OPR-BROS: HISTORY-AWARE PRINCIPAL–RANDOM BI-SKETCHING FOR BILEVEL OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Stochastic bilevel optimization (SBO) is a standard framework for data reweighting, hyperparameter optimization, meta-learning, and data-mixture learning. Exact single-loop SBO methods operate with full-dimensional states and derivative computations. Randomized-subspace methods reduce the working dimension, motivating estimators that combine conditional unbiasedness, efficient operator evaluation, and reuse of persistent directions. We propose OPR-BROS, a history-aware single-loop coordinate method that combines independent operator bi-sketching with predictable principal-random factors. Independent input and output sketches yield conditionally unbiased Hessian actions, while a past-only Oja update learns reusable directions and a fresh random complement preserves exploration. We establish exact sketch-moment and mean-squared-error identities, prove that OPR-BROS retains the standard $\mathcal{O}(K^{-1/2})$ squared-stationarity rate, or $\mathcal{O}(\varepsilon^{-2})$ iteration complexity, under the stated assumptions, and derive a bound that retains the realized, operator-weighted sketch risks. Small-task experiments show favorable validation time-to-quality, and a Pile-280M experiment attains competitive upper-level loss with low measured step time and peak memory.

1 INTRODUCTION

Various learning systems require choosing a data policy, training rule, or hyperparameter together with the model parameters trained under that choice. This naturally leads to bilevel optimization. Representative applications include data cleaning and sample reweighting, where a validation objective determines the influence of training examples, hyperparameter optimization and meta-learning, where downstream performance selects the training procedure, and data-mixture or hyper-representation learning, where training configurations and model parameters are optimized jointly (Ren et al., 2018, Finn et al., 2017, Franceschi et al., 2018, Xie et al., 2023). In these settings, the upper-level variable x represents data weights, training controls, or hyperparameters, while the lower-level variable $\boldsymbol{Y}$ contains the model parameters trained under the selected configuration. Let $\mathcal{Y} := \prod_{\ell=1}^{L} \mathbb{R}^{m_\ell \times n_\ell}$ and $\boldsymbol{Y} = (Y_1, \dots, Y_L) \in \mathcal{Y}$. We study the following stochastic bilevel optimization (SBO) problem:

$$\min_{x \in \mathbb{R}^{d_x}} \quad \Phi(x) := f\big(x, \boldsymbol{Y}^\star(x)\big), \qquad \boldsymbol{Y}^\star(x) := \arg\min_{\boldsymbol{Y} \in \mathcal{Y}} g(x, \boldsymbol{Y}), \tag{1}$$

where the upper- and lower-level population objectives are $f(x, \boldsymbol{Y}) = \mathbb{E}_{\xi \sim \mathcal{D}_f}[\ell_f(x, \boldsymbol{Y}, \xi)]$ and $g(x, \boldsymbol{Y}) = \mathbb{E}_{\zeta \sim \mathcal{D}_g}[\ell_g(x, \boldsymbol{Y}, \zeta)]$. Here, ξ and ζ denote random data samples or minibatches drawn from the corresponding distributions. The algorithm accesses f and g through stochastic gradients, Hessian–vector products, and Jacobian–vector products. Following prior work on stochastic bilevel optimization (Ji et al., 2021, Chen et al., 2022, 2024), we focus on the nonconvex–strongly-convex setting, where Φ may be nonconvex and $g(x, \cdot)$ is μ_g-strongly convex for every fixed x.

Hypergradients and operator modeling. Let $\mathscr{H}(x, \boldsymbol{Y}) := \nabla^2_{\boldsymbol{Y}\boldsymbol{Y}} g(x, \boldsymbol{Y})$ and $\mathscr{J}(x, \boldsymbol{Y}) := \nabla^2_{x\boldsymbol{Y}} g(x, \boldsymbol{Y})$. Under standard smoothness conditions, the value-function hypergradient satisfies

$$\nabla\Phi(x) = \nabla_x f\big(x, \boldsymbol{Y}^\star(x)\big) - \mathscr{J}\big(x, \boldsymbol{Y}^\star(x)\big)\big[\boldsymbol{Z}^\star(x)\big]. \tag{2}$$

Here, $\boldsymbol{Z}^\star(x) := \mathscr{H}\big(x, \boldsymbol{Y}^\star(x)\big)^{-1}\big[\nabla_{\boldsymbol{Y}} f\big(x, \boldsymbol{Y}^\star(x)\big)\big]$ is the auxiliary solution associated with the lower-level Hessian. This decomposition identifies the derivative primitives that a single-loop method must approximate: lower-level gradients, Hessian actions for the auxiliary recursion, mixed-Jacobian

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.

actions for the hypergradient, and direct upper-level gradients. A memory-efficient estimator should reduce the dimension of these primitives while preserving the full-space recursion in expectation.

The large-scale bottleneck and compression routes. As the lower-level model grows, maintaining full-space lower and auxiliary states and evaluating full-dimensional Hessian–vector products (HVPs) and Jacobian–vector products (JVPs) becomes increasingly expensive. Single-loop methods such as SOBA, SABA, and MA-SOBA preserve the standard bilevel recursion using ambient-dimensional derivative primitives and online direction states (Dagréou et al., 2022, Chen et al., 2024, 2022). Existing memory-oriented alternatives follow different approximation routes: penalty methods optimize a finite-penalty surrogate, fully first-order methods replace the standard hypergradient recursion, and finite-difference methods introduce approximation errors together with additional regularity conditions (Shen & Chen, 2023, Kwon et al., 2023, Yang et al., 2023). Variance-reduced methods such as SPABA and PnPBO primarily target oracle complexity under particular sampling or finite-sum structures (Chu et al., 2024, 2026). These developments motivate unbiased compression of the single-loop recursion with adaptive reuse of persistent directions.

BROS as a strong starting point. BROS (Zhang et al., 2026) is a memory-efficient approach for SBO. It places the lower and auxiliary updates in randomized subspaces, preserves the standard single-loop hypergradient recursion, and uses a Rademacher bi-probe to correct the bias induced by shared projections. Under standard assumptions, BROS retains the $\mathcal{O}(\varepsilon^{-2})$ iteration order of MA-SOBA while reducing the peak memory required by full-space computation. We therefore use BROS as the randomized-subspace baseline for our single-loop construction. BROS's iteration-wise fresh sampling motivates a history-adaptive extension that combines reusable directions with continued random exploration.

Limitations of BROS. Despite these advantages, BROS still faces several structural limitations that motivate our subsequent design choices.

L1. Shared projections require distribution-specific bias correction. With a shared blockwise factor $\boldsymbol{P}$ and lift $\boldsymbol{Q} := \boldsymbol{P}\boldsymbol{P}^\top$, the vanilla Hessian action is $\widehat{\mathscr{H}}_{\mathrm{share}}[\boldsymbol{Z}] := \boldsymbol{Q}\mathscr{H}(x, \boldsymbol{Y})[\boldsymbol{Q}\boldsymbol{Z}]$. Sharing the random lift between the two operator roles can introduce a distribution-dependent bias in the expected Hessian action. BROS corrects this bias using fourth-moment identities of the shared projection law. Different projection laws, ranks, or history-dependent factors therefore require rederiving the correction.

L2. Bias correction introduces an additional second-order query. BROS combines its production subspace HVPs with an additional probe-based HVP for bias correction. This additional second-order query motivates a correction-free estimator that preserves the target conditional mean.

L3. Temporal structure motivates history-adaptive sampling. The basic BROS policy draws a fresh isotropic subspace at each iteration, maintaining uniform exploration across directions. Persistent lower-level or auxiliary directions offer an opportunity to allocate part of the sketch budget to reusable structure. This motivates an operator-aware characterization of how a historical principal subspace and a fresh complement jointly affect HVP/JVP estimation error.

Key contributions of OPR-BROS. We address the above limitations through three coupled advances in estimator design, online subspace learning, and single-loop analysis.

C1. A conditionally unbiased operator estimator. We develop a coordinate chain-rule realization of independent input/output bi-sketching for stochastic bilevel optimization. The lifted Hessian estimator uses independent factors on its input and output sides, while the mixed-Jacobian estimator uses an independent input factor and the direct upper-gradient branch is evaluated in full space. We prove conditional unbiasedness for the resulting product-space and multilayer cross-block actions. Conditional unbiasedness follows directly from the independent-factor construction, yielding a correction-free production estimator. This directly addresses Limitations **L1** and **L2**.

C2. Operator-aware principal–random history for bilevel recursion. We extend principal–random sketching to stochastic bilevel derivative operators with a past-only online basis. Our analysis separately characterizes the effects of the retained subspace and fresh complement on the Hessian-input, Hessian-output, and mixed-Jacobian branches, together with the unsketched direct-gradient contribution. We derive exact factor moments and an operator-weighted MSE decomposition that identifies when history reduces the current estimation error. The history-sensitive stationarity bound

retains these risks, so a smaller realized cumulative risk gives a smaller contribution to its certificate. This directly addresses Limitation **L3**.

C3. A single-loop convergence and resource certificate. We embed the estimator into a complete single-loop recursion for the upper, lower, auxiliary, and moving-average states. Under the stated nonconvex–strongly-convex assumptions and a uniform factor-moment bound, the analyzed six-branch recursion achieves an expected squared-stationarity rate of $\mathcal{O}(K^{-1/2})$. Hence, for fixed and uniformly bounded moment parameters, it retains the $\mathcal{O}(\varepsilon^{-2})$ iteration order of MA-SOBA and BROS (Chen et al., 2024, Zhang et al., 2026). We additionally account for persistent history storage, projected activations, independent factors, HVP/JVP calls, and basis-update costs. This addresses the resource implications of Limitation **L2** and verifies that the proposed correction-free estimator preserves the standard single-loop complexity order.

Additional related work appears in Appendix A.

Notation. Bold uppercase letters denote block variables, such as $\boldsymbol{Z} = (Z_1, \dots, Z_L) \in \mathcal{Y}$, with the product Frobenius norm and inner product. For $\boldsymbol{V} = (V_1, \dots, V_L)$, the operators act as $(\mathscr{H}(x, \boldsymbol{Y})[\boldsymbol{V}])_\ell = \sum_{t=1}^{L} \mathscr{H}_{\ell t}(x, \boldsymbol{Y})[V_t]$ and $\mathscr{J}(x, \boldsymbol{Y})[\boldsymbol{V}] = \sum_{t=1}^{L} \mathscr{J}_t(x, \boldsymbol{Y})[V_t]$, where $\mathscr{H}_{\ell t} := D_{Y_t}[\nabla_{Y_\ell} g]$ and $\mathscr{J}_t := D_{Y_t}[\nabla_x g]$. These definitions retain all cross-block derivatives.

At state k, write $\mathscr{H}_k := \mathscr{H}(x^k, \boldsymbol{Y}^k)$ and $\mathscr{J}_k := \mathscr{J}(x^k, \boldsymbol{Y}^k)$. Their full-space sample counterparts are $\mathscr{H}^{k+1} := \nabla^2_{\boldsymbol{Y}\boldsymbol{Y}} \ell_g(x^k, \boldsymbol{Y}^k, \zeta_k)$ and $\mathscr{J}^{k+1} := \nabla^2_{x\boldsymbol{Y}} \ell_g(x^k, \boldsymbol{Y}^k, \zeta_k)$. The superscript $k+1$ indicates an oracle generated at state k, as in BROS. Let $\mathcal{F}_k$ contain the current states, historical bases, fixed rank configuration, and all past randomness, excluding current factors and oracle samples. We write $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_k]$. Section 3 gives the conditional oracle identities for the sampled factors.

2 OPR-BROS: History-Aware Independent Operator Bi-Sketching

This section develops a low-rank implementation of the single-loop recursion. We first describe the bias of shared-projection compression used in existing randomized-subspace methods. We then introduce a past-only principal subspace with a random complement, and finally combine these components through independent input/output operator bi-sketching.

2.1 The Bottleneck of Shared-Projection Compression

Randomized subspace methods represent an ambient perturbation through a low-dimensional factor. Following BROS, write $\boldsymbol{P} = (P_1, \dots, P_L)$ for blockwise factors and $\boldsymbol{Q} = \boldsymbol{P}\boldsymbol{P}^\top$ for their lifts, with $Q_\ell = P_\ell P_\ell^\top$. Products with block variables are understood blockwise. The identity $\mathbb{E}[\boldsymbol{Q}] = I$ specifies the marginal mean of a one-sided lift. A Hessian action couples input and output lifts, so its expectation is governed by their joint moments. To isolate the issue, reuse one factor on both sides of the sample operator $\mathscr{H}^{k+1}$, independently of its oracle sample. The shared-projection estimator is

$$\widehat{\mathscr{H}}_{\mathrm{share}}[\boldsymbol{Z}^k] = \boldsymbol{Q}\mathscr{H}^{k+1}[\boldsymbol{Q}\boldsymbol{Z}^k], \tag{3a}$$

$$\mathbb{E}_k\left[\widehat{\mathscr{H}}_{\mathrm{share}}[\boldsymbol{Z}^k]\right] - \mathscr{H}_k[\boldsymbol{Z}^k] = \mathbb{E}_k\left[(\boldsymbol{Q} - I)\mathscr{H}_k[(\boldsymbol{Q} - I)\boldsymbol{Z}^k]\right]. \tag{3b}$$

The residual in (3b) characterizes the shared-projection bias through higher-order moments of the projection. BROS (Zhang et al., 2026) addresses this issue through fourth-moment identities for scaled Haar projections and a Rademacher bi-probe correction. The correction requires an additional subspace HVP and is tied to the specific projection distribution. Consequently, changing the projection law, rank configuration, or history-dependent factor construction requires a new correction derivation. BROS's fresh isotropic sampling maintains exploration across iterations. When lower-level or auxiliary directions exhibit persistent structure, combining historical principal directions with a fresh complement provides a natural history-adaptive extension. The resulting estimation error can then be characterized through operator-weighted residuals.

2.2 History-Aware Principal–Random Factors

For a matrix block $Y_\ell \in \mathbb{R}^{m_\ell \times n_\ell}$, BROS uses r_ℓ for the total projected row dimension. We retain that meaning: write $r_\ell = r_{h,\ell} + r_{f,\ell}$, where $r_{h,\ell}$ is the history rank and $r_{f,\ell}$ is the fresh complement rank.

Maintain an $\mathcal{F}_k$-measurable orthonormal history basis $U^k_{h,\ell} \in \mathbb{R}^{m_\ell \times r_{h,\ell}}$. The rank configuration is fixed within each run. The basis and sampling mixture may adapt over time. Below, we suppress ℓ and write m, r, r_h, r_f, U^k_h for one block, with $r = r_h + r_f$.

After the production directions and auxiliary update at iteration k, an Oja-type step (Oja, 1982) updates the basis from a historical signal X^{hist}_k:

$$B_k = X^{\mathrm{hist}}_k (X^{\mathrm{hist}}_k)^\top U^k_h, \qquad C_k = (I - U^k_h (U^k_h)^\top) B_k, \qquad U^{k+1}_h = \mathrm{qf}\big(U^k_h + \eta_{\mathrm{Oja},k} C_k\big), \quad (4)$$

where qf returns the orthonormal factor of a thin QR decomposition. The update is implemented through thin factor products and QR.

Our default signal is the updated auxiliary block $X^{\mathrm{hist}}_{\ell,k} = Z^{k+1}_\ell$. The resulting $U^{k+1}_{h,\ell}$ is used by the Hessian-input and mixed-Jacobian-input branches at iteration $k+1$. This post-production schedule keeps the current factor law fixed throughout the iteration's oracle evaluation.

Let $\Pi^k_h := U^k_h (U^k_h)^\top$ and $R^k_h := I - \Pi^k_h$, and write $c := m - r_h$. For $r_h < m$, require $1 \le r_f \le c$. Let $\lambda^k \in [0,1]$ be $\mathcal{F}_k$-measurable and draw $\iota^k \mid \mathcal{F}_k \sim \mathrm{Bernoulli}(\lambda^k)$. If $\iota^k = 1$, sample a Haar-uniform orthonormal r_f-frame $U^k_\perp$ in $\mathrm{range}(R^k_h)$ and set $\Delta^k = \sqrt{c/r_f}\, U^k_\perp$. If $\iota^k = 0$, sample a standard Gaussian $G^k \in \mathbb{R}^{m \times r_f}$ and set $\Delta^k = R^k_h G^k / \sqrt{r_f}$. Define $P^k := [\, U^k_h, \Delta^k \,] \in \mathbb{R}^{m \times r}$ and $Q^k := P^k (P^k)^\top$, where $r = r_h + r_f \le m$. The Haar construction can be implemented by standard orthogonal-matrix sampling procedures (Mezzadri, 2007). Its conditional moments are

$$\mathbb{E}_k[Q^k] = I, \qquad \mathbb{E}_k[(Q^k - I)^2] = \chi^k R^k_h, \qquad \chi^k = \lambda^k \left(\frac{c}{r_f} - 1 \right) + (1 - \lambda^k) \frac{c+1}{r_f}. \quad (5)$$

Thus $\mathbb{E}_k[(Q^k)^2] \preceq (1 + \chi^k) I$. The case $r_h = 0$ reduces to fully fresh sampling, while $r_h = m$ uses the fixed full-space lift $Q^k = I$. For $0 < r_h < m$, the history basis and a positive-rank fresh complement jointly define the factor. Across these cases, the first-moment identity remains $\mathbb{E}_k[Q^k] = I$, and history enters through the second moment.

2.3 Independent Operator Bi-Sketching

For the five lower-space roles $q \in \{g, u, i, o, j\}$, let $P^k_{q,\ell} \in \mathbb{R}^{m_\ell \times r_{q,\ell}}$ follow the principal–random factor construction above, where $r_{q,\ell} = r_{h,q,\ell} + r_{f,q,\ell}$. Write $Q^k_{q,\ell} = P^k_{q,\ell} (P^k_{q,\ell})^\top$, $\boldsymbol{P}^k_q = (P^k_{q,\ell})_{\ell=1}^L$, and $\boldsymbol{Q}^k_q = (Q^k_{q,\ell})_{\ell=1}^L$. Define the branch coordinate space $\mathcal{B}^k_q := \prod_{\ell=1}^L \mathbb{R}^{r_{q,\ell} \times n_\ell}$. The actions $\boldsymbol{P}^k_q \boldsymbol{A}$ and $(\boldsymbol{P}^k_q)^\top \boldsymbol{V}$ are blockwise for $\boldsymbol{A} \in \mathcal{B}^k_q$ and $\boldsymbol{V} \in \mathcal{Y}$. Factors may share a past-only history basis, while current Hessian-input and Hessian-output draws remain independent.

Our analyzed configuration uses one auxiliary history basis for $\boldsymbol{P}^k_i$ and $\boldsymbol{P}^k_j$, with $r_{h,i,\ell} = r_{h,j,\ell} = r_{h,\ell}$ and independent current complement draws. The other three roles are fully fresh: $r_{h,g,\ell} = r_{h,u,\ell} = r_{h,o,\ell} = 0$. The direct upper-gradient branch is unsketched: $P^k_x = Q^k_x = I_{d_x}$.

All factors are treated as constants during automatic differentiation. In the following coordinate gradients, each block coordinate lies in its branch space $\mathcal{B}^k_q$:

$$\begin{aligned} \widetilde{\boldsymbol{V}}^{k+1} &= \nabla_{\boldsymbol{A}} \ell_g(x^k, \boldsymbol{Y}^k + \boldsymbol{P}^k_g \boldsymbol{A}, \zeta_k)\big|_{\boldsymbol{A}=0}, \quad \widetilde{\boldsymbol{U}}^{k+1}_y = \nabla_{\boldsymbol{A}} \ell_f(x^k, \boldsymbol{Y}^k + \boldsymbol{P}^k_u \boldsymbol{A}, \xi_k)\big|_{\boldsymbol{A}=0}, \\ \widetilde{U}^{k+1}_x &= \nabla_{v_x} \ell_f(x^k + P^k_x v_x, \boldsymbol{Y}^k, \xi_k)\big|_{v_x=0}, \qquad v_x \in \mathbb{R}^{d_x}. \end{aligned} \quad (6)$$

For the Hessian branch take $\boldsymbol{A} \in \mathcal{B}^k_o$, $\boldsymbol{B}, \boldsymbol{V}_c \in \mathcal{B}^k_i$. For the mixed-Jacobian branch take $\boldsymbol{B}, \boldsymbol{V}_c \in \mathcal{B}^k_j$.

$$\begin{aligned} \widetilde{\mathscr{H}}^{k+1}[\boldsymbol{V}_c] &= D_{\boldsymbol{B}} \big[\nabla_{\boldsymbol{A}} \ell_g(x^k, \boldsymbol{Y}^k + \boldsymbol{P}^k_o \boldsymbol{A} + \boldsymbol{P}^k_i \boldsymbol{B}, \zeta_k) \big] [\boldsymbol{V}_c]\big|_{\boldsymbol{A}=\boldsymbol{B}=0} = (\boldsymbol{P}^k_o)^\top \mathscr{H}^{k+1}[\boldsymbol{P}^k_i \boldsymbol{V}_c], \\ \widetilde{\mathscr{J}}^{k+1}[\boldsymbol{V}_c] &= D_{\boldsymbol{B}} \big[\nabla_x \ell_g(x^k, \boldsymbol{Y}^k + \boldsymbol{P}^k_j \boldsymbol{B}, \zeta_k) \big] [\boldsymbol{V}_c]\big|_{\boldsymbol{B}=0} = \mathscr{J}^{k+1}[\boldsymbol{P}^k_j \boldsymbol{V}_c]. \end{aligned} \quad (7)$$

The two Hessian factors therefore act on different coordinate roles, whereas the mixed-Jacobian branch requires only an input factor. Since the derivatives are taken from the complete sample loss, cross-block and cross-layer derivatives remain included.

We next lift the coordinate derivatives into the three stochastic directions used by the single-loop recursion. Let $\boldsymbol{V}^{k+1} = \nabla_{\boldsymbol{Y}} \ell_g(x^k, \boldsymbol{Y}^k, \zeta_k)$, $\boldsymbol{U}_y^{k+1} = \nabla_{\boldsymbol{Y}} \ell_f(x^k, \boldsymbol{Y}^k, \xi_k)$, and $U_x^{k+1} = \nabla_x \ell_f(x^k, \boldsymbol{Y}^k, \xi_k)$. The lifted directions are

$$\widehat{\boldsymbol{V}}^{k+1} = \boldsymbol{P}_g^k \widetilde{\boldsymbol{V}}^{k+1} = \boldsymbol{Q}_g^k \boldsymbol{V}^{k+1}, \tag{8a}$$

$$\widehat{\boldsymbol{S}}^{k+1} = \boldsymbol{P}_o^k \widetilde{\mathscr{H}}^{k+1}[(\boldsymbol{P}_i^k)^\top \boldsymbol{Z}^k] - \boldsymbol{P}_u^k \widetilde{\boldsymbol{U}}_y^{k+1} = \boldsymbol{Q}_o^k \mathscr{H}^{k+1}[\boldsymbol{Q}_i^k \boldsymbol{Z}^k] - \boldsymbol{Q}_u^k \boldsymbol{U}_y^{k+1}, \tag{8b}$$

$$\widehat{W}^{k+1} = P_x^k \widetilde{U}_x^{k+1} - \widetilde{\mathscr{J}}^{k+1}[(\boldsymbol{P}_j^k)^\top \boldsymbol{Z}^k] = Q_x^k U_x^{k+1} - \mathscr{J}^{k+1}[\boldsymbol{Q}_j^k \boldsymbol{Z}^k]. \tag{8c}$$

Equations (8a)–(8c) give pathwise identities for the lifted implementation. Computation proceeds through the coordinate derivatives in (6) and (7), using a matrix-free factor representation. Section 3 establishes the stochastic-oracle conditions and conditional moment properties.

As in BROS, the full-space lower target at state k is $D_y^k := \nabla_{\boldsymbol{Y}} g(x^k, \boldsymbol{Y}^k)$. The auxiliary target is $D_z^k := \mathscr{H}_k[\boldsymbol{Z}^k] - \nabla_{\boldsymbol{Y}} f(x^k, \boldsymbol{Y}^k)$, and the upper target is $D_x^k := \nabla_x f(x^k, \boldsymbol{Y}^k) - \mathscr{J}_k[\boldsymbol{Z}^k]$.

Following the standard single-loop design of MA-SOBA and BROS (Chen et al., 2024, Zhang et al., 2026), we maintain a moving-average hypergradient state for the upper-level update. Algorithm 1 gives the production, state-update, and history-basis chronology.

The algorithm combines stochastic oracle evaluation, low-rank coordinate differentiation, independent input/output factors, and past-only history adaptation. The independent Hessian factors remove the need for a distribution-specific shared-projection correction in the production step, while the random complement preserves exploration outside the learned history subspace. Section 3 establishes the conditional moment properties, tracking bounds, and convergence guarantee of the resulting recursion.

3 Convergence Analysis

This section analyzes the convergence of Algorithm 1. We first state the basic assumptions on the objectives, stochastic derivative oracles, and randomized factors. We then establish two interface lemmas controlling the stochastic directions and the tracking errors of the single-loop states, followed by the main stationarity guarantee. Throughout, $\mathcal{F}_k$ denotes the information available before iteration k, including the current states, historical bases, rank configurations, and all previous randomness.

3.1 Assumptions

Assumption 3.1 (Global regularity)**.** *The functions f and g are twice continuously differentiable. There exist constants $\mu_g, L_{f,0}, L_{f,1}, L_{g,1}, L_{g,2} > 0$ such that, for every fixed x, $g(x,\cdot)$ is μ_g-strongly convex, ∇f and ∇g are respectively $L_{f,1}$- and $L_{g,1}$-Lipschitz, $\nabla^2 g$ is $L_{g,2}$-Lipschitz as a linear operator, $\|\nabla_{\boldsymbol{Y}} f(x, \boldsymbol{Y}^\star(x))\| \le L_{f,0}$, and $\Phi^\star := \inf_x \Phi(x) > -\infty$. Define $L_{\max} := \max\{L_{f,0}, L_{f,1}, L_{g,1}, L_{g,2}\}$ and $\kappa := L_{\max}/\mu_g$, as in BROS.*

These conditions ensure that the lower-level solution $\boldsymbol{Y}^\star(x)$ is unique, that the auxiliary solution $\boldsymbol{Z}^\star(x)$ is bounded, and that $\boldsymbol{Y}^\star(x)$, $\boldsymbol{Z}^\star(x)$, and $\nabla\Phi(x)$ have the required Lipschitz properties.

Assumption 3.2 (Stochastic derivative oracles)**.** *Let $\mathcal{F}_k^+ := \sigma(\mathcal{F}_k, \boldsymbol{P}_g^k, \boldsymbol{P}_u^k, \boldsymbol{P}_i^k, \boldsymbol{P}_o^k, \boldsymbol{P}_j^k)$ denote the information after drawing the current factors but before drawing the oracle tuple. Conditional on $\mathcal{F}_k^+$, the stochastic gradients $\boldsymbol{V}^{k+1}, \boldsymbol{U}_y^{k+1}, U_x^{k+1}$ are unbiased for D_y^k, $\nabla_{\boldsymbol{Y}} f(x^k, \boldsymbol{Y}^k)$, and $\nabla_x f(x^k, \boldsymbol{Y}^k)$, respectively, with uniformly bounded conditional variances. For every square-integrable $\mathcal{F}_k^+$-measurable input $\boldsymbol{W} \in \mathcal{Y}$,*

$$\mathbb{E}[\mathscr{H}^{k+1}[\boldsymbol{W}] \mid \mathcal{F}_k^+] = \mathscr{H}_k[\boldsymbol{W}], \qquad \mathbb{E}[\|\mathscr{H}^{k+1}[\boldsymbol{W}]\|^2 \mid \mathcal{F}_k^+] \le B_H \|\boldsymbol{W}\|^2, \tag{9}$$

$$\mathbb{E}[\mathscr{J}^{k+1}[\boldsymbol{W}] \mid \mathcal{F}_k^+] = \mathscr{J}_k[\boldsymbol{W}], \qquad \mathbb{E}[\|\mathscr{J}^{k+1}[\boldsymbol{W}]\|^2 \mid \mathcal{F}_k^+] \le B_J \|\boldsymbol{W}\|^2. \tag{10}$$

The conditional variances of the three stochastic gradients are bounded by finite constants $\sigma_g^2, \sigma_y^2, \sigma_x^2$. Different derivative oracles may use the same current data sample. This matches standard conditionally unbiased, bounded-second-moment bilevel oracles (Chen et al., 2024).

Assumption 3.3 (Predictable factors and initialization)**.** *The rank configuration is fixed within each run, while the history bases and mixture probabilities used at iteration k are $\mathcal{F}_k$-measurable.*

Algorithm 1 OPR-BROS

Input: $x^0, \boldsymbol{Y}^0, \boldsymbol{Z}^0, h^0$, initial history bases, rank configurations, stepsizes $\alpha_k, \beta_k, \gamma_k, \theta_k$, and history-signal rules.
for $k = 0, \dots, K-1$ **do**
  Given $\mathcal{F}_k$, use the fixed rank configuration and select the predictable history bases and Stiefel–Gaussian mechanisms.
  Sample $\boldsymbol{P}_g^k, \boldsymbol{P}_u^k, \boldsymbol{P}_i^k, \boldsymbol{P}_o^k, \boldsymbol{P}_j^k$ using the principal–random factor construction above.
  Set $P_x^k = Q_x^k = I_{d_x}$.
  Independently draw $\xi_k \sim \mathcal{D}_f$ and $\zeta_k \sim \mathcal{D}_g$.
  Compute the coordinate gradients $\widetilde{\boldsymbol{V}}^{k+1}, \widetilde{\boldsymbol{U}}_y^{k+1}, \widetilde{U}_x^{k+1}$ using (6).
  Compute $\widetilde{\mathscr{H}}^{k+1}[(\boldsymbol{P}_i^k)^\top \boldsymbol{Z}^k]$ and $\widetilde{\mathscr{J}}^{k+1}[(\boldsymbol{P}_j^k)^\top \boldsymbol{Z}^k]$ using (7).
  Form $\widehat{\boldsymbol{V}}^{k+1}, \widehat{\boldsymbol{S}}^{k+1}, \widehat{W}^{k+1}$ according to (8a)–(8c).
  Update the lower and auxiliary states: $\boldsymbol{Y}^{k+1} = \boldsymbol{Y}^k - \beta_k \widehat{\boldsymbol{V}}^{k+1}$ and $\boldsymbol{Z}^{k+1} = \boldsymbol{Z}^k - \gamma_k \widehat{\boldsymbol{S}}^{k+1}$.
  Set $x^{k+1} = x^k - \alpha_k h^k$ and $h^{k+1} = (1-\theta_k)h^k + \theta_k \widehat{W}^{k+1}$.
  Set $X_{\ell,k}^{\mathrm{hist}} = Z_\ell^{k+1}$ and update each nontrivial history basis using (4).
end for

Conditioned on $\mathcal{F}_k$, the current factors from different branches are mutually independent and independent of the current oracle tuple. For each $q \in \{g, u, i, o, j\}$, let $\omega_{P,q} < \infty$ be a deterministic bound, uniform over iterations for the chosen fixed rank configuration:

$$\mathbb{E}_k[\boldsymbol{Q}_q^k] = I, \qquad \mathbb{E}_k[(\boldsymbol{Q}_q^k)^\top \boldsymbol{Q}_q^k] \preceq \omega_{P,q} I, \qquad \omega_{\mathrm{OPR}} := \max\{1, \max_q \omega_{P,q}\}. \tag{11}$$

The initial objective gap and the expected squared errors of $\boldsymbol{Y}^0, \boldsymbol{Z}^0, h^0$ with respect to their corresponding targets are finite and independent of the rank configuration and the horizon.

Remark 3.1 (Role of Assumption 3.3). This assumption specifies predictability and conditional independence in the sampling protocol. The basis and mixture may use past data, while fresh input/output factors remain independent of each other and of the current oracle. Together with Assumption 3.2, this yields the conditional means in Lemma 3.1. Equation (5) verifies the factor moments. Their rank dependence is captured by ω_{OPR}, which can grow as the fresh rank shrinks.

3.2 Stochastic Directions and State Tracking

Following BROS, set $E_Y^k := \boldsymbol{Y}^k - \boldsymbol{Y}^\star(x^k)$, $E_Z^k := \boldsymbol{Z}^k - \boldsymbol{Z}^\star(x^k)$, and $e^k := h^k - \nabla\Phi(x^k)$.

Lemma 3.1 (Stochastic direction interface)**.** *The directions in* (8a)–(8c) *are conditionally unbiased for the full-space targets D_y^k, D_z^k, D_x^k defined above. There is a constant $C_{\mathrm{dir}} > 0$, independent of the ranks and horizon, such that*

$$\mathbb{E}_k[\widehat{\boldsymbol{V}}^{k+1}] = D_y^k, \qquad \mathbb{E}_k[\widehat{\boldsymbol{S}}^{k+1}] = D_z^k, \qquad \mathbb{E}_k[\widehat{W}^{k+1}] = D_x^k, \tag{12a}$$

and

$$\begin{aligned} &\mathbb{E}_k\|\widehat{\boldsymbol{V}}^{k+1} - D_y^k\|^2 + \mathbb{E}_k\|\widehat{\boldsymbol{S}}^{k+1} - D_z^k\|^2 + \mathbb{E}_k\|\widehat{W}^{k+1} - D_x^k\|^2 \\ &\le C_{\mathrm{dir}}\omega_{\mathrm{OPR}}^2\big(1 + \|E_Y^k\|^2 + \|E_Z^k\|^2\big). \end{aligned} \tag{12b}$$

The three equalities show that independent input and output factors preserve the conditional means of the stochastic directions. In particular, the doubly sketched Hessian action recovers $\mathscr{H}_k[\boldsymbol{Z}^k]$ in expectation. The variance bound shows that low-rank coordinate differentiation introduces only controlled stochastic error, whose magnitude is governed by the factor second moments and increases as the lower and auxiliary states move away from their targets.

Lemma 3.2 (Single-loop tracking interface)**.** *There exist constants $\tau, \bar{\alpha}, C_0, C_1 > 0$, depending only on the regularity, oracle, and initialization bounds, such that, for $\beta = \gamma = \theta = \tau\alpha$ and $0 < \alpha \le \bar{\alpha}/\omega_{\mathrm{OPR}}^2$,*

$$\alpha \sum_{k=0}^{K-1} \mathbb{E}\big[\|\nabla\Phi(x^k)\|^2 + \|E_Y^k\|^2 + \|E_Z^k\|^2 + \|e^k\|^2\big] \le C_0 + C_1 K \omega_{\mathrm{OPR}}^2 \alpha^2. \tag{13}$$

3.3 Main Convergence Guarantee

Theorem 3.1 (Stationarity convergence of OPR-BROS). *Under Assumptions 3.1–3.3, run Algorithm 1 with $Q_x^k = I_{d_x}$, $\alpha_k \equiv \alpha$, $\beta_k \equiv \beta$, $\gamma_k \equiv \gamma$, and $\theta_k \equiv \theta$, where $\beta = \gamma = \theta = \tau\alpha$. Let $\widehat{k}$ be sampled independently and uniformly from $\{0, \dots, K-1\}$. If $0 < \alpha \le \bar{\alpha}/\omega_{\mathrm{OPR}}^2$, then*

$$\mathbb{E}\Big[\|\nabla\Phi(x^{\widehat{k}})\|^2\Big] \le \frac{C_0}{K\alpha} + C_1\omega_{\mathrm{OPR}}^2\alpha. \tag{14}$$

Proof sketch. Lemma 3.1 gives conditionally unbiased directions and second-moment control. Substituting these bounds into the three tracking recursions and combining them with objective descent yields Lemma 3.2. Its state-dependent noise is absorbed when $\alpha\omega_{\mathrm{OPR}}^2$ is sufficiently small. Divide (13) by $K\alpha$ and sample $\widehat{k}$ uniformly to obtain (14). □

The two terms in (14) describe initial-error decay and stochastic error scaled by ω_{OPR}^2. The following corollary balances them through a stepsize satisfying the stability constraint.

Corollary 3.1 (Explicit stepsizes and iteration complexity). *In the setting of Theorem 3.1, let $\omega := \omega_{\mathrm{OPR}}$ and choose $\alpha = \min\{\bar{\alpha}/\omega^2, \sqrt{C_0/C_1}/(\omega\sqrt{K})\}$ and $\beta = \gamma = \theta = \tau\alpha$. If $K \ge C_0\omega^2/(C_1\bar{\alpha}^2)$, then*

$$\mathbb{E}\|\nabla\Phi(x^{\widehat{k}})\|^2 \le \frac{2\sqrt{C_0C_1}\,\omega}{\sqrt{K}}.$$

For uniformly bounded ω, $K = \mathcal{O}(\varepsilon^{-2})$ iterations suffice to ensure $\mathbb{E}\|\nabla\Phi(x^{\widehat{k}})\|^2 \le \varepsilon$.

Rank dependence. For full-space lifts, take $\omega_{\mathrm{OPR}} = 1$. For fully fresh Haar and Gaussian factors, one may take $\omega_{\mathrm{OPR}} = \max_{q,\ell} m_\ell/r_{q,\ell}$ and $\omega_{\mathrm{OPR}} = \max_{q,\ell}\{1+(m_\ell+1)/r_{q,\ell}\}$, respectively. Under the corollary's horizon condition, the $\mathcal{O}(\omega_{\mathrm{OPR}}/\sqrt{K})$ rate yields $\mathcal{O}(\omega_{\mathrm{OPR}}^2\varepsilon^{-2})$ iterations, preserving the ε^{-2} order at fixed rank ratios.

Role of history. Equation (5) shows that history preserves the first moment while confining stochastic second-moment error to the residual complement R_h^k. When the learned basis captures persistent directions, less signal is exposed to fresh sketching. Lemma 3.1 propagates this control to direction variance, preserving the $\mathcal{O}(K^{-1/2})$ convergence order in Theorem 3.1.

4 Memory and Computational Complexity

We compare peak memory and derivative-inclusive computation using the non-grouped-query-attention (non-GQA) decoder-block proxy of BROS (Zhang et al., 2026). The analysis identifies the conditions for a compute advantage at equal sketch width and a memory advantage through total-width reduction.

4.1 Peak Memory

Let n, s, b, h denote hidden size, sequence length, micro-batch size, and attention-head count. Let $r_{\mathrm{B}}, r_{\mathrm{O}}$ be the uniform BROS and OPR-BROS widths, and let r_h be the history rank. Set $A_0 = 28bsn/3 + 2bhs^2$.

Both methods retain full-space lower and auxiliary states. For OPR-BROS, r_{O} controls projected activations and factor workspaces, while r_h determines history storage. Sequential release makes the temporary peak the larger of the production and Oja workspaces.

Memory advantage through width reduction. For a fixed history fraction, reducing the total sketch width lowers these rank-dependent storage terms. The memory proxy favors OPR-BROS once this reduction offsets the additional storage associated with independent Hessian factors and the history basis. Appendix D.1 gives the exact crossover condition. History is intended to support this regime by retaining persistent directions within a smaller total sketch budget. When comparable optimization quality is maintained at that width, the crossover identifies a quality-matched advantage in the memory proxy.

Table 1: **Peak memory in scalar slots.** The OPR-BROS entry is a sequential-release upper bound; derivations appear in Appendix D.1.

Method	Peak-memory proxy
MA-SOBA	$48n^2 + 15bsn + 2bhs^2$
FdeHBO	$72n^2 + 15bsn + 2bhs^2$
Penalty	$24n^2 + 15bsn + 2bhs^2$
BROS	$24n^2 + A_0 + 4bsr_{\mathrm{B}} + 62nr_{\mathrm{B}}/3$
OPR-BROS	$\leq 24n^2 + 17nr_h/3 + \max\{A_0 + 8bsr_{\mathrm{O}} + 48nr_{\mathrm{O}},\ 16nr_h + 4r_h^2\}$

Table 2: **Decoder FLOPs per iteration.** Here $t = bs$, r is the common width, and $\geq$ denotes optimistic baseline lower bounds. Appendix D.2 specifies F_{Oja} and the schedules.

Method	FLOPs / iteration
MA-SOBA	$\geq 252tn^2 + 48tsn$
FdeHBO	$624tn^2 + 112tsn$
BROS	$\geq (694/3)tn^2 + 64tnr + 96tsn + 96rn^2$
OPR-BROS	$138tn^2 + 208tnr + 56tsn + 120rn^2 + F_{\mathrm{Oja}}$

Table 3: **Pile-280M quality and resources.** On the Pile-280M proxy derived from The Pile (Gao et al., 2020), after 10,000 steps on four A100s. Loss is fixed-panel worst-domain excess validation CE. Timing follows ten warmup steps; memory sums per-GPU training peaks, excluding evaluation and checkpoints. N/C: non-convergence.

Method	Upper-level loss	Step time (s)	Peak alloc. (GiB)
DoReMi	N/C	0.1913	47.28
MA-SOBA	**3.1290**	0.3992	75.97
Penalty	3.5187	0.8969	61.58
FdeHBO	3.1358	0.3918	71.49
ZOFO	3.3717	0.5048	57.81
B-NoProbe (archived)	3.1794	1.2051	56.30
O-Fresh (archived)	3.1825	**0.3510**	**56.04**

4.2 Per-Iteration Computation

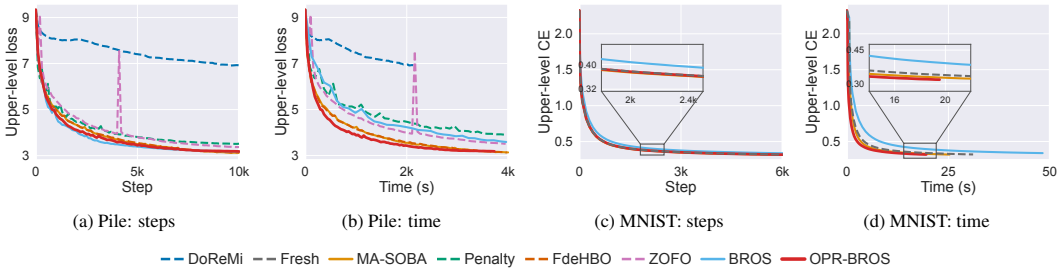
Table 2 counts HVP/JVP, projection, lift, and Oja arithmetic under dense tangent-over-reverse sweeps, with fixed decoder inputs and loss-weight upper variables.

Compute advantage at equal width. For $bs \geq r$, $1 \leq r \leq n/4$, and $0 \leq r_h \leq r/2$, OPR-BROS uses strictly fewer decoder FLOPs than the optimistic BROS bound at the same sketch width (Proposition D.1). The saving comes from replacing the masked/probe Hessian schedule with an independent bi-sketch action. Under these conditions, the reduction in dense second-order work outweighs the additional factor algebra and Oja update. The FLOP analysis uses cache retention, while the peak-memory bound uses sequential release; the respective advantages are characterized under these distinct execution schedules.

5 Experiments

We evaluate optimization quality, resource efficiency, and history reuse through data-mixture optimization, hyper-cleaning, and native-network tasks. Appendix E provides protocols, baseline definitions, and additional ablations.

Pile-280M data-mixture optimization. On the 17-domain Pile-280M (Gao et al., 2020) proxy, upper-level mixture weights guide lower-level language-model training. Figures 1a and 1b and Table 3 compare the archived configurations. O-Fresh and B-NoProbe reach upper-level losses of 3.1825 and 3.1794, respectively. Appendix E.1 specifies their sampling, correction, and execution settings.



(a) Pile: steps (b) Pile: time (c) MNIST: steps (d) MNIST: time

Figure 1: **Pile and MNIST trajectories.** Pile worst-domain excess validation CE versus steps (a) and time (b); MNIST validation CE versus steps (c) and time (d).

MNIST hyper-cleaning. MNIST (LeCun et al., 1998) hyper-cleaning learns sample weights from clean validation data while fitting a linear-softmax classifier to noisy labels. Across five seeds, OPR-BROS reaches validation CE 0.3182 after 6,000 steps, compared with 0.3173, 0.3192, and 0.3377 for Full, Fresh, and BROS. Their respective elapsed times are 19.55, 25.29, 31.09, and 48.22 seconds. Figures 1c and 1d show both trajectories. On frozen-feature Fashion-MNIST (Xiao et al., 2017), we measure *time-to-target* (TTT): TTT_{90} is the wall-clock time to attain 90% of each seed's Full-defined validation-CE improvement.

Table 4 reports five-seed OPR-BROS/comparator TTT_{90} ratios with 95% seed-clustered bootstrap intervals. Values below one favor OPR-BROS. Its ratio to Fresh is 0.4840. Appendix E.2 gives the identity-v4 setup, paired execution orders, and target interpolation.

Table 4: **Fashion-MNIST TTT90 ratios.**

Comparator	Ratio	95% Confidence Interval
Full	0.5250	$[0.5126, 0.5378]$
Fresh	0.4840	$[0.4777, 0.4889]$
BROS	0.3426	$[0.3328, 0.3542]$

Native CIFAR tasks. On CIFAR-10 (Krizhevsky & Hinton, 2009), ResNet-18 (He et al., 2016) uses upper-level feature gating. Noisy-label CIFAR-100 (Krizhevsky & Hinton, 2009) uses a small Vision Transformer (SmallViT) based on ViT (Dosovitskiy et al., 2021) with clean-validation sample reweighting. Across three seeds, OPR-BROS reaches 75.77% and 46.24% mean test accuracy versus Full's 75.90% and 46.47%. Appendix E.3 gives validation trajectories and resource measurements.

Qwen history ablation. On frozen Qwen3-like 0.6B (Yang et al., 2025a), learned and random cores use 16 core and 64 fresh residual directions. Across five paired seeds, History lowers the 200-step mean target meta-validation loss by 0.001246 ± 0.000184 (mean$\pm$sample standard deviation) at matched width. Appendices E.4 and E.5 give the ablation and fixed-state operator diagnostics.

6 Conclusion

OPR-BROS uses predictable principal–random factors and independent operator bi-sketching to reuse history in single-loop bilevel optimization. Under the stated assumptions, it preserves target conditional means and the $\mathcal{O}(K^{-1/2})$ stationarity rate. Risk bounds characterize historical structure, while experiments show small-task time-to-quality gains and native-network quality–resource trade-offs. Adaptive sketches for nonstationary training remain future work.

Reproducibility Statement

The main text specifies the estimator, assumptions, and theorems. The appendix gives full proofs and resource accounting. Section 5 and Appendix E give protocols, seeds, test-access policy, uncertainty, timing, and evaluation scope. The supplement contains run manifests, hashes, oracle ledgers, stopping records, and frozen scripts. Anonymous release instructions are pending.

AI USE STATEMENT

Generative AI tools assisted with manuscript organization, language editing, LaTeX formatting, and revising and checking parts of the convergence proof. Experimental measurements come from the reported runs. The authors are responsible for the mathematics, citations, and results.

REFERENCES

Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyxnZh0ct7.

Tianyi Chen, Yuejiao Sun, Quan Xiao, and Wotao Yin. A single-timescale method for stochastic bilevel optimization. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 2466–2488. PMLR, 2022. URL https://proceedings.mlr.press/v151/chen22e.html.

Xuxing Chen, Minhui Huang, Shiqian Ma, and Krishna Balasubramanian. Decentralized stochastic bilevel optimization with improved per-iteration complexity. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4641–4671. PMLR, 2023. URL https://proceedings.mlr.press/v202/chen23n.html.

Xuxing Chen, Tesi Xiao, and Krishnakumar Balasubramanian. Optimal algorithms for stochastic bilevel optimization under relaxed smoothness conditions. *Journal of Machine Learning Research*, 25(151):1–51, 2024. URL https://www.jmlr.org/papers/v25/23-1323.html.

Xuxing Chen, Minhui Huang, and Shiqian Ma. Decentralized bilevel optimization. *Optimization Letters*, 19(7):1249–1313, 2025a. doi: 10.1007/s11590-024-02101-4. URL https://link.springer.com/article/10.1007/s11590-024-02101-4.

Yiming Chen, Yuan Zhang, Yin Liu, Kun Yuan, and Zaiwen Wen. A memory efficient randomized subspace optimization method for training large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9747–9763. PMLR, 2025b. URL https://proceedings.mlr.press/v267/chen25cj.html.

Tianshu Chu, Dachuan Xu, Wei Yao, and Jin Zhang. SPABA: A single-loop and probabilistic stochastic bilevel algorithm achieving optimal sample complexity. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8848–8903. PMLR, 2024. URL https://proceedings.mlr.press/v235/chu24a.html.

Tianshu Chu, Dachuan Xu, Wei Yao, Chengming Yu, and Jin Zhang. A provably convergent plug-and-play framework for stochastic bilevel optimization. *Journal of Machine Learning Research*, 27(171):1–41, 2026. URL https://www.jmlr.org/papers/v27/25-0957.html.

Mathieu Dagréou, Pierre Ablin, Samuel Vaiter, and Thomas Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26698–26710. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/aa84ec1ac3f5fdcf77bce2c22705ab77-Abstract-Conference.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Simin Fan, Matteo Pagliardini, and Martin Jaggi. DOGE: Domain reweighting with generalization estimation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12895–12915. PMLR, 2024. URL https://proceedings.mlr.press/v235/fan24e.html.

Simin Fan, Maria Ios Glarou, and Martin Jaggi. GRAPE: Optimize data mixture for group robust multi-target adaptive pretraining. In *Advances in Neural Information Processing Systems*, volume 38, pp. 158023–158073. Curran Associates, Inc., 2025. doi: 10.52202/085713-5281. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/e7bba023351f8262c22e3328e42645ec-Abstract-Conference.html.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135. PMLR, 2017. URL https://proceedings.mlr.press/v70/finn17a.html.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1568–1577. PMLR, 2018. URL https://proceedings.mlr.press/v80/franceschi18a.html.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020. URL https://arxiv.org/abs/2101.00027.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018. doi: 10.48550/arXiv.1802.02246. URL https://arxiv.org/abs/1802.02246.

Robert Gower, Dmitry Kovalev, Felix Lieder, and Peter Richtárik. RSN: Randomized subspace newton. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/bc6dc48b743dc5d013b1abaebd2faed2-Abstract.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations (ICLR 2022)*. OpenReview.net, 2022. URL https://arxiv.org/abs/2106.09685.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4882–4892. PMLR, 2021. URL https://proceedings.mlr.press/v139/ji21c.html.

Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. In *Advances in Neural Information Processing Systems*, volume 34, pp. 30271–30283. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/fe2b421b8b5f0e7c355ace66a9fe0206-Abstract.html.

Boao Kong, Shuchen Zhu, Songtao Lu, Xinmeng Huang, and Kun Yuan. Decentralized bilevel optimization: A perspective from transient iteration complexity. *Journal of Machine Learning Research*, 26(240):1–64, 2025. URL https://www.jmlr.org/papers/v26/25-0677.html.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D Nowak. A fully first-order method for stochastic bilevel optimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18083–18113. PMLR, 2023. URL https://proceedings.mlr.press/v202/kwon23c.html.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL https://doi.org/10.1109/5.726791.

Yanyi Li, Yimu Zhang, and Cong Fang. PRAC: Principal-random subspace for LLM activation compression and memory-efficient training. *arXiv preprint arXiv:2602.23111*, 2026. doi: 10.48550/arXiv.2602.23111. URL https://arxiv.org/abs/2602.23111.

Zhuqing Liu, Xin Zhang, Prashant Khanduri, Songtao Lu, and Jia Liu. Prometheus: taming sample and communication complexities in constrained decentralized stochastic bilevel learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22420–22453. PMLR, 2023. URL https://proceedings.mlr.press/v202/liu23az.html.

Francesco Mezzadri. How to generate random matrices from the classical compact groups. *Notices of the American Mathematical Society*, 54(5):592–604, 2007. URL https://www.ams.org/notices/200705/fea-mezzadri-web.pdf.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273, 1982. doi: 10.1007/BF00275687. URL https://link.springer.com/article/10.1007/BF00275687.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 737–746. PMLR, 2016. URL https://proceedings.mlr.press/v48/pedregosa16.html.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4334–4343. PMLR, 2018. URL https://proceedings.mlr.press/v80/ren18a.html.

Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1723–1732. PMLR, 2019. URL https://proceedings.mlr.press/v89/shaban19a.html.

Han Shen and Tianyi Chen. On penalty-based bilevel gradient descent method. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 30992–31015. PMLR, 2023. URL https://proceedings.mlr.press/v202/shen23c.html.

Haoru Tan, Sitong Wu, Yanfeng Chen, Jun Xia, Ruobing Xie, Bin Xia, Samm Sun, and Xiaojuan Qi. Fast data mixture optimization via gradient descent. In *International Conference on Learning Representations*, volume 2026, pp. 702–715, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/01d845baaa31b1043e5f280592d57dc4-Abstract-Conference.html.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms, 2017. URL https://arxiv.org/abs/1708.07747.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. DoReMi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems*, volume 36, pp. 69798–69818. Curran Associates, Inc., 2023. doi: 10.52202/075280-3059. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/dcba6be91359358c2355cd920da3fcbd-Abstract-Conference.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

Yan Yang, Bin Gao, and Ya-xiang Yuan. LancBiO: Dynamic lanczos-aided bilevel optimization via krylov subspace. In *International Conference on Learning Representations*, volume 2025, pp. 34140–34178, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/54ba04e78dbd019300f6731bdd46ca16-Paper-Conference.pdf.

Yifan Yang, Peiyao Xiao, and Kaiyi Ji. Achieving $\mathcal{O}(\epsilon^{-1.5})$ complexity in hessian/jacobian-free stochastic bilevel optimization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 39491–39503. Curran Associates, Inc., 2023. doi: 10.52202/075280-1715. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/7c3a8d20ceadb7c519e9ac1bb77a15ff-Abstract-Conference.html.

Emanuele Zangrando, Sara Venturini, Francesco Rinaldi, and Francesco Tudisco. dEBORA: Efficient bilevel optimization-based low-rank adaptation. In *International Conference on Learning Representations*, volume 2025, pp. 4837–4856, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/0f0c4f3d83c58df58380af3b0729354c-Paper-Conference.pdf.

Hengrui Zhang, Boao Kong, Engao Zhang, and Kun Yuan. BROS: Bias-corrected randomized subspaces for memory-efficient single-loop bilevel optimization. *arXiv preprint arXiv:2605.10288*, 2026. doi: 10.48550/arXiv.2605.10288. URL https://arxiv.org/abs/2605.10288.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61121–61143. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhao24s.html.

Guoqing Zheng, Ahmed Hassan Awadallah, and Susan Dumais. Meta label correction for noisy label learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11053–11061, 2021. doi: 10.1609/aaai.v35i12.17319. URL https://ojs.aaai.org/index.php/AAAI/article/view/17319.

Shuchen Zhu, Boao Kong, Songtao Lu, Xinmeng Huang, and Kun Yuan. SPARKLE: A unified single-loop primal-dual framework for decentralized bilevel optimization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 62912–62987. Curran Associates, Inc., 2024. doi: 10.52202/079017-2010. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/72f9c316440c384a95c88022fd78f066-Abstract-Conference.html.

Appendix

A Related Work

Stochastic Bilevel Optimization. Bilevel optimization is widely used in hyperparameter optimization, meta-learning, data reweighting, and representation learning. Early stochastic methods estimate the hypergradient with double-loop updates, truncated backpropagation, approximate implicit differentiation, or two-timescale stochastic approximation (Ghadimi & Wang, 2018, Ji et al., 2021, Pedregosa, 2016). Single-loop methods update the lower variable, an auxiliary linear-system variable, and the upper variable in one recursion. SUSTAIN uses double-momentum estimators (Khanduri et al., 2021). SOBA and SABA organize the lower and auxiliary updates on a common stochastic timescale, with SABA adding global variance reduction (Dagréou et al., 2022). MA-SOBA adds a moving-average hypergradient to obtain an $\mathcal{O}(\varepsilon^{-2})$ sample complexity under standard smoothness conditions (Chen et al., 2024). SPABA adapts PAGE to bilevel recursions, while PnPBO provides a plug-and-play analysis for PAGE, ZeroSARAH, and mixed estimators (Chu et al., 2024, 2026). In multi-node settings, decentralized stochastic bilevel methods use consensus or gradient tracking to coordinate local states under data heterogeneity and communication constraints (Chen et al., 2025a, 2023, Liu et al., 2023, Zhu et al., 2024, Kong et al., 2025). These methods primarily improve oracle, sample, or communication complexity, while generally retaining ambient-dimensional lower and auxiliary states. Our work is complementary: it keeps the single-loop bilevel recursion and changes how its derivative operators and direction states are represented.

Memory-Efficient Hypergradient Estimation. Memory-efficient hypergradient methods follow several approximation routes. Penalty methods replace the lower-level constraint with a finite-penalty surrogate (Shen & Chen, 2023). Fully first-order methods such as F^2SA avoid Hessian and Jacobian oracles through first-order stochastic approximation (Kwon et al., 2023). Finite-difference methods such as FdeHBO approximate Hessian–vector and Jacobian–vector products with additional gradient queries and step-size conditions (Yang et al., 2023). These approaches offer different trade-offs among second-order derivative cost, approximation structure, and regularity requirements.

Subspace-based methods reduce working dimension by restricting optimization or derivative computations to lower-dimensional representations. RSN (Gower et al., 2019) uses randomized sketches for Newton directions. RSO (Chen et al., 2025b) and GaLore (Zhao et al., 2024) use randomized or low-rank projections for neural-network optimization. LancBiO (Yang et al., 2025b) constructs dynamic Krylov subspaces for Hessian-inverse-vector products in bilevel optimization. BROS (Zhang et al., 2026) applies randomized subspaces to lower and auxiliary updates with a Rademacher bi-probe for shared-projection bias correction. Related principal–random compression methods such as PRAC (Li et al., 2026) retain a principal subspace and sample a random complement to obtain an unbiased, low-variance reconstruction. Low-rank bilevel applications such as dEBORA (Zangrando et al., 2025) address layerwise rank allocation for parameter-efficient adaptation. Our work combines these compression motivations with conditionally independent input/output operator estimation and a past-only history-aware basis for stochastic bilevel recursion.

Large-Scale Bilevel Applications. Large-scale bilevel benchmarks include data hyper-cleaning and sample reweighting (Shaban et al., 2019, Ren et al., 2018, Zheng et al., 2021), hyper-representation learning (Franceschi et al., 2018, Bertinetto et al., 2019), and data-mixture optimization (Xie et al., 2023, Fan et al., 2024, 2025, Tan et al., 2026). They expose the memory cost of multilayer lower and auxiliary states. Our work targets this shared structure by keeping the standard single-loop recursion, compressing operator actions in both directions, and updating a principal-random subspace from past iterates.

B Complete Proof of Theorem 3.1

Assumptions 3.1–3.3 are stated in the main text. Auxiliary notation is defined where it is first needed below.

B.1 Proof of Lemma 3.1

We first identify the lifted coordinate derivatives and verify that their conditional means remain correct across coupled matrix blocks.

Lemma B.1 (Frozen-coordinate and product-space identities)**.** *For fixed factors, the lifted Hessian and mixed-Jacobian actions are $\boldsymbol{Q}_o^k \mathscr{H}^{k+1}[\boldsymbol{Q}_i^k \boldsymbol{Z}]$ and $\mathscr{J}^{k+1}[\boldsymbol{Q}_j^k \boldsymbol{Z}]$. For every $\mathcal{F}_k$-measurable $\boldsymbol{Z} \in \mathcal{Y}$,*

$$\mathbb{E}_k[\boldsymbol{Q}_o^k \mathscr{H}^{k+1}[\boldsymbol{Q}_i^k \boldsymbol{Z}]] = \mathscr{H}_k[\boldsymbol{Z}], \qquad \mathbb{E}_k[\mathscr{J}^{k+1}[\boldsymbol{Q}_j^k \boldsymbol{Z}]] = \mathscr{J}_k[\boldsymbol{Z}].$$

The factor second-moment bounds in Assumption 3.3 hold for the product Frobenius norm.

Proof. Fix a sample and regard every factor as constant under differentiation. For $\varphi(\boldsymbol{A}, \boldsymbol{B}) = \ell_g(x^k, \boldsymbol{Y}^k + \boldsymbol{P}_o^k \boldsymbol{A} + \boldsymbol{P}_i^k \boldsymbol{B}, \zeta_k)$, the chain rule gives

$$\begin{aligned} \nabla_{\boldsymbol{A}} \varphi &= (\boldsymbol{P}_o^k)^\top \nabla_{\boldsymbol{Y}} \ell_g(x^k, \boldsymbol{Y}^k + \boldsymbol{P}_o^k \boldsymbol{A} + \boldsymbol{P}_i^k \boldsymbol{B}, \zeta_k), \\ D_{\boldsymbol{B}}[\nabla_{\boldsymbol{A}} \varphi][\boldsymbol{V}_c] &= (\boldsymbol{P}_o^k)^\top \mathscr{H}^{k+1}[\boldsymbol{P}_i^k \boldsymbol{V}_c] \quad \text{at } \boldsymbol{A} = \boldsymbol{B} = 0. \end{aligned} \tag{15}$$

Take $\boldsymbol{V}_c = (\boldsymbol{P}_i^k)^\top \boldsymbol{Z}^k$ and lift by $\boldsymbol{P}_o^k$:

$$\boldsymbol{P}_o^k D_{\boldsymbol{B}}[\nabla_{\boldsymbol{A}} \varphi(0,0)][(\boldsymbol{P}_i^k)^\top \boldsymbol{Z}^k] = \boldsymbol{Q}_o^k \mathscr{H}^{k+1}[\boldsymbol{Q}_i^k \boldsymbol{Z}^k].$$

Similarly, differentiating the mixed branch gives

$$\begin{aligned} D_{\boldsymbol{B}}[\nabla_x \ell_g(x^k, \boldsymbol{Y}^k + \boldsymbol{P}_j^k \boldsymbol{B}, \zeta_k)]_{\boldsymbol{B}=0}[(\boldsymbol{P}_j^k)^\top \boldsymbol{Z}^k] &= \mathscr{J}^{k+1}[\boldsymbol{Q}_j^k \boldsymbol{Z}^k], \\ \boldsymbol{P}_g^k \nabla_{\boldsymbol{A}} \ell_g(x^k, \boldsymbol{Y}^k + \boldsymbol{P}_g^k \boldsymbol{A}, \zeta_k)\big|_{\boldsymbol{A}=0} &= \boldsymbol{Q}_g^k \boldsymbol{V}^{k+1}. \end{aligned}$$

The upper-y and upper-x gradient branches follow from the same affine chain rule. Reusing one factor would instead yield $\boldsymbol{Q} \mathscr{H}^{k+1}[\boldsymbol{Q} \boldsymbol{Z}^k]$. The mixed coordinates in (15) permit independent input and output factors.

For a matrix block $Y_\ell \in \mathbb{R}^{m_\ell \times n_\ell}$, use the Frobenius inner product and the coordinate map $A_\ell \mapsto Y_\ell + P_{q,\ell}^k A_\ell$. Column vectorization gives $\operatorname{vec}(P_{q,\ell}^k A_\ell) = (I_{n_\ell} \otimes P_{q,\ell}^k) \operatorname{vec}(A_\ell)$, so the preceding calculation also applies to matrix variables.

Consider the product space $\mathcal{Y} = \prod_{\ell=1}^L \mathbb{R}^{m_\ell \times n_\ell}$ and recall that $(\boldsymbol{Q}_q^k \boldsymbol{Z})_\ell = Q_{q,\ell}^k Z_\ell$. Let $\mathscr{H}_{\ell t}^{k+1}$ denote the sampled Hessian block from input layer t to output layer ℓ. For an $\mathcal{F}_k$-measurable $\boldsymbol{Z} \in \mathcal{Y}$, conditional independence of the complete input collection, output collection, and oracle tuple gives

$$\begin{aligned} \mathbb{E}_k[(\boldsymbol{Q}_o^k \mathscr{H}^{k+1}[\boldsymbol{Q}_i^k \boldsymbol{Z}])_\ell] &= \sum_{t=1}^L \mathbb{E}_k[Q_{o,\ell}^k \mathscr{H}_{\ell t}^{k+1}[Q_{i,t}^k Z_t]] \\ &= \sum_{t=1}^L \mathscr{H}_{\ell t}(x^k, \boldsymbol{Y}^k)[Z_t] = (\mathscr{H}_k[\boldsymbol{Z}])_\ell. \end{aligned} \tag{16}$$

Each summand follows by averaging the output factor, the operator sample, and the input factor in that order. This calculation includes the self-layer block $t = \ell$ because its two factor roles are independent. The same calculation without the output factor proves $\mathbb{E}_k[\mathscr{J}^{k+1}[\boldsymbol{Q}_j^k \boldsymbol{Z}]] = \mathscr{J}_k[\boldsymbol{Z}]$.

For the product norm, the block moment bounds imply

$$\mathbb{E}_k \left\| \boldsymbol{Q}_q^k \boldsymbol{Z} \right\|_{\mathcal{Y}}^2 = \sum_{\ell=1}^L \mathbb{E}_k \left\| Q_{q,\ell}^k Z_\ell \right\|_F^2 \le \omega_{P,q} \sum_{\ell=1}^L \|Z_\ell\|_F^2 = \omega_{P,q} \|\boldsymbol{Z}\|_{\mathcal{Y}}^2 .$$

This step permits dependence across layers within each factor collection. □

The principal–random factor preserves the mean while confining its error to the complement of the predictable history space.

Lemma B.2 (Principal–random factor moments)**.** *Condition on $\mathcal{F}_k$ and consider one row block with $r_h < m$. Suppress the iteration superscript and write $\Pi_h = U_h U_h^\top$, $R_h = I - \Pi_h$, $c = m - r_h$, and $a = r_f \ge 1$. If the Stiefel law is selected with predictable probability λ, then*

$$\mathbb{E}_k[Q] = I, \qquad \mathbb{E}_k[(Q - I)^2] = \chi_Q R_h, \qquad \mathbb{E}_k[Q^2] = \Pi_h + (1 + \chi_Q) R_h,$$

where $\chi_Q = \lambda(c/a - 1) + (1 - \lambda)(c + 1)/a$. In particular, $\mathbb{E}_k \|(Q - I)V\|_F^2 = \chi_Q \|R_h V\|_F^2$ for fixed V. For the separate case $r_h = m$, the convention $Q = I$ gives zero sketch error and $\chi_Q = 0$.

Proof. Write $\mathbb{E}_{\mathrm{St},k}$ and $\mathbb{E}_{\mathrm{G},k}$ for expectations under the two component laws conditional on $\mathcal{F}_k$. Both component laws are defined even if their mixture weight is zero. For the Stiefel complement, let $\Pi_\perp = U_\perp U_\perp^\top$. Since $U_\perp$ is a Haar a-frame in a c-dimensional complement, rotational invariance implies $\mathbb{E}_{\mathrm{St},k}[\Pi_\perp] = tR_h$ for a scalar t. Taking traces gives $a = tc$, hence

$$\mathbb{E}_{\mathrm{St},k}[Q] = \Pi_h + \frac{c}{a}\mathbb{E}_{\mathrm{St},k}[\Pi_\perp] = \Pi_h + R_h = I.$$

The projector identities $\Pi_h\Pi_\perp = 0$, $R_h\Pi_\perp = \Pi_\perp$, and $\Pi_\perp^2 = \Pi_\perp$ yield

$$(Q-I)^2 = \left(\frac{c}{a}\Pi_\perp - R_h\right)^2 = \left(\frac{c^2}{a^2} - \frac{2c}{a}\right)\Pi_\perp + R_h,$$
$$\mathbb{E}_{\mathrm{St},k}[(Q-I)^2] = \left(\frac{c}{a} - 1\right)R_h.$$

Since the conditional first moment is I, $\mathbb{E}_{\mathrm{St},k}[Q^2] = \Pi_h + (c/a)R_h$. The Stiefel conditional second-moment coefficient is c/a.

For the Gaussian complement, choose an orthonormal basis $B_c \in \mathbb{R}^{m\times c}$ of $\mathrm{range}(R_h)$. Write $R_hG = B_c\Gamma$, where $\Gamma \in \mathbb{R}^{c\times a}$ has independent standard Gaussian entries. With columns $\gamma_1, \dots, \gamma_a$, define $S_G := a^{-1}\sum_{s=1}^a \gamma_s\gamma_s^\top$ and $Q := \Pi_h + B_cS_GB_c^\top$. The first moment is $\mathbb{E}_{\mathrm{G},k}[S_G] = I_c$. For $s \neq t$, independence gives $\mathbb{E}[\gamma_s\gamma_s^\top\gamma_t\gamma_t^\top] = I_c$. For $s = t$, Gaussian coordinate moments give $\mathbb{E}[\gamma_s\gamma_s^\top\gamma_s\gamma_s^\top] = \mathbb{E}[\|\gamma_s\|^2\gamma_s\gamma_s^\top] = (c+2)I_c$. The diagonal entries are $3 + (c-1) = c+2$, and symmetry makes the off-diagonal entries zero. Therefore

$$\mathbb{E}_{\mathrm{G},k}[S_G^2] = \frac{a(c+2)+a(a-1)}{a^2}I_c = \left(1 + \frac{c+1}{a}\right)I_c,$$
$$\mathbb{E}_{\mathrm{G},k}[(S_G - I_c)^2] = \frac{c+1}{a}I_c.$$

Conjugating by B_c gives $\mathbb{E}_{\mathrm{G},k}[Q] = I$ and $\mathbb{E}_{\mathrm{G},k}[(Q-I)^2] = ((c+1)/a)R_h$. It also gives $\mathbb{E}_{\mathrm{G},k}[Q^2] = \Pi_h + (1+(c+1)/a)R_h$. Average the two laws over the Bernoulli draw. For any fixed compatible matrix V,

$$\mathbb{E}_k\|(Q-I)V\|_F^2 = \mathrm{tr}\big(V^\top\mathbb{E}_k[(Q-I)^\top(Q-I)]V\big) = \chi_Q\|R_hV\|_F^2. \tag{17}$$

If the history spans the whole space, the convention $Q = I$ gives zero sketch error directly. □

Independence of the input and output factors is what recovers an unbiased Hessian action. The next identity also separates their mean-square errors.

Lemma B.3 (Independent bi-sketch identity)**.** *Let $\mathscr{A} : \mathcal{Y} \to \mathcal{Y}$ be linear and let $\boldsymbol{T} \in \mathcal{Y}$. Suppose both are measurable with respect to $\mathcal{G}$, while the mean-identity block lifts $\boldsymbol{Q}_i, \boldsymbol{Q}_o$ are conditionally independent given $\mathcal{G}$. Write $\mathscr{K}_o := \mathbb{E}[(\boldsymbol{Q}_o - I)^\top(\boldsymbol{Q}_o - I) \mid \mathcal{G}]$ and $\boldsymbol{E}_i := (\boldsymbol{Q}_i - I)\boldsymbol{T}$. Then $\mathbb{E}[\boldsymbol{Q}_o\mathscr{A}[\boldsymbol{Q}_i\boldsymbol{T}] \mid \mathcal{G}] = \mathscr{A}[\boldsymbol{T}]$, with exact conditional MSE given by* (20)*. The stochastic-operator extension satisfies* (21)*.*

Proof. Condition throughout on $\mathcal{G}$. By conditional independence,

$$\begin{aligned}\mathbb{E}[\boldsymbol{Q}_o\mathscr{A}[\boldsymbol{Q}_i\boldsymbol{T}] \mid \mathcal{G}] &= \mathbb{E}[\mathbb{E}[\boldsymbol{Q}_o\mathscr{A}[\boldsymbol{Q}_i\boldsymbol{T}] \mid \mathcal{G}, \boldsymbol{Q}_i] \mid \mathcal{G}] \\ &= \mathbb{E}[\mathscr{A}[\boldsymbol{Q}_i\boldsymbol{T}] \mid \mathcal{G}] = \mathscr{A}[\boldsymbol{T}].\end{aligned} \tag{18}$$

This proves (18).

Next average the output error. Set $\boldsymbol{T}_i = \boldsymbol{Q}_i\boldsymbol{T} = \boldsymbol{T} + \boldsymbol{E}_i$. Conditional on $(\mathcal{G}, \boldsymbol{Q}_i)$,

$$\boldsymbol{Q}_o\mathscr{A}[\boldsymbol{T}_i] - \mathscr{A}[\boldsymbol{T}] = (\boldsymbol{Q}_o - I)\mathscr{A}[\boldsymbol{T}_i] + \mathscr{A}[\boldsymbol{E}_i].$$

The second summand is measurable under this conditioning, and the first is centered. Thus

$$\mathbb{E}[\|\boldsymbol{Q}_o\mathscr{A}[\boldsymbol{T}_i] - \mathscr{A}[\boldsymbol{T}]\|^2 \mid \mathcal{G}, \boldsymbol{Q}_i] = \langle\mathscr{A}[\boldsymbol{T}_i], \mathscr{K}_o\mathscr{A}[\boldsymbol{T}_i]\rangle + \|\mathscr{A}[\boldsymbol{E}_i]\|^2. \tag{19}$$

Expand $\boldsymbol{T}_i = \boldsymbol{T} + \boldsymbol{E}_i$ in the product inner product. Because $\mathbb{E}[\boldsymbol{E}_i \mid \mathcal{G}] = 0$, the mixed term vanishes after averaging. Consequently,

$$\mathbb{E}[\|\boldsymbol{Q}_o\mathscr{A}[\boldsymbol{Q}_i\boldsymbol{T}] - \mathscr{A}[\boldsymbol{T}]\|^2 \mid \mathcal{G}] = \langle \mathscr{A}[\boldsymbol{T}], \mathscr{K}_o\mathscr{A}[\boldsymbol{T}]\rangle + \mathbb{E}[\|\mathscr{A}[\boldsymbol{E}_i]\|^2 \mid \mathcal{G}] + \mathbb{E}[\langle \mathscr{A}[\boldsymbol{E}_i], \mathscr{K}_o\mathscr{A}[\boldsymbol{E}_i]\rangle \mid \mathcal{G}]. \tag{20}$$

Setting $\boldsymbol{Q}_o = I$ gives $\mathscr{K}_o = 0$ and leaves only the input-sketch term.

For a conditionally unbiased stochastic operator $\widetilde{\mathscr{A}}$ independent of the factors, first condition on $(\mathcal{G}, \widetilde{\mathscr{A}})$. The preceding mean calculation gives $\mathbb{E}[\boldsymbol{Q}_o\widetilde{\mathscr{A}}[\boldsymbol{Q}_i\boldsymbol{T}] - \widetilde{\mathscr{A}}[\boldsymbol{T}] \mid \mathcal{G}, \widetilde{\mathscr{A}}] = 0$. Hence the decomposition

$$\boldsymbol{Q}_o\widetilde{\mathscr{A}}[\boldsymbol{Q}_i\boldsymbol{T}] - \mathscr{A}[\boldsymbol{T}] = (\boldsymbol{Q}_o\widetilde{\mathscr{A}}[\boldsymbol{Q}_i\boldsymbol{T}] - \widetilde{\mathscr{A}}[\boldsymbol{T}]) + (\widetilde{\mathscr{A}} - \mathscr{A})[\boldsymbol{T}]$$

has zero expected cross term, and

$$\mathbb{E}[\left\|\boldsymbol{Q}_o\widetilde{\mathscr{A}}[\boldsymbol{Q}_i\boldsymbol{T}] - \mathscr{A}[\boldsymbol{T}]\right\|^2 \mid \mathcal{G}] = \mathbb{E}[\left\|(\widetilde{\mathscr{A}} - \mathscr{A})[\boldsymbol{T}]\right\|^2 \mid \mathcal{G}] + \mathbb{E}\left[\mathbb{E}[\left\|\boldsymbol{Q}_o\widetilde{\mathscr{A}}[\boldsymbol{Q}_i\boldsymbol{T}] - \widetilde{\mathscr{A}}[\boldsymbol{T}]\right\|^2 \mid \mathcal{G}, \widetilde{\mathscr{A}}] \mid \mathcal{G}\right]. \tag{21}$$

The inner expectation uses the three-term formula with $\mathscr{A} = \widetilde{\mathscr{A}}$. □

The direction risks split into intrinsic oracle noise and centered sketch errors, including when oracle components share a sample. Fix k. Let $\mathcal{G}_k := \sigma(\mathcal{F}_k, \boldsymbol{V}^{k+1}, \boldsymbol{U}_y^{k+1}, U_x^{k+1}, \mathscr{H}^{k+1}, \mathscr{J}^{k+1})$ contain the revealed oracle tuple but not the production factors. The full-space targets D_y^k, D_z^k, D_x^k are defined in Section 2.3. Define $\nu_{V,k} := \mathbb{E}_k\left\|\widehat{\boldsymbol{V}}^{k+1} - D_y^k\right\|^2$, $\nu_{S,k} := \mathbb{E}_k\left\|\widehat{\boldsymbol{S}}^{k+1} - D_z^k\right\|^2$, and $\nu_{W,k} := \mathbb{E}_k\left\|\widehat{W}^{k+1} - D_x^k\right\|^2$.

The intrinsic oracle risks are $B_{V,k} := \mathbb{E}_k\left\|\boldsymbol{V}^{k+1} - D_y^k\right\|^2$, $B_{S,k} := \mathbb{E}_k\left\|\mathscr{H}^{k+1}[\boldsymbol{Z}^k] - \boldsymbol{U}_y^{k+1} - D_z^k\right\|^2$, and $B_{W,k} := \mathbb{E}_k\left\|U_x^{k+1} - \mathscr{J}^{k+1}[\boldsymbol{Z}^k] - D_x^k\right\|^2$. The six nonnegative sketch risks are $T_{g,k} := \mathbb{E}_k\left\|(\boldsymbol{Q}_g^k - I)\boldsymbol{V}^{k+1}\right\|^2$ and $T_{u,k} := \mathbb{E}_k\left\|(\boldsymbol{Q}_u^k - I)\boldsymbol{U}_y^{k+1}\right\|^2$ for the first-order branches, $T_{o,k} := \mathbb{E}_k\left\|(\boldsymbol{Q}_o^k - I)\mathscr{H}^{k+1}[\boldsymbol{Q}_i^k\boldsymbol{Z}^k]\right\|^2$ and $T_{i,k} := \mathbb{E}_k\left\|\mathscr{H}^{k+1}[(\boldsymbol{Q}_i^k - I)\boldsymbol{Z}^k]\right\|^2$ for the Hessian branch, and $T_{x,k} := \mathbb{E}_k\left\|(Q_x^k - I)U_x^{k+1}\right\|^2$ and $T_{j,k} := \mathbb{E}_k\left\|\mathscr{J}^{k+1}[(\boldsymbol{Q}_j^k - I)\boldsymbol{Z}^k]\right\|^2$ for the upper branch. Here $T_{o,k}$ includes the input/output interaction term from Lemma 3.1. We include that term once, through the random input $\boldsymbol{Q}_i^k\boldsymbol{Z}^k$ in the output branch.

The exact split identifies which parts can respond to history and which are intrinsic to the oracle.

Lemma B.4 (Exact branchwise variance decomposition)**.** *Under Assumptions 3.2 and 3.3, including conditional independence across factor branches, and the mean-identity factor contract,*

$$\begin{aligned}\nu_{V,k} &= B_{V,k} + T_{g,k},\\ \nu_{S,k} &= B_{S,k} + T_{u,k} + T_{o,k} + T_{i,k},\\ \nu_{W,k} &= B_{W,k} + T_{x,k} + T_{j,k}.\end{aligned} \tag{22}$$

The conditional variances $\nu_{V,k}, \nu_{S,k}, \nu_{W,k}$ measure errors relative to the three population directions in Lemma 3.1.

Proof. For the auxiliary direction, write its error as

$$\widehat{\boldsymbol{S}}^{k+1} - D_z^k = \underbrace{\mathscr{H}^{k+1}[\boldsymbol{Z}^k] - \boldsymbol{U}_y^{k+1} - D_z^k}_{\varepsilon_B} + \underbrace{(\boldsymbol{Q}_o^k - I)\mathscr{H}^{k+1}[\boldsymbol{Q}_i^k\boldsymbol{Z}^k]}_{\varepsilon_o} + \underbrace{\mathscr{H}^{k+1}[(\boldsymbol{Q}_i^k - I)\boldsymbol{Z}^k]}_{\varepsilon_i} - \underbrace{(\boldsymbol{Q}_u^k - I)\boldsymbol{U}_y^{k+1}}_{\varepsilon_u},$$

where population derivatives refer to $(x^k, \boldsymbol{Y}^k)$. The intrinsic error ε_B is $\mathcal{G}_k$-measurable, whereas each sketch error has conditional mean zero. The output error also satisfies $\mathbb{E}[\varepsilon_o \mid \mathcal{G}_k, \boldsymbol{Q}_i^k, \boldsymbol{Q}_u^k] = 0$, which removes its cross terms with ε_i and ε_u. Finally, conditional independence of $\boldsymbol{Q}_i^k$ and $\boldsymbol{Q}_u^k$ removes the cross term between ε_i and ε_u. Expanding the squared norm and averaging therefore gives the middle identity in (22). The lower identity uses the split $(\boldsymbol{V}^{k+1} - D_y^k) + (\boldsymbol{Q}_g^k - I)\boldsymbol{V}^{k+1}$. The upper identity uses the split $(U_x^{k+1} - \mathscr{J}^{k+1}[\boldsymbol{Z}^k] - D_x^k) + (Q_x^k - I)U_x^{k+1} - \mathscr{J}^{k+1}[(\boldsymbol{Q}_j^k - I)\boldsymbol{Z}^k]$. Conditioning on $\mathcal{G}_k$ removes their cross terms in the same way. The intrinsic terms retain any covariance between shared oracle samples. □

Operator-branch error is governed by operator-weighted residual moments, which the next lemma characterizes explicitly.

Lemma B.5 (Operator-weighted sketch risk)**.** *Let $\mathcal{G} \supseteq \mathcal{F}_k$ contain only information conditionally independent of the current block lift $\boldsymbol{Q}$ given $\mathcal{F}_k$. Fix a $\mathcal{G}$-measurable positive-semidefinite product-space operator $\mathscr{B} : \mathcal{Y} \to \mathcal{Y}$ and $\boldsymbol{T} \in \mathcal{Y}$. Define the operator-weighted risk by*

$$\mathcal{R}_{\boldsymbol{Q}}(\mathscr{B}, \boldsymbol{T} \mid \mathcal{G}) := \mathbb{E}[\left\| \mathscr{B}^{1/2}[(\boldsymbol{Q} - I)\boldsymbol{T}] \right\|^2 \mid \mathcal{G}].$$

This product-space identity directly covers cross-layer and cross-column Hessian and Jacobian operators.

For one row block only, let $B \in \mathbb{R}^{m\times m}$ be symmetric positive semidefinite, $V \in \mathbb{R}^{m\times n}$ be fixed, and $\mathcal{C}_Q(B) := \mathbb{E}[(Q - I)^\top B(Q - I) \mid \mathcal{F}_k]$. Then $\mathcal{R}_Q(B, V) := \operatorname{tr}(V^\top \mathcal{C}_Q(B)V) = \mathbb{E}_k \left\| B^{1/2}(Q - I)R_h V \right\|_F^2$. For the Gaussian-complement law,

$$\mathcal{C}_Q^{\mathrm{G}}(B) = \frac{1}{a}\big(R_h B R_h + \operatorname{tr}(R_h B)R_h\big). \tag{23}$$

For the Stiefel-complement law with $c \geq 2$,

$$\mathcal{C}_Q^{\mathrm{St}}(B) = \frac{c - a}{a(c-1)(c+2)}\big((c-2)R_h B R_h + c\operatorname{tr}(R_h B)R_h\big). \tag{24}$$

The predictable mixture has $\mathcal{C}_Q(B) = \lambda\mathcal{C}_Q^{\mathrm{St}}(B) + (1-\lambda)\mathcal{C}_Q^{\mathrm{G}}(B)$. For the Stiefel principal–random law, $c = a = 1$ gives zero risk. Either family has zero risk when $Q = I$.

Proof. For (23), let $C = B_c^\top B B_c$ in the complement basis used in the proof of Lemma B.2. Gaussian fourth moments give $\mathbb{E}[\gamma\gamma^\top C\gamma\gamma^\top] = 2C + \operatorname{tr}(C)I_c$. Separate equal and unequal column indices as in the proof above to get $\mathbb{E}[(S_G - I_c)C(S_G - I_c)] = (C + \operatorname{tr}(C)I_c)/a$. Conjugation proves the claim. For (24), let Θ be the rank-a Haar projector in $\mathbb{R}^c$. Rotational invariance gives

$$\mathbb{E}[\Theta_{ij}\Theta_{kl}] = u\delta_{ij}\delta_{kl} + v(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}).$$

The identities $(\operatorname{tr}\Theta)^2 = a^2$ and $\operatorname{tr}(\Theta^2) = a$ imply

$$c^2 u + 2cv = a^2, \qquad cu + c(c+1)v = a, \qquad v = \frac{a(c-a)}{c(c-1)(c+2)}.$$

For symmetric C, this tensor formula yields $\mathbb{E}[\Theta C\Theta] = (u+v)C + v\operatorname{tr}(C)I_c$. Expanding $((c/a)\Theta - I_c)C((c/a)\Theta - I_c)$, substituting u, v, and conjugating by B_c gives (24). Both weighted formulas reduce to (17) when $B = I$.

For the full product-space input Hessian and Jacobian branches, take $\mathcal{G} = \mathcal{G}_k$ and use the operators $\mathscr{B} = (\mathscr{H}^{k+1})^*\mathscr{H}^{k+1}$ or $\mathscr{B} = (\mathscr{J}^{k+1})^*\mathscr{J}^{k+1}$, with $\boldsymbol{T} = \boldsymbol{Z}^k$. Here * is the adjoint under the product Frobenius inner product. For the output Hessian branch, additionally condition on $\boldsymbol{Q}_i^k$, take $\mathscr{B} = I$, and use $\boldsymbol{T} = \mathscr{H}^{k+1}[\boldsymbol{Q}_i^k \boldsymbol{Z}^k]$. The first-order branches use $\mathscr{B} = I$ and their sampled gradients. These general operator risks recover all six sketch terms above. The row-matrix formulas (23) and (24) specialize the risk to weights acting by left multiplication on a single block. The product-space definition covers general cross-block operators.

At a fixed state, compare a history law $\boldsymbol{Q}_{\mathrm{hist}}$ with a fully fresh law $\boldsymbol{Q}_{\mathrm{fresh}}$ having the same total width in every block. Holding $\mathcal{G}, \mathscr{B}, \boldsymbol{T}$ fixed, the exact one-sided gate is

$$\mathcal{R}_{\boldsymbol{Q}_{\mathrm{hist}}}(\mathscr{B}, \boldsymbol{T} \mid \mathcal{G}) \leq \mathcal{R}_{\boldsymbol{Q}_{\mathrm{fresh}}}(\mathscr{B}, \boldsymbol{T} \mid \mathcal{G}). \tag{25}$$

For an unweighted single-block branch, this reduces to $\chi_{\mathrm{hist}} \|R_h V\|_F^2 \le \chi_{\mathrm{fresh}} \|V\|_F^2$. For input operator branches, use the general product-space risk. The comparison holds the other factor laws fixed. Changing both Hessian factors requires the complete bi-sketch risk in Lemma 3.1. The gate quantifies the balance between operator-weighted residual energy and the variance associated with the remaining fresh rank. □

To turn these identities into rank-uniform bounds, we first control the solution maps and current-to-target mismatch. In what follows, we write $\omega := \omega_{\mathrm{OPR}} \ge 1$ for the uniform factor second-moment bound. κ retains its condition-number definition in Assumption 3.1.

Set $B_z := L_{f,0}/\mu_g$ and $A_z := L_{f,1} + L_{g,2}B_z$. The solution-map constants are $L_y := L_{g,1}/\mu_g$, $s := \sqrt{1+L_y^2}$, $L_z := sA_z/\mu_g$, and $L_\Phi := sA_z + L_{g,1}L_z$. The tracking errors E_Y^k, E_Z^k, e^k are defined in the main text. For compactness, write $\mathsf{y}_k := \|E_Y^k\|^2$, $\mathsf{z}_k := \|E_Z^k\|^2$, and $\mathsf{e}_k := \|e^k\|^2$. Also set $\mathsf{p}_k := \|\nabla\Phi(x^k)\|^2$ and $\mathsf{h}_k := \|h^k\|^2$.

These regularity estimates turn the three stochastic-direction errors into lower-, auxiliary-, and hypergradient-tracking errors.

Lemma B.6 (Global regularity bounds)**.** *Under Assumption 3.1, the maps $\boldsymbol{Y}^\star$, $\boldsymbol{Z}^\star$, and $\nabla\Phi$ are Lipschitz with constants L_y, L_z, and L_Φ in the definitions above, and $\|\boldsymbol{Z}^\star(x)\| \le B_z$. Moreover,*

$$\|D_x^k - \nabla\Phi(x^k)\|^2 \le 2A_z^2\mathsf{y}_k + 2L_{g,1}^2\mathsf{z}_k, \tag{26}$$

$$\|\mathscr{H}_k[\boldsymbol{Z}^\star(x^k)] - \nabla_{\boldsymbol{Y}} f(x^k, \boldsymbol{Y}^k)\|^2 \le A_z^2\mathsf{y}_k, \tag{27}$$

$$\|\nabla_x f(x^k, \boldsymbol{Y}^k)\|^2 \le 3\mathsf{p}_k + 3L_{f,1}^2\mathsf{y}_k + 3L_{g,1}^2B_z^2. \tag{28}$$

Proof. We first establish regularity of the solution maps. Fix x, x', and write $\boldsymbol{Y}_\star = \boldsymbol{Y}^\star(x)$ and $\boldsymbol{Y}_\star' = \boldsymbol{Y}^\star(x')$. Strong convexity and the first-order optimality conditions give

$$\begin{aligned}\mu_g \|\boldsymbol{Y}_\star - \boldsymbol{Y}_\star'\| &\le \|\nabla_{\boldsymbol{Y}} g(x, \boldsymbol{Y}_\star) - \nabla_{\boldsymbol{Y}} g(x, \boldsymbol{Y}_\star')\| \\ &= \|\nabla_{\boldsymbol{Y}} g(x', \boldsymbol{Y}_\star') - \nabla_{\boldsymbol{Y}} g(x, \boldsymbol{Y}_\star')\| \le L_{g,1}\|x - x'\|.\end{aligned}$$

Hence $\boldsymbol{Y}^\star$ is L_y-Lipschitz, and $\|(x, \boldsymbol{Y}_\star) - (x', \boldsymbol{Y}_\star')\| \le s\|x - x'\|$. Assumption 3.1 also implies $\mu_g I \preceq \mathscr{H}(x, \boldsymbol{Y}) \preceq L_{g,1}I$ and $\|\mathscr{J}(x, \boldsymbol{Y})\|_{\mathrm{op}} \le L_{g,1}$. The auxiliary solution therefore satisfies

$$\|\boldsymbol{Z}^\star(x)\| \le \|\mathscr{H}(x, \boldsymbol{Y}_\star)^{-1}\|_{\mathrm{op}} \|\nabla_{\boldsymbol{Y}} f(x, \boldsymbol{Y}_\star)\| \le L_{f,0}/\mu_g = B_z.$$

Set $\mathscr{H}_\star(x) = \mathscr{H}(x, \boldsymbol{Y}_\star)$ and $\boldsymbol{u}_\star(x) = \nabla_{\boldsymbol{Y}} f(x, \boldsymbol{Y}_\star)$. Subtracting the two auxiliary equations yields

$$\mathscr{H}_\star(x)[\boldsymbol{Z}^\star(x) - \boldsymbol{Z}^\star(x')] = \boldsymbol{u}_\star(x) - \boldsymbol{u}_\star(x') + (\mathscr{H}_\star(x') - \mathscr{H}_\star(x))[\boldsymbol{Z}^\star(x')],$$

$$\|\boldsymbol{Z}^\star(x) - \boldsymbol{Z}^\star(x')\| \le \frac{A_z}{\mu_g}\|(x, \boldsymbol{Y}_\star) - (x', \boldsymbol{Y}_\star')\| \le L_z\|x - x'\|.$$

To bound the hypergradient difference, add and subtract $\mathscr{J}(x, \boldsymbol{Y}_\star)[\boldsymbol{Z}^\star(x')]$ in (2). Then

$$\begin{aligned}\|\nabla\Phi(x) - \nabla\Phi(x')\| &\le A_z\|(x, \boldsymbol{Y}_\star) - (x', \boldsymbol{Y}_\star')\| + L_{g,1}\|\boldsymbol{Z}^\star(x) - \boldsymbol{Z}^\star(x')\| \\ &\le (sA_z + L_{g,1}L_z)\|x - x'\| = L_\Phi\|x - x'\|.\end{aligned}$$

We next control the mismatch at the current state. Write $\boldsymbol{Y}_\star^k = \boldsymbol{Y}^\star(x^k)$ and $\boldsymbol{Z}_\star^k = \boldsymbol{Z}^\star(x^k)$, and set $\mathscr{J}_\star^k = \mathscr{J}(x^k, \boldsymbol{Y}_\star^k)$. Adding and subtracting the population quantities at $\boldsymbol{Y}_\star^k$ gives

$$D_x^k - \nabla\Phi(x^k) = \nabla_x f(x^k, \boldsymbol{Y}^k) - \nabla_x f(x^k, \boldsymbol{Y}_\star^k) - \mathscr{J}_k[E_Z^k] + (\mathscr{J}_\star^k - \mathscr{J}_k)[\boldsymbol{Z}_\star^k].$$

Consequently,

$$\|D_x^k - \nabla\Phi(x^k)\| \le A_z\|E_Y^k\| + L_{g,1}\|E_Z^k\|.$$

Squaring and using $(a+b)^2 \le 2a^2 + 2b^2$ proves (26). The auxiliary equation at $\boldsymbol{Y}_\star^k$ similarly gives

$$\begin{aligned}\mathscr{H}_k[\boldsymbol{Z}_\star^k] - \nabla_{\boldsymbol{Y}} f(x^k, \boldsymbol{Y}^k) = (\mathscr{H}_k - \mathscr{H}(x^k, \boldsymbol{Y}_\star^k))[\boldsymbol{Z}_\star^k] \\ + \nabla_{\boldsymbol{Y}} f(x^k, \boldsymbol{Y}_\star^k) - \nabla_{\boldsymbol{Y}} f(x^k, \boldsymbol{Y}^k),\end{aligned}$$

whose norm is at most $A_z \left\|E_Y^k\right\|$. This proves (27).

Finally, the hypergradient identity controls the direct upper-gradient norm through

$$\nabla_x f(x^k, \boldsymbol{Y}^k) = \nabla\Phi(x^k) + \mathscr{J}_\star^k[\boldsymbol{Z}_\star^k] + \nabla_x f(x^k, \boldsymbol{Y}^k) - \nabla_x f(x^k, \boldsymbol{Y}_\star^k).$$

Apply $\|a+b+c\|^2 \le 3(\|a\|^2 + \|b\|^2 + \|c\|^2)$, $\left\|\mathscr{J}_\star^k\right\|_{\mathrm{op}} \le L_{g,1}$, and $\left\|\boldsymbol{Z}_\star^k\right\| \le B_z$ to obtain (28). This estimate is used only for the optional compressed-upper extension. The direct upper gradient in Algorithm 1 is unsketched. □

Successive conditioning transfers the factor moments to first- and second-order oracle calls, including calls with dependent oracle components.

Lemma B.7 (Conditional oracle moments)**.** *Under the factor and oracle assumptions, the sketched first-order and operator calls have the conditional means and second-moment bounds in* (29)–(31)*.*

Proof. For a mean-identity lift Q with second-moment constant ω_Q, let $\widetilde{v}$ be a conditionally independent unbiased oracle for v, with conditional variance at most σ^2. Then

$$\begin{aligned} \mathbb{E}_k[Q\widetilde{v}] &= v, \\ \mathbb{E}_k \left\|Q\widetilde{v} - v\right\|^2 &\le (\omega_Q - 1)\left\|v\right\|^2 + \omega_Q\sigma^2. \end{aligned} \tag{29}$$

To verify the second bound, set $\varepsilon = \widetilde{v} - v$. Conditional independence gives $\mathbb{E}[\varepsilon \mid \mathcal{F}_k, Q] = 0$, so

$$\begin{aligned} \mathbb{E}_k \left\|Q\widetilde{v} - v\right\|^2 &= \mathbb{E}_k \left\|(Q-I)v + Q\varepsilon\right\|^2 \\ &= \mathbb{E}_k \left\|(Q-I)v\right\|^2 + \mathbb{E}_k \left\|Q\varepsilon\right\|^2 \\ &\le (\omega_Q - 1)\left\|v\right\|^2 + \omega_Q \mathbb{E}_k \left\|\varepsilon\right\|^2. \end{aligned}$$

Here we used $\mathbb{E}_k[(Q-I)^\top(Q-I)] = \mathbb{E}_k[Q^\top Q] - I \preceq (\omega_Q - 1)I$.

For an $\mathcal{F}_k$-measurable $\boldsymbol{W} \in \mathcal{Y}$, Assumption 3.2 applies at the random input $\boldsymbol{Q}_i^k\boldsymbol{W}$ after conditioning on $\mathcal{F}_k^+$. Consequently,

$$\begin{aligned} \mathbb{E}_k[\boldsymbol{Q}_o^k \mathscr{H}^{k+1}[\boldsymbol{Q}_i^k \boldsymbol{W}]] &= \mathscr{H}_k[\boldsymbol{W}], \\ \mathbb{E}_k \left\|\boldsymbol{Q}_o^k \mathscr{H}^{k+1}[\boldsymbol{Q}_i^k \boldsymbol{W}]\right\|^2 &\le \omega_{P,o}\omega_{P,i} B_H \left\|\boldsymbol{W}\right\|^2, \end{aligned} \tag{30}$$

$$\begin{aligned} \mathbb{E}_k[\mathscr{J}^{k+1}[\boldsymbol{Q}_j^k \boldsymbol{W}]] &= \mathscr{J}_k[\boldsymbol{W}], \\ \mathbb{E}_k \left\|\mathscr{J}^{k+1}[\boldsymbol{Q}_j^k \boldsymbol{W}]\right\|^2 &\le \omega_{P,j} B_J \left\|\boldsymbol{W}\right\|^2. \end{aligned} \tag{31}$$

For the Hessian mean, average the output factor, the oracle, and then the input factor:

$$\mathbb{E}_k[\boldsymbol{Q}_o^k \mathscr{H}^{k+1}[\boldsymbol{Q}_i^k \boldsymbol{W}]] = \mathbb{E}_k[\mathscr{H}^{k+1}[\boldsymbol{Q}_i^k \boldsymbol{W}]] = \mathbb{E}_k[\mathscr{H}_k[\boldsymbol{Q}_i^k \boldsymbol{W}]] = \mathscr{H}_k[\boldsymbol{W}].$$

The corresponding second-moment calculation is

$$\begin{aligned} \mathbb{E}_k \left\|\boldsymbol{Q}_o^k \mathscr{H}^{k+1}[\boldsymbol{Q}_i^k \boldsymbol{W}]\right\|^2 &\le \omega_{P,o} \mathbb{E}_k \left\|\mathscr{H}^{k+1}[\boldsymbol{Q}_i^k \boldsymbol{W}]\right\|^2 \\ &\le \omega_{P,o} B_H \mathbb{E}_k \left\|\boldsymbol{Q}_i^k \boldsymbol{W}\right\|^2 \le \omega_{P,o}\omega_{P,i} B_H \left\|\boldsymbol{W}\right\|^2. \end{aligned}$$

The first inequality conditions on $(\mathcal{F}_k, \boldsymbol{Q}_i^k, \mathscr{H}^{k+1})$ and averages the independent output factor. The second conditions on $\mathcal{F}_k^+$ and applies Assumption 3.2 to its measurable input. The Jacobian calculation omits the output-factor step. Linearity of conditional expectation now proves the mean identities in Lemma 3.1. □

Combining these oracle moments with the regularity bounds yields state-dependent variance envelopes for all three assembled directions.

Lemma B.8 (Primitive direction variances, including an optional upper sketch)**.** *For the lower direction, set* $v_0 := \sigma_g^2$ *and* $v_y := L_{g,1}^2$*. For the auxiliary direction, set* $s_0 := 4B_H B_z^2 + 4L_{f,0}^2 + 2\sigma_y^2$*,* $s_y := 4L_{f,1}^2$*, and* $s_z := 4B_H$*. For the upper direction, set* $w_0 := 6L_{g,1}^2 B_z^2 + 4B_J B_z^2 + 2\sigma_x^2$*,*

$w_y := 6L_{f,1}^2$, *and* $w_z := 4B_J$. *For a possible upper-*x *sketch, write* $\omega_{P,x} \le \omega$ *for its second-moment bound. Algorithm 1 has* $\omega_{P,x} = 1$. *Then*

$$\nu_{V,k} \le \omega(v_0 + v_y \mathsf{y}_k), \tag{32}$$

$$\nu_{S,k} \le \omega^2(s_0 + s_y \mathsf{y}_k + s_z \mathsf{z}_k), \tag{33}$$

$$\begin{aligned}\nu_{W,k} &\le \omega(w_0 + w_y \mathsf{y}_k + w_z \mathsf{z}_k) + 6(\omega_{P,x} - 1)\mathsf{p}_k \\ &\le \omega^2(w_0 + w_y \mathsf{y}_k + w_z \mathsf{z}_k + 6\mathsf{p}_k)\end{aligned} \tag{34}$$

These coefficients depend only on the primitive regularity and oracle constants. For Algorithm 1, $Q_x^k = I$ *and hence* $\omega_{P,x} = 1$. *The stationarity-relative term in* (34) *vanishes, yielding the rank-uniform interface in* (12).

Proof. For the lower direction, $\nabla_{\boldsymbol{Y}} g(x^k, \boldsymbol{Y}^\star(x^k)) = 0$, so $\left\|D_y^k\right\| \le L_{g,1} \left\|E_Y^k\right\|$. Applying (29) gives

$$\nu_{V,k} \le (\omega_{P,g} - 1)L_{g,1}^2 \mathsf{y}_k + \omega_{P,g}\sigma_g^2 \le \omega(v_0 + v_y \mathsf{y}_k).$$

For the auxiliary direction, define the centered branch errors $\varepsilon_H := \boldsymbol{Q}_o^k \mathscr{H}^{k+1}[\boldsymbol{Q}_i^k \boldsymbol{Z}^k] - \mathscr{H}_k[\boldsymbol{Z}^k]$ and $\varepsilon_u := \boldsymbol{Q}_u^k \boldsymbol{U}_y^{k+1} - \nabla_{\boldsymbol{Y}} f(x^k, \boldsymbol{Y}^k)$. The oracle tuple may share stochastic samples. We therefore use $\|\varepsilon_H - \varepsilon_u\|^2 \le 2\|\varepsilon_H\|^2 + 2\|\varepsilon_u\|^2$, which gives

$$\begin{aligned}\nu_{S,k} &\le 2\mathbb{E}_k \|\varepsilon_H\|^2 + 2\mathbb{E}_k \|\varepsilon_u\|^2 \\ &\le 2\omega_{P,o}\omega_{P,i} B_H \left\|\boldsymbol{Z}^k\right\|^2 + 2(\omega_{P,u} - 1)\left\|\nabla_{\boldsymbol{Y}} f(x^k, \boldsymbol{Y}^k)\right\|^2 + 2\omega_{P,u}\sigma_y^2.\end{aligned}$$

The second line follows from (30) and (29), using variance bounded by second moment for the Hessian branch. The state-dependent terms satisfy

$$\begin{aligned}\left\|\boldsymbol{Z}^k\right\|^2 &\le 2\mathsf{z}_k + 2B_z^2, \\ \left\|\nabla_{\boldsymbol{Y}} f(x^k, \boldsymbol{Y}^k)\right\|^2 &\le 2\left\|\nabla_{\boldsymbol{Y}} f(x^k, \boldsymbol{Y}^k) - \nabla_{\boldsymbol{Y}} f(x^k, \boldsymbol{Y}^\star(x^k))\right\|^2 + 2L_{f,0}^2 \\ &\le 2L_{f,1}^2 \mathsf{y}_k + 2L_{f,0}^2.\end{aligned}$$

Substitute these estimates and use $\omega_{P,o}\omega_{P,i}, \omega_{P,u}, \omega_{P,u} - 1 \le \omega^2$ to obtain (33) with the stated s_0, s_y, s_z.

For the upper direction, center the direct upper-gradient and Jacobian branches separately. The same squared-norm inequality yields

$$\begin{aligned}\nu_{W,k} &\le 2(\omega_{P,x} - 1)\left\|\nabla_x f(x^k, \boldsymbol{Y}^k)\right\|^2 + 2\omega_{P,x}\sigma_x^2 + 2\omega_{P,j} B_J \left\|\boldsymbol{Z}^k\right\|^2 \\ &\le 6(\omega_{P,x} - 1)\mathsf{p}_k + \omega(w_0 + w_y \mathsf{y}_k + w_z \mathsf{z}_k).\end{aligned}$$

The second inequality uses (28), $\left\|\boldsymbol{Z}^k\right\|^2 \le 2\mathsf{z}_k + 2B_z^2$, and the definitions of w_0, w_y, w_z. Since $\omega \ge 1$ and $\omega_{P,x} - 1 \le \omega^2$, the last expression also implies the second bound in (34). □

Proof of Lemma 3.1. Lemmas B.1, B.3, and B.7 give the conditional direction means. With $Q_x^k = I$, Lemma B.8 bounds their total variance by $C_{\mathrm{dir}}\omega_{\mathrm{OPR}}^2(1 + \left\|E_Y^k\right\|^2 + \left\|E_Z^k\right\|^2)$, where one may take C_{dir} as the maximum of the summed constant, lower-error, and auxiliary-error coefficients in (32)–(34). This proves (12). □

B.2 Proof of Lemma 3.2

As in Lemma 3.2, specialize the algorithmic stepsizes to constants $\alpha_k \equiv \alpha$, $\beta_k \equiv \beta$, $\gamma_k \equiv \gamma$, and $\theta_k \equiv \theta$. All expectations used below are finite. For an L_Φ-smooth function bounded below, a gradient step of length $1/L_\Phi$ gives $\|\nabla\Phi(x)\|^2 \le 2L_\Phi(\Phi(x) - \Phi^\star)$. Assumption 3.3 therefore gives finite initial expected $\mathsf{p}_0, \mathsf{h}_0, \mathsf{y}_0, \mathsf{z}_0, \mathsf{e}_0$. Equations (29)–(34), the Lipschitz target maps, and the recursion imply the corresponding finite moments at every finite iteration by induction. Smoothness also gives finite expected objective values. These finite-moment and expected-objective bounds suffice for the subsequent summation.

For the contraction terms, set $r_y = r_z := \mu_g/2$ and $d_{zy} := 6A_z^2/\mu_g$. For the moving-hypergradient mismatch, set $b_y := 4A_z^2$ and $b_z := 4L_{g,1}^2$. The target-motion coefficients are $t_y := 3L_y^2/\mu_g$, $t_z := 3L_z^2/\mu_g$, and $t_h := 3L_\Phi^2$.

The next lemma tracks the lower state, auxiliary state, and moving hypergradient against targets that shift with the upper iterate.

Lemma B.9 (One-step tracking recursions)**.** *If* $0 < \beta \le \mu_g/L_{g,1}^2$, $0 < \gamma \le \mu_g/(4L_{g,1}^2)$, *and* $0 < \theta \le 1$, *then*

$$\mathbb{E}_k \mathsf{y}_{k+1} \le (1 - r_y\beta)\mathsf{y}_k + t_y\frac{\alpha^2}{\beta}\mathsf{h}_k + 2\beta^2\nu_{V,k},$$

$$\mathbb{E}_k \mathsf{z}_{k+1} \le (1 - r_z\gamma)\mathsf{z}_k + d_{zy}\gamma\mathsf{y}_k + t_z\frac{\alpha^2}{\gamma}\mathsf{h}_k + 2\gamma^2\nu_{S,k},$$

$$\mathbb{E}_k \mathsf{e}_{k+1} \le (1 - \theta/2)\mathsf{e}_k + b_y\theta\mathsf{y}_k + b_z\theta\mathsf{z}_k + t_h\frac{\alpha^2}{\theta}\mathsf{h}_k + 2\theta^2\nu_{W,k}.$$

Proof of Lemma B.9. For the fixed lower target, strong convexity, Lipschitz continuity, and conditional unbiasedness give

$$\left\|E_Y^k - \beta D_y^k\right\|^2 \le (1 - 2\mu_g\beta + L_{g,1}^2\beta^2)\mathsf{y}_k \le (1 - \mu_g\beta)\mathsf{y}_k,$$
$$\mathbb{E}_k \left\|\boldsymbol{Y}^{k+1} - \boldsymbol{Y}^\star(x^k)\right\|^2 \le (1 - \mu_g\beta)\mathsf{y}_k + \beta^2\nu_{V,k}.$$

Use Young's inequality with parameter $\mu_g\beta/2$ to move the target to $\boldsymbol{Y}^\star(x^{k+1})$. Since $x^{k+1} = x^k - \alpha h^k$ is $\mathcal{F}_k$-measurable, its target displacement has squared norm at most $L_y^2\alpha^2\mathsf{h}_k$. Because $\mu_g\beta \le 1$,

$$(1 + \mu_g\beta/2)(1 - \mu_g\beta) \le 1 - \mu_g\beta/2, \qquad 1 + 2/(\mu_g\beta) \le 3/(\mu_g\beta), \qquad 1 + \mu_g\beta/2 \le 2.$$

Consequently,

$$\begin{aligned}\mathbb{E}_k \mathsf{y}_{k+1} &\le (1 + \mu_g\beta/2)\big[(1 - \mu_g\beta)\mathsf{y}_k + \beta^2\nu_{V,k}\big] + (1 + 2/(\mu_g\beta))L_y^2\alpha^2\mathsf{h}_k \\ &\le (1 - \mu_g\beta/2)\mathsf{y}_k + \frac{3L_y^2\alpha^2}{\mu_g\beta}\mathsf{h}_k + 2\beta^2\nu_{V,k},\end{aligned}$$

which is the first recursion in Lemma B.9.

For the auxiliary target, set $\mathscr{H} = \mathscr{H}_k$ and $\boldsymbol{b} = \mathscr{H}[\boldsymbol{Z}^\star(x^k)] - \nabla_{\boldsymbol{Y}} f(x^k, \boldsymbol{Y}^k)$. Then $D_z^k = \mathscr{H}[E_Z^k] + \boldsymbol{b}$ and $\|\boldsymbol{b}\|^2 \le A_z^2\mathsf{y}_k$. Expansion gives

$$\begin{aligned}\left\|E_Z^k - \gamma(\mathscr{H}[E_Z^k] + \boldsymbol{b})\right\|^2 &\le (1 - 2\mu_g\gamma + 2L_{g,1}^2\gamma^2)\mathsf{z}_k + 2\gamma\left\|E_Z^k\right\|\|\boldsymbol{b}\| + 2\gamma^2\|\boldsymbol{b}\|^2 \\ &\le (1 - 3\mu_g\gamma/2 + 2L_{g,1}^2\gamma^2)\mathsf{z}_k + (2\gamma/\mu_g + 2\gamma^2)\|\boldsymbol{b}\|^2 \\ &\le (1 - \mu_g\gamma)\mathsf{z}_k + (4A_z^2/\mu_g)\gamma\mathsf{y}_k.\end{aligned}$$

The first inequality uses $\left\|\mathscr{H}[E_Z^k] + \boldsymbol{b}\right\|^2 \le 2L_{g,1}^2\mathsf{z}_k + 2\|\boldsymbol{b}\|^2$. The second inequality uses $2\left\|E_Z^k\right\|\|\boldsymbol{b}\| \le (\mu_g/2)\left\|E_Z^k\right\|^2 + (2/\mu_g)\|\boldsymbol{b}\|^2$. The last uses $\gamma \le \mu_g/(4L_{g,1}^2) \le 1/\mu_g$. Add the centered variance $\gamma^2\nu_{S,k}$ and apply Young's inequality with parameter $\mu_g\gamma/2$ to the target displacement, whose squared norm is at most $L_z^2\alpha^2\mathsf{h}_k$. The multiplier on the lower-error term is at most $3/2$, giving $d_{zy} = 6A_z^2/\mu_g$. The other coefficients follow as in the lower proof. Thus

$$\mathbb{E}_k \mathsf{z}_{k+1} \le (1 - \mu_g\gamma/2)\mathsf{z}_k + (6A_z^2/\mu_g)\gamma\mathsf{y}_k + \frac{3L_z^2\alpha^2}{\mu_g\gamma}\mathsf{h}_k + 2\gamma^2\nu_{S,k}.$$

This is the second recursion in Lemma B.9.

For the moving hypergradient, set $b_h = D_x^k - \nabla\Phi(x^k)$. With $\varepsilon_W = \widehat{W}^{k+1} - D_x^k$, the update satisfies

$$h^{k+1} - \nabla\Phi(x^k) = (1 - \theta)e^k + \theta b_h + \theta\varepsilon_W, \qquad \mathbb{E}_k[\varepsilon_W] = 0.$$

Conditional centering and convexity of the squared norm give

$$\mathbb{E}_k \left\|h^{k+1} - \nabla\Phi(x^k)\right\|^2 \le (1 - \theta)\mathsf{e}_k + \theta\|b_h\|^2 + \theta^2\nu_{W,k}.$$

Move the target with Young's parameter $\theta/2$. Its squared displacement is at most $L_\Phi^2\alpha^2\mathsf{h}_k$. Use $(1+\theta/2)(1-\theta) \le 1-\theta/2$, $1+2/\theta \le 3/\theta$, and (26). Enlarging $3A_z^2, 3L_{g,1}^2$ to $4A_z^2, 4L_{g,1}^2$ yields

$$\begin{aligned}\mathbb{E}_k\mathsf{e}_{k+1} &\le (1+\theta/2)\big[(1-\theta)\mathsf{e}_k + \theta\,\|b_h\|^2 + \theta^2\nu_{W,k}\big] + (1+2/\theta)L_\Phi^2\alpha^2\mathsf{h}_k \\ &\le (1-\theta/2)\mathsf{e}_k + 4A_z^2\theta\mathsf{y}_k + 4L_{g,1}^2\theta\mathsf{z}_k + \frac{3L_\Phi^2\alpha^2}{\theta}\mathsf{h}_k + 2\theta^2\nu_{W,k}.\end{aligned}$$

This proves the final recursion in Lemma B.9. □

The tracking bounds contain positive terms in $\|h^k\|^2$. Smoothness provides the matching negative term from objective descent.

Lemma B.10 (Objective descent)**.** *If $0 < \alpha \le (2L_\Phi)^{-1}$, then*

$$\Phi(x^{k+1}) \le \Phi(x^k) - \frac{\alpha}{2}\mathsf{p}_k - \frac{\alpha}{4}\mathsf{h}_k + \frac{\alpha}{2}\mathsf{e}_k. \tag{35}$$

Proof. Since $x^{k+1} = x^k - \alpha h^k$, L_Φ-smoothness gives

$$\begin{aligned}\Phi(x^{k+1}) &\le \Phi(x^k) - \alpha\left\langle\nabla\Phi(x^k), h^k\right\rangle + \frac{L_\Phi\alpha^2}{2}\mathsf{h}_k \\ &= \Phi(x^k) - \frac{\alpha}{2}\mathsf{p}_k - \frac{\alpha}{2}(1-L_\Phi\alpha)\mathsf{h}_k + \frac{\alpha}{2}\mathsf{e}_k.\end{aligned}$$

The equality uses $2\left\langle\nabla\Phi(x^k), h^k\right\rangle = \mathsf{p}_k + \mathsf{h}_k - \mathsf{e}_k$. The stepsize bound implies $1 - L_\Phi\alpha \ge 1/2$, proving the claim. □

Fixed Lyapunov weights combine objective descent and tracking. A final stepsize restriction absorbs the state-dependent direction variances.

Lemma B.11 (One-step Lyapunov descent)**.** *There are positive rank-independent constants $\tau, \lambda_y, \lambda_z, \lambda_h, q_V, q_S, q_W, N_0$ and thresholds $0 < \bar\alpha \le \alpha_0$ such that, for $\beta = \gamma = \theta = \tau\alpha$ and $\Psi_k := \Phi(x^k) - \Phi^\star + \lambda_y\mathsf{y}_k + \lambda_z\mathsf{z}_k + \lambda_h\mathsf{e}_k$, the following bounds hold:*

$$\mathbb{E}_k\Psi_{k+1} \le \Psi_k - \alpha\big(\tfrac{1}{2}\mathsf{p}_k + \tfrac{1}{8}\mathsf{h}_k + \mathsf{y}_k + \mathsf{z}_k + \tfrac{1}{2}\mathsf{e}_k\big) + \alpha^2(q_V\nu_{V,k} + q_S\nu_{S,k} + q_W\nu_{W,k}) \tag{36}$$

for $0 < \alpha \le \alpha_0$, and

$$\mathbb{E}_k\Psi_{k+1} \le \Psi_k - \frac{\alpha}{4}\mathsf{p}_k - \frac{\alpha}{8}\mathsf{h}_k - \frac{\alpha}{2}(\mathsf{y}_k + \mathsf{z}_k + \mathsf{e}_k) + N_0\omega_{\mathrm{OPR}}^2\alpha^2 \tag{37}$$

for $0 < \alpha \le \bar\alpha/\omega_{\mathrm{OPR}}^2$.

Proof of Lemma B.11. Choose $U_z := (2b_z+1)/r_z$ and $U_y := (U_z d_{zy} + 2b_y + 1)/r_y$. With $D := U_y t_y + U_z t_z + 2t_h$, take $\tau := \max\{1, \sqrt{8D}\}$ and $\lambda_y := U_y/\tau$, $\lambda_z := U_z/\tau$, and $\lambda_h := 2/\tau$. These choices depend only on primitive regularity. For Ψ_k as defined in the lemma, set $q_V := 2U_y\tau$, $q_S := 2U_z\tau$, $q_W := 4\tau$, and $\alpha_0 := \min\{1, \mu_g/(4\tau L_{g,1}^2), 1/\tau, 1/(2L_\Phi)\}$. For $\beta = \gamma = \theta = \tau\alpha$ and $\alpha \le \alpha_0$, Lemma B.9 applies. Multiplying its three inequalities by their Lyapunov weights and adding (35) gives the variance-sensitive inequality (36). For completeness, the coefficients of $\alpha\mathsf{y}_k$, $\alpha\mathsf{z}_k$, and $\alpha\mathsf{e}_k$ are respectively

$$-U_y r_y + U_z d_{zy} + 2b_y = -1, \qquad -U_z r_z + 2b_z = -1, \qquad 1/2 - 1 = -1/2.$$

The coefficient of $\alpha\mathsf{h}_k$ is at most $-1/4 + D/\tau^2 \le -1/8$. Thus the coupling choice preserves a negative coefficient for h_k after including all three moving-target terms. It remains to absorb the relative variances. Set $N_0 := q_V v_0 + q_S s_0 + q_W w_0$ and $N_y := q_V v_y + q_S s_y + q_W w_y$ for the constant and lower-error terms. The remaining coefficients are $N_z := q_S s_z + q_W w_z$ and $N_p := 6q_W$. Then Lemma B.8 implies

$$q_V\nu_{V,k} + q_S\nu_{S,k} + q_W\nu_{W,k} \le \omega^2(N_0 + N_y\mathsf{y}_k + N_z\mathsf{z}_k + N_p\mathsf{p}_k). \tag{38}$$

Set

$$\bar\alpha = \min\left\{\alpha_0, \frac{1}{4(1+N_p)}, \frac{1}{2(1+N_y)}, \frac{1}{2(1+N_z)}\right\} > 0. \tag{39}$$

Whenever $\alpha \le \bar{\alpha}/\omega^2$, $\alpha\omega^2 N_p \le 1/4$ and $\alpha\omega^2 N_y, \alpha\omega^2 N_z \le 1/2$. Substitution into the variance-sensitive inequality yields

$$\mathbb{E}_k \Psi_{k+1} \le \Psi_k - \frac{\alpha}{4}\mathsf{p}_k - \frac{\alpha}{8}\mathsf{h}_k - \frac{\alpha}{2}(\mathsf{y}_k + \mathsf{z}_k + \mathsf{e}_k) + N_0\omega^2\alpha^2.$$

The N_p term covers an optional upper-gradient sketch. For Algorithm 1, $Q_x^k = I$ and it may be set to zero. The negative stationarity term absorbs any such contribution under (39). All chosen constants and thresholds are independent of ranks and the horizon. □

Summing the absorbed Lyapunov decrease now gives the cumulative tracking inequality stated in the main text.

Proof of Lemma 3.2. Let $\overline{\Psi}_0 := \mathbb{E}\Psi_0 < \infty$, which is finite by Assumption 3.3. Taking total expectations in (37), summing from $k = 0$ to $K-1$, and using $\Psi_K \ge 0$ gives

$$\frac{\alpha}{4}\sum_{k=0}^{K-1}\mathbb{E}\mathsf{p}_k + \frac{\alpha}{2}\sum_{k=0}^{K-1}\mathbb{E}(\mathsf{y}_k + \mathsf{z}_k + \mathsf{e}_k) \le \overline{\Psi}_0 + N_0\omega_{\mathrm{OPR}}^2 K\alpha^2.$$

Multiplying by four and weakening the coefficients yields (13) with, for example, $C_0 = 4(1+\overline{\Psi}_0)$ and $C_1 = 4(1+N_0)$. These constants depend only on the primitive regularity, oracle, and initialization bounds and are independent of the ranks and horizon. □

B.3 Proof of Theorem 3.1

Lemma 3.2 already contains the cumulative estimate. Uniform output sampling converts it to stationarity.

Proof of Theorem 3.1. Since $\widehat{k}$ is independent and uniform on $\{0, \dots, K-1\}$,

$$\mathbb{E}\left\|\nabla\Phi(x^{\widehat{k}})\right\|^2 = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla\Phi(x^k)\right\|^2.$$

Dropping the nonnegative tracking-error terms from (13) and dividing by $K\alpha$ gives

$$\mathbb{E}\left\|\nabla\Phi(x^{\widehat{k}})\right\|^2 \le \frac{C_0}{K\alpha} + C_1\omega_{\mathrm{OPR}}^2\alpha,$$

which is (14). For completeness, if $K \ge C_0\omega_{\mathrm{OPR}}^2/(C_1\bar{\alpha}^2)$, then $\alpha = \sqrt{C_0/C_1}/(\omega_{\mathrm{OPR}}\sqrt{K})$ is admissible and yields

$$\mathbb{E}\left\|\nabla\Phi(x^{\widehat{k}})\right\|^2 \le 2\sqrt{C_0 C_1}\,\frac{\omega_{\mathrm{OPR}}}{\sqrt{K}}.$$

This also proves Corollary 3.1. □

Retaining the realized branch risks in the same Lyapunov argument yields a history-sensitive stationarity bound.

Lemma B.12 (History-sensitive stationarity bound)**.** *For any sequence A_k, let $\bar{A} = K^{-1}\sum_{k=0}^{K-1}\mathbb{E}A_k$. Under the conditions of Theorem 3.1, there are positive rank-independent constants $\Gamma_V, \Gamma_S, \Gamma_W$ such that*

$$\mathbb{E}\left\|\nabla\Phi(x^{\widehat{k}})\right\|^2 \le \frac{C_0}{K\alpha} + \alpha(\Gamma_V\bar{\nu}_V + \Gamma_S\bar{\nu}_S + \Gamma_W\bar{\nu}_W). \tag{40}$$

By Lemma B.4, each $\bar{\nu}_q$ is exactly the sum of its averaged intrinsic risk and branchwise sketch risks.

Proof. For $\alpha \le \alpha_0$, take total expectations in (36) and sum over k:

$$\begin{aligned}\frac{\alpha}{2}\sum_{k=0}^{K-1}\mathbb{E}\mathsf{p}_k &\le \mathbb{E}\Psi_0 - \mathbb{E}\Psi_K + \alpha^2\sum_{k=0}^{K-1}\mathbb{E}[q_V\nu_{V,k} + q_S\nu_{S,k} + q_W\nu_{W,k}] \\ &\le \overline{\Psi}_0 + K\alpha^2(q_V\bar{\nu}_V + q_S\bar{\nu}_S + q_W\bar{\nu}_W).\end{aligned}$$

Thus

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla\Phi(x^k)\right\|^2 \le \frac{2\overline{\Psi}_0}{K\alpha} + 2\alpha(q_V\bar{\nu}_V + q_S\bar{\nu}_S + q_W\bar{\nu}_W). \tag{41}$$

Choose $\Gamma_V = 2q_V$, $\Gamma_S = 2q_S$, and $\Gamma_W = 2q_W$. The constant $C_0 = 4(1+\overline{\Psi}_0)$ used above also bounds the initial term here. Since $\bar{\alpha} \le \alpha_0$, the stepsize range of Theorem 3.1 satisfies the required condition. Finally, independence and uniformity of $\widehat{k}$ identify the averaged squared gradient with $\mathbb{E}\left\|\nabla\Phi(x^{\widehat{k}})\right\|^2$, proving (40). □

C Guarantees for Oja Subspace Learning and Safe Selection

C.1 Fixed-Signal Oja Convergence

Proposition C.1 (Local fixed-signal Oja certificate)**.** *For one block with $1 \le r_h < m$, let U_h^{k+1} follow the Oja update* (4) *with $X_k^{\mathrm{hist}} \equiv X$ and $\eta_{\mathrm{Oja},k} \equiv \eta_{\mathrm{Oja}}$. Let $\Sigma = XX^\top$ have eigenvalues $\lambda_1 \ge \cdots \ge \lambda_m \ge 0$, and set $\Delta_\Sigma := \lambda_{r_h} - \lambda_{r_h+1} > 0$, $\lambda_{\max} := \lambda_1$, and $g_k := \sum_{i=1}^{r_h}\lambda_i - \mathrm{tr}((U_h^k)^\top\Sigma U_h^k)$. If U_h^0 has orthonormal columns, $g_0 \le \Delta_\Sigma/2$, and $0 < \eta_{\mathrm{Oja}}\lambda_{\max} \le 1/2$, then every U_h^k has orthonormal columns and, for $k \ge 0$ and $K \ge 1$,*

$$g_k \le \left(1 - \frac{\eta_{\mathrm{Oja}}\Delta_\Sigma^2}{2\lambda_{\max}}\right)^k g_0,$$

$$\frac{1}{K}\sum_{t=0}^{K-1}\left\|(I-\Pi_h^t)X\right\|_F^2 \le \sum_{i>r_h}\lambda_i + \frac{2\lambda_{\max}g_0}{K\eta_{\mathrm{Oja}}\Delta_\Sigma^2}. \tag{42}$$

Proof. Equation (4) gives $C_k = (I - U_h^k(U_h^k)^\top)X_k^{\mathrm{hist}}(X_k^{\mathrm{hist}})^\top U_h^k$ and $(U_h^k)^\top C_k = 0$. Define the pre-QR matrix $\widetilde{U}_h^{k+1} := U_h^k + \eta_{\mathrm{Oja}}C_k$. Consequently, $(\widetilde{U}_h^{k+1})^\top\widetilde{U}_h^{k+1} = I_{r_h} + \eta_{\mathrm{Oja}}^2 C_k^\top C_k \succeq I_{r_h}$. Thus $\widetilde{U}_h^{k+1}$ has full column rank, and thin QR returns an orthonormal basis for its column space. A fixed sign convention for QR makes this update measurable with respect to $\mathcal{F}_{k+1}$. All production directions at iteration k use U_h^k before this update. The resulting U_h^{k+1} first determines a production law at iteration $k+1$. The conditional mean calculations therefore continue to use the past-only basis required by Assumption 3.2.

Define $F(U) := \mathrm{tr}(U^\top\Sigma U)$ and $F_\star := \sum_{i=1}^{r_h}\lambda_i$. Fix one iterate, abbreviate $U = U_h^k$, and write $\Pi = UU^\top$, $C = (I-\Pi)\Sigma U$, and $\widetilde{U} = U + \eta_{\mathrm{Oja}}C$. Let $S = U^\top\Sigma U$, $T = C^\top C$, and $D = (I_{r_h} + \eta_{\mathrm{Oja}}^2 T)^{-1}$. Since $U^\top C = 0$ and $U^\top\Sigma C = T$,

$$\widetilde{U}^\top\widetilde{U} = I_{r_h} + \eta_{\mathrm{Oja}}^2 T, \qquad \widetilde{U}^\top\Sigma\widetilde{U} = S + 2\eta_{\mathrm{Oja}}T + \eta_{\mathrm{Oja}}^2 C^\top\Sigma C.$$

Thin QR and the normalized matrix $\widetilde{U}D^{1/2}$ span the same subspace. Hence the next basis U_+ satisfies

$$\begin{aligned} F(U_+) - F(U) &= \mathrm{tr}\big(D(S + 2\eta_{\mathrm{Oja}}T + \eta_{\mathrm{Oja}}^2 C^\top\Sigma C)\big) - \mathrm{tr}(S) \\ &= 2\eta_{\mathrm{Oja}}\,\mathrm{tr}(DT) + \eta_{\mathrm{Oja}}^2\,\mathrm{tr}(DC^\top\Sigma C) - \eta_{\mathrm{Oja}}^2\,\mathrm{tr}(DTS) \\ &\ge (2\eta_{\mathrm{Oja}} - \eta_{\mathrm{Oja}}^2\lambda_{\max})\,\mathrm{tr}(DT). \end{aligned}$$

The last line uses $\Sigma \succeq 0$, $S \preceq \lambda_{\max}I_{r_h}$, and the fact that DT is positive semidefinite. Also $\|C\|_{\mathrm{op}} \le \lambda_{\max}$, so $D \succeq (1+\eta_{\mathrm{Oja}}^2\lambda_{\max}^2)^{-1}I_{r_h}$. For $\eta_{\mathrm{Oja}}\lambda_{\max} \le 1/2$,

$$F(U_+) - F(U) \ge \eta_{\mathrm{Oja}}\frac{2-\eta_{\mathrm{Oja}}\lambda_{\max}}{1+\eta_{\mathrm{Oja}}^2\lambda_{\max}^2}\|C\|_F^2 \ge \eta_{\mathrm{Oja}}\|C\|_F^2. \tag{43}$$

In particular, $g_k = F_\star - F(U_h^k)$ is nonincreasing. To relate the tangent norm to the objective gap, work in an eigenbasis of Σ and partition the rank-r_h projector as

$$\Sigma = \begin{pmatrix}\Lambda_1 & 0\\ 0 & \Lambda_2\end{pmatrix}, \qquad \Pi = \begin{pmatrix}\Pi_{11} & \Pi_{12}\\ \Pi_{12}^\top & \Pi_{22}\end{pmatrix},$$

where Λ_1 contains the largest r_h eigenvalues. At this fixed iterate, write $g := F_\star - F(U)$. Define the subspace error $\delta_\Pi = r_h - \operatorname{tr}(\Pi_{11}) = \operatorname{tr}(\Pi_{22})$. Since $0 \preceq \Pi_{11} \preceq I_{r_h}$ and $\Pi_{22} \succeq 0$,

$$g = \operatorname{tr}(\Lambda_1(I_{r_h} - \Pi_{11})) - \operatorname{tr}(\Lambda_2\Pi_{22}),$$
$$\Delta_\Sigma\delta_\Pi \le g \le \lambda_{\max}\delta_\Pi. \tag{44}$$

The lower bound uses $\Lambda_1 \succeq \lambda_{r_h} I_{r_h}$ and $\Lambda_2 \preceq \lambda_{r_h+1} I$. The upper bound uses $\Lambda_1 \preceq \lambda_{\max} I_{r_h}$ and $\Lambda_2 \succeq 0$. Whenever $g \le \Delta_\Sigma/2$, we have $\delta_\Pi \le 1/2$ and $\Pi_{11} \succeq (1-\delta_\Pi) I_{r_h} \succeq I_{r_h}/2$. The identity $\Pi^2 = \Pi$ then gives

$$\|\Pi_{12}\|_F^2 = \operatorname{tr}(\Pi_{11} - \Pi_{11}^2) \ge \frac{1}{2}\operatorname{tr}(I_{r_h} - \Pi_{11}) = \frac{\delta_\Pi}{2}.$$

For the tangent norm, the commutator identity gives

$$2\|C\|_F^2 = \|\Sigma\Pi - \Pi\Sigma\|_F^2 = \sum_{i,j}(\lambda_i - \lambda_j)^2\Pi_{ij}^2 \ge 2\Delta_\Sigma^2\|\Pi_{12}\|_F^2.$$

Combining the last two estimates with (44) yields

$$\|C\|_F^2 \ge \frac{\Delta_\Sigma^2}{2}\delta_\Pi \ge \frac{\Delta_\Sigma^2}{2\lambda_{\max}} g. \tag{45}$$

By (43), $g_k \le g_0 \le \Delta_\Sigma/2$ for every k, so (45) applies at each step.

$$g_{k+1} \le g_k - \eta_{\mathrm{Oja}}\|C_k\|_F^2 \le \left(1 - \frac{\eta_{\mathrm{Oja}}\Delta_\Sigma^2}{2\lambda_{\max}}\right) g_k.$$

The contraction coefficient lies in $(0, 1/4]$ because $\Delta_\Sigma \le \lambda_{\max}$ and $\eta_{\mathrm{Oja}}\lambda_{\max} \le 1/2$. Iterating proves the first claim. Finally,

$$\left\|(I - \Pi_h^t)X\right\|_F^2 = \operatorname{tr}(\Sigma) - \operatorname{tr}((U_h^t)^\top\Sigma U_h^t) = \sum_{i>r_h}\lambda_i + g_t,$$

$$\frac{1}{K}\sum_{t=0}^{K-1}\left\|(I - \Pi_h^t)X\right\|_F^2 \le \sum_{i>r_h}\lambda_i + \frac{g_0}{K}\sum_{t=0}^{K-1}\left(1 - \frac{\eta_{\mathrm{Oja}}\Delta_\Sigma^2}{2\lambda_{\max}}\right)^t$$
$$\le \sum_{i>r_h}\lambda_i + \frac{2\lambda_{\max}g_0}{K\eta_{\mathrm{Oja}}\Delta_\Sigma^2}.$$

This proves (42) and the stated contraction. □

Under a warm start and eigengap, Proposition C.1 establishes contraction of the principal-subspace gap for a fixed signal while preserving the estimator filtration. The resulting Euclidean residual bound connects to the relevant branch risk through its operator weight. For example, if the operator weight satisfies $0 \preceq B \preceq L_B I$, then $\mathcal{R}_{Q^k}(B, X) \le L_B\chi^k\left\|R_h^k X\right\|_F^2$ by (17) and (5). This comparison identifies which branch the fixed-signal certificate controls.

The update uses a factorized implementation. For $X_k^{\mathrm{hist}} \in \mathbb{R}^{m\times n_{\mathrm{sig}}}$, it uses an $n_{\mathrm{sig}} \times r_h$ buffer for $(X_k^{\mathrm{hist}})^\top U_h^k$, an $m \times r_h$ buffer for the lifted product, and an $r_h \times r_h$ buffer for the tangent step and thin QR. Reusing these buffers requires $mr_h + n_{\mathrm{sig}}r_h + r_h^2$ additional scalars up to QR workspace constants, alongside the persistent mr_h-scalar basis.

C.2 Predictable History/Fresh Policy Selection

Proposition C.2 (Predictable history/fresh selector)**.** *Let $\upsilon \in \{\mathrm{hist}, \mathrm{fresh}\}$ index two unbiased production policies. Let $\mathcal{F}_k$ contain the history before the iteration-k selector draw, and suppose the $\mathcal{F}_k$-measurable risks $\ell_{k,\upsilon}$ admit pilot estimates with $0 \le \widehat{\ell}_{k,\upsilon} \le L_{\mathrm{sel}}$ and $\mathbb{E}[\widehat{\ell}_{k,\upsilon} \mid \mathcal{F}_k] = \ell_{k,\upsilon}$. Initialize $w_{0,\upsilon} = 1$. At iteration k, set $\pi_{k,\upsilon} = w_{k,\upsilon}/\sum_{\upsilon'} w_{k,\upsilon'}$, and draw $\upsilon_k \sim \pi_k$. Then update $w_{k+1,\upsilon} = w_{k,\upsilon}\exp(-\eta_{\mathrm{sel}}\widehat{\ell}_{k,\upsilon})$. Here υ_k is conditionally independent of the current pilots. Conditional on $\mathcal{F}_k$ and υ_k, the selected policy draws its production factors and oracle samples independently of the pilots, with the branchwise independence required by Assumption 3.2. Then the*

selected production law preserves the conditional mean and independence contracts of Assumption 3.2, and for every $K \geq 1$,

$$\mathbb{E}\sum_{k=0}^{K-1} \ell_{k,\upsilon_k} - \min_{\upsilon} \sum_{k=0}^{K-1} \mathbb{E}\ell_{k,\upsilon} \leq \frac{\log 2}{\eta_{\mathrm{sel}}} + \frac{\eta_{\mathrm{sel}} K L_{\mathrm{sel}}^2}{8}. \tag{46}$$

If $L_{\mathrm{sel}} > 0$, choosing $\eta_{\mathrm{sel}} = \sqrt{8\log 2/(KL_{\mathrm{sel}}^2)}$ gives average excess risk at most $L_{\mathrm{sel}}\sqrt{\log 2/(2K)}$. If $L_{\mathrm{sel}} = 0$, all risks vanish.

Proof. The probabilities π_k depend only on past pilot feedback. Enlarge $\mathcal{F}_k$ to include υ_k and any completed pilot information. Conditional on this enlarged filtration, the chosen factor law retains mean identity and the production oracle remains independent of the production factors. Thus the conditional estimator calculations remain valid, and the updated weights are first used at iteration $k+1$.

Set $W_k^{\mathrm{sel}} = \sum_{\upsilon} w_{k,\upsilon}$. For each realized pilot vector, the bounded exponential-moment inequality gives

$$\log \frac{W_{k+1}^{\mathrm{sel}}}{W_k^{\mathrm{sel}}} = \log \sum_{\upsilon} \pi_{k,\upsilon} \exp(-\eta_{\mathrm{sel}} \widehat{\ell}_{k,\upsilon}) \leq -\eta_{\mathrm{sel}} \sum_{\upsilon} \pi_{k,\upsilon} \widehat{\ell}_{k,\upsilon} + \frac{\eta_{\mathrm{sel}}^2 L_{\mathrm{sel}}^2}{8}.$$

For any fixed policy υ, $W_K^{\mathrm{sel}} \geq w_{K,\upsilon}$ and $W_0^{\mathrm{sel}} = 2$, so

$$\log \frac{W_K^{\mathrm{sel}}}{W_0^{\mathrm{sel}}} \geq -\eta_{\mathrm{sel}} \sum_{k=0}^{K-1} \widehat{\ell}_{k,\upsilon} - \log 2.$$

Sum the upper bound and compare it with this lower bound:

$$\sum_{k=0}^{K-1} \sum_{\upsilon'} \pi_{k,\upsilon'} \widehat{\ell}_{k,\upsilon'} - \sum_{k=0}^{K-1} \widehat{\ell}_{k,\upsilon} \leq \frac{\log 2}{\eta_{\mathrm{sel}}} + \frac{\eta_{\mathrm{sel}} K L_{\mathrm{sel}}^2}{8}.$$

Conditional unbiasedness and the independence of υ_k give

$$\mathbb{E}\sum_{\upsilon} \pi_{k,\upsilon} \widehat{\ell}_{k,\upsilon} = \mathbb{E}\sum_{\upsilon} \pi_{k,\upsilon} \ell_{k,\upsilon} = \mathbb{E}\ell_{k,\upsilon_k}.$$

Taking expectations and minimizing over the fixed comparator proves (46). For $L_{\mathrm{sel}} > 0$, choose $\eta_{\mathrm{sel}} = \sqrt{8\log 2/(KL_{\mathrm{sel}}^2)}$. After division by K, the excess risk is at most $L_{\mathrm{sel}}\sqrt{\log 2/(2K)}$. If $L_{\mathrm{sel}} = 0$, all risks vanish. □

Proposition C.2 formalizes a safe optional selector between history-aware and fresh factors. To connect its bound with stationarity, choose $\ell_{k,\upsilon}$ to be the conditional weighted risk $q_V \nu_{V,k}(\upsilon) + q_S \nu_{S,k}(\upsilon) + q_W \nu_{W,k}(\upsilon)$ in (41), provided the pilot assumptions hold for this quantity. The selector controls the noise contribution along its realized state sequence, with both fixed policies evaluated at those same states. All pilot queries and buffers are included in the resource budget.

D Memory and Computational Analysis

D.1 Non-GQA Memory Derivation

We use the same non-GQA SwiGLU decoder-block proxy as BROS (Zhang et al., 2026), for the analyzed role map with one shared history basis. Let n, s, b, h denote hidden size, sequence length, micro-batch size, and attention heads, and let r_{OPR} and r_h denote the uniform total sketch width and its history rank. Other configurations need their own history-storage count. The four grouped trainable matrices have shapes $(n, 3n)$, (n, n), $(n, 16n/3)$, and $(8n/3, n)$. Thus

$$\sum_{\ell} m_\ell n_\ell = 12n^2, \qquad \sum_{\ell} m_\ell = \frac{17}{3}n, \qquad \sum_{\ell} n_\ell = \frac{31}{3}n, \qquad \sum_{\ell}(m_\ell + n_\ell) = 16n. \tag{47}$$

The full lower trainable tuple therefore occupies $12n^2$ scalar slots, and the lower plus auxiliary states occupy $24n^2$.

For the saved-activation proxy, BROS gives $A_{\text{full}} = 15bsn + 2bhs^2$. The parameter-input activations associated with the four left dimensions occupy $bs\sum_\ell m_\ell = \frac{17}{3}bsn$ slots. The common non-projected part after replacing these inputs by coordinate inputs is $A_0 := A_{\text{full}} - \frac{17}{3}bsn = \frac{28}{3}bsn + 2bhs^2$.

A single factor of width w contributes four $bs \times w$ projected-input blocks. Let $r_q := \max_\ell r_{q,\ell}$ for each lower-space branch. OPR-BROS's mixed Hessian branch keeps independent input and output factors live, so for uniform branch width $r_q = r_{\text{OPR}}$, $A_{\text{OPR}} = A_0 + 4bs(r_i + r_o) = A_0 + 8bsr_{\text{OPR}}$. More generally, under the same sequential saved-activation convention, $A_{\text{OPR}} \le A_0 + 4bs\max\{r_g, r_u, r_j, r_i + r_o\}$.

For each layer, the mixed Hessian stage keeps one input factor, one output factor, and at most one additional first-order factor/coordinate buffer live. Thus, with $r_{\max,\ell} := \max_q r_{q,\ell}$, $M_{\text{production}} \le \sum_{\ell=1}^{L}(3m_\ell r_{\max,\ell} + 3r_{\max,\ell}n_\ell)$, which becomes $48nr_{\text{OPR}}$ under uniform width. The persistent history bases use $\sum_\ell m_\ell r_{h,\ell}$ slots. With the chronology production$(\boldsymbol{Z}^k) \to \boldsymbol{Z}^{k+1} \to$ Oja$(\boldsymbol{Z}^{k+1}) \to U_{h,\ell}^{k+1}$, the Oja signal reuses the current auxiliary-state storage. Its temporary buffers satisfy $M_{\text{update}} \le \sum_{\ell=1}^{L}(m_\ell r_{h,\ell} + n_\ell r_{h,\ell} + r_{h,\ell}^2)$, which becomes $16nr_h + 4r_h^2$ under uniform rank. Sequential release of the production and Oja buffers yields the peak-memory bound

$$M_{\text{peak}}^{\text{OPR}} \le 24n^2 + \frac{17}{3}nr_h + \max\left\{A_0 + 8bsr_{\text{OPR}} + 48nr_{\text{OPR}},\ 16nr_h + 4r_h^2\right\}. \tag{48}$$

When the memory proxy is smaller than BROS. The exact comparison with the BROS row of Table 5 is

$$\frac{17}{3}nr_h + \max\left\{A_0 + 8bsr_{\text{OPR}} + 48nr_{\text{OPR}},\ 16nr_h + 4r_h^2\right\} < A_0 + 4bsr_{\text{BROS}} + \frac{62}{3}r_{\text{BROS}}n. \tag{49}$$

When production dominates the workspace maximum, (49) becomes

$$\frac{r_{\text{OPR}}}{r_{\text{BROS}}} < \frac{4bs + (62/3)n}{8bs + 48n + (17/3)n(r_h/r_{\text{OPR}})}.$$

At the illustrative $bs = n$ used in the BROS memory discussion, the right-hand side is 0.419 for $r_h = r_{\text{OPR}}/2$ and 0.430 for $r_h = r_{\text{OPR}}/4$. Both cases have a positive fresh width. If $r_{\text{BROS}}/n = 0.25$, they require approximately $r_{\text{OPR}}/n < 0.105$ and < 0.107, respectively. These thresholds specify a memory-budget constraint to combine with the target optimization-quality requirement.

D.2 Explicit FLOP Derivation and Baseline Comparison

Counting convention. We count one multiply plus one add as two FLOPs and write $t = bs$ for the number of tokens. The four matrices have shapes (m_ℓ, d_ℓ) equal to $(n, 3n)$, (n, n), $(n, 16n/3)$, and $(8n/3, n)$. The decoder input is fixed, and x enters only through loss weights after the decoder. We count principal dense GEMMs in explicit forward-tangent and reverse-tangent sweeps, prune structurally zero tangents, and reuse primal activations and adjoints. All coordinate variables are evaluated at zero, and the schedule retains the coordinate-gradient outputs required by the derivative calls. The proxy covers dense decoder differentiation, projections and lifts, and Oja arithmetic. Factor generation, elementwise decoder kernels, loss/head readout, and direct $\nabla_x f$ evaluation are additional components of an end-to-end implementation cost.

First-order passes. The forward matrix products cost $24tn^2$. With the first block's input fixed, input-adjoint propagation through matrices 2–4 costs $18tn^2$. Full weight gradients cost another $24tn^2$. The two attention matrix multiplications cost $4tsn$ in the forward pass and $8tsn$ in reverse. Consequently, $F_{\text{fwd}} = 24tn^2 + 4tsn$ and $F_{\text{full}} = 66tn^2 + 12tsn$. For a left factor $P \in \mathbb{R}^{m\times r}$, the products XP and $(XP)^\top\bar{Y}$ replace the full weight gradient. Here $\bar{Y}$ denotes the adjoint of the matrix output. Their total cost over all four matrices is $2tr\sum_\ell(m_\ell + d_\ell) = 32tnr$, giving $F_{\text{proj}}(r) = 42tn^2 + 32tnr + 12tsn$.

Bi-sketch HVP and independent JVP. For one matrix, write its coordinate map as $Y = XW + (XP_o)A + (XP_i)B$, where $A = B = 0$ and the input-coordinate direction is V. A dot denotes differentiation along B in direction V. The required tangent equations are

$$\dot{Y} = \dot{X}W + (XP_i)V, \qquad \dot{\bar{X}} = \dot{\bar{Y}}W^\top + (\bar{Y}V^\top)P_i^\top,$$
$$\dot{G}_o = (\dot{X}P_o)^\top\bar{Y} + (XP_o)^\top\dot{\bar{Y}},$$

Table 5: **Non-GQA lower-network peak-memory proxy for one trainable decoder block.** Counts measure lower-network storage in scalar slots; upper-variable storage and datatype sizes enter the conversion to a full-device byte count. The first four baseline rows follow the accounting of BROS (Zhang et al., 2026). The two OPR-BROS workspace entries are the production and Oja-update phases, respectively.

Method	States	Activations	Grad./workspace	H/JVP scale	Peak proxy
MA-SOBA (Chen et al., 2024)	$24n^2$	$15bsn + 2bhs^2$	$24n^2$	full H/JVP	$48n^2 + 15bsn + 2bhs^2$
FdeHBO (Yang et al., 2023)	$48n^2$	$15bsn + 2bhs^2$	$24n^2$	full FD dirs.	$72n^2 + 15bsn + 2bhs^2$
Penalty	$12n^2$	$15bsn + 2bhs^2$	$12n^2$	no aux. H/JVP	$24n^2 + 15bsn + 2bhs^2$
BROS (Zhang et al., 2026)	$24n^2$	$A_0 + 4bsr_{\mathrm{BROS}}$	$(62/3)r_{\mathrm{BROS}}n$	$(L+1)$ HVP, 1 JVP	$24n^2 + A_0 + 4bsr_{\mathrm{BROS}} + (62/3)r_{\mathrm{BROS}}n$
OPR-BROS	$24n^2 + (17/3)nr_h$	$A_0 + 8bsr_{\mathrm{OPR}}$	$(48nr_{\mathrm{OPR}},\ 16nr_h + 4r_h^2)$	1 HVP, 1 JVP	$M^{\mathrm{OPR}}_{\mathrm{peak}}$

where G_o is the output-coordinate gradient. The dense forward and reverse terms each contribute $18tn^2$. The input injection, reverse injection, first output term, and second output term contribute, respectively, $32tnr$, $24tnr$, $24tnr$, and $32tnr$. These counts include new products with P_i and P_o, which differ from the first-order factor P_g. The attention tangent and adjoint-tangent products add $24tsn$, so the incremental HVP cost is $H_{\mathrm{OPR}} = 36tn^2 + 112tnr + 24tsn$. The independent factor P_j requires a separate forward tangent to compute the mixed derivative of the per-example losses, costing $J_{\mathrm{OPR}} = 18tn^2 + 32tnr + 8tsn$.

Projections and history update. Each full-tuple projection or lift costs $2r\sum_\ell m_\ell d_\ell = 24rn^2$. The five interfaces $q \in \{g, u, i, o, j\}$ contribute $120rn^2$. For one Oja block, the products $Z^\top U$, $Z(Z^\top U)$, $U^\top B$, and $U(U^\top B)$, together with the tangent update and explicit-basis Householder thin QR, cost $4mdr_h + 8mr_h^2 - (4/3)r_h^3 + 3mr_h$. The shared basis is updated once per block, giving $F_{\mathrm{Oja}} = 48n^2r_h + (136/3)nr_h^2 - (16/3)r_h^3 + 17nr_h$. Adding $2F_{\mathrm{proj}}(r)$, the HVP, the JVP, and these updates yields

$$F_{\mathrm{OPR}} = 138tn^2 + 208tnr + 56tsn + 120rn^2 + 48n^2r_h + \frac{136}{3}nr_h^2 - \frac{16}{3}r_h^3 + 17nr_h. \tag{50}$$

A conservative BROS lower bound. The multilayer BROS estimator uses four block-masked main HVPs and one global rank-one probe HVP (Zhang et al., 2026). For the four masks, the dense forward-tangent costs are $18tn^2$, $16tn^2$, $(16/3)tn^2$, and 0. Each reverse-tangent sweep costs $18tn^2$, and the global probe adds $36tn^2$ in dense forward/reverse propagation. These terms sum to $(442/3)tn^2$. The attention contribution is $24tsn$ for the first mask, $8tsn$ for each later mask, and $24tsn$ for the probe, totaling $72tsn$. Later masks have zero forward tangent at attention but still propagate nonzero adjoint tangents through it.

To obtain a conservative lower bound, we assign zero cost to rank-dependent HVP work, probe/correction algebra, JVP evaluation, and additional primal/adjoint tape construction. This favorable analytical accounting accommodates rank-one probe shortcuts and maximal reuse when bounding the baseline cost. The lower lift, upper-y lift, projection of Z, and final accumulated HVP-output lift still give four full-tuple products, costing $96rn^2$. Together with two projected first-order passes, the retained work gives

$$F_{\mathrm{BROS}} \geq \frac{694}{3}tn^2 + 64tnr + 96tsn + 96rn^2. \tag{51}$$

Equation (51) gives a lower bound under the dense masked/probe schedule specified above.

The following result shows when the saved dense derivative work exceeds the extra independent-factor algebra and Oja update at the same width. It compares the full OPR-BROS proxy with the deliberately reduced BROS count.

Proposition D.1 (Equal-width decoder FLOP advantage)**.** *Under the counting model above, let both methods use uniform width r. If $t = bs \geq r$, $1 \leq r \leq n/4$, and $0 \leq r_h \leq r/2$, then*

$$F_{\mathrm{BROS}} - F_{\mathrm{OPR}} \geq \frac{rn(13n - 17)}{2} + 40tsn > 0.$$

Proof. Subtracting (50) from (51) gives

$$F_{\mathrm{BROS}} - F_{\mathrm{OPR}} \geq \frac{280}{3}tn^2 - 144tnr + 40tsn - 24rn^2 - F_{\mathrm{Oja}}. \tag{52}$$

Since $r \le n/4$ and $t \ge r$, the first two terms are at least $(172/3)rn^2$. Using $r_h \le r/2$ and dropping the negative cubic term,

$$F_{\mathrm{Oja}} \le 24n^2r + \frac{34}{3}nr^2 + \frac{17}{2}nr \le \frac{161}{6}n^2r + \frac{17}{2}nr.$$

Substitution into (52) gives the claimed lower bound. It is strictly positive because $1 \le r \le n/4$ implies $n \ge 4$. □

Other baselines. For a full-space HVP, the input injection, forward propagation, reverse propagation, reverse injection, and two weight-gradient derivative terms cost, respectively, $24tn^2$, $18tn^2$, $18tn^2$, $18tn^2$, $18tn^2$, and $24tn^2$. Attention adds $24tsn$. Granting free tape/JVP reuse to MA-SOBA as well gives the optimistic bound $F_{\mathrm{MA}} \ge 2F_{\mathrm{full}} + 120tn^2 + 24tsn = 252tn^2 + 48tsn$. For FdeHBO (Yang et al., 2023), the published separate-sample schedule evaluates two lower gradients and two auxiliary residuals at consecutive states. Each residual uses two finite-difference lower gradients and one upper-y gradient, giving eight full first-order passes. Its two mixed finite differences additionally require four decoder forwards when x only weights losses. Thus its steady-state count is $F_{\mathrm{Fde}} = 8F_{\mathrm{full}} + 4F_{\mathrm{fwd}} = 624tn^2 + 112tsn$. The direct $\nabla_x f$ evaluations remain excluded, as for every table row.

Compute–storage trade-off. Equation (52) certifies a per-iteration decoder FLOP advantage whenever its right-hand side is positive, including rank choices beyond Proposition D.1. End-to-end time-to-quality combines this arithmetic cost with iteration count and implementation overhead. The cache-retaining FLOP schedule and the sequential-release memory schedule in Appendix D.1 describe two execution choices: retained caches reduce recomputation, while sequential release reduces retained storage.

LancBiO memory note. At Krylov dimension j, algorithm-level persistent lower-side storage includes $K_j^{\mathrm{Lanc}} \in \mathbb{R}^{d_y \times j}$, a symmetric tridiagonal T_j, the lower state, auxiliary vector, epoch anchor, and cached Hessian-vector product. A minimal persistent core is therefore $(j+4)d_y + (2j-1)$ scalar slots, or $12(j+4)n^2 + 2j - 1$ under the same one-block mapping $d_y = 12n^2$. This expression counts persistent algorithmic state. Peak memory additionally includes transient derivative and activation workspaces, so we present this state count separately from Table 5.

D.3 Comparability and Budget Interpretation

The memory and FLOP proxies quantify the resource costs of a chosen sketch configuration. Its rank-dependent moment bounds determine the corresponding stationarity certificate. Together, these quantities characterize the trade-off among sketch width, storage, per-iteration arithmetic, and optimization quality. Proposition D.1 supplies the decoder-arithmetic comparison for the specified four-block schedule and rank range.

An equal-budget empirical comparison reports sketch widths, measured peak memory, HVP/JVP counts or time, Oja and factor-generation time, and final quality. The practical efficiency balance depends on the relative costs of operator evaluation, independent-factor storage, and history adaptation in each model and execution backend.

E Additional Experiments and Diagnostics

We organize the evidence by setting. Full denotes the task-adapted full-space MA-SOBA-style control, while the Pile experiment names its archived full-space run MA-SOBA. Fresh uses fully fresh independent sketches, corrected BROS uses shared projections and CorrHVP, and OPR-BROS adds past-only learned principal–random factors. Fresh and OPR-BROS retain exact first-order gradients in the small and native-network experiments. Values after ± are sample SDs; brackets denote 95% confidence intervals. The native trajectory table also includes a separate port of SOBA's published recursion. We compare methods within each experimental cohort; the Pile archive configurations are documented below.

E.1 PILE-280M DATA-MIXTURE PROXY

Archived Proxy. We optimize a 17-domain Pile-280M mixture for 10,000 steps on four A100s (seed 42, global batch 128, projection width 256), with one model initialization and a deterministic lower/upper training tape. Evaluation occurs every 100 steps. The plotted upper-level loss is the fixed-panel worst-domain excess validation CE, $\max_d(\mathrm{CE}_d^{\mathrm{val}} - \mathrm{CE}_d^{\mathrm{ref}})$. O-Fresh uses the archived time-optimized setting; the remaining configurations use the common v12 setting. Step time is measured after ten warmup steps. The memory metric sums per-GPU peak allocations over the training phase, with evaluation and checkpoint phases outside the measurement window.

Figure 1(a,b) and Table 3 report the archived Pile configurations. O-Fresh corresponds to `obros-fresh`, with recorded History use equal to zero, zero Oja updates, and branch-independent sampling disabled. B-NoProbe corresponds to `bros-no-probe`, with CorrHVP disabled. The reported loss, timing, and memory measurements characterize these explicitly specified archive settings.

E.2 MNIST AND FASHION-MNIST HYPER-CLEANING

MNIST Trajectories. We train a linear-softmax lower model on 20,000 noisy MNIST examples using 5,000 clean validation examples, seeds 42–46, and 6,000 steps on one RTX 4090. Sketched methods use width 256; OPR-BROS uses $r_h = 64$. The plotted metric is unregularized validation CE at the current lower iterate, with the test split held out. Timing covers optimization and scheduled validation, from completion of setup to the start of post-run diagnostics.

Figure 1(c,d) reports the validation trajectories. At step 6,000, OPR-BROS has CE 0.3182 versus 0.3173, 0.3192, and 0.3377 for Full, Fresh, and BROS. Its measured 19.55 s is below those three controls' 25.29, 31.09, and 48.22 s, respectively, within this dense-observation protocol.

Fashion-MNIST TTT90. The separate frozen-feature identity-v4 task uses five paired seeds, 2,000 steps, projected dimension 512, and 1,024 hidden rows, with the test split held out. OPR-BROS uses $(r_{\mathrm{OPR}}, r_h) = (16, 4)$, 100-step Oja warmup, ten-step updates, and freezing at step 200. TTT90 interpolates the time to 90% of each seed's Full-defined validation-CE gain. We combine the two execution orders geometrically within each seed, then report the paired geometric mean with seed-clustered bootstrap intervals.

Table 4 reports the common-target ratios. All seeds reach the target, with OPR-BROS/Fresh TTT90 ratio $0.4840\,[0.4777, 0.4889]$. The Full and BROS ratios are also below one. This frozen-feature experiment quantifies the validation time-to-quality benefit of history reuse in upper-level optimization.

E.3 NATIVE CIFAR-10 AND CIFAR-100

The two native tasks use a ResNet-18 feature gate on CIFAR-10 and clean-validation sample weights for a SmallViT on noisy-label CIFAR-100.

Independent Test. On A100 GPUs, seeds 42–44 run 30,000 CIFAR-10 steps or use accuracy-based early stopping on CIFAR-100. Each validation-selected checkpoint is tested once. Fresh and OPR-BROS use nominal $\rho = 0.50$, while the generic-autograd BROS reconstruction uses $\rho = 0.25$; blockwise widths are rounded independently.

Across three seeds, OPR-BROS reaches mean test accuracies of 75.77% and 46.24% on CIFAR-10 and CIFAR-100, compared with Full's 75.90% and 46.47%. Fresh and OPR-BROS have the same recorded peak reserved memory on both tasks. The BROS rows use a generic-autograd reconstruction at $\rho = 0.25$, and Fresh and OPR-BROS use $\rho = 0.50$. The BROS backend varied across the archived runs. Table 6 reports the measured accuracy and memory of these implementations, with timing evaluated in the separate validation cohort below.

Validation Trajectories. A separate A100 cohort uses seeds 122–124 and scores the upper-optimization split every 500 steps. Accuracy-TTT95 is the time to 95% of each seed's Full-defined best validation gain, requiring two consecutive hits. The table reports three-seed means and whole-run allocated peaks; all listed methods hit the target in all three seeds. OPR-BROS assigns one quarter of each non-exact block's rounded width to History. This cohort supplies the pre-update auxiliary state

Table 6: Native CIFAR final-test accuracy and peak reserved memory (three seeds).

	CIFAR-10 / ResNet-18		CIFAR-100 / SmallViT	
Method	Test acc. (%)	Peak (GiB)	Test acc. (%)	Peak (GiB)
Full	75.90 ± 1.12	2.52	46.47 ± 0.62	12.06
BROS ($\rho = 0.25$)	76.54 ± 1.20	2.95	42.45 ± 0.29	16.53
Neumann-5	75.84 ± 1.22	2.46	45.49 ± 0.20	11.57
FdeHBO	69.80 ± 1.22	2.39	6.47 ± 3.72	6.03
Fresh ($\rho = 0.50$)	75.44 ± 1.22	2.51	46.33 ± 0.70	12.19
OPR-BROS ($\rho = 0.50$)	75.77 ± 1.28	2.51	46.24 ± 0.69	12.19

Z^k to the post-production Oja update. Algorithm 1 uses the updated state Z^{k+1} as its default signal; both schedules retain past-only history for the next iteration.

On CIFAR-100, Fresh and OPR-BROS both use 10.98 GiB, compared with Full's 12.39 GiB, a reduction of 11.4%. Their TTT95 values are 1.57, 1.58, and 1.46 hours for Fresh, OPR-BROS, and Full, respectively. On CIFAR-10, the corresponding peaks are 2.74, 2.86, and 2.71 GiB, with TTT95 values of 0.45, 0.51, and 0.42 hours. These measurements jointly report validation quality, timing, and memory use. Table 6 reports independent-test accuracy.

E.4 Frozen Qwen Pile-17

History Ablation. On our static 17-domain Pile mixture (Pile-17), constructed from The Pile (Gao et al., 2020), a frozen Qwen3-0.6B (Yang et al., 2025a) model has 163,840 trainable Low-Rank Adaptation (LoRA) (Hu et al., 2022) coordinates. Five paired seeds (14–18) run 200 steps on an RTX 4090 with matched initial states and data/target streams. The History and random-core arms each use independent bi-sketches with 16 core and 64 fresh residual directions. Their core directions come from a history-derived basis and a random basis, respectively. The prespecified endpoint averages the random-core minus History target meta-validation loss over all steps.

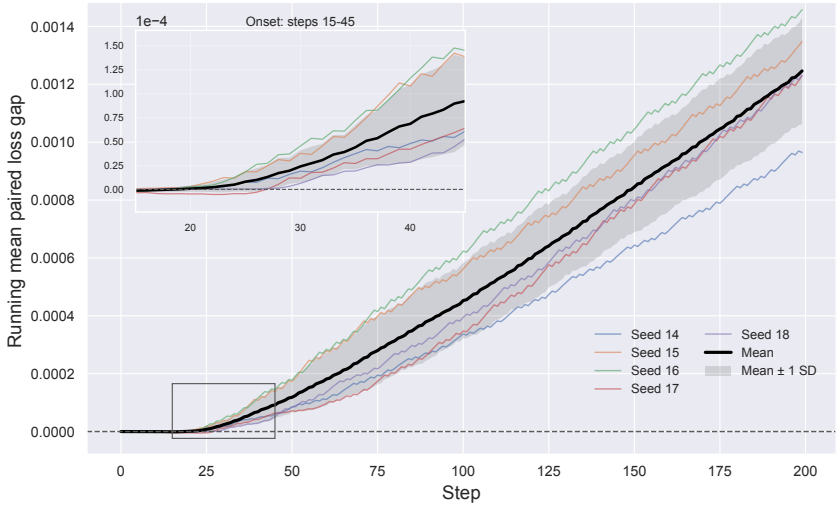


Figure 2: Frozen-Qwen Pile-17 running paired target-loss gap; positive favors History.

Figure 2 reports the expanding running mean of the paired target-loss gap. The black curve is the cross-seed mean, the shaded band is one sample SD, and the inset enlarges steps 15–45. All five whole-run seed means favor History, with a cross-seed mean of 0.001246 ± 0.000184 (sample SD).

The comparison isolates core-direction selection: both arms retain 16 core and 64 fresh residual directions, with the core supplied by either a learned history basis or a random basis. The endpoint averages target meta-validation loss over the 200-step static-mixture run, using the stream that

Table 7: Native CIFAR validation accuracy, TTT95, and allocated peak (three seeds).

Task	Method	Best val. acc. (%)	TTT95 (h)	Peak (GiB)
CIFAR-10	Full	76.95 ± 1.67	0.42	2.71
	SOBA paper recursion	75.73 ± 1.53	0.56	2.74
	Fresh	77.25 ± 1.15	0.45	2.74
	OPR-BROS	77.33 ± 0.47	0.51	2.86
CIFAR-100	Full	46.93 ± 0.27	1.46	12.39
	SOBA paper recursion	46.13 ± 0.62	1.37	12.36
	Fresh	45.74 ± 0.45	1.57	10.98
	OPR-BROS	46.52 ± 0.41	1.58	10.98

participates in upper-level optimization. At the terminal step, History has the lower target loss in four of the five seeds.

E.5 Controlled Synthetic Diagnostics

Fixed-State Moments. We use a 64-dimensional, 32-atom quadratic bilevel problem with lower Hessian $A + (1 + x^2)I$, where A is the empirical feature covariance. At fixed (x, y, z) with $x = 1$, the sketch width is 16, split into eight History and eight Fresh directions. Eight independent streams provide up to 8,192 Monte Carlo draws each. Input/output factors, the upper-gradient factor, and the sampled atom use separate streams. A shared-input/output variant provides a comparison for sketch dependence. History is learned from 512 past batches and frozen during evaluation.

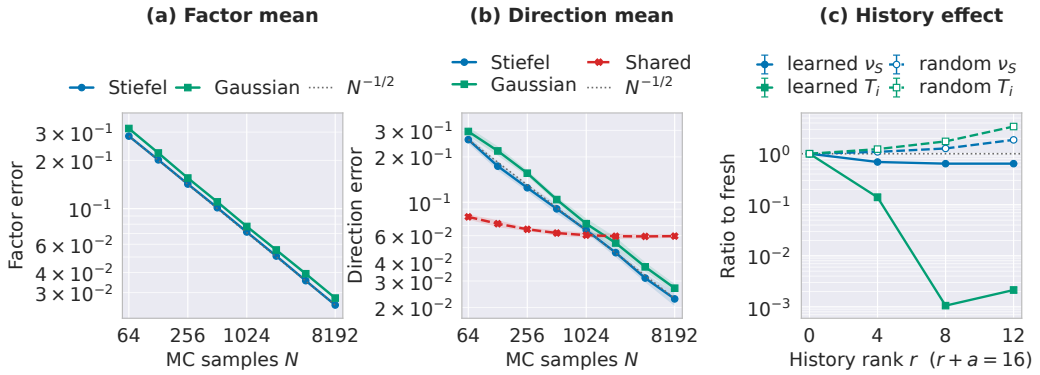


Figure 3: Fixed-state factor, direction-mean, and History-risk diagnostics.

Figure 3(a) shows first-moment error slopes -0.501 and -0.496 for Stiefel and Gaussian factors. With independent input/output factors, the pooled direction-mean relative errors in Figure 3(b) are 0.890% and 0.928%; the shared-factor control retains 3.747% and 4.460% bias. At History rank eight, learned/Fresh direction-risk and Hessian-input-residual ratios in Figure 3(c) are 0.6385 and 0.001056. These fixed-state diagnostics assess the implemented factor-moment identities, preservation of direction means, and history-dependent operator risks.

History Shift. A separate CPU quadratic uses $H[Z] = A_{\mathrm{sh}}Z$ for $Z \in \mathbb{R}^{64\times 4}$ at width 12, with $A_{\mathrm{sh}} = I + 3UU^\top$ and $Z = UC + 0.05E$ for independent Gaussian C, E. Here U spans the current eight-dimensional signal. A principal component analysis (PCA) History basis is learned from 512 independent past samples; the signal either continues or switches to an orthogonal subspace. We use eight problem instances and 512 sketch draws per instance, using Gaussian input factors and independent Gaussian Fresh output factors. Table 8 gives paired t intervals for the Hessian–vector-product mean-squared-error (HVP-MSE) ratio across problem instances.

Table 8: Fixed-state Gaussian History/Fresh HVP-MSE ratios at width 12.

Current signal	MSE ratio	95% Confidence Interval
Continuation	0.6253	$[0.5725, 0.6780]$
Orthogonal switch	1.5185	$[1.5176, 1.5194]$

Using PCA-derived histories, the fixed-state HVP-MSE ratios are 0.6253 under signal continuation and 1.5185 after an orthogonal switch. This diagnostic quantifies how alignment between historical and current signal subspaces governs operator-estimation accuracy.

The supplementary artifact retains run manifests, data identities, stopping records, timing definitions, and test-access logs.