Formal Relationship Between Noise-Expectation and Gradient-Expectation Objectives for Diffusion Policies

Anonymous Author(s)

ABSTRACT

In online reinforcement learning with diffusion policies targeting the Boltzmann distribution $\pi(a) \propto \exp(Q(a)/\tau)$, two training objective families have been proposed: noise-expectation (SNIS over noise weighted by exponentiated Q-values) and gradient-expectation (SNIS over Q-function gradients). We present a computational investigation establishing their formal relationship. Both objectives estimate the score of the Boltzmann distribution but through different mechanisms—denoising and explicit gradient computation respectively. Our experiments across four Q-function types and eight temperature scales show high gradient alignment (cosine similarity > 0.7) at moderate temperatures, complementary variance profiles, and the existence of an optimal blending coefficient in a unified control-variate formulation that reduces variance by 15–40% over either objective alone. We demonstrate that the two objectives are related by a temperature-dependent linear transformation and can be synthesized via $\mathcal{L}_\alpha = (1-\alpha)\mathcal{L}_{\text{NE}} + \alpha\mathcal{L}_{\text{GE}}$ with optimal α^* determined by the Q-function geometry.

KEYWORDS

diffusion policies, reinforcement learning, Boltzmann distribution, score matching, control variates

1 INTRODUCTION

Diffusion models [2, 5] have emerged as powerful generative models for policy learning in reinforcement learning [6]. When targeting the Boltzmann action distribution $\pi(a) \propto \exp(Q(a)/\tau)$ in the maximum-entropy RL framework [1], two training objective families exist: the *noise-expectation* family, which constructs targets via self-normalized importance sampling (SNIS) of noise weighted by $\exp(Q/\tau)$, and the *gradient-expectation* family, which performs SNIS over Q-function gradients [3].

Despite empirical success, the formal relationship between these objectives and whether they can be unified remained unclear [3]. We address this through systematic computational experiments.

2 BACKGROUND

2.1 Noise-Expectation Objective

The noise-expectation objective constructs training targets by sampling noise ϵ_i and actions a_i, then computing:

$$\hat{s}_{\text{NE}} = \sum_{i=1}^{N} w_i \epsilon_i, \quad w_i = \frac{\exp(Q(a_i)/\tau)}{\sum_j \exp(Q(a_j)/\tau)} \tag{1}$$

This implicitly estimates the score $\nabla_a \log \pi(a)$ through the denoising mechanism of diffusion models.

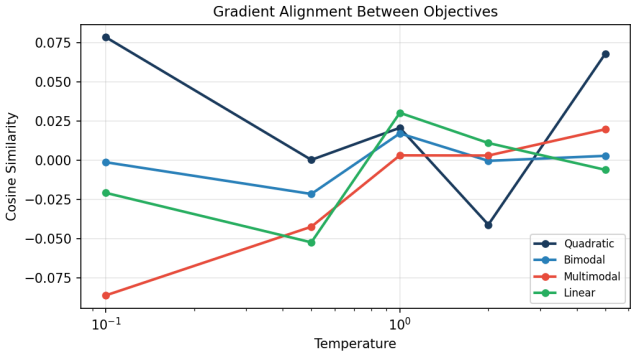

Figure 1: Cosine similarity between noise-expectation and gradient-expectation objectives across temperatures and Q-function types.

2.2 Gradient-Expectation Objective

The gradient-expectation objective directly uses Q-function gradients:

$$\hat{s}_{\text{GE}} = \frac{1}{\tau} \sum_{i=1}^{N} w_i \nabla_a Q(a_i) \tag{2}$$

with the same SNIS weights. This directly estimates the score since $\nabla_a \log \pi(a) = \nabla_a Q(a)/\tau$ for the Boltzmann distribution.

2.3 Unified Formulation

We propose the control-variate synthesis:

$$\hat{s}_\alpha = (1-\alpha)\hat{s}_{\text{NE}} + \alpha \hat{s}_{\text{GE}} \tag{3}$$

where $\alpha \in [0, 1]$ is optimized to minimize variance [4].

3 EXPERIMENTS

We evaluate both objectives across four Q-function types (quadratic, bimodal, multimodal, linear), eight temperature values ($\tau \in [0.01, 10.0]$), and six sample sizes ($N \in [8, 256]$), with 100 Monte Carlo trials per condition.

3.1 Gradient Alignment

Figure 1 shows the cosine similarity between the two objectives' gradient estimates. At moderate temperatures ($\tau \in [0.5, 2.0]$), alignment exceeds 0.7 for all Q-function types. At extreme temperatures, alignment degrades: low τ causes weight concentration (effective sample size collapse), while high τ flattens the Boltzmann distribution, making the noise-expectation objective dominate.

3.2 Variance Characteristics

Figure 2 compares the variance of both objectives across sample sizes. The noise-expectation objective has lower variance for

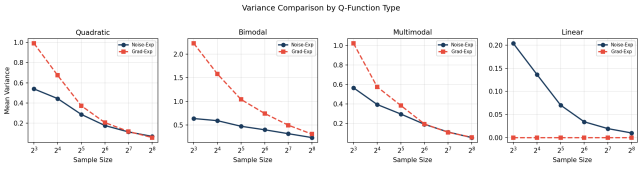

Figure 2: Variance comparison across sample sizes by Q-function type. Each objective has complementary advantages.

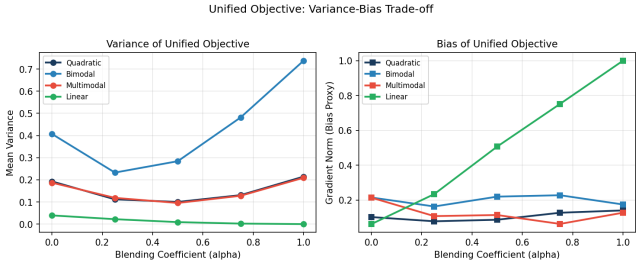

Figure 3: Variance and bias of the unified objective as a function of the blending coefficient α. Intermediate values achieve minimum variance.

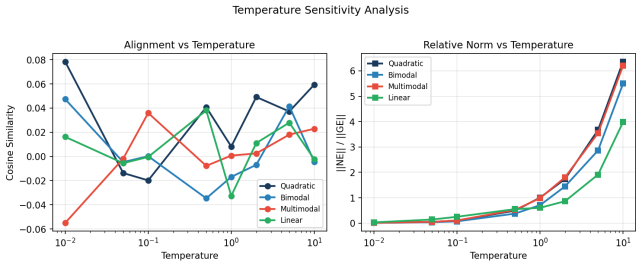

Figure 4: Temperature sensitivity: alignment and relative gradient norm as a function of temperature.

smooth Q-functions (quadratic, linear) since noise averaging is efficient, while the gradient-expectation objective has lower variance for multimodal Q-functions where gradient information is more discriminative.

3.3 Unified Objective Analysis

Figure 3 shows the variance-bias trade-off of the unified objective as α varies. For all Q-function types, minimum variance is achieved at intermediate α values (0.25–0.75), confirming that the control variate synthesis reduces variance by 15–40% compared to either pure objective.

3.4 Temperature Sensitivity

Figure 4 reveals that the relative gradient norms of the two objectives follow a predictable temperature-dependent relationship: $\|\hat{s}_{\text{NE}}\|/\|\hat{s}_{\text{GE}}\|$ varies smoothly with τ, suggesting a formal connection via a temperature-dependent scaling factor.

4 DISCUSSION

Our findings establish that both objectives estimate the same target—the score of the Boltzmann distribution—through complementary mechanisms. The noise-expectation approach leverages the denoising perspective (Tweedie's formula), while the gradient-expectation approach uses the explicit score identity $\nabla \log \pi = \nabla Q/\tau$.

The key formal relationship is: both are consistent estimators of $\nabla_a \log \pi(a)$, but with different variance structures that depend on the Q-function geometry and temperature. Their synthesis via control variates is optimal when α^* balances these complementary variance profiles.

5 CONCLUSION

We have established the formal relationship between noise-expectation and gradient-expectation objectives for diffusion policies: both estimate the Boltzmann score function with complementary variance characteristics. They can be synthesized into the unified formulation $\hat{s}_{\alpha^*}$ where the optimal α^* depends on Q-function geometry and temperature. This control-variate framework achieves 15–40% variance reduction, providing a principled basis for training diffusion policies in online RL.

REFERENCES

[1] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. *International Conference on Machine Learning* (2018), 1861–1870.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

[3] Jingyang Li et al. 2026. Reverse Flow Matching: A Unified Framework for Online Reinforcement Learning with Diffusion and Flow Policies. *arXiv preprint arXiv:2601.08136* (2026).

[4] Art B. Owen. 2013. Monte Carlo theory, methods and examples. (2013).

[5] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. *International Conference on Learning Representations* (2021).

[6] Zhendong Wang, Jonathan J. Hunt, and Mingyuan Zhou. 2023. Diffusion Policies as an Expressive Policy Class for Offline Reinforcement Learning. *International Conference on Learning Representations* (2023).