Error bounds for Bregman Denoising and Structured Natural Parameter Estimation

Amin Jalali
Wisconsin Institute for Discovery
amin.jalali@wisc.edu

James Saunderson
Monash University ECSE
james.saunderson@monash.edu

Maryam Fazel
University of Washington EE
mfazel@uw.edu

Babak Hassibi
Caltech EE
hassibi@caltech.edu

***Abstract*—We analyze an estimator based on the Bregman divergence for recovery of structured models from additive noise. The estimator can be seen as a regularized maximum likelihood estimator for an exponential family where the natural parameter is assumed to be structured. For all such Bregman denoising estimators, we provide an error bound for a natural associated error measure. Our error bound makes it possible to analyze a wide range of estimators, such as those in proximal denoising and inverse covariance matrix estimation, in a unified manner. In the case of proximal denoising, we exactly recover the existing tight normalized mean squared error bounds. In sparse precision matrix estimation, our bounds provide optimal scaling with interpretable constants in terms of the associated error measure.**

I. Introduction

Denoising is the pursuit of removing noise from an observed signal. Physical properties of a measurement mechanism may result in different distributional properties for measurement noise. For instance, in optical devices based on photon counting, the measurement noise is usually Poisson (shot noise). In statistical model selection, one can think of 'noise' as coming from finite sample approximations of population statistics [1]. Many such noise distributions can be treated in a uniform way using the formalism of *exponential families*.

On the other hand, having side information about the underlying signal usually allows for improved estimation. Different techniques have been developed for making use of such side information, or structure. In this work, we focus on estimation through optimization with penalty functions whose minimization tend to favor structured signals. Examples of such approaches are now prevalent in the literature [1], [2].

A. Proximal Denoising

The best understood case of *structured signal denoising* is when the noise has independent normally distributed entries with mean zero and known variance, namely $y = x_0 + \sigma z$ where $z \sim \mathcal{N}(0, I)$. In this case, the regularized maximum likelihood estimator is the proximal mapping given as

$$\widehat{x}(y) = \operatorname*{argmin}_{x} \ \frac{1}{2}\|x - y\|_2^2 + \sigma f(x) \tag{1}$$

where f is a convex, structure-inducing function. Authors in [3] studied this estimator for general norms f and provided upper bounds on the normalized mean squared error (NMSE) as

$$\frac{1}{\sigma^2}\mathbb{E}_z \|\widehat{x}(x_0 + \sigma z) - x_0\|_2^2 \leq \ \mathbb{E}_z \operatorname{dist}^2_{\ell_2}(z, \partial f(x_0)) \tag{2}$$

where $\partial f(x_0)$ denotes the subdifferential of f at x_0 (see Definition 5 in Section II) and

$$\operatorname{dist}_h(z, A) = \inf_{a \in A} \ h(z - a)\,.$$

By making use of Theorem 4.3 of [4] and other arguments, they showed that the above upper bound is achieved as $\sigma \to 0$.

The right hand side in (2) is nicely interpretable in the Gaussian case $z \sim \mathcal{N}(0, I)$. Namely, in common structured learning problems such as those with sparse vectors, low-rank matrices or row-sparse matrices, if the respective structure-inducing norm f is appropriately scaled [3], then the right hand side of (2) is scaling with the number of *degrees of freedom* in the model [4]. This quantity also arises as the required number of samples in time-data tradeoffs in denoising [1].

B. Bregman Denoising

The proximal mapping can be generalized by replacing the squared Euclidean distance with a more general divergence function; e.g., [5]–[8]. In this work, we consider *Bregman divergences* [9] (see Section II-A for background) and consider the following convex optimization program for denoising

$$\widehat{\theta}(\bar{x}) = P(\bar{x}; \Psi, f) \coloneqq \operatorname*{argmin}_{\theta} \ \mathtt{D}_\Psi(\theta, \nabla\phi(\bar{x})) + f(\theta) \tag{3}$$

where, $\mathtt{D}_\Psi$ is the Bregman divergence associated with Ψ, $f(\cdot)$ is a convex function which encodes the prior information on θ, and ϕ is the convex conjugate of Ψ. From now on, we will refer to the above formulation as the *Bregman proximal denoiser*, *Bregman denoiser* for short. It allows for measurements (denoted by $\bar{x} = x_0 + z$) to be taken in one domain and structure to be imposed in another (specifically on $\theta_0 \coloneqq \nabla\phi(x_0)$). We will discuss different interpretations of this estimation method in Section III. Most importantly, Bregman denoising can be seen as f-regularized maximum likelihood (ML) estimation when $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $x_1, \ldots, x_n$ are i.i.d. samples from the corresponding canonical exponential family (see Section II-B for background). Such regularized ML estimation has been previously studied for different examples of exponential families; e.g., see [10]–[14]. We discuss our contribution in the next section.

C. Our Contribution

In this work, we show that a natural error measure for Bregman denoising can be bounded by the same structure-dependent quantity that controls the mean-squared error in proximal denoising. The error measure is defined by the *symmetrized Bregman divergence* and is closely related to the *Fisher risk* [10] (see also Definition 4 in Section II-B). For example, see [15] for similar measures.

Proposition 1 (Upper Bound on Error): For the estimator in (3), and for $\widehat{\theta} = P(x_0 + z; \Psi, f)$ and $\theta_0 = \nabla\phi(x_0)$,

$$\frac{\mathtt{D}_\Psi(\widehat{\theta}, \theta_0) + \mathtt{D}_\Psi(\theta_0, \widehat{\theta})}{\|\widehat{\theta} - \theta_0\|_2} \leq \operatorname{dist}_{\ell_2}(z, \partial f(\theta_0)). \tag{4}$$

The proof of Proposition 1 is given in Section V. At this stage, we briefly comment on the statement of Proposition 1. Observe that the left-hand side of (4) is always nonnegative and symmetric, and is equal to zero only if $\widehat{\theta}$ and θ_0 are equal. Moreover, it reduces to the Euclidean distance when squared Euclidean norm is used for Ψ and, after squaring both sides, recovers NMSE and the bound in (2) exactly.

When considering the expected value of error over a noise distribution, the numerator in LHS of (4) can be thought of as the Fisher risk plus cubic and higher order terms in $\widehat{\theta} - \theta_0$:

$$\mathtt{D}_\Psi(\widehat{\theta}, \theta_0) + \mathtt{D}_\Psi(\theta_0, \widehat{\theta}) = \|\widehat{\theta} - \theta_0\|^2_{\mathcal{F}_0} + O(\|\widehat{\theta} - \theta_0\|^3) \tag{5}$$

where $\mathcal{F}_0$ is the Fisher information matrix at θ_0; see (10). Using $\|\cdot\|_{\mathcal{F}_0}$ and $\|\cdot\|_{\mathcal{F}_0^{-1}}$ as a pair of dual norms instead of the Euclidean norms in (4), gives the following more intrinsic alternative.

Corollary 2: For the estimator in (3), and for $\theta_0 = \nabla\phi(x_0)$ and $\widehat{\theta} = P(x_0 + z; \Psi, f)$, we have

$$\begin{aligned}\frac{\mathtt{D}_\Psi(\widehat{\theta}, \theta_0) + \mathtt{D}_\Psi(\theta_0, \widehat{\theta})}{\|\widehat{\theta} - \theta_0\|_{\mathcal{F}_0}} &\leq \operatorname{dist}_{\mathcal{F}_0^{-1}}(z, \partial f(\theta_0)) \qquad (6)\\ &= \inf_{g \in \partial f(\theta_0)} \sqrt{(z-g)\mathcal{F}_0^{-1}(z-g)}\end{aligned}$$

where $\mathcal{F}_0 = \mathcal{F}_{\theta_0}$ is the Fisher information matrix at θ_0. Combining (5) and (6) we get the following bound

$$\|\widehat{\theta} - \theta_0\|_{\mathcal{F}_0} \leq \operatorname{dist}_{\mathcal{F}_0^{-1}}(z, \partial f(\theta_0)) + O(\|\widehat{\theta} - \theta_0\|^2)$$

for the distance between $\widehat{\theta}$ and θ_0 in Fisher norm (at θ_0); see [10] for related discussions.

II. Background

A. Bregman Divergence

Definition 3 (Bregman divergence [16]): Let $\Psi : \Theta \to \mathbb{R}$ be a strictly convex function defined on a convex set $\Theta \subseteq \mathbb{R}^p$ such that Ψ is differentiable on the relative interior of Θ, denoted by $\mathrm{ri}\Theta$ and assumed to be nonempty. The *Bregman divergence* $\mathtt{D}_\Psi : \Theta \times \mathrm{ri}\Theta \mapsto [0, \infty)$ is defined as

$$\mathtt{D}_\Psi(\theta_1, \theta_2) := \Psi(\theta_1) - \Psi(\theta_2) - \langle \nabla\Psi(\theta_2), \theta_1 - \theta_2 \rangle \tag{7}$$

where $\nabla\Psi(\theta)$ is the gradient vector of Ψ evaluated at θ. The squared Euclidean distance, Mahalanobis distance, KL divergence, and logistic loss are all Bregman divergences.

Bregman divergences are nonnegative, convex in their first argument, and linear in the underlying convex function (Ψ in the above). Moreover, they enjoy a useful duality relationship. Denote by $\phi(\mu) := \sup_{\theta\in\Theta} \langle \mu, \theta \rangle - \Psi(\theta)$ the convex conjugate of Ψ. Then, $\nabla\Psi$ and $\nabla\phi$ are inverse maps, and the following duality relationship holds:

$$\mathtt{D}_\phi(\mu_1, \mu_2) = \mathtt{D}_\Psi(\nabla\phi(\mu_2), \nabla\phi(\mu_1)). \tag{8}$$

B. Exponential Families

In this section we provide a brief review of exponential families. We will mainly follow the presentation in [16, Section 4.1]. Consider a measure space $(\Omega, \mathcal{B}, P_0)$, and a measurable mapping $t(\omega)$ from Ω to $\mathbb{R}^p$ with $\mathrm{d}P_0(\omega) = p_0(t(\omega))\mathrm{d}t(\omega)$, where $t(\omega)$ does not satisfy any linear constraints with probability 1 (for the representation to be *minimal*). Let Θ (natural parameter space) be the (convex) set of $\theta \in \mathbb{R}^p$ such that $\int_{\omega\in\Omega} \exp(\langle \theta, t(\omega) \rangle)\mathrm{d}P_0(\omega) < \infty$ and define the log-partition function $\Psi : \Theta \to \mathbb{R}$ by $\Psi(\theta) = \log\left(\int_{\omega\in\Omega} \exp(\langle \theta, t(\omega)\rangle)\,\mathrm{d}P_0(\omega)\right)$. A family of probability distributions parametrized by $\theta \in \Theta$, such that the probability density functions with respect to the measure $\mathrm{d}t(\omega)$ can be written as

$$p(\omega; \theta) = \exp(\langle \theta, t(\omega) \rangle - \Psi(\theta))p_0(t(\omega)),$$

is called an *exponential family* with *natural statistic* $t(\omega)$, natural parameter θ, and natural parameter space Θ. We assume Θ is open, hence the exponential family is *regular*.

By the change of variables $x = t(\omega)$, a regular exponential family in its canonical form can be equivalently expressed as

$$p_{(\Psi,\theta)}(x) = \exp\left(\langle x, \theta \rangle - \Psi(\theta)\right) p_0(x) \tag{9}$$

where x is a minimal sufficient statistic for the family. For every regular exponential family, there exists a unique Bregman divergence associated to it such that (9) can be expressed via

$$\begin{aligned}\langle x, \theta \rangle - \Psi(\theta) &= (\langle \nabla\Psi(\theta), \theta \rangle - \Psi(\theta)) + \langle x - \nabla\Psi(\theta), \theta \rangle\\ &= \phi(\nabla\Psi(\theta)) + \langle x - \nabla\Psi(\theta), \theta \rangle\\ &= -\mathtt{D}_\phi(x, \nabla\Psi(\theta)) + \phi(x)\\ &= -\mathtt{D}_\Psi(\theta, \nabla\phi(x)) + \phi(x)\end{aligned}$$

where ϕ is the convex conjugate of Ψ and we used (8).

The *Fisher information* associated with an exponential family is defined as the following positive semidefinite matrix,

$$\mathcal{F}_\theta := -\mathbb{E}_\theta \nabla^2 \log p_{(\Psi,\theta)}(x)\,.$$

In a canonical exponential family model, $\nabla^2 \log p_{(\Psi,\theta)}(x) = -\nabla^2\Psi(\theta)$ is a constant, which gives

$$\mathcal{F}_\theta = \nabla^2\Psi(\theta). \tag{10}$$

Definition 4: The induced *Fisher risk* at $\theta_0 \in \mathrm{ri}\Theta$ is defined as $\|\theta - \theta_0\|^2_{\mathcal{F}_0} = (\theta - \theta_0)^T\mathcal{F}_0(\theta - \theta_0)$ where $\mathcal{F}_0 = \mathcal{F}_{\theta_0}$. This is a norm because Ψ is strictly convex, so its Hessian is invertible on the relative interior of its domain. The dual to the norm $\|\theta - \theta_0\|_{\mathcal{F}_0}$ is then given by $\|\theta - \theta_0\|_{\mathcal{F}_0^{-1}}$.

C. Regularization for Structured Learning

Definition 5 (subdifferential): For any proper, convex function f and any $x \in \mathrm{dom} f$, the subdifferential of f at x is

$$\partial f(\theta) = \{g : \ f(\theta') \geq f(\theta) + \langle g, \theta' - \theta \rangle \ \ \forall \theta'\} .$$

The expected squared distance to the subdifferential for Gaussian noise, i.e. $\mathbb{E}_z \mathrm{dist}^2(z, \partial f(x_0))$ where $z \sim \mathcal{N}(0, I)$, comes up frequently in the structured learning literature. It is tightly connected to the notions of *statistical dimension* [4] and *Gaussian width* [17]. For example, the statistical dimension of the tangent cone to f at x is equal to $\mathbb{E}_z \mathrm{dist}^2(z, \mathrm{cone}(\partial f(x_0)))$, where cone is the operation of taking the conic hull. This quantity, in the case of Gaussian noise, is in turn close to $\mathbb{E}_z \mathrm{dist}^2(z, \lambda \partial f(x_0))$ for an appropriate λ [4]. For example, in the case of $f(\cdot) = \lambda \|\cdot\|_1$, for $\lambda \geq \sqrt{2 \log p} \geq \mathbb{E}_z[\|z\|_\infty]$, the statistical dimension is the number of nonzeros in x.

III. Interpretations of Bregman Denoiser

A. Regularized Maximum Likelihood Estimator

Using the connection between exponential families and Bregman divergences (see Section II-B), the optimization problem in (3) can be seen as a f-regularized ML estimator for the natural parameter of an exponential family given i.i.d. samples. More specifically, given n samples, $x_1, x_2, \ldots, x_n$, drawn from the distribution $p_{(\Psi, \theta_0)}$, and defining $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, the ML estimate, $\widehat{\theta}^{\mathrm{ML}}$, for θ_0 is

$$\begin{aligned} \underset{\theta}{\operatorname{argmax}} \ \frac{1}{n} \sum_{i=1}^n \log \ell(x_i \mid \theta) &= \underset{\theta}{\operatorname{argmax}} \ \frac{1}{n} \sum_{i=1}^n \langle x_i, \theta \rangle - \Psi(\theta) \\ &= \underset{\theta}{\operatorname{argmin}} \ \mathtt{D}_\phi(\bar{x}, \nabla \Psi(\theta)) \\ &= \underset{\theta}{\operatorname{argmin}} \ \mathtt{D}_\Psi(\theta, \nabla \phi(\bar{x})) \quad (11) \end{aligned}$$

where the last equality holds by (8). Notice that (11) is a convex optimization problem. Regularizing (11) with a structure-inducing function f gives the Bregman denoiser.

Our main results (stated in Section I-C) bound the error between $\widehat{\theta}^{\mathrm{ML}}$ and θ_0, in terms the deviation between the true sufficient statistics $\nabla \Psi(\theta_0)$ and $\bar{x}$.

B. Regularized Loss Minimization

While (3) is *designed* for denoising a model contaminated with noise drawn from the corresponding exponential family (as discussed in Section III-A), *the performance guarantee of Proposition 1, in* (4)*, is valid for any* z. Therefore, (3) can be seen as a *regularized loss minimization* estimator for denoising a structured model with arbitrary noise. While we lose the maximum likelihood interpretation, the bounds still hold and are useful for understanding the performance of such an estimator. A similar approach has been proposed in [10].

C. Mean Estimation with Composite Regularization

The Bregman denoiser in (3) can be equivalently stated as

$$\widehat{x} = \underset{x}{\operatorname{argmin}} \ \mathtt{D}_\phi(\bar{x}, x) + f(\nabla \phi(x)) \qquad (12)$$

and can be thought of as a different (possibly nonconvex) regularized loss minimization, where $f(\theta) = f(\nabla \phi(x))$ is a *composite penalty* [18] and imposes the structure encoded by f to a transformation of x.

For example, when Ψ is a strictly convex quadratic function, namely $\Psi(\theta) = \frac{1}{2} \theta^T A \theta$ for a positive definite matrix A, (3) and (12) can be equivalently expressed as

$$\widehat{\theta} = \mathrm{argmin}_\theta \ \tfrac{1}{2}(\theta - \bar{\theta})^T A (\theta - \bar{\theta}) + f(\theta) \quad \text{or} \qquad (13)$$

$$\widehat{x} = \mathrm{argmin}_x \ \tfrac{1}{2}(x - \bar{x})^T A^{-1} (x - \bar{x}) + f(A^{-1} x) \qquad (14)$$

through $\bar{\theta} = A^{-1} \bar{x}$ and $\widehat{\theta} = A^{-1} \widehat{x}$. As an instance, one might think of the analysis-based problem in the compressive sensing literature [19] this way. While both estimators in the above are convex (in θ and x respectively), in general, (12) is not convex. However, in some cases, (see, e.g., [20]) we may still retain convexity over parts of the space. This may be desirable if we want to impose additional convex structural constraints on x.

IV. Example: Sparse Precision Matrix Estimation

In this section, we illustrate our main result in the case of ℓ_1-regularized ML estimator for sparse inverse covariance (precision) matrix estimation in Gaussian graphical models; e.g., see [2] for statement of the problem. Our intention is to show that this problem can be analyzed in a straightforward way using our general error bound for Bregman denoising.

Let $y_1, \ldots, y_n \sim \mathcal{N}(0, \Sigma_0)$ be i.i.d. samples from a p-dimensional Gaussian distribution. Since the population mean is assumed to be zero, the sample sufficient statistic is given by $S_n := \frac{1}{n} \sum_{i=1}^n y_i y_i^T$ and is distributed as a scaled Wishart matrix with n degrees of freedom and scale matrix Σ_0, namely $n S_n \sim \mathcal{W}(n, \Sigma_0)$. Since the distribution $\mathcal{W}(n, \Sigma_0)$ has mean $n \Sigma_0$, we have that $S_n = \Sigma_0 + \frac{1}{n} Z$ where $Z \sim \overline{\mathcal{W}}(n, \Sigma_0)$ has a *centered* Wishart distribution.

The ℓ_1-regularized ML estimator for the precision J is

$$\widehat{J}_n = \mathrm{argmax}_J \ \log\det(J) - \langle J, S_n \rangle - \tfrac{\lambda}{\sqrt{n}} \|J\|_1, \qquad (15)$$

which is exactly the Bregman denoiser in (3) for $f(\cdot) = \frac{\lambda}{\sqrt{n}} \|\cdot\|_1$ and $\Psi(\cdot) = -\log\det(\cdot)$. Applying Proposition 1 gives the following result. The proof is in Appendix A.

Proposition 6: The estimator in (15) satisfies

$$\frac{\langle \widehat{J}_n - \Sigma_0^{-1}, -\widehat{J}_n^{-1} + \Sigma_0 \rangle}{\|\widehat{J}_n - \Sigma_0^{-1}\|_F} \leq \frac{1}{n} \mathrm{dist}(Z, \lambda \sqrt{n} \partial \|\Sigma_0^{-1}\|_1) . \quad (16)$$

Denote by k the number of nonzero entries of Σ_0^{-1}. If $p \geq 9$, $C > 0$, $\lambda \geq 20 \|\Sigma_0\|_\infty \sqrt{(2C + 4) \log p}$, and $n > (2C + 4) \log p$, then with probability at least $1 - 2p^{-C}$,

$$\frac{\langle \widehat{J}_n - \Sigma_0^{-1}, -\widehat{J}_n^{-1} + \Sigma_0 \rangle}{\|\widehat{J}_n - \Sigma_0^{-1}\|_F} \leq 2\lambda \sqrt{\frac{k}{n}} . \qquad (17)$$

For $\Delta := \widehat{J}_n - \Sigma_0^{-1}$, observe that

$$\begin{aligned} \text{LHS of (17)} &= \langle \Delta, \Sigma_0 \Delta \widehat{J}_n^{-1} \rangle / \|\Delta\|_F \\ &= \langle \mathrm{vec}(\Delta), (\widehat{J}_n^{-1} \otimes \Sigma_0) \mathrm{vec}(\Delta) \rangle / \|\Delta\|_F \\ &\geq \frac{\sigma_{\min}(\Sigma_0)}{\sigma_{\max}(\widehat{J}_n)} \|\widehat{J}_n - \Sigma_0^{-1}\|_F . \end{aligned}$$

V. Proof of Main Result

Before analyzing the estimator in (3), and presenting the proof of Proposition 1, let us discuss three main ingredients.

a) Operator notation: After expanding the definition of Bregman divergence above, and by optimality, we get $0 \in \nabla\Psi(\widehat{\theta}) - \nabla\Psi(\nabla\phi(\bar{x})) + \partial f(\widehat{\theta})$ which implies

$$\widehat{\theta} = (\partial f + \nabla\Psi)^{-1}(\bar{x})\,. \tag{18}$$

For example, see Proposition 3.22 of [21].

*b) Exactly Denoisable Perturbations (*EDP*):* Here we define the set of perturbations to a given point x_0 which will be mapped back to the corresponding natural parameter for x_0 by $P(\cdot;\Psi,f)$. Define,

$$\begin{aligned}
&\mathrm{EDP}(x_0;\Psi,f) \coloneqq \{w:\ P(x_0+w) = \nabla\phi(x_0)\}\\
&= \{w:\ (\partial f + \nabla\Psi)^{-1}(x_0+w) = (\nabla\Psi)^{-1}(x_0)\}\\
&= \{w:\ \exists u \text{ s.t. } x_0 + w \in \partial f(u) + \nabla\Psi(u)\ ,\ x_0 = \nabla\Psi(u)\}\\
&= \partial f((\nabla\Psi)^{-1}(x_0)) = \partial f(\nabla\phi(x_0))
\end{aligned}$$

where the last equality can be established by showing that each side is included in the other.

c) $(\nabla\Psi)$-Firm Nonexpansiveness: For $\theta_1 = P(y_1)$ and $\theta_2 = P(y_2)$ we have

$$\langle y_1 - y_2, \theta_1 - \theta_2\rangle \geq \langle \nabla\Psi(\theta_1) - \nabla\Psi(\theta_2), \theta_1 - \theta_2\rangle\,. \tag{19}$$

The definition and proof can be found in Definition 3.4 and Proposition 3.22 of [21], where we additionally use the fact that $\nabla\Psi(\nabla\phi(y)) = y$.

We are now ready to prove Proposition 1 and provide an upper bound on the error of the estimator in (3).

Proof. [of Proposition 1] Consider any $w \in \mathrm{EDP}(x_0)$ and

$$\theta_0 \coloneqq \nabla\phi(x_0) = P(x_0 + w)\ , \qquad \widehat{\theta} = P(x_0 + z).$$

Then, from the $(\nabla\Psi)$-firm nonexpansiveness property with $y_1 = x_0 + z$, $y_2 = x_0 + w$, $\theta_1 = \widehat{\theta}$ and $\theta_2 = \theta_0$, we have

$$\langle z - w, \widehat{\theta} - \theta_0\rangle \geq \langle \nabla\Psi(\widehat{\theta}) - \nabla\Psi(\theta_0), \widehat{\theta} - \theta_0\rangle. \tag{20}$$

Using the Cauchy-Schwarz inequality, and taking the infimum over all possible $w \in \mathrm{EDP}(x_0)$, we get

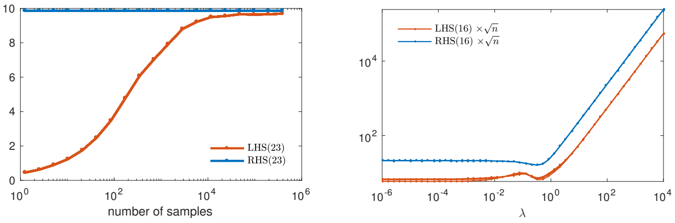
$$\frac{\langle \nabla\Psi(\widehat{\theta}) - \nabla\Psi(\theta_0), \widehat{\theta} - \theta_0\rangle}{\|\widehat{\theta} - \theta_0\|_2} \leq \mathrm{dist}(z, \partial f(\theta_0)).$$

The inner product in the numerator can be written as

$$\langle \nabla\Psi(\widehat{\theta}) - \nabla\Psi(\theta_0), \widehat{\theta} - \theta_0\rangle = \mathtt{D}_\Psi(\widehat{\theta}, \theta_0) + \mathtt{D}_\Psi(\theta_0, \widehat{\theta})\,,$$

giving the claimed bound. ■

The two terms in the inner product on the LHS of (20) can be separated in many different ways, each leading to a different final bound. We choose to use the Cauchy-Schwarz inequality here. Moreover, one can use the four-point property of Bregman divergence or the Fenchel inequality (for pairs of conjugate functions) to express the inner product on the LHS of (20), leading to different bounds.



(a) (b)

Fig. 1. Simulation results for (a) sparse mean estimation, and for (b) sparse precision matrix estimation in Section VI.

Corollary 7: For the estimator in (3), and for $\theta_0 = \nabla\phi(x_0)$ and $\widehat{\theta} = P(x_0 + z;\Psi,f)$, and dual norms $h(\cdot)$ and $h^d(\cdot)$,

$$\frac{\mathtt{D}_\Psi(\widehat{\theta}, \theta_0) + \mathtt{D}_\Psi(\theta_0, \widehat{\theta})}{h(\widehat{\theta} - \theta_0)} \leq \mathrm{dist}_{h^d}(z, \partial f(\theta_0))\,. \tag{21}$$

We briefly discuss two interesting choices for norm h. One is to use f itself. Then, as every subgradient have dual norm at most equal to 1, we can use the triangle inequality to get

$$\mathtt{D}_\Psi(\widehat{\theta}, \theta_0) + \mathtt{D}_\Psi(\theta_0, \widehat{\theta}) \leq (1 + f^d(z))f(\widehat{\theta} - \theta_0)\,. \tag{22}$$

The other choice is to use the norm associated to the Fisher risk which provides us with the bound in Corollary 2.

VI. Numerical Experiments

We present two numerical experiments. First, we consider structured mean estimation for a family of multivariate normal distributions with a given covariance matrix Σ. Given $x_1, \ldots, x_n \sim \mathcal{N}(x_0, \Sigma)$, we would like to estimate $x_0 \in \mathbb{R}^p$ using the prior information that $\theta_0 = \Sigma^{-1}x_0$ is sparse. The corresponding regularized log-likelihood estimator, where $\bar{\theta} \coloneqq \Sigma^{-1}\bar{x}$ and $\bar{x} \coloneqq \frac{1}{n}\sum_{i=1}^n x_i$ is given as $\widehat{\theta}_n = \mathrm{argmin}_\theta\ \frac{1}{2}(\theta - \bar{\theta})^T\Sigma(\theta - \bar{\theta}) + \frac{\lambda}{\sqrt{n}}\|\theta\|_1$, and corresponds to a Bregman denoiser with $\Psi(\theta) = \frac{1}{2}\theta^T\Sigma\theta$ and $f(\theta) = \frac{\lambda}{\sqrt{n}}\|\theta\|_1$. Note that $\bar{x} \sim \mathcal{N}(x_0, \frac{1}{n}\Sigma)$ and can be expressed as $\bar{x} = x_0 + \frac{1}{\sqrt{n}}z$ where $z \sim \mathcal{N}(x_0, \Sigma)$. Using our bound in Corollary 2 we get

$$(\widehat{\theta}_n - \theta_0)^T\Sigma(\widehat{\theta}_n - \theta_0) \leq \inf_{g \in \partial\|\theta_0\|_1} (z - g)^T\Sigma^{-1}(z - g). \tag{23}$$

Figure 1a plots the left and right hand side of (23), averaged over 100 instances of z, for different values of n. In our experiment, $p = 100$, $\mathrm{card}(x_0) = 10$, and $\lambda = \sqrt{2\log p}\|\Sigma\|_{\mathrm{op}}$.

For the second experiment, we consider sparse precision matrix estimation described in Section IV, where $p = 50$ and Σ_0^{-1} has 590 nonzero entries corresponding to the edges of a graph constructed as in Example 1 of [22]. We solve (15) using QUIC software [23], for 100 random trials. Figure 1b shows the left and right hand sides of (16) (scaled by $\sqrt{n}$) for different numbers of samples $n = 20, 30, 40$ (overlayed) and different values of λ.

The experiment suggests that the proposed error measure and our upper bound closely track each other across a wide range of λ. Note that, as $\lambda \to 0$, the error measure (LHS) remains bounded. One can see this from the experiment, or from (16) noting that the RHS approaches $\frac{1}{n}\|Z\|_F$ in the limit.

Acknowledgment

This work was supported in part by the NSF grant CCF-1409836 and ONR grant N00014-16-1-2789.

References

[1] V. Chandrasekaran and M. I. Jordan, "Computational and statistical tradeoffs via convex relaxation," *Proc. Natl. Acad. Sci. USA*, vol. 110, no. 13, pp. E1181–E1190, 2013.

[2] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[3] S. Oymak and B. Hassibi, "Sharp MSE bounds for proximal denoising," *Found. Comput. Math.*, pp. 1–65, 2013.

[4] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: Phase transitions in convex programs with random data," *Information and Inference*, pp. 224–294, 2014.

[5] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.

[6] Y. Censor and S. A. Zenios, "Proximal minimization algorithm with D-functions," *J. Optim. Theory Appl.*, vol. 73, no. 3, pp. 451–464, 1992.

[7] M. Collins, S. Dasgupta, and R. E. Schapire, "A generalization of principal component analysis to the exponential family," 2001.

[8] M. Basseville, "Divergence measures for statistical data processing—an annotated bibliography," *Signal Processing*, vol. 93, no. 4, pp. 621–633, 2013.

[9] L. M. Brègman, "A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming," *Ž. Vyčisl. Mat. i Mat. Fiz.*, vol. 7, pp. 620–631, 1967.

[10] S. M. Kakade, O. Shamir, K. Sridharan, and A. Tewari, "Learning exponential families in high-dimensions: Strong convexity and sparsity," *CoRR*, 2010.

[11] A. Agarwal, S. Negahban, and M. J. Wainwright, "Fast global convergence rates of gradient methods for high-dimensional statistical recovery," in *NIPS 23*, 2010, pp. 37–45.

[12] C. Zhang, Y. Jiang, and Y. Chai, "Penalized bregman divergence for large-dimensional regression and classification," *Biometrika*, 2010.

[13] C. Zhang, X. Guo, and Y. Chai, "Screening-based bregman divergence estimation with np-dimensionality," *Electronic Journal of Statistics*, vol. 10, no. 2, pp. 2039–2065, 2016.

[14] S. A. Van de Geer, "High-dimensional generalized linear models and the lasso," *The Annals of Statistics*, pp. 614–645, 2008.

[15] L. Wu, R. Jin, S. C. Hoi, J. Zhu, and N. Yu, "Learning bregman distance functions and its application for semi-supervised clustering," in *NIPS 22*, 2009, pp. 2089–2097.

[16] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *J. Mach. Learn. Res.*, vol. 6, pp. 1705–1749, 2005.

[17] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, no. 6, pp. 805–849, 2012.

[18] A. Jalali, "Convex optimization algorithms and statistical bounds for learning structured models," Ph.D. dissertation, 2016.

[19] S. Nam, M. E. Davies, M. Elad, and R. Gribonval, "The cosparse analysis model and algorithms," *Applied and Computational Harmonic Analysis*, vol. 34, no. 1, pp. 30–56, 2013.

[20] P. Zwiernik, C. Uhler, and D. Richards, "Maximum likelihood estimation for linear gaussian covariance models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

[21] H. H. Bauschke, J. M. Borwein, and P. L. Combettes, "Bregman monotone optimization algorithms," *SIAM J. Control Optim.*, vol. 42, no. 2, pp. 596–636, 2003.

[22] L. Li and K.-C. Toh, "An inexact interior point method for l1-regularized sparse covariance selection," *Mathematical Programming Computation*, vol. 2, no. 3, pp. 291–315, 2010.

[23] C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. K. Ravikumar, "Sparse inverse covariance matrix estimation using quadratic approximation," in *NIPS 24*, 2011, pp. 2330–2338.

[24] D. Hsu, S. M. Kakade, and T. Zhang, "Tail inequalities for sums of random matrices that depend on the intrinsic dimension," *Electronic Communications in Probability*, vol. 17, no. 14, pp. 1–13, 2012.

Appendix

A. Details: Sparse Precision Matrix Estimation in Section IV

Proof. [of Proposition 6] Define $J_0 := \Sigma_0^{-1}$, $\mathcal{S} := \mathrm{supp}(\Sigma_0^{-1})$ and $k := \mathrm{card}(\Sigma_0^{-1}) = |\mathcal{S}|$. For $G \in \partial\|J\|_1 \subset [-1,1]^{p\times p}$,

$$\begin{cases} G_{ij} = \mathrm{sign}(J_{ij}) & \text{if } J_{ij} \neq 0 \\ G_{ij} \in [-1,1] & \text{if } J_{ij} = 0\,. \end{cases}$$

Define $\lambda' := \lambda\sqrt{n}$. For a given Z, if $\lambda' \geq \|Z\|_\infty$ then

$$\mathrm{dist}^2(Z, \lambda'\partial\|J_0\|_1) = \inf_{G\in\lambda'\partial\|J_0\|_1} \|Z - G\|_F^2$$

$$= \inf_{G\in\lambda'\partial\|J_0\|_1} \sum_{(i,j)\in\mathcal{S}} (Z_{ij} - \lambda'\mathrm{sign}((J_0)_{ij}))^2 + \sum_{(i,j)\notin\mathcal{S}} (Z_{ij} - G_{ij})^2$$

$$= \sum_{(i,j)\in\mathcal{S}} (Z_{ij} - \lambda'\mathrm{sign}((J_0)_{ij}))^2$$

$$\leq k(\|Z_{\mathcal{S}}\|_\infty + \lambda')^2 \leq 4k(\lambda')^2 = 4k\lambda^2 n.$$

To conclude the proof, we apply Lemma 8, establishing a high probability bound on $\|Z\|_\infty$, when $Z \sim \overline{\mathcal{W}}(n, \Sigma_0)$, under the stated assumptions on n and p and C. ■

Lemma 8: Let $Z \sim \overline{\mathcal{W}}(n, \Sigma_0)$ be a $p \times p$ matrix with centered Wishart distribution. If $p \geq 9$, $C > 0$, and $n > (2C+4)\log p$ then

$$\Pr\Big[\|Z\|_\infty \geq\ 20\|\Sigma_0\|_\infty\sqrt{(2C+4)\log p}\,\sqrt{n}\Big] \leq 2p^{-C}.$$

Proof. For a $p \times p$ matrix X, let $X_{\{i,j\}}$ denote the principal submatrix indexed by rows i and j and columns i and j. Then $Z_{\{i,j\}} \sim \overline{\mathcal{W}}(n, (\Sigma_0)_{\{i,j\}})$. Observe that

$$\|Z\|_\infty = \max_{1\leq i,j\leq p}\|Z_{\{i,j\}}\|_\infty \leq \max_{1\leq i,j\leq p}\|Z_{\{i,j\}}\|_{\mathrm{op}}$$
$$\leq \max_{1\leq i,j\leq p}\|(\Sigma_0)_{\{i,j\}}\|_{\mathrm{op}} \max_{1\leq i,j\leq p}\|W_{\{i,j\}} - nI_{2\times 2}\|_{\mathrm{op}}$$

where $W \sim \mathcal{W}(n, I_{2\times 2})$ is a standard 2×2 Wishart matrix. The following tail bound for the extreme eigenvalues of Wishart matrices is from [24, Lemma A.1]:

$$\Pr\left[\|W_{\{i,j\}} - nI_{2\times 2}\|_{\mathrm{op}} \geq 2n\epsilon_{\delta,n}\right] \leq \delta$$

where $\epsilon_{\delta,n} = 4\sqrt{\nu} + \nu$, $\nu = \frac{2\log 9 + \log(\frac{2}{\delta})}{n}$. By a union bound,

$$\Pr\left[\max_{1\leq i,j\leq p}\|W_{\{i,j\}} - nI_{2\times 2}\|_{\mathrm{op}} \geq\ 2n\epsilon_{\delta,n}\right] \leq p^2\delta.$$

Putting $\delta = 2p^{-C-2}$ we observe that if $p \geq 9$ then $2/\delta > 81$ so that $2\log 9 + \log(2/\delta) < 2\log(2/\delta) = 2(C+2)\log p < n$. Hence $\epsilon_{\delta,n} < 5\sqrt{(2C+4)(\log p)/n}$. Since each $(\Sigma_0)_{\{i,j\}}$ is positive semidefinite,

$$\|(\Sigma_0)_{\{i,j\}}\|_{\mathrm{op}} \leq \mathrm{tr}(\Sigma_0)_{\{i,j\}} \leq 2\|(\Sigma_0)_{\{i,j\}}\|_\infty$$

and so $\max_{1\leq i,j\leq p}\|(\Sigma_0)_{\{i,j\}}\|_{\mathrm{op}} \leq 2\|\Sigma_0\|_\infty$. Putting these observations together gives the stated result. ■