Universal Compression of High Dimensional Gaussian Vectors with James-Stein shrinkage

Narayana Prasad Santhanam[1] and Mayank Bakshi[2]

[1]University of Hawaii, Manoa
[2]Arizona State University

2023 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-7554-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISIT54713.2023.10206798

***Abstract*—We study universal compression of n *i.i.d.* copies of a $k-$variate Gaussian random vector, when the mean is an unknown vector in an Euclidean ball of $\mathbb{R}^k$, and the covariance is known. We adopt the high dimensional scaling $k = \Theta(n)$ to bring out a compression perspective on the inadmissibility of unbiased estimates of a $k-$variate Gaussian (when $k \geq 3$), in particular focusing on the optimal unbiased Maximum Likelihood estimate. We use arguments based on the redundancy-capacity theorem to show that the redundancy of a universal compressor in this high dimensional setting must be lower bounded as $\Theta(n)$. We show that natural compression schemes based on the Maximum Likelihood estimate of the mean have suboptimal $\Theta(n \log n)$ redundancy, but a scheme based on the James-Stein biased estimate of the mean incurs redundancy that is also $\Theta(n)$.**

I. Introduction

The paper derives compression results on $k-$variate Gaussian vectors in the context of Stein's seminal result [1] showing that maximum likelihood estimation (MLE) is inadmissible when $k \geq 3$. In this paper, we primarily deal with real valued random variables described by probability measures absolutely continuous with respect to the Lebesgue measure. *Compression* refers simply to an assignment of a probability density function (pdf) on an appropriate set, a standard convention in vogue, see for example [2], [3], [4].

We consider *universally compressing* n *i.i.d.* Gaussian vectors $X_1, \ldots, X_n$, where $X_i \in \mathbb{R}^k$ are distributed as $k-$variate Gaussian random vectors with mean μ and covariance I. The mean μ is not revealed to us, we only know $\mu \in A$ for some $A \subset \mathbb{R}^k$. Therefore, we choose a *universal pdf* (or a universal scheme) q on $X_1, \ldots, X_n$ that is agnostic to μ. We will refer to $\mathbb{E} \log q(X_1, \ldots, X_n)$ (the expectation is taken with respect to the true distribution of $X_1, \ldots, X_n$) as the *description length* of q, in analogy with discrete random variables. Non-negativity of the KL divergence implies that the universal pdf q must have description length no smaller (usually longer) than the description length of the actual pdf of $X_1, \ldots, X_n$, the excess description length of the universal q being termed the *redundancy* of q.

The unknown mean is a $k-$coordinate vector. In the low-dimensional setting where k is held fixed and n increases, the redundancy of optimal universal schemes scales $\frac{k}{2} \log n$ as expected [4], [5]. We show this scaling can also be achieved by a natural compressor based on the MLE, which describes the symbols $X_1, \ldots, X_n$ one by one. For any $l \geq 2$, the scheme describes X_l using a normal distribution with mean equal to the the MLE of the mean using $X_1, \ldots, X_{l-1}$.

However, in the high dimensional setting $k = \Theta(n)$, we show that the redundancy of the best compressors scales as $\Theta(n)$. The MLE based estimator is suboptimal here (and still scales as $\frac{k}{2} \log n$). Instead we show that a similarly constructed compressor, but one that describes symbols using the James-Stein shrinkage estimate [1], [6] of the mean achieves the correct $\Theta(n)$ scaling.

This suboptimality of the MLE based compressor in the high dimensional setting is a reflection of the broader *inadmissibility* result on MLE of the mean (when $k \geq 3$) under mean square risk. A few salient points about estimating the unknown mean μ of a $k-$variate Gaussian with known covariance (which we will take to be I), under mean square risk, are in order.

Given a single observation $X \in \mathbb{R}^k$ from such a $k-$variate Gaussian with unknown mean μ, the natural maximum likelihood estimator is simply X. This is the minimum variance unbiased estimator under mean square risk, *i.e.,* for all unbiased estimators $\hat{\mu}(X)$ and for every vector μ,

$$\mathbb{E}||X - \mu||^2 \leq \mathbb{E}||\hat{\mu}(X) - \mu||^2,$$

and in addition, it is minimax and achieves the Cramer-Rao bound [7], [8]. For $k = 1, 2$, the MLE is *admissible*, *i.e.,* there is no other estimator that has risk $\leq$ the MLE risk for every μ, and in addition has risk strictly better than the MLE for at least some choice of μ.

From a Bayesian perspective, if the mean is chosen from a prior $\mu \sim \mathcal{N}(0, \Sigma)$ and $X|\mu \sim \mathcal{N}(\mu, I)$, then the minimum mean square estimator is

$$\arg\min_{\hat{\mu}(X)} \mathbb{E}||\hat{\mu}(X) - \mu||^2,$$

where the expectation is over the joint distribution of μ and X. Finding the Bayes optimal estimator corresponds to ridge regression and yields a biased linear estimate, $\hat{\mu}_\Sigma = (\Sigma^{-1} + I)^{-1} X$.

In particular, the biased Bayes optimal estimator beats the MLE when the mean square risk against the mean is averaged over the prior on the mean, and it is not surprising that it should be so. Furthermore, note that under the Bayes setting, the different components of X (after marginalizing out μ using its prior) are rendered dependent, and it makes sense that $\hat{\mu}_\Sigma$ uses all components of the vector X to estimate any single component of the vector μ.

This work was in part supported by the NSF Science & Technology Center for Science of Information Grant number CCF-0939370; and by the National Science Foundation under Grant No. CCF-2107526. Author emails: nsanthan@hawaii.edu, mayank.bakshi@ieee.org.

What is remarkable is that when $k \geq 3$ we can throw out the entire Bayes setup, but estimators conceptually related to the Bayes optimal estimator will still beat the MLE for *every* value of μ in the frequentist setup as well! Specifically, Stein showed in [1] that the MLE is *inadmissible* when $k \geq 3$, by constructing an estimator that is strictly better than the MLE for all choices of the parameter μ.

The estimator proposed by Stein can be surmised from several angles—Stein's argument in [1] or an empirical Bayes argument [9] that starts off with the Bayesian perspective described above. Formally, the James-Stein (JS) shrinkage estimator estimates the mean using

$$\hat{\mu}_{JS} = \left(1 - \frac{k-2}{||X||^2}\right) X,$$

and for all μ and $X \sim \mathcal{N}(\mu, I)$, it can be shown that

$$E||\hat{\mu}_{JS} - \mu||^2 = k - \mathbb{E}\left(\frac{k-2}{||X||^2}\right) < k = E||X - \mu||^2.$$

Remarkably, the JS shrinkage does something similar to the Bayes-optimal estimator: it uses every component of X to estimate any single component of μ, even though the components of X are *independent* in the frequentist setup above, and this achieves better mean square risk than the MLE.

The improvement is not trivial either, indeed when $\mu = \mathbf{0}$, the JS estimator has a mean square risk of 2 (no matter what k is), in contrast to the MLE mean square risk of k. In high dimensional settings where k is large, this is a substantial improvement.

As is to be expected, this thoroughly counterintuitive defeat of the MLE on its own turf shook the statistical world, and the JS shrinkage estimator is now extensively used in high dimensional settings where its gains can be substantial. See for example, [10], [11], [12].

We outline the formulation and formal statements of the main results in Section II, these involve some technical refinements of the standard formulations. One distraction that arises here is that universally describing a single $k-$variate Gaussian vector X whose unknown mean $\mu \in \mathbb{R}^k$ incurs infinite redundancy. This is however not a serious concern and can be circumvented by either choosing $\mu \in A$ for a compact subset $A \subset \mathbb{R}^k$ (reminiscent of the power constraint in Gaussian channel capacities) or by conditioning on a single training example. Regardless, our focus is the scaling of redundancy when $k = \Theta(n)$, and we adopt whichever solution highlights this aspect of the problem. The redundancy of the JS derived estimate is in general hard to analyze, since it depends on the expectation of the inverse of a non-central χ^2 random variable that does not have a closed form. Therefore, Section III develops a series of abstractions that reveal the correct order of the redundancy of the JS based compressors.

II. Problem Setup and Main Results

Let $n, k \in \mathbb{N}$. Let $X_1, , \ldots, X_n$, $X_i \in \mathbb{R}^k$ be drawn i.i.d. from $p_\mu = \mathcal{N}(\mu, I)$ for an unknown $\mu \in A$, where $A \subset \mathbb{R}^k$. We will use $\mathbb{P}(\mathbb{R}^k)$ to denote the set of all possible pdfs (not necessarily Gaussian) over the $k-$variate vectors, and $\mathbb{P}(\mathbb{R}^{kn})$ for the set of all possible pdfs over $\mathbb{R}^{kn}$. An Euclidean (ℓ_2) ball of radius R centered around a vector $\mathbf{m} \in \mathbb{R}^k$ is denoted by $B(\mathbf{m}, R)$ (balls in other spaces will come into play, so there is no k in the notation).

A universal length-n compressor for $X_1, \ldots, X_n$ is a pdf $q \in \mathbb{P}(\mathbb{R}^{kn})$, where the interpretation is that this pdf q is chosen to describe $X_1, \ldots, X_n$ no matter what μ is. We characterize the performance of such universal schemes using the (minimax, or average-case) *redundancy*, which has the interpretation of excess codelength in the discrete case.

The redundancy associated with a length n-compressor $q^{(n)} \in \mathbb{P}(\mathbb{R}^{kn})$ against a source with mean μ is defined as

$$R_n(\mu, q^{(n)}) = \mathbb{E}_{p_\mu} \ln \frac{p_\mu(X_1 X_2 \ldots X_n)}{q^{(n)}(X_1 X_2 \ldots X_n)}.$$

Note that we do not normalize by n as is sometimes done. For any subset $A \subset \mathbb{R}^k$,

$$R_n(A, q^{(n)}) = \sup_{\mu \in A} \mathbb{E}_{p_\mu} \ln \frac{p_\mu(X_1 X_2 \ldots X_n)}{q^{(n)}(X_1 X_2 \ldots X_n)},$$

and, the minimax redundancy of A is

$$R_n(A) = \inf_{q^{(n)} \in \mathbb{P}(\mathbb{R}^{kn})} \sup_{\mu \in A} R_n(\mu, q^{(n)}).$$

It is well known [13] that the inf above can be replaced by a min.

As mentioned in the introduction, we have to deal with one distraction in this problem setting:

Proposition 1: For all $n \geq 1$, $R_n(\mathbb{R}^k) = \infty$. □

While the above result seems negative, the situation is not quite as dire. Even with a "training sample" of 1—the universal compressor gets to observe X_0 from the unknown source, the infinity is no longer in play as we describe below. Meaning that the infinity is simply an artifact of having to describe the first sample. Similarly, if A is a compact set, say $B(\mathbf{0}, \sqrt{R})$ for a constant $R > 0$, the redundancy is again finite. Either of these workarounds continues to retain focus on the *scaling* of redundancy R_n when $k = \Theta(n)$.

A. Conditional redundancy

Let $X_i \in \mathbb{R}^k$ and let $X_0, X_1, X_2 \ldots$ be iid $p_\mu = \mathcal{N}(\mu, I)$, where $\mu \in \mathbb{R}^k$. Consider the quantity for $q \in \mathbb{P}(\mathbb{R}^{kn})$

$$\begin{aligned} R_n(\mu, q|X_0) &= \mathbb{E}\big[\ln \frac{p_\mu(X_1, \ldots, X_n|X_0)}{q(X_1, \ldots, X_n|X_0)}\big] \\ &= \mathbb{E}\Big[\mathbb{E}\big[\ln \frac{p_\mu(X_1, \ldots, X_n|X_0)}{q(X_1, \ldots, X_n|X_0)} \mid X_0\big]\Big]. \end{aligned}$$

Similarly, for any subset $A \subset \mathbb{R}^k$,

$$R_n(A|X_0) = \inf_{q \in \mathbb{P}(\mathbb{R}^{kn})} \sup_{\mu \in A} R_n(\mu, q|X_0).$$

One can interpret the above as the redundancy if we are allowed a training sample of size 1. Even though the redundancy $R_1(\mathbb{R}^k) = \infty$, Proposition 2 below shows that for all n, $R_n(\mathbb{R}^k, q_{ML}|X_0) < \infty$. We will often use analogously defined $R_1(\mu, q|X_0, \ldots, X_{n-1})$ and $R_1(A|X_0, \ldots, X_{n-1})$ as well.

B. Estimation based compressors

Suppose we have an estimate $f(X_0,\ldots,X_{n-1})$ of the mean. If we take the universal q_f to be the natural extension of an estimator: *i.e.,* $q_f(X_n|X_0,\ldots,X_{n-1})$ estimates the mean vector as $f(X_0,\ldots,X_{n-1})$ and then describes X_n using $\mathcal{N}(f(X_0,\ldots,X_{n-1}),I)$. Then

$$\mathbb{E}\big[\ln\frac{p_\mu(X_n|X_0,\ldots,X_{n-1})}{q_f(X_n|X_0,\ldots,X_{n-1})}\big]=\frac{1}{2}\mathbb{E}[||f(X_0,\ldots,X_{n-1})-\mu||^2].$$

To see the above, note that

$$\begin{aligned}&\mathbb{E}\big[\ln\frac{p_\mu(X_n|X_0,\ldots,X_{n-1})}{q(X_n|X_1,\ldots,X_{n-1})}\mid X_0,\ldots,X_{n-1}\big]\\&=-\mathbb{E}\frac{1}{2}||X_n-\mu||^2\\&\quad+\mathbb{E}\big[\frac{1}{2}||X_n-f(X_0,\ldots,X_{n-1})||^2\mid X_0,\ldots,X_{n-1}\big],\end{aligned}$$

Further simplifications are straightforward if a little monotonous. Thus the conditional redundancy is closely related to the regular mean square risk.

We will primarily deal with the Maximum Likelihood (ML) estimator and the James Stein (JS) estimators. Given $X_1,X_2,\ldots,X_n\in\mathbb{R}^k$, let

$$\bar{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$$

denote the sample mean, and this coincides with the **Maximum Likelihood** estimate.

Definition 1 (James-Stein shrinkage estimate): The **James-Stein** estimator shrunk towards a point $\mathbf{c}\in\mathbb{R}^k$ is

$$\hat{\mu}_{JS}^{(\mathbf{c})}=\mathbf{c}+\left(1-\frac{k-2}{n||\bar{X}-\mathbf{c}||^2}\right)(\overline{X}-\mathbf{c}).\tag{1}$$

In the above, when $\mathbf{c}=\mathbf{0}$, we drop the superscript for simplicity.

Let q_{ML} (q_{JS} respectively) be compressors based on the ML (JS shrunk to $\mathbf{0}$ respectively) estimator of the mean. The compressors based on $\hat{\mu}_{JS}^{(\mathbf{c})}$ will be denoted by $q_{JS}^{(\mathbf{c})}$.

The following proposition follows immediately.

Proposition 2: For all $n\geq 1$, and all $\mu\in\mathbb{R}^k$,

$$R_1(\mu,q_{ML}|X_0,\ldots,X_{n-1})=\frac{k}{2n},$$

and hence $R_n(\mathbb{R}^k,q_{ML}|X_0)\leq\frac{k}{2}\log(n+1)$. □

This redundancy corresponds to the scaling $\frac{k}{2}\ln n$ anticipated in [4], since the distributions are parameterized by the k coordinates of the unknown mean.

The connection with mean square risk immediately implies the following corollary.

Corollary 1: [1] For all $\mu\in\mathbb{R}^k$ and all n,

$$R_1(\mu,q_{JS}|X_0,\ldots,X_{n-1})<R_1(\mu,q_{ML}|X_0,\ldots,X_{n-1})=\frac{k}{2n}.$$

Thus, for all $\mu\in\mathbb{R}^k$ and all n, $R_n(\mu,q_{JS}|X_0)<R_n(\mu,q_{ML}|X_0)$. □

We can do marginally better in the ML case. Consider obtaining the distribution $q_{ML}(X_n|X_0,\ldots,X_{n-1})$ by first estimating the mean using $\bar{X}_n$, but the follow-up encoding of X_n is done using $\mathcal{N}(\bar{X}_n,\sigma^2 I)$. In general σ^2 can be chosen to be a function of $X_1,\ldots,X_n$. If we choose to only keep it as a function of n, the optimal choice can be seen to be $1+\frac{1}{n}$; however, this choice does not change the order of magnitude of $R_n(\mathbb{R}^k,q_{ML}|X_0)$. In this conference version we disregard this potential minor improvement for more transparency.

Let $B(\mathbf{c},\sqrt{R})$ be a ℓ_2 ball of radius $\sqrt{R}$. Our primary result is a universal compressor based on James-Stein shrinkage that achieves a redundancy of $\Theta(n)$.

Theorem 2: For all real $r>0$, there exists a sequential probability assignment on $\mathbb{R}^\infty$ that achieves for all $n\geq 1$,

$$\begin{aligned}&R_n(B(\mathbf{0},\sqrt{R}),q)\\&\quad\leq\frac{k}{2}\log\frac{r+1}{r+\frac{1}{n}}+\frac{r}{r+\frac{1}{n}}\log n+\frac{k}{2}\log\frac{2Rrn}{k}+\mathcal{O}(\log n).\end{aligned}$$

□

The compressor q noted in the Theorem above is a refinement of q_{JS} and the proof of the Theorem above is completed in Section III-C. Note that the redundancy above scales as $\Theta(n)$ when $k=\Theta(n)$. We have a lower bound of $\Theta(n)$ in Section IV. We also note that the redundancy of q_{ML} against sources in $B(\mathbf{0},\sqrt{R})$ is still $\frac{k}{2}\log n+\mathcal{O}(1)$, matching the order of magnitude of the result in Proposition 2 for $\mathbb{R}^k$.

III. James-Stein shrinkage based compressors

As noted earlier, in the low dimensional regime when k is held fixed and $n\to\infty$, both ML and JS based constructions achieve the $\frac{k}{2}\log n$ scaling. Indeed, the advantages of the JS estimator are revealed in the high dimensional setting, and this reflects the fact that it is in these settings that the estimator is commonly used.

Lemma 3: [1] When $X_0,\ldots,X_{n-1}\sim\mu$ (i.i.d.) and f_{JS} is the James-Stein estimator (shrunk to $\mathbf{0}$), then

$$E||f_{JS}(X_1,\ldots,X_n)-\mu||^2=\frac{k}{n}-\frac{(k-2)^2}{n^2}\mathbb{E}\frac{1}{||\bar{X}_n||^2}.$$

□

Note that if $\mu=\mathbf{0}$, $n||\bar{X}_n||^2$ is a (central) χ^2 distribution, and the expectation of its inverse is known[1]:

$$\mathbb{E}\frac{1}{n||\bar{X}_n||^2}=\frac{1}{k-2}.$$

Then the risk can be simplified to

$$\frac{k}{n}-\frac{(k-2)^2}{n^2}\mathbb{E}\frac{1}{||\bar{X}_n||^2}=\frac{k}{n}-\frac{k-2}{n}=\frac{2}{n},$$

which is independent of k altogether, as noted before in the Introduction.

When $\mu\neq\mathbf{0}$, $n||\bar{X}_n||^2$ is a non-central χ^2- variable with k degrees of freedom and non-centrality parameter $n||\mu||^2$. The expectation of its reciprocal is known, but not easily analyzed, and generally expressed as a hypergeometric series.

We construct a sequential compression scheme that achieves the right order of magnitude in the aforementioned high

[1]Note that the inverse chi-squared distribution with k degrees of freedom is a special case of the inverse gamma (a,b) distribution (*c.f.* [14, pp 254]) with $a=k/2$ and $b=1/2$.

dimensional regime $k = \Theta(n)$. We do this in multiple steps: We build the general estimator in three steps:

- Given a horizon n and any $r > 0$, we show that when $k = \Theta(n)$, $R_n(B(\mathbf{0}, \sqrt{\frac{k-2}{nr}}), q_{JS}) = \Theta(n)$ as well.
- Given a horizon n, we cover $B(\mathbf{0}, \sqrt{R})$ with balls of radius $\sqrt{\frac{k-2}{nr}}$, with the centers of the covering in $\mathcal{M}$. We then use $q_{JS}^{(\mathbf{c})}$, $\mathbf{c} \in \mathcal{M}$, to construct an estimator q_n. Again, when $k = \Theta(n)$ and r is constant, $R_n(B(\mathbf{0}, \sqrt{R}), q_n = \Theta(n)$
- We then construct a horizon-free, sequential estimator for $B(\mathbf{0}, \sqrt{R})$ that also incurs $\Theta(n)$ redundancy when compressing strings of length n, where $k = \Theta(n)$. This will prove Theorem 2

A. Estimator for $B\big(\mathbf{m}, \sqrt{\frac{k-2}{nr}}\big)$

We evaluate the redundancy of sources in $B(\mathbf{m}, \sqrt{\frac{k-2}{nr}})$ (with $\mathbf{m}$ known) against the James-Stein estimator shrunk towards $\mathbf{m}$. The redundancy remains the same regardless of what $\mathbf{m}$ is, therefore w.l.o.g., we assume $\mu = \mathbf{0}$.

A note on the setting: since the collection implicitly contains knowledge of n, we call this a known-horizon setup. However, the estimator q_{JS} does not use the knowledge of n, only our analysis does.

Proposition 3: For all $R > 0$, $R_1(B(\mathbf{0}, \sqrt{R})) < \infty$. □

Therefore, in what follows, we assume that we describe the first sample X_1 using a fixed q^*, and that this step incurs redundancy κ. The exact scheme is not pertinent, but we can assume this is an optimal compressor for the single letter redundancy, q^*.

Theorem 4: For all integers $k \geq 3$ and $n \geq 1$ and real $r > 0$

$$R_n(B(\mathbf{0}, \sqrt{\frac{k-2}{nr}}), q_{JS}) \leq \frac{k}{2} \log \frac{r+1}{r+\frac{1}{n}} + \frac{r}{r+\frac{1}{n}} \log n + \kappa.$$

Therefore when $k = \Theta(n)$, $R_n(B(\mathbf{0}, \sqrt{\frac{k-2}{nr}}), q_{JS})$ scales linearly with n or k.

Proof Recall that for $l \geq 1$, if $X_1, \ldots, X_l \sim$ iid $\mathcal{N}(\mu, I)$, then

$$E||f_{JS}(X_1, \ldots, X_l) - \mu||^2 = \frac{k}{l} - \frac{(k-2)^2}{l^2} \mathbb{E} \frac{1}{||\bar{X}_l||^2}$$

To bound the expectation of the inverse non-central χ^2 random variable, we note that if $Z \sim \mathcal{N}(\nu, I)$, then

$$\mathbb{E} \frac{1}{||Z||^2} = \frac{\exp\left(-\frac{||\nu||^2}{2}\right)}{k-2} \sum_{j \geq 0} \frac{k-2}{k-2+2j} \frac{||\nu||^{2j}}{2^j j!}$$

A simple convexity argument yields that

$$E \frac{1}{||Z||^2} \geq \frac{1}{\mathbb{E}||Z||^2} = \frac{1}{k + ||\nu||^2}.$$

This bound can be improved marginally as follows. Note that for all $j \geq 0$

$$\int_{t \geq 0} e^{-(k-2+2j)t} dt = \frac{1}{k-2+2j},$$

so that

$$\begin{aligned}
\mathbb{E} \frac{1}{||Z||^2} &= \frac{\exp\left(-\frac{||\nu||^2}{2}\right)}{k-2} \sum_{j \geq 0} \frac{k-2}{k-2+2j} \frac{||\nu||^{2j}}{2^j j!} \\
&= \exp\left(-\frac{||\nu||^2}{2}\right) \int_{t \geq 0} e^{-\left((k-2)t - \frac{||\nu||^2}{2} e^{-2t}\right)} dt \\
&\geq \exp\left(-\frac{||\nu||^2}{2}\right) \int_{t \geq 0} e^{-\left((k-2)t - \frac{||\nu||^2}{2}(1-2t)\right)} dt \\
&= \frac{1}{k-2+||\nu||^2}. \qquad (2)
\end{aligned}$$

Therefore, when we are describing X_{l+1} given $X_1, \ldots, X_l$,

$$\begin{aligned}
\mathbb{E}||f_{JS}(X_1, \ldots, X_l) - \mu||^2 &= \frac{k}{l} - \frac{(k-2)^2}{l^2} \mathbb{E} \frac{1}{||\bar{X}_l||^2} \\
&\leq \frac{k}{l} - \frac{(k-2)}{l} \frac{k-2}{k-2+l||\mu||^2}
\end{aligned}$$

where the second inequality notes that $\sqrt{l}||\bar{X}_l|| \sim N(\sqrt{l}\mu, I)$, and uses the bound from (2).

Now for any fixed $r > 0$, since $||\mu||^2 \leq \frac{k-2}{nr}$ it follows that

$$\begin{aligned}
E||f_{JS}(X_1, \ldots, X_l) - \mu||^2 &< \frac{k}{l} - \frac{(k-2)}{l} \frac{k-2}{k-2+l||\mu||^2} \\
&\leq \frac{k}{l} - \frac{(k-2)r}{l(r+\frac{l}{n})} \\
&= \frac{k}{n(r+\frac{l}{n})} + \frac{2r}{l(r+\frac{l}{n})}
\end{aligned}$$

Therefore

$$\begin{aligned}
R_n(\mu, q_{JS}) = &= \frac{1}{2} \sum_{l=1}^{n} \mathbb{E}||f_{JS}(X_1, \ldots, X_l) - \mu||^2 \\
&\leq \frac{k}{2} \log \frac{r+1}{r+\frac{1}{n}} + \frac{r}{r+\frac{1}{n}} \log n + \kappa,
\end{aligned}$$

where the κ term accounts for the redundancy incurred in describing X_1. The theorem now follows. □

B. Redundancy of $B(0, \sqrt{R})$

When the horizon n is known, universal compressors for $B(\mathbf{0}, \sqrt{R})$ can leverage the estimators we obtained for $B(\mathbf{0}, \sqrt{\frac{k-2}{rn}})$.

Theorem 5: For all integers $k \geq 3$ and $n \geq 1$, all real $r > 0$, there is $q_n \in \mathbb{P}(\mathbb{R}^{kn})$ depending on n satisfying

$$\begin{aligned}
&R_n(B(\mathbf{0}, \sqrt{R}), q_n) \leq \\
&\frac{k}{2} \log \frac{r+1}{r+\frac{1}{n}} + \frac{r}{r+\frac{1}{n}} \log n + \frac{k}{2} \log \frac{2Rrn}{k} + \kappa.
\end{aligned}$$

Proof We cover the ball $B(\mathbf{0}, \sqrt{R})$ with balls of radius $\sqrt{\frac{k-2}{rn}}$. Let $\mathcal{M} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_{|\mathcal{M}|}\}$ be the set of centers of the radius-$\sqrt{\frac{k-2}{rn}}$ balls in the covering. Standard volume based covering arguments imply that

$$|\mathcal{M}| \leq \left(\sqrt{\frac{2Rrn}{k-2}} + 1\right)^k.$$

Then we define for all $\mathbf{c} \in \mathcal{M}$,

$$q_{\mathbf{c}}(X_1,\ldots,X_n) = q^* \prod_{i=1}^{n-1} q_{\mathbf{c}}(X_{i+1}|X_1,\ldots,X_i),$$

where q^* is the optimal single letter encoder promised in Proposition 3 and $q_{\mathbf{c}}$ is the natural construction that uses the JS estimator shrunk towards $\mathbf{c}$. We define

$$q_n(X_1,\ldots,X_n) \stackrel{\text{def}}{=} \frac{1}{|\mathcal{M}|} \sum_{\mathbf{c}\in\mathcal{M}} q_{\mathbf{c}}(X_1,\ldots,X_n)).$$

For $\mu \in B(\mathbf{0},\sqrt{R})$, let $\mathbf{c}(\mu)$ be any one of the elements of $\mathcal{M}$ closest to μ in Euclidean distance. It follows that $||\mathbf{c}(\mu) - \mu||^2 \le \frac{k-2}{rn}$. To compute the redundancy of q_n, note that for any constant r,

$$\begin{aligned} &D_n(p_\mu(X_1,\ldots,X_n)||q_n(X_1,\ldots,X_n)) \\ &\quad\le D_n(p_\mu(X_1,\ldots,X_n)||q_{\mathbf{c}(\mu)}(X_1,\ldots,X_n)) + \log|\mathcal{M}| \\ &\quad\le \frac{k}{2}\log\frac{r+1}{r+\frac{1}{n}} + \frac{r}{r+\frac{1}{n}}\log n + \kappa + \frac{k}{2}\log\frac{2Rrn}{k}. \end{aligned}$$

which is $\Theta(n)$ when $k = \Theta(n)$. Also, in the low dimensional regime for any fixed k, and $n \to \infty$, the redundancy here scales asymptotically as $\frac{k}{2}\log n$ as expected. □

C. Horizon-free schemes

If the horizon n is not known in advance, it is still possible to achieve essentially the same redundany $R_n(B(0,\sqrt{R}), q_n)$ as if the horizon were known, with an additional penalty of roughly $2\log n$.

Proof of Theorem 2: We first extend each conditional probability density $q_n(X_1,\ldots,X_n|X_0)$ from Theorem 5 to a probability measure over semi-infinite real sequences $X_0, X_1, \ldots$. To do so, we assume a standard normal distribution over X_0. The assignment of probability density function values to sequences up to $n+1$ is settled by q_n defined above and its marginals. The density values for sequences longer than $n+1$ extend the values assigned to length n sequences in an arbitrary (but fixed) fashion. We will refer to this probability measure also as q_n.

Then, the universal measure q is the probability measure that extends the density assignment

$$q(X_1,\ldots,X_m) = \sum_{n\ge 1} \frac{1}{n(n+1)} q_n(X_1,\ldots,X_m)$$

over $\mathbb{R}^\infty$. Clearly now, for all n,

$$\begin{aligned} &D_n(p_\mu(X_1,\ldots,X_n)||q(X_1,\ldots,X_n)) \le \\ &\quad D_n(p_\mu(X_1,\ldots,X_n)||q_n(X_1,\ldots,X_n)) + \log(n(n+1)), \end{aligned}$$

and the theorem follows. □

IV. Lower bound on the redundancy

When $k = \Theta(n)$, the James-Stein based estimators obtain a redundancy of $\Theta(n)$ when compressing sequences $X_1,\ldots,X_n$ with X_i drawn *i.i.d.* from $\mathcal{N}(\mathbf{0}, I)$ distribution. This scaling is optimal, and no estimator can yield a redundancy that is $o(n)$.

To prove this, we use the following well known lower bound on the redundancy, see for example [15].

Lemma 6: Let $\mathcal{P}$ be a collection of pdfs on $\mathbb{R}$. For $1 \le i \le M$, let $S_i \subset \mathbb{R}$, and assume that these sets are pairwise disjoint. Suppose that for each i there exists $p_i \in \mathcal{P}$ such that $p_i(S_i) \ge \delta$. Then, for all pdfs q, we have

$$\sup_{p\in\mathcal{P}} D(p||q) \ge \delta\log(M) - 1. \qquad \square$$

For any r, we use the fact that the $B(\mathbf{0},\sqrt{R})$ ball can be packed with $\ge \left(\sqrt{\frac{Rrn}{k-2}}+1\right)^k$ balls with radius $\frac{k-2}{2nr}$. Let $\mathcal{N}$ be the set of centers of this packing. Then, we have for each $c \in \mathcal{N}$,

$$p_{\mathbf{c}}\left(\bar{X}_n \in B\left(\mathbf{c}, \sqrt{\frac{k-2}{2rn}}\right)\right) \ge 1-\gamma,$$

whenever $r \le \frac{1}{1+2\sqrt{\frac{\ln\frac{1}{\gamma}}{k}}+2\frac{\ln\frac{1}{\delta}}{k}}$, and therefore the redundancy of compressing length n sequences $X_1,\ldots,X_n$ generated *i.i.d.* p_μ, $\mu \in B(\mathbf{0},\sqrt{R})$ is at least

$$(1-\gamma)k\log\frac{Rrn}{k},$$

which, when $k = \Theta(n)$ grows as $\Theta(n)$.

Comparison with Pinsker's bound

The classical bound due to Pinker [16] establishes that the biased linear estimator, $f_L(\bar{X}_l) = \bar{X}_l \frac{\frac{l}{nr}}{\frac{l}{nr}+1}$, with risk

$$\sup_{||\mu||^2\le\frac{k}{rn}} \mathbb{E}||f_L(\bar{X}_l) - \mu||^2 = \frac{k}{n(r+\frac{l}{n})}$$

is asymptotically minimax when $\mu \in B(\mathbf{0},\sqrt{\frac{k}{rn}})$ and $k \to \infty$, *i.e.,* for any l and n,

$$\liminf_{k\to\infty} \inf_{\hat{f}} \sup_{||\mu||^2\le\frac{k}{rn}} \frac{1}{k}\mathbb{E}||\hat{f}(\bar{X}_l) - \mu||^2 \ge \frac{1}{n(r+\frac{l}{n})}.$$

Note that Pinsker's result is asymptotic. Unlike the James-Stein estimate, the biased linear estimator requires knowledge of the radius k/rn of the sphere from which the unknown mean is drawn. Indeed, one way to interpret the James-Stein estimate is that it is essentially universal over all values of rn. In addition, as we have seen, the bound suggested by Pinsker's bound is not just asymptotic, but holds for all lengths.

V. Conclusion

We showed how the classical result on inadmissibility of the Maximum Likelihood (ML) estimate under mean square risk applies to universal compression. We considered universal compression of n *i.i.d.* copies $k-$variate Gaussian vectors (mean unknown, covariance I). Focusing on the regime where $k = \Theta(n)$ sheds light on the suboptimality of the ML mean risk, wherein the ML based schemes achieve a worse order of magnitude than the James Stein (JS) shrinkage compressors, and where the latter are shown to be order optimal.

REFERENCES

[1] C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," *Berkeley Symp. on Math. Statist. and Prob.*, pp. 197–206, 1956.

[2] L. Davisson, "Universal noiseless coding," *IEEE Transactions on Information Theory*, vol. 19, no. 6, pp. 783–795, 1973.

[3] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2743–2760, 1998.

[4] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, 1996.

[5] M. Drmota and W. Szpankowski, "Precise minimax redundancy and regrets," *IEEE Trans. Information Theory*, vol. 50, pp. 2686–2707, 2004.

[6] W. James and C. Stein, "Estimation with quadratic loss," *Berkeley Symp. on Math. Statist. and Prob.*, pp. 361–379, 1961.

[7] Harald Cramer, *Mathematical Methods Of Statistics*. 1946.

[8] C. Radhakrishna Rao, "Information and the accuracy attainable in the estimation of statistical parameters," *Bulletin of the Calcutta Mathematical Society*, vol. 37, pp. 81–91, 1945.

[9] N. Reid. Lecture notes. Available at https://utstat.toronto.edu/reid/sta2212s/2021/LSIChapter1.pdf.

[10] "Presidential elections polling." FiveThirtyEight.com. https://perma.cc/E4FC-6WSL.

[11] J. Hausser and K. Strimmer, "Entropy inference and the james-stein estimator, with application to nonlinear gene association networks," *Journal of Machine Learning Research*, vol. 10, no. 50, pp. 1469–1484, 2009.

[12] R. I. Jennrich and S. D. Oman, "How much does stein estimation help in multiple linear regression?," *Technometrics*, vol. 28, no. 2, 1986.

[13] D. Haussler and M. Opper, "Mutual information, metric entropy and cumulative relative entropy risk," *The Annals of Statistics*, vol. 25, no. 6, pp. 2451 – 2492, 1997.

[14] P. D. Hoff, *A first course in Bayesian statistical methods*, vol. 580. Springer, 2009.

[15] N. Santhanam, V. Anantharam, and W. Szpankowski, "Data-derived weak universal consistency," *Journal of Machine Learning Research*, vol. 23, no. 27, pp. 1–55, 2022.

[16] M. S. Pinsker, "Optimal filtering of square-integrable signals in gaussian noise," *Problemy Peredachi Informatsii*, vol. 16, no. 2, pp. 52–68, 1980.