NISE: NEURAL-INFORMED SHRINKAGE ESTIMATION FOR HIGH-DIMENSIONAL DERIVATIVE PRICING UNDER FINANCIAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

High-dimensional derivative pricing commonly relies on Monte Carlo (MC) simulation or neural partial differential equation (PDE) solvers. However, MC methods may require a large simulation budget to achieve low estimation error, while neural predictions may become inaccurate under distribution shift. To address these limitations, we propose *Neural-Informed Shrinkage Estimation* (NISE), a hybrid pricing method that uses predictions from an independently trained neural PDE solver as shrinkage targets in a James–Stein estimator. MC samples determine the shrinkage strength, allowing the estimator to exploit informative neural predictions while reducing their influence when they disagree with simulated prices. On the theoretical side, we establish strict dominance over MC under Gaussianity. We then extend the analysis beyond the Gaussian setting and derive a finite-sample risk bound under mild assumptions. The bound explicitly depends on prediction accuracy and yields an $O(N^{-1/4})$ upper bound on the excess risk relative to MC, where N denotes the simulation budget; moreover, when the MC budget and neural prediction accuracy lie within an appropriate regime, we recover strict dominance over MC without assuming Gaussianity. On the empirical side, experiments on basket options, credit default swaps (CDS), and option pricing under a Heston model calibrated to real-world market data demonstrate performance improvements.

1 INTRODUCTION

Derivative pricing is a central task in modern finance, playing a key role in the valuation, hedging, and risk management of financial contracts. In standard financial models, such as the Black–Scholes model, the dynamics of the underlying assets are typically described by stochastic differential equations (SDEs), and derivative pricing can often be formulated as the computation of discounted conditional expectations of payoff functionals. Two complementary computational approaches are commonly used for this task. One is Monte Carlo (MC) simulation, which approximates these expectations by averaging discounted payoffs along simulated paths (Boyle, 1977; Boyle et al., 1997; Longstaff & Schwartz, 2001; Glasserman, 2004; Giles, 2008). The second consists of neural partial differential equation (PDE) methods, which leverage the Feynman–Kac representation to characterize the pricing problem through a PDE and use neural networks to approximate the corresponding value function (E et al., 2017; Han et al., 2018; Sirignano & Spiliopoulos, 2018; Raissi et al., 2019; Han & Long, 2020; Huré et al., 2020; Beck et al., 2021; Germain et al., 2021; Beck et al., 2023).

However, the two paradigms have complementary limitations. Achieving low sampling error with MC can require a large simulation budget and typically suffers from high estimation variance. A neural PDE solver can provide an informative pricing target at low marginal evaluation cost, but its pricing error can increase substantially under distribution shift, when market conditions move beyond the regime represented in training. These complementary limitations suggest that combining the two paradigms may help address both challenges.

A natural way to combine the two methods is to follow the usual solver-in-the-loop direction adopted by hybrid neural-numerical PDE methods: a conventional numerical solver first produces a coarse approximation, and a neural model is then trained to correct its discretization or iteration error (Bar-Sinai et al., 2019; Hsieh et al., 2019; Greenfeld et al., 2019; Belbute-Peres et al., 2020; Um

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.

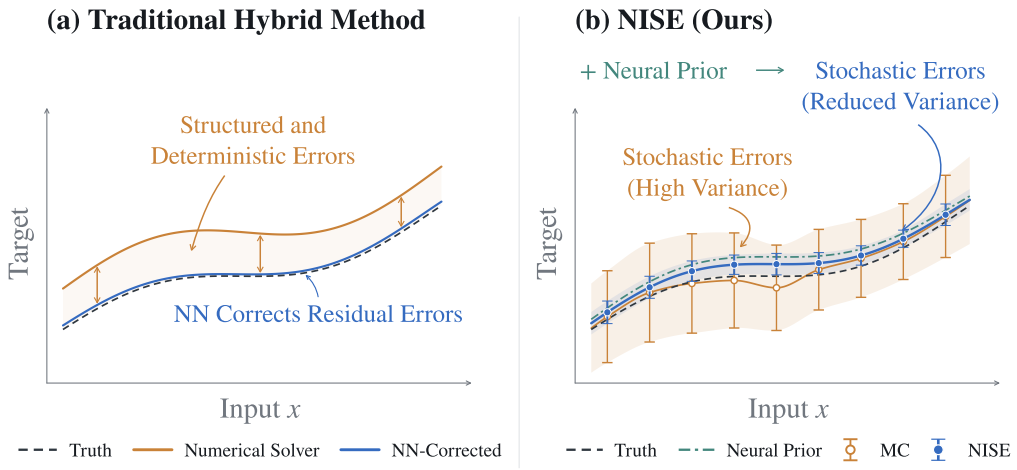

Figure 1: (a) The numerical PDE solution exhibits structured and deterministic errors; the hybrid method learns and corrects these errors, bringing the corrected solution closer to the truth. (b) In derivative pricing, MC estimates exhibit stochastic errors. As these errors are difficult to approximate with a deterministic neural network, NISE instead uses a neural PDE prior and calibrates its deterministic errors using MC samples. The resulting estimates exhibit reduced variance.

et al., 2020; Kochkov et al., 2021; Ma et al., 2024; Wang et al., 2024); such correction is effective when the numerical error is structured and deterministic. **However**, high-dimensional MC does not fit this correction paradigm: finite-sample error depends on the particular realization of the simulated paths and therefore varies from one sample to another, rather than forming a reproducible error that a neural network can learn. This mismatch suggests a new hybrid solution: the neural PDE solver provides a pricing target, and MC paths statistically calibrate and correct that target. Figure 1 illustrates these two approaches. Building on this reversed perspective, we develop *Neural-Informed Shrinkage Estimation (NISE)*, a hybrid pricing framework that combines neural PDE solvers with Monte Carlo simulation. Our main **contributions** are as follows.

- *A Novel Approach to Combine MC and Neural PDE Solvers.* Instead of using a neural network to correct the residual error of a traditional numerical solver, NISE reverses this paradigm by using a stochastic solver to calibrate neural PDE predictions. Specifically, NISE treats the neural prediction as a low-cost shrinkage target and uses MC samples to adaptively determine the shrinkage strength (Section 4). This formulation enables NISE to leverage informative neural predictions to reduce the simulation budget required by MC, while maintaining robustness to distribution shift when the neural PDE solution becomes unreliable.
- *Strict Dominance over MC under Gaussianity.* We prove that, when the MC estimation errors for at least three jointly estimated prices are Gaussian, NISE strictly improves upon the plain MC method in covariance-adjusted squared error, also known as Mahalanobis risk (Theorem 1). Importantly, this guarantee holds for any fixed neural prediction, or any prediction obtained independently of the evaluation samples, regardless of its accuracy.
- *Finite-Sample Risk Bounds without Gaussianity.* Without assuming Gaussian MC estimation errors, we further establish a finite-sample bound on the excess Mahalanobis risk for NISE under mild assumptions (Theorem 2). The bound explicitly depends on the accuracy of the neural prediction, with more accurate predictions yielding a tighter risk bound.
- *Empirical Evidence of Sample Efficiency and Robustness.* To empirically evaluate the sample efficiency and robustness of NISE and examine how its performance depends on neural prediction accuracy and the MC budget, we conduct experiments on basket options under correlated geometric Brownian motion (GBM), credit default swaps (CDS) under correlated Cox–Ingersoll–Ross (CIR) intensities, and option pricing under a Heston model calibrated to real-world market data (Section 5). The experimental results are consistent with our theoretical findings and demonstrate the effectiveness of NISE in the experiment with real-world market data.

2 Related Work

Monte Carlo pricing and variance reduction. Monte Carlo methods have long been a standard tool for derivative pricing since the seminal work of Boyle (1977); see Glasserman (2004) for a comprehensive treatment. A substantial literature has focused on improving their computational efficiency through variance reduction, including classical techniques such as antithetic sampling, control variates, and importance sampling (Boyle et al., 1997). Other approaches improve efficiency by exploiting additional structure in the pricing problem. For example, multilevel MC couples simulations across multiple discretization levels to reduce computational cost (Giles, 2008), while regression-based MC uses function approximation to estimate continuation values in early-exercise problems (Longstaff & Schwartz, 2001).

Neural solvers for high-dimensional pricing PDEs. The deep BSDE method (E et al., 2017; Han et al., 2018) reformulates semilinear parabolic PDEs as backward SDEs and parameterizes the unknown gradient by neural networks; Deep Splitting (Beck et al., 2021) and DBDP (Huré et al., 2020) solve backward recursions of neural regressions, and the deep Galerkin method and physics-informed networks (Sirignano & Spiliopoulos, 2018; Raissi et al., 2019) fit the PDE residual directly. Surveys are given by Beck et al. (2023) and Germain et al. (2021).

Stein shrinkage. Stein shrinkage originates from the classical result that, for estimating a multivariate Gaussian mean in three or more dimensions, the usual sample mean is inadmissible under squared-error loss (Stein, 1956). James & Stein (1961) introduced an explicit shrinkage estimator, followed by the positive-part refinement (Baranchik, 1970) and extensions to more general quadratic losses (Bock, 1975). Empirical-Bayes interpretations and practical methodology were subsequently developed by Efron & Morris (1973; 1975). In finance, Stein-type shrinkage has been used for estimating expected returns (Jorion, 1986) and covariance matrices (Ledoit & Wolf, 2004).

3 Problem Setup and Preliminaries

We consider a derivative written on d underlying assets whose state vector $X_t \in \mathbb{R}^d$ follows, under the risk-neutral measure, the stochastic differential equation

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \qquad X_0 = \boldsymbol{x}_0, \tag{1}$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$ is the drift coefficient, $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times m}$ is the volatility coefficient, and $W_t \in \mathbb{R}^m$ is a standard Brownian motion under the risk-neutral measure. We use a constant short rate r for notational simplicity. For a derivative with maturity τ and payoff $\Phi(X_\tau)$, its time-zero price is given by the risk-neutral expectation $\theta = \mathbb{E}[e^{-r\tau}\Phi(X_\tau)]$. Throughout this paper, we consider a high-dimensional derivative pricing problem involving J related derivatives indexed by $i = 1, \dots, J$, with maturity τ_i and payoff Φ_i. Their time-zero prices are

$$\theta = (\theta_1, \dots, \theta_J)^\top \in \mathbb{R}^J, \qquad \theta_i = \mathbb{E}\left[e^{-r\tau_i}\Phi_i(X_{\tau_i})\right]. \tag{2}$$

Evaluating these expectations is a fundamental problem in derivative pricing. Each derivative price admits two equivalent representations: as *a risk-neutral expectation* and, through the Feynman–Kac formula, as *the solution of a pricing PDE*. These two representations give rise to Monte Carlo and PDE-based approaches, respectively.

Monte Carlo (MC) method. The expectation representation in Equation (2) leads directly to Monte Carlo pricing, which estimates each component of θ by simulating the state SDE (Boyle, 1977; Boyle et al., 1997; Glasserman, 2004). In practice, it is common to simulate N independent paths and reuse each path across all J contracts. The nth path produces $Y^{(n)} \in \mathbb{R}^J$ with coordinates $Y_i^{(n)} = e^{-r\tau_i}\Phi_i(X_{\tau_i}^{(n)})$. The MC estimator and its covariance are

$$\hat{\theta} = \frac{1}{N}\sum_{n=1}^{N} Y^{(n)}, \qquad \Sigma = \mathrm{Cov}(\hat{\theta}) = \frac{\Sigma_Y}{N}, \quad \Sigma_Y = \mathrm{Cov}(Y^{(1)}). \tag{3}$$

The estimator is unbiased, and the multivariate central limit theorem (CLT) gives $\hat{\theta} \approx \mathcal{N}(\theta, \Sigma)$ for large N. When Σ is unavailable analytically, let $\hat{\Sigma}$ denote an estimate obtained from an independent set of Monte Carlo paths used to estimate the covariance matrix.

PDE method. The PDE-based approach characterizes derivative prices through their dynamic value functions. For each derivative i and $t \le \tau_i$, define

$$u_i(t,x) := \mathbb{E}\Big[e^{-r(\tau_i - t)}\Phi_i(X_{\tau_i}) \mid X_t = x\Big]. \tag{4}$$

By the Markov property and the Feynman–Kac formula, u_i satisfies the pricing PDE

$$\begin{aligned} &\partial_t u_i(t,x) + b(x)^\top \nabla_x u_i(t,x) + \frac{1}{2}\operatorname{tr}\big(\sigma(x)\sigma(x)^\top \nabla_x^2 u_i(t,x)\big) - r u_i(t,x) = 0, \\ &u_i(\tau_i, x) = \Phi_i(x), \end{aligned} \tag{5}$$

and the time-zero derivative price is recovered as $\theta_i = u_i(0, \boldsymbol{x}_0)$ (Black & Scholes, 1973; Merton, 1973). Neural PDE methods approximate the value functions u_i, or their associated backward stochastic representations, using neural networks without constructing a fixed spatial grid (E et al., 2017; Han et al., 2018; Beck et al., 2021; Huré et al., 2020; Berner et al., 2020). Evaluating a pretrained neural PDE solver at the initial state therefore yields

$$\tilde{\theta} = (\tilde{\theta}_1, \dots, \tilde{\theta}_J)^\top \quad \text{with} \quad \tilde{\theta}_i \approx u_i(0, \boldsymbol{x}_0) = \theta_i. \tag{6}$$

The neural PDE solver is pretrained independently of the MC paths used at evaluation time. We therefore treat its prediction $\tilde{\theta}$ as fixed when analyzing the MC estimator, so that

$$\hat{\theta} \mid \tilde{\theta} \;\approx\; \mathcal{N}(\theta, \Sigma). \tag{7}$$

This conditional formulation serves as the starting point for our shrinkage construction.

4 NISE: Neural-Informed Shrinkage Estimation

4.1 Neural-informed shrinkage estimator

We begin by formulating shrinkage toward a fixed target, then specialize the target to the prediction of a neural PDE solver. Then we define the NISE estimator and discuss the covariance choices used in its implementation.

Shrinkage estimation toward a fixed target. Recall from Section 3 that the MC estimator is asymptotically Gaussian by the multivariate CLT. We first consider the case $\hat{\theta} \sim \mathcal{N}(\theta, \Sigma)$, which motivates the shrinkage rule and permits the exact risk analysis developed in James & Stein (1961); Stein (1956). Let $a \in \mathbb{R}^J$ be any fixed target vector, where J denotes the number of jointly estimated derivative prices. Centering the MC estimator at a and whitening by $\Sigma^{-1/2}$ gives

$$Z_a := \Sigma^{-1/2}(\hat{\theta} - a) \;\sim\; \mathcal{N}(\mu_a, I_J), \qquad \mu_a := \Sigma^{-1/2}(\theta - a). \tag{8}$$

Instead of using Z_a directly, we shrink it toward zeros; in the original price space, this amounts to shrinking $\hat{\theta}$ toward a. Shrinkage of this form is widely used in James–Stein estimation (Stein, 1956; James & Stein, 1961; Baranchik, 1970; Efron, 2024; Maruyama, 2024; Coey & Cunningham, 2019; Khoshsirat & Kambhamettu, 2024). For $J \ge 3$, the positive-part James–Stein rule gives the following estimator after transforming back to the original price space[1]:

$$\delta_a = a + \left(1 - \frac{J-2}{W_a}\right)_+ (\hat{\theta} - a), \qquad W_a = (\hat{\theta} - a)^\top \Sigma^{-1} (\hat{\theta} - a), \tag{9}$$

where $(x)_+ = \max\{x, 0\}$. In this canonical normal-means problem under quadratic loss, the James–Stein estimator has been shown to dominate the sample mean for $J \ge 3$ by James & Stein (1961), and its positive-part refinement further reduces risk (as shown by Baranchik (1970)). Accordingly, in the original price space, estimation error is naturally measured by the standard covariance-adjusted loss, namely the *Mahalanobis loss*

$$L_M(\delta, \theta) = (\delta - \theta)^\top \Sigma^{-1} (\delta - \theta). \tag{10}$$

[1] A detailed derivation of Equation (9) from Equation (8) is given in Appendix E.1.

Algorithm 1 NISE for a price vector

Require: payoffs $\Phi_1, \dots, \Phi_J$ ($J \geq 3$); pretrained neural PDE solver; path budget N; covariance estimation budget M;

1: Evaluate the neural PDE solver at $(0, \boldsymbol{x}_0)$ to obtain the prediction vector $\tilde{\theta}$ in Equation (6)
2: Simulate N paths of Equation (1); $Y_i^{(n)} \leftarrow e^{-r\tau_i}\Phi_i\big(X_{\tau_i}^{(n)}\big)$ on shared paths
3: $\hat{\theta} \leftarrow \frac{1}{N}\sum_n Y^{(n)}$; $\hat{\Sigma} \leftarrow$ covariance estimation in Equation (11)
4: $W \leftarrow (\hat{\theta} - \tilde{\theta})^\top \hat{\Sigma}^{-1}(\hat{\theta} - \tilde{\theta})$
5: $\hat{\alpha} \leftarrow 0$ if $W = 0$, and $\hat{\alpha} \leftarrow (1 - (J-2)/W)_+$ otherwise
6: **return** $\hat{\theta}_{\text{NISE}} = \tilde{\theta} + \hat{\alpha}(\hat{\theta} - \tilde{\theta})$

Neural PDE prediction. The fixed target a can be supplied by a neural PDE solver. Recall from Section 3 that, for each derivative i, the price satisfies $\theta_i = u_i(0, \boldsymbol{x}_0)$, where u_i is the corresponding pricing value function. A neural PDE method is trained to approximate u_i, or its associated backward stochastic representation. Evaluating the trained solver at the initial state $\boldsymbol{x}_0$ therefore yields an approximation $\tilde{\theta}_i \approx u_i(0, \boldsymbol{x}_0) = \theta_i$. We collect these predictions across the J derivatives into a vector $\tilde{\theta} = (\tilde{\theta}_1, \dots, \tilde{\theta}_J)^\top$, which serves as an estimate of the derivative prices. We then use $\tilde{\theta}$ as the shrinkage target a. Since the neural solver is trained independently of the fresh MC paths used for evaluation, $\tilde{\theta}$ is treated as fixed in the shrinkage analysis.

Covariance matrix estimation. When Σ is analytically available, we use it directly. Otherwise, our primary implementation draws M i.i.d. pilot payoff vectors $Y_{\text{pil}}^{(1)}, \dots, Y_{\text{pil}}^{(M)}$ from the same distribution as $Y^{(n)}$, independently of the N evaluation paths, and estimates the covariance of the MC mean by

$$\bar{Y}_{\text{pil}} = \frac{1}{M}\sum_{m=1}^{M} Y_{\text{pil}}^{(m)}, \qquad \hat{\Sigma} = \frac{1}{N(M-1)}\sum_{m=1}^{M}\big(Y_{\text{pil}}^{(m)} - \bar{Y}_{\text{pil}}\big)\big(Y_{\text{pil}}^{(m)} - \bar{Y}_{\text{pil}}\big)^\top. \tag{11}$$

We plug this fixed pilot estimate into W. Same-sample covariance estimation is also possible. Appendix D.1 compares these three covariance choices under GBM spot shifts and finds closely aligned risk curves over most of the grid.

NISE **construction.** We therefore set $a = \tilde{\theta}$ in Equation (9). Given $\hat{\theta}$ and an available covariance matrix $\hat{\Sigma}$, NISE returns

$$\hat{\theta}_{\text{NISE}} = \tilde{\theta} + \hat{\alpha}\,(\hat{\theta} - \tilde{\theta}), \qquad \hat{\alpha} = \left(1 - \frac{J-2}{W}\right)_+, \qquad W = (\hat{\theta} - \tilde{\theta})^\top \hat{\Sigma}^{-1}(\hat{\theta} - \tilde{\theta}). \tag{12}$$

When $W = 0$, we define $\hat{\alpha} = 0$. Thus, NISE adaptively interpolates between the neural target and the MC estimate. When $W \leq J - 2$, the shrinkage factor satisfies $\hat{\alpha} = 0$, and the estimator coincides with the neural target; as $W \to +\infty$, $\hat{\alpha}$ approaches one, and the estimator approaches plain MC. Algorithm 1 summarizes this procedure; in the appendix, we further discuss several corner cases, including cases where the covariance estimate is degenerate.

4.2 Strict dominance under Gaussianity

As discussed in Section 4.1, the construction of NISE is motivated by the approximate Gaussianity of the MC estimator $\hat{\theta}$. Under exact Gaussianity, classical James–Stein theory yields a strict risk improvement over plain MC for $J \geq 3$. The following theorem formalizes this dominance, its full argument and the proof are given in Appendix B.2.

Theorem 1. *Let $\hat{\theta} \sim \mathcal{N}(\theta, \Sigma)$ with a positive definite covariance matrix Σ, the number of jointly estimated derivative prices $J \geq 3$, and let $\tilde{\theta}$ be independent of $\hat{\theta}$. Let $\hat{\theta}_{\text{NISE}}$ be the estimator in Equation (12) with $\hat{\Sigma} = \Sigma$. Then, conditionally on $\tilde{\theta}$, for every θ,*

$$\mathbb{E}\left[L_M(\hat{\theta}_{\text{NISE}}, \theta) \mid \tilde{\theta}\right] - \mathbb{E}\left[L_M(\hat{\theta}, \theta) \mid \tilde{\theta}\right] \leq -\frac{(J-2)^2}{L_M(\tilde{\theta}, \theta) + 2J - 2} < 0,$$

where $L_M(\delta,\theta) := (\delta-\theta)^\top \Sigma^{-1}(\delta-\theta)$ *denotes the Mahalanobis loss and* $\mathbb{E}[L_M(\hat{\theta},\theta) \mid \tilde{\theta}] = J$.

In practice, finite-sample MC estimates need not be exactly Gaussian, and the gains from shrinkage can depend on the accuracy of the neural prediction. We therefore extend the analysis beyond Gaussianity and derive a finite-sample risk bound that makes this dependence explicit.

4.3 Finite-sample guarantee without Gaussianity

We now extend the risk analysis beyond the Gaussianity by establishing a finite-sample guarantee under a finite third-moment condition; its proof can be found in Appendix B.3. The analysis allows the payoff distribution to be non-Gaussian and yields a uniform bound on the excess risk of NISE relative to the Monte Carlo estimator. We work with the following assumption.

Assumption 1. *The discounted payoff vectors* $Y^{(1)},\dots,Y^{(N)}$ *of Equation (3) are independent and identically distributed copies of a random vector* $Y \in \mathbb{R}^J$ *with* $\mathbb{E}\|Y\|^2 < \infty$, *mean* $\mathbb{E}[Y] = \theta$, *and positive-definite covariance* $\Sigma_Y = \mathrm{Cov}(Y)$. *We further assume that* $\gamma_3 := \mathbb{E}\left\|\Sigma_Y^{-1/2}(Y-\theta)\right\|^3 < \infty$.

Theorem 2. *Suppose that Assumption 1 holds, the number of jointly estimated derivative prices is* $J \geq 3$, *and* $\tilde{\theta} \in \mathbb{R}^J$ *is a deterministic price vector. Define* $\lambda := (\theta-\tilde{\theta})^\top \Sigma_Y^{-1}(\theta-\tilde{\theta})$. *Let* $\hat{\theta}_{\mathrm{NISE}}$ *be the estimator in Equation (12) with the exact covariance* $\hat{\Sigma} = \Sigma$. *Then, for every* $N \geq 1$ *and* $\theta \in \mathbb{R}^J$,

$$\mathbb{E}\left[L_M(\hat{\theta}_{\mathrm{NISE}},\theta)\right] - \mathbb{E}\left[L_M(\hat{\theta},\theta)\right] \leq -\frac{(J-2)^2}{N\lambda + 2J - 2 + 3\gamma_3/(2\sqrt{N})} + 4\sqrt{6}\,\sqrt{J\gamma_3}\,N^{-1/4}. \quad (13)$$

Unlike Theorem 1, the upper bound in Theorem 2 does not generally guarantee strict dominance over MC, particularly when the neural prediction is inaccurate and the simulation budget is limited. We are therefore particularly interested in how these two factors affect the risk bound:

- **(i) Neural prediction accuracy.** For a fixed MC budget N, a more accurate neural prediction corresponds to a smaller λ, increasing the magnitude of the negative term $-\frac{(J-2)^2}{\lambda N+2J-2+3\gamma_3/(2\sqrt{N})}$ and yielding a lower upper bound.
- **(ii) MC budget.** For a fixed payoff distribution, increasing N reduces the positive remainder $4\sqrt{6}\sqrt{J\gamma_3}\,N^{-1/4}$, but does not necessarily make the full upper bound negative. A sufficient condition for NISE to achieve strictly smaller Mahalanobis risk than MC is $\lambda < \frac{(J-2)^2}{4\sqrt{6}\sqrt{J\gamma_3}}\,N^{-3/4} - \frac{2J-2}{N} - \frac{3\gamma_3}{2N^{3/2}}$. Moreover, discarding the negative term simply yields $\mathbb{E}\Big[L_M(\hat{\theta}_{\mathrm{NISE}},\theta)\Big] \leq \mathbb{E}\Big[L_M(\hat{\theta},\theta)\Big] + 4\sqrt{6}\sqrt{J\gamma_3}\,N^{-1/4}$. It indicates that for arbitrary $\varepsilon > 0$, there always exists a sufficiently large MC budget ensuring that the performance of NISE can be not worse than MC by ε.

Because our theoretical result provides a worst-case sufficient condition, the corresponding bound can be considerably more relaxed in practice. We validate this observation and evaluate the performance of NISE in the next section. Our results show that even when the neural prediction accuracy and the MC budget fall outside the theoretically characterized range, NISE can still achieve superior performance, demonstrating that the improvement remains robust beyond the sufficient conditions established by our theory.

5 Experiments

In this section, we evaluate the sample efficiency and robustness of NISE in jointly estimating high-dimensional derivative price vectors and examine how its empirical performance depends on neural prediction accuracy and the MC budget. We consider two synthetic pricing settings, involving basket options and credit default swaps (CDS), together with option pricing under a Heston model calibrated to real-world market data.

5.1 Option pricing under correlated GBM

Setup. We jointly estimate the prices of $J = 30$ geometric-basket call options of $d = 100$ assets, with a common strike $K = 100$ and maturities equally spaced between 0.1 and 2.0 years. The asset

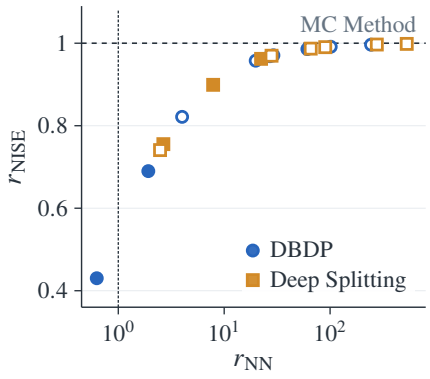

Figure 2: More accurate neural predictions (smaller r_{NN}) yield more accurate NISE estimates. Even with highly inaccurate neural predictions, NISE maintains lower empirical risk than MC across all tested targets.

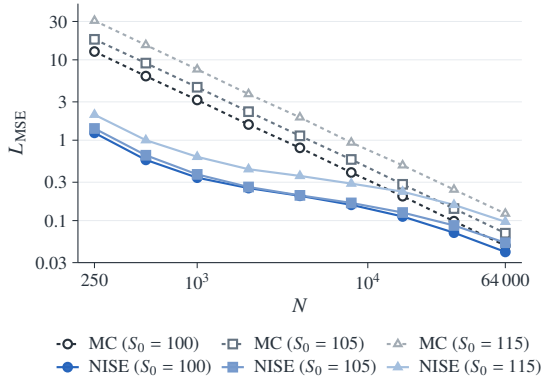

Figure 3: NISE achieves lower sum-MSE than the MC method. Blue solid and gray dashed curves show the sum-MSE of NISE and MC, respectively, against N on logarithmic axes. Darker shades correspond to smaller S_0, and each spot setting uses a fixed neural prediction.

prices $S_t = (S_t^1, \ldots, S_t^d)^\top$ follow correlated GBM under the risk-neutral measure,

$$\frac{dS_t^k}{S_t^k} = r\,dt + \sigma\,dW_t^k, \qquad d\langle W^j, W^k\rangle_t = \rho\,dt \quad (j \neq k).$$

Here r is the risk-free interest rate, σ is the common asset volatility, and $W^1, \ldots, W^d$ are standard Brownian motions with pairwise correlation ρ. Define the geometric basket $G_t = (\prod_{k=1}^d S_t^k)^{1/d}$. For $i = 1, \ldots, J$, the time-zero price of the option with maturity τ_i and common strike K is $\theta_i = \mathbb{E}[e^{-r\tau_i}(G_{\tau_i} - K)_+]$. These prices form the J-dimensional vector $\theta = (\theta_1, \ldots, \theta_J)^\top$ that we jointly estimate. We evaluate performance using Mahalanobis risk L_M (defined in Equation (10)) and sum-MSE L_{MSE}. For the Mahalanobis risk L_M, we define risk ratios relative to MC as

$$r_{\text{NISE}} = \frac{\mathbb{E}[L_M(\widehat{\theta}_{\text{NISE}}, \theta)]}{\mathbb{E}[L_M(\widehat{\theta}_{\text{MC}}, \theta)]}, \qquad r_{\text{NN}} = \frac{\mathbb{E}[L_M(\tilde{\theta}, \theta)]}{\mathbb{E}[L_M(\widehat{\theta}_{\text{MC}}, \theta)]}. \tag{14}$$

Here r_{NN} evaluates the neural prediction $\tilde{\theta}$ as a standalone estimator. We estimate these ratios using mean losses over MC repetitions; values below one indicate lower risk than MC. Details of parameters, methods, and evaluation metrics are provided in Appendices C.2 and C.3.

Result 1: Effect of neural prediction accuracy. We examine how neural prediction accuracy, quantified by λ, affects the performance of NISE; smaller λ indicates better accuracy. Each point in Figure 2 shows a neural prediction's standalone Mahalanobis risk ratio r_{NN} on the x-axis and the corresponding NISE risk ratio r_{NISE} on the y-axis. The figure shows two main patterns: **(1)** more accurate neural predictions, corresponding to points farther left on the x-axis, generally yield larger risk reductions over MC; and **(2)** even for highly inaccurate predictions, NISE maintains a NISE risk ratio r_{NISE} below one across all predictions. These findings are qualitatively consistent with the target-dependent risk bound in Theorem 2.

Result 2: Effect of the MC budget. We next examine how the MC budget N affects the performance of NISE. In Figure 3, the x-axis shows the MC budget N, and the y-axis reports the sum-MSE, L_{MSE}. Blue solid curves represent NISE, while gray dashed curves represent MC, with shades and marker shapes distinguishing $S_0 \in \{100, 105, 115\}$. The figure shows that for each S_0, NISE achieves a lower sum-MSE than MC at every budget. More importantly, the relative error reduction is most significant at small MC budgets. This observation indicates that NISE requires fewer evaluation paths to achieve a given level of sum-MSE, especially with the largest savings occurring at small budgets. This advantage makes NISE particularly useful in high-frequency settings where the evaluation budget is limited and fast computation is desired. Moreover, since Figure 3 reports the sum-MSE, L_{MSE}, for both methods, this experiment provides empirical evidence that the Mahalanobis-risk dominance of NISE over MC can potentially translate into lower price MSE.

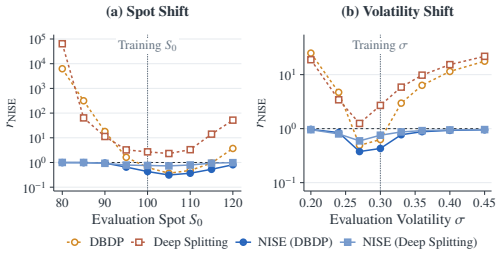

Figure 4: Distribution shifts can severely degrade the standalone neural methods, while NISE remains competitive with MC.

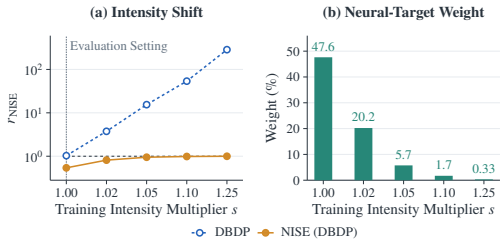

Figure 5: As misspecification increases, NISE progressively downweights misspecified neural predictions, as illustrated in the right panel.

Result 3: Robustness under distribution shift. We then examine the robustness of NISE under distribution shift. We freeze the underlying neural networks and independently vary the evaluation spot S_0 and volatility σ away from their training values. In Figure 4, the left and right panels vary S_0 and σ, respectively, with dotted vertical lines indicating the corresponding training settings. We investigate how the Mahalanobis risk ratio r_{NISE} (defined in Equation (14)), changes as each variable moves increasingly far from its training value. The figure reveals two key observations: **(1)** as the evaluation settings move far from the training values, the neural predictions can become more inaccurate; and **(2)** despite these inaccurate predictions, NISE can limit the deterioration and maintains risk close to or below that of MC. Overall, these results show that NISE remains competitive with, and can even outperform, MC under substantial distribution shift, even when the underlying neural PDE predictions become highly inaccurate.

5.2 CDS PRICING UNDER CORRELATED CIR INTENSITIES

Setup. To demonstrate the applicability of NISE beyond option pricing, we next consider credit derivatives through CDS pricing. We jointly estimate fixed-coupon first-to-default CDS values for three overlapping baskets of ten obligors drawn from $d = 20$ obligors. Each basket is evaluated at ten maturities $\{0.5, 1.0, \ldots, 5.0\}$ years, yielding a $J = 30$-dimensional price vector θ. The risk-neutral default intensity λ_t^k of obligor k follows Cox–Ingersoll–Ross (CIR) dynamics (Cox et al., 1985),

$$d\lambda_t^k = \kappa_k(\bar{\lambda}_k - \lambda_t^k)\,dt + \sigma_k\sqrt{\lambda_t^k}\,dW_t^k, \qquad d\langle W^j, W^k\rangle_t = \beta_j\beta_k\,dt \quad (j \neq k).$$

Here κ_k, $\bar{\lambda}_k$, and σ_k are the mean-reversion speed, long-run intensity, and volatility parameter, respectively, while $\beta_j\beta_k$ is the correlation coefficient between the Brownian motions W^j and W^k for $j \neq k$. Each coordinate of θ is the expected discounted net cash flow, consisting of protection payments minus premium payments, for one basket–maturity pair. To obtain neural predictions with different degrees of training–evaluation mismatch, we train DBDP models with $(\lambda^k_{0,\text{train}}, \bar{\lambda}_{k,\text{train}}, \sigma_{k,\text{train}}) = (s\lambda_0^k, s\bar{\lambda}_k, \sqrt{s}\sigma_k)$, where λ_0^k is the initial evaluation intensity and all other model parameters remain unchanged. We use $s \in \{1.00, 1.02, 1.05, 1.10, 1.25\}$, with $s = 1$ matching the evaluation model and larger values introducing greater parameter mismatch. Details of the methods and evaluation metrics are provided in Appendices C.2 and C.4.

Result 4: Applicability beyond option pricing. We keep the CDS evaluation dynamics and MC budget fixed and vary the training intensity multiplier s, with larger values above $s = 1$ introducing greater mismatch between training and evaluation. In Figure 5, we observe that NISE achieves lower Mahalanobis risk than MC in all cases, providing separate evidence of its effectiveness in credit derivative pricing. Moreover, as the parameter mismatch increases and neural predictions become less accurate, their mean weight decreases from 47.6% to 0.33%, illustrating how NISE adaptively reduces its reliance on unreliable predictions.

5.3 REAL-DATA OPTION PRICING UNDER CALIBRATED HESTON DYNAMICS

Setup. To assess the practical performance of NISE in a market-calibrated setting based on real-world market data, we consider European SPX call options using OptionsDX end-of-day option

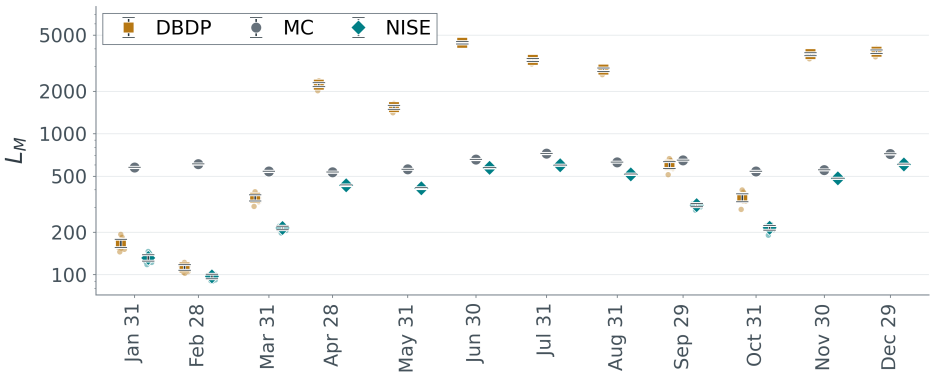

Figure 6: NISE achieves lower mean Mahalanobis loss than DBDP and MC at all twelve month-end valuation dates in 2023. Each point averages five independently evaluated runs with error bars showing one sample standard deviation across the five runs.

chains from the last available trading day of each month in 2023. For each of the twelve valuation dates, we hold out every fifth eligible expiry and retain all eligible strikes at those expiries, yielding a price vector θ of dimension J, where J ranges from 533 to 721 across valuation dates. The forward-normalized SPX index X_t and its instantaneous variance v_t follow risk-neutral Heston dynamics (Heston, 1993),

$$dX_t = X_t\sqrt{v_t}\,dW_t^S, \qquad dv_t = \kappa(\bar{v} - v_t)\,dt + \xi\sqrt{v_t}\,dW_t^v, \qquad d\langle W^S, W^v\rangle_t = \rho\,dt.$$

Here κ, $\bar{v}$, and ξ are the variance mean-reversion speed, long-run variance, and volatility-of-volatility parameter, respectively, while ρ is the instantaneous correlation between W^S and W^v. Each coordinate of θ is the time-zero price of a held-out call option with maturity T and strike K, $\theta_{(T,K)} = D(T)F(T)\,\mathbb{E}\left[\left(X_T - \frac{K}{F(T)}\right)_+\right]$, where $D(T)$ and $F(T)$ denote the discount factor and forward price for maturity T, respectively. For each valuation date, we calibrate the Heston model using market data from the non-held-out expiries and train five independent DBDP targets with different random seeds under the resulting calibrated dynamics to assess variability across training runs. We evaluate each target and its corresponding NISE estimator separately using $N = 2{,}000$ Monte Carlo paths per repetition. Experimental details are provided in Appendix C.5.

Result 5: Robustness with real-world market data. We lastly examine the performance of NISE under Heston dynamics calibrated to real-world market data across twelve month-end valuation dates in 2023. In Figure 6, the x-axis shows the valuation dates, while the y-axis reports the empirical mean Mahalanobis loss L_M for DBDP, MC, and NISE. As shown in the figure, the accuracy of the standalone DBDP predictions varies across valuation dates, with lower loss than MC on some dates but higher loss on others. More importantly, despite this variation in neural prediction accuracy, NISE achieves lower mean Mahalanobis loss than both DBDP and MC at all twelve valuation dates. Overall, these results empirically demonstrate the superior performance of NISE over MC in a real-world market setting.

6 Conclusion

In this work, we proposed Neural-Informed Shrinkage Estimation (NISE), a hybrid framework combining neural PDE solvers and Monte Carlo simulation for high-dimensional derivative pricing. Using independently trained neural predictions as shrinkage targets, NISE lets MC samples adaptively determine their influence through a positive-part James–Stein estimator. With exact covariance, we establish strict dominance over MC in Mahalanobis risk under Gaussianity for $J \geq 3$. Beyond Gaussianity, a finite third-moment condition yields an accuracy-dependent finite-sample bound with a target-uniform $O(N^{-1/4})$ upper bound on excess risk relative to MC. Experiments on basket options, CDS, and options under Heston dynamics calibrated to real-world market data demonstrate improved sample efficiency and robustness to inaccurate neural predictions. These findings highlight the value of statistically calibrating neural predictions for joint derivative pricing.

7 AI DISCLOSURE SECTION

In this work, we used generative AI tools for formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, design or provide feedback on research methodology or experiments, implement methods, support qualitative and thematic data analysis, interpret results. We have not used generative AI tools for propose or refine hypotheses, help develop theoretical models or conceptual frameworks, assist with translation, and generate synthetic data sets, clean and reformat dataset are not applicable to this work. Additionally, we used generative AI tools for identify relevant literature, format references, propose a title or keywords for a research paper. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Leif B. G. Andersen. Simple and efficient simulation of the Heston stochastic volatility model. *Journal of Computational Finance*, 11(3):1–42, 2008.

Yohai Bar-Sinai, Stephan Hoyer, Jason Hickey, and Michael P. Brenner. Learning data-driven discretizations for partial differential equations. *Proceedings of the National Academy of Sciences*, 116(31):15344–15349, 2019. doi: 10.1073/pnas.1814058116.

A. J. Baranchik. A family of minimax estimators of the mean of a multivariate normal distribution. *The Annals of Mathematical Statistics*, 41(2):642–645, 1970.

Christian Beck, Sebastian Becker, Patrick Cheridito, Arnulf Jentzen, and Ariel Neufeld. Deep splitting method for parabolic PDEs. *SIAM Journal on Scientific Computing*, 43(5):A3135–A3154, 2021.

Christian Beck, Martin Hutzenthaler, Arnulf Jentzen, and Benno Kuckuck. An overview on deep learning-based approximation methods for partial differential equations. *Discrete and Continuous Dynamical Systems – B*, 28(6):3697–3746, 2023.

Filipe De Avila Belbute-Peres, Thomas Economon, and Zico Kolter. Combining differentiable pde solvers and graph neural networks for fluid flow prediction. In *international conference on machine learning*, pp. 2402–2411. PMLR, 2020.

Julius Berner, Markus Dablander, and Philipp Grohs. Numerically solving parametric families of high-dimensional Kolmogorov partial differential equations via deep learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, 1973. doi: 10.1086/260062.

M. E. Bock. Minimax estimators of the mean of a multivariate normal distribution. *The Annals of Statistics*, 3(1):209–218, 1975.

Phelim Boyle, Mark Broadie, and Paul Glasserman. Monte Carlo methods for security pricing. *Journal of Economic Dynamics and Control*, 21(8–9):1267–1321, 1997.

Phelim P. Boyle. Options: A Monte Carlo approach. *Journal of Financial Economics*, 4(3):323–338, 1977.

Dominic Coey and Tom Cunningham. Improving treatment effect estimators through experiment splitting. In *The World Wide Web Conference*, pp. 285–295. ACM, 2019. doi: 10.1145/3308558.3313452.

John C Cox, Jonathan E Ingersoll, and Stephen A Ross. A theory of the term structure of interest rates. *Econometrica*, 53(2):385–407, 1985.

Darrell Duffie and Kenneth J. Singleton. Modeling term structures of defaultable bonds. *The Review of Financial Studies*, 12(4):687–720, 1999.

Weinan E, Jiequn Han, and Arnulf Jentzen. Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Communications in Mathematics and Statistics*, 5(4):349–380, 2017.

Bradley Efron. Machine learning and the James–Stein estimator. *Japanese Journal of Statistics and Data Science*, 7(1):257–266, 2024. doi: 10.1007/s42081-023-00209-y.

Bradley Efron and Carl Morris. Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

Bradley Efron and Carl Morris. Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

Maximilien Germain, Huyên Pham, and Xavier Warin. Neural networks-based algorithms for stochastic control and PDEs in finance. *arXiv preprint arXiv:2101.08068*, 2021.

Michael B. Giles. Multilevel Monte Carlo path simulation. *Operations Research*, 56(3):607–617, 2008.

Paul Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, New York, 2004.

Edwin J. Green and William E. Strawderman. A James-Stein type estimator for combining unbiased and possibly biased estimators. *Journal of the American Statistical Association*, 86(416):1001–1006, 1991. doi: 10.1080/01621459.1991.10475144.

Daniel Greenfeld, Meirav Galun, Ronen Basri, Irad Yavneh, and Ron Kimmel. Learning to optimize multigrid PDE solvers. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2415–2423, 2019.

Jeongmin Gu, Jose A. Iglesias-Guitian, and Bochang Moon. Neural James–Stein combiner for unbiased and biased renderings. *ACM Transactions on Graphics*, 41(6):1–14, 2022. doi: 10.1145/3550454.3555496.

Jiequn Han and Jihao Long. Convergence of the deep BSDE method for coupled FBSDEs. *Probability, Uncertainty and Quantitative Risk*, 5(1):5, 2020.

Jiequn Han, Arnulf Jentzen, and Weinan E. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34):8505–8510, 2018.

Steven L. Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2):327–343, 1993.

Jun-Ting Hsieh, Shengjia Zhao, Stephan Eismann, Lucia Mirabella, and Stefano Ermon. Learning neural PDE solvers with convergence guarantees. In *International Conference on Learning Representations*, 2019.

Côme Huré, Huyên Pham, and Xavier Warin. Deep backward schemes for high-dimensional nonlinear PDEs. *Mathematics of Computation*, 89(324):1547–1579, 2020.

W. James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 361–379. University of California Press, 1961.

Philippe Jorion. Bayes-Stein estimation for portfolio analysis. *Journal of Financial and Quantitative Analysis*, 21(3):279–292, 1986.

Seyedalireza Khoshsirat and Chandra Kambhamettu. Improving normalization with the James–Stein estimator. In *IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE, 2024.

Dmitrii Kochkov, Jamie A. Smith, Ayya Alieva, Qing Wang, Michael P. Brenner, and Stephan Hoyer. Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, 118(21):e2101784118, 2021. doi: 10.1073/pnas.2101784118.

David Lando. On Cox processes and credit risky securities. *Review of Derivatives Research*, 2(2–3):99–120, 1998.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Sida Li and Nikolaos Ignatiadis. Prediction-powered adaptive shrinkage estimation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34836–34875. PMLR, 2025.

Francis A. Longstaff and Eduardo S. Schwartz. Valuing American options by simulation: A simple least-squares approach. *The Review of Financial Studies*, 14(1):113–147, 2001.

Roger Lord, Remmert Koekkoek, and Dick van Dijk. A comparison of biased simulation schemes for stochastic volatility models. *Quantitative Finance*, 10(2):177–194, 2010.

Shaocong Ma, James Diffenderfer, Bhavya Kailkhura, and Yi Zhou. End-to-end mesh optimization of a hybrid deep learning black-box PDE solver. *arXiv preprint arXiv:2404.11766*, 2024.

Yuzo Maruyama. Special feature: Stein estimation and statistical shrinkage methods. *Japanese Journal of Statistics and Data Science*, 7(1):253–256, 2024. doi: 10.1007/s42081-024-00250-5.

Robert C. Merton. Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, 4(1):141–183, 1973. doi: 10.2307/3003143.

Thomas Müller, Fabrice Rousselle, Alexander Keller, and Jan Novák. Neural control variates. *ACM Transactions on Graphics*, 39(6):243:1–243:19, 2020.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Marc Sabate Vidales, David Šiška, and Lukasz Szpruch. Unbiased deep solvers for linear parametric PDEs. *Applied Mathematical Finance*, 28(4):299–329, 2021. doi: 10.1080/1350486X.2022.2030773. Published online 28 March 2022.

Justin Sirignano and Konstantinos Spiliopoulos. DGM: A deep learning algorithm for solving partial differential equations. *Journal of Computational Physics*, 375:1339–1364, 2018.

Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 197–206. University of California Press, 1956.

Kiwon Um, Robert Brand, Yun Raymond Fei, Philipp Holl, and Nils Thuerey. Solver-in-the-loop: Learning from differentiable physics to interact with iterative pde-solvers. *Advances in neural information processing systems*, 33:6111–6122, 2020.

Qi Wang, Pu Ren, Hao Zhou, Xin-Yang Liu, Zhiwen Deng, Yi Zhang, Ruizhi Chengze, Hongsheng Liu, Zidong Wang, Jian-Xun Wang, Ji-Rong Wen, Hao Sun, and Yang Liu. P^2c^2net: Pde-preserved coarse correction network for efficient prediction of spatiotemporal dynamics. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 68897–68925. Curran Associates, Inc., 2024. doi: 10.52202/079017-2201.

Appendix Contents

A Discussions on Relevant References **13**

B Main Theoretical Results **14**

B.1 Supporting Lemmas 14

B.2 Proof of Theorem 1 18

B.3 Proof of Theorem 2 18

C Experimental Details **19**

C.1 Discussion on Experimental Setup 20

C.2 Common evaluation protocol 20

C.3 Option pricing under correlated GBM 21

C.4 CDS pricing under correlated CIR intensities 21

C.5 Real-data option pricing under calibrated Heston dynamics 22

D Additional Experiments **23**

D.1 Covariance-source comparison under spot shifts 23

D.2 Effect of the independent-pilot sample size 24

D.3 Effect of the joint estimation dimension 25

D.4 Gaussian-risk approximation at fixed noncentrality 25

D.5 Supplementary diagnostics and robustness checks 27

E Additional Supplementary Materials **27**

E.1 Derivation from Equation (8) to Equation (9) 27

E.2 Difference from fixed convex combinations 28

E.3 Exact positive-part risk representation 29

A Discussions on Relevant References

Several existing methods also combine neural predictions with statistical estimation to exploit learned information while correcting or limiting the effects of prediction error (Green & Strawderman, 1991; Gu et al., 2022; Li & Ignatiadis, 2025). Related approaches also use neural approximations to construct control variates for Monte Carlo computation (Sabate Vidales et al., 2021; Müller et al., 2020). However, these results are substantially different from NISE in derivative pricing. We summarize these results as follows:

Green & Strawderman (1991); Gu et al. (2022). Gu et al. (2022) combines an unbiased but noisy rendered image with a biased denoised image using localized James–Stein shrinkage. Their framework employs a neural network to estimate local variance information and blending parameters, which are then used to construct the biased input to the combiner. However, the method is specifically tailored to image rendering and is not directly applicable to derivative pricing. In addition, our work provides a rigorous statistical analysis of the proposed methodology with and without Gaussianity, an aspect that is not addressed in Gu et al. (2022) or Green & Strawderman (1991).

Li & Ignatiadis (2025). This work proposes prediction-powered adaptive shrinkage, which first uses labeled observations to correct ML predictions within each mean-estimation task and then

shrinks the resulting estimates toward prediction-based targets across tasks. They tune shrinkage by minimizing a risk estimate that accounts for dependence between the corrected estimates and their targets, and demonstrate the method using neural predictions, including BERT-based predictions of product-review ratings. Compared to this method, NISE instead directly applies a positive-part James–Stein rule to a correlated MC price vector and an independent neural PDE target, with a closed-form shrinkage factor determined by their Mahalanobis discrepancy.

Sabate Vidales et al. (2021). This method learns parametric PDE solutions and their gradients, then use the learned gradients to construct martingale control variates for MC pricing. The martingale construction removes bias due to neural approximation under the stated integrability conditions, while the quality of the learned gradient determines the effectiveness of variance reduction. Instead, NISE uses only the neural price prediction at the valuation state as a shrinkage target, allowing bias in exchange for lower joint estimation risk without requiring neural-gradient evaluations along the simulated paths.

Müller et al. (2020). This work constructs neural control variates whose integrals are available by design, combining an integrable neural approximation with MC estimation of the remaining residual. Their core estimator preserves unbiasedness, and they additionally learn an importance sampler for the residual in light-transport applications. NISE does not require an integrable neural approximation to the payoff integrand, because the network output enters directly as a target for shrinking the estimated price vector. We also note that Müller et al. (2020) is also specially designed for graphics task and cannot potentially be transferred to the derivative pricing.

B Main Theoretical Results

Notation. For a matrix $A \in \mathbb{R}^{J\times J}$, $\|A\|_{\mathrm{op}} = \sup_{\|v\|=1}\|Av\|$ denotes the operator norm and $\operatorname{tr} A$ the trace. With $e_1, \dots, e_J$ denoting the standard basis,

$$|\operatorname{tr} A| = \Big|\sum_{i=1}^{J} e_i^\top A\, e_i\Big| \le \sum_{i=1}^{J}\|Ae_i\| \le J\,\|A\|_{\mathrm{op}}. \tag{15}$$

Use the following whitened, $\sqrt{N}$-standardized quantities:

$$\omega^{(n)} := \Sigma_Y^{-1/2}\big(Y^{(n)} - \theta\big), \qquad \zeta := \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\omega^{(n)} = \sqrt{N}\,\Sigma_Y^{-1/2}\big(\hat\theta - \theta\big), \tag{16}$$

$$\mu := \sqrt{N}\,\Sigma_Y^{-1/2}\big(\theta - \tilde\theta\big), \qquad z := \zeta + \mu = \sqrt{N}\,\Sigma_Y^{-1/2}\big(\hat\theta - \tilde\theta\big). \tag{17}$$

Then z is the whitened observation in the canonical problem of Section 4.1. By Assumption 1, the $\omega^{(n)}$ are i.i.d. with $\mathbb{E}[\omega^{(1)}] = 0$ and $\mathbb{E}[\omega^{(1)}\omega^{(1)\top}] = \Sigma_Y^{-1/2}\,\Sigma_Y\,\Sigma_Y^{-1/2} = I_J$, hence $\mathbb{E}[\zeta] = 0$ and $\mathrm{Cov}(\zeta) = I_J$. Jensen's inequality applied to $\mathbb{E}\|\omega^{(1)}\|^2 = \operatorname{tr} I_J = J$ gives the two moment bounds

$$\mathbb{E}\|\omega^{(1)}\| \le \big(\mathbb{E}\|\omega^{(1)}\|^2\big)^{1/2} = J^{1/2}, \qquad J^{3/2} = \big(\mathbb{E}\|\omega^{(1)}\|^2\big)^{3/2} \le \mathbb{E}\|\omega^{(1)}\|^3 = \gamma_3. \tag{18}$$

The second inequality follows because $s \mapsto s^{3/2}$ is convex on $[0,\infty)$. Likewise, $\mathbb{E}\|\zeta\| \le \big(\mathbb{E}\|\zeta\|^2\big)^{1/2} = J^{1/2}$. Define the shrinkage vector field $h : \mathbb{R}^J \to \mathbb{R}^J$ by

$$h(x) := -\min\Big(\frac{J-2}{\|x\|^2},\, 1\Big)\, x \quad (x \neq 0), \qquad h(0) := 0. \tag{19}$$

Let $\mathcal{S}_{J-2} := \{x \in \mathbb{R}^J : \|x\|^2 = J-2\}$. From Equation (19), $\|h(x)\| = \min\big(\|x\|, (J-2)/\|x\|\big) \le \sqrt{J-2}$ for every x.

B.1 Supporting Lemmas

We first restate the classical Gaussian shrinkage result using the notation needed for NISE.

Proposition 1 (James & Stein (1961); Baranchik (1970)). *Let $z \sim \mathcal{N}(\mu, I_J)$ with $J \geq 3$, and let $a \in \mathbb{R}^J$ be fixed. Set $q = \|z - a\|^2$ and*

$$\delta^+(z; a) := a + \left(1 - \frac{J-2}{q}\right)_+ (z - a).$$

For $q = 0$, define $\delta^+(z; a) = a$. Then $\mathbb{E}\|\delta^+(z; a) - \mu\|^2 < \mathbb{E}\|z - \mu\|^2 = J$ for every μ.

Lemma 1 (Whitened risk identities). *(a) $W = \|z\|^2$ and $\sqrt{N}\,\Sigma_Y^{-1/2}\big(\hat{\theta}_{\mathrm{NISE}} - \theta\big) = \zeta + h(z)$. (b) $L_M(\hat{\theta}, \theta) = \|\zeta\|^2$ and $\mathbb{E}\left[L_M(\hat{\theta}, \theta)\right] = J$. (c) $L_M(\hat{\theta}_{\mathrm{NISE}}, \theta) = \|\zeta + h(z)\|^2$, and its excess risk satisfies*

$$\mathbb{E}\left[L_M(\hat{\theta}_{\mathrm{NISE}}, \theta)\right] - J = \mathbb{E}\left[2\,\zeta^\top h(z)\right] + \mathbb{E}\left[\|h(z)\|^2\right].$$

Both expectations on the right-hand side are finite.

Proof. (a) Since $\Sigma = \Sigma_Y/N$, we have $\Sigma^{-1} = N\,\Sigma_Y^{-1}$, and therefore

$$W = (\hat{\theta} - \tilde{\theta})^\top \Sigma^{-1} (\hat{\theta} - \tilde{\theta}) = \left\|\sqrt{N}\,\Sigma_Y^{-1/2}(\hat{\theta} - \tilde{\theta})\right\|^2 = \|z\|^2.$$

Suppose first that $W > 0$. For every real $w > 0$, $(1 - (J-2)/w)_+ = 1 - \min((J-2)/w, 1)$. By Equation (12), $\hat{\theta}_{\mathrm{NISE}} - \theta = (\tilde{\theta} - \theta) + \big(1 - (J-2)/W\big)_+ (\hat{\theta} - \tilde{\theta})$, so multiplying by $\sqrt{N}\,\Sigma_Y^{-1/2}$ and using Equation (17) gives

$$\sqrt{N}\,\Sigma_Y^{-1/2}\big(\hat{\theta}_{\mathrm{NISE}} - \theta\big) = -\mu + \left(1 - \min\left(\frac{J-2}{\|z\|^2}, 1\right)\right) z = (z - \mu) + h(z) = \zeta + h(z).$$

If $W = 0$, then $z = 0$ and $\zeta = -\mu$; then $\hat{\theta}_{\mathrm{NISE}} = \tilde{\theta}$, so $\sqrt{N}\,\Sigma_Y^{-1/2}(\hat{\theta}_{\mathrm{NISE}} - \theta) = -\mu = \zeta + h(0) = \zeta + h(z)$.

(b) By Equation (16), $L_M(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^\top \Sigma^{-1} (\hat{\theta} - \theta) = \|\zeta\|^2$. Since $\mathbb{E}[\zeta] = 0$ and $\mathrm{Cov}(\zeta) = I_J$, $\mathbb{E}\|\zeta\|^2 = \mathrm{tr}\, I_J = J$.

(c) The first claim follows from (a) and $L_M(\hat{\theta}_{\mathrm{NISE}}, \theta) = \|\sqrt{N}\Sigma_Y^{-1/2}(\hat{\theta}_{\mathrm{NISE}} - \theta)\|^2$. Expanding the square gives $\|\zeta + h(z)\|^2 = \|\zeta\|^2 + 2\,\zeta^\top h(z) + \|h(z)\|^2$. Taking expectations term by term is legitimate because $\mathbb{E}|\zeta^\top h(z)| \leq \sqrt{J-2}\ \mathbb{E}\|\zeta\| \leq \sqrt{J(J-2)} < \infty$ and $\mathbb{E}\|h(z)\|^2 \leq J - 2 < \infty$. Subtracting $\mathbb{E}\|\zeta\|^2 = J$ gives the identity. □

Lemma 2 (The shrinkage field). *Let $J \geq 3$. (i) The field h is differentiable at every $x \notin \mathcal{S}_{J-2}$, with derivative*

$$Dh(x) = -I_J \quad (\|x\|^2 < J-2), \qquad Dh(x) = -\frac{J-2}{\|x\|^2}\left(I_J - \frac{2\,xx^\top}{\|x\|^2}\right) \quad (\|x\|^2 > J-2).$$

The derivative satisfies $\|Dh(x)\|_{\mathrm{op}} \leq 1$ at every such x, and $\|h(x) - h(y)\| \leq \|x - y\|$ for all $x, y \in \mathbb{R}^J$. (ii) For every $x \notin \mathcal{S}_{J-2}$, the risk integrand satisfies

$$2\ \mathrm{div}\, h(x) + \|h(x)\|^2 = \begin{cases} \|x\|^2 - 2J \leq -(J+2), & \|x\|^2 < J-2, \\ -(J-2)^2/\|x\|^2 < 0, & \|x\|^2 > J-2, \end{cases}.$$

Both branches satisfy the common quantitative bound

$$2\ \mathrm{div}\, h(x) + \|h(x)\|^2 \leq -\frac{(J-2)^2}{\|x\|^2 + J - 2}.$$

In addition, $|\mathrm{div}\, h(x)| \leq J$.

Proof. (i) On the open ball $\{\|x\|^2 < J-2\}$, which contains 0, $h(x) = -x$, so $Dh(x) = -I_J$ with operator norm 1. On the open set $\{\|x\|^2 > J-2\}$, $h(x) = -(J-2)x/\|x\|^2$, and differentiating gives

$$Dh(x) = -(J-2)\left(\frac{I_J}{\|x\|^2} - \frac{2\,xx^\top}{\|x\|^4}\right) = -\frac{J-2}{\|x\|^2}\left(I_J - \frac{2\,xx^\top}{\|x\|^2}\right).$$

The matrix $I_J - 2xx^\top/\|x\|^2$ is a reflection with operator norm 1, so $\|Dh(x)\|_{\mathrm{op}} = (J-2)/\|x\|^2 < 1$ there.

For the global Lipschitz bound, first note that h is continuous on $\mathbb{R}^J$ because its two formulas agree on $\mathcal{S}_{J-2}$. Fix $x \neq y$ and let $\phi(u) := h\big(x + u(y-x)\big)$ for $u \in [0,1]$. The map $u \mapsto \|x + u(y-x)\|^2$ is a quadratic polynomial, so it takes the value $J-2$ for at most two parameters u; these parameters split $[0,1]$ into at most three closed subintervals. On the interior of each subinterval, ϕ is continuously differentiable with $\|\phi'(u)\| = \|Dh\big(x + u(y-x)\big)\,(y-x)\| \le \|y-x\|$. Applying the mean value theorem to $u \mapsto a^\top \phi(u)$ for an arbitrary unit vector a shows that ϕ is Lipschitz on each closed subinterval with constant $\|y-x\|$. Summing over the subintervals via the triangle inequality gives $\|h(y) - h(x)\| \le \|y-x\| \sum_k (u_{k+1} - u_k) = \|y-x\|$.

(ii) On $\{\|x\|^2 < J-2\}$, $\operatorname{div} h(x) = -J$ and $\|h(x)\|^2 = \|x\|^2$, so $2\operatorname{div} h + \|h\|^2 \le (J-2) - 2J = -(J+2)$. On $\{\|x\|^2 > J-2\}$, $\operatorname{tr}\big(I_J - 2xx^\top/\|x\|^2\big) = J-2$, so $\operatorname{div} h(x) = -(J-2)^2/\|x\|^2$ and $\|h(x)\|^2 = (J-2)^2/\|x\|^2$. Therefore,

$$2\operatorname{div} h(x) + \|h(x)\|^2 = \frac{-2(J-2)^2 + (J-2)^2}{\|x\|^2} = -\frac{(J-2)^2}{\|x\|^2} < 0.$$

For the common quantitative bound, the outer branch obeys $-(J-2)^2/\|x\|^2 \le -(J-2)^2/(\|x\|^2 + J - 2)$. On the inner branch, using $\|x\|^2 < J-2$,

$$\|x\|^2 - 2J < -(J+2) < -(J-2) \le -\frac{(J-2)^2}{\|x\|^2 + J - 2}.$$

Both branches also give $|\operatorname{div} h(x)| \le J$: it equals J on the inner branch and is at most $J-2$ on the outer branch. □

Lemma 3 (Gaussian smoothing). *Fix $\varepsilon > 0$, let $g \sim \mathcal{N}(0, I_J)$, and define $h_\varepsilon(x) := \mathbb{E}\big[h(x + \varepsilon g)\big]$. Then: (i) h_ε is infinitely differentiable on $\mathbb{R}^J$; (ii) $\|h_\varepsilon(x)\| \le \sqrt{J-2}$ and $\|h_\varepsilon(x) - h(x)\| \le \varepsilon\sqrt{J}$ for all x; (iii) $Dh_\varepsilon(x) = \mathbb{E}\big[Dh(x+\varepsilon g)\big]$, $\|Dh_\varepsilon(x)\|_{\mathrm{op}} \le 1$, and $\|Dh_\varepsilon(x) - Dh_\varepsilon(y)\|_{\mathrm{op}} \le (2/\varepsilon)\,\|x-y\|$ for all x, y; (iv) for every $x \in \mathbb{R}^J$,*

$$2\operatorname{div} h_\varepsilon(x) + \|h_\varepsilon(x)\|^2 \le -(J-2)^2\,\mathbb{E}_g\left[\frac{1}{\|x+\varepsilon g\|^2 + J - 2}\right].$$

Proof. Write φ_ε for the $\mathcal{N}(0, \varepsilon^2 I_J)$ density, so that $h_\varepsilon(x) = \int_{\mathbb{R}^J} h(v)\,\varphi_\varepsilon(x-v)\,dv$ by symmetry of the density.

(i) The field h is bounded and measurable, with $\|h\| \le \sqrt{J-2}$, and every partial derivative of φ_ε is integrable and dominated locally uniformly in x by an integrable Gaussian envelope. Dominated convergence therefore permits differentiation under the integral sign to all orders, so h_ε is infinitely differentiable.

(ii) Jensen's inequality gives $\|h_\varepsilon(x)\| \le \mathbb{E}\|h(x+\varepsilon g)\| \le \sqrt{J-2}$. The Lipschitz bound of Lemma 2(i) gives $\|h_\varepsilon(x) - h(x)\| \le \mathbb{E}\|h(x+\varepsilon g) - h(x)\| \le \varepsilon\,\mathbb{E}\|g\| \le \varepsilon\,\big(\mathbb{E}\|g\|^2\big)^{1/2} = \varepsilon\sqrt{J}$.

(iii) For the identity, fix a direction v and $s \neq 0$; then

$$\frac{h_\varepsilon(x+sv) - h_\varepsilon(x)}{s} = \mathbb{E}\left[\frac{h(x+\varepsilon g + sv) - h(x+\varepsilon g)}{s}\right].$$

The integrand has norm at most $\|v\|$ by the Lipschitz bound, and as $s \to 0$ it converges almost surely to $Dh(x+\varepsilon g)\,v$ because $\mathcal{S}_{J-2}$ has Lebesgue measure zero and h is differentiable off this sphere by Lemma 2(i). Dominated convergence gives $Dh_\varepsilon(x)v = \mathbb{E}[Dh(x+\varepsilon g)v]$. The bound $\|Dh_\varepsilon(x)\|_{\mathrm{op}} \le 1$ follows because $\|Dh\|_{\mathrm{op}} \le 1$ almost everywhere.

For the Lipschitz bound on Dh_ε, use the convolution form $Dh_\varepsilon(x) = \int Dh(v)\,\varphi_\varepsilon(x-v)\,dv$, which equals $\mathbb{E}[Dh(x+\varepsilon g)]$ by a change of variables and symmetry of φ_ε. Fix a unit vector w. Differentiating under the integral sign, justified as in part (i), and using $\int \nabla\varphi_\varepsilon(u)\,du = 0$ gives

$$\partial_w Dh_\varepsilon(x) = \int Dh(v)\,\big(\nabla\varphi_\varepsilon(x-v)^\top w\big)\,dv = \int \big(Dh(v) + I_J\big)\,\big(\nabla\varphi_\varepsilon(x-v)^\top w\big)\,dv.$$

Since $\|Dh(v) + I_J\|_{\mathrm{op}} \le 2$ almost everywhere and $\nabla\varphi_\varepsilon(u) = -(u/\varepsilon^2)\,\varphi_\varepsilon(u)$,

$$\|\partial_w Dh_\varepsilon(x)\|_{\mathrm{op}} \le 2\int |\nabla\varphi_\varepsilon(u)^\top w|\,du = \frac{2}{\varepsilon^2}\,\mathbb{E}\left|(\varepsilon g)^\top w\right| = \frac{2}{\varepsilon}\,\mathbb{E}|g^\top w| = \frac{2}{\varepsilon}\sqrt{\frac{2}{\pi}} \;\le\; \frac{2}{\varepsilon}.$$

Here $g^\top w \sim \mathcal{N}(0,1)$ and $\mathbb{E}|\mathcal{N}(0,1)| = \sqrt{2/\pi} \le 1$. For arbitrary x, y, the fundamental theorem of calculus applied to $s \mapsto Dh_\varepsilon\big(y + s(x-y)\big)$ yields $\|Dh_\varepsilon(x) - Dh_\varepsilon(y)\|_{\mathrm{op}} \le (2/\varepsilon)\|x-y\|$.

(iv) Taking the trace in part (iii) and using linearity gives $\operatorname{div} h_\varepsilon(x) = \mathbb{E}\left[\operatorname{div} h(x+\varepsilon g)\right]$, with an integrable right-hand side because $|\operatorname{div} h| \le J$ by Lemma 2(ii). For any random vector V with $\mathbb{E}\|V\|^2 < \infty$, the identity $\|\mathbb{E}\,V\|^2 = \mathbb{E}\|V\|^2 - \mathbb{E}\|V - \mathbb{E}\,V\|^2$ gives $\|\mathbb{E}\,V\|^2 \le \mathbb{E}\|V\|^2$. Applying this inequality to $V = h(x+\varepsilon g)$ gives $\|h_\varepsilon(x)\|^2 \le \mathbb{E}\|h(x+\varepsilon g)\|^2$. Hence,

$$\begin{aligned} 2\,\operatorname{div} h_\varepsilon(x) + \|h_\varepsilon(x)\|^2 &\le \mathbb{E}\left[2\,\operatorname{div} h(x+\varepsilon g) + \|h(x+\varepsilon g)\|^2\right] \\ &\le -(J-2)^2\,\mathbb{E}_g\left[\frac{1}{\|x+\varepsilon g\|^2 + J - 2}\right]. \end{aligned}$$

The final inequality applies the common bound in Lemma 2(ii), since the event $x + \varepsilon g \in \mathcal{S}_{J-2}$ has probability zero. □

Lemma 4 (Leave-one-out bound on the Stein discrepancy)**.** *Let $M \ge 0$ and let $f : \mathbb{R}^J \to \mathbb{R}^J$ be bounded and continuously differentiable with $\|Df(x)\|_{\mathrm{op}} \le 1$ and $\|Df(x) - Df(y)\|_{\mathrm{op}} \le M\|x-y\|$ for all $x, y \in \mathbb{R}^J$. Then, with ζ as in Equation (16),*

$$\left|\,\mathbb{E}\left[\zeta^\top f(\zeta)\right] - \mathbb{E}\left[(\operatorname{div} f)(\zeta)\right]\right| \;\le\; \frac{3}{2}\,\frac{M\gamma_3}{\sqrt{N}}.$$

Proof. For $n = 1, \dots, N$, set $\zeta_{(n)} := \zeta - N^{-1/2}\omega^{(n)} = N^{-1/2}\sum_{m\neq n}\omega^{(m)}$, which is independent of $\omega^{(n)}$ by Assumption 1. All expectations below are finite because f and Df are bounded, $|\operatorname{div} f| \le J$ by Equation (15), and $\mathbb{E}\|\omega^{(1)}\|^3 = \gamma_3 < \infty$ by Assumption 1.

Step 1. By linearity, $\mathbb{E}[\zeta^\top f(\zeta)] = \sum_{n=1}^N N^{-1/2}\,\mathbb{E}\left[\omega^{(n)\top} f(\zeta)\right]$. Since $\omega^{(n)}$ has mean zero and is independent of $\zeta_{(n)}$, $\mathbb{E}\left[\omega^{(n)\top} f(\zeta_{(n)})\right] = \mathbb{E}\left[\omega^{(n)}\right]^\top \mathbb{E}\left[f(\zeta_{(n)})\right] = 0$, so

$$\mathbb{E}\left[\omega^{(n)\top} f(\zeta)\right] = \mathbb{E}\left[\omega^{(n)\top}\big(f(\zeta) - f(\zeta_{(n)})\big)\right]. \tag{20}$$

Step 2. Since f is continuously differentiable and $\zeta = \zeta_{(n)} + N^{-1/2}\omega^{(n)}$, the fundamental theorem of calculus gives

$$f(\zeta) - f(\zeta_{(n)}) = \int_0^1 Df\big(\zeta_{(n)} + uN^{-1/2}\omega^{(n)}\big)\,N^{-1/2}\omega^{(n)}\,du = N^{-1/2}\Big(Df(\zeta_{(n)}) + R_n\Big)\,\omega^{(n)},$$

where $R_n := \int_0^1 \left[Df\big(\zeta_{(n)} + uN^{-1/2}\omega^{(n)}\big) - Df(\zeta_{(n)})\right]du$ satisfies, by the Lipschitz hypothesis,

$$\|R_n\|_{\mathrm{op}} \le \int_0^1 M\,u\,N^{-1/2}\|\omega^{(n)}\|\,du = \frac{M}{2}\,N^{-1/2}\|\omega^{(n)}\|.$$

Step 3. For the leading term, independence of $\omega^{(n)}$ and $\zeta_{(n)}$ and $\mathbb{E}\left[\omega^{(n)}\omega^{(n)\top}\right] = I_J$ give, entrywise by Fubini's theorem,

$$\begin{aligned} &\mathbb{E}\left[\omega^{(n)\top} Df(\zeta_{(n)})\,\omega^{(n)}\right] \\ &\quad= \mathbb{E}\left[\operatorname{tr}\left(Df(\zeta_{(n)})\,\omega^{(n)}\omega^{(n)\top}\right)\right] \\ &\quad= \operatorname{tr}\Big(\mathbb{E}\left[Df(\zeta_{(n)})\right]\,\mathbb{E}\left[\omega^{(n)}\omega^{(n)\top}\right]\Big) \\ &\quad= \operatorname{tr}\Big(\mathbb{E}\left[Df(\zeta_{(n)})\right]\Big) = \mathbb{E}\left[(\operatorname{div} f)(\zeta_{(n)})\right]. \end{aligned}$$

Fubini's theorem applies because each entry is dominated by the integrable variable $\|\omega^{(n)}\|^2$. For the remainder, $|\mathbb{E}[\omega^{(n)\top} R_n\, \omega^{(n)}]| \le \mathbb{E}\left[\|\omega^{(n)}\|^2 \|R_n\|_{\mathrm{op}}\right] \le \frac{M}{2} N^{-1/2}\, \mathbb{E}\|\omega^{(n)}\|^3 = \frac{M}{2} N^{-1/2} \gamma_3$. Combining with Equation (20) and Step 2,

$$\left| \mathbb{E}\left[\omega^{(n)\top} f(\zeta)\right] - N^{-1/2}\, \mathbb{E}\left[(\operatorname{div} f)(\zeta_{(n)})\right] \right| \le N^{-1/2} \cdot \frac{M}{2}\, N^{-1/2} \gamma_3 = \frac{M\gamma_3}{2N}.$$

Summing over n and multiplying by $N^{-1/2}$,

$$\left| \mathbb{E}\left[\zeta^\top f(\zeta)\right] - \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}\left[(\operatorname{div} f)(\zeta_{(n)})\right] \right| \le N \cdot N^{-1/2} \cdot \frac{M\gamma_3}{2N} = \frac{M\gamma_3}{2\sqrt{N}}. \tag{21}$$

Step 4. By Equation (15) and the Lipschitz hypothesis, $|(\operatorname{div} f)(x) - (\operatorname{div} f)(y)| = |\operatorname{tr}\left(Df(x) - Df(y)\right)| \le J\, M\, \|x - y\|$, so Equation (18) gives, for each n,

$$\begin{aligned} &\left| \mathbb{E}\left[(\operatorname{div} f)(\zeta_{(n)})\right] - \mathbb{E}\left[(\operatorname{div} f)(\zeta)\right] \right| \\ &\quad \le JM\, \mathbb{E}\|\zeta - \zeta_{(n)}\| = JM\, N^{-1/2}\, \mathbb{E}\|\omega^{(n)}\| \\ &\quad \le \frac{M\, J^{3/2}}{\sqrt{N}} \le \frac{M\gamma_3}{\sqrt{N}}. \end{aligned}$$

Averaging over n and combining with Equation (21) via the triangle inequality gives

$$\left| \mathbb{E}\left[\zeta^\top f(\zeta)\right] - \mathbb{E}\left[(\operatorname{div} f)(\zeta)\right] \right| \le \frac{M\gamma_3}{2\sqrt{N}} + \frac{M\gamma_3}{\sqrt{N}} = \frac{3}{2} \frac{M\gamma_3}{\sqrt{N}}. \qquad \square$$

B.2 Proof of Theorem 1

Proof of Theorem 1. Condition on a realized target $\tilde{\theta}$; independence preserves the law $\hat{\theta} \sim \mathcal{N}(\theta, \Sigma)$. All expectations in this proof are conditional on $\tilde{\theta}$ until the final averaging step. Define the centered and whitened variables

$$z := \Sigma^{-1/2}(\hat{\theta} - \tilde{\theta}), \qquad \mu := \Sigma^{-1/2}(\theta - \tilde{\theta}), \qquad \zeta := z - \mu.$$

Then $z \sim \mathcal{N}(\mu, I_J)$, $\zeta \sim \mathcal{N}(0, I_J)$, $W = \|z\|^2$, and $\|\mu\|^2 = L_M(\tilde{\theta}, \theta)$. Using the field h in Equation (19), the transformed estimation error is

$$\Sigma^{-1/2}(\hat{\theta}_{\mathrm{NISE}} - \theta) = \zeta + h(z).$$

Consequently, $L_M(\hat{\theta}_{\mathrm{NISE}}, \theta) = \|\zeta + h(z)\|^2$ and $\mathbb{E}[L_M(\hat{\theta}, \theta)] = \mathbb{E}\|\zeta\|^2 = J$. By Lemma 2(i), h is bounded and globally Lipschitz, so coordinatewise Gaussian integration by parts gives $\mathbb{E}[\zeta^\top h(z)] = \mathbb{E}[(\operatorname{div} h)(z)]$. The sphere on which h is not differentiable has Gaussian probability zero, and the bounded field and derivatives ensure integrability. Expanding the squared loss and applying Lemma 2(ii) therefore yields

$$\begin{aligned} \mathbb{E}\left[L_M(\hat{\theta}_{\mathrm{NISE}}, \theta)\right] - \mathbb{E}\left[L_M(\hat{\theta}, \theta)\right] &= \mathbb{E}\left[2\,(\operatorname{div} h)(z) + \|h(z)\|^2\right] \\ &\le -(J-2)^2\, \mathbb{E}\left[\frac{1}{\|z\|^2 + J - 2}\right] \\ &\le -\frac{(J-2)^2}{L_M(\tilde{\theta}, \theta) + 2J - 2} < 0. \end{aligned}$$

The last inequality follows from Jensen's inequality for $u \mapsto u^{-1}$ and $\mathbb{E}\|z\|^2 = J + \|\mu\|^2$. This proves the conditional bound. For a random independent target, averaging this bound over $\tilde{\theta}$ also gives an unconditional excess-risk bound with right-hand side $-(J-2)^2\, \mathbb{E}[(L_M(\tilde{\theta}, \theta) + 2J - 2)^{-1}] < 0$. $\square$

B.3 Proof of Theorem 2

Proof of Theorem 2. Since $\hat{\Sigma} = \Sigma = \Sigma_Y / N$, the whitened quantities in Equations (16) and (17) satisfy

$$W = \|z\|^2, \qquad \|\mu\|^2 = N(\theta - \tilde{\theta})^\top \Sigma_Y^{-1} (\theta - \tilde{\theta}) = N\lambda = L_M(\tilde{\theta}, \theta).$$

The plain-MC risk identity follows from Lemma 1(b) and uses only Assumption 1.

To prove Equation (13), fix an auxiliary smoothing parameter $\varepsilon > 0$ and let h_ε be the smoothed field of Lemma 3. Define $f(x) := h_\varepsilon(x+\mu)$. By Lemma 3, f is bounded and infinitely differentiable, satisfies $\|Df(x)\|_{\mathrm{op}} \le 1$ and $\|Df(x)-Df(y)\|_{\mathrm{op}} \le (2/\varepsilon)\|x-y\|$, and obeys $(\operatorname{div} f)(x) = (\operatorname{div} h_\varepsilon)(x+\mu)$. Because $z = \zeta + \mu$, we also have $f(\zeta) = h_\varepsilon(z)$ and $(\operatorname{div} f)(\zeta) = (\operatorname{div} h_\varepsilon)(z)$.

Starting from Lemma 1(c), we bound the two terms. First, by Lemma 3(ii) and $\mathbb{E}\|\zeta\| \le \sqrt{J}$,

$$\begin{aligned}\mathbb{E}\left[2\,\zeta^\top h(z)\right] &\le \mathbb{E}\left[2\,\zeta^\top h_\varepsilon(z)\right] + 2\,\mathbb{E}\left[\|\zeta\|\,\|h(z)-h_\varepsilon(z)\|\right]\\ &\le \mathbb{E}\left[2\,\zeta^\top h_\varepsilon(z)\right] + 2\,\varepsilon\sqrt{J}\cdot\sqrt{J}\\ &= \mathbb{E}\left[2\,\zeta^\top h_\varepsilon(z)\right] + 2J\varepsilon.\end{aligned}$$

Second, since $\|h\| \le \sqrt{J-2}$ and $\|h_\varepsilon\| \le \sqrt{J-2}$ everywhere, the reverse triangle inequality and Lemma 3(ii) give $\big|\|h(z)\| - \|h_\varepsilon(z)\|\big| \le \varepsilon\sqrt{J}$.

$$\begin{aligned}\mathbb{E}\|h(z)\|^2 &\le \mathbb{E}\|h_\varepsilon(z)\|^2\\ &\quad + \mathbb{E}\left[\big(\|h(z)\| + \|h_\varepsilon(z)\|\big)\big|\|h(z)\| - \|h_\varepsilon(z)\|\big|\right]\\ &\le \mathbb{E}\|h_\varepsilon(z)\|^2 + 2\sqrt{J-2}\cdot\varepsilon\sqrt{J}\\ &\le \mathbb{E}\|h_\varepsilon(z)\|^2 + 2J\varepsilon.\end{aligned}$$

The last step uses $\sqrt{J-2} \le \sqrt{J}$. Third, applying Lemma 4 to f with $M = 2/\varepsilon$ gives

$$\mathbb{E}\left[2\,\zeta^\top h_\varepsilon(z)\right] = 2\,\mathbb{E}\left[\zeta^\top f(\zeta)\right] \le 2\,\mathbb{E}\left[(\operatorname{div} f)(\zeta)\right] + 3\cdot\frac{2}{\varepsilon}\cdot\frac{\gamma_3}{\sqrt{N}} = 2\,\mathbb{E}\left[(\operatorname{div} h_\varepsilon)(z)\right] + \frac{6\,\gamma_3}{\varepsilon\sqrt{N}}.$$

Finally, let the Gaussian vector g used for smoothing be independent of the payoff vectors. Applying Lemma 3(iv) first pointwise at z and then averaging over z gives

$$\begin{aligned}&\mathbb{E}\left[2\,(\operatorname{div} h_\varepsilon)(z) + \|h_\varepsilon(z)\|^2\right]\\ &\le -(J-2)^2\,\mathbb{E}\left[\frac{1}{\|z+\varepsilon g\|^2 + J - 2}\right]\\ &\le -\frac{(J-2)^2}{\mathbb{E}\|z+\varepsilon g\|^2 + J - 2}.\end{aligned}$$

The last inequality follows from Jensen's inequality for the convex map $u \mapsto u^{-1}$, followed by multiplication by $-(J-2)^2$. Because $z = \zeta + \mu$, both ζ and g are centered with covariance I_J, g is independent of ζ, and $\|\mu\|^2 = N\lambda$,

$$\mathbb{E}\|z+\varepsilon g\|^2 = J + N\lambda + J\varepsilon^2.$$

Collecting the preceding displays yields

$$\mathbb{E}\left[L_M(\hat\theta_{\mathrm{NISE}}, \theta)\right] - J \le 4J\varepsilon + \frac{6\gamma_3}{\varepsilon\sqrt{N}} - \frac{(J-2)^2}{N\lambda + 2J - 2 + J\varepsilon^2}, \qquad \varepsilon > 0. \tag{22}$$

The sum of the two positive error terms in Equation (22) is minimized at

$$\varepsilon_* := \left(\frac{3\gamma_3}{2J\sqrt{N}}\right)^{1/2},$$

which is positive by Equation (18). For this choice,

$$4J\varepsilon_* + \frac{6\gamma_3}{\varepsilon_*\sqrt{N}} = 4\sqrt{6}\,\sqrt{J\gamma_3}\,N^{-1/4}, \qquad J\varepsilon_*^2 = \frac{3\gamma_3}{2\sqrt{N}}.$$

Substitution into Equation (22) and $\mathbb{E}[L_M(\hat\theta, \theta)] = J$ prove Equation (13). □

C Experimental Details

This section provides the model, contract, training, and evaluation settings for the main-text experiments in Section 5. Additional experiments and robustness checks are collected separately in Appendix D.

C.1 Discussion on Experimental Setup

In this section, we clarify several experimental choices that are omitted from the main text due to space limitations.

Covariance estimation. Theorem 1 and Theorem 2 assume knowledge of the exact covariance $\Sigma = \Sigma_Y/N$. In practical implementations, however, the exact covariance may be unavailable and must instead be estimated, either from an independent pilot sample or from the same evaluation paths. To assess the impact of covariance estimation, we have separately compared these alternatives in Figure 7. The results show no significant differences among the covariance choices. Accordingly, in our main experiments, we use either a same-sample covariance estimate or an independent-pilot covariance estimate, depending on the experimental setting.

Training & Estimation Budget. N denotes the number of fresh MC paths used at evaluation time, and M denotes the covariance-estimation budget, while the neural target is trained independently and then frozen across repeated evaluations. Throughout the paper, our primary sample-efficiency comparisons focus on the number of fresh evaluation paths, rather than on the offline cost of training the neural target or estimating the covariance matrix. This choice reflects the intended deployment setting. In derivative-pricing applications, a market maker can train the neural solver offline, for example before the market opens using historical or simulated data, and then keep the trained model fixed throughout the trading session without additional fine-tuning. Consequently, the training cost is amortized over a potentially large number of subsequent pricing queries rather than incurred for each evaluation. Similarly, covariance estimation need not require an additional online simulation budget: the covariance of the MC estimator can be estimated from the same paths used to compute the MC price vector. Our covariance-source comparison in Appendix D.1 further verifies that this same-sample implementation yields behavior closely aligned with the exact- and independent-pilot covariance variants over most of the tested regime. Accordingly, the reported path-saving results are intended to measure online evaluation-sample efficiency, which is the computational resource most directly incurred for each new pricing problem.

C.2 Common evaluation protocol

For a given model and claim ladder, plain MC and NISE are evaluated on R paired repetitions using identical paths for the two estimators. In the GBM and CDS experiments, we estimate the ratio of expected losses

$$\frac{\mathbb{E}[L(\widehat{\theta}_{\mathrm{NISE}}, \theta)]}{\mathbb{E}[L(\widehat{\theta}_{\mathrm{MC}}, \theta)]}$$

under Mahalanobis loss or sum-MSE. For paired losses (X_r, Y_r) and $\widehat{\rho} = \bar{X}/\bar{Y}$, the reported ratio standard error is $\mathrm{sd}(X_r - \widehat{\rho} Y_r)/(\sqrt{R}\,\bar{Y})$, preserving the within-repetition covariance. Where stated, confidence intervals are paired normal intervals based on this standard error. The Heston experiment reports absolute mean losses with variability across five independently evaluated runs, as detailed in Appendix C.5. Target alone evaluates the frozen neural output as an estimator by itself. For Mahalanobis risk evaluation, GBM uses the analytic covariance, CDS uses a fixed covariance estimate from an independent reference simulation, and Heston uses covariance estimates from a separate independent pilot. Training and evaluation use disjoint seed namespaces. Neural targets are trained once and reused across repetitions.

Covariance implementations. To document how the empirical implementations relate to the exact-covariance theory, we distinguish the covariance used in shrinkage from the covariance used to score losses. The main GBM experiments in Figures 2 to 4 use the evaluation sample covariance in the shrinkage statistic. The CDS experiment uses the fixed reference covariance as an independent-pilot estimate in shrinkage. The Heston experiment applies shrinkage separately using a fixed pilot covariance. The estimated-covariance experiments examine the qualitative implications of Theorems 1 and 2 beyond their exact-covariance assumptions.

C.3 Option pricing under correlated GBM

Methods and evaluation metrics. MC estimates θ by averaging discounted payoffs over N independent simulated paths, with each path reused across all J maturities. For neural PDE pricing, we use deep splitting (Beck et al., 2021) and DBDP (Huré et al., 2020) to obtain price predictions $\tilde{\theta}$. The solvers are trained independently of the evaluation paths, and each prediction is held fixed across MC repetitions. NISE combines the MC estimate with a neural prediction through adaptive shrinkage. We note that NISE is agnostic to the choice of neural solver. In these experiments, we use a sample-covariance variant of NISE, estimating the covariance of the MC estimator from the same evaluation paths. We evaluate performance using Mahalanobis risk and sum-MSE, where Mahalanobis loss is computed using the analytic covariance, matching our theoretical analysis. For both metrics, we report risk ratios by dividing each estimator's mean loss by the corresponding loss of plain MC, so values below one indicate improvement over MC. This setup underlies the neural-accuracy, trajectory-budget, and distribution-shift experiments in Figures 2 to 4.

Model and reference quantities. The underlying is a $d = 100$ dimensional equicorrelated GBM under the risk-neutral measure,

$$\frac{dS_t^k}{S_t^k} = r\,dt + \sigma\,dW_t^k, \qquad d\langle W^j, W^k\rangle_t = \rho\,dt \quad (j \neq k),$$

with $r = 0.02$, $\sigma = 0.3$, $\rho = 0.3$, and $S_0^k = 100$. The primary claim vector consists of geometric-average basket calls at $J = 30$ maturities equally spaced on $[0.1, 2.0]$, with strike $K = 100$:

$$\theta_i = \mathbb{E}\left[e^{-r\tau_i}(G_{\tau_i} - K)_+\right], \qquad G_t = \left(\prod_{k=1}^{d} S_t^k\right)^{1/d}.$$

The geometric average is itself a GBM with $\sigma_G^2 = \sigma^2\{1 + (d-1)\rho\}/d$ and effective yield $q_G = (\sigma^2 - \sigma_G^2)/2$. Hence the price vector is available from Black–Scholes, shared-path payoff covariances are available from bivariate lognormal moments, and the dynamics are simulated exactly. The experiment therefore has closed-form truth, analytic covariance, and no time-discretization error.

Neural-target grid. The trained-target study contains 18 models: deep splitting and DBDP, each trained for $\{2, 10, 50\}$ epochs per backward step at training-time volatility multipliers $\{1.0, 1.2, 1.5\}$. The networks are two-layer MLPs of width 128, use fresh exact GBM samples at every epoch, warm-start across the backward recursion, and take a normalized log-geometric-mean input feature. Evaluation uses $N = 10^4$ and $R = 10^4$.

Trajectory-budget sweep. The in-distribution DBDP target is evaluated at spots $\{100, 105, 115\}$ over budgets from $N = 250$ to 64,000, with the target fixed within each spot regime. Figure 3 reports the sum-MSE of MC and NISE against N. For the path-savings calculation, let ρ_N denote the NISE sum-MSE divided by the MC sum-MSE at budget N; the corresponding path-saving factor is $1/\rho_N$. Since MC sum-MSE is proportional to $1/N$, the number of MC paths needed to match NISE's sum-MSE is N/ρ_N. These path-equivalent savings count evaluation trajectories and exclude the costs of training, reference simulation, target evaluation, and covariance computation.

Distribution-shift sweeps. Both solvers are trained at $S_0 = 100$ and $\sigma = 0.3$. Evaluation moves the spot over $S_0 \in [80, 120]$ and the market volatility over $\sigma \in [0.2, 0.45]$, while the solvers retain their training-time calibration. Each point uses $N = 10^4$ and $R = 10^4$.

C.4 CDS pricing under correlated CIR intensities

Methods and evaluation metrics. Conditional MC estimates θ by analytically integrating out default times given simulated intensity paths, with each path reused across all J contracts. We use independently trained DBDP (Huré et al., 2020) targets, held fixed across MC repetitions as in Section 5.1. MC, DBDP, and the reference simulations use the same monthly discretization, and estimation errors are evaluated against reference prices under the discretized evaluation model. For NISE, we use a fixed estimate of the MC estimator's covariance obtained from independent reference

simulations, both for shrinkage and for Mahalanobis scoring. Using $N = 2{,}000$ paths and $R = 2{,}000$ repetitions, we report the same Mahalanobis risk ratios as in Section 5.1, together with the mean neural-target weight.

Model and contracts. The credit experiment prices first-to-default basket CDS contracts on $d = 20$ obligors. Risk-neutral default intensities follow heterogeneous CIR processes with one-factor Brownian correlation (Cox et al., 1985) in the doubly stochastic framework (Lando, 1998; Duffie & Singleton, 1999):

$$d\lambda_t^k = \kappa_k(\bar{\lambda}_k - \lambda_t^k)\,dt + \sigma_k\sqrt{\lambda_t^k}\,dW_t^k, \qquad d\langle W^j, W^k\rangle_t = \beta_j\beta_k\,dt \quad (j \neq k).$$

The levels satisfy $\bar{\lambda}_k \in [0.012, 0.045]$, speeds $\kappa_k \in [0.7, 1.4]$, Feller ratios $2\kappa_k\bar{\lambda}_k/\sigma_k^2 = 2.5$, loadings $\beta_k \in [0.35, 0.65]$, and recoveries $R_k \in [0.35, 0.45]$.

The claim vector contains three overlapping ten-name baskets and ten maturities $\tau_M \in \{0.5, 1.0, \ldots, 5.0\}$, giving $J = 30$. Each coordinate is the fixed-coupon present value

$$\mathbb{E}\left[(1 - R_{k^*})e^{-r\eta(T^*)}\mathbf{1}\{T^* \le \tau_M\} - \frac{c}{4}\sum_{t_j \le \tau_M} e^{-rt_j}\mathbf{1}\{T^* > t_j\}\right] \times \text{notional},$$

with coupon $c = 5\%$, quarterly premium dates, notional 100, no accrued premium, and protection paid at the end of the monthly grid interval containing the first default. Conditional MC integrates the default clocks out analytically given the intensity paths.

Discretization and target construction. Because the correlated CIR system has no exact joint transition, the estimand is the discrete model generated by monthly full-truncation Euler steps (Lord et al., 2010); MC, the DBDP target, and the reference use the same transition. The target model scales intensities by $s \in \{1.00, 1.02, 1.05, 1.10, 1.25\}$ and volatilities by $\sqrt{s}$ to preserve the Feller ratios. Evaluation uses $N = 2{,}000$, $R = 2{,}000$, and a fixed independent-pilot covariance.

C.5 Real-data option pricing under calibrated Heston dynamics

Data and contract selection. We use OptionsDX SPX end-of-day option chains on the last observed trading day of each month in 2023. The twelve selected dates contain 97,575 raw quote rows. We fit discount factors and forwards by robust weighted put–call parity at each expiry, using at least eight valid call–put pairs with $|K/S_0 - 1| \le 0.12$. We retain expiry fits with normalized parity RMSE at most 5 and median bid–ask interval violation at most 0.50. Eligible option quotes have 21–365 calendar days to expiry, absolute log-forward moneyness at most 0.20, bid at least 0.05, midpoint at least 0.10, ask strictly above bid, and bid–ask spread at most 35% of the midpoint. We use the out-of-the-money instrument at each strike and convert put quotes to equivalent calls using the fitted parity relation. After these filters, 41,477 quotes remain. On each date, we sort eligible expiries and hold out the fifth, tenth, and subsequent every-fifth expiry from Heston parameter calibration. All eligible strikes at these expiries enter the evaluation vector, and all remaining eligible quotes enter calibration. This gives 34,221 calibration quotes and 7,256 evaluation contracts across the twelve dates; Table 1 reports the date-specific counts. The expiry holdout applies to the Heston parameter fit; discount and forward curves use eligible parity pairs, including those at evaluation expiries.

Calibration and numerical pricing target. For each date, we fit $(v_0, \kappa, \bar{v}, \xi, \rho)$ by robust nonlinear least squares with a soft-ℓ_1 loss, using two initializations and at most 150 objective evaluations per initialization. Price residuals are divided by the bid–ask spread, floored at 0.05, and weighted by the inverse square root of the number of calibration quotes at the expiry, up to a common normalization. The parameter bounds are $v_0, \bar{v} \in [0.0025, 0.25]$, $\kappa \in [0.05, 10]$, $\xi \in [0.05, 3]$, and $\rho \in [-0.999, 0]$. Calibration prices use Heston Fourier integration with 512 quadrature nodes and integration limit 400. Among converged fits, we select the one with the smallest mean squared weighted residual; all twelve selected fits meet the optimizer's convergence criterion. The computational pricing target is the expectation under the calibrated model discretized by the martingale-corrected Andersen quadratic-exponential scheme (QE-M) (Andersen, 2008). The grid has 96 steps per year with all evaluation expiries inserted, and uses $\psi_c = 1.5$ and $\gamma_1 = \gamma_2 = 0.5$. Neural training uses the appropriate prefix

Table 1: Market-data coverage for the twelve month-end pricing problems. Evaluation contracts are counted separately on each valuation date.

Valuation date	Calibration quotes	Evaluation contracts	Evaluation expiries
2023-01-31	2,611	574	5
2023-02-28	2,672	607	5
2023-03-31	2,939	539	6
2023-04-28	2,773	533	5
2023-05-31	2,729	557	5
2023-06-30	2,679	652	5
2023-07-31	2,605	721	5
2023-08-31	2,914	625	5
2023-09-29	3,015	646	6
2023-10-31	3,097	538	6
2023-11-30	2,954	550	6
2023-12-29	3,233	714	6
Total	34,221	7,256	65

of this same date-specific grid for each expiry. MC paths are shared across all contracts on a date. For strikes below the forward, we simulate discounted put payoffs and add the deterministic parity term $D(T)(F(T) - K)$ to obtain call estimates; for other strikes, we simulate discounted call payoffs directly. Reference prices use an independent simulation of 10^6 paths. All evaluation, covariance, and reference simulations use double precision. Thus the reported losses measure numerical estimation error conditional on the calibrated, discretized model. They do not measure errors against market quotes or the continuous-time model.

DBDP targets. For each date and held-out expiry, we train five DBDP solvers (Huré et al., 2020) with independent seeds, using the date's calibrated parameters and curves. Each solver learns the normalized call-price surface C/K as a function of log-spot moneyness and instantaneous variance, shared across all eligible strikes at that expiry. The networks have two hidden layers of 128 units with $\tanh$ activations. At each backward time step, training uses 20,000 sampled states per epoch, minibatches of 2,048, and Adam with initial learning rate 10^{-3}, reduced to 2×10^{-4} and 4×10^{-5} after 50% and 80% of the epochs, respectively. The first regression backward from maturity uses 300 epochs, and each subsequent regression uses 50 epochs with a warm start that accounts for changes in input and output normalization. Training uses single precision. At valuation, the resulting predictions satisfy the call-price bounds $D(T)(F(T) - K)_+ \leq \widetilde{\theta}_{(T,K)} \leq D(T)F(T)$. This formula parameterizes the output; training targets come from the Heston backward recursion, without Fourier prices or evaluation reference prices as labels. For each seed, we concatenate predictions across all held-out expiries into one complete date-specific target vector and construct NISE from that vector. The five prediction vectors are evaluated separately, without averaging them into an ensemble. All twelve dates and all five seeds are retained, giving 325 date–expiry–seed solvers and 60 evaluations of complete target vectors.

D Additional Experiments

In this section, we include additional experiments to examine how covariance estimation, the independent-pilot sample size, and the joint estimation dimension affect NISE, assess the Gaussian-risk approximation, and report supplementary checks for the main experiments.

D.1 Covariance-source comparison under spot shifts

Setup. To assess the effect of covariance estimation, we use the GBM option-pricing setup in Section 5.1, with model and training details given in Appendix C.3. We retain the 50-epoch DBDP and deep-splitting models trained at $S_0 = 100$ and $\sigma = 0.3$, and vary the evaluation spot over $S_0 \in \{80, 85, \ldots, 120\}$. At each spot, their scheme-consistent predictions are held fixed across covariance choices and MC repetitions. We compare three choices for the covariance of the MC estimator: the exact covariance Σ_Y/N, a sample covariance estimated from independent pilot paths

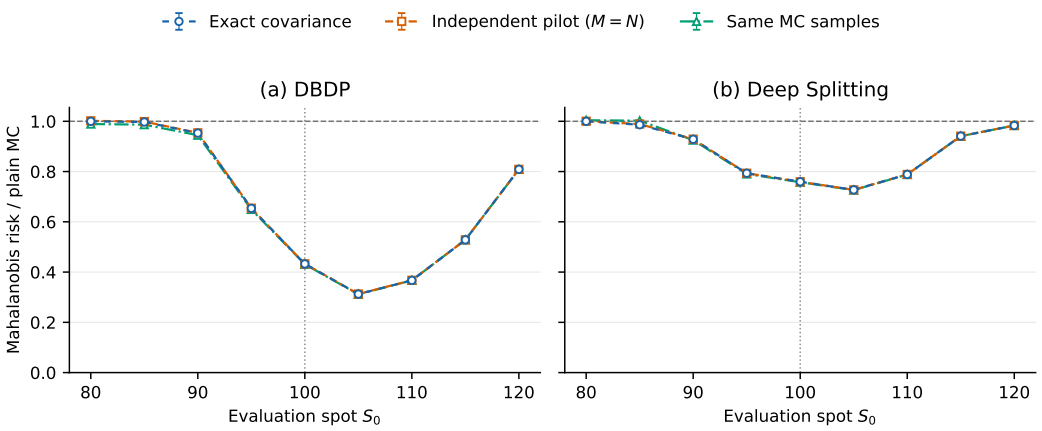

Figure 7: Exact, independent-pilot, and same-sample covariance choices yield similar risk reductions over most of the GBM spot grid. The left and right panels use frozen DBDP and deep-splitting targets, respectively; whiskers show pointwise paired 95% confidence intervals for the Mahalanobis risk ratio relative to MC.

and divided by N, and a sample covariance estimated from the MC evaluation paths and divided by N. Each spot uses $N = 10{,}000$ evaluation paths per repetition, $M = 10{,}000$ independent pilot paths, and $R = 10{,}000$ repetitions. The evaluation paths are shared across maturities, covariance choices, and solvers; a fresh pilot is generated in each repetition and shared across solvers. We evaluate the Mahalanobis risk ratio r_{NISE} from Equation (14) using the analytic covariance and closed-form prices at each spot for all methods. Pointwise paired 95% confidence intervals follow Appendix C.2 and are conditional on the frozen models.

Results. In Figure 7, the x-axis shows the evaluation spot S_0 and the y-axis reports r_{NISE}, with DBDP on the left and deep splitting on the right. The three covariance choices give closely aligned curves over most of the spot grid, with similar risk reductions relative to MC. At $S_0 = 100$, the exact, pilot, and same-sample risk ratios are respectively 0.432519, 0.432789, and 0.430923 for DBDP, and 0.759452, 0.759432, and 0.757572 for deep splitting. The same-sample variant has slightly lower risk than the exact-covariance variant at this spot: their paired differences are -0.001596 for DBDP and -0.001881 for deep splitting, with 54-comparison Bonferroni 95% intervals $[-0.001888, -0.001305]$ and $[-0.002052, -0.001710]$. These results show that empirical covariance estimates preserve most of the observed risk reduction and perform nearly as well as the exact covariance.

D.2 Effect of the independent-pilot sample size

Setup. We next examine how the number of independent pilot paths affects NISE. We use the setup in Appendix D.1, retaining its frozen neural models, spot grid, evaluation budget $N = 10{,}000$, and $R = 10{,}000$ repetitions, and vary only the pilot size over $M \in \{100, 500, 2000, 10{,}000\}$. Within each repetition, all sizes use the same evaluation mean and nested prefixes of a fresh independent pilot of $10{,}000$ paths; the covariance estimate is the pilot sample covariance divided by the evaluation budget N, irrespective of M.

Results. In Figure 8, the axes and panels follow Figure 7, with darker orange indicating a larger pilot size. Larger pilots yield greater risk reductions around the training spot, with diminishing gains as M increases. At $S_0 = 100$, increasing M from 100 to $10{,}000$ lowers the DBDP risk ratio from 0.579210 to 0.432789 and the deep-splitting ratio from 0.820008 to 0.759432. At $M = 2000$, the corresponding ratios are 0.434332 and 0.759933, already close to the largest-pilot results at this spot. The $M = 10{,}000$ losses reproduce the independent-pilot variant in Figure 7 at every repetition.

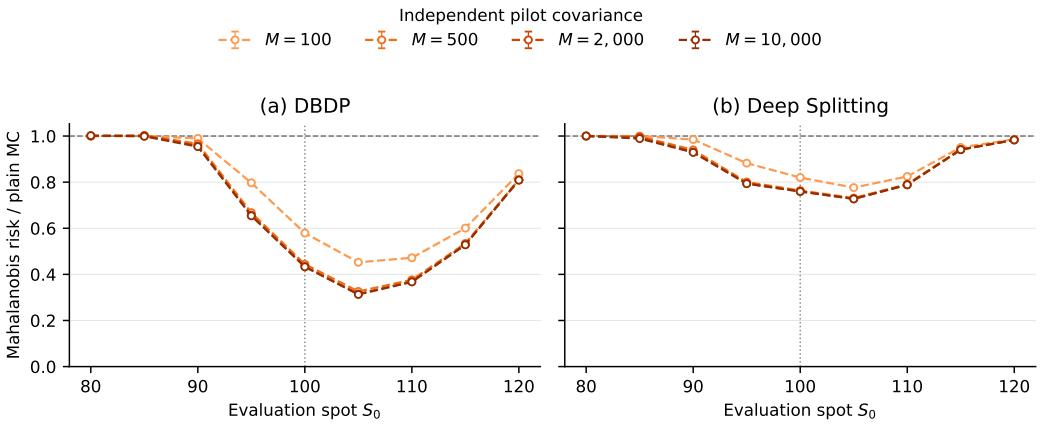

Figure 8: Larger independent pilots improve the risk reduction near the training spot, with diminishing gains as M increases. Darker orange denotes a larger pilot, $M \in \{100, 500, 2000, 10{,}000\}$, at fixed evaluation budget $N = 10{,}000$.

D.3 Effect of the joint estimation dimension

Setup. To examine the effect of jointly estimating more prices, we use the GBM dynamics and geometric-basket payoff in Section 5.1 and appendix C.3, keeping the asset dimension fixed at $d = 100$ and varying the number of contracts over $J \in \{3, 5, 10, 20, 30, 50, 100\}$. The maturities are predetermined nested subsets of a common grid of 100 equally spaced points on $[0.1, 2.0]$ years. Every subset includes both endpoints, and additional maturities are spread across the interval using a base-two van der Corput ordering rounded to grid indices, without observing neural predictions or evaluation errors. We train five new, independent DBDP models on the full grid and reuse each model's predictions across all subsets, so the training problem and the number of backward steps remain fixed as J changes. The network architecture follows Appendix C.3; each model uses 100,000 fresh paths per epoch, minibatches of 2,048, Adam with learning rate 10^{-3}, and 50 epochs per backward step, increased to 250 at the terminal step. Independent held-out residuals are monitored without selecting checkpoints or changing the training samples.

For evaluation, we use $N = 10{,}000$ paths per estimate and $R = 10{,}000$ independent repetitions, sharing the evaluation paths across dimensions and training seeds. Each subset uses its own analytic covariance Σ_Y/N for both shrinkage and Mahalanobis scoring. For each seed, we compute the ratio of mean losses relative to MC as in Equation (14); the plotted points average these ratios over the five seeds, and error bars show one sample standard deviation across seeds, rather than confidence intervals.

Results. In Figure 9, the x-axis shows the joint estimation dimension J and the y-axis reports Mahalanobis risk relative to MC, with both axes on logarithmic scales. The seed-averaged NISE risk is below both MC and standalone DBDP at every tested dimension. Its risk ratio decreases from 0.896 at $J = 3$ to 0.287 at $J = 30$ and 0.132 at $J = 100$, corresponding to progressively larger relative risk reductions along this nested contract family. The standalone DBDP ratio also decreases, from 2.099 at $J = 3$ to 0.137 at $J = 100$, and the two methods become close at large J.

D.4 Gaussian-risk approximation at fixed noncentrality

Setup. To assess the non-Gaussian correction error term $4\sqrt{6}\,\sqrt{J\gamma_3}\,N^{-1/4}$ presented in Theorem 2, we follow the same experiment setting of Section 5.1 and appendix C.3, replacing the neural prediction by a constructed target with fixed noncentrality, to validate the relation between the risk gap and the MC budget N. Let $\theta^{(0.05)}$ be the closed-form option ladder obtained after multiplying the GBM volatility by 1.05, and set $b = \theta^{(0.05)} - \theta$. For each $N \in \{250, 500, 1000, 2000, 4000, 8000\}$, we use the target $t_N = \theta + b\left\{\frac{J}{Nb^\top \Sigma_Y^{-1} b}\right\}^{1/2}$. The target's error direction is fixed, while its magnitude

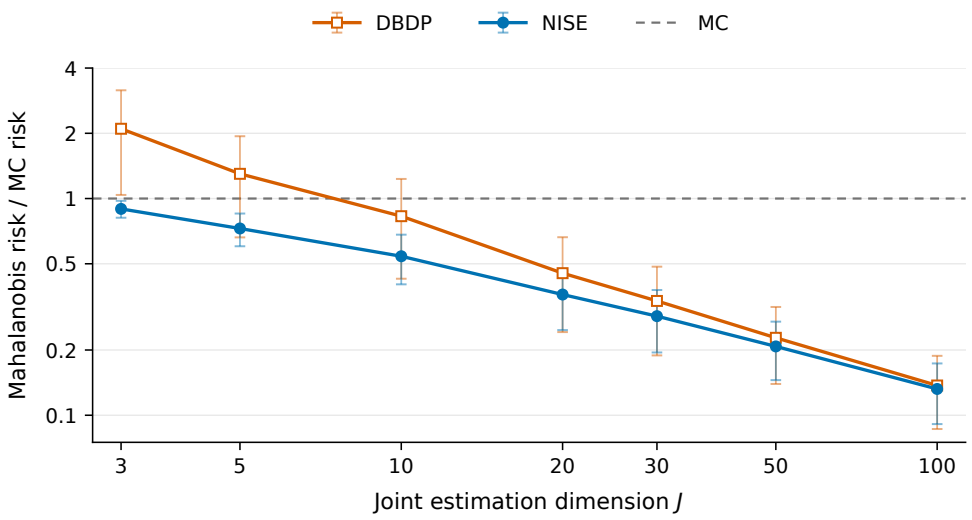

Figure 9: NISE achieves lower seed-averaged Mahalanobis risk than MC and standalone DBDP across the tested joint estimation dimensions, with asset dimension fixed at $d = 100$. Both axes are logarithmic, and each point averages five independently trained targets. Error bars represent the arithmetic mean plus or minus one sample standard deviation across training seeds.

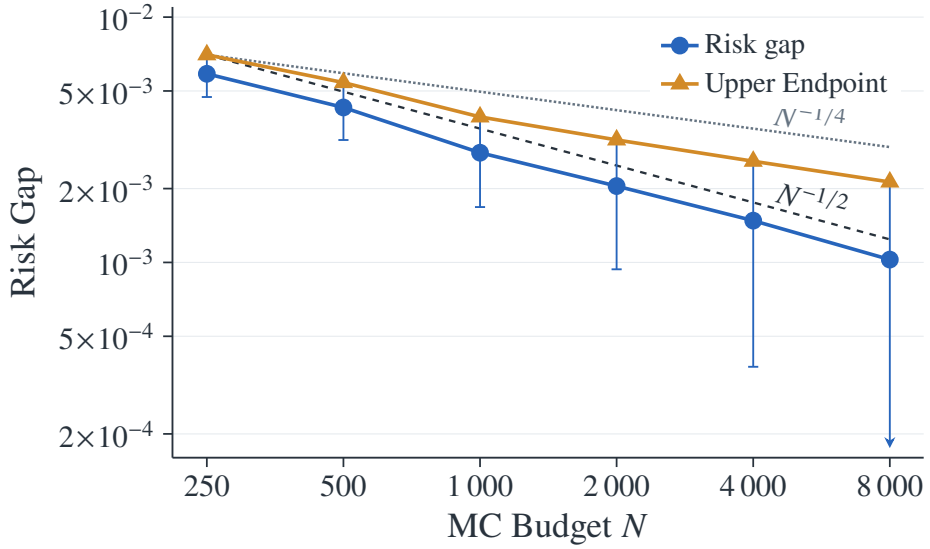

Figure 10: The Gaussian-risk approximation improves with the MC budget at fixed $\lambda = J$. The orange curve connects the upper endpoints; the $N^{-1/4}$ and $N^{-1/2}$ guides are anchored at $N = 250$.

changes with N so that $\lambda = N(t_N - \theta)^\top \Sigma_Y^{-1}(t_N - \theta) = J = 30$ at every budget. The estimator uses the analytic covariance Σ_Y/N. Write $R_N(t_N)$ for its finite-sample Mahalanobis risk and $R_+(J)$ for the exact positive-part Gaussian risk in Equation (37). We measure the risk gap by $|R_N(t_N) - R_+(J)|/J$, normalized by the exact plain-MC risk J. We use $R = 100{,}000$ independent repetitions in ten batches of 10,000; within a repetition, all budgets use nested prefixes of the same 8,000-path sample.

Results. In Figure 10, the x-axis shows the MC budget N and the y-axis reports the absolute normalized risk gap. The blue points show the estimated gaps, while the orange curve connects their simultaneous-interval upper endpoints.

D.5 Supplementary diagnostics and robustness checks

Setup. We use the GBM, CDS, and Heston setups in Sections 5.1 to 5.3 to examine estimation precision and sensitivity to numerical choices. For GBM, we retain the main experiment's covariance implementation in Appendix C.2, report uncertainty for the neural-accuracy and distribution-shift results in Figures 2 and 4, and evaluate the best target at the smaller budget $N = 10^3$. We also compare zero and grand-mean shrinkage centers to assess how much of the improvement depends on the neural prediction. For CDS, we refine the monthly simulation grid in Appendix C.4 to a semi-monthly grid and use a paired comparison to measure the change. For Heston, we repeat the scoring with a second independent 10^6-path reference, retaining all other settings in Appendix C.5.

Results. The GBM diagnostics confirm that the largest gains occur with informative predictions. For the best target in Figure 2, the paired standard error of r_{NISE} is 0.001058 and its 95% interval is 0.428201–0.432350; for the worst target, they are 0.000147 and 0.998068–0.998646. At $N = 10^3$, the best target gives a ratio of 0.099. By comparison, zero and grand-mean centers change risk relative to MC by less than 0.4%. The CDS grid refinement has Mahalanobis size 0.07 against an MC risk of 30, indicating a small discretization effect on the scale of the reported estimation risk. For Heston, NISE has lower mean Mahalanobis loss than MC in all 60 date–seed comparisons under both independent reference vectors. The observed ordering of the two estimators is therefore unchanged by the reference-price check.

E Additional Supplementary Materials

In this additional supplementary material, we provide several discussions omitted in our main text and previous sections.

E.1 Derivation from Equation (8) to Equation (9)

Suppose that $\hat{\theta} \sim \mathcal{N}(\theta, \Sigma)$, where Σ is positive definite, and let $a \in \mathbb{R}^J$ be fixed. Using the symmetric positive-definite square root of Σ, the centered and whitened estimator in Equation (8) satisfies

$$Z_a = \Sigma^{-1/2}(\hat{\theta} - a) \sim \mathcal{N}(\mu_a, I_J), \qquad \mu_a = \Sigma^{-1/2}(\theta - a). \tag{23}$$

Thus estimating θ by shrinking toward a is equivalent to estimating μ_a by shrinking Z_a toward the origin.

For $J \geq 3$, consider the canonical James–Stein estimator

$$d(z) = z + g(z), \qquad g(z) = -\frac{J-2}{\|z\|^2} z. \tag{24}$$

For $J \geq 3$ and $Z \sim \mathcal{N}(\mu, I_J)$, Stein's identity gives the risk representation (Stein, 1956; James & Stein, 1961)

$$\mathbb{E}\|d(Z) - \mu\|^2 = J + \mathbb{E}\left[\|g(Z)\|^2 + 2\ \mathrm{div}\, g(Z)\right]. \tag{25}$$

Writing $r^2 = \|z\|^2$, direct differentiation yields

$$\|g(z)\|^2 = \frac{(J-2)^2}{r^2}, \qquad \mathrm{div}\, g(z) = -(J-2)\sum_{i=1}^{J}\left(\frac{1}{r^2} - \frac{2z_i^2}{r^4}\right) = -\frac{(J-2)^2}{r^2}. \tag{26}$$

Substituting Equation (26) into Equation (25) gives

$$\mathbb{E}\|d(Z) - \mu\|^2 = J - (J-2)^2\,\mathbb{E}\left[\frac{1}{\|Z\|^2}\right]. \tag{27}$$

Truncating its shrinkage factor at zero gives the positive-part refinement

$$d^+(Z_a) = \left(1 - \frac{J-2}{\|Z_a\|^2}\right)_+ Z_a, \tag{28}$$

At $Z_a = 0$, we define $d^+(Z_a) = 0$. This truncation avoids reversing the direction of Z_a when $\|Z_a\|^2 < J - 2$ and improves upon the ordinary rule under quadratic loss (Baranchik, 1970).

The inverse transformation maps an estimator d of μ_a to the price-space estimator $a + \Sigma^{1/2}d$. Applying it to Equation (28) gives

$$\begin{aligned}\delta_a &= a + \Sigma^{1/2} d^+(Z_a) \\ &= a + \left(1 - \frac{J-2}{\|Z_a\|^2}\right)_+ (\hat{\theta} - a).\end{aligned} \tag{29}$$

Moreover,

$$\|Z_a\|^2 = (\hat{\theta} - a)^\top \Sigma^{-1} (\hat{\theta} - a) = W_a, \tag{30}$$

so Equation (29) is exactly Equation (9). The same change of coordinates maps Euclidean squared error in the whitened space to Mahalanobis loss in the price space:

$$\left\|\Sigma^{-1/2}(\delta - \theta)\right\|^2 = (\delta - \theta)^\top \Sigma^{-1} (\delta - \theta) = L_M(\delta, \theta). \tag{31}$$

Consequently, the classical James–Stein risk comparison in the whitened coordinates translates directly into the covariance-adjusted risk comparison used for NISE.

E.2 Difference from fixed convex combinations

Unlike a fixed convex combination, the estimator in Equation (9) is not affine in the target a and the MC estimate $\hat{\theta}$. For $W_a > 0$, define

$$\alpha_a = \left(1 - \frac{J-2}{W_a}\right)_+, \tag{32}$$

and set $\alpha_a = 0$ when $W_a = 0$. For each realized MC sample, $\alpha_a \in [0, 1]$, and the estimator can be written pointwise as

$$\delta_a = (1 - \alpha_a)a + \alpha_a \hat{\theta}. \tag{33}$$

However, α_a depends on both inputs through $W_a = (\hat{\theta} - a)^\top \Sigma^{-1} (\hat{\theta} - a)$. Consequently, the correction term

$$\alpha_a(\hat{\theta} - a) = \left(1 - \frac{J-2}{(\hat{\theta} - a)^\top \Sigma^{-1} (\hat{\theta} - a)}\right)_+ (\hat{\theta} - a)$$

is nonlinear in the discrepancy $\hat{\theta} - a$, so the map $(a, \hat{\theta}) \mapsto \delta_a$ is not affine even though each realized output lies on the line segment between a and $\hat{\theta}$.

This non-affine dependence produces the desired endpoint fallback. When $W_a \le J - 2$, the rule sets $\alpha_a = 0$ and returns the target exactly. As $W_a/(J-2) \to \infty$, the weight satisfies $\alpha_a \to 1$ and $\delta_a \to \hat{\theta}$, so a large standardized disagreement suppresses the target and recovers plain MC in the limit.

For comparison, consider the fixed-weight estimator

$$\delta_\rho = (1 - \rho)a + \rho\hat{\theta}, \qquad 0 \le \rho \le 1. \tag{34}$$

Let $\lambda = (\theta - a)^\top \Sigma^{-1} (\theta - a)$ denote the target's Mahalanobis loss. Under the Gaussian setting of this section, the risk of Equation (34) is

$$\mathbb{E}\, L_M(\delta_\rho, \theta) = \rho^2 J + (1 - \rho)^2 \lambda. \tag{35}$$

Its minimizing weight is $\rho^* = \lambda/(J + \lambda)$, which is unavailable because it depends on the unknown θ through λ. When $\lambda = 0$, this oracle weight is zero and the exact target is optimal; as $\lambda \to \infty$, it approaches one and plain MC becomes optimal in the limit within the fixed-weight class. These oracle endpoints explain why the two fallback regimes above are desirable: the James–Stein rule selects the target at small observed discrepancy and approaches MC at large observed discrepancy without requiring the unknown λ. Its Gaussian dominance follows from the specific data-dependent weight rather than from convex averaging alone.

E.3 EXACT POSITIVE-PART RISK REPRESENTATION

For a realized target $\tilde{\theta}$, define

$$Z_a := \Sigma^{-1/2}(\hat{\theta} - \tilde{\theta}), \qquad \mu_a := \Sigma^{-1/2}(\theta - \tilde{\theta}).$$

Under the assumptions of Theorem 1, $Z_a \sim \mathcal{N}(\mu_a, I_J)$. Define the target noncentrality $\lambda := \|\mu_a\|^2$ and the positive-part factor

$$A(0) := 0, \qquad A(w) := \left(1 - \frac{J-2}{w}\right)_+ \quad (w > 0). \tag{36}$$

Choose an orthogonal matrix U such that $U\mu_a = \sqrt{\lambda}\, e_1$, with an arbitrary orthogonal U when $\lambda = 0$, and write $\bar{Z} := UZ_a$. Rotational invariance gives $\bar{Z} \sim \mathcal{N}(\sqrt{\lambda}\, e_1, I_J)$. Therefore, with $Z_1 := \bar{Z}_1$ and $Q := \sum_{j=2}^{J} \bar{Z}_j^2$,

$$Z_1 \sim \mathcal{N}(\sqrt{\lambda}, 1), \qquad Q \sim \chi^2_{J-1}, \qquad Z_1 \perp Q, \qquad W = \|Z_a\|^2 = Z_1^2 + Q.$$

In these rotated coordinates, the positive-part estimator of $U\mu_a$ is $A(W)\bar{Z}$, where A is defined in Equation (36). Its Euclidean risk is equal to the conditional Mahalanobis risk of NISE. Both risks have the representation

$$\begin{aligned} R_+(\lambda) &= \mathbb{E}\left\| A(W)\bar{Z} - \sqrt{\lambda}\, e_1 \right\|^2 \\ &= \mathbb{E}\left[\{A(W)Z_1 - \sqrt{\lambda}\}^2 + A(W)^2 Q \right]. \end{aligned} \tag{37}$$