Learning Model-Based Sparsity via Projected Gradient Descent

Sohail Bahmani, *Student, IEEE*, Petros T. Boufounos, *Member, IEEE*, and Bhiksha Raj,

Abstract

Several convex formulation methods have been proposed previously for statistical estimation with structured sparsity as the prior. These methods often require a carefully tuned regularization parameter, often a cumbersome or heuristic exercise. Furthermore, the estimate that these methods produce might not belong to the desired sparsity model, albeit accurately approximating the true parameter. Therefore, greedy-type algorithms could often be more desirable in estimating structured-sparse parameters. So far, these greedy methods have mostly focused on linear statistical models. In this paper we study the projected gradient descent with non-convex structured-sparse parameter model as the constraint set. Should the cost function have a Stable Model-Restricted Hessian the algorithm converges to the desired minimizer up to an approximation error. As an example we elaborate on application of the main results to estimation in Generalized Linear Models.

I. Introduction

In a variety of applications such as bioinformatics, medical imaging, social networks, and astronomy there is a growing demand for computational methods that perform statistical inference on high-dimensional data. In the problems arising in these applications, p, the number of predictors in each sample is much larger than n, the number of observations. Although such problems are generally ill-posed, in many cases the data has known underlying structure such as sparsity that can be exploited to make the problem well-posed.

Beyond the ordinary, extensively studied, sparsity model, a variety of structured sparsity models have been proposed in the literature [1]–[8]. These sparsity models are designed to capture the interdependence of the locations of the non-zero components that is known *a priori* in certain applications. The models proposed for structured sparsity can be divided into two types. Models of the first type have a combinatorial construction and explicitly enforce the permitted "non-zero patterns" [4], [7], [9]. Greedy algorithms have been proposed for the least squares regression with true parameters belonging to such combinatorial sparsity models [4], [9]. Models of the second type capture sparsity patterns induced by the convex penalty functions tailored for specific estimation problems. For example, consistency of linear regression with mixed ℓ_1/ℓ_2-norm regularization in estimation of group sparse signals having non-overlapping groups is studied in [1]. Furthermore, a different convex penalty to induce group sparsity with overlapping groups is proposed in [3]. In [5], using submodular functions and their Lovász extension, a more general framework for design of convex penalties that induce given sparsity patterns is proposed. In [8] a convex signal model is proposed that is generated by a set of base signals called "atoms". The model can describe not only plain and structured sparsity, but also low-rank matrices and several other low-dimensional models. We refer readers to [10], [11] for extensive reviews on the estimation of signals with structured sparsity.

In addition to linear regression problems under structured sparsity assumptions, nonlinear statistical models have been studied in the convex optimization framework [1], [2], [6], [12]. For example, using the signal model introduced in [8], minimization of a convex function obeying a *restricted smoothness property* is studied in [12] where a coordinate-descent type of algorithm is shown to converge to the minimizer at a sublinear rate. In this formulation and other similar methods that rely on convex relaxation one needs to choose a regularization parameter to guarantee the desired statistical accuracy. However, choosing the appropriate value of this parameter may be intractable. Furthermore, the convex signal models usually provide an approximation of the ideal structures the estimates should have, while in certain tasks such as variable selection solutions are required to exhibit the exact structure considered. Therefore, in such tasks, convex optimization techniques may yield estimates that do not satisfy the desired structural properties , albeit accurately approximating the true parameter. These shortcomings motivate application of combinatorial sparsity structures in nonlinear statistical models, extending prior results such as [4], [9] that have focused exclusively on linear models.

Among the non-convex greedy algorithms, a generalization of Compressed Sensing is considered in [13] where the measurement operator is a nonlinear map and the union of subspaces is assumed as the signal model. This formulation, however, admits only a limited class of objective functions that are described using a norm. Furthermore, [14] proposes a generalization of the Orthogonal Matching Pursuit algorithm [15] that is specifically designed for estimation of group sparse parameters in Generalized Linear Models (GLMs). Also, [16] studies the problem of minimizing a generic objective function subject to sparsity constraint from the optimization perspective. Using certain necessary optimality conditions for the sparse minimizer, a few iterative algorithms are proposed in [16] that converge to the sparse minimizer, should the objective satisfies some conditions. However, that work does not address the minimization under structured sparsity.

S.B. is with the Department of Electrical and Computer Engineering at Carnegie Mellon University.
P.B. is with Mitsubishi Electric Research Labs.
B.R. is with the Language Technologies Institute and the Department of Electrical and Computer Engineering at Carnegie Mellon University.

In this paper we study the projected gradient descent method to approximate the minimizer of a cost function subject to a model-based sparsity constraint. The algorithm is described in Section II. The sparsity model considered in this paper is similar to the models in [4], [9] with minor differences in the definitions. To guarantee the accuracy of the algorithm our analysis requires the cost function to have a Stable Model-Restricted Hessian (SMRH) as defined in Section III. Using this property we show that for any given reference point in the considered model, each iteration shrinks the distance to the reference point up to an approximation error. As an example, Section III considers the cost functions that arise in Generalized Linear Models and discusses how the proposed sufficient condition (i.e., SMRH) can be verified and how large the approximation error of the algorithm is. To make precise statements on the SMRH and on the size of the approximation error we assume some extra properties on the cost function and/or the data distribution. Finally, we discuss and conclude in Section V.

Notation.: In the remainder of the paper we denote the positive part of a real number x by $(x)_+$. For a positive integer k, the set $\{1, 2, \ldots, k\}$ is denoted by $[k]$. Vectors and matrices are denoted by boldface characters and sets by calligraphic letters. The support set (i.e., the set of non-zero coordinates) of a vector $\mathbf{x}$ is denoted by $\operatorname{supp}(\mathbf{x})$. Restriction of a p-dimensional vector $\boldsymbol{v}$ to its entries corresponding to an index set $\mathcal{I} \subseteq [p]$ is denoted by $\mathbf{v}|_{\mathcal{I}}$. Similarly $\mathbf{A}_{\mathcal{I}}$ denotes the restriction of a matrix $\mathbf{A}$ to the rows enumerated by $\mathcal{I}$. For square matrices $\mathbf{A}$ and $\mathbf{B}$ we write $\mathbf{B} \preccurlyeq \mathbf{A}$ to state that $\mathbf{A} - \mathbf{B}$ is positive semidefinite. We denote the power set of a set $\mathcal{A}$ as $2^{\mathcal{A}}$. For two non-empty families of sets $\mathcal{F}_1$ and $\mathcal{F}_2$ we write $\mathcal{F}_1 \Cup \mathcal{F}_2$ to denote another family of sets given by $\{\mathcal{X}_1 \cup \mathcal{X}_2 \mid \mathcal{X}_1 \in \mathcal{F}_1 \text{ and } \mathcal{X}_2 \in \mathcal{F}_2\}$. Moreover, for any non-empty family of sets $\mathcal{F}$ for conciseness we set $\mathcal{F}^j = \mathcal{F} \Cup \ldots \Cup \mathcal{F}$ where the operation $\Cup$ is performed $j-1$ times. The inner product associated with a Hilbert space $\mathcal{H}$ is written as $\langle \cdot, \cdot \rangle$. The norm induced by this inner product is denoted by $\|\cdot\|$. We use $\nabla f(\cdot)$ and $\nabla^2 f(\cdot)$ to denote the gradient and the Hessian of a twice continuously differentiable function $f : \mathcal{H} \mapsto \mathbb{R}$. For an index set $\mathcal{I} \subset [p]$ with $p = \dim(\mathcal{H})$, the restriction of the gradient to the entries selected by $\mathcal{I}$ and the restriction of the Hessian to the entries selected by $\mathcal{I} \times \mathcal{I}$ are denoted by $\nabla_{\mathcal{I}} f(\cdot)$ and $\nabla^2_{\mathcal{I}} f(\cdot)$, respectively. Finally, numerical superscripts within parentheses denote the iteration index.

II. Problem Statement and Algorithm

To formulate the problem of minimizing a cost function subject to structured sparsity constraints, first we provide a definition of the sparsity model. This definition is an alternative way of describing the *Combinatorial Sparse Models* in [7]. In comparison, our definition merely emphasizes the role of a family of index sets as a *generator* of the sparsity model.

Definition 1. Suppose that p and k are two positive integers with $k \ll p$. Furthermore, denote by $\mathcal{C}_k$ a family of some non-empty subsets of $[p]$ that have cardinality at most k. The set $\bigcup_{\mathcal{S} \in \mathcal{C}_k} 2^{\mathcal{S}}$ is called a sparsity model of order k generated by $\mathcal{C}_k$ and denoted by $\mathcal{M}(\mathcal{C}_k)$.

Remark 1. Note that if a set $\mathcal{S} \in \mathcal{C}_k$ is a subset of another set in $\mathcal{C}_k$, then the same sparsity model can still be generated after removing $\mathcal{S}$ from $\mathcal{C}_k$ (i.e., $\mathcal{M}(\mathcal{C}_k) = \mathcal{M}(\mathcal{C}_k \backslash \{\mathcal{S}\})$). Thus, we can assume that there is no pair of distinct sets in $\mathcal{C}_k$ that one is a subset of the other.

In this paper we aim to approximate the solution to the optimization problem

$$\arg\min_{\boldsymbol{\theta} \in \mathcal{H}} f(\boldsymbol{\theta}) \quad \text{s.t. } \operatorname{supp}(\boldsymbol{\theta}) \in \mathcal{M}(\mathcal{C}_k), \tag{1}$$

where $f : \mathcal{H} \mapsto \mathbb{R}$ is a cost function with $\mathcal{H}$ being a p-dimensional real Hilbert space, and $\mathcal{M}(\mathcal{C}_k)$ a given sparsity model described by Def. 1. To approximate a solution $\widehat{\boldsymbol{\theta}}$ to (1) we use a *projected gradient descent* method summarized in Alg. 1. The only difference between Alg. 1 and standard projected gradient descent methods studied in convex optimization literature is that the projection, in line 3, is performed onto the generally non-convex set $\mathcal{M}(\mathcal{C}_k)$. The projection operator $\mathrm{P}_{\mathcal{C}_k, r} : \mathcal{H} \mapsto \mathcal{H}$ at any given point $\boldsymbol{\theta}_0 \in \mathcal{H}$ is defined as a solution to

$$\arg\min_{\boldsymbol{\theta} \in \mathcal{H}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \quad \text{s.t. } \operatorname{supp}(\boldsymbol{\theta}) \in \mathcal{M}(\mathcal{C}_k) \text{ and } \|\boldsymbol{\theta}\| \leq r. \tag{2}$$

Remark 2. In the context of statistical estimation, the cost function $f(\cdot)$ is usually the empirical loss associated with some observations generated by an underlying true parameter $\boldsymbol{\theta}^\star$. In these problems, it is more desired to estimate $\boldsymbol{\theta}^\star$ as it describes the data. The analysis presented in this paper allows evaluating the approximation error of the proposed algorithm with respect to any parameter vector in the considered sparsity model including $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^\star$. However, the approximation error with respect to the statistical truth $\boldsymbol{\theta}^\star$ can be simplified and interpreted to a greater extent. We elaborate more on this in Section III.

Remark 3. Assuming that for every $\mathcal{S} \in \mathcal{C}_k$ the cost function has a unique minimum over the set $\{\boldsymbol{\theta} \mid \operatorname{supp}(\boldsymbol{\theta}) \subseteq \mathcal{S} \text{ and } \|\boldsymbol{\theta}\| \leq r\}$, the operator $\mathrm{P}_{\mathcal{C}_k, r}[\cdot]$ can be defined without invoking *the axiom of choice* because there are only a finite number of choices for the set $\mathcal{S}$. One may also question the necessity of the constraint $\|\boldsymbol{\theta}\| \leq r$ in (2). As discussed later in Section IV, in statistical estimation problems where the cost function is not quadratic the sufficient condition we rely on cannot be guaranteed to hold unless the iterates and the true parameter lie in a bounded set. This shortcoming is typical for convergence proofs that use similar types of conditions (cf. [17]–[20]). Finally, the exact projection onto the sparsity model $\mathcal{M}(\mathcal{C}_k)$ might not be tractable. One may desire to show that accuracy can be guaranteed even using an inexact projection operator, at the cost of an extra error term. Existence and complexity of algorithms that find the desired exact or

Algorithm 1: Gradient Descent with Model Sparsity Constraint

input : $\mathcal{C}_k$, the family of possible supports,
r, the radius of feasible set
$i \longleftarrow 0$, $\theta^{(i)} \longleftarrow \mathbf{0}$
repeat
1 Choose step-size $\eta^{(i)} > 0$
2 $\chi^{(i)} \longleftarrow \theta^{(i)} - \eta^{(i)} \nabla f\left(\theta^{(i)}\right)$
3 $\theta^{(i+1)} \longleftarrow \mathrm{P}_{\mathcal{C}_k, r}\left[\chi^{(i)}\right]$
4 $i \longleftarrow i + 1$
until *halting condition holds*
return $\theta^{(i)}$

approximate projections, disregarding the length constraint in (2) (i.e., $\mathrm{P}_{\mathcal{C}_k,+\infty}[\cdot]$), are studied in [7], [9] for several interesting sparsity models. Also, in the general case where $r < +\infty$ one can derive a projection $\mathrm{P}_{\mathcal{C}_k,r}[\theta]$ from $\mathrm{P}_{\mathcal{C}_k,+\infty}[\theta]$ (see Lemma 2 in the Appendix). It is straightforward to generalize the guarantees in this paper to cases where only approximate projection is tractable. However, we do not attempt it here; our focus is to study the algorithm when the cost function is not necessarily quadratic. Instead, we apply the results to statistical estimation problems with non-linear models and we derive bounds on the statistical error of the estimate.

III. Theoretical Analysis

A. Stable Model-Restricted Hessian

In order to demonstrate accuracy of estimates obtained using Alg. 1 we require a variant of the *Stable Restricted Hessian* (SRH) condition proposed in [21] to hold. The SRH condition basically characterizes cost functions that have bounded curvature over canonical sparse subspaces. In this paper we require this condition to hold merely for the signals that belong to the considered model. Furthermore, we explicitly bound the length of the vectors at which the condition should hold. As will be discussed later, this restriction is necessary in general for non-quadratic cost functions. The condition we rely on, the Stable Model-Restricted Hessian (SMRH), can be formally defined as follows.

Definition 2. Let $f : \mathcal{H} \mapsto \mathbb{R}$ be a twice continuously differentiable function. Furthermore, let $\alpha_{\mathcal{C}_k}$ and $\beta_{\mathcal{C}_k}$ be in turn the largest and smallest real numbers such that

$$\beta_{\mathcal{C}_k} \|\boldsymbol{\Delta}\|^2 \leq \left\langle \boldsymbol{\Delta}, \nabla^2 f(\theta)\, \boldsymbol{\Delta} \right\rangle \leq \alpha_{\mathcal{C}_k} \|\boldsymbol{\Delta}\|^2, \tag{3}$$

holds for all $\boldsymbol{\Delta}$ and θ such that $\mathrm{supp}(\boldsymbol{\Delta}) \cup \mathrm{supp}(\theta) \in \mathcal{M}(\mathcal{C}_k)$ and $\|\theta\| \leq r$. Then f is said to have a Stable Model-Restricted Hessian with respect to the model $\mathcal{M}(\mathcal{C}_k)$ with constant $\mu_{\mathcal{C}_k} \geq 1$ in a sphere of radius $r > 0$, or in short ($\mu_{\mathcal{C}_k}$,r)-SMRH, if $1 \leq \alpha_{\mathcal{C}_k} / \beta_{\mathcal{C}_k} \leq \mu_{\mathcal{C}_k}$.

Remark 4. Typically in parametric estimation problems a sample loss function $l(\theta, \mathbf{x}, y)$ is associated with the covariate-response pair $(\mathbf{x}, y)$ and a parameter θ. Given n iid observations the empirical loss is formulated as $\widehat{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(\theta, \mathbf{x}_i, y_i)$. The estimator under study is the minimizer of the empirical loss, perhaps considering an extra regularization or constraint for the parameter θ. Most of the algorithms proposed for sparse estimation problems require that the cost function is strongly convex over a restricted but unbounded set of directions around the true parameter $\theta^\star$. It is known, however, that $\widehat{L}_n(\theta)$ as an empirical process is a good approximation of the expected loss $L(\theta) = \mathbb{E}[l(\theta, \mathbf{x}, y)]$ (see [22] and [23, Chapter 5]). If the required sufficient condition is not satisfied by $L(\theta)$ for a valid choice of $\theta^\star$, then in general it cannot be satisfied at the same $\theta^\star$ by $\widehat{L}_n(\theta)$ either. Thus, as also assumed in the prior work either explicitly [17] or implicitly [18]–[20], for a generic sample loss it is only possible to guarantee these types of sufficient conditions if the set of valid vectors $\theta^\star$ are further restriced, e.g., by bounding their length. This is the motivation behind the restriction imposed on the length of θ in Def. 2. Of course, if the true parameter violates this restriction we may incur an estimation bias as quantified in Theorem 1.

B. Accuracy Guarantee

Using the notion of SMRH we can now state the main theorem.

Theorem 1. *Consider the sparsity model $\mathcal{M}(\mathcal{C}_k)$ for some $k \in \mathbb{N}$ and a cost function $f : \mathcal{H} \mapsto \mathbb{R}$ that satisfies the $\left(\mu_{\mathcal{C}_k^3}, r\right)$-SMRH condition with parameters $\alpha_{\mathcal{C}_k^3}$ and $\beta_{\mathcal{C}_k^3}$ in (3). If $\eta^\star = 2/\left(\alpha_{\mathcal{C}_k^3} + \beta_{\mathcal{C}_k^3}\right)$ then for any $\overline{\theta} \in \mathcal{M}(\mathcal{C}_k)$ with $\left\|\overline{\theta}\right\| \leq r$ the iterates of Alg. 1 obey*

$$\left\|\theta^{(i+1)} - \overline{\theta}\right\| \leq 2\gamma^{(i)} \left\|\theta^{(i)} - \overline{\theta}\right\| + 2\eta^{(i)} \left\|\nabla_{\overline{\mathcal{I}}} f\left(\overline{\theta}\right)\right\|, \tag{4}$$

where $\gamma^{(i)} = \frac{\eta^{(i)}}{\eta^\star}\frac{\mu_{\mathcal{C}_k^3}-1}{\mu_{\mathcal{C}_k^3}+1} + \left|\frac{\eta^{(i)}}{\eta^\star} - 1\right|$ *and* $\overline{\mathcal{I}} = \operatorname{supp}\left(\mathrm{P}_{\mathcal{C}_k^2,r}\left[\nabla f\left(\overline{\boldsymbol{\theta}}\right)\right]\right)$.

Remark 5. One should choose the step size to achieve a contraction factor $2\gamma^{(i)}$ that is as small as possible. Straightforward algebra shows that the constant step-size $\eta^{(i)} = \eta^\star$ is optimal, but this choice may not be practical as the constants $\alpha_{\mathcal{C}_k^3}$ and $\beta_{\mathcal{C}_k^3}$ might not be known. Instead, we can always choose the step-size such that $1/\alpha_{\mathcal{C}_k^3} \leq \eta^{(i)} \leq 1/\beta_{\mathcal{C}_k^3}$ provided that the cost function obeys the SMRH condition by setting $\eta^{(i)} = 1/\left\langle \boldsymbol{\Delta}, \nabla^2 f\left(\boldsymbol{\theta}\right)\boldsymbol{\Delta}\right\rangle$ for some $\boldsymbol{\Delta}$,$\boldsymbol{\theta} \in \mathcal{H}$ such that $\operatorname{supp}\left(\boldsymbol{\Delta}\right) \cup \operatorname{supp}\left(\boldsymbol{\theta}\right) \in \mathcal{M}\left(\mathcal{C}_k^3\right)$. For this choice of $\eta^{(i)}$, we have $\gamma^{(i)} \leq \mu_{\mathcal{C}_k^3} - 1$.

Corollary 1. *A fixed step-size* $\eta > 0$ *coefficient corresponds to a fixed contraction coefficient* $\gamma = \frac{\eta}{\eta^\star}\frac{\mu_{\mathcal{C}_k^3}-1}{\mu_{\mathcal{C}_k^3}+1} + \left|\frac{\eta}{\eta^\star} - 1\right|$. *In this case, assuming that* $2\gamma \neq 1$, *the* i*-th iterate of Alg. 1 satisfies*

$$\left\|\boldsymbol{\theta}^{(i)} - \overline{\boldsymbol{\theta}}\right\| \leq (2\gamma)^i \left\|\overline{\boldsymbol{\theta}}\right\| + 2\eta\frac{1-(2\gamma)^i}{1-2\gamma}\left\|\nabla_{\overline{\mathcal{I}}} f\left(\overline{\boldsymbol{\theta}}\right)\right\|. \tag{5}$$

In particular,

(i) *if* $\mu_{\mathcal{C}_k^3} < 3$ *and* $\eta = \eta^\star = 2/\left(\alpha_{\mathcal{C}_k^3} + \beta_{\mathcal{C}_k^3}\right)$, *or*

(ii) *if* $\mu_{\mathcal{C}_k^3} < \frac{3}{2}$ *and* $\eta \in \left[1/\alpha_{\mathcal{C}_k^3}, 1/\beta_{\mathcal{C}_k^3}\right]$,

the iterates converge to $\overline{\boldsymbol{\theta}}$ *up to an approximation error bounded above by* $\frac{2\eta}{1-2\gamma}\left\|\nabla_{\overline{\mathcal{I}}} f\left(\overline{\boldsymbol{\theta}}\right)\right\|$ *with contraction factor* $2\gamma < 1$.

Proof: Applying (4) recursively under the assumptions of the corollary and using the identity $\sum_{j=0}^{i-1}(2\gamma)^j = \frac{1-(2\gamma)^i}{1-2\gamma}$ proves (5). In the first case, if $\mu_{\mathcal{C}_k^3} < 3$ and $\eta = \eta^\star = 2/\left(\alpha_{\mathcal{C}_k^3} + \beta_{\mathcal{C}_k^3}\right)$ we have $2\gamma < 1$ by definition of γ. In the second case, one can deduce from $\eta \in \left[1/\alpha_{\mathcal{C}_k^3}, 1/\beta_{\mathcal{C}_k^3}\right]$ that $|\eta/\eta^\star - 1| \leq \frac{\mu_{\mathcal{C}_k^3}-1}{2}$ and $\eta/\eta^\star \leq \frac{\mu_{\mathcal{C}_k^3}+1}{2}$ where equalities are attained simultaneously at $\eta = 1/\beta_{\mathcal{C}_k^3}$. Therefore, $\gamma \leq \mu_{\mathcal{C}_k^3} - 1 < 1/2$ and thus $2\gamma < 1$. Finally, in both cases it immediately follows from (5) that the approximation error converges to $\frac{2\eta}{1-2\gamma}\left\|\nabla_{\overline{\mathcal{I}}} f\left(\overline{\boldsymbol{\theta}}\right)\right\|$ from below as $i \to +\infty$. ■

IV. Application in Generalized Linear Models

Generalized Linear Models (GLMs) are among the most commonly used models for parametric estimation in variety of applications [24]. Linear, logistic, Poisson, and gamma models used in corresponding regression problems all belong to the family of GLMs. Given a covariate vector $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^p$ and a true parameter $\boldsymbol{\theta}^\star \in \mathbb{R}^p$, the response variable $y \in \mathcal{Y} \subseteq \mathbb{R}$ in canonical GLMs is assumed to follow an exponential family conditional distribution: $y \mid \mathbf{x}; \boldsymbol{\theta}^\star \sim Z\left(y\right)\exp\left(y\left\langle \mathbf{x}, \boldsymbol{\theta}^\star\right\rangle - \psi\left(\left\langle \mathbf{x}, \boldsymbol{\theta}^\star\right\rangle\right)\right)$, where $Z\left(y\right)$ is a positive function, and $\psi : \mathbb{R} \mapsto \mathbb{R}$ is the *log-partition function* that satisfies $\psi\left(t\right) = \log\int_{\mathcal{Y}} Z\left(y\right)\exp\left(ty\right)\mathrm{d}y$ for all $t \in \mathbb{R}$. Examples of the log-partition function include but are not limited to $\psi_{\text{lin}}\left(t\right) = t^2/2\sigma^2$, $\psi_{\log}\left(t\right) = \log\left(1+\exp\left(t\right)\right)$, and $\psi_{\text{Pois}}\left(t\right) = \exp\left(t\right)$ corresponding to linear, logistic, and Poisson models, respectively.

Suppose that n iid covariate-response pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are observed. In the Maximum Likelihood Estimation (MLE) framework the negative log likelihood is used as a measure of the discrepancy between the true parameter $\boldsymbol{\theta}^\star$ and an estimate $\boldsymbol{\theta}$ based on the observations. Formally, the average of negative log likelihoods is considered as the empirical loss

$$f\left(\boldsymbol{\theta}\right) = \frac{1}{n}\sum_{i=1}^n \psi\left(\left\langle \mathbf{x}_i, \boldsymbol{\theta}\right\rangle\right) - y_i\left\langle \mathbf{x}_i, \boldsymbol{\theta}\right\rangle,$$

and the MLE is performed by minimizing $f\left(\boldsymbol{\theta}\right)$ over the set of feasible $\boldsymbol{\theta}$. The constants c and Z that appear in the distribution are disregarded as they have no effect in the outcome.

A. Verifying SMRH for GLMs

Assuming that $\psi\left(\cdot\right)$ is twice continuously differentiable, the Hessian of $f\left(\cdot\right)$ is equal to

$$\nabla^2 f\left(\boldsymbol{\theta}\right) = \frac{1}{n}\sum_{i=1}^n \psi''\left(\left\langle \mathbf{x}_i, \boldsymbol{\theta}\right\rangle\right)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}.$$

Under the assumptions for GLMs, it can be shown that $\psi''\left(\cdot\right)$ is non-negative (i.e., $\psi\left(\cdot\right)$ is convex). For a given sparsity model generated by $\mathcal{C}_k$ let $\mathcal{S}$ be an arbitrary support set in $\mathcal{C}_k$ and suppose that $\operatorname{supp}\left(\boldsymbol{\theta}\right) \subseteq \mathcal{S}$ and $\|\boldsymbol{\theta}\| \leq r$. Furthermore, define

$$D_{\psi,r}\left(u\right) := \max_{t\in[-r,r]} \psi''\left(tu\right) \quad \text{and} \quad d_{\psi,r}\left(u\right) := \min_{t\in[-r,r]} \psi''\left(tu\right).$$

Using the Cauchy-Schwarz inequality we have $|\left\langle \mathbf{x}_i, \boldsymbol{\theta}\right\rangle| \leq r\left\|\mathbf{x}_i|_{\mathcal{S}}\right\|$ which implies

$$\frac{1}{n}\sum_{i=1}^n d_{\psi,r}\left(\left\|\mathbf{x}_i|_{\mathcal{S}}\right\|\right)\mathbf{x}_i|_{\mathcal{S}}\,\mathbf{x}_i|_{\mathcal{S}}^{\mathrm{T}} \preccurlyeq \nabla^2_{\mathcal{S}} f\left(\boldsymbol{\theta}\right) \preccurlyeq \frac{1}{n}\sum_{i=1}^n D_{\psi,r}\left(\left\|\mathbf{x}_i|_{\mathcal{S}}\right\|\right)\mathbf{x}_i|_{\mathcal{S}}\,\mathbf{x}_i|_{\mathcal{S}}^{\mathrm{T}}.$$

These matrix inequalities are precursors of (3). Imposing further restriction on the distribution of the covariate vectors $\{\mathbf{x}_i\}_{i=1}^n$ allows application of the results from random matrix theory regarding the extreme eigenvalues of random matrices (see e.g., [25] and [26]).

For example, in the logistic model where $\psi \equiv \psi_{\log}$ we can show that $D_{\psi,r}(u) = \frac{1}{4}$ and $d_{\psi,r}(u) = \frac{1}{4}\mathrm{sech}^2\left(\frac{ru}{2}\right)$. Assuming that the covariate vectors are iid instances of a random vectors whose length almost surely bounded by one, we obtain $d_{\psi,r}(u) \geq \frac{1}{4}\mathrm{sech}^2\left(\frac{r}{2}\right)$. Using the matrix Chernoff inequality [25] the extreme eigenvalues of $\frac{1}{n}\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}$ can be bounded with probability $1 - \exp(\log k - Cn)$ for some constant $C > 0$ (see [21] for detailed derivations). Using these results and taking the union bound over all $\mathcal{S} \in \mathcal{C}_k$ we obtain bounds for the extreme eigenvalues of $\nabla^2_{\mathcal{S}} f(\boldsymbol{\theta})$ that hold uniformly for all sets $\mathcal{S} \in \mathcal{C}_k$ with probability $1 - \exp(\log(k|\mathcal{C}_k|) - Cn)$. Thus (3) may hold if $n = O(\log(k|\mathcal{C}_k|))$.

B. Approximation Error for GLMs

Suppose that the approximation error is measured with respect to $\boldsymbol{\theta}^{\perp} = \mathrm{P}_{\mathcal{C}_k,r}[\boldsymbol{\theta}^{\star}]$ where $\boldsymbol{\theta}^{\star}$ is the statistical truth in the considered GLM. It is desirable to further simplify the approximation error bound provided in Corollary 1 which is related to the statistical precision of the estimation problem. The corollary provides an approximation error that is proportional to $\left\|\nabla_{\mathcal{T}} f\left(\boldsymbol{\theta}^{\perp}\right)\right\|$ where $\mathcal{T} = \mathrm{supp}\left(\mathrm{P}_{\mathcal{C}_k^2,r}\left[\nabla f\left(\boldsymbol{\theta}^{\perp}\right)\right]\right)$. We can write

$$\nabla_{\mathcal{T}} f\left(\boldsymbol{\theta}^{\perp}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\psi'\left(\left\langle \mathbf{x}_i, \boldsymbol{\theta}^{\perp}\right\rangle\right) - y_i\right)\mathbf{x}_i|_{\mathcal{T}},$$

which yields $\left\|\nabla_{\mathcal{T}} f\left(\boldsymbol{\theta}^{\perp}\right)\right\| = \|\mathbf{X}_{\mathcal{T}}\mathbf{z}\|$ where $\mathbf{X} = \frac{1}{\sqrt{n}}\left[\begin{array}{cccc}\mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n\end{array}\right]$ and $\mathbf{z}|_{\{i\}} = z_i = \frac{\psi'\left(\left\langle \mathbf{x}_i, \boldsymbol{\theta}^{\perp}\right\rangle\right) - y_i}{\sqrt{n}}$. Therefore,

$$\left\|\nabla_{\mathcal{T}} f\left(\boldsymbol{\theta}^{\perp}\right)\right\|^2 \leq \|\mathbf{X}_{\mathcal{T}}\|_{\mathrm{op}}^2 \|\mathbf{z}\|^2,$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm. Again using random matrix theory one can find an upper bound for $\|\mathbf{X}_{\mathcal{I}}\|_{\mathrm{op}}$ that holds uniformly for any $\mathcal{I} \in \mathcal{C}_k^2$ and in particular for $\mathcal{I} = \mathcal{T}$. Henceforth, $W > 0$ is used to denote this upper bound.

The second term in the bound can be written as

$$\|\mathbf{z}\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\psi'\left(\left\langle \mathbf{x}_i, \boldsymbol{\theta}^{\perp}\right\rangle\right) - y_i\right)^2.$$

To further simplify this term we need to make assumptions about the log-partition function $\psi(\cdot)$ and/or the distribution of the covariate-response pair $(\mathbf{x}, y)$. For instance, if $\psi'(\cdot)$ and the response variable y are bounded, as in the logistic model, then Hoeffding's inequality implies that for some small $\epsilon > 0$ we have $\|\mathbf{z}\|^2 \leq \mathbb{E}\left[\left(\psi'\left(\left\langle \mathbf{x}, \boldsymbol{\theta}^{\perp}\right\rangle\right) - y\right)^2\right] + \epsilon$ with probability at least $1 - \exp\left(-O\left(\epsilon^2 n^2\right)\right)$. Since in GLMs the true parameter $\boldsymbol{\theta}^{\star}$ is the minimizer of the expected loss $\mathbb{E}\left[\psi\left(\langle \mathbf{x}, \boldsymbol{\theta}\rangle\right) - y\langle \mathbf{x}, \boldsymbol{\theta}\rangle \mid \mathbf{x}\right]$ we deduce that $\mathbb{E}\left[\psi'\left(\langle \mathbf{x}, \boldsymbol{\theta}^{\star}\rangle\right) - y \mid \mathbf{x}\right] = 0$ and hence $\mathbb{E}\left[\psi'\left(\langle \mathbf{x}, \boldsymbol{\theta}^{\star}\rangle\right) - y\right] = 0$. Therefore,

$$\begin{aligned}\|\mathbf{z}\|^2 &\leq \mathbb{E}\left[\mathbb{E}\left[\left(\psi'\left(\left\langle \mathbf{x}, \boldsymbol{\theta}^{\perp}\right\rangle\right) - \psi'\left(\langle \mathbf{x}, \boldsymbol{\theta}^{\star}\rangle\right) \quad + \psi'\left(\langle \mathbf{x}, \boldsymbol{\theta}^{\star}\rangle\right) - y\right)^2 \mid \mathbf{x}\right]\right] + \epsilon \\ &\leq \underbrace{\mathbb{E}\left[\left(\psi'\left(\left\langle \mathbf{x}, \boldsymbol{\theta}^{\perp}\right\rangle\right) - \psi'\left(\langle \mathbf{x}, \boldsymbol{\theta}^{\star}\rangle\right)\right)^2\right]}_{\delta_1} + \underbrace{\mathbb{E}\left[\left(\psi'\left(\langle \mathbf{x}, \boldsymbol{\theta}^{\star}\rangle\right) - y\right)^2\right]}_{\sigma^2_{\mathrm{stat}}} + \epsilon.\end{aligned}$$

Then it follows from Corollary 1 and the fact that $\|\mathbf{X}|_{\mathcal{I}}\|_{\mathrm{op}} \leq W$ that

$$\begin{aligned}\left\|\boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}^{\star}\right\| &\leq \left\|\boldsymbol{\theta}^{(i)} - \boldsymbol{\theta}^{\perp}\right\| + \underbrace{\left\|\boldsymbol{\theta}^{\perp} - \boldsymbol{\theta}^{\star}\right\|}_{\delta_2} \\ &\leq (2\gamma)^i \left\|\boldsymbol{\theta}^{\perp}\right\| + \frac{2\eta W}{1-2\gamma}\sigma^2_{\mathrm{stat}} + \frac{2\eta W}{1-2\gamma}\delta_1 + \delta_2.\end{aligned}$$

Note that the total approximation error is comprised of two parts. The first part is due to statistical error that is given by $\frac{2\eta W}{1-2\gamma}\sigma^2_{\mathrm{stat}}$, and $\frac{2\eta W}{1-2\gamma}\delta_1 + \delta_2$ is the second part of the error due to the bias that occurs because of an infeasible true parameter. The bias vanishes if the true parameter lies in the considered bounded sparsity model (i.e., $\boldsymbol{\theta}^{\star} = \mathrm{P}_{\mathcal{C}_k,r}[\boldsymbol{\theta}^{\star}]$).

V. Conclusion

We studied the projected gradient descent method for minimization of a real valued cost function defined over a finite-dimensional Hilbert space, under structured sparsity constraints. Using previously known combinatorial sparsity models, we define a sufficient condition for accuracy of the algorithm, the SMRH. Under this condition the algorithm converges to the desired optimum at a linear rate up to an approximation error. Unlike the previous results on greedy-type methods that merely have focused on linear statistical models, our algorithm applies to a broader family of estimation problems. To provide an example, we examined application of the algorithm in estimation with GLMs. One can verify the SMRH for a specific statistical model. The approximation error can also be bounded by statistical precision and the potential bias. An interesting follow-up problem is to find whether the approximation error can be improved and the derived error is merely a by-product of requiring some form of restricted strong convexity through SMRH. Another problem of interest is to study the properties of the algorithm when the domain of the cost function is not finite-dimensional.

Appendix
Proofs

Lemma 1. *Suppose that f is a twice differentiable function that satisfies (3) for a given θ and all Δ such that* $\operatorname{supp}(\Delta) \cup \operatorname{supp}(\theta) \in \mathcal{M}(\mathcal{C}_k)$. *Then we have*

$$\left|\langle \mathbf{u}, \mathbf{v}\rangle - \eta \left\langle \mathbf{u}, \nabla^2 f(\theta)\, \mathbf{v}\right\rangle\right| \leq \left(\eta \frac{\alpha_{\mathcal{C}_k} - \beta_{\mathcal{C}_k}}{2} + \left|\eta \frac{\alpha_{\mathcal{C}_k} + \beta_{\mathcal{C}_k}}{2} - 1\right|\right) \|\mathbf{u}\|\, \|\mathbf{v}\|,$$

for all $\eta > 0$ and $\mathbf{u}, \mathbf{v} \in \mathcal{H}$ such that $\operatorname{supp}(\mathbf{u} \pm \mathbf{v}) \cup \operatorname{supp}(\theta) \in \mathcal{M}(\mathcal{C}_k)$.

Proof: We first the prove the lemma for unit-norm vectors $\mathbf{u}$ and $\mathbf{v}$. Since $\operatorname{supp}(\mathbf{u} \pm \mathbf{v}) \cup \operatorname{supp}(\theta) \in \mathcal{M}(\mathcal{C}_k)$ we can use (3) for $\Delta = \mathbf{u} \pm \mathbf{v}$ to obtain

$$\beta_{\mathcal{C}_k} \|\mathbf{u} \pm \mathbf{v}\|^2 \leq \left\langle \mathbf{u} \pm \mathbf{v}, \nabla^2 f(\theta)\, (\mathbf{u} \pm \mathbf{v})\right\rangle \leq \alpha_{\mathcal{C}_k} \|\mathbf{u} \pm \mathbf{v}\|^2.$$

These inequalities and the assumption $\|\mathbf{u}\| = \|\mathbf{v}\| = 1$ then yield

$$\frac{\beta_{\mathcal{C}_k} - \alpha_{\mathcal{C}_k}}{2} + \frac{\alpha_{\mathcal{C}_k} + \beta_{\mathcal{C}_k}}{2} \langle \mathbf{u}, \mathbf{v}\rangle \leq \left\langle \mathbf{u}, \nabla^2 f(\theta)\, \mathbf{v}\right\rangle \leq \frac{\alpha_{\mathcal{C}_k} - \beta_{\mathcal{C}_k}}{2} + \frac{\alpha_{\mathcal{C}_k} + \beta_{\mathcal{C}_k}}{2} \langle \mathbf{u}, \mathbf{v}\rangle,$$

where we used the fact that $\nabla^2 f(\theta)$ is symmetric since f is twice continuously differentiable. Multiplying all sides by η and rearranging the terms then imply

$$\begin{aligned}
\eta \frac{\alpha_{\mathcal{C}_k} - \beta_{\mathcal{C}_k}}{2} &\geq \left|\left(\eta \frac{\alpha_{\mathcal{C}_k} + \beta_{\mathcal{C}_k}}{2} - 1\right) \langle \mathbf{u}, \mathbf{v}\rangle + \langle \mathbf{u}, \mathbf{v}\rangle - \eta \left\langle \mathbf{u}, \nabla^2 f(\theta)\, \mathbf{v}\right\rangle\right| \\
&\geq \left|\langle \mathbf{u}, \mathbf{v}\rangle - \eta \left\langle \mathbf{u}, \nabla^2 f(\theta)\, \mathbf{v}\right\rangle\right| - \left|\left(\eta \frac{\alpha_{\mathcal{C}_k} + \beta_{\mathcal{C}_k}}{2} - 1\right) \langle \mathbf{u}, \mathbf{v}\rangle\right| \\
&\geq \left|\langle \mathbf{u}, \mathbf{v}\rangle - \eta \left\langle \mathbf{u}, \nabla^2 f(\theta)\, \mathbf{v}\right\rangle\right| - \left|\eta \frac{\alpha_{\mathcal{C}_k} + \beta_{\mathcal{C}_k}}{2} - 1\right|, \qquad (6)
\end{aligned}$$

which is equivalent to result for unit-norm $\mathbf{u}$ and $\mathbf{v}$ as desired. For the general case one can write $\mathbf{u} = \|\mathbf{u}\|\, \mathbf{u}'$ and $\mathbf{v} = \|\mathbf{v}\|\, \mathbf{v}'$ such that $\mathbf{u}'$ and $\mathbf{v}'$ are both unit-norm. It is straightforward to verify that using (6) for $\mathbf{u}'$ and $\mathbf{v}'$ as the unit-norm vectors and multiplying both sides of the resulting inequality by $\|\mathbf{u}\|\, \|\mathbf{v}\|$ yields the desired general case. ∎

Proof of Theorem 1: Using optimality of $\theta^{(i+1)}$ and feasibility of $\overline{\theta}$ one can deduce $\left\|\theta^{(i+1)} - \chi^{(i)}\right\|^2 \leq \left\|\overline{\theta} - \chi^{(i)}\right\|^2$, with $\chi^{(i)}$ as in line 2 of Alg. 1. Expanding the squared norms using the inner product of $\mathcal{H}$ then shows $0 \leq \left\langle \theta^{(i+1)} - \overline{\theta}, 2\chi^{(i)} - \theta^{(i+1)} - \overline{\theta}\right\rangle$ or equivalently $0 \leq \left\langle \Delta^{(i+1)}, 2\theta^{(i)} - 2\eta^{(i)} \nabla f\left(\overline{\theta} + \Delta^{(i)}\right) - \Delta^{(i+1)}\right\rangle$, where $\Delta^{(i)} = \theta^{(i)} - \overline{\theta}$ and $\Delta^{(i+1)} = \theta^{(i+1)} - \overline{\theta}$. Adding and subtracting $2\eta^{(i)} \left\langle \Delta^{(i+1)}, \nabla f\left(\overline{\theta}\right)\right\rangle$ and rearranging yields

$$\begin{aligned}
\left\|\Delta^{(i+1)}\right\|^2 \leq\ & 2\left\langle \Delta^{(i+1)}, \theta^{(i)}\right\rangle - 2\eta^{(i)} \left\langle \Delta^{(i+1)}, \nabla f\left(\overline{\theta} + \Delta^{(i)}\right) - \nabla f\left(\overline{\theta}\right)\right\rangle \\
& - 2\eta^{(i)} \left\langle \Delta^{(i+1)}, \nabla f\left(\overline{\theta}\right)\right\rangle \qquad (7)
\end{aligned}$$

Since f is twice continuously differentiable by assumption, it follows form the mean-value theorem that $\left\langle \Delta^{(i+1)}, \nabla f\left(\overline{\theta} + \Delta^{(i)}\right) - \nabla f\left(\overline{\theta}\right)\right\rangle = \left\langle \Delta^{(i+1)}, \nabla^2 f\left(\overline{\theta} + t\Delta^{(i)}\right) \Delta^{(i)}\right\rangle$, for some $t \in (0, 1)$. Furthermore, because $\overline{\theta}$, $\theta^{(i)}$, $\theta^{(i+1)}$ all belong to the model set $\mathcal{M}(\mathcal{C}_k)$ we have $\operatorname{supp}\left(\overline{\theta} + t\Delta^{(i)}\right) \in \mathcal{M}\left(\mathcal{C}_k^2\right)$ and thereby $\operatorname{supp}\left(\Delta^{(i+1)}\right) \cup$

$\operatorname{supp}\left(\overline{\theta}+t\boldsymbol{\Delta}^{(i)}\right) \in \mathcal{M}\left(\mathcal{C}_k^3\right)$. Invoking the $\left(\mu_{\mathcal{C}_k^3}, r\right)$-SMRH condition of the cost function and applying Lemma 1 with the sparsity model $\mathcal{M}\left(\mathcal{C}_k^3\right)$, $\theta = \overline{\theta}+t\boldsymbol{\Delta}^{(i)}$, and $\eta = \eta^{(i)}$ then yields

$$\left|\left\langle \boldsymbol{\Delta}^{(i+1)}, \boldsymbol{\Delta}^{(i)}\right\rangle - \eta^{(i)}\left\langle \boldsymbol{\Delta}^{(i+1)}, \nabla f\left(\overline{\theta}+\boldsymbol{\Delta}^{(i)}\right) - \nabla f\left(\overline{\theta}\right)\right\rangle\right| \leq \gamma^{(i)}\left\|\boldsymbol{\Delta}^{(i+1)}\right\|\left\|\boldsymbol{\Delta}^{(i)}\right\|.$$

Using the Cauchy-Schwarz inequality and the fact that $\left\|\nabla_{\operatorname{supp}\left(\boldsymbol{\Delta}^{(i+1)}\right)} f\left(\overline{\theta}\right)\right\| \leq \left\|\nabla_{\overline{\mathcal{I}}} f\left(\overline{\theta}\right)\right\|$ by the definition of $\overline{\mathcal{I}}$, (7) implies that $\left\|\boldsymbol{\Delta}^{(i+1)}\right\|^2 \leq 2\gamma^{(i)}\left\|\boldsymbol{\Delta}^{(i+1)}\right\|\left\|\boldsymbol{\Delta}^{(i)}\right\| + 2\eta^{(i)}\left\|\boldsymbol{\Delta}^{(i+1)}\right\|\left\|\nabla_{\overline{\mathcal{I}}} f\left(\overline{\theta}\right)\right\|$. Canceling $\left\|\boldsymbol{\Delta}^{(i+1)}\right\|$ from both sides proves the theorem. ■

Lemma 2 (Bounded Model Projection)**.** *Given an arbitrary $\mathbf{h}_0 \in \mathcal{H}$, a positive real number r, and a sparsity model generator $\mathcal{C}_k$, a projection $\mathrm{P}_{\mathcal{C}_k,r}\left[\mathbf{h}_0\right]$ can be obtained as the projection of $\mathrm{P}_{\mathcal{C}_k,+\infty}\left[\mathbf{h}_0\right]$ on to the sphere of radius r.*

Proof: To simplify the notation let $\widehat{\mathbf{h}} = \mathrm{P}_{\mathcal{C}_k,r}\left[\mathbf{h}_0\right]$ and $\widehat{\mathcal{S}} = \operatorname{supp}\left(\widehat{\mathbf{h}}\right)$. For $\mathcal{S} \subseteq [p]$ define

$$\mathbf{h}_0\left(\mathcal{S}\right) = \arg\min_{\mathbf{h}} \ \|\mathbf{h}-\mathbf{h}_0\| \quad \text{s.t. } \|\mathbf{h}\| \leq r \text{ and } \operatorname{supp}\left(\mathbf{h}\right) \subseteq \mathcal{S}.$$

It follows from the definition of $\mathrm{P}_{\mathcal{C}_k,r}\left[\mathbf{h}_0\right]$ that $\widehat{\mathcal{S}} \in \arg\min_{\mathcal{S}\in\mathcal{C}_k} \ \|\mathbf{h}_0\left(\mathcal{S}\right)-\mathbf{h}_0\|$. Using

$$\|\mathbf{h}_0\left(\mathcal{S}\right)-\mathbf{h}_0\|^2 = \|\mathbf{h}_0\left(\mathcal{S}\right)-\mathbf{h}_0|_{\mathcal{S}}-\mathbf{h}_0|_{\mathcal{S}^c}\|^2 = \|\mathbf{h}_0\left(\mathcal{S}\right)-\mathbf{h}_0|_{\mathcal{S}}\|^2 + \|\mathbf{h}_0|_{\mathcal{S}^c}\|^2,$$

we deduce that $\mathbf{h}_0\left(\mathcal{S}\right)$ is the projection of $\mathbf{h}_0|_{\mathcal{S}}$ onto the sphere of radius r. Therefore, we can write $\mathbf{h}_0\left(\mathcal{S}\right) = \min\left\{1, r/\|\mathbf{h}_0|_{\mathcal{S}}\|\right\}\mathbf{h}_0|_{\mathcal{S}}$ and from that

$$\begin{aligned}
\widehat{\mathcal{S}} &\in \arg\min_{\mathcal{S}\in\mathcal{C}_k} \ \|\min\left\{1, r/\|\mathbf{h}_0|_{\mathcal{S}}\|\right\}\mathbf{h}_0|_{\mathcal{S}}-\mathbf{h}_0\|^2 \\
&= \arg\min_{\mathcal{S}\in\mathcal{C}_k} \ \|\min\left\{0, r/\|\mathbf{h}_0|_{\mathcal{S}}\|-1\right\}\mathbf{h}_0|_{\mathcal{S}}\|^2 + \|\mathbf{h}_0|_{\mathcal{S}^c}\|^2 \\
&= \arg\min_{\mathcal{S}\in\mathcal{C}_k} \ \left((1-r/\|\mathbf{h}_0|_{\mathcal{S}}\|)_+^2 - 1\right)\|\mathbf{h}_0|_{\mathcal{S}}\|^2 \\
&= \arg\max_{\mathcal{S}\in\mathcal{C}_k} \ q\left(\mathcal{S}\right) := \|\mathbf{h}_0|_{\mathcal{S}}\|^2 - \left(\|\mathbf{h}_0|_{\mathcal{S}}\|-r\right)_+^2.
\end{aligned}$$

Furthermore, let

$$\mathcal{S}_0 = \operatorname{supp}\left(\mathrm{P}_{\mathcal{C}_k,+\infty}\left[\mathbf{h}_0\right]\right) = \arg\max_{\mathcal{S}\in\mathcal{C}_k} \ \|\mathbf{h}_0|_{\mathcal{S}}\|. \tag{8}$$

If $\left\|\mathbf{h}_0|_{\mathcal{S}_0}\right\| \leq r$ then $q\left(\mathcal{S}\right) = \|\mathbf{h}_0|_{\mathcal{S}}\| \leq q\left(\mathcal{S}_0\right)$ for any $\mathcal{S} \in \mathcal{C}_k$ and thereby $\widehat{\mathcal{S}} = \mathcal{S}_0$. Thus, we focus on cases that $\left\|\mathbf{h}_0|_{\mathcal{S}_0}\right\| > r$ which implies $q\left(\mathcal{S}_0\right) = 2\left\|\mathbf{h}_0|_{\mathcal{S}_0}\right\| r - r^2$. For any $\mathcal{S} \in \mathcal{C}_k$ if $\|\mathbf{h}_0|_{\mathcal{S}}\| \leq r$ we have $q\left(\mathcal{S}\right) = \|\mathbf{h}_0|_{\mathcal{S}}\|^2 \leq r^2 < 2\left\|\mathbf{h}_0|_{\mathcal{S}_0}\right\| r - r^2 = q\left(\mathcal{S}_0\right)$, and if $\|\mathbf{h}_0|_{\mathcal{S}}\| > r$ we have $q\left(\mathcal{S}\right) = 2\|\mathbf{h}_0|_{\mathcal{S}}\| r - r^2 \leq 2\left\|\mathbf{h}_0|_{\mathcal{S}_0}\right\| r - r^2 = q\left(\mathcal{S}_0\right)$ where (8) is applied. Therefore, we have shown that $\widehat{\mathcal{S}} = \mathcal{S}_0$. It is then straightforward to show the desired result that projecting $\mathrm{P}_{\mathcal{C}_k,+\infty}\left[\mathbf{h}_0\right]$ onto the centered sphere of radius r yields $\mathrm{P}_{\mathcal{C}_k,r}\left[\mathbf{h}_0\right]$. ■

REFERENCES

[1] F. R. Bach, "Consistency of the group lasso and multiple kernel learning," *Journal of Machine Learning Research*, vol. 9, pp. 1179–1225, Jun. 2008.

[2] V. Roth and B. Fischer, "The Group-Lasso for generalized linear models: uniqueness of solutions and efficient algorithms," in *Proceedings of the 25th International Conference on Machine learning*, ser. ICML '08, New York, NY, USA, 2008, pp. 848–855.

[3] L. Jacob, G. Obozinski, and J. Vert, "Group lasso with overlap and graph lasso," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09, New York, NY, USA, 2009, pp. 433–440.

[4] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[5] F. Bach, "Structured sparsity-inducing norms through submodular functions," in *Advances in Neural Information Processing Systems*, vol. 23, Vancouver, BC, Canada, Dec. 2010, pp. 118–126.

[6] R. Jenatton, J. Audibert, and F. Bach, "Structured variable selection with Sparsity-Inducing norms," *Journal of Machine Learning Research*, vol. 12, pp. 2777–2824, Oct. 2011.

[7] A. Kyrillidis and V. Cevher, "Combinatorial selection and least absolute shrinkage via the CLASH algorithm," Mar. 2012, preprint available at arXiv:1203.2936 [cs.IT].

[8] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Journal of Foundation of Computational Mathematics*, , to appear. Preprint available at arXiv:1012.0621.

[9] A. Kyrillidis and V. Cevher, "Sublinear time, approximate model-based sparse recovery for all," Mar. 2012, preprint available at arXiv:1203.4746 [cs.IT].

[10] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Structured sparsity through convex optimization," Sep. 2011, preprint available at arXiv:1109.2397.

[11] M. Duarte and Y. Eldar, "Structured compressed sensing: From theory to applications," *Signal Processing, IEEE Transactions on*, vol. 59, no. 9, pp. 4053–4085, Sep. 2011.

[12] A. Tewari, P. K. Ravikumar, and I. S. Dhillon, "Greedy algorithms for structurally constrained high dimensional problems," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds., 2011, pp. 882–890.

[13] T. Blumensath, "Compressed sensing with nonlinear observations," 2010, preprint. [Online]. Available: http://eprints.soton.ac.uk/164753/1/B_Nonlinear.pdf

[14] A. Lozano, G. Swirszcz, and N. Abe, "Group orthogonal matching pursuit for logistic regression," in *the Fourteenth International Conference on Artificial Intelligence and Statistics*, G. Gordon, D. Dunson, and M. Dudik, Eds., vol. 15. Ft. Lauderdale, FL, USA: JMLR W&CP, 2011, pp. 452–460.

[15] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Conference Record of the 27th Asilomar Conf. on Signals, Systems and Computers*, vol. 1, Pacific Grove, CA, Nov. 1993, pp. 40–44.

[16] A. Beck and Y. C. Eldar, "Sparsity constrained nonlinear optimization: Optimality conditions and algorithms," Mar. 2012, preprint available at arXiv:1203.4580.

[17] F. Bunea, "Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization," *Electronic Journal of Statistics*, vol. 2, pp. 1153–1194, 2008.

[18] S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," in *Advances in Neural Information Processing Systems*, vol. 22, Vancouver, BC, Canada, Dec. 2009, pp. 1348–1356, long version available at arXiv:1010.2731v1 [math.ST].

[19] A. Agarwal, S. Negahban, and M. Wainwright, "Fast global convergence rates of gradient methods for high-dimensional statistical recovery," in *Advances in Neural Information Processing Systems*, vol. 23, Vancouver, BC, Canada, 2010, pp. 37–45, long version available at arXiv:1104.4824v1 [stat.ML].

[20] S. M. Kakade, O. Shamir, K. Sridharan, and A. Tewari, "Learning exponential families in High-Dimensions: strong convexity and sparsity," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, ser. JMLR Workshop and Conference Proceedings, vol. 9, Sardinia, Italy, 2010, pp. 381–388.

[21] S. Bahmani, B. Raj, and P. Boufounos, "Greedy sparsity-constrained optimization," 2012, preprint available at arxiv:1203.5483v1 [stat.ML].

[22] S. A. van de Geer, *Empirical processes in M-estimation*. Cambridge, UK: Cambridge University Press, 2000.

[23] V. Vapnik, *Statistical learning theory*. New York, NY: Wiley, 1998.

[24] A. J. Dobson and A. Barnett, *An Introduction to Generalized Linear Models*, 3rd ed. Boca Reaton, FL: Chapman and Hall/CRC, May 2008.

[25] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of Computational Mathematics*, Aug. 2011.

[26] D. Hsu, S. Kakade, and T. Zhang, "Tail inequalities for sums of random matrices that depend on the intrinsic dimension," *Electron. Commun. Probab.*, vol. 17, no. 14, pp. 1–13, 2012.