SPARSE DICTIONARY LEARNING FROM 1-BIT DATA

Jarvis D. Haupt, Nikos D. Sidiropoulos, and Georgios B. Giannakis

Department of Electrical and Computer Engineering
University of Minnesota, Minneapolis MN

ABSTRACT

This work examines a sparse dictionary learning task – that of fitting a collection of data points, arranged as columns of a matrix, to a union of low-dimensional linear subspaces – in settings where only highly quantized (single bit) observations of the data matrix entries are available. We analyze a complexity penalized maximum likelihood estimation strategy, and obtain finite-sample bounds for the average per-element squared approximation error of the estimate produced by our approach. Our results are reminiscent of traditional parametric estimation tasks – we show here that despite the highly-quantized observations, the normalized per-element estimation error is bounded by the ratio between the number of "degrees of freedom" of the matrix and its dimension.

Index Terms— Sparse dictionary learning, complexity regularization, maximum likelihood estimation

1. INTRODUCTION

Our problem of interest here is, fundamentally, an estimation task – we aim to estimate mn real-valued elements $\{X^*_{i,j}\}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$, denoted collectively as the matrix $\mathbf{X}^* \in \mathbb{R}^{m \times n}$, from a total of mn observations corresponding to one per entry of the matrix. Such estimation tasks are, of course, trivial without further qualifications; here rather than observe the elements of $\mathbf{X}^*$ directly, we obtain only highly quantized (1-bit) observations, one per matrix entry. The question we address here is, can one still obtain a consistent estimate of $\mathbf{X}^*$ in these settings?

We establish below that the answer is affirmative whent the matrix $\mathbf{X}^*$ exhibits some form of intrinsic low-dimensional *structure*. Generally speaking, we are interested here in settings where the number of parameters or "degrees of freedom" required to specify or accurately model $\mathbf{X}^*$ is many fewer than the ambient or extrinsic dimension mn. Our particular focus here will be on sparse dictionary models for $\mathbf{X}^*$, where we assume that the unknown matrix $\mathbf{X}^*$ can be expressed as a product of an $m \times p$ matrix $\mathbf{D}^*$ (called a *dictionary*) and a $p \times n$ matrix $\mathbf{A}^*$ of coefficients comprised of n columns each having $k < k_{\max} < p$ nonzero elements. Note that even though the matrix has mn elements, the number of degrees of freedom associated with this parameterization is only $\mathcal{O}(m \cdot p + \|\mathbf{A}^*\|_0)$, where $\|\mathbf{A}^*\|_0$ denotes the number of non zeros in $\mathbf{A}^*$.

Our main results here establish that (under assumptions to be formalized) we can obtain an estimate $\widehat{\mathbf{X}}$ from the quantized data that satisfies $\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2/mn = \mathcal{O}\left((m \cdot p + \|\mathbf{A}^*\|_0)/mn\right)$ with high probability (over the randomness in our observation model). That the error rate exhibit characteristics of the well-known parametric estimation error rate is intuitively pleasing; that such rates are achievable from the highly quantized data is, perhaps, surprising.

Author emails: {jdhaupt, nikos, georgios}@umn.edu. This work is supported by the NSF EARS Project, Award No. AST-1247885.

Our investigation here is in the spirit of 1-bit compressed sensing works in the sparse inference literature, which examine tasks of sparse *vector* estimation from one-bit compressive measurements [1–7], though our approach here is not "compressive" per se since the number of observations is equal to the dimension of the object we aim to estimate. Closely related to ours is the recent work [8], which examined matrix completion tasks from a subset of highly quantized measurements of a matrix. We adopt here an observation model somewhat reminiscent of the observation model of [8] (although our approach here is based on full, not compressive, measurements), and our estimation approach is based on a maximum-likelihood strategy, as in [8]. That said, while the authors of [8] analyzed a *convex* program for their matrix completion estimation task, here we examine the sparse dictionary learning task which is well-known to be jointly non-convex in its parameters. Indeed, our proposed estimation strategy here is non-convex (in fact, it is combinatorial); in practice, one could solve our proposed estimation problem via greedy methods or convex relaxation, along the lines of existing efforts in sparse dictionary learning [9–13]. Several recent works that have established identifiability conditions for greedy [14] and convex [15–17] approaches to the dictionary learning problem.

Our analysis approach is based on techniques from complexity penalized maximum likelihood estimation, following along the lines of [18–23], as well as prior work employing such techniques in sparse inference tasks [24]. The complexity penalized maximum likelihood formulation is closely related to the minimum description length (MDL) principle [25]; in that sense, we note [26] that proposed MDL formulations for several dictionary learning tasks (but without theoretical performance guarantees, as is our focus here). Finally, we note several prior efforts that examined quantization as a form of bandwidth constraint in parametric estimation tasks [27–31], and investigated conventional (non-complexity-penalized) maximum likelihood estimation approaches, as well as universal approaches that were agnostic to the distribution of the underlying noises that contaminate each observation.

The remainder of this paper is organized as follows. Following the formalization of our problem in Section 2, we provide our main theoretical result, and its implication for the sparse dictionary learning task, in Section 3. The proof of our main result, along with several intermediate lemmata, are provided in Section 4. We briefly discuss a few conclusions in Section 5.

2. PROBLEM FORMULATION

The dictionary-based factorization model described above represents an ideal decomposition; in practical settings, rather than the data adhering *exactly* to such a model, it is more likely only well-approximated by the assumed model. The model "mismatch" in these cases could arise because of true modeling error, or some form of stochastic noise present in the data, or both. Here, we explicitly

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

model such nonidealities via the quantities

$$Y_{i,j} = X^*_{i,j} - W_{i,j},\ i \in [m],\ j \in [n],$$

where the $\{W_{i,j}\}_{i\in[m],j\in[n]}$ are iid continuous zero-mean real scalar random variables[1], and where $[n] = \{1, 2, \ldots, n\}$ denotes the set of non-negative integers less than or equal to n. For $w \in \mathbb{R}$, we denote by $f_W(w)$ and $F_W(w)$ the (common) probability density function and cumulative distribution function (cdf), respectively, of the W_i's. We use the shorthand $\mathbf{Y}$ to denote the collection $\{Y_{i,j}\}_{i\in[m],j\in[n]}$.

Rather than observe $\mathbf{X}^*$ (or even $\mathbf{Y}$, for that matter) directly, here we assume that we obtain observations that are each quantized to a single bit. Specifically, we make observations of the form

$$Z_{i,j} = \mathbf{1}_{\{Y_{i,j}\geq 0\}} = \begin{cases} 1, & \text{if } W_{i,j} \leq X^*_{i,j} \\ 0, & \text{otherwise} \end{cases}, \tag{1}$$

for $i \in [m]$ and $j \in [n]$, so that the collection of observations $\{Z_{i,j}\}_{i\in[m],j\in[n]}$ (denoted here by $\mathbf{Z}$, for shorthand) comprises a total of mn *bits*. Note that the independence of the $\{W_{i,j}\}_{i\in[m],j\in[n]}$ implies that the elements of $\mathbf{Z}$ are also independent.

Given this model, each $Z_{i,j}$ is easily seen to be a Bernoulli random variable. We denote $\pi(X^*_{i,j}) \triangleq \Pr(Z_{i,j} = 1) = \Pr(W_{i,j} \leq X^*_{i,j}) = F_W(X^*_{i,j})$ where, as discussed above, F_W denotes the cdf of the modeling error terms, and denote the joint pmf of $\mathbf{Z}$ by $p_{\pi(\mathbf{X}^*)}(\mathbf{z}) = \prod_{i\in[m],j\in[n]} p_{\pi(X^*_{i,j})}(z_{i,j})$, where $\mathbf{z}$ is shorthand for $\{z_{i,j}\}_{i\in[m],j\in[n]}$, and each scalar pmf is given by $p_{\pi(X^*_{i,j})}(z_{i,j}) = \left[F_W(X^*_{i,j})\right]^{z_{i,j}} \left[1 - F_W(X^*_{i,j})\right]^{1-z_{i,j}}$, for $z_{i,j} \in \{0, 1\}$ and $i \in [m], j \in [n]$.

3. MAIN RESULT

Our inference approach here will be based on a variant of the maximum likelihood approach, in which we regularize the negative log-likelihood of each candidate reconstruction with a term that quantifies its "complexity," so that more complicated candidates have a larger cost in the overall objective function. Our approach will be to construct a rich set of candidate reconstructions $\mathcal{X}$ where each $\mathbf{X} \in \mathcal{X}$ exhibits the type of structure that we assume is present in the data $\mathbf{X}^*$, but where the class $\mathcal{X}$ contains candidate reconstructions requiring varying numbers of parameters to specify. Then, we construct the corresponding penalties for each of the elements of $\mathcal{X}$ to encourage simple estimates over more "complicated" estimates. Formally, we construct a countable collection $\mathcal{X}$ of candidate reconstructions for $\mathbf{X}^*$, and assign to each $\mathbf{X} \in \mathcal{X}$ a penalty, $\text{pen}(\mathbf{X}) > 0$, such that

$$\sum_{\mathbf{X}\in\mathcal{X}} 2^{-\text{pen}(\mathbf{X})} \leq 1. \tag{2}$$

The condition (2) is the well-known Kraft Inequality from coding theory; using this interpretation we have that for any $\mathcal{X}$ we may satisfy the condition (2) by constructing any binary *prefix code* over $\mathcal{X}$. With this, we are in position to state our main result.

Theorem 3.1. *Suppose that the elements of the unknown matrix $\mathbf{X}^*$ are bounded in amplitude, so that $\max_{i\in[m],j\in[n]} |X^*_{i,j}| \leq X_{\max}$ for some finite $X_{\max} > 0$, and let $\mathcal{X}$ be a countable collection of candidate reconstructions $\mathbf{X}$ with corresponding penalty functions $\text{pen}(\mathbf{X})$ satisfying (2), constructed so that each $\mathbf{X} \in \mathcal{X}$ is comprised of elements satisfying the uniform bound $\max_{i\in[m],j\in[n]} |X_{i,j}| \leq X_{\max}$. Collect a total of mn independent random 1-bit observations $\mathbf{Z} = \{Z_{i,j}\}_{i\in[m],j\in[n]}$ of $\mathbf{X}^*$ according to the model (1), where the density f_W associated with the modeling errors satisfies $\inf_{x\in[-X_{\max},X_{\max}]} f_W(x) > 0$. There exists a positive (finite) constant $\lambda_{\min} = \lambda_{\min}(X_{\max}, f_W)$, such that for any $\lambda > \lambda_{\min}$ and any $\delta \in (0, 1)$, the penalized maximum likelihood estimate*

$$\widehat{\mathbf{X}} = \arg\min_{\mathbf{X}\in\mathcal{X}} \left\{ -\log p_{\pi(\mathbf{X})}(\mathbf{Z}) + \lambda \cdot \text{pen}(\mathbf{X}) \right\} \tag{3}$$

satisfies the oracle error bound

$$\frac{\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_2^2}{mn} \leq \tag{4}$$
$$c \cdot \left[\min_{\mathbf{X}\in\mathcal{X}} \left\{ c' \frac{\|\mathbf{X}^* - \mathbf{X}\|_2^2}{mn} + \frac{\lambda \cdot \text{pen}(\mathbf{X})}{mn} \right\} + \frac{2\lambda \log(\frac{1}{\delta})}{mn \log 2} \right],$$

with probability at least $1 - 2\delta$. Here, $c, c' > 0$ are finite constants that depend only on the signal amplitude bound $X_{\max}$, and properties of the density f_W and distribution F_W of the error terms[2].

In the context of our sparse dictionary learning problem, we state an implication of this result as a corollary.

Corollary 3.1. *Suppose $\mathbf{X}^*$ is an $m \times n$ matrix that satisfies the uniform entry-wise amplitude bound $\max_{i,j} |X^*_{i,j}| \leq X_{\max}/2$, and which admits a factorization of the form $\mathbf{X}^* = \mathbf{D}^*\mathbf{A}^*$, where the dictionary $\mathbf{D}^*$ is $m \times p$ for $p < n$ and has elements uniformly bounded by 1 in amplitude, and the coefficient matrix $\mathbf{A}^*$ is $p \times n$ is sparse with nonzero entries uniformly bounded by some constant $A_{\max} > 0$ in amplitude.*

Consider candidate reconstructions $\mathbf{X}$ of the form $\mathbf{X} = \mathbf{DA}$, where for a sufficiently large integer $q > 2$, each element $D_{i,j}$ takes values on one of $(mn)^q$ possible uniformly discretized values in the range $[-1, 1]$, and $\mathbf{A}$ is such that each nonzero element $A_{i,j}$ takes values one of $(mn)^q$ possible uniformly discretized values in the range $[-A_{\max}, A_{\max}]$. Take $\mathcal{X}$ to be the set of all such candidate reconstructions, and let the penalty function be given by $\text{pen}(\mathbf{X}) = q \cdot mp \cdot \log(mn) + (q + 1) \cdot \|\mathbf{A}\|_0 \cdot \log(mn)$.

If observations of $\mathbf{X}^*$ are acquired via the model (1), then for $\mathcal{X}$ constructed as above (with q sufficiently large) we have that for any $\delta \in (0, 1)$ and any λ sufficiently large (exceeding a constant, that does not depend on the problem dimensions) the complexity penalized estimate (3) obtained as above satisfies

$$\frac{\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_2^2}{mn} \preceq \lambda \left(\frac{p \log(mn)}{n} + \frac{\|\mathbf{A}^*\|_0 \log(mn)}{mn} + \frac{\log(\frac{1}{\delta})}{mn} \right)$$

with probability at least $1 - 2\delta$. Here, the notation $\preceq$ suppresses leading (finite) constants, for clarity of exposition.

We provide a sketch of a proof of the corollary below, but first, it is interesting to note the implications of this result in terms of the estimability of the problem parameters. Namely, to ensure the terms in the bound above are small, it is sufficient that $n \succeq p \log(mn)$, suggesting the number of columns in the matrix should exceed (by a logarithmic factor) the number of columns in the dictionary representation; $mn \succeq \|\mathbf{A}^*\|_0 \log(mn)$, so that the total number of measurements exceed (by a logarithmic factor) the number of nonzero elements in the coefficient matrix; and, of course, that mn be large relative to $\log(\frac{1}{\delta})$. Note that if each column of $\mathbf{A}^*$ has exactly k non zeros, the condition $mn \succeq \|\mathbf{A}^*\|_0 \log(mn)$ is satisfied when $m \succeq k \log(mn)$, which is reminiscent of sample complexities in other sparse inference tasks (e.g., in compressed sensing [32, 33]).

[1] The minus sign on the $W_{i,j}$'s is merely a modeling convenience here.

[2] In particular, the assumption that $\inf_{x\in[-X_{\max},X_{\max}]} f_W(x) > 0$ ensures that $c < \infty$; see the proof for the specific form of the constants here.

Proof. (Sketch) For a candidate $\mathbf{X} = \mathbf{DA}$, we encode each element of $\mathbf{D}$ using $q \log(mn)$ bits, so a total of $q \cdot mp \cdot \log(mn)$ bits suffice to encode $\mathbf{D}$. Further, we encode each nonzero element of $\mathbf{A}$ using $\log(pn) < \log(mn)$ bits to denote its location, and $q \log(mn)$ bits for its amplitude, so matrices $\mathbf{A}$ having $\|\mathbf{A}\|_0$ nonzero entries can be described using no more than $\|\mathbf{A}\|_0(q+1)\log(mn)$ bits. The overall code for $\mathbf{X}$ is the code for $\mathbf{D}$ concatenated with the code for $\mathbf{A}$, so $\mathrm{pen}(\mathbf{X}) = q \cdot mp \cdot \log(mn) + (q+1) \cdot \|\mathbf{A}\|_0 \cdot \log(mn)$ bits suffice. Such codes are prefix codes, so satisfy (2).

Now, suppose that the true parameter is $\mathbf{X}^* = \mathbf{D}^*\mathbf{A}^*$, and consider an estimate of the form $\mathbf{X}_Q^* = \mathbf{D}_Q^*\mathbf{A}_Q^*$ whose corresponding $\mathbf{D}_Q^*$ and $\mathbf{A}_Q^*$ denote the closest quantized surrogates of the parameters $\mathbf{D}^*$ and $\mathbf{A}^*$, and such that $\|\mathbf{A}^*\|_0 = \|\mathbf{A}_Q^*\|_0$. It is easy to show that $\mathbf{X}_Q^*$ is an element of $\mathcal{X}$ when q is sufficiently large (in particular, for q sufficiently large the quantization error is sufficiently small, so that the entries of $\mathbf{X}_Q^*$ are no larger than $X_{\max}$ in amplitude). Now, evaluating the oracle bound (4) at this particular candidate estimate, it is straightforward to show that the approximation error $\|\mathbf{X}^* - \mathbf{X}_Q^*\|_F^2 = \mathcal{O}(A_{\max}^2/(mn)^{2q-3})$, which is dominated by the $\log(\frac{1}{\delta})/mn$ term when the constant q is sufficiently large. Further, $\mathrm{pen}(\mathbf{X}_Q^*) = q \cdot mp \cdot \log(mn) + (q+1) \cdot \|\mathbf{A}^*\|_0 \cdot \log(mn)$. The result follows. □

4. USEFUL LEMMATA AND PROOF OF THEOREM 3.1

We begin with a few preliminaries. Let $p(\mathbf{z})$ and $q(\mathbf{z})$ be the (joint) probability mass functions of two discrete random variables taking values in a set $\mathcal{Z}$, the elements of which may be scalar or multivariate. The Kullback-Leibler divergence (or KL divergence) of q from p is denoted $\mathrm{D}(p\|q)$ and given by

$$\mathrm{D}(p\|q) = \begin{cases} \sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{z}) \log\left(\frac{p(\mathbf{z})}{q(\mathbf{z})}\right), & \text{if } p \ll q \\ +\infty, & \text{otherwise} \end{cases},$$

where log is the natural log. The notation $p \ll q$ means that the distribution associated with $p(\mathbf{z})$ is absolutely continuous with respect to the distribution associated with $q(\mathbf{z})$; here, this condition holds if $p(\mathbf{z}) = 0$ for all $\mathbf{z}$ at which $q(\mathbf{z}) = 0$. When $p(\mathbf{z})$ and $q(\mathbf{z})$ each take the form of a product, so that $p(\mathbf{z}) = \prod_{i=1}^n p_i(z_i)$ and $q(\mathbf{z}) = \prod_{i=1}^n q_i(z_i)$, where each $p_i(z_i)$ and each $q_i(z_i)$ is the pmf of a scalar random variable Z_i taking values in a set $\mathcal{Z}_i$, the KL divergence of q from p can be expressed as a sum, as $\mathrm{D}(p\|q) = \sum_{i=1}^n \mathrm{D}(p_i\|q_i)$, where $\mathrm{D}(p_i\|q_i) = \sum_{z_i\in\mathcal{Z}_i} p(z_i)\log\left(p_i(z_i)/q_i(z_i)\right)$.

4.1. Lemmata

Our first lemma establishes conditions under which the KL divergence between two univariate Bernoulli pmf's can be bounded by quadratic functions of the difference of their parameters.

Lemma 4.1. *Let p_π and $p_{\widetilde{\pi}}$ be Bernoulli pmfs whose parameters are bounded away from 0 and 1, in the sense that there exist constants c_ℓ and c_u, such that $0 < c_\ell \le \pi, \widetilde{\pi} \le c_u < 1$. Then,*

$$2(\pi - \widetilde{\pi})^2 \le \mathrm{D}(p_{\widetilde{\pi}}\|p_\pi) \le \frac{1}{2}\max\left\{\frac{1}{c_\ell(1-c_\ell)}, \frac{1}{c_u(1-c_u)}\right\}(\pi - \widetilde{\pi})^2.$$

Proof. First, note that the condition that each of the pmfs be bounded away from 0 and 1 implies that $p_{\widetilde{\pi}} \ll p_\pi$, so that the KL divergence is finite. Now, fix $\pi \in [c_\ell, c_u]$ and let $\widetilde{\pi} = \pi + \Delta$, where $\Delta \in [c_\ell - \pi, c_u - \pi]$. With this, we have $\mathrm{D}(p_{\widetilde{\pi}}\|p_\pi) = \mathrm{D}(p_{\pi+\Delta}\|p_\pi)$; we introduce the shorthand notation $g(\Delta) \triangleq \mathrm{D}(p_{\pi+\Delta}\|p_\pi)$, leaving the dependence on π implicit. Here, we have $g(\Delta) = (\pi+\Delta)\log\left(\frac{\pi+\Delta}{\pi}\right) + (1-\pi-\Delta)\log\left(\frac{1-\pi-\Delta}{1-\pi}\right)$. Now, on the domain $\Delta \in [c_\ell - \pi, c_u - \pi]$ we have that $g_\pi(\Delta)$ is twice differentiable with respect to Δ, where

$$g'(\Delta) \triangleq \frac{d}{d\Delta}g(\Delta) = \log\left(\frac{\pi+\Delta}{\pi}\right) - \log\left(\frac{1-\pi-\Delta}{1-\pi}\right), \quad (5)$$

and

$$g''(\Delta) \triangleq \frac{d^2}{d\Delta^2}g(\Delta) = \frac{1}{(\pi+\Delta)(1-\pi-\Delta)}. \quad (6)$$

It is easy to see that the denominator of (6) satisfies $0 < \min\{c_\ell(1-c_\ell), c_u(1-c_u)\} \le (\pi+\Delta)(1-\pi-\Delta) \le 1/4$, so that overall $4 \le g''(\Delta) \le \max\left\{\frac{1}{c_\ell(1-c_\ell)}, \frac{1}{c_u(1-c_u)}\right\}$. Together, these results imply that there exist upper and lower quadratic bounds for $g_\pi(\Delta)$ of the form $g(0) + g'(0)\Delta + 2\Delta^2 \le g(\Delta) \le g(0) + g'(0)\Delta + \frac{1}{2}\max\left\{\frac{1}{c_\ell(1-c_\ell)}, \frac{1}{c_u(1-c_u)}\right\}\Delta^2$. Now, since $g(0) = 0$ (a property of KL divergence) and $g'(0) = 0$ via (5), the quadratic upper and lower bounds follow. The same analysis holds (and the same bounds result) for any other choice of $\pi \in [c_\ell, c_u]$. □

Our next lemma establishes that, under certain conditions, the variance of a Bernoulli log-likelihood ratio can be upper-bounded in terms of the KL divergence of the corresponding pmf's.

Lemma 4.2. *As in the setting of Lemma 4.1, let p_π and $p_{\widetilde{\pi}}$ be Bernoulli pmfs whose parameters are bounded away from 0 and 1, in that there exist constants c_ℓ and c_u, such that $0 < c_\ell \le \pi, \widetilde{\pi} \le c_u < 1$. For Z distributed according to $p_{\widetilde{\pi}}$ (denoted $Z \sim p_{\widetilde{\pi}}$),*

$$var_{Z\sim p_{\widetilde{\pi}}}\left(\log\frac{p_{\widetilde{\pi}}(Z)}{p_\pi(Z)}\right) \le \frac{1}{2}\max\left\{\frac{1}{c_\ell(1-c_\ell)}, \frac{1}{c_u(1-c_u)}\right\}\mathrm{D}(p_{\widetilde{\pi}}\|p_\pi).$$

Proof. Our analysis borrows some of the essential ideas from the proof of Lemma 4.1. Namely, we begin by fixing $\pi \in [c_\ell, c_u]$ and letting $\widetilde{\pi} = \pi + \Delta$, where $\Delta \in [c_\ell - \pi, c_u - \pi]$. Now, for shorthand we let $g(\Delta) = \mathrm{var}_{Z\sim p_{\widetilde{\pi}}}(L(Z))$, with $L(Z) = \left(\log\frac{p_{\widetilde{\pi}}(Z)}{p_\pi(Z)}\right)$, again leaving the dependence on π implicit to simplify the notation. By the variance formula $g(\Delta) = \mathbb{E}_{Z\sim p_{\widetilde{\pi}}}\left[L^2(Z)\right] - \left(\mathbb{E}_{Z\sim p_{\widetilde{\pi}}}\left[L(Z)\right]\right)^2$. In terms of the notation we employ here, we have $g(\Delta) = (\pi+\Delta)\log^2\left(\frac{\pi+\Delta}{\pi}\right) + (1-\pi-\Delta)\log^2\left(\frac{1-\pi-\Delta}{1-\pi}\right) - \left((\pi+\Delta)\log\left(\frac{\pi+\Delta}{\pi}\right) + (1-\pi-\Delta)\log\left(\frac{1-\pi-\Delta}{1-\pi}\right)\right)^2$. Now, it is straightforward (though somewhat tedious) to verify that $g(\Delta)$ is twice differentiable on the domain $\Delta \in [c_\ell - \pi, c_u - \pi]$, with $g'(0) = 0$, and $g''(\Delta) \le \frac{2}{\pi(1-\pi)}$. This, along with the fact that $g(0) = 0$ implies that the quadratic upper bound $g(\Delta) \le \max\left\{\frac{1}{c_\ell(1-c_\ell)}, \frac{1}{c_u(1-c_u)}\right\}\Delta^2$ holds. Now, by Lemma 4.1 we have $\Delta^2 \le (1/2)\mathrm{D}(p_{\pi+\Delta}\|p_\pi)$, so

$$g(\Delta) \le \frac{1}{2}\max\left\{\frac{1}{c_\ell(1-c_\ell)}, \frac{1}{c_u(1-c_u)}\right\}\mathrm{D}(p_{\pi+\Delta}\|p_\pi).$$

The same analysis applies for each choice of $\pi \in [c_\ell, c_u]$. □

Finally, we provide (without proof) a lemma establishing that the quadratic difference in Bernoulli parameters can be related to an ℓ_2 distance between the actual parameters we aim to estimate.

Lemma 4.3. *Let π and $\widetilde{\pi}$ be related to underlying parameters X and $\widetilde{X}$ via $\pi = F_W(X)$ and $\widetilde{\pi} = F_W(\widetilde{X})$, where $F_W(\cdot)$ is the cdf of a continuous random variable with density f_W. If $|X|, |\widetilde{X}| \leq X_{\max}$,*

$$C_\ell^2 (X - \widetilde{X})^2 \leq (\pi - \widetilde{\pi})^2 \leq C_u^2 (X - \widetilde{X})^2,$$

where the bounding constants are $C_\ell = \inf_{x \in [-X_{\max}, X_{\max}]} f_W(x)$ and $C_u = \sup_{x \in [-X_{\max}, X_{\max}]} f_W(x)$.

4.2. Proof of Main Result

For any fixed $\mathbf{X} \in \mathcal{X}$, we define the *empirical risk* of $\mathbf{X}$ in terms of its negative log-likelihood, as $\widehat{r}_{\mathbf{X}}(\mathbf{Z}) \triangleq -\log p_{\pi(\mathbf{X})}(\mathbf{Z}) = -\sum_{i \in [m], j \in [n]} \log p_{\pi(X_{i,j})}(Z_{i,j})$. We define the *excess empirical risk* associated with $\mathbf{X}$ as $\widehat{r}_{\mathbf{X},\mathbf{X}^*}(\mathbf{Z}) \triangleq \widehat{r}_{\mathbf{X}}(\mathbf{Z}) - \widehat{r}_{\mathbf{X}^*}(\mathbf{Z})$. Likewise, we define the theoretical risk of $\mathbf{X} \in \mathcal{X}$ as $r_{\mathbf{X}} \triangleq \mathbb{E}_{\mathbf{Z} \sim p_{\pi(\mathbf{X}^*)}}[\widehat{r}_{\mathbf{X}}(\mathbf{Z})]$, and the theoretical excess risk $r_{\mathbf{X},\mathbf{X}^*} \triangleq \mathbb{E}_{\mathbf{Z} \sim p_{\pi(\mathbf{X}^*)}}[\widehat{r}_{\mathbf{X},\mathbf{X}^*}(\mathbf{Z})]$. Thus, we have that $\widehat{r}_{\mathbf{X},\mathbf{X}^*}(\mathbf{Z}) - r_{\mathbf{X},\mathbf{X}^*} = -\sum_{i \in [m], j \in [n]} (U_{i,j} - \mathbb{E}[U_{i,j}])$, where $U_{i,j} \triangleq -\log\left(p_{\pi(x^*_{i,j})}(Z_{i,j}) / p_{\pi(x_{i,j})}(Z_{i,j})\right)$.

Now, we use a result obtained by Craig [34] in his proof of Bernstein's Inequality, which for our purposes here may be stated as follows: let $U_{i,j}$, $i \in [m], j \in [n]$, be independent random variables each satisfying the moment condition, that for some $h > 0$,

$$\mathbb{E}\left[|U_{i,j} - \mathbb{E}[U_{i,j}]|^k\right] \leq \frac{\mathrm{var}(U_{i,j})}{2}\, k!\, h^{k-2},$$

for $k \geq 2$. For any $\tau > 0$ and $0 \leq \epsilon h \leq c < 1$, the probability that

$$\sum_{i \in [m], j \in [n]} (U_{i,j} - \mathbb{E}\left[U_{i,j}\right]) \geq \frac{\tau}{\epsilon} + \frac{\epsilon \sum_{i \in [m], j \in [n]} \mathrm{var}\left(U_{i,j}\right)}{2(1-c)} \quad (7)$$

is no larger than $e^{-\tau}$. In order to use the result (7) here we first must verify that the $U_{i,j}$'s satisfy the moment condition. To this end, we will use the (easy to verify) fact that *bounded* random variables with $|U_{i,j} - \mathbb{E}[U_{i,j}]| \leq \beta$ satisfy the moment condition with $h = \beta/3$.

Here, the assumption $|X_{i,j}|, |X^*_{i,j}| \leq X_{\max}$ ensures that for all $i \in [m], j \in [n]$, $0 < c_\ell \leq \pi(X_{i,j}), \pi(X^*_{i,j}) \leq c_u < 1$ with $c_\ell \triangleq F_W(-X_{\max})$ and $c_u \triangleq F_W(X_{\max})$. Thus, we may define

$$\beta \triangleq \max\left\{\log\left(\frac{1}{4c_\ell(1-c_\ell)}\right), \log\left(\frac{1}{4c_u(1-c_u))}\right)\right\}$$

so that the moment condition is satisfied for the $U_{i,j}$'s here with the choice $h = \beta/3$. Next, we use Lemma 4.2 to obtain that for each $i \in [m], j \in [n]$, $\mathrm{var}(U_{i,j}) \leq \gamma \mathrm{D}(p_{\pi^*_{i,j}} \| p_{\pi_{i,j}})$, where

$$\gamma \triangleq \frac{1}{2} \max\left\{\frac{1}{c_\ell(1-c_\ell)}, \frac{1}{c_u(1-c_u)}\right\}.$$

It follows that $\sum_{i \in [m], j \in [n]} \mathrm{var}(U_{i,j}) \leq \gamma \mathrm{D}(p_{\boldsymbol{\pi}^*} \| p_{\boldsymbol{\pi}})$. Using this, along with the fact that $\mathbb{E}\left[\sum_{i \in [m], j \in [n]} U_{i,j}\right] = -\mathrm{D}(p_{\boldsymbol{\pi}^*} \| p_{\boldsymbol{\pi}})$, we have (by (7)) that the excess empirical risk satisfies

$$\Pr\left(\widehat{r}_{\mathbf{X},\mathbf{X}^*}(\mathbf{Z}) + \frac{\tau}{\epsilon} \leq \left[1 - \frac{\epsilon\,\gamma}{2(1-c)}\right] \mathrm{D}(p_{\pi(\mathbf{X}^*)} \| p_{\pi(\mathbf{X})})\right) \leq e^{-\tau}$$

for any $\tau > 0$ and $0 \leq \epsilon\beta/3 \leq c < 1$. Now, let $a = \gamma\epsilon/2(1 - \epsilon\beta/3)$ and restrict that $0 < \epsilon < 6/(3\gamma + 2\beta)$ to ensure that $a < 1$. Letting $\delta = \exp(-\tau)$ we have that for any fixed $\mathbf{X} \in \mathcal{X}$ and any $\delta \in (0,1)$,

$$\Pr\left(\widehat{r}_{\mathbf{X},\mathbf{X}^*}(\mathbf{Z}) + \frac{\log(\frac{1}{\delta})}{\epsilon} \leq (1-a)\mathrm{D}(p_{\pi(\mathbf{X}^*)} \| p_{\pi(\mathbf{X})})\right) \leq \delta.$$

If we let $\delta_{\mathbf{X}} = \delta \cdot 2^{-\mathrm{pen}(\mathbf{X})}$ and apply the union bound, we obtain that for *all* $\mathbf{X} \in \mathcal{X}$,

$$(1-a)\mathrm{D}(p_{\pi(\mathbf{X}^*)} \| p_{\pi(\mathbf{X})}) \leq \widehat{r}_{\mathbf{X},\mathbf{X}^*}(\mathbf{Z}) + \frac{\mathrm{pen}(\mathbf{X}) \log 2 + \log(\frac{1}{\delta})}{\epsilon} \quad (8)$$

with probability at least $1 - \delta$.

Recalling the definition of the excess empirical risk, we see that

$$\widehat{\mathbf{X}} \triangleq \arg\min_{X \in \mathcal{X}} \left\{-\log p_{\pi(\mathbf{X})}(\mathbf{Z}) + \frac{\mathrm{pen}(\mathbf{X}) \log 2}{\epsilon}\right\} \quad (9)$$

minimizes the upper bound of (8). This implies, in particular, that

$$(1-a)\mathrm{D}(p_{\pi(\mathbf{X}^*)} \| p_{\pi(\widehat{\mathbf{X}})}) \leq \widehat{r}_{\widehat{\mathbf{X}}^*,\mathbf{X}^*}(\mathbf{Z}) + \frac{\mathrm{pen}(\widehat{\mathbf{X}}^*) \log 2 + \log(\frac{1}{\delta})}{\epsilon} \quad (10)$$

with probability at least $1 - \delta$, where the right-hand side is evaluated at $\widehat{\mathbf{X}}^* \triangleq \min_{X \in \mathcal{X}} \left\{\mathrm{D}(p_{\pi(\mathbf{X}^*)} \| p_{\pi(\mathbf{X})}) + \frac{\mathrm{pen}(\mathbf{X}) \log 2}{\epsilon}\right\}$. We apply Bernstein's inequality once again to $\sum_{i \in [m], j \in [n]} (\widetilde{U}_{i,j} - \mathbb{E}[\widetilde{U}_{i,j}])$, where $\widetilde{U}_{i,j} = \widehat{r}_{\widehat{\mathbf{X}}^*,\mathbf{X}^*}(\mathbf{Z})$ to obtain that for any $\delta \in (0,1)$, $\widehat{r}_{\widehat{\mathbf{X}}^*,\mathbf{X}^*}(\mathbf{Z}) \leq \log(\frac{1}{\delta})/\epsilon + (1+a)\mathrm{D}(p_{\pi(\mathbf{X}^*)} \| p_{\pi(\widehat{\mathbf{X}}^*)})$ with probability at least $1 - \delta$. Combining this with (10) (via another union bound) we have the estimate (9) is such that for any $\delta \in (0,1)$,

$$\begin{aligned}(1-a)\mathrm{D}(p_{\pi(\mathbf{X}^*)} \| p_{\pi(\widehat{\mathbf{X}})}) &\leq \frac{2\log(\frac{1}{\delta})}{\epsilon} \\ &+ (1+a) \min_{X \in \mathcal{X}} \left\{\mathrm{D}(p_{\pi(\mathbf{X}^*)} \| p_{\pi(\mathbf{X})}) + \frac{\mathrm{pen}(\mathbf{X}) \log 2}{\epsilon}\right\}\end{aligned}$$

with probability at least $1 - 2\delta$. Finally, we define $\lambda \triangleq \log(2)/\epsilon$, and use Lemma 4.3 and some straightforward bounding to obtain that for any $\lambda > \log(2)(3\gamma + 2\beta)/3$, with probability at least $1 - 2\delta$,

$$\begin{aligned}\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2 \leq \left(\frac{1}{2C_\ell^2}\right) \cdot \left(1 + \frac{6\gamma}{2\beta + 3\gamma}\right) \times \\ \left[\min_{X \in \mathcal{X}} \left\{\gamma C_u^2 \|\mathbf{X}^* - \mathbf{X}\|_F^2 + \lambda \mathrm{pen}(\mathbf{X})\right\} + \frac{2\lambda \log(\frac{1}{\delta})}{\log 2}\right].\end{aligned}$$

5. CONCLUSIONS

We conclude with a few brief comments. First, while our approach was discussed here in the context of a sparse dictionary-based estimation task, our analysis may be extended to other structured data approximation tasks; thus, it follows that our framework may be applied to problems of non-negative matrix factorization, structured sparse dictionary learning, low-rank matrix approximation, etc., in settings where the observations are quantized entry-wise, as here. Further, we note that the framework developed here can also be extended to treat settings where the observations may be quantized to any number $L \geq 2$ of levels. Indeed, this modification could be accounted for here by replacing the Bernoulli distributions with analogous categorical distributions, whose parameters would implicitly depend on the thresholds chosen to specify the quantization levels (the simple form of quantization employed here utilized an implicit threshold of value 0 for each of the observations). Finally, it is interesting to note that, even though we formulated our problem in terms of a matrix approximation task, our approach here was essentially agnostic to the actual data configuration (such notions only come in when constructing $\mathcal{X}$ and the corresponding penalties). Thus, the framework proposed here may also be applied to analogous tasks of higher-order tensor approximation from 1-bit data. We defer further investigations of these extensions to future efforts.

6. REFERENCES

[1] P. Boufounos and R. Baraniuk, "1-bit compressive sensing," in *Proc. Conference on Information Sciences and Systems*, 2008.

[2] P. Boufounos, "Greedy sparse signal reconstruction from sign measurements," in *Proc. Asilomar Conf. on Signals, Systems, and Computers*, 2009.

[3] A. Gupta, R. Nowak, and B. Recht, "Sample complexity for 1-bit compressed sensing and sparse classification," in *Proc. IEEE Intl. Symposium on Information Theory*, 2010.

[4] A. Zymnis, S. Boyd, and E. Candes, "Compressed sensing with quantized measurements," *IEEE Signal Processing Letters*, vol. 17, no. 2, pp. 149–152, 2010.

[5] J. Haupt and R. Baraniuk, "Robust support recovery using sparse compressive sensing matrices," in *Proc. Conference on Information Sciences and Systems*, 2011.

[6] Y. Plan and R. Vershynin, "Robust 1-bit compressive sensing and sparse logistic regression: A convex programming approach," *IEEE Trans. Information Theory*, vol. 59, no. 1, pp. 482–494, 2013.

[7] Y. Plan and R. Vershynin, "One-bit compressed sensing by linear programming," *Communications on Pure and Applied Mathematics*, 2013.

[8] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters, "1-bit matrix completion," *Submitted*, 2012, online at: `arxiv.org/abs/1209.3672`.

[9] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vision Research*, vol. 37, pp. 3311–3325, 1997.

[10] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Proc.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[11] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. ICML*, 2009.

[12] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.*, vol. 51, no. 1, pp. 34–81, 2009.

[13] R. Jenatton, J. Mairal, F. R. Bach, and G. R. Obozinski, "Proximal methods for sparse hierarchical dictionary learning," in *Proc International Conference on Machine Learning*, 2010, pp. 487–494.

[14] K. Schnass, "On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD," *Submitted*, 2013, online at: `arxiv.org/abs/1301.3375`.

[15] R. Gribonval and K. Schnass, "Dictionary identification - sparse matrix-factorization via ℓ_1 minimization," *IEEE Trans. Information Theory*, vol. 56, no. 7, pp. 3523–3539, 2010.

[16] Q. Geng, H. Wang, and J. Wright, "On the local correctness of ℓ^1 minimization for dictionary learning," *Submitted*, 2011, online at: `arxiv.org/abs/1101.5672`.

[17] R. Jenatton, R. Gribonval, and F. Bach, "Local stability and robustness of sparse dictionary learning in the presence of noise," *Submitted*, 2012, online at: `arxiv.org/abs/1210.0685`.

[18] A. R. Barron, "Complexity regularization with application to artificial neural networks," in *Nonparametric functional estimation and related topics*, pp. 561–576. Springer, 1991.

[19] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Information Theory*, vol. 37, no. 4, pp. 1034–1054, 1991.

[20] A. Barron, L. Birgé, and P. Massart, "Risk bounds for model selection via penalization," *Probability theory and related fields*, vol. 113, no. 3, pp. 301–413, 1999.

[21] J. Q. Li and A. R. Barron, "Mixture density estimation," in *Advances in Neural Information Processing Systems*, 1999.

[22] E. D. Kolaczyk and R. D. Nowak, "Multiscale likelihood analysis and complexity penalized estimation," *Annals of Statistics*, pp. 500–527, 2004.

[23] T. Zhang, "On the convergence of MDL density estimation," in *Learning Theory*, pp. 315–330. Springer, 2004.

[24] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *IEEE Trans. Inform. Theory*, vol. 52, no. 9, pp. 4036–4048, Sept. 2006.

[25] P. D. Grünwald, *The minimum description length principle*, MIT press, 2007.

[26] I. Ramírez and G. Sapiro, "An MDL framework for sparse coding and dictionary learning," *IEEE Trans. Signal Processing*, vol. 60, no. 6, pp. 2913–2927, 2012.

[27] Z.-Q. Luo, "Universal decentralized estimation in a bandwidth constrained sensor network," *IEEE Trans. Information Theory*, vol. 51, no. 6, pp. 2210–2219, 2005.

[28] Z.-Q. Luo, "An isotropic universal decentralized estimation scheme for a bandwidth constrained ad hoc sensor network," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 4, pp. 735–744, 2005.

[29] Z.-Q. Luo and J.-J. Xiao, "Decentralized estimation in an in homogenous sensing environment," *IEEE Trans. Information Theory*, vol. 51, no. 10, pp. 3564–3575, 2005.

[30] A. Ribeiro and G. Giannakis, "Bandwidth-constrained distributed estimation for wireless sensor networks – part I: Gaussian case," *IEEE Trans. Signal Processing*, vol. 54, no. 3, pp. 1131–1143, 2006.

[31] A. Ribeiro and G. Giannakis, "Bandwidth-constrained distributed estimation for wireless sensor networks – part II: Uknown pdf," *IEEE Trans. Signal Processing*, vol. 54, no. 7, pp. 2784–2796, 2006.

[32] D. Donoho, "Compressed sensing," *IEEE Trans. Information Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[33] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[34] C. Craig, "On the Tchebychef inequality of Bernstein," *Ann. Math. Statist.*, vol. 4, pp. 94–102, 1933.