ON THE CRITICAL BATCH SIZE FOR LOW PRECISION TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are increasingly trained with large batches and low-precision arithmetic to improve computational efficiency. The critical batch size (CBS) marks the onset of diminishing returns from larger batches, yet how this threshold changes with numerical precision remains poorly understood. We study the CBS of low-precision training through theoretical scaling laws for quantized stochastic gradient descent in sketched linear regression. Our analysis replaces the vector- and matrix-level error-alignment constraints of prior multiplicative quantization models with more practical bounds on the relative variance of independent, unbiased coordinate-wise rounding. Our theory predicts that lower precision increases the critical batch scale while preserving its scaling exponent with the data budget. The analysis explains this shift through a balance between quantization-induced noise and the cost of fewer optimization updates, accounting for parameter-rounding noise shared across a batch. We combine controlled regression experiments with matched-quality language-model comparisons to test the predicted scaling and its qualitative transfer to pretraining. This analysis motivates joint selection of numerical precision and batch size to exploit the useful range of data parallelism.

1 INTRODUCTION

How much data parallelism can low-precision training use efficiently? This question sits at the intersection of two central choices in large language model pretraining: numerical precision and global batch size. DeepSeek-V3 uses FP8 for the principal linear-layer matrix multiplications (DeepSeek-AI, 2024), while NVFP4 and UFP4 extend low-precision pretraining to four-bit arithmetic (NVIDIA et al., 2026; Zhao et al., 2026). At the same time, training recipes increasingly employ global batches of order 10^8 tokens (DeepSeek-AI, 2026; MiniMax et al., 2025). These developments motivate studying precision and batch size jointly: reducing the cost of an update is valuable only insofar as the resulting updates remain effective for learning.

The *critical batch size* (CBS) marks the onset of diminishing returns from batch scaling. Below this scale, increasing the batch reduces gradient noise and can reduce the number of updates needed to reach a target loss at roughly proportional rate. Beyond it, the savings in updates become increasingly small relative to the additional data processed per update (McCandlish et al., 2018; Shallue et al., 2019). For language-model pretraining, Zhang et al. (2025) report that the CBS scales primarily with the training data budget, with an approximately $D^{0.47}$ dependence in their experimental setting. This near-square-root dependence is an empirical fit, rather than a universal upper limit on batch size, and its applicability to large-batch low-precision training remains unresolved.

Precision-aware scaling laws characterize how quantization affects model quality as model size and data increase (Kumar et al., 2025; Zhang et al., 2026b), but they do not determine how batch size enters the scaling effects. In related work, larger batches have been shown to compensate for reduced backpropagation precision in some settings (Liu et al., 2025). These results, however, do not account for the fixed-budget trade-off: when the training data budget is fixed, increasing the batch size also reduces the number of parameter updates. This motivates our question:

How does numerical precision change the critical batch size and its scaling with data budget?

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.

Recent theoretical work on scaling laws has established Gaussian-sketched linear regression with stochastic gradient descent (SGD) as a useful and tractable proxy for several phenomenological features of deep learning, including the effects of batch size and numerical precision (Lin et al., 2024; 2025; Li et al., 2025; Yan et al., 2026; Zhang et al., 2026b). In particular, Li et al. (2025) establishes scaling laws under general batch size schedule while Zhang et al. (2026b) derives low-precision scaling laws under the matrix-level multiplicative quantization model with covariance-aligned errors. This quantization model captures the structured, scale-dependent perturbations induced by vector- and matrix-level quantization. However, it abstracts away the coordinate-wise rounding structure of practical quantizers.

We therefore study the critical batch size of low-precision training via one-pass, constant-step-size quantized mini-batch SGD in Gaussian-sketched linear regression, using a more practical model of independent, unbiased coordinate-wise rounding errors with bounded relative error variance and separate quantizers for data, sketches, features, labels, parameters, activations, and output gradients. Write the covariance eigenvalues as $\lambda_i \asymp i^{-a}$ and the target energy in those directions as $\mathbb{E}[\lambda_i w_i^{*2}] \asymp i^{-b}$. For a fixed sample budget $D = BN$ with batch size B and parameter updates N, our established scaling law gives the batch scaling prediction

$$B^* \eqsim \begin{cases} 1 + \bar{\epsilon}_p, & 1 < b < a, \\ \widetilde{\Theta}\left((1 + \bar{\epsilon}_a + \bar{\epsilon}_o + \bar{\epsilon}_p)^{a/\min\{b,2a+1\}} D^{\frac{\min\{b,2a+1\}-a}{\min\{b,2a+1\}}}\right), & 1 < a \le b. \end{cases} \tag{1}$$

Here $\bar{\epsilon}_a$, $\bar{\epsilon}_o$, and $\bar{\epsilon}_p$ bound the relative conditional variances of activation, output-gradient, and parameter quantization. This theoretical result implies that larger arithmetic-error levels increase the predicted batch scale while preserving its exponent in the data budget.

The mechanism is a precision-dependent balance between learning progress and statistical noise. In particular, generally representing the quantization error by q [1], the leading batch-dependent terms in the risk bound take the form

$$\underbrace{(B/D)^{(\min\{b,2a+1\}-1)/a}}_{\text{finite-update bias}} + \underbrace{(1+q)D^{1/a-1}B^{-1/a}}_{\text{effective noise}}. \tag{2}$$

A larger batch leaves fewer updates and therefore increases the finite-update bias, but reduces the effective noise. Quantization increases the noise coefficient, making noise reduction worthwhile over a larger range of batch sizes and shifting the balance toward larger batches. The data and target spectra determine the powers in this trade-off, so precision changes the batch-scale prefactor without changing its data-budget exponent in this regime. we summarize our contributions as follows:

- We develop excess risk upper bounds and corresponding batch-scaling laws under unbiased coordinate-wise quantization error model with bounded relative error variance. Separately, we establish a sharper bias bound without requiring commutativity or simultaneous diagonalization of the quantized and unquantized covariance operators compared with Zhang et al. (2026b) (see Theorem 4.1).
- We characterize the low-precision critical batch size as a function of data budget, covariance decay, target regularity, and arithmetic quantization, showing that quantization scales the prefactor of the critical batch size but preserves its data budget exponent and lower precision shifts the critical batch size upward (see Theorem 4.2).
- We validate the predicted critical batch size in empirical studies (see Section 4.3). Together, these results provide guidance for joint selecting numerical precision and batch size to exploit the useful range of data parallelism.

2 Related Work

Critical batch size. Early empirical studies characterize the diminishing returns of large-batch training and their dependence on optimizer and problem (McCandlish et al., 2018; Shallue et al., 2019; Naganuma et al., 2026). Theoretical explanations are complementary: smoothness-based analyses connect batch size to iteration and stochastic first-order oracle complexity (Imaizumi &

[1] q is exactly zero in full-precision settings.

Iiduka, 2024); noisy quadratic models expose curvature, momentum, and preconditioning (Zhang et al., 2019); and spectral regression analyses relate useful batch size to finite-budget generalization (Zhang et al., 2025). For language-model pretraining, Zhang et al. (2025) disentangle data and model size and define CBS via an operational target loss, while Merrill et al. (2025) measure local batch sensitivity during training and motivate batch-size warmup. Complementary scaling-law studies find that both loss-optimal and critical batch sizes scale primarily with training data rather than model size (Bergsma et al., 2025). Recent work studies scheduler-dependent batch selection (Zhou et al., 2026), predicts pretraining loss jointly from model size, batch size, and update count (Li & Maddison, 2026), and characterizes distinct benefits of Polyak and Nesterov momentum in one-pass regression (Wang et al., 2026). Fixed-token analyses further show that batch size is coupled to the optimization horizon: increasing the batch reduces stochastic noise but also reduces the number of parameter updates, leading to nontrivial token-optimal batch-size regimes (Shulgin et al., 2026; Islamov et al., 2026). These results show CBS depends on both its defining criterion and the training procedure. Our theory retains the fixed-data-budget risk perspective but makes arithmetic precision explicit, and our pretraining protocol follows the target-loss approach.

Low-precision training. Mixed-precision training combines inexpensive arithmetic with higher-precision quantities for stable optimization (Micikevicius et al., 2018). FP8-LM and subsequent trillion-token studies develop this for language models (Peng et al., 2023; Fishman et al., 2025), and DeepSeek-V3 uses FP8 in the forward, input-gradient, and weight-gradient matrix multiplications of linear layers (DeepSeek-AI, 2024). Recent FP4 approaches include spectral decomposition in Metis (Cao et al., 2026), large-scale NVFP4 pretraining (NVIDIA et al., 2026), and Ling's UFP4 recipe, which combines a uniform four-bit grid with Hadamard transforms and stochastic rounding of upstream gradients (Zhao et al., 2026). Alongside these systems developments, precision-aware scaling laws quantify accuracy–resource trade-offs (Kumar et al., 2025), while quantized optimization and stochastic-rounding analyses study finite-precision learning dynamics (Markov et al., 2023; Ozkara et al., 2025). The closest direct precision–batch comparison is Liu et al. (2025): they distinguish weight and activation quantization effects and show larger batches can compensate reduced backpropagation precision. Our contribution is not this compensation but a batch-scaling prediction that also accounts for reduced update count at fixed data and its dependence on spectral structure.

Theoretical scaling laws in linear regression. Linear and random-feature models provide tractable descriptions of how learning depends on data, model capacity, and training time (Bahri et al., 2024; Atanasov et al., 2024; Paquette et al., 2024; Bordelon et al., 2024; Lin et al., 2024; 2025; Yan et al., 2026; Li et al., 2025; Ding et al., 2025). A pivotal step toward finite-sample analysis is Lin et al. (2024), who derive joint compute, parameter, and data scaling laws for Gaussian-sketched linear regression. Subsequent work extends this framework: Lin et al. (2025) analyze data reuse (multi-pass SGD), showing that for relatively small multi-epoch count K, each epoch gives a linear gain in effective sample size, $N_{\text{eff}} \approx NK$; Yan et al. (2026) refine this for strongly convex or Zipf-distributed data, showing that for large K, the reuse rate N_{eff}/N plateaus at a problem-dependent value growing with N; and Li et al. (2025) establish functional scaling laws and analyze how learning-rate schedules shape these behaviors. Quantization extends this framework differently: Zhang et al. (2026a) distinguish effects of quantizing different computational sites, and Zhang et al. (2026b) study precision jointly with sketch dimension and sample size. We replace multiplicative covariance-alignment constraints with independent scalar rounding and ask which batch scale is selected at a fixed sample budget. This matters technically: coordinate-wise errors produce diagonal covariance terms that need not share data eigenspaces. Our analysis retains these terms and the covariance of shared parameter-rounding noise, yielding a batch-dependent risk bound and the associated precision-dependent scaling prediction.

3 Problem Setup

We assume access to M-dimensional sketched covariates and their responses, that is, $(\mathbf{S}\mathbf{x}, y)$, where $\mathbf{S} \in \mathbb{R}^M \times \mathbb{H}$ is a fixed sketch matrix, $\mathbf{x} \in \mathbb{H} \subseteq \mathbb{R}^p$ is the data vector and $\mathbb{H}$ is a Hilbert space that is either finite-dimensional or countably infinite-dimensional. We focus on the Gaussian sketch matrix (Lin et al., 2024; 2025; Chen et al., 2025; Ding et al., 2025; Zhang et al., 2026b). That is, entries of $\mathbf{S}$ are independently sampled from $\mathcal{N}(0, 1/M)$. We then consider linear model with M trainable parameters given by: $f_{\mathbf{v}} : \mathbb{H} \to \mathbb{R}, \quad \mathbf{x} \to \langle \mathbf{v}, \mathbf{S}\mathbf{x} \rangle$, where $\mathbf{v} \in \mathbb{R}^M$ are the trainable parameters. We consider training $f_{\mathbf{v}}$ via one-pass quantized SGD with constant step size and batch size schedule.

The parameter $\mathbf{v}_t$ is updated as follows: for $t = 1, ..., N$,

$$\mathbf{v}_t = \mathbf{v}_{t-1} + \frac{\gamma}{B}\mathcal{Q}_f\left(\mathcal{Q}_s(\mathbf{S})\mathcal{Q}_d(\mathbf{X}_t)^\top\right)\mathcal{Q}_o(\mathcal{Q}_l(\mathbf{y}_t) - \mathcal{Q}_a(\mathcal{Q}_f\left(\mathcal{Q}_s(\mathbf{S})\mathcal{Q}_d(\mathbf{X}_t)^\top\right)^\top\mathcal{Q}_p(\mathbf{v}_{t-1}))),$$

where $(\mathbf{X}_t, \mathbf{y}_t) \in \mathbb{R}^{B\times p} \times \mathbb{R}^B$ are independent sample batches with i.i.d. samples inside, B is the batch size, γ is the constant step size, $\mathcal{Q}_d, \mathcal{Q}_s, \mathcal{Q}_f, \mathcal{Q}_l, \mathcal{Q}_p, \mathcal{Q}_a, \mathcal{Q}_o$ are coordinate-wise quantization operations [2] for data, sketch matrix, feature, labels, model parameters, activations and output gradients respectively. Without loss of generality, we assume the initial parameter is $\mathbf{v}_0 = 0$. We also assume the sketch quantization $\mathcal{Q}_s$ is applied independently to $\mathbf{S}$ and is independent among samples, ensuring the independence of quantized samples given the sketch $\mathbf{S}$.

Our goal is first to analyze the population risk [3]

$$\mathcal{R}_M(\mathbf{v}) := \frac{1}{2}\mathbb{E}[(\langle \mathbf{S}\mathbf{x}, \mathbf{v}\rangle - y)^2], \quad \mathbf{v} \in \mathbb{R}^M$$

for $\overline{\mathbf{v}}_N := \frac{1}{N}\sum_{t=0}^{N-1}\mathbf{v}_t$, and then derive the critical batch size under fixed training budgets.

3.1 Quantization Model

For all quantization operations in quantized SGD, we employ the stochastic quantization method (Markov et al., 2023; Modoranu et al., 2024; Ozkara et al., 2025; Zhang et al., 2026a;b), which unbiasedly rounds values using randomly adjusted probabilities. We summarize this in the following assumption.

Assumption 3.1 (Unbiased quantization)**.** For $i \in \{d, s, f, l, p, a, o\}$, $\mathcal{Q}_i$ is unbiased, i.e., for any $u \in \mathbb{R}$, the quantization operation is unbiased: $\mathbb{E}\left[\mathcal{Q}_i(u) \mid u\right] = u$.

In practice, floating-point quantization (e.g., FP8, FP32) employs a value-aware bin length via its exponent and mantissa bits (e.g., E4M3 format in FP8) (Kuzmin et al., 2022). This structure causes the quantization error to scale with the magnitude of the value itself. We are thus motivated to establish the following multiplicative quantization model, which is consistent with Tang et al. (2026); Zhang et al. (2026c).

Definition 3.2 (Multiplicative quantization)**.** We call the unbiased quantization to $u \in \mathbb{R}$ is $\bar{\epsilon}$-multiplicative if the conditional second moment of quantization error is proportional to the product of raw data itself, i.e., $\mathbb{E}[(\mathcal{Q}(u) - u)^2 \mid u] \le \bar{\epsilon}u^2$.

In the multiplicative quantization model, we assume the absence of underflow and overflow Tang et al. (2026); Zhang et al. (2026a;b;c). This is well justified: low-precision LLM training commonly employs engineering techniques such as per-tensor or per-channel scaling, which ensure that post-quantization values remain within the representable range (Peng et al., 2023; Fishman et al., 2025).

We would like to remark that compared with the vector- and matrix-level models in Zhang et al. (2026a;b) which impose error constraints on the full vector or matrix, our model is more directly aligned with coordinate-wise stochastic rounding: it controls the relative variance of each scalar quantization error. In particular, current quantization retains the heterogeneity of the rounding errors across scalar entries, leading to explicit diagonal terms in the risk recursion.

3.2 Data Model

We then state the regularity assumptions on the data distribution, which align with those common in prior works (Zou et al., 2021; Wu et al., 2022a;b; 2023). As low-precision training is performed on quantized feature $\tilde{\mathbf{x}}^{(q)} = \mathcal{Q}_f\left(\mathcal{Q}_s(\mathbf{S})\mathcal{Q}_d(\mathbf{x})\right)$, we formulate these assumptions on the low-precision feature format following Zhang et al. (2026a;b).

Assumption 3.3 (Data covariance)**.** Let $\mathbf{H} := \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ be the data covariance and $\mathbf{H}_f^{(q)} := \mathbb{E}[\tilde{\mathbf{x}}^{(q)}(\tilde{\mathbf{x}}^{(q)})^\top]$ be the quantized feature covariance. Assume that $\mathrm{tr}(\mathbf{H}), \mathrm{tr}(\mathbf{H}_f^q)$ and all entries of $\mathbf{H}, \mathbf{H}_f^{(q)}$ are finite. For convenience, we assume that $\mathbf{H}$ is strictly positive definite.

[2] These quantization operations are mutually independent and each $\mathcal{Q}_i$ is independently coordinate-wise. Different index represents different quantization error level, namely, quantization bit width.

[3] In this paper, all expectations are conditioned on $\mathbf{S}$.

Let $\mathbf{H} = \sum_i \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$ be the eigen-decomposition of $\mathbf{H}$, where $\{\lambda_i\}_{i=1}^\infty$ are the eigenvalues of $\mathbf{H}$ sorted in non-increasing order and $\mathbf{v}_i$ are the corresponding eigenvectors. We denote $\mathbf{H}_{0:k} := \sum_{i=1}^k \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$, $\mathbf{H}_{k:\infty} := \sum_{i>k} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$, $\mathbf{I}_{0:k} := \sum_{i=1}^k \mathbf{v}_i \mathbf{v}_i^\top$, $\mathbf{I}_{k:\infty} := \sum_{i>k} \mathbf{v}_i \mathbf{v}_i^\top$. Similarly, we denote the eigen-decomposition of $\mathbf{H}_f^{(q)}$ as $\mathbf{H}_f^{(q)} = \sum_i \tilde{\lambda}_i^{(q)} \mathbf{v}_i^{(q)} \mathbf{v}_i^{(q)\top}$ and correspondingly obtain $\mathbf{H}_{f,0:k}^{(q)}, \mathbf{H}_{f,k:\infty}^{(q)}, \mathbf{I}_{f,0:k}^{(q)}, \mathbf{I}_{f,k:\infty}^{(q)}$, where $\{\tilde{\lambda}_i^{(q)}\}_{i=1}^\infty$ are the eigenvalues of $\mathbf{H}_f^{(q)}$. In line with Zhang et al. (2026a;b), we extend the fourth moment and noise assumptions (Zou et al., 2021; Wu et al., 2022b;a; 2023) to quantized features.

Assumption 3.4 (Fourth-moment conditions)**.** Assume that the fourth moment of $\tilde{\mathbf{x}}^{(q)}$ is finite and there exists constant $\alpha \geq 1$ such that for any positive semi-definite (PSD) matrix $\mathbf{A}$,

$$\mathbb{E}\left[\tilde{\mathbf{x}}^{(q)} (\tilde{\mathbf{x}}^{(q)})^\top \mathbf{A} \tilde{\mathbf{x}}^{(q)} (\tilde{\mathbf{x}}^{(q)})^\top\right] \preceq \alpha \operatorname{tr}(\mathbf{H}_f^{(q)} \mathbf{A}) \mathbf{H}_f^{(q)}.$$

Regarding the noise assumptions, we first define the population risk and global optimum in quantized feature space: $\mathcal{R}_M^{(q)}(\mathbf{v}) := \frac{1}{2}\mathbb{E}[(\langle \tilde{\mathbf{x}}^{(q)}, \mathbf{v}\rangle - \mathcal{Q}_l(y))^2],\ \mathbf{v} \in \mathbb{R}^M$, with $\mathbf{v}_q^* := \arg\min_{\mathbf{v}} \mathcal{R}_M^{(q)}(\mathbf{v})$.

Assumption 3.5 (Noise conditions)**.** Denote $\xi := \mathcal{Q}_l(y) - \langle \mathbf{v}^{(q)^*}, \tilde{\mathbf{x}}^{(q)} \rangle$. Assume there exists a positive constant $\overline{\sigma} > 0$ such that

$$\mathbb{E}[\xi^2 \tilde{\mathbf{x}}^{(q)} (\tilde{\mathbf{x}}^{(q)})^\top] \preceq \overline{\sigma}^2 \mathbf{H}_f^{(q)}.$$

To simplify the scaling-law behavior, we assume specific data distribution where the data spectrum satisfies a power law and the optimal parameter satisfies a prior (Lin et al., 2024; Zhang et al., 2026b). Specifically, we consider the population risk for $\mathbf{w} \in \mathbb{H}$: $\mathcal{R}(\mathbf{w}) := \frac{1}{2}\mathbb{E}[(\langle \mathbf{x}, \mathbf{w}\rangle - y)^2]$ with global optimum $\mathbf{w}^* := \arg\min_{\mathbf{w}} \mathcal{R}(\mathbf{w})$.

Assumption 3.6. We assume the well-specified model, i.e., $\mathbb{E}[y|\mathbf{x}, \mathbf{w}^*] = \mathbf{x}^\top \mathbf{w}^*$ and $\sigma^2 := \mathbb{E}\left[(y - \mathbf{x}^\top \mathbf{w}^*)^2\right]$, and the parameter prior, i.e., there exists $b > 1$, $\mathbb{E}_{\mathbf{w}^*}[\lambda_i \langle \mathbf{w}^*, \mathbf{v}_i\rangle^2] \eqsim i^{-b}$, $\mathbb{E}_{\mathbf{w}^*}[\langle \mathbf{w}^*, \mathbf{v}_i\rangle \langle \mathbf{w}^*, \mathbf{v}_j\rangle] = 0,\ i \neq j$. We also assume the data spectrum is polynomial, i.e., there exists $a > 1$ such that the eigenvalues of $\mathbf{H}$ satisfy $\lambda_i \eqsim i^{-a},\ i > 0$.

4 Main Theory

4.1 Population Risk

We derive population risk upper bounds to show how the coordinate-wise relative quantization (Definition 3.2) and batch size jointly affect the risk, which is beyond the scope of Zhang et al. (2026b).

Theorem 4.1. *Under Assumption 3.1, 3.3, 3.4, 3.5, 3.6, with probability at least* $1 - e^{-\Omega(M)}$*, if the step size satisfies* $\gamma \lesssim \frac{1}{\alpha \operatorname{tr}(\mathbf{H}_f^{(q)})}$*, for any* $i \in \{d, f, s, a, p, o\}$*, quantization* $\mathcal{Q}_i$ *is* $\overline{\epsilon}_i$*-multiplicative,*

$$\mathbb{E}[\mathcal{R}_M(\overline{\mathbf{v}}_N)] \lesssim \sigma^2 + \underbrace{\widetilde{\Theta}\left(M^{(1-b)\vee(-a)} + (N\gamma)^{\frac{(1-b)\vee(-a)}{a}}\right)}_{\text{approximation+bias}} + \underbrace{\frac{\epsilon_q}{\epsilon_q + M^{-a}}}_{\text{quantization floor}}$$
$$+ \underbrace{\left(\frac{1 + \overline{\epsilon}_a + \overline{\epsilon}_o + \epsilon_p}{B}\right) \min\left\{\frac{M}{N},\ \gamma^{1/a} N^{-(a-1)/a} + \gamma \epsilon_q\right\} + \frac{\epsilon_p}{N}}_{\text{variance}},$$

where $\epsilon_p = \overline{\epsilon}_p (1 + \epsilon_q M^a)(\min\{M, (N\gamma)^{1/a}\})^{(a-b)_+}$, $\epsilon_q = (1 + \overline{\epsilon}_d)(1 + \overline{\epsilon}_s)(1 + \overline{\epsilon}_f) - 1$.

The population-risk bound established in Theorem 4.1 has the same high-level decomposition as the population risk in Zhang et al. (2026b): irreducible noise, approximation and optimization bias, a quantization floor, and a variance term. However, the way that data, sketch, feature quantization errors affect the variance via spectrum is different. Unlike the matrix-level multiplicative quantization in Zhang et al. (2026b) that induces a positive semi-definite perturbation aligning with the eigenspaces of $\mathbf{SHS}^\top$, current coordinate-wise relative quantization yields diagonal perturbations that destruct the original spectrum structure. Thanks to the trace preserving property of diagonal

projection, the relative-error model gives trace-level control, resulting in an additional ϵ_q/j component in the spectral upper bound, which leads to the $\gamma\epsilon_q$ term in the variance bound (see Lemma B.14 for details).

Another difference arises from how parameter quantization enters the population risk. Because the parameter-rounding error is sampled once per update, it is shared across all examples in the batch. Expanding the resulting batch covariance therefore separates same-example terms, weighted by $1/B$, from cross-example terms, weighted by $(B-1)/B \eqsim 1$. Moreover, the coordinate-wise error model bounds the conditional rounding covariance by a diagonal matrix proportional to $\mathrm{diag}(\mathbf{v}_t\mathbf{v}_t^\top)$. We control this diagonal second moment along the iterates and tracking its contribution through the refined covariance recursion, which determines the scale factor $\epsilon_p/\overline{\epsilon}_p$ in the bound. We defer the technical details to Section 5.

We also remark that the bias error in Theorem 4.1 does not rely on the commutativity between $\mathbf{H}_f^{(q)}$ and $\mathbf{SHS}^\top$ used in Zhang et al. (2026b). This sharper bound is developed by comparing the filters generated by $\mathbf{H}_f^{(q)}$ and $\mathbf{SHS}^\top$ using a resolvent identity and a spectral head–tail decomposition. We refer to Lemma B.15 for technical details.

4.2 Critical Batch Size

For the CBS analysis, we work in the small-quantization regime $\epsilon_q M^a = o(1)$, so that the B-independent quantization floor does not mask the batch-dependent bias-variance trade-off. We further specialize to overparameterized regimes $N\gamma \lesssim M^a$, where the population-risk bound simplifies to an explicit leading-order bias–variance balance. We remark that when $b > a+1$, the generic bias estimate is not tight, even in the full-precision setting (Lin et al., 2024). We therefore derive a refined bias bound under a stronger overparameterization condition $(N\gamma)^{(\min\{b,2a+1\}-a)/a} \lesssim M/\log M$ (see Lemma B.18 for details and Section 5 for discussions) and characterize the CBS via the refined population risk in this regime.

Theorem 4.2. *Fix $D = BN$. Under Assumption 3.1, 3.3, 3.4, 3.5, 3.6, if the step size satisfies $\gamma \lesssim \frac{1}{\alpha \mathrm{tr}(\mathbf{H}_f^{(q)})}$, for any $i \in \{d, f, s, a, p, o\}$, quantization $\mathcal{Q}_i$ is $\overline{\epsilon}_i$-multiplicative, in the regime of small quantization $\epsilon_q M^a = o(1)$,*

- *under the over-parameterization $D \lesssim M^a$, with probability at least $1 - e^{-\Omega(M)}$,* [4]

$$B^* \eqsim \begin{cases} 1 + \gamma\overline{\epsilon}_p, & 1 < b < a, \\ \widetilde{\Theta}\left(\gamma(1+\overline{\epsilon}_a+\overline{\epsilon}_o+\overline{\epsilon}_p)^{a/b} D^{(b-a)/b}\right), & 1 < a \le b \le a+1, \end{cases}$$

- *under the over-parameterization $D^{(\min\{b,2a+1\}-a)/a} \lesssim M/\log M$ and $b > a+1 > 2$, with probability at least $1 - e^{-\Omega(M)} - M^{-r}$ for some constant $r > 0$,* [5]

$$B^* = \widetilde{\Theta}\left(\gamma(1+\overline{\epsilon}_a+\overline{\epsilon}_o+\overline{\epsilon}_p)^{a/\min\{b,2a+1\}} D^{\frac{\min\{b,2a+1\}-a}{\min\{b,2a+1\}}}\right).$$

Theorem 4.2 demonstrates how precision enters the critical batch size: quantization errors serve as a scaling factor and preserve the scaling exponent with respect to data budget D. In particular, given a fixed data budget D, lower precision favors larger critical batch size. The interpretation is that, activation, parameter and output gradient quantization amplify the label noise, necessitating larger batch size to balance the effective noise. This can be well-understood by population risk under the small quantization error and the overparameterization regime (see Theorem B.22 and Theorem B.23) where the excess risk in the non-trivial case $b \ge a$ is approximately [6]

$$(\frac{D}{B}\gamma)^{\frac{1-\min\{b,2a+1\}}{a}} + (\frac{1+\overline{\epsilon}_a+\overline{\epsilon}_o+\overline{\epsilon}_p}{D})(\frac{D\gamma}{B})^{1/a}.$$

Here the first term arises from the bias while the second term arises from the variance. The critical batch size $B^* \eqsim \gamma(1+\overline{\epsilon}_a+\overline{\epsilon}_o+\overline{\epsilon}_p)^{a/\min\{b,2a+1\}} D^{\frac{\min\{b,2a+1\}-a}{\min\{b,2a+1\}}}$ is solved by balancing the two

[4] The $\widetilde{\Theta}$ suppresses some logarithmic terms in case $b = a+1$.

[5] The $\widetilde{\Theta}$ suppresses some logarithmic terms in case $b = 2a+1$.

[6] Here we omit the logarithmic terms.

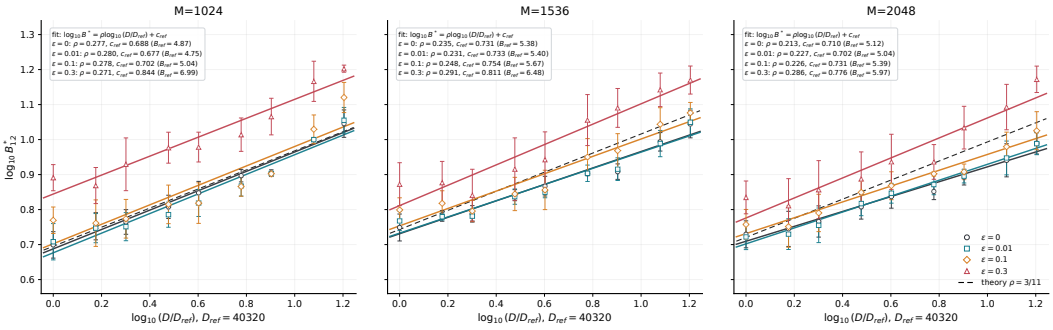

Figure 1: High-budget scaling of the critical batch size. Each panel fixes M and colors identify $\epsilon = \epsilon_p = \epsilon_a = \epsilon_o$. Colored lines are OLS fits, with (ρ, c_{ref}) shown in each panel; the dashed line has slope $3/11$.

terms: below B^*, increasing the batch substantially reduces gradient noise; beyond B^*, the loss of updates dominates and further batching has diminishing returns.

We remark that our results recovers the full-precision critical batch size established in (Zhang et al., 2025) when $\bar{\epsilon}_i = 0$. Moreover, we highlight that the overparameterization condition in $b \leq a + 1$ case is not necessary, and we can extend the critical batch size to the general regime of (M, D).

Corollary 4.3. *Fix $D = BN$. Under Assumption 3.1, 3.3, 3.4, 3.5, 3.6, with probability at least $1 - e^{-\Omega(M)}$, if the step size satisfies $\gamma \lesssim \frac{1}{\alpha \text{tr}(\mathbf{H}_f^{(q)})}$, for any $i \in \{d, f, s, a, p, o\}$, quantization $\mathcal{Q}_i$ is $\bar{\epsilon}_i$-multiplicative, in the regime of $\epsilon_q M^a = o(1)$ and $1 < b \leq a + 1$,*

$$B^* \eqsim \begin{cases} \max\left\{DM^{-a}\gamma, 1 + \gamma\bar{\epsilon}_p\right\}, & 1 < b < a, \\ \widetilde{\Theta}\left(\max\left\{DM^{-a}\gamma, \gamma(1 + \bar{\epsilon}_a + \bar{\epsilon}_o + \bar{\epsilon}_p)^{a/b} D^{(b-a)/b}\right\}\right), & 1 < a \leq b \leq a + 1, \end{cases}$$

The interpretation is that, if the M-dependent approximation or bias term dominates the optimized batch-dependent bias–variance contribution, the critical batch size is determined by increasing B until the batch-dependent contribution reaches the same order as this dominant M-dependent term, ensuring the highest data efficiency.

4.3 Empirical Study

Synthetic experiments. We study the finite-dimensional the quantized Gaussian-sketch linear-regression model with $p = 8192$, $(a, b) = (2, 2.75)$, $M \in \{1024, 1536, 2048\}$, step size $\gamma = 0.25$, and unit observation noise. For quantization level $\epsilon \in \{0, 0.01, 0.1, 0.3\}$, we set $\epsilon_p = \epsilon_a = \epsilon_o = \epsilon$ and disable the other quantizers. At each budget, we evaluate the largest tested batch whose mean risk is at most 1.2 times the minimum mean risk $B^*_{1.2}$. Points are means of $\log_{10} B^*_{1.2}$, with their standard deviations as error bars. For each (M, ϵ), we fit $\log_{10} B^*_{1.2} = \rho \log_{10}(D/D_{\text{ref}}) + c_{\text{ref}}$.

Figure 1 shows that, the slope fitted are close to $\rho = 3/11$ and the fitted intercept shows an increasing trend as epsilon increases, validating our theory of how precision changes the CBS.

LLM pre-training experiments. We further test the predicted precision dependence of CBS on a 302M decoder-only Transformer trained on FineWeb-Edu (Penedo et al., 2024), comparing BF16 with UFP4 (Zhao et al., 2026). We sweep global batch sizes $B \in \{256, 512, 1024, 2048, 4096, 8192\}$ and follow the constant-learning-rate plus exponential-weight-averaging protocol of Zhang et al. (2025); full training and quantization details are deferred to Appendix C.

For each precision, we estimate steps-to-target over five validation-loss targets and extract CBS using the 20% data-overhead criterion of Zhang et al. (2025). We then fit $B_{\text{CBS}}(D) = K(D/10^9)^{\beta}$. As shown in Fig. 2, the fitted laws are $447.56D^{0.4021}$ for BF16 and $494.78D^{0.4020}$ for UFP4. The nearly identical exponents, together with the 10.5% larger UFP4 prefactor, support our prediction that lower precision primarily shifts the CBS scale upward while preserving its data-budget exponent.

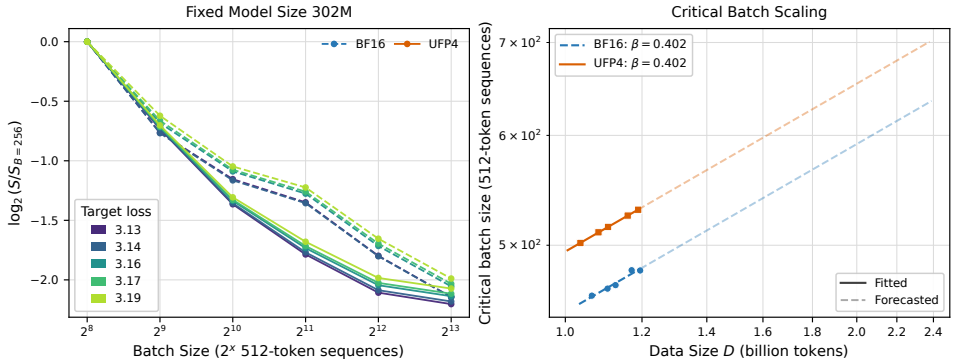

Figure 2: LLM pre-training CBS under BF16 and UFP4. Left: normalized steps-to-target versus batch size. Right: CBS versus data budget with power-law fits.

5 Technical Overview

Using risk decomposition as established in Lin et al. (2024); Zhang et al. (2026a;b) (see also our Lemma B.2, B.4 and (7), (8)) with $\boldsymbol{\eta}_t = \mathbf{v}_t - \mathbf{v}_q^*$, the main proof workload is to analyze the excess risk $\frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\langle(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{k-t}\mathbf{H}_f^{(q)}, \mathbb{E}[\boldsymbol{\eta}_t\boldsymbol{\eta}_t^\top]\rangle$ by bias-variance decomposition $\mathbb{E}[\boldsymbol{\eta}_t\otimes\boldsymbol{\eta}_t] \preceq 2\mathbf{B}_t + 2\mathbf{C}_t$ with [7] $\mathbf{B}_t = (\mathcal{I}-\gamma\mathcal{T}_B^{(q)})\circ\mathbf{B}_{t-1}$, $\mathbf{C}_t = (\mathcal{I}-\gamma\mathcal{T}_B^{(q)})\circ\mathbf{C}_{t-1}+\gamma^2\boldsymbol{\Sigma}_t$, $\mathbf{B}_0 = \boldsymbol{\eta}_0\boldsymbol{\eta}_0^\top$, $\mathbf{C}_0 = 0$.

Challenge 1: Diagonal quantization error matrices in optimization dynamics. To bound the variance error $\frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\langle(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{k-t}\mathbf{H}_f^{(q)}, \mathbf{C}_t\rangle$, we analyze the covariance recursion $\mathbf{C}_t$, whose driving term $\boldsymbol{\Sigma}_t$ is bounded using the full second moments of the parameter iterate and iterate error, together with the diagonal second moment of the parameter iterate (see Lemma B.7):

$$\begin{aligned}\boldsymbol{\Sigma}_t \preceq{}& \frac{(1+2\bar{\epsilon}_o)\bar{\sigma}^2}{B}\mathbf{H}_f^{(q)} + \frac{\bar{\epsilon}_a(1+2\bar{\epsilon}_o)}{B}\mathcal{M}^{(q)}\circ\mathbb{E}\left[\mathbf{v}_{t-1}\mathbf{v}_{t-1}^\top\right] + \frac{2\bar{\epsilon}_o}{B}\mathcal{M}^{(q)}\circ\mathbb{E}\left[\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top\right]\\ &+\frac{(1+\bar{\epsilon}_a)(1+2\bar{\epsilon}_o)\bar{\epsilon}_p}{B}\mathcal{M}^{(q)}\circ\mathbb{E}\left[\operatorname{diag}\left(\mathbf{v}_{t-1}\mathbf{v}_{t-1}^\top\right)\right] + \frac{B-1}{B}\bar{\epsilon}_p\mathbf{H}_f^{(q)}\operatorname{diag}\big(\mathbb{E}[\mathbf{v}_{t-1}\mathbf{v}_{t-1}^\top]\big)\,\mathbf{H}_f^{(q)}.\end{aligned} \tag{3}$$

The last two terms distinguish the same-example and cross-example effects of parameter rounding. Controlling the resulting diagonal second moment throughout the recursion is the main obstacle to obtaining the refined variance bound.

Rather than replacing the diagonal second moment by a uniform crude bound, as in the lower-bound analysis of $\mathbb{E}[\mathbf{v}_t\mathbf{v}_t^\top]$ in Zhang et al. (2026b), we control its contribution recursively. We first derive a crude upper bound on the bias second-moment matrix $\mathbf{B}_t$ and use it to bound $\mathbb{E}[\mathbf{v}_t\mathbf{v}_t^\top]$, which preserves the dependence on the evolving covariance $\mathbf{C}_t$. Taking the diagonal of this bound and substituting it into the covariance recursion (3) yields a recursive loop in terms of the evolving covariance $\mathbf{C}_t$, leading to a closed bound for the variance error. Specifically, Lemma B.8 controls a crude upper bound $\mathbf{B}_t \preceq (\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t\mathbf{B}_0(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t + \gamma\frac{\frac{\alpha}{B}\operatorname{tr}(\mathbf{H}_f^{(q)}\mathbf{v}_q^*(\mathbf{v}_q^*)^\top)}{1-\gamma\frac{\alpha}{B}\operatorname{tr}(\mathbf{H}_f^{(q)})}\mathbf{I}$. Using $\mathbb{E}[(\mathbf{v}_t - \mathbb{E}[\mathbf{v}_t])(\mathbf{v}_t-\mathbb{E}[\mathbf{v}_t])^\top] \preceq 2[\mathbf{B}_t - (\mathbb{E}[\mathbf{v}_t]-\mathbf{v}_q^*)(\mathbb{E}[\mathbf{v}_t]-\mathbf{v}_q^*)^\top] + 2\mathbf{C}_t$ gives

$$\mathbb{E}[\mathbf{v}_t\mathbf{v}_t^\top] \preceq [\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t]\mathbf{v}_q^*\mathbf{v}_q^{*\top}[\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t] + 2\mathbf{C}_t + \frac{2\gamma\alpha\|\mathbf{v}_q^*\|^2_{\mathbf{H}_f^{(q)}}}{B[1-\gamma\alpha\operatorname{tr}(\mathbf{H}_f^{(q)})/B]}\mathbf{I}.$$

Together with the identity $\operatorname{tr}(\mathbf{A}\operatorname{diag}(\mathbf{B})) = \operatorname{tr}(\mathbf{B}\operatorname{diag}(\mathbf{A}))$ and $\mathcal{M}^{(q)}\circ\mathbf{A} \preceq \alpha\operatorname{tr}(\mathbf{H}_f^{(q)}\mathbf{A})]\mathbf{H}_f^{(q)}$ from Assumption 3.4, all terms in $\boldsymbol{\Sigma}_t$ (3) that depend on the evolving covariance $\mathbf{C}_{t-1}$ are controlled by $\operatorname{tr}([\mathbf{H}_f^{(q)}+\operatorname{diag}(\mathbf{H}_f^{(q)})]\mathbf{C}_{t-1})$. Unrolling the recursion for $\mathbf{C}_t$ creates a recursive loop and allows

[7] Use $(\mathcal{I}-\gamma\mathcal{T}_B^{(q)})\circ\mathbf{A} = \mathbb{E}[(\mathbf{I}-\frac{\gamma}{B}\widetilde{\mathbf{X}}^{(q)}(\widetilde{\mathbf{X}}^{(q)})^\top)\mathbf{A}(\mathbf{I}-\frac{\gamma}{B}\widetilde{\mathbf{X}}^{(q)}(\widetilde{\mathbf{X}}^{(q)})^\top)]$, $\widetilde{\mathbf{X}}_t^{(q)} = \mathcal{Q}_f(\mathcal{Q}_s(\mathbf{S})\mathcal{Q}_d(\mathbf{X}_t)^\top)$, $\mathcal{M}^{(q)}\circ\mathbf{A} = \mathbb{E}[((\widetilde{\mathbf{x}}^{(q)})^\top\mathbf{A}\widetilde{\mathbf{x}}^{(q)})\widetilde{\mathbf{x}}^{(q)}(\widetilde{\mathbf{x}}^{(q)})^\top]$, and we refer to Appendix B.1 for detailed definitions.

us to close the bound:

$$\sum_{t=0}^{N-1} \operatorname{tr}\left([\mathbf{H}_f^{(q)} + \operatorname{diag}(\mathbf{H}_f^{(q)})]\mathbf{C}_t\right) \lesssim \gamma \operatorname{tr}(\mathbf{H}_f^{(q)}) \frac{N}{B}\left[(1+\overline{\epsilon}_o)\overline{\sigma}^2 + (1+\overline{\epsilon}_a+\overline{\epsilon}_o+\overline{\epsilon}_p)\|\mathbf{v}_q^*\|^2_{\mathbf{H}_f^{(q)}}\right] + \gamma \operatorname{tr}(\mathbf{H}_f^{(q)})\overline{\epsilon}_p \sum_{t=0}^{N-1} \left\|[\mathbf{I} - (\mathbf{I} - \gamma \mathbf{H}_f^{(q)})^t]\mathbf{v}_q^*\right\|^2_{\operatorname{diag}(\mathbf{H}_f^{(q)})}.$$

Substituting this aggregate bound into the unrolled variance representation yields the variance bound. We defer the detailed derivations to the proof for Theorem B.10.

Challenge 2: Sharp bias bounds in the smooth-target regime. Under Assumption 3.6, a key step in controlling the bias contribution $\frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\langle(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{k-t}\mathbf{H}_f^{(q)}, \mathbf{B}_t\rangle$ is to bound the filtered target energy $\mathbb{E}_{\mathbf{w}^*}\left[\frac{1}{\gamma^2 N^2}\|\mathbf{v}_q^*\|^2_{(\mathbf{H}^{(q)}_{f,0:k^*})^{-1}} + \|\mathbf{v}_q^*\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}\right]$. When $b > a+1$, the standard head–tail argument in the full-precision setting gives an upper bound of order $1/(N\gamma)$; see Lemma D.4 of Lin et al. (2024). As noted therein, this leaves a nonconstant gap between the available upper and lower rates. The gap arises because the standard argument replaces the direction-dependent weights on the head spectrum by the uniform bound $\frac{1}{\gamma^2N^2}\|\mathbf{v}^*\|^2_{(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}_{0:k^*}} \le \frac{1}{N\gamma}\|\mathbf{v}^*\|^2_{\mathbf{I}_{0:k^*}}$, thereby treating all head directions identically and discarding their individual eigenvalue dependence.[8]

To retain this eigenvalue dependence, we use $\min\{x, \theta^2/x\} \asymp \theta^2 x/(x+\theta)^2$, and obtain

$$\begin{aligned}&\mathbb{E}_{\mathbf{w}^*}\left[\|\mathbf{v}^*\|^2_{(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}_{0:k^*}} + \|\mathbf{v}^*\|^2_{(\mathbf{S}\mathbf{H}\mathbf{S}^\top)_{k^*:\infty}}\right]\\ \lesssim &\frac{1}{N^2\gamma^2}\sum_i \lambda_i^2(\mathbf{v}_i^\top \mathbb{E}[\mathbf{w}^*(\mathbf{w}^*)^\top]\mathbf{v}_i)\,\mathbf{v}_i^\top\mathbf{S}^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}(\mathbf{S}\mathbf{H}\mathbf{S}^\top + (N\gamma)^{-1}\mathbf{I})^{-2}\mathbf{S}\mathbf{v}_i.\end{aligned}$$

Unlike the uniform head bound, this representation retains the contribution of each population eigendirection. To control the resulting direction-wise quadratic forms, let $\mathbf{z}_i := \mathbf{S}\mathbf{v}_i$, use the resolvent identity together with the Sherman–Morrison formula, we have

$$\frac{\lambda_i^2}{\gamma^2N^2}\mathbf{z}_i^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}(\mathbf{S}\mathbf{H}\mathbf{S}^\top + (N\gamma)^{-1}\mathbf{I})^{-2}\mathbf{z}_i = \frac{\lambda_i + \lambda_i^2\mathbf{z}_i^\top\mathbf{R}_i(\mathbf{R}_i+(N\gamma)^{-1}\mathbf{I})^{-2}\mathbf{z}_i}{\left[1+\lambda_i\mathbf{z}_i^\top(\mathbf{R}_i+(N\gamma)^{-1}\mathbf{I})^{-1}\mathbf{z}_i\right]^2} - \frac{\lambda_i}{1+\lambda_i\mathbf{z}_i^\top\mathbf{R}_i^{-1}\mathbf{z}_i}.$$

We highlight here that to establish high-probability bounds, we introduce the leave-one-out matrix $\mathbf{R}_i := \mathbf{S}\mathbf{H}\mathbf{S}^\top - \lambda_i\mathbf{S}\mathbf{v}_i\mathbf{v}_i^\top\mathbf{S}^\top$ to ensure $\mathbf{R}_i$ is independent of $\mathbf{z}_i$. For $k = \lceil(N\gamma)^{1/a}\rceil$, this gives, simultaneously over $i \le k$ with polynomially high probability,

$$\mathbf{z}_i^\top\mathbf{R}_i(\mathbf{R}_i+(N\gamma)^{-1}\mathbf{I})^{-2}\mathbf{z}_i \lesssim \frac{N\gamma(k+\log M)}{M}.$$

To obtain a nondegenerate lower bound on the resolvent denominator, we impose the overparameterization condition $N\gamma \le c_0 M^a$. This condition ensures that the effective spectral dimension $k \asymp (N\gamma)^{1/a}$ remains sufficiently smaller than M, so that $\mathbf{R}_i$ retains sufficiently many directions below the regularization scale θ. Gaussian concentration then yields, uniformly over $i \le k$,

$$\mathbf{z}_i^\top(\mathbf{R}_i + \theta\mathbf{I})^{-1}\mathbf{z}_i \gtrsim N\gamma.$$

Summarizing these high-probability bounds recovers the desired direction-dependent head weight. We refer to Lemma B.18 for details.

6 Conclusion

We derive population risk bounds in the Gaussian-sketched linear regression setup with constant step-size quantized mini-batch SGD under unbiased coordinate-wise quantization error model with bounded relative error variance. Building on this framework, we characterize the low-precision critical batch size. Theoretical demonstrates that larger quantization-error levels increase the predicted batch scale while preserving its exponent in the data budget, implying larger critical batch size in lower precisions. Experiment results validate our theory.

Limitations. Future directions include: extending linear regression analysis to non-linear models, extending SGD to practical optimizers and developing a matching lower bounds.

[8] Here $\mathbf{v}^* := (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}\mathbf{w}^*$ denotes the full-precision optimum in the sketched feature space.

AI USE STATEMENT

Generative AI tools were employed to support manuscript drafting and editing, LaTeX formatting, and to provide feedback on the methodological descriptions and interpretation of experimental results. All AI-assisted suggestions were revised under the authors' direction. The authors take full responsibility for the final content, including all scientific claims, experimental results, and references.

REFERENCES

Alexander Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

Shane Bergsma, Nolan Simran Dey, Gurpreet Gosal, Gavia Gray, Daria Soboleva, and Joel Hestness. Power lines: Scaling laws for weight decay and batch size in LLM pre-training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=bFXbLQzRoZ.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 95–136. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.bigscience-1.9. URL https://aclanthology.org/2022.bigscience-1.9/.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *International Conference on Machine Learning*, pp. 4345–4382, 2024. URL https://proceedings.mlr.press/v235/bordelon24a.html.

Hengjie Cao, Mengyi Chen, Yifeng Yang, Fang Dong, Ruijun Huang, Jixian Zhou, Anrui Chen, Mingzhi Dong, Yujiang Wang, Jinlong Hou, Yuan Cheng, FAN WU, Fan Yang, Tun Lu, Ning Gu, and Li Shang. Metis: Training LLMs with FP4 quantization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=I2ZrCi5O84.

Yifang Chen, Xuyang Guo, Xiaoyu Li, Yingyu Liang, Zhenmei Shi, and Zhao Song. Scaling law phenomena across regression paradigms: Multiple and kernel approaches. *arXiv preprint arXiv:2503.01314*, 2025.

DeepSeek-AI. Deepseek-v3 technical report, 2024. URL https://arxiv.org/abs/2412.19437.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026. URL https://arxiv.org/html/2606.19348v1.

Shihong Ding, Haihan Zhang, Hanzhen Zhao, and Cong Fang. Scaling law for stochastic gradient descent in quadratically parameterized linear regression. *arXiv preprint arXiv:2502.09106*, 2025.

Maxim Fishman, Brian Chmiel, Ron Banner, and Daniel Soudry. Scaling FP8 training to trillion-token LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=E1EHO0imOb.

Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1737–1746. PMLR, 2015. URL https://proceedings.mlr.press/v37/gupta15.html.

Kento Imaizumi and Hideaki Iiduka. Iteration and stochastic first-order oracle complexities of stochastic gradient descent using constant and decaying learning rates. *arXiv preprint arXiv:2402.15344*, 2024. URL https://arxiv.org/abs/2402.15344.

Rustem Islamov, Roman Machacek, Aurelien Lucchi, Tony Silveti-Falls, Eduard Gorbunov, and Volkan Cevher. On the role of batch size in stochastic conditional gradient methods. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=yTFTLj7MoN.

Tanishq Kumar, Zachary Ankner, Benjamin F. Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Ré, and Aditi Raghunathan. Scaling laws for precision. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.04330.

Andrey Kuzmin, Mart Van Baalen, Yuwei Ren, Markus Nagel, Jorn Peters, and Tijmen Blankevoort. Fp8 quantization: The power of the exponent. *Advances in Neural Information Processing Systems*, 35:14651–14662, 2022.

Binghui Li, Fengling Chen, Zixun Huang, Lean Wang, and Lei Wu. Functional scaling laws in kernel regression: Loss dynamics and learning rate schedules. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=dpllevHMbc.

Chuning Li and Chris J. Maddison. Predicting large model test losses with a noisy quadratic system. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.09154.

Licong Lin, Jingfeng Wu, Sham M Kakade, Peter L Bartlett, and Jason D Lee. Scaling laws in linear regression: Compute, parameters, and data. *Advances in Neural Information Processing Systems*, 37:60556–60606, 2024.

Licong Lin, Jingfeng Wu, and Peter Bartlett. Improved scaling laws in linear regression via data reuse. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=jeen4x145W.

Taowen Liu, Marta Andronic, Deniz Gunduz, and George Anthony Constantinides. Training with fewer bits: Unlocking edge LLMs training with stochastic rounding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14531–14546, 2025. doi: 10.18653/v1/2025.findings-emnlp.784. URL https://aclanthology.org/2025.findings-emnlp.784/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Ilia Markov, Adrian Vladu, Qi Guo, and Dan Alistarh. Quantized distributed training of large models with convergence guarantees. In *International Conference on Machine Learning*, pp. 24020–24044. PMLR, 2023.

Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training. *arXiv preprint arXiv:1812.06162*, 2018. URL https://arxiv.org/abs/1812.06162.

William Merrill, Shane Arora, Dirk Groeneveld, and Hannaneh Hajishirzi. Critical batch size revisited: A simple empirical approach to large-batch language model training. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.23971.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1710.03740.

MiniMax, Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, Enwei Jiao, Gengxin Li, Guojun Zhang, Haohai Sun, Houze Dong, Jiadai Zhu, Jiaqi Zhuang, Jiayuan Song, Jin Zhu, Jingtao Han, Jingyang Li, Junbin Xie, Junhao Xu, Junjie Yan, Kaishun Zhang, Kecheng Xiao, Kexi Kang, Le Han, Leyang Wang, Lianfei Yu, Liheng Feng, Lin Zheng, Linbo Chai, Long Xing, Meizhi Ju, Mingyuan Chi, Mozhi Zhang, Peikai Huang, Pengcheng Niu, Pengfei Li, Pengyu Zhao, Qi Yang, Qidi Xu, Qiexiang Wang, Qin Wang, Qiuhui Li, Ruitao Leng, Shengmin Shi, Shuqi Yu, Sichen Li, Songquan Zhu, Tao Huang, Tianrun Liang, Weigao Sun, Weixuan Sun, Weiyu Cheng, Wenkai Li, Xiangjun Song, Xiao Su, Xiaodong Han, Xinjie Zhang, Xinzhu Hou, Xu Min, Xun Zou, Xuyang Shen, Yan Gong, Yingjie Zhu, Yipeng Zhou, Yiran Zhong, Yongyi Hu, Yuanxiang Fan, Yue Yu, Yufeng Yang, Yuhao Li, Yunan Huang, Yunji Li, Yunpeng Huang, Yunzhi Xu, Yuxin Mao, Zehan Li, Zekang Li, Zewei Tao, Zewen Ying, Zhaoyang Cong, Zhen Qin, Zhenhua Fan, Zhihang Yu, Zhuo Jiang, and Zijia Wu. Minimax-01: Scaling foundation models with lightning attention, 2025. URL https://arxiv.org/abs/2501.08313.

Ionut-Vlad Modoranu, Mher Safaryan, Grigory Malinovsky, Eldar Kurtić, Thomas Robert, Peter Richtárik, and Dan Alistarh. Microadam: Accurate adaptive optimization with low space overhead and provable convergence. *Advances in Neural Information Processing Systems*, 37:1–43, 2024.

Hiroki Naganuma, Shagun Gupta, Youssef Briki, Ioannis Mitliagkas, Irina Rish, Parameswaran Raman, and Hao-Jun Michael Shi. Adaptive batch sizes using non-euclidean gradient noise scales for stochastic sign and spectral descent. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=XMSaWRpEPS.

NVIDIA, Felix Abecassis, Anjulie Agrusa, Dong Ahn, Jonah Alben, Stefania Alborghetti, Michael Andersch, Sivakumar Arayandi, Alexis Bjorlin, Aaron Blakeman, Evan Briones, Ian Buck, Bryan Catanzaro, Muya Chang, Jinhang Choi, Mike Chrzanowski, Eric Chung, Victor Cui, Steve Dai, Bita Darvish Rouhani, Carlo del Mundo, Deena Donia, Burc Eryilmaz, Henry Estela, Abhinav Goel, Oleg Goncharov, Yugi Guvvala, Robert Hesse, Russell Hewett, Herbert Hum, Ujval Kapasi, Brucek Khailany, Mikail Khona, Nick Knight, Alex Kondratenko, Ronny Krashinsky, Ben Lanir, Simon Layton, Michael Lightstone, Daniel Lo, Paulius Micikevicius, Asit Mishra, Tim Moon, Deepak Narayanan, Chao Ni, Abhijit Paithankar, Satish Pasumarthi, Ankit Patel, Mostofa Patwary, Ashwin Poojary, Gargi Prasad, Sweta Priyadarshi, Yigong Qin, Xiaowei Ren, Oleg Rybakov, Charbel Sakr, Sanjeev Satheesh, Stas Sergienko, Pasha Shamis, Kirthi Shankar, Nishant Sharma, Mohammad Shoeybi, Michael Siu, Misha Smelyanskiy, Darko Stosic, Dusan Stosic, Bor-Yiing Su, Frank Sun, Nima Tajbakhsh, Shelby Thomas, Przemek Tredak, Evgeny Tsykunov, Gandhi Vaithilingam, Aditya Vavre, Rangharajan Venkatesan, Roger Waleffe, Qiyu Wan, Hexin Wang, Mengdi Wang, Lizzie Wei, Hao Wu, Evan Wu, Keith Wyss, Ning Xu, Jinze Xue, Charlene Yang, Yujia Zhai, Ruoxi Zhang, Jingyang Zhu, and Zhongbo Zhu. Pretraining large language models with nvfp4, 2026. URL https://arxiv.org/abs/2509.25149.

Kaan Ozkara, Tao Yu, and Youngsuk Park. Stochastic rounding for LLM training: Theory and practice. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=3j3NtXcc95.

Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+3 phases of compute-optimal neural scaling laws. *Advances in Neural Information Processing Systems*, 37:16459–16537, 2024.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024. URL https://arxiv.org/abs/2406.17557.

Houwen Peng, Kan Wu, Yixuan Wei, Guoshuai Zhao, Yuxiang Yang, Ze Liu, Yifan Xiong, Ziyue Yang, Bolin Ni, Jingcheng Hu, Ruihang Li, Miaosen Zhang, Chen Li, Jia Ning, Ruizhe Wang, Zheng Zhang, Shuguang Liu, Joe Chau, Han Hu, and Peng Cheng. Fp8-lm: Training fp8 large language models, 2023. URL https://arxiv.org/abs/2310.18313.

Christopher J. Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E. Dahl. Measuring the effects of data parallelism on neural network training. *Journal of Machine Learning Research*, 20(112):1–49, 2019. URL https://www.jmlr.org/papers/v20/18-789.html.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. URL https://arxiv.org/abs/2002.05202.

Egor Shulgin, Dimitri von Rütte, Tianyue H. Zhang, Niccolò Ajroldi, Bernhard Schölkopf, and Antonio Orvieto. Deriving hyperparameter scaling laws via modern optimization theory. In *Workshop on Scientific Methods for Understanding Deep Learning*, 2026. URL https://openreview.net/forum?id=4ERabDxDdl.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. URL https://arxiv.org/abs/2104.09864.

Xuan Tang, Jichu Li, and Difan Zou. A convergence analysis of adaptive optimizers under floating-point quantization. In *International Conference on Learning Representations*, volume 2026, pp. 4069–4127, 2026.

Jia-Nan Wang, Zixun Huang, Kairui Li, and Lei Wu. Momentum in large-batch training: Polyak enlarges the critical batch size, Nesterov improves data efficiency. *arXiv preprint arXiv:2609.02728*, 2026. URL https://arxiv.org/abs/2609.02728.

Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Last iterate risk bounds of sgd with decaying stepsize for overparameterized linear regression. In *International conference on machine learning*, pp. 24280–24314. PMLR, 2022a.

Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham Kakade. The power and limitation of pretraining-finetuning for linear regression under covariate shift. *Advances in Neural Information Processing Systems*, 35:33041–33053, 2022b.

Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. Finite-sample analysis of learning high-dimensional single relu neuron. In *International Conference on Machine Learning*, pp. 37919–37951. PMLR, 2023.

Tingkai Yan, Haodong Wen, Binghui Li, Kairong Luo, Wenguang Chen, and Kaifeng Lyu. Larger datasets can be repeated more: A theoretical analysis of multi-epoch scaling in linear regression. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0CXjpAxHUE.

Dechen Zhang, Junwei Su, and Difan Zou. Learning under quantization for high-dimensional linear regression. In *International Conference on Learning Representations*, volume 2026, pp. 26307–26374, 2026a.

Dechen Zhang, Xuan Tang, Yingyu Liang, and Difan Zou. Scaling laws for precision in high-dimensional linear regression. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=LyhBIrNBXv.

Dechen Zhang, Xuan Tang, and Difan Zou. On the convergence of low-precision lora training. In *High-dimensional Learning Dynamics 2026*, 2026c.

Guodong Zhang, Lala Li, Zachary Nado, James Martens, Sushant Sachdeva, George E. Dahl, Christopher J. Shallue, and Roger Grosse. Which algorithmic choices matter at which batch sizes? insights from a noisy quadratic model. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1907.04164.

Hanlin Zhang, Depen Morwani, Nikhil Vyas, Jingfeng Wu, Difan Zou, Udaya Ghai, Dean Foster, and Sham Kakade. How does critical batch size scale in pre-training? In *International Conference on Learning Representations*, 2025.

Qian Zhao, Kunlong Chen, Changxin Tian, Zhonghui Jiang, Haitao Zhang, Chaofan Yu, Peijie Jiang, Mingliang Gong, Jia Liu, Ziqi Liu, Zhiqiang Zhang, and Jun Zhou. Rethinking shrinkage bias in LLM FP4 pretraining: Geometric origin, systemic impact, and UFP4 recipe. *arXiv preprint arXiv:2606.20381*, 2026. URL https://arxiv.org/html/2606.20381v1.

Yunhua Zhou, Junhao Huang, Shuhao Xing, Yechen Zhang, Runyu Peng, Qiping Guo, and Xipeng Qiu. How to set the batch size for large-scale pre-training? *arXiv preprint arXiv:2601.05034*, 2026. URL https://arxiv.org/abs/2601.05034.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Benign overfitting of constant-stepsize sgd for linear regression. In *Conference on learning theory*, pp. 4633–4635. PMLR, 2021.

A Appendix

B Risk Analysis

B.1 Preliminaries

We first define the following linear operators:

$$\mathcal{I} = \mathbf{I} \otimes \mathbf{I}, \quad \mathcal{M}^{(q)} = \mathbb{E}[\widetilde{\mathbf{x}}^{(q)} \otimes \widetilde{\mathbf{x}}^{(q)} \otimes \widetilde{\mathbf{x}}^{(q)} \otimes \widetilde{\mathbf{x}}^{(q)}], \quad \widetilde{\mathcal{M}}^{(q)} = \mathbf{H}_f^{(q)} \otimes \mathbf{H}_f^{(q)},$$
$$\mathcal{T}^{(q)} = \mathbf{H}_f^{(q)} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{H}_f^{(q)} - \gamma \mathcal{M}^{(q)}, \quad \widetilde{\mathcal{T}}^{(q)} = \mathbf{H}_f^{(q)} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{H}_f^{(q)} - \gamma \widetilde{\mathcal{M}}^{(q)}.$$

For a positive semi-definite (PSD) matrix $\mathbf{A}$, the above definitions result in:

$$\mathcal{I} \circ \mathbf{A} = \mathbf{A}, \quad \mathcal{M}^{(q)} \circ \mathbf{A} = \mathbb{E}[((\widetilde{\mathbf{x}}^{(q)})^\top \mathbf{A} \widetilde{\mathbf{x}}^{(q)}) \widetilde{\mathbf{x}}^{(q)} (\widetilde{\mathbf{x}}^{(q)})^\top], \quad \widetilde{\mathcal{M}}^{(q)} \circ \mathbf{A} = \mathbf{H}_f^{(q)} \mathbf{A} \mathbf{H}_f^{(q)},$$
$$(\mathcal{I} - \gamma \mathcal{T}^{(q)}) \circ \mathbf{A} = \mathbb{E}[(\mathbf{I} - \gamma \widetilde{\mathbf{x}}^{(q)} (\widetilde{\mathbf{x}}^{(q)})^\top) \mathbf{A} (\mathbf{I} - \gamma \widetilde{\mathbf{x}}^{(q)} (\widetilde{\mathbf{x}}^{(q)})^\top)],$$
$$(\mathcal{I} - \gamma \widetilde{\mathcal{T}}^{(q)}) \circ \mathbf{A} = (\mathbf{I} - \gamma \mathbf{H}_f^{(q)}) \mathbf{A} (\mathbf{I} - \gamma \mathbf{H}_f^{(q)}).$$

Further, we generalize the linear operators to account for batch size effects. For a PSD matrix $\mathbf{A}$, we define

$$\mathcal{M}_B^{(q)} \circ \mathbf{A} = \mathbb{E}\left[\frac{1}{B^2} \widetilde{\mathbf{X}}^{(q)} (\widetilde{\mathbf{X}}^{(q)})^\top \mathbf{A} \widetilde{\mathbf{X}}^{(q)} (\widetilde{\mathbf{X}}^{(q)})^\top\right],$$
$$(\mathcal{I} - \gamma \mathcal{T}_B^{(q)}) \circ \mathbf{A} = \mathbb{E}\left[\left(\mathbf{I} - \gamma \frac{1}{B} \widetilde{\mathbf{X}}^{(q)} (\widetilde{\mathbf{X}}^{(q)})^\top\right) \mathbf{A} \left(\mathbf{I} - \gamma \frac{1}{B} \widetilde{\mathbf{X}}^{(q)} (\widetilde{\mathbf{X}}^{(q)})^\top\right)\right].$$

Lemma B.1. *Under Assumption 3.3, 3.4,*

$$\mathcal{M}_B^{(q)} \circ \mathbf{A} \preceq \alpha \mathrm{tr}\left(\mathbf{H}_f^{(q)} \mathbf{A}\right) \mathbf{H}_f^{(q)}, \quad (\mathcal{M}_B^{(q)} - \widetilde{\mathcal{M}}^{(q)}) \circ \mathbf{A} \preceq \frac{\alpha}{B} \mathrm{tr}\left(\mathbf{H}_f^{(q)} \mathbf{A}\right) \mathbf{H}_f^{(q)}.$$

Proof. For the first inequality, under Assumption 3.4,

$$\begin{aligned}
\mathcal{M}_B^{(q)} \circ \mathbf{A} =& \mathbb{E}\left[\frac{1}{B^2} \widetilde{\mathbf{X}}^{(q)} (\widetilde{\mathbf{X}}^{(q)})^\top \mathbf{A} \widetilde{\mathbf{X}}^{(q)} (\widetilde{\mathbf{X}}^{(q)})^\top\right] \\
=& \frac{B(B-1)}{B^2} \widetilde{\mathcal{M}}^{(q)} \circ \mathbf{A} + \frac{B}{B^2} \mathcal{M}^{(q)} \circ \mathbf{A} \\
\preceq& \frac{B-1}{B} \mathbf{H}_f^{(q)} \mathbf{A} \mathbf{H}_f^{(q)} + \frac{\alpha}{B} \mathrm{tr}\left(\mathbf{H}_f^{(q)} \mathbf{A}\right) \mathbf{H}_f^{(q)} \\
\preceq& \frac{B-1+\alpha}{B} \mathrm{tr}\left(\mathbf{H}_f^{(q)} \mathbf{A}\right) \mathbf{H}_f^{(q)} \\
\preceq& \alpha \mathrm{tr}\left(\mathbf{H}_f^{(q)} \mathbf{A}\right) \mathbf{H}_f^{(q)},
\end{aligned}$$

where the last inequality holds as $\alpha \geq 1$. For the second inequality, under Assumption 3.4, similarly,

$$(\mathcal{M}_B^{(q)} - \widetilde{\mathcal{M}}^{(q)}) \circ \mathbf{A} = \frac{1}{B} (\mathcal{M}^{(q)} - \widetilde{\mathcal{M}}^{(q)}) \circ \mathbf{A} \preceq \frac{1}{B} \mathcal{M}^{(q)} \circ \mathbf{A} \preceq \frac{\alpha}{B} \mathrm{tr}\left(\mathbf{H}_f^{(q)} \mathbf{A}\right) \mathbf{H}_f^{(q)}.$$

□

B.2 Risk Decomposition

Recall the population risk $\mathcal{R}_M(\mathbf{v}) = \frac{1}{2}\mathbb{E}\left[(\langle \mathbf{S}\mathbf{x}, \mathbf{v}\rangle - y)^2\right]$ and $\mathcal{R}(\mathbf{w}) = \frac{1}{2}\mathbb{E}\left[(\langle \mathbf{x}, \mathbf{w}\rangle - y)^2\right]$, we decompose

$$\mathcal{R}_M(\overline{\mathbf{v}}_N) = \underbrace{\min \mathcal{R}(\cdot)}_{\text{Irreducible}} + \underbrace{\min \mathcal{R}_M(\cdot) - \min \mathcal{R}(\cdot)}_{\text{Approx}} + \underbrace{\mathcal{R}_M(\overline{\mathbf{v}}_N) - \min \mathcal{R}_M(\cdot)}_{\text{Excess}}.$$

We first provide bounds for the irreducible risk. By the well-specified model Assumption 3.6,

$$\text{Irreducible} := \mathcal{R}(\mathbf{w}^*) = \frac{1}{2}\sigma^2. \tag{4}$$

We then provide upper bounds for the approximation risk which can be deduced from Lemma C.1 in Lin et al. (2024), under Assumption 3.6, with probability at least $1 - e^{-\Omega(M)}$,

$$\mathbb{E}_{\mathbf{w}^*}\text{Approx} \lesssim \begin{cases} M^{1-b}, & 1 < b < a+1, \\ M^{-a}\log M, & b = a+1, \\ M^{-a}, & b > a+1. \end{cases} \tag{5}$$

We then focus on the excess risk $\mathcal{R}_M(\overline{\mathbf{v}}_N) - \min \mathcal{R}_M(\cdot)$. Define the minimizer in the full-precision space

$$\mathbf{v}^* := \arg\min_{\mathbf{v}} \mathcal{R}_M(\mathbf{v}) = \arg\min_{\mathbf{v}} \frac{1}{2}\mathbb{E}\left[(\langle \mathbf{v}, \mathbf{S}\mathbf{x}\rangle - y)^2\right] = \left(\mathbf{S}\mathbf{H}\mathbf{S}^\top\right)^{-1}\mathbf{S}\mathbf{H}\mathbf{w}^*.$$

We analyze by utilizing the minimizer in the low-precision space

$$\mathbf{v}_q^* := \arg\min_{\mathbf{v}} \mathcal{R}_M^{(q)}(\mathbf{v}) = \arg\min_{\mathbf{v}} \frac{1}{2}\mathbb{E}\left[\left(\langle \tilde{\mathbf{x}}^{(q)}, \mathbf{v}\rangle - \mathcal{Q}_l(y)\right)^2\right] = (\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}\mathbf{w}^*.$$

The following excess risk decomposition is established in Zhang et al. (2026b), for completeness, we provide proofs here.

Lemma B.2 (Excess risk decomposition, Lemma B.2 in Zhang et al. (2026b))**.** *Under Assumption 3.1 and Assumption 3.3, it holds*

$$\mathbb{E}\left[\mathcal{R}_M(\overline{\mathbf{v}}_N) - \mathcal{R}_M(\mathbf{v}^*)\right] = \underbrace{\frac{1}{2}\left\langle \mathbf{S}\mathbf{H}\mathbf{S}^\top, \mathbb{E}\left[(\mathbf{v}_q^* - \overline{\mathbf{v}}_N)\otimes(\mathbf{v}_q^* - \overline{\mathbf{v}}_N)\right]\right\rangle}_{R_N} + \mathrm{R}_{\text{add}},$$

where the additive error

$$R_{\text{add}} = \left(\mathbf{v}_q^*\right)^\top \frac{1}{N\gamma}\left[\mathbf{I} - \left(\mathbf{I} - \gamma\mathbf{H}_f^{(q)}\right)^N\right]\left(\mathbf{H}_f^{(q)}\right)^{-1}\left(\mathbf{H}_f^{(q)} - \mathbf{S}\mathbf{H}\mathbf{S}^\top\right)\mathbf{v}_q^* + \frac{1}{2}\left\langle \mathbf{S}\mathbf{H}\mathbf{S}^\top, (\mathbf{v}_q^* - \mathbf{v}^*)\otimes(\mathbf{v}_q^* - \mathbf{v}^*)\right\rangle.$$

Proof. By definition,

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}_M(\overline{\mathbf{v}}_N) - \mathcal{R}_M(\mathbf{v}^*)\right] =& \underbrace{\mathbb{E}\left[\mathcal{R}_M(\overline{\mathbf{v}}_N)\right] - \mathbb{E}\left[\mathcal{R}_M^{(q)}(\overline{\mathbf{v}}_N)\right]}_{R_1} \\
&+ \underbrace{\mathbb{E}\left[\mathcal{R}_M^{(q)}(\overline{\mathbf{v}}_N)\right] - \mathbb{E}\left[\mathcal{R}_M^{(q)}(\mathbf{v}_q^*)\right]}_{R_2} \\
&+ \underbrace{\mathbb{E}\left[\mathcal{R}_M^{(q)}(\mathbf{v}_q^*)\right] - \mathbb{E}\left[\mathcal{R}_M(\mathbf{v}_q^*)\right]}_{R_3} \\
&+ \underbrace{\mathbb{E}\left[\mathcal{R}_M(\mathbf{v}_q^*)\right] - \mathcal{R}_M(\mathbf{v}^*)}_{R_4}.
\end{aligned}$$

By trivial computation and the unbiased quantization Assumption 3.1,

$$\begin{aligned}
R_1 =& \frac{1}{2}\mathbb{E}\left[(y - \langle\overline{\mathbf{v}}_N, \mathbf{S}\mathbf{x}\rangle)^2\right] - \frac{1}{2}\mathbb{E}\left[(\mathcal{Q}_l(y) - \langle\overline{\mathbf{v}}_N, \tilde{\mathbf{x}}^{(q)}\rangle)^2\right] \\
=& \frac{1}{2}\mathbb{E}\left[(y - \mathcal{Q}_l(y))(y + \mathcal{Q}_l(y))\right] - \frac{1}{2}\mathbb{E}\left[\langle\overline{\mathbf{v}}_N, \mathbf{S}\mathbf{x} - \tilde{\mathbf{x}}^{(q)}\rangle^2\right].
\end{aligned}$$

We can further expand the last negative term by $\overline{\mathbf{v}}_N = \overline{\mathbf{v}}_N - \mathbf{v}_q^* + \mathbf{v}_q^*$:

$$\begin{aligned}
\mathbb{E}\left[\langle\overline{\mathbf{v}}_N, \mathbf{S}\mathbf{x} - \tilde{\mathbf{x}}^{(q)}\rangle^2\right] =& \mathbb{E}\left[\overline{\mathbf{v}}_N^\top\left(\mathbf{H}_f^{(q)} - \mathbf{S}\mathbf{H}\mathbf{S}^\top\right)\overline{\mathbf{v}}_N\right] \\
=& \mathbb{E}\left[\left(\overline{\mathbf{v}}_N - \mathbf{v}_q^*\right)^\top\left(\mathbf{H}_f^{(q)} - \mathbf{S}\mathbf{H}\mathbf{S}^\top\right)\left(\overline{\mathbf{v}}_N - \mathbf{v}_q^*\right)\right] \\
&+ \mathbb{E}\left[\left(\mathbf{v}_q^*\right)^\top\left(\mathbf{H}_f^{(q)} - \mathbf{S}\mathbf{H}\mathbf{S}^\top\right)\mathbf{v}_q^*\right] \\
&+ 2\mathbb{E}\left[\left(\overline{\mathbf{v}}_N - \mathbf{v}_q^*\right)^\top\left(\mathbf{H}_f^{(q)} - \mathbf{S}\mathbf{H}\mathbf{S}^\top\right)\mathbf{v}_q^*\right].
\end{aligned}$$

By Lemma B.3, the parameter deviation $\boldsymbol{\eta}_t = \mathbf{v}_t - \mathbf{v}_q^*$ satisfies

$$\boldsymbol{\eta}_t = \left(\mathbf{I} - \frac{\gamma}{B}\widetilde{\mathbf{X}}_t^{(q)}(\widetilde{\mathbf{X}}_t^{(q)})^\top\right)\boldsymbol{\eta}_{t-1} + \frac{\gamma}{B}\widetilde{\mathbf{X}}_t^{(q)}\left(\boldsymbol{\xi}_t - (\widetilde{\mathbf{X}}_t^{(q)})^\top\boldsymbol{\epsilon}_{t-1}^{(p)} + \boldsymbol{\epsilon}_t^{(o)} - \boldsymbol{\epsilon}_t^{(a)}\right).$$

It follows that

$$\mathbb{E}[\boldsymbol{\eta}_t] = (\mathbf{I} - \gamma\mathbf{H}_f^{(q)})\mathbf{E}[\boldsymbol{\eta}_{t-1}] = (\mathbf{I} - \gamma\mathbf{H}_f^{(q)})^t\boldsymbol{\eta}_0, \tag{6}$$

where the first equality holds by Assumption 3.1 and the optimality $\mathbb{E}[(\mathcal{Q}_l(y) - \langle\mathbf{v}_q^*, \tilde{\mathbf{x}}^{(q)}\rangle)\tilde{\mathbf{x}}^{(q)}] = 0$. Hence,

$$\begin{aligned}
\mathbb{E}\left[\overline{\mathbf{v}}_N - \mathbf{v}_q^*\right]^\top\left(\mathbf{H}_f^{(q)} - \mathbf{SHS}^\top\right)\mathbf{v}_q^* =& \frac{1}{N}\sum_{t=0}^{N-1}\mathbb{E}[\boldsymbol{\eta}_t]^\top\left(\mathbf{H}_f^{(q)} - \mathbf{SHS}^\top\right)\mathbf{v}_q^* \\
=& \left[\frac{1}{N}\sum_{t=0}^{N-1}\left(\mathbf{I} - \gamma\mathbf{H}_f^{(q)}\right)^t\boldsymbol{\eta}_0\right]^\top\left(\mathbf{H}_f^{(q)} - \mathbf{SHS}^\top\right)\mathbf{v}_q^* \\
=& -\left(\mathbf{v}_q^*\right)^\top\frac{1}{N\gamma}\left[\mathbf{I} - \left(\mathbf{I} - \gamma\mathbf{H}_f^{(q)}\right)^N\right]\left(\mathbf{H}_f^{(q)}\right)^{-1}\left(\mathbf{H}_f^{(q)} - \mathbf{SHS}^\top\right)\mathbf{v}_q^*.
\end{aligned}$$

Notice that by the optimality $\mathbb{E}[(\mathcal{Q}_l(y) - \langle\mathbf{v}_q^*, \tilde{\mathbf{x}}^{(q)}\rangle)\tilde{\mathbf{x}}^{(q)}] = 0$ again, it holds

$$\begin{aligned}
R_2 =& \frac{1}{2}\mathbb{E}\left[(\mathcal{Q}_l(y) - \langle\overline{\mathbf{v}}_N, \tilde{\mathbf{x}}^{(q)}\rangle)^2\right] - \frac{1}{2}\mathbb{E}\left[(\mathcal{Q}_l(y) - \langle\mathbf{v}_q^*, \tilde{\mathbf{x}}^{(q)}\rangle)^2\right] \\
=& \frac{1}{2}\mathbb{E}\left[\langle\mathbf{v}_q^* - \overline{\mathbf{v}}_N, \tilde{\mathbf{x}}^{(q)}\rangle\left(2\mathcal{Q}_l(y) - \langle\mathbf{v}_q^* + \overline{\mathbf{v}}_N, \tilde{\mathbf{x}}^{(q)}\rangle\right)\right] \\
=& \frac{1}{2}\left\langle\mathbf{H}_f^{(q)}, \mathbb{E}\left[(\mathbf{v}_q^* - \overline{\mathbf{v}}_N)\otimes(\mathbf{v}_q^* - \overline{\mathbf{v}}_N)\right]\right\rangle.
\end{aligned}$$

Similarly, by Assumption 3.1,

$$\begin{aligned}
R_3 =& \frac{1}{2}\mathbb{E}\left[(\mathcal{Q}_l(y) - \langle\mathbf{v}_q^*, \tilde{\mathbf{x}}^{(q)}\rangle)^2\right] - \frac{1}{2}\mathbb{E}\left[\left(y - \langle\mathbf{v}_q^*, \mathbf{Sx}\rangle\right)^2\right] \\
=& \frac{1}{2}\mathbb{E}\left[(\mathcal{Q}_l(y) - y)(y + \mathcal{Q}_l(y))\right] + \frac{1}{2}\mathbb{E}\left[\langle\mathbf{v}_q^*, \mathbf{Sx} - \tilde{\mathbf{x}}^{(q)}\rangle^2\right].
\end{aligned}$$

By the optimality $\mathbb{E}[(y - \langle\mathbf{v}^*, \mathbf{Sx}\rangle)\mathbf{Sx}] = 0$, we have

$$\begin{aligned}
R_4 =& \frac{1}{2}\mathbb{E}\left[\left(y - \langle\mathbf{v}_q^*, \mathbf{Sx}\rangle\right)^2\right] - \frac{1}{2}\mathbb{E}\left[(y - \langle\mathbf{v}^*, \mathbf{Sx}\rangle)^2\right] \\
=& \frac{1}{2}\mathbb{E}\left[\left(\langle\mathbf{v}^* - \mathbf{v}_q^*, \mathbf{Sx}\rangle\right)\left(2y - \langle\mathbf{v}^* + \mathbf{v}_q^*, \mathbf{Sx}\rangle\right)\right] \\
=& \frac{1}{2}\left\langle\mathbf{SHS}^\top, (\mathbf{v}_q^* - \mathbf{v}^*)\otimes(\mathbf{v}_q^* - \mathbf{v}^*)\right\rangle.
\end{aligned}$$

Together with the derivations above, the excess risk

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}_M(\overline{\mathbf{v}}_N) - \mathcal{R}_M(\mathbf{v}^*)\right] =& \left(\mathbf{v}_q^*\right)^\top\frac{1}{N\gamma}\left[\mathbf{I} - \left(\mathbf{I} - \gamma\mathbf{H}_f^{(q)}\right)^N\right]\left(\mathbf{H}_f^{(q)}\right)^{-1}\left(\mathbf{H}_f^{(q)} - \mathbf{SHS}^\top\right)\mathbf{v}_q^* \\
&+ \frac{1}{2}\left\langle\mathbf{SHS}^\top, \mathbb{E}\left[(\mathbf{v}_q^* - \overline{\mathbf{v}}_N)\otimes(\mathbf{v}_q^* - \overline{\mathbf{v}}_N)\right]\right\rangle \\
&+ \frac{1}{2}\left\langle\mathbf{SHS}^\top, (\mathbf{v}_q^* - \mathbf{v}^*)\otimes(\mathbf{v}_q^* - \mathbf{v}^*)\right\rangle.
\end{aligned}$$

□

Lemma B.3 (Parameter update rule)**.** *Define $\boldsymbol{\eta}_t = \mathbf{v}_t - \mathbf{v}_q^*$. Denote the quantized sample batch $\widetilde{\mathbf{X}}_t^{(q)} = \mathcal{Q}_f\left(\mathcal{Q}_s(\mathbf{S})\mathcal{Q}_d(\mathbf{X}_t)^\top\right) \in \mathbb{R}^{M\times B}$, the sample noise $\boldsymbol{\xi}_t = \mathcal{Q}_l(\mathbf{y}_t) - (\widetilde{\mathbf{X}}_t^{(q)})^\top\mathbf{v}_q^*$, and the sample quantization errors: for $t = 0, 1, 2, ...,$*

$$\begin{aligned}
\boldsymbol{\epsilon}_t^{(o)} :=& \mathcal{Q}_o\left(\mathcal{Q}_l(\mathbf{y}_t) - \mathcal{Q}_a\left((\widetilde{\mathbf{X}}_t^{(q)})^\top\mathcal{Q}_p(\mathbf{v}_{t-1})\right)\right) - \left[\mathcal{Q}_l(\mathbf{y}_t) - \mathcal{Q}_a\left((\widetilde{\mathbf{X}}_t^{(q)})^\top\mathcal{Q}_p(\mathbf{v}_{t-1})\right)\right], \\
\boldsymbol{\epsilon}_t^{(a)} :=& \mathcal{Q}_a\left((\widetilde{\mathbf{X}}_t^{(q)})^\top\mathcal{Q}_p(\mathbf{v}_{t-1})\right) - (\widetilde{\mathbf{X}}_t^{(q)})^\top\mathcal{Q}_p(\mathbf{v}_{t-1}), \\
\boldsymbol{\epsilon}_t^{(p)} :=& \mathcal{Q}_p(\mathbf{v}_t) - \mathbf{v}_t.
\end{aligned}$$

The quantized SGD implies

$$\boldsymbol{\eta}_t = \left(\mathbf{I} - \frac{\gamma}{B}\widetilde{\mathbf{X}}_t^{(q)}(\widetilde{\mathbf{X}}_t^{(q)})^\top\right)\boldsymbol{\eta}_{t-1} + \frac{\gamma}{B}\widetilde{\mathbf{X}}_t^{(q)}\left(\boldsymbol{\xi}_t - (\widetilde{\mathbf{X}}_t^{(q)})^\top\boldsymbol{\epsilon}_{t-1}^{(p)} + \boldsymbol{\epsilon}_t^{(o)} - \boldsymbol{\epsilon}_t^{(a)}\right).$$

Proof. Via the quantized SGD, the following holds:

$$\begin{aligned}
\boldsymbol{\eta}_t =& \boldsymbol{\eta}_{t-1} + \frac{\gamma}{B}\widetilde{\mathbf{X}}_t^{(q)}\mathcal{Q}_o\left(\mathcal{Q}_l(\mathbf{y}_t) - \mathcal{Q}_a((\widetilde{\mathbf{X}}_t^{(q)})^\top\mathcal{Q}_p(\mathbf{v}_{t-1}))\right)\\
=& \boldsymbol{\eta}_{t-1} + \frac{\gamma}{B}\widetilde{\mathbf{X}}_t^{(q)}\left[\mathcal{Q}_l(\mathbf{y}_t) - (\widetilde{\mathbf{X}}_t^{(q)})^\top\mathcal{Q}_p(\mathbf{v}_{t-1}) + \boldsymbol{\epsilon}_t^{(o)} - \boldsymbol{\epsilon}_t^{(a)}\right]\\
=& \boldsymbol{\eta}_{t-1} + \frac{\gamma}{B}\widetilde{\mathbf{X}}_t^{(q)}\left[\boldsymbol{\xi}_t - (\widetilde{\mathbf{X}}_t^{(q)})^\top\left(\boldsymbol{\epsilon}_{t-1}^{(p)} + \mathbf{v}_{t-1} - \mathbf{v}_q^*\right) + \boldsymbol{\epsilon}_t^{(o)} - \boldsymbol{\epsilon}_t^{(a)}\right]\\
=& \left(\mathbf{I} - \frac{\gamma}{B}\widetilde{\mathbf{X}}_t^{(q)}(\widetilde{\mathbf{X}}_t^{(q)})^\top\right)\boldsymbol{\eta}_{t-1} + \frac{\gamma}{B}\widetilde{\mathbf{X}}_t^{(q)}\left(\boldsymbol{\xi}_t - (\widetilde{\mathbf{X}}_t^{(q)})^\top\boldsymbol{\epsilon}_{t-1}^{(p)} + \boldsymbol{\epsilon}_t^{(o)} - \boldsymbol{\epsilon}_t^{(a)}\right).
\end{aligned}$$

□

We provide upper bounds for excess risk by analyzing R_N and R_{add} respectively according to Lemma B.2. We first establish the bias-variance decomposition for $\mathbb{E}[\boldsymbol{\eta}_t \otimes \boldsymbol{\eta}_t]$. By Lemma B.3,

$$\boldsymbol{\eta}_t = \left(\mathbf{I} - \frac{\gamma}{B}\widetilde{\mathbf{X}}_t^{(q)}(\widetilde{\mathbf{X}}_t^{(q)})^\top\right)\boldsymbol{\eta}_{t-1} + \frac{\gamma}{B}\widetilde{\mathbf{X}}_t^{(q)}\left(\boldsymbol{\xi}_t - (\widetilde{\mathbf{X}}_t^{(q)})^\top\boldsymbol{\epsilon}_{t-1}^{(p)} + \boldsymbol{\epsilon}_t^{(o)} - \boldsymbol{\epsilon}_t^{(a)}\right).$$

Denote

$$\boldsymbol{\eta}_t^{\text{bias}} = \left(\mathbf{I} - \frac{\gamma}{B}\widetilde{\mathbf{X}}_t^{(q)}(\widetilde{\mathbf{X}}_t^{(q)})^\top\right)\boldsymbol{\eta}_{t-1}^{\text{bias}}, \quad \boldsymbol{\eta}_0^{\text{bias}} = \boldsymbol{\eta}_0,$$

$$\boldsymbol{\eta}_t^{\text{var}} = \left(\mathbf{I} - \frac{\gamma}{B}\widetilde{\mathbf{X}}_t^{(q)}(\widetilde{\mathbf{X}}_t^{(q)})^\top\right)\boldsymbol{\eta}_{t-1}^{\text{var}} + \frac{\gamma}{B}\widetilde{\mathbf{X}}_t^{(q)}\left(\boldsymbol{\xi}_t - (\widetilde{\mathbf{X}}_t^{(q)})^\top\boldsymbol{\epsilon}_{t-1}^{(p)} + \boldsymbol{\epsilon}_t^{(o)} - \boldsymbol{\epsilon}_t^{(a)}\right), \quad \boldsymbol{\eta}_0^{\text{var}} = \mathbf{0}.$$

Then it holds

$$\boldsymbol{\eta}_t = \boldsymbol{\eta}_t^{\text{bias}} + \boldsymbol{\eta}_t^{\text{var}}.$$

To proceed, we upper bound $\mathbb{E}\left[\boldsymbol{\eta}_t \otimes \boldsymbol{\eta}_t\right]$ by

$$\mathbb{E}\left[\boldsymbol{\eta}_t \otimes \boldsymbol{\eta}_t\right] \preceq 2\left(\underbrace{\mathbb{E}\left[\boldsymbol{\eta}_t^{\text{bias}} \otimes \boldsymbol{\eta}_t^{\text{bias}}\right]}_{\mathbf{B}_t} + \underbrace{\mathbb{E}\left[\boldsymbol{\eta}_t^{\text{var}} \otimes \boldsymbol{\eta}_t^{\text{var}}\right]}_{\mathbf{C}_t}\right), \tag{7}$$

where

$$\begin{aligned}
\mathbf{B}_t &= \mathbb{E}\left[\left(\mathbf{I} - \frac{\gamma}{B}\widetilde{\mathbf{X}}_t^{(q)}(\widetilde{\mathbf{X}}_t^{(q)})^\top\right)\mathbf{B}_{t-1}\left(\mathbf{I} - \frac{\gamma}{B}\widetilde{\mathbf{X}}_t^{(q)}(\widetilde{\mathbf{X}}_t^{(q)})^\top\right)\right].\\
\mathbf{C}_t &= \mathbb{E}\left[\left(\mathbf{I} - \frac{\gamma}{B}\widetilde{\mathbf{X}}_t^{(q)}(\widetilde{\mathbf{X}}_t^{(q)})^\top\right)\mathbf{C}_{t-1}\left(\mathbf{I} - \frac{\gamma}{B}\widetilde{\mathbf{X}}_t^{(q)}(\widetilde{\mathbf{X}}_t^{(q)})^\top\right)\right] + \gamma^2\boldsymbol{\Sigma}_t,
\end{aligned}$$

with [9]

$$\begin{aligned}
\boldsymbol{\Sigma}_t =& \frac{1}{B^2}\mathbb{E}\left[\widetilde{\mathbf{X}}_t^{(q)}\left(\boldsymbol{\xi}_t - (\widetilde{\mathbf{X}}_t^{(q)})^\top\boldsymbol{\epsilon}_{t-1}^{(p)} + \boldsymbol{\epsilon}_t^{(o)} - \boldsymbol{\epsilon}_t^{(a)}\right)\left(\boldsymbol{\xi}_t - (\widetilde{\mathbf{X}}_t^{(q)})^\top\boldsymbol{\epsilon}_{t-1}^{(p)} + \boldsymbol{\epsilon}_t^{(o)} - \boldsymbol{\epsilon}_t^{(a)}\right)^\top(\widetilde{\mathbf{X}}_t^{(q)})^\top\right]\\
=& \frac{1}{B^2}\mathbb{E}\left[\widetilde{\mathbf{X}}_t^{(q)}\boldsymbol{\xi}_t\boldsymbol{\xi}_t^\top(\widetilde{\mathbf{X}}_t^{(q)})^\top\right] + \frac{1}{B^2}\mathbb{E}\left[\widetilde{\mathbf{X}}_t^{(q)}(\widetilde{\mathbf{X}}_t^{(q)})^\top\boldsymbol{\epsilon}_{t-1}^{(p)}(\boldsymbol{\epsilon}_{t-1}^{(p)})^\top\widetilde{\mathbf{X}}_t^{(q)}(\widetilde{\mathbf{X}}_t^{(q)})^\top\right]\\
&+ \frac{1}{B^2}\mathbb{E}\left[\widetilde{\mathbf{X}}_t^{(q)}\boldsymbol{\epsilon}_t^{(a)}(\boldsymbol{\epsilon}_t^{(a)})^\top(\widetilde{\mathbf{X}}_t^{(q)})^\top\right] + \frac{1}{B^2}\mathbb{E}\left[\widetilde{\mathbf{X}}_t^{(q)}\boldsymbol{\epsilon}_t^{(o)}(\boldsymbol{\epsilon}_t^{(o)})^\top(\widetilde{\mathbf{X}}_t^{(q)})^\top\right].
\end{aligned} \tag{8}$$

Define $\overline{\boldsymbol{\eta}}_N = \frac{1}{N}\sum_{t=0}^{N-1}\mathbf{v}_t - \mathbf{v}_q^*$. The bias-variance decomposition is established below.

Lemma B.4. *Under Assumption 3.1 and Assumption 3.3, if* $\gamma < \frac{1}{\|\mathbf{H}_f^{(q)}\|}$*, then* R_N *satisfies*

$$R_N \leq \frac{\mu_{\max}\left((\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}\mathbf{S}^\top\right)}{N^2}\cdot\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle(\mathbf{I} - \gamma\mathbf{H}_f^{(q)})^{k-t}\mathbf{H}_f^{(q)}, \mathbb{E}[\boldsymbol{\eta}_t \otimes \boldsymbol{\eta}_t]\right\rangle.$$

Further, by (7),

$$R_N \leq \frac{2\mu_{\max}\left((\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}\mathbf{S}^\top\right)}{N^2}\cdot\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle(\mathbf{I} - \gamma\mathbf{H}_f^{(q)})^{k-t}\mathbf{H}_f^{(q)}, \mathbf{B}_t + \mathbf{C}_t\right\rangle.$$

[9] Here we use Assumption 3.1 and $\boldsymbol{\eta}_0^{\text{var}} = 0$ to derive the recursion of $\mathbf{C}_t$.

Proof. By definition $\overline{\boldsymbol{\eta}}_N = \frac{1}{N}\sum_{t=0}^{N-1}\boldsymbol{\eta}_t$, we have

$$\begin{aligned}\mathbb{E}[\overline{\boldsymbol{\eta}}_N \otimes \overline{\boldsymbol{\eta}}_N] =& \frac{1}{N^2}\cdot\left(\sum_{0\le k\le t\le N-1}\mathbb{E}[\boldsymbol{\eta}_t\otimes\boldsymbol{\eta}_k] + \sum_{0\le t<k\le N-1}\mathbb{E}[\boldsymbol{\eta}_t\otimes\boldsymbol{\eta}_k]\right)\\ \preceq& \frac{1}{N^2}\cdot\left(\sum_{0\le k\le t\le N-1}\mathbb{E}\left[\mathbb{E}[\boldsymbol{\eta}_t\otimes\boldsymbol{\eta}_k|\boldsymbol{\eta}_k]\right] + \sum_{0\le t\le k\le N-1}\mathbb{E}\left[\mathbb{E}[\boldsymbol{\eta}_t\otimes\boldsymbol{\eta}_k|\boldsymbol{\eta}_t]\right]\right).\end{aligned}$$

Proceed with (6), it follows

$$\begin{aligned}\mathbb{E}[\overline{\boldsymbol{\eta}}_N \otimes \overline{\boldsymbol{\eta}}_N] \preceq& \frac{1}{N^2}\cdot\left(\sum_{0\le k\le t\le N-1}(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{t-k}\mathbb{E}[\boldsymbol{\eta}_k\otimes\boldsymbol{\eta}_k] + \sum_{0\le t\le k\le N-1}\mathbb{E}[\boldsymbol{\eta}_t\otimes\boldsymbol{\eta}_t](\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{k-t}\right)\\ =& \frac{1}{N^2}\cdot\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left((\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{k-t}\mathbb{E}[\boldsymbol{\eta}_t\otimes\boldsymbol{\eta}_t] + \mathbb{E}[\boldsymbol{\eta}_t\otimes\boldsymbol{\eta}_t](\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{k-t}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}R_N \le& \frac{\mu_{\max}\left((\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}\mathbf{S}^\top\right)}{2}\left\langle\mathbf{H}_f^{(q)}, \mathbb{E}[\overline{\boldsymbol{\eta}}_N\otimes\overline{\boldsymbol{\eta}}_N]\right\rangle\\ \le& \frac{\mu_{\max}\left((\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}\mathbf{S}^\top\right)}{2N^2}\cdot\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle\mathbf{H}_f^{(q)}, (\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{k-t}\mathbb{E}[\boldsymbol{\eta}_t\otimes\boldsymbol{\eta}_t] + \mathbb{E}[\boldsymbol{\eta}_t\otimes\boldsymbol{\eta}_t](\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{k-t}\right\rangle\\ =& \frac{\mu_{\max}\left((\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}\mathbf{S}^\top\right)}{N^2}\cdot\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{k-t}\mathbf{H}_f^{(q)}, \mathbb{E}[\boldsymbol{\eta}_t\otimes\boldsymbol{\eta}_t]\right\rangle,\end{aligned}$$

where the last equality holds since $\mathbf{H}_f^{(q)}$ and $(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{k-t}$ commute. This completes the proof. □

B.3 Bias Error

Denote $\text{bias} = \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{k-t}\mathbf{H}_f^{(q)}, \mathbf{B}_t\right\rangle$, $\mathbf{S}_n = \sum_{t=0}^{n-1}\mathbf{B}_t$. By definition,

$$\begin{aligned}\mathbf{S}_t =& \sum_{k=0}^{t-1}(\mathcal{I}-\gamma\mathcal{T}_B^{(q)})^k\circ\mathbf{B}_0\\ =& (\mathcal{I}-\gamma\mathcal{T}_B^{(q)})\circ\left(\sum_{k=1}^{t-1}(\mathcal{I}-\gamma\mathcal{T}_B^{(q)})^{k-1}\circ\mathbf{B}_0\right)+\mathbf{B}_0 \qquad (9)\\ =& (\mathcal{I}-\gamma\mathcal{T}_B^{(q)})\circ\mathbf{S}_{t-1}+\mathbf{B}_0\\ =& (\mathcal{I}-\gamma\widetilde{\mathcal{T}}^{(q)})\circ\mathbf{S}_{t-1}+\gamma^2(\mathcal{M}_B^{(q)}-\widetilde{\mathcal{M}}^{(q)})\circ\mathbf{S}_{t-1}+\mathbf{B}_0.\end{aligned}$$

The next lemma provides an upper bound for $(\mathcal{M}_B^{(q)}-\widetilde{\mathcal{M}}^{(q)})\circ\mathbf{S}_{t-1}$.

Lemma B.5. *Under Assumption 3.3, 3.4, if step size* $\gamma < \frac{B}{\alpha\text{tr}(\mathbf{H}_f^{(q)})}\wedge\frac{1}{\mu_{\max}(\mathbf{H}_f^{(q)})}$,

$$(\mathcal{M}_B^{(q)}-\widetilde{\mathcal{M}}^{(q)})\circ\mathbf{S}_t \preceq \frac{\frac{\alpha}{B}\text{tr}\left((\mathcal{I}-(\mathcal{I}-\gamma\widetilde{\mathcal{T}}^{(q)})^t)\circ\mathbf{B}_0\right)}{\gamma(1-\gamma\frac{\alpha}{B}\text{tr}(\mathbf{H}_f^{(q)}))}\mathbf{H}_f^{(q)}.$$

Proof. Take summation via the update rule, we have

$$\begin{aligned}\mathbf{S}_t =& \sum_{k=0}^{t-1}(\mathcal{I}-\gamma\mathcal{T}_B^{(q)})^k\circ\mathbf{B}_0\\ =& \gamma^{-1}{\mathcal{T}_B^{(q)}}^{-1}\circ\left[\mathcal{I}-(\mathcal{I}-\gamma\mathcal{T}_B^{(q)})^t\right]\circ\mathbf{B}_0\\ \preceq& \gamma^{-1}{\mathcal{T}_B^{(q)}}^{-1}\circ(\mathcal{I}-(\mathcal{I}-\gamma\widetilde{\mathcal{T}}^{(q)})^t)\circ\mathbf{B}_0.\end{aligned}$$

Denote $\mathbf{A} := (\mathcal{I} - (\mathcal{I} - \gamma\widetilde{\mathcal{T}}^{(q)})^t) \circ \mathbf{B}_0$. As

$$\widetilde{\mathcal{T}}^{(q)} \circ {\mathcal{T}_B^{(q)}}^{-1} \circ \mathbf{A} = \gamma(\mathcal{M}_B^{(q)} - \widetilde{\mathcal{M}}^{(q)}) \circ {\mathcal{T}_B^{(q)}}^{-1} \circ \mathbf{A} + \mathbf{A},$$

it holds,

$${\mathcal{T}_B^{(q)}}^{-1} \circ \mathbf{A} \preceq \gamma{(\widetilde{\mathcal{T}}^{(q)})}^{-1} \circ (\mathcal{M}_B^{(q)} - \widetilde{\mathcal{M}}^{(q)}) \circ {\mathcal{T}_B^{(q)}}^{-1} \circ \mathbf{A} + {(\widetilde{\mathcal{T}}^{(q)})}^{-1} \circ \mathbf{A}.$$

Then we apply $\mathcal{M}_B^{(q)} - \widetilde{\mathcal{M}}^{(q)}$ on both sides.

$$\begin{aligned}
&(\mathcal{M}_B^{(q)} - \widetilde{\mathcal{M}}^{(q)}) \circ {\mathcal{T}_B^{(q)}}^{-1} \circ \mathbf{A} \\
\preceq&(\mathcal{M}_B^{(q)} - \widetilde{\mathcal{M}}^{(q)}) \circ \gamma{(\widetilde{\mathcal{T}}^{(q)})}^{-1} \circ (\mathcal{M}_B^{(q)} - \widetilde{\mathcal{M}}^{(q)}) \circ {\mathcal{T}_B^{(q)}}^{-1} \circ \mathbf{A} + (\mathcal{M}_B^{(q)} - \widetilde{\mathcal{M}}^{(q)}) \circ {(\widetilde{\mathcal{T}}^{(q)})}^{-1} \circ \mathbf{A} \\
\preceq&\sum_{t=0}^{\infty} \left[\gamma(\mathcal{M}_B^{(q)} - \widetilde{\mathcal{M}}^{(q)}) \circ {(\widetilde{\mathcal{T}}^{(q)})}^{-1}\right]^t \circ \left((\mathcal{M}_B^{(q)} - \widetilde{\mathcal{M}}^{(q)}) \circ {(\widetilde{\mathcal{T}}^{(q)})}^{-1} \circ \mathbf{A}\right) \text{ (By recursion).}
\end{aligned}$$

By Lemma B.1,

$$\begin{aligned}
(\mathcal{M}_B^{(q)} - \widetilde{\mathcal{M}}^{(q)}) \circ {(\widetilde{\mathcal{T}}^{(q)})}^{-1} \circ \mathbf{A} &\preceq \frac{\alpha}{B} \operatorname{tr}(\mathbf{H}_f^{(q)} {(\widetilde{\mathcal{T}}^{(q)})}^{-1} \circ \mathbf{A}) \mathbf{H}_f^{(q)} \\
&= \frac{\alpha}{B}\gamma \operatorname{tr}\left(\sum_{t=0}^{\infty} \mathbf{H}_f^{(q)} (\mathbf{I} - \gamma\mathbf{H}_f^{(q)})^t \mathbf{A} (\mathbf{I} - \gamma\mathbf{H}_f^{(q)})^t\right) \mathbf{H}_f^{(q)} \\
&= \frac{\alpha}{B} \operatorname{tr}\left(\mathbf{H}_f^{(q)} (2\mathbf{H}_f^{(q)} - \gamma(\mathbf{H}_f^{(q)})^2)^{-1} \mathbf{A}\right) \mathbf{H}_f^{(q)} \\
&\preceq \frac{\alpha}{B} \operatorname{tr}(\mathbf{A}) \mathbf{H}_f^{(q)},
\end{aligned}$$

where the last inequality holds as the condition that $\gamma < \frac{1}{\mu_{\max}(\mathbf{H}_f^{(q)})}$. Hence, together with ${(\widetilde{\mathcal{T}}^{(q)})}^{-1} \mathbf{H}_f^{(q)} \preceq \mathbf{I}$ and $(\mathcal{M}_B^{(q)} - \widetilde{\mathcal{M}}^{(q)}) \circ \mathbf{I} \preceq \frac{\alpha}{B} \operatorname{tr}(\mathbf{H}_f^{(q)}) \mathbf{H}_f^{(q)}$, we obtain

$$\begin{aligned}
(\mathcal{M}_B^{(q)} - \widetilde{\mathcal{M}}^{(q)}) \circ {\mathcal{T}_B^{(q)}}^{-1} \circ \mathbf{A} &\preceq \frac{\alpha}{B} \operatorname{tr}(\mathbf{A}) \sum_{t=0}^{\infty} (\gamma \frac{\alpha}{B} \operatorname{tr}(\mathbf{H}_f^{(q)}))^t \mathbf{H}_f^{(q)} \\
&\preceq \frac{\frac{\alpha}{B} \operatorname{tr}(\mathbf{A})}{1 - \gamma \frac{\alpha}{B} \operatorname{tr}(\mathbf{H}_f^{(q)})} \mathbf{H}_f^{(q)}.
\end{aligned}$$

Therefore,

$$(\mathcal{M}_B^{(q)} - \widetilde{\mathcal{M}}^{(q)}) \circ \mathbf{S}_t \preceq \gamma^{-1} (\mathcal{M}_B^{(q)} - \widetilde{\mathcal{M}}^{(q)}) \circ {\mathcal{T}_B^{(q)}}^{-1} \circ \mathbf{A} \preceq \frac{\frac{\alpha}{B} \operatorname{tr}\left((\mathcal{I} - (\mathcal{I} - \gamma\widetilde{\mathcal{T}}^{(q)})^t) \circ \mathbf{B}_0\right)}{\gamma(1 - \gamma \frac{\alpha}{B} \operatorname{tr}(\mathbf{H}_f^{(q)}))} \mathbf{H}_f^{(q)}.$$

□

Lemma B.6. *Under Assumption 3.3, Assumption 3.4, if the step size satisfies* $\gamma < \frac{B}{\alpha \operatorname{tr}(\mathbf{H}_f^{(q)})} \wedge \frac{1}{\mu_{\max}(\mathbf{H}_f^{(q)})}$, *then*

$$\begin{aligned}
\text{bias} \leq& \frac{2\frac{\alpha}{B}\left(\|\mathbf{v}_q^*\|^2_{\mathbf{I}_{f,0:k^*}^{(q)}} + N\gamma\|\mathbf{v}_q^*\|^2_{\mathbf{H}_{f,k^*:\infty}^{(q)}}\right)}{N\gamma(1 - \gamma\frac{\alpha}{B}\operatorname{tr}(\mathbf{H}_f^{(q)}))} \cdot \left(\frac{k^*}{N} + N\gamma^2 \sum_{i>k^*} (\tilde{\lambda}_i^{(q)})^2\right) \\
&+ \frac{1}{\gamma^2 N^2} \cdot \|\mathbf{v}_q^*\|^2_{(\mathbf{H}_{f,0:k^*}^{(q)})^{-1}} + \|\mathbf{v}_q^*\|^2_{\mathbf{H}_{f,k^*:\infty}^{(q)}}.
\end{aligned}$$

Proof. By (9) and Lemma B.5,

$$\begin{aligned}
\mathbf{S}_t \preceq&(\mathcal{I} - \gamma\tilde{\mathcal{T}}^{(q)}) \circ \mathbf{S}_{t-1} + \frac{\gamma\frac{\alpha}{B} \cdot \operatorname{tr}\left(\left[\mathcal{I} - (\mathcal{I} - \gamma\widetilde{\mathcal{T}}^{(q)})^N\right] \circ \mathbf{B}_0\right)}{(1 - \gamma\alpha/B \operatorname{tr}(\mathbf{H}_f^{(q)}))} \cdot \mathbf{H}_f^{(q)} + \mathbf{B}_0 \\
\preceq&\sum_{k=0}^{t-1} (\mathbf{I} - \gamma\mathbf{H}_f^{(q)})^k \left(\frac{\gamma\frac{\alpha}{B} \cdot \operatorname{tr}\left(\mathbf{B}_0 - (\mathbf{I} - \gamma\mathbf{H}_f^{(q)})^N \mathbf{B}_0 (\mathbf{I} - \gamma\mathbf{H}_f^{(q)})^N\right)}{(1 - \gamma\alpha/B \operatorname{tr}(\mathbf{H}_f^{(q)}))} \cdot \mathbf{H}_f^{(q)} + \mathbf{B}_0\right) (\mathbf{I} - \gamma\mathbf{H}_f^{(q)})^k.
\end{aligned}$$

Substituting into bias,

$$\begin{aligned}
\text{bias} =& \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle (\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{k-t}\mathbf{H}_f^{(q)}, \mathbf{B}_t\right\rangle \\
\leq& \frac{1}{\gamma N^2}\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^N, \sum_{t=0}^{N-1}\mathbf{B}_t\rangle \\
\leq& \frac{1}{\gamma N^2}\sum_{k=0}^{N-1}\left\langle (\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^k - (\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{N+k}, \frac{\gamma\frac{\alpha}{B}\cdot\operatorname{tr}\left(\mathbf{B}_0-(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^N\mathbf{B}_0(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^N\right)}{(1-\gamma\alpha/B\operatorname{tr}(\mathbf{H}_f^{(q)}))}\cdot\mathbf{H}_f^{(q)}+\mathbf{B}_0\right\rangle.
\end{aligned}$$

It suffices to upper bound the following two terms

$$\begin{aligned}
I_1 &= \frac{\frac{\alpha}{B}\cdot\operatorname{tr}(\mathbf{B}_{0,N})}{N^2(1-\gamma\alpha/B\operatorname{tr}(\mathbf{H}_f^{(q)}))}\sum_{k=0}^{N-1}\left\langle (\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^k-(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{N+k}, \mathbf{H}_f^{(q)}\right\rangle, \\
I_2 &= \frac{1}{\gamma N^2}\sum_{k=0}^{N-1}\left\langle (\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^k-(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{N+k}, \mathbf{B}_0\right\rangle,
\end{aligned}$$

where $\mathbf{B}_{a,b} \coloneqq \mathbf{B}_a - (\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{b-a}\mathbf{B}_a(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{b-a}$. For I_1,

$$\begin{aligned}
I_1 &= \frac{\frac{\alpha}{B}\operatorname{tr}(\mathbf{B}_{0,N})}{N^2(1-\gamma\frac{\alpha}{B}\operatorname{tr}(\mathbf{H}_f^{(q)}))}\sum_{k=0}^{N-1}\sum_i\left[(1-\gamma\tilde{\lambda}_i^{(q)})^k-(1-\gamma\tilde{\lambda}_i^{(q)})^{N+k}\right]\tilde{\lambda}_i^{(q)} \\
&= \frac{\frac{\alpha}{B}\operatorname{tr}(\mathbf{B}_{0,N})}{\gamma N^2(1-\gamma\frac{\alpha}{B}\operatorname{tr}(\mathbf{H}_f^{(q)}))}\sum_i\left[1-(1-\gamma\tilde{\lambda}_i^{(q)})^N\right]^2 \\
&\leq \frac{\frac{\alpha}{B}\operatorname{tr}(\mathbf{B}_{0,N})}{\gamma(1-\gamma\frac{\alpha}{B}\operatorname{tr}(\mathbf{H}_f^{(q)}))}\cdot\left(\frac{k^*}{N^2}+\gamma^2\sum_{i>k^*}(\tilde{\lambda}_i^{(q)})^2\right),
\end{aligned}$$

where $k^* = \max\{k : \tilde{\lambda}_k^{(q)} \geq \frac{1}{N\gamma}\}$ and $\{\tilde{\lambda}_i^{(q)}\}_{i=1}^M$ are eigenvalues of $\mathbf{H}_f^{(q)}$. Then we tackle $\operatorname{tr}(\mathbf{B}_{0,N})$.

$$\begin{aligned}
\operatorname{tr}(\mathbf{B}_{0,N}) &= \operatorname{tr}\left(\mathbf{B}_0-(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^N\mathbf{B}_0(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^N\right) \\
&= \sum_i\left(1-(1-\gamma\tilde{\lambda}_i^{(q)})^{2N}\right)\cdot\left(\langle\mathbf{v}_q^*, \mathbf{v}_i^{(q)}\rangle\right)^2 \\
&\leq 2\sum_i\min\{1, N\gamma\tilde{\lambda}_i^{(q)}\}\left(\langle\mathbf{v}_q^*, \mathbf{v}_i^{(q)}\rangle\right)^2 \\
&\leq 2\left(\|\mathbf{v}_q^*\|^2_{\mathbf{I}^{(q)}_{f,0:k^*}}+N\gamma\|\mathbf{v}_q^*\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}\right).
\end{aligned} \tag{10}$$

Hence,

$$I_1 \leq \frac{2\frac{\alpha}{B}\left(\|\mathbf{v}_q^*\|^2_{\mathbf{I}^{(q)}_{f,0:k^*}}+N\gamma\|\mathbf{v}_q^*\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}\right)}{N\gamma(1-\gamma\frac{\alpha}{B}\operatorname{tr}(\mathbf{H}_f^{(q)}))}\cdot\left(\frac{k^*}{N}+N\gamma^2\sum_{i>k^*}(\tilde{\lambda}_i^{(q)})^2\right).$$

For I_2, decompose $\mathbf{H}_f^{(q)} = \mathbf{V}^{(q)}\mathbf{\Lambda}^{(q)}\mathbf{V}^{(q)\top}$, then

$$I_2 = \frac{1}{\gamma N^2}\sum_{k=0}^{N-1}\left\langle (\mathbf{I}-\gamma\mathbf{\Lambda}^{(q)})^k-(\mathbf{I}-\gamma\mathbf{\Lambda}^{(q)})^{N+k}, \mathbf{V}^{(q)\top}\mathbf{B}_0\mathbf{V}^{(q)}\right\rangle.$$

Note that $\mathbf{B}_0 = \boldsymbol{\eta}_0\boldsymbol{\eta}_0^\top$, it can be shown that the diagonal entries of ${\mathbf{V}^{(q)}}^\top \mathbf{B}_0 \mathbf{V}^{(q)}$ are $\omega_1^2, \ldots,$ where $\omega_i = {\mathbf{v}_i^{(q)}}^\top \boldsymbol{\eta}_0 = {\mathbf{v}_i^{(q)}}^\top (\mathbf{v}_0 - \mathbf{v}_q^*)$. Hence,

$$
\begin{aligned}
I_2 &= \frac{1}{\gamma N^2} \sum_{k=0}^{N-1} \sum_i \left[(1-\gamma\tilde{\lambda}_i^{(q)})^k - (1-\gamma\tilde{\lambda}_i^{(q)})^{N+k} \right] \omega_i^2 \\
&= \frac{1}{\gamma^2 N^2} \sum_i \frac{\omega_i^2}{\tilde{\lambda}_i^{(q)}} \left[1 - (1-\gamma\tilde{\lambda}_i^{(q)})^N \right]^2 \\
&\leq \frac{1}{\gamma^2 N^2} \sum_i \frac{\omega_i^2}{\tilde{\lambda}_i^{(q)}} \min\left\{ 1, \gamma^2 N^2 (\tilde{\lambda}_i^{(q)})^2 \right\} \\
&\leq \frac{1}{\gamma^2 N^2} \cdot \sum_{i \leq k^*} \frac{\omega_i^2}{\tilde{\lambda}_i^{(q)}} + \sum_{i > k^*} \tilde{\lambda}_i^{(q)} \omega_i^2 \\
&= \frac{1}{\gamma^2 N^2} \cdot \|\mathbf{v}_q^*\|^2_{(\mathbf{H}^{(q)}_{f,0:k^*})^{-1}} + \|\mathbf{v}_q^*\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}.
\end{aligned}
$$

□

B.4 Variance Error

Lemma B.7. *Under Assumption 3.1, 3.3, 3.4, 3.5, if there exist $\bar{\epsilon}_p, \bar{\epsilon}_a$ and $\bar{\epsilon}_o$ such that for any $i \in \{p, a, o\}$, quantization $\mathcal{Q}_i$ is $\bar{\epsilon}_i$-multiplicative, then $\boldsymbol{\Sigma}_t$ defined in (8) satisfies*

$$
\begin{aligned}
\boldsymbol{\Sigma}_t \preceq & \frac{(1+2\bar{\epsilon}_o)\bar{\sigma}^2}{B} \mathbf{H}_f^{(q)} + \frac{\bar{\epsilon}_a(1+2\bar{\epsilon}_o)}{B} \mathbb{E}\left[\widetilde{\mathbf{x}}^{(q)} (\widetilde{\mathbf{x}}^{(q)})^\top \mathbf{v}_{t-1}\mathbf{v}_{t-1}^\top \widetilde{\mathbf{x}}^{(q)} (\widetilde{\mathbf{x}}^{(q)})^\top \right] \\
& + \frac{(1+\bar{\epsilon}_a)(1+2\bar{\epsilon}_o)\bar{\epsilon}_p}{B} \mathbb{E}\left[\widetilde{\mathbf{x}}^{(q)} (\widetilde{\mathbf{x}}^{(q)})^\top \operatorname{diag}\left(\mathbf{v}_{t-1}\mathbf{v}_{t-1}^\top \right) \widetilde{\mathbf{x}}^{(q)} (\widetilde{\mathbf{x}}^{(q)})^\top \right] \\
& + \frac{2\bar{\epsilon}_o}{B} \mathbb{E}\left[\widetilde{\mathbf{x}}^{(q)} (\widetilde{\mathbf{x}}^{(q)})^\top \boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top \widetilde{\mathbf{x}}^{(q)} (\widetilde{\mathbf{x}}^{(q)})^\top \right] + \frac{B-1}{B} \bar{\epsilon}_p \mathbf{H}_f^{(q)} \operatorname{diag}\big(\mathbb{E}[\mathbf{v}_{t-1}\mathbf{v}_{t-1}^\top] \big) \mathbf{H}_f^{(q)}.
\end{aligned}
$$

Proof. By Assumption 3.5,

$$
\begin{aligned}
\frac{1}{B^2} \mathbb{E}\left[\widetilde{\mathbf{X}}_t^{(q)} \boldsymbol{\xi}_t \boldsymbol{\xi}_t^\top (\widetilde{\mathbf{X}}_t^{(q)})^\top \right] &= \frac{1}{B^2} \mathbb{E}\left[\sum_{i=1}^B \sum_{j=1}^B (\widetilde{\mathbf{X}}_t^{(q)})_i (\boldsymbol{\xi}_t)_i (\boldsymbol{\xi}_t)_j (\widetilde{\mathbf{X}}_t^{(q)})_j^\top \right] \\
&= \frac{1}{B} \mathbb{E}\left[\xi^2 \widetilde{\mathbf{x}}^{(q)} (\widetilde{\mathbf{x}}^{(q)})^\top \right] \\
&\preceq \frac{\bar{\sigma}^2}{B} \mathbf{H}_f^{(q)},
\end{aligned} \tag{11}
$$

where the second equality holds as samples are independent. Utilizing the coordinate-wise independence in quantization,

$$
\begin{aligned}
& \frac{1}{B^2} \mathbb{E}\left[\widetilde{\mathbf{X}}_t^{(q)} (\widetilde{\mathbf{X}}_t^{(q)})^\top \boldsymbol{\epsilon}_{t-1}^{(p)} (\boldsymbol{\epsilon}_{t-1}^{(p)})^\top \widetilde{\mathbf{X}}_t^{(q)} (\widetilde{\mathbf{X}}_t^{(q)})^\top \right] \\
\preceq & \frac{\bar{\epsilon}_p}{B^2} \mathbb{E}\left[\widetilde{\mathbf{X}}_t^{(q)} (\widetilde{\mathbf{X}}_t^{(q)})^\top \operatorname{diag}\left(\mathbf{v}_{t-1}\mathbf{v}_{t-1}^\top \right) \widetilde{\mathbf{X}}_t^{(q)} (\widetilde{\mathbf{X}}_t^{(q)})^\top \right].
\end{aligned}
$$

For the variance of activation,

$$
\begin{aligned}
& \frac{1}{B^2} \mathbb{E}\left[\widetilde{\mathbf{X}}_t^{(q)} \boldsymbol{\epsilon}_t^{(a)} (\boldsymbol{\epsilon}_t^{(a)})^\top (\widetilde{\mathbf{X}}_t^{(q)})^\top \right] \\
\preceq & \frac{\bar{\epsilon}_a}{B^2} \mathbb{E}\left[\widetilde{\mathbf{X}}_t^{(q)} \operatorname{diag}\left((\widetilde{\mathbf{X}}_t^{(q)})^\top \mathcal{Q}_p(\mathbf{v}_{t-1}) \mathcal{Q}_p(\mathbf{v}_{t-1})^\top \widetilde{\mathbf{X}}_t^{(q)} \right) (\widetilde{\mathbf{X}}_t^{(q)})^\top \right] \\
= & \frac{\bar{\epsilon}_a}{B^2} \sum_{i=1}^B \mathbb{E}\left[(\widetilde{\mathbf{X}}_t^{(q)})_i (\widetilde{\mathbf{X}}_t^{(q)})_i^\top \mathcal{Q}_p(\mathbf{v}_{t-1}) \mathcal{Q}_p(\mathbf{v}_{t-1})^\top (\widetilde{\mathbf{X}}_t^{(q)})_i (\widetilde{\mathbf{X}}_t^{(q)})_i^\top \right] \\
= & \frac{\bar{\epsilon}_a}{B} \mathbb{E}\left[\widetilde{\mathbf{x}}^{(q)} (\widetilde{\mathbf{x}}^{(q)})^\top \mathcal{Q}_p(\mathbf{v}_{t-1}) \mathcal{Q}_p(\mathbf{v}_{t-1})^\top \widetilde{\mathbf{x}}^{(q)} (\widetilde{\mathbf{x}}^{(q)})^\top \right] \\
\preceq & \frac{\bar{\epsilon}_a}{B} \mathbb{E}\left[\widetilde{\mathbf{x}}^{(q)} (\widetilde{\mathbf{x}}^{(q)})^\top \left(\mathbf{v}_{t-1}\mathbf{v}_{t-1}^\top + \bar{\epsilon}_p \operatorname{diag}\left(\mathbf{v}_{t-1}\mathbf{v}_{t-1}^\top \right) \right) \widetilde{\mathbf{x}}^{(q)} (\widetilde{\mathbf{x}}^{(q)})^\top \right].
\end{aligned}
$$

For the variance of output gradient,

$$\begin{aligned}
&\frac{1}{B^2}\mathbb{E}\left[\widetilde{\mathbf{X}}_t^{(q)}\boldsymbol{\epsilon}_t^{(o)}(\boldsymbol{\epsilon}_t^{(o)})^\top(\widetilde{\mathbf{X}}_t^{(q)})^\top\right]\\
\preceq&\frac{\overline{\epsilon}_o}{B^2}\mathbb{E}\left[\widetilde{\mathbf{X}}_t^{(q)}\mathrm{diag}\left(\left[\mathcal{Q}_l(\mathbf{y}_t)-\mathcal{Q}_a\left((\widetilde{\mathbf{X}}_t^{(q)})^\top\mathcal{Q}_p(\mathbf{v}_{t-1})\right)\right]\left[\mathcal{Q}_l(\mathbf{y}_t)-\mathcal{Q}_a\left((\widetilde{\mathbf{X}}_t^{(q)})^\top\mathcal{Q}_p(\mathbf{v}_{t-1})\right)\right]^\top\right)(\widetilde{\mathbf{X}}_t^{(q)})^\top\right]\\
=&\frac{\overline{\epsilon}_o}{B^2}\mathbb{E}\left[\widetilde{\mathbf{X}}_t^{(q)}\mathrm{diag}\left(\left[\boldsymbol{\xi}_t-\boldsymbol{\epsilon}_t^{(a)}-(\widetilde{\mathbf{X}}_t^{(q)})^\top\boldsymbol{\epsilon}_{t-1}^{(p)}-(\widetilde{\mathbf{X}}_t^{(q)})^\top\boldsymbol{\eta}_{t-1}\right]\left[\boldsymbol{\xi}_t-\boldsymbol{\epsilon}_t^{(a)}-(\widetilde{\mathbf{X}}_t^{(q)})^\top\boldsymbol{\epsilon}_{t-1}^{(p)}-(\widetilde{\mathbf{X}}_t^{(q)})^\top\boldsymbol{\eta}_{t-1}\right]^\top\right)(\widetilde{\mathbf{X}}_t^{(q)})^\top\right]\\
\preceq&\frac{2\overline{\epsilon}_o}{B^2}\mathbb{E}\left[\widetilde{\mathbf{X}}_t^{(q)}\mathrm{diag}\left(\boldsymbol{\xi}_t\boldsymbol{\xi}_t^\top+\boldsymbol{\epsilon}_t^{(a)}(\boldsymbol{\epsilon}_t^{(a)})^\top+(\widetilde{\mathbf{X}}_t^{(q)})^\top\boldsymbol{\epsilon}_{t-1}^{(p)}(\boldsymbol{\epsilon}_{t-1}^{(p)})^\top\widetilde{\mathbf{X}}_t^{(q)}+(\widetilde{\mathbf{X}}_t^{(q)})^\top\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top\widetilde{\mathbf{X}}_t^{(q)}\right)(\widetilde{\mathbf{X}}_t^{(q)})^\top\right].
\end{aligned}$$

Notice that

$$\mathbb{E}\left[\widetilde{\mathbf{X}}_t^{(q)}\mathrm{diag}\left(\boldsymbol{\xi}_t\boldsymbol{\xi}_t^\top\right)(\widetilde{\mathbf{X}}_t^{(q)})^\top\right]=\sum_{i=1}^{B}\mathbb{E}\left[(\widetilde{\mathbf{X}}_t^{(q)})_i(\boldsymbol{\xi}_t)_i^2(\widetilde{\mathbf{X}}_t^{(q)})_i^\top\right]=B\mathbb{E}\left[\xi^2\widetilde{\mathbf{x}}^{(q)}(\widetilde{\mathbf{x}}^{(q)})^\top\right]\preceq B\overline{\sigma}^2\mathbf{H}_f^{(q)},$$

$$\begin{aligned}
&\mathbb{E}\left[\widetilde{\mathbf{X}}_t^{(q)}\mathrm{diag}\left((\widetilde{\mathbf{X}}_t^{(q)})^\top\boldsymbol{\epsilon}_{t-1}^{(p)}(\boldsymbol{\epsilon}_{t-1}^{(p)})^\top\widetilde{\mathbf{X}}_t^{(q)}\right)(\widetilde{\mathbf{X}}_t^{(q)})^\top\right]\\
=&\sum_{i=1}^{B}\mathbb{E}\left[(\widetilde{\mathbf{X}}_t^{(q)})_i(\widetilde{\mathbf{X}}_t^{(q)})_i^\top\boldsymbol{\epsilon}_{t-1}^{(p)}(\boldsymbol{\epsilon}_{t-1}^{(p)})^\top(\widetilde{\mathbf{X}}_t^{(q)})_i(\widetilde{\mathbf{X}}_t^{(q)})_i^\top\right]\\
=&B\mathbb{E}\left[\widetilde{\mathbf{x}}^{(q)}(\widetilde{\mathbf{x}}^{(q)})^\top\boldsymbol{\epsilon}_{t-1}^{(p)}(\boldsymbol{\epsilon}_{t-1}^{(p)})^\top\widetilde{\mathbf{x}}^{(q)}(\widetilde{\mathbf{x}}^{(q)})^\top\right]\\
\preceq&\overline{\epsilon}_pB\mathbb{E}\left[\widetilde{\mathbf{x}}^{(q)}(\widetilde{\mathbf{x}}^{(q)})^\top\mathrm{diag}\left(\mathbf{v}_{t-1}\mathbf{v}_{t-1}^\top\right)\widetilde{\mathbf{x}}^{(q)}(\widetilde{\mathbf{x}}^{(q)})^\top\right],
\end{aligned}$$

and similarly

$$\mathbb{E}\left[\widetilde{\mathbf{X}}_t^{(q)}\mathrm{diag}\left((\widetilde{\mathbf{X}}_t^{(q)})^\top\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top\widetilde{\mathbf{X}}_t^{(q)}\right)(\widetilde{\mathbf{X}}_t^{(q)})^\top\right]=B\mathbb{E}\left[\widetilde{\mathbf{x}}^{(q)}(\widetilde{\mathbf{x}}^{(q)})^\top\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top\widetilde{\mathbf{x}}^{(q)}(\widetilde{\mathbf{x}}^{(q)})^\top\right],$$

therefore,

$$\begin{aligned}
\boldsymbol{\Sigma}_t\preceq&\frac{(1+2\overline{\epsilon}_o)\overline{\sigma}^2}{B}\mathbf{H}_f^{(q)}+\frac{\overline{\epsilon}_a(1+2\overline{\epsilon}_o)}{B}\mathbb{E}\left[\widetilde{\mathbf{x}}^{(q)}(\widetilde{\mathbf{x}}^{(q)})^\top\mathbf{v}_{t-1}\mathbf{v}_{t-1}^\top\widetilde{\mathbf{x}}^{(q)}(\widetilde{\mathbf{x}}^{(q)})^\top\right]\\
&+\frac{(1+\overline{\epsilon}_a)(1+2\overline{\epsilon}_o)\overline{\epsilon}_p}{B}\mathbb{E}\left[\widetilde{\mathbf{x}}^{(q)}(\widetilde{\mathbf{x}}^{(q)})^\top\mathrm{diag}\left(\mathbf{v}_{t-1}\mathbf{v}_{t-1}^\top\right)\widetilde{\mathbf{x}}^{(q)}(\widetilde{\mathbf{x}}^{(q)})^\top\right]\\
&+\frac{2\overline{\epsilon}_o}{B}\mathbb{E}\left[\widetilde{\mathbf{x}}^{(q)}(\widetilde{\mathbf{x}}^{(q)})^\top\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top\widetilde{\mathbf{x}}^{(q)}(\widetilde{\mathbf{x}}^{(q)})^\top\right]+\frac{B-1}{B}\overline{\epsilon}_p\mathbf{H}_f^{(q)}\,\mathrm{diag}\big(\mathbb{E}[\mathbf{v}_{t-1}\mathbf{v}_{t-1}^\top]\big)\,\mathbf{H}_f^{(q)}.
\end{aligned}$$

□

Lemma B.8. *Under Assumption 3.3, 3.4, if* $\gamma<\frac{1}{\|\mathbf{H}_f^{(q)}\|}\wedge\frac{B}{\alpha\mathrm{tr}(\mathbf{H}_f^{(q)})}$,

$$\mathbf{B}_t\preceq(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t\mathbf{B}_0(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t+\gamma\frac{\frac{\alpha}{B}\mathrm{tr}\left(\mathbf{H}_f^{(q)}\mathbf{v}_q^*(\mathbf{v}_q^*)^\top\right)}{1-\gamma\frac{\alpha}{B}\mathrm{tr}(\mathbf{H}_f^{(q)})}\mathbf{I}.$$

Proof. Using

$$\mathcal{T}_B^{(q)}=\widetilde{\mathcal{T}}^{(q)}-\gamma(\mathcal{M}_B^{(q)}-\widetilde{\mathcal{M}}^{(q)}),$$

it holds by Lemma B.1 that

$$\begin{aligned}
\mathbf{B}_t\preceq&(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})\mathbf{B}_{t-1}(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})+\gamma^2(\mathcal{M}_B^{(q)}-\widetilde{\mathcal{M}}^{(q)})\circ\mathbf{B}_{t-1}\\
\preceq&(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})\mathbf{B}_{t-1}(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})+\gamma^2\frac{\alpha}{B}\mathrm{tr}(\mathbf{H}_f^{(q)}\mathbf{B}_{t-1})\mathbf{H}_f^{(q)}.
\end{aligned}$$

The iteration yields

$$\mathbf{B}_t\preceq(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t\mathbf{B}_0(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t+\gamma^2\frac{\alpha}{B}\sum_{s=0}^{t-1}\mathrm{tr}(\mathbf{H}_f^{(q)}\mathbf{B}_s)(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{t-1-s}\mathbf{H}_f^{(q)}(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{t-1-s}.$$

We then focus on bounding $\mathbf{E}_t := \gamma\frac{\alpha}{B}\sum_{s=0}^{t-1}\operatorname{tr}(\mathbf{H}_f^{(q)}\mathbf{B}_s)(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{t-1-s}\mathbf{H}_f^{(q)}(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{t-1-s}$ with $\mathbf{E}_0=\mathbf{0}$. This definition directly gives

$$\mathbf{E}_{t+1}=(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})\mathbf{E}_t(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})+\gamma\frac{\alpha}{B}\operatorname{tr}(\mathbf{H}_f^{(q)}\mathbf{B}_t)\mathbf{H}_f^{(q)}.$$

We proceed to prove $\mathbf{E}_t \preceq \frac{\frac{\alpha}{B}\operatorname{tr}\left(\mathbf{H}_f^{(q)}\mathbf{v}_q^*(\mathbf{v}_q^*)^\top\right)}{1-\gamma\frac{\alpha}{B}\operatorname{tr}(\mathbf{H}_f^{(q)})}\mathbf{I}$ by induction. The claim is immediate for $t=0$. If it holds at time t, then

$$\begin{aligned}\operatorname{tr}(\mathbf{H}_f^{(q)}\mathbf{B}_t)\leq&\operatorname{tr}\left(\mathbf{H}_f^{(q)}(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t\mathbf{B}_0(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t\right)+\gamma\frac{\frac{\alpha}{B}\operatorname{tr}\left(\mathbf{H}_f^{(q)}\mathbf{v}_q^*(\mathbf{v}_q^*)^\top\right)}{1-\gamma\frac{\alpha}{B}\operatorname{tr}(\mathbf{H}_f^{(q)})}\operatorname{tr}(\mathbf{H}_f^{(q)})\\ \leq&\operatorname{tr}(\mathbf{H}_f^{(q)}\mathbf{B}_0)+\gamma\frac{\frac{\alpha}{B}\operatorname{tr}\left(\mathbf{H}_f^{(q)}\mathbf{v}_q^*(\mathbf{v}_q^*)^\top\right)}{1-\gamma\frac{\alpha}{B}\operatorname{tr}(\mathbf{H}_f^{(q)})}\operatorname{tr}(\mathbf{H}_f^{(q)}).\end{aligned}$$

It follows that

$$\begin{aligned}\mathbf{E}_{t+1}\preceq&(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^2\frac{\frac{\alpha}{B}\operatorname{tr}\left(\mathbf{H}_f^{(q)}\mathbf{v}_q^*(\mathbf{v}_q^*)^\top\right)}{1-\gamma\frac{\alpha}{B}\operatorname{tr}(\mathbf{H}_f^{(q)})}+\frac{\gamma\alpha}{B}\operatorname{tr}\left(\mathbf{H}_f^{(q)}\mathbf{v}_q^*(\mathbf{v}_q^*)^\top\right)\mathbf{H}_f^{(q)}+\frac{\gamma^2\alpha}{B}\frac{\frac{\alpha}{B}\operatorname{tr}\left(\mathbf{H}_f^{(q)}\mathbf{v}_q^*(\mathbf{v}_q^*)^\top\right)}{1-\gamma\frac{\alpha}{B}\operatorname{tr}(\mathbf{H}_f^{(q)})}\operatorname{tr}(\mathbf{H}_f^{(q)})\mathbf{H}_f^{(q)}\\ =&\left[\gamma\mathbf{H}_f^{(q)}+(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^2\right]\frac{\frac{\alpha}{B}\operatorname{tr}\left(\mathbf{H}_f^{(q)}\mathbf{v}_q^*(\mathbf{v}_q^*)^\top\right)}{1-\gamma\frac{\alpha}{B}\operatorname{tr}(\mathbf{H}_f^{(q)})}\\ \preceq&\frac{\frac{\alpha}{B}\operatorname{tr}\left(\mathbf{H}_f^{(q)}\mathbf{v}_q^*(\mathbf{v}_q^*)^\top\right)}{1-\gamma\frac{\alpha}{B}\operatorname{tr}(\mathbf{H}_f^{(q)})}\mathbf{I},\end{aligned}$$

where the last inequality holds by $\gamma<\frac{1}{\|\mathbf{H}_f^{(q)}\|}$. □

Lemma B.9. *Under Assumptions 3.1, 3.3, 3.4, and 3.5, suppose that $\mathcal{Q}_i$ is $\bar{\epsilon}_i$-multiplicative for $i\in\{p,a,o\}$ and $\gamma<\frac{1}{\|\mathbf{H}_f^{(q)}\|}\wedge\frac{B}{\alpha\operatorname{tr}(\mathbf{H}_f^{(q)})}$. Then, for every $t\geq 0$,*

$$\mathbb{E}\left[\mathbf{v}_t\mathbf{v}_t^\top\right]\preceq\left[\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t\right]\mathbf{v}_q^*\mathbf{v}_q^{*\top}\left[\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t\right]+2\mathbf{C}_t+\frac{2\gamma\alpha\|\mathbf{v}_q^*\|^2_{\mathbf{H}_f^{(q)}}}{B[1-\gamma\alpha\operatorname{tr}(\mathbf{H}_f^{(q)})/B]}\mathbf{I}.$$

Moreover,

$$\max\left\{\operatorname{tr}\left(\mathbf{H}_f^{(q)}\mathbb{E}[\mathbf{v}_t\mathbf{v}_t^\top]\right),\operatorname{tr}\left(\mathbf{H}_f^{(q)}\mathbb{E}[\boldsymbol{\eta}_t\boldsymbol{\eta}_t^\top]\right)\right\}\leq\frac{2\|\mathbf{v}_q^*\|^2_{\mathbf{H}_f^{(q)}}}{1-\gamma\alpha\operatorname{tr}(\mathbf{H}_f^{(q)})/B}+2\operatorname{tr}(\mathbf{H}_f^{(q)}\mathbf{C}_t).$$

Proof. The first-moment recursion and the original bias–variance decomposition imply

$$\mathbb{E}[\boldsymbol{\eta}_t^{\text{bias}}]=-(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t\mathbf{v}_q^*,\qquad\mathbb{E}[\boldsymbol{\eta}_t^{\text{var}}]=0.$$

Consequently,

$$\mathbb{E}[\mathbf{v}_t]=[\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t]\mathbf{v}_q^*.$$

It follows that

$$\begin{aligned}\mathbb{E}\left[(\mathbf{v}_t-\mathbb{E}\mathbf{v}_t)(\mathbf{v}_t-\mathbb{E}\mathbf{v}_t)^\top\right]=&\mathbb{E}\left[\left(\boldsymbol{\eta}_t^{\text{bias}}+\boldsymbol{\eta}_t^{\text{var}}+(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t\mathbf{v}_q^*\right)\left(\boldsymbol{\eta}_t^{\text{bias}}+\boldsymbol{\eta}_t^{\text{var}}+(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t\mathbf{v}_q^*\right)^\top\right]\\ \preceq&2\left[\mathbf{B}_t-(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t\mathbf{B}_0(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t\right]+2\mathbf{C}_t.\end{aligned}$$

Applying Lemma B.8 and adding $\mathbb{E}[\mathbf{v}_t]\mathbb{E}[\mathbf{v}_t]^\top$ proves the first assertion.

The centered second moments of $\mathbf{v}_t$ and $\boldsymbol{\eta}_t$ coincide. Also,

$$\left\|[\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t]\mathbf{v}_q^*\right\|_{\mathbf{H}_f^{(q)}}^2 \le \|\mathbf{v}_q^*\|_{\mathbf{H}_f^{(q)}}^2,$$
$$\left\|(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t\mathbf{v}_q^*\right\|_{\mathbf{H}_f^{(q)}}^2 \le \|\mathbf{v}_q^*\|_{\mathbf{H}_f^{(q)}}^2.$$

Taking the $\mathbf{H}_f^{(q)}$-weighted trace of the covariance bound therefore bounds either second moment by

$$\left(1+\frac{2\gamma\alpha\,\mathrm{tr}(\mathbf{H}_f^{(q)})/B}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}_f^{(q)})/B}\right)\|\mathbf{v}_q^*\|_{\mathbf{H}_f^{(q)}}^2+2\,\mathrm{tr}(\mathbf{H}_f^{(q)}\mathbf{C}_t).$$

This implies the second assertion. □

We derive upper bounds for $\mathrm{var}=\frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{k-t}\mathbf{H}_f^{(q)},\mathbf{C}_t\right\rangle$. Denote

$$d_{\mathrm{eff}}=k^*+\gamma^2N^2\sum_{i>k^*}(\tilde{\lambda}_i^{(q)})^2. \tag{12}$$

Theorem B.10. *Under Assumptions 3.1, 3.3, 3.4, and 3.5, suppose that $\mathcal{Q}_i$ is $\bar{\epsilon}_i$-multiplicative for $i\in\{p,a,o\}$ and $\gamma\lesssim\frac{1}{\alpha\,\mathrm{tr}(\mathbf{H}_f^{(q)})}$,*

$$\begin{aligned}\mathrm{var}\lesssim&\frac{d_{\mathrm{eff}}}{BN}\left[(1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p)\overline{\sigma}^2+(1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p)\|\mathbf{v}_q^*\|_{\mathbf{H}_f^{(q)}}^2\right]\\&+\frac{\bar{\epsilon}_p}{N^2}\left(\frac{d_{\mathrm{eff}}}{B}+\min\{1,d_{\mathrm{eff}}\}\right)\sum_{t=0}^{N-1}\left\|[\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t]\mathbf{v}_q^*\right\|_{\mathrm{diag}(\mathbf{H}_f^{(q)})}^2.\end{aligned}$$

Proof. By Lemma B.9, Lemma B.7 and Lemma B.1,

$$\begin{aligned}&\boldsymbol{\Sigma}_{t+1}+(\mathcal{M}_B^{(q)}-\widetilde{\mathcal{M}}^{(q)})\circ\mathbf{C}_t\\\precsim&\frac{1}{B}\Bigg[(1+\bar{\epsilon}_o)\overline{\sigma}^2+(1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p)\left(\|\mathbf{v}_q^*\|_{\mathbf{H}_f^{(q)}}^2+\mathrm{tr}(\mathbf{H}_f^{(q)}\mathbf{C}_t)\right)\\&+\bar{\epsilon}_p\left(\left\|[\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t]\mathbf{v}_q^*\right\|_{\mathrm{diag}(\mathbf{H}_f^{(q)})}^2+\mathrm{tr}(\mathrm{diag}(\mathbf{H}_f^{(q)})\mathbf{C}_t)\right)\Bigg]\mathbf{H}_f^{(q)}\\&+\bar{\epsilon}_p\mathbf{H}_f^{(q)}\,\mathrm{diag}\left([\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t]\mathbf{B}_0[\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t]+\mathbf{C}_t+\frac{\gamma\|\mathbf{v}_q^*\|_{\mathbf{H}_f^{(q)}}^2}{B}\mathbf{I}\right)\mathbf{H}_f^{(q)}.\end{aligned} \tag{13}$$

By definition,

$$\begin{aligned}\mathbf{C}_t=&(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})\mathbf{C}_{t-1}(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})+\gamma^2\left[\boldsymbol{\Sigma}_t+(\mathcal{M}_B^{(q)}-\widetilde{\mathcal{M}}^{(q)})\circ\mathbf{C}_{t-1}\right]\\=&\gamma^2\sum_{s=0}^{t-1}(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{t-1-s}\left[\boldsymbol{\Sigma}_{s+1}+(\mathcal{M}_B^{(q)}-\widetilde{\mathcal{M}}^{(q)})\circ\mathbf{C}_s\right](\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{t-1-s}.\end{aligned}$$

It follows that

$$\begin{aligned}&\sum_{t=0}^{N-1}\mathrm{tr}\Big([\mathbf{H}_f^{(q)}+\mathrm{diag}(\mathbf{H}_f^{(q)})]\mathbf{C}_t\Big)\\=&\gamma^2\sum_{s=0}^{N-2}\sum_{j=0}^{N-2-s}\mathrm{tr}\left([\mathbf{H}_f^{(q)}+\mathrm{diag}(\mathbf{H}_f^{(q)})](\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^j\left[\boldsymbol{\Sigma}_{s+1}+(\mathcal{M}_B^{(q)}-\widetilde{\mathcal{M}}^{(q)})\circ\mathbf{C}_s\right](\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^j\right).\end{aligned}$$

Noticing that

$$\gamma\sum_{j=0}^{\infty}(\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^{2j}\mathbf{H}_f^{(q)}\preceq\mathbf{I},$$

it holds

$$\gamma^2 \sum_{j=0}^{\infty} \operatorname{tr}\Big([\mathbf{H}_f^{(q)} + \operatorname{diag}(\mathbf{H}_f^{(q)})](\mathbf{I} - \gamma \mathbf{H}_f^{(q)})^j \mathbf{H}_f^{(q)} (\mathbf{I} - \gamma \mathbf{H}_f^{(q)})^j\Big) \le 2\gamma \operatorname{tr}(\mathbf{H}_f^{(q)}).$$

Moreover, noticing that

$$\gamma \sum_{j=0}^{\infty} (\mathbf{I} - \gamma \mathbf{H}_f^{(q)})^{2j} (\mathbf{H}_f^{(q)})^2 \preceq \mathbf{H}_f^{(q)},$$

we have for every PSD matrix $\mathbf{X}$,

$$\gamma^2 \sum_{j=0}^{\infty} \operatorname{tr}\Big([\mathbf{H}_f^{(q)} + \operatorname{diag}(\mathbf{H}_f^{(q)})](\mathbf{I} - \gamma \mathbf{H}_f^{(q)})^j \mathbf{H}_f^{(q)} \operatorname{diag}(\mathbf{X}) \mathbf{H}_f^{(q)} (\mathbf{I} - \gamma \mathbf{H}_f^{(q)})^j\Big) \le 2\gamma \operatorname{tr}(\mathbf{H}_f^{(q)}) \operatorname{tr}(\operatorname{diag}(\mathbf{H}_f^{(q)})\mathbf{X}).$$

Therefore,

$$\begin{aligned}
&\sum_{t=0}^{N-1} \operatorname{tr}\Big([\mathbf{H}_f^{(q)} + \operatorname{diag}(\mathbf{H}_f^{(q)})]\mathbf{C}_t\Big) \\
\lesssim &\gamma \operatorname{tr}(\mathbf{H}_f^{(q)}) \Bigg\{ \frac{N}{B} \left[(1+\bar{\epsilon}_o)\bar{\sigma}^2 + (1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p) \|\mathbf{v}_q^*\|^2_{\mathbf{H}_f^{(q)}} \right] \\
&+ \bar{\epsilon}_p \sum_{t=0}^{N-1} \left\| [\mathbf{I} - (\mathbf{I} - \gamma \mathbf{H}_f^{(q)})^t] \mathbf{v}_q^* \right\|^2_{\operatorname{diag}(\mathbf{H}_f^{(q)})} \\
&+ \frac{1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p}{B} \sum_{t=0}^{N-1} \operatorname{tr}(\mathbf{H}_f^{(q)} \mathbf{C}_t) + \bar{\epsilon}_p \sum_{t=0}^{N-1} \operatorname{tr}(\operatorname{diag}(\mathbf{H}_f^{(q)}) \mathbf{C}_t) \Bigg\}.
\end{aligned}$$

Since $\gamma \operatorname{tr}(\mathbf{H}_f^{(q)}) \lesssim 1$, the sufficiently small step-size constant absorbs both feedback sums into the left-hand side. We have therefore proved

$$\begin{aligned}
&\sum_{t=0}^{N-1} \operatorname{tr}\Big([\mathbf{H}_f^{(q)} + \operatorname{diag}(\mathbf{H}_f^{(q)})]\mathbf{C}_t\Big) \\
\lesssim &\gamma \operatorname{tr}(\mathbf{H}_f^{(q)}) \left\{ \frac{N}{B} \left[(1+\bar{\epsilon}_o)\bar{\sigma}^2 + (1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p) \|\mathbf{v}_q^*\|^2_{\mathbf{H}_f^{(q)}} \right] + \bar{\epsilon}_p \sum_{t=0}^{N-1} \left\| [\mathbf{I} - (\mathbf{I} - \gamma \mathbf{H}_f^{(q)})^t] \mathbf{v}_q^* \right\|^2_{\operatorname{diag}(\mathbf{H}_f^{(q)})} \right\}.
\end{aligned} \tag{14}$$

Next, we turn to bound the variance error $\mathrm{var} = \frac{1}{\gamma N^2} \sum_{t=0}^{N-1} \left\langle \mathbf{I} - (\mathbf{I} - \gamma \mathbf{H}_f^{(q)})^{N-t}, \mathbf{C}_t \right\rangle$ by using $\mathbf{C}_t = \gamma^2 \sum_{s=0}^{t-1} (\mathbf{I} - \gamma \mathbf{H}_f^{(q)})^{t-1-s} \left[\mathbf{\Sigma}_{s+1} + (\mathcal{M}_B^{(q)} - \widetilde{\mathcal{M}}^{(q)}) \circ \mathbf{C}_s \right] (\mathbf{I} - \gamma \mathbf{H}_f^{(q)})^{t-1-s}$ again. It follows that

$$\begin{aligned}
\mathrm{var} =& \frac{\gamma}{N^2} \sum_{t=1}^{N-1} \sum_{s=0}^{t-1} \operatorname{tr}\left((\mathbf{I} - \gamma \mathbf{H}_f^{(q)})^{2(t-1-s)} [\mathbf{I} - (\mathbf{I} - \gamma \mathbf{H}_f^{(q)})^{N-t}] \left[\mathbf{\Sigma}_{s+1} + (\mathcal{M}_B^{(q)} - \widetilde{\mathcal{M}}^{(q)}) \circ \mathbf{C}_s \right] \right) \\
=& \sum_{s=0}^{N-2} \left\langle \frac{\gamma}{N^2} \sum_{j=0}^{N-2-s} (\mathbf{I} - \gamma \mathbf{H}_f^{(q)})^{2j} [\mathbf{I} - (\mathbf{I} - \gamma \mathbf{H}_f^{(q)})^{N-1-s-j}], \mathbf{\Sigma}_{s+1} + (\mathcal{M}_B^{(q)} - \widetilde{\mathcal{M}}^{(q)}) \circ \mathbf{C}_s \right\rangle.
\end{aligned}$$

Further note that

$$0 \le \gamma \tilde{\lambda}_i^{(q)} \sum_{j=0}^{N-2-s} (1 - \gamma \tilde{\lambda}_i^{(q)})^{2j} [1 - (1 - \gamma \tilde{\lambda}_i^{(q)})^{N-1-s-j}] \le \min\{1, (N\gamma \tilde{\lambda}_i^{(q)})^2\},$$

substitute into (13) gives

$$\begin{aligned}\mathrm{var} \lesssim\ & \frac{d_{\mathrm{eff}}}{BN}\left[(1+\bar{\epsilon}_o)\bar{\sigma}^2 + (1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p)\|\mathbf{v}_q^*\|^2_{\mathbf{H}_f^{(q)}}\right] \\ & + \frac{\bar{\epsilon}_p}{N^2}\left(\frac{d_{\mathrm{eff}}}{B} + \min\{1, d_{\mathrm{eff}}\}\right)\sum_{t=0}^{N-1}\left\|[\mathbf{I} - (\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t]\mathbf{v}_q^*\right\|^2_{\mathrm{diag}(\mathbf{H}_f^{(q)})} \\ & + \frac{1}{N^2}\left[\frac{d_{\mathrm{eff}}}{B}(1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p) + \bar{\epsilon}_p\min\{1,d_{\mathrm{eff}}\}\right]\cdot\sum_{t=0}^{N-1}\mathrm{tr}\Big([\mathbf{H}_f^{(q)} + \mathrm{diag}(\mathbf{H}_f^{(q)})]\mathbf{C}_t\Big) \\ & + \frac{\gamma\bar{\epsilon}_p\,\mathrm{tr}(\mathbf{H}_f^{(q)})\min\{1,d_{\mathrm{eff}}\}}{BN}\|\mathbf{v}_q^*\|^2_{\mathbf{H}_f^{(q)}}.\end{aligned}$$

Applying (14) gives

$$\begin{aligned}\mathrm{var} \lesssim & \frac{d_{\mathrm{eff}}}{BN}\left[(1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p)\bar{\sigma}^2 + (1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p)\|\mathbf{v}_q^*\|^2_{\mathbf{H}_f^{(q)}}\right] \\ & + \frac{\bar{\epsilon}_p}{N^2}\left(\frac{d_{\mathrm{eff}}}{B} + \min\{1, d_{\mathrm{eff}}\}\right)\sum_{t=0}^{N-1}\left\|[\mathbf{I} - (\mathbf{I}-\gamma\mathbf{H}_f^{(q)})^t]\mathbf{v}_q^*\right\|^2_{\mathrm{diag}(\mathbf{H}_f^{(q)})}.\end{aligned}$$

□

B.5 Power-Law Spectrum

Lemma B.11. *Under Assumption 3.1, 3.3, 3.6, if there exist $\bar{\epsilon}_f, \bar{\epsilon}_s$ and $\bar{\epsilon}_d$ such that for any $i \in \{d, f, s\}$, quantization $\mathcal{Q}_i$ is $\bar{\epsilon}_i$-multiplicative, then with probability at least $1 - e^{-\Omega(M)}$, for $1 \le j \le M$,*

$$\tilde{\lambda}_j^{(q)} \lesssim j^{-a} + \frac{\bar{\epsilon}_d + (1+\bar{\epsilon}_d)(\bar{\epsilon}_s\bar{\epsilon}_f + \bar{\epsilon}_s + \bar{\epsilon}_f)}{j}.$$

Proof. By the definition of multiplicative quantization,

$$\begin{aligned}\mathbf{H}_f^{(q)} =& \mathbb{E}\left[\mathcal{Q}_f\left(\mathcal{Q}_s(\mathbf{S})\mathcal{Q}_d(\mathbf{x})\right)\mathcal{Q}_f\left(\mathcal{Q}_s(\mathbf{S})\mathcal{Q}_d(\mathbf{x})\right)^\top\right] \\ \preceq& \mathbb{E}\left[\mathcal{Q}_s(\mathbf{S})\mathcal{Q}_d(\mathbf{x})\mathcal{Q}_d(\mathbf{x})^\top\mathcal{Q}_s(\mathbf{S})^\top\right] + \bar{\epsilon}_f\mathrm{diag}\left(\mathbb{E}\left[\mathcal{Q}_s(\mathbf{S})\mathcal{Q}_d(\mathbf{x})\mathcal{Q}_d(\mathbf{x})^\top\mathcal{Q}_s(\mathbf{S})^\top\right]\right) \\ \preceq& \mathbb{E}\left[\mathcal{Q}_s(\mathbf{S})\left(\mathbf{H} + \bar{\epsilon}_d\mathrm{diag}(\mathbf{H})\right)\mathcal{Q}_s(\mathbf{S})^\top\right] + \bar{\epsilon}_f\mathrm{diag}\left(\mathbb{E}\left[\mathcal{Q}_s(\mathbf{S})\left(\mathbf{H} + \bar{\epsilon}_d\mathrm{diag}(\mathbf{H})\right)\mathcal{Q}_s(\mathbf{S})^\top\right]\right) \\ \preceq& \mathbf{S}\left(\mathbf{H} + \bar{\epsilon}_d\mathrm{diag}(\mathbf{H})\right)\mathbf{S}^\top + \bar{\epsilon}_f\mathrm{diag}\left(\mathbf{S}\left(\mathbf{H} + \bar{\epsilon}_d\mathrm{diag}(\mathbf{H})\right)\mathbf{S}^\top\right) \\ & + (1+\bar{\epsilon}_f)\bar{\epsilon}_s\mathrm{diag}\left(\mathbf{S}\mathrm{diag}\left(\mathbf{H} + \bar{\epsilon}_d\mathrm{diag}(\mathbf{H})\right)\mathbf{S}^\top\right).\end{aligned}$$

Denote $\mathbf{E}_q := \mathbf{H}_f^{(q)} - \mathbf{S}\mathbf{H}\mathbf{S}^\top$, it follows that

$$\mathbf{E}_q \preceq \bar{\epsilon}_d\mathbf{S}\mathrm{diag}(\mathbf{H})\mathbf{S}^\top + \bar{\epsilon}_f\mathrm{diag}\left(\mathbf{S}\left(\mathbf{H} + \bar{\epsilon}_d\mathrm{diag}(\mathbf{H})\right)\mathbf{S}^\top\right) + (1+\bar{\epsilon}_f)\bar{\epsilon}_s\mathrm{diag}\left(\mathbf{S}\mathrm{diag}\left(\mathbf{H} + \bar{\epsilon}_d\mathrm{diag}(\mathbf{H})\right)\mathbf{S}^\top\right).$$

For a trace-class PSD operator $\mathbf{R}$, the standard Gaussian sketch concentration bounds give

$$\mathrm{tr}(\mathbf{S}\mathbf{R}\mathbf{S}^\top) \lesssim \mathrm{tr}(\mathbf{R})$$

with probability at least $1 - e^{-\Omega(M)}$. Hence, together with $\mathrm{tr}(\mathbf{H}) = \mathrm{tr}(\mathrm{diag}(\mathbf{H}))$, it holds

$$\mathrm{tr}(\mathbf{E}_q) \lesssim [\bar{\epsilon}_d + (1+\bar{\epsilon}_d)(\bar{\epsilon}_s\bar{\epsilon}_f + \bar{\epsilon}_s + \bar{\epsilon}_f)]\mathrm{tr}(\mathbf{H}),$$

with probability at least $1 - e^{-\Omega(M)}$.

Under Assumption 3.6, Weyl's inequality thus gives, uniformly over $1 \le j \le M$,

$$\tilde{\lambda}_j^{(q)} = \lambda_j(\mathbf{S}\mathbf{H}\mathbf{S}^\top + \mathbf{E}_q) \lesssim j^{-a} + \frac{\bar{\epsilon}_d + (1+\bar{\epsilon}_d)(\bar{\epsilon}_s\bar{\epsilon}_f + \bar{\epsilon}_s + \bar{\epsilon}_f)}{j}.$$

□

Lemma B.12. *Under Assumptions 3.1, 3.3, 3.4, 3.5 and 3.6, suppose that $\mathcal{Q}_i$ is $\bar{\epsilon}_i$-multiplicative for $i \in \{p, a, o, d, s, f\}$ and $\gamma \lesssim \frac{1}{\alpha \operatorname{tr}(\mathbf{H}_f^{(q)})}$, $\overline{\sigma}^2 \lesssim 1$, with probability at least $1 - e^{-\Omega(M)}$,*

$$\mathbb{E}_{\mathbf{w}^*}[\mathrm{var}] \lesssim \frac{d_{\mathrm{eff}}}{BN}\Big[1 + \bar{\epsilon}_a + \bar{\epsilon}_o + \bar{\epsilon}_p(1+\epsilon_q M^a)\min\{M, \max\{1, N\gamma\}^{1/a}\}^{(a-b)_+}\Big] + \frac{\bar{\epsilon}_p}{N}(1+\epsilon_q M^a)\min\{M, \max\{1, N\gamma\}^{1/a}\}^{(a-b)_+}.$$

Here d_{eff} is defined in (12), $\epsilon_q = \bar{\epsilon}_d + (1+\bar{\epsilon}_d)(\bar{\epsilon}_s\bar{\epsilon}_f + \bar{\epsilon}_s + \bar{\epsilon}_f)$.

Proof. Define $\mathbf{H}_w = \mathbb{E}_{\mathbf{w}^*}[\mathbf{w}^*\mathbf{w}^{*\top}]$ and $\mathbf{E}_q = \mathbf{H}_f^{(q)} - \mathbf{S}\mathbf{H}\mathbf{S}^\top$. Lemma C.22 (applied at $k = \lfloor M/2 \rfloor$) and Lemma G.1 of Zhang et al. (2026b), together with the covariance perturbation and Gaussian sketch trace estimates in the proof of Lemma B.11, give an event of probability at least $1 - e^{-\Omega(M)}$ on which

$$\big\|(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}_{0:k}\big\| \lesssim 1 \quad (1 \le k \le \lfloor M/2 \rfloor),$$
$$\mu_{\min}(\mathbf{S}\mathbf{H}\mathbf{S}^\top) \eqsim M^{-a}, \qquad 0 \preceq \mathbf{E}_q \preceq \mathbf{H}_f^{(q)}, \qquad \operatorname{tr}(\mathbf{E}_q) \lesssim \epsilon_q, \qquad \operatorname{tr}(\mathbf{H}_f^{(q)}) \lesssim 1 + \epsilon_q.$$

Firstly, $\mathbf{H}^{1/2}\mathbf{S}^\top(\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}^{1/2} \preceq \mathbf{I}$ gives

$$\mathbb{E}_{\mathbf{w}^*}\|\mathbf{v}_q^*\|^2_{\mathbf{H}_f^{(q)}} = \operatorname{tr}\Big(\mathbf{H}_w^{1/2}\mathbf{H}\mathbf{S}^\top(\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}\mathbf{H}_w^{1/2}\Big) \le \operatorname{tr}(\mathbf{H}\mathbf{H}_w) \lesssim \sum_{i\ge 1} i^{-b} \lesssim 1.$$

Secondly, we analyze

$$\begin{aligned}
&\mathbb{E}_{\mathbf{w}^*}\Big\|[\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H}_f^{(q)})^t]\mathbf{v}_q^*\Big\|^2_{\mathrm{diag}(\mathbf{H}_f^{(q)})} \\
=\ &\operatorname{tr}\Big(\mathrm{diag}(\mathbf{H}_f^{(q)})[\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H}_f^{(q)})^t]\mathbb{E}_{\mathbf{w}^*}[\mathbf{v}_q^*\mathbf{v}_q^{*\top}][\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H}_f^{(q)})^t]\Big) \\
\le\ &\operatorname{tr}(\mathbf{H}_f^{(q)})\Big\|[\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H}_f^{(q)})^t](\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}\mathbf{H}_w^{1/2}\Big\|^2.
\end{aligned}$$

We bound the last norm by splitting at an integer $1 \le k \le \lfloor M/2 \rfloor$. For the head, use

$$(\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}\mathbf{S}^\top = \mathbf{I} - (\mathbf{H}_f^{(q)})^{-1}\mathbf{E}_q$$

and the contraction $\|\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H}_f^{(q)})^t\| \le 1$. These yield

$$\begin{aligned}
&\Big\|[\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H}_f^{(q)})^t](\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}_{0:k}\mathbf{H}_w^{1/2}\Big\|^2 \\
\le\ &\Big\|\mathbf{I} - (\mathbf{H}_f^{(q)})^{-1}\mathbf{E}_q\Big\|^2\big\|(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}_{0:k}\big\|^2\|\mathbf{I}_{0:k}\mathbf{H}_w\mathbf{I}_{0:k}\| \\
\lesssim\ &(1+\epsilon_q M^a)k^{(a-b)_+}.
\end{aligned}$$

The last inequality follows without any commutativity assumption: $\mathbf{E}_q(\mathbf{H}_f^{(q)})^{-1}\mathbf{E}_q \preceq \mathbf{E}_q$, so

$$\Big\|(\mathbf{H}_f^{(q)})^{-1}\mathbf{E}_q\Big\|^2 = \Big\|\mathbf{E}_q(\mathbf{H}_f^{(q)})^{-2}\mathbf{E}_q\Big\| \le \frac{\|\mathbf{E}_q(\mathbf{H}_f^{(q)})^{-1}\mathbf{E}_q\|}{\mu_{\min}(\mathbf{H}_f^{(q)})} \le \frac{\|\mathbf{E}_q\|}{\mu_{\min}(\mathbf{H}_f^{(q)})} \lesssim \epsilon_q M^a.$$

For the tail, Assumption 3.6 gives $\mathbf{H}_{k:\infty}\mathbf{H}_w\mathbf{H}_{k:\infty} \precsim k^{-b}\mathbf{H}_{k:\infty}$. Combining this with $\mathbf{S}\mathbf{H}_{k:\infty}\mathbf{S}^\top \preceq \mathbf{H}_f^{(q)}$ gives

$$\begin{aligned}
&\Big\|[\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H}_f^{(q)})^t](\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}_{k:\infty}\mathbf{H}_w^{1/2}\Big\|^2 \\
\lesssim\ &k^{-b}\Big\|[\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H}_f^{(q)})^t]^2(\mathbf{H}_f^{(q)})^{-1}\Big\| \\
\lesssim\ &k^{-b}\min\left\{\gamma t, \frac{1}{\mu_{\min}(\mathbf{H}_f^{(q)})}\right\} \lesssim k^{-b}\min\{N\gamma, M^a\}.
\end{aligned}$$

Choose

$$k = \min\left\{\left\lceil \max\{1, N\gamma\}^{1/a}\right\rceil, \lfloor M/2\rfloor\right\}.$$

Then $k \asymp \min\{M, \max\{1, N\gamma\}^{1/a}\}$ and $\min\{N\gamma, M^a\} \lesssim k^a$. The squared norm of the head plus tail is at most twice the sum of their squared norms, and $k^{a-b} \le k^{(a-b)_+}$. Therefore

$$\begin{aligned}&\mathbb{E}_{\mathbf{w}^*} \sum_{t=0}^{N-1} \left\| [\mathbf{I} - (\mathbf{I} - \gamma \mathbf{H}_f^{(q)})^t] \mathbf{v}_q^* \right\|^2_{\mathrm{diag}(\mathbf{H}_f^{(q)})} \\ &\lesssim N \operatorname{tr}(\mathbf{H}_f^{(q)})(1 + \epsilon_q M^a) \min\{M, \max\{1, N\gamma\}^{1/a}\}^{(a-b)_+}.\end{aligned}$$

Therefore,

$$\mathbb{E}_{\mathbf{w}^*}[\mathrm{var}] \lesssim \frac{d_{\mathrm{eff}}}{BN}(1 + \overline{\epsilon}_a + \overline{\epsilon}_o) + \frac{\overline{\epsilon}_p}{N}\left(\frac{d_{\mathrm{eff}}}{B} + 1\right)(1 + \epsilon_q M^a) \min\{M, \max\{1, N\gamma\}^{1/a}\}^{(a-b)_+}.$$

□

Lemma B.13 (Refined variance bound in $b > a + 1$)**.** *Under the assumptions of Lemma B.12, suppose that $b > a + 1$. For every fixed $r > 0$, with probability at least $1 - M^{-r} - e^{-\Omega(M)}$,*

$$\begin{aligned}\mathbb{E}_{\mathbf{w}^*}[\mathrm{var}] \lesssim\ & \frac{(1 + \overline{\epsilon}_a + \overline{\epsilon}_o + \overline{\epsilon}_p) d_{\mathrm{eff}}}{BN} \\ &+ \frac{\overline{\epsilon}_p (1 + \epsilon_q M^a) \log M}{MN}\left(\frac{d_{\mathrm{eff}}}{B} + \min\{1, d_{\mathrm{eff}}\}\right).\end{aligned}$$

Here d_{eff} and ϵ_q are as in Lemma B.12, and the implicit constants may depend on a, b, r. In particular, if $\epsilon_q M^a \lesssim 1$, then on the same event,

$$\mathbb{E}_{\mathbf{w}^*}[\mathrm{var}] \lesssim \frac{(1 + \overline{\epsilon}_a + \overline{\epsilon}_o + \overline{\epsilon}_p) d_{\mathrm{eff}}}{BN} + \frac{\overline{\epsilon}_p \log M}{MN}. \tag{15}$$

Proof. The key ingredient is the notice that

$$\mathbb{E}_{\mathbf{w}^*} \|[\mathbf{I} - (\mathbf{I} - \gamma \mathbf{H}_f^{(q)})^t] \mathbf{v}_q^*\|^2_{\mathrm{diag}(\mathbf{H}_f^{(q)})} \le \|\operatorname{diag}(\mathbf{H}_f^{(q)})\| \cdot \mathbb{E}_{\mathbf{w}^*} \|\mathbf{v}_q^*\|^2.$$

We first bound the largest diagonal entry of $\mathbf{H}_f^{(q)}$. The coordinate-wise covariance estimate in the proof of Lemma B.11 gives, for every $1 \le j \le M$,

$$\begin{aligned}(\mathbf{H}_f^{(q)})_{jj} \le\ & (1 + \overline{\epsilon}_f)(\mathbf{S}\mathbf{H}\mathbf{S}^\top)_{jj} \\ &+ (1 + \overline{\epsilon}_f)\left[\overline{\epsilon}_d + (1 + \overline{\epsilon}_d)\overline{\epsilon}_s\right](\mathbf{S} \operatorname{diag}(\mathbf{H}) \mathbf{S}^\top)_{jj}.\end{aligned}$$

The sum of the two coefficients is $(1 + \overline{\epsilon}_d)(1 + \overline{\epsilon}_s)(1 + \overline{\epsilon}_f) = 1 + \epsilon_q$. For $\mathbf{R} \in \{\mathbf{H}, \operatorname{diag}(\mathbf{H})\}$, the Gaussian quadratic-form upper-tail bound gives, for every $t > 0$,

$$\mathbb{P}\left((\mathbf{S}\mathbf{R}\mathbf{S}^\top)_{jj} > \frac{\operatorname{tr}(\mathbf{R}) + 2\sqrt{\operatorname{tr}(\mathbf{R}^2) t} + 2\|\mathbf{R}\| t}{M}\right) \le e^{-t}.$$

Both choices of $\mathbf{R}$ have trace $\operatorname{tr}(\mathbf{H}) \eqsim 1$, and $\operatorname{tr}(\mathbf{R}^2) \le \operatorname{tr}(\mathbf{R})^2$, $\|\mathbf{R}\| \le \operatorname{tr}(\mathbf{R})$. Taking $t = (r + 3)\log M$ and applying a union bound over the two choices and all M rows gives failure probability at most $2M^{-r-2}$. Consequently, simultaneously over all j,

$$\|\operatorname{diag}(\mathbf{H}_f^{(q)})\| \lesssim \frac{(1 + \epsilon_q)\log M}{M}. \tag{16}$$

Next, recall $\mathbf{H}_w = \mathbb{E}_{\mathbf{w}^*}[\mathbf{w}^* \mathbf{w}^{*\top}]$ and $\mathbf{E}_q = \mathbf{H}_f^{(q)} - \mathbf{S}\mathbf{H}\mathbf{S}^\top$. Take $k = \lfloor M/2 \rfloor$. On the sketch event used in the proof of Lemma B.12, whose failure probability is $e^{-\Omega(M)}$,

$$\begin{aligned}\|(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} \mathbf{S}\mathbf{H}_{0:k}\| \lesssim 1, \qquad \mu_{\min}(\mathbf{H}_f^{(q)}) \gtrsim M^{-a}, \\ 0 \preceq \mathbf{E}_q \preceq \mathbf{H}_f^{(q)}, \qquad \|\mathbf{E}_q\| \le \operatorname{tr}(\mathbf{E}_q) \lesssim \epsilon_q.\end{aligned}$$

Since $\mathbf{E}_q(\mathbf{H}_f^{(q)})^{-1}\mathbf{E}_q \preceq \mathbf{E}_q$, we have

$$\|(\mathbf{H}_f^{(q)})^{-1}\mathbf{E}_q\|^2 \le \frac{\|\mathbf{E}_q(\mathbf{H}_f^{(q)})^{-1}\mathbf{E}_q\|}{\mu_{\min}(\mathbf{H}_f^{(q)})} \lesssim \epsilon_q M^a,$$
$$\|(\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}\mathbf{S}^\top\|^2 = \|\mathbf{I} - (\mathbf{H}_f^{(q)})^{-1}\mathbf{E}_q\|^2 \lesssim 1 + \epsilon_q M^a.$$

By Assumption 3.6, the head therefore satisfies

$$\begin{aligned}&\|(\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}_{0:k}\mathbf{H}_w^{1/2}\|_{\mathrm{F}}^2\\ \le&\|(\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}\mathbf{S}^\top\|^2\|(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}_{0:k}\|^2\,\mathrm{tr}(\mathbf{I}_{0:k}\mathbf{H}_w\mathbf{I}_{0:k})\\ \lesssim&(1+\epsilon_q M^a)\sum_{i=1}^{k} i^{a-b}.\end{aligned}$$

For the tail, use

$$(\mathbf{H}_f^{(q)})^{-2} \preceq \frac{(\mathbf{H}_f^{(q)})^{-1}}{\mu_{\min}(\mathbf{H}_f^{(q)})}, \qquad \lambda_i \mathbf{v}_i^\top \mathbf{S}^\top (\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{v}_i \le 1.$$

The second inequality follows from $\lambda_i \mathbf{S}\mathbf{v}_i\mathbf{v}_i^\top\mathbf{S}^\top \preceq \mathbf{H}_f^{(q)}$. Hence,

$$\begin{aligned}&\|(\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}_{k:\infty}\mathbf{H}_w^{1/2}\|_{\mathrm{F}}^2\\ =&\sum_{i>k}\lambda_i^2\mathbb{E}_{\mathbf{w}^*}[\langle\mathbf{w}^*,\mathbf{v}_i\rangle^2]\mathbf{v}_i^\top\mathbf{S}^\top(\mathbf{H}_f^{(q)})^{-2}\mathbf{S}\mathbf{v}_i\\ \lesssim&M^a\sum_{i>k} i^{-b} \lesssim M^{a+1-b}.\end{aligned}$$

The head and tail have no cross term in the prior expectation, since $\mathbf{H}_w$ is diagonal in the eigenbasis of $\mathbf{H}$. Using $\mathbf{v}_q^* = (\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}\mathbf{w}^*$ and $b > a+1$, we obtain

$$\mathbb{E}_{\mathbf{w}^*}\|\mathbf{v}_q^*\|^2 \lesssim (1+\epsilon_q M^a)\sum_{i=1}^{k} i^{a-b} + M^{a+1-b} \lesssim 1 + \epsilon_q M^a.$$

The first assertion is immediately proved by Theorem B.10. □

Lemma B.14. *If there exist constants $\overline{\epsilon}_s, \overline{\epsilon}_d, \overline{\epsilon}_f$ such that for $i \in \{s, d, f\}$, $\mathcal{Q}_i(\cdot)$ is $\overline{\epsilon}_i$-multiplicative, under Assumption 3.1, Assumption 3.3 and Assumption 3.6, with probability at least $1 - e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$, with d_{eff} defined in (12), it holds*

$$\frac{d_{\mathrm{eff}}}{N} \lesssim \min\left\{\frac{M}{N},\ \gamma^{1/a}N^{-(a-1)/a} + \gamma[\overline{\epsilon}_d + (1+\overline{\epsilon}_d)(\overline{\epsilon}_s\overline{\epsilon}_f + \overline{\epsilon}_s + \overline{\epsilon}_f)]\right\}.$$

Proof. Denote $\epsilon_q = \overline{\epsilon}_d + (1+\overline{\epsilon}_d)(\overline{\epsilon}_s\overline{\epsilon}_f + \overline{\epsilon}_s + \overline{\epsilon}_f)$. Let $k_0 = \lceil \max\{(\gamma N)^{1/a}, \gamma N\epsilon_q\}\rceil$.

If $k_0 \le \lfloor M/2\rfloor$, Lemma B.11 gives

$$\begin{aligned}\frac{d_{\mathrm{eff}}}{N} &\le \frac{k_0 + \gamma^2N^2\sum_{i>k_0}(\tilde{\lambda}_i^{(q)})^2}{N}\\ &\lesssim \frac{k_0}{N} + \gamma^2 N k_0^{1-2a} + \frac{\gamma^2 N\epsilon_q^2}{k_0}\\ &\lesssim \gamma^{1/a}N^{-(a-1)/a} + \gamma\epsilon_q.\end{aligned}$$

If $k_0 > \lfloor M/2\rfloor$, $\frac{d_{\mathrm{eff}}}{N} \le \frac{M}{N}$. Overall,

$$\frac{d_{\mathrm{eff}}}{N} \lesssim \min\left\{\frac{M}{N},\ \gamma^{1/a}N^{-(a-1)/a} + \gamma\epsilon_q\right\}.$$

□

Lemma B.15 (Bias bound when $b < a+1$)**.** *Under Assumption 3.1, Assumption 3.3, and Assumption 3.6, for any $i \in \{s, d, f\}$, if there exist $\bar{\epsilon}_i$ such that quantization $\mathcal{Q}_i$ is $\bar{\epsilon}_i$-multiplicative, then in the regime of $b < a + 1$, with probability at least $1 - e^{-\Omega(M)}$ over $\mathbf{S}$, simultaneously for every integer $1 \le k \le \lfloor M/2 \rfloor$, it holds*

$$\mathbb{E}_{\mathbf{w}^*}\left[\frac{1}{\gamma N}\cdot\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{I}^{(q)}_{f,0:k^*}} + \left\|\mathbf{v}_q^*\right\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}\right] \lesssim \frac{k^{1+a-b}}{\gamma N} + k^{1-b} + \epsilon_q k^{(a-b)_+},$$

with $\epsilon_q = \bar{\epsilon}_d + (1+\bar{\epsilon}_d)(\bar{\epsilon}_s\bar{\epsilon}_f + \bar{\epsilon}_s + \bar{\epsilon}_f)$. Consequently, by selecting k,

$$\mathbb{E}_{\mathbf{w}^*}\left[\frac{1}{\gamma N}\cdot\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{I}^{(q)}_{f,0:k^*}} + \left\|\mathbf{v}_q^*\right\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}\right] \lesssim \min\left\{1, \max\left\{(N\gamma)^{-(b-1)/a}, M^{1-b}\right\} + \epsilon_q \min\{(N\gamma)^{1/a}, M\}^{(a-b)_+}\right\}.$$

Proof. Denote $\mathbf{H}_w = \mathbb{E}[\mathbf{w}^*\mathbf{w}^{*\top}]$. Notice that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{w}^*}\left[\frac{1}{N\gamma}\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{I}^{(q)}_{f,0:k^*}} + \left\|\mathbf{v}_q^*\right\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}\right] \\
=&\mathbb{E}_{\mathbf{w}^*}\left\|(\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}\mathbf{w}^*\right\|^2_{\frac{1}{N\gamma}\mathbf{I}^{(q)}_{f,0:k^*}+\mathbf{H}^{(q)}_{f,k^*:\infty}} \\
=&\left\|\left(\frac{1}{N\gamma}\mathbf{I}^{(q)}_{f,0:k^*} + \mathbf{H}^{(q)}_{f,k^*:\infty}\right)^{1/2}(\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}\mathbf{H}_w^{1/2}\right\|^2_{\mathrm{F}}.
\end{aligned}$$

Utilizing Lemma C.22 of Zhang et al. (2026b), for $k \le M/2$, with probability at least $1 - e^{-\Omega(M)}$, it holds

$$\left\|(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}_{0:k}\right\|^2 \lesssim 1. \tag{17}$$

It follows that

$$\left\|(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}_{0:k}\right\|^2_F \lesssim k.$$

Recall that $\mathbf{E}_q = \mathbf{H}_f^{(q)} - \mathbf{S}\mathbf{H}\mathbf{S}^\top$, it follows that

$$(\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}\mathbf{S}^\top = \mathbf{I} - (\mathbf{H}_f^{(q)})^{-1}\mathbf{E}_q,$$

then using $\frac{1}{N\gamma}\mathbf{I}^{(q)}_{f,0:k^*} + \mathbf{H}^{(q)}_{f,k^*:\infty} \preceq \frac{1}{N\gamma}\mathbf{I}$ gives

$$\begin{aligned}
&\left\|\left(\frac{1}{N\gamma}\mathbf{I}^{(q)}_{f,0:k^*} + \mathbf{H}^{(q)}_{f,k^*:\infty}\right)^{1/2}(\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}_{0:k}\mathbf{H}_w^{1/2}\right\|^2_{\mathrm{F}} \\
\le& 2\left\|\left(\frac{1}{N\gamma}\mathbf{I}^{(q)}_{f,0:k^*} + \mathbf{H}^{(q)}_{f,k^*:\infty}\right)^{1/2}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}_{0:k}\mathbf{H}_w^{1/2}\right\|^2_{\mathrm{F}} \\
&+2\left\|\left(\frac{1}{N\gamma}\mathbf{I}^{(q)}_{f,0:k^*} + \mathbf{H}^{(q)}_{f,k^*:\infty}\right)^{1/2}(\mathbf{H}_f^{(q)})^{-1}\mathbf{E}_q(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}_{0:k}\mathbf{H}_w^{1/2}\right\|^2_{\mathrm{F}} \\
\lesssim& \frac{\sum_{i=1}^k i^{a-b}}{N\gamma} + \mathrm{tr}\left(\mathbf{E}_q(\mathbf{H}_f^{(q)})^{-1}\left(\frac{1}{N\gamma}\mathbf{I}^{(q)}_{f,0:k^*} + \mathbf{H}^{(q)}_{f,k^*:\infty}\right)(\mathbf{H}_f^{(q)})^{-1}\mathbf{E}_q\right)\left\|(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}_{0:k}\mathbf{H}_w^{1/2}\right\|^2 \\
\lesssim& \frac{k^{1+a-b}}{N\gamma} + \mathrm{tr}(\mathbf{E}_q)\|\mathbf{I}_{0:k}\mathbf{H}_w\mathbf{I}_{0:k}\| \\
\eqsim& \frac{k^{1+a-b}}{N\gamma} + \mathrm{tr}(\mathbf{E}_q)k^{(a-b)_+},
\end{aligned} \tag{18}$$

where the third inequality holds as $\frac{1}{N\gamma}\mathbf{I}^{(q)}_{f,0:k^*} + \mathbf{H}^{(q)}_{f,k^*:\infty} \preceq \mathbf{H}_f^{(q)}$ and $\mathbf{E}_q \preceq \mathbf{H}_f^{(q)}$, $(a-b)_+ := \max\{a-b, 0\}$. For the tail, use $\frac{1}{N\gamma}\mathbf{I}^{(q)}_{f,0:k^*} + \mathbf{H}^{(q)}_{f,k^*:\infty} \preceq \mathbf{H}_f^{(q)}$ again,

$$\left\|\left(\frac{1}{N\gamma}\mathbf{I}^{(q)}_{f,0:k^*} + \mathbf{H}^{(q)}_{f,k^*:\infty}\right)^{1/2}(\mathbf{H}_f^{(q)})^{-1}\mathbf{S}\mathbf{H}_{k:\infty}\mathbf{H}_w^{1/2}\right\|^2_{\mathrm{F}} \lesssim \sum_{i>k} i^{-b} \lesssim k^{1-b}. \tag{19}$$

Further by Lemma B.11, $\mathrm{tr}(\mathbf{E}_q) \lesssim \epsilon_q$,

$$\mathbb{E}_{\mathbf{w}^*}\left[\frac{1}{\gamma N}\cdot\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{I}^{(q)}_{f,0:k^*}}+\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}\right]\lesssim\frac{k^{1+a-b}}{\gamma N}+k^{1-b}+\epsilon_q k^{(a-b)_+}.$$

Selecting

$$k=\min\left\{\left\lceil(N\gamma)^{1/a}\right\rceil,\lfloor M/2\rfloor\right\}$$

when $N\gamma\geq 1$, yields

$$\frac{k^{1+a-b}}{N\gamma}+k^{1-b}\lesssim\max\left\{(N\gamma)^{-(b-1)/a},M^{1-b}\right\}.$$

For $N\gamma<1$, use the preceding k-independent bound $\mathrm{tr}(\mathbf{H}\mathbf{H}_w)\lesssim 1$. This proves the consequence. □

Lemma B.16 (Bias bound when $b=a+1$)**.** *Under Assumption 3.1, Assumption 3.3, and Assumption 3.6, for any $i\in\{s,d,f\}$, if there exist $\overline{\epsilon}_i$ such that quantization $\mathcal{Q}_i$ is $\overline{\epsilon}_i$-multiplicative, then in the regime of $b=a+1$ and $N\gamma\gtrsim 1$, with probability at least $1-e^{-\Omega(M)}$ over $\mathbf{S}$, simultaneously for every integer $1\leq k\leq\lfloor M/2\rfloor$, it holds*

$$\mathbb{E}_{\mathbf{w}^*}\left[\frac{1}{\gamma N}\cdot\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{I}^{(q)}_{f,0:k^*}}+\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}\right]\lesssim\frac{\log k}{\gamma N}+k^{1-b}+\epsilon_q,$$

with $\epsilon_q=\overline{\epsilon}_d+(1+\overline{\epsilon}_d)(\overline{\epsilon}_s\overline{\epsilon}_f+\overline{\epsilon}_s+\overline{\epsilon}_f)$. Consequently, by selecting k,

$$\mathbb{E}_{\mathbf{w}^*}\left[\frac{1}{\gamma N}\cdot\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{I}^{(q)}_{f,0:k^*}}+\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}\right]\lesssim\max\left\{\frac{\log(N\gamma)}{N\gamma},M^{1-b}\right\}+\epsilon_q.$$

Proof. By (18) and (19), for $1\leq k\leq\lfloor M/2\rfloor$,

$$\begin{aligned}\mathbb{E}_{\mathbf{w}^*}\left[\frac{1}{\gamma N}\cdot\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{I}^{(q)}_{f,0:k^*}}+\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}\right]\lesssim&\frac{\sum_{i=1}^k i^{a-b}}{N\gamma}+\epsilon_q k^{(a-b)_+}+k^{1-b}\\ \eqsim&\frac{\log k}{N\gamma}+k^{1-b}+\epsilon_q.\end{aligned}$$

Selecting

$$k=\min\left\{\left\lceil(N\gamma)^{1/a}\right\rceil,\lfloor M/2\rfloor\right\}$$

yields

$$\mathbb{E}_{\mathbf{w}^*}\left[\frac{1}{\gamma N}\cdot\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{I}^{(q)}_{f,0:k^*}}+\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}\right]\lesssim\frac{\log(N\gamma)}{N\gamma}+M^{1-b}+\epsilon_q.$$

□

Lemma B.17 (Bias bound when $b>a+1$)**.** *Under Assumption 3.1, Assumption 3.3, and Assumption 3.6, for any $i\in\{s,d,f\}$, if there exist $\overline{\epsilon}_i$ such that quantization $\mathcal{Q}_i$ is $\overline{\epsilon}_i$-multiplicative, then in the regime of $b>a+1$ and $N\gamma\gtrsim 1$, with probability at least $1-e^{-\Omega(M)}$ over $\mathbf{S}$, simultaneously for every integer $1\leq k\leq\lfloor M/2\rfloor$, it holds*

$$\mathbb{E}_{\mathbf{w}^*}\left[\frac{1}{\gamma N}\cdot\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{I}^{(q)}_{f,0:k^*}}+\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}\right]\lesssim\frac{1}{\gamma N}+k^{1-b}+\epsilon_q,$$

with $\epsilon_q=\overline{\epsilon}_d+(1+\overline{\epsilon}_d)(\overline{\epsilon}_s\overline{\epsilon}_f+\overline{\epsilon}_s+\overline{\epsilon}_f)$. Consequently, by selecting k,

$$\mathbb{E}_{\mathbf{w}^*}\left[\frac{1}{\gamma N}\cdot\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{I}^{(q)}_{f,0:k^*}}+\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}\right]\lesssim\max\left\{\frac{1}{N\gamma},M^{1-b}\right\}+\epsilon_q.$$

Proof. By (18) and (19), for $1\leq k\leq\lfloor M/2\rfloor$,

$$\begin{aligned}\mathbb{E}_{\mathbf{w}^*}\left[\frac{1}{\gamma N}\cdot\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{I}^{(q)}_{f,0:k^*}}+\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}\right]\lesssim&\frac{\sum_{i=1}^k i^{a-b}}{N\gamma}+\epsilon_q k^{(a-b)_+}+k^{1-b}\\ \eqsim&\frac{1}{N\gamma}+k^{1-b}+\epsilon_q.\end{aligned}$$

Selecting $k = \lfloor M/2 \rfloor$ yields

$$\mathbb{E}_{\mathbf{w}^*}\left[\frac{1}{\gamma N}\cdot\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{I}^{(q)}_{f,0:k^*}} + \left\|\mathbf{v}_q^*\right\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}\right] \lesssim \frac{1}{N\gamma} + M^{1-b} + \epsilon_q.$$

□

B.5.1 Tight Bounds of Bias Error

We improve the bias bounds established in Lemma B.17 under $b > a+1$ under specific regime of overparameterization $N\gamma \lesssim M^a$.

Lemma B.18. *Under Assumption 3.1, 3.3 and 3.6, suppose that $N\gamma \le c_0 M^a$ for some small constant $c_0 > 0$, and that $\mathcal{Q}_i$ is $\bar{\epsilon}_i$-multiplicative for $i \in \{d, s, f\}$, let $k = \lceil (N\gamma)^{1/a} \rceil$, for some fixed $r > 0$, with probability at least $1 - e^{-\Omega(M)} - M^{-r}$, if $b > a+1$,*

$$\mathbb{E}_{\mathbf{w}^*}\left[\frac{1}{\gamma^2 N^2}\left\|\mathbf{v}_q^*\right\|^2_{(\mathbf{H}^{(q)}_{f,0:k^*})^{-1}} + \left\|\mathbf{v}_q^*\right\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}\right] \lesssim \min\left\{\frac{1}{\gamma^2 N^2}\sum_{i=1}^{k} i^{2a-b} + \frac{k+\log M}{MN\gamma} + \epsilon_q + \sum_{i>k} i^{-b}, 1\right\},$$

where $\epsilon_q = \bar{\epsilon}_d + (1+\bar{\epsilon}_d)(\bar{\epsilon}_s\bar{\epsilon}_f + \bar{\epsilon}_s + \bar{\epsilon}_f)$. Moreover, if also $(N\gamma)^{(\min\{b,2a+1\}-a)/a} \lesssim M/\log M$,

$$\mathbb{E}_{\mathbf{w}^*}\left[\frac{1}{\gamma^2 N^2}\left\|\mathbf{v}_q^*\right\|^2_{(\mathbf{H}^{(q)}_{f,0:k^*})^{-1}} + \left\|\mathbf{v}_q^*\right\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}\right] \lesssim \min\left\{\widetilde{\Theta}\left((N\gamma)^{\frac{1-\min\{b,2a+1\}}{a}}\right) + \epsilon_q, 1\right\}.$$

Proof. Recall $\mathbf{E}_q = \mathbf{H}_f^{(q)} - \mathbf{S}\mathbf{H}\mathbf{S}^\top \succeq 0$ and

$$\mathbf{v}^* = (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}\mathbf{w}^*, \qquad \mathbf{v}_q^* = \mathbf{v}^* - (\mathbf{H}_f^{(q)})^{-1}\mathbf{E}_q\mathbf{v}^*.$$

For $\theta > 0$, the resolvent identity $(\mathbf{S}\mathbf{H}\mathbf{S}^\top + \mathbf{E}_q + \theta\mathbf{I})^{-1} - (\mathbf{S}\mathbf{H}\mathbf{S}^\top + \theta\mathbf{I})^{-1} = -(\mathbf{S}\mathbf{H}\mathbf{S}^\top + \mathbf{E}_q + \theta\mathbf{I})^{-1}\mathbf{E}_q(\mathbf{S}\mathbf{H}\mathbf{S}^\top + \theta\mathbf{I})^{-1}$ gives

$$\left\|(\mathbf{H}_f^{(q)} + \theta\mathbf{I})^{-1} - (\mathbf{S}\mathbf{H}\mathbf{S}^\top + \theta\mathbf{I})^{-1}\right\| \le \theta^{-2}\|\mathbf{E}_q\|,$$
$$\left\|(\mathbf{H}_f^{(q)} + \theta\mathbf{I})^{-2} - (\mathbf{S}\mathbf{H}\mathbf{S}^\top + \theta\mathbf{I})^{-2}\right\| \le 2\theta^{-3}\|\mathbf{E}_q\|.$$

Expanding each covariance times its squared resolvent as the difference of its resolvent and θ times its squared resolvent yields

$$\left\|\theta^2\mathbf{H}_f^{(q)}(\mathbf{H}_f^{(q)} + \theta\mathbf{I})^{-2} - \theta^2\mathbf{S}\mathbf{H}\mathbf{S}^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top + \theta\mathbf{I})^{-2}\right\| \le 3\|\mathbf{E}_q\|.$$

Since $\theta^2(\mathbf{H}_f^{(q)} + \theta\mathbf{I})^{-2} \preceq \mathbf{I}$ and $\mathbf{E}_q(\mathbf{H}_f^{(q)})^{-1}\mathbf{E}_q \preceq \mathbf{E}_q$,

$$\begin{aligned}
&\theta^2(\mathbf{v}_q^*)^\top\mathbf{H}_f^{(q)}(\mathbf{H}_f^{(q)} + \theta\mathbf{I})^{-2}\mathbf{v}_q^* \\
\le& 2\theta^2(\mathbf{v}^*)^\top\mathbf{H}_f^{(q)}(\mathbf{H}_f^{(q)} + \theta\mathbf{I})^{-2}\mathbf{v}^* + 2(\mathbf{v}^*)^\top\mathbf{E}_q\mathbf{v}^* \\
\le& 2\theta^2(\mathbf{v}^*)^\top\mathbf{S}\mathbf{H}\mathbf{S}^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top + \theta\mathbf{I})^{-2}\mathbf{v}^* + 8\|\mathbf{E}_q\|\,\|\mathbf{v}^*\|^2.
\end{aligned}$$

The exact-sketch head and tail estimates in the proof of Lemma B.13 give, since $b > a+1$,

$$\mathbb{E}_{\mathbf{w}^*}\|\mathbf{v}^*\|^2 \lesssim \sum_{i\le\lfloor M/2\rfloor} i^{a-b} + M^{a+1-b} \lesssim 1.$$

Taking $\theta = (N\gamma)^{-1}$ and using $\|\mathbf{E}_q\| \lesssim \epsilon_q$ from Lemma B.11, the diagonality of $\mathbf{H}_w$ gives

$$\begin{aligned}
&\mathbb{E}_{\mathbf{w}^*}\left[\frac{1}{\gamma^2 N^2}\|\mathbf{v}_q^*\|^2_{(\mathbf{H}^{(q)}_{f,0:k^*})^{-1}} + \|\mathbf{v}_q^*\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}\right] \\
\lesssim& \frac{1}{N^2\gamma^2}\sum_i \lambda_i^2\mathbf{v}_i^\top\mathbf{H}_w\mathbf{v}_i\,\mathbf{v}_i^\top\mathbf{S}^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}(\mathbf{S}\mathbf{H}\mathbf{S}^\top + (N\gamma)^{-1}\mathbf{I})^{-2}\mathbf{S}\mathbf{v}_i + \epsilon_q,
\end{aligned}$$

where we use the fact that for every $x > 0$, the scalar filter satisfies

$$\frac{(N\gamma)^{-2}x}{(x+(N\gamma)^{-1})^2} \le \min\left\{x, \frac{1}{\gamma^2 N^2 x}\right\} \le \frac{4(N\gamma)^{-2}x}{(x+(N\gamma)^{-1})^2}.$$

By the definition of k^*, the middle expression evaluated on $\mathbf{H}_f^{(q)}$ equals $(\gamma^2N^2)^{-1}(\mathbf{H}_{f,0:k^*}^{(q)})^{-1} + \mathbf{H}_{f,k^*:\infty}^{(q)}$.

Since $\mathbf{S}\mathbf{H}\mathbf{S}^\top \preceq \mathbf{H}_f^{(q)}$, we have

$$\lambda_i \mathbf{v}_i^\top \mathbf{S}^\top (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{v}_i \le 1.$$

Noticing that $(N\gamma)^{-2}(\mathbf{S}\mathbf{H}\mathbf{S}^\top + (N\gamma)^{-1}\mathbf{I})^{-2} \preceq \mathbf{I}$, then the part of $i > k$ is bounded by

$$\sum_{i>k} \lambda_i(\mathbf{v}_i^\top \mathbf{H}_w \mathbf{v}_i) \lesssim \sum_{i>k} i^{-b}. \tag{20}$$

For the head $i \le k$, define the exact-sketch leave-one-out matrix $\mathbf{R}_i := \sum_{j\neq i}\lambda_j \mathbf{S}\mathbf{v}_j\mathbf{v}_j^\top\mathbf{S}^\top = \mathbf{S}\mathbf{H}\mathbf{S}^\top - \lambda_i\mathbf{S}\mathbf{v}_i\mathbf{v}_i^\top\mathbf{S}^\top$ and $\mathbf{B}_i := \mathbf{R}_i(\mathbf{R}_i + (N\gamma)^{-1}\mathbf{I})^{-2}$. We first upper bound $\mathbf{v}_i^\top\mathbf{S}^\top\mathbf{B}_i\mathbf{S}\mathbf{v}_i$. Using the Gaussian quadratic-form inequality, for every $t > 0$, with probability at least $1 - e^{-t}$,

$$\mathbf{v}_i^\top\mathbf{S}^\top\mathbf{B}_i\mathbf{S}\mathbf{v}_i \le \frac{1}{M}\left[\mathrm{tr}(\mathbf{B}_i) + 2\sqrt{\mathrm{tr}(\mathbf{B}_i^2)t} + 2\|\mathbf{B}_i\|t\right].$$

Noticing that for every $u \ge 0$,

$$\frac{u}{(u+(N\gamma)^{-1})^2} \le \frac{N\gamma}{4}, \qquad \frac{u}{(u+(N\gamma)^{-1})^2} \le (N\gamma)^2 u,$$

and for $1 \le j \le M$, $\lambda_j(\mathbf{R}_i) \le \lambda_j(\mathbf{S}\mathbf{H}\mathbf{S}^\top) \lesssim j^{-a}$ with probability at least $1 - e^{-\Omega(M)}$, we have, by taking $t = (r+2)\log M$, with probability at least $1 - M^{-r} - e^{-\Omega(M)}$, simultaneously for all $i \le k$,

$$\mathbf{v}_i^\top\mathbf{S}^\top\mathbf{R}_i(\mathbf{R}_i + (N\gamma)^{-1}\mathbf{I})^{-2}\mathbf{S}\mathbf{v}_i \lesssim \frac{N\gamma(k+\log M)}{M}, \tag{21}$$

where we use

$$\|\mathbf{B}_i\| \le \frac{N\gamma}{4}, \quad \mathrm{tr}(\mathbf{B}_i) \lesssim N\gamma k + (N\gamma)^2\sum_{j>k} j^{-a} \lesssim N\gamma k,$$

with $k = \lceil (N\gamma)^{1/a}\rceil$. We then lower bound $\mathbf{v}_i^\top\mathbf{S}^\top(\mathbf{R}_i + (N\gamma)^{-1}\mathbf{I})^{-1}\mathbf{S}\mathbf{v}_i$. Consider $(N\gamma)^{1/a} < \frac{M}{2}$, then with probability at least $1 - e^{-\Omega(M)}$, at least $m_i \ge M/2$ eigenvalues of $\mathbf{R}_i$ satisfy are not larger than $(2N\gamma)^{-1}$. If $\mathbf{P}_i$ denotes the orthogonal projection onto this eigenspace, then

$$(\mathbf{R}_i + (N\gamma)^{-1}\mathbf{I})^{-1} \succeq \frac{N\gamma}{2}\mathbf{P}_i.$$

Since $M\|\mathbf{P}_i\mathbf{S}\mathbf{v}_i\|^2$ has a $\chi^2_{m_i}$ distribution, the standard chi-square lower-tail bound gives

$$\mathbb{P}\left(\|\mathbf{P}_i\mathbf{S}\mathbf{v}_i\|^2 \le \frac{m_i}{2M}\,\middle|\,\mathbf{R}_i\right) \le e^{-\Omega(m_i)} \le e^{-\Omega(M)}.$$

Hence, with probability at least $1 - e^{-\Omega(M)}$, simultaneously for all $i \le k$,

$$\mathbf{v}_i^\top\mathbf{S}^\top(\mathbf{R}_i + (N\gamma)^{-1}\mathbf{I})^{-1}\mathbf{S}\mathbf{v}_i \gtrsim N\gamma. \tag{22}$$

Finally, set $\mathbf{z}_i = \mathbf{S}\mathbf{v}_i$, then $\mathbf{S}\mathbf{H}\mathbf{S}^\top = \mathbf{R}_i + \lambda_i\mathbf{z}_i\mathbf{z}_i^\top$. By

$$(\mathbf{S}\mathbf{H}\mathbf{S}^\top + s\mathbf{I})^{-1} = (\mathbf{R}_i + s\mathbf{I})^{-1} - \frac{\lambda_i(\mathbf{R}_i + s\mathbf{I})^{-1}\mathbf{z}_i\mathbf{z}_i^\top(\mathbf{R}_i + s\mathbf{I})^{-1}}{1 + \lambda_i\mathbf{z}_i^\top(\mathbf{R}_i + s\mathbf{I})^{-1}\mathbf{z}_i},$$

we have

$$\mathbf{z}_i^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top + s\mathbf{I})^{-1}\mathbf{z}_i = \frac{\mathbf{z}_i^\top(\mathbf{R}_i + s\mathbf{I})^{-1}\mathbf{z}_i}{1 + \lambda_i\mathbf{z}_i^\top(\mathbf{R}_i + s\mathbf{I})^{-1}\mathbf{z}_i},$$

$$\mathbf{z}_i^\top(\mathbf{S}\mathbf{H}\mathbf{S}^\top + s\mathbf{I})^{-2}\mathbf{z}_i = \frac{\mathbf{z}_i^\top(\mathbf{R}_i + s\mathbf{I})^{-2}\mathbf{z}_i}{\left[1 + \lambda_i\mathbf{z}_i^\top(\mathbf{R}_i + s\mathbf{I})^{-1}\mathbf{z}_i\right]^2}.$$

Together with the identity

$$\theta^2(\mathbf{SHS}^\top)^{-1}(\mathbf{SHS}^\top+\theta\mathbf{I})^{-2}=(\mathbf{SHS}^\top)^{-1}-(\mathbf{SHS}^\top+\theta\mathbf{I})^{-1}-\theta(\mathbf{SHS}^\top+\theta\mathbf{I})^{-2},\qquad \theta>0,$$

we have

$$\begin{aligned}&\frac{\lambda_i^2}{\gamma^2N^2}\mathbf{z}_i^\top(\mathbf{SHS}^\top)^{-1}(\mathbf{SHS}^\top+(N\gamma)^{-1}\mathbf{I})^{-2}\mathbf{z}_i\\ =&\frac{\lambda_i+\lambda_i^2\mathbf{z}_i^\top\mathbf{R}_i(\mathbf{R}_i+(N\gamma)^{-1}\mathbf{I})^{-2}\mathbf{z}_i}{\left[1+\lambda_i\mathbf{z}_i^\top(\mathbf{R}_i+(N\gamma)^{-1}\mathbf{I})^{-1}\mathbf{z}_i\right]^2}\\ -&\frac{\lambda_i}{1+\lambda_i\mathbf{z}_i^\top\mathbf{R}_i^{-1}\mathbf{z}_i}.\end{aligned}$$

Using (22) and (21) yields

$$\begin{aligned}&\frac{\lambda_i^2}{\gamma^2N^2}\mathbf{v}_i^\top\mathbf{S}^\top(\mathbf{SHS}^\top)^{-1}(\mathbf{SHS}^\top+(N\gamma)^{-1}\mathbf{I})^{-2}\mathbf{S}\mathbf{v}_i\\ \lesssim&\frac{1}{\lambda_i^2(N\gamma)^2}\left[\lambda_i+\lambda_i^2\frac{N\gamma(k+\log M)}{M}\right]\\ \lesssim&\frac{1}{\gamma^2N^2\lambda_i}+\frac{k+\log M}{MN\gamma}.\end{aligned}\tag{23}$$

Multiplying by $\mathbf{v}_i^\top\mathbf{H}_w\mathbf{v}_i$, summing over $i\le k$, together with the tail (20), the preceding comparison of the quantized and exact-sketch filters, and Assumption 3.6, if $b>a+1$,

$$\mathbb{E}_{\mathbf{w}^*}\left[\frac{1}{\gamma^2N^2}\left\|\mathbf{v}_q^*\right\|^2_{(\mathbf{H}^{(q)}_{f,0:k^*})^{-1}}+\left\|\mathbf{v}_q^*\right\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}\right]\lesssim\frac{1}{\gamma^2N^2}\sum_{i=1}^k i^{2a-b}+\frac{k+\log M}{MN\gamma}+\epsilon_q+\sum_{i>k}i^{-b}.$$

The same scalar filter is bounded above by $\mathbf{H}_f^{(q)}$, and $\mathbb{E}_{\mathbf{w}^*}\|\mathbf{v}_q^*\|^2_{\mathbf{H}_f^{(q)}}\le\mathrm{tr}(\mathbf{H}\mathbf{H}_w)\lesssim 1$, which gives the truncation at 1. Evaluating the sums and using $(N\gamma)^{(\min\{b,2a+1\}-a)/a}\lesssim M/\log M$ gives the second bound. □

B.6 Additive Error

In this subsection, we analyze the additive error R_{add} introduced in Lemma B.2.

$$R_{\mathrm{add}}=\underbrace{\left(\mathbf{v}_q^*\right)^\top\frac{1}{N\gamma}\left[\mathbf{I}-\left(\mathbf{I}-\gamma\mathbf{H}_f^{(q)}\right)^N\right]\left(\mathbf{H}_f^{(q)}\right)^{-1}\left(\mathbf{H}_f^{(q)}-\mathbf{SHS}^\top\right)\mathbf{v}_q^*}_{R^{(1)}_{\mathrm{add}}}+\frac{1}{2}\underbrace{\left\langle\mathbf{SHS}^\top,(\mathbf{v}_q^*-\mathbf{v}^*)\otimes(\mathbf{v}_q^*-\mathbf{v}^*)\right\rangle}_{R^{(2)}_{\mathrm{add}}}.$$

We can adapt Lemma C.18 in Zhang et al. (2026b) by replacing $\bar\epsilon_s+\bar\epsilon_f+\bar\epsilon_s\bar\epsilon_d p+\bar\epsilon_d\frac{p}{M}$ with $\epsilon_q=\bar\epsilon_d+(1+\bar\epsilon_d)(\bar\epsilon_s\bar\epsilon_f+\bar\epsilon_s+\bar\epsilon_f)$, which is summarized below.

Lemma B.19. *Under Assumption 3.1, 3.3, 3.6, for any $i\in\{s,d,f\}$, if there exist $\bar\epsilon_i$ such that quantization $\mathcal{Q}_i$ is $\bar\epsilon_i$-multiplicative, $\gamma\lesssim\frac{1}{\alpha\mathrm{tr}(\mathbf{H}_f^{(q)})}$ then with probability at least $1-e^{-\Omega(M)}$,*

$$\mathbb{E}_{\mathbf{w}^*}R^{(1)}_{\mathrm{add}}\lesssim\frac{\epsilon_q}{\epsilon_q+M^{-a}},\quad \mathbb{E}_{\mathbf{w}^*}R^{(2)}_{\mathrm{add}}\lesssim\frac{\epsilon_q^2}{(M^{-a}+\epsilon_q)^2}.$$

We consider the additive error upper bounds in small quantization regime of $\epsilon_qM^a\lesssim 1$.

Lemma B.20. *Under Assumption 3.1, 3.3, 3.6, for any $i\in\{s,d,f\}$, if there exist $\bar\epsilon_i$ such that quantization $\mathcal{Q}_i$ is $\bar\epsilon_i$-multiplicative, $\gamma\lesssim\frac{1}{\alpha\mathrm{tr}(\mathbf{H}_f^{(q)})}$ and $\epsilon_qM^a\lesssim 1$, then with probability at least $1-e^{-\Omega(M)}$,*

$$\mathbb{E}_{\mathbf{w}^*}R^{(2)}_{\mathrm{add}}\lesssim\epsilon_q^2M^{a+(a-b)_+},\quad \mathbb{E}_{\mathbf{w}^*}R^{(1)}_{\mathrm{add}}\lesssim\epsilon_qM^{(a-b)_+}.$$

Proof. Set $k = \lfloor M/2 \rfloor$. With the probability at least $1 - e^{-\Omega(M)}$, $\|(\mathbf{SHS}^\top)^{-1}\mathbf{SH}_{0:k}\| \lesssim 1$. It follows under Assumption 3.6 that

$$\left\|(\mathbf{SHS}^\top)^{-1}\mathbf{SH}_{0:k}\mathbf{H}_w^{1/2}\right\|^2 \le \left\|(\mathbf{SHS}^\top)^{-1}\mathbf{SH}_{0:k}\right\|^2 \|\mathbf{I}_{0:k}\mathbf{H}_w\mathbf{I}_{0:k}\| \lesssim k^{(a-b)_+}.$$

For the tail, $\lambda_i \mathbf{v}_i^\top \mathbf{H}_w \mathbf{v}_i \lesssim i^{-b} \le k^{-b}$ for $i > k$ gives

$$\mathbf{H}_{k:\infty}\mathbf{H}_w\mathbf{H}_{k:\infty} = \sum_{i>k} \lambda_i^2 (\mathbf{v}_i^\top \mathbf{H}_w \mathbf{v}_i)\mathbf{v}_i\mathbf{v}_i^\top \preceq k^{-b}\mathbf{H}_{k:\infty}.$$

Thus

$$\left\|(\mathbf{SHS}^\top)^{-1}\mathbf{SH}_{k:\infty}\mathbf{H}_w^{1/2}\right\|^2 \lesssim k^{-b}\left\|(\mathbf{SHS}^\top)^{-1}\mathbf{SH}_{k:\infty}\mathbf{S}^\top(\mathbf{SHS}^\top)^{-1}\right\| \le k^{-b}\|(\mathbf{SHS}^\top)^{-1}\| \lesssim M^a k^{-b}.$$

Hence,

$$\left\|\mathbb{E}_{\mathbf{w}^*}[\mathbf{v}^*(\mathbf{v}^*)^\top]\right\| = \left\|(\mathbf{SHS}^\top)^{-1}\mathbf{SHH}_w^{1/2}\right\|^2 \lesssim k^{(a-b)_+} + M^a k^{-b} \lesssim M^{(a-b)_+}.$$

The identity $\mathbf{v}_q^* = [\mathbf{I} - (\mathbf{H}_f^{(q)})^{-1}\mathbf{E}_q]\mathbf{v}^*$ and $\|(\mathbf{H}_f^{(q)})^{-1}\mathbf{E}_q\| \lesssim \epsilon_q M^a \lesssim 1$ therefore imply

$$\left\|\mathbb{E}_{\mathbf{w}^*}[\mathbf{v}_q^*(\mathbf{v}_q^*)^\top]\right\| \lesssim M^{(a-b)_+}.$$

Together with $(\mathbf{H}_f^{(q)})^{-1}\mathbf{SHS}^\top(\mathbf{H}_f^{(q)})^{-1} \preceq (\mathbf{H}_f^{(q)})^{-1}$ we have

$$\mathbb{E}_{\mathbf{w}^*} R_{\text{add}}^{(2)} \le \left\|\mathbb{E}_{\mathbf{w}^*}[\mathbf{v}^*(\mathbf{v}^*)^\top]\right\| \operatorname{tr}\left[\mathbf{E}_q(\mathbf{H}_f^{(q)})^{-1}\mathbf{E}_q\right] \lesssim M^{a+(a-b)_+} \operatorname{tr}(\mathbf{E}_q^2) \lesssim \epsilon_q^2 M^{a+(a-b)_+}.$$

Regarding $R_{\text{add}}^{(1)}$, the admissible step size also gives the contraction

$$\left\|\frac{\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H}_f^{(q)})^N}{N\gamma}(\mathbf{H}_f^{(q)})^{-1}\right\| = \left\|\frac{1}{N}\sum_{t=0}^{N-1}(\mathbf{I} - \gamma\mathbf{H}_f^{(q)})^t\right\| \le 1.$$

By cyclicity of trace and the operator-norm trace bound, the absolute value of the cross term is at most

$$\operatorname{tr}(\mathbf{E}_q)\left\|\mathbb{E}_{\mathbf{w}^*}[\mathbf{v}_q^*(\mathbf{v}_q^*)^\top]\right\| \lesssim \epsilon_q M^{(a-b)_+}.$$

Combining the two bounds and using $\epsilon_q M^a \lesssim 1$ proves the consequence. □

B.7 Population Risk and Critical Batch Size

Recall $\epsilon_q = \bar{\epsilon}_d + (1 + \bar{\epsilon}_d)(\bar{\epsilon}_s\bar{\epsilon}_f + \bar{\epsilon}_s + \bar{\epsilon}_f)$. Here we consider $\overline{\sigma}^2 \lesssim 1$.

Theorem B.21 (General population risk upper bounds)**.** *Under Assumption 3.1, 3.3, 3.4, 3.5, 3.6, with probability at least $1 - e^{-\Omega(M)}$, if the step size satisfies $\gamma \lesssim \frac{1}{\alpha \operatorname{tr}(\mathbf{H}_f^{(q)})}$, for any $i \in \{d, f, s, a, p, o\}$, quantization $\mathcal{Q}_i$ is $\bar{\epsilon}_i$-multiplicative,*

- *in the regime of $b < a + 1$,*

$$\mathbb{E}[\mathcal{R}_M(\overline{\mathbf{v}}_N)] \lesssim \sigma^2 + M^{1-b} + (N\gamma)^{-(b-1)/a} + \frac{\epsilon_q}{\epsilon_q + M^{-a}} + \left(\frac{1 + \bar{\epsilon}_a + \bar{\epsilon}_o}{B}\right)\min\left\{\frac{M}{N},\ \gamma^{1/a}N^{-(a-1)/a} + \gamma\epsilon_q\right\}$$
$$+\bar{\epsilon}_p(1 + \epsilon_q M^a)(\min\{M, (N\gamma)^{1/a}\})^{(a-b)_+}\max\left\{\frac{\min\left\{\frac{M}{N},\ \gamma^{1/a}N^{-(a-1)/a} + \gamma\epsilon_q\right\}}{B}, \frac{1}{N}\right\},$$

- *in the regime of $b = a + 1$,*

$$\mathbb{E}[\mathcal{R}_M(\overline{\mathbf{v}}_N)] \lesssim \sigma^2 + M^{-a}\log M + \frac{\log(N\gamma)}{N\gamma} + \frac{\epsilon_q}{\epsilon_q + M^{-a}} + \left(\frac{1 + \bar{\epsilon}_a + \bar{\epsilon}_o}{B}\right)\min\left\{\frac{M}{N},\ \gamma^{1/a}N^{-(a-1)/a} + \gamma\epsilon_q\right\}$$
$$+\bar{\epsilon}_p(1 + \epsilon_q M^a)\max\left\{\frac{\min\left\{\frac{M}{N},\ \gamma^{1/a}N^{-(a-1)/a} + \gamma\epsilon_q\right\}}{B}, \frac{1}{N}\right\},$$

- *in the regime of* $b > a+1$,

$$\mathbb{E}[\mathcal{R}_M(\overline{\mathbf{v}}_N)] \lesssim \sigma^2 + M^{-a} + \frac{1}{N\gamma} + \frac{\epsilon_q}{\epsilon_q + M^{-a}} + \left(\frac{1+\overline{\epsilon}_a+\overline{\epsilon}_o}{B}\right)\min\left\{\frac{M}{N},\ \gamma^{1/a}N^{-(a-1)/a} + \gamma\epsilon_q\right\}$$
$$+\overline{\epsilon}_p(1+\epsilon_q M^a)\max\left\{\frac{\min\left\{\frac{M}{N},\ \gamma^{1/a}N^{-(a-1)/a}+\gamma\epsilon_q\right\}}{B}, \frac{1}{N}\right\}.$$

Proof. Combine (4), (5), Lemma B.2, Lemma B.4, Lemma B.6, Lemma B.12, Lemma B.14, Lemma B.15, Lemma B.19, and

$$\min\left\{1, \max\left\{(N\gamma)^{-(b-1)/a}, M^{1-b}\right\} + \epsilon_q \min\{(N\gamma)^{1/a}, M\}^{(a-b)_+}\right\}$$
$$\lesssim \max\left\{(N\gamma)^{-(b-1)/a}, M^{1-b}\right\} + \frac{\epsilon_q}{\epsilon_q + M^{-a}}.$$

The conclusion for $b = a+1$ and $b > a+1$ are completed by Lemma B.16 and Lemma B.17 respectively. □

Theorem B.22 (General population risk upper bounds under small quantization)**.** *Under Assumption 3.1, 3.3, 3.4, 3.5, 3.6, with probability at least* $1 - e^{-\Omega(M)}$*, if the step size satisfies* $\gamma \lesssim \frac{1}{\alpha \mathrm{tr}(\mathbf{H}_f^{(q)})}$ *and* $\epsilon_q M^a = o(1)$*, for any* $i \in \{d, f, s, a, p, o\}$*, quantization* $\mathcal{Q}_i$ *is* $\overline{\epsilon}_i$*-multiplicative,*

- *in the regime of* $b < a+1$,

$$\mathbb{E}[\mathcal{R}_M(\overline{\mathbf{v}}_N)] \lesssim \sigma^2 + M^{1-b} + (N\gamma)^{-(b-1)/a} + \left(\frac{1+\overline{\epsilon}_a+\overline{\epsilon}_o}{B}\right)\min\left\{\frac{M}{N},\ \gamma^{1/a}N^{-(a-1)/a}\right\}$$
$$+\overline{\epsilon}_p(\min\{M, (N\gamma)^{1/a}\})^{(a-b)_+}\frac{\min\left\{\frac{M}{N},\ \gamma^{1/a}N^{-(a-1)/a}\right\}}{B},$$

- *in the regime of* $b = a+1$,

$$\mathbb{E}[\mathcal{R}_M(\overline{\mathbf{v}}_N)] \lesssim \sigma^2 + M^{-a}\log M + \frac{\log(N\gamma)}{N\gamma} + \left(\frac{1+\overline{\epsilon}_a+\overline{\epsilon}_o+\overline{\epsilon}_p}{B}\right)\min\left\{\frac{M}{N},\ \gamma^{1/a}N^{-(a-1)/a}\right\},$$

- *in the regime of* $b > a+1$,

$$\mathbb{E}[\mathcal{R}_M(\overline{\mathbf{v}}_N)] \lesssim \sigma^2 + M^{-a} + \frac{1}{N\gamma} + \left(\frac{1+\overline{\epsilon}_a+\overline{\epsilon}_o}{B}\right)\min\left\{\frac{M}{N},\ \gamma^{1/a}N^{-(a-1)/a}\right\}$$
$$+\overline{\epsilon}_p\max\left\{\frac{\min\left\{\frac{M}{N},\ \gamma^{1/a}N^{-(a-1)/a}\right\}}{B}, \frac{1}{N}\right\}.$$

Proof. Combine (4), (5), Lemma B.2, Lemma B.4, Lemma B.6, Lemma B.12, Lemma B.14, Lemma B.15, Lemma B.20, and

$$\epsilon_q \min\{(N\gamma)^{1/a}, M\}^{(a-b)_+} \le \epsilon_q M^{(a-b)_+} \le \begin{cases} o(M^{1-b}), & 1 < b < a+1, \\ o(M^{-a})\log M, & b = a+1, \\ o(M^{-a}), & b > a+1, \end{cases}$$

$$\min\left\{\frac{M}{N},\ \gamma^{1/a}N^{-(a-1)/a} + \gamma\epsilon_q\right\} \lesssim \min\left\{\frac{M}{N},\ \gamma^{1/a}N^{-(a-1)/a}\right\},$$

$$\frac{1}{N}\min\{M, (N\gamma)^{1/a}\}^{(a-b)_+} \lesssim \begin{cases} \frac{1}{N\gamma}, & 1 < a \le b < a+1, \\ (N\gamma)^{-b/a}, & 1 < b < a. \end{cases} = o\left((N\gamma)^{-(b-1)/a}\right).$$

The conclusion for $b = a+1$ and $b > a+1$ are completed by Lemma B.16 and Lemma B.17 respectively. □

Theorem B.23 (Tight population risk upper bounds in $b > a+1$ under overparameterization)**.** *Under Assumption 3.1, 3.3, 3.4, 3.5, 3.6, in the regime of $b > a+1$, with probability at least $1 - M^{-r} - e^{-\Omega(M)}$ for some constant $r > 0$, suppose that $(N\gamma)^{(\min\{b,2a+1\}-a)/a} \lesssim M/\log M$, if the step size satisfies $\gamma \lesssim \frac{1}{\alpha \mathrm{tr}(\mathbf{H}_f^{(q)})}$ and $\epsilon_q M^a = o(1)$, for any $i \in \{d, f, s, a, p, o\}$, quantization $\mathcal{Q}_i$ is $\overline{\epsilon}_i$-multiplicative,*

$$\mathbb{E}[\mathcal{R}_M(\overline{\mathbf{v}}_N)] \lesssim \sigma^2 + \widetilde{\Theta}\left((N\gamma)^{\frac{1-\min\{b,2a+1\}}{a}}\right) + \left(\frac{1+\overline{\epsilon}_a+\overline{\epsilon}_o+\overline{\epsilon}_p}{B}\right)\gamma^{1/a}N^{-(a-1)/a}.$$

Proof. By Lemma B.18,

$$\mathbb{E}_{\mathbf{w}^*}\left[\frac{1}{\gamma^2 N^2}\left\|\mathbf{v}_q^*\right\|^2_{(\mathbf{H}^{(q)}_{f,0:k^*})^{-1}} + \left\|\mathbf{v}_q^*\right\|^2_{\mathbf{H}^{(q)}_{f,k^*:\infty}}\right] \lesssim \widetilde{\Theta}\left((N\gamma)^{\frac{1-\min\{b,2a+1\}}{a}} + \epsilon_q\right) \wedge 1.$$

Combine (4), (5), Lemma B.2, Lemma B.4, Lemma B.6, Lemma B.13, Lemma B.14, Lemma B.20,

$$\begin{aligned}\mathbb{E}[\mathcal{R}_M(\overline{\mathbf{v}}_N)] \lesssim& \sigma^2 + \widetilde{\Theta}\left((N\gamma)^{\frac{1-\min\{b,2a+1\}}{a}}\right) + \epsilon_q \\ &+ \left(\frac{1+\overline{\epsilon}_a+\overline{\epsilon}_o+\overline{\epsilon}_p}{B}\right)\min\left\{\frac{M}{N},\ \gamma^{1/a}N^{-(a-1)/a} + \gamma\epsilon_q\right\} \\ &+ \frac{\log M}{M}\overline{\epsilon}_p(1+\epsilon_q M^a)\max\left\{\frac{\min\left\{\frac{M}{N},\ \gamma^{1/a}N^{-(a-1)/a}+\gamma\epsilon_q\right\}}{B}, \frac{1}{N}\right\}.\end{aligned}$$

The fact that

$$\epsilon_q \le o(M^{-a}) \lesssim o\left((N\gamma)^{a-\min\{b,2a+1\}}\right) \lesssim o\left((N\gamma)^{\frac{1-\min\{b,2a+1\}}{a}}\right),$$
$$M \gtrsim (N\gamma)^{(\min\{b,2a+1\}-a)/a} \gtrsim (N\gamma)^{1/a}$$

and

$$\frac{\log M/(MN)}{(N\gamma)^{\frac{1-\min\{b,2a+1\}}{a}}} \lesssim \frac{(N\gamma)^{(a-\min\{b,2a+1\})/a}}{(N\gamma)^{\frac{a+1-\min\{b,2a+1\}}{a}}} \lesssim (N\gamma)^{-1/a}$$

prove the theorem. □

B.7.1 Critical Batch Size

To obtain the critical batch size, we fix data budget $D = BN$ in the small quantization error regime $\epsilon_q M^a = o(1)$. To characterize the effects of quantization error, we retain the first-order quantization effects in leading orders.

Theorem B.24 (General critical batch size in $b \le a+1$)**.** *Fix $D = BN$. Under Assumption 3.1, 3.3, 3.4, 3.5, 3.6, with probability at least $1 - e^{-\Omega(M)}$, if the step size satisfies $\gamma \lesssim \frac{1}{\alpha \mathrm{tr}(\mathbf{H}_f^{(q)})}$, for any $i \in \{d, f, s, a, p, o\}$, quantization $\mathcal{Q}_i$ is $\overline{\epsilon}_i$-multiplicative, in the regime of $\epsilon_q M^a = o(1)$ and $\gamma \eqsim 1$,*

- *When $1 < b < a$,*
$$B^* \eqsim \max\left\{D\gamma M^{-a}, 1 + \gamma\overline{\epsilon}_p\right\}.$$
- *When $1 < b = a$,*
$$B^* \eqsim \max\left\{D\gamma M^{-a}, \gamma(1+\overline{\epsilon}_a+\overline{\epsilon}_o+\overline{\epsilon}_p)\right\}.$$
- *When $1 < a < b < a+1$,*
$$B^* \eqsim \max\left\{D\gamma M^{-a}, \gamma(1+\overline{\epsilon}_a+\overline{\epsilon}_o+\overline{\epsilon}_p)^{a/b}D^{(b-a)/b}\right\}.$$
- *When $1 < b = a+1$,*
$$B^* = \widetilde{\Theta}\left(\max\left\{D\gamma M^{-a}, \gamma(1+\overline{\epsilon}_a+\overline{\epsilon}_o+\overline{\epsilon}_p)^{\frac{a}{a+1}}D^{\frac{1}{a+1}}\right\}\right).$$

Proof. We prove from the scaling laws established in Theorem B.22 in case of small quantization regime, $\epsilon_q M^a = o(1)$ and $\gamma \eqsim 1$.

1. **Case** $1 < a < b < a + 1$

$$\mathbb{E}[\mathcal{R}_M(\overline{\mathbf{v}}_N)] \lesssim \sigma^2 + M^{1-b} + \left(\frac{D}{B}\gamma\right)^{\frac{1-b}{a}} + (1 + \bar{\epsilon}_a + \bar{\epsilon}_o + \bar{\epsilon}_p)\frac{\min\left\{M, \left(\frac{D}{B}\gamma\right)^{1/a}\right\}}{D}.$$

- If $M \lesssim \left(\frac{D}{B}\gamma\right)^{1/a}$, then $\left(\frac{D}{B}\gamma\right)^{\frac{1-b}{a}} \lesssim M^{1-b}$, and thus the risk

$$\mathbb{E}[\mathcal{R}_M(\overline{\mathbf{v}}_N)] - \sigma^2/2 \lesssim M^{1-b} + (1 + \bar{\epsilon}_a + \bar{\epsilon}_o + \bar{\epsilon}_p)\frac{M}{D}$$

  is independent of the batch size B. The critical batch size is therefore the largest batch size $B^* \eqsim D\gamma M^{-a}$.
- If $\left(\frac{D}{B}\gamma\right)^{1/a} \lesssim M$, then $M^{1-b} \lesssim \left(\frac{D}{B}\gamma\right)^{\frac{1-b}{a}}$, and thus the risk

$$\mathbb{E}[\mathcal{R}_M(\overline{\mathbf{v}}_N)] - \sigma^2/2 \lesssim \left(\frac{D}{B}\gamma\right)^{\frac{1-b}{a}} + (1 + \bar{\epsilon}_a + \bar{\epsilon}_o + \bar{\epsilon}_p)\frac{\left(\frac{D}{B}\gamma\right)^{\frac{1}{a}}}{D}.$$

  Trading the B-dependent risk off gives

$$B_0^* \eqsim \gamma(1 + \bar{\epsilon}_a + \bar{\epsilon}_o + \bar{\epsilon}_p)^{a/b} D^{(b-a)/b}.$$

Overall, the critical batch size is

$$B^* \eqsim \max\left\{D\gamma M^{-a}, \gamma(1 + \bar{\epsilon}_a + \bar{\epsilon}_o + \bar{\epsilon}_p)^{a/b} D^{(b-a)/b}\right\}.$$

2. **Case** $1 < b < a$

$$\mathbb{E}[\mathcal{R}_M(\overline{\mathbf{v}}_N)] \lesssim \sigma^2 + M^{1-b} + \left(\frac{D\gamma}{B}\right)^{\frac{1-b}{a}} + (1 + \bar{\epsilon}_a + \bar{\epsilon}_o)\frac{\min\left\{M, \left(\frac{D}{B}\gamma\right)^{1/a}\right\}}{D} + \bar{\epsilon}_p\frac{\min\left\{M^{a+1-b}, \left(\frac{D}{B}\gamma\right)^{\frac{a+1-b}{a}}\right\}}{D}.$$

- If $M \lesssim (\frac{D}{B}\gamma)^{1/a}$, then $\left(\frac{D}{B}\gamma\right)^{\frac{1-b}{a}} \lesssim M^{1-b}$, and thus the risk

$$\mathbb{E}[\mathcal{R}_M(\overline{\mathbf{v}}_N)] - \sigma^2/2 \lesssim M^{1-b} + (1 + \bar{\epsilon}_a + \bar{\epsilon}_o)\frac{M}{D} + \bar{\epsilon}_p\frac{M^{a+1-b}}{D}$$

  is independent of the batch size B. The critical batch size is therefore the largest batch size $B^* \eqsim D\gamma M^{-a}$.
- If $\left(\frac{D}{B}\gamma\right)^{1/a} \lesssim M$, then $M^{1-b} \lesssim \left(\frac{D}{B}\gamma\right)^{\frac{1-b}{a}}$, and thus the risk

$$\mathbb{E}[\mathcal{R}_M(\overline{\mathbf{v}}_N)] - \sigma^2/2 \lesssim \left(\frac{D}{B}\gamma\right)^{\frac{1-b}{a}} + (1 + \bar{\epsilon}_a + \bar{\epsilon}_o)\frac{\left(\frac{D}{B}\gamma\right)^{\frac{1}{a}}}{D} + \bar{\epsilon}_p\frac{\left(\frac{D}{B}\gamma\right)^{\frac{a+1-b}{a}}}{D}$$
$$\lesssim \left(\frac{D}{B}\gamma\right)^{\frac{1-b}{a}}\left(1 + \frac{\gamma\bar{\epsilon}_p}{B}\right).$$

  Hence, $B_0^* \eqsim \max\{1, \gamma\bar{\epsilon}_p\} \eqsim 1 + \gamma\bar{\epsilon}_p$.

Overall, the critical batch size is

$$B^* \eqsim \max\left\{D\gamma M^{-a}, 1 + \gamma\bar{\epsilon}_p\right\}.$$

3. **Case:** $1 < b = a$

$$\mathbb{E}[\mathcal{R}_M(\overline{\mathbf{v}}_N)] \lesssim \sigma^2 + M^{1-b} + \left(\frac{D}{B}\gamma\right)^{\frac{1-b}{a}} + (1 + \bar{\epsilon}_a + \bar{\epsilon}_o + \bar{\epsilon}_p)\frac{\min\left\{M, \left(\frac{D}{B}\gamma\right)^{1/a}\right\}}{D}.$$

- If $M \lesssim (\frac{D}{B}\gamma)^{1/a}$, then $\left(\frac{D}{B}\gamma\right)^{\frac{1-b}{a}} \lesssim M^{1-b}$, and thus the risk is independent of B, yielding $B^* \eqsim D\gamma M^{-a}$.
- If $\left(\frac{D}{B}\gamma\right)^{1/a} \lesssim M$, then $M^{1-b} \lesssim \left(\frac{D}{B}\gamma\right)^{\frac{1-b}{a}}$, and thus the risk

$$\begin{aligned}\mathbb{E}[\mathcal{R}_M(\overline{\mathbf{v}}_N)] - \sigma^2/2 &\lesssim \left(\frac{D}{B}\gamma\right)^{\frac{1-b}{a}} + (1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p)\,\frac{\left(\frac{D}{B}\gamma\right)^{\frac{1}{a}}}{D} \\ &\lesssim \left(\frac{D}{B}\gamma\right)^{\frac{1-b}{a}}\left(1+\frac{\gamma(1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p)}{B}\right).\end{aligned}$$

  Hence, $B_0^* \eqsim \gamma(1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p)$.

Overall, the critical batch size is

$$B^* \eqsim \max\left\{D\gamma M^{-a}, \gamma(1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p)\right\}.$$

4. **Case** $1 < b = a+1$

$$\mathbb{E}[\mathcal{R}_M(\overline{\mathbf{v}}_N)] \lesssim \sigma^2 + M^{-a}\log M + \frac{\log(D\gamma/B)}{D\gamma/B} + (1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p)\,\frac{\min\left\{M,\ \left(\frac{D\gamma}{B}\right)^{1/a}\right\}}{D}.$$

- If $M \lesssim \left(\frac{D\gamma}{B}\right)^{1/a}$, $B/(D\gamma) \lesssim M^{-a}$, then $\frac{\log(D\gamma/B)}{D\gamma/B} \lesssim M^{-a}\log M$, and thus the risk is independent of B, yielding $B^* \eqsim D\gamma M^{-a}$.
- If $\left(\frac{D\gamma}{B}\right)^{1/a} \lesssim M$, the risk satisfies

$$\mathbb{E}[\mathcal{R}_M(\overline{\mathbf{v}}_N)] \lesssim \sigma^2 + \frac{\log(D\gamma/B)}{D\gamma/B} + \frac{1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p}{D}\left(\frac{D\gamma}{B}\right)^{1/a}.$$

  Balancing the two B-dependent terms yields

$$\left(\frac{D\gamma}{B}\right)^{\frac{a+1}{a}} \eqsim \frac{D}{1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p}\log\left(\frac{D\gamma}{B}\right).$$

  Hence, asymptotically,

$$B_0^* \eqsim \gamma(1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p)^{\frac{a}{a+1}} D^{\frac{1}{a+1}} \times \left[\log\left(\frac{D}{1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p}\right)\right]^{-\frac{a}{a+1}}.$$

Overall,

$$B^* = \widetilde{\Theta}\left(\max\left\{D\gamma M^{-a}, \gamma(1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p)^{\frac{a}{a+1}} D^{\frac{1}{a+1}}\right\}\right).$$

□

Corollary B.25 (Critical batch size under overparameterization in $b \le a+1$)**.** *Fix $D = BN$. Under Assumption 3.1, 3.3, 3.4, 3.5, 3.6, with probability at least $1 - e^{-\Omega(M)}$, if the step size satisfies $\gamma \lesssim \frac{1}{\alpha\mathrm{tr}(\mathbf{H}_f^{(q)})}$, for any $i \in \{d, f, s, a, p, o\}$, quantization $\mathcal{Q}_i$ is $\bar{\epsilon}_i$-multiplicative, in the regime of $\epsilon_q M^a = o(1),\ D \lesssim M^a$ and $\gamma \eqsim 1$,*

- *When $1 < b < a$, $B^* \eqsim 1 + \gamma\bar{\epsilon}_p$.*
- *When $1 < a \le b \le a+1$, $B^* = \widetilde{\Theta}\left(\gamma(1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p)^{a/b} D^{(b-a)/b}\right)$.*

Theorem B.26 (Critical batch size under overparameterization in $b > a+1$)**.** *Fix $D = BN$. Under Assumption 3.1, 3.3, 3.4, 3.5, 3.6, in the regime of $b > a+1$, with probability at least $1 - M^{-r} - e^{-\Omega(M)}$ for some constant $r > 0$, if the step size satisfies $\gamma \lesssim \frac{1}{\alpha\mathrm{tr}(\mathbf{H}_f^{(q)})}$, for any $i \in \{d, f, s, a, p, o\}$, quantization $\mathcal{Q}_i$ is $\bar{\epsilon}_i$-multiplicative, in the regime of $\epsilon_q M^a = o(1)$, $D^{(\min\{b,2a+1\}-a)/a} \lesssim M/\log M$, it holds*

$$B^* = \widetilde{\Theta}\left(\gamma(1+\bar{\epsilon}_a+\bar{\epsilon}_o+\bar{\epsilon}_p)^{a/\min\{b,2a+1\}} D^{\frac{\min\{b,2a+1\}-a}{\min\{b,2a+1\}}}\right).$$

Proof. By Theorem B.23,

$$\mathbb{E}[\mathcal{R}_M(\overline{\mathbf{v}}_N)] - \sigma^2/2 \lesssim \widetilde{\Theta}\left(\left(\frac{D}{B}\gamma\right)^{\frac{1-\min\{b,2a+1\}}{a}}\right) + \left(\frac{1+\overline{\epsilon}_a+\overline{\epsilon}_o+\overline{\epsilon}_p}{D}\right)\left(\frac{D\gamma}{B}\right)^{1/a}.$$

Solving

$$B^* = \widetilde{\Theta}\left(\gamma(1+\overline{\epsilon}_a+\overline{\epsilon}_o+\overline{\epsilon}_p)^{a/\min\{b,2a+1\}} D^{\frac{\min\{b,2a+1\}-a}{\min\{b,2a+1\}}}\right).$$

□

C LLM PRE-TRAINING EXPERIMENTS

Setup. We compare BF16 and UFP4 pre-training on FineWeb-Edu-100B (Penedo et al., 2024), using the GPT-NeoX tokenizer (Black et al., 2022) with a vocabulary of 50,280 tokens. The final data shard is held out for validation. Our decoder-only Transformer has 24 layers, width 1,024, 16 attention heads, SwiGLU feed-forward layers with hidden width 4,096 (Shazeer, 2020), and a context length of 512. We use RoPE (Su et al., 2021), without dropout, QK normalization, or Transformer biases. Batch sizes below count global sequences; each update consumes $512B$ tokens.

Optimization. Following the constant-learning-rate and exponential weight averaging (EWA) protocol of Zhang et al. (2025), we use AdamW (Loshchilov & Hutter, 2019) with $\beta_1 = 0.95$ and zero weight decay. The learning rate increases linearly from zero over 300×2^{20} tokens and remains constant thereafter: 3.16×10^{-3} for BF16 and 1.00×10^{-3} for UFP4, with the latter chosen for stability. Both precisions share the batch-dependent coefficients in Table 1. EWA is initialized at zero, updated after each optimizer step, and debiased by dividing by $1 - \tau^t$ at step t. We evaluate its token-weighted validation loss at powers of two, every 1,000 updates, and at token-budget endpoints.

Table 1: Batch-dependent optimizer and EWA coefficients, shared across precisions.

B	256	512	1024	2048	4096	8192
β_2	.99	.99	.99	.95	.95	.95
τ	.998	.998	.998	.998	.995	.99

Low-precision computation. For FP4 training, we adopt UFP4 (Zhao et al., 2026) to mitigate the shrinkage bias induced by non-uniform E2M1 grids, which can accumulate across layers and degrade training stability. UFP4 instead uses a uniform grid, together with randomized 16-point Hadamard transforms and 1×16 blockwise absmax scaling. Forward inputs and weights use round-to-nearest-even, while output gradients use stochastic rounding (Gupta et al., 2015). Attention and the output head remain in BF16, while master weights, optimizer states, logits, and loss computations remain in FP32.

CBS estimation and scaling-law fit. We use the same five validation-loss targets for both precisions, $\{3.13, 3.14, 3.16, 3.17, 3.19\}$. Every target is reached by all six global batch sizes $B \in \{256, 512, 1024, 2048, 4096, 8192\}$. For a target ℓ and precision p, $S_p(B;\ell)$ is the first descending crossing in the recorded validation trajectory, with linear interpolation only inside the bracketing validation interval. The data budget for each target is defined by the batch size $B = 1024$, $D = S_p(1024;\ell)$ tokens; thus D is measured in actual tokens and is independent of the batch used in the stage fit.

Following the Zhang et al. (2025) protocol, we fit the stage relation in log-step space,

$$\log S_p(B;\ell) = \log\left(a_{p,\ell} + \frac{c_{p,\ell}}{B}\right), \qquad a_{p,\ell}, c_{p,\ell} > 0, \tag{24}$$

using all six batch sizes. The CBS is the batch size at which the fitted data cost is 20% above the smallest-batch reference. With the stage model in Eq. equation 24, this is

$$B_{\mathrm{CBS}} = 1.2B_{\min} + \frac{c_{p,\ell}}{5a_{p,\ell}}, \qquad B_{\min} = 256. \tag{25}$$

Table 2: Target-wise CBS estimates used in Fig. 2. D is in billions of actual training tokens and CBS is in 512-token sequences.

ℓ	$\mathrm{CBS}_{\mathrm{BF16}}$	$\mathrm{CBS}_{\mathrm{UFP4}}$
3.13	479.5	530.3
3.14	479.7	525.2
3.16	468.0	515.4
3.17	465.3	510.9
3.19	459.9	501.9

Finally, we fit the resulting target-wise points to

$$B_{\mathrm{CBS},p}(D) = K_p \left(\frac{D}{10^9 \text{ tokens}} \right)^{\beta_p}, \tag{26}$$

with ordinary least squares in $(\log D, \log B_{\mathrm{CBS}})$.

Interpretation. The fitted CBS scaling laws are $B_{\mathrm{CBS}}^{\mathrm{BF16}}(D) = 447.56 D^{0.4021}$ and $B_{\mathrm{CBS}}^{\mathrm{UFP4}}(D) = 494.78 D^{0.4020}$, where D is measured in billions of tokens. The nearly identical exponents suggest that precision leaves the data-scaling exponent largely unchanged, while UFP4 increases the multiplicative coefficient by 10.5%. Consistently, UFP4 yields a larger CBS at every matched target (about 502–530 versus 460–480 sequences). This agrees with our theoretical prediction that lower precision primarily shifts the scale of CBS upward rather than changing its data-budget exponent.