

CCSGD: a reading guide

Aligned English edition | 11 October 2026

1. What problem does the paper solve?

The target is the clean objective $F(\theta) = \mathbb{E}\ell(\theta; Z)$, but only $W = Z + U$ is observed. Even centered input noise generally biases $g(\theta; W)$. Decreasing the learning rate does not remove this population discrepancy. The paper builds an executable stochastic-gradient update for the clean objective.

The main theory assumes Gaussian measurement error with known positive semidefinite covariance, sub-Gaussian latent inputs, and fresh independent samples. Entire extension and input-growth control concern the score, while convexity concerns the objective as a function of the parameter. These are different requirements.

2. How are the problems classified?

First separate correction from optimization. Exact Gaussian correction needs an entire input extension with integrability; it needs no convexity. To obtain optimization guarantees, derive a relative variance bound and then select the appropriate objective class.

Convex objectives: Theorem 4.1 bounds the objective gap of a step-size-weighted average. Strongly convex objectives: Theorem 4.2 proves almost-sure and mean-square convergence to θ^* ; Theorems 4.3–4.4 give last-iterate and uniformly averaged CLTs under extra local conditions. Nonconvex objectives: Theorem 4.5 controls the weighted mean squared gradient, not distance to a global optimizer.

Then classify the noise. Gaussian noise gives the main theorem package. Uniform and symmetric two-point noise admit explicit conjugate perturbations, discussed in Section 7.1. Their exponential tails restrict admissible score growth. Entire scores with excessive growth and arbitrary non-entire scores are outside the automatic guarantee.

3. What is the key mechanism?

Draw $V_t \sim N(0, \Omega)$ independently and set $G_t = \operatorname{Re} g^C(\theta_t; W_t + iV_t)$, $\theta_{t+1} = \theta_t - \eta_t G_t$. The classical complex corrected-score identity makes each update exactly unbiased. The paper connects this oracle to optimization through explicit moment estimates; no closed-form integration is required.

The key envelope is $\|D_z g^C(\theta; q)\| \leq M(\theta)e^{a\|q\|}$ with $M(\theta)^2 \leq A + B\|\nabla F(\theta)\|^2$. Conditional orthogonality separates clean sampling noise from correction error, giving $\mathbb{E}_t\|\xi_t\|^2 \leq \sigma^2 + \rho\|\nabla F(\theta_t)\|^2$. Higher moments are derived from growth and tails, not from a second-moment assumption alone.

4. Which conclusions are easy to reuse?

Use $\eta_t = c/(t + \gamma)^\alpha$, $1/2 < \alpha \leq 1$. Convex objective gaps and nonconvex weighted stationarity are $O(T^{\alpha-1})$ for $\alpha < 1$, and $O(1/\log T)$ at $\alpha = 1$, with the stated step-size cap. Strongly convex MSE is $O(t^{-\alpha})$ for $\alpha < 1$; at $\alpha = 1$ it is $O(t^{-1})$ when $2\mu c > 1$, with logarithmic and slower cases at and below the threshold.

For CLTs, first use convergence, then localize and verify Lindeberg using derived $(2 + \delta)$ moments. Covariance continuity and local Hessian regularity are still required. The endpoint $\alpha = 1$ needs $c\lambda_{\min}(H) > 1/2$ and has a different PR covariance; do not substitute the interior formula $H^{-1}S_*H^{-1}$.

For non-Gaussian conjugates, $a < \lambda$ suffices for the identity and $2a < \lambda$ for the moment reduction, where $\lambda = \pi/b$ for uniform error and $\lambda = \pi/(2b)$ for two-point error. The latter strict inequality also allows a suitable $\delta > 0$. Polynomial scores readily satisfy input moment requirements; amplitude and objective assumptions must still be verified.

5. What do the experiments establish?

Nonlinear-feature regression is strongly convex in the coefficients and has a separate joint-response moment derivation in Appendix A.9. The parameter MSE slope is -0.978 , consistent with $O(t^{-1})$. Dashed reference lines are population asymptotic squared biases of the specified updates, not finite-time lower bounds.

The sine-network experiment measures clean-input prediction MSE and is empirical. Fresh measurement noise is drawn at each presentation of fixed latent training pairs; unbiasedness targets the clean empirical objective. No global network-convergence or zero prediction-error limit is claimed. SIMEX has slightly smaller reported MSE at 200,000 updates. WH denotes their deterministic Eq. (5), not all methods in that work.

6. Recommended reading routes

For the algorithm: Section 3 and Algorithm 1, then Section 5 examples. For a reusable proof: Lemma 3.4 $\rightarrow$ Appendix A.2 $\rightarrow$ the relevant theorem in Section 4. For CLTs: Theorem 4.2 $\rightarrow$ Appendix A.5 $\rightarrow$ A.6 or A.7. For non-Gaussian implementation: Section 7.1 and Algorithm 2. For experimental scope: Section 6, Appendix A.9, and Appendix B.

Agents should start with the five-page AI edition, then resolve stable result IDs in the theorem graph. The original paper is authoritative. The companion conjugate criterion is an external dependency, and the author-provided 6 Astra review reports no major mathematical gaps for main(6).pdf; file-level correspondence with this snapshot is unconfirmed.