^{2} \frac{\eta^2}{c_{r,k} + \eta^2} \exp\left\{-\frac{\|m_{r,k}(a) - \nu_j\|_2^2}{2(c_{r,k} + \eta^2)}\right\}.$$

The Tweedie posterior mean used by the plug–in baseline is

$$\widehat{x}_{0,r}(a) = \mathbb{E}[X^\star \mid X_r = a] = \sum_{k=1}^{3} \rho_{r,k}(a) m_{r,k}(a),$$

and the exact posterior predictive transition is the Gaussian mixture

$$K_r(a, \cdot) = \sum_{k=1}^{3} \rho_{r,k}(a) \mathcal{N}\big(\alpha_{r+1} m_{r,k}(a), (\alpha_{r+1}^2 c_{r,k} + \sigma_{r+1}^2) I_2\big).$$

At the clean step $r = T$, $\theta_T(a) = h(a)$.

Proof. Conditionally on $Z = k$, the marginal distribution of X_r is Gaussian with mean $\alpha_r \mu_k$ and covariance $(\alpha_r^2 s_k^2 + \sigma_r^2) I_2 = v_{r,k} I_2$. Bayes' rule therefore gives the responsibilities $\rho_{r,k}(a)$. The pair $(X^\star, X_r)$ is jointly Gaussian conditional on $Z = k$, and the standard Gaussian conditioning formula gives posterior covariance $(s_k^{-2} + \alpha_r^2 \sigma_r^{-2})^{-1} I_2 = c_{r,k} I_2$ and posterior mean $m_{r,k}(a)$.

For $X \sim \mathcal{N}(m, c I_2)$,

$$\mathbb{E} \exp\{-\|X - \nu\|_2^2/(2\eta^2)\} = \frac{\eta^2}{c + \eta^2} \exp\left\{-\frac{\|m - \nu\|_2^2}{2(c + \eta^2)}\right\},$$

which follows by completing the square in the Gaussian integral. Averaging this identity over the posterior mixture components yields the displayed formula for $\theta_r(a)$. The posterior mean is obtained by averaging the component means $m_{r,k}(a)$ with weights $\rho_{r,k}(a)$. Finally, conditional on $(X_r = a, Z = k)$, the next noised state satisfies $X_{r+1} = \alpha_{r+1} X^\star + \sigma_{r+1} \varepsilon_{r+1}$ with $X^\star \sim \mathcal{N}(m_{r,k}(a), c_{r,k} I_2)$ and independent $\varepsilon_{r+1} \sim \mathcal{N}(0, I_2)$, giving the stated predictive Gaussian component. Mixing over Z completes the proof. □

E. Confidence–Sequence Construction for Safe Particle Selection

This appendix gives the complete construction of the anytime–valid confidence sequences used by Safe Particle Selection. The construction is an inversion of nonnegative betting martingales for bounded means. It uses Ville's inequality and the standard confidence–sequence/e–process recipe (Ville, 1939; Ramdas et al., 2023; Howard et al., 2021). The Tweedie plug–in enters only as a predictable guide for hedging and allocation; it is not treated as statistical evidence. This separation is what makes the certificate robust even when the plug–in value is biased for the nonlinear reward (Efron, 2011).

E.1. One Diffusion Step as a Bounded–Mean Bandit Problem

Fix diffusion step ℓ and condition on the realised child pool $Y_{1:M} = (Y_1, \dots, Y_M)$. For child $a \in [M]$, a rollout completes the trajectory from Y_a to the terminal time and returns

$$R_{a,j} = h(X_1^{a,j}) \in [0,1], \qquad \mathbb{E}[R_{a,j} \mid Y_{1:M}] = \theta_a.$$

The unknown mean θ_a is the child's look–ahead value. At inner round m, the algorithm chooses A_m predictably from past rollouts and observes R_m. The global count is

$$N_a(m) = \sum_{s=1}^{m} \mathbf{1}\{A_s = a\},$$

and the per–child reward sequence is indexed by the order in which rewards from that child are observed:

$$\mathcal{G}_{a,n} = \sigma(Y_{1:M}, R_{a,1}, \dots, R_{a,n}).$$

The goal is to maintain intervals $I_a(m) = [L_a(m), U_a(m)]$ satisfying

$$\mathbb{P}\Big(\forall a \in [M],\ \forall m \geq 0 : \theta_a \in I_a(m) \ \Big|\ Y_{1:M}\Big) \geq 1 - \delta_\ell.$$

A convenient choice is $\alpha_a = \delta_\ell / M$ for all arms, although any allocation with $\sum_a \alpha_a \leq \delta_\ell$ is valid.

E.2. Tweedie Proxy and Hedge Weights

For each child, compute the posterior–mean reconstruction

$$\widehat{x}_{0,a} := \mathbb{E}[X_1 \mid X_{t_{\ell+1}} = Y_a]$$

using Tweedie's formula or the model's equivalent denoising estimator. Define the cheap proxy score

$$p_a := h(\widehat{x}_{0,a}) \in [0,1].$$

Because p_a is a deterministic function of the realised child pool and the model, it is fixed before the rollout rewards are sampled. For a candidate mean $q \in (0,1)$, use the proxy to choose a hedge weight

$$w_a(q) = \varepsilon + (1 - 2\varepsilon)\frac{1}{1 + \exp\{-\gamma(p_a - q)\}}, \qquad \varepsilon \in (0, 1/2), \quad \gamma \geq 0.$$

If $p_a > q$, more weight is assigned to the upper–sided betting process; if $p_a < q$, more weight is assigned to the lower–sided betting process. The lower bound ε prevents the hedge from placing all mass on one side. The special case $\gamma = 0$ gives a symmetric hedge.

E.3. One–Sided Betting Capital Processes

For each $q \in (0,1)$, define predictable bets $\lambda^+_{a,n}(q)$ and $\lambda^-_{a,n}(q)$ satisfying

$$0 \leq \lambda^+_{a,n}(q) \leq \frac{1}{q}, \qquad 0 \leq \lambda^-_{a,n}(q) \leq \frac{1}{1-q},$$

with $\lambda^{\pm}_{a,n}(q)$ measurable with respect to $\mathcal{G}_{a,n-1}$. The upper– and lower–sided capital processes are

$$K^{+}_{a,n}(q) = \prod_{i=1}^{n}\big(1 + \lambda^{+}_{a,i}(q)(R_{a,i} - q)\big),$$

$$K^{-}_{a,n}(q) = \prod_{i=1}^{n}\big(1 - \lambda^{-}_{a,i}(q)(R_{a,i} - q)\big).$$

The clipping constraints imply nonnegative factors for all rewards in $[0, 1]$. The Tweedie–guided hedged capital is

$$K_{a,n}(q) = w_a(q)K^{+}_{a,n}(q) + \big(1 - w_a(q)\big)K^{-}_{a,n}(q).$$

Under the null hypothesis $H_0 : \theta_a = q$, $K_{a,n}(q)$ is a nonnegative martingale with initial value one. Therefore Ville's inequality gives

$$\mathbb{P}_{\theta_a=q}\left(\sup_{n\geq 0} K_{a,n}(q) \geq \frac{1}{\alpha_a} \;\middle|\; Y_{1:M}\right) \leq \alpha_a.$$

The per–arm confidence sequence is obtained by inversion:

$$\widetilde{I}_a(n) = \mathrm{cl}\left\{q \in (0, 1) : K_{a,n}(q) < \frac{1}{\alpha_a}\right\}, \qquad I_a(m) = \widetilde{I}_a(N_a(m)).$$

Here cl is closure in $[0, 1]$, which handles endpoint means by one–sided limits.

E.4. Predictable Plug–In Bets Used in the Experiments

Validity holds for any predictable bets satisfying the clipping constraints. In our experiments we use a predictable empirical–Bernstein–style plug–in rule, similar in spirit to the bounded–mean betting strategies of Waudby-Smith & Ramdas (2024). Initialise

$$\widehat{\mu}_{a,0} = \frac{1}{2}, \qquad \widehat{v}_{a,0} = \frac{1}{4}.$$

After n rewards from arm a, define the smoothed mean and predictable variance estimate

$$\widehat{\mu}_{a,n} = \frac{1/2 + \sum_{i=1}^{n} R_{a,i}}{n + 1},$$

$$\widehat{v}_{a,n} = \frac{1/4 + \sum_{i=1}^{n}(R_{a,i} - \widehat{\mu}_{a,i-1})^2}{n + 1}.$$

For the nth rollout from arm a, set

$$\beta_{a,n} = \sqrt{\frac{2\log(2/\alpha_a)}{\widehat{v}_{a,n-1}\, n \log(n + 1)}}$$

and, for a truncation constant $c \in (0, 1)$,

$$\lambda^{+}_{a,n}(q) = \min\left\{\beta_{a,n}, \frac{c}{q}\right\}, \qquad \lambda^{-}_{a,n}(q) = \min\left\{\beta_{a,n}, \frac{c}{1 - q}\right\}.$$

These bets are $\mathcal{G}_{a,n-1}$–measurable and satisfy the required constraints. The factor $c < 1$ leaves a margin of safety in the nonnegativity condition. Note that this particular formula controls tightness, not validity; changing it to any other predictable clipped betting rule leaves the proofs unchanged.

E.5. Numerical Inversion

The mathematically valid confidence sequence is the continuous inversion above. A practical implementation proceeds as follows.

- Choose an interior grid $\mathcal{Q} = \{q_g = (g - 1/2)/G : g = 1, \ldots, G\} \subset (0, 1)$.

- For each child a and each q_g, store $K^+_{a,n}(q_g)$ and $K^-_{a,n}(q_g)$, initialised to one.
- Pre–compute $w_a(q_g)$ for all children and grid points at the start of the diffusion step.
- When the nth rollout from child a returns $R_{a,n}$, update $\lambda^{\pm}_{a,n}(q_g)$ and multiply

$$K^+_{a,n}(q_g) \leftarrow K^+_{a,n-1}(q_g)\big(1 + \lambda^+_{a,n}(q_g)(R_{a,n} - q_g)\big),$$

$$K^-_{a,n}(q_g) \leftarrow K^-_{a,n-1}(q_g)\big(1 - \lambda^-_{a,n}(q_g)(R_{a,n} - q_g)\big).$$

- Form

$$\widehat{\mathcal{Q}}_{a,n} := \left\{q_g \in \mathcal{Q} : w_a(q_g)K^+_{a,n}(q_g) + (1 - w_a(q_g))K^-_{a,n}(q_g) < 1/\alpha_a\right\}.$$

- Set

$$L_a(n) \;:=\; \min \widehat{Q}_{a,n}, \qquad U_a(n) \;:=\; \max \widehat{Q}_{a,n},$$

  with the convention $\min \emptyset = 1$, $\max \emptyset = 0$. These are the discrete-grid analogues of the convex-hull endpoints $\inf \widetilde{I}_a(n)$ and $\sup \widetilde{I}_a(n)$ from the remark in §D.3, and they are the quantities plugged into the pruning rule (15).

The grid weights $w_a(q_g)$ are computed once per child pool because they depend only on the realised children and the candidate means.

E.6. Proxy–Regularised Allocation

The confidence sequence is valid for any predictable allocation rule. The following rule focuses computation on the boundary between the current selected set and its strongest "challenger".

Let

$$\bar{R}_a(m) = \begin{cases} N_a(m)^{-1} \sum_{j=1}^{N_a(m)} R_{a,j}, & N_a(m) > 0, \\ 0, & N_a(m) = 0, \end{cases}$$

and define the proxy–regularised score

$$\widetilde{\mu}_a(m) = \frac{N_a(m)\bar{R}_a(m) + \tau p_a}{N_a(m) + \tau}, \qquad \tau \geq 0.$$

When $N_a(m) = 0$ and $\tau > 0$, this equals p_a. Larger τ makes the early ranking more proxy–driven; $\tau = 0$ recovers the empirical mean for sampled arms. Since p_a is $\sigma(Y_{1:M})$–measurable and the empirical rewards are in $\mathcal{F}_m$, the score is $\mathcal{F}_m$–measurable.

Let $\widehat{S}_m = \arg\operatorname{top}_N\{\widetilde{\mu}_a(m) : a \text{ active}\}$ be a provisional allocation set. Define the inside boundary arm and outside challenger by

$$a_{\text{in}}(m) \in \arg\min_{a \in \widehat{S}_m} L_a(m), \qquad a_{\text{out}}(m) \in \arg\max_{b \notin \widehat{S}_m} U_b(m).$$

The next rollout is allocated to the wider of these two intervals:

$$A_{m+1} \in \arg\max_{a \in \{a_{\text{in}}(m), a_{\text{out}}(m)\}} \big(U_a(m) - L_a(m)\big).$$

This rule is predictable because it is a deterministic function of $\mathcal{F}_m$. It affects only efficiency. The pruning certificate below uses confidence bounds, not the proxy–regularised score.

E.7. Certified Pruning and Stopping

At any inner round m, define the certificate set by lower confidence bounds

$$S_m = \arg\operatorname{top}_N\{L_a(m) : a \text{ active}\},$$

with deterministic tie–breaking. Let

$$\ell_m = \min_{a \in S_m} L_a(m), \qquad u_m = \max_{b \notin S_m} U_b(m).$$

Only the following pruning rule is certified:

$$\text{prune child } b \notin S_m \quad \text{if} \quad U_b(m) < \ell_m.$$

If $\ell_m > u_m$, the confidence intervals certify that every child in S_m has larger value than every child outside S_m, so the algorithm may stop and return S_m. If pruning reduces the active set to exactly N children, the algorithm may also stop and return the active set. If a fixed rollout budget is exhausted before either certificate occurs, the returned set can be chosen by a practical fallback rule, but the formal no–false–prune guarantee applies only to children pruned by the certified rule above.

E.8. Stagewise and Whole–Run Error Budgets

For a total failure probability δ over L diffusion steps, choose stagewise budgets $\delta_0, \ldots, \delta_{L-1}$ satisfying

$$\sum_{\ell=0}^{L-1} \delta_\ell \leq \delta.$$

A default choice is $\delta_\ell = \delta/L$. At step ℓ, choose per–arm budgets $\alpha_{\ell,a}$ satisfying

$$\sum_{a=1}^{M_\ell} \alpha_{\ell,a} \leq \delta_\ell,$$

for example $\alpha_{\ell,a} = \delta_\ell / M_\ell$. The confidence sequences then satisfy simultaneous stagewise coverage with probability at least $1 - \delta_\ell$, and a union bound over steps gives whole–run validity at level $1 - \delta$.