# The Rate-Distortion Risk in Estimation from Compressed Data

Alon Kipnis*, Stefano Rini† and Andrea J. Goldsmith‡

* Department of Statistics, Stanford University, Stanford, CA 94305 USA

‡ Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA

† Department of Electrical Engineering, National Chiao-Tung University, Hsinchu, 30010, Taiwan



**Abstract**

We consider the problem of estimating a latent signal from a lossy compressed version of the data. We assume that the data is generated by an underlying signal and compressed using a lossy compression scheme that is agnostic to this signal. In reconstruction, the underlying signal is estimated so as to minimize a prescribed loss measure. For the above setting and an arbitrary distortion measure between the data and its compressed version, we define the rate-distortion (RD) risk of an estimator as its risk with respect to the distribution achieving Shannon's RD function with respect to this distortion. We derive conditions under which the RD risk describes the risk in estimating from the compressed data. The main theoretical tools to obtain these conditions are transportation-cost inequalities in conjunction with properties of of source codes achieving Shannon's RD function. We show that these conditions hold in various settings, including settings where the alphabet of the underlying signal is finite or when the RD achieving distribution is multivariate normal. We evaluate the RD risk in special cases under these settings. This risk provides an achievable loss in compress-and-estimate settings, i.e., when the data is first compressed, communicated or stored using a procedure that is agnostic to the underlying signal, which is later estimated from the compressed version of the data. Our results imply the following general procedure for designing estimators from datasets undergoing lossy compression without specifying the actual compression technique; train the estimator based on a perturbation of the data according to the RD achieving distribution. Under general conditions, this estimator achieves the RD risk when applied to the lossy compressed version of the data.

## I. Introduction

### *Motivation*

Digital systems are subject to constraints on the number of bits they can store, communicate, or process. Necessarily, any data acquired is compressed in a lossy manner before any procedure to extract information can take place. As a result, the performance of such procedures is intrinsically dictated by the quality of the compression applied to the raw data and, in particular, by the number of bits available for its compressed representation. In the following, we study the effect of lossy compression on the performance of estimation procedures through the

This paper was presented in part at the IEEE International Symposium on Information Theory (ISIT), 2016 [1] and 2017 [2], and at the Information Theory Workshop (ITW), 2015 [3].

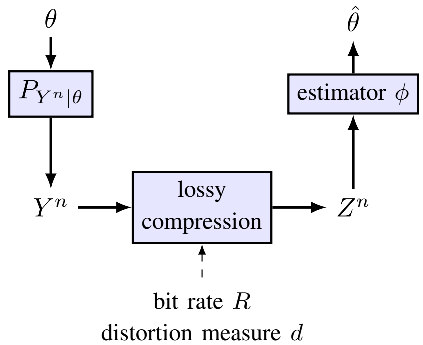


Fig. 1. Estimation from compressed data.

formulation in Figure 1. In this figure, the signal is represented by the latent variable $\theta$, while the data is represented by the $n$-dimensional sequence $Y^n$. The sequence $Y^n$ is compressed at bitrate $R$, i.e., using $nR$ bits, yielding a compressed representation of $Y^n$ denoted by $Z^n$. Finally, the sequence $Z^n$ is utilized to produce an estimate of the underlying signal $\theta = \phi(Z^n)$. For this setting, the optimal tradeoff between compression bitrate and loss in estimation is well known. Classical results in source coding [4], [5], [6] show that this optimal tradeoff is achieved by an *estimate-and-compress* approach. In this approach, we first compute an estimate or a sufficient statistic of the underlying signal $\theta$ from the data $Y^n$ and then compress this estimate according to the available bit budget. In many modern applications, however, the estimate-and-compress approach is infeasible. For example, in most data science and machine learning applications, the data is acquired and compressed for general-purpose use and without any assumption on possible future uses. Infeasibility of estimation before compression arises also in low complexity signal acquisition scenarios, as simple hardware cannot employ complex estimation procedures before quantization. In all scenarios above, the compression of the data in Figure 1 may be designed by considering a distortion measure $d$ between $Y^n$ and $Z^n$ (e.g., Euclidean distance or empirical mutual information) that is agnostic to $\theta$, rather than considering the end estimation task which is unknown to the compressor. In this paper, we strive to characterize the performance in estimating $\theta$ for a class of such lossy compression schemes.

Consider the setting of Figure 1. Given a loss function $\ell$, define the risk of the estimator $\phi$ as

$$\mathbb{E}\left[\ell(\theta, \phi(Z^n))\right]. \tag{1}$$

Equation 1 is the risk with respect to $\ell$ when using $\phi$ in estimating the underlying signal $\theta$ from a compressed version of the data. In this paper we assume that this compressed version is designed to attain a prescribed expected distortion level with respect to the distortion measure $d$. Since the underlying signal is estimated from the compressed representation of the data, we denote this setting as *compress-and-estimate* (CE) source coding. Consequently, (1) is refereed to as the CE risk of $\phi$.

*The Rate-Distortion Risk*

Consider the lossy compression problem of determining the minimal coding rate $R$ for describing the data $Y^n$ such that $\mathbb{E}\left[d(Y^n, Z^n)\right]/n \leq D$. The RD function $R(D)$ and its inverse, the distortion-rate (DR) function $D(R)$, describe the optimal tradeoff between compression rate $R$ and expected distortion $D$ in this problem. A *good* RD code for this problem is said to be any sequence of codes $Y^n \to Z^n$ that satisfies the target distortion $D$ and whose bitrate converges to $R(D)$.

Shannon [7] provided a variational characterization of $R(D)$ as the minimal mutual information over a family of transition probability kernels from the alphabet of the data $Y^n$ to the alphabet of its compressed representation $Z^n$. Whenever this minimizer $P_{Z^{*n}|Y^n}$ is unique, we define the ***RD risk*** of the estimator $\phi$, with respect to a loss function $\ell$, fidelity criteria $d$, and bitrate $R$, as the expected risk with respect to the joint distribution $P_{Y^n|\theta}P_{Z^{*n}|Y^n}$. Namely,

$$D^*_{\phi,d}(R,\theta) \triangleq \mathbb{E}\left[\ell(\theta, \phi(Z^{*n}))\right]. \tag{2}$$

In this paper, we characterize the risk in estimating $\theta$ from the compressed data $Z^n$ using the RD risk in (2). Specifically, we derive conditions under which the CE risk (1) converges to the RD risk (2), and show that these conditions hold in various important cases whenever $Y^n$ is compressed using a good RD code with respect to $d$. The benefit in such characterization is twofold; we can evaluate the risk in estimation from datasets undergoing lossy compression, as well as design estimators from compressed data by minimizing their RD risk.

As an application example where such characterization is useful, assume that $Y^n$ represents a high-resolution image and $Z^n$ is its bitrate $R$ lossy compressed version using some standard image compression technique. Consider the problem of estimating latent information about the image, represented by $\theta$, such as the identities of a person in this image. Traditionally, an estimator for this task depends on the particular lossy compression technique, e.g., it is obtained by training a convolutional neural network on the compressed data. Our results imply that one can derive an estimator $\phi$ that works well for any compression technique, by designing it to estimate $\theta$ from $Z^{*n}$ assuming a random perturbation $\theta \to Z^{*n}$ with distribution $P_{Y^n|\theta}P_{Z^{*n}|Y^n}$. Under conditions derived later in the paper, this estimator is guaranteed to attain its RD risk under any bitrate $R$ lossy compression technique approaching the Shannon compression limit with respect to a given distortion $d$.

*Related Works*

The problem of estimation from compressed data has a long history in source coding. When the underlying signal $\theta$ is an ergodic information source, the problem of encoding $Y^n$ for estimating $\theta$ is known as the *indirect* (a.k.a. *remote* or *noisy*) source coding problem [5, Ch 3.5], [4], [8], [6]. Consequently, the infimum of all achievable distortions $D$ using a compression code of rate not exceeding $R$ is denoted as the indirect distortion-rate function (I-DRF). When $\theta$ is a finite vector of parameters or a distribution governing the generation of the (random) data $Y^n$, the problem of compressing $Y^n$ for recovering $\theta$ is sometimes referred to as *compression for estimation* [9], [10] or *task-specific compression* [11], [12], [13]. These problems, and in particular their multi-terminal versions, have received much recent interest mostly due to their relevance in machine learning [14], [15], [16].

The CE setting departs from these works since the compression is assumed optimal only with respect to the data $Y^n$ rather than the underlying signal or parameter $\theta$. The works of [17], [18] study a multi-terminal version of this problem, and are motivated by the robustness in performance due to encoders oblivious to other system components. These works provided achievable coding bitrate regions under random codebook generation and joint typicality encoding for a prescribed distortion level. To our knowledge, the term CE distortion was coined in [19] and is inspired by the compress-and-forward coding scheme for the relay channel [20]. Special settings with exact CE characterization include compressing the samples of the Wiener process [21] and a Gaussian random walk [22], as well as compressing measurements obtained via a sequence of random linear projections [23]. Recently, the work of [24] provided conditions under which the risk of an estimator from data compressed using a random spherical code equals the risk of the same estimator from a corrupted version of the data using an additive white Gaussian noise (AWGN). This result implies that under this form of compression, the RD risk of any estimator converges to its CE risk (1) whenever the RD achieving distribution $P_{Z^{*n},Y^n}$ is Gaussian. Our current work generalizes this last statement to a broad class of RD achieving distributions and other compression procedures attaining the RD function with respect to the data. Specifically, our setup applies to any situation where the RD achieving distribution satisfies a suitable transportation-cost (TC) inequality [25], [26]. As we explain below, this condition holds quite broadly when the model is separable across dimensions in the sense that the data is i.i.d. and $d$ is a single letter distortion measure.

In what follows, we comment on the connection between the CE setting and other problems in source coding. The work of [27] considered the ability of a wiretap to recover an underlying signal from data undergoing lossy compression, and derived the minimal distortion of the wiretap that one must tolerate in order to guarantee a target distortion with respect to the data. In this case, the encoder chooses a code that is most pessimistic for estimating the underlying signal, hence this distortion from [27] bounds from above the risk of the Bayes-optimal estimator. Consequently, the RD risk provides a lower bound for this distortion whenever the RD equals the risk in estimating from $Z^n$. One may also view the CE scheme as an instance of mismatched source coding considered in [28] and [29]. Indeed, as pointed out in [5], [6], in some situations we can attain the I-DRF by compressing the data according to a fidelity criterion

$$\tilde{d}(y^n, \tilde{z}^n) \triangleq \inf_{\phi} \mathbb{E}\left[\ell(\theta, \phi(z^n)) | Y^n = y^n\right], \tag{3}$$

and thus reducing the indirect source coding problem to a standard one. The mismatch arises when the data is encoded so as to minimize the fidelity criterion $d$ instead of $\tilde{d}$ of (3). Finally, Donoho [30] considered the risk in estimating an $n$-dimensional vector of means $\theta^n = (\theta_1, \ldots, \theta_n)$ when

$$Y^n = \theta^n + \sigma W^n, \tag{4}$$

where $W^n$ is a standard normal vector, using the shortest computer program that produces a summary $Z^n$ of $Y^n$ with expected quadratic distortion $\sigma^2$. He showed that when $\theta^n$ is generated i.i.d. and $Y^n$ is compressed using any optimal lossy compression code that attains MSE equal to the true underlying noise level $\sigma^2$, the coordinates of $Z^n$ are distributed according to the true posterior $P_{\theta_1|Y_1}$ (even though this posterior is never revealed to the compressor!). Consequently, the resulting expected quadratic loss is the risk in sampling from this posterior which is $2\mathbb{E}\left[\mathbb{V}\mathrm{ar}(\theta_1|Y_1)\right]$. Translated to our terminology, Donoho's results says that under the model (4), the CE risk with respect to the quadratic loss of the maximum likelihood estimator $\phi(x) = x$ at coding rate $R = \frac{1}{2}\log_2\left(\frac{\mathbb{V}\mathrm{ar}(Y_1)}{\sigma^2}\right)$ and compression distortion $d = \|\cdot\|^2$, converges to

$$\mathbb{E}_\theta\left[D^*_{\phi,d}(R,\theta)\right] = 2\mathbb{E}\left[\mathbb{V}\mathrm{ar}(\theta_1|Y_1)\right]$$

as the problem dimension $n$ goes to infinity. Our results, specialized to the model (4), imply that $2\mathbb{E}\left[\mathbb{V}\mathrm{ar}(\theta_1|Y_1)\right]$ is the risk of $\phi$ under the quadratic loss when applied to a compressed version of $Y^n$ satisfying the following two conditions: (i) The expected quadratic distortion in this compression between $Y^n$ and $Z^n$ is $\sigma^2$, and (ii) the coding rate approaches $R(\sigma^2)$.

*Contributions*

The main result of this paper, as described in Section IV below, relates to cases where the RD achieving distribution $P_{Z^{*n},Y^n}$ satisfies a transportation-cost (TC) inequality with a constant $c^*$ that is $o(n)$. In this case, the difference between the RD risk of any $L$-Lipschitz estimator of $\theta$ and its CE risk when fed with the output of any rate-$R_n$ code $P_{Z^n|Y^n}$ satisfying the target distortion $D$, is bounded by a constant times $\sqrt{R_n - R(D)}$ plus the relative entropy between the marginals $P_Z^n$ and $P_{Z^{*n}}$. In particular, this difference vanishes whenever this relative entropy vanishes and $R_n \to R(D)$. The last condition indicates that the code utilized for compression is a good rate-distortion code for the target distortion $D$. Next, in Sections V and VI, we show that the TC condition holds quite broadly whenever $P_{Z^{*n},Y^n}$ is a product measure over a discrete domain, and when $P_{Z^{*n}|Y^n}$ is a Gaussian measure on $\mathbb{R}^n$. In these sections, we also evaluate the CE risk in three special cases: (i) when $\theta$ is an i.i.d. binary signal, $Y^n$ is its noisy observation under a binary symmetric channel, and the estimator $\phi$ is the Bayes-optimal estimator (ii) when $\theta$ is a Gaussian i.i.d. signal and each $Y_i$ is an $m$ dimensional vector representing a sequence of its noisy observations, and, finally, (iii) when $\theta$ is a parameter representing the unknown mean of a normal distribution and the estimator $\phi$ is the averaging operation. In the first two examples, (i) and (ii), we compare the CE risk to the optimal source coding performance described by the I-DRF. In example (iii), we derive the relative asymptotic efficiency of the mean estimator from the compressed data $Z^n$ compared to an efficient estimator having access to the uncompressed data $Y^n$.

*Paper Organization*

The remainder of the paper is organized as follows: Section II presents the problem formulation. Section III reviews relevant results in optimal transport theory. Our main results are given in Section IV. In Sections V and VI we specialize our main result to the settings where the data is discrete and Gaussian, respectively. Concluding remarks are provided in Section VII.

*Notation*

We use capital and calligraphic letters, respectively, to denote the random variable $U$ over an alphabet $\mathcal{U}$. Therefore, if not explicitly provided, we implicitly assume a probability space and a topology on $\mathcal{U}$ such that

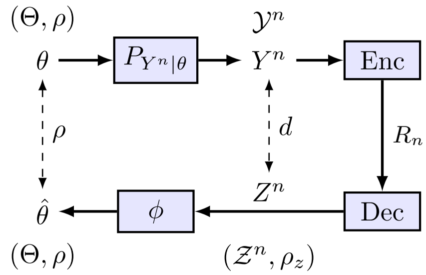


Fig. 2. $\theta$, $Y^n$ and $Z^n$ are the signal, data/measurements and the compressed representation, respectively. $\phi$ is Lipschitz between the metric spaces $(\Theta,\rho)$ and $(\mathcal{Z},\rho_z)$. $d$ is a sub-additive distortion measure on $\mathcal{Y}^n \times \mathcal{Z}^n$. $Z^n$ is obtained by compressing the data $Y^n$ using $nR_n$ bits according to the fidelity criterion $d$. $\phi$ is an estimator of $\theta$ from the compressed representation of the data.

$U$ is measurable. We denote an $n$-length sequence over $\mathcal{U}$ as $u^n = (u_1,\ldots,u_n) \in \mathcal{U}^n$. In some specific case when $\mathcal{U} = \mathbb{R}^m$, we denote $\mathbf{u}^n = (\mathbf{u}_1,\ldots,\mathbf{u}_n) \in \mathbb{R}^{m\times n}$. Matrices are also denoted by bold capital letters, hence, depending on the context, $\mathbf{U}$ may denote either a random vector over $\mathbb{R}^m$ or a matrix. We also denote by $\|\cdot\|$ the Euclidean norm in $\mathbb{R}^n$ and by $\rho_H$ the normalized Hamming distance over a discrete space. The multivariate normal distribution with mean $\mu$ and covariance matrix $\Sigma$ is denoted by $\mathcal{N}(\mu,\Sigma)$. The Bernoulli distribution with probability of success $p$ is denote by $\mathcal{B}(p)$. We use the notation $X^n \overset{\text{iid}}{\sim} P_X$ to denote that $X^n$ is a random sequence obtained by sampling independently $n$ times from the distribution $P_X$. The support of the distribution $P_X$ is denoted as $\mathsf{supp}P_X$.

## II. Problem Formulation

We consider the lossy compression and estimation problem illustrated in Figure 2. This figure is a detailed version of Figure 1; it separates the lossy compression operation into an encoder and decoder and indicates the alphabet and loss function associated with some of the variables. The data is represented by the random $n$-dimensional vector $Y^n$ over the alphabet $\mathcal{Y}^n$. The distribution of $Y^n$ depends on an underlying signal or a parameter vector $\theta \in \Theta$ through a conditional distribution kernel $P_{Y^n|\theta}$. The encoder encodes the data $Y^n$ using $nR_n$ bits, and the decoder produces a bit-restricted representation $Z^n$ of the data $Y^n$. This representation is over the alphabet $\mathcal{Z}^n$. Throughout this paper, we assume that for any $\theta$, the induced $Y^n$ is a stationary ergodic process and that the metric spaces $(\Theta,\rho)$ and $(\mathcal{Z}^n,\rho_z)$ are complete, and hence Polish.

For an estimator $\phi : \mathcal{Z}^n \to \Theta$ mensurable with respect to the Borel $\sigma$-algebras on $(\mathcal{Z}^n,\rho_z)$, its $p^{\text{th}}$ CE risk is defined as

$$\mathbb{E}\left[\rho^p(\theta,\phi(Z^n))\right], \tag{5}$$

where the expectation is with respect to $P_{Y^n|\theta}$ and possible randomness in the encoding and decoding $Y^n \to Z^n$ as explained below. Whenever a prior $P_\theta$ on $\Theta$ is given, the expectation above is also with respect to this prior, in which case (5) is refereed to as the $p^{\text{th}}$ Bayes risk with respect to $P_\theta$. The CE distortion-rate function is defined as the Bayes-optimal risk given $Z^n$, namely

$$\inf_{\phi} \mathbb{E}\left[\rho^p(\theta,\phi(Z^n))\right]. \tag{6}$$

Note that the loss function $\ell$ in (1) is the $p^{\mathrm{th}}$ power of the metric $\rho$ in the notation of (5). Our setting is restricted to loss functions of this form.

Given $n, M_n \in \mathbb{N}$, a *lossy-compression $n$-block code* $(f_{\mathsf{enc}}, f_{\mathrm{dec}}, M_n)$, or simply a $n$-block code, is defined as the pair of mappings

$$f_{\mathsf{enc}} : \mathcal{Y}^n \to \{1, \dots, M_n\} \tag{7a}$$

$$f_{\mathrm{dec}} : \{1, \dots, M_n\} \to \mathcal{Z}^n. \tag{7b}$$

We define $R_n \triangleq \log_2(M_n)/n$ as the *rate* of the code in bits. In the following, we consider codes for which $f_{\mathsf{enc}}$ and $f_{\mathrm{dec}}$ are random such that the mapping

$$Y^n \to Z^n = f_{\mathrm{dec}}(f_{\mathsf{enc}}(Y^n)) \tag{8}$$

is measurable[1]. In this case, it is more convenient to refer to the code $(f, g, M_n)$ by the transition probability kernel $P_{Z^n|Y^n}$ on $\mathcal{Y}^n \times \mathcal{Z}^n$ defined by (8). Namely,

$$\mathbb{P}(f_{\mathrm{dec}}(f_{\mathsf{enc}}(Y^n)) \in A | Y^n = y^n) = \int_A P_{Z^n|Y^n}(\mathrm{d}z^n | y^n)$$

for any Borel measurable $A \subset \mathcal{Z}^n$. The *output distribution* of the code $P_{Z^n|Y^n}$ is its marginal distribution $P_{Z^n}$ over $\mathcal{Z}^n$, i.e.,

$$P_{Z^n}(\mathrm{d}z^n) = \int_{\mathcal{Y}^n} P_{Z^n|Y^n}(\mathrm{d}z^n | y^n) P_{Y^n}(\mathrm{d}y^n).$$

For each $n \in \mathbb{N}$, the distortion between $Y^n$ and $Z^n$ is measured using the function[2] $d : \mathcal{Y}^n \times \mathcal{Z}^n \to [0, \infty)$, which is further assumed sub-additive in the sense that

$$d(y^{k+m}, z^{k+m}) \leq d(y^k, z^k) + d(y_{k+1}^{m+k}, z_{k+1}^{m+k})$$

for any $y^{m+k} \in \mathcal{Y}^{m+k}$, $z^{m+k} \in \mathcal{Z}^{m+k}$ with $k \leq m$, where $u_k^m = (u_k, \dots, u_m)$. We also set

$$d_{\min} \triangleq \inf_{y \in \mathcal{Y}, z \in \mathcal{Z}} d(y, z)$$

so that the sub-additivity of $d$ also implies $\inf_{z^n, y^n} d(y^n, z^n) \leq n d_{\min}$.

A code $(f_{\mathsf{enc}}, f_{\mathrm{dec}}, M)$ is said to be *$D$-admissible* if its expected distortion satisfies

$$\mathbb{E}\left[d\left(Y^n, Z^n\right)\right] \leq nD.$$

For a target distortion $D \geq d_{\min}$ and a distribution $P_{Y^n}$ on $\mathcal{Y}^n$, Shannon's RD function of $P_{Y^n}$ is defined as [7]

$$\mathcal{R}_{Y^n}(D) \triangleq \min_{\substack{P_{Z^n|Y^n} \\ \mathbb{E}[d(Y^n, Z^n)] \leq Dn}} I(Y^n; Z^n), \tag{9}$$

[1]We only specify the sigma algebras on $(\mathcal{Z}^n, \rho_z)$ and $(\Theta, \rho)$ which are the Borel sigma algebras. The sigma algebra on $\mathcal{Y}^n$ is implicit in our notation.

[2]The dependency of $d$ from $n$ is not explicitly indicated

where the minimum is over all conditional distribution kernels $P_{Z^n|Y^n}$ satisfying the prescribed distortion constraint. We also define the single letter RD function with respect to $Y^n$ as

$$R(D) = \lim_{n\to\infty} \frac{1}{n}\mathcal{R}_{Y^n}(D). \tag{10}$$

Note that the sub-additivity of $d$ guarantees that this limit in the right-hand side (RHS) of (10) exits [31].

The single letter distortion-rate function $D(R)$ is defined as the inverse function of $R(D)$ in (10) for $R$ smaller than the entropy rate of $Y^n$ and zero otherwise. In particular, $D(R)$ is defined for any $R \in [0,\infty)$ whenever $P_{Y^n|\theta}$ is absolutely continuous with respect to the Lebesgue measure. For a wide class of distributions $P_{Y^n}$, $R(D)$ describes the infimum rate of all $D$-admissible codes, and $D(R)$ describes the infimum over $\mathbb{E}\left[d(Y^n,Z^n\right]/n$ for all codes $P_{Z^n|Y^n}$ with $R_n \leq R$. In other words, for such $P_{Y^n}$, there exists a sequences of codes $P_{Z^n|Y^n}$ such that

$$\mathbb{E}\left[d\left(Y^n,Z^n\right))\right] \leq nD \quad \text{while} \quad \lim_{n\to\infty} R_n = R(D). \tag{11}$$

A sequence of codes satisfying (11) is refereed to as a *good* rate-distortion code for $(Y^n,d)$ at distortion level $D$ [32], [33], [34], [35], [36]. Throughout this paper we consider the following condition:

(C1) The minimum in (9) is achieved by $P_{Z^{*n}|Y^n}$ which also satisfies $\mathbb{E}\left[d(Y^n,Z^n)\right] = nD$.

This condition is relatively mild and usually holds whenever $d_{\min} \leq D \leq \inf_{z^n} \mathbb{E}\left[d(Y^n,z^n)\right]$, i.e., for distortion levels for which that source coding problem with respect to $Y^n$ is non-trivial.

Under (C1), Kostina and Verdu [36] provided a characterization of conditional relative entropy between any good code $P_{Z^n,Y^n}$ and the RD achieving distribution $P_{Z^{*n},Y}$.

**Proposition 2.1 (Properties of good RD codes):** [36, Thm. 1] Assume that (C1) holds. For any code $P_{Z^n|Y^n}$ with $\mathsf{supp}P_{Z^n} \subset \mathsf{supp}P_{Z^{*n}}$ and $\mathbb{E}\left[d(Y^n,Z^n)\right] \leq nD$, we have

$$D\left(P_{Y^n|Z^n}||P_{Y^n|Z^{*n}} \mid P_{Z^n}\right) = I(Y^n;Z^n) - \mathcal{R}_{Y^n}(D),$$

for $\mathcal{R}_{Y^n}(D)$ in (9).

We recall that the relative entropy between two distributions $P_U$ and $P_V$ on $\mathcal{X}$, with $P_U \ll P_V$, is defined as

$$D(P_U||P_V) \triangleq \int_{\mathcal{X}} \log_2\left(\frac{\mathrm{d}P_U}{\mathrm{d}P_V}(x)\right)\mathrm{d}P_U(x),$$

where $\frac{\mathrm{d}P_U}{\mathrm{d}P_V}$ is the Radon-Nikodym derivative of $P_U$ with respect to $P_V$.

For an estimator $\phi : \mathcal{Z}^n \to \Theta$ measurable with respect to the Borel $\sigma$-algebras on $(\mathcal{Z}^n,\rho_z)$ and $(\Theta,\rho)$, we define its $p^{\text{th}}$ RD risk with respect to $\rho$ as

$$D^*_\phi(R,\theta) \triangleq D^*_{\phi,d,\rho^p}(R,\theta) \triangleq \mathbb{E}\left[\rho^p(\theta,\phi(Z^{*n})\right], \tag{12}$$

where the expectation is with respect to $P_{Y^n|\theta}P_{Z^{*n},Y^n}$. Whenever a prior $P_\theta$ is provided, we consider the Bayes RD risk

$$D^*_\phi(R) \triangleq D^*_{\phi,d,\rho^p,P_\theta}(R) \triangleq \mathbb{E}\left[\rho^p(\theta,\phi(Z^{*n})\right],$$

and the Bayes-optimal RD risk

$$D^*(R) \triangleq D^*_{d,\rho^p,P_\theta}(R) \triangleq \inf_{\phi} \mathbb{E}\left[\rho^p(\theta,\phi(Z^{*n})\right], \tag{13}$$

where the expectation is with respect to $P_\theta P_{Y^n|\theta} P_{Z^{*n},Y^n}$.

The goal of this work is to establish the connection between $D^*_\phi(R,\theta)$ and the risk of $\phi$ given by (5) under an encoding of $Y^n$ that defines a good code for $R(D)$.

Throughout this paper we make frequent use in the following assumptions on the code $P_{Z^n|Y^n}$:

(A1) For $P_{Y^n}$-almost every $y^n$, $P_{Z^n|Y^n=y^n}$ is absolutely continuous with respect to $P_{Z^{*n}|Y^n=y^n}$. This is indicated as $P_{Z|Y=y} \ll P_{Z^*|Y=y}$.

(A2) The output distribution of $P_{Z^n|Y^n}$ satisfies

$$\lim_{n\to\infty} \frac{1}{n} D(P_{Z^n}||P_{Z^{*n}}) = 0.$$

Both (A1) and (A2) do not hold, in general, for deterministic codes. Indeed, for such codes, $P_{Z^n|Y^n=y^n}$ is a mass distribution at $z^n = f(y^n)$ hence (A1) cannot hold when $P_{Z^n|Y^n}$ is absolutely continuous with respect to the Lebesgue measure. Examples for deterministic codes violating (A2) are provided in [36] and [37, Prop. 2]. Nevertheless, (A1) and (A2) hold in the following important random coding scenarios:

(i) The codewords are randomly drawn from $P_{Z^{*n}}$ and encoding is done using joint typically.

(ii) Whenever $Y^n$ is i.i.d. over a discrete alphabet, the distortion is separable, and the output distribution of the code $P_{Z^n}$ is uniform over the codewords [35], [36].

(iii) Whenever $Y^n$ is standard normal, the codewords are drawn from a spherically symmetric distribution in $\mathbb{R}^n$ that is absolutely continuous with respect to the Lebesgue measure, and encoding is done by selecting the codeword of maximal cosine similarity with the input [24], [38].

All cases (i)-(iii) incorporate codes attaining the optimal source coding performance. It should be noted that by analyzing random coding results we confirm the existence of deterministic codes satisfying the average performance guarantee. This is relevant since in practice deterministic codes are needed.

The problem formulation above encompasses two classical settings: the case in which $\theta$ is an unknown parameter and the case in which $\theta$ is an information source. The two scenarios are introduced next, together with some examples that are revisited later in the paper.

### A. Parameter Estimation Setting

One of the scenarios encompassed by Figure 2 is the one in which $\theta$ is an unknown parameter for a family of distributions to be estimated for lossy compressed observations. The focus in this scenario is with respect to the efficiency of $\phi$ as an estimator of $\theta$ from the compressed data compared to an efficient estimator from the uncompressed data $Y^n$ [39], [10], [40], [16]. We can explore this efficiency using the setting in Figure 2 by letting $\rho(\theta,\hat{\theta}) = \sqrt{n}|\theta-\hat{\theta}|$, that is by considering the scaling of the estimation error with the number of observations available.

*1) Scalar Gaussian Location Example:* One example of the parameter estimation setting is a normal location model. In this example, $\theta$ is the unknown mean of a normal distribution and the data consists of independent draws from this distribution, namely:

$$Y_i = \theta + \sigma W_i, \quad i = 1, \ldots, n,$$

where $W_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $\theta \in \mathbb{R}$. The data is compressed using some procedure aiming to minimize the quadratic distortion:

$$d(y^n, z^n) = \|y^n - z^n\|^2 = \sum_{i=1}^{n} (y_i - z_i)^2,$$

i.e., the squared Euclidean distance, between $Y^n$ and its compressed representation $Z^n$. Such compression is said to be optimal with respect to the data if $\|Y^n - Z^n\|^2/n$ approaches $D(R) = \sigma^2 2^{-2R}$, which is the quadratic Gaussian DRF with variance $\sigma^2$. Importantly, our setting does not require the description of the actual compression procedure, so long as it is optimal.

## B. Memoryless Information Source Setting

Another example of the setting in Figure 2 is the case where $\theta$ represents $n$ independent samples from a distribution $P_X$ over $\mathcal{X}$. In this case the underlying $\theta$ is the sequence $X^n$. Whenever the $P_{Y^n|\theta} = P_{Y^n|X^n}$ decomposes as the product $P^n_{Y_i|X_i}$ and the distortion $d$ is additive, the Bayes-optimal RD risk corresponds to the Bayes-optimal risk in estimating $X_i$ from $Z_i^*$, where $P_{X_i|Z_i^*} = P_{X_i|Y_i} P^*_{Y_i|Z_i}$ and $P_{Y_i|Z_i^*}$ is the single-letter rate-distortion achieving distribution of $Y^n$ under a prescribed distortion. In Sections V and VI we explore the case in which $\theta$ is of the aforementioned form in the following specific examples:

*1) Binary signal observed under bitflip noise :* In this setting $X^n$ is an $n$-dimensional Bernoulli i.i.d. sequence while the data $Y^n$ is obtained by passing $X^n$ through a binary symmetric channel as in

$$Y_i = X_i \oplus W_i, \quad i = 1, \ldots, n, \tag{14}$$

where $\oplus$ denotes addition modulo 2 and $W^n$ is also a Bernoulli i.i.d. sequence. Namely, in this setting, we have $X = \{0,1\}^n$ and $\mathcal{Z} = \mathcal{Y} = \{0,1\}$. The estimation of $X^n$ from the compressed version of $Y^n$ is according to the metric $\rho$, which in this case is taken to be the normalized Hamming distance:

$$\rho_H(u^n, v^n) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{u_i \neq v_i\}}. \tag{15}$$

The distortion measure $d$ is also taken as the Hamming distance: namely, we consider the estimation of the binary sequence $X^n$ from $Z^n$ under the Hamming distance, when the latter is obtained by compressing $Y^n$ so as to attain a target expected Hamming distance between $Y^n$ and $Z^n$ under the prescribed rate constraint.

*2) Single input multiple output Gaussian Channel :* In this setting $X^n$ is generated by sampling from the standard Gaussian distribution and each data point is an $m$-dimensional jointly Gaussian vector

$$\mathbf{Y}_i = (Y_{i,1}, \ldots, Y_{i,m}), \tag{16}$$

where

$$Y_{i,j} = \sqrt{\gamma_j} X_i + W_{i,j}, \quad j = 1, \dots, m, \quad i = 1, \dots, n. \tag{17}$$

Here $W_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $\gamma_1, \dots, \gamma_m$ are positive constants determining the channel strength between each realization of $X_i$ and the observation at the $j$th output channel, $Y_{i,j}$. Accordingly, we have $\mathcal{Y} = \mathcal{Z} = \mathbb{R}^m$. Further, we assume that the data is compressed so as to minimize the quadratic distortion between $Y^n$ and its compressed representation $Z^n$, i.e.,

$$\mathbb{E}\left[d(Y^n, Z^n)\right] = \mathbb{E}\left[\|Y^n - Z^n\|^2\right].$$

In this scenario, the RD achieving distribution $P_{\mathbf{Z}^{*n}, \mathbf{Y}^n}$ is the product distribution of two $m$-dimensional multivariate Gaussian according to Pinsker and Kolmogorov waterfilling formula [41].

## III. Transportation Theory Interlude

Let us review, in this section, some transportation theoretic notions that will be useful in the remainder of the paper. Transportation cost-information inequalities are expressed through the Wasserstein distance: this distance is widely used in source coding and ergodic theory since many relevant quantities are continuous with respect to it, such as optimal quantizer performance and Shannon distortion-rate function [42], [43]. Transportation cost or distance-divergence inequalities were discovered by Marton [44] as an easy proof for the concentration of measure property. This property is a generalization of the *blowing up* property [45] which was used for providing various converse channel coding results [46]. We refer to [25], [26] for background and applications of transport theory in information theory and probability.

**Definition 3.1 (Wasserstein distance):** Let $P_U$, $P_{\bar{U}}$ be two Borel probability measures with respect to a metric $\mu$ on the Polish space $\mathcal{U}$. The $p$-Wassesrstein distance between $P_U$ and $P_{\bar{U}}$, for $p \geq 1$, is defined as

$$W_p(P_U, P_{\bar{U}}) \triangleq \left( \inf_{P_{U,\bar{U}}} \mathbb{E}\left[\mu^p(U, \bar{U})\right] \right)^{1/p},$$

where the infimum is over all joint probability distributions $P_{U,\bar{U}}$ on the product space $\mathcal{U} \times \mathcal{U}$ with marginals $P_U$ and $P_{\bar{U}}$.

Similarly to Definition 3.1, the $p$-Wasserstein distance between $U$ and $\bar{U}$ conditioned on $V$ is defined as

$$W_p(P_{U|V}, P_{\bar{U}|V}) \triangleq \left( \inf_{P_{V,U,\bar{U}}} \mathbb{E}\left[\mu^p(U, \bar{U})\right] \right)^{1/p}, \tag{18}$$

where $P_{V,U}, P_{V,\bar{U}}$ are defined over the product space $\mathcal{V} \times \mathcal{U}$ and the infimum is over all joint probability distributions $P_{U,\bar{U},V}$ over the product space $\mathcal{U} \times \mathcal{U} \times \mathcal{V}$ with marginals $P_{V,U}$ and $P_{V,\bar{U}}$.

The motivation for using the Wasserstein distance in our setting follows from the following continuity argument, stated here for an arbitrary triplet of random vectors $(X, Y, Z)$ such that $(\mathcal{X}, \rho)$ and $(\mathcal{Z}, \rho_z)$ are metric spaces:

**Proposition 3.1 (continuity of risk in Wassesrstein distance):** Let $P_{X,Y}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$ and let $P_{Z|Y}$ and $P_{\bar{Z}|Y}$ be two conditional distributions kernels from $\mathcal{Y}$ to $\mathcal{Z}$. For any $L$-Lipschitz $\phi : (\mathcal{Z}, \rho_z) \to (\mathcal{X}, \rho)$

and $p \geq 1$ such that $\mathbb{E}\left[\rho^p\left(X, \phi(Z)\right)\right] < \infty$ and $\mathbb{E}\left[\rho^p\left(X, \phi(\bar{Z})\right)\right] < \infty$, we have

$$\left|\left(\mathbb{E}\left[\rho^p\left(X, \phi(Z)\right)\right]\right)^{1/p} - \left(\mathbb{E}\left[\rho^p\left(X, \phi(\bar{Z})\right)\right]\right)^{1/p}\right| \leq L W_p(P_{Z|Y}, P_{\bar{Z}|Y}).$$

*Proof:* For any Markov chain $X \to Y \to (Z, Z^*)$, the triangle inequality and Lipschitz continuity of $\phi$ imply

$$\begin{aligned}\left(\mathbb{E}\left[\rho^p\left(X, \phi(Z)\right)\right]\right)^{1/p} &\leq \left(\mathbb{E}\left[\rho^p\left(X, \phi(\bar{Z})\right)\right]\right)^{1/p} + \left(\mathbb{E}\left[\rho^p\left(\phi(\bar{Z}), \phi(Z)\right)\right]\right)^{1/p} \\ &\leq \left(\mathbb{E}\left[\rho^p\left(X, \phi(\bar{Z})\right)\right]\right)^{1/p} + L\left(\mathbb{E}\left[\rho_z^p(\bar{Z}, Z)\right]\right)^{1/p}.\end{aligned} \tag{19}$$

One side of the inequality is now obtained by taking the minimum over all joint distributions $P_{Y,Z,Z^*}$ with marginals $P_{Y,Z}$ and $P_{Y,Z^*}$. The second inequality follows by interchanging the role of $Z$ and $Z^*$ in (19). □

Probability measures on a metric space whose Wasserstein distance with respect to any other measure is bounded by the relative entropy is said to satisfy a transportation cost inequality, as per the following definition:

**Definition 3.2 (transportation-cost (TC) inequality [26, Def. 22.1], [25, Def. 3.4.2]):** Let $(\mathcal{X}, \rho)$ be a Polish space and $p \in [1, \infty)$. It is said that a probability measure $P$ on $(\mathcal{X}, \rho)$ satisfies the TC inequality with constant $c > 0$ if, for every Borel probability measure $Q$ on $(\mathcal{X}, \rho)$,

$$W_p(P, Q) \leq \sqrt{cD(Q||P)}. \tag{20}$$

When (20) holds, we say that $(\mathcal{X}, \rho, P)$ is $\mathsf{TC}(c)$.

The following result by Marton shows that any product measure on a discrete space is $\mathsf{TC}(1/2n)$ with respect to the normalized Hamming distance on this space.

**Proposition 3.2:** [47] Let $\mathcal{Z}$ be a countable set and let $Q_{Z^n}$ and $P_{Z^n}$ be two probability measures on $\mathcal{Z}^n$. Consider the space $(\rho_H, \mathcal{Z}^n)$ where $\rho_H$ is the normalized Hamming distance. If $Q_{Z^n}(z^n) = \prod_{i=1}^n Q_{Z_1}(z_i)$, then

$$W_1(Q_{Z^n}, P_{Z^n}) \leq \sqrt{\frac{1}{2n} D(P_{Z^n}||Q_{Z^n})}. \tag{21}$$

The following result by Talagrand [48] shows that the standard Gaussian measure in $\mathbb{R}^n$ is $\mathsf{TC}(2)$ with respect to the Euclidean distance in $\mathbb{R}^n$.

**Proposition 3.3:** [48, Thm. 1.1] Assume $\mu(z^n, \bar{z}^n) = \|z^n - \bar{z}^n\|$. Let $P^n$ be a measure on $\mathbb{R}^n$, absolutely continuous with respect to

$$\gamma^n(\mathrm{d}z^n) = (2\pi)^{-n/2} e^{-\frac{1}{2}\|z^n\|^2} \mathrm{d}z^n,$$

then

$$W_2(P^n, \gamma^n) \leq \sqrt{2D(P^n||\gamma^n)}.$$

A straightforward extension of Proposition 3.3 to the case where $\gamma^n$ is a shifted and scaled Gaussian measure is as follows:

**Proposition 3.4:** Assume $\mu(z^n, \bar{z}^n) = \|z^n - \bar{z}^n\|$. Let $P^n$ be a Borel measure on $\mathbb{R}^n$ absolutely continuous with respect to

$$\overline{\gamma}^n \triangleq \gamma^n(\mathrm{d}z^n|y^n) = \frac{(2\pi)^{-n/2}}{\sqrt{\det(\Sigma)}} \exp\{-\frac{1}{2}(z^n - y^n)^*\Sigma^{-1}(z^n - y^n)\}\mathrm{d}z^n,$$

where $\Sigma$ is positive definite. Then

$$W_2^2(P^n, \overline{\gamma}^n) \leq 2\left\|\Sigma\right\|_2^2 D(P^n||\overline{\gamma}^n), \tag{22}$$

where $\left\|\Sigma\right\|_2$ is the operator norm of $\Sigma$.

*Proof:* See Appendix A. □

## IV. Main Result

In this section we present our main result, that is, a characterization of the difference between the CE risk (5) and the RD risk (2) whenever $Y^n$ is compressed using a good rate-distortion code.

**Theorem 4.1:** Consider a distribution $P_{\theta,Y^n}$ on $\Theta \times \mathcal{Y}^n$, a sub-additive distortion measure $d$ on $\mathcal{Y}^n \times \mathcal{Z}^n$, and a distortion level $D > d_{\min}$. Assume that the following conditions hold:

- The minimum in (9) is achieved by $P_{Z^{*n}|Y^n}$ which also satisfies $\mathbb{E}\left[d(Y^n, Z^n)\right] = nD$.
- For $P_{Y^n}$-almost every $y^n \in \mathcal{Y}$, there exists $c > 0$ such that $P_{Z^{*n}|Y^n=y^n}$ is $\mathsf{TC}(c)$.

Let $P_{Z^n|Y^n} : \mathcal{Y}^n \to \mathcal{Z}^n$ be any $D$-admissible random code satisfying (A1). Then for any $L$-Lipschitz estimator $\phi : (\mathcal{Z}^n, \rho_z) \to (\Theta, \rho)$,

$$\begin{aligned} &\left|\left(\mathbb{E}\left[\rho^p\left(\theta, \phi(Z^n)\right)\right]\right)^{1/p} - \left(D^*_\phi(R(D), \theta)\right)^{1/p}\right| \\ &\leq L\sqrt{n\mathbb{E}\left[c(Y^n)\right]\left(R_n - R(D) + \frac{1}{n}D(Z^n||Z^{*n})\right)}, \end{aligned} \tag{23}$$

provided $D_\phi(Z^n, \theta) < \infty$ and $D^*_\phi(R(D), \theta) < \infty$.

*Proof:* For the sake of brevity and clarity, we omit the superscripts $n$ as it can be deduced from the context. Assumption (A2) implies that $\mathsf{supp}P_{Z|Y=y} \subseteq \mathsf{supp}P_{Z^*|Y=y}$, and hence

$$D(P_{Z,Y}||P_{Y|Z^*}P_Z) = D(P_{Z,Y}||P_{Y,Z^*}) - D(P_Z||P_{Z^*}),$$

where

$$D(P_{Z,Y}||P_{Y,Z^*}) = \int_{\mathcal{Y}} D(P_{Z|Y=y}||P_{Z^*|Y=y})\mathrm{d}P_Y(y).$$

Therefore, if for almost every $y$,

$$W_p\left(P_{Z|Y=y}, P_{Z^*|Y=y}\right) \leq \sqrt{c(y)D(P_{Z|Y=y}||P_{Z^*|Y=y})},$$

then, by Proposition 3.1,

$$
\begin{aligned}
&= \left| \left( \mathbb{E} \left[ \rho^p \left( \theta, \phi(Z) \right) \right] \right)^{1/p} - \left( D^*_{\phi}(R(D), \theta) \right)^{1/p} \right| \\
&\leq L W_p(P_{Z|Y}, P_{Z^*|Y}) = L \int_{\mathcal{Y}} W_p \left( P_{Z|Y=y}, P_{Z^*|Y=y} \right) \mathrm{d}P_Y(y) \\
&\leq L \int_{\mathcal{Y}} \sqrt{c(y) D(P_{Z|Y=y} || P_{Z^*|Y=y})} \mathrm{d}P_Y(y) \\
&\leq L \sqrt{\mathbb{E} \left[ c(Y) \right]} \sqrt{D(P_{Z,Y} || P_{Z^*,Y})}, \qquad (24)
\end{aligned}
$$

where (24) follows from the Cauchy-Schwartz inequality. Next, we use Theorem 2.1 to obtain

$$
\begin{aligned}
D(P_{Z,Y} || P_{Z^*,Y}) qq &= D(P_{Y|Z} || P_{Y|Z^*} | P_Z) + D(P_Z || P_{Z^*}) \\
&\leq I(Y^n; Z^n) - \mathcal{R}_{Y^n}(D) - D(P_Z || P_{Z^*}).
\end{aligned}
$$

This implies

$$
\begin{aligned}
&\left| \left( \mathbb{E} \left[ \rho^p \left( \theta, \phi(Z^n) \right) \right] \right)^{1/p} - \left( D^*_{\phi}(R(D), \theta) \right)^{1/p} \right| \\
&\leq L \sqrt{\mathbb{E} \left[ c(Y^n) \right] \left( I(Y^n; Z^n) - \mathcal{R}_{Y^n}(D) + D(Z^n || Z^{*n}) \right)}. \qquad (25)
\end{aligned}
$$

Equation (23) now follows from (25) by the definition of $\mathcal{R}_{Y^n}(D)$ in (9) and the data processing inequality as any $D$-admissible rate $R_n$ code also satisfies $I(Y^n, Z^n) \leq nR_n$. $\square$

Theorem 4.1 bounds the distance between the $p^{\text{th}}$ risk of any Lipschitz estimator of $\theta$ from the compressed data and the $p^{\text{th}}$ RD risk of this estimator, using the expected TC constant of the measure $P_{Z^{*n}|Y^n=y^n}$ over the metric space $(\mathcal{Z}^n, \rho)$. A typical use case of Theorem 4.1 that we explore in the next section is situations where $L = \mathcal{O}(1)$ and $\mathbb{E}\left[c(Y)\right] = \mathcal{O}(1/n)$. In such cases, Theorem 4.1 provides sufficient conditions so that, for good rate-distortion codes satisfying (11) and (A2), we have

$$
n^{-1/p} \left| \left( \mathbb{E} \left[ \rho^p \left( \theta, \phi(Z^n) \right) \right] \right)^{1/p} - \left( D^*_{\phi}(R(D), \theta) \right)^{1/p} \right| \to 0.
$$

In the next two sections we consider special cases where these conditions are met.

## V. Separable discrete information sources

In this section we focus on special cases of the setting of Section II-B in which the space $\mathcal{Z}$ is discrete and the problem is separable across the information source dimension. As shown next, this setting satisfies many of the conditions of Theorem 4.1, leading to a precise characterization of the RD risk when the data is encoded using a good rate-distortion code as defined in Prop. 2.1.

## A. Estimation in Discrete Separable Models

Let us denote as a *separable information model* the setting of Section II-B in which

$$P_{Y^n|X^n}(y^n, x^n) = \prod_{i=1}^{n} P_{Y_i|X_i}(y_i, x_i) = \prod_{i=1}^{n} P_{Y_1|X_1}(y_i, x_i), \tag{26}$$

and

$$\rho(x^n, \hat{x}^n) = \frac{1}{n}\sum_{i=1}^{n} \rho(x_i, \hat{x}_i) \tag{27a}$$

$$d(y^n, z^n) = \sum_{i=1}^{n} d(y_i, z_i). \tag{27b}$$

Due to the additivity of the mutual information over product measures, it follows that $P_{Z^{*n},Y^n}$ that satisfies (9) with equality is also in product form, that is

$$P_{Z^{*n},Y^n}(z^n, y^n) = \prod_{i=1}^{n} P_{Z_i^*,Y_i}(z_i, y_i),$$

and thus

$$R(D) = \inf_{\substack{P_{Z_1|Y_1} \\ \mathbb{E}[d(Y_1,Z_1)]\leq D}} I(Y_1; Z_1). \tag{28}$$

Also note that the Bayes-optimal estimator of $X^n$ from $Z^{*n}$ with respect to $\rho$ is also separable, and its $i^{\text{th}}$ coordinate is given by

$$\tilde{\psi}_i(z) = \operatorname*{argmin}_{x\in\mathcal{X}} \mathbb{E}\left[\rho_i\left(X_i, x\right) \mid Z_i^* = z\right]. \tag{29}$$

In addition to the assumption that the information model is separable, we also assume in this section that the alphabet $\mathcal{Z}$ is countable. Under this assumption, the normalized Hamming distance $\rho_H$ of (15) provides a canonical metric on $\mathcal{Z}$. Proposition 3.2 thus implies that for any $y \in \mathcal{Y}$, $P_{Z_1^*|Y_1=y}$ satisfies a transportation-cost inequality. By combining the inequality from Proposition 3.2 with the result of Theorem 4.1, we obtain the following result.

**Theorem 5.1:** Consider a discrete separable model. Assume that $D \geq d_{\min}$ such that (C1) holds. For every $n \in \mathbb{N}$, let $P_{Z^n|Y^n} : \mathcal{Y}^n \to \mathcal{Z}^n$ be a code of rate $R_n$ satisfying (A1) and $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[d(Y_i, Z_i)\right] \leq D$. If the sequence of codes $\{P_{Z^n|Y^n}\}_{n=1}^{\infty}$ satisfies $R_n \to R$ and (A2), then:

(i) For any Lipschitz estimator $\psi : \mathcal{Z} \to (\mathcal{X}, \rho)$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\rho\left(X_i, \psi\left(Z_i\right)\right)\right] = \mathbb{E}\left[\rho\left(X_1, \psi(Z_1^*)\right)\right] = D_{\psi}^*(R(D)). \tag{30}$$

(ii) For any sequence of Lipschitz estimators $\phi_n : \mathcal{Z}^n \to \mathcal{X}^n$ whose Lipschitz constants are uniformly bounded in $n$, we have

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\rho\left(X_i, [\phi_n\left(Z^n\right)]_i\right)\right] \geq D^*(R(D)). \tag{31}$$

*Proof:* See Appendix C. □

We emphasize that Theorem 5.1 provides both positive, i.e. achievability, and negative, i.e. converse, coding results. Specifically, (30) implies that $D^*(R)$ in (13) is achievable by employing the Bayes-optimal estimator (29) at each coordinate. On the other hand, (31) implies that $D^*(R)$ bounds from below the distortion of any other estimator of $X^n$ (separable or not) from the output of a good sequence of codes for $Y^n$ satisfying (A1) and (A2). Another important conclusion of Theorem 5.1 is that, among all estimators of $X^n$ from the compressed representation $Z^n$, the separable estimator $\tilde{\psi}$ of (29) has the minimal asymptotic risk.

### *B. Example: Binary Signal under Bitflip Noise*

Let us return to the setting in Section II-B1 and utilize Theorem 5.1 to characterize the distortion for this scenario. Let $\mathcal{B}(\pi)$ indicated the Bernoulli distribution with probability of success $\pi$ so that $X^n \overset{\text{iid}}{\sim} \mathcal{B}(\pi)$ and $W^n \overset{\text{iid}}{\sim} \mathcal{B}(\alpha)$. Under this setting, we have that $Y^n \overset{\text{iid}}{\sim} \mathcal{B}(\beta)$ with

$$\beta = \pi \star \alpha \triangleq \pi(1-\alpha) + \alpha(1-\pi).$$

Henceforth, we assume for simplicity that $\alpha, \pi \leq 1/2$ which also implies that[3] $\beta \leq 1/2$. For $\pi \star \alpha \geq D \geq 0$, the conditional distribution $P_{Z_1^*|Y_1}$ obtained from (28) is described by the memoryless channel [5], [49]

$$Y_i = Z_i \oplus V_i, \quad i = 1, \dots, n, \tag{32}$$

where $V^n \overset{\text{iid}}{\sim} \mathcal{B}(D)$ and independent of $Z^n$. The resulting Shannon RD function of $Y^n$ is given by

$$R(D) = \begin{cases} h_2(\pi \star \alpha) - h(D), & \pi \star \alpha \geq D \geq 0 \\ 0, & D > \pi \star \alpha, \end{cases} \tag{33}$$

where $h_2(x)$ is the binary entropy function. Next, consider the estimation of the binary sequence $X^n$ from $Z^{*n}$ under the normalized Hamming distortion in (15): The Bayes-optimal estimator of $X_i$ from $Z_i$ in this setting is

$$\tilde{\phi}(z_i) = \underset{x \in \{0,1\}^n}{\operatorname{argmin}} \mathbb{P}(X_i = x \mid Z_i = z_i). \tag{34}$$

The estimator in (34) is Lipschitz with respect to the normalized Hamming distance, as is the case for any estimator from and into spaces with finite alphabet. In Appendix E we show that for such $D$, the RD risk of $\tilde{\phi}$ at a target distortion $D$ with respect to the data $Y^n$, i.e., the Bayes-optimal RD risk at bitrate $R(D)$, is given by

$$D^*(R(D)) \triangleq \alpha \star D. \tag{35}$$

Under the conditions of Theorem 5.1, the CE risk of the estimator $\tilde{\phi}$, applied to the output of any RD code for the data $Y^n$ at a target distortion $D$ whose rate converges to $R(D)$, converges to $\alpha \star D$.

[3]Note that the complementary cases are obtained by replacing $\alpha$ with $\min\{\alpha, 1-\alpha\}$ if $1 \geq \alpha > 1/2$ and $\pi$ with $\min\{\alpha, 1-\alpha\}$ if $1 \geq \pi > 1/2$

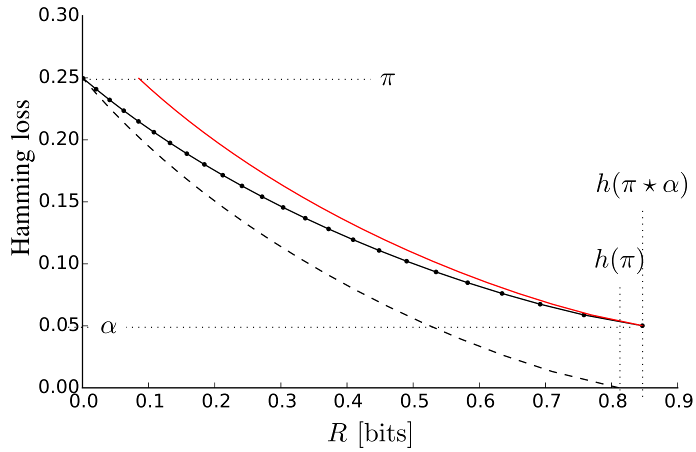


Fig. 3. The data is obtained by observing a binary $\pi = 0.25$ signal under a binary symmetric channel with bitflip probability $\alpha = 0.05$. $D^*(R)$ (red) is minimal risk in a CE setting when the data is compressed using a good RD rate-$R$ code (with respect to the data). I-DRF (dotted) is the minimal risk in estimating the signal under any rate-$R$ encoding of the data. Also shown is Shannon DRF (dashed) achievable by compressing a noiseless version of the binary signal, as in the case $\alpha = 0$.

From the expression in (33) we also obtain the number of bits $R^*$ required to attain a prescribed distortion $D^*$ with respect to $X^n$ as

$$R^*(D^*) = h(\pi \star \alpha) - h\left(D\right) = h(\pi \star \alpha) - h\left(\frac{D^* - \alpha}{1 - 2\alpha}\right). \tag{36}$$

Note that, when $\alpha = 0$, (36) coincides with the classical rate-distortion function of a binary i.i.d. source estimated under Hamming distortion. Furthermore, for $\pi = 1/2$, (36) reduces to the indirect rate-distortion function (the inverse function the the I-DRF) of a binary source observed through a binary symmetric channel with crossover probability $\alpha$ [5, Exercise 3.8]:

$$R_{\mathsf{Ind}}(D) \triangleq 1 - h\left(\frac{D - \alpha}{1 - 2\alpha}\right). \tag{37}$$

The equality between (37) and (35) for $\pi = 1/2$ implies that, asymptotically, there is no loss in compression performance if the encoder uses an optimal lossy compression code with respect to $Y^n$ rather than first estimating $\theta$ and then compressing this estimate. Nevertheless, $R^*(D)$ is strictly greater than $R_{\mathsf{Ind}}(D)$ whenever $\alpha \neq 1/2$ [3]. The difference between $D^*(R)$ and $D_{\mathsf{Ind}}(R)$ (the inverse of $R_{\mathsf{Ind}}(D)$) is illustrated in Figure 3.

Note that both the I-DRF and the Bayes-optimal RD risk attain the minimum distortions $d_{\min} = \alpha$ at rate $h(\pi \star \alpha)$: this rate is the entropy of the observation sequence $Y^n$, so the latter can be described almost losslessly by the decoder at this rate. Indeed, $d_{\min} = \alpha$ is the probability of making an error in estimating $\theta$ from the output of the channel (14). On the other hand, the I-DRF attains the maximum distortion $\pi$ at rate zero, while $D^*(R) = \pi$ already at $R = h(\alpha)$. This difference is rather interesting as it highlights that, for a rate less than $h(\alpha)$, the encoded representation of $Y^n$ provides no information on $\theta$.

## VI. Quadratic Gaussian Setting

In this section we focus on the case where the data $Y^n$ follows a normal distribution and the distortion measure $d$ is the squared Euclidean distance. The counterpart of Theorem 5.1 in this setting is as follows.

**Theorem 6.1 (main results for Gaussian data and Euclidean distance):** Let $Y^n \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$, and let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of $\Sigma_\theta$. For $1/n \operatorname{Tr}(\Sigma_\theta) \geq D \geq 0$, define $R(D) = R(D)$ by

$$R(D_\eta) \triangleq \frac{1}{2n} \sum_{i=1}^{n} \max\{\log_2(\lambda_i/\eta), 0\}, \tag{38a}$$

$$D_\eta \triangleq \frac{1}{n} \sum_{i=1}^{n} \min\{\lambda_i, \eta\}, \tag{38b}$$

Finally, define a joint distribution $P_{Y^n, Z^{*n}}$ by

$$Y^n = Z^{*n} + \sqrt{d_i} \mathbf{U} W^n, \tag{39}$$

where $d_i = \min\{\lambda_i, \eta\}$ and $W^n \sim \mathcal{N}(0, \mathbf{I})$, and $\mathbf{U}$ is the matrix of right eigenvectors of $\Sigma_\theta$. For each $n \in \mathbb{N}$, let $P_{Z^n|Y^n} : \mathbb{R}^n \to \mathbb{R}^n$ be any $D$-admissible code of rate $R_n$. Then

(i) For any $L$-Lipschtiz $\phi : (\mathbb{R}^n, \|\cdot\|) \to (\Theta, \rho)$,

$$\begin{aligned} &\frac{1}{\sqrt{n}} \left| \sqrt{\mathbb{E}\left[\rho^2\left(\theta, \phi(Z^n)\right)\right]} - \sqrt{D^*_\phi(R(D), \theta)} \right| \\ &\leq L \left\|\Sigma_\theta\right\|_2 \sqrt{2\left(R_n - R(D) + \frac{1}{n} D(Z^n || Z^{*n})\right)}, \end{aligned} \tag{40}$$

provided $\mathbb{E}\left[\rho^2(\theta, \phi(Z^n))\right] < \infty$ and $D^*_\phi(R(D), \theta) < \infty$.

(ii) Assume that $\limsup_n \|\Sigma_\theta\| < \infty$, and that the sequence of codes $\{P_{Z^n|Y^n}\}_{n=1}^\infty$ satisfies (A2) and $R_n \to R(D)$. For any sequence of estimators $\phi_n : (\mathbb{R}^n, \|\cdot\|) \to (\Theta, \rho)$ with $\mathbb{E}\left[\rho^2(\theta, \phi_n(Z^n))\right] < \infty$ whose Lipschtiz constants are uniformly bounded in $n$, we have

$$\lim_{n \to \infty} \frac{1}{n} \left| \mathbb{E}\left[\rho^2\left(\theta, \phi_n\left(Z^n\right)\right)\right] - D^*_\phi(R(D), \theta) \right| = 0. \tag{41}$$

*Proof:* See Appendix D. □

Theorem 6.1 provides a powerful characterization for the asymptotic performance of estimators from Gaussian data compressed using any good rate-distortion code satisfying (A2). Part (i) of this Theorem provides a bound for the risk of any Lipschitz estimator from the output of a $D$-admissible code in terms of the excess coding bitrate $R_n - R(D)$. In particular, it implies that the risk in estimation from the compressed representation converges to a finite limit, provided the limit of the RD risk exists. Part (ii) says that such estimators attain the same risk as if applied to the random vector $Z^{*n}$ defined by (39). The work of [24] provides a result similar to Theorem 6.1 for the case where the data is not necessarily Gaussian, but the encoder is limited to use a random spherical code.

Next, we apply the result of Theorem 6.1 to the examples in Section II-A and Section II-B2, respectively.

### A. Example: Location Parametric Estimation

We now consider the setting described in Subsection II-A where $\theta$ is a location parameter to be estimated from a lossy compressed version of $Y^n$.

**Proposition 6.2:** Assume that $Y^n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2)$ and let $R > 0$. For any rate-$R_n$ code $P_{Z^n|Y^n} : \mathbb{R}^n \to \mathbb{R}^n$ satisfying (A2) for which $\mathbb{E}\left[\|Y^n - Z^n\|^2\right]/n \le \sigma^2 2^{-2R}$, there exists an estimator $\phi : \mathcal{Z}^n \to \Theta$ such that

$$\mathbb{E}\left[(\theta - \phi(Z^n))^2\right] \le \frac{1}{n}\left(\sqrt{\frac{R_n - R + \frac{1}{n}D(Z^n||Z^{*n})}{1-2^{-2R}}} + \sigma\sqrt{\frac{1+2^{-2R}}{1-2^{-2R}}}\right)^2. \tag{42}$$

*Proof:* See Appendix F. □

Proposition 6.2 shows that there exists an estimator of the parameter $\theta$ with mean squared error (MSE) decreasing as $1/n$ from the output of any code that attains a distortion target with respect to the data that equals to the DR function of the data. A natural measure for the loss in efficiency due to compressing the measurements is the asymptotic relative efficiency (ARE) of this estimator compared to an efficient estimator of $\theta$ from $Y^n$, i.e., an estimator attaining MSE of order $\sigma^2/n$ [50]. In our setting, we define this ARE as[4]

$$\mathrm{ARE}(\phi) = \limsup_{n\to\infty} \frac{\sigma^2}{n\mathbb{E}\left[(\theta - \phi(Z^n))^2\right]}.$$

Intuitively, the ARE indicates the increase in the number of samples from the distribution required to attain a prescribed accuracy in estimating $\theta$ compared to an efficient estimator that has access to the uncompressed data. e.g., the mean of the data. Proposition 6.2 implies that

$$\mathrm{ARE}(\phi) \ge \left(\frac{1}{\sigma}\sqrt{\frac{R_n - R + \frac{1}{n}D(Z^n||Z^{*n})}{1-2^{-2R}}} + \sqrt{\frac{1+2^{-2R}}{1-2^{-2R}}}\right)^{-2}. \tag{43}$$

In particular, for rate-distortion codes for which $R_n \to R$ and $\frac{1}{n}D(Z^n||Z^{*n}) \to 0$, the RHS of (43) reduces to

$$\eta(R) \triangleq \frac{1-2^{-2R}}{1+2^{-2R}}.$$

For example, since $\eta(1) = 3/5$, we conclude that compressing $Y^n$ using a good rate-distortion code with one bit per sample increases the sample size required to attain a prescribed MSE by at most $5/3$.

It is important to note that it is possible to design a code especially for estimating $\theta$, rather than compressing the measurements, that attains an ARE of 1. Indeed, it is straightforward to represent the average of $Y^n$ using $nR$ bits with MSE exponentially small in $n$ [10]. This point highlights the difference between the estimate-and-compress approach in which we use a source code tailored for the specific end estimation problem and the CE approach where the compressor of $Y^n$ is agnostic to $\theta$.

[4]This definition of the ARE coincides with the standard definition in [50] whenever $\phi(Z^n)$ is asymptotically normal with a finite second moment.

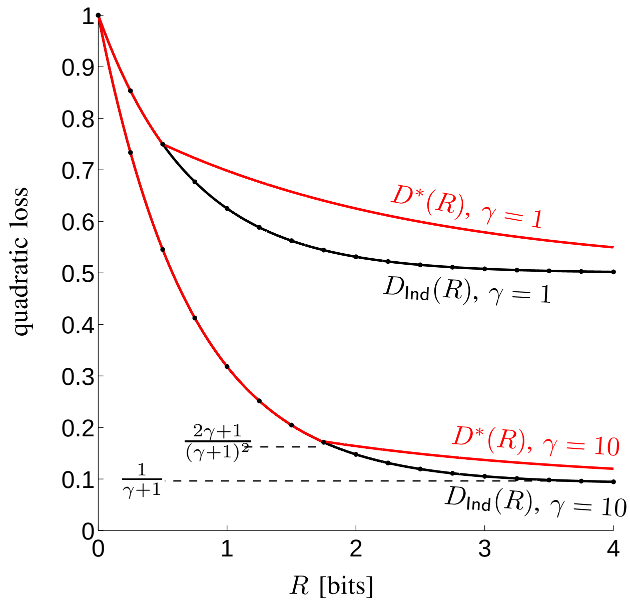


Fig. 4. Quadratic Gaussian measurement model (17) with $m = 3$ and $\gamma \in \{1, 10\}$, Gaussian signal prior, and optimal Bayes estimation. $D^*(R)$ and $D_{\mathsf{Ind}}(R)$ describe the RD and the optimal source coding risks, respectively.

### *B. Example: Single Input Multiple Output Gaussian Channel*

Let us return next to the setting in Section II-B2. For this setting, we have the following evaluation of the Bayes-optimal RD risk.

**Proposition 6.3:** Let $X^n \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ be input to the $m$ independent AWGN channels defined in (17) with $d = \|\cdot\|^2$, $\rho = \|\cdot\|$, and $p = 2$. Then

$$D^*(R) = \frac{1}{\gamma+1} + \frac{\gamma}{\gamma+1}\eta, \tag{44}$$

where $\gamma \triangleq \sum_{j=1}^{m} \gamma_j$ and where $\eta$ is obtained as

$$\eta = \begin{cases} 2^{-\frac{2R}{m} - \frac{m-1}{m}\log(1+\gamma)} & R > \frac{1}{2}\log(1+\gamma), \\ 2^{-2R} & R \leq \frac{1}{2}\log(1+\gamma). \end{cases} \tag{45}$$

*Proof:* See Appendix B. ■

Note that in this example, the data $\mathbf{Y}^n$ is defined as a sequence of $m$-dimensional vectors. Therefore, in order to use Theorem 6.1 for characterizing the CE risk in terms of $D^*(R)$, we may stack these vectors as a single sequence of length $mn$. Indeed, this transformation of the data does not change the rate-distortion function of $\mathbf{Y}^n$, and only leads to a similar transformation in the coordinates of the rate-distortion achieving distribution.

For comparison, the I-DRF in the setting of (17), describing the minimal distortion under an optimal rate $R$

source code applied to $Y^n$, is given as [51, Eq. 10]:

$$D_{\mathsf{Ind}}(R) = \frac{1}{1+\gamma} + \frac{\gamma}{1+\gamma} 2^{-2R}. \tag{46}$$

By comparing (44) and (46) we recover two observations first made in [19]: The RD risk coincides with the I-DRF either when $R$ is smaller than $0.5\log(1+\gamma)$ or when $m=1$. In other words, for this choice of the observation distortion, it is possible to attain the optimal source coding performance despite not having knowledge of the joint distribution between the source and the observations. On the other hand, when the rate $R$ is larger than $0.5\log(1+\gamma)$, the minimal distortion under CE decreases roughly $m$ times slower than the equivalent term in $D_{\mathsf{Ind}}(R)$. A comparison between $D^*(R)$ and $D_{\mathsf{Ind}}(R)$ is illustrated in Figure 4 for $m=3$ and different values of the parameter $\gamma$.

## VII. Conclusions

We considered the problem of estimation from compressed data, which encapsulates indirect source coding and compression for estimation settings. In this setup, we considered a compress-and-estimate (CE) approach in which the data is first compressed using a good rate-distortion code, i.e., a code attaining the Shannon rate-distortion function of the data with equality. The underlying signal or parameter is ultimately estimated from this compressed version of the data. We showed that whenever the rate-distortion achieving distribution satisfies a suitable transportation cost inequality, the performance of any robust enough estimation procedure from the data is given by the RD risk of this estimator which is the risk of this estimator as evaluated on the data sampled from the rate-distortion achieving distribution. In particular, these conditions hold, broadly speaking, whenever the compressed representation is embedded in a discrete space, or when the rate-distortion achieving distribution is Gaussian. In order to illustrate the usefulness of this characterization, we evaluated the RD risk in the case of a Bernoulli source observed under a binary symmetric channel, a Gaussian source observed through multiple AWGN channels, and a parametric estimation problem in a Gaussian location model.

The work presented here leaves a number of interesting research directions. First, it is unclear whether the equivalence between the RD risk and the true risk in estimating from the compressed data follows directly from the concentration of measure property. Another line of future work is the extension of the results provided here to a multi-terminal setting, i.e. when the data is compressed at multiple locations. Finally, good rate-distortion codes that violate condition (A2) were given in [37] and [36]. Our analysis leaves open the question whether the true risk converges to the RD risk in such cases.

## Acknowledgment

This work was supported in part by the NSF Center for Science of Information (CSoI) under grant CCF-0939370, the NSF grants DMS-1418362 and DMS-1407813, and the NSF-BSF under grant 1609695.

## References

[1] A. Kipnis, S. Rini, and A. J. Goldsmith, "Multiterminal compress-and-estimate source coding," in *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 540–544.

[2] ——, “Coding theorems for the compress and estimate source coding problem,” in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 2568–2572.

[3] ——, “The indirect rate-distortion function of a binary i.i.d source,” in *Information Theory Workshop - Fall (ITW), 2015 IEEE*, Oct 2015, pp. 352–356.

[4] R. Dobrushin and B. Tsybakov, “Information transmission with additional noise,” *IRE Transactions on Information Theory*, vol. 8, no. 5, pp. 293–304, 1962.

[5] T. Berger, *Rate-distortion theory: A mathematical basis for data compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[6] H. S. Witsenhausen, “Indirect rate distortion problems,” *Information Theory, IEEE Transactions on*, vol. 26, no. 5, pp. 518–521, 1980.

[7] C. E. Shannon, “A mathematical theory of communication,” *Bell System Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[8] J. Wolf and J. Ziv, “Transmission of noisy information to a noisy receiver with minimum distortion,” *IEEE Transactions on Information Theory*, vol. 16, no. 4, pp. 406–411, 1970.

[9] T. S. Han, “Hypothesis testing with multiterminal data compression,” *IEEE Transactions on Information Theory*, vol. 33, no. 6, pp. 759–772, 1987.

[10] T. Han and S. Amari, “Statistical inference under multiterminal data compression,” *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2300–2324, Oct 1998.

[11] S. Kassam, “Optimum quantization for signal detection,” *IEEE Transactions on Communications*, vol. 25, no. 5, pp. 479–484, May 1977.

[12] H. Poor and J. Thomas, “Applications of ali-silvey distance measures in the design generalized quantizers for binary decision systems,” *IEEE Transactions on Communications*, vol. 25, no. 9, pp. 893–900, Sep. 1977.

[13] R. M. Gray, “Quantization in task-driven sensing and distributed processing,” in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 5, May 2006, pp. V–V.

[14] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and Y. Zhang, “Optimality guarantees for distributed statistical estimation,” *arXiv preprint arXiv:1405.0782*, 2014.

[15] A. Kipnis and J. C. Duchi, “Mean estimation from one-bit measurements,” *CoRR*, vol. abs/1901.03403, 2019.

[16] L. P. Barnes, Y. Han, and A. Ozgur, “Learning distributions from their samples under communication constraints,” 2019.

[17] P. Ishwar, R. Puri, K. Ramchandran, and S. S. Pradhan, “On rate-constrained distributed estimation in unreliable sensor networks,” *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 4, pp. 765–775, April 2005.

[18] Y. Wang, L. Xie, X. Zhang, and J. Chen, “Robust distributed compression of symmetrically correlated Gaussian sources,” *CoRR*, vol. abs/1807.06799, 2018.

[19] I. D. Schizas, G. B. Giannakis, and N. Jindal, “Distortion-rate bounds for distributed estimation using wireless sensor networks,” *EURASIP J. Adv. Signal Process*, vol. 2008, pp. 118:1–118:12, Jan. 2008. [Online]. Available: https://doi.org/10.1155/2008/748605

[20] T. Cover and A. E. Gamal, “Capacity theorems for the relay channel,” *IEEE Transactions on information theory*, vol. 25, no. 5, pp. 572–584, 1979.

[21] A. Kipnis, A. J. Goldsmith, and Y. C. Eldar, “The distortion-rate function of sampled Wiener processes,” *IEEE Transactions on Information Theory*, vol. 65, no. 1, pp. 482–499, Jan 2019.

[22] G. Murray, A. Kipnis, and A. J. Goldsmith, “Lossy compression of decimated gaussian random walks,” in *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, March 2018, pp. 1–6.

[23] A. Kipnis, G. Reeves, and Y. C. Eldar, “Single letter formulas for quantized compressed sensing with gaussian codebooks,” in *2018 IEEE International Symposium on Information Theory (ISIT)*, June 2018, pp. 71–75.

[24] A. Kipnis and G. Reeves, “Gaussian approximation of quantization noise for estimation from compressed data,” 2019, submitted to ISIT2019.

[25] M. Raginsky, I. Sason *et al.*, “Concentration of measure inequalities in information theory, communications, and coding,” *Foundations and Trends® in Communications and Information Theory*, vol. 10, no. 1-2, pp. 1–246, 2013.

[26] C. Villani, *Optimal transport: old and new*. Springer Science & Business Media, 2008, vol. 338.

[27] H. Yamamoto, “A rate-distortion problem for a communication system with a secondary decoder to be hindered,” *IEEE Transactions on Information Theory*, vol. 34, no. 4, pp. 835–842, 1988.

[28] A. Lapidoth, “On the role of mismatch in rate distortion theory,” *IEEE Transactions on Information Theory*, vol. 43, no. 1, pp. 38–47, 1997.

[29] I. Kontoyiannis and R. Zamir, “Mismatched codebooks and the role of entropy coding in lossy data compression,” *IEEE transactions on information theory*, vol. 52, no. 5, pp. 1922–1938, 2006.

[30] D. L. Donoho, *The Kolmogorov sampler*. Department of Statistics, Stanford University, 2002.

[31] R. M. Gray, *Entropy and information theory*. Springer Science & Business Media, 2011.

[32] A. Kanlis, "Compression and transmission of information at multiple resolutions," Ph.D. dissertation, University of Maryland, 1997.

[33] S. Shamai and S. Verdú, "The empirical distribution of good codes," *IEEE Transactions on Information Theory*, vol. 43, no. 3, pp. 836–846, 1997.

[34] T. Weissman and E. Ordentlich, "The empirical distribution of rate-constrained source codes," *Information Theory, IEEE Transactions on*, vol. 51, no. 11, pp. 3718–3733, 2005.

[35] C. Schieler and P. Cuff, "A connection between good rate-distortion codes and backward dmcs," in *Information Theory Workshop (ITW), 2013 IEEE*. IEEE, 2013, pp. 1–5.

[36] V. Kostina and S. Verdu, "The output distribution of good lossy source codes," in *Information Theory and Applications Workshop (ITA), 2015*, Feb 2015, pp. 308–312.

[37] A. Kanlis, S. Khudanpur, and P. Narayan, "Typicality of a good rate-distortion code," *Problems of Information Transmission*, vol. 32, no. 1, pp. 96–103, 1996.

[38] D. Sakrison, "A geometric treatment of the source encoding of a gaussian random variable," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 481–486, 1968.

[39] Z. Zhang and T. Berger, "Estimation via compressed information," *IEEE Transactions on Information Theory*, vol. 34, no. 2, pp. 198–211, 1988.

[40] A. Kipnis and J. C. Duchi, "Mean estimation from adaptive one-bit measurements," in *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct 2017, pp. 1000–1007.

[41] A. Kolmogorov, "On the shannon theory of information transmission in the case of continuous signals," *IRE Transactions on Information Theory*, vol. 2, no. 4, pp. 102–108, December 1956.

[42] R. Gray, *Probability, Random Processes, and Ergodic Properties*. Springer, 2009.

[43] R. M. Gray, D. L. Neuhoff, and P. C. Shields, "A generalization of ornstein's d distance with applications to information theory," *The Annals of Probability*, pp. 315–328, 1975.

[44] K. Marton, "A simple proof of the blowing-up lemma (corresp.)," *IEEE Transactions on Information Theory*, vol. 32, no. 3, pp. 445–446, 1986.

[45] R. Ahlswede, P. Gács, and J. Körner, "Bounds on conditional probabilities with applications in multi-user communication," *Probability Theory and Related Fields*, vol. 34, no. 2, pp. 157–177, 1976.

[46] A. El Gamal and Y.-H. Kim, *Network information theory*. Cambridge university press, 2011.

[47] K. Marton *et al.*, "Bounding $\rho$-bar distance by informational divergence: A method to prove measure concentration," *The Annals of Probability*, vol. 24, no. 2, pp. 857–866, 1996.

[48] M. Talagrand, "Transportation cost for Gaussian and other product measures," *Geometric & Functional Analysis GAFA*, vol. 6, no. 3, pp. 587–600, 1996.

[49] T. M. Cover and J. A. Thomas, *Elements of information theory (2nd ed.)*. Wiley, 2006.

[50] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer Science & Business Media, 2006.

[51] M. Gastpar, "A lower bound to the AWGN remote rate-distortion function," in *Statistical Signal Processing, 2005 IEEE/SP 13th Workshop on*. IEEE, 2005, pp. 1176–1181.


# APPENDIX A

## PROOF OF PROPOSITION 3.4

Let $W^n \sim \gamma^n = \mathcal{N}(0, \mathbf{I}_n)$ and $V^n \sim P^n$ arbitrary absolutely continuous with respect to $\gamma'^n$. Write $\Sigma = \mathbf{U}\Lambda\mathbf{U}^*$ with $\mathbf{U}$ unitary and $\Lambda$ diagonal. For any coupling of $W^n$ and $V^n$ and for any rotation matrix $\mathbf{U}$, we have

$$\begin{aligned}
&\|W^n - V^n\|^2 = \frac{\left\|\Lambda^{1/2}\right\|_2^2}{\left\|\Lambda^{1/2}\right\|_2^2} \|W^n - y^n - V^n - y^n\|^2 \\
&\geq \frac{1}{\left\|\Lambda^{1/2}\right\|_2^2} \left\|\Lambda^{1/2}(W^n - y^n) - \Lambda^{1/2}(V^n - y^n)\right\|^2 \\
&= \frac{1}{\left\|\Lambda^{1/2}\right\|_2^2} \left\|\mathbf{U}\Lambda^{1/2}(W^n - y^n) - \mathbf{U}\Lambda^{1/2}(V^n - y^n)\right\|^2 \\
&= \frac{1}{\left\|\Lambda^{1/2}\right\|_2^2} \|W'^n - V'^n\|^2,
\end{aligned}$$

where $W'^n \sim \gamma'^n$ and $V'^n \sim P'^n$ where $P'^n$ is arbitrary absolutely continuous measure with respect to $\gamma'^n$. By taking the infimum over all coupling of $\gamma'^n$ and $P'^n$, Proposition 3.3 implies

$$2D(P^n||\gamma^n) \geq \frac{1}{\left\|\Lambda^{1/2}\right\|_2^2} W_2^2(P'^n, \gamma'^n).$$

Finally, (22) follows since $P'^n$ and $\gamma'^n$ are obtained from $P^n$ and $\gamma^n$ by the same parameter transformation, to which $D(P^n||\gamma^n)$ is invariant.

## APPENDIX B
## PROOF FOR PROPOSITION 6.3

The data $Y^n$ has distribution independent over blocks of length $m$, where the distribution of each such $m$-length block is jointly Gaussian with mean $0$ and covariane matrix

$$\Sigma_{\mathbf{Y}_1} = \mathbf{a}\mathbf{a}^* + \mathbf{I},$$

where $\mathbf{a} = (\gamma_1, \ldots, \gamma_m)$.

The distribution $P_{\mathbf{Z}_1^*, \mathbf{Y}_1}$ achieving RDF of $\mathbf{Y}^n$ satisfies the backward channel [5]:

$$\mathbf{Y}_1 = \mathbf{Z}_1 + \mathbf{U}\mathbf{W}, \tag{47}$$

where:

(i) $\mathbf{U}$ is a unitary matrix such that

$$\mathbf{U}^*\Sigma\mathbf{U} = \operatorname{diag}(\lambda_1, \ldots, \lambda_L),$$

and $\lambda_1, \ldots, \lambda_m$ are the eigenvalues of $\Sigma$ These eigenvalues are given by

$$\lambda_1 = \mathbf{a}^*\mathbf{a} + 1, \tag{48a}$$

$$\lambda_i = 1, \quad i = 2, \ldots, m. \tag{48b}$$

(ii) $\mathbf{W}$ is a Gaussian vector independent of $\mathbf{Z}_1$ whose $m$-th coordinate has variance $D_i \triangleq \min\{\eta, \lambda_i\}$, and $\eta$ is chosen such that

$$R = \sum_{i=1}^{m} R_i \triangleq \frac{1}{2} \sum_{i=1}^{m} \log^+ \left( \frac{\lambda_i}{\eta} \right). \tag{49}$$

Define $\tilde{\mathbf{Z}}_1 \triangleq \mathbf{U}^* \mathbf{Z}_1$, so that

$$\mathbf{U}^* \mathbf{Y}_1 = \tilde{\mathbf{Z}}_1 + \mathbf{W}_1,$$

is equivalent to (47). We also have

$$\begin{aligned} D(R) &= \mathrm{Tr}\mathbb{E}\left[ (\mathbf{Y}_1 - \mathbf{Z}_1)^* (\mathbf{Y}_1 - \mathbf{Z}_1) \right] \\ &= \mathrm{Tr}\mathbb{E}\left[ (\mathbf{U}^*\mathbf{Y}_1 - \tilde{\mathbf{Z}}_1)^* (\mathbf{U}^*\mathbf{Y}_1 - \tilde{\mathbf{Z}}_1) \right] \\ &= \mathrm{Tr}\mathbb{E}\left[ (\mathbf{U}^*\mathbf{Y}_1 - \tilde{\mathbf{Z}}_1)^2 \right] = \sum_{i=1}^{m} D_i. \end{aligned}$$

An equivalent representation to $P_{\mathbf{Z}_1^*, \mathbf{Y}_1}$ is as follows:

$$\begin{aligned} \tilde{Z}_{1,i} &= \left(1 - 2^{-2R_i}\right) [\mathbf{U}^* \mathbf{Y}_1]_i + \sqrt{D_i (1 - 2^{-2R_i})} G_i, \\ &= \left(1 - 2^{-2R_i}\right) \mathbf{u}_i^* \mathbf{a} \theta_1 + \left(1 - 2^{-2R_i}\right) [\mathbf{u}_1^* \mathbf{W}]_i \\ &\quad + \sqrt{D_i (1 - 2^{-2R_i})} G_i & \text{(50a)} \\ &= \left(1 - 2^{-2R_i}\right) \mathbf{u}_i^* \mathbf{a} \theta_1 + \sqrt{D_i (1 - 2^{2R_i}) + (1 - 2^{-2R})^2} V_i, & \text{(50b)} \end{aligned}$$

where $G_1, \ldots, G_m$ and $V_1, \ldots, V_m$ are i.i.d. and standard normal and independent of $\mathbf{W}$ and $\theta_1$. We write (50b) in the matrix form

$$\tilde{\mathbf{Z}}_1 = \mathbf{B} \theta_1 + \mathbf{C} \mathbf{V}.$$

The Bayes-optimal estimator of $\theta_1$ from $\mathbf{Z}_1$ is the conditional expectation. The expected quadratic risk is the minimal mean squared error, given by

$$\begin{aligned} \mathsf{mmse}(\theta_1 | \mathbf{Z}_1) &= \mathsf{mmse}(\theta_1 | \tilde{\mathbf{Z}}_1) \\ &= \left(1 + \mathbf{B}^* \Sigma_{\mathbf{C}}^{-1} \mathbf{B}\right)^{-1} \\ &= \left(1 + \sum_{i=1}^{m} \frac{(1 - 2^{-2R_i})(\mathbf{a}^* \mathbf{u}_i)^2}{(1 - 2^{-2R_i}) + D_i}\right)^{-1} \\ &= \left(1 + \sum_{i=1}^{m} \frac{(1 - 2^{-2R_i})(\mathbf{a}^* \mathbf{u}_i)^2}{(1 - 2^{-2R_i}) + \min\{\theta, \lambda_i\}}\right)^{-1}. \end{aligned}$$

Since $\mathbf{u}_1 = \mathbf{a}/\|\mathbf{a}\|$, $R_1 = \frac{1}{2}\log\left((1+\mathbf{a}^*\mathbf{a})/\theta\right)$ and $\{\mathbf{u}_i\}_{i=2,\ldots,m}$ are orthogonal to $\mathbf{a}$, we get

$$\begin{aligned}\mathsf{mmse}(\theta|\tilde{\mathbf{Z}}_1) &= \left(1+\mathbf{a}^*\mathbf{a}\frac{1+\mathbf{a}^*\mathbf{a}-\theta}{1+\mathbf{a}^*\mathbf{a}(\theta+1)}\right)^{-1} \\ &= \frac{1}{\mathbf{a}^*\mathbf{a}+1} + \frac{\mathbf{a}^*\mathbf{a}}{(\mathbf{a}^*\mathbf{a}+1)^2}\theta,\end{aligned}$$

and (44) is obtained by setting $\theta' = \theta/(1+\mathbf{a}^*\mathbf{a})$.

## APPENDIX C
## PROOF OF THEOREM 5.1

We first note that any separable $B$-bounded estimator is $B$-Lipschtiz with respect to the Hamming distance, as per the following proposition:

**Proposition C.1:** For each $i = 1,\ldots,n$, let $\phi_i : \mathcal{Z} \to \mathcal{X}$ such that $\phi_i$ is $B$-bounded. Define $\phi : \mathcal{Z}^n \to \mathcal{X}^n$ by $\phi(z^n) = (\phi_1(z_1),\ldots,\phi_n(z_n))$. Then $\phi$ is $B$-Lipschtiz with respect to the Hamming distance $\rho_H$ on $\mathcal{Z}^n$.

*Proof of Proposition C.1:* For any $z^n, \hat{z}^n \in \mathcal{Z}^n$,

$$\begin{aligned}&\rho\left(\phi(z^n),\phi(\hat{z}^n)\right) = \frac{1}{n}\sum_{i=1}^n \rho(\phi(z_i),\phi(\hat{z}_i)) \\ &\le \frac{1}{n}\sum_{z_i\neq\hat{z}_i}\rho(\phi(z_i),\phi(\hat{z}_i)) + \frac{1}{n}\sum_{z_i=\hat{z}_i}\rho(\phi(z_i),\phi(\hat{z}_i)) \\ &\le B\rho_H(z^n,\hat{z}^n)+0.\end{aligned}$$

We show that the conditions of Theorem 4.1 hold in this special case with $\Theta = \mathcal{X}^n$, $\theta = X^n$, and $P_\theta = P^n_{X_1}$. Since $\psi$ is bounded, we have that $\mathbb{E}\left[\rho(X^n,\psi(Z^{*n}))\right] < \infty$ and $\mathbb{E}\left[\rho(X^n,\psi(Z^n))\right] < \infty$. Note that, in the notation of Theorem 4.1, we use $\rho_z = \rho_H$, and $\rho$ given by

$$\rho(\theta,\hat{\theta}) \triangleq \frac{1}{n}\sum_{i=1}^n \rho(x_i,\hat{x}_i).$$

Finally, Proposition 3.2 implies that for any $y^n \in \mathcal{Y}^n$, $P_{Z^{*n}|Y^n=y^n}$ is $\mathsf{TC}(1/2n)$. We now apply Theorem 4.1 with parameter space $\Theta = \mathcal{X}^n$, $p=1$, and $\phi : \mathcal{Z}^n \to \Theta$ whose $i^{\text{th}}$ coordinate $[\phi(z^n)]_i$ equals $\psi(z_i)$. Since $\psi$ is assumed bounded, Proposition C.1 implies that it is Lipschitz and hence $\phi$ is Lipschitz. We denote its Lipschitz constant by $L$. It follows that

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\rho\left(X_i,[\phi(Z_i)]_i\right)\right] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\rho\left(X_i,\phi_i(Z_i^*)\right)\right]\right| \\ &\le L\sqrt{R_n - R(D) + \frac{1}{n}D(Z^n||Z^{*n})}.\end{aligned}$$

Equation (30) follows since $R_n \to R(D)$, $[\phi(z^n)]_i = \psi(z_i)$, and (A2) holds by assumption. This complete the proof of (i). In order to show (ii), first note that (31) trivially holds whenever $\mathbb{E}\left[\rho(X_i,[\phi_n(Z^n)]_i)\right] = \infty$, for some $i,n\in\mathbb{N}$, hence we assume that

$$\mathbb{E}\left[\rho(X_i,[\phi(Z^n)]_i)\right] < \infty, \qquad i,n\in\mathbb{N}.$$

We use Theorem 4.1 assuming $\phi$ is an arbitrary $L$-Lipschitz from $\mathcal{Z}^n$ to $\mathcal{X}^n$ (not necessarily separable). We obtain:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\rho\left(X_i,[\phi(Z^{*n})]_i\right)\right] \le L\sqrt{R_n - R(D) + \frac{1}{n}D(Z^n||Z^{*n})} \tag{51}$$
$$+\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}\left[\rho\left(X_i,[\phi(Z^n)]_i\right)\right]\right).$$

Since $P_{X^n,Y^n,Z^{*n}} = \prod_{i=1}^n P_{X_i,Y_i,Z_i^*}$, for each $i = 1,\ldots,n$, we have that $\mathbb{E}\left[\rho\left(X_i,[\phi(Z^{*n})]_i\right)\right]$ is bounded from below by

$$\mathbb{E}\left[\rho\left(X_1,\tilde{\psi}(Z_1^*)\right)\right] \triangleq D^*(R(D)),$$

where $\tilde{\psi}$ is the Bayes optimal estimator of $X_1$ from $Z_1^*$. Consequently, the left hand side of (51) is bounded from below by

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\rho\left(X_i,[\phi(Z^{*n})]_i\right)\right] = D^*(R(D)).$$

Inequality (31) now follows since $L\sqrt{R_n - R(D)} \to 0$ and $n^{-1}D(Z^n||Z^{*n}) \to 0$ as $n \to \infty$.

## APPENDIX D
## PROOF OF THEOREM 6.1

We show that the conditions of Theorem 4.1 are met in this case: First note that a single letter distortion measure is sub-additive, and that in our case $d_{\min} = 0$. In addition, (C1) holds for $0 \le D \le \frac{1}{n}\mathrm{Tr}\Sigma$, since in this case $P_{Z^{*n},Y^n}$ as defined by (39) is the unique solution to the minimization problem (9) when $P_{Y^n} = \mathcal{N}(\mu,\Sigma)$ [5]. For any $y^n \in \mathbb{R}^n$, $P_{Z^*|Y^n=y^n}$ is a Gaussian measure on $\mathbb{R}^n$, and by Proposition 3.4 it is $\mathsf{TC}(\|\sigma\|_2^2)$ with respect to the Euclidean distance and $p = 2$. We now use Theorem 4.1 with $p = 2$, where we also note that $\mathsf{supp}P_{Z^n} \subset \mathsf{supp}P_{Z^{*n}} = \mathbb{R}^n$. Under these conditions, part (i) follows from (23). Part (ii) now follows from (40) due to the additional condition $R_n \to R(D)$, $n^{-1}D(Z^n||Z^{*n}) \to 0$ (A2), and a uniformly bounded Lipschitz constant for the sequence $\{\phi_n\}$.

## APPENDIX E
## DERIVATION OF EQUATION (35)

Consider the RD achieving distribution $P_{Z_1^*,Y_1}$ defined by (32). Denote $X \triangleq \mathbb{P}(\theta_1 = Z_1)$. Since $\alpha, D(R) \le \pi$, we have

$$X = \alpha D(R) + (1-\alpha)\left(1 - D(R)\right),$$

where $D(R)$ is the DRF of $Y^n$ given by

$$D(R) = h^{-1}\left([h(\pi \star \alpha) - R]^+\right).$$

Since $\alpha, \pi \le 1/2$, we also have that $D(R) \le 1/2$, and thus the Bayes-optimal estimator of $\theta_1$ from $Z_1$ equals $\phi^*(z) = z$. It follows that the risk of $\phi^*$ equals to the probability of a bitflip in the channel from $\theta_1$ to $Z_1$. This probability is $\alpha \star D(R)$.

## Appendix F
## Proof of Proposition 6.2

For $\alpha > 0$, consider the estimator

$$\phi(z^n) = \frac{\alpha}{n} \sum_{i=1}^{n} z_i,$$

and the metric

$$\rho(\theta, \hat{\theta}) = \sqrt{n} \left| \theta - \hat{\theta} \right|.$$

For $z^n, \hat{z}^n \in \mathbb{R}^n$, we have

$$\begin{aligned} \rho(\phi(z^n), \phi(\hat{z}^n)) &= \sqrt{n} \frac{\alpha}{n} \left| \sum_{i=1}^{n} (z_i - \hat{z}_i) \right| \\ &\leq \frac{\alpha}{\sqrt{n}} \sum_{i=1}^{n} |z - \hat{z}_i| \leq \alpha \left\| z^n - \hat{z}^n \right\|, \end{aligned} \tag{52}$$

where in the last transition we used the inequality $\sum_{i=1}^n a_i \leq \sqrt{n} \sqrt{\sum_{i=1}^n a_i^2}$ that follows from the Cauchy-Schwartz inequality. (52) says that $\phi$ is $\alpha$-Lipschitz from $(\mathbb{R}^n, \|\cdot\|) \to (\mathbb{R}, \rho)$. We now use Theorem 6.1 with $(\Theta, \rho) = (\mathbb{R}, \rho)$ and $D = D(R) = \sigma^2 2^{-2R}$. We obtain

$$\left| \sqrt{n \mathbb{E}\left[ (\theta - \phi(Z^n))^2 \right]} - \sqrt{n \mathbb{E}\left[ (\theta - \phi(Z^{*n}))^2 \right]} \right| \leq \sqrt{\alpha} \sqrt{R_n - R + \frac{1}{n} D(Z^n || Z^{*n})}, \tag{53}$$

and hence

$$n \mathbb{E}\left[ (\theta - \phi(Z^n))^2 \right] \leq \left( \sqrt{\alpha} \sqrt{R_n - R + \frac{1}{n} D(Z^n || Z^{*n})} + \sqrt{n \mathbb{E}\left[ (\theta - \phi(Z^{*n}))^2 \right]} \right)^2. \tag{54}$$

Next, we evaluate the term $\mathbb{E}\left[ (\theta - \phi(Z^{*n}))^2 \right]$ in (53). Set $\bar{d} = 1 - 2^{-2R}$. The inverse channel to (39) is given by

$$\begin{aligned} Z_i^* &= \bar{d} Y_i + \sigma \sqrt{2^{-2R} \bar{d}} \mathcal{N}(0,1) \\ &= \bar{d}(\theta + \sigma \mathcal{N}(0,1)) + \sigma \sqrt{\bar{d} 2^{-2R}} \mathcal{N}(0,1) \\ &= \bar{d} \theta + \sigma \sqrt{\bar{d}(1 + 2^{-2R})} \mathcal{N}(0,1). \end{aligned}$$

Set $\alpha = 1/\bar{d} = 1/(1 - 2^{-2R})$. It follows that

$$\phi(Z^{*n}) = \theta + \sigma \sqrt{\frac{1 + 2^{-2R}}{\bar{d}}} \frac{1}{n} \sum_{i=1}^{n} W_i,$$

where $W^n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, and thus

$$n \mathbb{E}\left[ (\theta - \phi(Z^{*n}))^2 \right] = \sigma^2 \frac{1 + 2^{-2R}}{1 - 2^{-2R}}.$$

The proof is completed by substituting $\alpha$ and the last expression in (54).