Price of metric universality in vector quantization is at most 0.11 bit

Alina Harbuzova
Massachusetts Institute of Technology. Supported by MathWorks Fellowship and Siebel Scholarship

Or Ordentlich
Hebrew University of Jerusalem. Supported by the Israel Science Foundation (ISF), grant No. 2878/25

Yury Polyanskiy
Massachusetts Institute of Technology. Supported by NSF Grant No. 2112665 via subaward KR 704702 from the University of California, San Diego

Editors: Steve Hanneke and Tor Lattimore

Abstract

Fast computation of a matrix product $W^\top X$ is a workhorse of modern LLMs. To make their deployment more efficient, a popular approach is that of using a low-precision approximation $\widehat{W}$ in place of true W ("weight-only quantization"). Information theory demonstrates that an optimal algorithm for reducing precision of W depends on the (second order) statistics of X and requires a careful alignment of vector quantization codebook with PCA directions of X (a process known as "waterfilling allocation"). Dependence of the codebook on statistics of X, however, is highly impractical. This paper proves that there exist a universal codebook that is simultaneously near-optimal for all possible statistics of X, in the sense of being at least as good as an X-adapted waterfilling codebook with rate reduced by 0.11 bit per dimension in the case when W is Gaussian. Such universal codebook would be an ideal candidate for the low-precision storage format, a topic of active modern research, but alas the existence proof is non-constructive.

Equivalently, our result shows existence of a net in $\mathbb{R}^n$ that is a nearly-optimal covering of a sphere simultaneously with respect to all Hilbert norms.

Keywords: vector quantization, oracle bounds, rate-distortion, waterfilling, regret, universality

1. Introduction

The most basic element of all modern AI is a neural unit: given a (dynamically changing) activation vector $X \in \mathbb{R}^n$ the unit needs to compute the output

$$Y = W^\top X\,,$$

where $W \in \mathbb{R}^n$ is a (static) weight. The problem, rapidly becoming central for economic deployment and continued evolution of large language models (LLMs), is to reduce storage/communication requirement by saving W in "low-precision". (The symmetric question of also converting X to low-precision is outside of scope of this work, though see Ordentlich and Polyanskiy (2025) for some recent theoretical analysis.)

Early deep learning models saved W in full-precision (known as FP32, and corresponding to $R = 32$ bit / coordinate), but soon moved to half-precision (FP16/BF16, or $R = 16$ bit). In the domain of LLMs, pioneering work Dettmers et al. (2022) showed that very little degradation is introduced if W is approximated by rescaling it to appropriate range and then rounding each coordinate of normalized W to nearest integer in $\{-128, -127, \dots, 127\}$, a so called *INT8* quantization.

Subsequently, more sophisticated low-precision storage formats were introduced, with currently the most popular being NVFP4 (rate $R = 4.5$ bit) and MXFP4 (rate $R = 4.25$ bit), see NVIDIA et al. (2025); Open Compute Project (2023).

In this paper we are focusing on a fundamental question: what is the best way of reducing precision of W? That is, how to replace W by a version $\widehat{W}$ that incurs minimal degradation of performance, i.e. $\widehat{W}^\top X \approx W^\top X$, while admitting a short bit-length description. A natural way to do that, known as *vector quantization*, would be to pre-define a *codebook* $\mathbf{C} \subset \mathbb{R}^n$ of size $|\mathbf{C}| = 2^{nR}$. Clearly, any element of $\mathbf{C}$ can be described by nR bits, hence achieving rate R of bits / coordinate. Given $\mathbf{C}$ we approximate W as

$$\widehat{W} = \operatorname*{argmin}_{c \in \mathbf{C}} d(W, c)$$

for some distance metric $d(\cdot, \cdot)$. When d is a standard Euclidean metric, then the problem reduces to a classical vector quantization problem in $\mathbb{R}^n$, with many classical solutions including lattices and trellis-coded constructions Gersho and Gray (2012). However, as was brilliantly shown by Frantar et al. (2023) large savings can be made if metric $d(\cdot, \cdot)$ is chosen with the knowledge of statistics of X in mind.

Indeed, if X is modeled as random with second-order statistics $\Sigma_X = \mathbb{E}[XX^\top]$ then

$$\mathbb{E}_X\left[(Y - \widehat{Y})^2\right] = \mathbb{E}_X\left[(W^\top X - \widehat{W}^\top X)^2\right] = (W - \widehat{W})^\top \Sigma_X (W - \widehat{W})\,.$$

Thus, we see that a natural choice of metric (given knowledge of Σ_X) is

$$d_{\Sigma_X}(W, \widehat{W}) = \mathbb{E}_X\left[(W^\top X - \widehat{W}^\top X)^2\right] = (W - \widehat{W})^\top \Sigma_X (W - \widehat{W})\,. \qquad (d_{\Sigma_X})$$

The easiest way to demonstrate how adaptation to Σ_X can significantly improve rate-distortion tradeoff is to consider a rank-1 case, i.e. when X is always collinear with a fixed vector $v \in \mathbb{R}^n$. In this case, a clever choice of the codebook $\mathbf{C}$ is $\{0, \pm\epsilon v, \pm 2\epsilon v, \ldots\}$, i.e. very fine quantization along a single direction v in $\mathbb{R}^n$. Indeed, by not needing to spread the points of $\mathbf{C}$ among all n dimensions, one can get exponential improvement in quality of approximation of $W^\top X$, since only the scalar value $W^\top v$ affects the result. Since activations in LLMs are notoriously low-rank, this adaptation of $\mathbf{C}$ to directions of principal variation (PCA) of X understandably improves performance.

Herein, however, lies the main problem that we are trying to address: while adapting $\mathbf{C}$ to the statistics of the input X is desirable, it may not be generally possible due to restrictions of hardware. Indeed, the mapping from actual bits (loaded from memory) to elements of $\mathbf{C}$ needs to be fixed at hardware design stage and cannot depend on statistics of X (in particular, because the same hardware is used for implementing different neurons, facing different types of X). Below we call this requirement, alternatively, as *universal codebook* or *a Σ_X-oblivious decoder*, to reflect the fact that $\mathbf{C}$ has to be universal across all possible choices of Σ_X.

To continue with more quantitative investigation, let us make a modeling assumption (well justified by empirical statistics of LLM weight matrices) that $W \sim \mathcal{N}(0, I_n)$. In this case a given codebook $\mathbf{C}$ under Σ_X statistics attains *distortion*:

$$D(\mathbf{C}, \Sigma_X) := \frac{1}{n}\mathbb{E}_W\left[\min_{c \in \mathbf{C}} d_{\Sigma_X}(W, c)\right]\,. \qquad (1)$$

Classical information-theoretic field, known as rate-distortion theory, establishes that for a codebook $\mathbf{C}$ to achieve distortion $D(\mathbf{C}, \Sigma_X) \le D$ one must have

$$\log |\mathbf{C}| \ge n\mathbf{R}_{\text{wf}}(\Sigma_X, D)\,,$$

where $\mathbf{R}_{\text{wf}}$ is given by a so-called *waterfilling formula*, see Prop. 2.

The main question of this work: *How much does the requirement of universality cost in terms of performance?* For example, in the rank-1 case above the waterfilling codebook would allocate its elements along a single direction v. This codebook, however, would be grossly suboptimal for another rank-1 Σ_X which has its PCA direction orthogonal to v. Somewhat surprisingly, thus, we show that nevertheless the answer is *not much*. The main result of this work is demonstration of existence of a universal $\mathbf{C}$, which is simultaneously near optimal for all possible Σ_X. Informally, we can state our main result as follows.

Theorem 1 (Informal: Universality costs ≤ 0.11 Bits) *Let us assume that $W \sim \mathcal{N}(0, I_n)$ and let $\mathbf{R}_{\text{wf}}(\Sigma_X, D)$ denote the information-theoretic (waterfilling) lower bound on rate needed to achieve distortion at most D in the* oracle *setting, where codebook is optimized for a fixed Σ_X. There exists a universal codebook $\mathbf{C}$ with 2^{nR} points such that its distortion* simultaneously *for all $\Sigma_X \in \mathbb{S}_+^n$ satisfies:*

$$R \le \mathbf{R}_{\text{wf}}(\Sigma_X, D(\mathbf{C}, \Sigma_X)) + 0.11\,\text{bit}\,.$$

The implication for hardware design is clear: it is possible to create a universal low-precision storage format (for W) that is optimal (up to rate gap of at most 0.11 bit) simultaneously for all kinds of distributions of statistics of the other factor (X) in the inner-product. Our result can also be interpreted as a statement about metric entropy: *There exists a universal net on a unit sphere, which covers unit sphere near-optimally simultaneously for all possible Hilbert norms on $\mathbb{R}^n$.*

Paper organization. The following Section 2 formalizes weight-only quantization for inner products under the distortion d_{Σ_X}. We then introduce the oracle benchmark given by the Gaussian rate–distortion tradeoff under weighted MSE, attained by the waterfilling solution in the setting where both encoder and decoder know the second-order statistics Σ_X (Prop. 2). Our main results are stated in Theorems 3 and 5. Theorem 3 shows the existence of a universal codebook achieving the explicit rate-distortion tradeoff (RDRC) over all Σ_X. Theorem 5 upper bounds the worst-case rate overhead incurred by this universal decoder relative to the oracle waterfilling benchmark.

Section 3 provides a proof sketch of Theorem 3 and the main geometric ideas: Section 3.1 describes the universal codebook construction and intuition, and Section 3.2 outlines the random-coding analysis leading to (RDRC).

Complete proofs are given in the Appendix. While the results in Section 2 are presented for $W \sim \mathcal{N}(0, I_n)$, we first establish a general bound for a fixed (non-random) W in Section B. This result is then specialized to $W \sim \mathcal{N}(0, I_n)$ using concentration and covering argument in Section C, completing the proof of Theorem 3. Theorem 5 is proved in Section D: we derive explicit expressions for the rate-gap and prove that the maximum gap occurs at spectra with at most 2 distinct eigenvalues and vanishing distortions.

2. Main results and discussion

Consider an arbitrary $\Sigma_X \succeq 0$ and define a (square of) Hilbert metric with respect to Σ_X as in (d_{Σ_X}). We consider the problem of obtaining a low-precision (at rate R bits per coordinate) representation

$\hat{W}$ of a random vector $W \in \mathbb{R}^n$ with the goal of minimizing $d_{\Sigma_X}(\hat{W}, W)$. We focus presentation of results on the the standard setting in which W is an isotropic Gaussian vector:

$$W \sim \mathcal{N}(0, I_n)\,,$$

though the key technical results hold for general W (Section B). An (n, R) quantization scheme consists of an encoder $f : \mathbb{R}^n \to \left[2^{nR}\right]$ and decoder $g : \left[2^{nR}\right] \to \mathbb{R}^n$ and we set $\widehat{W} = g(f(W))$. The image of g is called the codebook $\mathbf{C} := \operatorname{im} g$. The optimal encoder consists of finding a nearest to W element of $\mathbf{C}$, and hence we can equivalently think of a quantization scheme as completely defined by $\mathbf{C}$. The distortion of $\mathbf{C}$ for a given Σ_X is denoted $D(\mathbf{C}, \Sigma_X)$, cf. (1).

Let us start with a simple case of Σ_X fixed (and hence known to both encoder f_{Σ_X} and decoder g_{Σ_X}). In this case, the optimal tradeoff between the distortion D and rate R is given by waterfilling, which we review.

Let $\Sigma_X = U\Lambda U^\top$ be the eigendecomposition with $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Consider a parametric curve

$$D_{\mathrm{wf}}(\Sigma_X, t) = \frac{1}{n}\sum_{i=1}^{n} \min\{\lambda_i, t\} \qquad R_{\mathrm{wf}}(\Sigma_X, t) = \frac{1}{2n}\sum_{i=1}^{n} \max\{0, \log(\lambda_i/t)\}\,, \tag{WF}$$

where t is the parameter (waterfilling level). The above implicitly defines waterfilling distortion as a function of rate:

$$\mathbf{D}_{\mathrm{wf}}(R, \Sigma_X) \triangleq D_{\mathrm{wf}}(\Sigma_X, t_{\mathrm{wf}}(R))\,, \quad \text{where } R_{\mathrm{wf}}(\Sigma_X, t_{\mathrm{wf}}(R)) = R\,. \tag{$\mathbf{D}_{\mathrm{wf}}$}$$

It turns out that this function indeed determines the fundamental limits in the case of oracle-knowledge of Σ_X. More exactly, we have the following.

Proposition 2 (Waterfilling) *Let $\Sigma_X \in \mathbb{S}_+^n$. For any compression scheme of rate R we have*

$$\mathbb{E}\left[d_{\Sigma_X}(W, g(f(W)))\right] \geq n\mathbf{D}_{\mathrm{wf}}(R, \Sigma_X)\,. \tag{2}$$

Conversely, for any $B > 0$ there exist a $c = c(B) > 0$ such that for any Σ_X there exist f and g (both depending on Σ_X) such that

$$\mathbb{E}\left[d_{\Sigma_X}(W, g(f(W)))\right] \leq n\mathbf{D}_{\mathrm{wf}}\left(R - c\sqrt{\frac{\log n}{n}}, \Sigma_X\right) + cn^{-B}\operatorname{tr}\Sigma_X\,. \tag{3}$$

Proof We give the proof of the lower bound (2) below; since the upper bound (3) is not relevant for the rest of this paper, we defer the brief of sketch of the proof to Appendix F. Though we do emphasize that the upper bound does not follow from classical theory, which concerns with separable (additive over coordinates) distortion measures, and we have to invoke more modern single-shot bounds, cf. (Polyanskiy and Wu, 2024, Chapter 25).

Let $\Sigma_X = U\Lambda U^\top$ with $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, and define $W' \triangleq U^\top W$, $\widehat{W}' \triangleq U^\top \widehat{W}$. Then

$$d_{\Sigma_X}(W, \widehat{W}) = (W - \widehat{W})^\top \Sigma_X (W - \widehat{W}) = (W' - \widehat{W}')^\top \Lambda (W' - \widehat{W}') = \sum_{i=1}^{n} \lambda_i (W_i' - \widehat{W}_i')^2\,.$$

In the oracle setting (where both the encoder and decoder know Σ_X), we may equivalently compress W', reconstruct $\widehat{W}'$, and output $\widehat{W} = U\widehat{W}'$. After this coordinate change, the problem becomes that of a weighted mean squared error. Since $W' \sim \mathcal{N}(0, I_n)$ has independent coordinates, a standard data-processing argument (see (Polyanskiy and Wu, 2024, Section 23.4 and Theorem 6.1)) gives:

$$nR \geq I(W'; \widehat{W}') \geq \sum_{i=1}^{n} I(W_i'; \widehat{W}_i')\,,$$

Writing $D_i \triangleq \mathbb{E}\left[(W_i' - \widehat{W}_i')^2\right]$, the smallest $I(W_i'; \hat{W}_i') = \frac{1}{2}\log\frac{1}{D_i}$ (given the value D_i) is then attained under Gaussian coupling, cf. (Polyanskiy and Wu, 2024, Section 26.1.2), which results in

$$R \geq \frac{1}{2n}\sum_{i=1}^{n}\log\frac{1}{D_i} \qquad \text{and} \qquad \frac{1}{n}\mathbb{E}\left[d_{\Sigma_X}(W, \widehat{W})\right] = \frac{1}{n}\sum_{i=1}^{n}\lambda_i D_i\,.$$

Minimizing $\frac{1}{n}\sum_i \lambda_i D_i$ subject to the rate constraint via Lagrange multipliers gives the (reverse) waterfilling optimum, summarized by the parametric curve in (WF). ∎

Now, as we discussed above, Σ_X describes distribution of activations and, practically speaking, is usually unavailable to decoder. Indeed, even if the eigenvalues Λ were known, optimal waterfilling requires knowledge of the eigenbasis U, since the decoder g_{Σ_X} essentially computes $U \cdot \widehat{U^\top W}$, where $\widehat{U^\top W}$ is the closest codeword to $U^\top W$. Communicating $U \in \mathcal{O}(n)$ to the decoder is expensive: it requires approximating $\Theta(n^2)$ real parameters, which will consume much larger than $\Theta(n)$ bits allocated for communicating W itself.[1] Thus, the tradeoff in (WF) is unattainable via a naïve "send W and Σ_X" scheme when decoder lacks Σ_X.

To capture the limitation above, we need to assume that Σ_X is available at *encoding* time, but unavailable at *decoding* time (since the deployed dequantizer is fixed, possible even in hardware). So, formally we define a universal (n, R) quantization scheme as a pair

$$f : \mathbb{R}^n \times \mathbb{S}_+^n \to [2^{nR}], \qquad g : [2^{nR}] \to \mathbb{R}^n\,, \tag{f, g}$$

where the encoder takes Σ_X as an input, while the decoder has no access to Σ_X and outputs $\widehat{W} = g(f(W, \Sigma_X))$. Again, the image of g is called the codebook $\mathbf{C} = \operatorname{im} g$, which is independent of Σ_X.

The goal is to design a pair (f, g) such that $D(\mathbf{C}, \Sigma_X)$ were low simultaneously for all Σ_X. Our first main result proves existence of a universal codebook $\mathbf{C}$ with an explicit guarantee on the achieved distortion. To define that guarantee, again let $\Sigma_X \in \mathbb{S}_+^n$ with $\operatorname{tr}(\Sigma_X) = n$ and spectrum $\lambda = (\lambda_1, \ldots, \lambda_n)$. The *random-coding rate-distortion function* is a parametric curve (with parameter $T > 0$) given by

$$D_{\mathrm{rc}}(\lambda, T) = \frac{1}{n}\sum_{i=1}^{n}\frac{\lambda_i}{1+\lambda_i T} \qquad R_{\mathrm{rc}}(\lambda, T) = \frac{1}{2n}\sum_{i=1}^{n}\log(1+\lambda_i T)\,. \tag{RDRC}$$

Denote

$$\mathbf{D}_{\mathrm{rc}}(\lambda, R) =: D_{\mathrm{rc}}(\lambda, T_{\mathrm{rc}}(\lambda, R))\,, \quad \text{where } R_{\mathrm{rc}}(\lambda, T_{\mathrm{rc}}(\lambda, R)) = R\,. \tag{$\mathbf{D}_{\mathrm{rc}}$}$$

1. In practice, GPU computes not a single inner product but $W^\top X$ for $W \in \mathbb{R}^{n\times a}$ being a matrix with $a \asymp n$. Thus, some of the cost of sending U ammortizes over a, but still makes it a highly suboptimal choice.

Theorem 3 (Main Result I: Universal Quantization Scheme for Gaussian Input) *Fix any constants $R^\star, \varepsilon, \eta, B > 0$. There exists an encoder $f : \mathbb{R}^n \times \mathbb{S}_+^n \times [0,1] \to [2^{nR}]$, a decoder $g : [2^{nR}] \times [0,1] \to \mathbb{R}^n$ with $R \le R^\star + \varepsilon$, and a (shared) random variable $S \in [0,1]$ with the following property. For $W \sim \mathcal{N}(0, I_n)$ and $\Sigma_X \in \mathbb{S}_+^n$ we set $\widehat{W}(\Sigma_X) = g\left(f(W, \Sigma_X, S), S\right)$. Then, for sufficiently large $n \ge n_0 = n_0(\varepsilon, \eta, R^\star, B)$, we have with probability at least $1 - \exp\left(-n^B\right)$ over $S \sim \mathrm{Unif}[0,1]$ that*

$$\frac{1}{n}\mathbb{E}_W\left[d_{\Sigma_X}(W, \widehat{W}(\Sigma_X))\right] \le \mathbf{D}_{\mathrm{rc}}(\mathrm{spec}(\Sigma_X), R^\star) + \eta$$

simultaneously for all Σ_X with $\operatorname{tr}\Sigma_X = n$.

Theorem 3 uses S to generate an entire codebook $\mathbf{C}$ and shows that with high probability it achieves $\approx \mathbf{D}_{\mathrm{rc}}(\Sigma_X, R)$ distortion. Of course, by fixing a value of S it implies *existence* of a single codebook $\mathbf{C}$ with the same property. Formally, we have a corollary.

Corollary 4 *Under the same assumptions as in Thm 3, there exists a codebook $\mathbf{C}$ of size $\log_2|\mathbf{C}| \le n(R^\star + \varepsilon)$ and an encoder-decoder pair f, g (see Eq. (f, g)) with the following property. Set $\widehat{W}(\Sigma_X) = g(f(W, \Sigma_X))$. Then, simultaneously for all Σ_X we have*

$$\frac{1}{n}\mathbb{E}_W\left[d_{\Sigma_X}(W, \widehat{W}(\Sigma_X))\right] \le \mathbf{D}_{\mathrm{rc}}(\mathrm{spec}(\Sigma_X), R^\star) + \eta\frac{\operatorname{tr}\Sigma_X}{n}\,.$$

The description of the scheme and intuition are in Sec. 3 and the full proof in Appendix C. While the results are presented for isotropic Gaussian W, our proof proceeds by showing a more general result for a fixed non-random W (see Sec. B). Then, the result for Gaussian W is obtained by using concentration of measure (Sec. C).

What Theorem 3 shows is that, roughly speaking, there exists a universal codebook of rate R which achieves distortion $\mathbf{D}_{\mathrm{rc}}(\mathrm{spec}(\Sigma_X), R)$ simultaneously for all Σ_X. A natural question is: *how far is $\mathbf{D}_{\mathrm{rc}}(\mathrm{spec}(\Sigma_X), R)$ from the oracle waterfilling benchmark $\mathbf{D}_{\mathrm{wf}}(\mathrm{spec}(\Sigma_X), R)$?*

To make the comparison easier to interpret, we will phrase it in terms of rate overhead (of our codebook) compared to Σ_X-fine-tuned optimal codebook. Specifically, for a fixed spectrum $\lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \succeq 0$ and a distortion level $D^\star$, let $\mathbf{R}_{\mathrm{wf}}(\lambda, D^\star) = R_{\mathrm{wf}}(\lambda, t)$ be the minimum oracle rate achieving $D_{\mathrm{wf}}(\lambda, t) = D^\star$ (see Eq. (WF)), and let $\mathbf{R}_{\mathrm{rc}}(\lambda, D^\star) = R_{\mathrm{rc}}(\lambda, T)$ be the minimum random-coding rate achieving $D_{\mathrm{rc}}(\lambda, T) = D^\star$ (see Eq. (RDRC)). The difference $\mathbf{R}_{\mathrm{rc}}(\lambda, D^\star) - \mathbf{R}_{\mathrm{wf}}(\lambda, D^\star)$ is the rate overhead incurred by using a universal decoder g that is agnostic to the covariance matrix Σ_X with $\mathrm{spec}(\Sigma_X) = \lambda$.

Our second main result in Theorem 5 shows that this overhead is uniformly bounded by 0.11 bit. Specifically, we derive precise expressions for $\mathbf{R}_{\mathrm{rc}}(\lambda, D^\star) - \mathbf{R}_{\mathrm{wf}}(\lambda, D^\star)$ that depend on $\lambda = \mathrm{spec}(\Sigma_X)$ and $D^\star$ and prove that the maximum gaps occur at the spectra with at most 2 distinct eigenvalues and vanishing distortions (see Fig. 1 for the worst-case rate gap found at each $R = \mathbf{R}_{\mathrm{rc}}(\lambda, D^\star)$). See Sec. D for the full proof.

Theorem 5 (Main Result II: Worst-Case Rate Gap to Oracle Setting)

$$\sup_{D^\star \in (0,1)} \sup_{\Lambda} \left\{\mathbf{R}_{\mathrm{rc}}(\Lambda, D^\star) - \mathbf{R}_{\mathrm{wf}}(\Lambda, D^\star)\right\} \le 0.11\,,$$

where the supremum is over $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \succeq 0$ with $\operatorname{tr}(\Lambda) = n$.

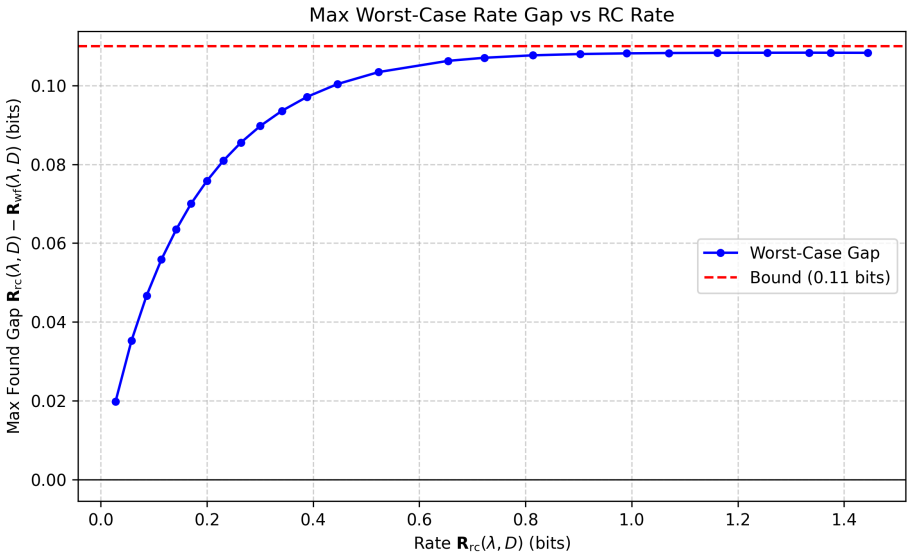

Figure 1: Maximum rate gap found at each rate R.

An intuitive way to see why the gap is bounded is to consider the most extreme of $\operatorname{tr}\Sigma_X = n$ cases: the identity and the rank-1 case, for which $\mathbf{R}_{\mathrm{rc}}(\lambda, D^\star) = \mathbf{R}_{\mathrm{wf}}(\lambda, D^\star)$ by a simple computation. In fact, the gap vanishes for all matrices with semi-flat spectra $\lambda = (n/m, \ldots, n/m, 0, \ldots, 0), m \leq n$, a phenomenon we further discuss in Sec. 3.1.1.

Discussion and Open Questions. Our results prove the *existence* of a universal codebook whose performance is uniformly within a constant rate gap of the Σ_X-aware (waterfilling) oracle benchmark, showing that universality is not, by itself, an information-theoretic bottleneck. Our random-coding-based proof is nonconstructive and does not yield an explicit codebook design or efficient encoder/decoder pair. Designing explicit and computationally efficient constructions is the most immediate open direction.

Often, low complexity quantizers are based on lattices. Indeed, for any fixed Σ_X a lattice randomly drawn from the natural Haar-Siegel measure will be a good quantizer with probability $1 - e^{-\Omega(n)}$. This follows from the results of Ordentlich et al. (2022) that show that the covering radius of a random lattice is typically near-optimal with respect to any (fixed) norm. However, *no lattice can be near-optimal simultaneously for all* Σ_X, and the reason is simple: for any fixed lattice $L \subset \mathbb{R}^n$ there is a rotation $U^\top L$ that aligns its directions with the natural basis. Consequently, for this rotation $U^\top L = \prod_{i=1}^n (\alpha_i \mathbb{Z})$ for some $\alpha_1, \ldots, \alpha_n$. The integer lattice is a bad quantizer, and therefore any lattice quantizer must lose at least $\frac{1}{2}\log\frac{2\pi e}{12} \approx 0.254$ bits with respect to the waterfilling benchmark for some Σ_X (see Ordentlich and Polyanskiy (2026) for more details). It therefore follows that Σ_X-universal near optimal schemes cannot solely rely on lattice quantizers.

One example of a lattice-based algorithm that provides a practical solution to the problem studied here is the GPTQ algorithm Frantar et al. (2023) (with appropriate shaping/entropy coding Ordentlich and Polyanskiy (2026)). Its high-rate gap to the waterfilling benchmark, for particular Σ_X, is $\frac{1}{2}\log\frac{2\pi e}{12} + \frac{1}{2}\log(\mathrm{AM} - \mathrm{GM}(\Sigma_X))$ where $\mathrm{AM} - \mathrm{GM}(\Sigma_X)$ is the ratio between arithmetic-mean and geometric-mean of the squared diagonal elements in the Cholesky decomposition of Σ_X, which can be unbounded in general Ordentlich and Polyanskiy (2026). When W is a matrix consisting of $a \gg 1$ rows (rather than the vector case considered here), some of this gap can be reduced by send-

ing to the decoder $o(na)$ bits of information on Σ_X (which has negligible effect on the rate). One such example is the WaterSIC quantization scheme Lifar et al. (2026).

Related literature. The problem of vector quantization is classical Gersho and Gray (2012), and its asymptotic behavior (for iid sources and additive distortion) is given by the famous rate-distortion formulas, e.g. (Polyanskiy and Wu, 2024, Part V). A recent wave of interest, however, focuses on computing quantized inner-product and matrix multiplication.

On the practical side, the pioneering work of Dettmers et al. (2022) demonstrated that substantial compression is possible via scaling plus uniform rounding (INT8 weight quantization), thus establishing the field of post-training quantization (PTQ). Notable PTQ works include SmoothQuant Xiao et al. (2024), which introduced calibration-based methods, i.e. those which depend on statistics of activations X via Σ_X. GPTQ Frantar et al. (2023) and LDLQ Chee et al. (2024), which are equivalent, simultaneously introduced an algorithm for Σ_X-dependent quantization. Going beyond simple integer-rounding, Tseng et al. (2024) consider lattice, Savkin et al. (2025) consider nested lattice and Tseng et al. (2025a) consider trellis quantization methods, respectively. Going beyond simple quadratic losses, are Tseng et al. (2025b) and Badri and Shaji (2023). We note that Tseng et al. (2024) also reintroduced random Hadamard transform (RHT) as a way of mitigating outliers, following earlier usage in quantization of gradients, and classically. See Ashkboos et al. (2024); Liu et al. (2025); Chen et al. (2025a) for other applications of RHT in PTQ.

On the theoretical side, the work Ordentlich and Polyanskiy (2025) established fundamental limits of quantized matrix multiplication by leveraging nested lattice quantization. The GPTQ/LDLQ algorithm was understood as Babai's nearest-plane algorithm applied after a Cholesky factorization of Σ_X in Chen et al. (2025b); Birnick (2025), and as a successive interference cancellation (SIC) algorithm in Ordentlich and Polyanskiy (2026), thus connecting weight-only quantization to lattice decoding and approximate closest vector problems Conway and Sloane (1982); Babai (1986). Authors of Ordentlich and Polyanskiy (2026) developed theoretical high-rate analysis of GPTQ, showed it can be arbitrarily far from waterfilling and proposed an improved algorithm which provably matches waterfilling to within 0.255 bit in the high-rate regime.

The question considered here (Σ_X-oblivious quantization) in the special case of diagonal Σ_X falls under the umbrella of the *compression with distortion as side-information* proposed in Martinian et al. (2008). This viewpoint connects modern LLM quantization to a long tradition in lattice decoding and approximate closest vector problems Conway and Sloane (1982); Babai (1986).

3. Technical Overview

As in any source-coding problem, once the codebook $\mathbf{C} = \{c_1, \dots, c_M\} \subset \mathbb{R}^n$ is fixed, the optimal encoder computes $i^* = i^*(W) = \operatorname{argmin}_{i\in[M]} d_{\Sigma_X}(W, c_i)$, and sends the index i^* to the decoder, which in turn outputs $\widehat{W} = c_{i^*}$. Thus, the distortion of the codebook $\mathbf{C}$ is

$$nD(\mathbf{C}, \Sigma_X) = \mathbb{E}\left[\min_{i\in[M]} d_{\Sigma_X}(W, c_i)\right], \tag{4}$$

where the expectation is with respect to $W \sim \mathcal{N}(0, I_n)$. Even if we could design $\mathbf{C}$ based on Σ_X, the distortion $D(\mathbf{C}, \Sigma_X)$ must satisfy (see Sec. 2) the waterfilling lower bound

$$D(\mathbf{C}, \Sigma_X) \geq \mathbf{D}_{\text{wf}}(R, \Sigma_X)\,;$$

the lower bound is also asymptotically achievable in the limit of large n (see Proposition 2). The challenge is to find a single codebook $\mathbf{C}$ with $M = 2^{nR}$ codewords in $\mathbb{R}^n$ that attains small $D(\mathbf{C}, \Sigma_X) - \mathbf{D}_{\mathrm{wf}}(R, \Sigma_X)$ simultaneously for all Σ_X.

As is standard, we prove the existence of such a codebook $\mathbf{C}$ by drawing a random code with $M = 2^{nR}$ iid codewords from a distribution $P_{\widehat{W}}$. We show in Theorem 3 that for appropriate choice of $P_{\widehat{W}}$ we have that

$$\Pr_{\mathbf{C}}\left[\sup_{\Sigma_X}\left(D(\mathbf{C}, \Sigma_X) - \mathbf{D}_{\mathrm{rc}}(\mathrm{spec}(\Sigma_X), R - \varepsilon)\right) < \eta\right] \geq 1 - \exp\left(-\mathrm{poly}(n)\right) \tag{5}$$

holds for any $\eta, \varepsilon > 0$ and n large enough, where $\mathrm{spec}(\Sigma_X)$ is the vector of the eigenvalues of Σ_X and $\mathbf{D}_{\mathrm{rc}}$ is defined in Eq. ($\mathbf{D}_{\mathrm{rc}}$). Consequently there must exist a fixed rate-R codebook $\mathbf{C}$ with

$$D(\mathbf{C}, \Sigma_X) \leq \mathbf{D}_{\mathrm{rc}}(\mathrm{spec}(\Sigma_X), R - \varepsilon) + \eta, \quad \forall \Sigma_X \in \mathbb{S}_+^n \text{ with } \mathrm{tr}(\Sigma_X) = n. \tag{6}$$

3.1. Codebook Distribution

How should we choose $P_{\widehat{W}}$? For a given Σ_X with spectral decomposition $\Sigma_X = U \Lambda U^\top$, the optimal $P_{\widehat{W}}$ follows from the waterfilling solution. Specifically, for water-level $1/t$ chosen so that R_{wf} defined in (WF) equals R, the optimal distribution is

$$\begin{aligned} P^*_{\widehat{W}}(\Sigma_X, R) = P^*_{\widehat{W}}(\Sigma_X, t) &= \mathcal{N}(0, U\,\Gamma(\Lambda, t)\,U^\top), \\ \text{where } \Gamma(\Lambda, t) &= \mathrm{diag}\left(\max\left\{1 - \frac{t}{\lambda_1}, 0\right\}, \cdots, \max\left\{1 - \frac{t}{\lambda_n}, 0\right\}\right). \end{aligned} \tag{7}$$

We need to choose a single $P_{\widehat{W}}$ that "works well" for all Σ_X. Since there is no preference to any $U \in \mathcal{O}_n$, it makes sense to take an isotropic Gaussian $P_{\widehat{W}}$. Observing that for any fixed Σ_X and $t > 0$ the covariance matrix for $P^*_{\widehat{W}}(\Sigma_X, t)$ satisfies $U\,\Gamma(\Lambda, t)\,U^\top \preceq I_n$, we will take

$$P_{\widehat{W}}(\tau) = \mathcal{N}(0, \tau^2 I_n), \tag{8}$$

for some $0 < \tau < 1$.

An appealing feature of the isotropic Gaussian distribution is that drawing M iid vectors from $\mathcal{N}(0, \tau^2 I_n)$ is equivalent to first drawing them iid from $\mathcal{N}(0, I_n)$ and then scaling all of them by τ. The consequence of this simple fact is that while we cannot perfectly match our codebook distribution to $P^*_{\widehat{W}}(\Sigma_X, R)$, the flexibility in the choice of $\tau = \tau(\Sigma_X, R)$ allows for a better match. Consequently, we draw the M codewords of $\mathbf{C}$ from the $\mathcal{N}(0, I_n)$ distribution. The encoder, that knows Σ_X, computes $\tau(\Sigma_X, R)$ that provides the smallest expected distortion, and sends a description of τ to the decoder.[2] With this procedure, the effective codebook $\tilde{\mathbf{C}} = \tau \mathbf{C}$ is drawn from $P_{\widehat{W}} = \mathcal{N}(0, \tau^2(\Sigma_X, R) I_n)$. The encoder then finds $i^* = \mathrm{argmin}_{i \in [M]}\, d_{\Sigma_X}(W, \tilde{c}_i) = \mathrm{argmin}_{i \in [M]}\, d_{\Sigma_X}(W, \tau c_i)$ and sends i^* as well as a high-resolution description of τ in bits to the decoder.

2. As we will see below, some further gain can be attained by allowing τ to also depend on the source realization $w \in \mathbb{R}^n$.

3.1.1. GEOMETRIC INTUITION

To get some intuition to why the "universal" codebook distribution $P_{\widehat{W}} = \mathcal{N}(0, \tau^2(\Sigma_X, R) I_n)$ works well simultaneously for all Σ_X, let us restrict attention to the family of covariance matrices with semi-flat spectrum. In particular, for $m \le n$ let

$$\mathcal{S}_m^n = \left\{ \Sigma_X = U \Lambda U^\top \ : \ U \in \mathcal{O}_n, \ \lambda_1 = \cdots = \lambda_m = \frac{n}{m}, \ \lambda_{m+1} = \cdots = \lambda_n = 0 \right\}, \tag{9}$$

be the collection of PSD matrices with $m \le n$ equal and non-zero singular values and $m - n$ zero singular values. From (7) we see that the Σ_X-matched optimal codebook distribution is of the form

$$P^*_{\widehat{W}}(\Sigma_X, R) = \mathcal{N}\left(0, U \cdot \operatorname{diag}\left(\tau^2, \cdots, \tau^2, 0, \cdots, 0\right) U^\top\right) \tag{10}$$

for some $\tau > 0$. Thus the optimal procedure for random coding is to draw iid Gaussian codewords within the subspace $U_{[m]}$ spanned by the first m singular vectors. Our universal distribution, on the other hand, draws isotropic iid Gaussian codewords, and is hence very far from the optimal distribution. However, since the encoder searches for the nearest codeword under the d_{Σ_X} metric, it effectively projects both W and the codebook $\mathbf{C}$ to $U_{[m]}$ and finds the nearest codeword in ℓ_2-metric within this subspace. It therefore follows that what dictates performance of a random code $\mathbf{C}$ under d_{Σ_X} metric (for Σ_X with semi-flat spectrum) is the distribution of $U_{[m]}^\top c$, where $c \sim P_{\widehat{W}}$. Thus, our universal $P_{\widehat{W}} = \mathcal{N}(0, \tau^2 I_n)$ is simultaneously optimal for all semi-flat Σ_X (with all possible $m \in [n]$), provided that we judiciously choose $\tau = \tau(\Sigma_X, R)$. Inspection of our Σ_X-universal rate-distortion tradeoff (RDRC) shows that indeed

$$\mathbf{D}_{\mathrm{rc}}(\operatorname{spec}(\Sigma_X), R) = \mathbf{D}_{\mathrm{wf}}(\operatorname{spec}(\Sigma_X), R),$$

for all semi-flat Σ_X. Whenever the spectrum of Σ_X is not semi-flat the distribution of $U^\top c$ for $c \sim P_{\widehat{W}}$ does not match that of $U^\top c$ under the optimal $P^*_{\widehat{W}}(\Sigma_X, R)$, and consequently in these cases our Σ_X-universal rate-distortion tradeoff (RDRC) is worse than the waterfilling rate-distortion tradeoff. Nevertheless, it turns out that the loss for this mismatch is at most 0.11 bits, as shown in Theorem 5.

3.2. Sketch of Proof

After we obtained intuition for the choice of using a random iid isotropic Gaussian codebook, with scale τ determined by the encoder, we move on to giving an overview of the proof of Theorem 3. The detailed rigorous proof is given in Appendix C.

Let us first fix $\Sigma_X = U \Lambda U^\top$, and analyze the performance of a codebook $\mathbf{C} = \{c_1, \ldots, c_M\}$, $M = 2^{nR}$, with $c_i \overset{iid}{\sim} \mathcal{N}(0, I_n)$. We will show that $\mathbf{C}$ is "good" for Σ_X with probability $1 - \exp(-e^{\Omega(n)})$, and from this we will deduce that a random $\mathbf{C}$ is "good" for all Σ_X by a covering argument.

The codebook $\mathbf{C}$ can describe a fixed $w \in \mathbb{R}^n$ with distortion $\le D$ if at least one of its τ-scaled codewords is inside the region $w + \sqrt{D}\mathcal{B}_{\Sigma_X}$, where

$$\mathcal{B}_{\Sigma_X} = \left\{ e \in \mathbb{R}^n \ : e^\top \Sigma_X e \le n \right\}. \tag{11}$$

Thus, the key to analyzing the tradeoff between rate and distortion is understanding how the success probability of a single codeword behaves as a function of D. Since the codewords are $\mathcal{N}(0, I_n)$, their distribution is invariant to rotation, and therefore the success probability of a single codeword is

$$\begin{aligned} p_{\text{success}}(w, \tau, \Sigma_X, D) &= \phi_{\tau^2}\left(w + \sqrt{D}\mathcal{B}_{\Sigma_X}\right) \\ &= \phi_{\tau^2}\left(U^\top w + \sqrt{D}\mathcal{B}_{\Lambda}\right) = p_{\text{success}}(U^\top w, \tau, \Lambda, D), \end{aligned} \tag{12}$$

where ϕ_{τ^2} denotes the probability distribution for $\mathcal{N}(0, \tau^2 I_n)$. When $p_{\text{success}} < 2^{-n(R+\varepsilon)}$, it is very unlikely to find a codeword in $w + \sqrt{D}\mathcal{B}_{\Sigma_X}$, and on the other hand, when $p_{\text{success}} > 2^{-n(R-\varepsilon)}$ we are very likely to find a codeword in $w + \sqrt{D}\mathcal{B}_{\Sigma_X}$. Thus, what we are looking for is the critical D for which $-\frac{1}{n}\log p_{\text{success}}(U^\top w, \tau, \Lambda, D) \approx R$. Since the codebook's scale τ needs to be sent from the encoder to the decoder anyway, we may let it depend not only on Λ but also on $U^\top w$. Therefore, given $U^\top w, \Lambda$ and R we choose $\tau = \tau(U^\top w, \Lambda, R)$ for which the critical D is small. A tedious but straightforward calculation shows that the optimal choice is

$$\tau = \tau(U^\top w, \Lambda, R) = \left(T\sum_j \frac{(U^\top w)_j^2 \lambda_j^2}{(1+\lambda_j T)^2}\right)^{1/2} \left(\sum_j \frac{\lambda_j}{1+\lambda_j T}\right)^{-1/2}, \tag{13}$$

where $T = T(\Lambda, R)$ is such that $R_{\text{rc}}(\Lambda, T) = R$, and $R_{\text{rc}}(\Lambda, T)$ is defined in (RDRC). Let

$$D_{\text{rc}}(U^\top w) = D_{\text{rc}}(\Lambda, R, U^\top w) = \frac{1}{n}\sum_{i=1}^n \frac{(U^\top w)_i^2 \lambda_i}{1+\lambda_i T}, \tag{14}$$

where here as well $T = T(\Lambda, R)$. In Lemma 10 we prove that

$$-\log p_{\text{success}}(U^\top w, \tau, \Lambda, D_{\text{rc}}(U^\top w)) = -\log \phi_{\tau^2}\left(U^\top w + \sqrt{D_{\text{rc}}(U^\top w)}\mathcal{B}_{\Lambda}\right) \approx nR. \tag{15}$$

Thus, $D_{\text{rc}}(U^\top w)$ is the critical distortion for fixed $w \in \mathbb{R}^n$ and Σ_X. Note that for $W \sim \mathcal{N}(0, I_n)$ we have

$$\mathbb{E}_W\left[D_{\text{rc}}(U^\top W)\right] = \frac{1}{n}\sum_{i=1}^n \frac{\lambda_i}{1+\lambda_i T} = D_{\text{rc}}(\Lambda, R), \tag{16}$$

where $D_{\text{rc}}(\Lambda, R)$ is the distortion in ($\mathbf{D}_{\text{rc}}$).

The proof of Lemma 10 uses standard large deviation techniques, but note that what we really need for the analysis is a quantitative lower bound on p_{success}, and this requires some more work beyond Chernoff bound. The choice of τ from (13) is the one that minimizes the large deviations exponent, whereas the parameter T in our rate-distortion tradeoff (RDRC) is just a rescaling of the parameter t in Chernoff's bound $\Pr(X > D) \le e^{-tD}\mathbb{E}[e^{tX}]$.

By (15), for $\eta > 0$ we have that $p_{\text{success}}(U^\top w, \tau, \Lambda, D_{\text{rc}}(U^\top w) + \eta/2) \ge 2^{-n(R-\varepsilon_\eta)}$ for some $\varepsilon_\eta > 0$ (and n large enough). Thus, $\forall w \in \mathbb{R}^n$

$$\begin{aligned} P_{\text{failure}}(w) &= \Pr_{\mathbf{C}}\left(\frac{1}{n}\min_{i\in[M]} d_{\Sigma_X}(w, \tau c_i) > D_{\text{rc}}(U^\top w) + \eta/2\right) \\ &= 1 - \left(1 - p_{\text{success}}\left(U^\top w, \tau, \Lambda, D_{\text{rc}}(U^\top w) + \eta/2\right)\right)^M \le \exp(-2^{n\varepsilon}), \end{aligned} \tag{17}$$

for some $\varepsilon > 0$. For a *fixed* code $\mathbf{C}$ define

$$\mathcal{E}_{\text{failure}}(\mathbf{C}) = \left\{ w \in \mathbb{R}^n \ : \ \frac{1}{n} \min_{i \in [M]} d_{\Sigma_X}(w, \tau c_i) > D_{\text{rc}}(U^\top w) + \eta/2 \right\}. \tag{18}$$

Assuming we can always encode w to 0 (e.g., by setting $\tau = 0$ if needed), for any $W \sim P_W$ we have

$$\begin{aligned} \frac{1}{n} D(\mathbf{C}, \Lambda) &= \mathbb{E}\left[\min_{i \in [M]} d_{\Sigma_X}(W, \tau c_i) \right] \le \mathbb{E}\left[D_{\text{rc}}(U^\top W) + \frac{\eta}{2} + \mathbb{1}\{W \in \mathcal{E}_{\text{failure}}(\mathbf{C})\} d_{\Sigma_X}(W, 0) \right] \\ &\le \mathbb{E}\left[D_{\text{rc}}(U^\top W) \right] + \frac{\eta}{2} + \sqrt{P_W[\mathcal{E}_{\text{failure}}(\mathbf{C})]} \cdot \sqrt{\mathbb{E}[d^2_{\Sigma_X}(W, 0)]}, \end{aligned} \tag{19}$$

where the last inequality is Cauchy-Schwarz. If $\mathbb{E}_W[d^2_{\Sigma_X}(W, 0)] \le \text{poly}(n)$ under $W \sim P_W$,[3] it follows that whenever $P_W[\mathcal{E}_{\text{failure}}(\mathbf{C})]$ is sufficiently small, say $\le e^{-n}$, the codebook $\mathbf{C}$ attains the desired

$$\frac{1}{n} D(\mathbf{C}, \Lambda) \le \mathbb{E}\left[D_{\text{rc}}(U^\top W) \right] + \eta \,.$$

$P_W[\mathcal{E}_{\text{failure}}(\mathbf{C})]$ can indeed be bounded: using (17) and Markov's inequality we obtain that this holds for the vast majority of codebooks:

$$\begin{aligned} \Pr_{\mathbf{C}}\left(P_W[\mathcal{E}_{\text{failure}}(\mathbf{C})] > e^{-n} \right) &\le e^n \mathbb{E}_{\mathbf{C}}\left[P_W[\mathcal{E}_{\text{failure}}(\mathbf{C})] \right] = e^n \mathbb{E}_{\mathbf{C},W}[P_{\text{failure}}(W)] \\ &\le \exp(-2^{n\varepsilon} + n). \end{aligned} \tag{20}$$

From this we conclude that for fixed P_W and any fixed Σ_X

$$\Pr_{\mathbf{C}}\left(\left[\frac{1}{n} D(\mathbf{C}, \Lambda) - \mathbb{E}_W\left[D_{\text{rc}}(U^\top W) \right] > \eta \right] \right) \le \exp(-2^{n\varepsilon'}). \tag{21}$$

Specializing this to $P_W = \mathcal{N}(0, I_n)$ we obtain

$$\Pr_{\mathbf{C}}\left(\left[\frac{1}{n} D(\mathbf{C}, \Lambda) - D_{\text{rc}}(\Lambda, R) \right] > \eta \right) \le \exp(-2^{n\varepsilon'}). \tag{22}$$

We have therefore obtained that for any fixed Σ_X the probability of drawing a "bad" codebook with $\left[\frac{1}{n} D(\mathbf{C}, \Lambda) - D_{\text{rc}}(\Lambda, R)\right] > \eta$ is double-exponentially small. From here, it is clear how to prove that a randomly drawn $\mathbf{C}$ will have

$$\Pr_{\mathbf{C}}\left(\sup_{\Sigma_X} \left[\frac{1}{n} D(\mathbf{C}, \Lambda_X) - D_{\text{rc}}(\Lambda, R) \right] > \eta \right) < \exp(-2^{n\varepsilon''}). \tag{23}$$

All we need is to find a dense cover of PSD matrices in $R^{n \times n}$ with trace n whose size is $\exp(-2^{o(n)})$. In particular, we need to find a collection of $N = \exp(-2^{o(n)})$ PSD matrices with the property that for any valid PSD matrix Σ_X there exists a matrix Σ_i, $i \in [N]$ such that:

$$D_{\text{rc}}(\text{spec}(\Sigma_X), R) \approx D_{\text{rc}}(\text{spec}(\Sigma_i), R) \qquad \text{and} \qquad \frac{1}{n} D(\mathbf{C}, \Sigma_X) \approx \frac{1}{n} D(\mathbf{C}, \Sigma_i) \,. \tag{24}$$

3. e.g., $\mathbb{E}_W[d^2_{\Sigma_X}(W, 0)] \le O(n^2)$ for $W \sim \mathcal{N}(0, I_n)$

Since N is allowed to be so large, these two constraints can be met to arbitrary resolution (though the proof for this requires a lot of bookkeeping and is somewhat technical, see Sec. C.2).

We end this overview with listing some of the technical issues that our sketch of proof above ignored, and briefly mention how addressing them, as we of course do in the actual proofs, affects the results.

τ quantization. While our sketch assumed that τ can be conveyed to the decoder in perfect resolution, in reality, some of our nR bits budget is allocated to the description of τ. In order to compress τ that depends on $U^\top W$, we require a high probability bound on $\|U^\top W\|_\infty$. This norm constraint is also needed for the proof of our large deviations result (Lemma 10).

Norm bound for codewords. The perturbation argument (24) for the dense cover must account for quantization and change in τ. Since these errors are multiplied by codewords from $\mathbf{C}$, we require a uniform norm bound $\|c_i\|_2 < n^B$ for some $B > 1$ (say $B = 10$) to make their contribution to end-to-end distortion negligible. While the probability that this occurs is overwhelmingly large, it is only $1 - \exp(-\mathrm{poly}(n))$ rather than double exponential. For this reason the probability of drawing a Σ_X-universal codebook is only $1 - \exp(-\mathrm{poly}(n))$ (see Eq. (5)) rather than the $1 - \exp(-2^{\Omega(n)})$ that our sketch of proof gives. Because event $\{\|c_i\|_2 < n^B \ \ \forall i \in [M]\}$ is independent of Σ_X, it contributes only a single term to the union bound. Consequently, our dense grid is still allowed to be of size $\exp(-2^{o(n)})$.

Rate-penalty. In the overview above, we assumed that if $-\ln p_{\mathrm{success}}(D) = R$, then for any $\eta > 0$ we have $-\ln p_{\mathrm{success}}(D + \eta) = R - \varepsilon_\eta$ for some $\varepsilon_\eta > 0$. This is indeed the case whenever $D'(R)$ is finite. However, our results are for the supremum over all Σ_X with trace n, and this cannot be guaranteed for all such matrices at all rates. Consequently, in Theorem 3 there is both a rate-penalty $\varepsilon > 0$ and a distortion penalty $\eta > 0$ (which both can be made arbitrarily small for n large enough), whereas in the sketch above we only had a distortion penalty.

Disclosure of LLM Assistance

The authors used ChatGPT to assist with editing, code generation for Fig. 1, and technical steps (e.g., perturbation analysis and derivative computations) in the proofs of main theorems. The final optimization step after the spectrum reduction in Theorem 5 was also suggested by ChatGPT. All model-generated code, computations, and proof steps were independently verified and adjusted by the authors. The authors take full responsibility for the correctness of all analytical and numerical results in the paper.

References

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms, 2024. URL https://arxiv.org/abs/2404.00456.

László Babai. On lovász'lattice reduction and the nearest lattice point problem. *Combinatorica*, 6 (1):1–13, 1986.

Hicham Badri and Appu Shaji. Half-quadratic quantization of large machine learning models, November 2023. URL https://mobiusml.github.io/hqq_blog/.

Johann Birnick. The lattice geometry of neural network quantization–a short equivalence proof of gptq and babai's algorithm. *arXiv preprint arXiv:2508.01077*, 2025.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. Quip: 2-bit quantization of large language models with guarantees, 2024. URL https://arxiv.org/abs/2307.13304.

Jiale Chen, Vage Egiazarian, Torsten Hoefler, and Dan Alistarh. Wush: Near-optimal adaptive transforms for llm quantization. *arXiv preprint arXiv:2512.00956*, 2025a.

Jiale Chen, Yalda Shabanzadeh, Elvir Crnčević, Torsten Hoefler, and Dan Alistarh. The geometry of llm quantization: Gptq as babai's nearest plane algorithm. *arXiv preprint arXiv:2507.18553*, 2025b.

John Conway and Neil Sloane. Fast quantizing and decoding and algorithms for lattice quantizers and codes. *IEEE Transactions on Information Theory*, 28(2):227–232, 1982.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35: 30318–30332, 2022.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers, 2023. URL https://arxiv.org/abs/2210.17323.

Allen Gersho and Robert M Gray. *Vector quantization and signal compression*, volume 159. Springer Science & Business Media, 2012.

Egor Lifar, Semyon Savkin, Or Ordentlich, and Yury Polyanskiy. Watersic: information-theoretically (near) optimal linear layer quantization. *arXiv preprint arXiv:2603.04956*, 2026.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. Spinquant: Llm quantization with learned rotations, 2025. URL https://arxiv.org/abs/2405.16406.

Emin Martinian, Gregory W. Wornell, and Ram Zamir. Source coding with distortion side information. *IEEE Transactions on Information Theory*, 54(10):4638–4665, 2008. doi: 10.1109/TIT.2008.928983.

NVIDIA et al. Pretraining large language models with NVFP4. *arXiv preprint arXiv:2509.25149*, 2025.

Open Compute Project. OCP microscaling formats (MX) specification. Technical report, Open Compute Project, 2023. URL https://www.opencompute.org/documents/ocp-microscaling-formats-mx-v1-0-spec-final-pdf.

Or Ordentlich and Yury Polyanskiy. Optimal quantization for matrix multiplication. *IEEE Transactions on Information Theory*, 2025.

Or Ordentlich and Yury Polyanskiy. High-rate quantized matrix multiplication II. *arXiv preprint arXiv:2605.13768*, 2026.

Or Ordentlich, Oded Regev, and Barak Weiss. New bounds on the density of lattice coverings. *Journal of the American Mathematical Society*, 35(1):295–308, 2022.

Jan Ostergaard and Ram Zamir. Incremental refinement using a gaussian test channel. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2233–2237, 2011. doi: 10.1109/ISIT.2011.6033957.

Yury Polyanskiy and Yihong Wu. *Information theory: From coding to learning*. Cambridge university press, 2024.

Semyon Savkin, Eitan Porat, Or Ordentlich, and Yury Polyanskiy. Nestquant: Nested lattice quantization for matrix products and llms, 2025. URL https://arxiv.org/abs/2502.09720.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks, 2024. URL https://arxiv.org/abs/2402.04396.

Albert Tseng, Qingyao Sun, David Hou, and Christopher De Sa. Qtip: Quantization with trellises and incoherence processing, 2025a. URL https://arxiv.org/abs/2406.11235.

Albert Tseng, Zhaofeng Sun, and Christopher De Sa. Model-preserving adaptive rounding, 2025b. URL https://arxiv.org/abs/2505.22988.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models, 2024. URL https://arxiv.org/abs/2211.10438.

Ram Zamir. The rate loss in the wyner-ziv problem. *IEEE Transactions on Information Theory*, 42 (6):2073–2084, 2002.

Ram Zamir and Toby Berger. Multiterminal source coding with high resolution. *IEEE Transactions on Information Theory*, 45(1):106–117, 2002.

Appendix A. Preliminaries and Notation

Notation. We write $\Sigma \succeq 0$ to denote that the matrix $\Sigma \in \mathbb{R}^{n\times n}$ is positive semidefinite; we define the set of positive semidefinite matrices as $\mathbb{S}_+^n = \{\Sigma \in \mathbb{R}^{n\times n} : \Sigma \succeq 0\}$.

For $\Sigma \in \mathbb{S}_+^n$ we denote the spectral decomposition as

$$\mathrm{EVD}(\Sigma) = U\Lambda U^\top,$$

where $U \in \mathcal{O}_n$ is orthogonal and Λ is diagonal.

$\mathrm{diag}(v)$ for $v \in \mathbb{R}^n$ denotes an $n \times n$ diagonal matrix with $\mathrm{diag}(v)_{ii} = v_i$. $\mathrm{spec}(A)$ for $A \in \mathbb{R}^{n\times n}$ denotes the vector of the eigenvalues of A.

For $A \in \mathbb{R}^{n\times a}$ and $\Sigma \in \mathbb{S}_+^n$, denote

$$\|A\|_\Sigma \triangleq \sqrt{\mathrm{tr}(A^\top \Sigma A)}\,.$$

Proposition 6 (Hanson-Wright Concentration Inequality) *For $X \sim \mathcal{N}(0, I_n)$ and $A \in \mathbb{R}^{n\times n}$, for every $t \geq 0$,*

$$\Pr\left[|X^\top AX - \mathrm{tr}\, A| > t\right] \leq 2\exp\left(-c\min\left(\frac{t^2}{K^4\|A\|_F^2}, \frac{t}{K^2\|A\|_{op}}\right)\right),$$

for universal constants c, K.

Appendix B. Random Coding: Worst-Case W

In this section, for fixed $\Sigma_X \in \mathbb{S}_+^n$, we characterize the rate-distortion of quantizing a fixed vector $W \in \mathbb{R}^n$ under a distortion metric d_{Σ_X} using a random coding scheme. In particular, for a given $\Sigma_X = U^\top \Lambda U$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, fixed vector $W \in \mathbb{R}^n$, and constant $R^\star > 0$, we define a generalized distortion function $\mathbf{D}_{\mathrm{rc}}^\lambda(U^\top W, R^\star)$ and demonstrate that it is achievable in Theorem 7.

General Rate-Distortion Function. Let $V, \lambda = (\lambda_1, \ldots, \lambda_n)^\top \in \mathbb{R}^n$ be such that $\lambda_i \geq 0$ for all $i \in [n]$ and $\sum_i \lambda_i = n$. We define *general random-coding rate-distortion function* in dimension n to be the following parametric curve for $T > 0$:

$$D_{\mathrm{rc}}^\lambda(V, T) = \frac{1}{n}\sum_{i=1}^n \frac{V_i^2\lambda_i}{1+\lambda_i T} \qquad \text{and} \qquad R_{\mathrm{rc}}^\lambda(T) = \frac{1}{2n}\sum_{i=1}^n \log(1+\lambda_i T)\,. \tag{RDRC}$$

Throughout, log denotes $\log_2$ and ln denotes the natural logarithm. Denote $T_{\mathrm{rc}}^\lambda(R)$ be a unique value T, s.t. $R_{\mathrm{rc}}^\lambda(T) = R$ (note that $T_{\mathrm{rc}}^\lambda(R)$ is independent of V). Let

$$\mathbf{D}_{\mathrm{rc}}^\lambda(V, R) \triangleq D_{\mathrm{rc}}^\lambda(V, T_{\mathrm{rc}}^\lambda(R))\,. \tag{$\mathbf{D}_{\mathrm{rc}}$}$$

We consider a task of quantizing a given vector $W \in \mathbb{R}^n$ under a distortion function (d_{Σ_X}).

Condition 1 (Admissible W, Σ_X) *We consider $W \in \mathbb{R}^n, \Sigma_X \in \mathbb{S}_+^n$ satisfying*

1. $\mathrm{EVD}(\Sigma_X) = U\Lambda U^\top$ *for $U \in \mathcal{O}_n$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \succeq 0$ with $\mathrm{tr}(\Lambda) = n$;*
2. $\|U^\top W\|_\infty \leq n^\alpha$ *for a known constant α.*

In Theorem 7 we show that there exists an encoder-decoder pair that, for any constant target rate $R^\star > 0$ and fixed admissible W, Σ_X (Cond. 1), achieves rate $R^\star$ and distortion $\mathbf{D}_{\mathrm{rc}}^{U^\top W,\lambda}(R^\star)$ asymptotically, with high probability over the randomness S shared between encoder and decoder (from which a codebook is generated).[4]

Theorem 7 (Achievability of Random-Coding Rate-Distortion: Nonasymptotic Guarantee) *Fix any constants $R^\star, \varepsilon_\circ, \eta_\circ > 0$ and $\alpha > 0$. There exists an encoder $f : \mathbb{R}^n \times \mathbb{S}_+^n \times [0,1] \to [2^{nR}]$, a decoder $g : [2^{nR}] \times [0,1] \to \mathbb{R}^n$ with $R \le R^\star + \varepsilon_\circ$ and a (shared) random variable $S \in [0,1]$ with the following property. For any fixed W, Σ_X (satisfying Cond. 1) we set $\widehat{W} = g(f(W, \Sigma_X, S), S)$. Then for any sufficiently large $n \ge n_0 = n_0(\varepsilon_\circ, \eta_\circ, R^\star, \alpha)$ and any $\beta > \alpha - 1/4$ in case $\alpha \ge 1/4$ and $\beta = 0$ otherwise, we have*

$$\Pr_S\left[\frac{1}{n} d_{\Sigma_X}(W, \widehat{W}) \le \mathbf{D}_{\mathrm{rc}}^{\lambda}(U^\top W, R^\star) + n^{2\beta}\eta_\circ\right] \ge 1 - \exp\left(-2^{n\varepsilon_\circ(1 - cn^{4(\alpha-\beta)-1})}\right),$$

where $d_{\Sigma_X}(W, \widehat{W})$ is the distortion function in Eq. (d_{Σ_X}) and $c = c(\varepsilon_\circ, \eta_\circ, R^\star, \alpha)$.

Remark 8 *In case of $\alpha < 1/4$, the distortion bound above simplifies to $\mathbf{D}_{\mathrm{rc}}^{\lambda}(U^\top W, R^\star) + \eta$ (since $\beta = 0$).*

B.1. Proof of Theorem 7

Quantization Scheme. From the shared randomness S, f, g generate a Gaussian codebook $\mathbf{C} = \{c_1, \ldots, c_{M=2^{n(R^\star+\varepsilon)}}\}$ for $\varepsilon = \varepsilon(\varepsilon_\circ)$ to be chosen later. Denote $\widetilde{W} = U^\top W$. We define:

- **Encoder f:** Let $T := T_{\mathrm{rc}}^{\lambda}(R^\star)$ and define the scaling parameter $\tau = \tau(\widetilde{W}, \lambda)$ as

$$\tau = \left(T \sum_j \frac{\widetilde{W}_j^2 \lambda_j^2}{(1 + \lambda_j T)^2}\right)^{1/2} \left(\sum_j \frac{\lambda_j}{1 + \lambda_j T}\right)^{-1/2}. \qquad (\tau \text{ def.})$$

  We set f to be a tuple

$$f(W, \Sigma_X, S) = \left(\operatorname*{argmin}_{i \in [M]} d_{\Sigma_X}(W, \tau c_i), q(\tau)\right),$$

  where $q(\tau) = \delta \|\widetilde{W}\|_\infty \lfloor \tau / (\delta \|\widetilde{W}\|_\infty) \rceil$ is a rounding quantization scheme of precision δ and recall the distortion function

$$d_{\Sigma_X}(W, C) := (W - C)^\top \Sigma_X (W - C).$$

- **Decoder g:**

$$g(i, q(\tau), S) = q(\tau) \cdot c_i.$$

4. In the regime $\alpha < 1/4$, see the statement of Thm. 7.

Rate-Distortion Bound. In the quantization scheme above,

$$R = \underbrace{R^\star + \varepsilon}_{\text{Gauss. codebook}} + \underbrace{\frac{1}{n}\log(1/\delta)}_{\tau \text{ quant.}} .$$

The rest of the proof is to obtain a high probability bound on the resulting distortion that, given $\widehat{W} = g(f(W, \Sigma_X, S), S) = q(\tau) \cdot c_i$, can be expressed as:

$$d_{\Sigma_X}(W, \widehat{W}) = d_{\Sigma_X}(W, q(\tau)c_i) = (W - q(\tau)c_i)^\top \Sigma_X (W - q(\tau)c_i) .$$

In what follows, denote $D^\star := \mathbf{D}^\lambda_{\rm rc}(\widetilde{W}, R^\star)$.

Before giving the proof we state two helpful claims that bound the effect of quantizing τ in the scheme above. Proofs of Claim 1 and 2 are found in this subsection below.

Claim 1 (Bound on τ.) *The value τ in Eq.(τ def.) satisfies*

$$0 \le \tau \le \|\widetilde{W}\|_\infty \le n^\alpha .$$

Claim 2 (Bound on distortion from τ quantization) *Given $|q(\tau) - \tau| = \delta_\tau \le \delta n^\alpha$, with probability at least $1 - 2^{n(R^\star+\varepsilon)} \exp(-Ct)$ for a universal constant C and any $t > 1/n$,*

$$d_{\Sigma_X}(W, q(\tau)c_i) \le \left(\sqrt{d_{\Sigma_X}(W, \tau c_i)} + \delta n^\alpha \sqrt{n(1+t)} \right)^2 .$$

The main part of the argument is essentially contained in the following Lemma 9. The proof, which also explains the expressions for $D_{\rm rc}$ and $R_{\rm rc}$ is proven in a separate section B.2 due to its importance.

Lemma 9 (Gaussian Book Success) *Let $\varepsilon, \eta > 0$ be constants and $\mathbf{C} = \{c_1, \dots, c_{M=2^{n(R^\star+\varepsilon)}}\}$, $c_i \sim_{i.i.d.} \mathcal{N}(0, I_n)$ be a randomly generated Gaussian codebook.*

For admissible W, Σ_X (Cond. 1) with $\alpha < 1/4$,

$$\Pr_{\mathbf{C}}\left[\frac{1}{n} \min_{i \in [M]} d_{\Sigma_X}(W, \tau(W, \Sigma_X) \cdot c_i) \le \mathbf{D}^\lambda_{\rm rc}(U^\top W, R^\star) + \eta \right] \ge 1 - \exp\left(-2^{n\varepsilon(1 - cn^{4\alpha-1})}\right) ,$$

where $c = c(\varepsilon, \eta, R^\star, \alpha)$ is an explicit constant function, $\mathbf{D}^\lambda_{\rm rc}(U^\top W, R^\star)$ is defined in Eq. ($\mathbf{D}_{\rm rc}$), and $\tau(W, \Sigma_X)$ is defined in Eq. (τ def.).

Proof (of Theorem 7) We apply Lemma 9 to $W' = n^{-\beta} W$ for $\eta = \eta(\eta_\circ)$ to be chosen later and any constant $\beta > \alpha - 1/4$ in case $\alpha \ge 1/4$ and $\beta = 0$ otherwise. We have $\|U^\top W'\|_\infty = n^{-\beta}\|U^\top W\|_\infty \le n^{\alpha-\beta} = o(n^{1/4})$, and therefore,

$$\Pr_{\mathbf{C}}\left[\min_{i \in [M]} d_{\Sigma_X}(W, \tau c_i) \le n\big(\underbrace{\mathbf{D}^\lambda_{\rm rc}(\widetilde{W}, R^\star)}_{D^\star} + n^{2\beta}\eta \big) \right] = \Pr_{\mathbf{C}}\left[\min_{i \in [M]} d_{\Sigma_X}(W', \tau n^{-\beta} c_i) \le n\left(\mathbf{D}^\lambda_{\rm rc}(U^\top W', R^\star) + \eta \right) \right]$$
$$\ge 1 - \exp\left(-2^{n\varepsilon(1 - cn^{4(\alpha-\beta)-1})}\right) .$$

From Claim 1, the simple rounding quantizer $q(\tau) = \delta\|\widetilde{W}\|_\infty \lfloor \tau/(\delta\|\widetilde{W}\|_\infty)\rceil$ achieves $|q(\tau)-\tau| \le \delta n^\alpha$, and therefore, by Claim 2 and a union bound, with probability at least $1-\exp\left(-2^{n\varepsilon(1-cn^{4(\alpha-\beta)-1})}\right) - 2^{n(R^\star+\varepsilon)}\exp(-Ct)$,

$$d_{\Sigma_X}(W, q(\tau)c_i) \le \left(\sqrt{n(D^\star + n^{2\beta}\eta)} + \delta n^\alpha \sqrt{n(1+t)}\right)^2 . \tag{25}$$

It remains to simplify the expression in Eq. (25). Let $A > 0$ be any constant and set

$$\delta =: \min\left\{ \frac{1}{4\sqrt{1+t}} \left(\frac{1}{2\ln 2 \cdot R^\star} + \eta\right)^{-1/2} n^{-A-2\alpha-1}, \frac{1}{\sqrt{2(1+t)}} n^{-(A+2\alpha+1)/2} \right\} . \tag{26}$$

Notice that since $\forall i$, $\frac{\lambda_i}{1+2T\lambda_i} \le \frac{1}{2T}$, we have $D^\star \le \frac{n^{2\alpha}}{2T} \le \frac{n^{2\alpha}}{2\ln 2 \cdot R^\star}$, where the last inequality is derived by $\ln 2 \cdot R^\star = \frac{1}{2n}\sum_i \ln(1+2\lambda_i T) \le T$. Plugging in the δ value in Eq. (26) into Eq. (25), we obtain

$$\begin{aligned} d_{\Sigma_X}(W, q(\tau)c_i) &\le n\left(D^\star + n^{2\beta}\eta + 2\delta\sqrt{D^\star + n^{2\beta}\eta}\cdot n^\alpha\sqrt{1+t} + \delta^2 n^{2\alpha}(1+t)\right) \\ &\le n\left(D^\star + n^{2\beta}\eta + \frac{1}{2}n^{-A-1} + \frac{1}{2}n^{-A-1}\right) = n(D^\star + n^{2\beta}\eta) + n^{-A} . \end{aligned}$$

We denote $c', C_1, C_2, C_3, C_4, C_{12}$ to be explicit constants depending on $\varepsilon, \eta, R^\star, \alpha$ (but not W or Σ_X). Now plug in $t = 2^{n\varepsilon} + C^{-1} n \ln 2 (R^\star + \varepsilon)$. The condition in Eq. (25) holds with probability at least

$$1-\exp\left(-2^{n\varepsilon(1-cn^{4(\alpha-\beta)-1})}\right) - \exp\left(n\ln 2(R^\star+\varepsilon)\right)\exp(-Ct) = 1-\exp\left(-2^{n\varepsilon(1-c'n^{4(\alpha-\beta)-1})}\right) .$$

Finally, the rate of this quantization scheme is $R = R^\star + \varepsilon + \frac{1}{n}\log(1/\delta)$, which we now bound:

$$\begin{aligned} \log(1/\delta) &\le \max\left\{ C_1 \log n + \frac{1}{2}\log(1+t) + 2\alpha\log n, C_2 \log n + \frac{1}{2}\log(1+t) + \alpha \log n \right\} \\ &\le C_{12}\log n + \frac{1}{2}\log(1+t) + 2\alpha\log n . \end{aligned}$$

Plugging in the expression for t, we obtain

$$\log(1+t) \le C_3 + n\varepsilon + C_4 \log n ,$$

and therefore,

$$R \le R^\star + \frac{3}{2}\varepsilon + C \cdot \frac{\log n + \alpha \log n}{n} .$$

For sufficiently large n, the RHS is $\le R^\star + 2\varepsilon$. Moreover, for sufficiently large n, our final distortion bound simplifies to $n(D^\star + n^{2\beta}\cdot 2\eta)$. Setting $\varepsilon = \varepsilon_\circ/2, \eta = \eta_\circ/2$, we conclude the proof. ■

Proof of Claim 1. Since for all $j \in [n]$, $\widetilde{W}_j^2 \le \|\widetilde{W}\|_\infty^2$ and $\frac{\lambda_j}{1+\lambda_j T} \le \frac{1}{T}$,

$$T\sum_j \frac{\widetilde{W}_j^2 \lambda_j^2}{(1+\lambda_j T)^2} \le \|\widetilde{W}\|_\infty^2 \sum_j \frac{\lambda_j}{1+\lambda_j T} ,$$

yielding $0 \le \tau \le \|\widetilde{W}\|_\infty$. □

Proof of Claim 2.

$$\begin{aligned}\sqrt{d_{\Sigma_X}(W, q(\tau)c_i)} &= \sqrt{\mathbb{E}_X\left[\left\|(W - q(\tau)c_i)^T X\right\|_2^2\right]} \\ &\le \sqrt{d_{\Sigma_X}(W, \tau c_i)} + \sqrt{\mathbb{E}_X\left[\left\|(\delta_\tau \cdot c_i)^T X\right\|_2^2\right]} \\ &= \sqrt{d_{\Sigma_X}(W, \tau c_i)} + \delta_\tau \sqrt{c_i^T \Sigma_X c_i}\,,\end{aligned}$$

where $\delta_\tau = |q(\tau) - \tau| \le \delta n^\alpha$. By Hanson-Wright inequality (Prop. 6) for any codeword j, the second term can be bounded as

$$\Pr\left[|c_j^T \Sigma_X c_j - \operatorname{tr}\Sigma_X| > nt\right] \le 2\exp\left(-c\min\left(\frac{t^2 n^2}{K^4\|\Sigma_X\|_F^2}, \frac{tn}{K^2\|\Sigma_X\|_{op}}\right)\right),$$

for universal constants c, K, and therefore, with probability at least $1 - \exp(-Ctn/\|\Sigma_X\|_{op}) \ge 1 - \exp(-Ct)$,

$$\delta_\tau \sqrt{c_j^T \Sigma_X c_j} \le \delta n^\alpha \sqrt{n(1+t)}\,.$$

The statement of the Claim follows by a union bound over all $2^{n(R^\star+\varepsilon)}$ codewords. □

B.2. Proof of Lemma 9: Success Probability of Random Gaussianå Code

Condition 1 (Admissible W, Σ_X) *We consider $W \in \mathbb{R}^n, \Sigma_X \in \mathbb{S}_+^n$ satisfying*

1. $\operatorname{SVD}(\Sigma_X) = U\Lambda U^\top$ *for $U \in \mathcal{O}_n$ and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \succeq 0$ with $\operatorname{tr}(\Lambda) = n$;*
2. $\|U^\top W\|_\infty \le n^\alpha$ *for a known $\alpha > 0$.*

For $W, C \in \mathbb{R}^n$, $\tau \in \mathbb{R}$, and $\Sigma \in \mathbb{S}_+^n$, recall that we define the distortion function to be

$$d_{\Sigma_X}(W, \tau C) = (W - \tau C)^\top \Sigma_X (W - \tau C)\,. \qquad (E \text{ def.})$$

For admissible (W, Σ_X) (Cond. 1), denote $T = T_{\mathrm{rc}}^\lambda(R^\star)$ to be a unique solution to $R_{\mathrm{rc}}^\lambda(T) = R^\star$ and define

$$\tau(W, \Sigma_X) = \left(T\sum_j \frac{(U^\top W)_j^2 \lambda_j^2}{(1+\lambda_j T)^2}\right)^{1/2} \left(\sum_j \frac{\lambda_j}{1+\lambda_j T}\right)^{-1/2}. \qquad (\tau \text{ def.})$$

Lemma 9 (Gaussian Book Success) *Let $\varepsilon, \eta > 0$ be constants and $\mathbf{C} = \{c_1, \ldots, c_{M=2^{n(R^\star+\varepsilon)}}\}$, $c_i \sim_{i.i.d.} \mathcal{N}(0, I_n)$ be a randomly generated Gaussian codebook.*

For admissible W, Σ_X (Cond. 1) with $\alpha < 1/4$,

$$\Pr_{\mathbf{C}}\left[\frac{1}{n}\min_{i\in[M]} d_{\Sigma_X}(W, \tau(W,\Sigma_X)\cdot c_i) \le \mathbf{D}_{\mathrm{rc}}^\lambda(U^\top W, R^\star) + \eta\right] \ge 1 - \exp\left(-2^{n\varepsilon(1-cn^{4\alpha-1})}\right),$$

where $c = c(\varepsilon, \eta, R^\star, \alpha)$ is an explicit constant function, $\mathbf{D}_{\mathrm{rc}}^\lambda(U^\top W, R^\star)$ is defined in Eq. ($\mathbf{D}_{\mathrm{rc}}$), and $\tau(W, \Sigma_X)$ is defined in Eq. (τ def.).

The proof of Lemma 9 relies on the following bound on a probability that a single codeword $c_i \sim \mathcal{N}(0, I_n)$ achieves small distortion d_{Σ_X}.

Lemma 10 *In the setting of Lemma 9, denote*

$$p_n \coloneqq \Pr_{c_i \sim \mathcal{N}(0,I_n)} \left[\frac{1}{n} d_{\Sigma_X}(W, \tau(W, \Sigma_X) \cdot c_i) \le \mathbf{D}^{\lambda}_{\mathrm{rc}}(U^\top W, R^\star) + \eta \right] .$$

Then,

$$\ln p_n \ge -nR^\star(1 + O(n^{4\alpha-1})) = -nR^\star(1 + o(1)) .$$

Proof (of Lemma 9) Denote $D^\star = \mathbf{D}^{\lambda}_{\mathrm{rc}}(U^\top W, R^\star)$, and $\tau = \tau(W, \Sigma_X)$. Given the EVD $(\Sigma_X) = U \Lambda U^\top$, by definition of d_{Σ_X}, for all $i \in [M]$,

$$\begin{aligned} d_{\Sigma_X}(W, \tau \cdot c_i) &= (W - \tau c_i)^\top U \Lambda U^\top (W - \tau c_i) \\ &= (U^\top W - \tau U^\top c_i)^\top \Lambda (U^\top W - \tau U^\top c_i) \\ &=: (\widetilde{W} - \tau \tilde{c}_i)^\top \Lambda (\widetilde{W} - \tau \tilde{c}_i) , \end{aligned}$$

where we denote $\widetilde{W} = U^\top W$ and $\tilde{c}_i = U^\top c_i$. Note that $\tilde{c}_i \sim \mathcal{N}(0, I_n)$ are independent.

We show that for the optimal choice of τ in Eq. (τ def.) and $M = 2^{n(R^\star + \varepsilon)}$, with probability at least $1 - \exp\left(-2^{n\varepsilon(1 - c\|\widetilde{W}\|_\infty^4/n)}\right)$ (where $\|\widetilde{W}\|_\infty^4/n \le n^{4\alpha-1}$),

$$\frac{1}{n} \min_{i \in [M]} d_{\Sigma_X}(W, \tau c_i) = \frac{1}{n} \min_{i \in [M]} \sum_j \lambda_j \left(\widetilde{W}_j - \tau \tilde{c}_{ij} \right)^2 \le D^\star + \eta .$$

By Lemma 10, for any $i \in [M]$,

$$\ln p_n = \ln \Pr \left[\frac{1}{n} \sum_j \lambda_j \left(\widetilde{W}_j - \tau \tilde{c}_{ij} \right)^2 \le D^\star + \eta \right] \ge -nR^\star(1 + O(n^{4\alpha-1})) .$$

Then, since $c_1, \ldots, c_M$ are independent,

$$\begin{aligned} \Pr \left[\frac{1}{n} \min_{i \in M} d_{\Sigma_X}(W, \tau c_i) \le D^\star + \eta \right] \ge 1 - (1 - p_n)^M &\ge 1 - \exp\left(-2^{-nR^\star(1 + O(\|\widetilde{W}\|^4/n))} 2^{nR^\star + n\varepsilon} \right) \\ &= 1 - \exp\left(-2^{n\varepsilon(1 - O(\|\widetilde{W}\|^4/n))} \right) \\ &\ge 1 - \exp\left(-2^{n\varepsilon(1 - O(n^{4\alpha-1}))} \right) , \end{aligned}$$

which concludes the proof of Lemma 9. ■

Proof (of Lemma 10) Denote $R_{\mathrm{nat}} = R^\star \ln 2$, $D^\star = \mathbf{D}^{\lambda}_{\mathrm{rc}}(U^\top W, R^\star)$, and $\tau = \tau(W, \Sigma_X)$. First, for all $j \in [n]$ denote $\mu_j = \mathcal{L}\left(\lambda_j \left(\widetilde{W}_j - \tau \tilde{c}_{ij} \right)^2 \right)$, where $\tilde{c}_{ij} \sim \mathcal{N}(0, 1)$ and rewrite

$$p_n = \Pr_{X_j \sim \mu_j} \left[\frac{1}{n} \sum_j X_j \le D^\star + \eta \right] .$$

Let $t = -T/(2\tau^2)$, where T, τ are defined in Eq. (τ def.). We have

$$\ln \mathbb{E}_{X\sim\mu_j}\left[e^{tX}\right] = \frac{t\widetilde{W}_j^2\lambda_j}{1-2t\lambda_j\tau^2} - \frac{1}{2}\ln\left(1-2t\lambda_j\tau^2\right) =: \Phi_j(t)\,.$$

A direct computation yields

$$\Phi_j'(t) = \frac{\widetilde{W}_j^2\lambda_j}{(1-2t\lambda_j\tau^2)^2} + \frac{\lambda_j\tau^2}{1-2t\lambda_j\tau^2}\,. \tag{27}$$

For each $j \in [m]$, define a new probability measure $\tilde{\mu}_j$ as $\frac{d\tilde{\mu}_j}{d\mu_j}(x) = e^{tx-\Phi_j(t)}$. We have $\int_{\mathbb{R}} d\tilde{\mu}_j = e^{-\Phi_j(t)}\int_{\mathbb{R}} e^{tx}d\mu_j = 1$. Since the exponential tilting above is applied to a noncentral χ^2 distribution μ, the transformation simply shifts/rescales the parameters of the χ^2 distribution. We can explicitly compute

$$\tilde{\mu}_j = \mathcal{L}\left(\mathcal{N}\left(m_j, s_j^2\right)^2\right)\,, \qquad \text{where } m_j = \frac{\widetilde{W}_j\lambda_j^{1/2}}{1-2t\lambda_j\tau^2} \text{ and } s_j^2 = \frac{\lambda_j\tau^2}{1-2t\lambda_j\tau^2}\,.$$

Plugging in the choice of τ in Eq. (τ def.) and $T = -2t\tau^2$ into Eq. (27), we obtain

$$\begin{aligned}\mathbb{E}_{X\sim\tilde{\mu}}\sum_j X_j &= \sum_j \int x e^{tx-\Phi_j(t)}d\mu_j(x) = \sum_j \frac{d}{dt}\ln\int e^{tx}d\mu_j(x) = \sum_j \Phi_j'(t)\\ &= \sum_j \frac{\widetilde{W}_j^2\lambda_j}{1-2t\lambda_j\tau^2} = \sum_j \frac{\widetilde{W}_j^2\lambda_j}{1+\lambda_j T} = nD^\star\,.\end{aligned}$$

Then,

$$D(\tilde{\mu}\|\mu) = \sum_j D(\tilde{\mu}_j\|\mu_j) = \sum_j t\Phi_j'(t) - \Phi_j = \frac{1}{2}\sum_j \ln(1-2t\lambda_j\tau^2) = \frac{1}{2}\sum_j \ln(1+\lambda_j T) = nR_{\text{nat}}\,.$$

Recall that we defined $p_n = \Pr_{X_j\sim\mu_j}\left[\sum_j X_j \le n\left(D^\star+\eta\right)\right]$ and let $\tilde{p}_n = \Pr_{X_j\sim\tilde{\mu}_j}\left[\sum_j X_j \le n\left(D^\star+\eta\right)\right]$. By DPI,

$$nR_{\text{nat}} = D(\tilde{\mu}\|\mu) \ge d\left(\tilde{p}_n\|p_n\right) \ge -h\left(\tilde{p}_n\right) + \tilde{p}_n\ln\frac{1}{p_n} \ge -\ln 2 + \tilde{p}_n\ln\frac{1}{p_n}\,,$$

which yields

$$\ln p_n \ge \frac{-nR_{\text{nat}} - \ln 2}{\tilde{p}_n}\,. \tag{28}$$

By Chebyshev's inequality,

$$1-\tilde{p}_n = 1 - \Pr_{X_j\sim\tilde{\mu}_j}\left[\sum_j X_j \le \mathbb{E}\sum_j X_j + n\eta\right] \le \frac{\text{Var}\left[\sum_j X_j\right]}{n^2\eta^2}\,.$$

In what follows we show that

$$\operatorname{Var}\left[\sum_j X_j\right] = O\left(n\|\widetilde{W}\|_\infty^4\right) = o(n^2)\,, \tag{Var}$$

yielding $1-\tilde{p}_n = O\left(\|\widetilde{W}\|_\infty^4/n\right) = o(1)$. Plugging this into Eq. (28), we get

$$\ln p_n \geq \frac{-nR_{\text{nat}} - \ln 2}{1 - O\left(\|\widetilde{W}\|_\infty^4/n\right)} = -nR_{\text{nat}}(1 + O\left(\|\widetilde{W}\|_\infty^4/n\right))$$

for sufficiently large n.

It remains to obtain the bound in Eq. (Var). From the definition of R_{nat} and $\frac{x}{1+x} \leq \ln(1+x) \leq x$ for $x > -1$,

$$|t|\tau^2 \sum_j \frac{\lambda_j}{1-2t\lambda_j\tau^2} \leq nR_{\text{nat}} = \frac{1}{2}\sum_j \ln(1-2t\lambda_j\tau^2) \leq |t|\tau^2 n\,.$$

Using $\frac{\lambda_j}{1-2t\lambda_j\tau^2} \leq \frac{1}{2|t|\tau^2}$ and $(1-2t\lambda_j\tau^2)^2 \geq 4|t|\lambda_j\tau^2$, we obtain

$$\begin{aligned}
\operatorname{Var}[X_j] = \Lambda_j''(t) &= \frac{4\widetilde{W}_j^2\lambda_j^2\tau^2}{(1-2t\lambda_j\tau^2)^3} + \frac{2\lambda_j^2\tau^4}{(1-2t\lambda_j\tau^2)^2} \\
&\leq \frac{4\widetilde{W}_j^2\lambda_j^2\tau^2}{(1-2t\lambda_j\tau^2)\cdot 4|t|\lambda_j\tau^2} + \frac{\tau^2}{|t|}\cdot\frac{\lambda_j}{1-2t\lambda_j\tau^2} \\
&= \frac{1}{|t|}\cdot\frac{\widetilde{W}_j^2\lambda_j}{1-2t\lambda_j\tau^2} + \frac{\tau^2}{|t|}\cdot\frac{\lambda_j}{1-2t\lambda_j\tau^2}\,,
\end{aligned}$$

and therefore,

$$\operatorname{Var}\left[\sum_j X_j\right] \leq n\cdot\left(\frac{D^\star}{|t|} + \frac{R_{\text{nat}}}{|t|^2}\right)\,. \tag{29}$$

We show a lower bound on parameter $|t|$. From the way we choose τ in Eq. (τ def.),

$$|t| = \frac{\sum_j \frac{\lambda_j}{1-2t\lambda_j\tau^2}}{\sum_j \frac{2\widetilde{W}_j^2\lambda_j^2}{(1-2t\lambda_j\tau^2)^2}} \geq \frac{2R_{\text{nat}}}{\|\widetilde{W}\|_\infty^2}\frac{\sum_j \frac{\lambda_j\|\widetilde{W}\|_\infty^2}{1-2t\lambda_j\tau^2}}{\sum_j \frac{2\widetilde{W}_j^2\lambda_j}{1-2t\lambda_j\tau^2}} \geq \frac{R_{\text{nat}}}{\|\widetilde{W}\|_\infty^2}\,, \tag{30}$$

where we used $\frac{\lambda_j}{1-2t\lambda_j\tau^2} \leq \frac{1}{2|t|\tau^2} \leq \frac{1}{2R_{\text{nat}}}$. Plugging (30) into the bound for variance in (29) and using $D^\star \leq \frac{\|\widetilde{W}\|_\infty^2}{T} \leq \frac{\|\widetilde{W}\|_\infty^2}{2R_{\text{nat}}}$, we obtain

$$\operatorname{Var}\left[\sum_j X_j\right] \leq n\cdot\|\widetilde{W}\|_\infty^4\left(\frac{1}{2(R_{\text{nat}})^2} + \frac{1}{R_{\text{nat}}}\right) = O\left(n\|\widetilde{W}\|_\infty^4\right)\,,$$

which is the desired inequality in Eq. (Var). Here we used that $R_{\text{nat}} = \Theta(1)$. □

Appendix C. Random Coding: Gaussian Isotropic W

In this section we describe quantization of a Gaussian isotropic vector $W \sim \mathcal{N}(0, I_n) \in \mathbb{R}^n$ using random coding.

Rate-Distortion Function. Let $\lambda = (\lambda_1, \ldots, \lambda_n)^\top \in \mathbb{R}^n$ be such that $\lambda_i \geq 0$ for all $i \in [n]$ and $\sum_i \lambda_i = n$. We define *random-coding rate-distortion function* in dimension n to be the following parametric curve for $T > 0$:

$$D_{\mathrm{rc}}^{\lambda}(T) = \frac{1}{n}\sum_{i=1}^{n} \frac{\lambda_i}{1+\lambda_i T} \qquad \text{and} \qquad R_{\mathrm{rc}}^{\lambda}(T) = \frac{1}{2n}\sum_{i=1}^{n} \log(1+\lambda_i T)\,. \tag{RDRC}$$

Throughout, $\log$ denotes $\log_2$ and $\ln$ denotes the natural logarithm. Denote $T_{\mathrm{rc}}^{\lambda}(R)$ to be a unique value T, s.t. $R_{\mathrm{rc}}^{\lambda}(T) = R$. Let

$$\mathbf{D}_{\mathrm{rc}}^{\lambda}(R) \triangleq D_{\mathrm{rc}}^{\lambda}(T_{\mathrm{rc}}^{\lambda}(R))\,. \tag{$\mathbf{D}_{\mathrm{rc}}$}$$

We consider a task of quantizing a vector $W \sim \mathcal{N}(0, I_n) \in \mathbb{R}^n$ under a distortion function

$$d_{\Sigma_X}(W, \widehat{W}) = \mathbb{E}_X\left[\left\| \left(W - \hat{W}\right)^T X\right\|_2^2\right] = (W-\widehat{W})^\top \Sigma_X (W-\widehat{W})\,, \tag{d_{Σ_X}}$$

where $X \in \mathbb{R}^n$ is a random vector with $\mathbb{E}XX^\top = \Sigma_X \in \mathbb{S}_+^n$. Here we assume $\mathrm{EVD}\,(\Sigma_X) = U\Lambda U^\top$ for $U \in \mathcal{O}_n$ and $\Lambda = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_n) \succeq 0$ with $\mathrm{tr}(\Lambda) = n$. Our goal is to obtain an upper bound on $\mathbb{E}_W\left[d_{\Sigma_X}(W, \widehat{W})\right]$.

In Theorem 11 we show that there exists an encoder-decoder that, for any constant target rate $R^\star > 0$ and admissible Σ_X as above, achieves rate $R^\star$ and expected distortion $\mathbf{D}_{\mathrm{rc}}^{\lambda}(R^\star)$ asymptotically, with high probability over the randomness S shared between encoder and decoder (from which the codebook is generated).

Theorem 11 (Achievability of RC RD for Gaussian Input: Nonasymptotic Guarantee) *Fix any constants $R^\star, \varepsilon_\circ, \eta_\circ, B > 0$. There exists an encoder $f : \mathbb{R}^n \times \mathbb{S}_+^n \times [0,1] \to \left[2^{nR}\right]$, a decoder $g : \left[2^{nR}\right] \times [0,1] \to \mathbb{R}^n$ with $R \leq R^\star + \varepsilon_\circ$, and a (shared) random variable $S \in [0,1]$ with the following property.*

For $W \sim \mathcal{N}(0, I_n)$, any Σ_X (with $\mathrm{EVD}(\Sigma_X) = U\Lambda U^\top$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$) as above we set $\widehat{W} = g(f(W, \Sigma_X, S), S)$. Then for any sufficiently large $n \geq n_0 = n_0(\varepsilon_\circ, \eta_\circ, R^\star, B)$ we have

$$\Pr_S\left[\mathbb{E}_W\left[\frac{1}{n} d_{\Sigma_X}(W, \widehat{W})\right] \leq \mathbf{D}_{\mathrm{rc}}^{\lambda}(R^\star) + \eta_\circ\right] \geq 1 - \exp\left(-n^B\right)\,,$$

where $d_{\Sigma_X}(W, \widehat{W})$ is the distortion function in Eq. (d_{Σ_X}).

In Theorem 3 we show that, in fact, the quantization scheme in Theorem 11 achieves the rate-distortion guarantee simultaneously for all $\Sigma_X \in \mathbb{S}_+^n$ with $\mathrm{tr}(\Sigma_X) = n$, with high probability over the shared randomness (i.e., the codebook $\mathbf{C}$)

Theorem 3 (Achievability of RC RD for Gaussian Input: Worst-Case Σ_X) *Fix any constants $R^\star, \varepsilon_\circ, \eta_\circ, B > 0$. There exist an encoder $f : \mathbb{R}^n \times \mathbb{S}_+^n \times [0,1] \to \left[2^{nR}\right]$, a decoder $g : \left[2^{nR}\right] \times$*

$[0,1] \to \mathbb{R}^n$ *with* $R \le R^\star + \varepsilon_\circ$*, and a (shared) random variable* $S \in [0,1]$ *with the following property.*

For $W \sim \mathcal{N}(0, I_n)$ *and* $\Sigma_X \in \mathbb{S}_+^n$ *we set* $\widehat{W}(\Sigma_X) = g\left(f(W, \Sigma_X, S), S\right)$*. Then, for sufficiently large* $n \ge n_0 = n_0(\varepsilon_\circ, \eta_\circ, R^\star, B)$*, we have*

$$\Pr_S\left[\sup_{\substack{\Sigma_X \in \mathbb{S}_+^n \\ \mathrm{tr}(\Sigma_X)=n}} \left(\frac{1}{n}\mathbb{E}_W\left[d_{\Sigma_X}(W, \widehat{W}(\Sigma_X))\right] - \mathbf{D}_{\mathrm{rc}}^{\mathrm{spec}(\Sigma_X)}(R^\star)\right) \le \eta_\circ\right] \ge 1 - \exp\left(-n^B\right),$$

where $d_{\Sigma_X}(W, \widehat{W})$ *is the distortion function in Eq. (*d_{Σ_X}*).*

C.1. Proof of Theorem 11

The proof uses the result of Theorem 7, which obtains a distortion guarantee for any fixed vector $W \in \mathbb{R}^n$. We adjust the quantization scheme to not rely on the Theorem 7 assumption $\|U^\top W\|_\infty \le n^\alpha$ (Cond. 1): in the unlikely case of large $\|U^\top W\|_\infty$, the decoder returns 0.

Quantization Scheme. Fix any constant $\alpha \in (0, 1/4)$. From the shared randomness S, f, g generate a Gaussian codebook $\mathbf{C} = \{c_1, \dots, c_{M=2^{n(R^\star+\varepsilon)}}\}$ for $\varepsilon = \varepsilon(\varepsilon_\circ)$ to be chosen later. Denote $\widetilde{W} = U^\top W$. We define:

- **Encoder** f**:** Let $T \coloneqq T_{\mathrm{rc}}^{\lambda}(R^\star)$ and define the scaling parameter $\tau = \tau(\widetilde{W}, \lambda)$ as

$$\tau = \begin{cases} \left(T\sum_j \frac{\widetilde{W}_j^2 \lambda_j^2}{(1+\lambda_j T)^2}\right)^{1/2} \left(\sum_j \frac{\lambda_j}{1+\lambda_j T}\right)^{-1/2} & \text{if } \|\widetilde{W}\|_\infty \le n^\alpha \\ 0 & \text{otherwise.} \end{cases} \qquad (\tau \text{ def.})$$

  We set f to be a tuple

$$f(W, \Sigma_X, S) = \left(\operatorname*{argmin}_{i \in [M]} d_{\Sigma_X}(W, \tau c_i), q(\tau)\right),$$

  where $q(\tau) = \delta\|\widetilde{W}\|_\infty \lfloor \tau/(\delta\|\widetilde{W}\|_\infty)\rceil$ is a rounding quantization scheme of precision δ and recall the distortion function

$$d_{\Sigma_X}(W, C) = (W - C)^\top \Sigma_X (W - C).$$

- **Decoder** g**:**

$$g(i, q(\tau), S) = q(\tau) \cdot c_i.$$

Rate-Distortion Bound. Denote In the quantization scheme above,

$$R = \underbrace{R^\star + \varepsilon}_{\text{Gauss. codebook}} + \underbrace{\frac{1}{n}\log(1/\delta)}_{\tau \text{ quant.}}.$$

The rest of the proof is to obtain a high probability bound on the resulting distortion that, given $\widehat{W} = g(f(W, \Sigma_X, S), S) = q(\tau) \cdot c_i$, can be expressed as:

$$d_{\Sigma_X}(W, \widehat{W}) = d_{\Sigma_X}(W, q(\tau)c_i) = (W - q(\tau)c_i)^\top \Sigma_X (W - q(\tau)c_i)\,.$$

In what follows, denote $D^\star = \mathbf{D}^\lambda_{\rm rc}(R^\star)$. Let $\eta = \eta(\eta_\circ)$ be a constant to be chosen later. By Theorem 7, with an appropriate choice of parameter $\delta = \delta(\varepsilon, \eta, R^\star, \alpha, n)$ in the quantization scheme above, we have for any fixed W such that $\|U^\top W\|_\infty \le n^\alpha$ and sufficiently large $n \ge n_0 = n_0(\varepsilon, \eta, \alpha, R^\star)$,

$$\Pr_{\mathbf{C}}\left[\frac{1}{n} d_{\Sigma_X}(W, \widehat{W}) \le \mathbf{D}^\lambda_{\rm rc}(U^\top W, R^\star) + \eta\right] \ge 1 - \exp\left(-2^{n\varepsilon(1-cn^{4\alpha-1})}\right)\,,$$

where $c = c(\varepsilon, \eta, \alpha, R^\star)$. At the same time, the rate can be bounded as $R \le R^\star + \varepsilon_\circ$ (for an appropriately chosen $\varepsilon = \varepsilon(\varepsilon_\circ)$). Denote the event of the codebook failure as $F(W, \mathbf{C}) : \mathbb{R}^n \times (\mathbb{R}^n)^M \to \{0, 1\}$:

$$F(W, \mathbf{C}) = \left\{\|U^\top W\|_\infty \le n^\alpha \text{ and } \frac{1}{n} d_{\Sigma_X}(W, \widehat{W}) > \mathbf{D}^\lambda_{\rm rc}(U^\top W, R^\star) + \eta\right\}\,. \qquad (F \text{ def.})$$

Rewriting the result of Theorem 7, we obtain that

$$\forall W \in \mathbb{R}^n :\ \|U^\top W\|_\infty \le n^\alpha : \qquad \Pr_{\mathbf{C}}\left[F(W, \mathbf{C})\right] \le \exp\left(-2^{n\varepsilon(1-cn^{4\alpha-1})}\right)\,.$$

Denote $E_{\mathbf{C}}$ to be the event $E_{\mathbf{C}} = \left\{\forall i \in [M] :\ \|c_i\|_2 \le n^B\right\}$ for some fixed constant $B > 10$ and note that

$$\Pr_{\mathbf{C}}\left[E^c_{\mathbf{C}}\right] \le M \cdot e^{-Cn^{2B}} = e^{n \ln 2 (R^\star + \varepsilon) - Cn^{2B}} \le \exp\left(-c' n^{2B}\right)\,,$$

for constant $c' = c'(\varepsilon, R^\star)$ and sufficiently large (constant) n. We denote $\mathbf{C}|E_{\mathbf{C}}$ to be the distribution of $\mathbf{C}$ conditioned on $E_{\mathbf{C}}$ and note that, since $\Pr_{\mathbf{C}}\left[E_{\mathbf{C}}\right] \ge 1/2$, Theorem 7 in fact yields

$$\forall W \in \mathbb{R}^n :\ \|U^\top W\|_\infty \le n^\alpha : \qquad \Pr_{\mathbf{C}|E_{\mathbf{C}}}\left[F(W, \mathbf{C})\right] \le 2\exp\left(-2^{n\varepsilon(1-cn^{4\alpha-1})}\right)\,.$$

We show in steps 1-2 that for any constant $A > 0$ and sufficiently large n,

$$\Pr_{\mathbf{C}|E_{\mathbf{C}}}\left[\mathbb{E}_W\left[d_{\Sigma_X}(W, \widehat{W})\right] \le n(\mathbf{D}^\lambda_{\rm rc}(R^\star) + \eta) + n^{-A}\right] \ge 1 - \exp\left(-2^{n\varepsilon(1-c'n^{4\alpha-1})}\right)\,,$$

and then conclude the proof in step 3 via a union bound.

Step 1: Bound on $\Pr_{\mathbf{C}|E_{\mathbf{C}}}\left[\Pr_{W\sim\mathcal{N}(0,I_n)}\left[F(W, \mathbf{C})|\mathbf{C}\right] > p^\star\right]$. Denote the event $E_W = \left\{\|U^\top W\|_\infty \le n^\alpha\right\}$. All expressions below assume $W \sim \mathcal{N}(0, I_n)$.

$$\mathbb{E}_{\mathbf{C}|E_{\mathbf{C}}}\left[\Pr_W\left[F(W, \mathbf{C})|\mathbf{C}\right]\right] = \Pr_{W, \mathbf{C}|E_{\mathbf{C}}}\left[F(W, \mathbf{C})\right] = \Pr_W\left[\Pr_{\mathbf{C}|E_{\mathbf{C}}}\left[F(W, \mathbf{C})|W\right]\right] \le \Pr_W\left[E_W\right] \cdot 2\exp\left(-2^{n\varepsilon(1-cn^{4\alpha-1})}\right)\,.$$

Then, by Markov's inequality,

$$\Pr_{\mathbf{C}|E_{\mathbf{C}}}\left[\Pr_W\left[F(W, \mathbf{C})|\mathbf{C}\right] > p^\star\right] \le \frac{\Pr_W\left[E_W\right] \cdot 2\exp\left(-2^{n\varepsilon(1-cn^{4\alpha-1})}\right)}{p^\star}\,. \qquad (31)$$

Step 2: Bound on $\mathbb{E}_W\left[d_{\Sigma_X}(W,\widehat{W})\right]$ in case of codebook success in Eq. (31). Fix any codebook $\mathbf{C}$ for which $\Pr_W\left[F(W,\mathbf{C})|\mathbf{C}\right] \le p^\star$ and $E_{\mathbf{C}}$ hold. Expand

$$\mathbb{E}_W\left[d_{\Sigma_X}(W,\widehat{W})\right] = \underbrace{\mathbb{E}_W\left[d_{\Sigma_X}(W,\widehat{W})\mathbb{1}_{E_W}\right]}_{\mathrm{I}} + \underbrace{\mathbb{E}_W\left[d_{\Sigma_X}(W,\widehat{W})\mathbb{1}_{E_W^c}\right]}_{\mathrm{II}}.$$

First,

$$\begin{aligned}E_W\left[d_{\Sigma_X}(W,\widehat{W})\mathbb{1}_{E_W}\right] \le & p^\star \mathbb{E}_W\left[d_{\Sigma_X}(W,\widehat{W})\mathbb{1}_{E_W}|F(W,\mathbf{C})\right] + \\ & \Pr_W\left[F(W,\mathbf{C})^c\right]\mathbb{E}_W\left[d_{\Sigma_X}(W,\widehat{W})\mathbb{1}_{E_W}|F(W,\mathbf{C})^c\right].\end{aligned}$$

For any W, we have, since $q(\tau) \le n^\alpha$ by Claim 1,

$$d_{\Sigma_X}(W,\widehat{W}) = (W - q(\tau)c_i)^\top \Sigma_X (W - q(\tau)c_i) \le (\|U^\top W\|_\infty + n^\alpha \|U^\top c_i\|_\infty)^2 n\,,$$

and therefore, since the above is $\le n^{3B}$ for $W \in E_W$ and $\mathbf{C} \in E_{\mathbf{C}}$,

$$p^\star E_W\left[d_{\Sigma_X}(W,\widehat{W})\mathbb{1}_{E_W}|F(W,\mathbf{C})\right] \le p^\star \cdot n^{3B}\,. \tag{I.1}$$

Moreover, by the definition of $F(W,\mathbf{C})$ in Eq. (F def.),

$$\begin{aligned}\Pr_W\left[F(W,\mathbf{C})^c\right]\mathbb{E}_W\left[d_{\Sigma_X}(W,\widehat{W})\mathbb{1}_{E_W}|F(W,\mathbf{C})^c\right] &\le n\mathbb{E}_W\left[\mathbf{D}^\lambda_{\mathrm{rc}}(U^\top W, R^\star)\cdot \mathbb{1}_{F(W,\mathbf{C})^c}\mathbb{1}_{E_W}\right] + n\eta \\ &\le \mathbb{E}_W\left[\sum_i \frac{(U^\top W)_i^2 \lambda_i}{1+T\lambda_i}\mathbb{1}_{E_W}\right] + n\eta \\ &\le \mathbb{E}_W\left[\sum_i \frac{(U^\top W)_i^2 \lambda_i}{1+T\lambda_i}\right] + n\eta \\ &= n(D^\star + \eta)\,. \end{aligned} \tag{I.2}$$

In the above we used that $R^\lambda_{\mathrm{rc}}(T)$ is independent of vector $U^\top W$, and therefore both $\mathbf{D}^\lambda_{\mathrm{rc}}(U^\top W, R^\star)$ and $\mathbf{D}^\lambda_{\mathrm{rc}}(R^\star)$ share the same parameter $T > 0$. Combining (I.1) and (I.2), we obtain

$$\mathrm{I} \le n(D^\star + \eta) + p^\star \cdot n^{3B}\,. \tag{I}$$

Now in the case of E_W^c, we have $\tau = 0$, and therefore, $d_{\Sigma_X}(W,\widehat{W}) = W^\top \Sigma_X W \le n\|U^\top W\|_\infty^2$. Note that $U^\top W \sim \mathcal{N}(0, I_n)$, so by the standard Gaussian tail bound,

$$\mathrm{II} \le n\mathbb{E}_W\left[\|U^\top W\|_\infty^2 \cdot \mathbb{1}_{\|U^\top W\|_\infty > n^\alpha}\right] = n\int_{t=n^\alpha}^\infty t^2 \cdot 2ne^{-t^2/2} \le 4n^{2+\alpha}e^{-n^{2\alpha}/2}\,. \tag{II}$$

Combining (I) and (II) together, we obtain

$$\mathbb{E}_W\left[d_{\Sigma_X}(W,\widehat{W})\right] \le n(D^\star+\eta) + p^\star \cdot n^{3B} + 4n^{2+\alpha}e^{-n^{2\alpha}/2} \le n(D^\star+\eta) + n^{-A}\,,$$

for sufficiently large constant n and the choice $p^\star = e^{-n}$.

Step 3: Bound on the probability of successful codebook C. Recall that $\Pr_{\mathbf{C}}[E_{\mathbf{C}}^c] \le \exp\left(-c' n^{2B}\right)$. From Eq. (31) and the choice of $p^\star$ above, we have for sufficiently large constant n,

$$\Pr_{\mathbf{C}|E_{\mathbf{C}}}\left[\Pr_W\left[F(W,\mathbf{C})|\mathbf{C}\right] > p^\star\right] \le \exp\left(-2^{n\varepsilon(1-cn^{4\alpha-1})} + n\right) \le \exp\left(-c' n^{2B}\right) .$$

Then, by a union bound, the probability to select a good codebook is at least $1 - \exp\left(-c'' n^{2B}\right)$, so we conclude (since for sufficiently large n, $n^{-A} < n\eta$),

$$\Pr_{\mathbf{C}}\left[\mathbb{E}_W\left[\frac{1}{n} d_{\Sigma_X}(W,\widehat{W})\right] \le D^\star + 2\eta\right] \ge 1 - \exp\left(-c'' n^{2B}\right) ,$$

from which the statement of the theorem follows.

C.2. Proof of Theorem 3

We show Theorem 3 by applying the results of Theorem 11 for fixed Σ_X with a covering argument. The quantization scheme is the same as in the proof of Theorem 11; the technical challenge of Theorem 3 is to show that is succeeds simultaneously with high probability for all $\Sigma_X \in \mathbb{S}_+^n$ with $\mathrm{tr}(\Sigma_X) = n$.

We first recall some notation. In the quantization scheme of Theorem 11 we have a Gaussian codebook $\mathbf{C} = \{c_1, \ldots, c_M\}$ for $M = 2^{n(R^\star+\varepsilon)}$ (where $\varepsilon = \varepsilon(\varepsilon_\circ)$ is a constant) that is generated from the shared randomness $S \in [0,1]$. We denote $\mathrm{EVD}(\Sigma_X) = U\Lambda U^\top$ for $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, and recall the distortion

$$d_{\Sigma_X}(W,C) = (W-C)^\top \Sigma_X (W-C) .$$

The quantizer then sets the optimal scaling factor $\tau = \tau(U^\top W, \Lambda)$ in Eq. (τ def.) and returns a tuple

$$f(W,\Sigma_X,S) = \left(\operatorname*{argmin}_{i\in[M]} d_{\Sigma_X}(W,\tau c_i), q(\tau)\right) ,$$

where $q(\tau)$ is a simple rounding quantizer. The decoder is defined to be

$$g((i,q(\tau)),S) = q(\tau)\cdot c_i .$$

Finally, for some constant $B > 10$, we denote

$$E_{\mathbf{C}} = \left\{\forall i \in [M] :\ \|c_i\|_2 \le n^B\right\} , \qquad \Pr_{\mathbf{C}}[E_{\mathbf{C}}^c] \le \exp\left(-c' n^{2B}\right)$$

and, denoting the conditional distribution $\mathbf{C}|E_{\mathbf{C}}$, the proof of Theorem 11 shows for any $\alpha \in (0,1/4)$ and $\eta = \eta(\eta_\circ)$ to be chosen later,

$$\Pr_{\mathbf{C}|E_{\mathbf{C}}}\left[\mathbb{E}_W\left[\frac{1}{n} d_{\Sigma_X}(W,\widehat{W})\right] \le \mathbf{D}_{\mathrm{rc}}^{\lambda}(R^\star) + \eta\right] \ge 1 - \exp\left(-2^{n\varepsilon(1-c' n^{4\alpha-1})}\right) , \tag{32}$$

for sufficiently large $n \ge n_0 = n_0(\varepsilon,\eta,R^\star,\alpha)$. Simultaneously, it is shown that the scheme above achieves final rate $R \le R + \varepsilon_\circ$ for the appropriately chosen $\varepsilon = \varepsilon(\varepsilon_\circ)$.

Step 1: Covering for $U \in \mathcal{O}_n$ and Λ.

Fact 1 (Covering of $\mathcal{O}_n$) *For a universal constant c,*

$$N(\gamma, \mathcal{O}_n, \|\cdot\|_{op}) \le (c/\gamma)^{n^2}.$$

Consequently, there exists a γ-covering $U_1, \dots, U_N \in \mathcal{O}_n$ such that for all $U \in \mathcal{O}_n$,

$$\min_{i \in N} \|U_i - U\|_{op} \le \gamma\,,$$

and $N \le (c/\gamma)^{n^2}$.

Fact 2 (Covering of Λ) *For a universal constant c,*

$$N\left(\delta/\sqrt{d}; B_1^d, \|\cdot\|_2\right) \le (c + c/\delta)^d\ .$$

Consequently, there exists a $(\tilde{\gamma}\sqrt{n})$-covering $\tilde{s}_1, \dots, \tilde{s}_{N'} \in \left\{s \in \mathbb{R}_+^n : \sum_j s_j \le n\right\}$ such that for all $\lambda \in \mathbb{R}_+^n$ with $\sum_j \lambda_j \le n$,

$$\min_{i \in N'} \|\tilde{s}_i - \lambda\|_2 \le \tilde{\gamma}\sqrt{n}\,,$$

and $N' \le (c/\tilde{\gamma})^n$ for a universal constant c.

Setting $s_i \coloneqq \operatorname{argmin}_{s \in \mathbb{R}_+^n : \sum_j s_j = n} \|s - \tilde{s}_i\|_2$ and applying Fact 2 with $\tilde{\gamma} = \gamma/2$, we obtain:

Corollary 12 (Of Fact 2) *There exists a $(\gamma\sqrt{n})$-covering $s_1, \dots, s_{N'} \in \left\{s \in \mathbb{R}_+^n : \sum_j s_j = n\right\}$ such that for all $\lambda \in \mathbb{R}_+^n$ with $\sum_j \lambda_j = n$,*

$$\min_{i \in N'} \|s_i - \lambda\|_2 = \gamma\sqrt{n}\,,$$

and $N' \le (c/\gamma)^n$ for a universal constant c.

Consider the γ-covering $U_1, \dots, U_N \in \mathcal{O}_n$ and $(\gamma_2\sqrt{n})$-covering $s_1, \dots, s_{N'} \in \left\{s \in \mathbb{R}_+^n : \sum_j s_j = n\right\}$ from the facts above for γ to be chosen later. Denote $\Lambda_i \coloneqq \operatorname{diag}(s_i)$. By a union bound applied to Eq. (32),

$$\Pr_{\mathbf{C}|E_{\mathbf{C}}}\left[\sup_{\Sigma_X = U_i \Lambda_j U_i^\top}\left(\mathbb{E}_W\left[\frac{1}{n} d_{\Sigma_X}(W, \widehat{W})\right] - \mathbf{D}_{\mathrm{rc}}^{s_j}(R^\star)\right) \le \eta\right] \ge 1 - NN' \exp\left(-2^{n\varepsilon(1 - c' n^{4\alpha - 1})}\right)\ .$$

Step 2: Perturbation bound on $\mathbb{E}_W\left[d_{\Sigma_X}(W, \widehat{W})\right]$. Fix any constant $A > 0$ (say $A = 10$). We show the following distortion perturbation results.

Claim 3 (Distortion Perturbation) *Let $\Lambda = \operatorname{diag}(\lambda), \widetilde{\Lambda} = \operatorname{diag}(\widetilde{\lambda})$ be s.t. $\Lambda, \widetilde{\Lambda} \succeq 0$ and $\operatorname{tr}(\Lambda) = \operatorname{tr}(\widetilde{\Lambda}) = n$ and $U, \widetilde{U} \in \mathcal{O}_n$ satisfy*

$$\|\Lambda - \widetilde{\Lambda}\|_2 \le \gamma\sqrt{n} \qquad \textit{and} \qquad \|U - \widetilde{U}\|_{op} \le \gamma\,.$$

Denote

$$\Sigma_X = U\Lambda U^\top \quad \textit{and} \quad \widetilde{\Sigma}_X = \widetilde{U}\widetilde{\Lambda}\widetilde{U}^\top .$$

If $\mathbf{C} = \{c_1, \dots, c_M\} \in (\mathbb{R}^n)^M$ *satisfies* $\max_{i\in[M]} \|c_i\|_2 \le n^B$*, then, in the setup above,*

$$\mathbb{E}_{W\sim\mathcal{N}(0,I_n)}\left[d_{\widetilde{\Sigma}_X}(W,\widehat{W})\right] \le \mathbb{E}_{W\sim\mathcal{N}(0,I_n)}\left[d_{\Sigma_X}(W,\widehat{W})\right] + \frac{1}{2}n^{-A} + \gamma n^{O(B)}\cdot 2^{4nR^\star} ,$$

for any sufficiently large (constant) n*, where* $O(B)$ *hides constants in* $A, B, R^\star$*.*

Plugging in $\gamma = 2^{-5nR^\star}$, the bound in Claim 3 simplifies to n^{-A}, since for sufficiently large (constant) n, $\gamma n^{O(B)}\cdot 2^{4nR^\star} \le \frac{1}{2}n^{-A}$. Moreover, as we will shortly see, $|\mathbf{D}^{\lambda}_{\mathrm{rc}}(R^\star) - \mathbf{D}^{s_j}_{\mathrm{rc}}(R^\star)| \le \mathrm{poly}(n)\cdot\|\lambda - s_j\|_2 \le \mathrm{poly}(n)\cdot\gamma$, so, for a fixed codebook $\mathbf{C}$ and sufficiently large n,

$$\sup_{\Sigma_X = U\Lambda U^\top, U\in\mathcal{O}_n}\left(\mathbb{E}_W\left[d_{\Sigma_X}(W,\widehat{W})\right] - n\mathbf{D}^{\lambda}_{\mathrm{rc}}(R^\star)\right) \le \sup_{\Sigma_X = U_i\Lambda_j U_i^\top}\left(\mathbb{E}_W\left[d_{\Sigma_X}(W,\widehat{W})\right] - n\mathbf{D}^{s_j}_{\mathrm{rc}}(R^\star)\right) + n^{-A} .$$

Then,

$$\Pr_{\mathbf{C}|E_{\mathbf{C}}}\left[\sup_{\Sigma_X = U\Lambda U^\top, U\in\mathcal{O}_n}\left(\mathbb{E}_W\left[\frac{1}{n}d_{\Sigma_X}(W,\widehat{W})\right] - \mathbf{D}^{\lambda}_{\mathrm{rc}}(R^\star)\right) \le \eta + n^{-A-1}\right] \ge 1 - NN'\exp\left(-2^{n\varepsilon(1-c'n^{4\alpha-1})}\right) .$$

In the above, $NN' \le (c/\gamma)^{n^2+n} = 2^{5(n^3+n^2)R^\star}$, so the probability on the RHS of the equation above is $\ge 1 - \exp\left(-2^{n\varepsilon(1-cn^{4\alpha-1})}\right)$ for a constant c. The statement of Theorem 3 follows by a union bound and the fact that $\Pr_{\mathbf{C}}[E^c_{\mathbf{C}}] \le \exp\left(-c'n^{2B}\right)$ and choosing n such that $n^{-A-1} \le \eta$.

It remains to verify that $|\mathbf{D}^{\lambda}_{\mathrm{rc}}(R^\star) - \mathbf{D}^{s_j}_{\mathrm{rc}}(R^\star)| \le \mathrm{poly}(n)\cdot\|\lambda - s_j\|_2$. An explicit calculation (using the implicit function theorem) gives for all $j \in [n]$

$$\begin{aligned}\frac{\partial n\mathbf{D}^{\lambda}_{\mathrm{rc}}(R^\star)}{\partial\lambda_j} &= \frac{1}{(1+\lambda_j T)^2} + \sum_i \frac{\lambda_i^2}{(1+\lambda_i T)^2}\cdot\frac{\frac{T}{1+\lambda_j T}}{\sum_i \frac{\lambda_i}{1+\lambda_i T}} \\ &\le \frac{1}{(1+\lambda_j T)^2} + \frac{1}{1+\lambda_j T} \le 2 ,\end{aligned}$$

which yields the desired statement.

Proof (of Claim 3) Denote $\tau = \tau(W,\Sigma_X)$ and $\widetilde{\tau} = \tau(W,\widetilde{\Sigma}_X)$. In our notation the claim statement is equivalent to, for $W\sim\mathcal{N}(0,I_n)$,

$$\mathbb{E}_W\left[d_{\widetilde{\Sigma}_X}(W, q(\widetilde{\tau})c_i)\right] \le \mathbb{E}_W\left[d_{\Sigma_X}(W, q(\tau)c_j)\right] + \frac{1}{2}n^{-A} + \gamma n^{O(B)}\cdot 2^{4nR^\star} ,$$

where $i = \mathrm{argmin}_k\, d_{\widetilde{\Sigma}_X}(W,\widetilde{\tau}c_k)$ and $j = \mathrm{argmin}_k\, d_{\Sigma_X}(W,\tau c_k)$.

Error from quantizing $\tau,\widetilde{\tau}$. We first uniformly bound the effect of quantizing $\tau,\widetilde{\tau}$. Similarly to Claim 2,

$$\left|\sqrt{d_{\Sigma_X}(W,q(\tau)c_j)} - \sqrt{d_{\Sigma_X}(W,\tau c_j)}\right| \le \delta_\tau\sqrt{c_j^\top\Sigma_X c_j} \le \delta_\tau n^{B+1} .$$

A standard Gaussian tail argument, combined with a bound $\max\{q(\tau),\tau\} \le \|U^\top W\|_\infty$, shows that $\mathbb{E}_W\left[\sqrt{d_{\Sigma_X}(W,q(\tau)c_j)} + \sqrt{d_{\Sigma_X}(W,\tau c_j)}\right] \le n^P$ for a sufficiently large constant P, so setting $\delta_\tau \le \frac{1}{8}n^{-B-1-P-A}$ incurs $o_n(1)$ factors in R (see proof of Theorem 7) and achieves

$$\left|d_{\Sigma_X}(W,q(\tau)c_j) - d_{\Sigma_X}(W,\tau c_j)\right| \le \frac{1}{8}n^{-A} \quad \text{and} \quad \left|d_{\widetilde{\Sigma}_X}(W,q(\widetilde{\tau})c_i) - d_{\widetilde{\Sigma}_X}(W,\widetilde{\tau}c_i)\right| \le \frac{1}{8}n^{-A} .$$

Thus, it is sufficient to show

$$\mathbb{E}_W\left[\min_i d_{\widetilde{\Sigma}_X}(W,\widetilde{\tau}c_i)\right] \le \mathbb{E}_W\left[\min_i d_{\Sigma_X}(W,\tau c_i)\right] + \frac{1}{4}n^{-A} + \gamma n^{O(B)}\cdot 2^{4nR^\star}\,.$$

Error from the tail W event. First, denote the event $E_W = \left\{\|W\|_2 \le n^B\right\}$. We can expand

$$\mathbb{E}_W\left[\min_i d_{\Sigma_X}(W,\tau c_i)\right] = \mathbb{E}_W\left[\min_i d_{\Sigma_X}(W,\tau c_i)\cdot \mathbb{1}_{E_W}\right] + \mathbb{E}_W\left[\min_i d_{\Sigma_X}(W,\tau c_i)\cdot \mathbb{1}_{E_W^c}\right],$$

and, by definition of $d_{\Sigma_X}(W,\tau c_i)$, the second term is

$$\begin{aligned}
\mathbb{E}_W\left[\min_i d_{\Sigma_X}(W,\tau c_i)\cdot \mathbb{1}_{E_W^c}\right] &= \mathbb{E}_W\left[\min_i (W-\tau c_i)^\top \Sigma_X (W-\tau c_i)\cdot \mathbb{1}_{E_W^c}\right] \\
&\le \mathbb{E}_W\left[n\|U^\top W\|_\infty^2 (1+\|U^\top c_i\|_\infty)\cdot \mathbb{1}_{E_W^c}\right] \\
&\le 2n^{21}\mathbb{E}_W\left[\|U^\top W\|_\infty^2\cdot \mathbb{1}_{E_W^c}\right] \\
&\le 2n^{2B+1}\mathbb{E}_W\left[\|W\|_2^2\cdot \mathbb{1}_{E_W^c}\right] \\
&\le 2n^{2B+1}\int_{t=n^B}^{\infty} te^{-c_\chi t} \le c' n^{3B} e^{-c' n^B} \le \frac{1}{4}n^{-A}
\end{aligned}$$

for sufficiently large n, where we used $\tau \le \|U^\top W\|_\infty$, see Claim 1, and c_χ, c' are universal constants.

Error from Σ_X perturbation. We now show that

$$\mathbb{E}_W\left[\min_i d_{\widetilde{\Sigma}_X}(W,\widetilde{\tau}c_i)\cdot \mathbb{1}_{E_W}\right] \le \mathbb{E}_W\left[\min_i d_{\Sigma_X}(W,\tau c_i)\cdot \mathbb{1}_{E_W}\right] + \gamma n^{O(B)}\cdot 2^{4nR^\star}\,, \tag{33}$$

which, combined with the bound above yields the Claim. We use Lemma 14 (proved in Sec. E), which bounds $\left|\tau(W,\Sigma_X) - \tau(W,\widetilde{\Sigma}_X)\right|$ in terms of $\|U^\top W - \widetilde{U}^\top W\|_\infty$ and $\|\Lambda - \widetilde{\Lambda}\|_2$. Given that $\|U^\top W\|_\infty, \|\widetilde{U}^\top W\|_\infty \le \|W\|_2 \le n^B$ under the E_W event, Lemma 14 gives a (crude) bound:

$$|\tau - \widetilde{\tau}| = \left|\tau(W,\Sigma_X) - \tau(W,\widetilde{\Sigma}_X)\right| \le \|U^\top W - \widetilde{U}^\top W\|_\infty + n^{O(B)}\|\Lambda - \widetilde{\Lambda}\|_2 \cdot 2^{4nR^\star}\,.$$

Plugging in $\|U^\top W - \widetilde{U}^\top W\|_\infty \le \|U-\widetilde{U}\|_{op}\|W\|_2 \le \gamma n^B$ and $\|\Lambda - \widetilde{\Lambda}\|_2 \le \gamma\sqrt{n}$, we obtain

$$|\tau - \widetilde{\tau}| \le \gamma n^{O(B)}\cdot 2^{4nR^\star}\,.$$

Additionally,

$$\begin{aligned}
\|\widetilde{\Sigma}_X - \Sigma_X\|_{op} &\le \|U^\top(\Lambda - \widetilde{\Lambda})U\|_{op} + \|U^\top\widetilde{\Lambda}U - \widetilde{U}^\top\widetilde{\Lambda}\widetilde{U}\|_{op} \\
&\le \|\Lambda - \widetilde{\Lambda}\|_2 + \|(U-\widetilde{U})^\top\widetilde{\Lambda}\widetilde{U}\|_{op} + \|\widetilde{U}^\top\widetilde{\Lambda}(U-\widetilde{U})\|_{op} + \|(U-\widetilde{U})^\top\widetilde{\Lambda}(U-\widetilde{U})\|_{op} \\
&\le \gamma\sqrt{n} + 2\gamma n + \gamma^2 n \le 4\gamma n\,.
\end{aligned}$$

Again denote $j = \operatorname{argmin}_k d_{\Sigma_X}(W,\tau c_k)$, and let $x = W - \tau C_j$, $\widetilde{x} = W - \widetilde{\tau}C_j$. Then,

$$\begin{aligned}
\min_i d_{\widetilde{\Sigma}_X}(W,\widetilde{\tau}c_i)\cdot \mathbb{1}_{E_W} &\le d_{\widetilde{\Sigma}_X}(W,\widetilde{\tau}c_j)\cdot \mathbb{1}_{E_W} \\
&= (W-\widetilde{\tau}c_j)^\top\widetilde{\Sigma}_X(W-\widetilde{\tau}c_j)\cdot \mathbb{1}_{E_W} \\
&= \widetilde{x}^\top\widetilde{\Sigma}_X\widetilde{x}\cdot \mathbb{1}_{E_W} \\
&= x^\top\Sigma_X x\cdot \mathbb{1}_{E_W} + (\widetilde{x}-x)^\top\widetilde{\Sigma}_X\widetilde{x}\cdot \mathbb{1}_{E_W} + x^\top\Sigma_X(\widetilde{x}-x)\cdot \mathbb{1}_{E_W} \\
&\qquad + x^\top(\widetilde{\Sigma}_X - \Sigma_X)\widetilde{x}\cdot \mathbb{1}_{E_W}\,.
\end{aligned}$$

Since $\max\{\tau, \widetilde{\tau}\} \le \max\left\{\|U^\top W\|_\infty, \|\widetilde{U}^\top W\|_\infty\right\} \le \|W\|_2$, we can bound

$$\|x\|_2 \cdot \mathbb{1}_{E_W}, \|\widetilde{x}\|_2 \cdot \mathbb{1}_{E_W} \le \|W\|_2(1 + \|C_j\|_2) \cdot \mathbb{1}_{E_W} \le 2n^{2B}$$

and, using the bound on $|\tau - \widetilde{\tau}|$ above,

$$\|x - \widetilde{x}\|_2 \le |\tau - \widetilde{\tau}| \cdot n^B \le \gamma n^{O(B)} \cdot 2^{4nR^\star} .$$

Plugging this into our bound, we obtain

$$\begin{aligned}\min_i d_{\widetilde{\Sigma}_X}(W, \widetilde{\tau} c_i) \cdot \mathbb{1}_{E_W} &\le \min_i d_{\Sigma_X}(W, \tau c_i) \cdot \mathbb{1}_{E_W} + 2\gamma n^{O(B)} \cdot 2^{4nR^\star} + 4n^{4B} \cdot 4\gamma n \\ &\le \min_i d_{\Sigma_X}(W, \tau c_i) \cdot \mathbb{1}_{E_W} + \gamma n^{O(B)} \cdot 2^{4nR^\star} ,\end{aligned}$$

which concludes the proof of this step and the claim. ■

□

Appendix D. Worst-Case Gap Between Waterfilling and Random Coding: Theorem 5

Recall that for $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n) \succeq 0$ with $\operatorname{tr}(\Lambda) = n$, the random-coding rate-distortion is given by a parametric relationship

$$D_{\mathrm{rc}}(\Lambda, T) = \frac{1}{n}\sum_{i=1}^n \frac{\lambda_i}{1 + \lambda_i T} \qquad R_{\mathrm{rc}}(\Lambda, T) = \frac{1}{2n}\sum_{i=1}^n \log(1 + \lambda_i T) , \tag{RDRC}$$

and the waterfilling rate-distortion by

$$D_{\mathrm{wf}}(\Lambda, t) = \frac{1}{n}\sum_{i=1}^n \min\{\lambda_i, t\} \qquad R_{\mathrm{wf}}(\Lambda, t) = \frac{1}{2n}\sum_{i=1}^n \max\{0, \log(\lambda_i/t)\} . \tag{RDWF}$$

Throughout, $\log$ denotes the base-2 logarithm, and $\ln$ denotes the natural logarithm. The goal of this section is to quantify, at a fixed target distortion $D^\star$, the rate overhead of our universal codebook, whose rate-distortion is given in Eq. (RDRC), compared to the Σ_X-fine-tuned optimal codebook, whose rate-distortion is given in Eq. (RDWF).

Denote[5] the implicit functions $T_{\mathrm{rc}}(\Lambda, D^\star) = T^\star$ s.t. $D_{\mathrm{rc}}(\Lambda, T^\star) = D^\star$ and $t_{\mathrm{wf}}(\Lambda, D^\star) = t^\star$ s.t. $D_{\mathrm{wf}}(\Lambda, t^\star) = D^\star$. Denote

$$\mathbf{R}_{\mathrm{rc}}(\Lambda, D^\star) = R_{\mathrm{rc}}(\Lambda, T_{\mathrm{rc}}(\Lambda, D^\star)) \qquad \text{and} \qquad \mathbf{R}_{\mathrm{wf}}(\Lambda, D^\star) = R_{\mathrm{wf}}(\Lambda, t_{\mathrm{wf}}(\Lambda, D^\star)) .$$

In this notation, the rate overhead incurred by our universal codebook is

$$\sup_\Lambda \{\mathbf{R}_{\mathrm{rc}}(\Lambda, D^\star) - \mathbf{R}_{\mathrm{wf}}(\Lambda, D^\star)\} ,$$

where the supremum is over $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n) \succeq 0$ with $\operatorname{tr}(\Lambda) = n$. In Theorem 5, we prove

$$\sup_{D^\star \in (0,1)} \sup_\Lambda \{\mathbf{R}_{\mathrm{rc}}(\Lambda, D^\star) - \mathbf{R}_{\mathrm{wf}}(\Lambda, D^\star)\} \le 0.11 .$$

Concretely, we first verify that the maximum rate gap is approached at spectra containing at most 2 distinct eigenvalue and at vanishing distortions. The resulting gap expression can be directly bounded by 0.11 bit.

5. All of the $D_{\mathrm{rc}}, D_{\mathrm{wf}}, R_{\mathrm{rc}}, R_{\mathrm{wf}}$ are monotone in T, t.

D.1. Proof of Theorem 5

Theorem 5 (Worst-Case Gap Between Waterfilling and Random Coding)

$$\sup_{D^\star\in(0,1)} \sup_{\Lambda} \{\mathbf{R}_{\rm rc}(\Lambda, D^\star) - \mathbf{R}_{\rm wf}(\Lambda, D^\star)\} \le 0.11\,,$$

where the supremum is over $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n) \succeq 0$ *with* $\mathrm{tr}(\Lambda) = n$.

Theorem 5 follows immediately from a more general version of the same statement in Theorem 13. Let $\mathcal{P}([0,\infty))$ denote the space of Borel probability measures on $[0,\infty)$, and define

$$\mathcal{P}_1([0,\infty)) = \left\{\lambda \in \mathcal{P}([0,\infty)) : \int x d\lambda(x) = 1\right\}.$$

We extend the finite-dimension rate-distortion curves in Eq. (RDRC) and (RDWF) as follows: for $\mu \in \mathcal{P}_1([0,\infty))$, let

$$D_{\rm rc}(\mu, T) = \int_0^\infty \frac{x}{1+xT} d\mu(x), \qquad R_{\rm rc}(\mu, T) = \frac{1}{2}\int_0^\infty \log(1+xT) d\mu(x),$$

and

$$D_{\rm wf}(\mu, t) = \int_0^\infty \min\{x, t\}\, d\mu(x), \qquad R_{\rm wf}(\mu, t) = \frac{1}{2}\int_0^\infty \max\{0, \log(x/t)\}\, d\mu(x).$$

Similarly to above, for each $D^\star \in (0,1)$, let $T_{\rm rc}(\mu, D^\star)$ and $t_{\rm wf}(\mu, D^\star)$ be defined by $D_{\rm rc}(\mu, T_{\rm rc}(\mu, D^\star)) = D^\star$ and $D_{\rm wf}(\mu, t_{\rm wf}(\mu, D^\star)) = D^\star$, and set

$$\mathbf{R}_{\rm rc}(\mu, D^\star) = R_{\rm rc}(\mu, T_{\rm rc}(\mu, D^\star)), \qquad \mathbf{R}_{\rm wf}(\mu, D^\star) = R_{\rm wf}(\mu, t_{\rm wf}(\mu, D^\star)).$$

Theorem 13 (Worst-Case Gap Between Waterfilling and Random Coding)

$$\sup_{D^\star\in(0,1)} \sup_{\mu} \{\mathbf{R}_{\rm rc}(\mu, D^\star) - \mathbf{R}_{\rm wf}(\mu, D^\star)\} \le 0.11\,,$$

where the supremum is over probability measures $\mu \in \mathcal{P}_1([0,\infty))$.

Indeed, for any $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n) \succeq 0$ with $\mathrm{tr}(\Lambda) = n$, the empirical measure $\mu_\Lambda = \frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i}$ belongs to $\mathcal{P}_1([0,\infty))$.

Proof (of Theorem 13). Denote the rate-gap

$$\Delta(\mu, D^\star) = \frac{1}{2}\int r_{T_{\rm rc}(\mu,D^\star), t_{\rm wf}(\mu,D^\star)}(x) d\mu(x)\,, \qquad r_{T,t}(x) = \begin{cases} \log(1+xT) & \text{if } x \le t \\ \log\left(\frac{1+xT}{x/t}\right) & \text{if } x > t\,. \end{cases}$$

It will be convenient to consider the following rescaling:

$$\widetilde{\mu} := (x \to x/D^\star)_{\sharp\mu}\,, \tag{$\widetilde{\mu}$}$$

so that $\int x d\widetilde{\mu}(x) = (D^\star)^{-1} > 1$. Accordingly, we define $D^\star(\widetilde{\mu}) = \left(\int x d\widetilde{\mu}(x)\right)^{-1}$. Additionally, rescale the implicitly defined parameters $t_{\rm wf}$ and $T_{\rm rc}$:

$$\widetilde{t}_{\rm wf}(\widetilde{\mu}) \triangleq t_{\rm wf}\left((x \to x\cdot D^\star)_{\sharp\widetilde{\mu}}, D^\star(\widetilde{\mu})\right)/D^\star(\widetilde{\mu}) \qquad \text{and} \qquad \widetilde{T}_{\rm rc}(\widetilde{\mu}) \triangleq T_{\rm rc}\left((x \to x\cdot D^\star)_{\sharp\widetilde{\mu}}, D^\star(\widetilde{\mu})\right)\cdot D^\star(\widetilde{\mu})\,,$$

i.e., $\widetilde{t}_{\rm wf}(\widetilde{\mu}) = t_{\rm wf}(\mu, D^\star)/D^\star$ and $\widetilde{T}_{\rm rc}(\widetilde{\mu}) = T_{\rm rc}(\mu, D^\star) \cdot D^\star$ in the original parameters. The distortion conditions then correspond to

$$1 = \int \frac{x}{1+x\widetilde{T}} d\widetilde{\mu}(x) = \int \min\left\{x, \widetilde{t}\right\} d\widetilde{\mu}(x)\,, \quad \text{where} \quad \widetilde{t} = \widetilde{t}_{\rm wf}(\widetilde{\mu}), \widetilde{T} = \widetilde{T}_{\rm rc}(\widetilde{\mu})\,. \tag{Dist}$$

We can now rewrite the optimization objective in terms of $\widetilde{\mu}$:

$$\Phi(\widetilde{\mu}) \triangleq \Delta\left((x \to x \cdot D^\star)_{\sharp\widetilde{\mu}}, D^\star(\widetilde{\mu})\right) = \Delta(\mu, D^\star) = \frac{1}{2}\int r_{\widetilde{T},\widetilde{t}}(x) d\widetilde{\mu}(x)\,. \tag{Φ}$$

Finally, for technical reasons, we will consider $\widetilde{\mu} \in \mathcal{P}([0,\infty]) \supseteq \mathcal{P}([0,\infty))$ with the natural extensions of functions $\frac{x}{1+x\widetilde{T}}, \log\left(\frac{1+x\widetilde{T}}{x/\widetilde{t}}\right)$. To prove the theorem it is then sufficient to show

$$\sup_{\widetilde{\mu}\in\mathcal{A}} \Phi(\widetilde{\mu}) \le 0.11\,, \quad \text{where} \quad \mathcal{A} \triangleq \left\{\widetilde{\mu} \in \mathcal{P}([0,\infty]) : D^\star(\widetilde{\mu}) \in (0,1)\right\}.$$

In what follows, we will restrict the optimization scope $\mathcal{A}$ in three steps. For $\widetilde{\mu}$ as above, to simplify notation, when clear from the context, we write $D^\star = D^\star(\widetilde{\mu})$, $\widetilde{t} = \widetilde{t}_{\rm wf}(\widetilde{\mu})$, $\widetilde{T} = \widetilde{T}_{\rm rc}(\widetilde{\mu})$.

Step I. Denote, for $c = 0.11 \ln 2$,

$$\mathcal{A}_{\rm bd} = \left\{\widetilde{\mu} \in \mathcal{A} :\ \widetilde{t}_{\rm wf}(\widetilde{\mu}) \in \left[1, c^{-1}\right], \widetilde{T}_{\rm rc}(\widetilde{\mu}) \in [c, 1]\,, \text{ and } \widetilde{t}_{\rm wf}(\widetilde{\mu}) \cdot \widetilde{T}_{\rm rc}(\widetilde{\mu}) \le 1\right\}.$$

As a first step, we show

$$\sup_{\widetilde{\mu}\in\mathcal{A}} \Phi(\widetilde{\mu}) \le \max\left\{0.11, \sup_{\widetilde{\mu}\in\mathcal{A}_{\rm bd}} \Phi(\widetilde{\mu})\right\}.$$

Let $\kappa = \kappa(\widetilde{\mu}) \triangleq \int_{\widetilde{t}}^{+\infty} d\widetilde{\mu}(x)$ be the mass $\widetilde{\mu}$ puts to the waterfilling-active modes.[6] Expanding $\int \min\left\{x, \widetilde{t}\right\} d\widetilde{\mu}(x) = \int_0^{\widetilde{t}} x d\widetilde{\mu}(x) + \kappa\widetilde{t}$, from Eq. (Dist) we conclude $\widetilde{t} \ge 1$, $\widetilde{T} \le 1$, and $\widetilde{t}\widetilde{T} \le 1$, where the last inequality is derived as follows:

$$1 = \int \frac{x}{1+x\widetilde{T}} d\widetilde{\mu}(x) \le \int_0^{\widetilde{t}} x d\widetilde{\mu}(x) + \frac{\kappa}{\widetilde{T}} = (1-\kappa\widetilde{t}) + \frac{\kappa}{\widetilde{T}} = 1 + \kappa\left(-\widetilde{t} + \frac{1}{\widetilde{T}}\right).$$

We now obtain bounds on $\Phi(\widetilde{\mu})$ in terms of $\widetilde{t}, \widetilde{T}$:

$$\Phi(\widetilde{\mu}) = \underbrace{\frac{1}{2}\int_0^{\widetilde{t}} \log(1+x\widetilde{T}) d\widetilde{\mu}(x)}_{\rm I} + \underbrace{\frac{1}{2}\int_{\widetilde{t}}^{+\infty} \log\left(\frac{1+x\widetilde{T}}{x/\widetilde{t}}\right) d\widetilde{\mu}(x)}_{\rm II}.$$

From $\int \min\left\{x, \widetilde{t}\right\} d\widetilde{\mu}(x) = 1$ and $\log(1+x) \le x/\ln 2$, we bound ${\rm I} \le \frac{(1-\kappa)\widetilde{T}}{2\ln 2}$. To bound II, notice that for $x \ge \widetilde{t}$, $\log\left(\frac{1+x\widetilde{T}}{x/\widetilde{t}}\right) = \log(\widetilde{t}\widetilde{T} + \widetilde{t}/x) \le \log(\widetilde{t}\widetilde{T}+1) \le \widetilde{t}\widetilde{T}/\ln 2$, so overall,

$${\rm I} + {\rm II} \le \frac{(1-\kappa)\widetilde{T}}{2\ln 2} + \frac{\kappa\widetilde{t}\widetilde{T}}{2\ln 2} \le \frac{\widetilde{T}}{2\ln 2} + \frac{\widetilde{T}}{2\ln 2} = \frac{\widetilde{T}}{\ln 2} \le \frac{1}{\widetilde{t}\ln 2}\,, \tag{34}$$

6. $\kappa > 0$ from the water-filling distortion condition.

where the last inequality used previously derived $\widetilde{t}\widetilde{T} \le 1$. Then, if $\widetilde{t} \ge c^{-1}$ or $\widetilde{T} \le c$, we obtain the desired bound $\Phi(\widetilde{\mu}) \le 0.11$, which concludes this step.

Step II. Second, we show that the largest gap is obtained at spectra, whose mass in the active tail (i.e., values $\ge \widetilde{t}$) is distributed between endpoints $\widetilde{t}$ and $+\infty$. Concretely, we show that

$$\sup_{\widetilde{\mu}\in\mathcal{A}_{\rm bd}} \Phi(\widetilde{\mu}) \le \sup_{\widetilde{\mu}\in\mathcal{A}_{\rm tail}} \Phi(\widetilde{\mu})\,,$$

where

$$\mathcal{A}_{\rm tail} = \left\{\widetilde{\mu}\in\mathcal{A}_{\rm bd}:\ \text{for } x \ge \widetilde{t}_{\rm wf}(\widetilde{\mu}), \theta\in[0,1],\ \widetilde{\mu}(x) \propto \theta\cdot\delta_{\widetilde{t}}(x) + (1-\theta)\cdot\delta_{+\infty}(x)\right\}\,.$$

Recall the contribution to the rate gap of the active modes (those $\ge \widetilde{t}$) is

$$\frac{1}{2}\int_{\widetilde{t}}^{+\infty} \log\left(\frac{1+x\widetilde{T}}{x/\widetilde{t}}\right) d\widetilde{\mu}(x) = \frac{1}{2}\int_{\widetilde{t}}^{+\infty} \log\left(\widetilde{t}\cdot\left(\frac{x}{1+x\widetilde{T}}\right)^{-1}\right) d\widetilde{\mu}(x)\,.$$

The distortion constraints in Eq. (Dist) yield an equality $\int_{\widetilde{t}}^{\infty} \frac{x}{1+x\widetilde{T}} d\widetilde{\mu}(x) = 1 - \int_0^{\widetilde{t}} \frac{x}{1+x\widetilde{T}} d\widetilde{\mu}(x)$. Note that for $x\in[\widetilde{t},+\infty]$, $\frac{x}{1+x\widetilde{T}} \in \left[\frac{\widetilde{t}}{1+\widetilde{t}\widetilde{T}}, \frac{1}{\widetilde{T}}\right]$ and is a monotone function. Moreover, both the objective $f(y) = \log(\widetilde{t}y^{-1})$ and the first moment constraint are convex, so the supremum is obtained at

$$\widetilde{\mu}(x) \propto \theta\cdot\delta_{\widetilde{t}}(x) + (1-\theta)\cdot\delta_{+\infty}(x) \qquad \text{for } x \ge \widetilde{t}$$

for some $\theta\in[0,1]$. Note that this suggests that the extremal regime occurs at vanishing distortions.

Step III. Finally, we show that the inactive modes (those $\le \widetilde{t}$) of the worst-case spectra $\widetilde{\mu}$ equalize, i.e.,

$$\sup_{\widetilde{\mu}\in\mathcal{A}_{\rm tail}} \Phi(\widetilde{\mu}) \le \sup_{\widetilde{\mu}\in\mathcal{A}_{\rm 2pt}} \Phi(\widetilde{\mu})\,,$$

where

$$\mathcal{A}_{\rm 2pt} = \left\{\widetilde{\mu}\in\mathcal{A}_{\rm tail}:\ \text{for } \theta\in[0,1], \widetilde{\mu}_0 \le \widetilde{t}_{\rm wf}(\widetilde{\mu}),\ \widetilde{\mu}(x) \propto \theta\cdot\delta_{\widetilde{\mu}_0}(x) + (1-\theta)\cdot\delta_{+\infty}(x)\right\}\,.$$

For a given $\widetilde{\mu}\in\mathcal{A}_{\rm tail}$ with $\widetilde{\mu} \propto (1-\theta)\cdot\delta_{+\infty}(x)$ for $x > \widetilde{t}$, define a corresponding measure $\nu\in\mathcal{A}_{\rm 2pt}$ as

$$\nu \propto \theta\cdot\delta_{\widetilde{\mu}_0} + (1-\theta)\cdot\delta_{+\infty}(x)\,, \quad \text{where} \quad \widetilde{\mu}_0 = \frac{\int_0^{\widetilde{t}} x d\widetilde{\mu}(x)}{\theta}\,.$$

To complete Step III, it is sufficient to show that $\Phi(\widetilde{\mu}) \le \Phi(\nu)$. Define a path $\mu_s = s\cdot\widetilde{\mu} + (1-s)\cdot\nu$ for $s\in[0,1]$, and $J(s) \triangleq \Phi(\mu_s)$. We will show that $\frac{\partial J(s)}{\partial s} \le 0$ for all $s\in[0,1]$, yielding $\Phi(\widetilde{\mu}) = J(1) \le J(0) = \Phi(\nu)$.

For this, we compute:

$$\frac{\partial \widetilde{T}_{\rm rc}(\mu_s)}{\partial s} = -\frac{\frac{\partial \int \frac{x}{1+x\widetilde{T}} d\mu_s(x)}{\partial s}}{\frac{\partial \int \frac{x}{1+x\widetilde{T}} d\mu_s(x)}{\partial \widetilde{T}}} = \frac{\int \frac{x}{1+x\widetilde{T}} d(\widetilde{\mu}(x) - \nu(x))}{\int \frac{x^2}{(1+x\widetilde{T})^2} d\mu_s(x)}\,.$$

Note that for all $s \in [0,1]$, $\widetilde{t}_{\rm wf}(\mu_s)$ is unchanged. Then,

$$\begin{aligned}\frac{\partial 2J(s)}{\partial s} &= \frac{\partial\left(\int_0^{\widetilde{t}}\log(1+x\widetilde{T})d\mu_s(x) + (1-\theta)\log(\widetilde{T}\widetilde{t})\right)}{\partial s}\\ &= \int_0^{\widetilde{t}}\log(1+x\widetilde{T})d(\widetilde{\mu}(x)-\nu(x)) + \frac{1}{\ln 2}\cdot\left(\int_0^{\widetilde{t}}\frac{x}{1+x\widetilde{T}}d\mu_s(x) + \frac{1-\theta}{\widetilde{T}}\right)\cdot\frac{\int\frac{x}{1+x\widetilde{T}}d(\widetilde{\mu}(x)-\nu(x))}{\int\frac{x^2}{(1+x\widetilde{T})^2}d\mu_s(x)}\\ &= \int_0^{\widetilde{t}}\left(\log(1+x\widetilde{T}) + A_s\cdot\frac{x}{1+x\widetilde{T}}\right)d(\widetilde{\mu}(x)-\nu(x))\,,\end{aligned}$$

where $A_s \geq 0$. The function $\log(1+x\widetilde{T}) + A_s\cdot\frac{x}{1+x\widetilde{T}}$ is concave on $x\in[0,\widetilde{t}]$, yielding the final inequality $\frac{\partial J(s)}{\partial s}\leq 0$. It remains to show $\sup_{\widetilde{\mu}\in\mathcal{A}_{\rm 2pt}}\Phi(\widetilde{\mu})\leq 0.11$. Recall that every $\widetilde{\mu}\in\mathcal{A}_{\rm 2pt}$ can be expressed for some $\widetilde{\mu}_0\leq\widetilde{t}_{\rm wf}(\widetilde{\mu})$ as

$$\widetilde{\mu}(x) = \theta\cdot\delta_{\widetilde{\mu}_0}(x) + (1-\theta)\cdot\delta_{+\infty}(x)\,.$$

We can then express

$$2\Phi(\widetilde{\mu}) = \theta\log(1+\widetilde{\mu}_0\widetilde{T}) + (1-\theta)\log(\widetilde{t}\widetilde{T})\,,$$

where from Eq. (Dist),

$$1 = \theta\widetilde{\mu}_0 + (1-\theta)\widetilde{t} = \frac{\theta\widetilde{\mu}_0}{1+\widetilde{\mu}_0\widetilde{T}} + \frac{1-\theta}{\widetilde{T}}\,.$$

The case $\theta = 1$ is impossible for $\widetilde{T} > 0$, so we will focus on the $\theta\in[0,1)$ case. We now further simplify this optimization problem to obtain the final 0.11 bound.

From the first distortion constraint, we express $\widetilde{t} = (1-\theta)^{-1} - \left((1-\theta)^{-1}-1\right)\widetilde{\mu}_0$. From the second distortion constraint, $\widetilde{T} = \frac{\theta\widetilde{\mu}_0\widetilde{T}}{1+\widetilde{\mu}_0\widetilde{T}} + 1 - \theta$. Denoting $\alpha = \widetilde{\mu}_0\widetilde{T}$, we then simplify $\widetilde{t}\widetilde{T} = (1-\theta)^{-1}\widetilde{T} - \left((1-\theta)^{-1}-1\right)\alpha$. Plugging into the rate-gap expression:

$$\begin{aligned}2\Phi(\widetilde{\mu}) &\leq \theta\log(1+\alpha) + (1-\theta)\log\left(1 + \frac{\theta\alpha}{(1+\alpha)(1-\theta)} - \frac{\alpha\theta}{1-\theta}\right)\\ &= \theta\log(1+\alpha) + (1-\theta)\log\left(1 - \frac{\alpha^2\theta}{(1+\alpha)(1-\theta)}\right)\,,\end{aligned}$$

where from the positivity of $\widetilde{t},\widetilde{T}$ we have $0\leq\theta<\frac{1+\alpha}{1+\alpha+\alpha^2}$ and $0\leq\alpha=\widetilde{\mu}_0\widetilde{T}\leq\widetilde{t}\widetilde{T}\leq 1$. Denoting $\rho = \frac{\theta}{1-\theta}$ and differentiating with respect to α gives

$$\begin{aligned}\ln 2(1-\theta)^{-1}\frac{\partial(2\Phi(\widetilde{\mu}))}{\partial\alpha} &= \frac{\rho}{1+\alpha} - \frac{1}{1-\frac{\alpha^2\rho}{1+\alpha}}\cdot\rho\cdot\frac{2\alpha+\alpha^2}{(1+\alpha)^2}\\ &= \frac{\rho}{1+\alpha}\left(1 - \frac{2\alpha+\alpha^2}{1+\alpha-\alpha^2\rho}\right)\,.\end{aligned}$$

Setting this to 0 yields $\rho = \frac{1-\alpha-\alpha^2}{\alpha^2}$. At the boundary points $\alpha\in\left\{0, \frac{\sqrt{5}-1}{2}\right\}$, the objective function is below the objective of the Theorem. Plugging into the objective, we derive a one-parameter

function

$$2\Delta^\star \le \frac{1-\alpha-\alpha^2}{1-\alpha}\log(1+\alpha) + \frac{\alpha^2}{1-\alpha}\log\left(1-\frac{1-\alpha-\alpha^2}{1+\alpha}\right).$$

The maximum of this objective is achieved at $\alpha \approx 0.35$ and is $\approx 2\cdot 0.108 < 2\cdot 0.11$, which concludes the proof.

■

Appendix E. Proof of Lemma 14

Lemma 14 (τ Perturbation) *Let $R^\star$ be a fixed constant, $C_W = C_W(n) \in \mathbb{R}$, $W \in \mathbb{R}^n, s \in \mathbb{R}^n_+$, such that $\|W\|_\infty \le C_W$ and $\sum_j s_j = n$. Denote*

$$\tau(W,s) = \left(T\sum_j \frac{W_j^2 s_j^2}{(1+s_jT)^2}\right)^{1/2}\left(\sum_j \frac{s_j}{1+s_jT}\right)^{-1/2},$$

where $T = T(s)$ is a unique solution to $R^s_{\rm rc}(T) = R^\star$ (see Eq. (RDRC) for definition of $R^s_{\rm rc}(T)$). If $W, \widehat{W}, s, \widehat{s} \in \mathbb{R}^n$ satisfy $\|W\|_\infty, \|\widehat{W}\|_\infty \le C_W$ and $\sum_j s_j = \sum_j \widehat{s}_j = n$ and

$$\|W-\widehat{W}\|_\infty \le \gamma_1 C_W \quad \text{and} \quad \|s-\widehat{s}\|_2 \le \gamma_2\sqrt{n},$$

then

$$\left|\tau(W,s) - \tau(\widehat{W},\widehat{s})\right| \le \gamma_1 C_W + c\cdot C_W\gamma_2 n^4 2^{4nR^\star}$$

for sufficiently large n and a constant $c = c(R^\star)$.

Proof of Lemma 14. Denote $T, \widehat{T}$ to be the solutions to $R^s_{\rm rc}(T) = R^\star$ and $R^{\widehat{s}}_{\rm rc}(\widehat{T}) = R^\star$ respectively, where recall that $R^{V,s}_{\rm rc} = R^s_{\rm rc}$ does not depend on V.

Step 1: Bound on $|T-\widehat{T}|$. First, we show that under the Lemma conditions,

$$|T-\widehat{T}| \le \gamma_2 n\cdot 2^{4nR^\star}. \tag{35}$$

First, since for each $j\in[n]$, $\left|\log(1+s_jT) - \log(1+\widehat{s}_jT)\right| \le \frac{1}{\ln 2}\cdot\frac{T}{1+\min(s_j,\widehat{s}_j)T}\cdot|s_j-\widehat{s}_j| \le T|s_j-\widehat{s}_j|/\ln 2$, we have

$$\left|\underbrace{R^{\widehat{s}}_{\rm rc}(\widehat{T})}_{R^\star} - R^{\widehat{s}}_{\rm rc}(T)\right| = \left|R^s_{\rm rc}(T) - R^{\widehat{s}}_{\rm rc}(T)\right| \le \frac{T\|s-\widehat{s}\|_1}{2n\ln 2} \le \frac{T\|s-\widehat{s}\|_2}{2\sqrt{n}\ln 2} \le \frac{T\gamma_2}{2\ln 2}.$$

By the mean value theorem, we have for some $\bar{T} \in \left[\min(T,\widehat{T}), \max(T,\widehat{T})\right]$,

$$|T-\widehat{T}| = \frac{\left|R^{\widehat{s}}_{\rm rc}(\widehat{T}) - R^{\widehat{s}}_{\rm rc}(T)\right|}{(R^{\widehat{s}}_{\rm rc})'(\bar{T})} \le \frac{T\gamma_2}{2\ln 2\cdot(R^{\widehat{s}}_{\rm rc})'(\bar{T})}.$$

A direct calculation gives $(R^{\widehat{s}}_{\rm rc})'(\bar{T}) = \frac{1}{2n\ln 2}\sum_j \frac{\widehat{s}_j}{1+\widehat{s}_j\bar{T}} \ge \frac{1}{2n\ln 2(1+\bar{T})} \ge \frac{1}{2n\ln 2(1+\max(T,\widehat{T}))}$, where we used that $\max_j \widehat{s}_j \ge 1$. Moreover, for any $s\in\mathbb{R}^n_+$ with $\|s\|_1 = n$, $2nR^s_{\rm rc}(T) \ge \log(1+T)$, and therefore, $1+\max(T,\widehat{T}) \le 2^{2nR^\star}$. This yields

$$|T-\widehat{T}| \le \gamma_2 n\cdot T(1+\max(T,\widehat{T})) \le \gamma_2 n\cdot 2^{4nR^\star}.$$

Step 2: Bound on $|\tau(W,s)-\tau(\widehat{W},s)|$. Here we show that

$$|\tau(W,s)-\tau(\widehat{W},s)| \le \gamma_1 C_W\,. \tag{36}$$

By triangle inequality and $\frac{s_j}{1+s_jT} \le \frac{1}{T}$,

$$\begin{aligned}|\tau(W,s)-\tau(\widehat{W},s)| &\le \left(T\sum_j \frac{(W_j-\widehat{W}_j)^2 s_j^2}{(1+s_jT)^2}\right)^{1/2}\left(\sum_j \frac{s_j}{1+s_jT}\right)^{-1/2}\\ &\le \gamma_1 C_W\cdot\left(T\sum_j \frac{s_j^2}{(1+s_jT)^2}\right)^{1/2}\left(\sum_j \frac{s_j}{1+s_jT}\right)^{-1/2} \le \gamma_1 C_W\,.\end{aligned}$$

Step 3: Bound on $|\tau(W,s)-\tau(W,\widehat{s})|$. Here we bound $|\tau(W,s)-\tau(W,\widehat{s})|$. For convenience, denote $d_j = \frac{s_j}{1+s_jT}$ $A(s,T) =: \sum_j W_j^2 d_j^2$, $B(s,T) =: \sum_j d_j$ and let $A = A(s,T)$, $\widehat{A} = A(\widehat{s},\widehat{T})$ and $B = B(s,T)$, $\widehat{B} = B(\widehat{s},\widehat{T})$, so that

$$|\tau(W,s)-\tau(W,\widehat{s})| = \Big|\frac{\sqrt{TA}}{\sqrt{B}} - \frac{\sqrt{\widehat{T}\widehat{A}}}{\sqrt{\widehat{B}}}\Big| \le \sqrt{T}\Big|\frac{\sqrt{A}}{\sqrt{B}} - \frac{\sqrt{\widehat{A}}}{\sqrt{\widehat{B}}}\Big| + \big|\sqrt{T}-\sqrt{\widehat{T}}\big|\cdot\frac{\sqrt{\widehat{A}}}{\sqrt{\widehat{B}}} =: \mathrm{I}+\mathrm{II}$$

by triangle inequality. To bound I, we use a triangle inequality $\mathrm{I} = \sqrt{T}\Big|\frac{\sqrt{A}}{\sqrt{B}} - \frac{\sqrt{\widehat{A}}}{\sqrt{\widehat{B}}}\Big| \le \sqrt{T}\frac{\big|\sqrt{A}-\sqrt{\widehat{A}}\big|}{\sqrt{B}} + \sqrt{T\widehat{A}}\Big|\frac{1}{\sqrt{B}} - \frac{1}{\sqrt{\widehat{B}}}\Big|$ and notice

$$\big|\sqrt{A}-\sqrt{\widehat{A}}\big| = \big|\|W\odot d\|_2 - \|W\odot\widehat{d}\|_2\big| \le \|W\|_\infty\|d-\widehat{d}\|_2 \le C_W\|d-\widehat{d}\|_2\,.$$

Moreover,

$$\big|\sqrt{B}-\sqrt{\widehat{B}}\big| = \big|\|d\|_1^{1/2} - \|\widehat{d}\|_1^{1/2}\big| \le \frac{\sqrt{n}\|d-\widehat{d}\|_2}{\sqrt{B}+\sqrt{\widehat{B}}}\,.$$

For a uniform lower bound on $B, \widehat{B}$ we use $\min(B,\widehat{B}) \ge \frac{1}{1+T}$, so the combined bound is

$$\begin{aligned}\mathrm{I} &\le \sqrt{T}\|d-\widehat{d}\|_2\left[C_W(1+T)^{1/2} + \sqrt{\widehat{A}n}(1+T)^{3/2}\right]\\ &\le 2\|d-\widehat{d}\|_2 C_W n 2^{2nR^\star}\,.\end{aligned}$$

where we used $\widehat{A} \le C_W^2\frac{n}{T^2}$ and $1+T \le 2^{2nR^\star}$. It remains to bound $\|d-\widehat{d}\|_2$. Denoting $d(s,T) = \left(\frac{s_j}{1+s_jT}\right)_j$ (so that $d = d(s,T)$ and $\widehat{d} = d(\widehat{s},\widehat{T})$), we have

$$\|d-\widehat{d}\|_2 \le \|d(s,T)-d(\widehat{s},T)\|_2 + \|d(\widehat{s},T)-d(\widehat{s},\widehat{T})\|_2\,.$$

Since $\Big|\frac{s_j}{1+s_jT} - \frac{\widehat{s}_j}{1+\widehat{s}_jT}\Big| \le |s_j-\widehat{s}_j|$, we have $\|d(s,T)-d(\widehat{s},T)\|_2 \le \|d(s,T)-d(\widehat{s},T)\|_1 \le \|s-\widehat{s}\|_1 \le \gamma_2 n$. Moreover, $\Big|\frac{s_j}{1+s_jT} - \frac{s_j}{1+s_j\widehat{T}}\Big| \le n^2|T-\widehat{T}|$, so we obtain

$$\|d-\widehat{d}\|_2 \le \gamma_2 n + 2\gamma_2 n^4 2^{4nR^\star} \le 3\gamma_2 n^4 2^{4nR^\star}\,,$$

where we used $|T - \widehat{T}| \le 2\gamma_2 n 2^{4nR^\star}$ in Eq. (35). To bound II, we use the bound in Eq. (35) and the lower bound $T \ge 2R^\star \ln 2$ obtained in the proof of Thm. 7:

$$\left|\sqrt{T} - \sqrt{\widehat{T}}\right| = \frac{\left|T - \widehat{T}\right|}{\sqrt{T} + \sqrt{\widehat{T}}} \le \frac{\gamma_2 n 2^{4nR^\star}}{\sqrt{2R^\star \ln 2}} .$$

Then,

$$\mathrm{II} \le \frac{\gamma_2 n 2^{4nR^\star}}{\sqrt{2R^\star \ln 2}} \cdot \frac{C_W \sqrt{n}}{2T} \cdot \sqrt{1+T} \le \gamma_2 C_W n^{3/2} 2^{4nR^\star} / (2R^\star \ln 2) .$$

We have for the final bound

$$\begin{aligned} |\tau(W,s) - \tau(W,\widehat{s})| \le \mathrm{I} + \mathrm{II} &\le 6\gamma_2 n^4 2^{4nR^\star} C_W n 2^{2nR^\star} + \gamma_2 C_W n^{3/2} 2^{4nR^\star} / (2R^\star) \\ &\le c \cdot C_W \gamma_2 n^4 2^{4nR^\star} \end{aligned}$$

for sufficiently large n and a constant $c = c(R^\star)$. Analogous bound holds for $|\tau(\widehat{W}, s) - \tau(\widehat{W}, \widehat{s})|$.

Step 4: Final bound. Combining Steps 2,3 above, we obtain

$$\begin{aligned} \left|\tau(W,s) - \tau(\widehat{W},\widehat{s})\right| &\le \left|\tau(W,s) - \tau(\widehat{W},s)\right| + \left|\tau(\widehat{W},s) - \tau(\widehat{W},\widehat{s})\right| \\ &\le \gamma_1 C_W + c \cdot C_W \gamma_2 n^4 2^{4nR^\star} . \end{aligned}$$

□

Appendix F. Proof of upper bound in oracle Proposition 2

Here we sketch proof of (3). First, since f and g are allowed to depend on Σ_X we can rotate Σ_X to eigenbasis, and thus assume from now on that $\Sigma_X = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$. Fix t and let $D_i = \min\{t/\lambda_i, 1\}$. Let $R_0 = R_{\mathrm{wf}}(\Sigma_X, t)$. Fix arbitrary $\epsilon > 0$ and set rate $R = R_0 + 2\epsilon$. We will show that by generating codebook $\mathbf{C}$ randomly via sampling $1 + 2^{nR}$ codewords from distribution $\prod_{i=1}^n \mathcal{N}(0, 1 - D_i)$ one can attain distortion

$$\mathbb{E}_W \left[\min_{c \in \mathbf{C}} d_{\Sigma_X}(W, c)\right] \le n D_{\mathrm{wf}}(\Sigma_X, t) + \left(e^{-2^{n\epsilon}} + c_1 e^{-c_2 n \epsilon^2}\right) \mathrm{tr}\, \Sigma_X , \tag{37}$$

where $c_1, c_2 > 0$ are some absolute constants. From here the statement of the theorem follows by setting $\epsilon_n = c\sqrt{\frac{\log n}{n}}$ with an apropriate $c > 0$.

To show (37) let $Y_i \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, 1 - D_i)$ independently. Let also $Z_i \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,1)$. Set

$$W_i = Y_i + \sqrt{D_i} Z_i$$

and notice that $W = (W_1, \dots, W_n) \sim \mathcal{N}(0, I_n)$. This coupling of W to Y satisfies a useful property:

$$\mathbb{E}_W[d_{\Sigma_X}(W, Y)] = \mathbb{E}_W \left[\sum_{i=1}^n \lambda_i D_i Z_i^2\right] = D_{\mathrm{wf}}(\Sigma_X, t) .$$

On the other hand, we have *information density*

$$i(W;Y) := \log \frac{dP_{W,Y}}{d(P_W \times P_Y)}(W,Y) = nR_0 + \frac{\log e}{2} \sum_{i=1}^n W_i^2 - Z_i^2 .$$

Note that $\mathbb{E}[i(W;Y)] = I(W;Y) = nR_0$. Note that while W_i, Z_i are jointly correlated Gaussians, they are independent for different i's. A fascinating property of information density (underlying information stability) of Gaussian processes is that its variance is uniformly bounded regardless of the distribution, see (Polyanskiy and Wu, 2024, (19.32)). We will exploit this below to show "local" subgaussian estimate on the concentration of $i(W;Y)$.

Define log-MGF function

$$f(z) := \frac{1}{n} \ln \mathbb{E}_{W,Z}\left[e^{z\sum_{i=1}^n W_i^2 - Z_i^2}\right] = \frac{1}{n}\sum_{i=1}^n \ln \mathbb{E}_{W_i,Z_i}\left[e^{z(W_i^2-Z_i^2)}\right] .$$

It is not hard to show that there exists a neighborhood $\mathcal{S}$ of 0 on the complex plane $\mathbb{C}$ and a constant c such that

$$\sup_{z\in\mathcal{S}} |f(z)| \le c\,,$$

and crucially $\mathcal{S}$ and c can be chosen independent of $\{D_i\}$ (and hence of t, λ_i). For this, one only needs to apply Cauchy-Scwharz to reduce to analysis of log-MGF of W_i^2 and Z_i^2, which are just squares of $\mathcal{N}(0,1)$.

In addition to being analytic, f also satisfies $f(0) = f'(0) = 0$ and $f''(0) \le 2\mathbb{E}[W_i^4 + Z_i^4] \le 12$. Thus, by Cauchy formula we can uniformly bound $f''(z)$ inside any compact subset of $\mathcal{S}$. Consequently, there exists a universal $c' > 0$ and $z_0 > 0$ such that for all real $-z_0 < z < z_0$ we have

$$f(z) \le 2c'z^2\,.$$

Applying Chernoff estimate we find that for all $\epsilon < \epsilon_0$ we have

$$\mathbb{P}[i(W;Y) > I(W;Y) + n\epsilon] \le e^{-2c_2 n\epsilon^2}\,, \tag{38}$$

where crucially ϵ_0, c_2 are absolute constants.

We are now ready to apply standard finite blocklength rate-distortion upper bound (Polyanskiy and Wu, 2024, Theorem 25.2), which claims existence of codebook $\mathbf{C}$ with distortion

$$\mathbb{E}_W\left[\min_{c\in\mathbf{C}} d_{\Sigma_X}(W,c)\right] \le \mathbb{E}[d_{\Sigma_X}(W,Y)] + \mathbb{E}[d_{\Sigma_X}(W,0)]e^{-2^{nR}/\gamma} + \mathbb{E}[d_{\Sigma_X}(W,Y)1\{i(W;Y) > \log_2\gamma\}]\,, \tag{39}$$

where γ is arbitrary, but we set it to

$$\log_2\gamma = nR_0 + n\epsilon\,.$$

Applying Cauchy-Schwarz and (38) to the last term in (39) we obtain (37).

Appendix G. Additive rate-distortion for quantization of a colored X

It is worth mentioning that the rate-distortion region (RDRC) we obtained also characterizes the tradeoff between rate and distortion of a particular quantization scheme in a different, but closely related setup.

In particular, let $X \sim \mathcal{N}(0, \Sigma_X)$ be a Gaussian vector in $\mathbb{R}^n$ to be quantized under the standard quadratic distortion measure $D = \frac{1}{n}\mathbb{E}\|\hat{X} - X\|_2^2$. Clearly the optimal rate-distortion tradeoff for this problem is given by

$$R(D) = \frac{1}{n} \min I(X; \hat{X}) \tag{40}$$

where the minimum is over all $P_{\hat{X}|X}$ for which $\frac{1}{n}\mathbb{E}\|\hat{X} - X\|_2^2 \leq D$. As we already discussed, the optimal $P_{\hat{X}|X}$ is determined by the reverse waterfilling solution, and is given precisely by (WF). In fact, we derived the oracle lower bound by showing that in the oracle setup, where the decoder also knows Σ_X, our problem is equivalent to that of quantizing X under standard quadratic loss.

While the waterfilling solution gives the optimal tradeoff $R^*(D)$ function, any other valid choice of $P_{\hat{X}|X}$ yields an achievable $R(D)$. A very simple choice is to construct $\hat{X}$ by first adding independent noise $Z \sim \mathcal{N}(0, \frac{1}{T} I_n)$ to X and then performing minimum mean squared error (MMSE) estimation of X from $X+Z$. The $R(D)$ tradeoff attained by this particular choice of $P_{\hat{X}|X}$ has been studied in the information theory literature (for sources that are not necessarily Gaussian) under the name *additive rate-distortion function (ARDF)* Zamir (2002); Zamir and Berger (2002); Ostergaard and Zamir (2011).

The MMSE estimator of X from $X + Z$ is linear and therefore $\hat{X} = F(X + Z)$, where

$$F = \Sigma_X \left(\Sigma_X + \frac{1}{T} I_n \right)^{-1} = U \cdot \operatorname{diag}\left(\frac{\lambda_1 T}{1 + \lambda_1 T}, \ldots, \frac{\lambda_n T}{1 + \lambda_n T} \right) \cdot U^\top \tag{41}$$

and

$$\frac{1}{n}\mathbb{E}\|\hat{X} - X\|_2^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{\lambda_i}{1 + \lambda_i T} = D_{\mathrm{rc}}(\lambda, T). \tag{42}$$

Furthermore, if all singular values of Σ_X are positive, F is invertible and

$$\frac{1}{n} I(X; \hat{X}) = \frac{1}{n} I(X; X + Z) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \log(1 + \lambda_i T) = R_{\mathrm{rc}}(\lambda, T). \tag{43}$$

Thus, the Σ_X universal rate distortion tradeoff we derived for the problem of quantizing a white source $\mathcal{N}(0, I_n)$ under d_{Σ_X} metric known only to the encoder is precisely the additive rate-distortion function for quantizing $X \sim \mathcal{N}(0, \Sigma_X)$ under standard quadratic loss.