SSTQ:PRIVACY-PRESERVING VECTOR QUANTIZATION VIA SUBSAMPLED STOCHASTIC TURBOQUANT

Anonymous authors
Paper under double-blind review

ABSTRACT

Achieving local differential privacy (LDP) in distributed optimization at low communication cost remains challenging: existing vector quantization methods such as vqSGD rely on high-dimensional geometric constructions and incur unfavorable dimension-dependent variance. We propose Subsampled Stochastic TurboQuant (SSTQ), combining a bounded Kashin representation, data-independent coordinate subsampling, and privacy-aware one-dimensional quantization. SSTQ has two variants: (1) a Flat Randomized Response variant that is unbiased and, for fixed codebook bit-width b, frame redundancy, and dimension-independent Kashin level, achieves reconstruction MSE linear in the ambient dimension d using only $\lceil \log_2 N \rceil + b$ bits per message, where $N = \Theta(d)$ is the Kashin frame size; and (2) a metric-aware truncated-Laplace variant that removes the exponential bit-width dependence at the cost of a non-vanishing bias. We also derive a convex uniform-surrogate codebook objective whose worst-case codebook-dependent bound improves from $\mathcal{O}(4^b)$ to $\mathcal{O}(2^b)$. Experiments on synthetic regression, Fashion-MNIST, and CIFAR-10 compare the per-message privacy–utility and uplink-communication trade-offs of SSTQ against established baselines, demonstrating favorable utility and communication efficiency.

1 INTRODUCTION AND MOTIVATION

The communication bottleneck in federated learning (FL) and large-scale distributed optimization has driven strong interest in gradient compression (Li et al., 2020; Kairouz & McMahan, 2021), while protecting sensitive client data against adversarial or untrusted aggregators requires Local Differential Privacy (LDP) (Dwork et al., 2006; Kasiviswanathan et al., 2011; Warner, 1965). The two objectives are inherently at odds: compression reduces high-dimensional continuous signals to compact discrete representations, whereas LDP requires adding noise and maintaining sufficient support over the same space for plausible deniability.

Existing approaches to private vector release fall into three paradigms, each with distinct limitations. Geometric quantization methods such as vqSGD (Gandikota et al., 2021) encode vectors using structured high-dimensional polytopes and apply randomized response over the resulting discrete set; though communication-efficient, their reliance on high-dimensional geometry yields unfavorable variance scaling that can grow cubically with the dimension. Appendix B shows their mean-squared error (equal to the variance, as the estimators are unbiased) is $\Theta_\epsilon(d^3)$ for fixed ϵ, and $\Theta(d^3/\epsilon^2)$ uniformly in the high-privacy regime $0 < \epsilon \leq 1$.

A second class attains optimal variance scaling under LDP: SQKR (Chen et al., 2020) and PrivUnit (Bhowmick et al., 2018) reach the information-theoretic limit of $O(d/\epsilon^2)$ by operating over bounded representations such as Kashin frames or continuous distributions. However, they either impose restrictive communication constraints—1-bit quantization and shared randomness in SQKR—or transmit dense, uncompressed vectors, as in PrivUnit, limiting practicality under bandwidth constraints.

A third class consists of data-oblivious quantization schemes combining randomized orthogonal transformations with dataset-independent scalar codebooks, avoiding data-dependent training while attaining strong compression across a broad range of dimensions and rates. Early examples are DRIVE and EDEN, rotation-based schemes for distributed mean estimation with unbiased reconstruction and

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.

flexible rates (Vargaftik et al., 2021; 2022); more recently TurboQuant (Zandieh et al., 2026) achieves near-optimal distortion over a wide range of bit-widths and dimensions, and RaBitQ applies the same template to distance and inner-product estimation (Gao & Long, 2024). Effective as they are for compression, these methods give no privacy guarantees and can suffer directional bias under naive privatization. Moreover, TurboQuant randomly rotates the input, then exploits near-independence of coordinates in high dimensions to quantize each with an optimal scalar quantizer, so its output bits scale with the input dimension (constant bit-width per coordinate); by contrast, we target a representation whose bit count grows only logarithmically with the dimension.

Building on these insights, we develop a data-oblivious, differentially private quantization framework based on one-dimensional scalar quantization of Kashin representations, deriving the corresponding mean squared error and designing codebooks optimized for the privatized setting, which enables efficient communication while preserving strong privacy guarantees. *Although our goal differs, we emphasize this connection by incorporating "TurboQuant" directly into the name of our procedure.* Correlation-aware compression schemes exploiting similarities between the vectors being averaged in federated learning have also been proposed (Suresh et al., 2022; Jiang et al., 2023; Vardhan & Mazumdar, 2026), as have low-complexity encoding/decoding procedures (Ben-Basat et al., 2024). Prior Kashin-based LDP mechanisms and other related work are discussed in Appendix A.

Table 1: Single-message reconstruction MSE for normalized inputs $x \in B_2^d(1)$. Constants depending on fixed frame parameters are suppressed. SQKR's displayed risk is the per-user contribution to mean-estimation risk before division by the number of users.

Mechanism	Topological / DP Geometry	Number of Bits per Client	Pure LDP MSE
vqSGD (cross-polytope RR)	Cross-polytope	$\lceil \log_2 d \rceil + 1$	$d\left(\frac{e^\epsilon + 2d - 1}{e^\epsilon - 1}\right)^2 - 1$
SQKR	Binary Kashin frame	k public-coin; $k(\lceil \log_2 N \rceil + 1)$ private-coin	$\mathcal{O}\left(\frac{d}{\epsilon^2 \wedge \epsilon \wedge b_0}\right)$
PrivUnit	Continuous sphere	continuous output; $32d$ under float32	$\mathcal{O}\left(\frac{d}{\epsilon \wedge \epsilon^2}\right)$
SSTQ (Flat-RR)	Scalar Kashin codebook	$\lceil \log_2 N \rceil + b$	$\mathcal{O}\left(\frac{dK^2}{\rho}\left[1 + \frac{4^b}{\epsilon \wedge \epsilon^2}\right]\right)$
SSTQ (Flat-RR, optimized surrogate)	Scalar Kashin codebook	$\lceil \log_2 N \rceil + b$	$\mathcal{O}\left(\frac{dK^2}{\rho}\left[1 + \frac{2^b}{\epsilon \wedge \epsilon^2}\right]\right)$
SSTQ (Metric-Aware)	Scalar Kashin codebook	$\lceil \log_2 N \rceil + b$	$\mathcal{O}\left(\frac{dK^2}{\rho}\left[1 + \frac{C'^2}{(2^b-1)^2} + \frac{1}{\epsilon \wedge \epsilon^2}\right]\right)$

Here $\rho = N/d$. For SQKR, $k = \min\{\lceil \epsilon \log_2 e \rceil, b_0\}$. A public coin refers to randomness shared in advance between the client and server, while a private coin is randomness generated locally by the client and should be communicated to the server for reconstruction.

1.1 Our Contributions

We present a unified framework connecting information-theoretic quantization with local differential privacy. Our contributions are as follows.

- **SSTQ architecture.** We introduce Subsampled Stochastic TurboQuant (SSTQ), combining a scaled Parseval frame admitting a bounded Kashin representation with data-independent coordinate subsampling. For fixed bit-width b, redundancy $\rho = N/d$, and dimension-independent Kashin constant K, it achieves MSE scaling $\mathcal{O}(d/(1 \wedge \epsilon \wedge \epsilon^2))$—optimal in the constant-ϵ regime (Duchi & Rogers, 2019)—using only $\lceil \log_2 N \rceil + b$ uplink bits in the private-coin setting.
- **Privacy-aware codebook optimization.** We derive a surrogate privacy-aware stochastic quantization loss $\mathcal{L}_{\text{SSTQ}}$, a convex formulation for codebook design; optimizing it improves the codebook-dependent worst-case MSE scaling from $\mathcal{O}(4^b)$ (Theorem 3.2) to $\mathcal{O}(2^b)$ (Theorem 4.3).
- **Metric-aware mechanism for dense regimes.** We propose a metric-aware truncated-Laplace mechanism avoiding the exponential variance growth of standard randomized response at higher bit-widths, while preserving pure ϵ-LDP and enabling efficient multi-bit quantization.
- **Empirical evaluation.** We evaluate SSTQ (Flat-RR and Metric-Aware) against established baselines—PrivUnit, vqSGD, and SQKR—on federated CIFAR-10 and Fashion-MNIST tasks, across varying dimensions and codebook sizes, assessing utility, variance behavior, and communication efficiency.

2 Mathematical Preliminaries

We briefly review the notions of local differential privacy and overcomplete frame theory underlying the developments below.

Local Differential Privacy (LDP). LDP requires each client to perturb its data before transmission, so the released message stays private even against an untrusted aggregator: the server cannot reliably infer the original client data, regardless of auxiliary information or post-processing.

Definition 2.1 (ϵ-Local Differential Privacy)**.** *A randomized mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Z}$ satisfies pure ϵ-Local Differential Privacy (ϵ-LDP) if $\Pr[\mathcal{M}(x) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(x') \in S]$ for all pairs of possible inputs $x, x' \in \mathcal{X}$ and all measurable subsets of outputs $S \subseteq \mathcal{Z}$, where $\epsilon > 0$ is the strict privacy budget.*

Definition 2.2 (Scaled Parseval tight frame)**.** *Let $U \in \mathbb{R}^{N \times d}$, with rows $u_1^\top, \ldots, u_N^\top$, satisfy $U^\top U = \frac{N}{d} I_d$. Equivalently, $\Phi := \sqrt{\frac{d}{N}}\, U^\top \in \mathbb{R}^{d \times N}$ satisfies $\Phi\Phi^\top = I_d$, so Φ is a Parseval frame. We call U the corresponding scaled Parseval frame.*

Theorem 2.3 (Kashin's Representation Theorem)**.** *Let $\rho = N/d > 1$, let $U \in \mathbb{R}^{N \times d}$ be a scaled Parseval frame, and suppose that the associated Parseval frame $\Phi = \sqrt{d/N}\, U^\top$ satisfies the Lyubarskii–Vershynin (η, δ)-uncertainty principle (Lyubarskii & Vershynin, 2010), where $0 < \eta, \delta < 1$. Then every $x \in \mathbb{R}^d$ admits a coefficient vector $y \in \mathbb{R}^N$ such that*

$$x = \frac{d}{N} U^\top y, \qquad \|y\|_\infty \leq \frac{K}{\sqrt{N}} \|x\|_2, \qquad K = \frac{\sqrt{\rho}}{(1-\eta)\sqrt{\delta}}.$$

The theorem follows from the Parseval-frame formulation of (Lyubarskii & Vershynin, 2010) after rescaling, so K depends on the redundancy and the uncertainty-principle constants. For any fixed $\rho = N/d > 1$, a Haar-distributed $N \times d$ Stiefel matrix scaled so that $U^\top U = (N/d) I_d$ is a tight frame satisfying the uncertainty principle with high probability with constants depending only on ρ, so its Kashin level is independent of d. Structured partial Fourier or Hadamard constructions permit fast multiplication but currently guarantee only a polylogarithmic Kashin level; we therefore retain the dependence on K explicitly in presenting our bounds.

3 Subsampled Stochastic TurboQuant (SSTQ)

As shown in Appendix B, geometric vector quantization, most notably vqSGD (Gandikota et al., 2021), suffers a $\Theta(d^3/\epsilon^2)$ variance growth rate that severely limits its use for high-dimensional gradient embeddings. We therefore propose Subsampled Stochastic TurboQuant (SSTQ), which decouples the codebook size M from the ambient dimension d by combining overcomplete tight frames, data-oblivious coordinate subsampling, and one-dimensional stochastic quantization.

The SSTQ Algorithmic Pipeline. Fix a public input radius $R > 0$ and let $\mathcal{X} = B_2^d(R) = \{x \in \mathbb{R}^d : \|x\|_2 \leq R\}$. Let $U \in \mathbb{R}^{N \times d}$ be a public scaled Parseval frame satisfying Theorem 2.3, with $N/d = \rho$. The Kashin coefficient therefore is bounded by $B = \frac{KR}{\sqrt{N}}$. All protocol parameters, including R, U, and the codebook, are fixed independently of the protected input.

We represent the scalar codebook as an indexed list $\Gamma = (c_1, \ldots, c_M)$, with $M = 2^b$, rather than as a set, and refer to $b = \log_2(M)$ as the *"codebook bit-width"*. The labels $1, \ldots, M$ remain distinct even when some scalar values coincide, and we assume $-B = c_1 \leq c_2 \leq \cdots \leq c_M = B$ and $\sum_{i=1}^M c_i = 0$, where repeated entries are counted with multiplicity. Algorithm 1 takes such a codebook as input and, as stated, uses Flat Randomized Response; the design of Γ is discussed in Section 4 via a surrogate privacy-aware stochastic quantization loss. Section 5 introduces an alternative Metric-Aware truncated-Laplace variant, which replaces flat randomized response with continuous truncated Laplace noise injection followed by nearest-codeword quantization and is better suited to higher bit-width regimes. Because the protocol employs a 1-sparse projection over an N-dimensional Kashin frame, compressing the full d-dimensional target vector requires transmitting exactly two discrete components: the coordinate index $J \in [N]$ ($\lceil \log_2 N \rceil$ bits) and the randomized codeword index $Z \in [M]$ (b bits), so the payload per message is $\lceil \log_2 N \rceil + b$ bits. Notably, and in contrast to SQKR (Chen et al., 2020), this scheme does not rely on shared randomness (i.e., public-coin assumptions). The corresponding server-side reconstruction is given in Algorithm 2.

Theoretical Guarantees. We now state the theoretical guarantees on the utility and privacy of the SSTQ pipeline, with proofs deferred to Appendix C.

Algorithm 1 SSTQ Client Encoding Algorithm (Flat-RR)

Require: $x \in B_2^d(R)$, public scaled Parseval frame $U \in \mathbb{R}^{N\times d}$, and public indexed codebook $\Gamma = (c_1, \dots, c_M)$.

1: **Redundant Kashin Bounding:** Execute Lyubarskii's algorithm (Lyubarskii & Vershynin, 2010) to compute $y \in \mathbb{R}^N$ such that $\frac{d}{N}U^\top y = x$ and $\|y\|_\infty \le B = \frac{KR}{\sqrt{N}}$.

2: **Oblivious Subsampling:** Sample $J \sim \mathrm{Unif}([N])$, independently of x.

3: **Stochastic Interpolation on labels:** If $y_J = c_i$ for one or more indices, set

$$V = \min\{i : c_i = y_J\}.$$

Otherwise set $k_- = \max\{i : c_i < y_J\}$ and $k_+ = \min\{i : c_i > y_J\}$, and sample $V \in \{k_-, k_+\}$ according to

$$\Pr(V = k_+ \mid y_J) = \frac{y_J - c_{k_-}}{c_{k_+} - c_{k_-}}, \qquad \Pr(V = k_- \mid y_J) = \frac{c_{k_+} - y_J}{c_{k_+} - c_{k_-}}.$$

4: **Flat Randomized Response on labels:** Sample $Z \in [M]$ with

$$\Pr(Z = \ell \mid V = i) = \begin{cases} p = \frac{e^\epsilon}{e^\epsilon + M - 1}, & \ell = i, \\ q = \frac{1}{e^\epsilon + M - 1}, & \ell \neq i. \end{cases}$$

5: **return** Transmit (J, Z). **Cost:** $\lceil \log_2 N \rceil + b$ bits.

Algorithm 2 SSTQ Server Decoding Algorithm (Flat-RR)

Require: Message (J, Z), indexed codebook $\Gamma = (c_1, \dots, c_M)$, and scaled Parseval frame U.

1: **Scalar unbiasing:** Set $\widetilde{y}_J = c_Z/(p - q)$.

2: **Sparse reconstruction:** Set $\widetilde{y} = N\widetilde{y}_J e_J \in \mathbb{R}^N$.

3: **Frame projection:** Set

$$\hat{x} = \frac{d}{N}U^\top \widetilde{y} = d\,\widetilde{y}_J u_J.$$

4: **return** $\hat{x}$.

Theorem 3.1 (Strict ϵ-LDP and Unbiasedness)**.** *The SSTQ with flat randomized response, as presented in Algorithm 1, satisfies pure ϵ-Local Differential Privacy. Furthermore, the decoding algorithm given by Algorithm 2 returns an unbiased estimator, $\mathbb{E}[\hat{x} \mid x] = x$.*

Our next theorem bounds the mean-squared-error (MSE) of the SSTQ (with Flat-RR) algorithm.

Theorem 3.2 (MSE of SSTQ (Flat-RR))**.** *Let $x \in B_2^d(R)$, $M = 2^b$, and let Γ be any indexed, ordered, boundary-anchored, zero-sum codebook supported on $[-B, B]$. Then, for the reconstructed vector $\hat{x}$ by the Flat-RR SSTQ algorithm, we have*

$$\mathrm{MSE}_{\mathrm{SSTQ}}(x) := \mathbb{E}\Big[\|\hat{x} - x\|_2^2 \mid x\Big] \le \frac{d^2B^2}{(p-q)^2} - \|x\|_2^2 = \frac{dK^2R^2}{\rho}\left(1 + \frac{M}{e^\epsilon - 1}\right)^2 - \|x\|_2^2. \quad (1)$$

Consequently,

$$\mathrm{MSE}_{\mathrm{SSTQ}}(x) = O\left(\frac{dK^2R^2}{\rho}\left[1 + \frac{4^b}{\epsilon \wedge \epsilon^2}\right]\right).$$

For constant b and constant ϵ, SSTQ attains the optimal minimax rate $O(dR^2/(\epsilon \wedge \epsilon^2))$ for private mean estimation without quantization (Duchi & Rogers, 2019). For n independent clients whose unbiased messages are averaged, the mechanism-induced contribution is divided by n.

Remark 3.3. *In non-private quantization, increasing the bit-width b refines the $M = 2^b$-state grid and reduces quantization error. Under LDP, however, MSE also includes privatization variance: Flat Randomized Response replaces the true bin uniformly among the other $M - 1$, ignoring codebook geometry. As M grows, this dilutes the signal and increases reconstruction error. Indeed, unbiased decoding requires a correction factor that scales with $M = 2^b$, motivating the metric-aware mechanism in Section 5.*

4 UNIFORM-SURROGATE PRIVACY-AWARE CODEBOOK OPTIMIZATION ($\mathcal{L}_{\text{SSTQ}}$)

A key feature of SSTQ is that the Kashin representation distributes the signal across bounded scalar coordinates. By jointly modeling stochastic interpolation and randomized-response noise, we derive a data-oblivious, privacy-aware codebook optimization method suited to online applications.

Let Y have an arbitrary law μ on $[-B, B]$; μ may be discrete, as for the empirical coordinate law of a fixed vector. For the indexed interpolation rule in Algorithm 1, define $r_\Gamma(y) := \mathrm{Var}(c_V | Y = y)$, so $r_\Gamma(y) = (y - c_k)(c_{k+1} - y)$ if $c_k < y < c_{k+1}$ across a positive gap and $r_\Gamma(y) = 0$ if y equals a codeword value; thus r_Γ stays well defined even when codeword values repeat. Applying the law of total variance to interpolation followed by Flat RR gives $\mathrm{Var}(\widetilde{y}_j \mid Y = y) = \frac{r_\Gamma(y)}{p-q} + \left(\frac{1}{p-q} - 1\right) y^2 + \frac{q}{(p-q)^2} \sum_{i=1}^M c_i^2$, and since $\mathbb{E}[\widetilde{y}_j \mid Y = y] = y$ it follows that

$$\mathbb{E}[\widetilde{y}_j^2] = \frac{1}{p-q} \underbrace{\left[\int r_\Gamma(y)\, d\mu(y) + \frac{1}{e^\epsilon - 1} \sum_{i=1}^M c_i^2 \right]}_{:= \mathcal{L}^*_{\text{SSTQ}}(\Gamma; \mu)} + \frac{1}{p-q} \int y^2\, d\mu(y); \tag{2}$$

see Appendix D.1 for the full derivation. It is straightforward to see that $\text{MSE}_{\text{SSTQ}} < d^2 \mathbb{E}[\tilde{y}_j^2]$ (see Equation equation 43), and the last integral above does not depend on Γ. This motivates the codebook-dependent loss $\mathcal{L}^*_{\text{SSTQ}}(\Gamma; \mu)$ for a general coordinate law, which, if μ has a density f_Y, can equivalently be written as

$$\mathcal{L}^*_{\text{SSTQ}}(\Gamma; \mu) = \sum_{k:\, c_k < c_{k+1}} \int_{c_k}^{c_{k+1}} (y - c_k)(c_{k+1} - y) f_Y(y)\, dy + \frac{1}{e^\epsilon - 1} \sum_{i=1}^M c_i^2. \tag{3}$$

Definition 4.1 (Uniform-Surrogate Privacy-Aware Objective)**.** *For $\mu_{\text{unif}}(dy) = dy/(2B)$, define*

$$\mathcal{L}_{\text{SSTQ}}(\Gamma) := \mathcal{L}^*_{\text{SSTQ}}(\Gamma; \mu_{\text{unif}}) = \sum_{k=1}^{M-1} \frac{(c_{k+1} - c_k)^3}{12B} + \frac{1}{e^\epsilon - 1} \sum_{i=1}^M c_i^2, \tag{4}$$

subject to $c_1 = -B$, $c_M = B$, $c_1 \le \cdots \le c_M$, and $\sum_{i=1}^M c_i = 0$. Equalities between adjacent entries are permitted because the entries are indexed labels.

The privacy term is an explicit ℓ_2 penalty on all labeled codeword positions, including multiplicities; since the boundary values are fixed, optimization moves interior entries toward the origin while balancing the interpolation penalty. The next theorem proves strong convexity of the loss $\mathcal{L}_{\text{SSTQ}}$.

Theorem 4.2 (Strong Convexity of $\mathcal{L}_{\text{SSTQ}}$ on the Feasible Set)**.** *For $B > 0$, $\epsilon > 0$, and $M \ge 2$, the feasible set is nonempty, compact, and convex. The objective in equation 4 is $2/(e^\epsilon - 1)$-strongly convex on this set and therefore has a unique global minimizer Γ^*.*

The proof is given in Appendix D.

Improved Worst-Case Bound via Surrogate Codebook Optimization. While Theorem 3.2 gives a worst-case bound quadratic in $M = 2^b$, optimizing the uniform-surrogate objective improves the codebook-dependent part of this bound to linear in M.

Theorem 4.3. *Let Γ^* minimize $\mathcal{L}_{\text{SSTQ}}$ over the feasible indexed codebooks. Suppose $B = KR/\sqrt{N}$ and write $\rho = N/d$. Assuming $M > (4(e^\epsilon - 1))^{1/3} + 1$, the Flat-RR SSTQ with Γ^* satisfies*

$$\text{MSE}_{\text{SSTQ}} = \mathcal{O}\left(\frac{dK^2R^2}{\rho} \left[1 + \frac{2^b}{\epsilon \wedge \epsilon^2} \right] \right). \tag{5}$$

In particular, the bound is linear in d when K, ρ, b, and R are dimension-independent constants.

The proof (Appendix D) compares the surrogate loss at the optimized codebook with the minimum over an explicit family $\tilde{\Gamma}_m$ spanning m contiguous gap intervals of size B/m and collapsing the remaining $M - 2m - 1$ indices symmetrically to 0. For comparison, a uniform grid obeys $\sum_{i=1}^M c_i^2 = B^2 \frac{M(M+1)}{3(M-1)} = \Theta(B^2 M)$, so at fixed ϵ its privacy contribution scales as $\Theta_\epsilon(M^2)$ after the Flat-RR multiplier, whereas Theorem 4.3 gives an $\mathcal{O}_\epsilon(M)$ bound for the optimized surrogate.

5 CODEBOOK BIT-WIDTH SCALABILITY: THE METRIC-AWARE LAPLACE MECHANISM

As noted in Remark 3.3, Flat Randomized Response incurs a variance penalty growing rapidly with b, since the probability gap $p - q$ shrinks on the order of 2^{-b}. To remove this alphabet-size multiplier from the privacy-noise term, we introduce a *Metric-Aware Laplace Mechanism*. Given an interpolated label $V = k$ with scalar value c_k, it (i) draws a continuous sample T from a truncated Laplace distribution on $[-B, B]$ centered at c_k,

$$f_{T|V=k}(t) = \frac{1}{Z_k} \exp\left(-\frac{\epsilon|t - c_k|}{2D_B}\right), \quad t \in [-B, B], \tag{6}$$

where $D_B = 2B$ is the domain width and $Z_k = \int_{-B}^{B} \exp(-\epsilon|t - c_k|/(2D_B))dt$ normalizes; and (ii) applies nearest-codeword quantization under the deterministic measurable tie rule $\tau(t) := \min \arg\min_{i\in[M]} |t - c_i|$, releasing $Z = \tau(T)$ and decoding it as c_Z. With $\mathrm{cell}_i := \{t \in [-B, B] : \tau(t) = i\}$—a disjoint measurable partition, including when codeword values coincide—the effective transition probabilities are

$$P_{i|k} = \Pr(Z = i \mid V = k) = \frac{1}{Z_k} \int_{\mathrm{cell}_i} \exp\left(-\frac{\epsilon|t - c_k|}{2D_B}\right) dt. \tag{7}$$

Theorem 5.1 (Pure ϵ-LDP of Metric-Aware Laplace)**.** *For any codebook bounded in $[-B, B]$, the Metric-Aware Laplace Mechanism defined above satisfies pure ϵ-LDP.*

The metric-aware variant relaxes coordinate-level unbiasedness at decoding: on receiving the ϵ-LDP codeword c_Z, the aggregator applies *identity decoding*, setting $\widetilde{y}_J = c_Z$ and projecting into the 1-sparse vector $\widetilde{y} = N\widetilde{y}_J e_J = Nc_Z e_J$, with J the index sampled uniformly by the encoder. This trades a localized bias for reduced variance, enabling seamless integration with arbitrarily optimized, non-uniform continuous codebooks. We refer to Algorithms 3–4 in Appendix E for the formal description of the encoder and decoder.

Theorem 5.2 (MSE and Bias of metric-aware SSTQ)**.** *We assume a codebook of size $M = 2^b$ with its maximum codebook gap satisfying $\Delta_{\max} = \max_k(c_{k+1} - c_k) \le 2BC/(2^b - 1)$, with $B = KR/\sqrt{N}$ for a structural constant $C \ge 1$. Define the row-coherence parameter $\mu_U := \max_{j\in[N]} \|u_j\|_2^2$. Then,*

$$\textit{MSE}_{\textit{MA}} := \mathbb{E}\left[\|\widehat{x} - x\|_2^2\right] \le \frac{dK^2R^2}{\rho}\left(1 + \frac{2C^2}{(2^b - 1)^2} + \frac{16(\mu_U + 1/d)}{\epsilon} + \frac{256}{\epsilon^2}\right) - R^2,$$

$$\zeta^2 := \|\mathbb{E}[\widehat{x}] - x\|_2^2 \le \frac{64K^2R^2}{\rho\epsilon^2},$$

where $\rho = N/d > 1$. For any scaled Parseval frame (Definition 2.2), $\mu_U \le \rho$.

Structural Implications. The theorem, proved in Appendix E.2, gives a bias bound independent of dimension and bit-width. By allowing this bias, the metric-aware mechanism reduces variance and avoids the Flat-RR variant's $O(2^b)$ dependence. In the MSE bound, $\mathcal{O}(4^{-b})$ reflects deterministic quantization error, while $\mathcal{O}(\epsilon^{-1})$ and $\mathcal{O}(\epsilon^{-2})$ capture noise introduced for local differential privacy. Together, these terms impose an algebraic error floor that denser grids cannot overcome. Balancing the spatial and privacy-dependent terms gives the saturation bit-width $b_{\mathrm{sat}} = \max\{1, \lceil \log_2(1 + \Theta(C\sqrt{\epsilon \wedge \epsilon^2}))\rceil\}$. For $\epsilon \le 1$, the bound is saturated with one bit; for $\epsilon \gg 1$, the threshold scales as $\frac{1}{2}\log_2 \epsilon$, up to additive constants. See Appendix E.3 for the derivation.

5.1 CODEBOOK OPTIMIZATION

Throughout, $Y = y_J(X, U)$ is the coordinate seen by the encoder when X is a random population input and the frame U is fixed and public (or publicly randomized independently of X), with $J \sim \mathrm{Unif}[N]$ independent. Any density f_Y below is thus the population law over (X, U, J), not one generated by frame randomness for a single private vector; Appendix E.4 makes this precise and notes that, unlike the uniform surrogate of Definition 4.1 (which uses only the public interval $[-B, B]$), optimization under the true f_Y is an oracle/public-distribution result.

Similar to the flat randomization mechanism, we can write a data-dependent objective loss that captures the codebook-dependent part of MSE bound. For the metric-aware mechanism, the loss is

derived in Appendix E.2.1 and is given by

$$\mathcal{L}^*_{\text{MA-SSTQ}}(\Gamma) = \sum_{k=1}^{M-1} \int_{c_k}^{c_{k+1}} W_k(y-c_k)(c_{k+1}-y) f_Y(y)\, dy \;\; + \;\; \sum_{k=1}^{M} \pi_k \Phi_k \,,$$

where $\pi_k = \int_{-B}^{B} \Pr[V = k \mid Y = y] f_Y(y)\, dy$ defines the assignment probability of the interpolated label V. In addition, $\mu_k = \mathbb{E}[c_Z \mid V = k]$ (the expected output codeword) and $\Phi_k = \mathbb{E}[(c_Z - c_k)^2 \mid V = k]$ (the metric-aware variance), both computed under the effective transition probabilities equation 7, and $W_k = \frac{2(\mu_{k+1}-\mu_k)}{c_{k+1}-c_k} - 1$.

The flat mechanism's objective includes a rigid regularization term from uniform point-selection variance, $G(\Gamma) \propto \sum_{i=1}^{M} c_i^2$. For a uniform grid over $[-B, B]$ this grows as $\mathcal{O}(M)$, making the MSE bound vacuous as M grows. Optimizing against the uniform surrogate $f_Y(y) = \frac{1}{2B}$ instead clusters the points near the origin, suppressing this penalty to $\mathcal{O}(1)$—a strict improvement over the uniform grid, independent of the target distribution. Identity decoding, by contrast, localizes the metric mechanism's variance: $\sum_k \pi_k \Phi_k \le \max_k \Phi_k \approx \mathcal{O}(\epsilon^{-2})$, independently of codebook geometry and size. Distribution-aware optimization can thus improve the spatial term, as discussed below.

Optimization under the True Distribution. We formalize this limited improvement in Appendix E.2.3 (Theorem E.1) via the Panter–Dite shape factor $C_f = \frac{1}{4B^2}\left(\int_{-B}^{B} f_Y(y)^{1/3} dy\right)^3$, which by Hölder's inequality satisfies $C_f \le 1$, with equality iff f_Y is uniform. Minimizing $\mathcal{L}^*_{\text{MA-SSTQ}}$ under a known, strictly positive f_Y leaves the bias bound $\zeta^2_{\text{pop}} \le 64K^2R^2/(\rho\epsilon^2)$ and the $\mathcal{O}(\epsilon^{-1})$, $\mathcal{O}(\epsilon^{-2})$ privacy terms unchanged, but replaces the structural constant $C \ge 1$ of Theorem 5.2 by $C_0^2 \le \min\{1, \frac{2}{3} C_f R_M\}$, where $R_M = (M-1)^{-1} \sum_k \sup_{I_k} f_Y / \inf_{I_k} f_Y \ge 1$ is the average local oscillation ratio across $I_k = [c_k, c_{k+1}]$. Since $C_f < 1$ for any strictly non-uniform density and $R_M \to 1$ as $b \to \infty$, $C_0^2 < 1$ at high resolutions, explaining the strict improvement over Theorem 5.2.

The Asymptotic Limit ($\epsilon \to \infty$) and the Panter–Dite Framework. As the privacy constraint relaxes, the truncated Laplace density concentrates at a Dirac delta at the interpolated codeword, so the localized variance vanishes ($\Phi_k \to 0$) and the spatial scale modifier satisfies $W_k \to 1$, while the stochastic interpolation step remains active and reduces to the classical dithered quantizer. The loss therefore degenerates to the classical spatial quantization distortion $\sum_k \int_{c_k}^{c_{k+1}} (y - c_k)(c_{k+1} - y) f_Y(y)\, dy$, whose high-resolution minimizer allocates codewords with point density proportional to $f_Y(y)^{1/3}$—exactly the Panter–Dite rule (Panter & Dite, 1951). Thus, as LDP constraints vanish, the optimal metric-aware codebook abandons variance regularization and recovers the classical foundations of scalar source coding; Appendix E.5 gives the full derivation.

6 Federated Learning Convergence Analysis

We next apply SSTQ to private federated learning under a standard setup with common assumptions of non-convex optimization and bounded statistical variance. Consider n clients, with $f_i(w)$ the local loss for client i and $F(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$ the global objective. Assumption F.1, stated in full in Appendix F.1, collects the standing conditions. All of them are conditional on the natural filtration $\mathcal{F}_t = \sigma(w_0,\ \hat{g}_{i,s} : i \in [n],\ s < t)$ generated by the initial point and the gradients transmitted before step t, with respect to which the iterate w_t is measurable; we write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$. The conditions are: (i) *objective smoothness*—F is L-smooth and bounded below by F^*, with $\Delta_0 := F(w_0) - F^*$; (ii) *stochastic gradient bounds*—the minibatch gradients $g_{i,t} = \nabla f_i(w_t, \xi_{i,t})$ satisfy $\|g_{i,t}\|_2 \le G$ almost surely and $\mathbb{E}_t[\|g_{i,t} - \nabla f_i(w_t)\|_2^2] \le \sigma^2$; (iii) *compression noise*—the transmitted gradient $\hat{g}_{i,t}$ obeys $\mathbb{E}_t[\|\hat{g}_{i,t} - g_{i,t}\|_2^2] \le \omega G^2$ for all i, t; (iv) *bounded bias*—the server estimator $\hat{g}_t = \frac{1}{n}\sum_{i=1}^{n} \hat{g}_{i,t}$ has bias $b_t := \mathbb{E}_t[\hat{g}_t] - \nabla F(w_t)$ with $\|b_t\|_2^2 \le \zeta^2$ almost surely for all t; and (v) *conditional independence*—given $\mathcal{F}_t$, the client estimator errors $\{\hat{g}_{i,t} - \nabla f_i(w_t)\}_{i=1}^n$ are mutually independent.

Our next theorem is a general convergence result for distributed SGD under Assumption F.1.

Theorem 6.1 (Non-Convex DSGD Convergence Rate)**.** *Under Assumption F.1, consider the iterates $w_{t+1} = w_t - \eta\, \hat{g}_t$ run for T steps across n clients, and define $V := \frac{2(\omega G^2 + \sigma^2)}{n} + \zeta^2$.* (Non-asymptotic

bound.) *Setting* $\eta = \min\{\frac{1}{4L}, \sqrt{\frac{\Delta_0}{LVT}}\}$ *guarantees*

$$\min_{0\le t\le T-1} \mathbb{E}\big[\|\nabla F(w_t)\|_2^2\big] \;\le\; \frac{16L\Delta_0}{T} + \frac{8\sqrt{\Delta_0 LV}}{\sqrt{T}} + 2\zeta^2. \tag{8}$$

(Asymptotic rate.) *For* $T \ge 16L\Delta_0/V$*, the above simplifies to*

$$\min_{0\le t\le T-1} \mathbb{E}\big[\|\nabla F(w_t)\|_2^2\big] \;\le\; \mathcal{O}\left(\frac{\sqrt{\omega G^2+\sigma^2}}{\sqrt{nT}} + \zeta^2\right), \tag{9}$$

where $\mathcal{O}(\cdot)$ *hides dependence on the problem constants* Δ_0 *and* L*.*

The proof, which uses the L-smooth descent lemma and controls the bias via Young's inequality, is in Appendix F. We next specialize this result to the two SSTQ variants.

Corollary 6.2 (Convergence under Flat Randomized Response)**.** *SSTQ with Flat RR is unbiased* ($\zeta = 0$) *by Theorem 3.1 and* $\omega = \mathcal{O}\Big(\frac{dK^2}{\rho}[1 + 2^b/(\epsilon \wedge \epsilon^2)]\Big)$ *under the condition in Theorem 4.3. Hence*

$$\min_{0\le t\le T-1} \mathbb{E}[\|\nabla F(w_t)\|_2^2] \le \mathcal{O}\left(\frac{1}{T} + \frac{\sqrt{\frac{dK^2}{\rho}[1 + 2^b/(\epsilon \wedge \epsilon^2)]G^2 + \sigma^2}}{\sqrt{nT}}\right). \tag{10}$$

When the $\mathcal{O}(2^b)$ penalty becomes a computational bottleneck at higher bit-widths, we switch to the Metric-Aware Laplace mechanism with identity decoding, trading absolute unbiasedness for a smaller variance bound.

Corollary 6.3 (Convergence under Metric-Aware Mechanism)**.** *Under the max-gap condition and fixed row-coherence constant stated in Theorem 5.2, the metric-aware mechanism has* $\omega = \mathcal{O}((dK^2/\rho)(1 + \epsilon^{-1} + \epsilon^{-2}))$*, and the bias bound* $\zeta \le 8KG/(\sqrt{\rho}\,\epsilon)$*. Hence*

$$\min_{0\le t\le T-1} \mathbb{E}[\|\nabla F(w_t)\|_2^2] \le \mathcal{O}\left(\frac{1}{T} + \frac{\sqrt{\frac{dK^2}{\rho}G^2(1 + \epsilon^{-1} + \epsilon^{-2}) + \sigma^2}}{\sqrt{nT}} + \frac{K^2G^2}{\rho\epsilon^2}\right). \tag{11}$$

7 Neural Network Experiments

We evaluate the LDP mechanisms in a federated learning setting with neural network models on two standard image classification benchmarks: Fashion-MNIST and CIFAR-10.

Experimental Setup. We train a fully connected neural network with a single hidden layer of 32 units and ReLU activations: for Fashion-MNIST (10 classes, 784-dimensional inputs) the model has $d = 25{,}450$ parameters, and for CIFAR-10 (10 classes, 3,072-dimensional inputs) it has $d = 98{,}666$ parameters. For all Kashin-based methods (SQKR, SSTQ), we construct the frame using a randomized partial Hadamard transform of size $N = 2^{\lceil \log_2(2.5\,d)\rceil}$. We simulate a federated setting with $W = 100$ workers, each holding a disjoint partition of the training data. At each of the $T = 100$ rounds, every worker computes a stochastic gradient on a mini-batch of size 64, clips it to $\|g\|_2 \le C = 0.1$, applies the LDP mechanism, and transmits the privatized message; the server aggregates, applies a gradient norm cap of 10.0 for stability, and updates the global model with a constant learning rate ($\eta = 0.2$ for Fashion-MNIST, $\eta = 0.5$ for CIFAR-10). All methods use pure ϵ-LDP with per-round $\epsilon = 3$, and SSTQ uses bit-width $b = 4$. For fair comparison, a common random seed fixes mini-batch selection across all methods, so every method trains on exactly the same sequence of batches and the only differing randomness is the LDP noise; test accuracy is evaluated each round on the full held-out test set (10,000 images). Appendix G gives additional details as well as remarks on the effect of clipping and the cumulative privacy budget.

Results. Figure 1 presents the training loss, test accuracy, and accuracy-vs-cost tradeoff for both datasets; per-client per-round communication cost is summarized in Table 2 of Appendix G.

Fashion-MNIST. On training loss (Figure 1, top-left), Clean SGD and PrivUnit decrease steadily to ≈ 1.2 and ≈ 1.25; SQKR and SSTQ (Flat-RR) settle slightly higher, around 1.3, reflecting added variance from quantization and privacy noise, while SSTQ (Metric-Aware) reaches ≈ 1.7, consistent

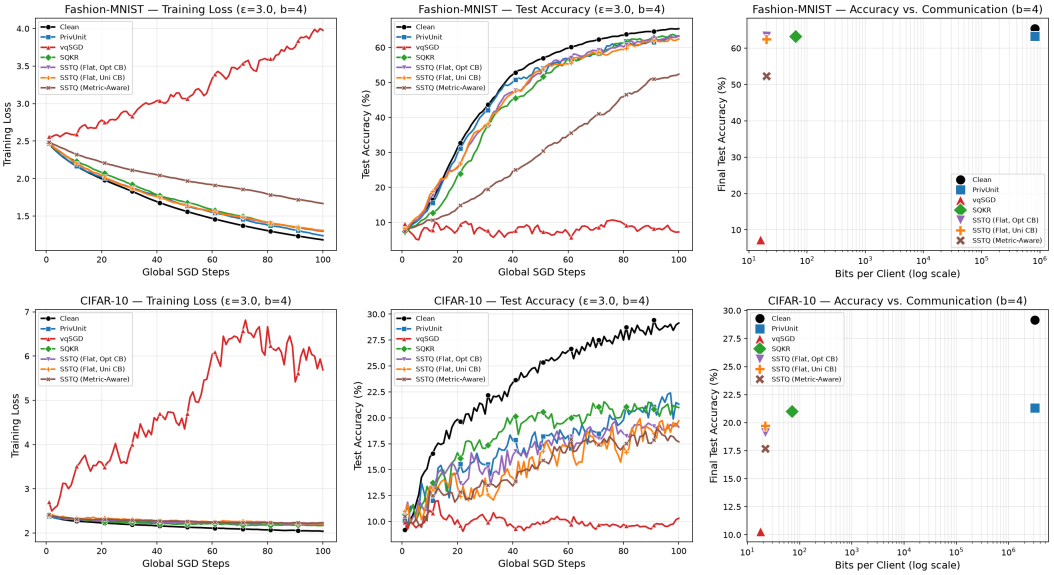

Figure 1: Neural network experiments under ϵ-LDP ($\epsilon = 3$, $b = 4$, $W = 100$ workers). **Top:** Fashion-MNIST ($d = 25{,}450$); **bottom:** CIFAR-10 ($d = 98{,}666$). **Left:** training cross-entropy loss vs. SGD steps; **center:** test accuracy vs. SGD steps; **right:** final test accuracy vs. bits per client (log scale).

with the small bias ζ^2 of the identity decoder (Corollary 6.3). By contrast vqSGD is *divergent*, its loss rising from 2.5 to ≈ 4.0, confirming that its $O(d^3/\epsilon^2)$ variance scaling is prohibitive at $d = 25{,}450$. On test accuracy (top-center), Clean SGD reaches 65.4% and PrivUnit tracks closely at 63.3%; among quantized methods SQKR attains 63.2% using 68 bits per client, whereas SSTQ (Flat-RR) attains 63.3% with the optimized codebook and 62.4% with the uniform one using only 20 bits, a $3.2\times$ reduction. All three quantized methods are thus within 1% of PrivUnit, while vqSGD fails to learn (7.2%, near chance). SSTQ (Metric-Aware) trails Flat-RR here because of its bias; its advantage emerges at larger bit widths b, as shown in Figure 4 of Appendix H.

CIFAR-10. The higher dimension ($d = 98{,}666$) amplifies variance. For training loss (bottom-left), Clean SGD and PrivUnit steadily decrease while SQKR and the SSTQ variants remain around 2.2, and vqSGD again diverges dramatically, reaching ≈ 5.7. On test accuracy (bottom-center), Clean SGD reaches 29.1% and PrivUnit 21.3%; among quantized methods SQKR achieves 21.0% (76 bits), then SSTQ (Flat-RR, uniform codebook) at 19.7% and SSTQ (Flat-RR, optimized codebook) at 19.1% (both 22 bits), SSTQ (Metric-Aware) at 17.7%, and vqSGD at 10.3% (18 bits).

Accuracy vs. Communication Cost. The rightmost panels of Figure 1 and Table 2 (Appendix G) summarize the accuracy-versus-bits trade-off: SSTQ matches SQKR's accuracy at a fraction of the payload. On Fashion-MNIST, SSTQ (Flat-RR, optimized) achieves 63.3%—essentially matching SQKR's 63.2%—using $3.4\times$ fewer bits (20 vs. 68); on CIFAR-10 it reaches 19.1% against SQKR's 21.0%, a gap of only 2%, using $3.45\times$ fewer bits (22 vs. 76). This arises because SQKR sends $k = \min\{\lceil \epsilon \log_2 e \rceil, b_0\}$ coordinates at $\lceil \log_2 N \rceil + 1$ bits each, whereas SSTQ sends one coordinate index plus a b-bit level, i.e. $\lceil \log_2 N \rceil + b$ bits total. vqSGD uses the fewest bits (16–18) but fails to converge, confirming its cubic variance scaling is impractical in high dimensions, while PrivUnit is the most accurate private method but needs full-precision transmission (>800K bits); its $32d$ cost reflects float32 encoding, and lower precision (e.g., $16d$ or $8d$) shrinks the constant but not the $\Theta(d)$ scaling inherent to any mechanism emitting a d-dimensional vector. Per-client computational cost is compared in Appendix G.

Conclusion and Limitations. SSTQ is a principled framework for private quantization. By combining discrete quantization with adaptive, privacy-aware codebook optimization, it avoids the variance penalties associated with fixed-geometry schemes. Its current formulation assumes a fixed communication budget and known data range; future work could address nonstationary streams and develop robust, range-agnostic adaptive mechanisms for more flexible real-world deployment.

AI USE STATEMENT

In this work, we used generative AI tools, specifically the Long Proof harness in Google Antigravity 2.0, to assist in developing initial drafts of certain analytical proofs. These proofs were subsequently reviewed, refined, and finalized through human intervention. Additionally, we used generative AI tools to improve the presentation and writing of the manuscript. All AI-assisted material was carefully reviewed and verified by the authors. We accept full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

REFERENCES

Hilal Asi, Vitaly Feldman, and Kunal Talwar. Optimal algorithms for mean estimation under local differential privacy. In *International Conference on Machine Learning*, pp. 1046–1056. PMLR, 2022.

Hilal Asi, Vitaly Feldman, Jelani Nelson, Huy Nguyen, and Kunal Talwar. Fast optimal locally private mean estimation via random projections. *Advances in Neural Information Processing Systems*, 36: 16271–16282, 2023.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in neural information processing systems*, 31, 2018.

Ran Ben-Basat, Amit Portnoy, Gil Einziger, Yaniv Ben-Itzhak, Michael Mitzenmacher, et al. Accelerating federated learning with quick distributed mean estimation. In *Forty-first International Conference on Machine Learning*, 2024.

Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018.

Wei-Ning Chen, Peter Kairouz, and Ayfer Ozgur. Breaking the communication-privacy-accuracy trilemma. *Advances in Neural Information Processing Systems*, 33:3312–3324, 2020.

Wei-Ning Chen, Peter Kairouz, and Ayfer Ozgur. Breaking the dimension dependence in sparse distribution estimation under communication constraints. In *Conference on Learning Theory*, pp. 1028–1059. PMLR, 2021.

Wei-Ning Chen, Dan Song, Ayfer Ozgur, and Peter Kairouz. Privacy amplification via compression: Achieving the optimal privacy-accuracy-communication trade-off in distributed mean estimation. *Advances in Neural Information Processing Systems*, 36:69202–69227, 2023.

Wei-Ning Chen, Berivan Isik, Peter Kairouz, Albert No, Sewoong Oh, and Zheng Xu. Improved communication-privacy trade-offs in ℓ_2 mean estimation under streaming differential privacy. *arXiv preprint arXiv:2405.02341*, 2024.

John Duchi and Ryan Rogers. Lower bounds for locally private estimation via communication complexity. In *Conference on Learning Theory*, pp. 1161–1191. PMLR, 2019.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Venkata Gandikota, Daniel Kane, Raj Kumar Maity, and Arya Mazumdar. vqSGD: Vector quantized stochastic gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pp. 2195–2205. PMLR, 2021.

Jianyang Gao and Cheng Long. Rabitq: Quantizing high-dimensional vectors with a theoretical error bound for approximate nearest neighbor search. *Proceedings of the ACM on Management of Data*, 2(3):1–27, 2024.

Shuli Jiang, Pranay Sharma, and Gauri Joshi. Correlation aware sparsified mean estimation using random projection. *Advances in Neural Information Processing Systems*, 36:53892–53923, 2023.

Peter Kairouz and H Brendan McMahan. Advances and open problems in federated learning. *Foundations and trends in machine learning*, 14(1-2):1–210, 2021.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International Conference on Machine Learning*, pp. 1376–1385. PMLR, 2015.

Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Natalie Lang, Elad Sofer, Tomer Shaked, and Nir Shlezinger. Joint privacy enhancement and quantization in federated learning. *IEEE Transactions on Signal Processing*, 71:295–310, 2023.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

Zhize Li, Haoyu Zhao, Boyue Li, and Yuejie Chi. SoteriaFL: A unified framework for private federated learning with communication compression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yurii Lyubarskii and Roman Vershynin. Uncertainty principles and vector quantization. *IEEE Transactions on Information Theory*, 56(7):3491–3501, 2010.

PF Panter and Wu Dite. Quantization distortion in pulse-count modulation with nonuniform spacing of levels. *Proceedings of the IRE*, 39(1):44–48, 1951.

Ananda Theertha Suresh, Ziteng Sun, Jae Ro, and Felix Yu. Correlated quantization for distributed mean estimation and optimization. In *International Conference on Machine Learning*, pp. 20856–20876. PMLR, 2022.

Harsh Vardhan and Arya Mazumdar. Collaborative compressors in distributed mean estimation with limited communication budget. *arXiv preprint arXiv:2601.18950*, 2026.

Shay Vargaftik, Ran Ben Basat, Amit Portnoy, Gal Mendelson, Yaniv Ben-Itzhak, and Michael Mitzenmacher. Drive: One-bit distributed mean estimation. *Advances in Neural Information Processing Systems*, 34:362–377, 2021.

Shay Vargaftik, Ran Ben Basat, Amit Portnoy, Gal Mendelson, Yaniv Ben-Itzhak, and Michael Mitzenmacher. Eden: Communication-efficient and robust distributed mean estimation for federated learning. In *International Conference on Machine Learning*, pp. 21984–22014. PMLR, 2022.

Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American statistical association*, 60(309):63–69, 1965.

Amir Zandieh, Majid Daliri, Majid Hadian, and Vahab Mirrokni. Turboquant: Online vector quantization with near-optimal distortion rate. In *International Conference on Learning Representations*, volume 2026, pp. 56418–56439, 2026.

A Related Work

A.1 Distinctions from Prior Kashin-Based LDP Mechanisms

Overcomplete tight frames (Kashin representations) have been used in prior work, notably SQKR (Chen et al., 2020), to control sensitivity under joint privacy and communication constraints. While SQKR achieves the optimal $\mathcal{O}(d/(\epsilon \wedge \epsilon^2))$ mean squared error scaling, its design imposes several structural limitations. SSTQ differs in three key aspects:

First, SSTQ removes the effective bandwidth restriction present in SQKR. In SQKR, dense encoding requires transmitting one bit per Kashin coefficient, leading to $\Theta(d)$ bits. In the subsampled setting, grouping k coordinates with Flat Randomized Response incurs an exponential variance penalty $\mathcal{O}(4^k)$ (see Claim C.1 in (Chen et al., 2020)), which forces k to be capped by $\min(\lceil \epsilon \log_2 e \rceil, b_0)$. In contrast, SSTQ allocates the full b-bit budget to a single coordinate. In its metric-aware variant, it operates over a one-dimensional codebook, thereby fully utilizing the available bandwidth without incurring exponential penalties.

Second, SSTQ replaces 1-bit quantization with a continuous surrogate optimization. SQKR maps Kashin coefficients to their extreme values prior to privatization, which fixes the baseline variance at a high level. SSTQ instead performs stochastic quantization over a multi-bit codebook optimized for the privatized setting, effectively reducing quantization error while accommodating LDP noise.

Third, SSTQ operates as a private-coin protocol with explicit communication cost. SQKR relies on shared randomness to coordinate subsampling; without it, transmitting indices increases the communication cost to $\mathcal{O}(k \log d)$ to send the indices of the k sampled coordinates. SSTQ directly transmits the sampled index along with the quantized value, resulting in a fixed cost of $\lceil \log_2 N \rceil + b$ bits, independent of shared randomness assumptions.

A.2 LDP Mean Estimation under Communication Constraints

The fundamental tension between Local Differential Privacy (LDP) and extreme communication bandwidth in distributed mean estimation has motivated a rich recent literature. Our Subsampled Stochastic TurboQuant (SSTQ) framework differs from prior approaches along three axes: privacy threat model, quantization efficiency, and geometric optimization.

Geometric Quantization and Dimensional Curses. Foundational mechanisms such as PrivUnit and PrivUnitG (Bhowmick et al., 2018; Asi et al., 2022) achieve the information-theoretic optimal LDP variance of $\mathcal{O}(d/\epsilon^2)$ but require transmitting uncompressed dense continuous vectors. By contrast, geometric quantizers such as vqSGD (Gandikota et al., 2021) reduce bandwidth logarithmically by mapping vectors to high-dimensional cross-polytopes. However, this rigid geometry shrinks the LDP probability gap and triggers the $\Theta(d^3/\epsilon^2)$ variance curse proven in Appendix B. While recent work has broken dimension dependence for *sparse* discrete distribution estimation under communication constraints (Chen et al., 2021), SSTQ achieves a similar breakthrough for dense, high-dimensional continuous gradient quantization.

Random Projections and Subspace LDP. To reduce the $\mathcal{O}(d)$ communication bottleneck of optimal LDP mechanisms, methods such as ProjUnit (Asi et al., 2023) project gradients into a lower k-dimensional subspace before applying continuous LDP. Although ProjUnit achieves order-optimal expected MSE, it still outputs continuous vectors that require further quantization and relies on dense matrix-vector multiplications during decoding. SSTQ instead natively unifies LDP with discrete quantization: by coupling scaled Parseval frames with 1-sparse subsampling, it reduces the problem to exact 1D scalar quantization, with payload $\lceil \log_2 N \rceil + b$ bits.

Central/Shuffle DP and Coordinate Subsampling. Coordinate subsampling has also been used to amplify privacy under relaxed trust models. The Coordinate Subsampled Gaussian Mechanism (CSGM) (Chen et al., 2023) combines Kashin's representation with random subsampling to achieve optimal $\mathcal{O}(d/n^2\epsilon^2)$ error for the estimation of the aggregated mean across n clients under Central/Shuffle DP, but it depends on a trusted central server or a secure shuffler. SSTQ instead operates in the strictly stronger pure ϵ-Local DP setting, providing message-level plausible deniability without trusted aggregation.

Geometric Constants and Continuous Optimization. The streaming variant L_2-CSGM (Chen et al., 2024) shows that pushing L_2 geometries through L_∞ Kashin bounds with rigid 1-bit quantization, as in SQKR (Chen et al., 2020), leads to suboptimal MSE constants. SSTQ demonstrates that this limitation stems from the rigid binarization, not from the overcomplete frame itself: by replacing 1-bit quantization with a multi-bit continuous surrogate geometry optimized via $\mathcal{L}_{\text{SSTQ}}$, it factors out these spatial penalties and reduces the codebook-dependent MSE scaling from $\mathcal{O}(4^b)$ to $\mathcal{O}(2^b)$ (Theorem 4.3), all within pure LDP.

Joint Quantization–Privacy Co-Design. A complementary line of work co-designs compression and privatization rather than composing them. JoPEQ (Lang et al., 2023) applies universal vector quantization over a shared random lattice with subtractive dithering, so that the quantization distortion is statistically equivalent to additive noise, and augments this distortion with dedicated multivariate Laplace-type noise to meet a target LDP budget, whereas SoteriaFL (Li et al., 2022) pairs a general family of (shifted) compression operators with Gaussian-mechanism privatization and characterizes the resulting privacy–utility–communication trade-offs, improving on the naive CDP-SGD composition of compression with DP-SGD. Both are complementary to our setting but differ in threat model and rate: SoteriaFL's Gaussian noise yields the (ϵ, δ) relaxation rather than pure ϵ-LDP, while JoPEQ's subtractive dither presumes a lattice shared between client and server, a public-coin protocol, with a rate specified per entry and hence $\Theta(d)$ uplink bits. SSTQ instead privatizes a *finite* label alphabet under pure ϵ-LDP in the private-coin model at a fixed cost of $\lceil \log_2 N \rceil + b$ bits per message, and, unlike a lattice or compression operator fixed a priori, optimizes the codebook geometry itself against the privatized MSE (Section 4).

B Curse of Dimensionality in Fixed-Geometry LDP

Previous studies in geometric vector quantization, most notably vqSGD (Gandikota et al., 2021), have claimed that quantizing vectors to a cross-polytope and applying Randomized Response yields a variance of $\mathcal{O}(d^2/\epsilon^2)$ for the proposed unbiased estimator. We challenge this assertion and demonstrate that this formulation actually incurs a variance of $\Theta(d^3/\epsilon^2)$. This fundamental scaling issue critically impairs the performance of high-dimensional gradient embeddings in privacy-constrained settings.

Formalization of vqSGD and Cross-Polytope LDP. Geometric vector quantization methods, such as vqSGD, typically map continuous vectors $x \in \mathbb{S}^{d-1}$ to the vertices of a discrete geometric object. A canonical choice is the cross-polytope codebook, defined as:

$$\mathcal{C}_{cp} = \{\pm\sqrt{d}e_i\}_{i=1}^{d}, \tag{12}$$

where e_i is the i-th standard basis vector in $\mathbb{R}^d$. Note that the total size of this discrete domain is exactly $M = 2d$ vertices.

To achieve pure ϵ-Local Differential Privacy, vqSGD applies a Flat Randomized Response (RR) to a selected vertex $v \in \mathcal{C}_{cp}$. The mechanism outputs a noisy vertex $z \in \mathcal{C}_{cp}$ such that the transition probabilities are:

$$p = \Pr(z = v \mid v) = \frac{e^\epsilon}{e^\epsilon + 2d - 1}, \tag{13}$$

$$q = \Pr(z = c \neq v \mid v) = \frac{1}{e^\epsilon + 2d - 1}. \tag{14}$$

Using the fact that the sum of vertices is zero, the server debiases the output z using a simple scaling factor of $\frac{1}{p-q}$ and outputs the estimator $\hat{x} = \frac{z}{p-q}$.

The High-Dimensional Variance Growth. Many analyses conceptually treat the unbiasing scalar $\frac{1}{(p-q)^2}$ as an $\mathcal{O}(1/\epsilon^2)$ constant when evaluating the expected Mean Squared Error (MSE). We explicitly show that because the codebook size $M = 2d$ is intertwined with the ambient dimension, the probability gap $p - q$ shrinks as $\mathcal{O}(1/d)$.

Theorem B.1 (Exact cross-polytope RR exhibits the dimensionality curse)**.** *For any input vector $x \in \mathbb{S}^{d-1}$ quantized via the Cross-Polytope $\mathcal{C}_{cp}$ and perturbed via Flat Randomized Response for pure ϵ-LDP, the exact Mean Squared Error (variance) of the unbiased estimator scales asymptotically*

as:

$$\mathit{MSE}_{\mathrm{vqSGD}} = d - 1 + \frac{4d^2}{e^\epsilon - 1} + \frac{4d^3}{(e^\epsilon - 1)^2}. \tag{15}$$

Consequently it is $\Theta_\epsilon(d^3)$ for fixed $\epsilon > 0$, and $\Theta(d^3/\epsilon^2)$ jointly when $0 < \epsilon \leq 1$. If ϵ grows faster than $\log d$, the error instead approaches the nonprivate stochastic-quantization rate of $\Theta(d)$.

Before proceeding to the proof, we provide a remark on alternative polytopes, noting that they are subject to analogous limitations.

Remark B.2 (Scope beyond the cross-polytope)**.** *The same cubic conclusion applies to other codebook families, provided they share three ingredients: $M = \Theta(d)$ labels, codeword squared norm $\Theta(d)$, and a centered RR decoder whose debiasing gap is $\Theta(1/d)$ at fixed privacy. Conversely, mapping to dense uniform hypercubes bypasses this geometric penalty but necessitates transmitting $\mathcal{O}(d)$ bits per client, which fails to satisfy the fundamental requirements of federated communication compression.*

Proof of Theorem B.1. Let $x \in \mathbb{S}^{d-1}$ be an input unit vector. Assume the client stochastically quantizes x to a true vertex v in the cross-polytope codebook $\mathcal{C}_{cp} = \{\pm\sqrt{d}e_i\}_{i=1}^d$ such that $\mathbb{E}[v \mid x] = x$. The pure ϵ-LDP Flat Randomized Response mechanism then outputs a noisy token $z \in \mathcal{C}_{cp}$ given v. The server constructs the unbiased estimator $\hat{x} = \frac{z}{p-q}$.

To compute the Mean Square Error (MSE), we use the unbiasedness property of the estimator to write $\mathbb{E}[\hat{x}] = \mathbb{E}_v[\mathbb{E}_{z|v}[\frac{z}{p-q} \mid v] \mid x] = \mathbb{E}[v] = x$, and therefore:

$$\begin{aligned}
\mathrm{MSE}_{\mathrm{vqSGD}} &= \mathbb{E}[\|\hat{x} - x\|_2^2] \\
&= \mathbb{E}[\|\hat{x}\|_2^2] - 2\mathbb{E}[\langle \hat{x}, x\rangle] + \|x\|_2^2 \\
&= \mathbb{E}[\|\hat{x}\|_2^2] - 2\langle x, x\rangle + \|x\|_2^2 \\
&= \mathbb{E}[\|\hat{x}\|_2^2] - \|x\|_2^2 \\
&= \mathbb{E}\left[\left\|\frac{z}{p-q}\right\|_2^2\right] - 1 \\
&= \frac{1}{(p-q)^2}\mathbb{E}[\|z\|_2^2] - 1.
\end{aligned} \tag{16}$$

By definition of the cross-polytope codebook $\mathcal{C}_{cp}$, every vertex $z \in \mathcal{C}_{cp}$ takes the form $\pm\sqrt{d}e_i$, and so $\|z\|_2^2 = d$.

Next, we evaluate the unbiasing scalar. The cross-polytope $\mathcal{C}_{cp}$ consists of exactly $M = 2d$ discrete vertices. Under Flat Randomized Response for pure ϵ-LDP over M elements, the transition probabilities are defined as:

$$p = \frac{e^\epsilon}{e^\epsilon + 2d - 1}, \quad \text{and} \quad q = \frac{1}{e^\epsilon + 2d - 1}. \tag{17}$$

The probability gap evaluates to:

$$p - q = \frac{e^\epsilon - 1}{e^\epsilon + 2d - 1}. \tag{18}$$

Substituting this multiplier and $\mathbb{E}[\|z\|_2^2] = d$ directly back into Equation equation 16:

$$\mathrm{MSE}_{\mathrm{vqSGD}} = \frac{d}{(p-q)^2} - 1 = d\left(\frac{e^\epsilon + 2d - 1}{e^\epsilon - 1}\right)^2 - 1. \tag{19}$$

We expand the numerator by writing $(e^\epsilon + 2d - 1)^2$ as $((e^\epsilon - 1) + 2d)^2$:

$$\begin{aligned}\text{MSE}_{\text{vqSGD}} &= \frac{d}{(e^\epsilon - 1)^2}\left((e^\epsilon - 1)^2 + 4d(e^\epsilon - 1) + 4d^2\right) - 1 \\ &= d\left(1 + \frac{4d}{e^\epsilon - 1} + \frac{4d^2}{(e^\epsilon - 1)^2}\right) - 1 \\ &= d - 1 + \frac{4d^2}{e^\epsilon - 1} + \frac{4d^3}{(e^\epsilon - 1)^2}.\end{aligned}$$

For fixed $\epsilon > 0$, the cubic term dominates as $d \to \infty$, giving $\Theta_\epsilon(d^3)$. Uniformly for $0 < \epsilon \leq 1$, $e^\epsilon - 1 = \Theta(\epsilon)$ and the expression is $\Theta(d^3/\epsilon^2)$. This completes the proof.

□

C Proofs for the SSTQ Mechanism

This appendix provides the formal mathematical proofs for the theoretical guarantees of the Subsampled Stochastic TurboQuant (SSTQ) architecture with Flat randomization response introduced in Section 3. We first prove that SSTQ satisfies pure ϵ-Local Differential Privacy and guarantees an unbiased gradient estimator. We then establish a bound on the conditional variance of the 1D randomized response mechanism, by which we prove that SSTQ achieves the information-theoretic optimal $\mathcal{O}(d/\epsilon^2)$ Mean Squared Error.

C.1 Proof of Strict ϵ-LDP and Unbiasedness

Proof of Theorem 3.1. We establish the privacy and utility guarantees in two distinct parts.

Part 1: Strict pure ϵ-LDP. The SSTQ client encoding algorithm transmits a pair (J, Z), where $J \in [N]$ is the subsampled coordinate index and $Z \in [M]$ is a randomized-reponse label. First, the subsampling index J is generated uniformly at random from $[N]$, totally independent of the client's private data x. Because the distribution of J does not depend on x, its transmission consumes exactly $\epsilon = 0$ privacy budget.

Second, the stochastic quantization step maps the continuous Kashin coordinate y_J into a finite, discrete codebook $\Gamma = \{c_1, \ldots, c_M\}$ bounded in $[-B, B]$, yielding a codeword label $V \in [M]$. This step is a local data-independent mapping applied to the already bounded coordinate.

Third, the algorithm applies Flat Randomized Response over the constant domain $M = 2^b$ to produce the codeword label $Z \in [M]$. By the definition of the mechanism, the transition probabilities are $p = \frac{e^\epsilon}{e^\epsilon + M - 1}$ if $Z = V$, and $q = \frac{1}{e^\epsilon + M - 1}$ if $Z \neq V$. For any two possible labels $i, i' \in [M]$ and any label $\ell \in [M]$, the maximum probability ratio is strictly bounded by:

$$\frac{\Pr(Z = \ell \mid V = i)}{\Pr(Z = \ell \mid V = i')} \leq \frac{p}{q} = \frac{\frac{e^\epsilon}{e^\epsilon + M - 1}}{\frac{1}{e^\epsilon + M - 1}} = e^\epsilon. \tag{20}$$

Because the maximum probability ratio between any two inputs is exactly e^ϵ, the mechanism satisfies pure ϵ-LDP on the client device, without requiring a trusted shuffler or shared public coin.

Part 2: Unbiasedness. We trace the expectations backwards from the server's estimator $\hat{x}$ to the true vector x. First, the server receives (J, Z) and constructs $\widetilde{y}_J = \frac{c_Z}{p - q}$. Given the zero-mean codebook constraint ($\sum_{i=1}^M c_i = 0$), the expected value of c_Z given the interpolated label V is:

$$\mathbb{E}[c_Z \mid V = i] = pc_i + q\sum_{\ell \neq i} c_\ell = pc_i + q\left(\sum_{\ell=1}^M c_\ell - c_i\right) = pc_i + q(0 - c_i) = (p - q)c_i. \tag{21}$$

Thus, $\mathbb{E}[\widetilde{y}_J \mid V = i] = \mathbb{E}\left[\frac{c_Z}{p-q} \;\middle|\; V = i\right] = c_i$.

Second, the label V is generated via exact linear stochastic interpolation of the coordinate $y_J \in [c_k, c_{k+1}]$. The expected value is:

$$\mathbb{E}[c_V \mid y_J] = c_{k+1}\Pr(V = k{+}1){+}c_k \Pr(V = k) = c_{k+1}\left(\frac{y_J - c_k}{c_{k+1} - c_k}\right){+}c_k\left(\frac{c_{k+1} - y_J}{c_{k+1} - c_k}\right) = y_J. \tag{22}$$

Third, taking the expectation over the uniform oblivious subsampling index $J \in [N]$, the server's reconstructed 1-sparse vector $\widetilde{y} = N\widetilde{y}_J e_J$ yields:

$$\mathbb{E}_J[\widetilde{y}] = \sum_{j=1}^{N} \Pr(J = j)(N y_j e_j) = \sum_{j=1}^{N} \frac{1}{N}(N \cdot y_j \cdot e_j) = \sum_{j=1}^{N} y_j e_j = y. \tag{23}$$

Finally, the server projects $\widetilde{y}$ through the scaled-Parseval frame U. Because the Kashin representation guarantees exact reconstruction $\frac{d}{N}U^T y = x$ (Theorem 2.3), applying linearity of expectation yields:

$$\mathbb{E}[\hat{x}] = \mathbb{E}\left[\frac{d}{N}U^T\widetilde{y}\right] = \frac{d}{N}U^T\mathbb{E}[\widetilde{y}] = \frac{d}{N}U^T y = x. \tag{24}$$

The global estimator is therefore strictly unbiased. □

C.2 MSE BOUND FOR SSTQ (FLAT-RR)

We begin with an elementary trace identity for scaled Parseval frames (Definition 2.2), used repeatedly below.

Lemma C.1 (Aggregate frame energy)**.** *Every scaled Parseval frame satisfies*

$$\sum_{j=1}^{N} \|u_j\|_2^2 = \mathrm{Tr}(U^\top U) = N.$$

Proof of Lemma C.1. Consider the trace of the matrix $UU^\top \in \mathbb{R}^{N\times N}$. By the cyclic permutation property of the trace operator:

$$\mathrm{Tr}(UU^\top) = \mathrm{Tr}(U^\top U). \tag{25}$$

We are given $U^\top U = \frac{N}{d} I_d$. Therefore:

$$\mathrm{Tr}(U^\top U) = \mathrm{Tr}\left(\frac{N}{d} I_d\right) = \frac{N}{d} \times d = N. \tag{26}$$

The diagonal elements of the outer product $UU^\top$ are the squared Euclidean norms of the rows of U. Thus, $\sum_{j=1}^{N} \|u_j\|_2^2 = \mathrm{Tr}(UU^\top) = N$. □

To formalize how SSTQ addresses the curse of dimensionality, we first bound the conditional variance of the 1D randomized response mechanism operating over the bounded Kashin coefficients.

Lemma C.2. *For any stochastically quantized client token bounded by $|c_V| \le B = \frac{KR}{\sqrt{N}}$, the server's debiased scalar satisfies:*

$$\mathbb{E}[\widetilde{y}_J^2 \mid V] \le \frac{B^2}{(p-q)^2}. \tag{27}$$

Because the discrete domain size M over which the randomized response mechanism operates is now defined by the constant payload bit-width ($M = 2^b$), the unbiasing probability gap $(p - q)$ is structurally decoupled from the high-dimensional geometry. This decoupling enables to bypass the curse of dimensionality.

Proof of Lemma C.2. We bound the uncentered conditional second moment $\tilde{y}_J = \frac{c_Z}{p-q}$. The second moment of the Flat Randomized Response output c_Z is:

$$\mathbb{E}[c_Z^2 \mid V = i] = p c_i^2 + q \sum_{\ell \neq i} c_\ell^2 = (p - q)c_i^2 + q \sum_{\ell=1}^{M} c_\ell^2. \tag{28}$$

Dividing by the unbiasing scalar squared $(p-q)^2$ gives:

$$\mathbb{E}[\widetilde{y}_J^2 \mid V=i] = \mathbb{E}\left[\left(\frac{c_Z}{p-q}\right)^2 \,\middle|\, V=i\right] = \frac{1}{p-q}c_i^2 + \frac{q}{(p-q)^2}\sum_{\ell=1}^{M} c_\ell^2. \tag{29}$$

By $|c_i|, |c_\ell| \le B$, substituting the common bound B^2 gives

$$\mathbb{E}[\widetilde{y}_J^2 \mid V] \le \frac{B^2}{p-q} + \frac{qMB^2}{(p-q)^2} = \frac{B^2}{(p-q)^2}, \tag{30}$$

where we used that the randomization probabilities sum to one: $p + (M-1)q = 1$.

□

We next proceed to bound the MSE achieved by SSTQ (Flat-RR variant).

Proof of Theorem 3.2. We evaluate the expected Mean Squared Error (total variance) of the full d-dimensional reconstructed gradient $\hat{x}$. Let $U \in \mathbb{R}^{N\times d}$ be a scaled Parseval frame where $U^TU = \frac{N}{d}I_d$.

The server reconstructs the target gradient via the projection $\hat{x} = \frac{d}{N}U^T\widetilde{y}$. The 1-sparse vector is $\widetilde{y} = N\widetilde{y}_j e_j$. Thus, the projection evaluates to:

$$\hat{x} = \frac{d}{N}U^T(N\widetilde{y}_j e_j) = d\widetilde{y}_j U^T e_j = d\widetilde{y}_j u_j, \tag{31}$$

where u_j^T is the j-th row of the frame matrix U.

We calculate the expected squared ℓ_2 norm of the estimator $\hat{x}$:

$$\mathbb{E}[\|\hat{x}\|_2^2] = \mathbb{E}\left[\|d\widetilde{y}_J u_J\|_2^2\right] = d^2\mathbb{E}[\widetilde{y}_J^2\,\|u_J\|_2^2]. \tag{32}$$

applying Lemmas C.2 and C.1,

$$\mathbb{E}\|\hat{x}\|_2^2 = \frac{d^2}{N}\sum_{j=1}^{N}\|u_j\|_2^2\mathbb{E}[\tilde{y}_j^2 \mid J=j] \le \frac{d^2B^2}{(p-q)^2} = \frac{dK^2R^2}{\rho(p-q)^2}. \tag{33}$$

Because the estimator is unbiased, the MSE is $\mathbb{E}\|\hat{x}\|_2^2 - \|x\|_2^2$. For $\|x\|_2 = R$, Equation equation 33 gives

$$\mathrm{MSE}_{\mathrm{SSTQ}} \le \frac{dK^2R^2}{\rho}\left(1 + \frac{M}{e^\epsilon - 1}\right)^2 - R^2. \tag{34}$$

Finally, we evaluate the unbiasing scalar. Because the algorithm operates over a 1D scalar codebook, the codebook size $M = 2^b$ is an $\mathcal{O}(1)$ constant completely independent of d. The probability gap $(p-q)$ for Flat Randomized Response is:

$$p - q = \frac{e^\epsilon - 1}{e^\epsilon + 2^b - 1}. \tag{35}$$

Therefore,

$$\frac{1}{(p-q)^2} = \left(1 + \frac{2^b}{e^\epsilon - 1}\right)^2 < 1 + \frac{2^{b+1}}{\epsilon} + \frac{4^b}{\epsilon^2} \le 1 + 4^b\left(\frac{1}{\epsilon} + \frac{1}{\epsilon^2}\right) \le 1 + \frac{2\cdot 4^b}{\epsilon \wedge \epsilon^2}.$$

Using this bound into equation 34, we get

$$\mathrm{MSE}_{\mathrm{SSTQ}} = \mathcal{O}\left(\frac{dK^2R^2}{\rho}\left(1 + \frac{4^b}{\epsilon \wedge \epsilon^2}\right)\right),$$

which completes the proof.

□

D Proofs for Surrogate Codebook Optimization

This appendix provides the analytical derivations and formal proofs for the surrogate continuous codebook optimization framework ($\mathcal{L}_{SSTQ}$) introduced in Section 4.

D.1 Derivation of the Surrogate Minimax Objective

We start by deriving the exact closed-form relation between the expected Mean Squared Error of the Local Differential Privacy mechanism and the continuous codebook geometry.

Derivation of Equation equation 2 and Definition 4.1. Let $y \in [-B, B]$ be a bounded Kashin coordinate, and assume it lies in the interval $[c_k, c_{k+1}]$ between two adjacent centroids in the codebook Γ. As defined in Algorithm 1, y is stochastically quantized to a codeword with label $V \in \{c_k, c_{k+1}\}$ via exact linear interpolation. The conditional variance of this stochastic interpolation step is:

$$\begin{aligned} r_\Gamma(y) &:= \mathrm{Var}(c_V \mid Y = y) = \mathbb{E}[c_V^2 \mid y] - y^2 \\ &= c_{k+1}^2 \left(\frac{y - c_k}{c_{k+1} - c_k} \right) + c_k^2 \left(\frac{c_{k+1} - y}{c_{k+1} - c_k} \right) - y^2 \\ &= (y - c_k)(c_{k+1} - y). \end{aligned} \tag{36}$$

Next, the Flat Randomized Response mechanism maps the label V to a noisy label $Z \in [M]$. The server computes the debiased estimator $\widetilde{y}_J = \frac{c_Z}{p-q}$. Because Flat RR operates over the M centroids in Γ, its second moment evaluates to $\mathbb{E}[c_Z^2 \mid V] = (p-q)c_V^2 + q\sum_{\ell=1}^M c_\ell^2$. Dividing by $(p-q)^2$, we obtain:

$$\mathbb{E}[\widetilde{y}_J^2 \mid V] = \frac{1}{p-q} c_V^2 + \frac{q}{(p-q)^2} \sum_{\ell=1}^M c_\ell^2. \tag{37}$$

Since $\mathbb{E}[\widetilde{y}_J \mid V] = c_V$, the conditional variance is:

$$\mathrm{Var}(\widetilde{y}_J \mid V) = \mathbb{E}[\widetilde{y}_J^2 \mid V] - c_V^2 = \left(\frac{1}{p-q} - 1 \right) c_V^2 + \frac{q}{(p-q)^2} \sum_{\ell=1}^M c_\ell^2. \tag{38}$$

By invoking the Law of Total Variance over the two cascaded stochastic mechanisms (quantization followed by LDP), the variance of the estimator given the continuous coordinate y is:

$$\begin{aligned} \mathrm{Var}(\widetilde{y}_J \mid Y = y) &= \mathbb{E}_V[\mathrm{Var}(\widetilde{y}_J \mid V) \mid Y = y] + \mathrm{Var}_V(\mathbb{E}[\widetilde{y}_J \mid V] \mid Y = y) \\ &= \mathbb{E}_V \left[\left(\frac{1}{p-q} - 1 \right) c_V^2 + \frac{q}{(p-q)^2} \sum_{\ell=1}^M c_\ell^2 \,\middle|\, Y = y \right] + \mathrm{Var}(c_V \mid Y = y). \end{aligned} \tag{39}$$

Substituting $\mathbb{E}_V[c_V^2 \mid Y = y] = \mathrm{Var}(c_V \mid Y = y) + y^2$, we get:

$$\begin{aligned} \mathrm{Var}(\widetilde{y}_J \mid Y = y) &= \left(\frac{1}{p-q} - 1 \right) \left(\mathrm{Var}(c_V \mid Y = y) + y^2 \right) + \frac{q}{(p-q)^2} \sum_{\ell=1}^M c_\ell^2 + \mathrm{Var}(c_V \mid Y = y) \\ &= \frac{r_\Gamma(y)}{p-q} + \left(\frac{1}{p-q} - 1 \right) y^2 + \frac{q}{(p-q)^2} \sum_{\ell=1}^M c_\ell^2. \end{aligned} \tag{40}$$

Because the estimator is locally unbiased ($\mathbb{E}[\widetilde{y}_J \mid y] = y$), we have:

$$\mathbb{E}[\widetilde{y}_J^2 \mid Y = y] = \mathrm{Var}(\widetilde{y}_J \mid Y = y) + y^2 = \frac{r_\Gamma(y)}{p-q} + \frac{1}{p-q} y^2 + \frac{q}{(p-q)^2} \sum_{\ell=1}^M c_\ell^2. \tag{41}$$

Taking the unconditional expectation over $Y \sim \mu$, factoring out $\frac{1}{p-q}$, and applying the Flat RR probability identity $\frac{q}{p-q} = \frac{1}{e^\epsilon - 1}$, we derive the surrogate loss formulation $\mathcal{L}^*_{\text{SSTQ}}(\Gamma; \mu)$:

$$\begin{aligned}\mathbb{E}_Y[\widetilde{y}_J^2] &= \frac{1}{p-q}\left[\mathbb{E}_Y[r_\Gamma(y)] + \frac{q}{p-q}\sum_{\ell=1}^{M} c_\ell^2\right] + \frac{1}{p-q}\mathbb{E}_Y[y^2] \\ &= \frac{1}{p-q}\underbrace{\left[\mathbb{E}_Y[r_\Gamma(y)] + \frac{1}{e^\epsilon - 1}\sum_{\ell=1}^{M} c_\ell^2\right]}_{=\mathcal{L}^*_{\text{SSTQ}}(\Gamma;\mu)} + \frac{1}{p-q}\mathbb{E}_Y[y^2]. \end{aligned} \tag{42}$$

If μ has a density f_Y, by writing the two expectations above as integration against f_Y, we obtain the characterization of $\mathcal{L}^*_{\text{SSTQ}}(\Gamma; \mu)$ given by equation 3. □

D.2 Proof of Global Convexity

Proof of Theorem 4.2. We decompose the objective function into the spatial quantization penalty $F(\Gamma)$ and the variance regularization penalty $G(\Gamma)$:

$$\mathcal{L}_{\text{SSTQ}}(\Gamma) = F(\Gamma) + G(\Gamma)$$

where

$$F(\Gamma) = \sum_{k=1}^{M-1} \frac{(c_{k+1} - c_k)^3}{12B}, \quad \text{and} \quad G(\Gamma) = \lambda \sum_{i=1}^{M} c_i^2$$

with the privacy-dependent scalar $\lambda = \frac{1}{e^\epsilon - 1}$. For any valid privacy budget $\epsilon > 0$, the parameter λ is strictly positive. We show that F is convex and G is strictly convex and hence the loss is a strictly convex function of Γ.

Let $D \in \mathbb{R}^{(M-1)\times M}$ be the finite difference matrix defined such that the k-th element of the matrix-vector product is $(D\Gamma)_k = c_{k+1} - c_k$. The feasible set $\mathcal{D}$ corresponds to the region where $D\Gamma \geq 0$ element-wise.

We define the scalar function $\phi(x) = \frac{x^3}{12B}$. For $x \geq 0$, its second derivative is $\phi''(x) = \frac{x}{2B} \geq 0$. The term $F(\Gamma)$ can be written as $F(\Gamma) = \sum_{k=1}^{M-1} \phi((D\Gamma)_k)$. By the chain rule for vectors, the Hessian matrix of $F(\Gamma)$ is given by:

$$\nabla^2 F(\Gamma) = D^\top \operatorname{diag}\Big(\phi''((D\Gamma)_1), \ldots, \phi''((D\Gamma)_{M-1})\Big) D$$

Because $(D\Gamma)_k \geq 0$ for all $\Gamma \in \mathcal{D}$, the diagonal matrix contains solely non-negative entries. Consequently, for any vector $v \in \mathbb{R}^M$, the quadratic form $v^\top \nabla^2 F(\Gamma) v \geq 0$. This establishes that the Hessian $\nabla^2 F(\Gamma)$ is positive semi-definite ($\nabla^2 F(\Gamma) \succeq 0$), and thus $F(\Gamma)$ is convex on $\mathcal{D}$.

The regularization term evaluates to a scaled squared Euclidean norm, $G(\Gamma) = \lambda \|\Gamma\|_2^2$. The Hessian matrix of this canonical quadratic form is:

$$\nabla^2 G(\Gamma) = 2\lambda I_M$$

where I_M is the $M \times M$ identity matrix. Because $\lambda > 0$, the matrix $2\lambda I_M$ is strictly positive definite ($\nabla^2 G(\Gamma) \succ 0$). This establishes that $G(\Gamma)$ is strongly convex on $\mathbb{R}^M$.

The Hessian of the total objective function is the sum of the constituent Hessians:

$$\nabla^2 \mathcal{L}_{\text{SSTQ}}(\Gamma) = \nabla^2 F(\Gamma) + \nabla^2 G(\Gamma)$$

The sum of a positive semi-definite matrix and a strictly positive definite matrix is strictly positive definite. Thus, $\nabla^2 \mathcal{L}_{\text{SSTQ}}(\Gamma) \succeq 2\lambda I_M \succ 0$ for all $\Gamma \in \mathcal{D}$. This rigorously guarantees that $\mathcal{L}_{\text{SSTQ}}(\Gamma)$ is strictly globally convex (and formally, strongly convex) over the feasible domain.

We next prove the existence of the global minimizer via coercivity. A continuous function defined on a closed set is guaranteed to attain a global minimum if it is coercive, meaning the function value diverges to infinity as the norm of the input diverges ($\lim_{\|\Gamma\|_2 \to \infty} \mathcal{L}_{\text{SSTQ}}(\Gamma) = \infty$).

For any $\Gamma \in \mathcal{D}$, the condition $c_{k+1} - c_k \geq 0$ guarantees that $F(\Gamma) \geq 0$. Consequently, the total loss function is strictly bounded from below by its quadratic regularization term:

$$\mathcal{L}_{\text{SSTQ}}(\Gamma) \geq G(\Gamma) = \lambda \|\Gamma\|_2^2$$

Taking the limit as the norm of the codebook grows yields:

$$\lim_{\|\Gamma\|_2 \to \infty} \mathcal{L}_{\text{SSTQ}}(\Gamma) \geq \lim_{\|\Gamma\|_2 \to \infty} \lambda \|\Gamma\|_2^2 = \infty$$

The strong coercivity of the objective function on the closed convex set $\mathcal{D}$ ensures that the infimum is finite and attained. Thus, there exists at least one global minimizer $\Gamma^* \in \mathcal{D}$.

Note that if explicit boundary constraints such as $c_1 = -B$ and $c_M = B$ are enforced, the monotonic ordering requirement confines all coordinates strictly to the interval $[-B, B]$. The feasible set $\mathcal{D}$ then becomes a compact subset of $\mathbb{R}^M$, and existence is directly guaranteed by the Weierstrass extreme value theorem without requiring the coercivity argument.

The uniqueness of global minimizer follows from the strict convexity (if there are two distinct global minimizers, their average will attain a smaller function value which is a contradiction.)

□

D.3 Improved MSE Bound Using Optimized Codebook

The proof proceeds by constructing a family of candidate codebooks. For an integer parameter $1 \leq m \leq \lfloor (M-1)/2 \rfloor$, we define $\tilde{\Gamma}_m \in \mathcal{D}$ to span exactly m contiguous gap intervals of size B/m, mapping the corresponding values from $-B$ to B, while collapsing the remaining $M - 2m - 1$ indices symmetrically to 0 (repeated codewords). We then compare the surrogate loss evaluated at the optimized codebook with the minimum loss over this family, which yields a tighter dependence on M. For comparison, a uniform grid obeys

$$\sum_{i=1}^{M} c_i^2 = B^2 \frac{M(M+1)}{3(M-1)} = \Theta(B^2 M).$$

At fixed ϵ, its privacy contribution scales as $\Theta_\epsilon(M^2)$ after the Flat-RR multiplier, whereas Theorem 4.3 gives an $\mathcal{O}_\epsilon(M)$ upper bound for the optimized surrogate.

Proof of Theorem 4.3. By recalling Equation equation 33, we have

$$\text{MSE}_{\text{SSTQ}} = \mathbb{E}[\|\hat{x}\|_2^2] - \|x\|_2^2 < \mathbb{E}[\|\hat{x}\|_2^2] = \frac{d^2}{N} \sum_{j=1}^{N} \|u_j\|_2^2 \mathbb{E}[\tilde{y}_j^2 \mid J = j]\,. \tag{43}$$

Next by combining equation 3, equation 2, and using Lemma C.1, we get

$$\text{MSE}_{\text{SSTQ}} \leq \frac{d^2}{p-q} \left[\sum_{k=1}^{M-1} \int_{c_k}^{c_{k+1}} (y - c_k)(c_{k+1} - y) f_Y(y)\, dy \;\; + \;\; \frac{1}{e^\epsilon - 1} \sum_{i=1}^{M} c_i^2 + \int_{-B}^{B} y^2 f_Y(y)\, dy \right] \tag{44}$$

We denote the right-hand side (without the scaling factor $\frac{d^2}{p-q}$) by $U(\Gamma)$ as a function of the codebook Γ, and decompose it as the data-dependent part $I(\Gamma)$ and the regularization part $G(\Gamma)$ given by

$$I(\Gamma) := \sum_{k=1}^{M-1} \int_{c_k}^{c_{k+1}} [(y - c_k)(c_{k+1} - y) + y^2] f_Y(y)\, dy \tag{45}$$

$$G(\Gamma) := \lambda \sum_{i=1}^{M} c_i^2\,, \tag{46}$$

with $\lambda = (e^\epsilon - 1)^{-1}$, and so $U(\Gamma) = I(\Gamma) + G(\Gamma)$.

We next upper bound $I(\Gamma)$. Define the polynomial within the brackets as $h_k(y)$ for $y \in [c_k, c_{k+1}]$:

$$h_k(y) = (y - c_k)(c_{k+1} - y) + y^2$$

Expanding $h_k(y)$ yields:

$$h_k(y) = yc_{k+1} - y^2 - c_k c_{k+1} + c_k y + y^2 = (c_k + c_{k+1})y - c_k c_{k+1}$$

The function $h_k(y)$ is an affine function with respect to y. By the properties of affine functions on closed intervals, $h_k(y)$ attains its maximum at one of the boundaries of the interval $[c_k, c_{k+1}]$. Evaluating $h_k(y)$ at the endpoints gives:

$$h_k(c_k) = (c_k + c_{k+1})c_k - c_k c_{k+1} = c_k^2$$
$$h_k(c_{k+1}) = (c_k + c_{k+1})c_{k+1} - c_k c_{k+1} = c_{k+1}^2$$

Thus, for all $y \in [c_k, c_{k+1}]$, the function is bounded by:

$$h_k(y) \leq \max(c_k^2, c_{k+1}^2)$$

Because the codebook elements are constrained to the interval $[-B, B]$, it follows that $c_i^2 \leq B^2$ for all $i \in \{1, \ldots, M\}$. Therefore, $h_k(y) \leq B^2$ uniformly for all $k \in \{1, \ldots, M-1\}$ and $y \in [c_k, c_{k+1}]$.

Applying this pointwise upper bound to the integrals results in:

$$I(\Gamma) \leq \sum_{k=1}^{M-1} \int_{c_k}^{c_{k+1}} B^2 f_Y(y)\, dy = B^2 \int_{-B}^{B} f_Y(y)\, dy$$

Because $f_Y(y)$ is a valid probability density function supported on $[-B, B]$, its integral over this domain evaluates to 1. Hence, $I(\Gamma) \leq B^2$. This bound is independent of the distribution $f_Y(y)$ and holds for any feasible codebook, including Γ^*.

We next upper bound the regularization term $G(\Gamma^*)$. To constrain $G(\Gamma^*)$, we evaluate $J(\Gamma) = F(\Gamma) + G(\Gamma)$ where $F(\Gamma) = \sum_{k=1}^{M-1} \frac{\Delta_k^3}{12B}$. By the continuous global optimality of Γ^*, $J(\Gamma^*) \leq J(\tilde{\Gamma})$ for any valid proxy codebook sequence. For an integer parameter $1 \leq m \leq \lfloor \frac{M-1}{2} \rfloor$, define $\tilde{\Gamma}_m \in \mathcal{D}$ to span exactly m contiguous gap intervals of size B/m mapping respectively from $-B$ and B, collapsing the remaining $M - 2m - 1$ indices symmetrically to 0. This sequence computes strictly to $F(\tilde{\Gamma}_m) = 2m\frac{(B/m)^3}{12B} = \frac{B^2}{6m^2}$ and $G(\tilde{\Gamma}_m) = 2\lambda \sum_{j=1}^{m} \left(\frac{jB}{m}\right)^2 = \lambda B^2 \frac{(m+1)(2m+1)}{3m}$. Applying Jensen's inequality to the objective yields a strict minimal floor $F(\Gamma^*) \geq \frac{2B^2}{3(M-1)^2}$. Subtracting this extracts the uniform limit:

$$G(\Gamma^*) \leq F(\tilde{\Gamma}_m) + G(\tilde{\Gamma}_m) - F(\Gamma^*) \leq \frac{B^2}{6m^2} + \frac{B^2(m+1)(2m+1)}{3m(e^\epsilon - 1)} - \frac{2B^2}{3(M-1)^2}$$

Combining the bounds on $I(\Gamma^*)$ and $G(\Gamma^*)$ we obtain

$$U(\Gamma^*) \leq B^2 + \min_{1 \leq m \leq \lfloor \frac{M-1}{2} \rfloor} B^2 \left(\frac{1}{6m^2} + \frac{(m+1)(2m+1)}{3m(e^\epsilon - 1)} \right) - \frac{2B^2}{3(M-1)^2}$$

To quantify the term with minimum over m on the right-hand side, consider the continuous relaxation of the objective function, defined approximately as $g(m) \approx B^2 \left(\frac{1}{6m^2} + \frac{2\lambda m}{3}\right)$. Setting the first derivative to zero provides the unconstrained optimum $m^* \approx (2\lambda)^{-1/3}$. We proceed by setting $m = (2\lambda)^{-1/3} \vee 1$, since we should have $m \geq 1$. In addition, by our assumption, $M > (4(e^\epsilon - 1))^{1/3} + 1$, which implies that $\frac{M-1}{2} \geq m$, and so this is a valid choice.

Substituting m back into the objective yields the asymptotic behavior:

$$\frac{1}{6m^2} + \frac{(m+1)(2m+1)}{3m(e^\epsilon - 1)} = \mathcal{O}(\lambda^{2/3} \vee \lambda) = \mathcal{O}(\epsilon^{-1} \vee \epsilon^{-2/3}),$$

using the fact that $\lambda < 1/\epsilon$. This also implies that $\frac{2B^2}{3(M-1)^2} = \mathcal{O}(\lambda^{2/3} B^2) = \mathcal{O}(B^2 \epsilon^{-2/3})$. Hence,

$$U(\Gamma^*) = \mathcal{O}(B^2(1 + \epsilon^{-1} + \epsilon^{-2/3})) = \mathcal{O}(B^2(1 + \epsilon^{-1})).$$

Consequently,

$$\text{MSE}_{\text{SSTQ}} = \mathcal{O}\left(\frac{d^2}{p-q} B^2(1 + \epsilon^{-1}))\right) = \mathcal{O}\left(\frac{dK^2R^2}{\rho}(1 + M\epsilon^{-1} + M\epsilon^{-2})\right),$$

given that $B = KR/\sqrt{N}$, $\rho = N/d$, $p - q = 1/(1 + M\lambda)$, and so $1/(p - q) < 1 + M\epsilon^{-1}$. This completes the proof. □

E PROOFS FOR THE METRIC-AWARE LAPLACE MECHANISM

This appendix provides the formal mathematical proofs for the Metric-Aware Laplace mechanism introduced in Section 5. The mechanism operates in two stages: (1) continuous Laplace sampling, and (2) nearest-codeword quantization. We first prove that the mechanism satisfies pure ϵ-LDP via the post-processing theorem. We then formulate the surrogate objective $\mathcal{L}^{\mathrm{MA}}_{\mathrm{SSTQ}}$ and establish upper bounds on the variance and the bias of the estimator under identity decoding. The global bias is entirely independent of the ambient dimension.

Algorithm 3 SSTQ Client Encoding Algorithm (Metric-Aware)

Require: $x \in B_2^d(R)$, public scaled Parseval frame $U \in \mathbb{R}^{N\times d}$, and public indexed codebook $\Gamma = (c_1, \ldots, c_M)$ with $M = 2^b$ and maximum gap $\Delta_{\max} = \max_k(c_{k+1} - c_k) \le 2BC/(2^b - 1)$.

1: **Redundant Kashin Bounding:** Execute Lyubarskii's algorithm (Lyubarskii & Vershynin, 2010) to compute $y \in \mathbb{R}^N$ such that $\frac{d}{N}U^\top y = x$ and $\|y\|_\infty \le B = \frac{KR}{\sqrt{N}}$.

2: **Oblivious Subsampling:** Sample $J \sim \mathrm{Unif}([N])$, independently of x.

3: **Stochastic Interpolation on labels:** If $y_J = c_i$ for one or more indices, set

$$V = \min\{i : c_i = y_J\}.$$

Otherwise set $k_- = \max\{i : c_i < y_J\}$ and $k_+ = \min\{i : c_i > y_J\}$, and sample $V \in \{k_-, k_+\}$ according to

$$\Pr(V = k_+ \mid y_J) = \frac{y_J - c_{k_-}}{c_{k_+} - c_{k_-}}, \qquad \Pr(V = k_- \mid y_J) = \frac{c_{k_+} - y_J}{c_{k_+} - c_{k_-}}.$$

4: **Truncated Laplace perturbation:** Given $V = k$, draw a continuous sample $T \in [-B, B]$ from

$$f_{T|V=k}(t) = \frac{1}{Z_k}\exp\left(-\frac{\epsilon|t - c_k|}{2D_B}\right), \qquad D_B = 2B, \quad Z_k = \int_{-B}^{B}\exp\left(-\frac{\epsilon|t - c_k|}{2D_B}\right)dt.$$

5: **Nearest-codeword quantization:** Release the label

$$Z = \tau(T), \qquad \tau(t) := \min\arg\min_{i\in[M]} |t - c_i|,$$

where the tie rule makes τ deterministic and measurable, so that the cells $\mathrm{cell}_i = \{t \in [-B, B] : \tau(t) = i\}$ partition $[-B, B]$ even when codeword values coincide.

6: **return** Transmit (J, Z). **Cost:** $\lceil \log_2 N \rceil + b$ bits.

Algorithm 4 SSTQ Server Decoding Algorithm (Metric-Aware, Identity Decoding)

Require: Message (J, Z), indexed codebook $\Gamma = (c_1, \ldots, c_M)$, and scaled Parseval frame U.

1: **Identity decoding:** Set $\widetilde{y}_J = c_Z$ (no $1/(p - q)$ unbiasing factor is applied).

2: **Sparse reconstruction:** Set $\widetilde{y} = N\widetilde{y}_J e_J = Nc_Z e_J \in \mathbb{R}^N$.

3: **Frame projection:** Set

$$\hat{x} = \frac{d}{N}U^\top \widetilde{y} = d\, c_Z\, u_J.$$

4: **return** $\hat{x}$.

E.1 PROOF OF ϵ-LDP

Proof of Theorem 5.1. The proof follows from the post-processing theorem of differential privacy. The mechanism consists of two stages:

Stage 1 (Continuous Laplace Sampling). The Kashin coefficient y_J is mapped to a codeword c_V via stochastic rounding. The mechanism draws T from a truncated Laplace distribution on $[-B, B]$

with density $f_{T|V=k}(t) = \frac{1}{Z_k} \exp\left(-\frac{\epsilon|t-c_k|}{2D_B}\right)$ for $t \in [-B, B]$, where $D_B = 2B$. This continuous mechanism satisfies ϵ-LDP: for any two codewords $c_k, c_m \in \Gamma$ and any measurable set $S \subseteq [-B, B]$,

$$\frac{\Pr(T \in S \mid V = k)}{\Pr(T \in S \mid V = m)} = \frac{\int_S \frac{1}{Z_k} e^{-\frac{\epsilon|t-c_k|}{2D_B}}\, dt}{\int_S \frac{1}{Z_m} e^{-\frac{\epsilon|t-c_m|}{2D_B}}\, dt}. \tag{47}$$

For the numerator, using the triangle inequality $|t - c_k| \geq |t - c_m| - |c_m - c_k|$:

$$e^{-\frac{\epsilon|t-c_k|}{2D_B}} \leq e^{\frac{\epsilon|c_m-c_k|}{2D_B}} \cdot e^{-\frac{\epsilon|t-c_m|}{2D_B}}. \tag{48}$$

Integrating over S yields $\int_S \frac{1}{Z_k} e^{-\frac{\epsilon|t-c_k|}{2D_B}} dt \leq \frac{Z_m}{Z_k} e^{\frac{\epsilon|c_m-c_k|}{2D_B}} \int_S \frac{1}{Z_m} e^{-\frac{\epsilon|t-c_m|}{2D_B}} dt$.

An analogous bound shows $\frac{Z_m}{Z_k} \leq e^{\frac{\epsilon|c_m-c_k|}{2D_B}}$. Multiplying yields:

$$\frac{\Pr(T \in S \mid V = k)}{\Pr(T \in S \mid V = m)} \leq \exp\left(\frac{\epsilon|c_k - c_m|}{D_B}\right) \leq e^{\epsilon}, \tag{49}$$

where the final step uses $|c_k - c_m| \leq 2B = D_B$.

Stage 2 (Nearest-Codeword Quantization). The output $Z = \tau(T)$ with $\tau(t) := \min \arg\min_{i \in [M]} |t - c_i|$, is a deterministic function of T. By the post-processing theorem of differential privacy, applying a deterministic function to the output of an ϵ-LDP mechanism preserves ϵ-LDP.

In addition, since the subsampled index J is chosen independently of the input x, its transmission does not add any privacy loss. Therefore, the composed mechanism satisfies pure ϵ-LDP for any codebook geometry. □

E.2 MSE and Bias Bounds under Identity Decoding

E.2.1 Derivation of Surrogate Loss Objective for the Metric-Aware Mechanism

To optimize the codebook $\Gamma = \{c_1, \ldots, c_M\}$ under the Metric-Aware Laplace mechanism, the loss objective must evaluate the expected mean squared error. We define

$$\mathcal{L}^*_{\text{MA-SSTQ}}(\Gamma) := \mathbb{E}[(y - c_Z)^2], \tag{50}$$

utilizing identity decoding.

We first recall the pipeline. Let y denote the continuous coordinate drawn from the density $f_Y(y)$. The mechanism pipeline operates as follows:

1. **Stochastic Quantization:** The Kashin coordinate $y \in [c_k, c_{k+1}]$ is stochastically assigned to a codeword c_V with label $V \in \{k, k+1\}$ to enforce the unbiasedness property $\mathbb{E}[c_V|Y = y] = y$. The transition probabilities are $p_k(y) = \frac{c_{k+1}-y}{c_{k+1}-c_k}$ and $p_{k+1}(y) = \frac{y-c_k}{c_{k+1}-c_k}$.
2. **Metric-Aware Privatization:** Given the label $V = j$, the mechanism draws a continuous sample T from the truncated Laplace distribution $f_{T|V=j}(t) = \frac{1}{Z_j} \exp\left(-\frac{\epsilon|t-c_j|}{4B}\right)$ for $t \in [-B, B]$, and outputs the nearest codeword c_Z, where $Z = \tau(T)$ with $\tau(t) := \min \arg\min_{i \in [M]} |t - c_i|$. The effective transition probability $P_{i|j} = \Pr[Z = i \mid V = j]$ is given by Eq. 7.
3. **Identity Decoding:** The decoder assigns the reconstructed coordinate directly as $\widetilde{y} = c_Z$.

By the Law of Total Expectation, the conditional error decomposes as:

$$\mathbb{E}[(y - c_Z)^2 \mid Y = y] = \mathbb{E}_{V|Y}\left[\mathbb{E}_{Z|V}[(y - c_Z)^2]\right]$$

For a fixed intermediate codeword label $V = j$, we define two geometric statistics representing the mechanism's behavior:

- **Expected Structural Bias:** $\mu_j = \mathbb{E}[c_Z \mid V = j]$, the expected output codeword under the effective transition probabilities equation 7.

- **Localized Mechanism Variance:** $\Phi_j = \mathbb{E}[(c_Z - c_j)^2 \mid V = j]$.

Expanding the inner expectation explicitly around c_j gives:

$$\mathbb{E}_{Z|V}[(y - c_j + c_j - c_Z)^2] = (y - c_j)^2 + 2(y - c_j)(c_j - \mu_j) + \Phi_j$$

Evaluating the outer expectation over the label $V \in \{k, k+1\}$ strictly partitions the polynomial into three distinct components:

1. Base Quantization Error: Applying the probabilities yields the standard error polynomial:

$$p_k(y)(y - c_k)^2 + p_{k+1}(y)(y - c_{k+1})^2 = \frac{c_{k+1} - y}{c_{k+1} - c_k}(y - c_k)^2 + \frac{y - c_k}{c_{k+1} - c_k}(y - c_{k+1})^2$$

   Factoring out the term $(y - c_k)(c_{k+1} - y)$ leaves:

$$\frac{(y - c_k)(c_{k+1} - y)}{c_{k+1} - c_k}\left[(y - c_k) + (c_{k+1} - y)\right] = (y - c_k)(c_{k+1} - y)$$

2. Cross-Correlation Bias Term: The expected cross-term captures the interaction between the quantization interval bounds and the mechanism's structural bias:

$$2p_k(y)(y - c_k)(c_k - \mu_k) + 2p_{k+1}(y)(y - c_{k+1})(c_{k+1} - \mu_{k+1})$$

   Applying the sign substitution $y - c_{k+1} = -(c_{k+1} - y)$ and replacing $p_k(y), p_{k+1}(y)$ transforms this to:

$$2\frac{c_{k+1} - y}{c_{k+1} - c_k}(y - c_k)(c_k - \mu_k) - 2\frac{y - c_k}{c_{k+1} - c_k}(c_{k+1} - y)(c_{k+1} - \mu_{k+1})$$

   Factoring out the data-dependent polynomial provides the algebraic simplification:

$$2\frac{(y - c_k)(c_{k+1} - y)}{c_{k+1} - c_k}\left[c_k - \mu_k - (c_{k+1} - \mu_{k+1})\right] = 2(y - c_k)(c_{k+1} - y)\left(\frac{\mu_{k+1} - \mu_k}{c_{k+1} - c_k} - 1\right)$$

3. Expected Privacy Variance: The expected structural variance interpolates linearly between the two valid codeword states:

$$\mathbb{E}_{V|Y}[\Phi_V] = p_k(y)\Phi_k + p_{k+1}(y)\Phi_{k+1}$$

Summing the base quantization error and the cross-correlation bias exactly consolidates the y-dependent portion of the loss. We define the scale modifier W_k:

$$W_k = 1 + 2\left(\frac{\mu_{k+1} - \mu_k}{c_{k+1} - c_k} - 1\right) = \frac{2(\mu_{k+1} - \mu_k)}{c_{k+1} - c_k} - 1$$

Integrating the components over the density $f_Y(y)$ establishes the intermediate objective:

$$\mathcal{L}^*_{\text{MA-SSTQ}}(\Gamma) = \sum_{k=1}^{M-1} \int_{c_k}^{c_{k+1}} \left[W_k(y - c_k)(c_{k+1} - y) + p_k(y)\Phi_k + p_{k+1}(y)\Phi_{k+1}\right] f_Y(y)\, dy$$

By regrouping the local variance assignments globally across the entire support domain $[-B, B]$, the continuous expected mean squared error minimizes to its definitive analytical form:

$$\mathcal{L}^*_{\text{MA-SSTQ}}(\Gamma) = \sum_{k=1}^{M-1} \int_{c_k}^{c_{k+1}} W_k(y - c_k)(c_{k+1} - y) f_Y(y)\, dy \;\; + \;\; \sum_{k=1}^{M} \pi_k \Phi_k \,,$$

where $\pi_k = \int_{-B}^{B} \Pr[V = k \mid Y = y] f_Y(y)\, dy$ defines the assignment probability of the interpolated label V.

E.2.2 Proof of Theorem 5.2

Let $u_j^\top$ denote the j-th row of U, and set

$$w_j := \|u_j\|_2^2, \qquad D_U := \mathrm{diag}(w_1, \dots, w_N), \qquad \mu_U := \max_{j\in[N]} w_j.$$

Since U is a scaled Parseval frame, $U^\top U = \frac{N}{d} I_d = \rho I_d$, and the trace identity (Lemma C.1) gives $\sum_{j=1}^N w_j = \mathrm{Tr}(U^\top U) = N$.

Let $y \in \mathbb{R}^N$ be a Kashin representation of x, so that

$$x = \frac{d}{N} U^\top y, \qquad \|y\|_\infty \le B = \frac{KR}{\sqrt{N}}.$$

For each coordinate y_j, let $c_{V,j}$ denote the output of stochastic endpoint interpolation and let $c_{Z,j}$ denote the scalar obtained after applying the metric-aware mechanism and identity decoding. Here, we made the dependence on j explicit in the index. Thus, if $y_j \in [c_k, c_{k+1}]$, then

$$\mathbb{E}[c_{V,j} \mid y_j] = y_j$$

and

$$\mathbb{E}[(c_{V,j} - y_j)^2 \mid y_j] = (y_j - c_k)(c_{k+1} - y_j) \le \frac{\Delta_{\max}^2}{4}.$$

The scalar metric-aware mechanism satisfies the uniform moment bounds

$$\max_i |\mathbb{E}[c_Z - c_i \mid V = i]| \le \frac{8B}{\epsilon}, \qquad \max_i \mathbb{E}[(c_Z - c_i)^2 \mid V = i] \le \frac{128B^2}{\epsilon^2}.$$

Define

$$r_j := \mathbb{E}[c_{Z,j} \mid y_j] - y_j, \qquad a_j := \mathbb{E}[(c_{Z,j} - y_j)^2 \mid y_j].$$

Because stochastic endpoint interpolation is unbiased, r_j is a convex combination of the conditional mechanism biases at the two neighboring codewords. Hence

$$|r_j| \le \frac{8B}{\epsilon}.$$

Moreover, using

$$c_{Z,j} - y_j = (c_{V,j} - y_j) + (c_{Z,j} - c_{V,j})$$

and the inequality $(s+t)^2 \le 2s^2 + 2t^2$, we obtain

$$\begin{aligned} a_j &\le 2\mathbb{E}[(c_{V,j} - y_j)^2 \mid y_j] + 2\mathbb{E}[(c_{Z,j} - c_{V,j})^2 \mid y_j] \\ &\le \frac{\Delta_{\max}^2}{2} + \frac{256B^2}{\epsilon^2} \\ &\le \frac{2B^2C^2}{(2^b-1)^2} + \frac{256B^2}{\epsilon^2}. \end{aligned}$$

Denote the last upper bound by

$$A := \frac{2B^2C^2}{(2^b-1)^2} + \frac{256B^2}{\epsilon^2}.$$

The SSTQ estimator can be written as $\widehat{x} = d c_{Z,J} u_J$, with $J \sim \mathrm{Unif}[N]$.

We first bound its bias. Writing $r := (r_1, \dots, r_N)^\top$, the unbiased reconstruction identity $x = (d/N) U^\top y$ gives

$$\mathbb{E}[\widehat{x}] - x = \frac{d}{N} U^\top r.$$

Since

$$\|UU^\top\|_{\mathrm{op}} = \|U^\top U\|_{\mathrm{op}} = \frac{N}{d},$$

we have

$$\|\mathbb{E}[\widehat{x}] - x\|_2^2 = \frac{d^2}{N^2} \|U^\top r\|_2^2 \le \frac{d}{N} \|r\|_2^2 \le d \left(\frac{8B}{\epsilon}\right)^2 = \frac{64K^2R^2}{\rho\epsilon^2}.$$

This proves the claimed bias bound.

We next consider the mean-squared error. Define $P := \frac{d}{N} UU^\top$. The scaled Parseval identity implies that P is an orthogonal projection. Furthermore,

$$Ux = Py, \qquad \|x\|_2^2 = \frac{d}{N} y^\top P y.$$

Since

$$\mathbb{E}[c_{Z,j}^2 \mid y_j] = a_j + y_j^2 + 2y_j r_j,$$

a direct expansion gives

$$\begin{aligned}\mathbb{E}\big[\|\widehat{x} - x\|_2^2\big] = \frac{d^2}{N} \sum_{j=1}^N w_j a_j + \frac{d^2}{N} y^\top D_U y - \|x\|_2^2 \\ + \frac{2d}{N} y^\top (dD_U - P) r.\end{aligned}$$

For the first term, $w_j \le \mu_U$ and $a_j \le A$ imply

$$\frac{1}{N} \sum_{j=1}^N w_j a_j \le A \frac{1}{N} \sum_{j=1}^N w_j \le A.$$

Therefore,

$$\frac{d^2}{N} \sum_{j=1}^N w_j a_j \le d^2 B^2 \left(\frac{2C^2}{(2^b - 1)^2} + \frac{256}{\epsilon^2} \right).$$

For the second and the third terms, using $\|y\|_\infty \le \frac{K\|x\|_2}{\sqrt{N}}$ and $\sum_j w_j = N$, we have

$$y^\top D_U y = \sum_{j=1}^N w_j y_j^2 \le \|y\|_\infty^2 \sum_{j=1}^N w_j = N \|y\|_\infty^2 \le K^2 \|x\|_2^2 .$$

Thus,

$$\frac{d^2}{N} y^\top D_U y - \|x\|_2^2 \le \left(\frac{K^2 d}{\rho} - 1 \right) \|x\|_2^2 \le \left(\frac{K^2 d}{\rho} - 1 \right) R^2, \tag{51}$$

because $dK^2/\rho \ge 1$ and $\|x\|_2 \le R$. Finally,

$$\|y\|_2 \le \sqrt{N} B, \qquad \|r\|_2 \le \frac{8\sqrt{N} B}{\epsilon},$$

and

$$\|dD_U - P\|_{\mathrm{op}} \le d\|D_U\|_{\mathrm{op}} + \|P\|_{\mathrm{op}} \le d\mu_U + 1.$$

It follows that

$$\begin{aligned}\left| \frac{2d}{N} y^\top (dD_U - P) r \right| &\le \frac{2d}{N} \|y\|_2 \|dD_U - P\|_{\mathrm{op}} \|r\|_2 \\ &\le \frac{16 dB^2}{\epsilon} (d\mu_U + 1) \\ &= \frac{16 d^2 B^2}{\epsilon} \left(\mu_U + \frac{1}{d} \right).\end{aligned}$$

Combining the preceding bounds and

$$d^2 B^2 = \frac{d^2 K^2 R^2}{N} = \frac{dK^2 R^2}{\rho},$$

gives

$$\mathbb{E}\big[\|\widehat{x} - x\|_2^2\big] \le \frac{dK^2 R^2}{\rho} \left(1 + \frac{2C^2}{(2^b - 1)^2} + \frac{16(\mu_U + 1/d)}{\epsilon} + \frac{256}{\epsilon^2} \right) - R^2.$$

To complete the proof of theorem, it remains to bound μ_U. For every j,

$$\begin{aligned}\|u_j\|_2^2 &= e_j^\top U U^\top e_j \\ &\le \|UU^\top\|_{\mathrm{op}} \\ &= \|U^\top U\|_{\mathrm{op}} = \rho.\end{aligned}$$

Taking the maximum over j proves that $\mu_U \le \rho$.

E.2.3 Tight Bounds under the True Distribution

We formalize the improvement available from distribution-aware codebook optimization by introducing the Panter–Dite functional, which quantifies the compressibility of a continuous density. Let the distribution shape factor be defined as:

$$C_f = \frac{1}{4B^2}\left(\int_{-B}^{B} f_Y(y)^{1/3}\,dy\right)^3$$

By Hölder's inequality, $C_f \le 1$ for all valid densities supported on $[-B, B]$, with equality holding if and only if $f_Y(y)$ is the uniform distribution.

Theorem E.1. *Let X be a random vector with $\|X\|_2 \le R$ almost surely, and U be a fixed and public Parseval frame or publicly randomized independently of X. Define the row-coherence parameter $\mu_U := \max_{j\in[N]} \|u_j\|_2^2$. Let $Y = y_J(X, U)$ have a known, strictly positive continuous density f_Y on $[-B, B]$, with $B = KR/\sqrt{N}$, and put $\rho = N/d$. For $M = 2^b \ge 2$, minimize $\mathcal{L}^*_{\mathrm{MA-SSTQ}}$ over the compact set $\{-B = c_1 \le \cdots \le c_M = B\}$, and let Γ^* be a minimizer. Then the population bias and MSE obey*

$$\varsigma^2_{\mathrm{pop}} := \|\mathbb{E}[\hat{X} - X]\|_2^2 \le \frac{64K^2R^2}{\rho\epsilon^2},$$

$$\mathbb{E}[\|\hat{X} - X\|_2^2] \le \frac{dK^2R^2}{\rho}\left(1 + \frac{2\mu_U C_0^2}{(2^b-1)^2} + \frac{16(\mu_U + 1/d)}{\epsilon} + \frac{256\mu_U}{\epsilon^2}\right) - R^2,$$

where $C_0^2 \le \min\{1, \frac{2}{3}C_f R_M\}$.

Here $R_M = (M-1)^{-1}\sum_k \sup_{I_k} f_Y / \inf_{I_k} f_Y \ge 1$ is the average local oscillation ratio of the density across the intervals $I_k = [c_k, c_{k+1}]$. Since $C_f < 1$ for any strictly non-uniform density, and $R_M \to 1$ as $b \to \infty$, we have $C_0^2 < 1$, for sufficiently high resolutions. Also, for standard isotropic random frames, such as a Gaussian matrix followed by Parseval normalization, the row norms concentrate sharply around their common average, and so μ_U is typically close to one. This explains strict improvement in the MSE bound compared to the bound derived in Theorem 5.2 (Recall that there $C \ge 1$).

E.2.4 Proof of Theorem E.1

Let X be the population input, with $\|X\|_2 = R$ almost surely. The frame $U \in \mathbb{R}^{N\times d}$ is fixed and public, or publicly randomized independently of X. For every realization under consideration,

$$U^\top U = \frac{N}{d} I_d = \rho I_d,$$

where $u_j^\top$ is the j-th row of U. Write

$$w_j := \|u_j\|_2^2, \qquad D_U := \mathrm{diag}(w_1, \ldots, w_N), \qquad \mu_U := \max_{j\in[N]} w_j.$$

If U is publicly randomized, assume the same deterministic upper bound $\max_j w_j(U) \le \mu_U$ holds almost surely. The trace identity (Lemma C.1) gives $\sum_{j=1}^N w_j = \mathrm{Tr}(U^\top U) = N$.

A measurable Kashin-selection rule produces $y = y(X, U) \in \mathbb{R}^N$ satisfying

$$X = \frac{d}{N} U^\top y, \qquad \|y\|_\infty \le B := \frac{KR}{\sqrt{N}}.$$

Let $J \sim \mathrm{Unif}[N]$ be independent of (X, U), and define $Y = y_J(X, U)$. This is the index used by the SSTQ mechanism.

1. Conditional and population coordinate laws. For fixed (x, u), the conditional law of the selected coordinate is

$$P_{Y|X=x,U=u} = \frac{1}{N}\sum_{j=1}^{N} \delta_{y_j(x,u)}. \tag{52}$$

It is therefore an atomic measure, with no density, while conditioning on (X, U).

For a feasible codebook $\Gamma = (-B = c_1 \le \cdots \le c_M = B)$, let $Z_\Gamma(y)$ denote the output of stochastic endpoint interpolation followed by the metric-aware mechanism and identity decoding, when the scalar input is y. Define

$$a_\Gamma(y) := \mathbb{E}\big[(c_Z(y) - y)^2\big], \qquad r_\Gamma(y) := \mathbb{E}[c_Z(y)] - y,$$

where these expectations are only over interpolation and mechanism randomness.

The codebook in the stated algorithm minimizes the original uniform-coordinate population objective

$$\begin{aligned}\mathcal{L}(\Gamma) &:= \mathcal{L}^*_{\text{MA-SSTQ}}(\Gamma) := \mathbb{E}\big[(Y - c_Z(Y))^2\big] \\ &= \mathbb{E}_{X,U}\left[\frac{1}{N}\sum_{j=1}^{N} a_\Gamma(y_j(X, U))\right].\end{aligned}$$

Here the expectation is over the population law of (X, U), the uniform index J, and the internal scalar-mechanism randomness. Assume that the unconditional law P_Y has a strictly positive continuous density f_Y on $[-B, B]$. This density is introduced only after averaging the conditional atomic law in equation 52, and

$$\mathcal{L}(\Gamma) = \int_{-B}^{B} a_\Gamma(y) f_Y(y)\, dy.$$

Let Γ^* be a population minimizer of this objective. Thus, for every feasible deterministic competitor Γ',

$$\mathcal{L}(\Gamma^*) \le \mathcal{L}(\Gamma').$$

The reconstruction error involves the distinct row-norm-weighted loss

$$\mathcal{L}^{(w)}(\Gamma) := \mathbb{E}_{X,U}\left[\frac{1}{N}\sum_{j=1}^{N} w_j a_\Gamma(y_j(X, U))\right]. \tag{53}$$

This is not the objective minimized by Γ^*. Nevertheless, $w_j \le \mu_U$ and nonnegativity of a_Γ imply

$$\mathcal{L}^{(w)}(\Gamma^*) \le \mu_U \mathcal{L}(\Gamma^*). \tag{54}$$

2. Scalar mechanism bounds. For an intermediate codeword label $V = i$, set

$$\beta_i := \mathbb{E}[c_Z \mid V = i] - c_i, \qquad \Phi_i := \mathbb{E}[(c_Z - c_i)^2 \mid V = i].$$

The continuous truncated-Laplace variable has scale $4B/\epsilon$. Since c_i is itself a possible nearest-codeword output,

$$|c_Z - c_i| \le 2|T - c_i|.$$

The first two absolute moments of the corresponding untruncated Laplace variable therefore give

$$\beta_{\max} := \max_i |\beta_i| \le \frac{8B}{\epsilon}, \qquad \Phi_{\max} := \max_i \Phi_i \le \frac{128B^2}{\epsilon^2}. \tag{55}$$

If $y \in [c_k, c_{k+1}]$, stochastic endpoint interpolation is unbiased and $r_\Gamma(y)$ is a convex combination of β_k and β_{k+1}. Consequently,

$$|r_\Gamma(y)| \le \beta_{\max}. \tag{56}$$

Writing $S_\Gamma(y) := \mathbb{E}[(c_V - y)^2]$, we also have

$$S_\Gamma(y) = (y - c_k)(c_{k+1} - y) \quad \text{for } y \in [c_k, c_{k+1}]. \tag{57}$$

The inequality $(a+b)^2 \le 2a^2 + 2b^2$, applied to $c_Z - y = (c_V - y) + (c_Z - c_V)$, yields

$$a_\Gamma(y) \le 2S_\Gamma(y) + 2\Phi_{\max}. \tag{58}$$

3. Population loss of the optimized codebook. First consider the uniform grid. Its gap is $\Delta = 2B/(M-1)$, and (12) gives

$$S_{\Gamma_{\text{unif}}}(y) \le \frac{\Delta^2}{4} = \frac{B^2}{(M-1)^2}.$$

Using equation 58, equation 55, and population optimality,

$$\mathcal{L}(\Gamma^*) \le \frac{2B^2}{(M-1)^2} + \frac{256B^2}{\epsilon^2}. \tag{59}$$

For the distribution-dependent competitor, define

$$S := \int_{-B}^{B} f_Y(y)^{1/3}\, dy, \qquad C_f := \frac{S^3}{4B^2}.$$

Choose the companding partition $-B = c_1 < \cdots < c_M = B$ so that

$$\int_{c_k}^{c_{k+1}} f_Y(y)^{1/3}\, dy = \frac{S}{M-1}. \tag{60}$$

Let

$$M_k := \sup_{[c_k, c_{k+1}]} f_Y, \qquad m_k := \inf_{[c_k, c_{k+1}]} f_Y.$$

Strict positivity and continuity ensure $m_k > 0$. From equation 60,

$$c_{k+1} - c_k \le \frac{S}{(M-1)m_k^{1/3}}.$$

It follows that

$$\begin{aligned} &\int_{c_k}^{c_{k+1}} (y - c_k)(c_{k+1} - y) f_Y(y)\, dy \\ &\qquad \le M_k \frac{(c_{k+1} - c_k)^3}{6} \le \frac{S^3}{6(M-1)^3} \frac{M_k}{m_k}. \end{aligned} \tag{61}$$

With

$$R_M := \frac{1}{M-1} \sum_{k=1}^{M-1} \frac{M_k}{m_k},$$

summing equation 61 and using $S^3 = 4B^2 C_f$ gives

$$\int_{-B}^{B} S_{\Gamma_{\text{PD}}}(y) f_Y(y)\, dy \le \frac{B^2}{(M-1)^2} \left(\frac{2}{3} C_f R_M \right). \tag{62}$$

Combining the better of equation 59 and equation 62 with equation 58, define

$$C_0^2 := \min \left\{ 1, \frac{2}{3} C_f R_M \right\}$$

and obtain

$$\mathcal{L}(\Gamma^*) \le \frac{2B^2 C_0^2}{(M-1)^2} + \frac{256B^2}{\epsilon^2}. \tag{63}$$

4. Population bias. Fix (x, u), write $y = y(x, u)$, and define

$$r := \big(r_{\Gamma^*}(y_1), \ldots, r_{\Gamma^*}(y_N) \big)^\top.$$

The conditional vector bias is

$$b(x, u) := \mathbb{E}[\widehat{X} - X \mid X = x, U = u] = \frac{d}{N} U^\top r.$$

Since $\|UU^\top\|_{\mathrm{op}} = N/d$, equations equation 55-equation 56 give

$$\|b(x,u)\|_2^2 \le \frac{d}{N}\|r\|_2^2 \le d\beta_{\max}^2 \le \frac{64dB^2}{\epsilon^2} = \frac{64K^2R^2}{\rho\epsilon^2}.$$

Jensen's inequality now yields the claimed population-bias bound:

$$\left\|\mathbb{E}[\widehat{X} - X]\right\|_2^2 = \|\mathbb{E}_{X,U}[b(X,U)]\|_2^2 \le \mathbb{E}_{X,U}\|b(X,U)\|_2^2 \le \frac{64K^2R^2}{\rho\epsilon^2}.$$

5. Population mean-squared error. For fixed (x,u), let

$$a_j := a_{\Gamma^*}(y_j), \qquad r_j := r_{\Gamma^*}(y_j), \qquad P := \frac{d}{N}UU^\top.$$

The matrix P is an orthogonal projection. Using $\widehat{X} = dc_{Z,J}u_J$ and $\mathbb{E}[c_{Z,j}^2] = a_j + y_j^2 + 2y_jr_j$, direct expansion gives the exact conditional identity

$$\mathbb{E}\left[\left\|\widehat{X} - x\right\|_2^2 \middle| x,u\right] = \frac{d^2}{N}\sum_{j=1}^N w_ja_j + \frac{d^2}{N}y^\top D_U y - \|x\|_2^2 + \frac{2d}{N}y^\top(dD_U - P)r. \tag{64}$$

The Kashin envelope and equation 56 imply

$$\|y\|_2 \le \sqrt{N}B, \qquad \|r\|_2 \le \sqrt{N}\beta_{\max}.$$

In addition, $y^\top D_U y \le B^2\sum_j w_j = NB^2$. Moreover,

$$\|dD_U - P\|_{\mathrm{op}} \le d\|D_U\|_{\mathrm{op}} + \|P\|_{\mathrm{op}} \le d\mu_U + 1.$$

Hence the last term in equation 64 satisfy

$$\left|\frac{2d}{N}y^\top(dD_U - P)r\right| \le 2dB\beta_{\max}(d\mu_U + 1) \le \frac{16d^2B^2}{\epsilon}\left(\mu_U + \frac{1}{d}\right).$$

Also, following the same derivation as in equation 51, we have

$$\frac{d^2}{N}y^\top D_U y - \|x\|_2^2 \le \left(\frac{K^2d}{\rho} - 1\right)R^2,$$

Taking expectation over (X,U) in equation 64, equations equation 53-equation 54 give

$$\mathbb{E}_{X,U}\left[\frac{1}{N}\sum_{j=1}^N w_ja_j\right] = \mathcal{L}^{(w)}(\Gamma^*) \le \mu_U\mathcal{L}(\Gamma^*). \tag{65}$$

Combining equation 63 and equation 64-equation 65, and using $d^2B^2 = dK^2R^2/\rho$ and $M = 2^b$, yields

$$\mathbb{E}\left\|\widehat{X} - X\right\|_2^2 \le \frac{dK^2R^2}{\rho}\left(1 + \frac{2\mu_UC_0^2}{(2^b-1)^2} + \frac{16(\mu_U + 1/d)}{\epsilon} + \frac{256\mu_U}{\epsilon^2}\right) - R^2.$$

This concludes the proof.

E.3 Optimal Codebook Bit-Width Allocation

Balancing the spatial term in Theorem 5.2 with the privacy-dependent terms gives, up to constants,

$$\frac{C^2}{(2^b-1)^2} \asymp \frac{1}{\epsilon \wedge \epsilon^2} \Longrightarrow 2^b - 1 \asymp C\sqrt{\epsilon \wedge \epsilon^2}.$$

Thus a convenient integer scaling is

$$b_{\mathrm{sat}} = \max\left\{1, \left\lceil \log_2\left(1 + \Theta\left(C\sqrt{\epsilon \wedge \epsilon^2}\right)\right)\right\rceil\right\}.$$

For $\epsilon \le 1$, one bit is already sufficient at the level of this bound; for $\epsilon \gg 1$, the threshold grows as $\frac{1}{2}\log_2\epsilon$ up to additive constants.

E.4 Population Law Underlying Codebook Optimization

Let X denote a random population input. The frame U is either fixed and public, or publicly randomized independently of X; a measurable Kashin-selection rule produces $y(X, U)$, and $J \sim \mathrm{Unif}[N]$ is independent. We write $Y = y_J(X, U)$. This distinction is essential: for fixed (x, U), the law induced only by J is the atomic measure $N^{-1} \sum_j \delta_{y_j}$ and need not have a density. Any density f_Y used in the codebook-optimization statements is therefore the population law over (X, U, J), not a density generated by frame randomness for one private vector.

The uniform-surrogate objective in Definition 4.1 uses only the public interval $[-B, B]$, where $B = KR/\sqrt{N}$. Optimization under the true f_Y is instead an oracle/public-distribution result. A codebook estimated from private client data would itself be data dependent and requires a separate privacy analysis (or a genuinely public auxiliary sample).

E.5 The Asymptotic Limit ($\epsilon \to \infty$) and the Panter–Dite Framework

The benefits of the codebook optimization formulation become most apparent when considering the limit as privacy constraints relax ($\epsilon \to \infty$). In this regime, the continuous Laplace density $f_{T|v}(t)$ concentrates at a Dirac delta centered at v, so the localized variance vanishes ($\Phi_k \to 0$) and the spatial scale modifier $W_k \to 1$. Note that the initial stochastic interpolation step remains active even in this limit; a non-codeword input is still rounded randomly to its two endpoints followed by deterministic identity output, which is the classical dithered quantizer.

Substituting these limits, the loss simplifies to the classical spatial quantization distortion:

$$\lim_{\epsilon \to \infty} \mathcal{L}_{\text{MA-SSTQ}}(\Gamma) = \sum_{k=1}^{M-1} \int_{c_k}^{c_{k+1}} (y - c_k)(c_{k+1} - y) f_Y(y)\, dy$$

In the high-resolution limit ($M \to \infty$) of classical quantization theory, the discrete codebook intervals can be approximated by a continuous point density function $\lambda(y)$. The spatial sum converges to the continuous integral $\frac{1}{6(M-1)^2} \int_{-B}^{B} \frac{f_Y(y)}{\lambda(y)^2}\, dy$.

The Panter-Dite framework (1951) (Panter & Dite, 1951) establishes that minimizing this spatial integral via Hölder's inequality strictly requires the continuous point density to satisfy: $\lambda^*(y) = \frac{f_Y(y)^{1/3}}{\int_{-B}^{B} f_Y(z)^{1/3}\, dz}$. Therefore, as local differential privacy constraints vanish, the optimal metric-aware codebook abandons variance regularization and allocates discrete coordinates exactly proportional to $f_Y(y)^{1/3}$. This formally bridges optimal differentially private randomization with the classical foundations of scalar source coding.

F Proofs for Federated Learning Convergence Analysis

This appendix provides the proofs for the non-convex Distributed Stochastic Gradient Descent (DSGD) convergence rate established in Theorem 6.1, as well as its specific operational corollaries for our SSTQ framework.

F.1 Formal Statement of the Optimization Conditions

We restate here in full the standing assumptions summarized in Section 6.

Assumption F.1 (Optimization Conditions)**.** *Let $\mathcal{F}_t = \sigma\big(w_0,\ \hat{g}_{i,s} : i \in [n],\ s < t\big)$ denote the σ-algebra generated by the initial point and all transmitted gradients prior to step t. Since w_t is a deterministic function of $(w_0, \hat{g}_{i,0}, \ldots, \hat{g}_{i,t-1})_{i=1}^n$, the iterate w_t is $\mathcal{F}_t$-measurable. Write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$. We assume:*

(i) ***Objective Smoothness.*** *The global objective $F(w)$ is L-smooth: $\|\nabla F(x) - \nabla F(y)\|_2 \le L \|x - y\|_2$ for all x, y, and is bounded below by F^*. Denote $\Delta_0 := F(w_0) - F^*$.*

(ii) ***Stochastic gradient bounds.*** *Let $g_{i,t} = \nabla f_i(w_t, \xi_{i,t})$ be the stochastic minibatch gradient sampled by client i at step t. Then, $\|g_{i,t}\|_2 \le G$ almost surely and $\mathbb{E}_t\big[\|g_{i,t} - \nabla f_i(w_t)\|_2^2\big] \le \sigma^2$.*

(iii) ***Compression noise.*** *Let $\hat{g}_{i,t}$ be the (private quantized) gradient transmitted by client i. We assume $\forall\, i,\ t$, $\mathbb{E}_t\big[\|\hat{g}_{i,t} - g_{i,t}\|_2^2\big] \le \omega\, G^2$.*

(iv) ***Bounded bias.*** *Define the global server estimator $\hat{g}_t = \frac{1}{n}\sum_{i=1}^n \hat{g}_{i,t}$ and the bias $b_t := \mathbb{E}_t[\hat{g}_t] - \nabla F(w_t)$. We assume $\|b_t\|_2^2 \le \zeta^2$, almost surely, for all t.*

(v) ***Conditional independence.*** *Conditioned on $\mathcal{F}_t$, the client estimator errors $\{\hat{g}_{i,t} - \nabla f_i(w_t)\}_{i=1}^n$ are mutually independent.*

F.2 Proof of Theorem 6.1

The proof proceeds in four parts: decomposing the Mean Squared Error (MSE) of the global estimator, establishing the descent lemma with bias, absorbing the bias via Young's inequality, and telescoping to yield the final convergence rate.

Throughout, we let $\mathcal{F}_t = \sigma\big(w_0,\ \hat{g}_{i,s} : i \in [n],\ s < t\big)$ denote the σ-algebra generated by the initial point and all transmitted gradients prior to step t. Since w_t is a deterministic function of $(w_0, \hat{g}_{i,0}, \dots, \hat{g}_{i,t-1})_{i=1}^n$, the iterate w_t is $\mathcal{F}_t$-measurable. We write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$ for the conditional expectation given $\mathcal{F}_t$.

Step 1: Gradient and MSE Decomposition.
Define the client-level error $X_{i,t} = \hat{g}_{i,t} - \nabla f_i(w_t)$. The global estimator error decomposes as

$$\hat{g}_t - \nabla F(w_t) = \frac{1}{n}\sum_{i=1}^n X_{i,t}.$$

Applying the standard bias-variance identity $\mathbb{E}[\|Y\|_2^2] = \mathbb{E}[\|Y - \mathbb{E}[Y]\|_2^2] + \|\mathbb{E}[Y]\|_2^2$ conditionally on $\mathcal{F}_t$, with $Y = \frac{1}{n}\sum_{i=1}^n X_{i,t}$:

$$\mathbb{E}_t\left[\left\|\frac{1}{n}\sum_{i=1}^n X_{i,t}\right\|_2^2\right] = \mathbb{E}_t\left[\left\|\frac{1}{n}\sum_{i=1}^n \big(X_{i,t} - \mathbb{E}_t[X_{i,t}]\big)\right\|_2^2\right] + \left\|\mathbb{E}_t\left[\frac{1}{n}\sum_{i=1}^n X_{i,t}\right]\right\|_2^2. \tag{66}$$

Variance term. By assumption (v), the centered variables $X_{i,t} - \mathbb{E}_t[X_{i,t}]$ are conditionally independent and mean-zero given $\mathcal{F}_t$. Hence, for the variance of their average:

$$\mathbb{E}_t\left[\left\|\frac{1}{n}\sum_{i=1}^n \big(X_{i,t} - \mathbb{E}_t[X_{i,t}]\big)\right\|_2^2\right] = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}_t\big[\|X_{i,t} - \mathbb{E}_t[X_{i,t}]\|_2^2\big] \le \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}_t\big[\|X_{i,t}\|_2^2\big], \tag{67}$$

where the inequality uses $\mathrm{Var}(Z) \le \mathbb{E}[\|Z\|_2^2]$ for any random variable Z.

Bias term. By linearity, the squared bias equals

$$\left\|\mathbb{E}_t\left[\frac{1}{n}\sum_{i=1}^n X_{i,t}\right]\right\|_2^2 = \|\mathbb{E}_t[\hat{g}_t] - \nabla F(w_t)\|_2^2 = \|b_t\|_2^2 \le \zeta^2 \quad \text{a.s.},$$

by assumption (iv).

Combining equation 66–equation 67:

$$\mathbb{E}_t\Big[\|\hat{g}_t - \nabla F(w_t)\|_2^2\Big] \le \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}_t\big[\|X_{i,t}\|_2^2\big] + \zeta^2. \tag{68}$$

Bounding the local MSE. We bound each $\mathbb{E}_t[\|X_{i,t}\|_2^2]$ by splitting the compression and stochastic errors. Using $\|A + B\|_2^2 \le 2\,\|A\|_2^2 + 2\,\|B\|_2^2$ (which holds for all vectors A, B without any

independence requirement):

$$\begin{aligned}\mathbb{E}_t\big[\|X_{i,t}\|_2^2\big] &= \mathbb{E}_t\big[\|(\hat{g}_{i,t} - g_{i,t}) + (g_{i,t} - \nabla f_i(w_t))\|_2^2\big] \\ &\le 2\,\mathbb{E}_t\big[\|\hat{g}_{i,t} - g_{i,t}\|_2^2\big] + 2\,\mathbb{E}_t\big[\|g_{i,t} - \nabla f_i(w_t)\|_2^2\big] \\ &\le 2\omega G^2 + 2\sigma^2 = 2(\omega G^2 + \sigma^2), \end{aligned} \tag{69}$$

where the last line applies assumptions (ii) and (iii).

Substituting equation 69 into equation 68:

$$\mathbb{E}_t\Big[\|\hat{g}_t - \nabla F(w_t)\|_2^2\Big] \le \frac{2(\omega G^2 + \sigma^2)}{n} + \zeta^2. \tag{70}$$

Second-moment bound. We additionally bound the conditional second moment of the update direction, which will be needed in Step 2. By the same inequality $\|A + B\|_2^2 \le 2\,\|A\|_2^2 + 2\,\|B\|_2^2$:

$$\begin{aligned}\mathbb{E}_t\big[\|\hat{g}_t\|_2^2\big] &= \mathbb{E}_t\big[\|(\hat{g}_t - \nabla F(w_t)) + \nabla F(w_t)\|_2^2\big] \\ &\le 2\,\mathbb{E}_t\big[\|\hat{g}_t - \nabla F(w_t)\|_2^2\big] + 2\,\|\nabla F(w_t)\|_2^2 \\ &\le 2\left(\frac{2(\omega G^2 + \sigma^2)}{n} + \zeta^2\right) + 2\,\|\nabla F(w_t)\|_2^2, \end{aligned} \tag{71}$$

where the last step substitutes equation 70. Note that $\nabla F(w_t)$ is $\mathcal{F}_t$-measurable and thus passes through $\mathbb{E}_t[\cdot]$ as a constant.

Step 2: The Descent Lemma with Bias.
By L-smoothness (assumption (i)) and the update rule $w_{t+1} = w_t - \eta\,\hat{g}_t$:

$$F(w_{t+1}) \le F(w_t) - \eta\,\langle \nabla F(w_t),\, \hat{g}_t\rangle + \frac{\eta^2 L}{2}\,\|\hat{g}_t\|_2^2. \tag{72}$$

Taking the conditional expectation $\mathbb{E}_t[\cdot]$ of both sides (noting $F(w_t)$ and $\nabla F(w_t)$ are $\mathcal{F}_t$-measurable):

$$\begin{aligned}\mathbb{E}_t[F(w_{t+1})] &\le F(w_t) - \eta\,\big\langle \nabla F(w_t),\, \mathbb{E}_t[\hat{g}_t]\big\rangle + \frac{\eta^2 L}{2}\,\mathbb{E}_t\big[\|\hat{g}_t\|_2^2\big] \\ &= F(w_t) - \eta\,\big\langle \nabla F(w_t),\, \nabla F(w_t) + b_t\big\rangle + \frac{\eta^2 L}{2}\,\mathbb{E}_t\big[\|\hat{g}_t\|_2^2\big] \\ &= F(w_t) - \eta\,\|\nabla F(w_t)\|_2^2 - \eta\,\langle \nabla F(w_t),\, b_t\rangle + \frac{\eta^2 L}{2}\,\mathbb{E}_t\big[\|\hat{g}_t\|_2^2\big]. \end{aligned} \tag{73}$$

Step 3: Absorbing the Bias via Young's Inequality.
We bound the inner-product bias term using the Cauchy–Schwarz inequality followed by Young's inequality ($ab \le \frac{a^2}{2} + \frac{b^2}{2}$ for $a, b \ge 0$):

$$-\langle \nabla F(w_t),\, b_t\rangle \le \|\nabla F(w_t)\|_2\,\|b_t\|_2 \le \frac{1}{2}\,\|\nabla F(w_t)\|_2^2 + \frac{1}{2}\,\|b_t\|_2^2 \le \frac{1}{2}\,\|\nabla F(w_t)\|_2^2 + \frac{\zeta^2}{2}, \tag{74}$$

where the last inequality uses $\|b_t\|_2^2 \le \zeta^2$ a.s. (assumption (iv)).

Substituting equation 74 and the second-moment bound equation 71 into equation 73:

$$\begin{aligned}\mathbb{E}_t[F(w_{t+1})] &\le F(w_t) - \frac{\eta}{2}\,\|\nabla F(w_t)\|_2^2 + \frac{\eta\,\zeta^2}{2} + \frac{\eta^2 L}{2}\left[2\,\|\nabla F(w_t)\|_2^2 + 2\left(\frac{2(\omega G^2 + \sigma^2)}{n} + \zeta^2\right)\right] \\ &= F(w_t) - \underbrace{\left(\frac{\eta}{2} - \eta^2 L\right)}_{\text{gradient coefficient}}\,\|\nabla F(w_t)\|_2^2 + \frac{\eta\,\zeta^2}{2} + \eta^2 L\left(\frac{2(\omega G^2 + \sigma^2)}{n} + \zeta^2\right). \end{aligned} \tag{75}$$

Step 4: Telescoping and Convergence.
Learning rate constraint. We require $\eta \le \frac{1}{4L}$, which ensures $\eta^2 L \le \frac{\eta}{4}$ and hence:

$$\frac{\eta}{2} - \eta^2 L \ge \frac{\eta}{2} - \frac{\eta}{4} = \frac{\eta}{4} > 0. \tag{76}$$

Substituting into equation 75:

$$\mathbb{E}_t[F(w_{t+1})] \leq F(w_t) - \frac{\eta}{4}\,\|\nabla F(w_t)\|_2^2 + \frac{\eta\,\zeta^2}{2} + \eta^2 L\left(\frac{2(\omega G^2+\sigma^2)}{n} + \zeta^2\right). \tag{77}$$

Telescoping. Rearranging equation 77:

$$\frac{\eta}{4}\,\|\nabla F(w_t)\|_2^2 \leq F(w_t) - \mathbb{E}_t[F(w_{t+1})] + \frac{\eta\,\zeta^2}{2} + \eta^2 L\left(\frac{2(\omega G^2+\sigma^2)}{n} + \zeta^2\right).$$

Taking the full (unconditional) expectation on both sides and applying the tower property $\mathbb{E}[\mathbb{E}_t[\cdot]] = \mathbb{E}[\cdot]$:

$$\frac{\eta}{4}\,\mathbb{E}\Big[\|\nabla F(w_t)\|_2^2\Big] \leq \mathbb{E}[F(w_t)] - \mathbb{E}[F(w_{t+1})] + \frac{\eta\,\zeta^2}{2} + \eta^2 L\left(\frac{2(\omega G^2+\sigma^2)}{n} + \zeta^2\right).$$

Summing over $t = 0, 1, \ldots, T-1$ (the right-hand side telescopes):

$$\frac{\eta}{4}\sum_{t=0}^{T-1}\mathbb{E}\Big[\|\nabla F(w_t)\|_2^2\Big] \leq F(w_0) - \mathbb{E}[F(w_T)] + \frac{T\eta\,\zeta^2}{2} + T\eta^2 L\left(\frac{2(\omega G^2+\sigma^2)}{n} + \zeta^2\right).$$

Using $\mathbb{E}[F(w_T)] \geq F^*$ (Assumption F.1(i)) and dividing both sides by $\frac{\eta T}{4}$:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\|\nabla F(w_t)\|_2^2\Big] \leq \frac{4\Delta_0}{\eta T} + 2\zeta^2 + 4\eta L\left(\frac{2(\omega G^2+\sigma^2)}{n} + \zeta^2\right), \tag{78}$$

where $\Delta_0 := F(w_0) - F^*$.

Optimal learning rate. Define the noise-plus-bias parameter $V := \frac{2(\omega G^2+\sigma^2)}{n} + \zeta^2$. The bound equation 78 has the form $\frac{4\Delta_0}{\eta T} + 2\zeta^2 + 4\eta L V$. We set $\eta = \min\{\frac{1}{4L},\ \sqrt{\frac{\Delta_0}{LVT}}\}$ and consider both cases.

Case 1: $T \geq T_0 := 16L\Delta_0/V$. In this regime, $\eta^* = \sqrt{\Delta_0/(LVT)} \leq \frac{1}{4L}$, and substituting into equation 78:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\|\nabla F(w_t)\|_2^2\Big] &\leq \frac{4\Delta_0}{\sqrt{\Delta_0/(LVT)}\cdot T} + 2\zeta^2 + 4\sqrt{\frac{\Delta_0}{LVT}}\cdot LV \\ &= \frac{8\sqrt{\Delta_0 LV}}{\sqrt{T}} + 2\zeta^2.\end{aligned} \tag{79}$$

Case 2: $T < T_0$. Here $\eta = \frac{1}{4L}$, and substituting into equation 78:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\|\nabla F(w_t)\|_2^2\Big] \leq \frac{16L\Delta_0}{T} + 2\zeta^2 + V. \tag{80}$$

For $T < T_0$, we have $\frac{8\sqrt{\Delta_0 LV}}{\sqrt{T}} > \frac{8\sqrt{\Delta_0 LV}}{\sqrt{T_0}} = \frac{8\sqrt{\Delta_0 LV}\cdot\sqrt{V}}{4\sqrt{L\Delta_0}} = 2V$, so the V term in the burn-in bound is dominated by the square-root term in Case 1 evaluated at T. Therefore, the unified bound valid for all $T \geq 1$ is:

$$\min_{0\leq t\leq T-1}\mathbb{E}\Big[\|\nabla F(w_t)\|_2^2\Big] \leq \frac{16L\Delta_0}{T} + \frac{8\sqrt{\Delta_0 LV}}{\sqrt{T}} + 2\zeta^2. \tag{81}$$

Expanding V and simplifying. Recall $V = \frac{2(\omega G^2+\sigma^2)}{n} + \zeta^2$. Thus $\sqrt{V} \leq \sqrt{\frac{2(\omega G^2+\sigma^2)}{n}} + \sqrt{\zeta^2}$, where we have used $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$. Substituting:

$$\begin{aligned}\frac{8\sqrt{\Delta_0 LV}}{\sqrt{T}} &\leq \frac{8\sqrt{\Delta_0 L}}{\sqrt{T}}\left(\sqrt{\frac{2(\omega G^2+\sigma^2)}{n}} + \zeta\right) \\ &= \frac{8\sqrt{2\,\Delta_0 L}\cdot\sqrt{\omega G^2+\sigma^2}}{\sqrt{nT}} + \frac{8\sqrt{\Delta_0 L}\cdot\zeta}{\sqrt{T}}.\end{aligned} \tag{82}$$

Combining the unified bound with equation 82, and using $\min_t \leq \frac{1}{T}\sum_t$:

$$\min_{0\leq t\leq T-1} \mathbb{E}\Big[\|\nabla F(w_t)\|_2^2\Big] \leq \frac{16L\Delta_0}{T} + \frac{8\sqrt{2\,\Delta_0 L}\cdot\sqrt{\omega G^2+\sigma^2}}{\sqrt{nT}} + \frac{8\sqrt{\Delta_0 L}\cdot\zeta}{\sqrt{T}} + 2\zeta^2. \tag{83}$$

Asymptotic simplification. For $T \geq T_0 = 16L\Delta_0/V$, the $\frac{16L\Delta_0}{T}$ burn-in term is dominated by $\frac{8\sqrt{\Delta_0 LV}}{\sqrt{T}}$. For the $\frac{8\sqrt{\Delta_0 L}\cdot\zeta}{\sqrt{T}}$ term, we apply the AM–GM inequality $\frac{\zeta}{\sqrt{T}} \leq \frac{\zeta^2}{2} + \frac{1}{2T}$, giving $\frac{8\sqrt{\Delta_0 L}\cdot\zeta}{\sqrt{T}} \leq 4\sqrt{\Delta_0 L}\,\zeta^2 + \frac{4\sqrt{\Delta_0 L}}{T}$. Since $T \geq 16L\Delta_0/V \geq 1$, the second term is $\mathcal{O}(1/T)$ and absorbed into the burn-in. The first term is absorbed into $\mathcal{O}(\zeta^2)$ with the problem-dependent constant $4\sqrt{\Delta_0 L}$. Therefore, the asymptotic rate (for $T \geq 16L\Delta_0/V$) is:

$$\min_{0\leq t\leq T-1} \mathbb{E}\Big[\|\nabla F(w_t)\|_2^2\Big] \leq \mathcal{O}\left(\frac{\sqrt{\omega G^2+\sigma^2}}{\sqrt{nT}} + \zeta^2\right), \tag{84}$$

where the $\mathcal{O}(\cdot)$ hides dependence on the problem constants Δ_0 and L. This completes the proof.

F.3 Proof of Corollary 6.2

Under Flat Randomized Response, the SSTQ mechanism employs the unbiasing scalar $\frac{1}{p-q}$ on the server. As established in Theorem 3.1, this renders the global estimator perfectly unbiased in expectation: $b_t = \mathbf{0}$, which mathematically enforces $\zeta = 0$.

By Theorem 3.2 (or more precisely, by Theorem 4.3 when $M \geq (4(e^\epsilon - 1))^{1/3} + 1$), the total variance multiplier under Flat RR is bounded by $\omega = \mathcal{O}((dK^2/\rho)[1 + 2^b/(\epsilon \wedge \epsilon^2)])$. Substituting $\zeta = 0$ and this variance multiplier directly into the general convergence bound of Theorem 6.1 gives:

$$\min_{0\leq t\leq T-1} \mathbb{E}[\|\nabla F(w_t)\|_2^2] \leq \mathcal{O}\left(\frac{1}{T} + \frac{\sqrt{\frac{dK^2}{\rho}[1+2^b/(\epsilon\wedge\epsilon^2)]G^2+\sigma^2}}{\sqrt{nT}}\right). \tag{85}$$

Because the bias is exactly zero, the gradient norm converges to zero as $T \to \infty$.

F.4 Proof of Corollary 6.3

The Metric-Aware Laplace framework trades absolute unbiasedness to address the $\mathcal{O}(2^b)$ Flat RR variance inflation. Under the max-gap and row-coherence assumptions, Theorem 5.2 gives $\omega = \mathcal{O}((dK^2/\rho)(1+\epsilon^{-1}+\epsilon^{-2}))$ and $\zeta^2 \leq 64K^2G^2/(\rho\epsilon^2)$. Using these bounds along with the bound established by Theorem 6.1, we get:

$$\min_{0\leq t\leq T-1} \mathbb{E}[\|\nabla F(w_t)\|_2^2] \leq \mathcal{O}\left(\frac{1}{T} + \frac{\sqrt{(dK^2/\rho)G^2(1+\epsilon^{-1}+\epsilon^{-2})+\sigma^2}}{\sqrt{nT}} + \frac{K^2G^2}{\rho\epsilon^2}\right). \tag{86}$$

This confirms that while the optimization trajectory converges to a non-zero stationary neighborhood dictated by ζ^2, the size of this neighborhood is explicitly independent of the massive parameter scale d. This algebraic decoupling breaks the dimensionality curse, ensuring that the gradient trajectory remains highly robust and computationally viable even for extremely large neural architectures.

G Experimental Details

Table 2: Communication cost per client per round ($b = 4$, $\epsilon = 3$).

Method	**Formula**	**Fashion-MNIST**	**CIFAR-10**
Clean / PrivUnit	$32d$	814,400	3,157,312
vqSGD	$\lceil \log_2 d \rceil + 1$	16	18
SQKR	$k(\lceil \log_2 N \rceil + 1)$	68	76
SSTQ (all variants)	$\lceil \log_2 N \rceil + b$	20	22

To efficiently implement the high-dimensional transformations required by SQKR-style constructions, we use a randomized partial Hadamard transform, which enables computation over an overcomplete frame of dimension $N = 2^{\lceil \log_2(2.5\,d) \rceil}$ in $O(N \log N)$ time without explicit dense matrix storage.

For the Kashin representation, we apply the exact-reconstruction variant of the iterative Lyubarskii–Vershynin algorithm with parameter $K = 2$. We run $r = 30$ truncated iterations at nominal level $KC/\sqrt{N}$, then add the exact residual correction. If a certified residual contraction $\|e_r\|_2 \le \eta^r C$ holds, the correction Ue_r can enlarge a coordinate by at most $\sqrt{\mu_U}\eta^r C$, so the implemented envelope is

$$B_r = \frac{KC}{\sqrt{N}} + \sqrt{\mu_U}\,\eta^r C.$$

For the randomized partial Hadamard frame with redundancy $N/d \approx 2.5$, we have $\eta \approx 0.4$, giving a residual of order $\eta^{30} \approx 10^{-12}$ after 30 iterations—well below machine precision. Consequently, the untruncated final coefficients have magnitude at most $\eta^{30} \cdot \sqrt{\mu_U} C$, which is negligible, and the effective Kashin level remains $K = 2$ to numerical precision.

On the server side, after averaging the W privatized messages, we apply gradient norm clipping with a threshold of 10.0 to the aggregate. Because clipping is a nonlinear operation, it can introduce a small bias even when the underlying per-client estimator is unbiased (see Remark 7.1 in the main text). In practice the cap activates infrequently: the client-side clipping at norm C combined with the frame scaling $B = K/\sqrt{N}$ keeps the expected aggregate norm well below 10.0, so the induced bias is negligible.

Remark G.1 (Effect of Learning Rate on High-Variance Mechanisms)**.** *A natural question is whether reducing the learning rate can compensate for the high variance introduced by certain privatization mechanisms. The convergence bound in Theorem 6.1 provides a precise answer. With step size η, the averaged gradient norm is bounded by $\frac{4\Delta_0}{\eta T} + 2\zeta^2 + 4\eta LV$, where $V = \frac{2(\omega G^2 + \sigma^2)}{n} + \zeta^2$. Reducing η shrinks the variance term $4\eta LV$ but inflates the optimization term $\frac{4\Delta_0}{\eta T}$; the optimal trade-off $\eta^* = \sqrt{\Delta_0/(LVT)}$ yields a rate of $\mathcal{O}(\sqrt{V/T})$. Consequently, a mechanism with ω times larger variance multiplier requires $\Theta(\omega)$ times more communication rounds to reach the same accuracy. For vqSGD, whose variance scales as $\Theta(d^3/\epsilon^2)$, this translates to an impractical number of rounds in high dimensions, regardless of step-size tuning.*

For the optimized codebook variant of SSTQ (Flat-RR), we solve the loss minimization problem

$$\mathcal{L}_{\mathrm{SSTQ}}(\Gamma) \;=\; \sum_{k=1}^{M-1} \frac{(c_{k+1} - c_k)^3}{12B} \;+\; \frac{1}{e^\epsilon - 1} \sum_{i=1}^{M} c_i^2$$

subject to $c_1 = -B,\ c_M = B,\ c_1 \le c_2 \le \cdots \le c_M,\ \sum_i c_i = 0$, using Sequential Least-Squares Programming (SLSQP). The codebook is precomputed once per configuration (d, b, ϵ) using the worst-case bound $B = K\,C/\sqrt{N}$, which is constant across all clipped gradients.

Remark G.2 (Effect of server-side clipping)**.** *Server-side gradient clipping is a nonlinear operation: $\mathbb{E}[\mathrm{clip}(\hat{g})] \neq \mathrm{clip}(\mathbb{E}[\hat{g}])$ in general, so clipping introduces a small bias even when the underlying estimator is unbiased. In the convergence framework of Theorem 6.1, this clipping-induced bias is absorbed into the ζ^2 term. In our experiments, the client-side clipping at $C = 0.1$ combined with the frame scaling $B = KR/\sqrt{N}$ ensures that the reconstructed one-coordinate estimator has bounded norm, and the server-side cap of 10.0 activates infrequently. The theoretical convergence rates in Corollaries 6.2 and 6.3 hold exactly for the unclipped algorithm; the clipped variant converges to a neighborhood that includes the additional clipping bias.*

Remark G.3 (Per-round vs. cumulative privacy budget)**.** *All reported privacy parameters are* per-round *guarantees: each client's message in each round satisfies ϵ-LDP independently. Under basic composition with T rounds of full participation, the cumulative budget is $\epsilon_{\mathrm{total}} = T\epsilon$. We report per-round ϵ (i.e., message-level privacy) because it isolates the mechanism's intrinsic privacy–utility trade-off and is the standard metric across all baselines compared (vqSGD, SQKR, PrivUnit). In practice, client subsampling (where each client participates in only a fraction of rounds) provides privacy amplification by subsampling (Balle et al., 2018), and advanced composition (Kairouz et al., 2015) yields $\epsilon_{\mathrm{total}} = \mathcal{O}(\epsilon\sqrt{T\log(1/\delta)})$ under (ϵ, δ)-DP. These orthogonal techniques apply equally to all mechanisms and are independent of the quantization scheme.*

Computational cost. All Kashin-based methods (SQKR, SSTQ) share the same computational bottleneck: the iterative Kashin representation, which involves repeated applications of the randomized partial Hadamard transform on a vector of dimension N, costing $O(N \log N)$ per client per round. In our experiments, each SGD step takes approximately 18 s on Fashion-MNIST and 65 s on CIFAR-10 (single CPU), with the Kashin transform accounting for over 99% of the per-worker computation. The remaining operations (quantization, RR or Laplace sampling) are $O(M)$ for SSTQ and $O(N)$ for SQKR. The server-side reconstruction requires applying a single column of the frame transform, costing $O(d)$ per client, which is negligible compared to the Kashin encoding.

In contrast, vqSGD avoids the Kashin representation entirely: it operates directly on the gradient vector using simplex sampling, costing only $O(d)$ per client. This makes vqSGD significantly faster per step (approximately 5–10$\times$ compared to Kashin-based methods), but as demonstrated in Section 7, this computational advantage is negated by its prohibitive variance scaling in high dimensions.

PrivUnit also avoids Kashin representations ($O(d)$ per client) but requires transmitting full-dimensional vectors, making it communication-efficient only in a per-dimension sense.

All experiments were implemented in Python using NumPy and SciPy, and were run on a single CPU.

H Additional Experiments

We conduct additional experiments designed to directly test the variance bounds, communication efficiency, and downstream model utility across varying dimensions d and codebook sizes. In these experiments, we compare seven methods: Clean (no privacy), PrivUnit, vqSGD, SQKR, SSTQ (Flat-RR with optimized codebook), SSTQ (Flat-RR with uniform codebook), and SSTQ (Metric-Aware with uniform codebook).

Empirical Variance Scaling (Theorem A.1 vs. Theorem 3.2). In this experiment, we isolate the client-to-server gradient transmission step and focus on computing the raw Mean Squared Error (MSE) between the true vector and the reconstructed estimator. We generate a single gradient vector as $g \sim \mathcal{N}(0, I_d)$ and normalize it to have norm $C = 0.2$, to ensure bounded sensitivity under LDP. We pass the same vector g into each algorithm, which quantizes the vector, adds LDP noise, and reconstructs an estimated vector $\hat{g}$. We sweep the dimension d over $\{100, 500, 1000, 5000, 10000\}$ and plot the MSE $\mathbb{E}[\|g - \hat{g}\|_2^2]$ versus d on a log-log scale, by averaging over 1000 independent trials per dimension. All methods use per-round privacy parameter $\epsilon = 3$ and bit budget $b = 8$ (codebook size $M = 2^8 = 256$).

The 95% confidence intervals are sufficiently narrow that they are not visually discernible on the log-log scale, reflecting the reliability of the Monte Carlo estimates at this sample size.

The results, reported in Figure 2, confirm the theoretical predictions. vqSGD exhibits devastating cubic $O(d^3/\epsilon^2)$ scaling, with MSE growing from ≈ 561 at $d = 100$ to $\approx 4.4 \times 10^8$ at $d = 10{,}000$—an increase consistent with the cubic ratio. In contrast, PrivUnit, SQKR, and SSTQ (Metric-Aware) all exhibit linear scaling in d, consistent with the $O(d/\epsilon^2)$ bound of Theorem 3.2. Among these, SSTQ (Metric-Aware) achieves the lowest MSE across all dimensions (≈ 2.0 at $d = 100$ and ≈ 171 at $d = 10{,}000$), followed by SQKR (≈ 4.2 to ≈ 328) and PrivUnit (≈ 2.9 to ≈ 767).

The two Flat-RR variants exhibit substantially higher MSE, now revealed by the exact second-moment computation. SSTQ (Flat-RR, uniform codebook) has MSE ranging from ≈ 446 at $d = 100$ to $\approx 35{,}746$ at $d = 10{,}000$, reflecting the $O(M^2)$ variance inflation from the RR debiasing scalar at large $M = 256$. SSTQ (Flat-RR, optimized codebook) reduces this to ≈ 53 at $d = 100$ and $\approx 5{,}072$ at $d = 10{,}000$—a 7–8$\times$ improvement over the uniform variant, confirming that codebook optimization mitigates but does not eliminate the RR variance. Both Flat-RR variants still incur MSE far above the Metric-Aware variant (by 27–30$\times$), demonstrating the decisive advantage of the Metric-Aware Discrete Laplace mechanism in eliminating the exponential 2^b dependence.

Comparison of LDP mechanisms for linear regression across varying dimensions. We study the convergence of distributed stochastic gradient descent (SGD) on a synthetic least-squares problem with objective

$$L(\theta) = \frac{1}{2n}\|A\theta - b\|_2^2 .$$

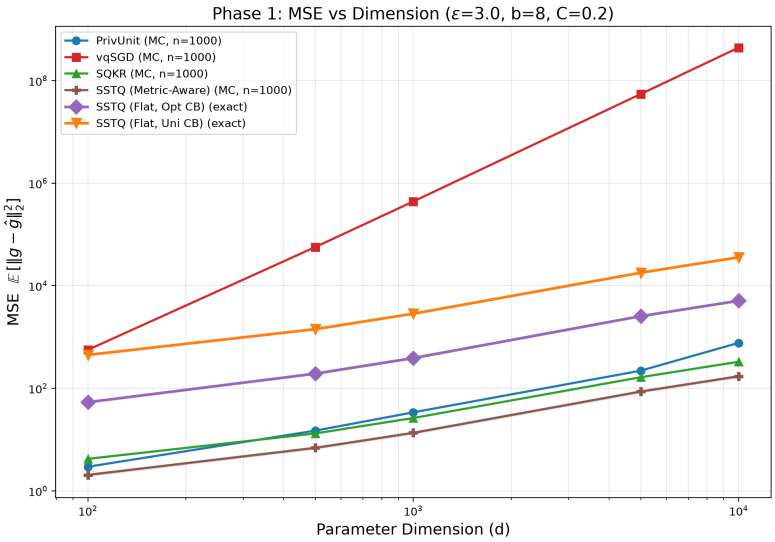

Figure 2: MSE versus dimension for all methods (per-round $\epsilon = 3$, $b = 8$, $M = 256$, $C = 0.2$). Flat-RR variants are computed exactly (no Monte Carlo); other methods use 1000 MC trials with 95% confidence bands.

We denote by $\theta^* \in \mathbb{R}^d$ the ground-truth parameter. The design matrix $A \in \mathbb{R}^{n\times d}$ has entries drawn i.i.d. from $\mathcal{N}(0, 1/d)$, ensuring stable conditioning across dimensions. We generate $\theta^* \sim \mathcal{N}(0, I_d)$ and set $b = A\theta^*$, yielding a realizable model.

We simulate a federated setting with $n = 4{,}000$ samples distributed uniformly across $W = 200$ workers. Each worker computes local gradients on batches of size 20 over $T = 100$ communication rounds. Gradients are clipped to $\|g\|_2 \le C = 1$, to ensure bounded sensitivity under LDP. Then privatized and transmitted to the server, which performs aggregation and updates using a constant step size $\eta = 0.5$. Performance is measured by the squared estimation error $\|\theta_t - \theta^*\|_2^2$.

All private methods operate under pure ϵ-LDP with per-round $\epsilon = 3$. We first evaluate convergence at bit budget $b = 4$ across dimensions $d \in \{100, 200, 500\}$ (Figure 3). For methods requiring frame representations (SSTQ and SQKR), we use an overcomplete frame of dimension $N = 2^{\lceil \log_2(2.5d)\rceil}$.

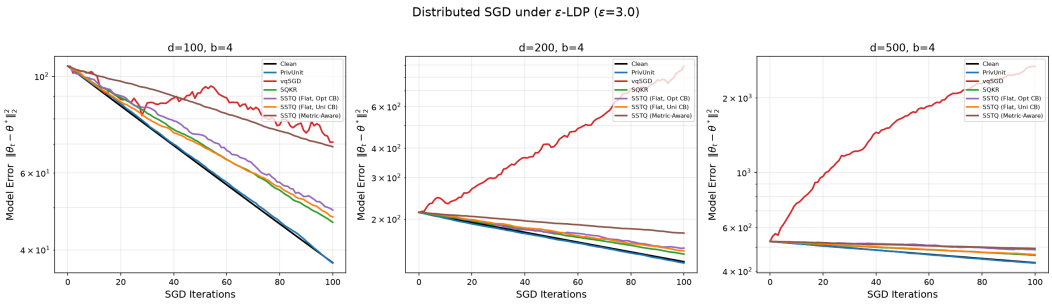

Figure 3: Distributed SGD convergence under ϵ-LDP ($\epsilon = 3$, $b = 4$). Comparison of the two variants of our proposed framework—SSTQ (Flat-RR) and SSTQ (Metric-Aware)—against other vector quantization algorithms across $d \in \{100, 200, 500\}$.

Figure 3 illustrates distributed least-squares SGD under pure ϵ-LDP ($\epsilon = 3$). Consistent with the observations in Section 7, vqSGD exhibits unstable optimization behavior that becomes more pronounced as the dimension increases: the model error *increases* monotonically, reaching ≈ 2685 at $d = 500$ compared to ≈ 432.6 for the clean baseline. While reducing the learning rate can stabilize individual updates, by Remark G.1 this trades variance reduction against slower optimization: the $\Theta(d^3/\epsilon^2)$ variance multiplier of vqSGD would require a proportionally larger number of communication rounds to converge, rendering it impractical in high dimensions. In contrast, the 1-sparse structure of SSTQ mitigates this dependence on the ambient dimension, resulting in stable and efficient convergence.

At $b = 4$ (moderate codebook size $M = 16$), the Flat-RR variants converge closely to the clean baseline, with SSTQ (Flat-RR, uniform codebook) achieving final error ≈ 132.8 at $d = 200$ versus ≈ 131.7 for the clean update—a modest overhead. SSTQ (Flat-RR, optimized codebook) achieves final error ≈ 140.5 at $d = 200$. The Metric-Aware variant converges to a slightly higher error floor of ≈ 162 at $d = 200$, consistent with the non-zero bias ζ^2 introduced by the identity decoder (Corollary 6.3). PrivUnit tracks the clean baseline most closely (≈ 130.0), as it incurs no quantization overhead.

Comparison of SSTQ variants across codebook sizes. To isolate the effect of codebook size on the Flat-RR variance, we fix $d = 200$ and compare convergence at $b = 4$ ($M = 16$) versus $b = 8$ ($M = 256$) in Figure 4.

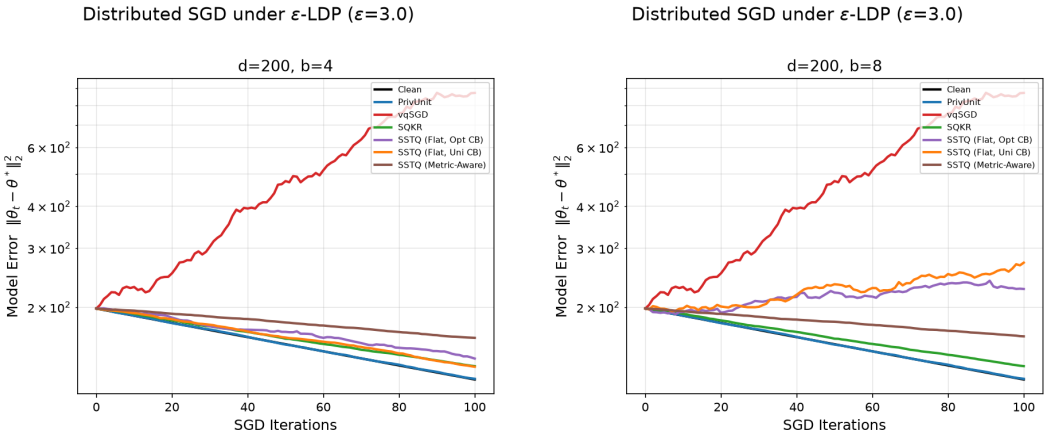

Figure 4: Flat-RR explosion versus Metric-Aware stability at $d = 200$, $\epsilon = 3$. **Left:** Low bit-width ($b = 4$, $M = 16$). **Right:** Dense bit-width ($b = 8$, $M = 256$).

As shown in Figure 4, increasing the codebook size from $M = 16$ to $M = 256$ causes a dramatic degradation in both Flat-RR variants. SSTQ (Flat-RR, uniform codebook) sees its final error grow from ≈ 132.8 to ≈ 271.6 ($2.0\times$ increase), while SSTQ (Flat-RR, optimized codebook) grows from ≈ 140.5 to ≈ 226.7 ($1.6\times$ increase). This is consistent with the predicted $O(M^2)$ scaling of the unbiasing factor: the debiasing scalar $1/(p - q) = (e^\epsilon + M - 1)/(e^\epsilon - 1)$ grows from ≈ 1.8 at $M = 16$ to ≈ 14.4 at $M = 256$, inflating the variance quadratically.

In stark contrast, SSTQ (Metric-Aware) remains essentially unchanged, with final error growing only marginally from ≈ 162.0 to ≈ 163.7. This stability arises because the Metric-Aware Discrete Laplace mechanism does not require the uniform debiasing correction—its noise distribution concentrates naturally around the true quantization level regardless of M. This behavior highlights the limitations of Flat-RR in dense quantization regimes and motivates the use of Metric-Aware estimators for maintaining stable performance at larger codebook sizes.