

Complex-Corrected Stochastic Gradient Descent for Noisy Observations

Abstract. We study stochastic optimization with Gaussian measurement error in the inputs. For scores admitting an entire extension, complex-corrected stochastic gradient descent evaluates the score at a noisy observation plus an independent imaginary Gaussian perturbation. Building on the classical complex corrected-score identity, this gives a directly implementable learning rule whose gradient is exactly unbiased at every iteration, rather than corrected only to finite order. Under an exponential derivative bound with parameter-dependent amplitude, a gradient-based amplitude control, and sub-Gaussian latent inputs, we derive the conditional moment estimates that place this recursion within classical stochastic approximation. We obtain convex objective bounds, almost-sure convergence to the clean optimizer θ^* and mean-square rates under strong convexity, and nonconvex stationarity bounds. We also establish last-iterate and averaged central limit laws for step sizes $c/(t + \gamma)^\alpha$, $1/2 < \alpha \leq 1$, including the distinct averaging covariance at $\alpha = 1$. The construction allows any known positive semidefinite noise covariance. We discuss exact correction for admissible non-Gaussian errors, including uniform and symmetric two-point noise, under suitable integrability conditions. Explicit examples and numerical experiments illustrate parameter convergence and improved prediction over naive training.

1 Introduction

Consider the stochastic optimization problem

$$\min_{\theta \in \mathbb{R}^p} F(\theta), \quad F(\theta) = \mathbb{E} \ell(\theta; Z),$$

when the latent input Z is observed only through $W = Z + U$. Even for centered measurement error, the score $g(\theta; W) = \nabla_\theta \ell(\theta; W)$ is generally biased for $\nabla F(\theta)$. A decreasing step size controls stochastic fluctuations but does not eliminate this discrepancy. The question is whether noisy observations can nevertheless support an explicit stochastic-gradient method for the clean objective.

For Gaussian error with known covariance, the classical complex corrected-score identity provides a construction [14, 10]. If the score has an integrable entire extension, an independent imaginary Gaussian perturbation cancels the measurement error in expectation. This leads to complex-corrected stochastic gradient descent (CCSGD):

$$\theta_{t+1} = \theta_t - \eta_t \operatorname{Re} g^C(\theta_t; W_t + iV_t), \quad V_t \sim N(0, \Omega).$$

Each iteration uses one noisy observation and one complex score evaluation. Only the noise covariance is required; it may be correlated or singular.

The central analytical issue is control of the corrected gradient along the stochastic trajectory. Pointwise unbiasedness and finite variance do not by themselves give the conditional moment bounds required for convergence. We address this issue through three results.

1. *Exact correction.* We formulate the Gaussian correction for vector-valued entire scores under an explicit exponential growth condition. The construction is independent of the convexity of the objective.
2. *Relative variance control.* We bound the correction error using the score's input derivative and the tails of the latent and Gaussian variables. The derivative amplitude may grow with the parameter, provided its square is controlled by the squared objective gradient.

Conditional orthogonality then yields the classical SGD relative variance condition and local higher-moment control for limit laws.

3. *Convergence and fluctuations.* For polynomially decaying step sizes with exponent $1/2 < \alpha \leq 1$, we establish convex objective bounds, strongly convex mean-square rates and almost-sure convergence, and nonconvex stationarity bounds. Under local regularity, we identify the last-iterate and averaged Gaussian limits, including the $\alpha = 1$ endpoint.

The analysis connects the classical Gaussian identity to stochastic approximation through explicit moment estimates [10, 5, 11]. Examples and numerical studies illustrate the method; a separate moment argument covers nonlinear-feature regression.

2 Related work

Table 1. Selected approaches to measurement-error correction. Exactness is understood within each method’s stated model and function class.

Work	Main result	Relation to CCSGD
Nakamura (1990) [9]	Corrected estimating equations	Foundational score-correction principle.
Novick–Stefanski (2002) [10]	Exact Gaussian complex Monte Carlo corrected scores and batch M -estimation	Our Gaussian correction uses their identity; we study its per-sample use in SGD.
Stefanski (1989) [14]	Analyticity restriction for globally unbiased Gaussian correction	Delimits exact correction for scores with complex singularities.
Konstantopoulos–Patie–Sarkar (2024) [7]	van Dantzig pairs and distributional constructions	Supplies candidate conjugate perturbations for non-Gaussian noise; score integrability remains to be checked.
Woolsey–Huang (2026) [15]	Statistical estimation with non-Gaussian error through finite-order bias cancellation	CCSGD gives an explicit stochastic-gradient update and optimization guarantees; exact correction for admissible non-Gaussian noise is discussed in Further work.
Cook–Stefanski (1994); Spicker et al. (2024) [4, 13]	Parametric and nonparametric SIMEX	Correct by extrapolation rather than an exact per-sample gradient identity.
Belomestny–Goldenshluger (2021) [1]	Density deconvolution allowing zeros of the error characteristic function	Related inverse problem; its target is density estimation.
Loh–Wainwright (2012) [8]	Corrected high-dimensional sparse linear regression with a convergent projected-gradient algorithm	Addresses linear regression; CCSGD treats scores of more general analytic losses in an online stochastic update.
Carroll et al. (2006) [2]	Regression calibration and broader measurement-error methodology	Context for plug-in correction.

Nakamura’s corrected-score principle and the complex Monte Carlo construction of Novick and Stefanski [9, 10] provide the statistical foundation. CCSGD studies the associated online gradient recursion and derives its noise conditions before applying classical optimization theory. Loh and Wainwright [8] already provide a computationally tractable corrected optimization method for sparse linear regression; our analysis allows more general analytic scores un-

der explicit growth conditions. Woolsey and Huang [15] develop non-Gaussian corrections for statistical estimation by cancelling finitely many bias terms, which generally leaves a residual bias. Our focus is stochastic optimization: CCSGD gives a directly implementable learning rule using one noisy observation and one simulated complex perturbation per update, with convergence rates and asymptotic guarantees. Further work discusses exact correction for non-Gaussian noise admitting a suitable conjugate perturbation, including uniform and symmetric two-point noise, under the required score-integrability conditions. In our experiments, the Woolsey–Huang baseline is their deterministic correction in Eq. (5).

Our contributions

1. *An explicit optimization method.* CCSGD turns complex correction into a directly implementable stochastic optimization algorithm for a broad class of analytic losses with noisy inputs. Each update evaluates the score at one noisy observation plus one simulated complex perturbation. Experiments demonstrate convergence to the true parameter in nonlinear-feature regression and improved prediction over naive training in neural-network regression.
2. *Exact correction beyond Gaussian noise.* The conjugate-variable criterion in our companion work [3] extends the correction principle to admissible non-Gaussian noise, including uniform and symmetric two-point distributions. Further work describes these constructions and the integrability conditions required for exact unbiasedness; the main optimization theorems here concern Gaussian measurement error.
3. *Per-step unbiasedness and asymptotic guarantees.* The corrected gradient is exactly unbiased at every iteration, rather than merely cancelling finitely many bias terms. We derive the classical SGD relative variance bound from input growth and sampling conditions, then establish convex and nonconvex guarantees. Under strong convexity, CCSGD converges to the clean optimizer θ^* almost surely and in mean square, with explicit MSE rates; under the stated local regularity conditions, both the last iterate and the Polyak–Ruppert average are asymptotically normal.

3 Complex-Corrected SGD

Let $\theta \in \mathbb{R}^p$. A fixed response, when present, is suppressed in the notation. We observe $W = Z + U$, where $U \sim N(0, \Omega)$ is independent of Z and the covariance $\Omega \succeq 0$ is known. No invertibility or diagonal structure is required. Assume that $Z \in \mathbb{R}^d$ is sub-Gaussian about its mean m_Z : for some $\sigma_Z < \infty$,

$$\mathbb{E} e^{s v^\top (Z - m_Z)} \leq e^{\sigma_Z^2 s^2 / 2} \quad (s \in \mathbb{R}, \|v\| = 1).$$

For a loss $\ell(\theta; z)$ differentiable in θ , with $\mathbb{E}|\ell(\theta; Z)| < \infty$ for every θ , write

$$F(\theta) = \mathbb{E}_Z \ell(\theta; Z), \quad g(\theta; z) = \nabla_\theta \ell(\theta; z).$$

The following assumptions separate input regularity from clean sampling noise. All norms are Euclidean, with $\|\cdot\|_{\text{op}}$ denoting the induced operator norm.

Assumption 3.1 (Holomorphic score). For every θ , $z \mapsto g(\theta; z)$ admits an entire extension $g^{\mathbb{C}}(\theta; \cdot)$ on $\mathbb{C}^d$, real-valued on $\mathbb{R}^d$.

Assumption 3.2 (Exponential envelope). There are a finite constant $a \geq 0$ and a measurable function $M : \mathbb{R}^p \rightarrow [0, \infty)$ such that

$$\|D_q g^{\mathbb{C}}(\theta; q)\|_{\text{op}} \leq M(\theta) e^{a\|q\|}, \quad \theta \in \mathbb{R}^p, q \in \mathbb{C}^d. \quad (3.1)$$

The exponent a is independent of θ ; the amplitude $M(\theta)$ need not be bounded. In addition, for finite constants $A, B \geq 0$,

$$M(\theta)^2 \leq A + B\|\nabla F(\theta)\|^2, \quad \theta \in \mathbb{R}^p. \quad (3.2)$$

Assumption 3.3 (Clean sampling variance). Differentiation may be interchanged with the expectation defining F , and for constants $\sigma_0^2, \rho_0 \geq 0$, independent of θ ,

$$\mathbb{E}_Z \|g(\theta; Z) - \nabla F(\theta)\|^2 \leq \sigma_0^2 + \rho_0 \|\nabla F(\theta)\|^2.$$

This condition concerns the clean stochastic score; no variance bound is assumed on the corrected score.

The amplitude control transfers the input growth bound to the standard SGD relative variance condition. The correction identity itself uses only the entire extension and the bound (3.1); it requires neither amplitude control nor convexity.

Algorithm 1 Complex-Corrected SGD

Require: Initial point θ_1 , step sizes $\{\eta_t\}_{t=1}^T$, known noise covariance Ω

- 1: **for** $t = 1, \dots, T$ **do**
 - 2: Observe a fresh $W_t = Z_t + U_t$
 - 3: Draw $V_t \sim N(0, \Omega)$ independently of the observation and the past
 - 4: $G_t \leftarrow \text{Re } g^{\text{C}}(\theta_t; W_t + iV_t)$
 - 5: $\theta_{t+1} \leftarrow \theta_t - \eta_t G_t$
 - 6: **end for**
 - 7: **return** θ_{T+1}
-

Optional averaging. For the same observation W_t , one may draw m independent $V_{t,j} \sim N(0, \Omega)$ and replace G_t by

$$G_t^{(m)} = \frac{1}{m} \sum_{j=1}^m \text{Re } g^{\text{C}}(\theta_t; W_t + iV_{t,j}).$$

This preserves conditional unbiasedness with respect to the past. Given $(\mathcal{F}_{t-1}, W_t)$, the covariance is $1/m$ times that of a single draw, reducing simulation variability and stabilizing the update at the cost of additional score evaluations. Variability due to the shared observation remains.

Lemma 3.4 (Exact correction). Under Assumptions 3.1–3.2, for every (z, θ) ,

$$\mathbb{E}_{U,V} [\text{Re } g^{\text{C}}(\theta; z + U + iV)] = g(\theta; z).$$

Consequently, if θ_t is measurable with respect to the past $\mathcal{F}_{t-1}$, (Z_t, U_t, V_t) is independent of that past, and differentiation may be interchanged with the expectation defining F , then $\mathbb{E}[\text{Re } g^{\text{C}}(\theta_t; W_t + iV_t) \mid \mathcal{F}_{t-1}] = \nabla F(\theta_t)$.

Observed responses. For noise-free responses $Y = r(Z)$, keep Y fixed when extending $z \mapsto g(\theta; z, Y)$ to complex inputs. The same argument gives

$$\mathbb{E}_{U,V} [\text{Re } g^{\text{C}}(\theta; Z + U + iV, Y) \mid Z, Y] = g(\theta; Z, Y),$$

provided the score satisfies the entire and growth conditions at each fixed (θ, y) . For the objective $F(\theta) = \mathbb{E}\ell(\theta; Z, Y)$, sampling and moment conditions must hold for the joint pair (Z, Y) . Appendix A.9 derives these bounds for the regression experiment.

4 Main results

Throughout this section, the assumptions of Section 3 hold. The initial point is deterministic, and (Z_t, U_t, V_t) are i.i.d. triples, independent of the past $\mathcal{F}_{t-1}$. Write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$ and

$$\Delta_t = G_t - g(\theta_t; Z_t), \quad \zeta_t = G_t - \nabla F(\theta_t).$$

Exact correction gives $\mathbb{E}_t \zeta_t = 0$. The exponential envelope and independence of the fresh data imply

$$\mathbb{E}_t \|\Delta_t\|^2 \leq K_a M(\theta_t)^2 \leq K_a A + K_a B \|\nabla F(\theta_t)\|^2,$$

where

$$K_a = \mathbb{E} e^{2a\|Z\|} \mathbb{E} \left[\|U + iV\|^2 e^{2a\|U+iV\|} \right] < \infty.$$

Since $\mathbb{E}[\Delta_t \mid \mathcal{F}_{t-1}, Z_t] = 0$, the cross term vanishes. Assumption 3.3 therefore yields

$$\mathbb{E}_t \|\zeta_t\|^2 = \mathbb{E}_t \|\Delta_t\|^2 + \mathbb{E}_t \|g(\theta_t; Z_t) - \nabla F(\theta_t)\|^2 \leq \sigma^2 + \rho \|\nabla F(\theta_t)\|^2, \quad (4.1)$$

with

$$\sigma^2 = \sigma_0^2 + K_a A, \quad \rho = \rho_0 + K_a B.$$

These constants are independent of t and θ . Thus the correction preserves the classical relative variance structure without requiring a bounded amplitude M . All proofs, including the local higher-moment estimates for the CLTs, are given in Appendix A.

Mini-batching and Monte Carlo averaging. With n independent fresh observations and m independent imaginary perturbations per observation, use

$$G_t^{(n,m)} = \frac{1}{n} \sum_{i=1}^n \frac{1}{m} \sum_{j=1}^m \operatorname{Re} g^C(\theta_t; W_{t,i} + iV_{t,i,j}).$$

All perturbations are independent of the observations and the past. For fixed θ , write $h_\theta(W) = \mathbb{E}_V[\operatorname{Re} g^C(\theta; W + iV) \mid W]$ and define

$$\begin{aligned} v_{\text{obs}}(\theta) &= \mathbb{E} \|h_\theta(W) - \nabla F(\theta)\|^2, \\ v_{\text{sim}}(\theta) &= \mathbb{E} \|G(\theta) - h_\theta(W)\|^2. \end{aligned}$$

Conditional centering makes the observation and simulation errors orthogonal. Independence across observations and simulations therefore gives

$$\mathbb{E}_t \|G_t^{(n,m)} - \nabla F(\theta_t)\|^2 = \frac{v_{\text{obs}}(\theta_t)}{n} + \frac{v_{\text{sim}}(\theta_t)}{nm} \leq \frac{\sigma^2 + \rho \|\nabla F(\theta_t)\|^2}{n}, \quad (4.2)$$

since $v_{\text{obs}} + v_{\text{sim}} = \mathbb{E} \|G - \nabla F\|^2$. Thus mini-batching reduces both components by n , while repeated imaginary perturbations further reduce only the simulation component by m . The bounds below apply with variance constants σ^2/n and ρ/n ; for the CLTs, the corrected-oracle covariance is that of $G^{(n,m)}$ when n and m are fixed. The analysis below requires only the relative variance bound; its precise constants affect the quantitative bounds, but not the stated convergence rates or the validity of the CLTs. Averaging is primarily a practical option for reducing variability and improving training stability.

Theorem 4.1 (Convex objective). Suppose $F : \mathbb{R}^p \rightarrow \mathbb{R}$ is convex and L -smooth and has a minimizer θ^* . Let $\eta_t = c/(t + \gamma)^\alpha$ with $c > 0$, $\gamma \geq 0$, $1/2 < \alpha \leq 1$, and $\eta_1 \leq 1/[L(1 + \rho)]$. Define $\bar{\theta}_T^\eta = (\sum_{t=1}^T \eta_t \theta_t) / (\sum_{t=1}^T \eta_t)$. Then

$$\mathbb{E}[F(\bar{\theta}_T^\eta)] - F(\theta^*) \leq \frac{\|\theta_1 - \theta^*\|^2 + \sigma^2 \sum_{t=1}^T \eta_t^2}{2 \sum_{t=1}^T \eta_t}.$$

Consequently, the expected objective gap is $O(T^{\alpha-1})$ if $\alpha < 1$, and $O(1/\log T)$ if $\alpha = 1$.

Theorem 4.2 (Strong convexity). Suppose $F : \mathbb{R}^p \rightarrow \mathbb{R}$ is μ -strongly convex and L -smooth, with minimizer θ^* . Let $\eta_t = c/(t + \gamma)^\alpha$ with $c > 0$, $\gamma \geq 0$, and $1/2 < \alpha \leq 1$. Then $\theta_t \rightarrow \theta^*$ almost surely and in mean square. For $\alpha < 1$,

$$\mathbb{E}\|\theta_t - \theta^*\|^2 = O(t^{-\alpha}).$$

For $\alpha = 1$,

$$\mathbb{E}\|\theta_t - \theta^*\|^2 = \begin{cases} O(t^{-1}), & 2\mu c > 1, \\ O(t^{-1} \log t), & 2\mu c = 1, \\ O(t^{-2\mu c}), & 2\mu c < 1. \end{cases}$$

For the two central limit theorems, write

$$S(\theta) = \mathbb{E}[(G(\theta) - \nabla F(\theta))(G(\theta) - \nabla F(\theta))^\top], \quad G(\theta) = \text{Re } g^C(\theta; Z + U + iV).$$

Theorem 4.3 (Last-iterate fluctuations). In addition to the strong convexity and smoothness conditions of Theorem 4.2, suppose F is twice continuously differentiable near θ^* , its Hessian is locally Lipschitz there, and $S(\theta) \rightarrow S_* := S(\theta^*)$ as $\theta \rightarrow \theta^*$. Let $H = \nabla^2 F(\theta^*)$ and $\eta_t = c/(t + \gamma)^\alpha$, where $c > 0$, $\gamma \geq 0$, and $1/2 < \alpha \leq 1$. If $\alpha < 1$, then

$$t^{\alpha/2}(\theta_t - \theta^*) \Rightarrow N(0, \Sigma_c), \quad H\Sigma_c + \Sigma_c H = cS_*.$$

If $\alpha = 1$ and $c\lambda_{\min}(H) > 1/2$, then

$$\sqrt{t}(\theta_t - \theta^*) \Rightarrow N(0, \Sigma_c), \quad (cH - \frac{1}{2}I)\Sigma_c + \Sigma_c(cH - \frac{1}{2}I) = c^2S_*.$$

Each Lyapunov equation has a unique positive semidefinite solution.

Theorem 4.4 (Polyak–Ruppert averaging). Under the objective and covariance conditions of Theorem 4.3, let $\eta_t = c/(t + \gamma)^\alpha$ with $c > 0$, $\gamma \geq 0$, and $1/2 < \alpha \leq 1$. For $\bar{\theta}_T = T^{-1} \sum_{t=1}^T \theta_t$, if $\alpha < 1$,

$$\sqrt{T}(\bar{\theta}_T - \theta^*) \Rightarrow N(0, H^{-1}S_*H^{-1}).$$

If $\alpha = 1$ and $c\lambda_{\min}(H) > 1/2$, the same scaling holds with covariance

$$\Xi_c = \int_0^1 K_c(u) S_* K_c(u)^\top du, \quad K_c(u) = c u^{cH-I} \int_u^1 v^{-cH} dv,$$

where $u^A = e^{(\log u)A}$ for a matrix A .

The two averages have different roles: $\bar{\theta}_T^\eta$ gives the convex objective bound, whereas $\bar{\theta}_T$ gives the averaged CLT. At $\alpha = 1$, Ξ_c generally differs from $H^{-1}S_*H^{-1}$. In every CLT, S_* is the covariance of the corrected oracle, including measurement and simulation noise.

Theorem 4.5 (Nonconvex stationarity). Suppose $F : \mathbb{R}^p \rightarrow \mathbb{R}$ is L -smooth and bounded below by $F_{\inf}$. Let $\eta_t = c/(t + \gamma)^\alpha$ with $c > 0$, $\gamma \geq 0$, $1/2 < \alpha \leq 1$, and $\eta_1 \leq 1/[L(1 + \rho)]$. Then, for any $T \geq 1$,

$$\frac{\sum_{t=1}^T \eta_t \mathbb{E}\|\nabla F(\theta_t)\|^2}{\sum_{t=1}^T \eta_t} \leq \frac{2(F(\theta_1) - F_{\inf}) + L\sigma^2 \sum_{t=1}^T \eta_t^2}{\sum_{t=1}^T \eta_t}.$$

Consequently, the weighted mean squared gradient is $O(T^{\alpha-1})$ if $\alpha < 1$, and $O(1/\log T)$ if $\alpha = 1$.

5 Explicit corrections

These formulas illustrate the correction identity; the algorithm evaluates the complex score directly and does not require symbolic integration.

Linear prediction with squared loss. Let $\theta, Z, W, U, V \in \mathbb{R}^p$, with $W = Z + U$ and independent $U, V \sim N(0, \Omega)$. For a fixed response y and the squared linear-prediction loss $\ell(\theta; z) = \frac{1}{2}(\theta^\top z - y)^2$, the clean score is

$$g(\theta; z) = (\theta^\top z - y)z.$$

The complex-corrected score and its conditional mean over the simulated perturbation are

$$G(\theta; W, V) = \text{Re}[(\theta^\top (W + iV) - y)(W + iV)] = (\theta^\top W - y)W - (\theta^\top V)V,$$

$$\tilde{G}(\theta; W) = \mathbb{E}_V[G(\theta; W, V) \mid W] = (\theta^\top W - y)W - \Omega\theta.$$

In particular, $\mathbb{E}\tilde{G}(\theta; W) = \nabla F(\theta)$. The identity is exact for every θ . Indeed,

$$D_z g^C(\theta; q) = q\theta^\top + (\theta^\top q - y)I_p, \quad \|D_z g^C(\theta; q)\|_{\text{op}} \leq 2\|\theta\|\|q\| + |y|.$$

Thus Assumption 3.2 holds with $a = 1$ and $M(\theta) = 2\|\theta\| + |y|$. If $H = \mathbb{E}[ZZ^\top] \succ 0$, let $\mu = \lambda_{\min}(H)$ and $\theta^* = H^{-1}y\mathbb{E}Z$. Then $\nabla F(\theta) = H(\theta - \theta^*)$ and

$$M(\theta)^2 \leq 2(2\|\theta^*\| + |y|)^2 + 8\mu^{-2}\|\nabla F(\theta)\|^2.$$

The clean score is affine in θ with square-integrable random coefficients, so its centered variance also has this relative form. Consequently the main results apply without bounding the parameter set.

Nonlinear analytic loss. Let $\theta, z \in \mathbb{R}^p$ and $\ell(\theta; z) = \frac{1}{2}\|\theta - \sin z\|^2$, where $\sin$ acts coordinate-wise. Then $g(\theta; z) = \theta - \sin z$, and the complex update is

$$G(\theta; W, V) = \theta - \sin W \odot \cosh V.$$

Since $V_j \sim N(0, \Omega_{jj})$, integrating out V gives

$$\tilde{G}(\theta; W) = \theta - (e^{\Omega_{jj}/2} \sin W_j)_{j=1}^p, \quad \mathbb{E}\tilde{G}(\theta; W) = \nabla F(\theta).$$

Here $D_z g^C(\theta; q) = -\text{diag}(\cos q_j)$ has an exponential bound independent of θ , while $\nabla^2 F = I_p$. The assumptions therefore hold on all of $\mathbb{R}^p$.

6 Numerical experiments

We compare CCSGD with naive SGD, second- and fourth-order Taylor correction, quadratic SIMEX, the deterministic Woolsey–Huang Eq. (5) correction, and regression calibration. Responses are noise-free; Gaussian measurement noise is added only to the inputs. Appendix B specifies the baseline updates and population-bias calculations.

6.1 Nonlinear-feature regression

We use

$$\psi(z) = (\sin z, \cos(0.7z), z^2)^\top, \quad Y = \theta_{\text{true}}^\top \psi(Z), \quad \ell(\theta; z, y) = \frac{1}{2}(\theta^\top \psi(z) - y)^2,$$

where $\theta_{\text{true}} = (0.8, -0.6, 0.2)^\top$ and $Z \sim 0.8N(-0.5, 0.8^2) + 0.2N(2, 0.8^2)$. The algorithms observe (W, Y) with $W = Z + U$ and $U \sim N(0, 1.2^2)$ independent of Z . CCSGD evaluates the full score at $W + iV$, with an independent $V \sim N(0, 1.2^2)$ at each update, using complex arithmetic. All runs use fresh latent samples and fresh measurement noise. The objective is quadratic and

strongly convex in θ , despite its nonlinear dependence on z . Appendix A.9 verifies the relative variance bound directly and gives the $O(t^{-1})$ MSE rate. Figure 1 shows a fitted slope of -0.978 , consistent with this rate, while the specified approximate updates approach nonzero population biases.

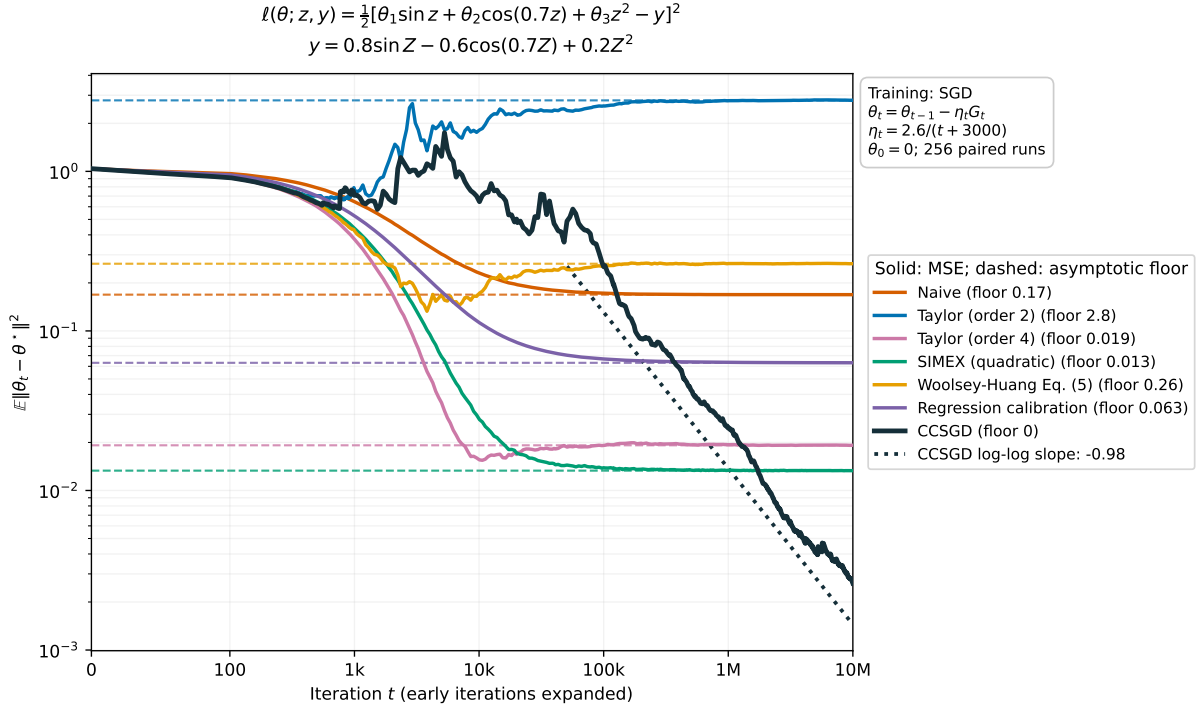


Figure 1. Parameter MSE averaged over 256 paired runs, starting at $\theta_0 = 0$ with $\eta_t = 2.6/(t + 3000)$. Dashed lines show the calculated asymptotic squared biases of the specified baseline updates. Quadratic SIMEX extrapolates three separately optimized parameter trajectories; its additional Gaussian noise is integrated exactly. The Woolsey–Huang baseline uses their deterministic Eq. (5) correction. The dotted CCSGD guide has fitted log–log slope -0.978 over $5 \times 10^4 \leq t \leq 10^7$.

6.2 Neural-network regression

We generate noise-free responses from

$$Z \sim \text{Unif}[-2, 2], \quad Y = f_0(Z) = e^{-Z^2} \sin(3Z) + 0.3Z^2,$$

and fit a single-hidden-layer network

$$h_\theta(z) = c + \sum_{j=1}^{32} a_j \sin(b_j z + d_j), \quad \ell(\theta; z, y) = \frac{1}{2}(h_\theta(z) - y)^2,$$

training all 97 parameters. For each fixed finite parameter vector, the score is entire and has at most exponential growth in the input; its Gaussian complex correction is therefore unbiased with finite second moment. The exponential rate can depend on the input weights; neither a parameter-independent exponent nor the amplitude control in (3.2) is established for this network. Its training curves are therefore empirical and do not invoke the nonconvex theorem.

We use 8,192 latent training pairs and 4,096 independent clean prediction pairs, with no validation set. At every minibatch presentation, we generate fresh $U \sim N(0, 0.3^2)$ and observe $W = Z + U$; all methods share these observations, minibatch indices, and initialization. CCSGD independently samples $V \sim N(0, 0.3^2)$ and computes

$$\text{Re}[(h_\theta(W + iV) - Y) \nabla_\theta h_\theta(W + iV)]$$

using complex arithmetic and the analytic squared loss, without complex conjugation. Training uses minibatches of size 128 and $\eta_t = 0.01(1 + t/2000)^{-0.6}$.

We compare naive SGD, Taylor corrections of orders two and four, quadratic SIMEX, the Woolsey–Huang Eq. (5) correction, and regression calibration. Taylor corrections differentiate the full score with respect to its input, holding the response fixed. SIMEX trains separate networks at additional noise levels $\lambda = 0, 1, 2$ and extrapolates predictions as $3h^{(0)} - 3h^{(1)} + h^{(2)}$. Regression calibration uses the exact $\mathbb{E}[Z | W]$ under the specified uniform latent distribution. The prediction set is used only to report $n_{\text{pred}}^{-1} \sum_i (h_{\theta_t}(Z_i) - Y_i)^2$.

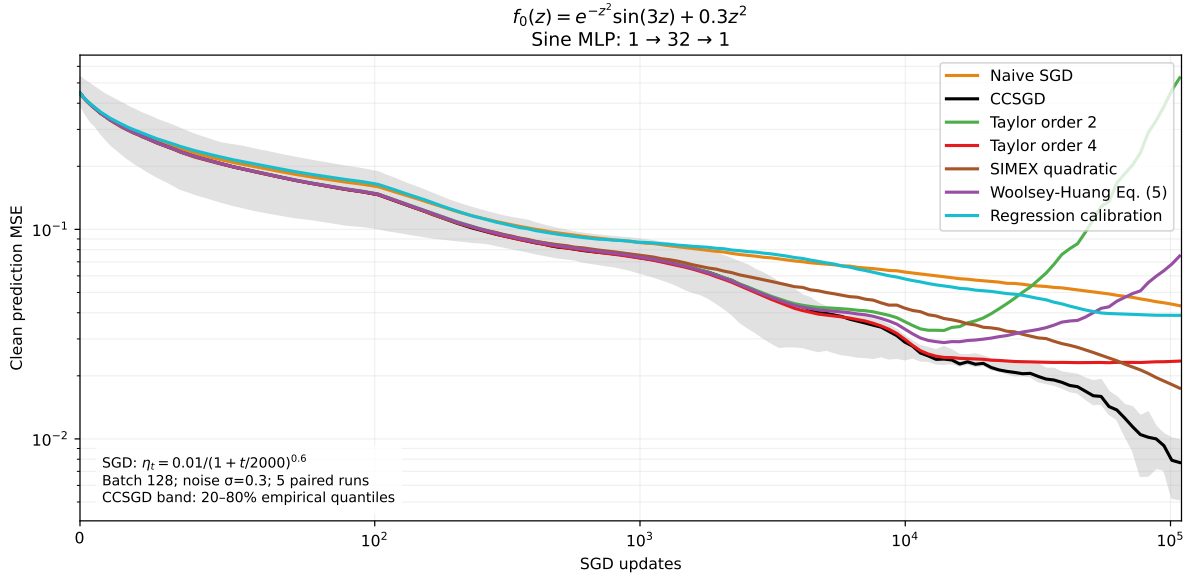


Figure 2. Clean-input prediction MSE averaged over five paired runs. The shaded CCSGD band gives empirical 20–80% quantiles. Training runs for 200,000 updates; the figure displays the trajectory through 1.1×10^5 updates. CCSGD attains lower mean prediction error than the specified baselines near the displayed endpoint.

This experiment targets the clean empirical training objective conditional on the latent training set. Fresh measurement noise preserves conditional unbiasedness for that objective; it does not remove finite-sample or approximation error on the independent prediction set. The methods’ ordering changes during training: at 200,000 updates, mean prediction MSE is 0.0142 for CCSGD and 0.0125 for SIMEX. These results illustrate finite-time behavior without asserting a zero prediction-error limit.

7 Further work

7.1 Noise generalization

The Gaussian perturbation is one instance of a more general conjugate construction. Our companion manuscript [3] characterizes symmetric scalar errors U admitting an independent real variable V such that

$$M_U(t)\varphi_V(t) = 1, \quad t \in \mathbb{R}, \quad (7.1)$$

where $M_U(t) = \mathbb{E}e^{tU}$ is finite on $\mathbb{R}$ and $\varphi_V(t) = \mathbb{E}e^{itV}$. Existence is equivalent to $1/M_U$ being a characteristic function; the companion manuscript gives Fourier and moment criteria and constructs the conjugate law. The imaginary perturbation must follow this law, which generally differs from the measurement-error law.

Uniform measurement error. For $U \sim \text{Unif}[-b, b]$, with known $b > 0$,

$$M_U(t) = \frac{\sinh(bt)}{bt}, \quad \varphi_V(t) = \frac{bt}{\sinh(bt)},$$

with both ratios interpreted as 1 at $t = 0$. The conjugate V has a centered logistic distribution with scale b/π :

$$f_V(v) = \frac{\pi}{4b \cosh^2(\pi v/(2b))}, \quad F_V(v) = \frac{1}{1 + e^{-\pi v/b}}.$$

It can be sampled directly from $R \sim \text{Unif}(0, 1)$ via

$$V = \frac{b}{\pi} \log \frac{R}{1-R}. \quad (7.2)$$

Thus a uniform measurement error is corrected by an independent logistic imaginary perturbation.

Symmetric two-point measurement error. For $\mathbb{P}(U = b) = \mathbb{P}(U = -b) = 1/2$,

$$M_U(t) = \cosh(bt), \quad \varphi_V(t) = \frac{1}{\cosh(bt)}.$$

Here V has a scaled hyperbolic-secant distribution:

$$f_V(v) = \frac{1}{2b \cosh(\pi v/(2b))}, \quad F_V(v) = \frac{2}{\pi} \arctan(e^{\pi v/(2b)}).$$

An exact sampler is

$$V = \frac{2b}{\pi} \log \tan\left(\frac{\pi R}{2}\right), \quad R \sim \text{Unif}(0, 1). \quad (7.3)$$

The imaginary perturbation is continuous even though the measurement error is discrete.

Exact correction and implementation. Set $D = U + iV$. Equation (7.1) implies $\mathbb{E}D^k = 0$ for every integer $k \geq 1$ [3]. Hence, for fixed (θ, z) ,

$$\mathbb{E}_{U,V} \text{Re } g^{\mathbb{C}}(\theta; z + U + iV) = g(\theta; z) \quad (7.4)$$

for polynomial scores. The identity also holds for entire scores when

$$\sum_{k=0}^{\infty} \frac{\|\partial_z^k g^{\mathbb{C}}(\theta; z)\|}{k!} \mathbb{E}|D|^k < \infty. \quad (7.5)$$

Indeed, this condition permits termwise expectation of the Taylor series; all nonconstant terms vanish. Thus correction is exact, with no finite-order truncation. Algorithm 2 implements it using the explicit inverse-distribution samplers above. It requires the error law and its scale, but not the latent input distribution.

Algorithm 2 CCSGD for uniform or symmetric two-point measurement error**Require:** Initial point θ_1 , step sizes $\{\eta_t\}_{t=1}^T$, known error type and scale $b > 0$

```

1: for  $t = 1, \dots, T$  do
2:   Observe a fresh  $W_t = Z_t + U_t$ 
3:   Draw  $R_t \sim \text{Unif}(0, 1)$  independently of the observation and the past
4:   if  $U_t \sim \text{Unif}[-b, b]$  then
5:      $V_t \leftarrow (b/\pi) \log\{R_t/(1 - R_t)\}$ 
6:   else  $\triangleright \mathbb{P}(U_t = \pm b) = 1/2$ 
7:      $V_t \leftarrow (2b/\pi) \log \tan(\pi R_t/2)$ 
8:   end if
9:    $G_t \leftarrow \text{Re } g^{\mathbb{C}}(\theta_t; W_t + iV_t)$ 
10:   $\theta_{t+1} \leftarrow \theta_t - \eta_t G_t$ 
11: end for
12: return  $\theta_{T+1}$ 

```

For noise-free responses, the observed response is held fixed during the complex evaluation, as in Section 3. Independent replicates of V_t may be averaged using the same observation. For independent input-error coordinates, the construction can also be applied coordinatewise with independent conjugate perturbations, subject to the corresponding multivariate integrability condition.

Moment conditions for optimization. Unlike a Gaussian conjugate, these conjugates have exponential moments only below a finite threshold. Specifically,

$$\mathbb{E}e^{s|V|} < \infty \iff 0 \leq s < \lambda, \quad \lambda = \begin{cases} \pi/b, & U \sim \text{Unif}[-b, b], \\ \pi/(2b), & \mathbb{P}(U = \pm b) = 1/2. \end{cases}$$

Under the derivative envelope (3.1), the strict condition $a < \lambda$ suffices for (7.5): the score has exponential growth of any rate slightly larger than a , and the conjugate has an exponential moment at a still larger rate. If $2a < \lambda$, boundedness of U gives

$$K_a = \mathbb{E}e^{2a|Z|} \mathbb{E}[|D|^2 e^{2a|D|}] < \infty.$$

Together with the clean sampling and amplitude bounds, the same argument as in Section 4 yields the relative variance condition (4.1). For the CLTs, choose $\delta > 0$ with $(2 + \delta)a < \lambda$; the higher-moment proof then applies as well. Thus the optimization analysis transfers when these moment conditions and the corresponding objective and covariance assumptions are verified. Polynomial scores satisfy the required input moment conditions for both pairs. Arbitrary entire scores still require growth control, and random responses require the corresponding joint moment bounds.

7.2 Function generalization

Exact Gaussian correction restricts the admissible score class. In the scalar Gaussian location model with positive noise variance, Stefanski [14] shows that a target admitting an integrable unbiased estimator for every real mean must extend to an entire function. Applied componentwise at fixed θ , this rules out globally exact unbiased correction of non-entire scores in that setting. Entire analyticity alone does not ensure the integrability required by CCSGD; growth control remains essential. For non-entire scores, a natural direction is approximate or regularized correction with explicit bias, variance, and optimization-error bounds. With singular Ω , the analyticity restriction concerns the directions affected by noise.

A Proofs of the main results

A.1 Proof of Lemma 3.4

Proof. Fix θ and $z \in \mathbb{R}^d$, and put $C = U + iV$. The fundamental theorem of calculus and the exponential envelope give, for every $q \in \mathbb{C}^d$,

$$\|g^{\mathbb{C}}(\theta; q) - g(\theta; 0)\| \leq M(\theta)\|q\|e^{a\|q\|}. \quad (\text{A.1})$$

Since C is complex Gaussian, its norm has exponential moments of every linear order. Thus $g^{\mathbb{C}}(\theta; z + C)$ is absolutely integrable. The same bound permits exchange of the expectation and angular integral below.

For every $\phi \in [0, 2\pi]$, the real pair

$$(\operatorname{Re}(e^{i\phi}C), \operatorname{Im}(e^{i\phi}C)) = (U \cos \phi - V \sin \phi, U \sin \phi + V \cos \phi)$$

has the same centered Gaussian law as (U, V) : its covariance is $\operatorname{diag}(\Omega, \Omega)$. Hence $e^{i\phi}C \stackrel{d}{=} C$. On the other hand, for each fixed $c \in \mathbb{C}^d$, the one-variable function $s \mapsto g^{\mathbb{C}}(\theta; z + sc)$ is entire. Cauchy's mean-value formula, applied componentwise, gives

$$\frac{1}{2\pi} \int_0^{2\pi} g^{\mathbb{C}}(\theta; z + e^{i\phi}c) d\phi = g^{\mathbb{C}}(\theta; z) = g(\theta; z).$$

Taking expectation over C and using rotational invariance yields

$$\mathbb{E} g^{\mathbb{C}}(\theta; z + C) = \mathbb{E} \left[\frac{1}{2\pi} \int_0^{2\pi} g^{\mathbb{C}}(\theta; z + e^{i\phi}C) d\phi \right] = g(\theta; z).$$

Taking real parts proves the first claim. For the second, condition on $\mathcal{F}_{t-1}$, apply the first claim conditional on Z_t , and then average over Z_t . Predictability and the assumed interchange of gradient and expectation give $\nabla F(\theta_t)$. $\square$

A.2 Reduction to classical SGD noise conditions

Proof. Put $D = U + iV$. The fundamental theorem of calculus gives

$$\|\operatorname{Re} g^{\mathbb{C}}(\theta; Z + D) - g(\theta; Z)\| \leq M(\theta)\|D\|e^{a(\|Z\| + \|D\|)}.$$

Independence of Z and D gives $\mathbb{E}\|\Delta(\theta)\|^2 \leq K_a M(\theta)^2$. Equation (3.2) then gives the correction-error bound in the main text. Sub-Gaussian vectors in finite dimension have exponential moments of their norm of every linear order; the same holds for the Gaussian pair (U, V) , regardless of whether Ω is singular. Hence $K_a < \infty$. Conditioning first on Z_t and using exact correction shows that Δ_t is orthogonal in L^2 to $g(\theta_t; Z_t) - \nabla F(\theta_t)$. This proves (4.1). In particular,

$$\mathbb{E}_t \|G_t\|^2 \leq \sigma^2 + (1 + \rho)\|\nabla F(\theta_t)\|^2. \quad (\text{A.2})$$

For the CLTs, stronger moments follow from the same input bound. Set $b_a(q) = \|q\|e^{a\|q\|}$ and $Q = Z + U + iV$. Centering cancels $g(\theta; 0)$ and gives

$$\|G(\theta) - \nabla F(\theta)\| \leq M(\theta)[b_a(Q) + \mathbb{E}b_a(Z)].$$

The bracket has moments of every finite order. Hence, for every $r \geq 2$,

$$\mathbb{E}_t \|\xi_t\|^r \leq C_r M(\theta_t)^r \leq C_r (A + B\|\nabla F(\theta_t)\|^2)^{r/2}, \quad (\text{A.3})$$

where C_r is deterministic and finite. Under the strong convexity and smoothness conditions used for the CLTs, almost-sure convergence gives

$$\sup_t \|\nabla F(\theta_t)\| < \infty \quad \text{almost surely.}$$

Thus the conditional higher moments are bounded along each convergent trajectory by a finite random constant. On a fixed neighborhood of θ^* the bound is deterministic, since $\|\nabla F(\theta)\| \leq L\|\theta - \theta^*\|$. These bounds supply the Lindeberg control below without a global uniform bound on M or a moment assumption on the corrected score. $\square$

A.3 Convex convergence

Proof. Write $e_t = \theta_t - \theta^*$ and $f_t = F(\theta_t) - F(\theta^*)$. For a convex, L -smooth function,

$$\langle \nabla F(\theta_t), e_t \rangle \geq f_t + \frac{1}{2L} \|\nabla F(\theta_t)\|^2.$$

Expanding the squared distance, using unbiasedness and (A.2), gives

$$\mathbb{E}_t \|e_{t+1}\|^2 \leq \|e_t\|^2 - 2\eta_t f_t + ((1 + \rho)\eta_t^2 - \eta_t/L) \|\nabla F(\theta_t)\|^2 + \sigma^2 \eta_t^2.$$

The gradient term is nonpositive under the stated step-size bound. Taking expectations, summing, and applying convexity to the weighted average proves the displayed bound. The rates follow from $\sum_t \eta_t^2 < \infty$ and the growth of $\sum_{t \leq T} \eta_t$. $\square$

A.4 Strong convexity and MSE

Proof. Write $e_t = \theta_t - \theta^*$. Strong convexity and smoothness give $\langle \nabla F(\theta_t), e_t \rangle \geq \mu \|e_t\|^2$ and $\|\nabla F(\theta_t)\| \leq L\|e_t\|$. Hence

$$\mathbb{E}_t \|e_{t+1}\|^2 \leq (1 - 2\mu\eta_t + (1 + \rho)L^2\eta_t^2) \|e_t\|^2 + \sigma^2 \eta_t^2. \quad (\text{A.4})$$

This inequality first ensures finite second moments at every finite time. The Robbins–Siegmund theorem [12], applied with summable coefficients $(1 + \rho)L^2\eta_t^2$ and $\sigma^2\eta_t^2$, then shows that $\|e_t\|^2$ converges almost surely and $\sum_t \eta_t \|e_t\|^2 < \infty$. Since $\sum_t \eta_t = \infty$, the limit is zero.

For the rates, put $u_t = \mathbb{E} \|e_t\|^2$, $h_t = t + \gamma$, $A = 2\mu c$, $B = (1 + \rho)L^2 c^2$, and $D = \sigma^2 c^2$. Taking expectations gives

$$u_{t+1} \leq (1 - Ah_t^{-\alpha} + Bh_t^{-2\alpha})u_t + Dh_t^{-2\alpha}. \quad (\text{A.5})$$

If $\alpha < 1$, choose N so that the coefficient is nonnegative and at most $1 - (A/2)h_t^{-\alpha}$ for $t \geq N$. Choose $M \geq 4D/A$ with $u_N \leq Mh_N^{-\alpha}$. Increasing N if needed, $h_t^{1-\alpha} \geq 4\alpha/A$. The induction hypothesis $u_t \leq Mh_t^{-\alpha}$ then implies

$$u_{t+1} \leq Mh_t^{-\alpha} - \frac{AM}{4} h_t^{-2\alpha} \leq M(h_t + 1)^{-\alpha},$$

where the last inequality uses $h^{-\alpha} - (h+1)^{-\alpha} \leq \alpha h^{-\alpha-1}$. After adjusting M at this fixed starting index, induction proves $u_t = O(t^{-\alpha})$.

If $\alpha = 1$, choose N so that $q_j = 1 - A/h_j + B/h_j^2 > 0$ for $j \geq N$. Since $\log q_j = -A/h_j + O(h_j^{-2})$,

$$\prod_{j=s}^{t-1} q_j \leq C(h_s/h_t)^A \quad (t > s \geq N),$$

with a constant independent of s, t . Iterating (A.5) therefore gives

$$u_t \leq Ch_t^{-A} \left(h_N^A u_N + D \sum_{s=N}^{t-1} h_s^{A-2} \right).$$

The sum is $O(t^{A-1})$ for $A > 1$, $O(\log t)$ for $A = 1$, and $O(1)$ for $A < 1$. These are the three asserted rates, and all converge to zero. $\square$

A.5 Preparatory estimates for the central limit theorems

Throughout the CLT proofs, write $\bar{e}_T = T^{-1} \sum_{t=1}^T e_t = \bar{\theta}_T - \theta^*$, and put $h = \lambda_{\min}(H) > 0$ and

$$r(e) = \nabla F(\theta^* + e) - He.$$

Local Lipschitz continuity of the Hessian gives a radius $\varepsilon_0 > 0$ and a constant C_r such that $\|r(e)\| \leq C_r \|e\|^2$ for $\|e\| \leq \varepsilon_0$. Globally, smoothness gives $\|r(e)\| \leq (L + \|H\|)\|e\|$.

Localization. Choose $0 < \varepsilon \leq \varepsilon_0$ and define $\tau_N = \inf\{t \geq N : \|e_t\| > \varepsilon\}$, with $\inf \emptyset = \infty$. Almost-sure convergence implies

$$\mathbb{P}(\tau_N < \infty) \longrightarrow 0 \quad (N \rightarrow \infty). \quad (\text{A.6})$$

For $\alpha < 1$, the global MSE bound yields

$$\mathbb{E}[\|r(e_t)\| \mathbf{1}_{\{\tau_N > t\}}] \leq C t^{-\alpha}.$$

For $\alpha = 1$, the local spectral condition $ch > 1/2$ permits a choice $m \in (1/(2c), h)$ and a sufficiently small ε such that $\langle \nabla F(\theta^* + e), e \rangle \geq m \|e\|^2$ on this ball. The event $\{\tau_N > t\}$ is measurable with respect to $\mathcal{F}_{t-1}$. Setting $v_t = \mathbb{E}[\|e_t\|^2 \mathbf{1}_{\{\tau_N > t\}}]$, discarding paths that exit at the next step gives

$$v_{t+1} \leq (1 - 2m\eta_t + (1 + \rho)L^2\eta_t^2)v_t + \sigma^2\eta_t^2.$$

The preceding scalar comparison, with $2mc > 1$, gives

$$v_t = O_N(t^{-1}), \quad \mathbb{E}[\|r(e_t)\| \mathbf{1}_{\{\tau_N > t\}}] = O_N(t^{-1}). \quad (\text{A.7})$$

Here constants may depend on the fixed starting index N . We first take the iteration horizon to infinity and then let $N \rightarrow \infty$ in (A.6). Thus a remainder that vanishes on these localized paths is $o_p(1)$ for the original process.

Matrix products. For $b \geq a$, let $P_{a,b} = \prod_{j=a}^b (I - \eta_j H)$, and set $P_{a,b} = I$ when $a > b$. Choose N large enough that $\eta_j \|H\| < 1$ for $j \geq N$. Spectral decomposition gives

$$\begin{aligned} \|P_{s+1,t-1}\| &\leq \exp\left(-h \sum_{j=s+1}^{t-1} \eta_j\right), \quad \alpha < 1, \\ \|P_{s+1,t-1}\| &\leq C_\nu \left(\frac{s+\gamma}{t+\gamma}\right)^\nu, \quad \alpha = 1, \quad 0 < \nu < ch, \end{aligned} \quad (\text{A.8})$$

for $t > s \geq N$. For the second bound, apply $\log(1 - c\lambda/(j + \gamma)) = -c\lambda/(j + \gamma) + O(j^{-2})$ to each eigenvalue λ of H and sum over j .

A common martingale CLT argument. Equation (A.3), almost-sure convergence, and covariance continuity imply, for some $\delta > 0$,

$$\mathbb{E}_t \tilde{\zeta}_t = 0, \quad \sup_t \mathbb{E}_t \|\tilde{\zeta}_t\|^{2+\delta} \leq C_\delta(\omega) < \infty, \quad S(\theta_t) \longrightarrow S_* \quad \text{almost surely.}$$

Let $D_{T,s}$ be deterministic matrix weights such that

$$\max_s \|D_{T,s}\| \rightarrow 0, \quad \sup_T \sum_s \|D_{T,s}\|^2 < \infty, \quad \sum_s D_{T,s} S_* D_{T,s}^\top \rightarrow \Sigma. \quad (\text{A.9})$$

Splitting the sum at a fixed index and then sending that index to infinity shows that replacing S_* by $S(\theta_s)$ changes the last limit by $o(1)$ almost surely. For each $\epsilon > 0$, conditional Lindeberg follows from

$$\sum_s \mathbb{E}_s \left[\|D_{T,s} \xi_s\|^2 \mathbf{1}_{\{\|D_{T,s} \xi_s\| > \epsilon\}} \right] \leq C_\delta(\omega) \epsilon^{-\delta} \max_s \|D_{T,s}\|^\delta \sum_s \|D_{T,s}\|^2 \longrightarrow 0 \quad \text{almost surely.}$$

The martingale array CLT and the Cramér–Wold argument therefore give $\sum_s D_{T,s} \xi_s \Rightarrow N(0, \Sigma)$ [6, Chapter 3]. This verifies explicitly the probabilistic conditions used in the following two proofs.

A.6 Last-iterate fluctuations

Proof of Theorem 4.3. The exact recursion, started at a fixed sufficiently large N , is

$$e_t = P_{N,t-1} e_N - \sum_{s=N}^{t-1} \eta_s P_{s+1,t-1} \xi_s - \sum_{s=N}^{t-1} \eta_s P_{s+1,t-1} r(e_s). \quad (\text{A.10})$$

We verify directly the linearization and covariance limits underlying the stochastic approximation CLT of Fabian [5].

If $\alpha < 1$, the first term multiplied by $t^{\alpha/2}$ vanishes by (A.8). On localized paths, the expected norm of the scaled last term is bounded by

$$C t^{\alpha/2} \sum_{s=N}^{t-1} \eta_s \exp \left(-h \sum_{j=s+1}^{t-1} \eta_j \right) s^{-\alpha} = O(t^{-\alpha/2}).$$

To see this, split at $t/2$. The earlier terms have an exponentially small factor $e^{-c_1 t^{1-\alpha}}$; for the later terms, $s^{-\alpha} \leq C t^{-\alpha}$ and the remaining weighted geometric sum is bounded. Localization then makes this term $o_p(1)$.

Set $D_{t,s} = t^{\alpha/2} \eta_s P_{s+1,t-1}$. The same product bound gives $\max_s \|D_{t,s}\| = O(t^{-\alpha/2})$ and $\sum_s \|D_{t,s}\|^2 = O(1)$. On the window $t - s = O(\eta_t^{-1})$, uniformly on bounded rescaled intervals, $\eta_s / \eta_t \rightarrow 1$ and $P_{s+1,t-1} = e^{-H\eta_t(t-s)} + o(1)$. Riemann summation, using $t^\alpha \eta_t \rightarrow c$, gives

$$\sum_{s=N}^{t-1} D_{t,s} S_* D_{t,s}^\top \longrightarrow c \int_0^\infty e^{-Hv} S_* e^{-Hv} dv.$$

The omitted tail beyond $t - s > R/\eta_t$ is bounded by $Ce^{-c_2 R}$ after the negligible range $s < t/2$ is removed. This justifies sending $R \rightarrow \infty$. The common martingale argument proves the first CLT.

If $\alpha = 1$, fix $\nu \in (1/2, \min\{ch, 1\})$. The scaled initial term is $O_p(t^{1/2-\nu}) = o_p(1)$. By (A.7), the localized expected norm of the scaled nonlinear term is at most

$$C_N t^{1/2-\nu} \sum_{s=N}^{t-1} s^{\nu-2} = O_N(t^{1/2-\nu}) = o(1).$$

Set $D_{t,s} = \sqrt{t} \eta_s P_{s+1,t-1}$. Then $\|D_{t,s}\| \leq C t^{-1/2} (s/t)^{\nu-1}$, so the first two conditions in (A.9) hold. Uniformly for s/t in compact subsets of $(0, 1]$,

$$\sqrt{t} D_{t,s} \longrightarrow c u^{cH-I} \quad \text{when } s/t \rightarrow u.$$

The squared-weight sum over $N \leq s \leq \epsilon t$ is bounded by $C\epsilon^{2\nu-1}$. Hence Riemann summation gives

$$\sum_{s=N}^{t-1} D_{t,s} S_* D_{t,s}^\top \longrightarrow c^2 \int_0^1 u^{cH-I} S_* u^{cH-I} du.$$

The martingale argument proves the endpoint CLT.

Finally, the first integral covariance solves $H\Sigma_c + \Sigma_c H = cS_*$. In the second integral, the substitution $u = e^{-v}$ gives $c^2 \int_0^\infty e^{-(cH-I/2)v} S_* e^{-(cH-I/2)v} dv$, which solves the stated endpoint equation. Positive definiteness of H and $cH - I/2$, respectively, gives uniqueness and positive semidefiniteness of these solutions. $\square$

A.7 Polyak–Ruppert averaging

Proof of Theorem 4.4. Suppose first that $\alpha < 1$. Summing $He_t = (e_t - e_{t+1})/\eta_t - \xi_t - r(e_t)$ from N to T gives

$$H \sum_{t=N}^T e_t = - \sum_{t=N}^T \xi_t - \sum_{t=N}^T r(e_t) + R_{N,T},$$

where summation by parts yields

$$R_{N,T} = \frac{e_N}{\eta_N} - \frac{e_{T+1}}{\eta_T} + \sum_{t=N+1}^T (\eta_t^{-1} - \eta_{t-1}^{-1}) e_t.$$

Since $\mathbb{E}\|e_t\| = O(t^{-\alpha/2})$ and $\eta_t^{-1} - \eta_{t-1}^{-1} = O(t^{\alpha-1})$,

$$T^{-1/2} \mathbb{E}\|R_{N,T}\| = O_N(T^{-1/2}) + O(T^{(\alpha-1)/2}) = o(1).$$

The localized nonlinear sum, divided by $\sqrt{T}$, has expected norm at most

$$CT^{-1/2} \sum_{t=N}^T t^{-\alpha} = O(T^{1/2-\alpha}) = o(1).$$

Removing localization and the finite initial segment gives the classical averaging representation [11]

$$\sqrt{T} \bar{e}_T = -H^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^T \xi_t + o_p(1).$$

Taking $D_{T,s} = H^{-1}/\sqrt{T}$ in (A.9) proves the first assertion.

For $\alpha = 1$, use the product representation (A.10) and define

$$B_{T,s} = \eta_s \sum_{t=s+1}^T P_{s+1,t-1}, \quad N \leq s \leq T,$$

Then

$$\sqrt{T} \bar{e}_T = -\frac{1}{\sqrt{T}} \sum_{s=N}^{T-1} B_{T,s} \xi_s - \frac{1}{\sqrt{T}} \sum_{s=N}^{T-1} B_{T,s} r(e_s) + I_{N,T},$$

where $B_{T,T} = 0$ and $I_{N,T}$ contains the finite initial segment and the propagated initial value. For $\nu \in (1/2, \min\{ch, 1\})$, (A.8) implies

$$\|B_{T,s}\| \leq Cs^{\nu-1} \sum_{t=s+1}^T t^{-\nu} \leq C(s/T)^{\nu-1}, \quad I_{N,T} = O_p(T^{1/2-\nu}). \quad (\text{A.11})$$

By (A.7), the localized nonlinear sum is bounded in expected norm by

$$\frac{C_N}{\sqrt{T}} \sum_{s=N}^{T-1} (s/T)^{\nu-1} s^{-1} = O_N(T^{1/2-\nu}) = o(1).$$

Thus both remainders vanish in probability.

To identify the weights, the eigenvalue expansion used in (A.8) gives, uniformly when $\varepsilon T \leq s < t \leq T$,

$$P_{s+1,t-1} = ((s + \gamma)/(t + \gamma))^{cH} + O_\varepsilon(T^{-1}).$$

Consequently, uniformly on compact subsets of $(0, 1]$,

$$B_{T, \lfloor Tu \rfloor} \longrightarrow K_c(u) = c u^{cH-1} \int_u^1 v^{-cH} dv.$$

Both $B_{T,s}$ and $K_c(s/T)$ satisfy the bound $C(s/T)^{\nu-1}$. In particular,

$$\frac{1}{T} \sum_{N \leq s \leq \varepsilon T} (\|B_{T,s}\|^2 + \|K_c(s/T)\|^2) \leq C\varepsilon^{2\nu-1}.$$

Compact convergence and this bound establish

$$\frac{1}{T} \sum_{s=N}^{T-1} \|B_{T,s} - K_c(s/T)\|^2 \longrightarrow 0. \quad (\text{A.12})$$

The relative variance bound, smoothness, and the strong-convexity MSE estimate imply $\sup_s \mathbb{E}\|\xi_s\|^2 < \infty$. Martingale orthogonality therefore shows that replacing $B_{T,s}$ by $K_c(s/T)$ in the noise sum produces an $o_{L^2}(1)$ error. The finitely many terms with $s < N$ vanish because $T^{-1/2}\|K_c(s/T)\| = O(T^{1/2-\nu})$ for fixed s . We have therefore proved

$$\sqrt{T} \bar{e}_T = -\frac{1}{\sqrt{T}} \sum_{s=1}^{T-1} K_c(s/T) \xi_s + o_p(1).$$

For $D_{T,s} = T^{-1/2}K_c(s/T)$, (A.11) gives vanishing maximum weight and a bounded squared-weight sum. The same compact/tail decomposition gives

$$\sum_{s=1}^{T-1} D_{T,s} S_* D_{T,s}^\top \longrightarrow \int_0^1 K_c(u) S_* K_c(u)^\top du.$$

The common martingale CLT argument completes the proof. $\square$

A.8 Nonconvex stationarity

Proof. Smoothness, unbiasedness, and (A.2) give

$$\mathbb{E}_t F(\theta_{t+1}) \leq F(\theta_t) - \eta_t \left(1 - \frac{\ell}{2}(1 + \rho)\eta_t\right) \|\nabla F(\theta_t)\|^2 + \frac{\ell}{2}\sigma^2\eta_t^2.$$

The stated step-size bound makes the descent coefficient at least $\eta_t/2$. Taking expectations, summing, and using $F(\theta_{T+1}) \geq F_{\inf}$ proves the bound and the asserted rates. $\square$

A.9 Moment bounds for nonlinear-feature regression

For the regression model in Section 6, write $Q = Z + U + iV$, $\theta^* = \theta_{\text{true}}$, and

$$A = \text{Re}\{\psi(Q)\psi(Q)^\top\}, \quad b = Y \text{Re}\psi(Q), \quad H = \mathbb{E}[\psi(Z)\psi(Z)^\top].$$

The corrected gradient is $G(\theta) = A\theta - b$. Gaussian tails, the polynomial growth of Y , and the exponential growth of the sine and cosine factors imply that A and b have moments of every finite order. Conditional exact correction gives $\mathbb{E}A = H$ and $\mathbb{E}b = H\theta^*$. The functions $\sin z$, $\cos(0.7z)$, and z^2 are linearly independent on $\mathbb{R}$, and the latent density is positive everywhere. Consequently $H \succ 0$ and

$$F(\theta) = \frac{1}{2}(\theta - \theta^*)^\top H(\theta - \theta^*).$$

Putting $e = \theta - \theta^*$ and $\mu = \lambda_{\min}(H)$, we obtain

$$\xi(\theta) = (A - H)e + (A\theta^* - b), \quad \mathbb{E}\|\xi(\theta)\|^2 \leq 2\mathbb{E}\|A\theta^* - b\|^2 + 2\mathbb{E}\|A - H\|_{\text{op}}^2\|e\|^2.$$

Since $\|e\| \leq \mu^{-1}\|\nabla F(\theta)\|$, this derives (4.1) directly from the data and score moments, with finite constants. This argument verifies the required variance condition for the joint response model, whose response-dependent envelope is discussed in Section 3. The proof of Theorem 4.2 therefore applies. Here $\mu \approx 0.382$ and $c = 2.6$, so $2\mu c > 1$ and the theoretical MSE rate is $O(t^{-1})$. The numerical value of μ is obtained from population quadrature.

B Experimental details

Baseline updates. Both experiments use scalar inputs. Write s^2 for the measurement-noise variance and suppress the iteration index. Taylor corrections use

$$G_{T2} = g(\theta; W, Y) - \frac{s^2}{2}\partial_w^2 g(\theta; W, Y), \quad G_{T4} = G_{T2} + \frac{s^4}{8}\partial_w^4 g(\theta; W, Y),$$

with the response held fixed. The deterministic Woolsey–Huang baseline uses $\text{Re } g^{\mathbb{C}}(\theta; W + is, Y)$; CCSGD instead samples $V \sim N(0, s^2)$ at every update. Regression calibration substitutes the exact conditional mean $\mathbb{E}[Z | W]$ under the stated latent law. Quadratic SIMEX trains three models with additional Gaussian noise variances λs^2 , $\lambda = 0, 1, 2$. It extrapolates parameters in nonlinear-feature regression and predictions in neural-network regression, with weights $(3, -3, 1)$. In the first experiment its additional-noise expectation is integrated analytically; in the second it is sampled. These comparisons use common data streams and step-size schedules; SIMEX uses three training trajectories, so update counts do not equalize computational cost.

Population reference lines. In nonlinear-feature regression, each direct baseline has a population score $H_k\theta - r_k$. The dashed line is $\|H_k^{-1}r_k - \theta^*\|^2$; for SIMEX it is the squared distance between the extrapolation of the three population minimizers and θ^* . We evaluate H_k by 160-node Gauss–Hermite quadrature and r_k by 128×128 joint quadrature over (Z, U) , retaining the dependence between Y and W . Thus the reference lines are numerical evaluations of population squared biases, not fits to the simulated trajectories or lower bounds at every finite iteration.

References

- [1] Denis Belomestny and Alexander Goldenshluger. Density deconvolution under general assumptions on the distribution of measurement errors. *The Annals of Statistics*, 49(2):615–649, 2021.
- [2] Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC, 2 edition, 2006.
- [3] Constructive Characterizations and New Families of van Dantzig Pairs, 2026. Companion manuscript, October 10, 2026.
- [4] J. R. Cook and Leonard A. Stefanski. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328, 1994.
- [5] Václav Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4):1327–1332, 1968.
- [6] Peter Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, 1980.
- [7] Takis Konstantopoulos, Pierre Patie, and Rohan Sarkar. A new class of solutions to the van dantzig problem, the lee–yang property, and the riemann hypothesis. *Annales de l’Institut Fourier*, 74(1):377–421, 2024.

- [8] Po-Ling Loh and Martin J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.
- [9] Takashi Nakamura. Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika*, 77(1):127–137, 1990.
- [10] Steven J. Novick and Leonard A. Stefanski. Corrected score estimation via complex variable simulation extrapolation. *Journal of the American Statistical Association*, 97(458):472–481, 2002.
- [11] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.
- [12] Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In Jagdish S. Rustagi, editor, *Optimizing Methods in Statistics*, pages 233–257. Academic Press, 1971.
- [13] Dylan Spicker, Michael P. Wallace, and Grace Y. Yi. Nonparametric simulation extrapolation for measurement-error models. *Canadian Journal of Statistics*, 52(2):477–499, 2024. Published online in 2023.
- [14] Leonard A. Stefanski. Unbiased estimation of a nonlinear function of a normal mean with application to measurement-error models. *Communications in Statistics—Theory and Methods*, 18(12):4335–4358, 1989.
- [15] Nicholas Woolsey and Xianzheng Huang. Debiased inference in errors-in-variables problems with non-gaussian measurement error. *Annals of the Institute of Statistical Mathematics*, 2026. Published online, 18 June 2026. <https://doi.org/10.1007/s10463-026-00984-x>.