Gaussian Approximation of Quantization Error for Estimation From Compressed Data

Alon Kipnis, *Member, IEEE*, and Galen Reeves, *Member, IEEE*

***Abstract*— We consider the distributional connection between the lossy compressed representation of a high-dimensional signal X using a random spherical code and the observation of X under an additive white Gaussian noise (AWGN). We show that the Wasserstein distance between a bitrate-R compressed version of X and its observation under an AWGN-channel of signal-to-noise ratio $2^{2R} - 1$ is bounded in the problem dimension. We utilize this fact to connect the risk of an estimator based on the compressed version of X to the risk attained by the same estimator when fed the AWGN-corrupted version of X. We demonstrate the usefulness of this connection by deriving various novel results for inference problems under compression constraints, including minimax estimation, sparse regression, compressed sensing, and universality of linear estimation in remote source coding.**

***Index Terms*— Lossy source coding, spherical coding, Gaussian noise, parameter estimation, indirect source coding, sparse regression, approximate message passing.**

I. Introduction

DUE to the disproportionate size of modern datasets compared to available computing and communication resources, many inference techniques are applied to a compressed representation of the data rather than the data itself (Figure 1). In the attempt to develop and analyze inference techniques based on a degraded version of the data, it is conceptually appealing to model inaccuracies resulting from lossy compression as additive noise. Indeed, there exists a rich literature devoted to the characterization of this "noise", i.e., the difference between the original data and its compressed representation [2]. Nevertheless, because of the difficulty of analyzing non-linear compression operations, this characterization is generally limited to the high-bit compression regime and other restrictions on the distribution of the data [3]–[7].

Manuscript received March 11, 2020; revised January 22, 2021; accepted May 19, 2021. Date of publication May 24, 2021; date of current version July 14, 2021. The work of Alon Kipnis was supported in part by the Koret Foundation and in part by NSF under Grant DMS-1816114. The work of Galen Reeves was supported in part by NSF under Grant 1750362. This article was presented in part at the 2019 IEEE International Symposium on Information Theory (ISIT) [1]. *(Corresponding author: Alon Kipnis.)*

Alon Kipnis is with the Department of Statistics, Stanford University, Stanford, CA 94305 USA (e-mail: kipnisal@stanford.edu).

Galen Reeves is with the Department of Electrical and Computer Engineering and the Department of Statistical Science, Duke University, Durham, NC 27708 USA.

Communicated by R. Venkataramanan, Associate Editor for Machine Learning.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2021.3083271.

Digital Object Identifier 10.1109/TIT.2021.3083271

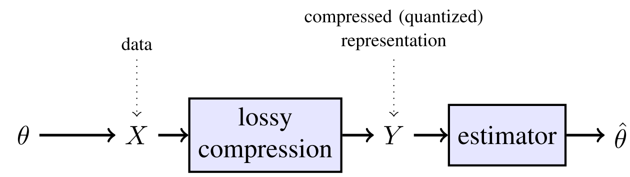

Fig. 1. Inference about the latent signal θ is based on degraded observations Y of the data X.

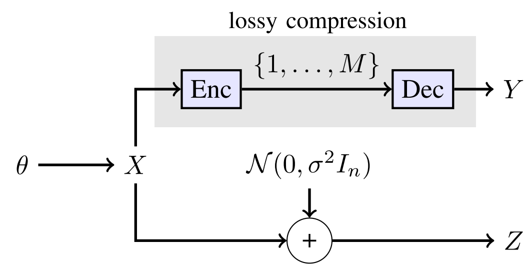

Fig. 2. The effect of a bitrate constraint is compared to the effect of additive Gaussian noise by studying the Wasserstein distance between P_Y and P_Z. Under random spherical encoding, we show that this distance is bounded in the problem dimension n, hence estimating θ from Y is equivalent to estimating it from Z.

In this paper, we establish a strong and relatively simple characterization of the distribution of quantization error corresponding to a random spherical code. Specifically, we show that, in the sense of the Wasserstein distance, this error can be approximated by additive white Gaussian noise (AWGN) whose variance σ^2 is inversely proportional to $2^{2R} - 1$ where R is the bitrate of the code (Figure 2). This approximation implies that the expected error of an estimator applied to the compressed representation of the data is asymptotically equivalent to the expected error of the same estimator applied to a Gaussian noise-corrupted version of the data. The benefit from such approximation is twofold: (1) inference techniques from observations corrupted by Gaussian noise can now be applied directly to the compressed representation; and (2) it provides a mechanism to characterize the performance of inference using such techniques.

A. Overview of Main Contributions

The equivalence illustrated in Figure 2 allows us to derive various novel results for two closely related inference settings, both of which are performed on a lossy compressed representation of the observed data $X = (X_1, \dots, X_n)$.

- *Parameter Estimation:* (Section III) The data are drawn according to a distribution indexed by an unknown d-dimensional parameter vector θ and the goal is to estimate the parameter vector under the squared error loss. In the high-dimensional setting, the number of parameters d is possibly much larger than the number of observations n. This problem is also related to learning distributions under communication constraints [8]–[17].
- *Indirect Source Coding:* (Section IV) The data are distributed jointly with an unknown random (source) vector $U = (U_1, \ldots, U_n)$ and the goal is to reconstruct this vector from the compressed representation of X [18]–[21].

At a high level, the main difference between these inference tasks is that the source coding problem assumes a joint distribution over the data and the quantities of interest. Beyond these settings, one may also consider minimax and universal source coding formulation [22], [23] as well as hypotheses testing [8], [10].

In the parameter estimation setting, we consider the minimax mean-squared error (MSE)

$$\mathcal{M}_n^* := \inf_{\phi,\psi} \sup_{\theta \in \Theta_n} \frac{1}{d_n} \mathbb{E}\left[\|\theta - \psi(\phi(X))\|^2\right],$$

where the infimum is over all encoding functions $\phi\colon \mathbb{R}^n \to \{1, \ldots, M\}$ and decoding functions $\psi\colon \{1, \ldots, M\} \to \mathbb{R}^d$ with $M = \lceil 2^{nR} \rceil$. Zhu and Lafferty [24] provided an asymptotic expression for $\mathcal{M}_n^*$ in the special case of the Gaussian location model $X \sim \mathcal{N}(\theta, \epsilon^2 \; I_n)$ where the parameter space Θ_n is an n-dimensional ball. Under a similar setting, our main results yield a non-asymptotic upper bound to $\mathcal{M}_n^*$. Furthermore, under the additional assumption that θ is k-sparse, our main results implies that $\mathcal{M}_n^*$ is upper bounded by a univariate function describing the minimax risk of soft-thresholding in sparse estimation [25], [26]. Finally, we consider the case where the data X and the parameter θ are described by the model $X \sim \mathcal{N}(A\theta, \epsilon^2 \; I)$, where $A \in \mathbb{R}^{d \times n}$ is a random matrix with i.i.d. Gaussian entries. This setting with θ sparse and d_n much larger than n was studied in the context of the compressed sensing signal acquisition frameworks [27]. By applying our main results to estimation with the approximate message passing (AMP) algorithm [28], we provide an exact asymptotic characterization of the MSE in recovering θ from a lossy compressed version of X obtained using bitrate-R random spherical coding. Versions of this compression and estimation problem for other type of lossy compression codes and estimators were considered in [29]–[32].

The indirect source coding setting (other names are *remote* or *noisy* source coding and *rate-constrained denoising*) corresponds to the case where $\{(U_i, X_i)\}_{i=1}^\infty$ is an ergodic process with a finite second moment. A bitrate-R spherical code is applied to $X = \{X_i\}_{i=1}^n$ while the goal is to estimate $\{U_n\}_{i=1}^n$ from the output Y of this code [18], [20] [19, Ch 3.5] [21]. For data normalized as $\mathbb{E}\left[\|X\|^2\right] = n$, our main results imply that the MSE attained by any sequence of Lipschtiz estimators converges to the MSE attained by these estimators when applied to $\{Z_i\}_{i=1}^n$, where

$$Z_i = X_i + \sigma W_i, \quad \sigma^2 = \frac{1}{2^{2R} - 1}, \tag{1}$$

Specialized to the case $X_i \mid U_i \sim \mathcal{N}(U_i, \epsilon^2)$, our result implies an interesting universality property of spherical coding followed by linear estimation: the resulting MSE equals the minimal MSE, over all encoding and estimation schemes, when a Gaussian source of the same second moment is estimated from a bitrate-R encoded version of its observations under AWGN. This fact can be seen as a direct extension of the saddle point property of the Gaussian distribution in the standard (direct) source coding setting discussed in [33]–[36].

B. Background and Related Works

Spherical codes have multiple theoretical and practical uses in numerous fields [37]. In the context of information theory, Sakrison [33] and Wyner [34] provided a geometric understanding of random spherical coding in a Gaussian setting; our main result extends their insights. Specifically, consider the representation of an n-dimensional standard Gaussian vector X using $M = \lceil 2^{nR} \rceil$ codewords uniformly distributed over the sphere of radius $r = \sqrt{n(1 - 2^{-2R})}$. The left side of Fig. 3, adapted from [33], shows a conceptual relation between X and its nearest codeword $\hat{X}$: As n increases, the angle α^* between the two concentrates so that $\sin(\alpha^*)$ converges to 2^{-R} in probability, hence the quantized representation of X and the error $X - \hat{X}$ become orthogonal. Consequently, the MSE between X and its quantized representation, averaged over all random codebooks, converges to the Gaussian distortion-rate function 2^{-2R}. In fact, as noted in [35], this Gaussian coding scheme[1] achieves the Gaussian DRF when X is generated by any ergodic information source of unit variance, implying that the second moments of $X - \hat{X}$ are independent of the distribution of X as the problem dimension n goes to infinity.

In this paper, we show that a much stronger statement holds for a properly scaled version of the quantized representation (Y in Fig. 3): in the limit of high dimension, the distribution of $Y - X$ is independent of the distribution of X and is approximately Gaussian. This property of $Y - X$ suggests that the underlying quantities of interest (e.g., the parameter vector θ or the sequence $\{U_n\}_{n=1}^\infty$) can now be estimated as if X is observed under additive Gaussian noise. This paper formalizes this intuition by showing that estimators from the Gaussian-noise corrupted version of X (Z in Fig. 3) attain similar performances if applied to the scaled representation Y.

In general, the radius of the codebook under which the distribution of Y and Z are close depends on the magnitude of X. This magnitude is only needed at the decoder, while the encoder can represent its input X using codewords living, say, on the unit sphere. In particular, such an encoder is agnostic to the relationship between X and θ. This situation is in contrast to optimal quantization schemes in indirect source coding [18], [19] and in problems involving estimation from compressed

[1] We denote this scheme as *Gaussian* since r is chosen according to the distribution attaining the Gaussian DRF [38, Ch. 10.5].

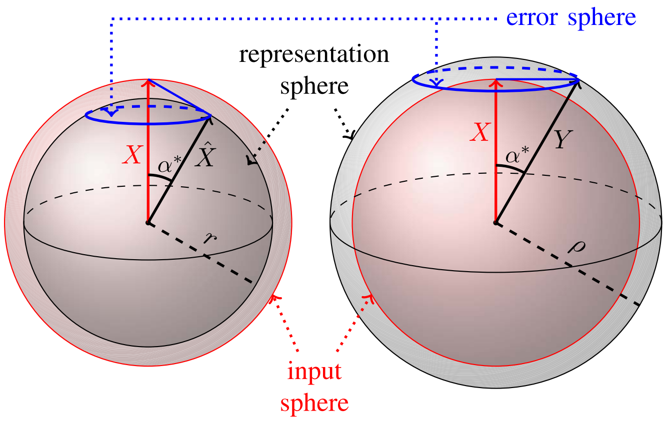

Fig. 3. Conceptual 3D illustration of random spherical coding in high-dimension. The norm of the input vector X concentrates around the input sphere. Left Sphere: Geometric interpretation of standard source coding from [33], [34]. The representation sphere is chosen such that the error vector $\hat{X} - X$ is orthogonal to the reconstruction vector $\hat{X}$. Right Sphere: Geometric description of the quantization error considered in this paper. The representation sphere is chosen such that $Y - X$ is orthogonal to X.

data [9], [10], [13], [14], [39], [40], where the specification of the model $\theta \to X$ is crucial for designing the compression and estimation schemes. As a result, the random spherical coding scheme we rely on is sub-optimal in general, although it can be applied in situations where the model $\theta \to X$ is unknown at the compressor. Coding schemes with similar properties were studied in the context of indirect source coding under the name *compress-and-estimate* in [41], [42].

The equivalence between quantization noise and AWGN we provide in this paper is given in terms of the Wasserstein distance between the distributions of these vectors. We refer to [43]–[45] for properties, applications, and the long list of alternative names of the Wasserstein distance. In the context of information theory, the Wasserstein distance has been used to establish consistency of some quantization procedures [46], [47] and to define a class of channels over which communication is possible without assuming synchronization [48], [49]. One of the core results of this paper is a novel coupling of the distributions of Y and Z given X, leading to a bound on the Wasserstein distance between them. This bound, in combination with the fact that the L_p risk of a Lipschitz estimator is continuous with respect to the Wasserstein distance, implies that the risk of such an estimator, when used at the output of a random spherical code, converges to the risk when used at the output of a Gaussian channel.

Our work is similar in spirit to the work of Zamir and Feder [50, Section III], who provide a Gaussian approximation for the quantization error of a random dithered lattice quantizer. In the setting of their paper, the quantization error is independent of the input and distributed uniformly over the basic cell of the lattice. They show that as the dimension n increases, there exists a sequence of lattice quantizers such that the relative entropy between the distribution of the quantization noise and the isotropic Gaussian distribution with matched power is bounded from above by $c \cdot \log(n)$ where c is a positive constant. Normalizing by n, they conclude that the relative entropy per dimension converges to zero in the large-n limit. To interpret their results in the setting of this paper, we can use the Gaussian transportation inequality [51], which leads to an upper bound on the 2-Wasserstein distance that is order $\sigma\sqrt{\log n}$ where σ^2 is the variance of the additive noise. By contrast, for quantization using a random spherical code, our results provide an upper bound on p-Wasserstein distance that is order σp. Namely, both bounds are proportional to σ but the bound following from [50] is unbounded in the problem dimension.

Another related setting is the problem of channel simulation, the goal of which is to design a random code that induces a particular target distribution between the data and the compressed representation [52]–[54].

The rest of this paper is organized as follows. In Section II we provide our main results on the distributional connection between spherical coding and AWGN. In Sections III and IV we apply these results to parameter estimation and source coding, respectively. Section V provides the proofs of the main results. Concluding remark are provided in Section VI.

II. Main Results

The main result of this paper is a comparison between the quantization error under random spherical coding and independent Gaussian noise.

Definition 1 (Random Spherical Code): An (n, M) random spherical code is a collection of M codewords $\mathcal{C} = \{C(1), \dots, C(M)\}$ drawn independently from the uniform distribution on the unit sphere in $\mathbb{R}^n$. The encoder maps an input vector $x \in \mathbb{R}^n \setminus \{0\}$ to the index $i^* \in \{1, \dots, M\}$ of a codeword that maximizes the cosine similarity

$$i^* \in \arg\max_{1 \le i \le M} \langle C(i), x \rangle. \tag{2}$$

Given the index i^* and knowledge of the codebook $\mathcal{C}$, the decoder outputs the compressed representation

$$Y := \rho\, C(i^*), \tag{3}$$

where $\rho \ge 0$ is a scaling parameter.

Our results focus on the distribution of the compressed representation Y induced by the randomness in the codebook. Note that this distribution is parameterized by the input x and has a density with respect to the surface measure on the sphere of radius ρ. We also define the maximal cosine similarity according to

$$\cos(\alpha^*) := \frac{\langle x, C(i^*) \rangle}{\|x\|}, \qquad \alpha^* \in [0, \pi]. \tag{4}$$

It is well known (see e.g., [55]) that the distribution of α^* does not depend on x and is given by

$$\mathbb{P}\left[(\cos(\alpha^*) \le s)\right] = \left(1 - Q_n(s)\right)^M, \tag{5}$$

where

$$Q_n(s) := \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi}\Gamma(\frac{n}{2})} \int_s^1 (1 - t^2)^{\frac{n-3}{2}}\, dt, \tag{6}$$

and where $\Gamma(z) = \int_0^\infty x^{z-1} \exp(-z)\, dz$ is the Gamma function.

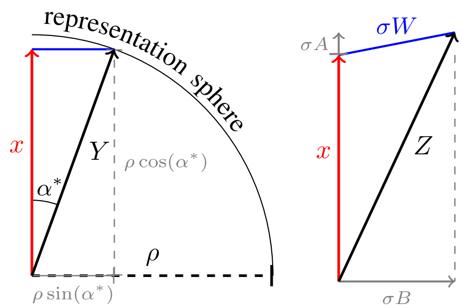

Fig. 4. Conceptual 2D description of the coupling in Theorem 1 when the representation sphere is matched to $\|x\|$. The quantization error $Y - x$ (Left) is compared to the standard n-dimensional Gaussian noise vector W (Right). The random variable Y is the nearest codeword to x in the random codebook ensemble. The standard Gaussian random variable A is the normalized component of W in the direction of x. The random variable B describes the magnitude of the projection of W onto the $(n-1)$-dimensional space orthogonal to x.

A. Approximation Using AWGN

The fundamental question we address is the extent to which the quantization error $Y - x$ can be approximated by an isotropic zero-mean Gaussian noise. To answer this question we introduce the AWGN-corrupted observation model

$$Z = x + \sigma W, \qquad W \sim \mathcal{N}(0, I_n). \tag{7}$$

Our main results are based on a coupling argument. Specifically, we show that there exists a joint distribution on the pair (Y, Z) under which the distribution of $\|Y - Z\|$ can be described exactly in terms of the tuple (ρ, σ, n, M) and the magnitude of the input. The proof of the following result is in Section V.

Theorem 1: For any $x \in \mathbb{R}^n \backslash \{0\}$, positive integer M, and real numbers $\rho, \sigma > 0$, there exists a joint distribution on (Y, Z) such that Y has the distribution of the compressed representation of magnitude ρ obtained from an (n, M) random spherical code with input x, $Z \sim \mathcal{N}(x, \sigma^2)$, and

$$\|Y - Z\|^2 = (\rho\cos(\alpha^*) - \|x\| - \sigma A)^2 + (\rho\sin(\alpha^*) - \sigma B)^2, \tag{8}$$

where:

- A, B, α^* are independent,
- $A \sim \mathcal{N}(0, 1)$,
- B has a chi distribution with $n - 1$ degrees of freedom.

Figure 4 provides a conceptual illustration of α^*, A, and B in the comparison between Y and Z provided in Theorem 1. The proof of this theorem in Section V provides the exact description of these random variables and vectors.

The coupling described in Theorem 1 holds for any choice of the parameters (ρ, σ). The next step is to show that the term $\|Y - Z\|$ in (8) is negligible compared to the magnitude of the quantization error under a proper specification of these parameters. To provide a sense of scale, observe that the error due to AWGN satisfies $\|Z - x\| = \sigma\|W\|$ where $\|W\|$ concentrates about $\sqrt{n}$ in the large-n limit. For comparison, we consider the upper bound given by

$$\|Y - Z\| \le |\|x\| - \gamma| + \sigma\Delta \tag{9}$$

where γ is an estimate of $\|x\|$ and the normalized error term

$$\Delta := \frac{1}{\sigma}\sqrt{(\rho\cos(\alpha^*) - \gamma - \sigma A)^2 + (\rho\sin(\alpha^*) - \sigma B)^2} \tag{10}$$

does not depend on x. In the following we show that (ρ, σ) can be chosen as a function of (n, M, γ) such that the distribution of Δ is bounded independently of the dimension n.

First we consider the setting where the number of codewords is given by $M = \lceil 2^{nR} \rceil$ for a fixed bitrate $R > 0$. For a given $\gamma \ge 0$ we use the specification

$$\rho = \frac{\gamma}{\sqrt{1 - 2^{-2R}}}, \qquad \sigma = \frac{\gamma}{\sqrt{n}\sqrt{2^{2R} - 1}}. \tag{11}$$

Theorem 2: Suppose that $M = \lceil 2^{nR} \rceil$ for a fixed bitrate $R > 0$. Under the specification given in (11), the normalized error term Δ defined in (10) has a sub-Gaussian distribution with parameters that depend only on the bitrate R. In particular, there exists a positive number C_R such that

$$\mathbb{E}\left[\Delta^p\right]^{1/p} \le C_R\sqrt{p}, \qquad p \ge 1. \tag{12}$$

The significance of Theorem 2 is that the distribution of Δ is bounded uniformly with respect to n and thus the term $\sigma\Delta$ in (9) is order one. By comparison, the magnitude of the additive noise $\sigma\|W\|$ scales at rate $n^{1/2}$. This means that if the estimate of $\|x\|$ is accurate in the sense that $\|x\|/\gamma \to 1$ as $n \to \infty$, then the relative difference between the mismatch $\|Y - Z\|$ and the quantization error $\|Y - x\|$ converges to zero.

Next, we provide a result that holds in the high-rate setting where $\frac{1}{n}\log M$ diverges. This regime requires a more precise estimate of the max-cosine similarity and we use the specification

$$\tilde{\rho} = \frac{\gamma}{\sqrt{1 - 2^{-2\tilde{R}}}}, \quad \tilde{\sigma} = \frac{\gamma}{\sqrt{n}\sqrt{2^{2\tilde{R}} - 1}}, \quad \tilde{R} = \frac{1}{n-1}\log_2 M. \tag{13}$$

Note that $\tilde{R}$ is normalized by $(n-1)$ instead of n. Further, we will require that the number of codewords is bounded from below by

$$M_\beta(n) := \sqrt{n}(\csc\beta)^{n-1}, \tag{14}$$

for some fixed constant $\beta \in (0, \pi/2)$.

Theorem 3: Suppose that $M \ge M_\beta(n)$ for some constant $\beta \in (0, \pi/2)$. Under the specification given in (13), the normalized error term Δ defined in (10) has a sub-exponential distribution with parameters that depend only on β. In particular, there exists a positive number C_β such that

$$\mathbb{E}\left[\Delta^p\right]^{1/p} \le C_\beta\, p, \qquad p \ge 1. \tag{15}$$

Theorem 3 is stronger than Theorem 2 in the sense that the bound holds uniformly for all (n, M) satisfying the constraint $M \ge M_\beta(n)$. This is important for the high-bitrate setting where σ converges to zero. The price that is paid is that the sub-Gaussian tail condition is replaced with the weaker sub-exponential condition. An explicit value for the constant C_β and its dependence on β can be found in the proof of Theorem 3, which is given in Appendix A.

B. Bounds on Wasserstein Distance

Our results can also be stated in terms of the Wasserstein distance on distributions. The p-Wasserstein distance between distributions P and Q on $\mathbb{R}^n$ is defined by

$$W_p(P,Q) := \inf \left(\mathbb{E}\left[\|U-V\|^p\right]\right)^{1/p},$$

where the infimum is over all joint distributions on (U,V) satisfying the marginal constraints $U \sim P$ and $V \sim Q$. For $p \geq 1$, the p-Wasserstein distance is a metric on the space of distributions with finite p-th moments.

Theorems 2 and 3 imply upper bounds on the Wasserstein distance between the distribution of the compressed representation obtained using a random spherical code and the distribution of the AWGN-corrupted version of the input.

Theorem 4: Let X be a random vector in $\mathbb{R}^n$ with $\mathbb{E}\left[\|X\|^p\right] < \infty$ for some $p \geq 1$. Let P_Y be the distribution of the compressed representation of magnitude ρ obtained from an (n, M) random spherical code with input X and let P_Z be the distribution of $Z = X + \sigma W$ where W is an independent standard Gaussian vector.

(i) If (ρ, σ) are defined as in (11), then

$$W_p(P_Y, P_Z) \leq \left(\mathbb{E}\left[|\|X\| - \gamma|^p\right]\right)^{1/p} + \sigma C_R \sqrt{p}, \quad (16)$$

where C_R is a positive number that depends only on R.

(ii) If $M \geq M_n(\beta)$ for some $\beta \in (0, \pi/2)$ and (ρ, σ) are defined as in (13), then

$$W_p(P_Y, P_Z) \leq \left(\mathbb{E}\left[|\|X\| - \gamma|^p\right]\right)^{1/p} + \sigma C_\beta p, \quad (17)$$

where C_β is a positive number that depends only on β.

Proof: The p-th power of the Wasserstein distance is convex in the pair (P,Q) [45, Theorem 4.8], and thus

$$W_p^p(P_Y, P_Z) \leq \int W_p^p(P_{Y|X=x}, P_{Z|X=x})\, \mathrm{d}P_X(x), \quad (18)$$

where $P_{Y|X=x}$ and $P_{Z|X=x}$ denote the conditional distributions of Y and Z, respectively. In view of (9) and (12), it follows that

$$W_p(P_{Y|X=x}, P_{Z|X=x}) \leq |\|x\| - \gamma| + \sigma C_R \sqrt{p}. \quad (19)$$

Combining these displays with Minkowski's inequality leads to (16). Inequality (17) is obtained in a similar manner from (14). □

A useful property of the Wasserstein distance is that it controls the expectations of Lipschitz continuous functions. Recall that a mapping $f : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous if there exists a constant L such that

$$\|f(u) - f(v)\| \leq L\|u - v\|, \quad \text{for all } u, v \in \mathbb{R}^n. \quad (20)$$

The infimum over all L is call the Lipschitz constant and is denoted by $\|f\|_{\mathrm{Lip}}$. The 1-Wasserstein distance, which is also known as the Kantorovich-Rubinstein distance, can be expressed equivalently as

$$W_1(P,Q) = \sup\left\{\mathbb{E}\left[f(U)\right] - \mathbb{E}\left[f(V)\right] \mid \|f\|_{\mathrm{Lip}} \leq 1\right\}, \quad (21)$$

where $U \sim P$ and $V \sim Q$. More generally, the p-Wasserstein can be used to bound the difference between p-th moments.

Proposition 5: Let $U \sim P$ and $V \sim Q$ be random vectors on $\mathbb{R}^n$. For any Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}^m$,

$$\left|\left(\mathbb{E}\left[\|f(U)\|^p\right]\right)^{1/p} - \left(\mathbb{E}\left[\|f(V)\|^p\right]\right)^{1/p}\right| \leq \|f\|_{\mathrm{Lip}}\, W_p(P,Q), \quad (22)$$

provided that the expectations exist.

Proof: For any coupling of (U,V), Minkowski's inequality and the Lipschitz assumption on f yield

$$\begin{aligned}\left(\mathbb{E}\left[\|f(U)\|^p\right]\right)^{1/p} &\leq \left(\mathbb{E}\left[\|f(V)\|^p\right]\right)^{1/p} + \left(\mathbb{E}\left[\|f(U) - f(V)\|^p\right]\right)^{1/p} \\ &\leq \left(\mathbb{E}\left[\|f(V)\|^p\right]\right)^{1/p} + \|f\|_{\mathrm{Lip}} \left(\mathbb{E}\left[\|U - V\|^p\right]\right)^{1/p}.\end{aligned}$$

Taking the infimum over all possible couplings leads to one side of the inequality. Interchanging the role of U and V and repeating the same steps gives the other side. □

C. Concentration of the Norm

The bounds in Theorems 2 and 3 simplify further when the parameter γ in (10) is matched to the magnitude of the input. As a specific example, suppose that the data is known to lie on the sphere of radius $\sqrt{n}$ and that $M = \lceil 2^{nR} \rceil$. By setting $\gamma = \sqrt{n}$, we see that there exists a coupling of Y and Z under which the quantization error $\|Y - Z\|$ is bounded by a sub-Gaussian random variables that is independent of n.

For many applications, however, the assumption that the data lay on a sphere of known radius is too restrictive. Therefore, in this paper, we assume that the data at the input to the compressor is a random vector X in $\mathbb{R}^n$ whose magnitude $\|X\|$ concentrates about a known value γ. This assumption is reasonable for high-dimensional settings where the entries of X are weakly correlated. More generally, there are a number of other approaches that can be used to deal with the fact that $\|X\|$ is unknown. One approach is to use additional bits to encode the magnitude of X, as is done in [24]. For example, if $\|X\| \leq \kappa\sqrt{n}$ almost surely where κ is a known constant, then $\log_2 \sqrt{n}$ bits are sufficient to encode $\|X\|$ with absolute error less than κ, such that

$$\left(\mathbb{E}\left[|\|X\| - \gamma|^p\right]\right)^{1/p} \leq \kappa.$$

When n is large, the logarithmic number of bits used to encode the magnitude of X is negligible compared to the nR bits used to encode its direction. An alternative approach is to compare the compressed representation with a noisy version of X after it has been projected onto the unit sphere in $\mathbb{R}^n$. This can be achieved, by setting $\gamma = 1$ and redefining the input to be $\tilde{X} = X/\|X\|$ such that the magnitude is equal to one almost surely. In both of the approaches described above, the noise variance σ^2 is scaled in such a way that the signal-to-noise ratio in the AWGN observation model (7) depends only on (n, R) and is given by $(\rho/\sigma)^2$. One may also consider a variable-length coding strategy that adapts the number of bits to the magnitude of X such that the effective noise power is constant and the signal-to-noise ratio is proportional to $\|X\|^2$. We leave this as a direction for future work.

III. Application to Parameter Estimation

In this section, we apply our main results to the problem of estimating an unknown parameter vector θ from a compressed representation of the data X. For each integer n, let $\mathcal{P}_n = \{P_{n,\theta} \,:\, \theta \in \Theta_n\}$ be a family of distributions on $\mathbb{R}^n$ with index set $\Theta_n \subseteq \mathbb{R}^{d_n}$. For the purposes of exposition we will focus on the squared error loss. Our approach is quite general, however, and can be extended to other loss functions.

An important performance benchmark in estimating θ from a bitrate-R compressed representation of X is the minimax MSE:

$$\mathcal{M}_n^* := \inf_{\phi,\psi} \sup_{\theta\in\Theta_n} \frac{1}{d_n}\mathbb{E}_{P_{n,\theta}}\left[\|\theta - \psi(\phi(X))\|^2\right], \qquad (23)$$

where the minimum is over all encoding functions $\phi : \mathbb{R}^n \to \{1, \dots, M\}$ and decoding functions $\psi : \{1, \dots, M\} \to \mathbb{R}^d$ with $M = \lceil 2^{nR} \rceil$.

Zhu and Lafferty [24] studied the asymptotic minimax MSE for the Gaussian location model $X \sim \mathcal{N}(\theta, \epsilon^2\, I_n)$ with Θ_n the n-dimensional Euclidean ball of radius $\kappa\sqrt{n}$, and showed that

$$\limsup_{n\to\infty} \mathcal{M}_n^* = \frac{\kappa^2\epsilon^2}{\kappa^2+\epsilon^2} + \frac{\kappa^4}{\epsilon^2+\kappa^2}2^{-2R}. \qquad (24)$$

Their achievability result is based on random spherical coding while devoting a number of bits sublinear in n to encode the magnitude of the X, as discussed in Section II-C above.

The comparison between quantization error and Gaussian noise in Theorem 4 provides a straightforward method for obtaining non-asymptotic upper bounds on the minimax MSE that can be applied to a large class of models. The basic idea is to study the MSE of Lipschitz estimators applied to the AWGN-corrupted data. We use the following assumption, which says that $P_{n,\theta}$ concentrates on a spherical shell whose radius does not depend on θ.

Assumption 1 (Concentration of Magnitude): There exists a sequence of positive numbers $\{(\gamma_n, \tau_n)\}_{n\in\mathbb{N}}$ such that

$$\sup_{\theta\in\Theta_n} \mathbb{E}_{P_{n,\theta}}\left[\left|\|X\| - \gamma_n\right|^2\right] \le \tau_n^2 \qquad (25)$$

Assumption 1 provides a way to formulate many cases of interest in terms of the radius of the shell γ_n and its width τ_n.

The next result uses this assumption to bound the difference in root MSE between an estimator applied to the compressed representation Y and the same estimator applied to the AWGN-corrupted version Z.

Theorem 6: Let $\{P_{n,\theta}\}_{n\in\mathbb{N}}$ be a sequence of models that satisfies Assumption 1. Given $X \sim P_{n,\theta}$, let Y be the output of an $(n, \lceil 2^{nR} \rceil)$ random spherical code with input X and output scaling ρ and let $Z = X + \sigma W$ where $W \sim \mathcal{N}(0, I_n)$ is independent of X and ρ is according to (11). For any Lipschitz estimator $\hat\theta : \mathbb{R}^n \to \mathbb{R}^{d_n}$, the root MSE satisfies

$$\left|\sqrt{\mathbb{E}\left[\|\hat\theta(Y) - \theta\|^2\right]} - \sqrt{\mathbb{E}\left[\|\hat\theta(Z) - \theta\|^2\right]}\right| \le C\,\|\hat\theta\|_{\mathrm{Lip}}\,\beta_n \qquad (26)$$

where C is a constant that depends only on the bitrate R and $\beta_n = \tau_n \vee (\gamma_n/\sqrt{n})$. Furthermore, for all $t > 0$, the minimax MSE satisfies

$$\mathcal{M}_n^* \le \frac{1}{n}\mathbb{E}\left[\|\hat\theta(Y) - \theta\|^2\right] \le (1+t)\frac{1}{d_n}\mathbb{E}\left[\|\hat\theta(Z) - \theta\|^2\right] \qquad (27)$$
$$+ \left(1 + \frac{1}{t}\right) C^2 \frac{\|\hat\theta\|_{\mathrm{Lip}}^2 \beta_n^2}{d_n}.$$

Proof: For each $\theta \in \Theta_n$, let $Q_{n,\theta}$ and $Q'_{n,\theta}$ denote the corresponding distributions of Y and Z. Proposition 5 evaluated with $f(u) = \|\hat\theta(u) - \theta\|$ gives,

$$\left|\left(\mathbb{E}\left[\|\hat\theta(Y) - \theta\|^2\right]\right)^{1/2} - \left(\mathbb{E}\left[\|\hat\theta(Z) - \theta\|^2\right]\right)^{1/2}\right| \le \|\hat\theta\|_{\mathrm{Lip}} W_2(Q_{n,\theta}, Q'_{n,\theta}),$$

where we have used the fact that f is the composition of $\hat\theta$ with the 1-Lipschitz function $\|\cdot - \theta\|$, and thus $\|f\|_{\mathrm{Lip}} = \|\hat\theta\|_{\mathrm{Lip}}$. By Theorem 4 and Assumption 1, the Wasserstein distance is upper bounded by $\tau_n + \sigma\sqrt{2}C_R$, where σ is given in (11). It follows that

$$\frac{1}{n}\mathbb{E}\left[\|\hat\theta(Y) - \theta\|^2\right] \le \left(\frac{1}{d_n}\mathbb{E}\left[\|\hat\theta(Z) - \theta\|^2\right] + C'\frac{\gamma}{\sqrt{n}}\|\hat\theta\|_{\mathrm{Lip}}\right)^2,$$

where C' only depends on R. The upper bound on the minimax MSE follows from the inequality $(a+b)^2 \le a^2(1+t) + b^2(1+1/t)$ for all $t > 0$. □

One takeaway from Theorem 6 is that the MSE obtained from the compressed representation is asymptotically equivalent to that of the AWGN-corrupted observation, provided that the Lipschitz constant of the estimator is small enough. To gain insight into the interplay between the Lipschitz constant of the estimator, the magnitude of the data, and the typical size of the squared error, it is useful to consider some concrete examples.

A. Gaussian Location Model

For our first example, we consider the Gaussian location model $X \sim \mathcal{N}(\theta, \epsilon^2\, I_n)$. Assume that the parameter set Θ_n is a subset of the spherical shell:

$$\mathcal{S}_n := \{\theta \in \mathbb{R}^n \,:\, \kappa\sqrt{n} - \omega_n \le \|\theta\| \le \kappa\sqrt{n} + \omega_n\}. \qquad (28)$$

As an intuition for this notation, one may think about $\sqrt{n}\kappa$ as an estimate for the magnitude of θ and ω_n as the uncertainty in this estimate. For example, if the entries of θ are sampled independently from a sub-Gaussian distribution with a second moment κ^2, then $\|\theta\| - \sqrt{n}\kappa$ is sub-Gaussian [56, Thm 3.1.1]. In this case, there exists a constant C independent of n such that $\theta \in \mathcal{S}_n$ for $\omega_n = C\sqrt{2\log n}$ with probability at least $1 - 1/n$.

The next statement says that under a Gaussian location model, X concentrates whenever θ is restricted.

Proposition 7: Consider the model $X \sim \mathcal{N}(\theta, \epsilon^2\, I_n)$ with $\Theta_n \subseteq \mathcal{S}_n$. Assumption 1 is satisfied with $\gamma_n = \sqrt{n(\kappa^2+\epsilon^2)}$ and $\tau_n = \omega_n + 2\epsilon$.

Proof: Let $\mu = \mathbb{E}\left[\|X\|^2\right] = \|\theta\|^2 + n\epsilon^2$. By the triangle inequality,

$$\sqrt{\mathbb{E}\left[|\|X\| - \gamma_n|^2\right]} \le \sqrt{\mathbb{E}\left[|\|X\| - \sqrt{\mu}|^2\right]} + |\sqrt{\mu} - \gamma_n| . \tag{29}$$

The assumption $\Theta_n \subseteq \mathcal{S}_n$ implies

$$\begin{aligned} |\sqrt{\mu} - \gamma_n| &= |\sqrt{\|\theta\|^2 + n\epsilon^2} - \sqrt{n(\kappa^2 + \epsilon^2)}| \\ &= \left|\|\theta\| - \kappa\sqrt{n}\right| \frac{\|\theta\| + \sqrt{n}\kappa}{\sqrt{\|\theta\|^2 + n\epsilon^2} + \sqrt{n(\kappa^2 + \epsilon^2)}} \\ &\le \left|\|\theta\| - \kappa\sqrt{n}\right| \le \omega_n. \end{aligned}$$

Consequently, the second term on the right-hand side of (29) is upper bounded by ω_n. For the first term, we write

$$\begin{aligned} \mathbb{E}\left[|\|X\| - \sqrt{\mu}|^2\right] &\le \mathbb{E}\left[|\|X\| - \sqrt{\mu}|^2 \left(1 + \frac{\|X\|}{\sqrt{\mu}}\right)^2\right] \\ &= \frac{\mathrm{Var}(\|X\|^2)}{\mu} = \frac{2\epsilon^2(2\|\theta\|^2 + n\epsilon^2)}{\|\theta\|^2 + n\epsilon^2} \le 4\epsilon^2, \end{aligned}$$

where we have used the fact that $\|X\|^2/\epsilon^2$ has a non-central chi-squared distribution with n degrees of freedom and non-centrality parameter $\|\theta\|^2/\epsilon^2$ □

In this setting, the AWGN-corrupted data Z, corresponding to a bitrate R and magnitude γ_n, is drawn according to the Gaussian location model whose noise variance depends on the original noise level ϵ^2 and the bitrate R:

$$Z \sim \mathcal{N}(\theta, (\epsilon^2 + \sigma^2)I_n), \qquad \sigma^2 = \frac{\kappa^2 + \epsilon^2}{2^{2R} - 1}. \tag{30}$$

The MSE in the Gaussian location model has been studied extensively. If we restrict our attention to linear estimators of the form $\hat{\theta}(z) = \lambda z$ then a standard calculation (see e.g., [26, Ch. 4.8]) gives

$$\inf_{\lambda \ge 0} \sup_{\theta \,:\, \|\theta\| \le \kappa\sqrt{n}} \frac{1}{n}\mathbb{E}\left[\|\theta - \lambda Z\|^2\right] = \frac{\kappa^2(\epsilon^2 + \sigma^2)}{\kappa^2 + \epsilon^2 + \sigma^2}, \tag{31}$$

where the minimum over λ is attained at $\lambda^* = k^2/(\kappa^2 + \epsilon^2 + \sigma^2)$. By expressing the right-hand side as a function of R and combining with Theorem 6, we obtain a non-asymptotic upper bound on the minimax MSE.

Proposition 8: Consider the model $X \sim \mathcal{N}(\theta, \epsilon^2\ I_n)$ with $\Theta_n \subseteq \mathcal{S}_n$. Let Y be the output of a bitrate-R random spherical code applied to X and scaled to the radius $\sqrt{n(\kappa^2 + \epsilon^2)/(1 - 2^{-2R})}$. Then

$$\frac{1}{n}\mathbb{E}\left[\|\theta - \lambda^* Y\|^2\right] \le \frac{\kappa^2\epsilon^2}{\kappa^2 + \epsilon^2} + \frac{\kappa^4}{\epsilon^2 + \kappa^2}2^{-2R} + C\,\frac{1 \vee \omega_n}{\sqrt{n}}, \tag{32}$$

where C is a constant that depends on (κ, ϵ, R) but not n.

Proof: We have $\gamma_n/\sqrt{n} = \sqrt{\kappa^2 + \epsilon^2}$, and $\|\hat{\theta}\|_{\mathrm{Lip}} \le 1$ for the linear estimator $\hat{\theta}(z) = \lambda^* z$. Following Proposition 7, Assumption 1 is satisfied with $\tau_n = \omega_n + 2\epsilon$. We use Theorem 6 with $t = (1 \vee \omega_n)/\sqrt{n}$ and $\beta_n = (\omega_n + 2\epsilon) \vee \sqrt{\kappa^2 + \epsilon^2}$. It follows that there exists a constant c such that

$$\begin{aligned} \frac{1}{n}\mathbb{E}\left[\|\theta - \lambda^* Y\|^2\right] \le& \left(1 + \frac{1 \vee \omega_n}{\sqrt{n}}\right)\left(\frac{\kappa^2\epsilon^2}{\kappa^2 + \epsilon^2} + \frac{\kappa^4}{\epsilon^2 + \kappa^2}2^{-2R}\right) \\ &+ \left(1 + \frac{\sqrt{n}}{1 \vee \omega_n}\right) c^2\, \frac{(\kappa^2 + \epsilon^2) \vee (\omega_n + 2\epsilon)^2}{n}. \end{aligned}$$

Grouping $1/\sqrt{n}$ factors leads to (32). □

Since

$$\mathcal{M}_n^* \le \frac{1}{n}\mathbb{E}\left[\|\theta - \lambda^* Y\|^2\right],$$

Proposition 8 recovers parts of the results in [24] by showing that there exists a bitrate-R coding scheme with minimax risk approaching (24). Furthermore, Proposition 8 shows that the minimax risk under such scheme converges at rate $1/\sqrt{n}$, which is faster than the convergence rate established in [24] by a factor of $1/\sqrt{\log n}$.

More generally, we can also provide bounds for non-linear estimators. The case of a k-sparse parameter vector can be modeled as

$$\Theta_n = \mathcal{S}_n \cap \{\theta \in \mathbb{R}^n \,:\, \|\theta\|_0 \le k\}$$

where $\|\theta\|_0$ denotes the number of nonzero entries in θ. A great deal of work has studied the MSE of the soft-thresholding estimator

$$\hat{\theta}_\lambda(z) := \begin{cases} z - \lambda, & z > \lambda, \\ 0, & |z| \le \lambda, \\ z + \lambda, & z < \lambda. \end{cases} \tag{33}$$

in the model (30) [25], [26]. Specifically, we have

$$\inf_{\lambda \ge 0} \sup_{\theta \,:\, \|\theta\|_0 \le k} \frac{1}{n}\mathbb{E}\left[\|\theta - \hat{\theta}_\lambda(Z)\|^2\right] \le (\epsilon^2 + \sigma^2)\beta_0\left(\frac{k}{n}\right), \tag{34}$$

where, for $\nu > 0$,

$$\beta_0(\nu) = \inf_{\lambda \ge 0} \left\{(1 - \nu)[2(1 + \lambda^2)\Phi(-\lambda) - 2\lambda\phi(\lambda)] + \nu(1 + \lambda^2)\right\} \tag{35}$$

where $\Phi(z)$ and $\phi(z)$ are the cumulative and density functions of the standard Gaussian distribution, respectively. Let λ^* be the minimizer in (34).

Proposition 9: Let $X \sim (\theta, \epsilon^2\ I_n)$ where $\theta \in \mathcal{S}_n \cap \{\theta \in \mathbb{R}^n \ :\ \|\theta\|_0 \le k\}$. Let Y be the output of a bitrate-R random spherical code applied to X and scaled to the radius $\sqrt{n(\kappa^2 + \epsilon^2)/(1 - 2^{-2R})}$. Then

$$\frac{1}{n}\mathbb{E}\left[\left\|\theta - \hat{\theta}_{\lambda^*}(Y)\right\|^2\right] \le \left(\epsilon^2 + \frac{\kappa^2 + \epsilon^2}{2^{2R} - 1}\right)\beta_0\left(\frac{k}{n}\right) + C\frac{1 \vee \omega_n}{\sqrt{n}}, \tag{36}$$

where C is a constant that depends on (κ, ϵ, R) but neither k or n.

Proof: For any $\lambda > 0$ we have $\|\hat{\theta}_\lambda\|_{\mathrm{Lip}} = 1$. Equation (36) follows from Theorem 6 by using $t = (1 \vee \omega_n)/\sqrt{n}$ and grouping $1/\sqrt{n}$ factors. □

Corollary 10: Assume that $X \sim (\theta, \epsilon^2 I_n)$. The bitrate-$R$ constrained minimax risk over $\theta \in \mathcal{S}_n \cap \{\theta \in \mathbb{R}^n : \|\theta\|_0 \leq k\}$, with $\omega_n/\sqrt{n} \to 0$, satisfies

$$\mathcal{M}_n^* \leq \left(\epsilon^2 + \frac{\kappa^2 + \epsilon^2}{2^{2R}-1}\right) \beta_0\left(\frac{k}{n}\right).$$

B. Linear Model With IID Matrix

For the next example, we consider the linear model

$$X \sim \mathcal{N}(A\theta, \epsilon^2 I_n), \tag{37}$$

where A is a known $n \times d$ matrix, θ is an unknown d-dimensional vector, and ϵ^2 a known noise variance. In this setting, the AWGN-corrupted version of X given by $Z = X + \epsilon W$ with $W \sim \mathcal{N}(0, \sigma^2 I_n)$ corresponds to a linear model with larger noise variance, that is

$$Z \sim \mathcal{N}(A\theta, \xi^2 I_n), \qquad \xi^2 = \epsilon^2 + \sigma^2. \tag{38}$$

We study the approximate message passing (AMP) algorithm [28] to estimate θ from Z. AMP is an iterative algorithm that can be defined by a sequence of scalar denoising functions $\{\eta_t\}_{t\geq 1}$ with $\eta_t : \mathbb{R} \to \mathbb{R}$ that are assumed to be Lipschitz continuous, and hence differentiable almost everywhere. Starting with an initial points $\hat{\theta}^0 = 0_{d\times 1}$ and $r^0 = 0_{n \times 1}$, a sequence of estimates $\hat{\theta}^t$ is generated according to

$$\hat{\theta}^{t+1} = \eta_t\left(A^\top r^t + \hat{\theta}^t\right), \tag{39}$$

$$r^t = Z - A\hat{\theta}^t + \frac{d}{n} r^{t-1} \operatorname{div}\left(\eta_t(A^\top r^{t-1} + \hat{\theta}^{t-1})\right) \tag{40}$$

where $\eta_t(\cdot)$ is applied comopontwise and $\operatorname{div}(\eta_t(z)) = \frac{1}{n}\sum_{i=1}^n \eta_t'(z_i)$ with $\eta_t'(z) = \frac{\mathrm{d}}{\mathrm{d}z}\eta_t(z)$.

The main result of [28], [57] says that the MSE of each iteration of AMP can be characterized precisely in the high-dimensional limit when A is a realization of a random matrix with i.i.d. zero-mean Gaussian entries. To formally state and use this result, we need to the following assumptions:

Assumption 2: $\{\theta(n)\}_{n\in\mathbb{N}}$ is a sequence of d_n-dimensional vectors such that $n/d_n \to \delta \in (0,\infty)$ as n goes to infinity. The empirical distributions of $\theta(1), \theta(2), \ldots,$ i.e., the probability distribution that puts a point mass $1/d_n$ at each of the d_n entries of $\theta(n)$, converges weakly to a distribution π on $\mathbb{R}$ with finite second moment κ^2. Furthermore, $\|\theta(n)\|^2/n$ converges to κ^2 as $n \to \infty$.

Assumption 3: $\{P_{\theta(n),n}\}_{n\in\mathbb{N}}$ is a sequence of models defined by $X \sim (A\theta(n), \epsilon^2 I_n)$, where the entries of A are i.i.d. $\mathcal{N}(0, 1/n)$.

For a fixed n, we further consider a sequence of estimators for $\theta(n)$ defined as follows:

Assumption 4: $\{\eta_t\}_{t\in\mathbb{N}}$ is a sequence of scalar, Lipschitz continuous, and differentiable denoisers $\eta_t : \mathbb{R} \to \mathbb{R}$. For every $n, d_n \in \mathbb{N}$, the approximate message-passing (AMP) estimator $\theta^t_{\mathsf{AMP}}(z)$ is defined as the results of t iterations of (39) and (40).

The characterization of the MSE of the estimator θ^t_{AMP} in the high-dimensional limit is given by the *state evolution* recursion. This recursion is defined in terms of a distribution π on $\mathbb{R}$, sampling ratio $\delta \in (0,\infty)$, and initial noise level τ_0, as

$$\tau_{t+1}^2 = \xi^2 + \frac{1}{\delta}\mathbb{E}\left[\left(\eta_t\left(\theta_0 + \tau_t W\right) - \theta_0\right)^2\right], \quad t = 1, 2, \ldots, \tag{41}$$

where $\theta_0 \sim \pi$ and $W \sim \mathcal{N}(0,1)$. Finally, define

$$\mathcal{M}^t_{\mathsf{AMP}}(\xi^2) := \mathbb{E}\left[\left(\eta_t\left(\theta_0 + \tau_t W\right) - \theta_0\right)^2\right],$$

where τ_t is given by t iterations of (41). Under assumptions 2-4 above, [57, Thm. 1] implies that

$$\lim_{n\to\infty} \frac{1}{d_n}\left\|\theta(n) - \theta^t_{\mathsf{AMP}}(Z)\right\|^2 = \mathcal{M}^t_{\mathsf{AMP}}(\xi^2). \tag{42}$$

Combining this result with Theorem 6, we conclude the following:

Theorem 11: Consider a sequence of problems satisfying Assumptions 2 and 3. Let Y be the output of an $(n, \lceil 2^{nR}\rceil)$ random spherical code applied to X with radius ρ for some $R > 0$. Let θ^t_{AMP} be an estimator satisfying Assumption 4. Then

$$\lim_{n\to\infty} \frac{1}{d_n}\mathbb{E}\left[\left\|\theta(n) - \theta^T_{\mathsf{AMP}}(Y)\right\|^2 \mid A\right] = \mathcal{M}^t_{\mathsf{AMP}}(\xi_R^2), \tag{43}$$

almost surely, where

$$\xi_R^2 = \epsilon^2 + \frac{\epsilon^2 + \kappa^2/\delta}{2^{2R}-1}.$$

Proof: Set $\gamma_n^2 = n(\epsilon^2 + \kappa^2/\delta)$ and $\sigma^2 = \gamma_n/(n(2^{2R}-1))$. We first show that X and γ_n satisfy Assumption 1. Since A has i.i.d. entries $\mathcal{N}(0,1/n)$, then $X \sim \mathcal{N}\left(0, (\frac{1}{n}\|\theta(n)\|^2 + \epsilon^2)I_n\right)$. Using similar arguments as in Proposition 7, we get

$$\begin{aligned}&\left(\mathbb{E}\left[|\|X\| - \gamma_n|^2\right]\right)^{1/2} \\ &\quad \leq \left(\mathbb{E}\left[\left|\|X\| - \sqrt{\|A\theta(n)\|^2 + n\epsilon^2}\right|^2\right]\right)^{1/2} + \omega_n.\end{aligned}$$

Assumption 2 implies that $\theta = \theta(n) \in \mathcal{S}_{d_n}$ with $\omega_n = o(\sqrt{d_n})$. We conclude that

$$\begin{aligned}&\mathbb{E}\left[\left|\|X\| - \sqrt{\|A\theta(n)\|^2 + n\epsilon^2}\right|^2\right] \\ &\quad\leq \mathbb{E}\left[\frac{\left(\|X\|^2 - \|A\theta(n)\|^2 + n\epsilon^2\right)^2}{\|A\theta\|^2 + n\epsilon^2}\right] \\ &\quad\leq \frac{\operatorname{Var}(\|X\|^2)}{n\epsilon^2} = \frac{2n(\frac{1}{n}\|\theta(n)\|^2 + \epsilon^2)^2}{n\epsilon^2} \\ &\quad\leq \frac{2(\frac{(\omega_n + \sqrt{d_n}\kappa)^2}{n} + \epsilon^2)^2}{\epsilon^2} = O(1),\end{aligned}$$

and thus Assumption 1 is satisfied for some $\tau_n = o(\sqrt{n})$. Let $L_{n,t} := \|\theta^t_{\mathsf{AMP}}\|_{\mathrm{Lip}}$. In Appendix E we show that $\sup_n L_{n,t} < \infty$ almost surely.

Applied to our setting, (42) says that

$$\left|\left(\frac{1}{d_n}\left\|\theta(n) - \theta^t_{\mathsf{AMP}}(Z)\right\|^2\right)^{1/2} - \sqrt{\mathcal{M}^t_{\mathsf{AMP}}(\xi_R^2)}\right| = o(1).$$

Using the triangle inequality once with the last display, Theorem 6 implies that there exists C, that depends only on R and $\kappa^2/\delta + \epsilon^2$, such that

$$\left| \left(\frac{1}{d_n} \mathbb{E}\left[\|\theta(n) - \theta^t_{\mathsf{AMP}}(Y)\|^2 \right] \right)^{1/2} - \sqrt{\mathcal{M}^t_{\mathsf{AMP}}(\xi_R^2)} \right| \leq \frac{C \, \|\theta^t_{\mathsf{AMP}}\|_{\mathrm{Lip}} \beta_n}{\sqrt{d_n}},$$

with $\beta_n = o(\sqrt{d_n})$. □

IV. Application to Indirect Source Coding

For the second application, we consider an indirect source coding setting where the observed data is a degraded version of the realization of an information source. The goal is to compress this version at bitrate R and recover the source realization. Traditionally, both the encoder and decoder are designed with full knowledge of the joint distribution of the source and the data [19]. In this section, we study an encoding-decoding scheme where the encoder uses a random spherical code and the decoder is described by a Lipschitz estimator, which may be designed with partial or full knowledge of the distribution of the source and the data. Leveraging the results in Section II, we show that the asymptotic performance can be described in terms of an AWGN model.

Throughout this section, the source and the data are modeled as a stochastic process $\{(U_n, X_n)\}_{n\in\mathbb{N}}$. The first n terms in this sequence are denoted by $U^n = (U_1, \dots, U_n)$ and $X^n = (X_1, \dots, X_n)$. We focus on the squared error loss (or distortion function)

$$d(u^n, \hat{u}^n) = \frac{1}{n} \sum_{i=1}^n (u_i - \hat{u}_i)^2, \tag{44}$$

and assume the following regularity condition:

Assumption 5: The process $\{(U_n, X_n)\}_{n\in\mathbb{N}}$ is stationary and second-order ergodic with finite second moments. In particular, this means that the empirical second moments converge in mean:

$$\frac{1}{n} \sum_{i=1}^n \begin{pmatrix} U_i \\ X_i \end{pmatrix} \begin{pmatrix} U_i \\ X_i \end{pmatrix}^T \to \mathbb{E}\left[\begin{pmatrix} U_1 \\ X_1 \end{pmatrix} \begin{pmatrix} U_1 \\ X_1 \end{pmatrix}^T \right]. \tag{45}$$

A. The Indirect Distortion-Rate Function

We begin by reviewing some basic properties of the indirect distortion-rate function, which describes the fundamental tradeoff between the bitrate R and the expected distortion in our source coding setting. For each problem of size n, the indirect distortion-rate function is given by

$$D_n(R) := \min_{\phi,\psi} \mathbb{E}\left[d\left(U^n, \psi(\phi(X^n))\right) \right], \tag{46}$$

where the minimum is over all encoding functions $\phi : \mathbb{R}^n \to \{1, \dots, M\}$ and decoding functions $\psi : \{1, \dots, M\} \to \mathbb{R}^n$ with $M = \lceil 2^{nR} \rceil$. The standard source coding setting corresponds to the special case where the source equals the data. When the source and data are stationary, as we assume in this paper, $nD_n(R)$ is sub-additive in n, and the limit

$$D(R) := \lim_{n\to\infty} D_n(R) \tag{47}$$

is well-defined [58, Lem. 10.6.2].

For some classes of processes, $D(R)$ can be expressed equivalently in terms of an optimization problem over a family of probability distributions subject to a mutual information constraint [18], [19]. Specifically, we have

$$D(R) = \lim_{n\to\infty} \min_{I(X^n;\hat{U}^n)\leq nR} \mathbb{E}\left[d\left(U^n, \hat{U}^n \right) \right], \tag{48}$$

where the minimum is over all joint distributions on $(U^n, X^n, \hat{U}^n)$ such that (U^n, X^n) satisfy their marginal constraints, $U^n \to X^n \to \hat{U}^n$ forms a Markov chain, and $I(X^n; \hat{U}^n) \leq nR$. For example, a representation of the form (48) exists for memoryless processes [19], [59] and in cases where the direct (standard) distortion-rate function of the sequence of random vectors $\tilde{U}^n = \mathbb{E}[X^n \mid U^n]$ has a representation of the form (48) by setting $X^n = U^n = \tilde{U}^n$ [60, Ch. 3.2].

There are a few cases where the distortion-rate function has simple closed-form expressions. For example if $\{(U_n, X_n)\}$ are i.i.d. from bivariate Gaussian distribution with zero mean, then the distortion-rate function is given by $D(R) = D_G(R)$ where

$$D_G(R) := \mathbb{E}\left[|U_1|^2 \right] - \frac{\mathbb{E}\left[U_1 X_1\right]^2}{\mathbb{E}\left[|X_1|^2\right]} \left(1 - 2^{-2R}\right). \tag{49}$$

This characterization was obtained in [18] and also [20]. Note that the limiting case $R \to \infty$ corresponds to the minimum MSE in estimating U_n from X_n. Moreover, for the direct source coding problem where U_n is equal to X_n, this expression reduces to the standard distortion-rate function for an i.i.d. Gaussian source, $\mathbb{E}\left[|X_1|^2\right] 2^{-2R}$.

B. Achievability Using Spherical Coding

We now consider the distortion that can be achieved when X^n is compressed using a random spherical code. For each problem of size n, let Y^n be the output of a bitrate-R random spherical code with input X^n and squared magnitude $n\mathbb{E}\left[X_1^2\right]/(1 - 2^{-2R})$. The distortion-rate function associated with random spherical coding and estimator $f : \mathbb{R}^n \to \mathbb{R}^n$ is defined as

$$D_n^{\mathsf{sp}}(R, f) := \mathbb{E}\left[d\left(U^n, f(Y^n)\right) \right], \tag{50}$$

where the expectation is with respect to the joint distribution of (U^n, Y^n). Under the squared error distortion, the minimum with respect to f is achieved by the conditional expectation $f(y) = \mathbb{E}[U^n \mid Y^n = y]$. We note that this formulation of the distortion-rate function does not necessarily describe the optimal performance that is possible using a random spherical code, because the estimation stage is based only on the compressed representation Y^n and does not use any other information about the realization of the codebook.

Following the central theme of this paper, our results are described in terms of an AWGN counterpart to the distortion-rate function. Given noise variance σ^2, define the sequence $\{Z_n\}_{n\in\mathbb{N}}$ by

$$Z_n = X_n + \sigma W_n, \tag{51}$$

where W_n is an independent standard Gaussian noise. The MSE associated with an estimator $f : \mathbb{R}^n \to \mathbb{R}^n$ is defined by

$$\mathcal{M}_n(\sigma^2, f) := \mathbb{E}\left[d\left(U^n, f(Z^n)\right)\right], \tag{52}$$

The minimum over f is attained by the conditional expectation $f(z) = \mathbb{E}\left[U^n \mid Z^n = z\right]$ and is denoted by $\mathcal{M}_n(\sigma^2) := \min_f \mathcal{M}_n(\sigma^2, f)$. Stationarity of the sequence $\{(U_n, Z_n)\}$ implies that $n\mathcal{M}_n(\sigma^2)$ is sub-additive in n, and thus the following limit is well-defined

$$\mathcal{M}(\sigma^2) := \lim_{n\to\infty} \mathcal{M}_n(\sigma^2). \tag{53}$$

We refer to $\mathcal{M}(\sigma^2)$ as the minimum MSE function associated with the AWGN model. The next result establishes the formal equivalence between the distortion-rate function associated with random spherical coding and $\mathcal{M}(\sigma^2)$. The proof is based on the Gaussian approximation of quantization error in Theorem 4 as well as some further properties of the AWGN model.

Theorem 12: Suppose that $\{(U_n, X_n)\}$ is a random process satisfying Assumption 5. Let $\{f_n\}_{n\in\mathbb{N}}$ be a sequence of estimators $f_n : \mathbb{R}^n \to \mathbb{R}^n$ satisfying $\|f_n\|_{\mathrm{Lip}} \le L$ and $\|f_n(0)\| \le \sqrt{n}C$ for all n where L, C are positive constants. Then, for each $R > 0$,

$$\lim_{n\to\infty} \left| D_n^{\mathsf{sp}}(R, f_n) - \mathcal{M}_n(\sigma_R^2, f_n)\right| = 0, \tag{54}$$

where $\sigma_R^2 = \mathbb{E}\left[|X_1|^2\right]/(2^{2R}-1)$. Furthermore, there exists a sequence of estimators $\{f_n\}_{n\in\mathbb{N}}$ such that

$$\lim_{n\to\infty} \frac{1}{n} D^{\mathsf{sp}}(R, f_n) = \lim_{n\to\infty} \frac{1}{n}\mathcal{M}(\sigma_R^2, f_n) = \mathcal{M}(\sigma_R^2). \tag{55}$$

Proof: Set $\gamma = \sqrt{n}\sqrt{\mathbb{E}\left[|X_1|^2\right]}$, $M = \lceil 2^{nR}\rceil$, and (ρ, σ) as in (11). Note that $\sigma^2 = \sigma_R^2$. Following the same steps as in the proof of Proposition 5, we have

$$\left|\sqrt{D_n^{\mathsf{sp}}(R, f_n)} - \sqrt{\mathcal{M}_n(\sigma_R^2, f_n)}\right| \le \frac{L \cdot W_2(P_{Y^n}, P_{Z^n})}{\sqrt{n}}. \tag{56}$$

By Theorem 4, the normalized Wasserstein distance can be upper bound as

$$\frac{W_2(P_{Y^n}, P_{Z^n})}{\sqrt{n}} \le \left(\mathbb{E}\left[\left|\frac{1}{\sqrt{n}}\|X^n\| - \sqrt{\mathbb{E}\left[|X_1|^2\right]}\right|^2\right]\right)^{1/2} + \frac{\sqrt{2}C_R}{\sqrt{2^{2R}-1}} \frac{\sqrt{\mathbb{E}\left[|X_1|^2\right]}}{\sqrt{n}},$$

where C_R is a constant that depends only on R. The second term in this bound converges to zero at a rate $1/\sqrt{n}$. Combining the inequality $|\sqrt{a} - \sqrt{b}| \le \sqrt{|a-b|}$ with the assumption that $\{X_n\}$ is second order ergodic, one finds that the first term also converges to zero. Putting everything together, we conclude that

$$\lim_{n\to\infty}\left|\sqrt{D_n^{\mathsf{sp}}(R, f_n)} - \sqrt{\mathcal{M}_n(\sigma_R^2, f_n)}\right| = 0. \tag{57}$$

To prove that this comparison holds without the square roots, it is sufficient to show that $D_n^{\mathsf{sp}}(R, f_n)$ and $\mathcal{M}_n(R, f_n)$ are bounded uniformly with respect to n. To this end, we can use the triangle inequality and the assumptions on f_n to write:

$$\begin{aligned}\|U^n - f_n(Y^n)\| &\le \|U^n - f_n(0)\| + \|f_n(0)\| + \|f(Y^n)\| \\ &\le \|U^n\| + \sqrt{n}C + L\|Y^n\|.\end{aligned}$$

Combining this bound with the assumptions on U^n and Y^n establishes that $D_n^{\mathsf{sp}}(R, f_n)$ is bounded uniformly, and the same approach also works for $\mathcal{M}_n(\sigma_R^2, f_n)$.

To prove the second part, we will show that for each $\epsilon > 0$, there exists a sequence of estimators f_n satisfying $\sup_n \|f_n\|_{\mathrm{Lip}} < \infty$ and $\limsup_{n\to\infty} \mathcal{M}(\sigma^2, f_n) \le \mathcal{M}(\sigma^2) + \epsilon$. The existence of the limit in the definition of $M(\sigma^2)$ means that for each $\epsilon > 0$, there exists an integer N such that $|\mathcal{M}_n(\sigma^2) - M(\sigma^2)| \le \epsilon$ for all $n \ge N$. By Lemma 21 in the Appendix, there exists a Lipschitz continuous function $g : \mathbb{R}^N \to \mathbb{R}^N$ such that $|\mathcal{M}_n(\sigma^2) - \mathcal{M}_n(\sigma^2, g)| \le \epsilon$. For $n \ge N$, let $f_n : \mathbb{R}^n \to \mathbb{R}^n$ be defined by applying g to the first $\lfloor n/N\rfloor$ successive length-N blocks of Z^n and setting any remaining entries to zero. Then, we have $\|f_n\|_{\mathrm{Lip}} = \|g\|_{\mathrm{Lip}}$ and

$$\begin{aligned}&\mathcal{M}_n(\sigma^2, f_n) \\ &\quad = \frac{1}{n}\lfloor n/N\rfloor \mathcal{M}_n(\sigma^2, g) + \frac{1}{n}\left(n - \lfloor n/N\rfloor\right)\mathbb{E}\left[|U_1|^2\right].\end{aligned}$$

Putting everything together, we have $|\mathcal{M}_n(\sigma^2, f_n) - M(\sigma^2)| \le 3\epsilon$ for all n large enough. As ϵ can be chosen arbitrarily small, the proof is complete. □

The significance Theorem 12 is that it provides a link between the problem of estimation from compressed data, which is often difficult to study directly, and the better-understood problem of estimation in Gaussian noise. We emphasize that the assumptions on the source and data are quite general, particularly in comparison to many of the existing results in the literature.

Compared to optimal encoding schemes that attain the indirect distortion-rate function $D(R)$, a useful property of random spherical coding is that it can be implemented without any knowledge of the underlying source distribution. Therefore, the coding scheme described in this paper can be employed in typical data acquisition situations where the distribution of the data and the source of interest is learned *after* the data are collected and quantized.

C. Universality of Linear Estimation

We now consider the performance of linear estimators. Given a bitrate $R > 0$, define the scalar

$$\alpha_R = \left(1 - 2^{-2R}\right)\frac{\mathbb{E}\left[U_1 X_1\right]}{\mathbb{E}\left[|X_1|^2\right]}. \tag{58}$$

A standard calculation reveals that under the AWGN model, the MSE of the linear estimator $f(y) = \alpha_R y$ is independent of the problem dimension and is given by

$$\mathcal{M}_n(\sigma_R^2, f) = \frac{1}{n}\mathbb{E}\left[\|U^n - \alpha_R Z^n\|^2\right] = D_G(R), \quad (59)$$

where we recall that $D_G(R)$ of (49) is the distortion-rate function associated with a zero-mean Gaussian source. In view of Theorem 12, this correspondence between the Gaussian distortion-rate function and the MSE of linear estimators in the AWGN model implies an achievable result for random spherical codinig.

Proposition 13: Let $\{(U_n, X_n)\}$ be a process satisfying Assumption 5. For each integer n, let Y^n be the output of a bitrate-R random spherical code with input X^n and squared magnitude $n\mathbb{E}\left[|X_1|^2\right]/(2^{2R}-1)$. Then,

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\|U^n - \alpha_R Y^n\|^2\right] = D_G(R), \quad (60)$$

where α_R is given by (58).

Applied to the special case of direct source coding $X^n = U^n$, Proposition 13 recovers the results in [33] and [35], which showed that squared error distortion of a (properly scaled) random spherical code depends only on the second-order statistics of the source and is equal to the Gaussian distortion-rate function. The contribution of Proposition 13 is to show that this result carries over naturally to the indirect source coding setting. Moreover, if $\{(U_n, X_n)\}$ are i.i.d. zero-mean Gaussian, then we have the equivalence:

$$D(R) = \mathcal{M}(\sigma_R^2) = D_G(R). \quad (61)$$

We note that, in general, codebooks approaching the optimal trade-off between bitrate and MSE described by $D(R)$ depend on the joint distribution of $\{(U_n, X_n)\}$. This is because such codebooks essentially encode the sequence obtained by estimating U^n from X^n [21], [41], i.e., estimation precedes encoding in this case. When U^n and X^n are i.i.d. and jointly Gaussian, this estimation is obtained by multiplying X^n by α_R, and there is essentially no difference if this multiplication is performed pre- or post-encoding. To summarize, for i.i.d. Gaussian and zero mean $\{(U_n, X_n)\}$, the equality $D(R) = \mathcal{M}(\sigma_R^2)$ is due to two factors: (1) The optimal estimator is a scalar multiple of the data, and (2) random spherical coding is optimal for encoding Gaussian sources.

D. Non-Linear Estimation

Next, we consider the performance of non-linear estimators when the source and the data are non-Gaussian. Suppose that the source and the data are memoryless, that is the pairs (U_n, X_n) are i.i.d. from a distribution $P_{U,X}$ with finite second moments. Under this assumption, the indirect distortion-rate function $D(R)$ can be expressed as [18], [19]

$$D(R) = \min_{I(X;\hat{U})\le R} \mathbb{E}\left[(U - \hat{U})^2\right], \quad (62)$$

where the minimum is over all distributions on $(X, \hat{U})$ such that $X \sim P_X$, and $I(X;\hat{U}) \le R$. Noting that

$$\min_{I(X;\hat{U})\le R} \mathbb{E}\left[(U - \hat{U})^2\right] = \min_{I(X;\hat{X})\le R} \mathbb{E}\left[d(X, \hat{X})\right],$$

where $d(x, \hat{x}) := \mathbb{E}\left[(U - \hat{U})^2 \mid X = x, \hat{U} = \hat{x}\right]$, $D(R)$ can be approximated numerically using [61].

In the setting of the AWGN model, the memoryless assumption means that the problem of estimating U^n from Z^n decouples into n independent estimation problems, and the minimum MSE function is given by

$$\mathcal{M}(\sigma^2) = \mathbb{E}\left[(U - \mathbb{E}[U \mid Z])^2\right], \quad Z = X + \sigma W, \quad (63)$$

where $(U, X) \sim P_{U,X}$ and $W \sim \mathcal{N}(0, 1)$ are independent. This expression can be approximated numerically using standard techniques.

An interesting special case of the indirect source coding problem occurs when the data is an AWGN corrupted version of the source, that is

$$X = U + \epsilon W', \quad (64)$$

with $U \sim P_U$ independent of $W' \sim \mathcal{N}(0, 1)$. In this case, the Gaussian noise in the data can be combined with the independent Gaussian noise in the AWGN model such that

$$Z = U + \sqrt{\epsilon^2 + \sigma^2}W'', \quad (65)$$

where $U \sim P_U$ independent of $W'' \sim \mathcal{N}(0, 1)$. In Figure 5, we provide a comparison of the indirect distortion-rate function $D(R)$ and the upper bound on the distortion obtained using random spherical coding $\mathcal{M}(\sigma_R^2)$ in the setting where U is uniform on $\{-1, 1\}$ and X is drawn according to (64). For comparison, we also plot the upper bound $D_G(R)$ corresponding to linear estimation, as well as the asymptotes of all MSE functions as the noise variance ϵ vanishes.

V. Proof of Main Result

The proof of Theorem 1 requires several lemmas.

Lemma 14: Suppose that V is distributed uniformly on the unit sphere in $\mathbb{R}^n$ with $n \ge 2$. For any $x \in \mathbb{R}^n \backslash \{0\}$, the distribution on V can be decomposed as

$$V = G\frac{x}{\|x\|} + \sqrt{1 - G^2}H \quad (66)$$

where $G = \langle x, V\rangle/\|x\|$ is a random variable supported on $[-1, 1]$ with complementary cumulative distribution function $\mathbb{P}[G \ge g] = Q_n(g)$ of (6), and H is an independent random vector distributed uniformly on the set $\{h \in \mathbb{R}^n : \|h\| = 1 \text{ and } \langle x, h\rangle = 0\}$.

Proof: By the orthogonal invariance of the distribution on V, we may assume without loss of generality that x is a unit vector of the form $x = (1, 0, \dots 0)$. Then, $G = V_1$ and $H = (0, V_2, \dots, V_n)/\sqrt{\sum_{i=2}^n V_i^2}$. The joint distribution of (G, H) follows from the joint distribution on the entries of a random spherical vector [55, Eq. (3)]. □

Lemma 15: For $n \ge 2$, let Y be the output of an (n, M)-random spherical code with input $x \in \mathbb{R}^n \backslash \{0\}$ and magnitude ρ. The distribution of Y can be decomposed as

$$Y = \rho\left(\frac{x}{\|x\|}\cos(\alpha^*) + \sqrt{1 - S^2}H\right) \quad (67)$$

where $\cos(\alpha^*)$ has the cumulative distribution function (5), and H is an independent random vector distributed uniformly on the set $\{h \in \mathbb{R}^n : \|h\| = 1, \text{ and } \langle x, h\rangle = 0\}$.

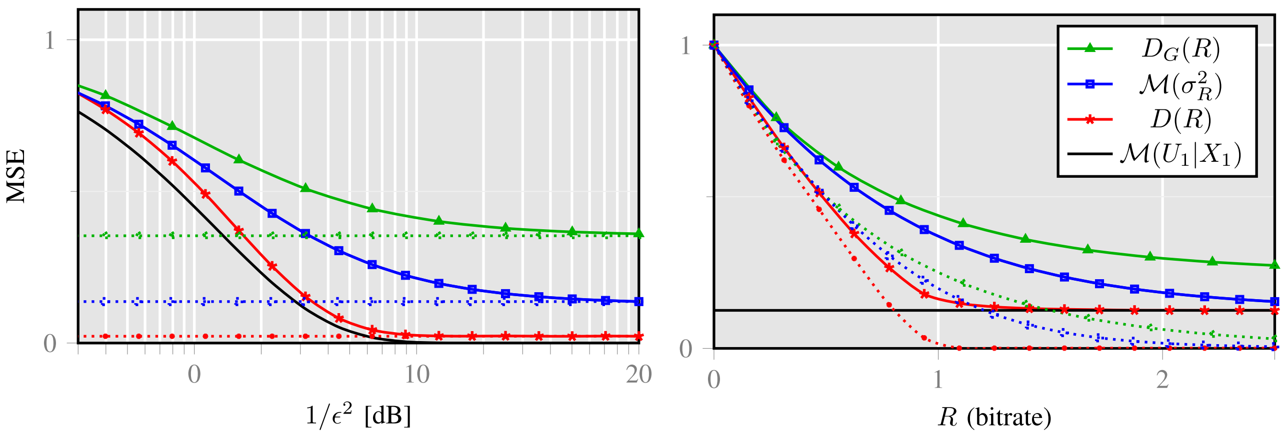

Fig. 5. Mean square error (MSE) in estimating an i.i.d. signal equiprobable on $\{-1, 1\}$ from a bitrate-R encoding of its AWGN-corrupted version. Left Panel: MSE versus noise variance ϵ^2 with a fixed encoding bitrate $R = 1$. Right Panel: MSE versus bitrate R with noise variance $\epsilon^2 = 1/3$. $\mathcal{M}(\sigma_R^2)$ is achievable using a random spherical code followed by a scalar Bayes estimator. $D_G(R)$ is achievable using random spherical coding followed by a scalar linear estimator. $D(R)$ is the indirect distortion-rate function corresponding to the optimal encoding scheme. The dashed lines indicate the asymptotic MSE as $\epsilon \to 0$. Also shown is $\mathcal{M}(U_1|X_1)$, which is the minimal MSE in estimating the signal from its corrupted version corresponding to the limit $R \to \infty$.

Proof: For each code word $C(i)$ we apply the decomposition in Lemma 14 to obtain

$$C(i) = G(i) + \sqrt{1 - G(i)^2} H(i),$$

where $G(i) = \langle x, C(i) \rangle / \|x\| \|C(i)\|$ is the cosine similarity of the i-th codeword. Recall that the index i^* corresponds to the code word that maximizes the cosine similarity $\cos(\alpha^*) = G(i^*) := \max\{G(1), \dots G(M)\}$. Therefore, the distribution of S follows from the fact that $G(1), \dots G(M)$ are i.i.d. with complementary cumulative distribution function given by (6). Furthermore, because i^* depends only on the terms $G(1), \dots G(M)$, it follows from Lemma 14 that $H := H(i^*)$ is independent of α^* and uniform on the subset of the unit sphere that is orthogonal to x. Noting that $Y = \rho C(i^*)$ completes the proof. □

Lemma 16: Suppose that W is a standard Gaussian vector on $\mathbb{R}^n$ with $n \geq 2$. For any $x \in \mathbb{R}^n \backslash \{0\}$ the random vector $Z = x + \sigma W$ can be decomposed as

$$Z = x + \sigma A \frac{x}{\|x\|} + \sigma B H \tag{68}$$

where (A, B, H) are independent, $A \sim \mathcal{N}(0, 1)$ has a standard Gaussian distribution, $B \sim \chi_{n-1}$ has a chi distribution with $n - 1$ degrees of freedom, and H is distributed uniformly on the set $\{h \in \mathbb{R}^n : \|h\| = 1, \text{ and } \langle x, h \rangle = 0\}$.

Proof: By the orthogonal invariance of the Gaussian distribution on W, we may assume without loss of generality that x is a unit vector of the form $x = (1, 0, \dots, 0)$. Letting $A = W_1$, $B = \sqrt{\sum_{i=2}^n W_i^2}$, and $H = (0, W_2, \dots, W_n)/B$ yields $W = Ax/\|x\| + BH$. By construction, A is a standard Gaussian variable that is independent of $W_2, \dots, W_n$. The distribution of (B, H) follows from the fact that $B(H_2, \dots, H_n)$ is the polar decomposition of the $(n-1)$-dimensional standard Gaussian vector $(W_2, \dots W_n)$. □

Using the characterizations of Y and Z given in Lemma 15 and Lemma 16, respectively, we see that for every $x \in \mathbb{R}^n \backslash \{0\}$, there exists a coupling on (Y, Z) such that

$$Y = x + (\rho \cos(\alpha^*) - \|x\|) \frac{x}{\|x\|} + \rho \sin(\alpha^*) H,$$
$$Z = x + \sigma A \frac{x}{\|x\|} + \sigma B H,$$

where (A, B, H, α^*) are independent. By the orthogonality of x and H, the magnitude of the difference between Y and Z depends only on the tuple (A, B, α^*) and is given by (8).

VI. Conclusion

We considered the problem of estimating an underlying signal or parameter from the lossy compressed version of another high dimensional signal. For compression codes defined by a random spherical code of bitrate R, we showed that the distribution of the output codeword is close in Wasserstein distance to the conditional distribution of the output of an AWGN channel with SNR $2^{2R} - 1$. This equivalence between the noise associated with lossy compression and an AWGN channel allows us to adapt existing techniques for inference from AWGN-corrupted measurements to estimate the underlying signal from the compressed measurements, as well as to characterize their asymptotic performance.

We demonstrated the usefulness of this equivalence by deriving novel expressions for the achievable risk in various source coding and parameter estimation settings. These include bitrate-constrained sparse parameter estimation using soft thresholding, bitrate-constrained parameter estimation in high-dimensional linear models, and indirect source coding with linear and non-linear decoders. In each of these settings, our results yielded achievable MSE and provided the equivalent noise level required to tune the estimator to attain this MSE.

We believe that the characterization of lossy compression error developed in this paper can be useful in numerous important cases aside from the ones we explored. Examples of such cases include hypothesis testing based on compressed data, signal estimation in distributed lossy compression settings, and the study of convergence rates and accuracies of first-order optimization procedures employing gradient compression.

Appendix A
Proofs of Theorem 2 and 3

The proof of Theorems 2 and 3 require a characterization of the moments of the random variable

$$\Delta := \frac{1}{\sigma}\sqrt{(\rho\cos(\alpha^*) - \gamma - \sigma A)^2 + (\rho\sin(\alpha^*) - \sigma B)^2} \tag{69}$$

where (ρ, σ, γ) are deterministic parameters and (α^*, A, B) are independent random variables whose distributions are described in Theorem 1.

Given $\alpha \in (0, \pi/2)$ let

$$\rho = \gamma \sec\alpha, \qquad \sigma = \frac{\gamma\tan\alpha}{\sqrt{n}}. \tag{70}$$

Evaluating Δ with these values and then using the triangle inequality as well as basic trigonometric identities leads to

$$\Delta \le \Delta_1 + \Delta_2, \tag{71}$$

where

$$\Delta_1 := 2\sqrt{n}\csc(\alpha)\left|\sin\left(\frac{\alpha^* - \alpha}{2}\right)\right|, \tag{72}$$

$$\Delta_2 := \sqrt{A^2 + (B - \sqrt{n})^2}. \tag{73}$$

The term Δ_2 is sub-Gaussian with mean and variance parameter independent of n. An estimate for its sub-Gaussian constant is provided in Lemma 20. For the term, Δ_1, we use the following result, which is proved in Appendix B.

Lemma 17: Suppose that $M = M_\alpha(n) := \sqrt{n}(\csc\alpha)^{n-1}$ for $\alpha \in (0, \pi/2)$. Then, for $p \ge 1$,

$$\mathbb{E}\left[\left|\sin\left(\frac{\alpha^* - \alpha}{2}\right)\right|^p\right]^{1/p} \tag{74}$$

$$\le C\frac{\tan(\alpha)[\log(n \wedge \sec(\alpha)) + p]}{n} + 2^{-M/p} \tag{75}$$

where C is a numerical constant.

Proof of Theorem 2: In view of (71) and the fact that Δ_2 is sub-Gaussian with a constant that does not depend on n all that remains is to establish the desired upper bound on the moments of Δ_1. Note that if $p \ge n$ then this term is bounded almost surely according to

$$\Delta_1 \le 2\csc(\alpha)\sqrt{n} \le 2\csc(\alpha)\sqrt{p},$$

where $2\csc(\alpha)$ depends only on R. The remainder of the proof focuses on the case $1 \le p \le n$.

Recall that the specification in (11) corresponds to the choice $\alpha = \arcsin(2^{-R})$. Define

$$N_R = \min\left\{n \in \mathbb{N} \,:\, 2^{R+1} \le \sqrt{n} \le 2^{\frac{n+1}{2}R}\right\} \tag{76}$$

For $n \ge N_R$ it can be verified that there exists a unique value $\alpha_n \in [\alpha, \pi/2)$ such that $M = \sqrt{n}\csc(\alpha_n)^{(n-1)}$ and $\sin(\alpha_n) \le \sqrt{\sin(\alpha)}$. Noting the sin function is Lipschitz and non-decreasing on $[0, \pi/2]$ we can write

$$\left|\sin\left(\frac{\alpha^* - \alpha}{2}\right)\right| \le \left|\sin\left(\frac{\alpha^* - \alpha_n}{2}\right)\right| + \sin\left(\frac{\alpha_n - \alpha}{2}\right). \tag{77}$$

The second term in (77) is deterministic and satisfies

$$\sin\left(\frac{\alpha_n - \alpha}{2}\right) = \cos\left(\frac{\alpha_n - \alpha}{2}\right)\frac{\sin(\alpha_n) - \sin(\alpha)}{\cos(\alpha_n) + \cos(\alpha)} \tag{78}$$

$$\le \frac{\tan(\alpha_n)}{2}\left(1 - \frac{\sin(\alpha)}{\sin(\alpha_n)}\right) \tag{79}$$

$$\le \frac{\tan(\alpha_n)}{2}\log\left(\frac{\sin(\alpha_n)}{\sin(\alpha)}\right) \tag{80}$$

$$= \frac{\tan(\alpha_n)}{2}\left[\frac{\frac{1}{2}\log n - \log M}{n-1} + R\log(2)\right] \tag{81}$$

$$\le \frac{\tan(\alpha_n)\log n}{2n}. \tag{82}$$

To bound the moments of the first term in (77) we can use Lemma 17. Finally, recalling that $\sin(\alpha_n) \le \sqrt{\sin\alpha}$, it follows that

$$\tan(\alpha_n) = \frac{1}{\sqrt{\sin(\alpha_n)^{-2} - 1}} \tag{83}$$

$$\le \frac{1}{\sqrt{\sin(\alpha)^{-1} - 1}}. \tag{84}$$

In view of (74) and $n > p$, it follows that

$$\mathbb{E}\left[\Delta_1^p\right]^{1/p} \le C_R\left(\frac{[\log(n) + p]}{\sqrt{n}} + \sqrt{n}2^{-M/p}\right) \tag{85}$$

$$\le C_R\left(1 + \sqrt{p} + \sqrt{n}2^{-M/p}\right). \tag{86}$$

Finally, for the term $\sqrt{n}2^{-M/p}$, recalling that $M_n = \lceil 2^{nR}\rceil$, we see that for $p \le \sqrt{n}$,

$$\sqrt{n}2^{-M/p} \le \sqrt{n}2^{-2^{nR}/p} \le \sqrt{n}2^{-2^{nR}/n} \le C_R'' \tag{87}$$

for some positive constant C_R''. This completes the proof of Theorem 2. □

Proof of Theorem 3: The specification given (13) corresponds to the choice

$$\alpha = \arcsin\left(M^{1/(n-1)}\right) \tag{88}$$

Under the assumption $M \ge M_\beta(n)$, there exists $\tilde{\alpha} \in (\alpha, \beta]$ such that $M = \sqrt{n}(\csc(\tilde{\alpha}))^{n-1}$. Using the same approach as in the proof Theorem 2 leads to

$$\left|\sin\left(\frac{\alpha^* - \alpha}{2}\right)\right| \le \left|\sin\left(\frac{\alpha^* - \tilde{\alpha}}{2}\right)\right| + \frac{\tan(\alpha_n)\log n}{2n}. \tag{89}$$

By Lemma 17 it follows that

$$\mathbb{E}\left[\Delta_1^p\right]^{1/p} \le C'\frac{\sec(\tilde{\alpha})[\log(n) + p]}{\sqrt{n}} + 2\sqrt{n}\csc(\tilde{\alpha})2^{-M/p} \tag{90}$$

where C' is a numerical constant. Since the secant function is non-decreasing on $[0, \pi/2]$ we have $\sec(\alpha_n) \le \sec(\beta)$, and so the first term is bounded from above by $C'_\beta p$ for some number C'_β. The second term satisfies

$$\begin{aligned}\sqrt{n}\csc(\tilde\alpha)2^{-M/p} &= \sqrt{n}\csc(\tilde\alpha)\exp\left\{\frac{\sqrt{n}}{p}(\csc\tilde\alpha)^{n-1}\log 2\right\}\\ &\le \sqrt{n}\csc(\tilde\alpha)\exp\left\{-\frac{\sqrt{n}}{p}\csc(\tilde\alpha)\log 2\right\}\\ &\le \frac{p}{e\log 2}.\end{aligned}$$

□

Appendix B
Proof of Lemma 17

Let integers n, M and $\alpha \in (0, \pi/2)$ be such that $M = \sqrt{n}(\csc\alpha)^{n-1}$. Recall that the goal is to bound the absolute moments of the random variable $\sin((\alpha^* - \alpha)/2)$ where α^* is drawn according to (5). We consider two cases. First, on the event $\alpha^* > \pi/2$, we can use the trivial upper bound of one. Note that $Q_n(0) = 1/2$ and so the probability of this event is 2^{-M}. Alternatively, on the event $\alpha^* \le \pi/2$, we use basic trigonometric identities to write

$$\sin\left(\frac{\alpha^*-\alpha}{2}\right) = \cos\left(\frac{\alpha^*-\alpha}{2}\right)\frac{(\sin(\alpha^*)-\sin(\alpha))}{(\cos(\alpha^*)+\cos(\alpha))}. \quad (91)$$

Noting that $\cos(\alpha^*) \ge 0$ leads to the upper bound $|\sin((\alpha^*-\alpha)/2)| \le \tan(\alpha)\,\xi$ where

$$\xi := \left|\frac{\sin(\alpha^*)}{\sin(\alpha)} - 1\right|. \quad (92)$$

Combining these two cases yields

$$\mathbb{E}\left[\left|\sin\left(\frac{\alpha^*-\alpha}{2}\right)\right|^p\right]^{1/p} \le \tan(\alpha)\mathbb{E}\left[\xi^p\right]^{1/p} + 2^{-M/p}. \quad (93)$$

The remaining step in the proof is to provide an upper bound moments of ξ. We need the following bounds on the function $Q_n(s)$ defined in (6).

Lemma 18: For integer $n \ge 2$ and $s \in [0,1]$,

$$\frac{(1-s^2)^{(n-1)/2}}{\sqrt{2\pi}\mu_n} \le Q_n(s) \le \frac{(1-s^2)^{(n-1)/2}}{\sqrt{2\pi}\mu_n\max(s, n^{-1})}, \quad (94)$$

where

$$\mu_n := \frac{\sqrt{2}\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \quad (95)$$

is the mean of the chi distribution with n degrees of freedom.

Proof: Making the change of variables $u = (1-t^2)$ and using the relation $\mu_{n-1}\mu_n = n-1$ leads to

$$Q_n(s) = \frac{1}{\sqrt{2\pi}\mu_n}\int_0^{1-s^2}\frac{(n-1)u^{(n-3)/2}}{2\sqrt{1-u}}\,du.$$

Noting that the denominator in the integral satisfies $s \le \sqrt{1-u} \le 1$ and then integrating gives the double inequality:

$$\frac{(1-s^2)^{(n-1)/2}}{\sqrt{2\pi}\mu_n} \le Q_n(s) \le \frac{(1-s^2)^{(n-1)/2}}{\sqrt{2\pi}\mu_n s}. \quad (96)$$

To prove an upper bound with s replaced by n^{-1} in the denominator, observe that for $n \ge 3$,

$$\begin{aligned}Q_n(s) &= \frac{(n-1)}{\sqrt{2\pi}\mu_n}\int_s^1(1-t^2)^{(n-3)/2}\,dt\\ &\le \frac{(n-1)}{\sqrt{2\pi}\mu_n}(1-s^2)^{(n-3)/2}(1-s)\\ &= \frac{(n-1)(1-s^2)^{(n-1)/2}}{\sqrt{2\pi}\mu_n(1+s)}.\end{aligned}$$

Meanwhile, for $n = 2$, direct calculation reveals

$$\begin{aligned}Q_2(s) &= \frac{1}{\sqrt{2\pi}\mu_2}\left(\frac{\pi}{2} - \arcsin(s)\right)\\ &\le \frac{2(1-s^2)^{1/2}}{\sqrt{2\pi}\mu_2}.\end{aligned}$$

Combining these upper bounds completes the proof. □

Lemma 19: For integers $n, M \ge 2$ and $u \in [0,1]$,

$$\mathbb{P}\left[\sin(\alpha^*) \le u\right] \le \frac{u^{n-1}M}{\sqrt{2\pi}\mu_n\max(\sqrt{1-u^2}, n^{-1})} \quad (97)$$

$$\mathbb{P}\left[\sin(\alpha^*) \ge u\right] \le \exp\left(-\frac{u^{n-1}M}{\sqrt{2\pi}\mu_n}\right). \quad (98)$$

Proof: Let $s = \sqrt{1-u^2}$. Expressing the event $\{\sin(\alpha^*) < u\}$ in terms of $\cos(\alpha^*)$, we can write

$$\begin{aligned}\mathbb{P}\left[\sin(\alpha^*) \le u\right] &= \mathbb{P}\left[\cos(\alpha^*) \ge s\right] + \mathbb{P}\left[\cos(\alpha^*) \le -s\right]\\ &= 1 - (1-Q_n(s))^M + (1-Q_n(-s))^M\\ &= 1 - (1-Q_n(s))^M + Q_n(s)^M\\ &\le MQ_n(s).\end{aligned}$$

Here, the third step follows because $Q_n(-s) = 1 - Q_n(s)$ and the last step is due to the basic inequality $1-(1-q)^p - q^p \le pq$ for $q \in [0,1]$ and $p \ge 2$. Combining this inequality with the upper bound on $Q_n(s)$ in (94) gives (97). For the complementary event, we have

$$\begin{aligned}\mathbb{P}\left[\sin(\alpha^*) \ge u\right] &= (1-Q_n(s))^M - (1-Q_n(-s))^M\\ &\le \exp(-MQ_n(s)),\end{aligned}$$

where we have used the basic inequality $1 - q \le e^{-q}$ for all $q \in [0,1]$. Combining with the lower bound on $Q_n(s)$ in (94) gives (98). □

To bound the distribution of ξ we consider two cases. First, conditional on the event $\sin(\alpha^*) \le \sin(\alpha)$, we can write

$$\xi = 1 - \frac{\sin(\alpha^*)}{\sin(\alpha)} \le \log\frac{\sin(\alpha)}{\sin(\alpha^*)}, \quad (99)$$

where we used the basic inequality $1 - 1/x \le \log x$ for $x > 0$. In view of (97) and the assumption $M = \sqrt{n}(\csc\alpha)^{n-1}$, it follows that for $t \ge 0$,

$$\begin{aligned}\mathbb{P}\left[\xi \ge t, \sin(\alpha^*) < \sin(\alpha)\right] &\le \mathbb{P}\left[\sin(\alpha^*) \le \sin(\alpha)e^{-t}\right]\\ &\le \frac{\sqrt{n}\exp\left(-(n-1)t\right)}{\sqrt{2\pi}\mu_n\max(\cos(\alpha), n^{-1})}\\ &\le \min(\csc(\alpha), n)\exp\left(-(n-1)t\right),\end{aligned}$$

where the last step uses the lower bound $\mu_n \ge \sqrt{n - 1/2}$.

Next, conditional on the event $\sin(\alpha^*) \geq \sin(\alpha)$, we can write

$$\xi = \frac{\sin(\alpha^*)}{\sin(\alpha)} - 1 \leq \frac{1}{n-1}\left[\left(\frac{\sin(\alpha^*)}{\sin(\alpha)}\right)^{n-1} - 1\right], \quad (100)$$

where we used the basic inequality $x - 1 \leq (x^p - 1)/p$ for all $x \geq 0$ and $p \geq 1$. In view of (98) and the assumption $M = \sqrt{n}(\csc\alpha)^{n-1}$, it then follows that for $t \geq 0$,

$$\begin{aligned}
&\mathbb{P}[\xi \geq t, \sin(\alpha^*) \geq \sin(\beta)] \\
&\quad \leq \mathbb{P}\left[\sin(\alpha^*) \geq (\sin(\beta))\,[1 + (n-1)t]^{\frac{1}{n-1}}\right] \\
&\quad \leq \exp\left(-\frac{\sqrt{n}[1 + (n-1)t]}{\sqrt{2\pi}\mu_n}\right) \\
&\quad \leq \exp\left(-\frac{(n-1)t}{\sqrt{2\pi}}\right), \quad (101)
\end{aligned}$$

where the last step uses the upper bound $\mu_n \leq \sqrt{n}$.

With (100) and (101) in hand, we can upper bound the moments of ξ according to

$$\begin{aligned}
\mathbb{E}[\xi^p] &= \int_0^\infty \mathbb{P}\left[\xi \geq t^{1/p}\right]\, dt \\
&\leq \frac{\log(\min(\sec\alpha, n))}{n-1} + \frac{1 + \sqrt{2\pi}}{n-1}\Gamma(p+1),
\end{aligned}$$

where $\Gamma(z) = \int_0^\infty x^{z-1}\exp(-x)\, dx$ is the gamma function. Combining this inequality with (93) and using the basic upper bound $\Gamma(z+1) \leq z^z$ completes the proof of Lemma 17.

Appendix C
Proof of Lemma 20

Lemma 20: Let $A \sim \mathcal{N}(0,1)$ and $B \sim \chi_{n-1}$. For all $p \geq 1$ and $n \geq 2$,

$$\left(\mathbb{E}\left[\left|\sqrt{A^2 + (B - \sqrt{n})^2}\right|^p\right]\right)^{1/p} \leq C\sqrt{p},$$

where $C \leq \sqrt{2} + e^{1/e}$.

Define $\Delta_2 := \sqrt{A^2 + (B - \sqrt{n})^2}$. Minkowski's inequality implies

$$(\mathbb{E}[\Delta_2^p])^{1/p} \leq \mathbb{E}[\Delta_2] + (\mathbb{E}[|\Delta_2 - \mathbb{E}[\Delta_2]|^p])^{1/p} \quad (102)$$

Recall that μ_n of (95) is the mean of the chi-distribution with n degrees of freedom. By Jensen's inequality, $\mathbb{E}[\Delta_2]$ can be upper bounded as

$$\begin{aligned}
\mathbb{E}[\Delta_2]^2 &\leq \mathbb{E}\left[\Delta_2^2\right] \\
&= \mathsf{Var}(A) + \mathsf{Var}(B) + (\mu_{n-1} - \sqrt{n})^2 \\
&= 1 + (n-1) - \mu_{n-1}^2 + (\mu_{n-1} - \sqrt{n})^2 \leq 2,
\end{aligned}$$

where we used that $\sqrt{n-1} \leq \mu_n < \sqrt{n}$. To bound the deviation about the mean, we use the fact that B can be expressed as $B = \|W\|$ where W is an $(n-1)$-dimensional vector with i.i.d. standard Gaussian entries. This allows us to write $\Delta_2 = \psi(A, W)$, where $\psi : \mathbb{R} \times \mathbb{R}^{n-1} \to \mathbb{R}$ is given by $\psi(a, w) = (a^2 + (\|w\| - \sqrt{n})^2)^{1/2}$. The function ψ is 1-Lipschitz continuous because it is the composition of 1-Lipschitz functions. Therefore, we can apply the Tsirelson-Ibragimov-Sudakov Gaussian concentration inequality (see e.g., [56, Theorem 5.2.2]) to conclude that $\Delta_2 - \mathbb{E}[\Delta_2]$ is sub-Gaussian with variance parameter one. Consequently $(\mathbb{E}[|\Delta_2 - \mathbb{E}[\Delta_2]|^p])^{1/p} \leq e^{1/e}\sqrt{p}$. □

Appendix D
Lipschitz Estimation in AWGN Model

Lemma 21: Suppose that (U, X) are random vectors in $\mathbb{R}^m \times \mathbb{R}^n$ with finite second moments. Let $Z = X + \sigma W$ where $\sigma > 0$ is known and $W \sim \mathcal{N}(0, I_n)$ is independent Gaussian noise. Define $f^*(z) = \mathbb{E}[U \mid Z = z]$. For each $\epsilon > 0$ there exists a number L and estimator $f : \mathbb{R}^n \to \mathbb{R}^m$ with $\|f\|_{\text{Lip}} \leq L$ such that

$$\mathbb{E}\left[\|U - f(Z)\|^2\right] \leq \mathbb{E}\left[\|U - f^*(Z)\|^2\right] + \epsilon. \quad (103)$$

Proof: Given $T > 0$, let B be a binary random variable that is equal to zero if $\{\|U\| \vee \|X\| \leq T\}$ and one otherwise. Define $g(z, b) = \mathbb{E}[U \mid Z = z, B = b]$ to be the conditional expectation given (Z, B) and let $f(z) = g(z, 0)$. Two different applications of the law of total variance yields

$$\mathbb{E}\left[\|U - f^*(Z)\|^2\right] = \mathbb{E}\left[\|U - g(Z,B)\|^2\right] + \mathbb{E}\left[\|g(Z,B) - f^*(Z)\|^2\right]$$
$$\mathbb{E}\left[\|U - f(Z)\|^2\right] = \mathbb{E}\left[\|U - g(Z,B)\|^2\right] + \mathbb{E}\left[\|g(Z,B) - f(Z)\|^2\right].$$

Meanwhile, noting that $f(Z) = g(Z, B)$ whenever $B = 0$, we have

$$\begin{aligned}
\|g(Z,B) - f(Z)\| &\leq B\|g(Z,B) - f(Z)\| \\
&\leq B\|g(Z,B)\| + BT
\end{aligned}$$

where the second step follows from the triangle inequality and that fact that $f(y)$ lies in a Euclidean ball of radius T. Combining the above displays with the inequality $(a+b)^2 \leq 2\,a^2 + 2\,b^2$ leads to

$$\begin{aligned}
&\mathbb{E}\left[\|U - f(Z)\|^2\right] - \mathbb{E}\left[\|U - f^*(Z)\|^2\right] \\
&\quad \leq 2\mathbb{E}\left[B\|g(Z,B)\|^2\right] + 2\mathbb{E}[B]\,T^2 \\
&\quad \leq 2\mathbb{E}\left[\mathbf{1}(\|U\| > T)\|U\|^2\right] + 2\mathbb{P}[\|U\| > T]\,T^2.
\end{aligned}$$

By the assumption that $\|U\|$ has finite second moment, this upper bound converges to zero as T increases. Thus, for each $\epsilon > 0$, there exists T large enough such that (103) holds.

Next, we will verify that f has a finite Lipschitz constant. Lemma 22 below implies that the Jacobian of f is given by

$$\frac{\partial f(z)}{\partial z} = \frac{\mathsf{Cov}(U, X \mid Z = z, B = 0)}{\sigma^2}.$$

By the Cauchy-Schwarz inequality and the definition of B, it follows that $\|\mathsf{Cov}(U, X \mid Z = z, B = 0)\| \leq T^2$, uniformly for all z, and thus $\|f\|_{\text{Lip}} \leq T^2/\sigma^2$. □

Lemma 22: Let X be a n-dimensional random vector with $\mathbb{E}[\rho(X)] < \infty$, where $\rho(x)$ is the standard Gaussian density in n dimensions, and let $Y \sim \mathcal{N}(X, \sigma^2 I_n)$. Let $h : \mathbb{R}^n \to \mathbb{R}^m$ be a measurable function such that

$$\phi(y) := \mathbb{E}[h(X) | Y = y], \quad y \in \mathbb{R}^n,$$

is defined for any $y \in \mathbb{R}^n$. The Jacobian of ϕ is given by

$$J_\phi(y) = \frac{1}{\sigma^2}\,\mathsf{Cov}\,(X, h(X) \mid Y = y)\,.$$

Proof of Lemma 22: Set $\rho_\sigma(x) := \rho(x/\sigma)/\sigma$, $P_Y(y) := \mathbb{E}[\rho_\sigma(X-y)]$, and $x_y := \mathbb{E}[X|Y=y]$. From Bayes rule we have

$$\phi(y) = \frac{\mathbb{E}[h(X)\rho_\sigma(X-y)]}{\mathbb{E}[\rho_\sigma(X-y)]} = \frac{\mathbb{E}[h(X)\rho_\sigma(X-y)]}{P_Y(y)}.$$

It follows from [62, Thm. 2.7.1] that we may differentiate with respect to y_j within the expectation. We get

$$\begin{aligned}[J_\phi(y)]_{i,j} &= \frac{\partial \phi_i}{\partial y_j}\\ &= \frac{\mathbb{E}[h_i(X)\rho_\sigma(X-y)(X_j-y_j)]}{\sigma^2\ P_Y(y)}\\ &\quad - \frac{\mathbb{E}[h_i(X)\rho_\sigma(X-y)]\,\mathbb{E}[\rho_\sigma(X-y)(X_j-y_j)]}{\sigma^2\ P_Y^2(y)}\\ &= \frac{\mathbb{E}[\rho_\sigma(X-y)(h_i(X)-\phi_i(y))(X_j-y_j)]}{\sigma^2\ P_Y(y)}\\ &= \frac{\mathbb{E}\left[\rho_\sigma^2(X-y)(h_i(X)-\phi_i(y))X_j\right]}{\sigma^2\ P_Y(y)}.\end{aligned}$$

The last transition implies

$$\frac{\partial\phi_i}{\partial y_j} = \mathbb{E}\left[\frac{\rho_\sigma(X-y)}{\sigma^2\ P_Y(y)}(h_i(X)-\phi_i(y))(X_j-a)\right],$$

for any constant $a \in \mathbb{R}$. It follows that

$$\begin{aligned}J_\phi(y) &= \mathbb{E}\left[(X-x_y)(h(X)-\phi(y))^T\frac{\rho_\sigma(X-y)}{\sigma^2\ P_Y(y)}\right]\\ &\frac{1}{\sigma^2}\int_{\mathbb{R}^n}(x-x_y)(h(x)-\phi(y))^T P_{X|Y}(dx|Y=y)\\ &= \frac{1}{\sigma^2}\mathbb{E}\left[(X-x_y)(h(X)-\phi(y))^T \mid Y=y\right]\\ &= \frac{1}{\sigma^2}\mathsf{Cov}(X,h(X)|Y=y).\end{aligned}$$

APPENDIX E
LIPSCHTIZ CONTINUITY OF AMP

Proposition 23: Let $A \in \mathbb{R}^{n\times d_n}$ be a random matrix with i.i.d. entries $\mathcal{N}(0,1/n)$. Assume that $n/d_n \to \delta \in (0,\infty)$. Denote by $z \to \theta^t_{\mathsf{AMP}}(z)$ the result of t iterations of AMP using a sequence of local non-linearity functions $\{\eta_1,\dots,\eta_t\}$ as in (41) and set $L^{\mathsf{AMP}}_{n,t} = \|\theta^t_{\mathsf{AMP}}\|_{\mathrm{Lip}}$. If η_k is L_k-Lipschitz for $k=1,\dots,t$, then with probability one there exists K_t such that $\sup_n L^{\mathsf{AMP}}_{n,t} \le K_t$.

Proof: Use the tail bound on the maximal eigenvalue of a random matrix with sub-Gaussian entries from [56, Thm 4.4.5] to deduce that there exists a constant c, independent of n, such that

$$\Pr\left(\sqrt{n}\|A\|_2 \le c(\sqrt{n}(1+\sqrt{\delta})+a)\right) \ge 1-2e^{-a^2}.$$

For $C = c(2+\sqrt{\delta})$, define the event

$$E_n = \{\|A\|_2 \le C\}.$$

Using $a=\sqrt{n}$, the Borel-Cantelli Lemma applied to the sequence $\{E_n\}$ implies that the event

$$G := \{\exists n_0 : \|A\|_2 \le C, \quad \forall n \ge n_0\}$$

occurs with probability one. Conditioning on G and given such n_0, we consider the t-th iteration of AMP for reconstructing θ from $z = A\theta + W$ as given by (39) and (40). For η_t applied element-wise to vectors $x, r \in \mathbb{R}^n$, we have

$$\|\eta_t(u)-\eta_t(x)\| \le L_t\|u-x\|,$$

and

$$|\langle\eta_t'(x)\rangle| \le L_t.$$

For $x \in \mathbb{R}^n$, denote by $\theta^t(x)$ and $r^t(x)$ the result of applying t iterations of (40) to x. We have

$$\begin{aligned}&\left\|\theta^{t+1}(x)-\theta^{t+1}(\tilde{x})\right\|\\ &= \left\|\eta_t\left(A^\top r^t(x)+\theta^t(x)\right)-\eta_t\left(A^\top r^t(\tilde{x})+\theta^t(\tilde{x})\right)\right\|\\ &\le L_t\left\|A^\top\left(r^t(x)-r^t(\tilde{x})\right)+\theta^t(x)-\theta^t(\tilde{x})\right\|\\ &\le L_t\left\|A^\top\left(r^t(x)-r^t(\tilde{x})\right)\right\|+L_t\left\|\theta^t(x)-\theta^t(\tilde{x})\right\|\\ &\le L_tC\left\|r^t(x)-r^t(\tilde{x})\right\|+L_t\left\|\theta^t(x)-\theta^t(\tilde{x})\right\|. \qquad (104)\end{aligned}$$

Furthermore,

$$\begin{aligned}&\left\|r^t(x)-r^t(\tilde{x})\right\| \le \|x-\tilde{x}\|+\left\|A\left(\theta^t(x)-\theta^t(\tilde{x})\right)\right\|\\ &\quad+\frac{1}{n}\left\|r^{t-1}(x)\sum_{i=1}^n\eta'_{t-1}\left([A^\top r^{t-1}(x)+\theta^t(x)]_i\right)\right.\\ &\qquad\left.-\,r^{t-1}(\tilde{x})\sum_{i=1}^n\eta'_{t-1}\left([A^\top r^{t-1}(\tilde{x})+\theta^t(\tilde{x})]_i\right)\right\|\\ &\le \|x-\tilde{x}\|+C\left\|\theta^t(x)-\theta^t(\tilde{x})\right\|\\ &\quad+L_{t-1}\left\|r^{t-1}(x)-r^{t-1}(\tilde{x})\right\|. \qquad (105)\end{aligned}$$

We now prove by induction that for $t=1,\dots,$ there exists K_t and R_t such that

$$\left\|\theta^t(x)-\theta^t(\tilde{x})\right\| \le K_t\|x-\tilde{x}\| \qquad (106)$$

$$\left\|r^{t-1}(x)-r^{t-1}(\tilde{x})\right\| \le R_{t-1}\|x-\tilde{x}\|. \qquad (107)$$

For $t=1$, we have

$$\begin{aligned}&\left\|\theta^0(x)-\theta^0(\tilde{x})\right\| = 0,\\ &\left\|u^0(x)-u^0(\tilde{x})\right\| = \|x-\tilde{x}\|,\\ &\left\|\theta^1(x)-\theta^1(\tilde{x})\right\| \le L_0\ C\|x-\tilde{x}\|,\end{aligned}$$

and for the second inequality we take $R_0=0$. Assume now that for all $k=1,\dots,t-1$, there are K_k and R_k such that

$$\begin{aligned}&\left\|\theta^k(x)-\theta^k(\tilde{x})\right\| \le K_k\|x-\tilde{x}\|,\\ &\left\|r^{k-1}(x)-r^{k-1}(\tilde{x})\right\| \le R_{k-1}\|x-\tilde{x}\|.\end{aligned}$$

From (104), (106), and (107), we obtain

$$\begin{aligned}&\left\|\theta^t(x)-\theta^t(\tilde{x})\right\| \le L_{t-1}C\left\|r^{t-1}(x)-r^{t-1}(\tilde{x})\right\|\\ &\quad+L_{t-1}\left\|\theta^{t-1}(x)-\theta^{t-1}(\tilde{x})\right\|,\\ &\le L_{t-1}CR_{t-1}\|x-\tilde{x}\|+L_{t-1}K_{t-1}\|x-\tilde{x}\|\\ &= (L_{t-1}CR_{t-1}+L_{t-1}K_{t-1})\|x-\tilde{x}\|.\end{aligned}$$

From (105), (106), and (107), we obtain

$$\begin{aligned}&\left\|r^{t-1}(x)-r^{t-1}(\tilde{x})\right\| \le \|x-\tilde{x}\|+C\left\|\theta^{t-1}(x)-\theta^{t-1}(\tilde{X})\right\|\\ &\quad+L_{t-2}\left\|r^{t-2}(x)-r^{t-2}(\tilde{x})\right\|\\ &\le (1+C_AK_{t-1}+L_{t-2}R_{t-2})\|x-\tilde{x}\|.\end{aligned}$$

It follows that both (106) and (107) hold with $k = t$.

We have shown that with probability one there exists n_0 such that, for each $t \in \mathbb{N}$, there exists K_t for which

$$\left\|\theta^t(x) - \theta^t(\tilde{x})\right\| \leq K_t \left\|x - \tilde{x}\right\|, \quad \forall n \geq n_0. \tag{108}$$

It follows that for each t, $\sup_n L_{n,t}^{\mathsf{AMP}} \leq K_t$. $\square$

Acknowledgment

The authors would like to thank Cynthia Rush, David Donoho, and Robert Gray for helpful discussions and comments. They would also grateful for the two anonymous reviewers that encouraged us to derive a much stronger version of the main results than we initially reported.

References

[1] A. Kipnis and G. Reeves, "Gaussian approximation of quantization error for estimation from compressed data," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 2029–2033.

[2] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[3] W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 446–472, Jul. 1948.

[4] D. Marco and D. L. Neuhoff, "The validity of the additive noise model for uniform scalar quantizers," *IEEE Trans. Inf. Theory*, vol. 51, no. 5, pp. 1739–1755, May 2005.

[5] R. Zamir and M. Feder, "Rate-distortion performance in coding bandlimited sources by sampling and dithered quantization," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 141–154, Jan. 1995.

[6] D. H. Lee and D. L. Neuhoff, "Asymptotic distribution of the errors in scalar and vector quantizers," *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 446–460, Mar. 1996.

[7] I. Kontoyiannis and R. Zamir, "Mismatched codebooks and the role of entropy coding in lossy data compression," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 1922–1938, May 2006.

[8] J. N. Tsitsiklis, "Decentralized detection by a large number of sensors," *Math. Control, Signals, Syst.*, vol. 1, no. 2, pp. 167–182, Jun. 1988.

[9] Z. Zhang and T. Berger, "Estimation via compressed information," *IEEE Trans. Inf. Theory*, vol. IT-34, no. 2, pp. 198–211, Mar. 1988.

[10] T. S. Han and S. Amari, "Statistical inference under multiterminal data compression," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2300–2324, Oct. 1998.

[11] J. Steinhardt and J. Duchi, "Minimax rates for memory-bounded sparse linear regression," in *Proc. Conf. Learn. Theory*, 2015, pp. 1564–1587.

[12] Y. Zhu and J. Lafferty, "Quantized minimax estimation over sobolev ellipsoids," *Inf. Inference A, J. IMA*, vol. 7, no. 1, pp. 31–82, Mar. 2018.

[13] A. Kipnis and J. C. Duchi, "Mean estimation from one-bit measurements," *CoRR*, vol. abs/1901.03403, Jan. 2019. [Online]. Available: http://arxiv.org/abs/1901.03403

[14] Y. Han, P. Mukherjee, A. Ozgur, and T. Weissman, "Distributed statistical estimation of high-dimensional and nonparametric distributions," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 506–510.

[15] Y. Dagan and O. Shamir, "Detecting correlations with little memory and communication," in *Proc. Conf. Learn. Theory*, 2018, pp. 1145–1198.

[16] B. Szabo and H. van Zanten, "Adaptive distributed methods under communication constraints," 2018, *arXiv:1804.00864*. [Online]. Available: http://arxiv.org/abs/1804.00864

[17] L. P. Barnes, Y. Han, and A. Ozgur, "Lower bounds for learning distributions under communication constraints via Fisher information," *J. Mach. Learn. Res.*, vol. 21, no. 236, pp. 1–30, 2020.

[18] R. Dobrushin and B. Tsybakov, "Information transmission with additional noise," *IEEE Trans. Inf. Theory*, vol. IT-8, no. 5, pp. 293–304, Sep. 1962.

[19] T. Berger, *Rate-Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1971.

[20] J. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 4, pp. 406–411, Jul. 1970.

[21] H. Witsenhausen, "Indirect rate distortion problems," *IEEE Trans. Inf. Theory*, vol. IT-26, no. 5, pp. 518–521, Sep. 1980.

[22] A. Dembo and T. Weissman, "The minimax distortion redundancy in noisy source coding," *IEEE Trans. Inf. Theory*, vol. 49, no. 11, pp. 3020–3030, Nov. 2003.

[23] T. Weissman, "Universally attainable error exponents for rate-distortion coding of noisy sources," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1229–1246, Jun. 2004.

[24] Y. Zhu and J. Lafferty, "Quantized estimation of Gaussian sequence models in Euclidean balls," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3662–3670.

[25] D. L. Donoho and I. M. Johnstone, "Minimax risk over ℓ_p-balls for ℓ_p-error," *Probab. Theory Rel. Fields*, vol. 99, no. 2, pp. 277–303, Jun. 1994.

[26] I. M. Johnstone, "Gaussian estimation: Sequence and multiresolution models," to be published.

[27] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[28] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 45, pp. 18914–18919, Nov. 2009.

[29] V. K. Goyal, A. K. Fletcher, and S. Rangan, "Compressive sampling and lossy compression," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 48–56, Mar. 2008.

[30] R. G. Baraniuk, S. Foucart, D. Needell, Y. Plan, and M. Wootters, "Exponential decay of reconstruction error from binary measurements of sparse signals," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3368–3385, Jun. 2017.

[31] A. Kipnis, G. Reeves, Y. C. Eldar, and A. J. Goldsmith, "Compressed sensing under optimal quantization," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 2148–2152.

[32] M. Leinonen, M. Codreanu, M. Juntti, and G. Kramer, "Rate-distortion performance of lossy compressed sensing of sparse sources," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4498–4512, Oct. 2018.

[33] D. Sakrison, "A geometric treatment of the source encoding of a Gaussian random variable," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 3, pp. 481–486, May 1968.

[34] A. D. Wyner, "Communication of analog data from a Gaussian source over a noisy channel," *Bell Syst. Tech. J.*, vol. 47, no. 5, pp. 801–812, May 1968.

[35] A. Lapidoth, "On the role of mismatch in rate distortion theory," *IEEE Trans. Inf. Theory*, vol. 43, no. 1, pp. 38–47, Jan. 1997.

[36] L. Zhou, V. Y. F. Tan, and M. Motani, "Refined asymptotics for rate-distortion using Gaussian codebooks for arbitrary sources," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3145–3159, May 2019.

[37] P. Delsarte, J.-M. Goethals, and J. J. Seidel, "Spherical codes and designs," in *Geometry and Combinatorics*. Amsterdam, The Netherlands: Elsevier, 1991, pp. 68–93.

[38] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[39] T. Han, "Hypothesis testing with multiterminal data compression," *IEEE Trans. Inf. Theory*, vol. IT-33, no. 6, pp. 759–772, Nov. 1987.

[40] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright, "Information-theoretic lower bounds for distributed statistical estimation with communication constraints," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2328–2336.

[41] A. Kipnis, S. Rini, and A. J. Goldsmith, "The rate-distortion risk in estimation from compressed data," *IEEE Trans. Inf. Theory*, vol. 67, no. 5, pp. 2910–2924, May 2021.

[42] A. Kipnis, A. J. Goldsmith, and Y. C. Eldar, "The distortion-rate function of sampled Wiener processes," *IEEE Trans. Inf. Theory*, vol. 65, no. 1, pp. 482–499, Jan. 2019.

[43] S. T. Rachev and L. Rüschendorf, *Mass Transportation Problems: Theory*, vol. 1. New York, NY, USA: Springer, 1998.

[44] L. Ambrosio, Y. Brenier, G. Buttazzo, and C. Villani, *Optimal Transportation and Applications: Lectures Given at the CIME Summer School Held in Martina Franca, Italy, September 2–8, 2001*. Berlin, Germany: Springer, 2003.

[45] C. Villani, *Optimal Transport: Old and New*, vol. 338. Berlin, Germany: Springer, 2008.

[46] D. Pollard, "Quantization and the method of k-means," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 199–205, Mar. 1982.

[47] T. Linder, "Lagrangian empirical design of variable-rate vector quantizers: Consistency and convergence rates," *IEEE Trans. Inf. Theory*, vol. 48, no. 11, pp. 2998–3003, Nov. 2002.

[48] R. Gray and D. Ornstein, "Block coding for discrete stationary d-continuous noisy channels," *IEEE Trans. Inf. Theory*, vol. IT-25, no. 3, pp. 292–306, May 1979.

[49] R. M. Gray, D. S. Ornstein, and R. L. Dobrushin, "Block synchronization, sliding-block coding, invulnerable sources and zero error codes for discrete noisy channels," *Ann. Probab.*, vol. 8, no. 4, pp. 639–674, Aug. 1980.

[50] R. Zamir and M. Feder, "On lattice quantization noise," *IEEE Trans. Inf. Theory*, vol. 42, no. 4, pp. 1152–1159, Jul. 1996.
[51] M. Raginsky and I. Sason, "Concentration of measure inequalities in information theory, communications, and coding," *Found. Trends Commun. Inf. Theory*, vol. 10, nos. 1–2, pp. 1–246, 2013.
[52] P. Harsha, R. Jain, D. McAllester, and J. Radhakrishnan, "The communication complexity of correlation," in *Proc. 22nd Annu. IEEE Conf. Comput. Complex. (CCC)*, Jun. 2007, pp. 10–23.
[53] P. Cuff, "Distributed channel synthesis," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7071–7096, Nov. 2013.
[54] C. T. Li and A. El Gamal, "A universal coding scheme for remote generation of continuous random variables," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2583–2592, Apr. 2018.
[55] A. J. Stam, "Limit theorems for uniform distributions on spheres in high-dimensional Euclidean spaces," *J. Appl. Probab.*, vol. 19, no. 1, pp. 221–228, Mar. 1982.
[56] R. Vershynin, *High-Dimensional Probability: An Introduction With Applications in Data Science*, vol. 47. Cambridge, U.K.: Cambridge Univ. Press, 2018.
[57] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 764–785, Feb. 2011.
[58] R. M. Gray, *Entropy and Information Theory*, vol. 1. New York, NY, USA: Springer, 2011.
[59] V. Kostina and S. Verdú, "Nonasymptotic noisy lossy source coding," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6111–6123, Nov. 2016.
[60] A. Kipnis, "Fundamental distortion limits of analog-to-digital compression," Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford, CA, USA, 2017.
[61] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 4, pp. 460–473, Jul. 1972.
[62] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*. New York, NY, USA: Springer, 2006.

Alon Kipnis (Member, IEEE) received the Ph.D. degree in electrical engineering from Stanford University in 2017, under the supervision of Andrea Goldsmith. He is currently a Post-Doctoral Research Fellow and a Lecturer with the Department of Statistics, Stanford University, advised by David Donoho. His research interests include mathematical statistics, information theory, and data science. He was a recipient of numerous fellowships and awards, including the Koret Foundation Postdoctoral Fellowship, the Stanford's Electrical Engineering Department Fellowship, and a prize from the Wolf Foundation.

Galen Reeves (Member, IEEE) received the Ph.D. degree in electrical engineering and computer sciences from the University of California at Berkeley in 2011. He was a Post-Doctoral Associate with the Department of Statistics, Stanford University, from 2011 to 2013. In Fall 2013, he joined Duke University as a Faculty Member. He is currently working as an Associate Professor with a joint appointment at the Department of Electrical Computer Engineering and the Department of Statistical Science. His research interests include information theory and high-dimensional statistics. He received the NSF CAREER Award in 2017.