CORRECTING OPTIMIZER-INDUCED BIAS IN ADAPTIVE ZEROTH-ORDER OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Zeroth-order (ZO) optimization adapts first-order optimizers by substituting exact gradients with zeroth-order gradient estimators (ZOGEs). This plug-in approach propagates ZOGE noise through the optimizer, thereby affecting both the update direction and the states retained across iterations. We derive joint bias and covariance expansions that characterize the resulting update and state errors. These expansions separate finite-radius bias from finite sampling bias, whose leading term scales as q^{-1} (where q is the number of independent sampled directions), and identify a leading q^{-1} covariance term. We introduce *Corrected Same-Query Extrapolation (C-SQE)*, which reduces finite sampling bias from $\mathcal{O}(q^{-1})$ to $\mathcal{O}(q^{-2})$ while preserving the leading covariance term by extrapolating optimizer outputs from existing ZOGEs, with a rational correction extending these guarantees to constrained states. Our convergence bounds for Adam, RMSProp, regularized Muon, and R-AdaZO quantify the improved q-dependence of sampling bias under C-SQE while retaining the leading stochastic terms. Synthetic experiments validate the predicted bias and covariance scaling and demonstrate faster optimization across all four optimizer families, while OPT-125M soft-prompt tuning on SST-2 and COPA shows lower mean training loss across all four families on both tasks.

1 INTRODUCTION

Zeroth-order (ZO) optimization enables learning from function evaluations when gradients are unavailable or expensive to compute. Its applications include black-box adversarial attacks (Chen et al., 2019), policy search and simulated control (Salimans et al., 2017), and hyperparameter optimization (Gu et al., 2021). For large language models (LLMs), forward-only ZO updates enable memory-efficient fine-tuning, providing an alternative when the memory cost of backpropagation limits adaptation (Malladi et al., 2023).

A substantial body of ZO research focuses on improving the accuracy of zeroth-order gradient estimation (ZOGE) under limited function-query budgets. ZOGE methods typically combine directional finite differences, whose errors arise from both finite perturbation radii and random direction sampling. Higher-order smoothing kernels (Bach & Perchet, 2016) and extrapolation (Dong et al., 2026) reduce finite-radius bias, while variance-reduction methods improve the accuracy of sampled gradient information (Liu et al., 2018; Ji et al., 2019; Lang et al., 2026). These advances improve the gradient information available for optimization.

These improved gradient estimates serve as plug-in substitutes for exact gradients in adaptive optimizers, which compute update directions and update persistent states such as momentum and second-moment estimates. Existing work on ZO optimizer design emphasizes efficient update rules and their convergence properties, as exemplified by MeZO's memory-efficient stochastic-gradient updates (Malladi et al., 2023) and ZO-AdaMM's adaptive momentum method (Chen et al., 2019). More recently, R-AdaZO improves second-moment estimation through first-moment averaging to refine adaptive normalization (Shu et al., 2025). We study how ZOGE noise affects adaptive optimizers through a joint analysis of the update direction and persistent state. This motivates the following central question:

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.

> ***Q1*: How does ZOGE noise affect both the update direction and the persistent states of an adaptive optimizer?**

We answer *Q1* by jointly characterizing the conditional bias and covariance of the optimizer's update and state errors, measured relative to the corresponding exact-gradient outputs. The bias contains a finite-radius component inherited from gradient estimation and a sampling component with a leading q^{-1} term, where q is the number of independent sampled directions. The finite-radius component is unaffected by increasing q. The sampling component arises from the interaction between directional-estimate covariance and the curvature of the optimizer's update and state mappings, so even unbiased gradient estimates can produce biased updates and states. The leading covariance scales as q^{-1} and reflects how the optimizer's Jacobian propagates directional-estimate noise to its update and state outputs.

At a fixed number of directions q, the sampling bias can remain nonzero as the perturbation radius vanishes. Reducing this bias by increasing q requires additional function evaluations. This motivates the following question:

> ***Q2*: Can we reduce the finite sampling bias with the fixed gradient estimates?**

We answer *Q2* by proposing *Corrected Same-Query Extrapolation (C-SQE)*, which combines extrapolation of optimizer updates and states with a rational correction that keeps second-moment states nonnegative. C-SQE combines optimizer outputs computed from the full set of directional estimates and two complementary halves. The extrapolation cancels the leading q^{-1} sampling-bias term, leaving an $\mathcal{O}(q^{-2})$ sampling residual in the extrapolated outputs while preserving their leading covariance. Reusing the same directional estimates for all three optimizer evaluations keeps the function-query budget unchanged.

Our contributions are as follows.

• **Joint analysis of plug-in update and state errors.** Our conditional bias and covariance expansions separate finite-radius from sampling bias while capturing the dependence between the update and state errors introduced by the plug-in approach.

• **C-SQE: Bias correction using existing directional estimates.** For optimizers with unconstrained states, including regularized Muon (Jordan et al., 2024), C-SQE reduces sampling bias from $\mathcal{O}(q^{-1})$ to $\mathcal{O}(q^{-2})$ while preserving the leading covariance. For Adam, RMSProp, and R-AdaZO (Kingma & Ba, 2014; Hinton et al., 2012; Shu et al., 2025), a rational correction keeps second-moment states nonnegative and preserves these guarantees when the second moments evaluated at the mean gradient input are strictly positive.

• **Improved convergence bounds.** We establish a common convergence bound for Adam, RMSProp, and regularized Muon, together with a corresponding bound for R-AdaZO. In these bounds, C-SQE improves the q-dependence of the squared update-bias contribution from q^{-2} to q^{-4} while preserving the leading stochastic terms.

• **Empirical validation.** Experiments on the Quadratic function validate the predicted bias and covariance scaling. Optimization experiments on the Quadratic, Cubic, Lévy, and Rosenbrock functions show faster objective decrease across four optimizer families at matched function-query budgets. Soft-prompt tuning of a frozen OPT-125M model yields lower mean training loss across all four families on both SST-2 and COPA, with higher mean test accuracy in all four families on SST-2 and three on COPA.

2 Preliminaries and Related Work

2.1 Zeroth-Order Optimization

Problem formulation. We consider the stochastic optimization problem $\min_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x}) := \mathbb{E}_\xi[f(\boldsymbol{x};\xi)]$, where $\boldsymbol{x} \in \mathbb{R}^d$ is the parameter vector, ξ denotes the random variable representing oracle noise, and $f(\boldsymbol{x};\xi)$ is measurable in ξ for each fixed $\boldsymbol{x}$. The objective function can only be accessed through noisy function evaluations.

ZO gradient estimator. Let $\mathcal{F}_t$ denote the information available before the queries at iteration t, including the current parameter and any incoming optimizer state. Draw q i.i.d. pairs $(\boldsymbol{z}_i, \xi_i)$ independently of $\mathcal{F}_t$, where $\boldsymbol{z}_i \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ is independent of ξ_i and $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$. For a radius $r > 0$, use the same oracle sample for both the positive and negative evaluations, we define the central zeroth-order gradient estimator (ZOGE) as

$$\widehat{\boldsymbol{g}}_i(r) := \frac{d}{2r}\left(f(\boldsymbol{x}_t + r\boldsymbol{z}_i; \xi_i) - f(\boldsymbol{x}_t - r\boldsymbol{z}_i; \xi_i)\right)\boldsymbol{z}_i. \tag{1}$$

The averaged ZOGE is $\widehat{\boldsymbol{g}}_{q,r} := q^{-1}\sum_{i=1}^{q}\widehat{\boldsymbol{g}}_i(r)$, which requires $2q$ function evaluations.

We impose the following regularity conditions on the objective and oracle.

Assumption 2.1 (Objective regularity)**.** *The objective F is twice continuously differentiable on $\mathbb{R}^d$ and bounded below by $\underline{F}$. For all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,*

$$\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\|_2 \le L\|\boldsymbol{x} - \boldsymbol{y}\|_2, \qquad \|\nabla^2 F(\boldsymbol{x}) - \nabla^2 F(\boldsymbol{y})\|_{\mathrm{op}} \le L_H\|\boldsymbol{x} - \boldsymbol{y}\|_2.,$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the induced operator norm.

Assumption 2.2 (Oracle regularity)**.** *Almost surely, $f(\cdot;\xi)$ is G-Lipschitz on $\mathbb{R}^d$, with G independent of ξ, i.e., for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,*

$$|f(\boldsymbol{x};\xi) - f(\boldsymbol{y};\xi)| \le G\|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

Because the two evaluations in (1) share the same oracle sample, Assumption 2.2 implies that $\|\widehat{\boldsymbol{g}}_i(r)\|_2 \le d\,G$ almost surely, uniformly in $r > 0$.

2.2 Optimizer Transitions

Stateful optimizers combine a gradient input with persistent states to produce an update direction and the next state. Examples include Adam, RMSProp, and Muon (Kingma & Ba, 2014; Hinton et al., 2012; Jordan et al., 2024). With the current parameter $\boldsymbol{x}_t$, incoming state $\boldsymbol{s}_t$, and optimizer coefficients held fixed, define the joint update–state transition

$$\mathcal{T}_t(\boldsymbol{g}) := [\mathcal{U}_t(\boldsymbol{g}); \mathcal{S}_t(\boldsymbol{g})] \in \mathbb{R}^p, \tag{2}$$

where $\mathcal{U}_t(\boldsymbol{g}) \in \mathbb{R}^d$ is the update direction and $\mathcal{S}_t(\boldsymbol{g}) \in \mathbb{R}^{p-d}$ is the next state. For a learning rate $\eta_t > 0$, the optimizer executes

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t\,\mathcal{U}_t(\boldsymbol{g}), \qquad \boldsymbol{s}_{t+1} = \mathcal{S}_t(\boldsymbol{g}). \tag{3}$$

Plug-in ZO methods, including ZO-AdaMM and ZO-Muon (Chen et al., 2019; Petrov et al., 2025), replace $\boldsymbol{g}$ with $\widehat{\boldsymbol{g}}_{q,r}$ in this transition.

We specialize this framework to Adam, RMSProp, and regularized Muon in the convergence analysis, and also include R-AdaZO (Shu et al., 2025) to study its interaction with moment averaging.

Adam and R-AdaZO. We adopt the Adam formulation as used in ZO-AdaMM, with a positive stabilizer added inside the square root. For the incoming state $\boldsymbol{s}_t = [\boldsymbol{m}_t; \boldsymbol{v}_t]$, with $\boldsymbol{v}_t \succeq \boldsymbol{0}$, define

$$\mathcal{M}_t(\boldsymbol{g}) := \beta_1\boldsymbol{m}_t + (1-\beta_1)\boldsymbol{g},\ \mathcal{W}_t(\boldsymbol{g}) := \beta_2\boldsymbol{v}_t + (1-\beta_2)\boldsymbol{g}^{\odot 2},\ \mathcal{U}_t(\boldsymbol{g}) := \mathcal{M}_t(\boldsymbol{g}) \oslash \sqrt{\mathcal{W}_t(\boldsymbol{g}) + \epsilon_A^2\boldsymbol{1}}. \tag{4}$$

The next-state map is $\mathcal{S}_t(\boldsymbol{g}) := [\mathcal{M}_t(\boldsymbol{g}); \mathcal{W}_t(\boldsymbol{g})]$, with $\beta_1, \beta_2 \in [0, 1)$ and $\epsilon_A > 0$. Here $\odot$ and $\oslash$ denote coordinatewise multiplication and division; powers and square roots are also taken coordinatewise. The moments are initialized at $\boldsymbol{m}_0 = \boldsymbol{v}_0 = \boldsymbol{0}$. R-AdaZO retains the same first-moment map and normalization but replaces $\boldsymbol{g}^{\odot 2}$ in the second-moment recursion with $\mathcal{M}_t(\boldsymbol{g})^{\odot 2}$.

RMSProp. RMSProp normalizes the gradient input directly and retains only a nonnegative second-moment state. For the incoming state $\boldsymbol{s}_t = \boldsymbol{w}_t \succeq \boldsymbol{0}$, define

$$\mathcal{W}_t^{\mathrm{RMS}}(\boldsymbol{g}) := \beta_R\boldsymbol{w}_t + (1-\beta_R)\boldsymbol{g}^{\odot 2}, \qquad \mathcal{U}_t^{\mathrm{RMS}}(\boldsymbol{g}) := \boldsymbol{g} \oslash \sqrt{\mathcal{W}_t^{\mathrm{RMS}}(\boldsymbol{g}) + \epsilon_R^2\boldsymbol{1}}. \tag{5}$$

The next-state map is $\mathcal{S}_t^{\mathrm{RMS}} = \mathcal{W}_t^{\mathrm{RMS}}$, with $\beta_R \in [0, 1)$, $\epsilon_R > 0$, and $\boldsymbol{w}_0 = \boldsymbol{0}$.

Regularized Muon. Muon transforms a matrix momentum into an update via orthogonalization. For matrix parameters in $\mathbb{R}^{d_1 \times d_2}$, the incoming state is $\boldsymbol{s}_t = \mathrm{vec}(\boldsymbol{M}_t)$, where $\boldsymbol{M}_t$ is the unrestricted momentum matrix. Using the smooth polar map $\Psi_{\epsilon_M}(\boldsymbol{Z}) := \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z} + \epsilon_M^2 \boldsymbol{I})^{-1/2}$, define

$$\mathcal{M}_t^{\mathrm{Mu}}(\boldsymbol{G}) := \beta_M \boldsymbol{M}_t + (1-\beta_M)\boldsymbol{G}, \qquad \mathcal{U}_t^{\mathrm{Mu}}(\boldsymbol{G}) := \Psi_{\epsilon_M}\big(\mathcal{M}_t^{\mathrm{Mu}}(\boldsymbol{G})\big). \tag{6}$$

The next-state map in matrix form is $\mathcal{S}_t^{\mathrm{Mu}}(\boldsymbol{G}) = \mathcal{M}_t^{\mathrm{Mu}}(\boldsymbol{G})$, with $\beta_M \in [0,1)$, $\epsilon_M > 0$, and $\boldsymbol{M}_0 = \boldsymbol{0}$. Vectorizing the update and state blocks yields (2) with $d = d_1 d_2$; Euclidean norms then coincide with Frobenius norms.

The local transition analysis uses the following regularity condition.

Assumption 2.3 (Optimizer regularity)**.** *For fixed t, r, and history $\mathcal{F}_t$, the map $\mathcal{T}_t$ is K times continuously differentiable, with $K \geq 1$, on an open convex domain $\mathcal{D}$ containing $\nabla F(\boldsymbol{x}_t)$ and the conditional support of $\widehat{\boldsymbol{g}}_i(r)$. The derivatives of $\mathcal{T}_t$ satisfy*

$$\|\nabla \boldsymbol{g}^k \mathcal{T}_t(\boldsymbol{g})\|_{\mathrm{op}} \leq M_k, \qquad \boldsymbol{g} \in \mathcal{D}, \quad k = 1, \ldots, K,$$

where $\mathcal{D}$ and M_k are independent of q.

Remark 2.1. *The Adam/R-AdaZO, RMSProp, and regularized Muon transitions above satisfy Assumption 2.3 for every finite K; see Proposition A.1 in Appendix A.4.*

2.3 Related Work

Zeroth-order gradient estimation. Classical analyses characterize the smoothing bias and dimension-dependent estimation error of randomized ZOGEs (Duchi et al., 2015; Nesterov & Spokoiny, 2017). Higher-order smoothing kernels (Bach & Perchet, 2016; Akhavan et al., 2024) and extrapolation (Dong et al., 2026) improve the bias order. Variance reduction methods use orthogonal directions (Feng & Wang, 2023), control variates (Liu et al., 2018; Ji et al., 2019), and incremental Jacobian estimates for finite-sum composite problems (Zhang & Tang, 2026). Other approaches incorporate temporal gradient priors (Ilyas et al., 2018), reuse historical coordinate estimates through partial refreshing (Liang et al., 2026), or infer subspace-projected gradients through Bayesian estimation (Feng & Huang, 2026). These methods improve the gradient information supplied to an optimizer.

Zeroth-order optimizer design. ZO methods adapt first-order optimizers by substituting ZOGEs into stochastic gradient updates (Ghadimi & Lan, 2013; Malladi et al., 2023), Adam-style normalization (Chen et al., 2019), and Muon-style matrix orthogonalization (Petrov et al., 2025; Lang et al., 2026). Further designs employ sign-based updates (Liu et al., 2019), normalized updates (Dang et al., 2026), curvature-based preconditioning (Zhao et al., 2025; Seung et al., 2026), and sharpness-aware optimization (Liu et al., 2025). These approaches develop update rules for function-based optimization, motivating the analysis of how estimation noise propagates through the resulting optimizer transitions.

Noise in adaptive ZO optimization. R-AdaZO constructs its second moment from the averaged first moment and establishes moment-error and variance-aware convergence guarantees (Shu et al., 2025). Related work identifies an isotropic noise contribution to second-moment estimates and motivates scalar adaptation in high dimensions (Dbouk et al., 2026). These analyses explain how ZOGE noise affects moment estimation and adaptive normalization.

Extrapolation bias correction. Extrapolation methods combine estimates computed at different sample sizes to cancel leading terms in bias expansions (Quenouille, 1956; Jiao & Han, 2020). The half-panel extrapolation combines full- and half-panel estimates to remove first-order bias while retaining the leading asymptotic variance under regularity conditions (Dhaene & Jochmans, 2015). C-SQE applies this principle to optimizer updates and next states, combining same-query extrapolation with a rational correction for constrained states to ensure feasibility.

3 Plug-in Transition Error and Proposed Method

A nonlinear optimizer can convert gradient-estimation noise into systematic error in both its update direction and persistent states. We characterize this joint error and introduce Corrected Same-Query

Extrapolation (C-SQE), which cancels the leading sampling bias using the original function evaluations.

3.1 Plug-in Transition Error

Fix an iteration index t and radius $r > 0$, and condition on the pre-query history $\mathcal{F}_t$, which fixes the current parameter, the incoming optimizer state, and the optimizer coefficients. Write $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$ and $\mathrm{Cov}_t(\cdot) := \mathrm{Cov}(\cdot \mid \mathcal{F}_t)$ for conditional expectation and covariance, respectively, and let $\boldsymbol{g}_t := \nabla F(\boldsymbol{x}_t)$. The following theorem separates the finite-radius error from the optimizer-induced sampling bias and characterizes the covariance of the joint transition.

Theorem 3.1 (Joint bias and covariance of the plug-in transition)**.** *Under Assumptions 2.1, 2.2, and 2.3 with $K \geq 4$, there exist constants P_b, P_V independent of $q \geq 2$ such that*

$$\begin{aligned} &\text{Bias:} && \left\|\{\mathbb{E}_t[\mathcal{T}_t(\widehat{\boldsymbol{g}}_{q,r})] - \mathcal{T}_t(\boldsymbol{g}_t)\} - \{\mathcal{T}_t(\boldsymbol{g}_r) - \mathcal{T}_t(\boldsymbol{g}_t)\} - q^{-1}\boldsymbol{B}_r\right\|_2 \leq P_b q^{-2}, \\ &\text{Covariance:} && \left\|\mathrm{Cov}_t(\mathcal{T}_t(\widehat{\boldsymbol{g}}_{q,r})) - q^{-1}\boldsymbol{J}_r\boldsymbol{\Sigma}_r\boldsymbol{J}_r^\top\right\|_{\mathrm{op}} \leq P_V q^{-2}, \end{aligned} \tag{7}$$

where $\boldsymbol{g}_r := \mathbb{E}_t[\widehat{\boldsymbol{g}}_i(r)]$, $\boldsymbol{\Sigma}_r := \mathrm{Cov}_t(\widehat{\boldsymbol{g}}_i(r))$, $\boldsymbol{J}_r := \nabla_{\boldsymbol{g}}\mathcal{T}_t(\boldsymbol{g}_r)$, and the bias coefficient has components $[\boldsymbol{B}_r]_\ell := \frac{1}{2}\,\mathrm{tr}\big(\nabla^2_{\boldsymbol{g}}[\mathcal{T}_t]_\ell(\boldsymbol{g}_r)\boldsymbol{\Sigma}_r\big)$ for $\ell = 1, \ldots, p$.

The finite-radius term $\mathcal{T}_t(\boldsymbol{g}_r) - \mathcal{T}_t(\boldsymbol{g}_t)$ is inherited from the ZOGE and vanishes when $\boldsymbol{g}_r = \boldsymbol{g}_t$. The sampling-bias term $q^{-1}\boldsymbol{B}_r$ couples transition curvature with the directional-estimate covariance and can remain nonzero even when the ZOGE is unbiased. The proof is given in Appendix A.1.

3.2 Corrected Same-Query Extrapolation

C-SQE exploits the inverse scaling of the leading bias with the sample size by combining full- and half-sample transitions. Each transition evaluation keeps the current parameter, incoming state, and optimizer coefficients fixed.

For even $q \geq 2$, choose a balanced partition (I_1, I_2) of $\{1, \ldots, q\}$ before collecting the queries, with $|I_1| = |I_2| = q/2$. The half-sample estimates satisfy

$$\widehat{\boldsymbol{g}}_j := \frac{2}{q}\sum_{i \in I_j}\widehat{\boldsymbol{g}}_i(r), \qquad j \in \{1, 2\}, \qquad \frac{\widehat{\boldsymbol{g}}_1 + \widehat{\boldsymbol{g}}_2}{2} = \widehat{\boldsymbol{g}}_{q,r}. \tag{8}$$

By Theorem 3.1, each half-sample transition carries twice the leading sampling bias of the full-sample transition. Subtracting their average from twice the full-sample transition therefore cancels this term.

At each iteration, let P_t be an optimizer-specific state operator that preserves the update block and enforces state feasibility, using the full-sample state $\boldsymbol{s}_F := \mathcal{S}_t(\widehat{\boldsymbol{g}}_{q,r})$ as a reference. Its dependence on this reference is suppressed in the notation. For unrestricted states, P_t reduces to the identity operator E on the joint output. The executed C-SQE transition is

$$\mathcal{T}^{\mathrm{C}}_{t,q,r} := \mathrm{P}_t\left(2\mathcal{T}_t(\widehat{\boldsymbol{g}}_{q,r}) - \frac{1}{2}\sum_{j=1}^{2}\mathcal{T}_t(\widehat{\boldsymbol{g}}_j)\right). \tag{9}$$

Writing $\mathcal{T}^{\mathrm{C}}_{t,q,r} = [\boldsymbol{a}^{\mathrm{C}}_{t,q,r}; \boldsymbol{s}_{t+1}]$, the algorithm retains $\boldsymbol{s}_{t+1}$ and directly applies the extrapolated update direction:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \boldsymbol{a}^{\mathrm{C}}_{t,q,r}. \tag{10}$$

The averaging identity in (8) preserves every affine transition block and the leading linear fluctuation of the extrapolation within P_t. Theorem 4.1 establishes the joint bias and covariance guarantees for identity processing, and Corollary 4.1 extends these guarantees to the specified constrained-state rules under the conditions stated therein.

Remark 3.1 (Optimizer-specific state handling)**.** *Regularized Muon has an unrestricted momentum state, so $\mathrm{P}_t = \mathrm{E}$. Adam and RMSProp require nonnegative second moments; P_t maps the candidate second moment $\boldsymbol{v}$ to $\boldsymbol{v}_F^{\odot 2} \oslash (2\boldsymbol{v}_F - \boldsymbol{v})$, where $\boldsymbol{v}_F$ is the full-sample second moment. R-AdaZO follows the same rule with its own second-moment map. Detailed derivations are given in Appendix A.3.*

C-SQE requires $2q$ function evaluations and three frozen transition evaluations per iteration. A streaming implementation can use $\mathcal{O}(d)$ auxiliary buffers for subset gradients and candidate states. The additional computation is $\mathcal{O}(d)$ for Adam, R-AdaZO, and RMSProp, while Muon requires two additional matrix orthogonalization evaluations and their temporary workspace. Odd query counts, higher-order extrapolation, and partition averaging are detailed in Appendix B.

4 THEORY

We establish joint bias and covariance guarantees for the executed C-SQE transition, along with a unified convergence bound for Adam, RMSProp, and regularized Muon. An additional result characterizes the interaction between the correction and R-AdaZO's moment averaging. Proofs and more detailed optimizer-specific bounds are deferred to the appendix A.

4.1 JOINT TRANSITION GUARANTEES

The executed transition $\mathcal{T}^{\mathrm{C}}_{t,q,r}$ contains both the update direction and the state retained for the next iteration. We begin with unrestricted states, for which $\mathrm{P}_t = \mathrm{E}$ and the extrapolation is executed directly. The following theorem characterizes its sampling bias and joint covariance, with $\boldsymbol{J}_r = \nabla_{\boldsymbol{g}} \mathcal{T}_t(\boldsymbol{g}_r)$.

Theorem 4.1 (Joint bias and covariance of C-SQE)**.** *Under Assumptions 2.1, 2.2, and 2.3 with $K \geq 4$, suppose the states are unrestricted and $\mathrm{P}_t = \mathrm{E}$. Then there exist constants C_b and C_V independent of $q \geq 2$ such that the executed transition in (9) satisfies*

$$\begin{aligned} &\text{Bias:} && \left\| \mathbb{E}_t[\mathcal{T}^{\mathrm{C}}_{t,q,r}] - \mathcal{T}_t(\boldsymbol{g}_r) \right\|_2 \leq C_b q^{-2}, \\ &\text{Covariance:} && \left\| \mathrm{Cov}_t(\mathcal{T}^{\mathrm{C}}_{t,q,r}) - q^{-1} \boldsymbol{J}_r \boldsymbol{\Sigma}_r \boldsymbol{J}_r^\top \right\|_{\mathrm{op}} \leq C_V q^{-2}. \end{aligned} \tag{11}$$

For odd q, the transition uses the weighted construction in Appendix B.1.

Compared to Theorem 3.1, C-SQE improves the sampling-bias order from q^{-1} to q^{-2} using the same function evaluations. The exact averaging identity preserves the linear fluctuation of the full-sample transition, yielding the same leading covariance for the update and states, including their cross-covariance. These guarantees are stated relative to $\mathcal{T}_t(\boldsymbol{g}_r)$; the finite-radius error relative to the exact-gradient transition remains $\mathcal{T}_t(\boldsymbol{g}_r) - \mathcal{T}_t(\boldsymbol{g}_t)$. The proof is given in Appendix A.2.

For constrained states, P_t preserves the update direction and processes the state before it is retained. The following corollary covers the rational second-moment rules for Adam, R-AdaZO, and RMSProp.

Corollary 4.1 (Optimizer-specific joint guarantees)**.** *Under Assumptions 2.1, 2.2, and 2.3 with $K \geq 4$, the bounds in (11) hold with possibly different constants for Adam, R-AdaZO, and RMSProp with rational state processing whenever $\mathcal{W}_t(\boldsymbol{g}_r) \succ \mathbf{0}$, where $\mathcal{W}_t$ denotes the corresponding second-moment map.*

Strictly positive $\mathcal{W}_t(\boldsymbol{g}_r)$ ensures that rational state processing preserves the joint bias and covariance orders, with constants that may depend on the smallest coordinate of $\mathcal{W}_t(\boldsymbol{g}_r)$. The proof is given in Appendix A.3. This positivity condition concerns the local accuracy of the retained second moments. The convergence analysis below uses only their nonnegativity and full-sample upper bound, so Adam and RMSProp may start from zero moments. Muon retains its unconstrained momentum exactly and requires no state processing.

4.2 CONVERGENCE GUARANTEES

We now analyze the actual iterates of Adam, RMSProp, and regularized Muon defined in Section 2.2. Adam and RMSProp use rational second-moment processing, while Muon uses $\mathrm{P}_t = \mathrm{E}$ and retains its affine momentum state. Each plug-in baseline and its C-SQE variant start from the same parameter with zero states. Directional estimates are conditionally i.i.d., and the balanced partition is chosen before the current queries. Write $\mathcal{G}_J := T^{-1} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla F(\boldsymbol{x}_t^J)\|_2^2$ for the average squared gradient norm along method J's trajectory, using the Frobenius norm for matrix parameters.

A common convergence bound. The following theorem separates the shared stochastic term from the residual squared update-bias term for all three optimizers. The constants and admissible step-size factors may depend on the dimension d through ZOGE moments, transition derivatives, and normalization bounds.

Theorem 4.2 (Convergence with Adam, RMSProp, and regularized Muon)**.** *Under Assumptions 2.1 and 2.2, let $q \geq 2$ be even, $T \geq q$, and $r \in (0, 1]$. Use fixed stabilizers $\epsilon_A, \epsilon_R, \epsilon_M > 0$ and fixed second-moment coefficients $\beta_2, \beta_R \in [0, 1)$ for Adam and RMSProp, respectively. Set $\eta = c\sqrt{q/T}$ for a sufficiently small constant $c > 0$, with $\beta_1 = 1 - \sqrt{q/T}$ for Adam and $\beta_M = 1 - \sqrt{q/T}$ for Muon; RMSProp has no first-moment state. For each optimizer, its plug-in and C-SQE iterates satisfy*

$$\mathcal{G}_{\mathrm{P}} \leq C_0\big[(qT)^{-1/2} + q^{-2} + r^4\big], \qquad \mathcal{G}_{\mathrm{C}} \leq C_0\big[(qT)^{-1/2} + q^{-4} + r^4\big], \tag{12}$$

where the common constant C_0 may depend on the fixed problem and optimizer parameters but is independent of q, T, r.

Both bounds share the leading stochastic term $(qT)^{-1/2}$ and the finite-radius contribution r^4. C-SQE improves the residual squared update-bias contribution in this common bound from q^{-2} to q^{-4} at the same function-query budget. The result covers constrained adaptive states in Adam and RMSProp as well as unconstrained momentum in Muon. The proof for all three optimizers is given in Appendix A.4.

Combining C-SQE with R-AdaZO. R-AdaZO reduces the noise entering the second moment through first-moment averaging. C-SQE further corrects the resulting nonlinear transition. The following corollary quantifies this additional bias reduction under a common variance bound.

Corollary 4.2 (Convergence with R-AdaZO)**.** *Under Assumptions 2.1 and 2.2, let $q \geq 2$ be even, $T \geq 1$, and $r \in (0, 1]$. Use zero initial moments, fixed $\beta_2 \in [0, 1)$ and $\epsilon_A > 0$, and rational state processing for C-SQE-R-AdaZO. Write $\gamma := 1 - \beta_1 \in (0, 1]$ and set $\eta = c\gamma$ for a sufficiently small constant $c > 0$. R-AdaZO and C-SQE-R-AdaZO satisfy*

$$\mathcal{G}_{\mathrm{R}} \leq C_1\big[(\gamma T)^{-1} + \gamma/q + \gamma^4/q^2 + r^4\big], \qquad \mathcal{G}_{\mathrm{CR}} \leq C_1\big[(\gamma T)^{-1} + \gamma/q + \gamma^6/q^4 + r^4\big], \tag{13}$$

where the common constant C_1 is independent of q, T, r, and γ.

The bounds retain the same optimization term $(\gamma T)^{-1}$, stochastic term γ/q, and finite-radius contribution r^4. C-SQE improves the residual squared update-bias contribution from γ^4/q^2 to γ^6/q^4, giving the same improvement in its q dependence as for Adam at fixed γ. Both methods use $2q$ function evaluations per iteration. The proof is given in Appendix A.5.

5 Experiments

We evaluate C-SQE through transition diagnostics, synthetic optimization, and LLM fine-tuning at matched function-query budgets. All C-SQE variants use a single partition and three frozen transition evaluations per iteration, without additional function queries. Implementation differences from the analyzed transitions, full protocols, and supplementary diagnostics are provided in Appendix C.

5.1 Synthetic Experiments

Theory validation. We evaluate the joint bias and covariance of ZO-AdaMM, ZO-RMSProp, R-AdaZO, regularized ZO-Muon, and their C-SQE variants, testing Theorems 3.1 and 4.1 and Corollary 4.1. For C-SQE, we measure the executed transition $\mathcal{T}^{\mathrm{C}}_{t,q,r} = [\boldsymbol{a}^{\mathrm{C}}_{t,q,r}; \boldsymbol{s}_{t+1}]$, including rational state processing for Adam, RMSProp, and R-AdaZO and identity processing for Muon. Each optimizer pair shares the same directional observations and frozen incoming states. We use $F(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$ with $d = 128$ and $\|\nabla F(\boldsymbol{x}_t)\|_2 = 1$, so central differences introduce no finite-radius bias. Each configuration uses $32{,}768$ independent repetitions, with $q \in \{64, 128, 256, 512\}$. The frozen states, optimizer coefficients, and 16×8 matrix parameterization for Muon are specified in Appendix C.1.

For either joint output $\boldsymbol{Y}_q$, we measure the bias $\|\mathbb{E}_t[\boldsymbol{Y}_q] - \mathcal{T}_t(\boldsymbol{g}_r)\|_2$ and the covariance remainder $\|\operatorname{Cov}_t(\boldsymbol{Y}_q) - \boldsymbol{V}_q\|_{\mathrm{F}}$, where $\boldsymbol{V}_q = q^{-1}\boldsymbol{J}_r\boldsymbol{\Sigma}_r\boldsymbol{J}_r^\top$ is computed separately for each optimizer.

Table 1: Joint transition accuracy at $d = 128$. Slopes are fitted over $q \in \{64, 128, 256, 512\}$; bias and approximate pointwise 95% Monte Carlo upper bounds on relative covariance error are reported at $q = 128$. Boldface marks the smaller bias within each pair.

		Bias		Covariance remainder	
Optimizer	Method	Slope	Bias ↓	Slope	Rel. error upper (%)
ZO-AdaMM	Plug-in	-1.000	8.802×10^{-4}	-2.000	0.201
ZO-AdaMM	C-SQE	-2.004	$\mathbf{7.673 \times 10^{-8}}$	-2.000	0.0257
ZO-RMSProp	Plug-in	-1.000	8.807×10^{-4}	-2.000	0.0445
ZO-RMSProp	C-SQE	-1.999	$\mathbf{5.580 \times 10^{-7}}$	-2.021	0.00130
R-AdaZO	Plug-in	-1.000	8.803×10^{-6}	-2.000	0.00197
R-AdaZO	C-SQE	-2.000	$\mathbf{5.742 \times 10^{-10}}$	-2.058	2.66×10^{-6}
ZO-Muon	Plug-in	-0.996	1.536×10^{-3}	-2.002	0.659
ZO-Muon	C-SQE	-2.009	$\mathbf{1.518 \times 10^{-5}}$	-2.083	0.0251

We fit log–log slopes over the four query counts and report the covariance remainder relative to $\|\boldsymbol{V}_q\|_{\mathrm{F}}$. Polynomial control variates with exact finite-sample moments and symmetry averaging reduce Monte Carlo error while retaining the complete nonlinear joint output in the estimates. The statistical procedure, confidence intervals, and scaling plots are provided in Appendix C.1.

Table 1 shows bias slopes near -1 for the plug-in transitions and -2 for their C-SQE counterparts across all four optimizer families. The covariance-remainder slopes range from -2.000 to -2.083, consistent with the predicted second-order remainder. At $q = 128$, C-SQE reduces the measured joint bias by roughly two to four orders of magnitude, while the relative covariance-error upper limits remain below 0.66% for all methods. All measured biases and covariance remainders are resolved above their Monte Carlo uncertainty. These results support leading-bias cancellation and preservation of the leading joint covariance for both constrained and unrestricted optimizer states.

Optimization performance. We compare ZO-AdaMM, R-AdaZO, ZO-RMSProp, ZO-Muon, and their C-SQE variants on Quadratic, Cubic, Lévy, and Rosenbrock objectives with $d = 1{,}000$. Each method uses $q = 10$ central-difference directions and runs for $2{,}000$ iterations, totaling $40{,}000$ function evaluations. We use 10 paired seeds with shared initializations and direction streams, and identical hyperparameters within each baseline–C-SQE pair. Detailed settings are provided in Appendix C.2.

Figure 1 shows the natural logarithm of the mean optimality gap across seeds. At the same query budget, all four C-SQE variants achieve faster objective decrease than their paired baselines. ZO-AdaMM and ZO-RMSProp follow nearly identical trajectories and converge most slowly, with substantial improvements from their C-SQE variants. R-AdaZO achieves faster decrease, and C-SQE provides a further improvement. ZO-Muon attains the lowest final gap among the baselines, while C-SQE-Muon achieves the lowest final gap on all four objectives.

5.2 LLM Fine-Tuning

We evaluate C-SQE on SST-2 (Socher et al., 2013) and COPA (Roemmele et al., 2011) by optimizing a 10-token soft prompt in a 64-dimensional intrinsic parameter space for a frozen OPT-125M model (Zhang et al., 2022). Each baseline and its C-SQE variant share the same hyperparameters and run for $20{,}000$ steps with batch size 16 and $q = 2$. This requires $80{,}000$ mini-batch loss evaluations for optimization per run, excluding logging and test evaluation. We report means and standard deviations over three seeds; detailed protocols, additional results, peak GPU memory measurements and wall-clock time are provided in Appendix C.3.

Table 2 shows that C-SQE lowers mean training loss across all four optimizer families on both tasks. On SST-2, C-SQE raises mean test accuracy by 4.060–13.410 percentage points; C-SQE-Muon achieves the lowest mean training loss, while C-SQE-RMSProp attains the highest mean test accuracy (78.820%). On COPA, C-SQE raises mean test accuracy by 0.670–6.670 percentage points for R-AdaZO, RMSProp, and Muon, and lowers it by 0.330 points for AdaMM. C-SQE-R-AdaZO

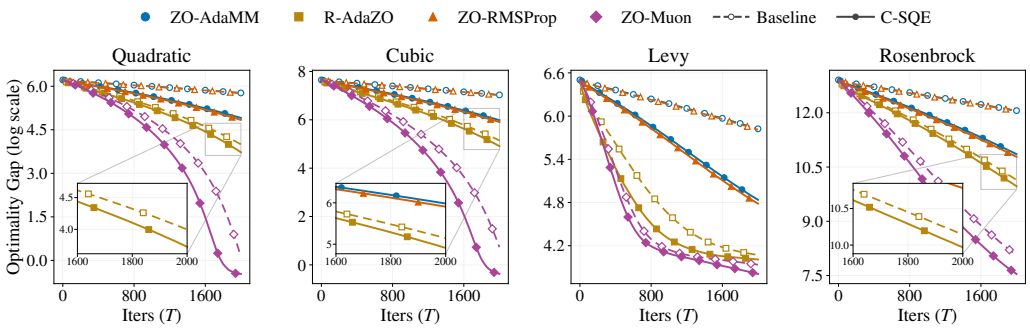

Figure 1: Synthetic optimization with $d = 1{,}000$ and $q = 10$ over 10 paired seeds. Each run uses 40,000 function evaluations.

achieves the lowest mean training loss on COPA, while C-SQE-RMSProp attains the highest mean test accuracy (64.670%).

Table 2: OPT-125M soft-prompt tuning after 20,000 steps with $q = 2$ (mean $\pm$ sample standard deviation over three seeds). Training loss averages the unperturbed mini-batch losses over the final 500 steps. Boldface marks the better mean within each baseline/C-SQE pair.

	SST-2		COPA	
Method	Train loss ↓	Test acc. (%) ↑	Train loss ↓	Test acc. (%) ↑
ZO-AdaMM	0.541 ± 0.061	72.550 ± 5.640	0.624 ± 0.031	$\mathbf{61.330 \pm 0.580}$
C-SQE-AdaMM	$\mathbf{0.466 \pm 0.024}$	$\mathbf{76.610 \pm 0.410}$	$\mathbf{0.608 \pm 0.025}$	61.000 ± 3.460
R-AdaZO	0.587 ± 0.109	65.140 ± 11.990	0.610 ± 0.017	59.330 ± 0.580
C-SQE-R-AdaZO	$\mathbf{0.479 \pm 0.049}$	$\mathbf{77.370 \pm 3.760}$	$\mathbf{0.601 \pm 0.014}$	$\mathbf{60.000 \pm 2.000}$
ZO-RMSProp	0.579 ± 0.148	65.410 ± 15.190	0.619 ± 0.019	58.000 ± 3.000
C-SQE-RMSProp	$\mathbf{0.447 \pm 0.018}$	$\mathbf{78.820 \pm 2.370}$	$\mathbf{0.615 \pm 0.019}$	$\mathbf{64.670 \pm 2.080}$
ZO-Muon	0.544 ± 0.086	67.780 ± 9.720	0.629 ± 0.016	60.330 ± 2.890
C-SQE-Muon	$\mathbf{0.440 \pm 0.018}$	$\mathbf{78.020 \pm 1.860}$	$\mathbf{0.608 \pm 0.008}$	$\mathbf{64.000 \pm 1.730}$

6 Conclusion

Our bias and covariance expansions characterize how noise from plug-in zeroth-order gradient estimators affects both update directions and persistent states in adaptive optimizers. To address the finite sampling bias, we proposed C-SQE, which extrapolates optimizer outputs from existing ZOGEs to reduce this bias from $\mathcal{O}(q^{-1})$ to $\mathcal{O}(q^{-2})$ while preserving the leading covariance term, with a rational correction extending these guarantees to constrained states. The convergence bounds quantify the improved q-dependence of sampling bias under C-SQE while retaining the leading stochastic terms. Synthetic experiments validate the predicted scaling, and OPT-125M soft-prompt tuning on SST-2 and COPA yields lower mean training loss across all four optimizer families on both tasks.

Limitations. C-SQE requires $q \geq 2$ and inherits the dimension dependence of the underlying ZOGE. The rational correction is currently designed only for Adam-style second-moment constraints, leaving extensions to other constrained optimizers for future work. Our LLM evaluation covers 64-dimensional soft-prompt tuning on OPT-125M; larger models and larger trainable subspaces remain to be explored.

AI USE STATEMENT

Generative AI assisted with experimental code generation and manuscript polishing. The authors manually reviewed the entire manuscript and take full responsibility for its content.

REPRODUCIBILITY STATEMENT

We provide the code to reproduce all experiments reported in this paper.

ETHICS STATEMENT

The reported experiments use synthetic objectives and involve no personal data or human participants. Applications to pretrained models should follow the licenses and usage terms of the models and evaluation data.

REFERENCES

Arya Akhavan, Evgenii Chzhen, Massimiliano Pontil, and Alexandre B Tsybakov. Gradient-free optimization of highly smooth functions: improved analysis and a new algorithm. *Journal of Machine Learning Research*, 25(370):1–50, 2024.

Francis Bach and Vianney Perchet. Highly-smooth zero-th order online optimization. In *Conference on Learning Theory*, pp. 257–283. PMLR, 2016.

Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, and David Cox. Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization. *Advances in neural information processing systems*, 32, 2019.

Sizhe Dang, Yanjun Zhao, Xiaodong Zheng, Guang Dai, Ivor Tsang, and Haishan Ye. Fzoo: Fast zeroth-order optimizer for fine-tuning large language models towards adam-scale speed. In *International Conference on Learning Representations*, volume 2026, pp. 113760–113792, 2026.

Hassan Dbouk, Nidham Gazagnadou, Matthias Reisser, and Christos Louizos. On adaptivity in zeroth-order optimization. *arXiv preprint arXiv:2605.03869*, 2026.

Geert Dhaene and Koen Jochmans. Split-panel jackknife estimation of fixed-effect models. *The Review of Economic Studies*, 82(3):991–1030, 2015.

Hongcheng Dong, Wenqiang Pu, Licheng Zhao, Rui Zhou, and Feng Yin. Romberg-extrapolated zeroth-order gradient estimator: Higher-order bias reduction with preserved leading directional variance. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=FiuJVpxuSX.

John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

Jian Feng and Zhihong Huang. Robust and efficient zeroth-order llm fine-tuning via adaptive bayesian subspace optimizer. *arXiv preprint arXiv:2601.01452*, 2026.

Yasong Feng and Tianyu Wang. Stochastic zeroth-order gradient and hessian estimators: variance reduction and refined bias bounds. *Information and Inference: A Journal of the IMA*, 12(3): 1514–1545, 2023.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

Bin Gu, Guodong Liu, Yanfu Zhang, Xiang Geng, and Heng Huang. Optimizing large-scale hyperparameters via automated learning algorithm. *arXiv preprint arXiv:2102.09026*, 2021.

G Hinton, S Nitish, and K Swersky. Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 2012.

Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. *arXiv preprint arXiv:1807.07978*, 2018.

Kaiyi Ji, Zhe Wang, Yi Zhou, and Yingbin Liang. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *International conference on machine learning*, pp. 3100–3109. PMLR, 2019.

Jiantao Jiao and Yanjun Han. Bias correction with jackknife, bootstrap, and taylor series. *IEEE Transactions on Information Theory*, 66(7):4392–4418, 2020. doi: 10.1109/TIT.2020.2969439.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. *URL https://kellerjordan. github. io/posts/muon*, 6(3):4, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yicheng Lang, Changsheng Wang, Yihua Zhang, Mingyi Hong, Zheng Zhang, Wotao Yin, and Sijia Liu. Powering up zeroth-order training via subspace gradient orthogonalization. *arXiv preprint arXiv:2602.17155*, 2026.

Chen Liang, Xiatao Sun, Qian Wang, and Daniel Rakita. Turning stale gradients into stable gradients: Coherent coordinate descent with implicit landscape smoothing for lightweight zeroth-order optimization. *arXiv preprint arXiv:2605.14373*, 2026.

Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, and Lisa Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. *Advances in neural information processing systems*, 31, 2018.

Sijia Liu, Pin-Yu Chen, Xiangyi Chen, and Mingyi Hong. signsgd via zeroth-order oracle. In *International conference on learning representations*, 2019.

Yang Liu, Chuan Zhou, Yuhan Lin, Shuai Zhang, Yang Gao, Zhao Li, and Shirui Pan. Sharpness-aware zeroth-order optimization for graph transformers. In *Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*. International Joint Conferences on Artificial Intelligence Organization, 2025.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*, 36:53038–53075, 2023.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Egor Petrov, Grigoriy Evseev, Aleksey Antonov, Andrey Veprikov, Nikolay Bushkov, Stanislav Moiseev, and Aleksandr Beznosikov. Leveraging coordinate momentum in signsgd and muon: Memory-optimized zero-order. *arXiv preprint arXiv:2506.04430*, 2025.

Maurice H Quenouille. Notes on bias in estimation. *Biometrika*, 43(3/4):353–360, 1956.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pp. 90–95, 2011.

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

Hyunseok Seung, Jaewoo Lee, and Hyunsuk Ko. Low-rank curvature for zeroth-order optimization in llm fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 25235–25242, 2026.

Yao Shu, Qixin Zhang, Kun He, and Zhongxiang Dai. Refining adaptive zeroth-order optimization at ease. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=NpIIbrg361.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Silan Zhang and Yujie Tang. Zivr: An incremental variance reduction technique for zeroth-order composite problems. *arXiv preprint arXiv:2601.05056*, 2026.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Yanjun Zhao, Sizhe Dang, Haishan Ye, Guang Dai, Yi Qian, and Ivor Tsang. Second-order fine-tuning without pain for llms: A hessian informed zeroth-order optimizer. In *International Conference on Learning Representations*, volume 2025, pp. 43496–43520, 2025.

A PROOFS

We prove the transition and convergence results stated in the main text. Unless stated otherwise, expectations and covariances condition on the pre-query history, including the selected partition.

We first establish two lemmas on smoothing properties and the moments of centered sample means.

Define $F_r(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{v}}[F(\boldsymbol{x} + r\boldsymbol{v})]$ for $\boldsymbol{v} \sim \mathrm{Unif}(\mathbb{B}^d)$, and let $\boldsymbol{g}_{t,r} := \mathbb{E}_t[\widehat{\boldsymbol{g}}_i(r)]$. Write $U_0 := dG$, $b_r := L_H dr^2/[2(d+2)]$, and $\Delta_r := F_r(\boldsymbol{x}_0) - \underline{F}$.

Lemma A.1 (Directional estimates and smoothing)**.** *Under Assumptions 2.1 and 2.2, $\|\widehat{\boldsymbol{g}}_i(r)\|_2 \le U_0$ almost surely and $\boldsymbol{g}_{t,r} = \nabla F_r(\boldsymbol{x}_t)$. The function F_r satisfies $F_r \ge \underline{F}$ and $\|\nabla F_r(\boldsymbol{x})\|_2 \le G$, inherits the gradient and Hessian Lipschitz constants L and L_H, and obeys $\sup_{\boldsymbol{x}} \|\nabla F_r(\boldsymbol{x}) - \nabla F(\boldsymbol{x})\|_2 \le b_r$. For $0 < r \le 1$, $\Delta_r \le F(\boldsymbol{x}_0) - \underline{F} + G$.*

Proof. Oracle Lipschitz continuity gives $\|\widehat{\boldsymbol{g}}_i(r)\|_2 \le d(2Gr)/(2r) = U_0$ and makes F G-Lipschitz. Sphere symmetry and the divergence theorem give

$$\mathbb{E}_t[\widehat{\boldsymbol{g}}_i(r)] = \frac{d}{r}\mathbb{E}_{\boldsymbol{z}}[F(\boldsymbol{x}_t + r\boldsymbol{z})\boldsymbol{z}] = \nabla F_r(\boldsymbol{x}_t).$$

Moreover,

$$\begin{aligned}\|\nabla F_r(\boldsymbol{x}) - \nabla F(\boldsymbol{x})\|_2 &= \big\|\mathbb{E}_{\boldsymbol{v}}[\nabla F(\boldsymbol{x} + r\boldsymbol{v}) - \nabla F(\boldsymbol{x}) - r\nabla^2 F(\boldsymbol{x})\boldsymbol{v}]\big\|_2 \\ &\le \frac{L_H r^2}{2}\mathbb{E}_{\boldsymbol{v}}\|\boldsymbol{v}\|_2^2 = \frac{L_H d r^2}{2(d+2)} = b_r.\end{aligned}$$

Here the equality uses $\mathbb{E}\boldsymbol{v} = \boldsymbol{0}$, and the inequality follows from Jensen's inequality and the Taylor remainder. Averaging preserves the lower bound, gradient bound, and derivative Lipschitz constants.

Finally, for $0 < r \le 1$, $\Delta_r \le F(\boldsymbol{x}_0) - \underline{F} + Gr\mathbb{E}\|\boldsymbol{v}\|_2 \le F(\boldsymbol{x}_0) - \underline{F} + G$. □

Fix the history and radius, so $\boldsymbol{g}_r = \boldsymbol{g}_{t,r}$. Write $\boldsymbol{Z}_i := \widehat{\boldsymbol{g}}_i(r) - \boldsymbol{g}_r$, $\boldsymbol{\delta}_n := n^{-1}\sum_{i=1}^n \boldsymbol{Z}_i$, $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}_r$, and $m_k := \mathbb{E}_t\|\boldsymbol{Z}_i\|_2^k$. Let $\nu_2 := U_0^2$, $\nu_k := (2U_0)^k$ for $k \ge 3$, $A_4 := m_4 + 3m_2^2$, and $A_6 := m_6 + 15m_4m_2 + 10m_3^2 + 15m_2^3$.

Lemma A.2 (Moments of centered sample means)**.** *Under the conditions of Lemma A.1, $m_k \le \nu_k$ for $k = 2, 3, \ldots$, and*

$$\mathbb{E}_t\boldsymbol{\delta}_n = \boldsymbol{0}, \quad \mathbb{E}_t[\boldsymbol{\delta}_n\boldsymbol{\delta}_n^\top] = \frac{\boldsymbol{\Sigma}}{n}, \quad \mathbb{E}_t[\boldsymbol{\delta}_n^{\otimes 3}] = \frac{\mathbb{E}_t[\boldsymbol{Z}_i^{\otimes 3}]}{n^2}, \quad \mathbb{E}_t\|\boldsymbol{\delta}_n\|_2^4 \le \frac{A_4}{n^2}, \quad \mathbb{E}_t\|\boldsymbol{\delta}_n\|_2^6 \le \frac{A_6}{n^3}. \tag{14}$$

More generally, $\mathbb{E}_t\|\boldsymbol{\delta}_n\|_2^{2k} = \mathcal{O}(n^{-k})$ for every fixed integer $k \ge 1$.

Proof. The directional bound gives

$$m_2 = \mathbb{E}_t\|\widehat{\boldsymbol{g}}_i(r)\|_2^2 - \|\boldsymbol{g}_r\|_2^2 \le U_0^2, \qquad m_k \le (2U_0)^k, \quad k \ge 3.$$

Expand the sample-mean moments and use conditional independence and centering. Every term containing a singleton sample index vanishes, yielding (14). The remaining fourth- and sixth-order index partitions give

$$\begin{aligned}&n^4\mathbb{E}_t\|\boldsymbol{\delta}_n\|_2^4 \le nm_4 + 3n(n-1)m_2^2 \le n^2A_4, \\ &n^6\mathbb{E}_t\|\boldsymbol{\delta}_n\|_2^6 \le nm_6 + n(n-1)(15m_4m_2 + 10m_3^2) + 15n(n-1)(n-2)m_2^3 \le n^3A_6.\end{aligned}$$

For order $2k$, at most k distinct indices survive, so boundedness gives $\mathbb{E}_t\|\boldsymbol{\delta}_n\|_2^{2k} = n^{-2k}\mathcal{O}(n^k) = \mathcal{O}(n^{-k})$. □

A.1 PROOF OF THEOREM 3.1

Fix the pre-query history and radius, and use the notation of Lemma A.2. Write $\boldsymbol{\theta} := \mathcal{T}_t(\boldsymbol{g}_r)$, $\boldsymbol{J} := \nabla_{\boldsymbol{g}}\mathcal{T}_t(\boldsymbol{g}_r)$, $\mathsf{H} := \nabla_{\boldsymbol{g}}^2\mathcal{T}_t(\boldsymbol{g}_r)$, and $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}_r$. The bias coefficient is $\boldsymbol{B} := \frac{1}{2}\mathbb{E}_t\mathsf{H}[\boldsymbol{Z}_i, \boldsymbol{Z}_i] = \boldsymbol{B}_r$, since $\mathbb{E}_t[\boldsymbol{Z}_i\boldsymbol{Z}_i^\top] = \boldsymbol{\Sigma}_r$. Let M_k denote the derivative bounds in Assumption 2.3. All Taylor segments lie in its convex domain, since they join the conditional mean and sample averages of admissible inputs.

Bias. A third-order Taylor expansion and the centered-moment identities in (14) give

$$\begin{aligned}\mathbb{E}_t[\mathcal{T}_t(\boldsymbol{g}_r+\boldsymbol{\delta}_n)] &= \boldsymbol{\theta}+\tfrac{1}{2}\mathbb{E}_t\mathsf{H}[\boldsymbol{\delta}_n,\boldsymbol{\delta}_n]+\tfrac{1}{6}\mathbb{E}_t\nabla^3_{\boldsymbol{g}}\mathcal{T}_t(\boldsymbol{g}_r)[\boldsymbol{\delta}_n,\boldsymbol{\delta}_n,\boldsymbol{\delta}_n]+\mathbb{E}_t\boldsymbol{R}_{4,n}\\ &= \boldsymbol{\theta}+\frac{\boldsymbol{B}}{n}+\frac{1}{6n^2}\mathbb{E}_t\nabla^3_{\boldsymbol{g}}\mathcal{T}_t(\boldsymbol{g}_r)[\boldsymbol{Z}_i,\boldsymbol{Z}_i,\boldsymbol{Z}_i]+\mathbb{E}_t\boldsymbol{R}_{4,n}\\ &= \boldsymbol{\theta}+\frac{\boldsymbol{B}}{n}+\boldsymbol{\rho}_n,\end{aligned} \tag{15}$$

where $\|\boldsymbol{R}_{4,n}\|_2 \le M_4\|\boldsymbol{\delta}_n\|_2^4/24$. Consequently,

$$\|\boldsymbol{\rho}_n\|_2 \le \frac{M_3m_3}{6n^2}+\frac{M_4}{24}\mathbb{E}_t\|\boldsymbol{\delta}_n\|_2^4 \le \frac{M_3m_3/6+M_4A_4/24}{n^2}=\frac{P_b}{n^2},$$

with

$$P_b := \frac{M_3m_3}{6}+\frac{M_4A_4}{24}, \qquad R_{\mathrm{b}} := P_b.$$

Setting $n=q$ in (15) and subtracting $\mathcal{T}_t(\boldsymbol{g}_t)$ separates the finite-radius term $\boldsymbol{\theta}-\mathcal{T}_t(\boldsymbol{g}_t)$ from $q^{-1}\boldsymbol{B}_r+\boldsymbol{\rho}_q$, proving the bias inequality in (7).

Covariance. Let $\boldsymbol{L}_n := \boldsymbol{J}\boldsymbol{\delta}_n$ and $\boldsymbol{Q}_n := \frac{1}{2}\mathsf{H}[\boldsymbol{\delta}_n,\boldsymbol{\delta}_n]$. The quadratic expansion gives

$$\mathcal{T}_t(\boldsymbol{g}_r+\boldsymbol{\delta}_n)-\boldsymbol{\theta}=\boldsymbol{L}_n+\boldsymbol{Q}_n+\boldsymbol{R}_n, \qquad \|\boldsymbol{R}_n\|_2 \le \frac{M_3}{6}\|\boldsymbol{\delta}_n\|_2^3. \tag{16}$$

Conditional independence and centering imply

$$\mathbb{E}_t[\boldsymbol{L}_n\boldsymbol{Q}_n^\top]=\frac{1}{2n^3}\sum_{i,j,k=1}^{n}\mathbb{E}_t\big[(\boldsymbol{J}\boldsymbol{Z}_i)\mathsf{H}[\boldsymbol{Z}_j,\boldsymbol{Z}_k]^\top\big]=\frac{1}{2n^2}\mathbb{E}_t\big[(\boldsymbol{J}\boldsymbol{Z}_1)\mathsf{H}[\boldsymbol{Z}_1,\boldsymbol{Z}_1]^\top\big],$$

because only $i=j=k$ survives. Hence,

$$\big\|\mathbb{E}_t[\boldsymbol{L}_n\boldsymbol{Q}_n^\top]\big\|_{\mathrm{op}} \le \frac{M_1M_2m_3}{2n^2}.$$

The moment bounds further give

$$\mathbb{E}_t\|\boldsymbol{L}_n\|_2^2 \le \frac{M_1^2m_2}{n}, \qquad \mathbb{E}_t\|\boldsymbol{Q}_n\|_2^2 \le \frac{M_2^2A_4}{4n^2}, \qquad \mathbb{E}_t\|\boldsymbol{R}_n\|_2^2 \le \frac{M_3^2A_6}{36n^3}.$$

Since $\mathbb{E}_t\boldsymbol{L}_n=\boldsymbol{0}$ and $\mathrm{Cov}_t(\boldsymbol{L}_n)=n^{-1}\boldsymbol{J}\boldsymbol{\Sigma}\boldsymbol{J}^\top$, covariance expansion and Cauchy–Schwarz yield

$$\begin{aligned}&\big\|\mathrm{Cov}_t(\mathcal{T}_t(\boldsymbol{g}_r+\boldsymbol{\delta}_n))-n^{-1}\boldsymbol{J}\boldsymbol{\Sigma}\boldsymbol{J}^\top\big\|_{\mathrm{op}}\\ &\le \big\|\mathbb{E}_t[\boldsymbol{L}_n\boldsymbol{Q}_n^\top]\big\|_{\mathrm{op}}+2\sqrt{\mathbb{E}_t\|\boldsymbol{L}_n\|_2^2\,\mathbb{E}_t\|\boldsymbol{R}_n\|_2^2}+2\mathbb{E}_t\|\boldsymbol{Q}_n\|_2^2+2\mathbb{E}_t\|\boldsymbol{R}_n\|_2^2\\ &\le \frac{M_1M_2m_3}{n^2}+\frac{M_1M_3\sqrt{m_2A_6}}{3n^2}+\frac{M_2^2A_4}{2n^2}+\frac{M_3^2A_6}{18n^3}\\ &\le \frac{P_V}{n^2},\end{aligned}$$

where we use $n^{-3}\le n^{-2}$ and choose

$$P_V := M_1M_2m_3+\frac{M_1M_3}{3}\sqrt{m_2A_6}+\frac{M_2^2A_4}{2}+\frac{M_3^2A_6}{18}.$$

Here the quadratic and remainder terms are controlled by $\|\mathrm{Cov}_t(\boldsymbol{Q}_n+\boldsymbol{R}_n)\|_{\mathrm{op}} \le \mathbb{E}_t\|\boldsymbol{Q}_n+\boldsymbol{R}_n\|_2^2 \le 2\mathbb{E}_t\|\boldsymbol{Q}_n\|_2^2+2\mathbb{E}_t\|\boldsymbol{R}_n\|_2^2$. Setting $n=q$ proves the covariance inequality in (7). Both P_b and P_V depend only on the fixed conditional moments and derivative bounds and are independent of $q\ge 2$.

A.2 Proof of Theorem 4.1

Let $q_1 = \lfloor q/2 \rfloor$, $q_2 = \lceil q/2 \rceil$, and $\omega_j = q_j/q$. For the selected partition, write $\boldsymbol{\delta}_j := \frac{1}{q_j}\sum_{i\in I_j} \boldsymbol{Z}_i$ and $\boldsymbol{\delta} := \frac{1}{q}\sum_{i=1}^{q} \boldsymbol{Z}_i = \sum_{j=1}^{2} \omega_j \boldsymbol{\delta}_j$. For both parities of q, the unprocessed extrapolation is $\boldsymbol{Z}^{\mathrm{C}} := 2\mathcal{T}_t(\boldsymbol{g}_r + \boldsymbol{\delta}) - \sum_{j=1}^{2} \omega_j \mathcal{T}_t(\boldsymbol{g}_r + \boldsymbol{\delta}_j)$. Under identity processing, $\boldsymbol{Z}^{\mathrm{C}} = \mathcal{T}^{\mathrm{C}}_{t,q,r}$.

Using the full- and subsample expansions in (15),

$$\begin{aligned}\|\mathbb{E}_t \boldsymbol{Z}^{\mathrm{C}} - \boldsymbol{\theta}\|_2 &= \left\| \left(\frac{2}{q} - \sum_{j=1}^{2} \frac{\omega_j}{q_j} \right) \boldsymbol{B} + 2\boldsymbol{\rho}_q - \sum_{j=1}^{2} \omega_j \boldsymbol{\rho}_{q_j} \right\|_2 \overset{(a)}{=} \left\| 2\boldsymbol{\rho}_q - \sum_{j=1}^{2} \omega_j \boldsymbol{\rho}_{q_j} \right\|_2 \\ &\le R_{\mathrm{b}} \left(\frac{2}{q^2} + \sum_{j=1}^{2} \frac{\omega_j}{q_j^2} \right) \overset{(b)}{\le} \frac{8R_{\mathrm{b}}}{q^2},\end{aligned}$$

where (a) uses $\omega_j/q_j = 1/q$ and (b) uses $q_j \ge q/3$. This proves the bias bound.

For the covariance, (16) and the averaging identity give

$$\boldsymbol{Z}^{\mathrm{C}} - \boldsymbol{\theta} = \boldsymbol{J} \left(2\boldsymbol{\delta} - \sum_j \omega_j \boldsymbol{\delta}_j \right) + \boldsymbol{Q}^{\mathrm{C}} + \boldsymbol{R}^{\mathrm{C}} = \boldsymbol{J}\boldsymbol{\delta} + \boldsymbol{Q}^{\mathrm{C}} + \boldsymbol{R}^{\mathrm{C}},$$

where $\boldsymbol{Q}^{\mathrm{C}} := \mathsf{H}[\boldsymbol{\delta}, \boldsymbol{\delta}] - \frac{1}{2}\sum_j \omega_j \mathsf{H}[\boldsymbol{\delta}_j, \boldsymbol{\delta}_j]$. Let $\boldsymbol{L} := \boldsymbol{J}\boldsymbol{\delta}$ and $\boldsymbol{K}_3 := \mathbb{E}_t[(\boldsymbol{J}\boldsymbol{Z}_i)\mathsf{H}[\boldsymbol{Z}_i, \boldsymbol{Z}_i]^\top]$. Independence of a subset and its complement yields

$$\mathbb{E}_t[(\boldsymbol{J}\boldsymbol{\delta})\mathsf{H}[\boldsymbol{\delta}_j, \boldsymbol{\delta}_j]^\top] = \omega_j \mathbb{E}_t[(\boldsymbol{J}\boldsymbol{\delta}_j)\mathsf{H}[\boldsymbol{\delta}_j, \boldsymbol{\delta}_j]^\top] = \frac{\boldsymbol{K}_3}{q q_j}.$$

Consequently,

$$\mathbb{E}_t[\boldsymbol{L}(\boldsymbol{Q}^{\mathrm{C}})^\top] = \frac{\boldsymbol{K}_3}{q^2} - \frac{1}{2}\sum_{j=1}^{2} \frac{\omega_j \boldsymbol{K}_3}{q q_j} = \frac{\boldsymbol{K}_3}{q^2} - \frac{\boldsymbol{K}_3}{q^2} = \boldsymbol{0}.$$

The second equality again uses $\omega_j/q_j = 1/q$. Weighted Cauchy–Schwarz and (14) imply

$$\mathbb{E}_t\|\boldsymbol{L}\|_2^2 = \mathcal{O}(q^{-1}), \qquad \mathbb{E}_t\|\boldsymbol{Q}^{\mathrm{C}}\|_2^2 = \mathcal{O}(q^{-2}), \qquad \mathbb{E}_t\|\boldsymbol{R}^{\mathrm{C}}\|_2^2 = \mathcal{O}(q^{-3}).$$

Expanding the covariance and using the exact cancellation above gives

$$\left\|\mathrm{Cov}_t(\boldsymbol{Z}^{\mathrm{C}}) - q^{-1}\boldsymbol{J}\boldsymbol{\Sigma}\boldsymbol{J}^\top\right\|_{\mathrm{op}} \le 2\sqrt{\mathbb{E}_t\|\boldsymbol{L}\|_2^2\, \mathbb{E}_t\|\boldsymbol{R}^{\mathrm{C}}\|_2^2} + 2\mathbb{E}_t\|\boldsymbol{Q}^{\mathrm{C}}\|_2^2 + 2\mathbb{E}_t\|\boldsymbol{R}^{\mathrm{C}}\|_2^2 = \mathcal{O}(q^{-2}). \quad (17)$$

This controls the full joint covariance, including the update–state blocks. The joint law is the same for every fixed partition of these sizes, so averaging over pre-query partition randomness preserves the result.

A.3 Proof of Corollary 4.1

For the subset means above, define

$$\boldsymbol{V}_{t,q,r} := \omega_1\omega_2(\widehat{\boldsymbol{g}}_1 - \widehat{\boldsymbol{g}}_2)^{\odot 2}, \qquad \mathbb{E}_t \boldsymbol{V}_{t,q,r} = q^{-1}\,\mathrm{diag}(\boldsymbol{\Sigma}_r).$$

The expectation follows from independence and $\omega_1\omega_2(q_1^{-1} + q_2^{-1}) = q^{-1}$. The constrained second-moment maps share the form $\mathcal{W}_t(\boldsymbol{g}) = \beta\boldsymbol{v}_t + \kappa(\boldsymbol{h} + \lambda\boldsymbol{g})^{\odot 2}$. For Adam and RMSProp, $\boldsymbol{h} = \boldsymbol{0}$ and $\lambda = 1$, with $(\beta, \kappa) = (\beta_2, 1-\beta_2)$ and $(\beta_R, 1-\beta_R)$, respectively. For R-AdaZO, $(\beta, \kappa) = (\beta_2, 1-\beta_2)$, $\boldsymbol{h} = \beta_1\boldsymbol{m}_t$, and $\lambda = 1-\beta_1$. For RMSProp, $\boldsymbol{v}_t$ denotes $\boldsymbol{w}_t$. Write $\boldsymbol{v}_F := \mathcal{W}_t(\widehat{\boldsymbol{g}}_{q,r})$, $\boldsymbol{C} := \kappa\lambda^2\boldsymbol{V}_{t,q,r}$, and $\boldsymbol{D} := \mathrm{P}_t(\boldsymbol{Z}^{\mathrm{C}}) - \boldsymbol{Z}^{\mathrm{C}}$.

The quadratic variance identity gives $\widetilde{\boldsymbol{v}}_{t+1} = \boldsymbol{v}_F - \boldsymbol{C}$. The rational rule therefore satisfies

$$\boldsymbol{v}_{t+1} = \boldsymbol{v}_F^{\odot 2} \oslash (\boldsymbol{v}_F + \boldsymbol{C}) = \widetilde{\boldsymbol{v}}_{t+1} + \boldsymbol{C}^{\odot 2} \oslash (\boldsymbol{v}_F + \boldsymbol{C}), \quad (18)$$

where $0/0 = 0$, implemented using masked division with zero-initialized outputs. Thus $\boldsymbol{D}$ has only one nonzero block, lying coordinatewise between $\boldsymbol{0}$ and $\boldsymbol{C}$.

Let $\nu := \min_i [\mathcal{W}_t(\boldsymbol{g}_r)]_i > 0$ and $\mathcal{A} := \{\min_i v_{F,i} \geq \nu/2\}$. Choose a Lipschitz bound $L_W > 0$ for $\mathcal{W}_t$ on the bounded input domain and a uniform bound $C_* \geq \|\boldsymbol{C}\|_2$. The eighth-moment sample-mean bound gives

$$\Pr_t(\mathcal{A}^c) \leq \Pr_t\left(\|\widehat{\boldsymbol{g}}_{q,r} - \boldsymbol{g}_r\|_2 \geq \frac{\nu}{2L_W} \right) \leq \left(\frac{2L_W}{\nu} \right)^8 \mathbb{E}_t \|\widehat{\boldsymbol{g}}_{q,r} - \boldsymbol{g}_r\|_2^8 = \mathcal{O}(q^{-4}).$$

The first inequality uses Lipschitz continuity, and the second is Markov's inequality. Also, $\mathbb{E}_t\|\boldsymbol{C}\|_2^4 = \mathcal{O}(q^{-4})$, since $\boldsymbol{C}$ is quadratic in the difference of the two sample means. Splitting over $\mathcal{A}$ and its complement now yields

$$\mathbb{E}_t\|\boldsymbol{D}\|_2^2 = \mathbb{E}_t[\|\boldsymbol{D}\|_2^2 \mathbf{1}_{\mathcal{A}}] + \mathbb{E}_t[\|\boldsymbol{D}\|_2^2 \mathbf{1}_{\mathcal{A}^c}] \leq \frac{4}{\nu^2}\mathbb{E}_t\|\boldsymbol{C}\|_2^4 + C_*^2 \Pr_t(\mathcal{A}^c) \leq C_{\mathrm{s}}^2 q^{-4}.$$

Here the denominator in (18) is at least $\nu/2$ on $\mathcal{A}$, while $\|\boldsymbol{D}\|_2 \leq \|\boldsymbol{C}\|_2$ everywhere. The constants may depend on ν but are independent of q.

The executed mean obeys

$$\|\mathbb{E}_t[\boldsymbol{Z}^{\mathrm{C}} + \boldsymbol{D}] - \boldsymbol{\theta}\|_2 \leq \|\mathbb{E}_t \boldsymbol{Z}^{\mathrm{C}} - \boldsymbol{\theta}\|_2 + (\mathbb{E}_t\|\boldsymbol{D}\|_2^2)^{1/2} = \mathcal{O}(q^{-2}).$$

For its covariance,

$$\begin{aligned} \|\operatorname{Cov}_t(\boldsymbol{Z}^{\mathrm{C}} + \boldsymbol{D}) - \operatorname{Cov}_t(\boldsymbol{Z}^{\mathrm{C}})\|_{\mathrm{op}} &\leq 2\sqrt{\mathbb{E}_t\|\boldsymbol{Z}^{\mathrm{C}} - \mathbb{E}_t\boldsymbol{Z}^{\mathrm{C}}\|_2^2\, \mathbb{E}_t\|\boldsymbol{D}\|_2^2} + \mathbb{E}_t\|\boldsymbol{D}\|_2^2 \\ &= \mathcal{O}(q^{-5/2}) + \mathcal{O}(q^{-4}). \end{aligned}$$

This follows from covariance expansion, Cauchy–Schwarz, and the $\mathcal{O}(q^{-1})$ centered second moment of $\boldsymbol{Z}^{\mathrm{C}}$. Combining with (17) proves the executed joint bounds. Regularized Muon has $\boldsymbol{D} = \mathbf{0}$, so its bounds follow directly.

A.4 Proof of Theorem 4.2

We first establish the state and derivative bounds used by all three optimizers. Throughout, $\gamma_M := 1 - \beta_M$. Write $\widehat{\boldsymbol{g}}_t := \widehat{\boldsymbol{g}}_{q,r}$ for the averaged directional estimate at iteration t and $\widehat{\boldsymbol{G}}_t$ for its matrix form.

Uniform regularity and state bounds.

Proposition A.1 (Regularity of the stabilized transitions)**.** *Under Assumption 2.2, the unprocessed transitions in (4), (5), and (6), together with the R-AdaZO modification described after (4), satisfy Assumption 2.3 for every finite K, provided the incoming second moments are nonnegative. With zero initial states, their derivative bounds on the gradient ball of radius U_0 are uniform over iterations and incoming histories.*

Proof. The moment recursions, affine preservation, and rational state domination give

$$\begin{aligned} \|\boldsymbol{m}_t\|_2 &\leq (1-\beta_1)\sum_{s<t} \beta_1^{t-1-s}\|\widehat{\boldsymbol{g}}_s\|_2 \leq U_0(1-\beta_1^t), \\ \mathbf{0} \preceq \boldsymbol{v}_t &\preceq (1-\beta_2)\sum_{s<t} \beta_2^{t-1-s} U_0^2 \mathbf{1} = U_0^2(1-\beta_2^t)\mathbf{1}, \\ \|\boldsymbol{M}_t\|_{\mathrm{F}} &\leq \gamma_M \sum_{s<t} \beta_M^{t-1-s}\|\widehat{\boldsymbol{G}}_s\|_{\mathrm{F}} \leq U_0. \end{aligned}$$

The second line uses $\|\boldsymbol{m}_{s+1}\|_2 \leq U_0$ for R-AdaZO and holds with $(\boldsymbol{v}_t, \beta_2)$ replaced by $(\boldsymbol{w}_t, \beta_R)$ for RMSProp. All reachable states therefore lie in fixed compact sets. The scalar normalizations are smooth because their squared denominators are bounded below by the positive stabilizers. For Muon's regularized Gram matrix $\boldsymbol{A} \succeq \epsilon_M^2 \boldsymbol{I}$, the resolvent representation gives

$$\boldsymbol{A}^{-1/2} = \frac{1}{\pi}\int_0^\infty s^{-1/2}(\boldsymbol{A} + s\boldsymbol{I})^{-1}\, ds, \quad \|\nabla_{\boldsymbol{A}}^k(\boldsymbol{A}^{-1/2})\|_{\mathrm{op}} \leq \frac{k!}{\pi}\int_0^\infty \frac{s^{-1/2}}{(\epsilon_M^2 + s)^{k+1}}\, ds < \infty.$$

The derivative bound follows by differentiating the resolvent. Composition with the affine and quadratic moment maps now gives uniform derivatives of every finite order on a slightly enlarged gradient ball, as required. □

Proof of Theorem 4.2. Let $J \in \{\mathrm{P}, \mathrm{C}\}$ index the two methods along their respective trajectories, and set $s_{\mathrm{P}} := 1$ and $s_{\mathrm{C}} := 2$. Write $\boldsymbol{a}_t^J$ for the executed direction, so $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \boldsymbol{a}_t^J$. Use $\boldsymbol{g}_{t,r} = \nabla F_r(\boldsymbol{x}_t)$ and let $\sigma^2 := U_0^2$ bound the variance of one directional estimate. Constants below are uniform in q, T, r, and the first-moment parameter $\gamma \in (0, 1]$. Until the step-size factor c is fixed, constants denoted by C are also independent of $c \in (0, 1]$.

Adam. Write $\gamma := 1 - \beta_1$, $\gamma_2 := 1 - \beta_2$, and $a := \epsilon_A$. Define the predictable metric, reference direction, and conditional bias by

$$\boldsymbol{D}_t := \operatorname{diag}\Big((\beta_2 \boldsymbol{v}_t + \gamma_2 \boldsymbol{g}_{t,r}^{\odot 2} + a^2 \mathbf{1})^{-1/2}\Big),\ \boldsymbol{a}_t^0 := \boldsymbol{D}_t(\beta_1 \boldsymbol{m}_t + \gamma \boldsymbol{g}_{t,r}),\ \overline{\boldsymbol{a}}_t := \mathbb{E}_t \boldsymbol{a}_t^J,\ \boldsymbol{b}_t := \overline{\boldsymbol{a}}_t - \boldsymbol{a}_t^0.$$

For $\boldsymbol{e}_t := \boldsymbol{m}_t - \boldsymbol{g}_{t,r}$, write

$$W_J := \frac{1}{T}\sum_{t<T} \mathbb{E}\langle \boldsymbol{g}_{t,r}, \boldsymbol{D}_t \boldsymbol{g}_{t,r}\rangle, \quad E_J := \frac{1}{T}\sum_{t<T} \mathbb{E}\|\boldsymbol{e}_t\|_2^2,$$
$$M_J := \frac{1}{T}\sum_{t<T} \mathbb{E}\|\overline{\boldsymbol{a}}_t\|_2^2, \qquad N_J := \frac{1}{T}\sum_{t<T} \mathbb{E}\|\boldsymbol{a}_t^J\|_2^2.$$

Both methods retain the affine first moment and satisfy the second-moment domination in Proposition A.1. Together with the absolute extrapolation weights, this gives

$$\|\boldsymbol{m}_t\|_2 \le U_0, \qquad \mathbf{0} \preceq \boldsymbol{v}_t \preceq U_0^2 \mathbf{1}, \qquad \|\boldsymbol{a}_t^J\|_2 \le A_* := 3U_0/a.$$

Hence $(U_0^2 + a^2)^{-1/2} \boldsymbol{I} \preceq \boldsymbol{D}_t \preceq a^{-1} \boldsymbol{I}$, and the uniform transition expansions yield

$$\|\boldsymbol{D}_t^{-1/2} \boldsymbol{b}_t\|_2 \le B_J := b_J q^{-s_J}, \quad \mathbb{E}_t\|\boldsymbol{a}_t^J - \overline{\boldsymbol{a}}_t\|_2^2 \le V_J := \frac{v_1}{q} + \mathbf{1}_{\{J=\mathrm{C}\}} \frac{v_2}{q^2}. \tag{19}$$

The bias orders follow from Theorems 3.1 and 4.1 applied to the update block, which state processing leaves unchanged. The variance bound uses Lipschitz continuity for Plug-in and the preserved linear term for C-SQE; $\boldsymbol{D}_t$ remains fixed under differentiation.

Since $\overline{\boldsymbol{a}}_t = \boldsymbol{D}_t(\boldsymbol{g}_{t,r} + \beta_1 \boldsymbol{e}_t) + \boldsymbol{b}_t$, the conditional second moments satisfy

$$\|\overline{\boldsymbol{a}}_t\|_2^2 \le 3\big(\|\boldsymbol{D}_t \boldsymbol{g}_{t,r}\|_2^2 + \beta_1^2 \|\boldsymbol{D}_t \boldsymbol{e}_t\|_2^2 + \|\boldsymbol{b}_t\|_2^2\big) \le C\big(\langle \boldsymbol{g}_{t,r}, \boldsymbol{D}_t \boldsymbol{g}_{t,r}\rangle + \|\boldsymbol{e}_t\|_2^2 + B_J^2\big),$$
$$\mathbb{E}_t\|\boldsymbol{a}_t^J\|_2^2 = \|\overline{\boldsymbol{a}}_t\|_2^2 + \mathbb{E}_t\|\boldsymbol{a}_t^J - \overline{\boldsymbol{a}}_t\|_2^2 \le \|\overline{\boldsymbol{a}}_t\|_2^2 + V_J.$$

Averaging over time and histories gives

$$M_J \le C(W_J + E_J + B_J^2), \qquad N_J \le C(W_J + E_J + B_J^2) + V_J. \tag{20}$$

To control momentum tracking, decompose $\boldsymbol{g}_{t+1,r} - \boldsymbol{g}_{t,r} = \boldsymbol{h}_t + \boldsymbol{z}_t$, where $\boldsymbol{h}_t := \mathbb{E}_t(\boldsymbol{g}_{t+1,r} - \boldsymbol{g}_{t,r})$ and $\mathbb{E}_t \boldsymbol{z}_t = \mathbf{0}$. With $\boldsymbol{\xi}_t := \widehat{\boldsymbol{g}}_t - \boldsymbol{g}_{t,r}$, conditional centering gives

$$\begin{aligned}
\mathbb{E}_t\|\boldsymbol{e}_{t+1}\|_2^2 &= \|(1-\gamma)\boldsymbol{e}_t - \boldsymbol{h}_t\|_2^2 + \mathbb{E}_t\|\gamma \boldsymbol{\xi}_t - \boldsymbol{z}_t\|_2^2 \\
&\le (1-\gamma)\|\boldsymbol{e}_t\|_2^2 + \gamma^{-1}\|\boldsymbol{h}_t\|_2^2 + 2\gamma^2 \mathbb{E}_t\|\boldsymbol{\xi}_t\|_2^2 + 2\mathbb{E}_t\|\boldsymbol{z}_t\|_2^2 \\
&\le (1-\gamma)\|\boldsymbol{e}_t\|_2^2 + \gamma^{-1}\|\boldsymbol{h}_t\|_2^2 + 2\gamma^2\sigma^2/q + 2\mathbb{E}_t\|\boldsymbol{z}_t\|_2^2.
\end{aligned}$$

The second line uses weighted Young's inequality for the predictable terms and $\|\boldsymbol{u}-\boldsymbol{v}\|_2^2 \le 2\|\boldsymbol{u}\|_2^2 + 2\|\boldsymbol{v}\|_2^2$ for the centered terms. Taylor expansion and the bound $\|\boldsymbol{a}_t^J\|_2 \le A_*$ further give

$$\|\boldsymbol{h}_t\|_2^2 \le 2L^2\eta^2\|\overline{\boldsymbol{a}}_t\|_2^2 + \tfrac{1}{2}L_H^2\eta^4\big(\mathbb{E}_t\|\boldsymbol{a}_t^J\|_2^2\big)^2 \le 2L^2\eta^2\|\overline{\boldsymbol{a}}_t\|_2^2 + \tfrac{1}{2}L_H^2\eta^4 A_*^2 \mathbb{E}_t\|\boldsymbol{a}_t^J\|_2^2,$$
$$\mathbb{E}_t\|\boldsymbol{z}_t\|_2^2 \le \tfrac{1}{2}L^2\eta^2 \mathbb{E}_t\|\boldsymbol{a}_t^J - (\boldsymbol{a}_t^J)'\|_2^2 = L^2\eta^2 \mathbb{E}_t\|\boldsymbol{a}_t^J - \overline{\boldsymbol{a}}_t^J\|_2^2 \le L^2\eta^2 V_J.$$

Here $(\boldsymbol{a}_t^J)'$ is a conditionally independent copy, so the last line follows from the variance identity and the L-Lipschitz gradient. Summing the tracking recursion, using $\boldsymbol{e}_0 = -\boldsymbol{g}_{0,r}$, and setting $\eta = c\gamma$ yields

$$\begin{aligned}
E_J &\le \frac{\|\boldsymbol{g}_{0,r}\|_2^2}{\gamma T} + \frac{2\gamma\sigma^2}{q} + \frac{2L^2\eta^2}{\gamma^2} M_J + \frac{L_H^2\eta^4 A_*^2}{2\gamma^2} N_J + \frac{2L^2\eta^2}{\gamma} V_J \\
&= \frac{\|\boldsymbol{g}_{0,r}\|_2^2}{\gamma T} + \frac{2\gamma\sigma^2}{q} + 2L^2c^2 M_J + \tfrac{1}{2}L_H^2 A_*^2 c^4 \gamma^2 N_J + 2L^2c^2\gamma V_J.
\end{aligned} \tag{21}$$

For descent, let $w_t := \langle \boldsymbol{g}_{t,r}, \boldsymbol{D}_t \boldsymbol{g}_{t,r} \rangle$. Weighted Young's inequality gives

$$\begin{aligned}\langle \boldsymbol{g}_{t,r}, \overline{\boldsymbol{a}}_t \rangle &= w_t + \beta_1 \langle \boldsymbol{g}_{t,r}, \boldsymbol{D}_t \boldsymbol{e}_t \rangle + \langle \boldsymbol{g}_{t,r}, \boldsymbol{b}_t \rangle \\ &\geq w_t - \tfrac{1}{4} w_t - \beta_1^2 \|\boldsymbol{D}_t^{1/2} \boldsymbol{e}_t\|_2^2 - \tfrac{1}{4} w_t - B_J^2 \\ &\geq \tfrac{1}{2} w_t - a^{-1} \|\boldsymbol{e}_t\|_2^2 - B_J^2.\end{aligned}$$

Consequently, smoothness and telescoping imply

$$\begin{aligned}\tfrac{1}{2} W_J &\leq \frac{1}{\eta T} \sum_{t<T} \mathbb{E}\big[F_r(\boldsymbol{x}_t) - F_r(\boldsymbol{x}_{t+1})\big] + a^{-1} E_J + B_J^2 + \tfrac{1}{2} L \eta N_J \\ &\leq \frac{\Delta_*}{c\gamma T} + a^{-1} E_J + B_J^2 + \tfrac{1}{2} L c \gamma N_J,\end{aligned} \tag{22}$$

where $\Delta_* := F(\boldsymbol{x}_0) - \underline{F} + G$ bounds the initial smoothed gap for $r \leq 1$.

Substitute (20) into (21) and take $c \in (0, 1]$ small enough to absorb the coefficient of E_J:

$$\begin{aligned}E_J &\leq C \left(\frac{1}{\gamma T} + \frac{\gamma}{q} \right) + C c^2 (W_J + E_J + B_J^2 + \gamma V_J) \\ &\leq \tfrac{1}{2} E_J + C \left(\frac{1}{\gamma T} + \frac{\gamma}{q} + c^2 W_J + c^2 B_J^2 + c^2 \gamma V_J \right).\end{aligned}$$

Using the resulting bound for E_J in (22) gives

$$\begin{aligned}W_J &\leq C \left(\frac{1}{c\gamma T} + \frac{\gamma}{q} + B_J^2 + \gamma V_J \right) + C(c + c^2) W_J \\ &\leq \tfrac{1}{2} W_J + C \left(\frac{1}{c\gamma T} + \frac{\gamma}{q} + B_J^2 + \gamma V_J \right), \\ W_J &\leq C \left((\gamma T)^{-1} + \gamma/q + B_J^2 + \gamma V_J \right).\end{aligned} \tag{23}$$

The second line follows by decreasing c once more; after fixing c, its reciprocal is absorbed into C. All absorption choices are uniform in γ, q, T, r.

For $T \geq q$, substitute $\gamma = \sqrt{q/T}$ and (19) to obtain

$$\begin{aligned}W_J &\leq C \left(\frac{1}{\sqrt{qT}} + q^{-2s_J} + \mathbf{1}_{\{J=\mathrm{C}\}} \frac{1}{q\sqrt{qT}} \right) \leq C_0 (qT)^{-1/2} + C_1 q^{-2s_J}, \\ \mathcal{G}_J &\leq \frac{2}{T} \sum_{t<T} \mathbb{E}\|\boldsymbol{g}_{t,r}\|_2^2 + 2b_r^2 \leq 2\sqrt{U_0^2 + a^2}\, W_J + 2b_r^2.\end{aligned}$$

The last line uses the metric lower bound and Lemma A.1, so $b_r^2 = \mathcal{O}(r^4)$ gives the Adam bound in (12).

RMSProp. Set $D_R := \sqrt{U_0^2 + \epsilon_R^2}$, $\boldsymbol{D}_t := \operatorname{diag}\big((\mathcal{W}_t^{\mathrm{RMS}}(\boldsymbol{g}_{t,r}) + \epsilon_R^2 \mathbf{1})^{-1/2}\big)$, and $\overline{\boldsymbol{a}}_t := \mathbb{E}_t \boldsymbol{a}_t^J$. Proposition A.1 gives $D_R^{-1} \boldsymbol{I} \preceq \boldsymbol{D}_t \preceq \epsilon_R^{-1} \boldsymbol{I}$ and the local bounds (19) with reference direction $\boldsymbol{D}_t \boldsymbol{g}_{t,r}$. For $w_t := \langle \boldsymbol{g}_{t,r}, \boldsymbol{D}_t \boldsymbol{g}_{t,r} \rangle$, weighted Young's inequality gives

$$\begin{aligned}\langle \boldsymbol{g}_{t,r}, \overline{\boldsymbol{a}}_t \rangle &\geq w_t - \sqrt{w_t} B_J \geq \tfrac{1}{2} w_t - \tfrac{1}{2} B_J^2, \\ \mathbb{E}_t \|\boldsymbol{a}_t^J\|_2^2 &\leq 2\|\boldsymbol{D}_t \boldsymbol{g}_{t,r}\|_2^2 + 2\|\overline{\boldsymbol{a}}_t - \boldsymbol{D}_t \boldsymbol{g}_{t,r}\|_2^2 + V_J \leq 2\epsilon_R^{-1}(w_t + B_J^2) + V_J, \\ \mathbb{E}_t F_r(\boldsymbol{x}_{t+1}) &\leq F_r(\boldsymbol{x}_t) - \eta \langle \boldsymbol{g}_{t,r}, \overline{\boldsymbol{a}}_t \rangle + \tfrac{1}{2} L \eta^2 \mathbb{E}_t \|\boldsymbol{a}_t^J\|_2^2 \\ &\leq F_r(\boldsymbol{x}_t) - \eta \left(\tfrac{1}{2} - \frac{L\eta}{\epsilon_R} \right) w_t + \eta \left(\tfrac{1}{2} + \frac{L\eta}{\epsilon_R} \right) B_J^2 + \tfrac{1}{2} L \eta^2 V_J.\end{aligned}$$

Choose $c \leq 1$ so that $Lc/\epsilon_R \leq 1/4$ and set $\eta = c\sqrt{q/T} \leq c$. Telescoping and the metric lower bound then yield

$$\frac{1}{T} \sum_{t<T} \mathbb{E} w_t \leq \frac{4\Delta_r}{\eta T} + 3B_J^2 + 2L\eta V_J,$$

and

$$\mathcal{G}_J \le \frac{2D_R}{T}\sum_{t<T}\mathbb{E}w_t + 2b_r^2 \le \frac{8D_R\Delta_r}{\eta T} + 6D_R B_J^2 + 4LD_R\eta V_J + 2b_r^2$$
$$\le C\big[(qT)^{-1/2} + q^{-2s_J} + r^4\big].$$

Here $\Delta_r \le \Delta_*$ by Lemma A.1, and (19) gives $\eta V_J \le C(qT)^{-1/2}$. This proves (12) for RMSProp.

Regularized Muon. For matrix parameters, write $\boldsymbol{G}_{t,r} := \nabla F_r(\boldsymbol{X}_t)$ and $\overline{\boldsymbol{M}}_t := \beta_M \boldsymbol{M}_t + \gamma_M \boldsymbol{G}_{t,r}$ along each method's history. Define the predictable metric by

$$\boldsymbol{Q}_t := (\overline{\boldsymbol{M}}_t^\top \overline{\boldsymbol{M}}_t + \epsilon_M^2 \boldsymbol{I})^{-1/2}, \qquad \boldsymbol{D}_t := \boldsymbol{Q}_t^\top \otimes \boldsymbol{I}_{d_1}.$$

Under columnwise vectorization, $\boldsymbol{D}_t$ represents right multiplication by $\boldsymbol{Q}_t$. The state bound $\|\overline{\boldsymbol{M}}_t\|_{\mathrm{F}} \le U_0$ implies

$$\epsilon_M^2 \boldsymbol{I} \preceq \overline{\boldsymbol{M}}_t^\top \overline{\boldsymbol{M}}_t + \epsilon_M^2 \boldsymbol{I} \preceq (U_0^2 + \epsilon_M^2)\boldsymbol{I},$$
$$(U_0^2 + \epsilon_M^2)^{-1/2}\boldsymbol{I}_d \preceq \boldsymbol{D}_t \preceq \epsilon_M^{-1}\boldsymbol{I}_d, \qquad \|\Psi_{\epsilon_M}(\boldsymbol{M})\|_{\mathrm{F}} \le \|\boldsymbol{M}\|_{\mathrm{F}}/\epsilon_M.$$

Thus the executed direction $\boldsymbol{A}_t^J$ satisfies $\|\boldsymbol{A}_t^J\|_{\mathrm{F}} \le 3U_0/\epsilon_M$. For $\boldsymbol{e}_t := \mathrm{vec}(\boldsymbol{M}_t - \boldsymbol{G}_{t,r})$, the reference direction and retained momentum obey

$$\mathrm{vec}\,\Psi_{\epsilon_M}(\overline{\boldsymbol{M}}_t) = \boldsymbol{D}_t\,\mathrm{vec}\,\overline{\boldsymbol{M}}_t = \boldsymbol{D}_t(\mathrm{vec}\,\boldsymbol{G}_{t,r} + \beta_M \boldsymbol{e}_t), \quad \boldsymbol{M}_{t+1} = \beta_M \boldsymbol{M}_t + \gamma_M \widehat{\boldsymbol{G}}_t.$$

The second identity holds for both methods by affine preservation. Proposition A.1 and the update-block expansions give (19) with this reference direction and constants independent of $\gamma_M \in (0,1]$. The metric bounds, uniform direction bound, and affine momentum identity are precisely the properties used in (20)–(22). Applying those steps after vectorization, with $\gamma = \gamma_M$, $a = \epsilon_M$, and $\eta = c\gamma_M$, therefore gives (23) for

$$W_J := \frac{1}{T}\sum_{t<T}\mathbb{E}\langle \mathrm{vec}\,\boldsymbol{G}_{t,r}, \boldsymbol{D}_t\,\mathrm{vec}\,\boldsymbol{G}_{t,r}\rangle.$$

Substituting $\gamma_M = \sqrt{q/T}$ and converting back to the Frobenius norm yields

$$W_J \le C\big[(\gamma_M T)^{-1} + \gamma_M/q + B_J^2 + \gamma_M V_J\big] \le C\big[(qT)^{-1/2} + q^{-2s_J}\big],$$
$$\mathcal{G}_J \le \frac{2}{T}\sum_{t<T}\mathbb{E}\|\boldsymbol{G}_{t,r}\|_{\mathrm{F}}^2 + 2b_r^2 \le 2\sqrt{U_0^2 + \epsilon_M^2}\,W_J + 2b_r^2 \le C\big[(qT)^{-1/2} + q^{-2s_J} + r^4\big].$$

Choosing a common $c > 0$ below the three step-size thresholds and taking a common upper constant proves (12) for all three optimizers. The state bounds used throughout hold from zero initialization. □

A.5 Proof of Corollary 4.2

Write $a := \epsilon_A$, $\gamma_2 := 1 - \beta_2$, and $D_* := \sqrt{U_0^2 + a^2}$. Use $\mathcal{U}_t^{\mathrm{R}}$ and $\mathcal{W}_t^{\mathrm{R}}$ for the update and second-moment maps of the R-AdaZO modification described after (4), evaluated along each method's trajectory. Let $\widehat{\boldsymbol{g}}_t$ and $\widehat{\boldsymbol{g}}_{t,j}$ be the full-sample and balanced half-sample means, respectively, and let

$$\boldsymbol{a}_t^{\mathrm{R}} := \mathcal{U}_t^{\mathrm{R}}(\widehat{\boldsymbol{g}}_t), \qquad \boldsymbol{a}_t^{\mathrm{CR}} := 2\mathcal{U}_t^{\mathrm{R}}(\widehat{\boldsymbol{g}}_t) - \tfrac{1}{2}\sum_{j=1}^{2}\mathcal{U}_t^{\mathrm{R}}(\widehat{\boldsymbol{g}}_{t,j}).$$

All conditional expectations below fix the incoming history and the partition selected before the current i.i.d. queries. Constants denoted by C are common to both methods and independent of q, T, r, γ and, until c is fixed, of $c \in (0,1]$.

Proof. Lemma A.1 supplies the bound $\|\widehat{\boldsymbol{g}}_i(r)\|_2 \le U_0$, the identity $\boldsymbol{g}_{t,r} = \nabla F_r(\boldsymbol{x}_t)$, and uniform smoothness, initial-gap, and smoothing-error bounds for $0 < r \le 1$. Define $\boldsymbol{v}_{F,t} := \mathcal{W}_t^{\mathrm{R}}(\widehat{\boldsymbol{g}}_t)$ and $\boldsymbol{C}_t := \gamma_2\gamma^2(\widehat{\boldsymbol{g}}_{t,1} - \widehat{\boldsymbol{g}}_{t,2})^{\odot 2}/4$. Affine preservation and rational state processing give

$$\boldsymbol{m}_{t+1} = \mathcal{M}_t(\widehat{\boldsymbol{g}}_t) = \beta_1 \boldsymbol{m}_t + \gamma\widehat{\boldsymbol{g}}_t, \quad \boldsymbol{0} \preceq \boldsymbol{v}_{t+1}^{\mathrm{CR}} = \boldsymbol{v}_{F,t}^{\odot 2} \oslash (\boldsymbol{v}_{F,t} + \boldsymbol{C}_t) \preceq \boldsymbol{v}_{F,t} = \beta_2 \boldsymbol{v}_t + \gamma_2 \boldsymbol{m}_{t+1}^{\odot 2},$$
$$\|\boldsymbol{m}_t\|_2 \le U_0, \qquad \boldsymbol{0} \preceq \boldsymbol{v}_t \preceq U_0^2\boldsymbol{1}, \qquad \|\boldsymbol{a}_t^J\|_2 \le 3U_0/a, \quad J \in \{\mathrm{R}, \mathrm{CR}\}.$$

The rational quotient uses $0/0 = 0$, and the baseline retains $\boldsymbol{v}_{F,t}$. The last line follows by induction from the zero initial states and from the absolute extrapolation weights.

Set $\overline{\boldsymbol{m}}_{t+1} := \beta_1 \boldsymbol{m}_t + \gamma \boldsymbol{g}_{t,r}$ and define the predictable metric and conditional update bias by

$$\boldsymbol{D}_t := \operatorname{diag}\Big((\beta_2 \boldsymbol{v}_t + \gamma_2 \overline{\boldsymbol{m}}_{t+1}^{\odot 2} + a^2 \mathbf{1})^{-1/2}\Big), \quad D_*^{-1} \boldsymbol{I} \preceq \boldsymbol{D}_t \preceq a^{-1} \boldsymbol{I},$$
$$\boldsymbol{a}_t^0 := \boldsymbol{D}_t \overline{\boldsymbol{m}}_{t+1}, \qquad \overline{\boldsymbol{a}}_t := \mathbb{E}_t \boldsymbol{a}_t^J, \qquad \boldsymbol{b}_t := \overline{\boldsymbol{a}}_t - \boldsymbol{a}_t^0.$$

For $\boldsymbol{\Psi}_t(\boldsymbol{z}) := \boldsymbol{z} \oslash \sqrt{\beta_2 \boldsymbol{v}_t + \gamma_2 \boldsymbol{z}^{\odot 2} + a^2 \mathbf{1}}$, the chain rule yields

$$\mathcal{U}_t^{\mathrm{R}}(\boldsymbol{g}) = \boldsymbol{\Psi}_t(\beta_1 \boldsymbol{m}_t + \gamma \boldsymbol{g}), \qquad \nabla_{\boldsymbol{g}}^k \mathcal{U}_t^{\mathrm{R}}(\boldsymbol{g}) = \gamma^k \nabla_{\boldsymbol{z}}^k \boldsymbol{\Psi}_t(\beta_1 \boldsymbol{m}_t + \gamma \boldsymbol{g}), \quad 1 \le k \le 4.$$

The positive stabilizer and the state bounds give uniform bounds on these derivatives after removing the factors γ^k. The same holds after multiplication by $\boldsymbol{D}_t^{-1/2}$, because $\|\boldsymbol{D}_t^{-1/2}\|_{\mathrm{op}} \le \sqrt{D_*}$ and the metric is held fixed under differentiation.

Use the centered sample-mean identities and moment bounds from Lemma A.2, with $\sigma^2 := U_0^2$ bounding the conditional total variance of one record. Write $\boldsymbol{\delta}^{(j)} := \widehat{\boldsymbol{g}}_{t,j} - \boldsymbol{g}_{t,r}$ and $\boldsymbol{\delta} := (\boldsymbol{\delta}^{(1)} + \boldsymbol{\delta}^{(2)})/2$. Taylor expansion at $\boldsymbol{g}_{t,r}$, with the displayed derivative factors, gives

$$\|\boldsymbol{D}_t^{-1/2} \boldsymbol{b}_t^{\mathrm{R}}\|_2 \le C\gamma^2 \mathbb{E}_t \|\boldsymbol{\delta}\|_2^2 \le C\gamma^2/q, \quad 2\mathbb{E}_t \boldsymbol{\delta}^{\otimes 2} - \tfrac{1}{2} \sum_{j=1}^{2} \mathbb{E}_t (\boldsymbol{\delta}^{(j)})^{\otimes 2} = \mathbf{0},$$
$$\|\boldsymbol{D}_t^{-1/2} \boldsymbol{b}_t^{\mathrm{CR}}\|_2 \le C\gamma^3/q^2 + C\gamma^4/q^2 \le C\gamma^3/q^2.$$

The corrected linear expectation vanishes, the second line cancels its quadratic expectation, and the third-moment identity and fourth-moment bound control the remaining cubic and fourth-order terms.

For the corrected variance, expand to second order as

$$\boldsymbol{a}_t^{\mathrm{CR}} - \boldsymbol{a}_t^0 = \boldsymbol{L}_t + \boldsymbol{Q}_t + \boldsymbol{R}_t, \qquad \boldsymbol{L}_t := \nabla_{\boldsymbol{g}} \mathcal{U}_t^{\mathrm{R}}(\boldsymbol{g}_{t,r}) \boldsymbol{\delta}, \quad \boldsymbol{Q}_t := \tfrac{1}{2} \nabla_{\boldsymbol{g}}^2 \mathcal{U}_t^{\mathrm{R}}(\boldsymbol{g}_{t,r})[\boldsymbol{\delta}^{(1)}, \boldsymbol{\delta}^{(2)}],$$
$$\mathbb{E}_t \|\boldsymbol{L}_t\|_2^2 \le v_1 \gamma^2/q, \qquad \mathbb{E}_t \|\boldsymbol{Q}_t\|_2^2 \le C\gamma^4/q^2, \quad \mathbb{E}_t \|\boldsymbol{R}_t\|_2^2 \le C\gamma^6/q^3.$$

The remainder bound uses the third-derivative factor and the sixth-moment bound. The independent centered halves imply $\mathbb{E}_t \boldsymbol{L}_t = \mathbb{E}_t \boldsymbol{Q}_t = \mathbf{0}$ and $\mathbb{E}_t \langle \boldsymbol{L}_t, \boldsymbol{Q}_t \rangle = 0$. Consequently,

$$\begin{aligned} \mathbb{E}_t \|\boldsymbol{a}_t^{\mathrm{CR}} - \overline{\boldsymbol{a}}_t\|_2^2 &\le \mathbb{E}_t \|\boldsymbol{L}_t + \boldsymbol{Q}_t + \boldsymbol{R}_t\|_2^2 = \mathbb{E}_t \|\boldsymbol{L}_t\|_2^2 + 2\mathbb{E}_t \langle \boldsymbol{L}_t, \boldsymbol{R}_t \rangle + \mathbb{E}_t \|\boldsymbol{Q}_t + \boldsymbol{R}_t\|_2^2 \\ &\le v_1 \gamma^2/q + C\gamma^4/q^2 + C\gamma^6/q^3 \le v_1 \gamma^2/q + v_2 \gamma^4/q^2. \end{aligned}$$

Cauchy–Schwarz bounds the linear–remainder term by $C\gamma^4/q^2$. Lipschitz continuity gives the baseline variance bound $v_1\gamma^2/q$ with the same sufficiently large v_1. Thus the local errors satisfy

$$\|\boldsymbol{D}_t^{-1/2} \boldsymbol{b}_t\|_2 \le B_J, \quad B_{\mathrm{R}} := b_{\mathrm{R}} \gamma^2/q, \quad B_{\mathrm{CR}} := b_{\mathrm{CR}} \gamma^3/q^2,$$
$$\mathbb{E}_t \|\boldsymbol{a}_t^J - \overline{\boldsymbol{a}}_t\|_2^2 \le V_J, \quad V_{\mathrm{R}} := v_1 \gamma^2/q, \quad V_{\mathrm{CR}} := v_1 \gamma^2/q + v_2 \gamma^4/q^2.$$

To apply the tracking and descent bound from Appendix A.4, let $\boldsymbol{e}_t := \boldsymbol{m}_t - \boldsymbol{g}_{t,r}$ and $w_t := \langle \boldsymbol{g}_{t,r}, \boldsymbol{D}_t \boldsymbol{g}_{t,r} \rangle$. The affine first-moment recursion, uniform direction bound, and metric bounds established above give

$$\boldsymbol{a}_t^0 = \boldsymbol{D}_t(\boldsymbol{g}_{t,r} + \beta_1 \boldsymbol{e}_t), \quad \|\overline{\boldsymbol{a}}_t\|_2^2 \le C(w_t + \|\boldsymbol{e}_t\|_2^2 + B_J^2), \quad \mathbb{E}_t \|\boldsymbol{a}_t^J\|_2^2 \le C(w_t + \|\boldsymbol{e}_t\|_2^2 + B_J^2) + V_J.$$

These are the properties used in (23); hence, with $\eta = c\gamma$,

$$W_J := \frac{1}{T} \sum_{t<T} \mathbb{E} w_t \le C\left((\gamma T)^{-1} + \gamma/q + B_J^2 + \gamma V_J\right).$$

The tracking and descent argument absorbs coefficients of order c^2 and $c + c^2$, respectively; all bounds above are uniform in q, T, r, γ, so one fixed sufficiently small $c > 0$ works for both methods.

Finally, $\gamma \le 1$ and $q \ge 2$ give $\gamma^3/q \le \gamma/q$ and $\gamma^5/q^2 \le \gamma/(2q)$. The local-error bounds and the metric lower bound therefore yield

$$\mathcal{G}_{\mathrm{R}} \le 2D_* W_{\mathrm{R}} + 2b_r^2 \le C_1\left((\gamma T)^{-1} + \gamma/q + \gamma^4/q^2 + r^4\right),$$
$$\mathcal{G}_{\mathrm{CR}} \le 2D_* W_{\mathrm{CR}} + 2b_r^2 \le C_1\left((\gamma T)^{-1} + \gamma/q + \gamma^6/q^4 + r^4\right).$$

Here Lemma A.1 gives $b_r = L_H d r^2/[2(d+2)]$, and C_1 is independent of q, T, r, γ. □

B SAME-QUERY EXTENSIONS

We extend C-SQE to odd numbers of directional estimates, higher-order extrapolation, and partition averaging. All constructions reuse the same $2q$ function evaluations and hold the parameter, incoming state, and optimizer coefficients fixed. Conditional moments in this section average over both the current query record and the partition draws.

B.1 ODD NUMBERS OF DIRECTIONAL ESTIMATES

For odd $q \geq 3$, choose a partition (I_1, I_2) independently of the query record, with $q_1 := |I_1| = \lfloor q/2 \rfloor$ and $q_2 := |I_2| = \lceil q/2 \rceil$. Define $\widehat{\boldsymbol{g}}_j := q_j^{-1} \sum_{i \in I_j} \widehat{\boldsymbol{g}}_i(r)$ and $\omega_j := q_j/q$. The weighted C-SQE transition is

$$\mathcal{T}^{\mathrm{C}}_{t,q,r} := \mathrm{P}_t \left(2\mathcal{T}_t(\widehat{\boldsymbol{g}}_{q,r}) - \sum_{j=1}^{2} \omega_j \mathcal{T}_t(\widehat{\boldsymbol{g}}_j) \right). \tag{24}$$

The identities

$$\sum_{j=1}^{2} \omega_j \widehat{\boldsymbol{g}}_j = \widehat{\boldsymbol{g}}_{q,r}, \qquad \sum_{j=1}^{2} \frac{\omega_j}{q_j} = \frac{2}{q}$$

preserve affine blocks and the leading linear fluctuation before state processing, and cancel the leading sampling bias. Since $q_j \geq q/3$, the remainder bounds in Theorem 4.1 retain their orders uniformly in q. The same state processing applies, with the joint guarantees for constrained states subject to Corollary 4.1.

B.2 HIGHER-ORDER EXTRAPOLATION

Fix an extrapolation order $L \geq 2$ and let q be divisible by 2^{L-1}. At level $\ell = 0, \ldots, L-1$, partition the query indices into 2^ℓ blocks of size $n_\ell := q/2^\ell$, independently of the query record. For each block, define $\widehat{\boldsymbol{g}}_{\ell,j} := n_\ell^{-1} \sum_{i \in I_{\ell,j}} \widehat{\boldsymbol{g}}_i(r)$ and average the corresponding transitions as $\boldsymbol{Z}_\ell := 2^{-\ell} \sum_{j=1}^{2^\ell} \mathcal{T}_t(\widehat{\boldsymbol{g}}_{\ell,j})$. The order-$L$ extrapolation is

$$\boldsymbol{Z}^{(L)} := \sum_{\ell=0}^{L-1} c_\ell \boldsymbol{Z}_\ell, \qquad c_\ell := \prod_{\substack{0 \leq h < L \\ h \neq \ell}} \frac{2^h}{2^h - 2^\ell}. \tag{25}$$

For $L = 2$, the weights are $(2, -1)$, recovering basic C-SQE. The construction requires at most $2^L - 1$ frozen transition evaluations. The following theorem characterizes its bias and covariance.

Theorem B.1 (Higher-order extrapolation)**.** *Under Assumptions 2.1, 2.2, and 2.3 with $K \geq 2L$, the construction in (25) satisfies*

$$\left\| \mathbb{E}_t \boldsymbol{Z}^{(L)} - \mathcal{T}_t(\boldsymbol{g}_r) \right\|_2 = \mathcal{O}(q^{-L}), \qquad \left\| \mathrm{Cov}_t(\boldsymbol{Z}^{(L)}) - q^{-1} \boldsymbol{J}_r \boldsymbol{\Sigma}_r \boldsymbol{J}_r^\top \right\|_{\mathrm{op}} = \mathcal{O}(q^{-2}).$$

Every affine transition block equals its full-sample value. For fixed history, radius, and L, the implied constants are independent of q and the partitions. Under identity state processing, these guarantees hold for the executed joint transition.

Proof. Condition on the partitions. Taylor expansion through degree $2L - 1$ and the centered-moment identities in Lemma A.2 give

$$\mathbb{E}_t \mathcal{T}_t(\widehat{\boldsymbol{g}}_{n,r}) = \mathcal{T}_t(\boldsymbol{g}_r) + \sum_{k=1}^{L-1} \boldsymbol{b}_{r,k} n^{-k} + \mathcal{O}(n^{-L}),$$

where the coefficients $\boldsymbol{b}_{r,k}$ are independent of n. The interpolation weights satisfy $\sum_\ell c_\ell = 1$ and $\sum_\ell c_\ell 2^{k\ell} = 0$ for $1 \leq k < L$, so

$$\mathbb{E}_t \boldsymbol{Z}^{(L)} - \mathcal{T}_t(\boldsymbol{g}_r) = \sum_{k=1}^{L-1} \frac{\boldsymbol{b}_{r,k}}{q^k} \sum_{\ell=0}^{L-1} c_\ell 2^{k\ell} + \mathcal{O}(q^{-L}) = \mathcal{O}(q^{-L}).$$

For the covariance, let $\boldsymbol{\delta} := \widehat{\boldsymbol{g}}_{q,r} - \boldsymbol{g}_r$. The exact block-averaging identity gives $\boldsymbol{Z}^{(L)} - \mathcal{T}_t(\boldsymbol{g}_r) = \boldsymbol{J}_r\boldsymbol{\delta} + \boldsymbol{Q} + \boldsymbol{R}$, where $\boldsymbol{Q}$ collects the quadratic Taylor terms and $\boldsymbol{R}$ is the remaining error. The sample-moment bounds and the identity $\sum_\ell c_\ell 2^\ell = 0$ yield

$$\mathbb{E}_t\|\boldsymbol{Q}\|_2^2 = \mathcal{O}(q^{-2}), \qquad \mathbb{E}_t\|\boldsymbol{R}\|_2^2 = \mathcal{O}(q^{-3}), \qquad \mathbb{E}_t[(\boldsymbol{J}_r\boldsymbol{\delta})\boldsymbol{Q}^\top] = \boldsymbol{0}.$$

Cauchy–Schwarz therefore gives

$$\left\|\mathrm{Cov}_t(\boldsymbol{Z}^{(L)}) - q^{-1}\boldsymbol{J}_r\boldsymbol{\Sigma}_r\boldsymbol{J}_r^\top\right\|_{\mathrm{op}} \le 2\|\,\mathrm{Cov}_t(\boldsymbol{J}_r\boldsymbol{\delta}, \boldsymbol{R})\|_{\mathrm{op}} + \mathbb{E}_t\|\boldsymbol{Q} + \boldsymbol{R}\|_2^2 = \mathcal{O}(q^{-2}).$$

Affine preservation follows from the same averaging identity. The conditional mean is independent of the partitions, so averaging over their draws preserves both bounds. □

Regularized Muon uses identity state processing. For Adam, R-AdaZO, and RMSProp, we pair the higher-order update with the basic rational state rule computed from the full sample and the level-1 halves. The executed update retains bias $\mathcal{O}(q^{-L})$, while the retained state has bias $\mathcal{O}(q^{-2})$ when $\mathcal{W}_t(\boldsymbol{g}_r) \succ \boldsymbol{0}$, by Corollary 4.1.

B.3 Partition Averaging

For even q, draw $B \ge 1$ independent uniform balanced partitions with replacement, independently of the query record. Define $\widehat{\boldsymbol{g}}_{b,j} := (2/q)\sum_{i\in I_{b,j}} \widehat{\boldsymbol{g}}_i(r)$ and $\boldsymbol{H}_b := \frac{1}{2}[\mathcal{T}_t(\widehat{\boldsymbol{g}}_{b,1}) + \mathcal{T}_t(\widehat{\boldsymbol{g}}_{b,2})]$. Average before applying the state operator:

$$\boldsymbol{Z}_B := 2\mathcal{T}_t(\widehat{\boldsymbol{g}}_{q,r}) - \frac{1}{B}\sum_{b=1}^{B}\boldsymbol{H}_b, \qquad \mathcal{T}_{t,q,r}^{\mathrm{C},B} := \mathrm{P}_t(\boldsymbol{Z}_B). \tag{26}$$

Here P_t uses the full-sample state as its reference, and $B = 1$ recovers basic C-SQE. The construction requires at most $2B + 1$ frozen transition evaluations.

For the query record $\mathcal{Q}_t := (\widehat{\boldsymbol{g}}_i(r))_{i=1}^q$, define

$$\boldsymbol{Z}_\infty := 2\mathcal{T}_t(\widehat{\boldsymbol{g}}_{q,r}) - \mathbb{E}_t[\boldsymbol{H}_1 \mid \mathcal{Q}_t], \qquad \boldsymbol{V}_{\mathrm{part}} := \mathbb{E}_t[\mathrm{Cov}_t(\boldsymbol{H}_1 \mid \mathcal{Q}_t)].$$

The following theorem isolates the covariance contribution controlled by B.

Theorem B.2 (Partition averaging)**.** *Under the oracle and transition assumptions of Theorem 4.1, the construction in (26) satisfies*

$$\mathbb{E}_t\boldsymbol{Z}_B = \mathbb{E}_t\boldsymbol{Z}_1, \qquad \mathrm{Cov}_t(\boldsymbol{Z}_B) = \mathrm{Cov}_t(\boldsymbol{Z}_\infty) + B^{-1}\boldsymbol{V}_{\mathrm{part}},$$

where $\boldsymbol{V}_{\mathrm{part}} \succeq \boldsymbol{0}$ and $\|\boldsymbol{V}_{\mathrm{part}}\|_{\mathrm{op}} = \mathcal{O}(q^{-2})$. Thus $\boldsymbol{Z}_B$ satisfies the bias and covariance bounds in (11) uniformly in B.

Increasing B preserves the mean and reduces the covariance due to partition randomness by a factor of $1/B$. The leading q^{-1} covariance remains unchanged.

Proof. Conditionally on $\mathcal{Q}_t$, the variables $\boldsymbol{H}_b$ are independent and identically distributed, giving

$$\mathbb{E}_t[\boldsymbol{Z}_B \mid \mathcal{Q}_t] = \boldsymbol{Z}_\infty, \qquad \mathrm{Cov}_t(\boldsymbol{Z}_B \mid \mathcal{Q}_t) = B^{-1}\,\mathrm{Cov}_t(\boldsymbol{H}_1 \mid \mathcal{Q}_t).$$

Total expectation and covariance give the two identities. The constant and linear Taylor terms of $\boldsymbol{H}_1$ are fixed by the query record, so, with $\boldsymbol{\delta} := \widehat{\boldsymbol{g}}_{q,r} - \boldsymbol{g}_r$,

$$\|\boldsymbol{V}_{\mathrm{part}}\|_{\mathrm{op}} \le \mathbb{E}_t\|\boldsymbol{H}_1 - \mathcal{T}_t(\boldsymbol{g}_r) - \boldsymbol{J}_r\boldsymbol{\delta}\|_2^2 = \mathcal{O}(q^{-2}).$$

Finally, $\mathrm{Cov}_t(\boldsymbol{Z}_B) = \mathrm{Cov}_t(\boldsymbol{Z}_1) - (1 - B^{-1})\boldsymbol{V}_{\mathrm{part}}$ transfers the bounds of Theorem 4.1 uniformly to all $B \ge 1$. □

Identity processing transfers all statements to the executed joint transition. For the rational Adam, R-AdaZO, and RMSProp rules, the joint bias and covariance bounds remain valid when $\mathcal{W}_t(\boldsymbol{g}_r) \succ \boldsymbol{0}$: Jensen's inequality bounds the moments of the averaged state correction by their single-partition counterparts, allowing the proof of Corollary 4.1 to apply uniformly in B. The exact mean and covariance identities also hold for the unchanged update block.

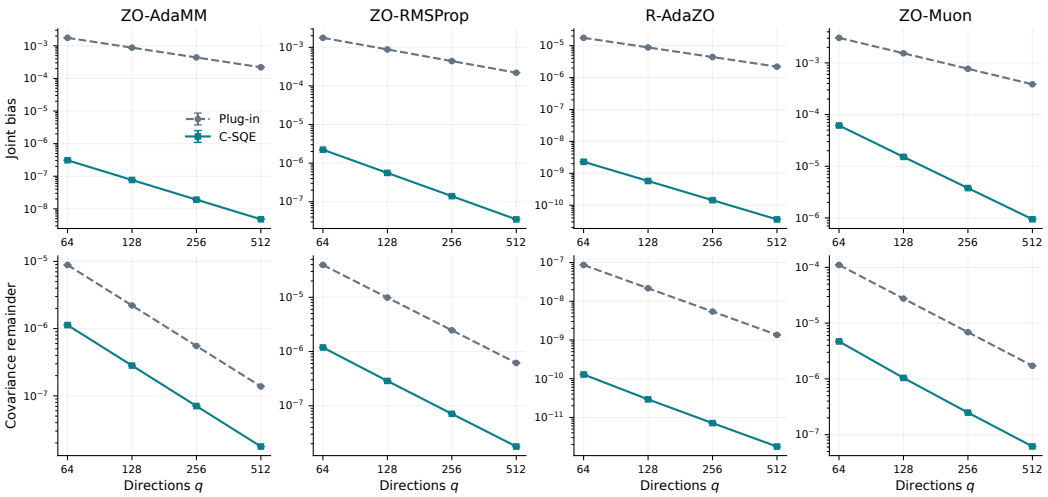

Figure 2: Joint bias (top) and covariance remainder (bottom) at $d = 128$. Error bars indicate approximate pointwise 95% Monte Carlo confidence radii.

C Experimental Details

We provide the settings and measurement procedures for the three experiments in Section 5.

C.1 Theory Validation

Configuration. We use $F(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$ with $d = 128$ and $\boldsymbol{x}_t = \boldsymbol{e}_1$, the first coordinate vector. The exact central-difference estimate is $\widehat{\boldsymbol{g}}_i(r) = d(\boldsymbol{e}_1^\top \boldsymbol{z}_i)\boldsymbol{z}_i$, with $\boldsymbol{z}_i$ uniform on the unit sphere; we evaluate this identity directly to avoid cancellation error. For each $q \in \{64, 128, 256, 512\}$, we generate 32,768 independent records, shared across all optimizer pairs, and use one balanced partition for C-SQE.

Adam and R-AdaZO use incoming moments $\boldsymbol{m}_t = 0.25\boldsymbol{e}_1$ and $\boldsymbol{v}_t = 4\mathbf{1}$, coefficients $(\beta_1, \beta_2) = (0.9, 0.99)$, and $\epsilon_A = 0.1$. RMSProp uses $\boldsymbol{w}_t = 4\mathbf{1}$, $\beta_R = 0.99$, and $\epsilon_R = 0.1$. Muon reshapes vectors in row-major order into 16×8 matrices and uses $\boldsymbol{M}_t = 0.25\boldsymbol{E}_{11}$, $\beta_M = 0.9$, and $\epsilon_M = 0.5$, where $\boldsymbol{E}_{11}$ is the first coordinate matrix. All measurements use the executed joint transition, including the state processing in Remark 3.1.

Measurement and uncertainty. We use the bias and covariance-remainder metrics in Section 5.1, with unweighted log–log slopes fitted over all four query counts. Cubic Taylor control variates with exact finite-sample moments through degree six and averaging over configuration symmetries reduce Monte Carlo error. Residual corrections retain the complete nonlinear output, rational state processing, and all update–state cross-covariances. These controls are used only for offline measurement. We form delete-block pseudovalues from 64 blocks of 512 records and use 2,048 multinomial bootstrap draws to obtain approximate pointwise 95% confidence radii in Euclidean norm for the mean and Frobenius norm for the covariance. The reported relative covariance-error upper limit is the remainder norm estimate plus its confidence radius, divided by $\|\boldsymbol{V}_q\|_{\mathrm{F}}$ and expressed as a percentage. Figure 2 supplements Table 1 with the complete query-count sweep.

C.2 Synthetic Optimization

Objectives and settings. We use the following objectives without dimension normalization, with $w_j := 1 + (x_j - 1)/4$.

Quadratic: $F_{\mathrm{Quadratic}}(\boldsymbol{x}) = \frac{1}{2}\sum_{j=1}^{d} x_j^2$.

Cubic: $F_{\mathrm{Cubic}}(\boldsymbol{x}) = \sum_{j=1}^{d}(|x_j|^3 + \frac{1}{2}x_j^2)$.

Rosenbrock: $F_{\text{Rosenbrock}}(\boldsymbol{x}) = \sum_{j=1}^{d-1}[100(x_{j+1} - x_j^2)^2 + (1 - x_j)^2]$.

Lévy: $F_{\text{Levy}}(\boldsymbol{x}) = \sin^2(\pi w_1) + \sum_{j=2}^{d-1}(w_j - 1)^2[1 + 10\sin^2(\pi w_j + 1)] + (w_d - 1)^2[1 + \sin^2(2\pi w_d)]$. All four objectives have minimum value zero. We set $d = 1{,}000$, $q = 10$, $r = 10^{-5}$, and $\eta = 10^{-3}$, and run each method for $2{,}000$ iterations. First- and second-moment coefficients are 0.9 and 0.99, respectively, where applicable, and the additive stabilizer for the coordinate optimizers is 10^{-8}. Initial parameters have independent standard normal coordinates, and optimizer states start at zero. We use the central-difference ZOGE in (1) and the C-SQE state handling in Remark 3.1.

Pairing and reporting. The ten paired seeds are $200, 201, \ldots, 208, 300$. For seed s, initialization and direction generators use seeds $500000 + s$ and $700000 + 100000s$, respectively. Methods share initializations and unit-sphere directions, with each method evaluating its own current parameter. Each run uses $40{,}000$ optimization queries; objective logging every 20 iterations is counted separately. Figure 1 reports the natural logarithm of the mean optimality gap across the ten seeds, without smoothing.

C.3 LLM Fine-Tuning

Model and prompt parameterization. We freeze OPT-125M and optimize a soft prompt with $L = 10$ tokens in a $d = 64$ dimensional intrinsic parameter space. The embedding dimension is $h = 768$, giving an ambient prompt dimension of $Lh = 7{,}680$. For each experimental seed, we initialize $\boldsymbol{P}_0 \in \mathbb{R}^{10\times 768}$ using the pretrained embeddings of 10 token IDs sampled uniformly with replacement from the non-special-token vocabulary. Independently, we sample a Gaussian matrix $\widetilde{\boldsymbol{A}} \in \mathbb{R}^{7680\times 64}$ and obtain an orthonormal basis $\boldsymbol{A}$ by reduced QR decomposition with deterministic sign fixing. The prompt is parameterized as

$$\boldsymbol{P}(\boldsymbol{z}) = \boldsymbol{P}_0 + \text{reshape}(\boldsymbol{A}\boldsymbol{z}, 10, 768), \qquad \boldsymbol{z} \in \mathbb{R}^{64}. \tag{27}$$

We initialize $\boldsymbol{z}_0 = \boldsymbol{0}$ and prepend the resulting prompt embeddings to the text-token embeddings. The maximum sequence length is 128 positions, including the 10 prompt positions, leaving at most 118 positions for text tokens. Optimization and evaluation use batch sizes of 16 and 8, respectively.

Task construction and scoring. For SST-2, we format each input as {`sentence`} `It was` and use `terrible` and `great`, each prefixed by a space, as the single-token verbalizers for the negative and positive classes. The frozen language-model head maps the hidden state at the final text position to vocabulary logits. We retain the two verbalizer logits and minimize their cross-entropy loss.

For COPA, we remove a trailing period from the premise, when present, and append `because` for cause questions or `so` for effect questions. Each candidate is appended with a leading space and its first character converted to lowercase. Let x denote the formatted context and c_j the tokenized completion for candidate j. Its score is the mean conditional log-probability over completion tokens:

$$s_j(\boldsymbol{z}) = \frac{1}{|c_j|}\sum_{t=1}^{|c_j|} \log p_\theta(c_{j,t} \mid \boldsymbol{P}(\boldsymbol{z}), x, c_{j,<t}) \,. \tag{28}$$

Only completion tokens contribute to the score; context and padding tokens are excluded. We minimize the cross-entropy loss over (s_1, s_2). The implementation rejects any example for which either candidate sequence, including the soft prompt, exceeds the 128-position budget.

Data construction. We construct disjoint training and held-out test subsets of $1{,}000/500$ examples for SST-2 and $320/80$ examples for COPA from their official training splits. Both tasks use deterministic label-stratified sampling without replacement, with class counts allocated by the largest-remainder rule to approximately preserve the source label distribution. Hyperparameters are selected using runs with seeds 9900 and 9901 and then fixed for the reported runs with seeds 9902–9904, each lasting $20{,}000$ optimization steps. All reported test accuracies use these held-out test subsets; the official validation and test splits are unused.

Table 3: Hyperparameters shared by each baseline and its C-SQE variant. Muon momentum is specified separately in the text.

Optimizer family	SST-2 η	COPA η	β_1	β_2 or α
AdaMM	0.0096	0.0064	0.90	0.99
R-AdaZO	0.0016	0.0032	0.80	0.995
RMSProp	0.0040	0.0060	—	0.99
Muon	0.0075	0.0600	—	—

Gradient estimation and query budget. At each step, we draw $q = 2$ independent directions uniformly from the unit sphere in $\mathbb{R}^{64}$. For mini-batch loss F, the gradient estimate is

$$\widehat{\boldsymbol{g}}(\boldsymbol{z}) = \frac{d}{q}\sum_{i=1}^{q}\frac{F(\boldsymbol{z}+r\boldsymbol{u}_i)-F(\boldsymbol{z}-r\boldsymbol{u}_i)}{2r}\,\boldsymbol{u}_i, \qquad d = 64. \tag{29}$$

The perturbation radius is $r = 10^{-5}$ for SST-2 and $r = 10^{-3}$ for COPA. C-SQE uses $K = 1$ complementary balanced bipartition, the unique partition $\{1\} \mid \{2\}$ for $q = 2$. The full-sample and two singleton transitions share the incoming optimizer state and coefficients and reuse the same directional estimates, requiring no additional objective queries.

Each step requires $2q = 4$ endpoint loss evaluations, totaling $80{,}000$ optimization queries per run. An additional $20{,}001$ unperturbed center-loss evaluations are performed for diagnostics: one at initialization and one per optimization step. Center-loss diagnostics and test evaluations are excluded from the optimization-query count.

Optimizer settings. Each baseline and its C-SQE variant share the hyperparameters in Table 3, which are fixed across seeds within each task–optimizer combination. AdaMM and R-AdaZO use raw moments without time-dependent bias correction and denominators $\sqrt{\boldsymbol{v}+\epsilon\mathbf{1}}$, with $\epsilon = 10^{-8}$ on both tasks. This corresponds to $\epsilon_A^2 = \epsilon$ in the theoretical notation. AdaMM accumulates squared gradient estimates in its second moment, while R-AdaZO accumulates the elementwise square of the updated first moment. RMSProp uses no first-moment momentum and updates $\boldsymbol{w}_t = \alpha\boldsymbol{w}_{t-1} + (1-\alpha)\widehat{\boldsymbol{g}}_t^2$, with direction $\widehat{\boldsymbol{g}}_t/(\sqrt{\boldsymbol{w}_t}+\epsilon\mathbf{1})$, $\alpha = 0.99$, and $\epsilon = 10^{-8}$ on both tasks. Its stabilizer is outside the square root; the analyzed transition in (5) places it inside.

Muon uses an 8×8 matrix representation of the intrinsic update, Nesterov momentum of 0.90 for SST-2 and 0.85 for COPA, and five Newton–Schulz iterations with coefficients $(a, b, c) = (3.4445, -4.775, 2.0315)$ and normalization stabilizer 10^{-7}. Newton–Schulz computations use float64; intrinsic parameters, momentum states, soft prompts, and frozen model weights use FP32 on both tasks. The implementation uses finite-step Newton–Schulz orthogonalization, while the analyzed transition in (6) uses the smooth regularized polar map. C-SQE-Muon retains the full-sample affine momentum state and corrects the update without second-moment or rational-state processing.

Randomness control. For experimental seed s, the direction, mini-batch, split, partition, prompt, and basis generators use seeds $s + 10{,}000$, $s + 20{,}000$, $s + 30{,}000$, $s + 40{,}000$, $s + 60{,}000$, and $s + 70{,}000$, respectively; the model seed is s. Within each task and seed, paired methods share the model checkpoint, base prompt, intrinsic basis, data split, mini-batch schedule, and direction stream. Together with the shared hyperparameters, these controls isolate the optimizer transition rule. For $q = 2$ and $K = 1$, the partition is deterministic, and its seed is retained in the run configuration.

Evaluation and reporting. For each seed, training loss averages the unsmoothed center mini-batch losses over the final 500 optimization steps, and the reported test accuracy is measured at the final checkpoint at step $20{,}000$. Table 2 reports the mean and sample standard deviation of the per-seed metrics over seeds 9902–9904.

For visualization, each seed's trajectory is smoothed using a centered moving average over 21 logged values, recorded every 10 optimization steps. At the boundaries, the average uses the available values within the window. We then compute the pointwise mean and sample standard deviation across the three smoothed trajectories.

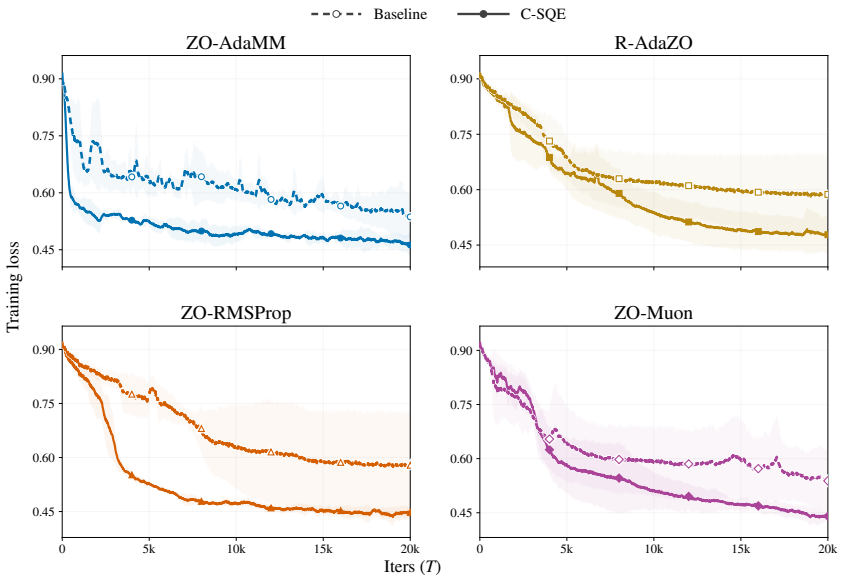

Figure 3: Smoothed training loss on SST-2 with OPT-125M, $q = 2$, and $r = 10^{-5}$. Curves show the mean across three seeds, and shaded regions indicate ± 1 sample standard deviation. Dashed and solid curves denote baselines and C-SQE variants, respectively.

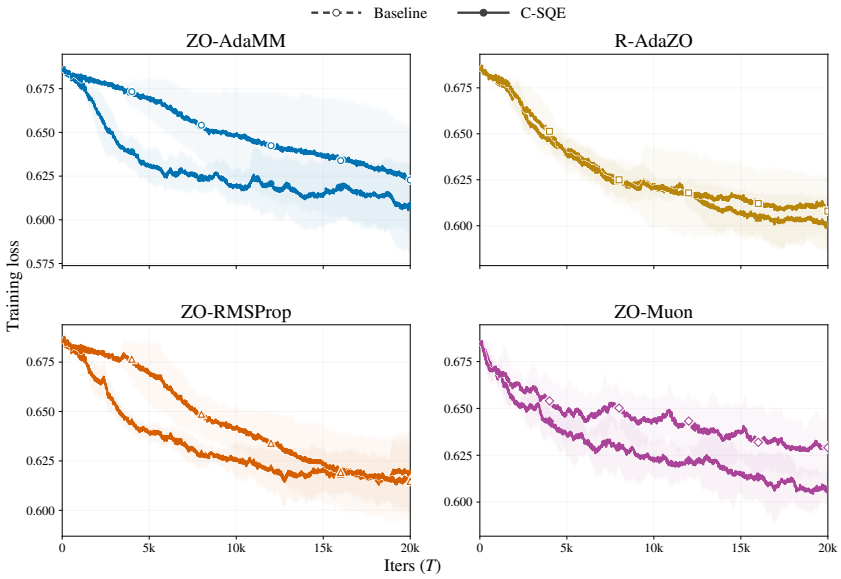

Figure 4: Smoothed training loss on COPA with OPT-125M, $q = 2$, and $r = 10^{-3}$. Curves show the mean across three seeds, and shaded regions indicate ± 1 sample standard deviation. Dashed and solid curves denote baselines and C-SQE variants, respectively.

Figures 3 and 4 compare the training-loss trajectories across four optimizer families. On both tasks, C-SQE achieves lower mean training loss by the end of training in every family, consistent with Table 2.

Peak GPU memory. We measure peak allocated GPU memory during optimization using the same GPU and model settings for each baseline/C-SQE pair. Each run uses a fresh process, with

Table 4: Peak allocated GPU memory (MiB) during OPT-125M soft-prompt tuning with $d = 64$, $q = 2$, and batch size 16. Each entry is the maximum training peak over seeds 9902–9904. Δ denotes the increase from the baseline to C-SQE.

Dataset	Optimizer family	Baseline	C-SQE	Δ
SST-2	AdaMM	561.05469	561.05664	0.00195
	R-AdaZO	561.05469	561.05664	0.00195
	RMSProp	561.05469	561.05664	0.00195
	Muon	561.05469	561.05664	0.00195
COPA	AdaMM	619.30811	619.31006	0.00195
	R-AdaZO	619.30811	619.31006	0.00195
	RMSProp	619.30811	619.31006	0.00195
	Muon	619.30811	619.31006	0.00195

Table 5: Wall-clock time for 20,000 optimization steps with OPT-125M, $q = 2$, and batch size 16 (mean $\pm$ sample standard deviation over three seeds). Times are in minutes. Relative overhead is $100(\overline{T}_{\text{C-SQE}}/\overline{T}_{\text{base}} - 1)$.

Dataset	Optimizer family	Baseline	C-SQE	Overhead (%)
SST-2	AdaMM	30.54 ± 0.28	31.12 ± 0.51	1.91
	R-AdaZO	30.40 ± 0.45	31.40 ± 0.63	3.29
	RMSProp	30.21 ± 0.56	31.23 ± 0.50	3.38
	Muon	30.61 ± 0.59	31.59 ± 0.30	3.22
COPA	AdaMM	28.27 ± 0.32	29.27 ± 0.60	3.51
	R-AdaZO	28.29 ± 0.50	29.05 ± 0.54	2.70
	RMSProp	28.33 ± 0.59	29.19 ± 1.16	3.03
	Muon	28.34 ± 0.20	29.75 ± 0.44	4.98

model loading and test evaluation outside the measurement interval. The reported peaks include the resident frozen model, forward-pass temporaries, optimizer states, and correction buffers.

Table 4 shows negligible additional peak-memory overhead from C-SQE: approximately 0.002 MiB (about 2 KiB), or less than 0.001%, across all evaluated configurations.

Wall-clock time. Experiments run on a Tesla V100S-PCIE-32GB GPU using Python 3.10.19, PyTorch `2.9.1+cu128`, CUDA 12.8, cuDNN 9.10.2, and Transformers 4.57.6. We measure the time required for 20,000 optimization steps under identical hardware, precision, and batch-size settings within each baseline/C-SQE pair. The timed interval includes mini-batch preparation, directional perturbations, endpoint loss evaluations, gradient estimation, and optimizer updates, including C-SQE transition evaluations and state processing. Model loading, data preprocessing, test evaluation, center-loss diagnostics, and file logging are excluded. Each run uses a fresh process, with warm-up followed by restoration of the initial parameters, optimizer states, and random-number generator states. We synchronize the GPU immediately before starting and stopping the timer and report the mean and sample standard deviation across seeds 9902–9904.

Table 5 shows that C-SQE adds 1.91%–4.98% to the optimization time across the evaluated configurations. The additional transition evaluations and state processing therefore incur a modest runtime overhead while preserving the function-query budget.