GENERALIZATION BEHAVIOR OF OPTQ AND THE ROLE OF REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large neural networks can be compressed by rounding or "quantizing" their weights to numbers that admit representations with fewer bits. One algorithm for quantization, OPTQ, progressively quantizes the weights of a neural network so that the squared quantization error on a specified calibration dataset is as small as possible. We study the performance of OPTQ and a variant algorithm, stochastic OPTQ, in a generalization setting and derive bounds for the expected squared error accrued by the algorithm when a test point is drawn from a fixed distribution. We prove two results. One result relates the generalization error to the error on a calibration dataset comprising independent samples from the same distribution as the test distribution. The other result bounds the generalization error of stochastic OPTQ for all sufficiently nice distributions, regardless of the calibration dataset. In both of these results, the regularization term λ plays an important role. We use insights from these results to make a new recommendation for the choice of λ and see that this choice of λ preforms favorably in experiments when compared to prior recommendations in the literature.

1 INTRODUCTION

Modern neural networks typically feature massive amounts of parameters. Early large language models (LLMs) used billions of parameters, such as GPT-3 (Brown et al., 2020) (175B) and LLaMa (Touvron et al., 2023) (7–65B). Recent AI models are even larger, with Deepseek-V4-Pro (DeepSeek-AI et al., 2026) having 1.6 trillion parameters for example. In turn, storing and using these models can require substantial amount of memory and compute time. Several paradigms have emerged recently to reduce these costs, including pruning, low-rank factorization, and knowledge distillation (Dantas et al., 2024). One technique of particular interest is to store the parameters of the neural network in representations that use fewer bits. This process is called *quantization*.

There are two major types of methods for quantization (Wei et al., 2024). One type, called quantization-aware training (QAT) modifies the optimization of neural networks to enable efficient quantization (Jacob et al., 2018; Zhang, 2025). The other major type of quantization algorithms is post-training quantization (PTQ). Methods for PTQ work on an already trained neural network and attempt to quantize the network with as minimal an impact on its performance as possible. In comparison to QAT methods, PTQ methods typically require less data and computational resources but are usually not able to achieve the same level of accuracy (Wei et al., 2024). A popular PTQ method that has emerged recently is OPTQ, also known as GPTQ (Frantar et al., 2023). OPTQ is able to successfully quantize many LLMs to 4-bit weights (Frantar et al., 2023; Jin et al., 2024) and is a component of more sophisticated quantization algorithms, such as QuIP (Chee et al., 2023) and Qronos (Zhang et al., 2026).

A line of work has been developed to analyze OPTQ from a theoretical perspective. Zhang et al. (2025) bound overall and coordinate-wise errors for OPTQ on the calibration dataset used to quantize a neural network model. Birnick (2026); Chen et al. (2026) establish connections between OPTQ and Babai's algorithm for the closest vector problem. Despite this progress, little is known about how OPTQ is able to generalize outside of the calibration dataset. This paper aims to address this problem by deriving explicit bounds for the generalization error of OPTQ in a variety of settings. We are motivated by the perspective that calibration error (the error on the dataset used to guide the quantization) and generalization error (the expected error on an unseen datapoint) are distinct

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.

quantities for data-dependent PTQ methods. In particular, OPTQ can make the error small on the calibration data by exploiting the empirical geometry of the sampled activations, but this empirical geometry may contain spurious correlations when the calibration set is small. Regularization plays an important role in our analysis, mainly to prevent this form of overfitting. The contributions of the paper should therefore be viewed in two layers. First, we prove a regularized empirical-to-population comparison inequality that is not intrinsically tied to OPTQ and can be applied to any data-dependent quantizer whose regularized empirical error is controlled. Second, we specialize this comparison to OPTQ using its update structure, and we prove a complementary positive-semidefinite quadratic-form bound for stochastic OPTQ.

1.1 NOTATION

Throughout this paper, we use the following notation. For a real number x, the quantity $\lfloor x \rfloor$ is the floor function, which maps its argument down to the largest interger not exceeding x. The abbreviation "i.i.d." stands for independent and identically distributed. For a random variable X, we use $\mathbb{E}_X$ and $\mathbb{P}_X$ to denote the expectation and probability over only the randomness in X. Boldfaced lowercase letters (e.g., $\boldsymbol{u}$) will denote vectors and boldfaced uppercase letters (e.g., $\mathbf{A}$) will denote matrices. For a vector $\boldsymbol{u}$, we use $\boldsymbol{u}_i$ to denote value of the i-th coordinate of $\boldsymbol{u}$. For a matrix $\mathbf{A}$, we use $\mathbf{A}_i$ to denote the i-th column of $\mathbf{A}$ (as a vector). We extend this indexing notation to allow taking multiple coordinates or rows at a time, for example we use $\mathbf{A}_{>i}$ to denote the submatrix formed by considering only the rows with index larger than i. The vector $\boldsymbol{e}^{(i)}$ is the i-th standard basis vector, which takes the value 1 and the i-th coordinate and 0 in all others. The matrix $\mathbf{I}$ is the identity matrix. We denote the psuedoinverse of a matrix $\mathbf{A}$ to be $\mathbf{A}^\dagger$. For vectors, $\|\cdot\|$ refers to the ℓ^2-norm and $\|\cdot\|_\infty$ refers to the ℓ^∞-norm. For matrices, $\|\cdot\|_{\mathrm{F}}$ refers to the Frobenius norm and $\|\cdot\|_{\mathrm{op}}$ refers to the $\ell^2 \to \ell^2$ operator norm. The function $\mathrm{tr}(\cdot)$ denotes the trace of its matrix argument. We use $\succeq$ to denote the Loewner order on positive semi definite matrices.

2 SETTING

Starting from a weight vector $\boldsymbol{w} \in \mathbb{R}^N$, the goal of post-training quantization algorithms is to obtain a quantized vector $\boldsymbol{q} \in \mathcal{A}^N$ that approximates $\boldsymbol{w}$. Here, $\mathcal{A}$ represents the quantization alphabet, which is a discrete set of real numbers. As an example, $\mathcal{A}$ can be the following set

$$\mathcal{A} = \{\delta k : k = -2^b, -2^b + 1, \ldots, 2^b - 2, 2^b - 1\} \tag{1}$$

which allows us to represent $\boldsymbol{q}$ as a vector of $(b+1)$-bit integers scaled by some value $\delta > 0$. Quantizing the real-valued vector $\boldsymbol{w}$ will then accrue errors in two ways: by rounding and by clipping. In this paper, we will only be interested in the error caused by the rounding process, so we will work with the alphabet $\mathcal{A} = \{\delta k : k \in \mathbb{Z}\}$. This is a model for an alphabet of the form given by Eq. (1) where b is chosen sufficiently large so that there is no clipping performed.

2.1 OPTQ

OPTQ is a data-dependent post-training quantization algorithm introduced by Frantar et al. (2023). Given a weight vector $\boldsymbol{w} \in \mathbb{R}^N$ and a calibration matrix $\mathbf{X} \in \mathbb{R}^{m \times N}$, OPTQ constructs a quantized vector $\boldsymbol{q}$ by sequentially quantizing the coordinates of $\boldsymbol{w}$ while attempting to keep the calibration error $\|\mathbf{X}\boldsymbol{w} - \mathbf{X}\boldsymbol{q}\|^2$ small. The key feature of OPTQ is that, after quantizing coordinate i, the algorithm updates the remaining unquantized coordinates $j > i$ to compensate for the error introduced at coordinate i.

We recall a mathematical formulation of this OPTQ update structure (Frantar et al., 2023; Zhang et al., 2026), first focusing on the unregularized version. OPTQ maintains a working vector, initialized as

$$\boldsymbol{w}^{(1)} = \boldsymbol{w}. \tag{2}$$

At step i, in the standard (deterministic) version of OPTQ, the current value of the i-th coordinate is quantized by setting

$$\boldsymbol{q}_i = \mathsf{Q}_\delta\left(\boldsymbol{w}_i^{(i)}\right), \tag{3}$$

where Q_δ denotes nearest-neighbor quantization onto the alphabet $\delta\mathbb{Z}$, i.e.,

$$\mathsf{Q}_\delta(t) \in \arg\min_{a\in\delta\mathbb{Z}} |t-a|. \tag{4}$$

The rounding error introduced at this step is

$$\alpha_i := \boldsymbol{q}_i - \boldsymbol{w}_i^{(i)}. \tag{5}$$

Thus, before any compensation is applied to obtain $\boldsymbol{w}^{(i+1)}$, the new contribution to the error vector $\boldsymbol{q}-\boldsymbol{w}$ is $\alpha_i \boldsymbol{e}^{(i)}$.

With the coordinates $1,\dots,i-1$ already quantized, OPTQ compensates for the error $\alpha_i\boldsymbol{e}^{(i)}$ using only the remaining coordinates $j>i$. If the tail compensation is written as $\alpha_i\mathbf{z}$ with $\mathbf{z}\in\mathbb{R}^{N-i}$, then the corresponding error to be minimized is $\left\|\mathbf{X}\left(\alpha_i\boldsymbol{e}^{(i)}+\alpha_i\begin{pmatrix}\mathbf{0}_i\\ \mathbf{z}\end{pmatrix}\right)\right\|^2$. After factoring out α_i^2, the optimization problem becomes

$$\min_{\mathbf{z}\in\mathbb{R}^{N-i}} \|\mathbf{X}_i + \mathbf{X}_{>i}\mathbf{z}\|^2. \tag{6}$$

Assuming for the moment that $\mathbf{X}$ has full column rank with $m>N$, the least-squares solution is

$$\mathbf{z} = -\mathbf{X}_{>i}^{\dagger}\mathbf{X}_i. \tag{7}$$

This gives the update direction

$$(\boldsymbol{v}^{(i)})_{<i} = 0, \qquad (\boldsymbol{v}^{(i)})_i = 1, \qquad (\boldsymbol{v}^{(i)})_{>i} = -\mathbf{X}_{>i}^{\dagger}\mathbf{X}_i. \tag{8}$$

Equivalently, $\boldsymbol{v}^{(i)}$ is characterized by the variational problem

$$\begin{aligned} \boldsymbol{v}^{(i)} := \arg\min_{\boldsymbol{v}\in\mathbb{R}^N} \quad & \|\mathbf{X}\boldsymbol{v}\|^2 \\ \text{subject to} \quad & \boldsymbol{v}_j = 0, \qquad j<i, \\ & \boldsymbol{v}_i = 1. \end{aligned} \tag{9}$$

The weight vector is then updated as

$$\boldsymbol{w}^{(i+1)} := \boldsymbol{w}^{(i)} + \alpha_i \boldsymbol{v}^{(i)}. \tag{10}$$

Since the coordinates $1,\dots,i$ are fixed after step i, after all N steps we obtain

$$\boldsymbol{r} := \boldsymbol{q}-\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i \boldsymbol{v}^{(i)}. \tag{11}$$

This representation of the OPTQ quantization error will be useful throughout the analysis.

We now introduce the regularization parameter λ. The procedure described above corresponds to OPTQ with $\lambda=0$. For $\lambda>0$, OPTQ is equivalent to running the same procedure with an augmented calibration matrix $\tilde{\mathbf{X}}$ in place of $\mathbf{X}$, where

$$\tilde{\mathbf{X}} := \begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda}\mathbf{I}\end{pmatrix}. \tag{12}$$

Note that $\tilde{\mathbf{X}}\in\mathbb{R}^{(m+N)\times N}$ is always full column rank. Therefore, with $\lambda>0$, the update directions are well-defined for any calibration matrix $\mathbf{X}$. In this case, the preceding formulas are interpreted with $\mathbf{X}$ replaced by $\tilde{\mathbf{X}}$. In particular,

$$(\boldsymbol{v}^{(i)})_{<i} = 0, \qquad (\boldsymbol{v}^{(i)})_i = 1, \qquad (\boldsymbol{v}^{(i)})_{>i} = -\tilde{\mathbf{X}}_{>i}^{\dagger}\tilde{\mathbf{X}}_i. \tag{13}$$

Let us now elaborate on how the coefficients α_i are chosen. For deterministic OPTQ, the i-th coordinate $\boldsymbol{w}_i^{(i)}$ is rounded to the nearest multiple of δ, and the corresponding residual is

$$\alpha_i := \delta\left\lfloor \frac{\boldsymbol{w}_i^{(i)}}{\delta} + \frac{1}{2}\right\rfloor - \boldsymbol{w}_i^{(i)}. \tag{14}$$

This choice gives $\boldsymbol{q}_i = \boldsymbol{w}_i^{(i)} + \alpha_i$ and $|\alpha_i| \le \delta/2$.

We will also consider a stochastic variant, studied by Chee et al. (2023); Zhang et al. (2025), which differs only in the resulting α_i. Let $p_i := \boldsymbol{w}_i^{(i)}/\delta - \lfloor \boldsymbol{w}_i^{(i)}/\delta \rfloor$ be the remainder when $\boldsymbol{w}_i^{(i)}$ is divided by δ. Stochastic OPTQ sets

$$\alpha_i := \begin{cases} -p_i\delta & \text{with probability } 1-p_i, \\ (1-p_i)\delta & \text{with probability } p_i. \end{cases} \tag{15}$$

With this choice, $\mathbb{E}[\alpha_i] = 0$ and $|\alpha_i| \le \delta$ almost surely. We will refer to the deterministic and stochastic variants collectively as OPTQ when the distinction is not important.

In what follows, the quantization error $\boldsymbol{r}$, update directions $\boldsymbol{v}^{(i)}$, and coefficients α_i will be central to our analysis. These quantities depend on the calibration matrix and on the regularization parameter through the augmented matrix $\tilde{\mathbf{X}}$. When this dependence is relevant, we will indicate it explicitly.

Finally, we note a geometric consequence of the variational characterization. Since $\boldsymbol{v}^{(i)}$ minimizes $\|\tilde{\mathbf{X}}\boldsymbol{v}\|^2$ among vectors whose first nonzero coordinate is a unit entry in position i, the vector $\tilde{\mathbf{X}}\boldsymbol{v}^{(i)}$ is the projection of $\tilde{\mathbf{X}}_i$ onto the orthogonal complement of the span of the columns of $\tilde{\mathbf{X}}_{>i}$. We denote this projection operator by $\mathbf{P}_{>i}^{\perp}$, so that $\mathbf{P}_{>i}^{\perp}\tilde{\mathbf{X}}_i = \tilde{\mathbf{X}}\boldsymbol{v}^{(i)}$.

3 Main results

We begin by isolating the step whereby we use an empirical result to obtain a population result from the particular quantization algorithm. The following theorem shows that, under regularization, the regularized empirical error controls the corresponding population error uniformly over all error vectors, including those that depend on the calibration data.

Theorem 3.1. *Suppose we sample m i.i.d. random samples $\boldsymbol{x}^{(i)}$ of a random vector $\boldsymbol{x} \in \mathbb{R}^N$ satisfying $\|\boldsymbol{x}\|_2 \le R$ almost surely. Denote the $m \times N$ matrix whose i-th row is ${\boldsymbol{x}^{(i)}}^{\mathsf{T}}$ by*

$$\mathbf{X} := \begin{pmatrix} \boldsymbol{x}^{(1)} & \cdots & \boldsymbol{x}^{(m)} \end{pmatrix}^{\mathsf{T}} \tag{16}$$

and let $\boldsymbol{\Sigma} := \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}]$. Fix $\epsilon \in (0,1)$ and assume that $\lambda > 2R^2\log(N/\epsilon)$. Then, with probability at least $1-\epsilon$ over the randomness of the sample $\mathbf{X}$, the following holds simultaneously for all $\boldsymbol{r} \in \mathbb{R}^N$:

$$\mathbb{E}_{\boldsymbol{x}}\left[|\boldsymbol{x}^{\mathsf{T}}\boldsymbol{r}|^2\right] \le \left(1 - \sqrt{\frac{2R^2\log(N/\epsilon)}{\lambda}}\right)^{-1} \frac{1}{m}\left(\|\mathbf{X}\boldsymbol{r}\|^2 + \lambda\|\boldsymbol{r}\|^2\right). \tag{17}$$

In particular, the conclusion applies to any possibly data-dependent vector $\boldsymbol{r} = \boldsymbol{r}(\mathbf{X})$.

The preceding theorem is independent of the particular quantization algorithm. It shows that any bound on the regularized empirical error $\|\tilde{\mathbf{X}}(\boldsymbol{q}-\boldsymbol{w})\|^2$ immediately yields a corresponding population error bound. We next illustrate the usefulness of this result for OPTQ by invoking OPTQ's bound for the augmented calibration matrix $\tilde{\mathbf{X}}$ (Zhang et al., 2025). Indeed, when applied to OPTQ, with $\boldsymbol{r} = \boldsymbol{q} - \boldsymbol{w}$, the term $\|\tilde{\mathbf{X}}\boldsymbol{r}\|^2/m$ that appears in the upper bound represents the average "training error" of OPTQ with an additional regularization term. That is,

$$\frac{1}{m}\|\mathbf{X}\boldsymbol{r}\|^2 + \frac{\lambda}{m}\|\boldsymbol{r}\|^2 = \left(\frac{1}{m}\sum_{i=1}^{N}|{\boldsymbol{x}^{(i)}}^{\mathsf{T}}\boldsymbol{w} - {\boldsymbol{x}^{(i)}}^{\mathsf{T}}\boldsymbol{q}|^2\right) + \frac{\lambda}{m}\|\boldsymbol{w}-\boldsymbol{q}\|^2. \tag{18}$$

This term is then multiplied by a factor which can be controlled by the choice of λ. Corollary 3.2, connecting this bound to the geometry of the calibration data, follows immediately by applying Eq. (3.5) of (Zhang et al., 2025).

Corollary 3.2. *Suppose the hypotheses of Theorem 3.1 hold. Let $\boldsymbol{q}$ be the output of OPTQ applied to $\boldsymbol{w}$ using the augmented calibration matrix $\tilde{\mathbf{X}} = \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I} \end{pmatrix}$, and let $\boldsymbol{r} := \boldsymbol{q} - \boldsymbol{w}$. Then, with probability at least $1-\epsilon$ over the randomness of the sample $\mathbf{X}$,*

$$\mathbb{E}_{\boldsymbol{x}}\left[|\boldsymbol{x}^{\mathsf{T}}\boldsymbol{w} - \boldsymbol{x}^{\mathsf{T}}\boldsymbol{q}|^2\right] \le \frac{C\delta^2}{m}\left(1 - \sqrt{\frac{2R^2\log(N/\epsilon)}{\lambda}}\right)^{-1}\sum_{i=1}^{N}\|\mathbf{P}_{>i}^{\perp}\tilde{\mathbf{X}}_i\|^2. \tag{19}$$

Here $C = 1/4$ for deterministic OPTQ and $C = 1$ for stochastic OPTQ.

The proofs of Theorem 3.1 and Corollary 3.2 are in Appendix A. These results demonstrate an important role of the regularization parameter. In the absence of the term $\lambda\|\boldsymbol{r}\|^2$, the empirical quadratic form $\|\mathbf{X}\boldsymbol{r}\|^2$ need not control $\boldsymbol{r}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{r}$ uniformly over data-dependent error vectors $\boldsymbol{r}$, since directions that are poorly represented (or absent) in the calibration sample may still have nonzero population variance. The regularization term makes every direction visible to the empirical objective, at the cost of introducing the additional contribution $\lambda\|\boldsymbol{r}\|^2$ into the bound.

We now proceed to state the second major result of this paper, which bounds arbitrary positive semi-definite quadratic forms of $\boldsymbol{r}$ when using stochastic OPTQ.

Theorem 3.3. *Let $\mathbf{A} \in \mathbb{R}^{(m+N)\times(m+N)}$ be an arbitrary positive semi-definite matrix. For stochastic OPTQ, we can bound*

$$\mathbb{E}_\alpha[(\tilde{\mathbf{X}}\boldsymbol{r})^\mathsf{T}\mathbf{A}\tilde{\mathbf{X}}\boldsymbol{r}] \leq \frac{\delta^2}{4}\left(\max_i \|\mathbf{P}_{>i}^\perp \tilde{\mathbf{X}}_i\|^2\right) \operatorname{tr}\mathbf{A}, \tag{20}$$

where $\mathbb{E}_\alpha[\cdot]$ denotes the expectation is taken only over the randomness in Eq. (15) *associated with choosing α, with the sample matrix $\mathbf{X}$ fixed.*

Furthermore, with probability at least $1-\epsilon$ over the randomness in α_i, we have

$$(\tilde{\mathbf{X}}\boldsymbol{r})^\mathsf{T}\mathbf{A}\tilde{\mathbf{X}}\boldsymbol{r} \leq 2\delta^2\log(2/\epsilon)\left(\max_i \|\mathbf{P}_{>i}^\perp \tilde{\mathbf{X}}_i\|^2\right) \operatorname{tr}\mathbf{A}. \tag{21}$$

This theorem is proven in Appendix B. Theorem 3.3 holds very broadly and has a variety of applications. An important application for this work is as follows.

Corollary 3.4. *Let $\boldsymbol{z} \in \mathbb{R}^N$ be a random vector with finite second moment. Then, when stochastic OPTQ with $\lambda > 0$ is run with a data matrix $\mathbf{X}$ we have*

$$\mathbb{E}_\alpha\mathbb{E}_{\boldsymbol{z}}[|\boldsymbol{z}^\mathsf{T}\boldsymbol{w} - \boldsymbol{z}^\mathsf{T}\boldsymbol{q}|^2] \leq \frac{\delta^2}{4}\left(\frac{\max_i \|\mathbf{P}_{>i}^\perp \tilde{\mathbf{X}}_i\|^2}{\lambda}\right)\mathbb{E}_{\boldsymbol{z}}\left[\|\boldsymbol{z}\|^2\right], \tag{22}$$

where $\mathbb{E}_{\boldsymbol{z}}$ denotes the expectation taken over $\boldsymbol{z}$. With probability at least $1-\epsilon$ over the randomness of α_i, we also have

$$\mathbb{E}_{\boldsymbol{z}}[|\boldsymbol{z}^\mathsf{T}\boldsymbol{w} - \boldsymbol{z}^\mathsf{T}\boldsymbol{q}|^2] \leq 2\delta^2\log(2/\epsilon)\left(\frac{\max_i \|\mathbf{P}_{>i}^\perp \tilde{\mathbf{X}}_i\|^2}{\lambda}\right)\mathbb{E}_{\boldsymbol{z}}\left[\|\boldsymbol{z}\|^2\right]. \tag{23}$$

Proof. Apply Theorem 3.3 with

$$\mathbf{A} = (\tilde{\mathbf{X}}^\dagger)^\mathsf{T}(\mathbb{E}_{\boldsymbol{z}}\boldsymbol{z}\boldsymbol{z}^\mathsf{T})\tilde{\mathbf{X}}^\dagger. \tag{24}$$

Since $\lambda > 0$, it follows that $\tilde{\mathbf{X}}$ has full column rank, so $\tilde{\mathbf{X}}^\dagger\tilde{\mathbf{X}} = \mathbf{I}$. Consequently,

$$(\tilde{\mathbf{X}}\boldsymbol{r})^\mathsf{T}\mathbf{A}(\tilde{\mathbf{X}}\boldsymbol{r}) = \boldsymbol{r}^\mathsf{T}(\mathbb{E}_{\boldsymbol{z}}\boldsymbol{z}\boldsymbol{z}^\mathsf{T})\boldsymbol{r} = \mathbb{E}_{\boldsymbol{z}}|\boldsymbol{z}^\mathsf{T}\boldsymbol{r}|^2. \tag{25}$$

Moreover,

$$\operatorname{tr}(\mathbf{A}) = \operatorname{tr}\left((\mathbb{E}_{\boldsymbol{z}}\boldsymbol{z}\boldsymbol{z}^\mathsf{T})(\tilde{\mathbf{X}}^\mathsf{T}\tilde{\mathbf{X}})^{-1}\right) \leq \frac{1}{\lambda}\mathbb{E}_{\boldsymbol{z}}\|\boldsymbol{z}\|^2, \tag{26}$$

because $\tilde{\mathbf{X}}^\mathsf{T}\tilde{\mathbf{X}} = \mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I} \succeq \lambda\mathbf{I}$. □

Note that this bound holds for each fixed calibration matrix $\mathbf{X}$, and the only probability in the second part of the corollary is the stochastic rounding probability. It is therefore distribution-agnostic in $\boldsymbol{z}$, even when the distribution of $\boldsymbol{z}$ differs from the distribution that generated $\mathbf{X}$. If we instead used memoryless scalar quantization (MSQ), which is often also referred to as round-to-nearest (RTN), we generically would expect

$$\mathbb{E}_{\boldsymbol{z}}[|\boldsymbol{z}^\mathsf{T}\boldsymbol{w} - \boldsymbol{z}^\mathsf{T}\boldsymbol{q}|^2] \gtrsim \delta^2\mathbb{E}_{\boldsymbol{z}}[\|\boldsymbol{z}\|^2]. \tag{27}$$

Therefore, using the data-dependent algorithm stochastic OPTQ with sufficiently large choice of λ will produce a quantized vector that performs only a bounded factor worse in the worst case compared to a data-independent algorithm. In Appendix C, we make the statement in Eq. (27) precise.

Another application of Theorem 3.3 is the following, which improves a result in (Zhang et al., 2025, Theorem 4.6) by a factor of $\sqrt{\pi}$.

Corollary 3.5. *When stochastic OPTQ with $\lambda > 0$ is run with a data matrix $\mathbf{X}$, with probability at least $1 - \epsilon$ over the randomness of α_i we have*

$$\|\boldsymbol{w} - \boldsymbol{q}\|_\infty \leq \delta\sqrt{2\log(2(m+N)/\epsilon)}\left(\frac{\max_i \|\mathbf{P}_{>i}^{\perp}\tilde{\mathbf{X}}_i\|}{\sqrt{\lambda}}\right). \tag{28}$$

Proof. Apply Eq. (21) of Theorem 3.3 to the $(m+N)\times(m+N)$ coordinate projectors $\boldsymbol{e}^{(1)}\boldsymbol{e}^{(1)\mathsf{T}}, \ldots, \boldsymbol{e}^{(m)}\boldsymbol{e}^{(m)\mathsf{T}}$ and use a union bound over the failure probabilities. This shows, with probability at least $1-\epsilon$, we have

$$\|\tilde{\mathbf{X}}\boldsymbol{w} - \tilde{\mathbf{X}}\boldsymbol{q}\|_\infty \leq \delta\sqrt{2\log(2(m+N)/\epsilon)}\left(\max_i \|\mathbf{P}_{>i}^{\perp}\tilde{\mathbf{X}}_i\|\right). \tag{29}$$

The desired result follows by recognizing the last N entries of $\tilde{\mathbf{X}}\boldsymbol{w} - \tilde{\mathbf{X}}\boldsymbol{q}$ as $\sqrt{\lambda}(\boldsymbol{w}-\boldsymbol{q})$. □

The utility of this result is that it bounds above the ℓ^∞-norm of $\boldsymbol{q}$ based on λ, $\mathbf{X}$, and $\boldsymbol{w}$. This allows us to use the alphabet in Eq. (1) and determine an appropriate value of b so that with high probability no clipping will need to be performed.

4 Choosing λ

4.1 Theory

Corollaries 3.2 and 3.4 bound the generalization error of OPTQ in two different ways, but both bounds feature a dependence on λ. Suppose each row of $\mathbf{X}$ is drawn i.i.d. from a distribution with compact support in the ball of radius of R. If the following two inequalities hold

$$\lambda \gtrsim R^2\log(N), \tag{30}$$

$$\lambda \gtrsim \max_i \|\mathbf{X}_i\|^2, \tag{31}$$

then with high probability we can bound for a random vector $\boldsymbol{z}$

$$\mathbb{E}_{\boldsymbol{z}}[|\boldsymbol{z}^\mathsf{T}\boldsymbol{w} - \boldsymbol{z}^\mathsf{T}\boldsymbol{q}|^2] \lesssim \frac{\|\tilde{\mathbf{X}}\boldsymbol{r}\|^2}{m} \tag{32}$$

whenever $\boldsymbol{z}$ is drawn from the same distribution that each row of $\mathbf{X}$ was and

$$\mathbb{E}_{\boldsymbol{z}}[|\boldsymbol{z}^\mathsf{T}\boldsymbol{w} - \boldsymbol{z}^\mathsf{T}\boldsymbol{q}|^2] \lesssim \delta^2\mathbb{E}_{\boldsymbol{z}}\left[\|\boldsymbol{z}\|^2\right]. \tag{33}$$

always. In general, we won't know the value of R in Eq. (30). However, a lower bound for R is the maximum row norm of $\mathbf{X}$. Therefore, if we wish to have both bounds Eqs. (32) and (33) we should ensure

$$\lambda \gtrsim \max\left\{\max_i \|\mathbf{X}_i\|^2, \max_i \|(\mathbf{X}^\mathsf{T})_i\|^2\log(N)\right\}. \tag{34}$$

We now turn our attention to Eq. (19) of Corollary 3.2. This bound depends on λ in two ways: through the regularized training error term $\|\tilde{\mathbf{X}}\boldsymbol{r}\|^2/m = (\|\mathbf{X}\boldsymbol{r}\|^2 + \lambda\|\boldsymbol{r}\|^2)/m$ and the multiplicative factor in front of it. We consider how to choose λ to minimize this bound.

Remark 4.1. *To minimize the bound in Eq.* (17)*, a good choice of λ is*

$$\lambda = \beta \cdot \frac{\|\mathbf{X}\|_F^2}{m^{1/3}\min\{m,N\}^{2/3}}, \tag{35}$$

where β is a scaling factor to be determined empirically.

We proceed to justify this remark. From Corollary 3.2, for $\lambda > 2R^2\log(N/\epsilon)$ we have

$$\mathbb{E}\left[|\boldsymbol{x}^\mathsf{T}\boldsymbol{w} - \boldsymbol{x}^\mathsf{T}\boldsymbol{q}|^2\right] \leq \frac{C\delta^2}{m}\left(1 - \sqrt{\frac{2R^2\log(N/\epsilon)}{\lambda}}\right)^{-1}\sum_{i=1}^{N}\|\tilde{\mathbf{X}}\boldsymbol{v}^{(i)}(\lambda)\|^2. \tag{36}$$

Here $C = 1$ for stochastic OPTQ and $C = 1/4$ for deterministic OPTQ, corresponding respectively to $|\alpha_i|^2 \leq \delta^2$ and $|\alpha_i|^2 \leq \delta^2/4$. We will attempt to minimize the bound in Eq. (36). Since the bound in Eq. (36) is strictly positive, minimizing it is equivalent to minimizing

$$f(\lambda) = \log\left(\sum_{i=1}^{N} \|\tilde{\mathbf{X}}\boldsymbol{v}^{(i)}(\lambda)\|^2\right) - \log\left(1 - \sqrt{\frac{2R^2\log(N/\epsilon)}{\lambda}}\right). \tag{37}$$

For each i, the function $g_i : \lambda \mapsto \|\tilde{\mathbf{X}}\boldsymbol{v}^{(i)}(\lambda)\|^2$ is differentiable for $\lambda > 0$, as can be seen by standard formulas for matrix-vector calculus. However, this function also has the following representation:

$$g_i(\lambda) = \min_{\boldsymbol{v} \in V_i} \left[\|\mathbf{X}\boldsymbol{v}\|^2 + \lambda\|\boldsymbol{v}\|^2\right], \tag{38}$$

where V_i is the affine subspace $V_i = \boldsymbol{e}^{(i)} + \mathrm{span}\{\boldsymbol{e}^{(j)} : i < j \leq N\}$. The minimum in Eq. (38) is attained at exactly one point $\boldsymbol{v} = \boldsymbol{v}^{(i)}(\lambda)$. In particular, this means that each function $g_i(\lambda)$ has derivative equal to $\|\boldsymbol{v}^{(i)}(\lambda)\|^2$. We therefore see that

$$f'(\lambda) = \frac{\sum_{i=1}^{N} \|\boldsymbol{v}^{(i)}(\lambda)\|^2}{\sum_{i=1}^{N} \|\tilde{\mathbf{X}}\boldsymbol{v}^{(i)}(\lambda)\|^2} - \frac{\sqrt{R^2\log(N/\epsilon)}}{\sqrt{2\lambda^3}\left(1 - \sqrt{\frac{2R^2\log(N/\epsilon)}{\lambda}}\right)}. \tag{39}$$

The first term of Eq. (39) is at most $1/\lambda$. If we let $\lambda = \alpha R^2\log(N/\epsilon)$. We can then bound

$$f'(\lambda) \leq \frac{1}{\lambda}\left(1 - \frac{1}{\sqrt{2\alpha} - 2}\right). \tag{40}$$

Thus $f'(\lambda) \leq 0$ whenever $\lambda \leq 4.5R^2\log(N/\epsilon)$.

On the other hand, the first term of Eq. (39) is at least $N/(N\lambda + \|\mathbf{X}\|_{\mathrm{F}}^2)$. Therefore, we can bound

$$f'(\lambda) \geq \frac{N\sqrt{2\lambda^3}\left(1 - \sqrt{\frac{2R^2\log(N/\epsilon)}{\lambda}}\right) - \sqrt{R^2\log(N/\epsilon)}\left(N\lambda + \|\mathbf{X}\|_{\mathrm{F}}^2\right)}{\left(N\lambda + \|\mathbf{X}\|_{\mathrm{F}}^2\right)\sqrt{2\lambda^3}\left(1 - \sqrt{\frac{2R^2\log(N/\epsilon)}{\lambda}}\right)} \tag{41}$$

$$= \frac{N\sqrt{2\lambda^3} - (3N\lambda + \|\mathbf{X}\|_{\mathrm{F}}^2)\sqrt{R^2\log(N/\epsilon)}}{\left(N\lambda + \|\mathbf{X}\|_{\mathrm{F}}^2\right)\sqrt{2\lambda^3}\left(1 - \sqrt{\frac{2R^2\log(N/\epsilon)}{\lambda}}\right)} \tag{42}$$

and so $f'(\lambda) \geq 0$ whenever

$$\lambda \geq \left(\sqrt{\frac{R^2\log(N/\epsilon)}{2}} \cdot \frac{3N\lambda + \|\mathbf{X}\|_{\mathrm{F}}^2}{N}\right)^{2/3}. \tag{43}$$

In particular, $f'(\lambda) \geq 0$ whenever

$$\lambda \geq \left(\sqrt{2R^2\log(N/\epsilon)}\max\left\{3\lambda, \frac{\|\mathbf{X}\|_{\mathrm{F}}^2}{N}\right\}\right)^{2/3}. \tag{44}$$

This means that the value λ_* of λ that minimizes $f(\lambda)$ satisfies

$$4.5R^2\log(N/\epsilon) \leq \lambda_* \leq \max\left\{18R^2\log(N/\epsilon), \left(\frac{\sqrt{2R^2\log(N/\epsilon)}\|\mathbf{X}\|_{\mathrm{F}}^2}{N}\right)^{2/3}\right\}. \tag{45}$$

In the regime where each row of $\mathbf{X}$ is an i.i.d. sample from the same distribution as a generic test point, we should choose λ satisfying Eq. (45). We now proceed to simplify this bound. Let K be such that $R^2 = K^2\mathbb{E}[\|\boldsymbol{z}\|^2]$. For distributions for which the empirical second moment is representative of its expectation, $\|\mathbf{X}\|_{\mathrm{F}}^2/m \approx \mathbb{E}\|\boldsymbol{z}\|^2$, and hence $R^2 \approx K^2\|\mathbf{X}\|_{\mathrm{F}}^2/m$. This replacement is, of course, heuristic unless accompanied by an explicit concentration assumption on the row norms. With this definition of K, Eq. (45) becomes

$$\frac{\|\mathbf{X}\|_{\mathrm{F}}^2}{m} \lesssim_{K,\log} \lambda_* \lesssim_{K,\log} \frac{\|\mathbf{X}\|_{\mathrm{F}}^2}{m^{1/3}\min\{m, N\}^{2/3}}, \tag{46}$$

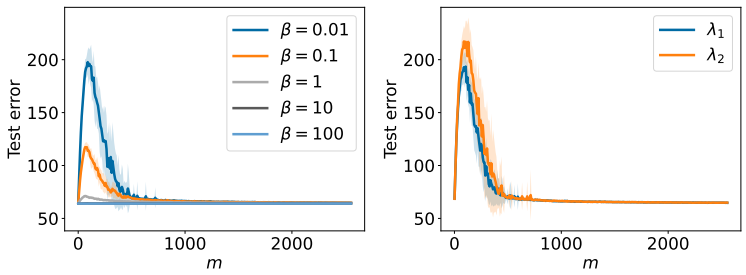


Figure 1: Test error versus m for various choices of λ when OPTQ is run with samples from the uniform Hadamard distribution.

where $\lesssim_{K,\log}$ hides the dependence on K and $\log$ factors. In light of Corollary 3.4, it is sensible to use the upper bound as a conservative estimate for λ. This leads to the scaling in Eq. (35). While this argument is not precise enough to determine an optimal scaling constant, in the next subsection we will empirically determine a good choice of constant.

4.2 Experiments

We now demonstrate how the recommendations from Section 4.1 work in practice. To do this, we run the output of OPTQ with varying choices of λ for data drawn from three different distributions and show the generalization error of the result.[1]

We consider a total of seven different choices of λ for OPTQ. Five of these choices correspond to Eq. (35) with different values of β. The remaining thwo are recommendations previously made in the literature. These are $\lambda_1 = 0.01\|\mathbf{X}\|_{\mathrm{F}}^2/N$ from Frantar et al. (2023) and $\lambda_2 = 10^{-3} \cdot \|\mathbf{X}\|_{\mathrm{op}}$ from Zhang et al. (2026).[2]

In all experiments, we sample the rows of the $m \times N$ data matrix $\mathbf{X}$ i.i.d. from the same distribution as a generic test point $\boldsymbol{z}$ and vary the number of rows m. We simplify the measures of the generalization error by reporting the value of

$$E = \sum_{i=1}^{N} \mathbb{E}[|\boldsymbol{z}^{\mathsf{T}} \boldsymbol{v}^{(i)}|^2] \tag{47}$$

as the test error. For stochastic OPTQ, this expression is precisely $\mathbb{E}[|\boldsymbol{z}^{\mathsf{T}} \boldsymbol{r}|^2]$ where the contribution from the α_i is ignored (see Eqs. (70) and (74) from the proof of Theorem 3.3). The contribution of each α_i depends on both the randomness of stochastic OPTQ and the individual quantization error for each coordinate. Up to constant factors of δ, Eq. (47) can be regarded as both a worst-case for the generalization error and an average-case for the generalization error (over the randomness of α_i and the initial position of $\boldsymbol{w}$). We approximate the expectations in Eq. (47) with 1000 samples of $\boldsymbol{z}$. For all the plots in this section, the mean and standard deviation over 30 trials (each an independent sample of $\mathbf{X}$) is shown.

Uniform Hadamard The first data distribution we consider is the uniform distribution on rows from Sylvester's construction of a 64×64 Hadamard matrix. In the full data distribution, each coordinate is independent of the others. However, in a low sampling regime, many spurious correlations between the columns of $\mathbf{X}$ can appear. As such, OPTQ performs poorly when improperly regularized. The results of these experiments are shown in Fig. 1.

Non-uniform Hadamard The second data distribution we consider again draws rows from a 64×64 Hadamard matrix, but with a non-uniform distribution on the rows. The probability of a row i being drawn is proportional to $1/i$. Similar to the uniform Hadamard distribution, this data distribution results in many spurious correlations appearing in a low sampling regime. However, the

[1] See [link removed for double-blind review] for the code used for these experiments.

[2] The recommendation λ_2 is for Qronos, but the effect of λ on OPTQ is identical.

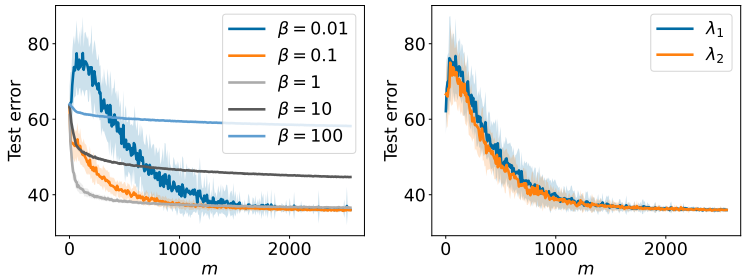


Figure 2: Test error versus m for various choices of λ when OPTQ is run with samples from the non-uniform Hadamard distribution.

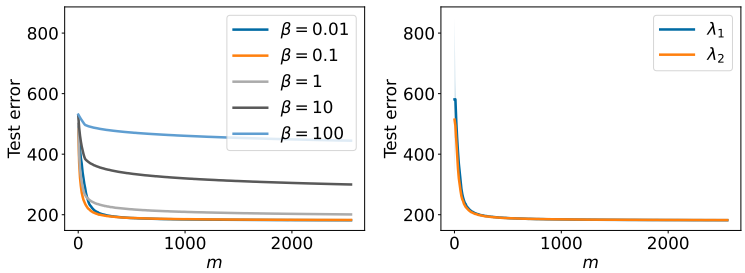


Figure 3: Test error versus m for various choices of λ when OPTQ is run with samples from the ReLU neural network distribution.

true data distribution does have a low-rank structure exploitable by OPTQ to reduce the error. The results of these experiments are shown in Fig. 2.

ReLU neural network The last data distribution we consider is a model of a ReLU neural network. First, we fix a 64×16 matrix $\mathbf{W}$ where each coordinate is sampled i.i.d. from a standard normal distribution. We generate a random vector $\boldsymbol{z} \in \mathbb{R}^{64}$ by first sampling a random vector $\boldsymbol{v} \in \mathbb{R}^{16}$ with i.i.d. standard normal coordinates and then applying the ReLU function[3] coordinate-wise to $\mathbf{W}\boldsymbol{v}$. This is a model of what the data distribution of the output of a single neural network layer may look like. Here, the low-rank structure comes from the overparameterization of the neural network layer and the positivity enforced by the ReLU function. The results of these experiments are in Fig. 3.

Across Figs. 1 to 3, a clear trend emerges. The existing recommendations for λ both do well in data distributions with low-rank structure in a well-sampled regime, but are susceptible to producing large errors in a low sampling regime. However, choosing λ according to Eq. (35) can result in the same error being achieved in the well-sampled regime without the large error present in the low sampling regime. Across all the experiments, choosing $\beta = 1$ is able to substantially reduce the error in Figs. 1 and 2 with low m while still achieving small values for the error in Figs. 2 and 3 for large m.

5 Conclusion

In this work, we bound the generalization error of OPTQ in two different settings: when the input data is drawn i.i.d. from the same distribution as a test point and when randomized rounding is used as the quantization procedure for OPTQ. Using these results, we derive a formula for choosing the quantization parameter λ that outperforms previous recommendations in a low-sampling regime while still achieving similarly good generalization performance in a high-sampling regime.

[3]The ReLU function ϕ is defined as $\phi(x) = \max\{x, 0\}$.

AI DISCLOSURE

In this work, we used generative AI tools to find references in the literature and to edit the paper. All AI-provided references and edits were then reviewed by the authors. We did not use generative AI tools for any other task.

REFERENCES

Johann Birnick. The lattice geometry of neural network quantization: A short equivalence proof of GPTQ and Babai's algorithm, 2026. URL https://arxiv.org/abs/2508.01077.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. Quip: 2-bit quantization of large language models with guarantees. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 4396–4429. Curran Associates, Inc., 2023. doi: 10.52202/075280-0196. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/0df38cd13520747e1e64e5b123a78ef8-Paper-Conference.pdf.

Jiale Chen, Yalda Shabanzadeh, Elvir Crnčević, Torsten Hoefler, and Dan Alistarh. The geometry of LLM quantization: GPTQ as Babai's nearest plane algorithm. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=NFB4QGGS65.

Pierre Vilar Dantas, Waldir Sabino da Silva, Lucas Carvalho Cordeiro, and Celso Barbosa Carvalho. A comprehensive review of model compression techniques in machine learning. *Applied Intelligence*, 54(22):11804–11844, November 2024. ISSN 1573-7497. doi: 10.1007/s10489-024-05747-w. URL https://doi.org/10.1007/s10489-024-05747-w.

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Donghao Li, Dongjie Ji, Erhang Li, Fang Wei, Fangyun Lin, Fangzhou Yuan, Feiyu Xia, Fucong Dai, et al. DeepSeek-V4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tcbBPnfwxS.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2704–2713, 2018. doi: 10.1109/CVPR.2018.00286.

Renren Jin, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. A comprehensive evaluation of quantization strategies for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12186–12215, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.726. URL https://aclanthology.org/2024.findings-acl.726/.

Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994. ISSN 00911798, 2168894X. URL http://www.jstor.org/stable/2244912.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(82):1–9, 2013.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, August 2012. ISSN 1615-3383. doi: 10.1007/s10208-011-9099-z. URL https://doi.org/10.1007/s10208-011-9099-z.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2nd edition, 2025. doi: 10.1017/9781108231596. URL https://doi.org/10.1017/9781108231596.

Lu Wei, Zhong Ma, Chaojie Yang, and Qin Yao. Advances in the neural network quantization: A comprehensive review. *Applied Sciences*, 14(17), 2024. ISSN 2076-3417. doi: 10.3390/app14177445. URL https://www.mdpi.com/2076-3417/14/17/7445.

Haoyu Zhang, Shihao Zhang, Ian Colbert, and Rayan Saab. Provable post-training quantization: Theoretical analysis of OPTQ and Qronos, 2025. URL https://arxiv.org/abs/2508.04853.

Jiaqi Zhang. Survey of quantization-aware training (qat) applications in deep learning quantization. In *Proceedings of the 2025 International Symposium on Artificial Intelligence and Computational Social Sciences*, AICSS '25, pp. 431–442, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400721007. doi: 10.1145/3776759.3776826. URL https://doi.org/10.1145/3776759.3776826.

Shihao Zhang, Haoyu Zhang, Ian Colbert, and Rayan Saab. Qronos: Correcting the past by shaping the future... in post-training quantization. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 87775–87798, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/8d22a09f18d9fc85dcbe6f482e13e1d7-Paper-Conference.pdf.

A Proof of Theorem 3.1 and Corollary 3.2

To prove Theorem 3.1, we will use the following result.

Theorem A.1 (Matrix Chernoff inequality (Tropp, 2012, Remark 5.3))**.** *Consider a finite sequence* $\{\mathbf{A}^{(k)}\}$ *of independent, random, self-adjoint matrices with dimension* d*. Assume that each random matrix is positive semidefinite and satisfies* $\|\mathbf{A}^{(k)}\|_{op} \leq R$ *almost surely.*

Let $\sigma_{\min}(\cdot)$ *be the smallest eigenvalue of its matrix argument and define*

$$\mu_{\min} := \sigma_{\min}\left(\sum_k \mathbb{E}[\mathbf{A}^{(k)}]\right). \tag{48}$$

Then for $t \in [0, 1]$*,*

$$\mathbb{P}\left[\sigma_{\min}\left(\sum_k \mathbf{A}^{(k)}\right) \leq (1-t)\mu_{\min}\right] \leq d\exp\left(-\frac{t^2\mu_{\min}}{2R}\right). \tag{49}$$

We now proceed to prove Theorem 3.1.

Theorem 3.1. *Suppose we sample m i.i.d. random samples $\boldsymbol{x}^{(i)}$ of a random vector $\boldsymbol{x} \in \mathbb{R}^N$ satisfying $\|\boldsymbol{x}\|_2 \le R$ almost surely. Denote the $m \times N$ matrix whose i-th row is ${\boldsymbol{x}^{(i)}}^\mathsf{T}$ by*

$$\mathbf{X} := \begin{pmatrix} \boldsymbol{x}^{(1)} & \cdots & \boldsymbol{x}^{(m)} \end{pmatrix}^\mathsf{T} \tag{16}$$

and let $\boldsymbol{\Sigma} := \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^\mathsf{T}]$. Fix $\epsilon \in (0,1)$ and assume that $\lambda > 2R^2 \log(N/\epsilon)$. Then, with probability at least $1-\epsilon$ over the randomness of the sample $\mathbf{X}$, the following holds simultaneously for all $\boldsymbol{r} \in \mathbb{R}^N$:

$$\mathbb{E}_{\boldsymbol{x}}\left[|\boldsymbol{x}^\mathsf{T}\boldsymbol{r}|^2\right] \le \left(1 - \sqrt{\frac{2R^2\log(N/\epsilon)}{\lambda}}\right)^{-1} \frac{1}{m}\left(\|\mathbf{X}\boldsymbol{r}\|^2 + \lambda\|\boldsymbol{r}\|^2\right). \tag{17}$$

In particular, the conclusion applies to any possibly data-dependent vector $\boldsymbol{r} = \boldsymbol{r}(\mathbf{X})$.

Proof. Let $\boldsymbol{\Sigma} := \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^\mathsf{T}]$. We first note that, for any $\boldsymbol{r} \in \mathbb{R}^N$,

$$\mathbb{E}_{\boldsymbol{x}}\left[|\boldsymbol{x}^\mathsf{T}\boldsymbol{r}|^2\right] = \boldsymbol{r}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{r}. \tag{50}$$

Let $\mathbf{P} := (\boldsymbol{\Sigma} + (\lambda/m)\mathbf{I})^{-1/2}$ and let $K = \lceil \lambda/R^2 \rceil$. Consider

$$\mathbf{S} := \mathbf{P}\frac{1}{m}\left(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I}\right)\mathbf{P}. \tag{51}$$

Then $\mathbb{E}[\mathbf{S}] = \mathbf{I}$. We write $\mathbf{S}$ as

$$\mathbf{S} = \left(\sum_{j=1}^{K} \frac{\lambda}{Km}\mathbf{P}^2\right) + \left(\sum_{i=1}^{m} \frac{1}{m}\mathbf{P}\boldsymbol{x}^{(i)}{\boldsymbol{x}^{(i)}}^\mathsf{T}\mathbf{P}\right). \tag{52}$$

Each summand has operator norm at most R^2/λ. Indeed, since $\boldsymbol{\Sigma} + (\lambda/m)\mathbf{I} \succeq (\lambda/m)\mathbf{I}$, we have $\|\mathbf{P}\|_{\mathrm{op}}^2 = \left\|(\boldsymbol{\Sigma} + (\lambda/m)\mathbf{I})^{-1/2}\right\|_{\mathrm{op}}^2 \le \frac{m}{\lambda}$. Therefore, using $\|\boldsymbol{x}^{(i)}\|_2 \le R$,

$$\left\|\frac{1}{m}\mathbf{P}\boldsymbol{x}^{(i)}{\boldsymbol{x}^{(i)}}^\mathsf{T}\mathbf{P}\right\|_{\mathrm{op}} = \frac{1}{m}\|\mathbf{P}\boldsymbol{x}^{(i)}\|_2^2 \le \frac{R^2}{\lambda}. \tag{53}$$

Similarly, for the artificial regularization summands, $\left\|\frac{\lambda}{Km}\mathbf{P}^2\right\|_{\mathrm{op}} \le \frac{1}{K} \le \frac{R^2}{\lambda}$, where the last inequality follows from $K = \lceil \lambda/R^2 \rceil$. Therefore, by Theorem A.1,

$$\mathbb{P}\left[\sigma_{\min}(\mathbf{S}) \le 1 - t\right] \le N\exp\left(-\frac{t^2\lambda}{2R^2}\right). \tag{54}$$

Taking $t = \sqrt{\frac{2R^2\log(N/\epsilon)}{\lambda}}$, we obtain that, with probability at least $1-\epsilon$,

$$\sigma_{\min}(\mathbf{S}) > 1 - \sqrt{\frac{2R^2\log(N/\epsilon)}{\lambda}}. \tag{55}$$

On this event,

$$\sup_{\boldsymbol{u}\in\mathbb{R}^N} \frac{\boldsymbol{u}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{u}}{(1/m)(\|\mathbf{X}\boldsymbol{u}\|^2 + \lambda\|\boldsymbol{u}\|^2)} = \sup_{\boldsymbol{u}\in\mathbb{R}^N} \frac{\boldsymbol{u}^\mathsf{T}(\mathbf{P}\boldsymbol{\Sigma}\mathbf{P})\boldsymbol{u}}{\boldsymbol{u}^\mathsf{T}\mathbf{S}\boldsymbol{u}} \tag{56}$$

$$\le \left(1 - \sqrt{\frac{2R^2\log(N/\epsilon)}{\lambda}}\right)^{-1}, \tag{57}$$

where we used $\|\mathbf{P}\boldsymbol{\Sigma}\mathbf{P}\|_{\mathrm{op}} \le 1$. Combining this estimate with Eq. (50) proves the result. □

Corollary 3.2. *Suppose the hypotheses of Theorem 3.1 hold. Let $\boldsymbol{q}$ be the output of OPTQ applied to $\boldsymbol{w}$ using the augmented calibration matrix $\tilde{\mathbf{X}} = \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I} \end{pmatrix}$, and let $\boldsymbol{r} := \boldsymbol{q} - \boldsymbol{w}$. Then, with probability at least $1-\epsilon$ over the randomness of the sample $\mathbf{X}$,*

$$\mathbb{E}_{\boldsymbol{x}}\left[|\boldsymbol{x}^\mathsf{T}\boldsymbol{w} - \boldsymbol{x}^\mathsf{T}\boldsymbol{q}|^2\right] \le \frac{C\delta^2}{m}\left(1 - \sqrt{\frac{2R^2\log(N/\epsilon)}{\lambda}}\right)^{-1} \sum_{i=1}^{N} \|\mathbf{P}_{>i}^{\perp}\tilde{\mathbf{X}}_i\|^2. \tag{19}$$

Here $C = 1/4$ for deterministic OPTQ and $C = 1$ for stochastic OPTQ.

Proof. Apply Theorem 3.1 to the OPTQ error vector

$$\boldsymbol{r} = \boldsymbol{q} - \boldsymbol{w}. \tag{58}$$

Then, to obtain Eq. (19), we use the OPTQ identity (Zhang et al., 2025, Eq. (3.5))

$$\|\tilde{\mathbf{X}}\boldsymbol{r}\|^2 = \sum_{i=1}^{N} |\alpha_i|^2 \|\mathbf{P}_{>i}^{\perp}\tilde{\mathbf{X}}_i\|^2. \tag{59}$$

For deterministic OPTQ, $|\alpha_i| \leq \delta/2$, and for stochastic OPTQ, $|\alpha_i| \leq \delta$ almost surely. Hence

$$|\alpha_i|^2 \leq C\delta^2, \tag{60}$$

where $C = 1/4$ for deterministic OPTQ and $C = 1$ for stochastic OPTQ. Substituting this into Eq. (59) and applying Eq. (17) completes the proof. □

B Proof of Theorem 3.3

We first will establish the following lemma.

Lemma B.1. *Let $\mathbf{A} \in \mathbb{R}^{(m+N)\times(m+N)}$ be a positive semi-definite matrix. If we run OPTQ, then*

$$\sum_{i=1}^{N} (\tilde{\mathbf{X}}\boldsymbol{v}^{(i)})^{\mathsf{T}} \mathbf{A} \tilde{\mathbf{X}}\boldsymbol{v}^{(i)} \leq \left(\max_i \|\mathbf{P}_{>i}^{\perp}\tilde{\mathbf{X}}_i\|^2\right) \operatorname{tr} \mathbf{A}. \tag{61}$$

Proof. Recall that by Eq. (8), each vector $\tilde{\mathbf{X}}\boldsymbol{v}^{(i)}$ is $\mathbf{P}_{>i}^{\perp}\tilde{\mathbf{X}}_i$, the projection of $\tilde{\mathbf{X}}_i$ onto the orthogonal complement of the range of $\tilde{\mathbf{X}}_{>i}$. Therefore, $\tilde{\mathbf{X}}\boldsymbol{v}^{(1)}, \ldots, \tilde{\mathbf{X}}\boldsymbol{v}^{(N)}$ is a collection of orthogonal vectors. Define

$$\boldsymbol{u}^{(i)} = \frac{\tilde{\mathbf{X}}\boldsymbol{v}^{(i)}}{\|\tilde{\mathbf{X}}\boldsymbol{v}^{(i)}\|} \tag{62}$$

for $i = 1, \ldots, N$ and let

$$\mathbf{U} = \begin{pmatrix} \boldsymbol{u}^{(1)} & \cdots & \boldsymbol{u}^{(N)} \end{pmatrix}. \tag{63}$$

We bound

$$\sum_{i=1}^{N} (\tilde{\mathbf{X}}\boldsymbol{v}^{(i)})^{\mathsf{T}} \mathbf{A} \tilde{\mathbf{X}}\boldsymbol{v}^{(i)} \leq \left(\max_i \|\mathbf{P}_{>i}^{\perp}\tilde{\mathbf{X}}_i\|^2\right) \sum_{i=1}^{N} {\boldsymbol{u}^{(i)}}^{\mathsf{T}} \mathbf{A} \boldsymbol{u}^{(i)} \tag{64}$$

and compute

$$\sum_{i=1}^{N} {\boldsymbol{u}^{(i)}}^{\mathsf{T}} \mathbf{A} \boldsymbol{u}^{(i)} = \operatorname{tr}(\mathbf{U}^{\mathsf{T}}\mathbf{A}\mathbf{U}) = \operatorname{tr}(\mathbf{A}). \tag{65}$$

□

We now proceed to prove Theorem 3.3.

Theorem 3.3. *Let $\mathbf{A} \in \mathbb{R}^{(m+N)\times(m+N)}$ be an arbitrary positive semi-definite matrix. For stochastic OPTQ, we can bound*

$$\mathbb{E}_{\alpha}[(\tilde{\mathbf{X}}\boldsymbol{r})^{\mathsf{T}} \mathbf{A} \tilde{\mathbf{X}}\boldsymbol{r}] \leq \frac{\delta^2}{4} \left(\max_i \|\mathbf{P}_{>i}^{\perp}\tilde{\mathbf{X}}_i\|^2\right) \operatorname{tr} \mathbf{A}, \tag{20}$$

where $\mathbb{E}_{\alpha}[\cdot]$ denotes the expectation is taken only over the randomness in Eq. (15) *associated with choosing α, with the sample matrix $\mathbf{X}$ fixed.*

Furthermore, with probability at least $1 - \epsilon$ over the randomness in α_i, we have

$$(\tilde{\mathbf{X}}\boldsymbol{r})^{\mathsf{T}} \mathbf{A} \tilde{\mathbf{X}}\boldsymbol{r} \leq 2\delta^2 \log(2/\epsilon) \left(\max_i \|\mathbf{P}_{>i}^{\perp}\tilde{\mathbf{X}}_i\|^2\right) \operatorname{tr} \mathbf{A}. \tag{21}$$

Proof. We first show Eq. (20). In the stochastic version of OPTQ, each α_i is a random variable that satisfies

$$\mathbb{E}[\alpha_i \mid \alpha_1, \ldots, \alpha_{i-1}] = 0 \quad \text{and} \quad \mathbb{E}[\alpha_i^2 \mid \alpha_1, \ldots, \alpha_{i-1}] \leq \frac{\delta^2}{4}. \tag{66}$$

By the law of total expectation,

$$\mathbb{E}[\alpha_i \alpha_j] = 0 \text{ if } i \neq j \text{ and} \tag{67}$$

$$\mathbb{E}[\alpha_i^2] \leq \frac{\delta^2}{4}. \tag{68}$$

We therefore see that

$$\mathbb{E}_\alpha[(\tilde{\mathbf{X}}\boldsymbol{r})^\mathsf{T}\mathbf{A}(\tilde{\mathbf{X}}\boldsymbol{r})] = \sum_{i=1}^N \sum_{j=1}^N \mathbb{E}[\alpha_i \alpha_j](\tilde{\mathbf{X}}\boldsymbol{v}^{(i)})^\mathsf{T}\mathbf{A}\tilde{\mathbf{X}}\boldsymbol{v}^{(j)} \tag{69}$$

$$\leq \sum_{i=1}^N \frac{\delta^2}{4}(\tilde{\mathbf{X}}\boldsymbol{v}^{(i)})^\mathsf{T}\mathbf{A}\tilde{\mathbf{X}}\boldsymbol{v}^{(i)}. \tag{70}$$

Equation (20) follows by applying Lemma B.1 to Eq. (70).

We now consider Eq. (21). From Eq. (66), we see that the sequence of vectors

$$\boldsymbol{u}^{(j)} = \sum_{i=1}^j \alpha_i \mathbf{A}^{1/2}\tilde{\mathbf{X}}\boldsymbol{v}^{(i)} \tag{71}$$

is a vector martingale. From (Pinelis, 1994, Theorem 3.5), we can bound

$$\mathbb{P}\left[\max_j \|\boldsymbol{u}^{(j)}\|^2 \geq t\right] \leq 2\exp\left(-\frac{t}{2\delta^2 \sum_{i=1}^N (\tilde{\mathbf{X}}\boldsymbol{v}^{(i)})^\mathsf{T}\mathbf{A}\tilde{\mathbf{X}}\boldsymbol{v}^{(i)}}\right). \tag{72}$$

Since $(\tilde{\mathbf{X}}\boldsymbol{r})^\mathsf{T}\mathbf{A}\tilde{\mathbf{X}}\boldsymbol{r} = \|\boldsymbol{u}^{(N)}\|^2$, we have

$$\mathbb{P}\left[(\tilde{\mathbf{X}}\boldsymbol{r})^\mathsf{T}\mathbf{A}\tilde{\mathbf{X}}\boldsymbol{r} \geq t\right] \leq 2\exp\left(-\frac{t}{2\delta^2 \sum_{i=1}^N (\tilde{\mathbf{X}}\boldsymbol{v}^{(i)})^\mathsf{T}\mathbf{A}\tilde{\mathbf{X}}\boldsymbol{v}^{(i)}}\right). \tag{73}$$

We set $t = 2\delta^2 \log(2/\epsilon) \sum_{i=1}^N (\tilde{\mathbf{X}}\boldsymbol{v}^{(i)})^\mathsf{T}\mathbf{A}\tilde{\mathbf{X}}\boldsymbol{v}^{(i)}$ to see that with probability at least $1-\epsilon$ over the randomness of choosing α_i we have

$$(\tilde{\mathbf{X}}\boldsymbol{r})^\mathsf{T}\mathbf{A}(\tilde{\mathbf{X}}\boldsymbol{r}) \leq 2\delta^2 \log(2/\epsilon) \sum_{i=1}^N (\tilde{\mathbf{X}}\boldsymbol{v}^{(i)})^\mathsf{T}\mathbf{A}\tilde{\mathbf{X}}\boldsymbol{v}^{(i)}. \tag{74}$$

We then again apply Lemma B.1 to Eq. (74) to conclude Eq. (21). □

C GENERALIZATION BEHAVIOR OF ROUND-TO-NEAREST ALGORITHMS

To precisely state a lower bound for MSQ, we first precisely define the algorithm. Similarly to OPTQ, we will consider two variants of MSQ: a deterministic variant and a stochastic variant. In deterministic MSQ, we obtain a quantized vector $\boldsymbol{q}_{\text{MSQ}}$ by rounding each coordinate of $\boldsymbol{w}$ to the nearest multiple of δ. That is,

$$(\boldsymbol{q}_{\text{MSQ}})_i := \delta \left\lfloor \frac{\boldsymbol{w}_i}{\delta} + \frac{1}{2} \right\rfloor. \tag{75}$$

In stochastic MSQ, we do a similar randomized rounding procedure as stochastic OPTQ to obtain a quantized vector $\boldsymbol{q}_{\text{SR}}$. First, for each coordinate we define

$$p_i := \boldsymbol{w}_i - \left\lfloor \frac{\boldsymbol{w}_i}{\delta} \right\rfloor \tag{76}$$

and randomly choose

$$\alpha_i = \begin{cases} -p_i\delta & \text{with probability } 1-p_i \\ (1-p_i)\delta & \text{with probability } p_i. \end{cases} \tag{77}$$

We then set $(\boldsymbol{q}_{\text{MSQ}})_i = \boldsymbol{w}_i + \alpha_i$.

Proposition C.1. *Let $\delta > 0$ and $\boldsymbol{z} \in \mathbb{R}^N$ be a vector with finite second moment. Suppose $\boldsymbol{w}$ is determined by sampling from the uniform distribution on $[-\delta/2, \delta/2]^N$. Choose either stochastic MSQ or deterministic MSQ and let $\boldsymbol{q}_{MSQ}$ be the quantized vector obtained by applying this algorithm to $\boldsymbol{w}$. There are absolute constants $c_1, c_2 > 0$ such that*

$$\mathbb{P}_{\boldsymbol{w},\alpha}\left[\mathbb{E}_{\boldsymbol{z}}\left[|\boldsymbol{z}^{\mathsf{T}}\boldsymbol{w} - \boldsymbol{z}^{\mathsf{T}}\boldsymbol{q}_{\mathrm{MSQ}}|^2\right] \geq c_1\delta^2\mathbb{E}_{\boldsymbol{z}}\|\boldsymbol{z}\|^2\right] \geq c_2. \tag{78}$$

Here, $\mathbb{E}_{\boldsymbol{z}}$ denotes that the expectation is taken only over $\boldsymbol{z}$ and $\mathbb{P}_{\boldsymbol{w},\alpha}$ denotes the probability with respect to the random choice of $\boldsymbol{w}$ and the randomness in stochastic MSQ, if chosen.

Proof. Denote $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^{\mathsf{T}}]$. Let $\boldsymbol{r}$ denote $\boldsymbol{w} - \boldsymbol{q}_{\mathrm{MSQ}}(\boldsymbol{w})$. With this notation,

$$\mathbb{E}_{\boldsymbol{z}}\left[|\boldsymbol{z}^{\mathsf{T}}\boldsymbol{w} - \boldsymbol{z}^{\mathsf{T}}\boldsymbol{q}_{\mathrm{MSQ}}|^2\right] = \boldsymbol{r}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{r} \tag{79}$$

and

$$c\delta^2\mathbb{E}_{\boldsymbol{z}}\|\boldsymbol{z}\|^2 = c\delta^2 \operatorname{tr}\boldsymbol{\Sigma}. \tag{80}$$

Therefore, we need only to demonstrate

$$\mathbb{P}_{\boldsymbol{w},\alpha}\left[\boldsymbol{r}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{r} \geq c\delta^2 \operatorname{tr}\boldsymbol{\Sigma}\right] \geq \frac{1}{2}. \tag{81}$$

We now proceed to do so. We see that $\boldsymbol{r}$ is a random vector independent of $\boldsymbol{z}$. This random vector consists of i.i.d. mean-zero coordinates such that each coordinate almost surely lies in the interval $[-\delta, \delta]$. Consequentially, $\boldsymbol{r}$ is a subgaussian random vector with subgaussian norm bounded above by $C_1\delta$, for some absolute constant C_1. We can now apply the Hanson–Wright inequality (Rudelson & Vershynin, 2013, Theorem 1.1) to the quadratic form $\boldsymbol{r}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{r}$ to conclude

$$\mathbb{P}\left[\left|\boldsymbol{r}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{r} - \mathbb{E}\left[\boldsymbol{r}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{r}\right]\right| \geq t\right] \leq 2\exp\left(-C_2\min\left\{\frac{t^2}{C_1^4\delta^4\|\boldsymbol{\Sigma}\|_{\mathrm{F}}^2}, \frac{t}{C_1^2\delta^2\|\boldsymbol{\Sigma}\|_{\mathrm{op}}}\right\}\right), \tag{82}$$

where C_2 is an absolute constant. We proceed to compute $\mathbb{E}[\boldsymbol{r}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{r}]$. This can be done as follows

$$\mathbb{E}[\boldsymbol{r}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{r}] = \mathbb{E}\left[\sum_{i=1}^N\sum_{j=1}^N \boldsymbol{r}_i\boldsymbol{r}_j\boldsymbol{\Sigma}_{ij}\right] \tag{83}$$

$$= \sum_{i=1}^N \mathbb{E}[\boldsymbol{r}_i\boldsymbol{r}_i]\boldsymbol{\Sigma}_{ii} \tag{84}$$

$$= \mathbb{E}[\boldsymbol{r}_1^2] \operatorname{tr}\boldsymbol{\Sigma}, \tag{85}$$

where in the last line we use that the coordinates of $\boldsymbol{r}$ are i.i.d. and mean-zero.

In the case of deterministic round-to-nearest $\boldsymbol{q}_1 = 0$ for all possible values of $\boldsymbol{w}_1$. Therefore, $\boldsymbol{r}_1 = \boldsymbol{w}_1$ is a uniform variable on $[-\delta/2, \delta]$, so

$$\mathbb{E}[\boldsymbol{r}_1^2] = \frac{1}{\delta}\int_{-\delta/2}^{\delta/2} \boldsymbol{w}_1^2 d\boldsymbol{w}_1 = \frac{\delta^2}{12}. \tag{86}$$

In the case of stochastic round-to-nearest, $\boldsymbol{r}_1 = \alpha_1$ by construction. This depends on the random variable p_1, whose marginal distribution is the uniform distribution on $[0, 1]$. Thus, we can compute

$$\mathbb{E}[\boldsymbol{r}_1^2] = \int_0^1 \left(p_1(1-p_1)^2\delta^2 + (1-p_1)p_1^2\delta^2\right) dp_1 = \int_0^1 p_1(1-p_1)\delta^2 dp_1 = \frac{\delta^2}{6}. \tag{87}$$

From Eqs. (85) to (87), we obtain the bound

$$\mathbb{E}[\boldsymbol{r}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{r}] \geq \frac{\delta^2}{12} \operatorname{tr}\boldsymbol{\Sigma}. \tag{88}$$

We now combine Eqs. (82) and (88) to obtain

$$\mathbb{P}\left[\boldsymbol{r}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{r} \geq \left(\frac{1}{6} + s\right)\delta^2 \operatorname{tr}\boldsymbol{\Sigma}\right] \leq 2\exp\left(-C_2\min\left\{\frac{s^2(\operatorname{tr}\boldsymbol{\Sigma})^2}{C_1^4\|\boldsymbol{\Sigma}\|_{\mathrm{F}}^2}, \frac{s\operatorname{tr}\boldsymbol{\Sigma}}{C_1^2\|\boldsymbol{\Sigma}\|_{\mathrm{op}}}\right\}\right), \tag{89}$$

Because $\mathbf{\Sigma}$ is a positive semi-definite matrix, we have

$$\|\mathbf{A}\|_{\text{op}} \leq \|\mathbf{A}\|_{\text{F}} \leq \operatorname{tr} \mathbf{A}. \tag{90}$$

Therefore,

$$\mathbb{P}\left[\boldsymbol{r}^{\mathsf{T}}\mathbf{\Sigma}\boldsymbol{r} \geq \left(\frac{1}{6} + s\right)\delta^2 \operatorname{tr}\mathbf{\Sigma}\right] \leq 2\exp\left(-C_2 \min\left\{\left(\frac{s}{C_1^2}\right)^2, \frac{s}{C_1}\right\}\right). \tag{91}$$

As $\boldsymbol{r}^{\mathsf{T}}\mathbf{\Sigma}\boldsymbol{r}$ is non-negative, Eq. (91) is enough to conclude that $\boldsymbol{r}^{\mathsf{T}}\mathbf{\Sigma}\boldsymbol{r}$ is subexponential with subexponential norm bounded by $C_3\delta^2 \operatorname{tr}\mathbf{\Sigma}$ for an absolute constant C_3. This yields the following bound for its second moment (Vershynin, 2025, Proposition 2.8.1)

$$\mathbb{E}\left[(\boldsymbol{r}^{\mathsf{T}}\mathbf{\Sigma}\boldsymbol{r})^2\right] \leq C_4(\delta^2 \operatorname{tr}\mathbf{\Sigma})^2 \tag{92}$$

for another absolute constant C_4. We now apply the Paley-Zygmund inequality to obtain

$$\mathbb{P}\left[\boldsymbol{r}^{\mathsf{T}}\mathbf{\Sigma}\boldsymbol{r} \geq \frac{\delta^2}{24}\operatorname{tr}\mathbf{\Sigma}\right] \geq \frac{1}{576C_4}. \tag{93}$$

□