

CCSGD: an agent reading edition

Source snapshot: 11 October 2026. Canonical mathematics: `paper/main.pdf`.

1. Contract, scope, and result lookup

This is a five-page extraction, not an additional paper or an independent review. Page 1 fixes the contract; pages 2–5 collect reusable reasoning and implementation lessons. Stable IDs resolve through `theorem_graph/theorem_graph.json` to the original source and proof.

Problem. Minimize $F(\theta) = \mathbb{E}\ell(\theta; Z)$, $\theta \in \mathbb{R}^p$, while observing $W = Z + U$. In the main theorems, Z is sub-Gaussian, $U \sim N(0, \Omega)$ is independent of Z , and the known $\Omega \succeq 0$ may be correlated or singular. Samples are fresh and independent of the past.

Executable oracle. With independent $V_t \sim N(0, \Omega)$,

$$G_t = \operatorname{Re} g^{\mathbb{C}}(\theta_t; W_t + iV_t), \quad \theta_{t+1} = \theta_t - \eta_t G_t.$$

$g = \nabla_{\theta} \ell$ is extended in the *input*, with real parameters. No symbolic integration is required.

Input assumptions. $g^{\mathbb{C}}(\theta; \cdot)$ is entire and real on real inputs;

$$\|D_z g^{\mathbb{C}}(\theta; q)\|_{\text{op}} \leq M(\theta) e^{a\|q\|}, \quad M(\theta)^2 \leq A + B\|\nabla F(\theta)\|^2.$$

a, A, B are fixed finite constants. Differentiation and expectation commute, and the *clean* score obeys $\mathbb{E}\|g(\theta; Z) - \nabla F(\theta)\|^2 \leq \sigma_0^2 + \rho_0\|\nabla F(\theta)\|^2$. The corrected-score variance is derived, not assumed. No convexity is required for correction.

Stable ID / paper	Use	Proof
R-EXACT / L3.4	Exact Gaussian correction	A.1
R-VAR, R-HM	Relative variance; higher moments	A.2
T-CONVEX / T4.1	Weighted objective-gap bound	A.3
T-STRONG / T4.2	Almost-sure convergence and MSE	A.4
T-LAST / T4.3	Last-iterate CLT	A.5–A.6
T-PR / T4.4	Uniform-average CLT	A.5, A.7
T-NONCONVEX / T4.5	Weighted stationarity bound	A.8
R-REG, R-NG	Joint-response example; conjugate extension	A.9; §7.1

Attribution and evidence. Gaussian complex corrected scores are classical (Novick–Stefanski, 2002). The paper connects an executable online oracle to SGD through explicit moment estimates. Non-Gaussian conjugate existence and its mean-value identity come from a companion manuscript. Neural-network curves are empirical evidence, not a global convergence theorem. A user-provided 6 Astra review reports no major mathematical gaps for `main(6).pdf`; its file-level match to this snapshot is not confirmed.

2. Takeaways: exact cancellation is an oracle design principle

Takeaway 1: restore the clean gradient, rather than shrinking a persistent bias. A smaller step size cannot generally repair an observed-data gradient with the wrong population mean. CCSGD changes the oracle so that $\mathbb{E}_t G_t = \nabla F(\theta_t)$, where $\mathbb{E}_t = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$. This separates the measurement-error problem from the choice of optimization geometry. The statement concerns exact expectation; a realized update remains noisy.

Takeaway 2: prove the identity through symmetry before manipulating infinite series. For $D = U + iV$, the common Gaussian covariance gives $e^{i\phi} D \stackrel{d}{=} D$. For each fixed d , $s \mapsto g^C(\theta; z + sd)$ is entire, hence

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

The derivative envelope ensures integrability. Average over D and use rotational invariance to obtain R-EXACT. This proof works for vector scores and singular Ω ; it does not require a Gaussian latent input or coordinatewise independent measurement errors. The generalizable technique is *invariance + holomorphic mean value + explicit domination*.

Takeaway 3: identify the conditioning that makes the correction error orthogonal. Write

$$\zeta_t = g(\theta_t; Z_t) - \nabla F(\theta_t), \quad \Delta_t = G_t - g(\theta_t; Z_t), \quad \tilde{\zeta}_t = \zeta_t + \Delta_t.$$

Exact correction gives $\mathbb{E}[\Delta_t \mid \mathcal{F}_{t-1}, Z_t] = 0$. Therefore

$$\mathbb{E}_t \langle \zeta_t, \Delta_t \rangle = 0, \quad \mathbb{E}_t \|\tilde{\zeta}_t\|^2 = \mathbb{E}_t \|\zeta_t\|^2 + \mathbb{E}_t \|\Delta_t\|^2.$$

This avoids a factor-two loss from a generic norm inequality. With a response, condition also on Y_t and keep it fixed in the complex evaluation.

Takeaway 4: distinguish simulation averaging from observation batching. For $h_\theta(W) = \mathbb{E}_V[G(\theta) \mid W]$, put $v_{\text{obs}} = \mathbb{E} \|h_\theta(W) - \nabla F(\theta)\|^2$ and $v_{\text{sim}} = \mathbb{E} \|G(\theta) - h_\theta(W)\|^2$. With n independent observations and m independent simulations per observation,

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

Repeating virtual noise at one W reduces only simulation variance. Taking $m \rightarrow \infty$ does not create a clean observation. The exact correction identity is conditioned on Z ; $h_\theta(W)$ is generally not $g(\theta; Z)$ for a realized pair.

Reuse boundary. A finite number of Monte Carlo draws preserves unbiasedness when each draw has the correct law and required integrability. A fixed imaginary shift, clipping, truncation, or a modulus-square loss changes the oracle and requires a fresh bias calculation.

3. Takeaways: a parameter-dependent envelope closes the variance argument

Takeaway 5: control input derivatives, and let their amplitude grow with the objective gradient. The fundamental theorem of calculus gives

$$\|\Delta(\theta)\| \leq M(\theta)\|D\|e^{a(\|Z\|+\|D\|)}, \quad K_a = \mathbb{E}e^{2a\|Z\|}\mathbb{E}[\|D\|^2e^{2a\|D\|}] < \infty.$$

Independence and orthogonality imply

$$\mathbb{E}_t\|\xi_t\|^2 \leq \sigma^2 + \rho\|\nabla F(\theta_t)\|^2, \quad \sigma^2 = \sigma_0^2 + K_a A, \quad \rho = \rho_0 + K_a B.$$

This is the reusable bridge R-VAR. A global constant bound on M is unnecessarily restrictive. Conversely, finiteness at each fixed θ does not supply deterministic relative-variance constants along an arbitrary trajectory. Verify the amplitude relation instead of assuming the desired corrected-oracle conclusion.

Takeaway 6: remove a deterministic anchor before estimating centered moments. Let $b_a(q) = \|q\|e^{a\|q\|}$ and $Q = Z + U + iV$. The anchor $g(\theta; 0)$ cancels:

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

Sub-Gaussian and Gaussian tails give every finite moment of the bracket. For $r \geq 2$,

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

The clean centered score has the same type of estimate using $Q = Z$. A second-moment assumption alone would not imply this.

Takeaway 7: obtain convergence before using trajectory bounds in a CLT. Strong convexity and R-VAR first yield almost-sure convergence via a squared-distance recursion and Robbins–Siegmund. Then $\sup_t \|\nabla F(\theta_t)\| < \infty$ almost surely. The resulting conditional $(2 + \delta)$ -moment bound may have a *random* finite constant. On a fixed neighborhood of θ^* it is deterministic. Vanishing maximum martingale weights and bounded sums of squared weights imply Lindeberg. This avoids circularly assuming stability to prove stability. Covariance continuity and local Hessian regularity remain separate requirements.

Takeaway 8: exploit affine parameter structure when responses enter the envelope. For nonlinear-feature regression, $G(\theta) = \mathcal{A}\theta - b$ and $\nabla F(\theta) = H(\theta - \theta^*)$. If $H \succeq \mu I$ and the coefficients have finite second moments,

$$\mathbb{E}\|\xi(\theta)\|^2 \leq 2\mathbb{E}\|\mathcal{A}\theta^* - b\|^2 + 2\mu^{-2}\mathbb{E}\|\mathcal{A} - H\|_{\text{op}}^2 \|\nabla F(\theta)\|^2.$$

Appendix A.9 verifies this for the actual joint (Z, Y) experiment. A response-dependent envelope must not silently be replaced by a deterministic function of θ alone.

4. Takeaways: endpoint step sizes change the limit laws

Common schedule. $\eta_t = c/(t + \gamma)^\alpha$, $c > 0$, $\gamma \geq 0$, $1/2 < \alpha \leq 1$. Convex and nonconvex bounds additionally impose $\eta_1 \leq [L(1 + \rho)]^{-1}$. These are upper bounds, not lower bounds or optimal-complexity claims.

Geometry / output	$1/2 < \alpha < 1$	$\alpha = 1$
Convex / η -weighted average	Objective gap $O(T^{\alpha-1})$	$O(1/\log T)$
Strongly convex / last iterate	MSE $O(t^{-\alpha})$	$O(t^{-1})$ if $2\mu c > 1$; $O(t^{-1} \log t)$ if equality; $O(t^{-2\mu c})$ otherwise
Nonconvex / weighted gradients	Mean squared gradient $O(T^{\alpha-1})$	$O(1/\log T)$

Nonconvex stationarity does not imply convergence to a global minimizer or a planted parameter.

Takeaway 9: the last-iterate covariance depends on the schedule. Let $H = \nabla^2 F(\theta^*)$ and $S_* = \mathbb{E}[\xi(\theta^*)\xi(\theta^*)^\top]$. For $\alpha < 1$, $t^{\alpha/2}(\theta_t - \theta^*) \Rightarrow N(0, \Sigma_c)$ with $H\Sigma_c + \Sigma_c H = cS_*$. For $\alpha = 1$, require $c\lambda_{\min}(H) > 1/2$ and replace this by

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

The $1/2$ shift comes from the $\sqrt{t}$ normalization. S_* includes measurement and simulation noise; exact unbiasedness does not equate it with the clean-oracle covariance.

Takeaway 10: uniform averaging has an exceptional $1/t$ endpoint. For $\bar{\theta}_T = T^{-1} \sum_{t=1}^T \theta_t$, the usual PR covariance is $H^{-1}S_*H^{-1}$ when $\alpha < 1$. At $\alpha = 1$, under the same spectral condition,

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

The proof retains the full time-dependent averaging weights; simply reusing the interior PR formula is wrong. For example, in one dimension with $cH = 1$, $K_c(u) = -c \log u$ and $\Xi_c = 2c^2S_* = 2H^{-2}S_*$. This scalar specialization is derived here from T4.4, not a separately numbered paper theorem.

Proof-reuse rule. Both CLTs use T-STRONG and the common localization/martingale module R-CLTPREP. T4.4 shares local assumptions with T4.3; it does not require T4.3 as an already-proved CLT. Distinguish the weighted convex average from the uniform PR average.

5. Takeaways: conjugate laws, growth thresholds, and empirical scope

Takeaway 11: non-Gaussian cancellation is a distribution-design problem. For symmetric scalar U with $M_U(t) = \mathbb{E}e^{tU} < \infty$ for all real t , an independent conjugate has $\varphi_V(t) = 1/M_U(t)$ when this is a characteristic function. The companion manuscript supplies existence criteria. Its moment cancellation gives exact correction for polynomials and entire scores with an absolutely integrable Taylor expansion. It does not authorize arbitrary entire scores.

For $R \sim \text{Unif}(0, 1)$ independent of all observations,

$$\begin{aligned} U \sim \text{Unif}[-b, b] : \quad V &= \frac{b}{\pi} \log \frac{R}{1-R}, \quad \lambda = \pi/b, \\ \Pr(U = \pm b) = 1/2 : \quad V &= \frac{2b}{\pi} \log \tan(\pi R/2), \quad \lambda = \pi/(2b). \end{aligned}$$

Use these V laws in the same complex-gradient update, not independent copies of U . Here $\mathbb{E}e^{s|V|} < \infty$ precisely for $0 \leq s < \lambda$.

Takeaway 12: matching growth to conjugate tails is indispensable. With the paper’s input derivative exponent a , $a < \lambda$ suffices for the entire-score identity; $2a < \lambda$ supplies finite correction variance. This strict inequality permits a $\delta > 0$ with $(2 + \delta)a < \lambda$, giving the needed higher moments. Polynomial scores permit arbitrarily small positive exponents, after adjusting M ; the amplitude condition still needs verification. Gaussian conjugates have no finite linear-exponential threshold. The thresholds are sufficient envelope criteria. Their failure alone does not prove divergence for a specific score. More simulations or a smaller step size cannot turn an actually infinite one-step variance into a finite one.

Implementation invariants.

- Extend the analytic score in the noisy input. For squared loss use an ordinary complex square, not a modulus square; apply the analytic chain rule and take the real part after the complete gradient.
- Generate fresh independent virtual noise at every update. For fixed latent training pairs, fresh measurement noise preserves unbiasedness for the clean *empirical* objective, not automatically a population objective.
- Entire input maps need not have a parameter-independent growth exponent. Trainable sine-network input weights change this exponent; stacked sine layers can grow faster than exponentially. Neither property can be inferred from real-input boundedness.

Experimental evidence and comparator boundary. The nonlinear-feature regression is strongly convex in its coefficients, with planted $\theta^* = (0.8, -0.6, 0.2)$ and a fitted MSE slope -0.978 over 5×10^4 – 10^7 updates. Dashed baselines are computed asymptotic squared biases, not finite-time lower bounds. The neural experiment is a trainable $1 \rightarrow 32 \rightarrow 1$ sine network; its curves do not establish a global theorem. At 200,000 updates SIMEX slightly outperforms CCSGD in reported mean prediction MSE. The WH comparison is specifically deterministic Eq. (5), not every correction in that paper. SIMEX uses three trajectories, so equal update counts do not equalize computation.