RC24836 (W0908-045) August 14, 2009
Mathematics

IBM Research Report

Sparse Signal Recovery with Exponential-Family Noise

Irina Rish
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Genady Grabarnik
CUNY

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Sparse Signal Recovery with Exponential-Family Noise

Irina Rish and Genady Grabarnik

Abstract

The problem of sparse signal recovery from a relatively small number of noisy measurements has been studied extensively in the recent literature on compressed sensing. However, the focus of those studies appears to be limited to the case of linear projections disturbed by *Gaussian* noise, and the sparse signal reconstruction problem is treated as *linear* regression with l_1-norm regularization constraint. A natural question to ask is whether one can accurately recover sparse signals under different noise assumptions. Herein, we extend the results of [13] to the more general case of *exponential-family noise* that includes Gaussian noise as a particular case, and yields l_1-regularized *Generalized Linear Model (GLM)* regression problem. We show that, under standard restricted isometry property (RIP) assumptions on the design matrix, l_1-minimization can provide a stable recovery of a sparse signal under exponential-family noise assumptions, and investigate (sufficient) recovery conditions for the general case, and for some specific members of the exponential family.

I. Introduction

Accurate and efficient recovery of sparse high-dimensional signals from low-dimensional linear measurements received much attention in the recent compressed sensing literature [4]–[7], [10], [12]. While the problem of finding the sparsest signal satisfying linear constraints is NP-hard as it involves a combinatorial problem of l_0-norm minimization, it turns out that using the l_1-norm instead can still accurately recover the original signal, under certain conditions, and yields efficient optimization algorithms. Particularly interesting for real-life applications is the case of signal recovery from *noisy* measurements [11], [13], which relates to practically all modern applications of compressed sensing in image processing, sensor networks, biology, and medical imaging, just to name a few (see [15] for a comprehensive list of references on compressed sensing and its recent applications).

The majority of work in compressed sensing assumes linear Gaussian noise model; specifically, given an input signal $\mathbf{x}$ and the design matrix A, the vector of measurements $\mathbf{y}$ is assumed to follow a Gaussian distribution with the mean $A\mathbf{x}$ and unit variance(s) (effectively, it is assumed that measurements are independent), yielding the sum-squared loss constraint in standard formulations of noisy compressed sensing [11], [13]. However, in many practical applications, non-Gaussian noise assumptions are more applicable: for example, Bernoulli or multinomial distributions are better suited for describing such measurements as (binary) failures or multilevel performance degradations of end-to-end test transactions ("probes") in a distributed computer systems [16], [18]; exponential distribution is better suited for describing nonnegative measurements such as end-to-end response time in such systems [3], [9]. Non-Gaussian observations, including binary, discrete, non-negative, etc., variables, are common in various other applications such as medical diagnosis, or computational biology (e.g., presence or absence of gene expression).

In this paper, we will consider a general class of exponential-family distributions that includes, besides Gaussian, a wide variety of other commonly used distributions, such as exponential, Bernoulli, multinomial, gamma, chi-square, beta, Weibull, Dirichlet, and Poisson, just to name a few. The corresponding regression problem of recovering the signal $\mathbf{x}$ from the measurements $\mathbf{y}$ that follow exponential-family noise, is to solve a *Generalized Linear Model (GLM)* regression, which essentially solves the log-likelihood optimization problem that for exponential-family likelihoods is equivalent to minimizing the corresponding *Bregman divergence* $d(\mathbf{y}, \mu(A\mathbf{x}))$, where μ is the mean parameter corresponding to the natural parameter $\theta = A\mathbf{x}$. In case of Gaussian likelihood, for example, $\mu = \theta$ and the corresponding Bregman divergence is simply the Euclidean distance $||\mathbf{y} - A = bx||$. Adding l_1-norm constraint to GLM regression allows for an efficient method of sparse signal recovery, and is often used in statistical literature [14]. Thus, a natural question to ask is to what extent stable signal recovery results from the compressed sensing literature apply to the linear measurements corrupted by an exponential-family noise? This work provides an initial investigation of this question, deriving some conditions for stable sparse signal recovery from exponential-family observations.

We show that accurate recovery of sparse signals under the exponential-family noise assumption is possible in many cases, and derive the conditions on such recovery, for a general case and for several individual exponential-family members. Essentially, given a sparse signal $\mathbf{x}^0$, we show that, if the measurement noise is small (expressed as a small Bregman divergence between the measurement y and the mean μ^0 of the distribution determined by the natural parameter $\theta^0 = A\mathbf{x}^0$) and the matrix A obeys the restricted isometry property (RIP) with appropriate RIP constant, then the solution to the GLM regression problem (i.e. l_1-norm minimization subject to Bregman-divergence constraint that replaces the sum-squared loss (Euclidean distance) constraint) approximates the true signal well. Moreover, we show that the results of [13] for a more general case of compressible, rather than sparse, signals can be also extended to the exponential-family noise.

Irina Rish is with IBM T.J. Watson Research Center, Yorktown, NY rish@us.ibm.com
Genady Grabarnik is with CUNY genady@gmail.com

II. BACKGROUND

A. Sparse Signal Recovery from Noisy Observations

We assume that $x^0 \in R^m$ is an s-sparse signal, i.e. a signal with no more than s nonzero entries, where $s << m$. Let A be an n by m matrix that produces a vector of linear projections $y^0 = Ax^0$, where $n << m$, and let and y be a vector of n noisy measurements that follow some noise distribution $P(y|Ax^0)$. It is often assumed that A satisfies the so-called "restricted isometry property" (RIP) at the sparsity level S (or S-restricted isometry property), that essentially says that every subset of columns of A with cardinality less than S behaves like an almost orthonormal system. Formally, following [8]

Definition 1 (Restricted Isometry Property) *Let A_T, where $T\ subset\{1, ..., m\}$ denote an $n \times T$ submatrix of A that contains columns with indexes in T. The S-restricted isometry constant δ_S of A is the smallest quantity such that*

$$(1 - \delta_S)||c||_{l_2}^2 \leq ||A_T c||_{l_2}^2 \leq (1 + \delta_S)||c||_{l_2}^2 \tag{1}$$

for any all subsets T with $|T| \leq S$ and for any vector $(c_j)_{j \in T}$ defined over coordinates in T. The matrix A is said to satisfy the restricted isometry property if there exists such constant δ_S that the eq. 1 is satisfied.
It was shown (e.g., in [8]) that if

$$\delta_S + \delta_{2S} + \delta_{3S} < 1,$$

then solving the l_1-minimization problem in eq. 2 below can recover any signal x that is S-sparse (contains no more than S non-zero entries).

Our question is: can one recover x^0 from y, given that noise is "sufficiently small" (to be defined precisely below)? This question has been answered in the compressed sensing literature for the particular case when the noise distribution is Gaussian. Indeed, [13] show that, if: (1) $||y - Ax^0||_{l_2} \leq \epsilon$ (small noise assumption), (2) x^0 is sufficiently sparse and the (3) matrix A obeys the restricted isometry property (RIP) with appropriate RIP constants, then the solution to the following l_1-optimization problem

$$x^* = \arg\min_x ||x||_{l_1} \qquad subject\ to\ ||y - Ax||_{l_2} \leq \epsilon \tag{2}$$

approximates the true signal well. More formally, Theorem 1 in [13] states:

Theorem 1 [13] *Let S be such that $\delta_{3S} + 3\delta_{4S} < 2$, where δ_S is the S-restricted isometry constant of the matrix A, as defined above. Then for any signal x^0 with the support $T^0 = \{t : x^0 \neq 0\}$, where $|T^0| \leq S$ and any noise vector (perturbation) e with $||e||_{l_2} \leq \epsilon$, the solution x^* to the problem in eq. 2 obeys*

$$||x^* - x^0||_{l_2} \leq C_S \cdot \epsilon, \tag{3}$$

where the constant C_S may only depend on δ_{4S}. For reasonable values of δ_{4S}, C_S is well-behaved; e.g. $C_S \approx 8.82$ for $\delta_{4S} = 1/5$ and $C_S \approx 10.47$ for $\delta_{4S} = 1/4$.

Moreover, [13] show that (1) no other recovery method "can perform fundamentally better for arbitrary perturbations of size ϵ, i.e. even if an oracle would make the actual support T^0 of x^0 available to us, making the problem well-posed, the least-squares solution $\hat{x}$ (i.e., the maximum-likelihood solution which is optimal in the absence of any other information) would approximate the true signal x^0 with the error proportional to ϵ".

Finally, [13] extend their result from sparse to approximately sparse vectors in the following

Theorem 2 [13] *Let $x^0 \in R^m$ be an arbitrary vector, and let $x_{0,S}$ be the truncated vector corresponding to the S largest values of x^0 (in absolute value). Under the assumptions of Theorem 1, the solution x^* to the problem in eq. 2 obeys*

$$||x^* - x^0||_{l_2} \leq C_{1,S} \cdot \epsilon + C_{2,S} \cdot \frac{||x^0 - x^0, S||_{l_1}}{\sqrt{S}}. \tag{4}$$

For reasonable values of δ_{4S} the constants above are well-behaved; e.g. $C_{1,S} \approx 12.04$ and $C_{2,S} \approx 8.77$ for $\delta_{4S} = 1/5$.

B. Exponential-family distributions and Bregman divergences

Herein, we will generalize the above results in the case of *exponential-family* noise. Note that $||y - Ax||_{l_2} \leq \epsilon$ is a constraint on the negative log-likelihood of a Gaussian variable $y \sim N(\mu, \Sigma)$ with $\mu = Ax$ and $\Sigma = I$ (assuming independent unit-variance noise), i.e.

$$-logP(y|Ax^0) = f(y) + \frac{1}{2}||y - Ax||_{l_2}^2. \tag{5}$$

Gaussian distribution is a particular member of the *exponential family* of distributions.

Definition 2 *An* **exponential family** *is a parametric family of probability distributions where the probability density has the form*

$$\log p_{\psi,\theta}(\mathbf{y}) = \mathbf{x}\theta - \psi(\theta) + \log p_0(\mathbf{y}), \tag{6}$$

where θ is called the natural parameter, *$\psi(\theta)$ is the (strictly convex and differentiable)* cumulant function, *or the* log-partition function, *that uniquely determines the member distribution of the exponential family, and $p_0(\mathbf{y})$ is a non-negative function called* base measure *that does not depend on the parameter θ.*

As shown by [2], there is a bijection between the exponential-family densities $p_{\psi,\theta}(\mathbf{y})$ and *Bregman divergences* $d_\phi(\mathbf{y},\mu)$, so that each exponential-family density can be also expressed as

$$p_{\psi,\theta}(\mathbf{y}) = exp(-d_\phi(\mathbf{y},\mu))f_\phi(\mathbf{y}), \tag{7}$$

where $\mu = \mu(\theta) = E_{p_{\psi,\theta}}(Y)$ is the *expectation parameter* corresponding to θ, ϕ is the (strictly convex and differentiable) Legendre conjugate of ψ, $f_\phi(\mathbf{y})$ is a uniquely determined function, and $d_\phi(\mathbf{y},\mu)$ is the corresponding Bregman divergence defined as follows.

Definition 3 *Given a strictly convex function $\phi : S \to \mathbb{R}$ defined on a convex set $S \subseteq \mathbb{R}$, and differentiable on the interior of S, $int(S)$ [17], the* **Bregman divergence** *$d_\phi : S \times int(S) \to [0,\infty)$ is defined as*

$$d_\phi(\mathbf{x},\mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - < (\mathbf{x}-\mathbf{y}), \nabla\phi(\mathbf{y}) >, \tag{8}$$

where $\nabla\phi(\mathbf{y})$ is the gradient of ϕ.

In other words, the Bregman divergence can be thought of as the difference between the value of ϕ at point $\mathbf{x}$ and the value of the first-order Taylor expansion of ϕ around point $\mathbf{y}$ evaluated at point $\mathbf{x}$ (see Figures 1 and 2 below, where $h(x) = \phi(y) + < (\mathbf{x}-\mathbf{y}), \nabla\phi(\mathbf{y}) >$**).**

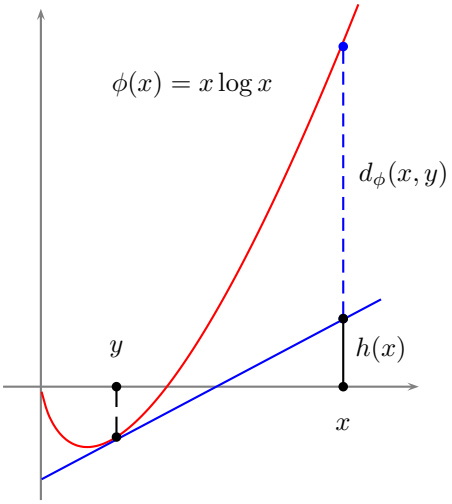

Fig. 1. Relative entropy (KL-divergence)

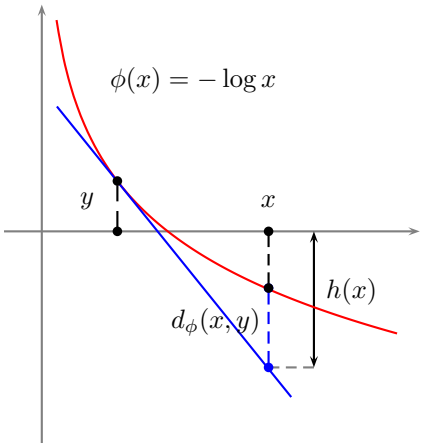

Fig. 2. Itakura-Saito distance (Burg divergence)

Table I (derived from Tables 1 and 2 in [1]) shows particular examples of commonly used exponential-family distributions and their corresponding Bregman divergences. For example, the unit-variance Gaussian distribution leads to square loss, multivariate spherical Gaussian (diagonal covariance/independent variables) gives rise to Euclidean distance, an multivariate Gaussian with the inverse-covariance (concentration) matrix C leads to Mahalanobis distance, Bernoulli distribution corresponds to logistic loss, exponential distribution leads to Itakura-Saito distance, while a multinomial distribution corresponds to the KL-divergence.

TABLE I. Examples of commonly-used exponential-family distributions and their corresponding Bregman divergences.

Domain	Distribution	$p_\theta(y)$	μ	$\phi(\mu)$	$d_\phi(\mathbf{y},\mu)$	Divergence
$\mathbb{R}$	$1D$ Gaussian	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-a)^2}{2\sigma^2}}$	a	$\frac{1}{2\sigma^2}\mu^2$	$\frac{1}{2\sigma^2}(y-\mu)^2$	square loss
$\{0,1\}$	Bernoulli	$q^y(1-q)^{1-y}$	q	$\mu\log\mu+$ $(1-\mu)\log(1-\mu)$	$y\log(\frac{y}{\mu})+$ $(1-y)\log(\frac{1-y}{1-\mu})$	logistic loss
R_{++}	Exponential	$\lambda e^{-\lambda y}$	$1/\lambda$	$-\log\mu-1$	$\frac{y}{\mu}-\log(\frac{y}{\mu})-1$	Itakura-Saito distance
n-simplex	nD Multinomial	$\frac{N!}{\prod_{j=1}^n y_j!}\prod_{j=1}^n q_j^{y_j}$	$[Nq_j]_{j=1}^{n-1}$	$\sum_{j=1}^n \mu_j\log(\frac{\mu_j}{N})$	$\sum_{j=1}^n y_j\log(\frac{y_j}{\mu_j})$	KL-divergence
$\mathbb{R}^n$	nD Sph. Gaussian	$\frac{1}{\sqrt{(2\pi\sigma^2)^n}}e^{-\frac{\|\|\mathbf{x}-\mathbf{a}\|\|_2^2}{2\sigma^2}}$	$\mathbf{a}$	$\frac{1}{2\sigma^2}\|\|\mu\|\|_2^2$	$\frac{1}{2\sigma^2}\|\|\mathbf{y}-\mu\|\|_2^2$	squared Euclidean distance
$\mathbb{R}^n$	nD Gaussian	$\frac{\sqrt{\det(C)}}{\sqrt{(2\pi)^n}}e^{-\frac{(\mathbf{y}-\mathbf{a})^T C(\mathbf{y}-\mathbf{a})}{2}}$	$\mathbf{a}$	$\frac{\mu^T C\mu}{2}$	$\frac{(\mathbf{y}-\mu)^T C(\mathbf{y}-\mu)}{2}$	Mahalanobis distance[1]

III. Our Contribution

We now extend the result in Theorem 1 to the case of exponential-family noise. Let us consider the following constrained l_1-regularization problem that generalizes the standard noisy compressed sensing problem of [13] to the following:

$$min||x||_1 \quad subject\ to \quad \sum_i d(y_i, \mu(A_i x)) \leq \epsilon, \tag{9}$$

where $d(y_i, \mu(A_i x))$ is Bregman divergence between the noisy observation y_i and the mean parameter of the corresponding exponential-family distribution with the natural parameter $\theta_i = A_i x$. Note that this problem corresponds to l_1-regularized *Generalized Linear Model (GLM)* regression, that includes as a particular case the standard compressed-sensing formulation, i.e. the l_1 regularized linear regression (in that case, Bregman divergence is simply the Euclidian distance, and $\mu(A_i x) = A_i x$).

We show that, if: (1) the noise is small, (2) x^0 is sufficiently sparse and the (3) matrix A obeys the restricted isometry property (RIP) with appropriate RIP constants, then the solution to the above problem approximates the true signal well. More formally,

Theorem 3 *Let S be such that $\delta_{3S} + 3\delta_{4S} < 2$, where δ_S is the S-restricted isometry constant of the matrix A, as defined above. Then for any signal x^0 with the support $T^0 = \{t : x^0 \neq 0\}$, where $|T^0| \leq S$, and for any vector $\mathbf{y} = (y_1, ..., y_n)$ of noisy linear measurements where*

1) *the noise follows exponential-family distributions $p_{\theta_i}(y_i)$, with the natural parameter $\theta_i = (A_{i,:}x^0)$, and*
2) *the noise is sufficiently small, i.e. $\forall i,\ d(y_i, \mu(A_{i,:}x^0)) \leq \epsilon$,*

the solution x^* to the problem in eq. 9 obeys

$$||x^* - x^0||_{l_2} \leq C_S \cdot \delta(\epsilon), \tag{10}$$

where C_S is the constant from Theorem 1 of [13], and $\delta(\epsilon)$ is a continuous monotone increasing function of ϵ s.t. $\delta(0) = 0$ (and thus $\delta(\epsilon)$ is small when ϵ is small). A particular form of this function depends on particular members of exponential family.

Proof: Following the proof of Theorem 1 in [13], we will only have to show that the "tube constraint" (condition 1) still holds (the rest of the proof remains unchanged), i.e. that

$$||Ax^* - Ax^0||_{l_2} \leq \delta(\epsilon) \tag{11}$$

where δ is some continuous monotone increasing function of ϵ, and $\delta(0) = 0$, so its small when ϵ is small. It was a trivial consequence of the triangle inequality in case of Euclidean distance; however, triangle inequality does not hold, in general, for Bregman divergences, and thus we must provide a different proof for the tube constraint, possibly for each type of Bregman divergence (exponential-family distribution). Since

$$||Ax^* - Ax^0||_{l_2}^2 = \sum_{i=1}^m (A_{i,:}x^* - A_{i,:}x^0)^2 = \sum_{i=1}^m (\theta_i^* - \theta_i^0)^2,$$

we will need to show that $|\theta_i^* - \theta_i^0| < \beta(\epsilon)$, where $\beta(\epsilon)$ is a continuous monotone increasing function of ϵ s.t. $\beta(0) = 0$ (and thus $\beta(\epsilon)$ is small when ϵ is small), then in eq. 11 we get $\delta(\epsilon) = \sqrt{m \cdot \beta(\epsilon)}$. The proof of this fact for the general case of exponential-family noise is provided by Lemma 1. However, for particular members of the exponential family, one may have to provide specific proofs since the assumptions required by the general proof do not always hold for specific cases. Thus, we provide separate proofs for several different members of the exponential family in Lemmas 2.1 and 2.2, and obtain

particular expressions for $\beta(\epsilon)$ in each case. Note that for simplicity sake, we only consider univariate exponential-family distributions, corresponding to the case of independent noise for each measurement y_i, which was effectively assumed in standard problem formulation that used Euclidean distance corresponding to a spherical Gaussian distribution, i.e. a vector of independent Gaussian variables. However, Lemma 1 below can be extended from scalar to vector case, i.e. to multivariate exponential-family distributions that do not necessarily imply independent noise. Lemma 2.3 will provide a specific case of such distribution - a multivariate Gaussian with concentration matrix C.

The "cone constraint" part of the proof in [13] remains intact; it is easy to see that it does not depend on the particular constraint in the l_1-minimization problem 9, and only makes use of the sparsity of x^0 and l_1-optimality of x^*. Thus, we can simply substitute $||Ah||_{l_2}||$ by $\delta(\epsilon)$ in eq. 13 on page 8 in the proof of Theorem 1 of [13], or, equivalently, replace 2ϵ (that was shown to bound $||Ah||_{l_2}||$) by $\delta(\epsilon)$ in the eq. 14.

■

Just like for the sparse signal case (Theorem 1 in [13]), the only change we have to make in the proof of the Theorem 2 (general case of approximable, rather than sparse, signals), when generalizing it from Euclidean distance to Bregman divergence in eq. 9, is the tube constraint. Thus, once we showed it for the Theorem 3 above, the generalization to approximable signals follows automatically:

Theorem 4 *Let $x^0 \in R^m$ be an arbitrary vector, and let $x_{0,S}$ be the truncated vector corresponding to the S largest values of x^0 (in absolute value). Under the assumptions of Theorem 3, the solution x^* to the problem in eq. 9 obeys*

$$||x^* - x^0||_{l_2} \leq C_{1,S} \cdot \delta(\epsilon) + C_{2,S} \cdot \frac{||x^0 - x^0, S||_{l_1}}{\sqrt{S}}. \tag{12}$$

where $C_{1,S}$ and $C_{2,S}$ are the constants from Theorem 2 of [13], and $\delta(\epsilon)$ is a continuous monotone increasing function of ϵ s.t. $\delta(0) = 0$ (and thus $\delta(\epsilon)$ is small when ϵ is small). A particular form of this function depends on particular members of exponential family.

The following lemma states the sufficient conditions for the "tube constraint" in eq. 11 to hold in general case of arbitrary exponential-family noise, provided that $\phi''(y)$ exists and is bounded on the appropriate intervals.

Lemma 1 *Let y denote a random variable following an exponential-family distribution $p_\theta(y)$, with the natural parameter θ, and the corresponding mean parameters $\mu(\theta)$. Let $d_\phi(y, \mu(\theta))$ denote the Bregman divergence associated with this distribution. If*

1) $d_\phi(y, \mu^0(\theta^0)) \leq \epsilon$ *(small noise),*
2) $d_\phi(y, \mu^*(\theta^*)) \leq \epsilon$ *(constraint in GLM problem eq. 9), and*
3) $\phi''(y)$ *exists and is bounded on* $[y_{min}, y_{max}]$*, where* $y_{min} = \min\{y, \mu^0, \mu^*\}$ *and* $y_{max} = \max\{y, \mu^0, \mu^*\}$,

then

$$|\mu^* - \mu^0| \leq \frac{2\sqrt{2\epsilon}}{\sqrt{\min_{\hat{y} \in [y_{min}; y_{max}]} \phi''(\hat{y})}}, \text{ and}$$

$$|\theta^* - \theta^0| \leq \frac{2\sqrt{2\epsilon}}{\sqrt{\min_{\hat{y} \in [y_{min}; y_{max}]} \phi''(\hat{y})}} \max_{\hat{\mu} \in [\mu^*; \mu^0]} |\phi''(\hat{\mu})|.$$

Proof: We prove the lemma in two steps: first, we show that $|\mu^*(\theta^*) - \mu^0(\theta^0)|$ is small if ϵ is small, and then infer $|\theta^* - \theta^0|$ is small.

1) By definition in eq. 8, Bregman divergence is the non-linear tail of the Taylor expansion of $\phi(y)$ at point μ, i.e., the *Lagrange remainder* of the linear approximation:

$$d_\phi(y, \mu) = \phi''(\hat{y})(y - \mu)^2/2, \quad \hat{y} \in [y_1; y_2], \text{ where } y_1 = \min\{y, \mu\},\ y_2 = \max\{y, \mu\}.$$

Let $y_1^0 = \min\{y, \mu^0\}$, $y_2^0 = \max\{y, \mu^0\}$ and $y_1^* = \min\{y, \mu^*\}$, $y_2^* = \max\{y, \mu^*\}$. Using the conditions $0 \leq d_\phi(y, \mu^0) \leq \epsilon$ and $0 \leq d_\phi(y, \mu^*) \leq \epsilon$, and observing that

$$\min_{\hat{y} \in [y_{min}; y_{max}]} \phi''(\hat{y}) \leq \min_{\hat{y} \in [y_1^0; y_2^0]} \phi''(\hat{y}) \text{ and } \min_{\hat{y} \in [y_{min}; y_{max}]} \phi''(\hat{y}) \leq \min_{\hat{y} \in [y_1^*; y_2^*]} \phi''(\hat{y}),$$

we get

$$\phi''(\hat{y})(y-\mu^0)^2/2 \le \epsilon \quad \Leftrightarrow \quad (y-\mu^0)^2 \le \frac{2\epsilon}{\phi''(\hat{y})} \;\Leftrightarrow |y-\mu^0| \le \frac{\sqrt{2\epsilon}}{\sqrt{\min_{\hat{y}\in[y_1^0;y_2^0]}\phi''(\hat{y})}} \le \frac{\sqrt{2\epsilon}}{\sqrt{\min_{\hat{y}\in[y_{min};y_{max}]}\phi''(\hat{y})}}$$

$$\text{and,} \quad \text{similarly,} \;\; |y-\mu^*| \le \frac{\sqrt{2\epsilon}}{\sqrt{\min_{\hat{y}\in[y_1^*;y_2^*]}\phi''(\hat{y})}} \le \frac{\sqrt{2\epsilon}}{\sqrt{\min_{\hat{y}\in[y_{min};y_{max}]}\phi''(\hat{y})}}, \tag{13}$$

from which, using the triangle inequality, we conclude

$$|\mu^*-\mu^0| \le |y-\mu^*| + |y-\mu^0| \le \frac{2\sqrt{2\epsilon}}{\sqrt{\min_{\hat{y}\in[y_{min};y_{max}]}\phi''(\hat{y})}} = \delta_1(\epsilon). \tag{14}$$

Note that $\phi''(\hat{y})$ under the square root is always nonnegative since ϕ us convex.

2) The mean and the natural parameters of an exponential-family distribution relate to each other as follows: $\theta(\mu) = \phi'(\mu)$ (respectively, $\theta(\mu) = \nabla\phi(\mu)$ for vector μ), where $\phi'(\mu)$ is called the *link function*. Therefore, we can write

$$|\theta^*-\theta^0| = |\phi'(\mu^*)-\phi'(\mu^0)| = |\phi''(\hat{\mu})(\mu^*-\mu^0)|, \quad \text{where} \;\; \hat{\mu} \in [\mu^*;\mu^0],$$

and thus, using the above result in eq. 14, we get

$$|\theta^*-\theta^0| \le \max_{\hat{\mu}\in[\mu^*;\mu^0]} |\phi''(\hat{\mu})|\delta_1(\epsilon), \tag{15}$$

which concludes the proof.

■

Note that the conditions (3) in the above lemma requires that $\phi(y)$ exists and is bounded on the intervals between y and both μ^0 and μ^*. However, even when this condition is not satisfied, as, for example, in case of logistic loss (where $\phi''(y) = \frac{1}{y(1-y)}$ is unbounded at 0 and 1) and several other Bregman divergences shown in Table 1 (e.g., , we can nevertheless prove a similar results using specific properties of each $\phi(y)$, as shown by the following lemmas.

Lemma 2.1 (Bernoulli noise / Logistic loss) *Let the conditions (1) and (2) of Lemma 1 be satisfied, and let $\phi(y) = y\log y + (1-y)\log(1-y)$, which corresponds to the logistic-loss Bregman divergence and Bernoulli distribution $p(y) =$, where the mean parameter $\mu = P(y=1)$. We assume that $0 < \mu^* < 1$, and $0 < \mu^0 < 1$. Then*

$$|\mu^0-\mu^*| \le 2\epsilon\cdot e^{2\epsilon}, \;\; \textit{and} \;\; |\theta^0-\theta^*| \le 4\epsilon.$$

Proof: Using the definition of the logistic-loss Bregman divergence from Table 1, and the above conditions, we can write:

$$d_\phi(y,\mu^0) = y\log(\frac{y}{\mu_0}) + (1-y)\log(\frac{1-y}{1-\mu^0}) \le \epsilon, d_\phi(y,\mu^*) = y\log(\frac{y}{\mu_*}) + (1-y)\log(\frac{1-y}{1-\mu^*}) \le \epsilon, \tag{16}$$

which implies

$$|d_\phi(y,\mu^0) - d_\phi(y,\mu^*)| \le 2\epsilon, \tag{17}$$

and, after substituting the expressions 16 into eq. 17, and simplifying, we get

$$|y\log(\frac{\mu^0}{\mu^*}) + (1-y)\log(\frac{1-\mu^0}{1-\mu^*})| \le 2\epsilon. \tag{18}$$

The above must be satisfied for each $y \in \{0,1\}$ (the domain of Bernoulli distribution). Thus, we get:

$$(1)\; |\log(\frac{1-\mu^0}{1-\mu^*})| \le 2\epsilon \;\; \text{if } y=0, \;\; \text{and} \;\; (2)\; |\log(\frac{\mu^0}{\mu^*})| \le 2\epsilon \;\; \text{if } y=1, \tag{19}$$

or, equivalently

$$(1)\; e^{-2\epsilon} \le \frac{1-\mu^0}{1-\mu^*} \le e^{2\epsilon} \;\; \text{if } y=0, \text{ and } (2)\; e^{-2\epsilon} \le \frac{\mu^0}{\mu^*} \le e^{2\epsilon} \;\; \text{if } y=1.$$

Let us first consider the case of $y=0$; subtracting 1 from the corresponding inequalities yields

$$e^{-2\epsilon}-1 \le \frac{\mu^*-\mu^0}{1-\mu^*} \le e^{2\epsilon}-1 \Leftrightarrow (1-\mu^*)(e^{-2\epsilon}-1) \le \mu^*-\mu^0 \le (1-\mu^*)(e^{2\epsilon}-1).$$

By the mean value theorem, $e^x - 1 = e^x - e^0 = \frac{d(e^x)}{dx}|_{\hat{x}} \cdot (x-0) = e^{\hat{x}}x$, for some $\hat{x} \in [0, x]$ if $x > 0$, or for some $\hat{x} \in [x, 0]$ if $x < 0$. Thus, $e^{-2\epsilon} - 1 = -e^{\hat{x}} \cdot 2\epsilon$, for some $\hat{x} \in [-2\epsilon, 0]$, and since e^x is a continuous monotone increasing function, $e^{\hat{x}} \leq 1$ and thus $e^{-2\epsilon} - 1 \geq -2\epsilon$. Similarly, $e^{2\epsilon} - 1 = e^{\hat{x}} \cdot 2\epsilon$, for some $\hat{x} \in [0, 2\epsilon]$, and since $e^{\hat{x}} \leq e^{2\epsilon}$, we get $e^{2\epsilon} - 1 \leq 2\epsilon \cdot e^{2\epsilon}$. Thus,

$$-2\epsilon(1-\mu^*) \leq \mu^* - \mu^0 \leq 2\epsilon e^{2\epsilon}(1-\mu^*) \Rightarrow |\mu^* - \mu^0| \leq 2\epsilon \cdot e^{2\epsilon}. \tag{20}$$

Similarly, in case of $y = 1$, we get

$$e^{-2\epsilon} - 1 \leq \frac{\mu^0 - \mu^*}{\mu^*} \leq e^{2\epsilon} - 1.$$

and can apply same derivation as above, and get same result for $|\mu^* - \mu^0|$ as in eq. 20. Finally, since $\theta(\mu) = \phi'(\mu) = \log(\frac{\mu}{1-\mu})$, we get

$$|\theta^0 - \theta^*| = |\log(\frac{\mu^0}{1-\mu^0}) - \log(\frac{\mu^*}{1-\mu^*})| = |\log(\frac{\mu^0}{\mu^*}) - \log(\frac{1-\mu^0}{1-\mu^*})|.$$

From the eq. 19 we get $|\log(\frac{\mu^0}{\mu^*})| \leq 2\epsilon$ and $|\log(\frac{1-\mu^0}{1-\mu^*})| \leq 2\epsilon$, thus

$$|\theta^0 - \theta^*| = |\log(\frac{\mu^0}{\mu^*}) - \log(\frac{1-\mu^0}{1-\mu^*})| \leq 4\epsilon.$$

■

Lemma 2.2 (Exponential noise/ Itakura-Saito distance) *Let the conditions (1) and (2) of Lemma 1 be satisfied, and let* $\phi(y) = -\log\mu - 1$*, which corresponds to the Itakura-Saito distance* $d_\phi(y, \mu) = \frac{y}{\mu} - \log(\frac{y}{\mu}) - 1$ *and exponential distribution* $p(y) = \lambda e^{\lambda y}$*, where the mean parameter* $\mu = 1/\lambda$*. We will also assume that the mean parameter is always separated from zero, i.e.* $\exists c_\mu > 0$ *such that* $\mu \geq c_\mu$*. Then*

$$|\mu^0 - \mu^*| \leq \sqrt{6\,\epsilon} \cdot \max\{\mu^0, \mu^*\}, \text{ and}$$

$$|\theta^* - \theta^0| \leq \frac{\sqrt{6\,\epsilon}}{c_\mu}.$$

Proof: To establish the result of the lemma we start with inequality $|u - \log u - 1| \leq \epsilon$, where u is $\frac{y}{\mu}$. Replacing u by $z = u - 1, z > -1$ gives us $|z - \log(1+z)| \leq \epsilon$. Without loss of generality, let us assume that $\epsilon \leq \frac{1}{18}$. Then the Taylor decomposition of function $z - \log(1+z)$ at the point $z = 0$

$$z - \log(1+z) = \frac{z^2}{2} - \frac{z^3}{3} + \frac{\theta^4}{4}, \text{ for } \theta \in [0, z] \text{ or } [z, 0]$$

implies that

$$\epsilon \geq z - \log(1+z) \geq \frac{z^2}{2} - \frac{z^3}{3} \text{ (since } \frac{\theta^4}{4} \geq 0 \text{).}$$

This, in turns, implies that $z \leq \frac{1}{3}$ and $\frac{z^2}{2} - \frac{z^3}{3} \geq \frac{z^2}{6}$ for $0 \leq z \leq \frac{1}{3}$.

Hence

$$z - \log(1+z) \geq \frac{z^2}{2} \quad \text{for} \quad -\frac{1}{3} \leq z \leq 0, \tag{21}$$

$$z - \log(1+z) \geq \frac{z^2}{6} \quad \text{for} \quad 0 \leq z \leq \frac{1}{3}. \tag{22}$$

Combining together both estimates we get $|z| \leq \sqrt{6\,\epsilon}$, or

$$|y - \mu| \leq \sqrt{6\,\epsilon} \cdot \mu,$$

and

$$|\mu^0 - \mu^*| \leq \sqrt{6\,\epsilon} \cdot \max\{\mu^0, \mu^*\}.$$

Then

$$|\theta^* - \theta^0| = |\frac{1}{\mu^0} - \frac{1}{\mu^*}| = |\frac{\mu^* - \mu^0}{\mu^*\,\mu^0}| \leq \frac{\sqrt{6\,\epsilon}}{\min\{\mu^*, \mu^0\}} \leq \frac{\sqrt{6\,\epsilon}}{c_\mu},$$

since by the assumption of the lemma $\min\{\mu^*, \mu^0\} \geq c_\mu$.

■

We now consider multivariate exponential-family distributions; the next lemma handles the general case of a multivariate Gaussian distribution (not necessarily spherical one that had a diagonal covariance matrix and corresponded to the standard Euclidean distance (see Table 1).

Lemma 2.3 (Non-i.i.d. Multivariate Gaussian noise / Mahalanobis distance) *Let $\phi(\mathbf{y}) = \mathbf{y}^T C\mathbf{y}$, which corresponds to the general multivariate Gaussian with concentration matrix C, and Mahalanobis distance $d_\phi(y,\mu) = \frac{1}{2}(\mathbf{y}-\mu)^T C(\mathbf{y}-\mu)$. If $d_\phi(y,\mu^0) \leq \epsilon$ and $d_\phi(y,\mu^*) \leq \epsilon$, then*

$$||\mu^0 - \mu^*|| \leq \frac{2\sqrt{\epsilon}}{2}||C^{-1}||^{1/2}, \text{ and } ||\theta^0 - \theta^*|| \leq \sqrt{2\epsilon}||C^{-1}||^{1/2} \cdot ||C||.$$

Proof: Since C is (symmetric) positive definite, it can be written as $C = L^T L$ where L defines a linear operator on $\mathbf{y}$ space, and thus

$$\epsilon/2 \geq (\mathbf{y}-\mu)^T C(\mathbf{y}-\mu) = (L(\mathbf{y}-\mu))^T(L(\mathbf{y}-\mu)) = ||L(\mathbf{y}-\mu)||^2.$$

Also, it is easy to show that $||C^{-1}||I \leq C \leq ||C||I$ (where $||B||$ will herein denote the operator norm of B), and that

$$\epsilon/2 \geq ||L(\mathbf{y}-\mu)||^2 \geq ||L^{-1}||^{-2}||\mathbf{y}-\mu||^2 \Rightarrow ||\mathbf{y}-\mu|| \leq \sqrt{\frac{\epsilon}{2}}||L^{-1}||.$$

Then, using triangle inequality, we get

$$||\mu^* - \mu^0|| \leq ||\mathbf{y}-\mu^0|| + ||\mathbf{y}-\mu^*|| \leq \sqrt{2\epsilon}||L^{-1}||.$$

Finally, since $\theta(\mu) = \nabla\phi(\mu) = C\mu$, we get

$$||\theta^0 - \theta^*|| = ||C\mu^0 - C\mu^*|| \leq ||C|| \cdot ||\mu^0 - \mu^*|| = ||C|| \cdot ||\mu^0 - \mu^*|| \leq \sqrt{2\epsilon}||L^{-1}|| \cdot ||C||.$$

Note that $||L^{-1}|| = ||C^{-1}||^{1/2}$, which concludes the proof. ■

IV. Summary

In this paper, we extend the results of [13] to the more general case of *exponential-family noise* that includes Gaussian noise as a particular case, and yields l_1-regularized *Generalized Linear Model (GLM)* regression problem. We show that, under standard restricted isometry property (RIP) assumptions on the design matrix, l_1-minimization can provide a stable recovery of a sparse signal under exponential-family noise assumptions, and investigate (sufficient) recovery conditions for the general case, and for some specific members of the exponential family. We also show that the results of [13] for a more general case of compressible (rather than sparse) signals can be extended to the exponential-family noise in a similar way.

Clearly, this is work in progress, since it only demonstrated the results for two specific members of the exponential family, and needs to be extended to others. We also show the general results; however, the conditions imposed by the Lemma 1 might be sometimes too restrictive, and thus we must further explore the sparse signal recovery conditions.

References

[1] A. Banerjee and S. Merugu and I. S. Dhillon and J. Ghosh. Clustering with Bregman Divergences. *Journal of Machine Learning Research*, 6:1705–1749, October 2005.

[2] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with Bregman divergences. In *Proceedings of the Fourth SIAM International Conference on Data Mining*, pages 234–245, April 2004.

[3] A. Beygelzimer, J. Kephart, and I. Rish. Evaluation of optimization methods for network bottleneck diagnosis. In *In Proc. of ICAC-07*, 2007.

[4] E. Candes. Compressive sampling. In *Int. Congress of Mathematics*, volume 3, pages 1433–1452, 2006.

[5] E. Candes and J. Romberg. Quantitative robust uncertainty principles and optimally sparse decompositions. *Foundations of Comput. Math.*, 6(2):227–254, April 2006.

[6] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. on Information Theory*, 52(2):489–509, February 2006.

[7] E. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. on Information Theory*, (51(12)):4203–4215, December 2005.

[8] E. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. on Information Theory*, 51(12):4203 – 4215, December 2005.

[9] G. Chandalia and I. Rish. Blind Source Separation Approach to Performance Diagnosis and Dependency Discovery. In *In Proceedings of IMC-2007*, 2007.

[10] D. Donoho. Compressed sensing. *IEEE Trans. on Information Theory*, 52(4):1289–1306, April 2006.

[11] D. Donoho. For most large underdetermined systems of linear equations, the minimal ell-1 norm near-solution approximates the sparsest near-solution. *Communications on Pure and Applied Mathematics*, 59(7):907–934, July 2006.

[12] D. Donoho. For most large underdetermined systems of linear equations, the minimal ell-1 norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, (59(6)):797–829, June 2006.

[13] E. Candes and J. Romberg and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, August 2006.

[14] Mee-Young Park and Trevor Hastie. An L1 Regularization-path Algorithm for Generalized Linear Models. *JRSSB*, 69(4):659–677, 2007.

[15] Compressive Sensing Resources. http://dsp.rice.edu/cs.

[16] I. Rish, M. Brodie, S. Ma, N. Odintsova, A. Beygelzimer, G. Grabarnik, and K. Hernandez. Adaptive diagnosis in distributed systems. *IEEE Transactions on Neural Networks (special issue on Adaptive Learning Systems in Communication Networks)*, 16(5):1088–1109, 2005.
[17] R.T. Rockafeller. *Convex Analysis*. Princeton University Press, 1970.
[18] A. Zheng, I. Rish, and A. Beygelzimer. Efficient Test Selection in Active Diagnosis via Entropy Approximation. In *Proceedings of UAI-05*, 2005.