$\ell_{1,p}$-Norm Regularization: Error Bounds and Convergence Rate Analysis of First-Order Methods

Zirui Zhou ZRZHOU@SE.CUHK.EDU.HK
Qi Zhang QZHANG@SE.CUHK.EDU.HK
Anthony Man-Cho So MANCHOSO@SE.CUHK.EDU.HK
Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong S.A.R., China

Abstract

In recent years, the $\ell_{1,p}$-regularizer has been widely used to induce structured sparsity in the solutions to various optimization problems. Currently, such $\ell_{1,p}$-regularized problems are typically solved by first-order methods. Motivated by the desire to analyze the convergence rates of these methods, we show that for a large class of $\ell_{1,p}$-regularized problems, an error bound condition is satisfied when $p \in [1,2]$ or $p = \infty$ but fails to hold for any $p \in (2,\infty)$. Based on this result, we show that many first-order methods enjoy an asymptotic linear rate of convergence when applied to $\ell_{1,p}$-regularized linear or logistic regression with $p \in [1,2]$ or $p = \infty$. By contrast, numerical experiments suggest that for the same class of problems with $p \in (2,\infty)$, the aforementioned methods may not converge linearly.

1. Introduction

Optimization with sparsity-inducing penalties has received increasing attention in various application domains such as machine learning, statistics, computational biology, and signal processing (Bach et al., 2012). As the convex envelope of ℓ_0-norm, the ℓ_1-norm has been widely used as a regularizer in sparse variable selection, such as LASSO (Tibshirani, 1996). Recently, ℓ_1-regularization has been extended to Group-Lasso regularization (Yuan & Lin, 2006; Bach, 2008; Meier et al., 2008), and more generally, to $\ell_{1,p}$-regularization with $1 \leq p \leq \infty$ (Fornasier & Rauhut, 2008; Kowalski, 2009; Vogt & Roth, 2012). Such extensions have been applied to sparse regression (Eldar et al., 2010), multiple kernel learning (Tomioka & Suzuki, 2010; Kloft et al., 2011), *etc.*, and have witnessed great success in yielding sparsity on the group level when $p > 1$. In these applications, one is interested in solving a convex optimization problem of the form

$$\min_{x \in \mathbb{R}^n} F(x) := f(x) + P(x). \tag{1}$$

Here, $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth convex function and $P : \mathbb{R}^n \to \mathbb{R}$ takes the form

$$P(x) = \sum_{J \in \mathcal{J}} \omega_J \|x_J\|_p,$$

where $\mathcal{J}$ is a non-overlapping partition of the coordinate index set $\{1, 2, \ldots, n\}$, $\omega_J > 0$ for each $J \in \mathcal{J}$, and $\|\cdot\|_p$ is the ℓ_p-norm. Note that ℓ_1-regularization and Group-Lasso regularization are special cases of (1), as they correspond to $p = 1$ and $p = 2$, respectively. Additionally, ℓ_p-regularization is also incorporated when no partition is made.

To cope with the rapidly growing size of datasets, recent researches on numerical algorithms for solving nonsmooth composite minimization problems such as (1) have chiefly been focusing on first-order methods, such as the proximal gradient method and its accelerated version (Beck & Teboulle, 2009), the coordinate descent method (Tseng, 2001), and the coordinate gradient descent method (Tseng & Yun, 2009). Since then, adaptations of these methods to the $\ell_{1,p}$-regularized problem (1) have been proposed in (Meier et al., 2008; Liu et al., 2009; Liu & Ye, 2010). To study the efficiency of these iterative algorithms, one approach is to analyze the rates at which the iterates generated by the algorithms converge to an optimal solution. Existing results in this line of research reveal that for smooth convex functions f, the aforementioned first-order methods for solving the $\ell_{1,p}$-regularized problem (1) converge at a sublinear rate, and a linear rate is achievable when f is additionally assumed to be strongly convex (Nesterov, 2004; Meier et al., 2008; Liu & Ye, 2010). However, for many applications, the

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

strong convexity assumption is too stringent. Moreover, various first-order methods for solving (1) have exhibited a linear rate of convergence in numerical experiments even when f is not strongly convex—a case in point is the proximal gradient method for solving ℓ_1-regularized linear regression problems (Hale et al., 2008; Xiao & Zhang, 2013). It is thus natural to ask whether such a phenomenon can be explained theoretically, and more generally, whether certain structures of the functions f and P can be exploited to establish faster convergence rates for the aforementioned first-order methods.

To address these questions, a powerful approach is to utilize a so-called error bound (EB) condition (Definition 1), which can be viewed as a relaxed notion of strong convexity. Indeed, assuming the EB condition holds, various first-order algorithms have been demonstrated to achieve a linear rate of convergence (Luo & Tseng, 1993; Hong & Luo, 2012; So, 2013; Wang & Lin, 2014). Moreover, it has been shown that the EB condition is satisfied by a number of optimization problems for which strong convexity fails to hold, such as linear regression with ℓ_1-regularizer (Luo & Tseng, 1992). However, verifying whether a given optimization problem satisfies the EB condition remains an intriguing issue.

In this paper, we consider the $\ell_{1,p}$-regularized problem (1) with $p \in [1,\infty]$ and study when the EB condition holds for this problem. Previous researches show that under some mild assumptions on the function f (which are satisfied by many machine learning applications), the EB condition holds when $p \in \{1,2,\infty\}$ (Luo & Tseng, 1992; Tseng, 2010; Zhang et al., 2013). However, to the best of our knowledge, it is not known whether the same is true for other values of p. In fact, this question does not seem to be amenable to the techniques developed in (Luo & Tseng, 1992; Tseng, 2010), as they require either the non-smooth function P to have a polyhedral epigraph, which merely corresponds to $p = 1$ and $p = \infty$, or an explicit expression of the residual function, which is only available when $p = 1$ and $p = 2$.

The contribution of this paper is twofold. First, by exploiting the notion of upper Lipschitz continuity of set-valued mappings, we establish a sufficient condition under which the EB condition holds for problem (1). In fact, our condition only requires the function P to be convex and thus can potentially be used to certify the EB condition for a wide range of regularizations. Second, based on our newly developed sufficient condition, we completely determine the values of p for which the $\ell_{1,p}$-regularized problem (1) satisfies the EB condition. Specifically, we show that under standard assumptions on the smooth convex function f (see Assumption 1), the EB condition holds when $p \in [1,2]$ and $p = \infty$. On the other hand, we show via a family of examples that without further assumptions, the EB condition can fail for any $p \in (2,\infty)$.

As a direct consequence of our results, we show that many first-order methods, including the proximal gradient algorithm and coordinate gradient descent method, enjoy an asymptotic linear rate of convergence when applied to $\ell_{1,p}$-regularized linear or logistic regression with $p \in [1,2]$ or $p = \infty$. By contrast, for the same class of problems with $p \in (2,\infty)$, our numerical results suggest that these methods may not converge linearly. Our results not only expand the repertoire of optimization problems that are known to satisfy the EB condition but also explain how the choice of p could affect the convergence rates of first-order methods.

In the sequel, we shall adopt the following notations. For any vector $x \in \mathbb{R}^n$, $x_J \in \mathbb{R}^{|J|}$ denotes the restriction of x onto the coordinate index set $J \subseteq \{1,\dots,n\}$; $\|x\|_p$, where $p \in [1,\infty]$, denotes the ℓ_p-norm of x. For simplicity, we write $\|x\|$ for $\|x\|_2$. For any matrix $B \in \mathbb{R}^{m\times n}$, $\|B\|$ is the matrix norm of B induced by the ℓ_2-norm; *i.e.*, $\|B\| = \max_{\|v\|=1} \|Bv\|$. For any scalar $a \in \mathbb{R}$, $\mathrm{sgn}(a)$ is the sign of a; *i.e.*, $\mathrm{sgn}(a) = 1$ if $a > 0$, $\mathrm{sgn}(a) = 0$ if $a = 0$, and $\mathrm{sgn}(a) = -1$ if $a < 0$. For any closed set $\mathcal{S}$, $d(x,\mathcal{S})$ is the distance of x to $\mathcal{S}$; *i.e.*, $d(x,\mathcal{S}) = \min_{v\in\mathcal{S}} \|v - x\|$.

2. Preliminaries

2.1. Basic Setup

Throughout the paper, we make the following assumptions regarding the $\ell_{1,p}$-regularized problem (1):

Assumption 1 *(a) The convex function f is of the form*

$$f(x) = h(Ax), \tag{2}$$

where $A \in \mathbb{R}^{m\times n}$ is a matrix and $h : \mathbb{R}^m \to \mathbb{R}$ is a continuously differentiable function with Lipschitz continuous gradient ∇h and is strongly convex over any compact subset of the effective domain $\mathrm{dom}(h)$ of h.

(b) The optimal solution set of (1), denoted by $\mathcal{X}$, is non-empty; i.e., $\mathcal{X} \neq \emptyset$.

The above assumption is satisfied by many optimization problems arising in machine learning. For instance, in linear models, the empirical risk takes the form $f(x) = \frac{1}{N}\sum_{i=1}^N \ell(\hat{y}^{(i)}, x^T\hat{z}^{(i)})$, where $\{(\hat{z}^{(i)},\hat{y}^{(i)}) \in \mathbb{R}^n \times \mathbb{R}^p \,|\, i = 1,\dots,N\}$ are sample points and $\ell : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a loss function. Such an f can be put into the form (2) by letting $A = [\hat{z}^{(1)},\dots,\hat{z}^{(N)}]^T$ and $h(y) = \frac{1}{N}\sum_{i=1}^N \ell(\hat{y}^{(i)}, y^{(i)})$ with $y = (y^{(1)},\dots,y^{(N)})$. Two commonly used loss functions are the square loss $\ell(\hat{y}^{(i)}, y^{(i)}) = \frac{1}{2}\|\hat{y}^{(i)} - y^{(i)}\|^2$ and the logistic loss $\ell(\hat{y}^{(i)}, y^{(i)}) = \sum_{j=1}^p \log(1 + \exp(-\hat{y}_j^{(i)} y_j^{(i)}))$. It can be verified that linear models with either the square loss or the logistic loss satisfy Assumption 1.

Assumption 1 implies some important properties of the optimal solution set of (1), which we summarize in the following proposition. The proof is given in the supplementary material.

Proposition 1 *Under Assumption 1, the optimal solution set $\mathcal{X}$ has the following properties:*

(i) There exist a pair of vectors $(\bar{y}, \bar{g}) \in \mathbb{R}^m \times \mathbb{R}^n$ with $\bar{g} = A^T \nabla h(\bar{y})$ such that for any $x \in \mathcal{X}$,

$$Ax = \bar{y}, \quad \nabla f(x) = \bar{g}.$$

(ii) $\mathcal{X}$ is a compact convex set.

2.2. Error Bound Condition

In the convergence analysis of numerical algorithms for (1), it is essential to measure the distance of any given iterate x^k to the optimal solution set $\mathcal{X}$; *i.e.*, $d(x^k, \mathcal{X})$. However, without actually solving (1), such a quantity is not easily accessible. As an alternative, let us define a function $R : \mathbb{R}^n \to \mathbb{R}^n$, which we call the *residual function* of (1), as follows:

$$R(x) := \operatorname*{argmin}_{d \in \mathbb{R}^n} \left\{ \langle \nabla f(x), d \rangle + P(x + d) + \frac{1}{2}\|d\|^2 \right\}. \tag{3}$$

It is easy to verify that $R(x) = \mathbf{0}$ if and only if $x \in \mathcal{X}$. Moreover, given any $x \in \mathbb{R}^n$, $R(x)$ is typically much easier to compute and analyze than $d(x, \mathcal{X})$. This suggests that $\|R(x)\|$ can serve as a surrogate measure of the proximity of x to $\mathcal{X}$. However, such a surrogate measure would not be very useful unless a relationship between $\|R(x)\|$ and $d(x, \mathcal{X})$ can be established. This motivates the exploration of the following error bound (EB) condition:

Definition 1 (EB Condition) *We say that problem (1) satisfies the EB condition if there exist a constant $\kappa > 0$ and a closed set $\mathcal{U} \subseteq dom(F)$, such that*

$$d(x, \mathcal{X}) \le \kappa \|R(x)\| \quad \textit{whenever} \quad x \in \mathcal{U}. \tag{4}$$

Moreover, we say the EB condition is global *if $\mathcal{U} = dom(F)$ and is* local *if $\mathcal{U}$ is the closure of some neighborhood of the optimal solution set $\mathcal{X}$.*

The EB condition can alternatively be viewed as a relaxed notion of strong convexity, as it is automatically satisfied if F is strongly convex (Pang, 1987). For illustration, consider the simple case of (1) where $P \equiv 0$. From (3), we see that $R(x) = -\nabla f(x)$. Hence, the EB condition is asking for a constant $\kappa > 0$ such that $d(x, \mathcal{X}) \le \kappa \|\nabla f(x)\|$, which holds globally when f is strongly convex.

2.3. Set-Valued Mappings and Upper Lipschitz Continuity

Our approach to establishing the EB condition is based on the notion of upper Lipschitz continuity of set-valued mappings, which features prominently in variational analysis. Let us begin with some definitions.

Let $\mathcal{Y}$ and $\mathcal{Z}$ be two Euclidean spaces. A mapping $\Gamma : \mathcal{Y} \to \mathcal{Z}$ is said to be a *set-valued mapping*, or equivalently, a *multifunction*, if for each element of $y \in \mathcal{Y}$, $\Gamma(y)$ is a subset of $\mathcal{Z}$. For example, let $B \in \mathbb{R}^{m \times n}$ be given and consider the solution set of the following linear system:

$$\mathcal{S}(b) = \{z \in \mathbb{R}^n \mid Bz = b\}.$$

Then, $\mathcal{S}$ is a set-valued mapping from $\mathbb{R}^m$ to $\mathbb{R}^n$, because for each $b \in \mathbb{R}^m$, $\mathcal{S}(b)$ is an affine subset of $\mathbb{R}^n$. The *graph* of a set-valued mapping $\Gamma : \mathcal{Y} \to \mathcal{Z}$, denoted by $\mathrm{gph}(\Gamma)$, is the subset of $\mathcal{Y} \times \mathcal{Z}$ defined by

$$\mathrm{gph}(\Gamma) := \{(y, z) \in \mathcal{Y} \times \mathcal{Z} \mid z \in \Gamma(y)\}.$$

For set-valued mappings, we can define a notion of continuity as follows:

Definition 2 *A set-valued mapping $\Gamma : \mathcal{Y} \to \mathcal{Z}$ is said to be upper Lipschitz continuous (ULC) at $\bar{y} \in \mathcal{Y}$ if $\Gamma(\bar{y})$ is non-empty and closed, and there exist constants $\theta > 0$ and $\delta > 0$ such that for all $y \in \mathcal{Y}$ with $\|y - \bar{y}\| \le \delta$,*

$$\Gamma(y) \subseteq \Gamma(\bar{y}) + \theta \|y - \bar{y}\| \mathcal{B},$$

where $\mathcal{B} = \{z \in \mathcal{Z} \mid \|z\| \le 1\}$ is the unit ℓ_2-norm ball of $\mathcal{Z}$ and "$+$" is the Minkowski sum of two sets.

The ULC property above can be viewed as an extension of the calmness property of single-valued functions to set-valued functions (Dontchev & Rockafellar, 2009).

Before leaving this section, we present an important lemma characterizing the ULC property of polyhedral multifunctions, the proof of which can be found in (Robinson, 1981). A set-valued mapping is called a polyhedral multifunction if its graph is a finite union of polyhedral convex sets.

Lemma 1 *Let $\Gamma : \mathcal{Y} \to \mathcal{Z}$ be a polyhedral multifunction. Then, Γ is ULC at any $\bar{y} \in \mathcal{Y}$ such that $\Gamma(\bar{y})$ is non-empty.*

3. A Sufficient Condition for the EB Condition

In this section, we prove a sufficient condition for the EB condition to hold, which forms the basis of our subsequent analysis. Let $\Sigma : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$ be the set-valued mapping defined by

$$\Sigma(y, g) := \{x \in \mathbb{R}^n \mid Ax = y, \ -g \in \partial P(x)\}. \tag{5}$$

The following proposition characterizes the relationship between the set-valued mapping Σ and the optimal solution set $\mathcal{X}$:

Proposition 2 *Under Assumption 1, we have*

$$\mathcal{X} = \Sigma(\bar{y}, \bar{g}),$$

where $(\bar{y}, \bar{g}) \in \mathbb{R}^m \times \mathbb{R}^n$ *are given in Proposition 1.*

Furthermore, as shown in the following theorem, the ULC property of Σ implies the EB condition for (1).

Theorem 1 *Under Assumption 1, the EB condition holds for* (1) *if the set-valued mapping* Σ *is ULC at* $(\bar{y}, \bar{g}) \in \mathbb{R}^m \times \mathbb{R}^n$.

The proofs of Proposition 2 and Theorem 1 are presented in the supplementary material. Theorem 1 gives an alternative analysis framework for establishing the EB condition. Indeed, instead of establishing the inequality (4) directly, we may turn to study the ULC property of the set-valued mapping Σ associated with the optimization problem. This approach can be advantageous, as it only relies on the properties of the subdifferential of the non-smooth function P, which are often simpler than those of the residual function R. In what follows, we will utilize this approach to study when the EB condition holds for the $\ell_{1,p}$-regularized problem (1).

4. EB Condition for $\ell_{1,p}$-Regularization

In this section, we consider the $\ell_{1,p}$-regularized problem (1) under Assumption 1 and investigate for which values of $p \in [1, \infty]$ will the EB condition hold. In view of Theorem 1, our strategy is to study when the set-valued mapping Σ possesses the ULC property. We divide our analysis into three cases: (a) $p = 1$ and $p = \infty$; (b) $p \in (1, 2]$; (c) $p \in (2, \infty)$.

4.1. EB Condition Holds when $p = 1$ and $p = \infty$

We first state a result concerning the set-valued mapping (5) when P has a polyhedral epigraph.

Lemma 2 *Suppose that* P *has a polyhedral epigraph; i.e., the set* $\{(x, t) \in \mathbb{R}^n \times \mathbb{R} \mid P(x) \le t\}$ *is a polyhedron. Then, the set-valued mapping* Σ *is a polyhedral multifunction.*

The proof is given in the supplementary material. Noting that both the ℓ_1-norm and ℓ_∞-norm have polyhedral epigraphs, by Lemma 2, the set-valued mapping Σ is a polyhedral multifunction when $p = 1$ and $p = \infty$. Hence, by Lemma 1, Σ is ULC at $(\bar{y}, \bar{g}) \in \mathbb{R}^m \times \mathbb{R}^n$ if $\Sigma(\bar{y}, \bar{g})$ is non-empty. The latter is ensured by Assumption 1. Upon applying Theorem 1, we have the following result:

Corollary 1 *Under Assumption 1, the EB condition holds for* (1) *when* $p = 1$ *and* $p = \infty$.

4.2. EB Condition Holds when $p \in (1, 2]$

Next, we show that under Assumption 1, the EB condition holds for (1) when $p \in (1, 2]$. Towards that end, let us first state several technical results that will be used to establish the ULC property of the set-valued mapping Σ. The proofs of these results can be found in the supplementary material.

Lemma 3 *Let* $B \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$, $d \in \mathbb{R}^n$, *and* $J \subseteq \{1, \ldots, n\}$ *be given. Define the sets*

$$\mathcal{P}_1 := \{x \in \mathbb{R}^n \mid Bx = b\},$$
$$\mathcal{P}_2 := \{x \in \mathbb{R}^n \mid x_J = a_J \cdot d_J,\ a_J \le 0\}.$$

Suppose that $\mathcal{P}_1$ *is non-empty. Then, there exists a constant* $\theta > 0$ *such that for any* $x \in \mathbb{R}^n$,

$$d(x, \mathcal{P}_1) \le \theta \|Bx - b\|.$$

Moreover, for any $x \in \mathbb{R}^n$ *and* $p \in [1, \infty]$,

$$d(x, \mathcal{P}_2) \le \|x_J\|_p \cdot \left\| \frac{d_J}{\|d_J\|_p} + \frac{x_J}{\|x_J\|_p} \right\|,$$

where we adopt the convention that $u/\|u\|_p = \mathbf{0}$ *if* $u = \mathbf{0}$.

The following result is the so-called linear regularity of a collection of polyhedral sets; see Corollary 5.26 of (Bauschke & Borwein, 1996).

Lemma 4 *Let* $\mathcal{C}_1, \ldots, \mathcal{C}_N$ *be polyhedra in* $\mathbb{R}^n$. *Then, there exists a constant* $\tau > 0$ *such that for any* $x \in \mathbb{R}^n$,

$$d\left(x, \bigcap_{i=1}^N \mathcal{C}_i\right) \le \tau \sum_{i=1}^N d\left(x, \mathcal{C}_i\right).$$

We next present a result concerning the subdifferential of the ℓ_p-norm when $p \in (1, \infty)$. Let q denote the Hölder conjugate of p; *i.e.*, $1/p + 1/q = 1$.

Proposition 3 *Let* $g \in \mathbb{R}^n$, $\omega > 0$, *and* $p \in (1, \infty)$ *be given. Define the set*

$$\mathcal{S} := \{x \in \mathbb{R}^n \mid -g \in \omega \partial \|x\|_p\}.$$

Then,

$$\mathcal{S} = \begin{cases} \emptyset & \text{if } \|g\|_q > \omega; \\ \{x \mid x = a \cdot v(g),\ a \le 0\} & \text{if } \|g\|_q = \omega; \\ \{\mathbf{0}\} & \text{if } \|g\|_q < \omega, \end{cases}$$

where the function $v : \mathbb{R}^n \to \mathbb{R}^n$ *is defined by*

$$v(g) := \left(sgn(g_1)|g_1|^{\frac{q}{p}}, \ldots, sgn(g_n)|g_n|^{\frac{q}{p}}\right).$$

In addition, when $p \in (1, 2]$, *for any* $g \in \mathbb{R}^n$, *there exist constants* $\delta > 0$ *and* $\nu > 0$ *such that*

$$\|v(g) - v(\tilde{g})\| \le \nu \|g - \tilde{g}\| \text{ whenever } \|g - \tilde{g}\| \le \delta. \quad (6)$$

Recall that $P(x) = \sum_{J \in \mathcal{J}} \omega_J \|x_J\|_p$, where $\mathcal{J}$ is a non-overlapping partition of the coordinate index set $\{1, \ldots, n\}$. Hence, for any $x, g \in \mathbb{R}^n$, $-g \in \partial P(x)$ if and

only if $-g_J \in \omega_J \partial \|x_J\|_p$ for all $J \in \mathcal{J}$. This, together with Proposition 3, implies that if $\Sigma(y,g)$ is non-empty, then $\|g_J\| \le \omega_J$ for all $J \in \mathcal{J}$. In particular, we may write

$$\Sigma(y,g) = \left\{ x \;\middle|\; \begin{array}{l} Ax = y, \\ x_J = a_J \cdot v(g_J),\ a_J \le 0,\ \forall J \in \mathcal{J}_1^g, \\ x_J = \mathbf{0},\ \forall J \in \mathcal{J}_2^g \end{array} \right\}, \tag{7}$$

where

$$\mathcal{J}_1^g := \{J \in \mathcal{J} \mid \|g_J\| = \omega_J\},$$
$$\mathcal{J}_2^g := \{J \in \mathcal{J} \mid \|g_J\| < \omega_J\}.$$

This shows that $\Sigma(y,g)$ is closed. The next lemma reveals that boundedness of Σ is a property that is stable under small perturbations.

Lemma 5 *Suppose that the set-valued mapping Σ is non-empty and bounded at $(y,g) \in \mathbb{R}^m \times \mathbb{R}^n$. Then, there exists a constant $\delta > 0$ such that $\Sigma(\tilde{y},\tilde{g})$ is bounded whenever $(\tilde{y},\tilde{g}) \in \mathbb{R}^m \times \mathbb{R}^n$ satisfies $\|\tilde{g} - g\| \le \delta$ and $\Sigma(\tilde{y},\tilde{g})$ is non-empty.*

Now, we are ready to study the ULC property of the set-valued mapping Σ.

Theorem 2 *Suppose that $p \in (1,2]$. Then, the set-valued mapping Σ is ULC at any $(y,g) \in \mathbb{R}^m \times \mathbb{R}^n$ such that $\Sigma(y,g)$ is non-empty and bounded.*

Proof Define the sets

$$\mathcal{C}_1(J) := \{x \in \mathbb{R}^n \mid x_J = a_J \cdot v(g_J),\ a_J \le 0\},\ \forall J \in \mathcal{J}_1^g,$$
$$\mathcal{C}_2 := \{x \in \mathbb{R}^n \mid x_J = \mathbf{0},\ \forall J \in \mathcal{J}_2^g\},$$
$$\mathcal{C}_3 := \{x \in \mathbb{R}^n \mid Ax = y\}.$$

Then, by (7), we have $\Sigma(y,g) = \left(\cap_{J \in \mathcal{J}_1^g} \mathcal{C}_1(J)\right) \cap \mathcal{C}_2 \cap \mathcal{C}_3$. Moreover, since $\mathcal{C}_1(J), \mathcal{C}_2, \mathcal{C}_3$, where $J \in \mathcal{J}_1^g$, are all polyhedral subsets of $\mathbb{R}^n$, by Lemma 4, there exists a constant $\tau > 0$ such that for any $x \in \mathbb{R}^n$,

$$d(x, \Sigma(y,g)) \le \tau \left(\sum_{J \in \mathcal{J}_1^g} d(x, \mathcal{C}_1(J)) + \sum_{i=2}^{3} d(x, \mathcal{C}_i) \right). \tag{8}$$

Thus, to prove Theorem 2, it suffices to bound the right-hand side of (8) for all $x \in \Sigma(\tilde{y},\tilde{g})$, where $(\tilde{y},\tilde{g}) \in \mathbb{R}^m \times \mathbb{R}^n$ lies in a neighborhood of $(y,g) \in \mathbb{R}^m \times \mathbb{R}^n$ and $\Sigma(\tilde{y},\tilde{g})$ is non-empty. Towards that end, we first note that since $\|g_J\| < \omega_J$ for $J \in \mathcal{J}_2^g$, there exists a constant $\delta_1 > 0$ such that

$$\|\tilde{g}_J\| < \omega_J, \quad \forall J \in \mathcal{J}_2^g \tag{9}$$

whenever $\|(\tilde{y},\tilde{g}) - (y,g)\| \le \delta_1$. Now, for any such pair $(\tilde{y},\tilde{g}) \in \mathbb{R}^m \times \mathbb{R}^n$ and any index set $J \in \mathcal{J}_1^g$, we either have (a) $\|\tilde{g}_J\| = \omega_J$ or (b) $\|\tilde{g}_J\| < \omega_J$. (The case $\|\tilde{g}_J\| > \omega_J$ cannot happen because $\Sigma(\tilde{y},\tilde{g})$ is assumed to be non-empty.) It follows that $\mathcal{J}_1^g = \mathcal{J}_1^g(a) \cup \mathcal{J}_1^g(b)$, where

$$\mathcal{J}_1^g(a) := \{J \in \mathcal{J}_1^g \mid \|\tilde{g}_J\| = \omega_J\}, \tag{10}$$
$$\mathcal{J}_1^g(b) := \{J \in \mathcal{J}_1^g \mid \|\tilde{g}_J\| < \omega_J\}. \tag{11}$$

Since $\Sigma(y,g)$ is non-empty and bounded, by Lemma 5, there exist constants $\delta_2 > 0$ and $R > 0$ such that for any $x \in \Sigma(\tilde{y},\tilde{g})$, we have

$$\|x\|_p \le R \quad \text{whenever} \quad \|(\tilde{y},\tilde{g}) - (y,g)\| \le \delta_2. \tag{12}$$

Therefore, in view of (9)–(12) and Proposition 3, every $x \in \Sigma(\tilde{y},\tilde{g})$ that satisfies $\|(\tilde{y},\tilde{g}) - (y,g)\| \le \min\{\delta_1, \delta_2\}$ must also satisfy the following conditions:

$$Ax = \tilde{y}, \tag{13}$$
$$x_J = a_J \cdot v(\tilde{g}_J) \text{ for some } a_J \le 0,\ \forall J \in \mathcal{J}_1^g(a), \tag{14}$$
$$x_J = \mathbf{0},\ \forall J \in \mathcal{J}_1^g(b) \cup \mathcal{J}_2^g, \tag{15}$$
$$\|x\|_p \le R. \tag{16}$$

Using (15), it is clear that

$$d(x, \mathcal{C}_2) = 0. \tag{17}$$

Moreover, by (13) and Lemma 3, there exists a constant $\theta' > 0$ such that

$$d(x, \mathcal{C}_3) \le \theta' \|Ax - y\| = \theta' \|\tilde{y} - y\|. \tag{18}$$

Now, by (14), (15), and Lemma 3, we have $d(x, \mathcal{C}_1(J)) = 0$ for $J \in \mathcal{J}_1^g(b)$ and

$$\begin{aligned} d(x, \mathcal{C}_1(J)) &\le \|x_J\|_p \cdot \left\| \frac{v(g_J)}{\|v(g_J)\|_p} + \frac{x_J}{\|x_J\|_p} \right\| \\ &= \|x_J\|_p \cdot \left\| \frac{v(g_J)}{\|v(g_J)\|_p} - \frac{v(\tilde{g}_J)}{\|v(\tilde{g}_J)\|_p} \right\| \\ &= \omega_J^{1-q} \cdot \|x_J\|_p \cdot \|v(g_J) - v(\tilde{g}_J)\| \\ &\le \nu_J \omega_J^{1-q} \cdot \|x_J\|_p \cdot \|g_J - \tilde{g}_J\| \end{aligned}$$

for $J \in \mathcal{J}_1^g(a)$, where the third line follows from the fact that $\|v(g)\|_p = \omega^{q-1}$ whenever $\|g\|_q = \omega$, and the fourth line is due to (6). Together with (16), the above yields

$$\begin{aligned} \sum_{J \in \mathcal{J}_1^g} d(x, \mathcal{C}_1(J)) &\le \sum_{J \in \mathcal{J}_1^g(a)} \nu_J \omega_J^{1-q} \cdot \|x_J\|_p \cdot \|g_J - \tilde{g}_J\| \\ &\le \left(R \sum_{J \in \mathcal{J}} \nu_J \omega_J^{1-q} \right) \|\tilde{g} - g\|. \end{aligned} \tag{19}$$

Substituting (17), (18), and (19) into (8), we obtain

$$d(x, \Sigma(y,g)) \le \theta \|(\tilde{y},\tilde{g}) - (y,g)\|$$

for any $x \in \Sigma(\tilde{y},\tilde{g})$ with $\|(\tilde{y},\tilde{g}) - (y,g)\| \le \min\{\delta_1, \delta_2\}$, where $\theta = \max\left\{\theta', R\sum_{J \in \mathcal{J}} \nu_J \omega_J^{1-q}\right\}$. It follows that Σ is ULC at $(y,g) \in \mathbb{R}^m \times \mathbb{R}^n$, as desired. □

Seeing that Assumption 1 and Propositions 1 and 2 ensure the boundedness of $\Sigma(\bar{y}, \bar{g})$, the following result is a direct combination of Theorems 1 and 2:

Corollary 2 *Under Assumption 1, the EB condition holds for (1) when $p \in (1, 2]$.*

4.3. EB Condition Fails when $p \in (2, \infty)$

It can be verified that in this scenario, the set-valued mapping Σ is not ULC at certain points. The key intuition is that when $p \in (2, \infty)$, we have $0 < q/p < 1$, which implies that the function $s \mapsto |s|^{q/p}$ is not Lipschitz continuous at $s = 0$. As such, the inequality (6) fails to hold, which means that Theorem 1 is no longer valid in this scenario. In what follows, we will construct an explicit example to demonstrate that under Assumption 1, the EB condition for problem (1) fails to hold for any $p \in (2, \infty)$.

Example. Consider the following problem:

$$\min_{x \in \mathbb{R}^2} \ \frac{1}{2}\|Ax - b\|^2 + \|x\|_p, \tag{20}$$

where $A = [1, 0]$, $b = 2$. It is obvious that this problem satisfies Assumption 1. In addition, the optimal value and optimal solution set of (20) can be calculated explicitly.

Proposition 4 *Consider problem (20) with $p \in (2, \infty)$. The optimal value is $v^* = 3/2$ and the optimal solution set is given by $\mathcal{X} = \{(1, 0)\}$.*

The proof of Proposition 4 can be found in the supplementary material. Now, let $\{\delta_k\}_{k\geq 0}$ be a sequence converging to zero; *i.e.*, $\delta_k = o(1)$. For simplicity, we assume that $\delta_k > 0$ for all $k \geq 0$. Consider the sequence $\{x^k\}_{k\geq 0}$ with

$$x_1^k := 2 - (1 - \delta_k)^{\frac{1}{q}}, \ x_2^k := \frac{2 - (1 - \delta_k)^{\frac{1}{q}}}{(1 - \delta_k)^{\frac{1}{p}}} \cdot \delta_k^{\frac{1}{p}} + \delta_k^{\frac{1}{q}},$$

where q is the Hölder conjugate of p. Since $\delta_k \to 0$, the sequence x^k converges to $\mathcal{X}$. Our goal now is to show that $\|R(x^k)\| = o(d(x^k, \mathcal{X}))$ when $p \in (2, \infty)$.

To begin, observe that x_1^k converges to 1 at the rate $\Theta(\delta_k)$ and x_2^k converges to 0 at the rate $\Theta(\delta_k^{1/p})$ (note that when $p \geq 1$, $\delta_k = O(\delta_k^{1/p})$). Thus, we have $d(x^k, \mathcal{X}) = \Theta(\delta_k^{1/p})$.

Next, we need to compute $R(x^k)$. This is done in the following lemma, whose proof can be found in the supplementary material.

Lemma 6 *For the sequence $\{x^k\}_{k\geq 0}$ defined above, we have $R(x^k) = (0, -\delta_k^{1/q})$.*

Since $1/p < 1/q$ when $p \in (2, \infty)$, we have $\delta_k^{1/q} = o(\delta_k^{1/p})$. It follows from Lemma 6 that when $p \in (2, \infty)$,

$$\|R(x^k)\| = \Theta\left(\delta_k^{\frac{1}{q}}\right) = o\left(\delta_k^{\frac{1}{p}}\right) = o\left(d\left(x^k, \mathcal{X}\right)\right),$$

which shows that the EB condition fails for problem (20).

5. Convergence Rates of First-Order Methods

As mentioned in the Introduction, the EB condition (4) can be used to derive strong convergence rate results for various first-order methods. In this section, we use the newly-established EB condition for $\ell_{1,p}$-regularization to analyze the convergence rates of the proximal gradient (PG) and block coordinate gradient descent (BCGD) methods when they are applied to solve problem (1). In what follows, we say that a sequence $\{w^k\}_{k\geq 0}$ converges Q-linearly (resp. R-linearly) to w^∞ if there exists a constant $\rho \in (0, 1)$ such that $\limsup_{k\to\infty}\{\|w^{k+1} - w^\infty\|/\|w^k - w^\infty\|\} \leq \rho$ (resp. if there exist constants $\gamma > 0$ and $\rho \in (0, 1)$ such that $\|w^k - w^\infty\| \leq \gamma \cdot \rho^k$ for all $k \geq 0$).

5.1. Proximal Gradient Method

The PG method is well suited for solving non-smooth composite optimization problems. Its adaptation for solving $\ell_{1,p}$-regularization is proposed in (Liu & Ye, 2009; 2010; Zhang et al., 2013). Each iteration of the PG method involves the computation of a proximal operator. For problem (1), the proximal operator is defined as

$$\text{prox}_P(x) := \operatorname*{argmin}_{z \in \mathbb{R}^n} \left\{ P(z) + \frac{1}{2}\|z - x\|^2 \right\}.$$

It can be verified that $x \in \mathbb{R}^n$ is an optimal solution to (1) if and only if it satisfies the following fixed-point equation:

$$x = \text{prox}_P(x - \nabla f(x)).$$

This motivates the following fixed-point iteration for solving (1):

$$x^{k+1} = \text{prox}_{\alpha_k P}\left(x^k - \alpha_k \nabla f(x^k)\right),$$

where $\alpha_k > 0$ is the stepsize. It has been shown that for $p \in [1, \infty]$, $\text{prox}_P(x)$ can be computed efficiently using the so-called $\ell_{1,p}$-regularized Euclidean projection (EP_{1p}) method (Liu & Ye, 2009; 2010). We summarize the PG method for solving (1) in Algorithm 1.

Algorithm 1 Proximal Gradient Method

Input: initial point x^0
for $k = 0$ **to** N **do**
1. choose a stepsize $\alpha_k > 0$
2. compute $y^k = x^k - \alpha_k \nabla f(x^k)$
3. compute $\text{prox}_{\alpha_k P}(y^k)$ using the EP_{1p} method
4. set $x^{k+1} = \text{prox}_{\alpha_k P}(y^k)$

end for

It is known that the sequence generated by the PG method converges linearly if the EB condition (4) is satisfied (Zhang et al., 2013). By invoking Corollaries 1 and 2, we obtain the following result:

Corollary 3 *Consider the $\ell_{1,p}$-regularized problem (1) with Assumption 1 satisfied. Let $L > 0$ be the Lipschitz constant of ∇f. Let $\{x^k\}_{k\geq 0}$ be the sequence generated by Algorithm 1. Suppose that the stepsizes $\{\alpha_k\}_{k\geq 0}$ satisfy*

$$\inf_k \alpha_k > 0, \quad \sup_k \alpha_k < \frac{1}{L}.$$

If $p \in [1,2]$ or $p = \infty$, then $\{f(x^k)\}_{k\geq 0}$ converges Q-linearly to the optimal value v^* and $\{x^k\}_{k\geq 0}$ converges R-linearly to an element in $\mathcal{X}$.

5.2. Block Coordinate Gradient Descent Method

The BCGD method is developed in (Tseng & Yun, 2009) and is applied to the $\ell_{1,p}$-regularized problem (1) in (Meier et al., 2008; Liu et al., 2009). In each iteration of the BCGD method, a block $J \in \mathcal{J}$ and a symmetric positive definite matrix H are chosen. Then, a search direction $v_H(x;J)$, which is defined as the minimizer of the problem

$$\begin{array}{ll} \min & \nabla f(x)^T d + \frac{1}{2} d^T H d + P(x+d) \\ \text{s.t.} & d_j = 0, \ \forall\, j \notin J, \end{array} \tag{21}$$

is computed. Finally, the iterate is updated by moving along the direction $v_H(x;J)$ with stepsize $\alpha > 0$, where α is chosen according to the Armijo rule (Tseng & Yun, 2009). We summarize the BCGD method in Algorithm 2.

Algorithm 2 Block Coordinate Gradient Descent Method

Input: initial point x^0
for $k = 0$ **to** N **do**
1. choose a block $J^k \in \mathcal{J}$ and a symmetric positive definite matrix H^k
2. solve problem (21) and obtain the search direction $v_{H^k}(x^k; J^k)$
3. choose a stepsize $\alpha_k > 0$ by the Armijo rule and update $x^{k+1} = x^k + \alpha_k v_{H^k}(x^k; J^k)$

end for

It has been shown in Theorem 2 of (Tseng & Yun, 2009) that Algorithm 2 attains a linear rate of convergence if the EB condition (4) is satisfied. By invoking again Corollaries 1 and 2, we obtain the following result:

Corollary 4 *Consider the $\ell_{1,p}$-regularized problem (1) with Assumption 1 satisfied. Let $\{x^k\}_{k\geq 0}$ be the sequence generated by Algorithm 2, where the blocks $\{J^k\}_{k\geq 0}$ cycle over $\mathcal{J}$ and the stepsizes $\{\alpha_k\}_{k\geq 0}$ satisfy*

$$\inf_k \alpha_k > 0, \quad \sup_k \alpha_k \leq 1.$$

If $p \in [1,2]$ or $p = \infty$, then $\{f(x^k)\}_{k\geq 0}$ converges Q-linearly to the optimal value v^* and $\{x^k\}_{k\geq 0}$ converges R-linearly to an element in $\mathcal{X}$.

As implied by Corollaries 3 and 4, the PG and BCGD methods for solving $\ell_{1,p}$-regularized linear regression or logistic regression are theoretically guaranteed to attain a linear rate of convergence when $p \in [1,2]$ or $p = \infty$. By contrast, since the EB condition fails to hold when $p \in (2,\infty)$, the PG and BCGD methods for solving the same class of problems may not converge linearly.

6. Numerical Experiments

In this section, we perform numerical experiments to study the convergence rates of the PG and BCGD methods for solving $\ell_{1,p}$-regularized linear regression and logistic regression on synthetic datasets. As we shall see, the results corroborate our theoretical analyses in previous sections.

6.1. Example for which the EB Condition Fails

Recall the example we constructed in Section 4.3; *i.e.*, problem (20). In spite of its small size, problem (20) is of particular interest in experiments of convergence rates due to the following reasons. First, it belongs to the class of $\ell_{1,p}$-regularized problems that satisfy Assumption 1. Second, the EB condition holds for (20) when $p \in [1,2]$ and $p = \infty$, while it fails when $p \in (2,\infty)$. Third, its optimal value v^* is known in advance (Proposition 4), so that we can trace the curve $\log(f(x^k) - v^*)$ precisely.

We implement the PG method (Algorithm 1) to solve (20) with $p = 1, 1.25, 1.5, 1.75, 2, 2.5, 3, 4, \infty$. The stepsize is chosen to be constant $\alpha_k \equiv 0.5$, which can be verified to satisfy the conditions stated in Corollary 3. The convergence performance of the objective value is presented in Figure 1. It is readily seen that when $p \in [1,2]$ or $p = \infty$, $\{f(x^k)\}_{k\geq 0}$ converges linearly to v^* (Figure 1(a)). By contrast, when $p \in (2,\infty)$, the objective value converges at a sublinear rate (Figure 1(b)). Our experiments suggest that for the $\ell_{1,p}$-regularized problem (1), a linear rate of convergence is in general not achievable if $p \in (2,\infty)$.

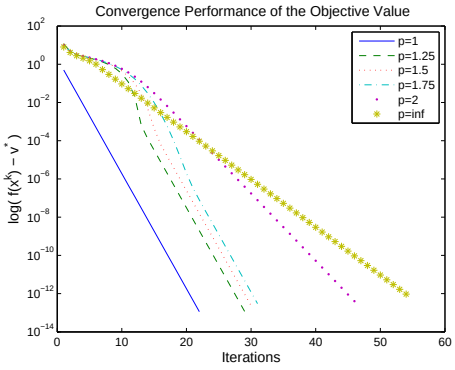
6.2. Synthetic Datasets

In this section, we test the convergence rates of first-order methods for solving $\ell_{1,p}$-regularized regression with $p \in [1,2]$ or $p = \infty$ on synthetic datasets. In particular, we consider the PG method (Algorithm 1) for solving $\ell_{1,p}$-regularized linear regression and the BCGD method (Algorithm 2) for solving $\ell_{1,p}$-regularized logistic regression.

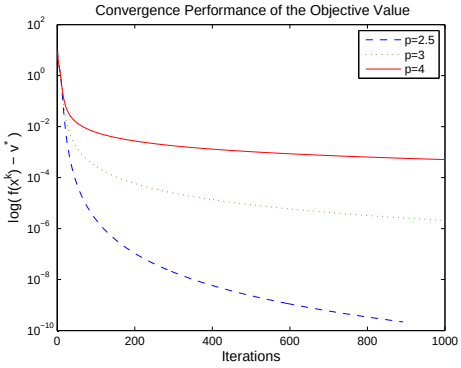
$\ell_{1,p}$-Regularized Linear Regression. We consider

$$\min_{X \in \mathbb{R}^{d\times k}} \ \frac{1}{2}\|AX - Y\|_F^2 + \tau \sum_{i=1}^{d} \|X^{(i)}\|_p, \tag{22}$$

where $A \in \mathbb{R}^{m\times d}$ is a measurement matrix, $Y \in \mathbb{R}^{m\times k}$ is the response matrix, and $\tau > 0$ is a regularization parameter. In addition, we treat each row of X as a group and



(a)



(b)

Figure 1. The PG method for solving problem (20).

use $X^{(i)}$ to denote the i-th row of X. We utilize the same strategy as the experiments in (Liu & Ye, 2010). Precisely, each entry of A is generated independently from the standard normal distribution. Moreover, we generate a jointly sparse matrix $X^* \in \mathbb{R}^{d\times k}$, where the entries of the first $d_0 < d$ rows are being sampled from the normal distribution and the remaining entries are all set to 0. Then, we let $Y = AX^* + Z$, where $Z \in \mathbb{R}^{m\times k}$ is the noise matrix whose entries are sampled from the normal distribution with mean zero and standard deviation 0.1. Figure 2 illustrates the convergence performance of the PG method (Algorithm 1) for solving (22) with $m = 50, d = 100, d_0 = 30, k = 20$, and $\tau = 50$. It reveals that the objective value converge linearly to the optimal value when $p \in [1, 2]$ or $p = \infty$. This confirms our result in Corollary 3.

$\ell_{1,p}$-Regularized Logistic Regression. We consider

$$\min_{X\in\mathbb{R}^{d\times k}} \sum_{s=1}^{S} \log(1 + \exp(-y_s\langle W_s, X\rangle)) + \tau \sum_{i=1}^{d} \|X^{(i)}\|_p, \tag{23}$$

where $W_s \in \mathbb{R}^{d\times k}$, $y_s \in \{-1, 1\}$, and $\tau > 0$ is a regularization parameter. Here, $\langle W_s, X\rangle = \text{trace}(W_s^T X)$ and $X^{(i)}$ denotes the i-th row of X. For the data generation, we

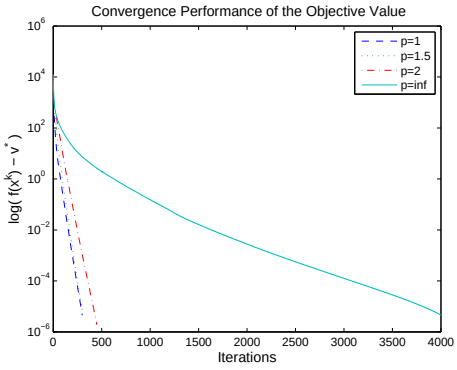


Figure 2. The performance of the PG method for solving $\ell_{1,p}$-regularized linear regression.

first sample S matrices $W_1, \ldots, W_S$ independently from the standard Wishart distribution. Then, a jointly sparse matrix X^* is generated in the same way as in the experiment of $\ell_{1,p}$-regularized linear regression. Finally, we let $y_s = \text{sgn}(\langle W_s, X\rangle)$, where $s = 1, \ldots, S$. Figure 3 shows the convergence performance of the BCGD method (Algorithm 2) for solving (23) with $d = k = 50$, $S = 100$, $d_0 = 10$, and $\tau = 20$. It is clear from the figure that the objective value of (23) converges linearly to the optimal value when $p \in [1, 2]$ or $p = \infty$. This corroborates our result in Corollary 4.

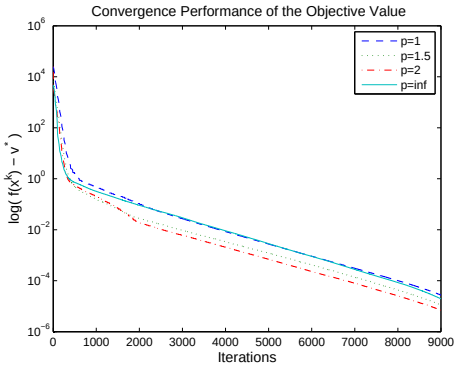


Figure 3. The performance of the BCGD method for solving $\ell_{1,p}$-regularized logistic regression.

Acknowledgements

This work is supported in part by the Hong Kong Research Grants Council (RGC) General Research Fund (GRF) Project CUHK 14206814 and in part by a gift grant from Microsoft Research Asia.

References

Bach, Francis, Jenatton, Rodolphe, Mairal, Julien, and Obozinski, Guillaume. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

Bach, Francis R. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

Bauschke, Heinz H and Borwein, Jonathan M. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3):367–426, 1996.

Beck, Amir and Teboulle, Marc. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Combettes, Patrick L and Wajs, Valérie R. Signal recovery by proximal forward–backward splitting. *Multiscale Modeling and Simulation*, 4(4):1168–1200, 2005.

Dontchev, Asen L and Rockafellar, R Tyrrell. *Implicit Functions and Solution Mappings*. Springer Monographs in Mathematics. Springer Science+Business Media, LLC, New York, 2009.

Eldar, Yonina C, Kuppinger, Patrick, and Bolcskei, Helmut. Block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Transactions on Signal Processing*, 58 (6):3042–3054, 2010.

Fornasier, Massimo and Rauhut, Holger. Recovery algorithms for vector-valued data with joint sparsity constraints. *SIAM Journal on Numerical Analysis*, 46(2): 577–613, 2008.

Hale, Elaine T, Yin, Wotao, and Zhang, Yin. Fixed-point continuation for ℓ_1-minimization: Methodology and convergence. *SIAM Journal on Optimization*, 19(3): 1107–1130, 2008.

Hong, Mingyi and Luo, Zhi-Quan. On the linear convergence of the alternating direction method of multipliers. *arXiv preprint arXiv:1208.3922*, 2012.

Kloft, Marius, Brefeld, Ulf, Sonnenburg, Sören, and Zien, Alexander. ℓ_p-norm multiple kernel learning. *Journal of Machine Learning Research*, 12:953–997, 2011.

Kowalski, Matthieu. Sparse regression using mixed norms. *Applied and Computational Harmonic Analysis*, 27(3): 303–324, 2009.

Liu, Han, Palatucci, Mark, and Zhang, Jian. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *Proceedings of the 26th International Conference on Machine Learning*, pp. 649–656, 2009.

Liu, Jun and Ye, Jieping. Efficient Euclidean projections in linear time. In *Proceedings of the 26th International Conference on Machine Learning*, pp. 657–664, 2009.

Liu, Jun and Ye, Jieping. Efficient ℓ_1/ℓ_q norm regularization. *arXiv preprint arXiv:1009.4766*, 2010.

Luo, Zhi-Quan and Tseng, Paul. On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM Journal on Control and Optimization*, 30(2):408–425, 1992.

Luo, Zhi-Quan and Tseng, Paul. Error bounds and convergence analysis of feasible descent methods: A general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

Meier, Lukas, van de Geer, Sara, and Bühlmann, Peter. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

Minty, George J. Monotone (nonlinear) operators in Hilbert space. *Duke Mathematical Journal*, 29(3):341–346, 1962.

Minty, George J. On the monotonicity of the gradient of a convex function. *Pacific Journal of Mathematics*, 14(1): 243–247, 1964.

Nesterov, Yurii. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, Boston, 2004.

Pang, Jong-Shi. A posteriori error bounds for the linearly-constrained variational inequality problem. *Mathematics of Operations Research*, 12(3):474–484, 1987.

Robinson, Stephen M. Some continuity properties of polyhedral multifunctions. In König, H, Korte, B, and Ritter, K (eds.), *Mathematical Programming at Oberwolfach*, volume 14 of *Mathematical Programming Study*, pp. 206–214. North-Holland Publishing Company, Amsterdam, 1981.

Rockafellar, R Tyrrell. *Convex Analysis*. Princeton University Press, Princeton, New Jersey, 1970.

So, Anthony Man-Cho. Non-asymptotic convergence analysis of inexact gradient methods for machine learning without strong convexity. *arXiv preprint arXiv:1309.0113*, 2013.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Tomioka, Ryota and Suzuki, Taiji. Sparsity-accuracy trade-off in MKL. *arXiv preprint arXiv:1001.2615*, 2010.

Tseng, Paul. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

Tseng, Paul. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming*, 125(2):263–295, 2010.

Tseng, Paul and Yun, Sangwoon. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1-2):387–423, 2009.

Vogt, Julia E and Roth, Volker. A complete analysis of the $\ell_{1,p}$ group-lasso. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Wang, Po-Wei and Lin, Chih-Jen. Iteration complexity of feasible descent methods for convex optimization. *Journal of Machine Learning Research*, 15:1523–1548, 2014.

Xiao, Lin and Zhang, Tong. A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization*, 23(2):1062–1091, 2013.

Yuan, Ming and Lin, Yi. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Zhang, Haibin, Jiang, Jiaojiao, and Luo, Zhi-Quan. On the linear convergence of a proximal gradient method for a class of nonsmooth convex minimization problems. *Journal of the Operations Research Society of China*, 1 (2):163–186, 2013.

7. Supplementary Material

7.1. Proof of Proposition 1

For arbitrary $x_1, x_2 \in \mathcal{X}$, let $y_1 = Ax_1, y_2 = Ax_2$ and suppose that $y_1 \neq y_2$. Assumption 1(a) implies that the function h is strongly convex on the line segment joining y_1 and y_2. Thus, there exists a constant $\sigma > 0$ such that

$$h\left(\frac{y_1+y_2}{2}\right) \leq \frac{1}{2}h(y_1) + \frac{1}{2}h(y_2) - \frac{\sigma}{2}\|y_1 - y_2\|^2.$$

Using (2), the above is equivalent to

$$f\left(\frac{x_1+x_2}{2}\right) \leq \frac{1}{2}f(x_1) + \frac{1}{2}f(x_2) - \frac{\sigma}{2}\|y_1 - y_2\|^2.$$

Moreover, by the convexity of P, we have

$$P\left(\frac{x_1+x_2}{2}\right) \leq \frac{1}{2}P(x_1) + \frac{1}{2}P(x_2).$$

Adding the above two inequalities and using $x_1, x_2 \in \mathcal{X}$ yield

$$F\left(\frac{x_1+x_2}{2}\right) \leq v^* - \frac{\sigma}{2}\|y_1 - y_2\|^2 < v^*,$$

where v^* is the optimal value of (1). However, this contradicts the optimality of v^*. Hence, we have $y_1 = y_2$; *i.e.*, Ax is invariant over $\mathcal{X}$. Since $\nabla f(x) = A^T\nabla h(Ax)$ by (2), we see that $\nabla f(x)$ is also invariant over $\mathcal{X}$. Therefore, there exists a vector $\bar{y} \in \mathbb{R}^m$ such that $\bar{y} = Ax$ and $\nabla f(x) = A^T\nabla h(\bar{y}) = \bar{g}$ for any $x \in \mathcal{X}$. Now, we can express the optimal solution set as

$$\mathcal{X} = \left\{ x \in \mathbb{R}^n \;\middle|\; Ax = \bar{y},\ \sum_{J\in\mathcal{J}} \omega_J \|x_J\|_p = v^* - h(\bar{y}) \right\}.$$

This shows that $\mathcal{X}$ is a compact convex set. □

7.2. Proof of Proposition 2

Since problem (1) is convex, its first-order optimality condition is both necessary and sufficient. Hence, we have

$$\mathcal{X} = \{x \in \mathbb{R}^n \mid \mathbf{0} \in \nabla f(x) + \partial P(x)\}. \tag{24}$$

Now, let $x \in \mathcal{X}$ be arbitrary. By Proposition 1, we have $Ax = \bar{y}$ and $\nabla f(x) = \bar{g}$. This, together with (24), leads to $x \in \Sigma(\bar{y}, \bar{g})$. On the other hand, for any $x \in \Sigma(\bar{y}, \bar{g})$, since $\bar{g} = A^T\nabla h(\bar{y}) = A^T\nabla h(Ax) = \nabla f(x)$, we conclude that $\mathbf{0} \in \nabla f(x) + \partial P(x)$; *i.e.*, $x \in \mathcal{X}$. □

7.3. Proof of Theorem 1

Since Σ is ULC at $(\bar{y}, \bar{g})$, there exist constants $\theta > 0$ and $\delta > 0$ such that for any (y, g) satisfying $\|(y, g) - (\bar{y}, \bar{g})\| \leq \delta$,

$$\Sigma(y, g) \subseteq \Sigma(\bar{y}, \bar{g}) + \theta\|(y, g) - (\bar{y}, \bar{g})\|\mathcal{B}. \tag{25}$$

Consider the functions $y^+ : \mathbb{R}^n \to \mathbb{R}^m$ and $g^+ : \mathbb{R}^n \to \mathbb{R}^n$ given by

$$y^+(x) = A(x + R(x)), \quad g^+(x) = \nabla f(x) + R(x). \tag{26}$$

It is easy to verify that $R(x) = \mathrm{prox}_P(x - \nabla f(x)) - x$, where $\mathrm{prox}_P : \mathbb{R}^n \to \mathbb{R}^n$ is the proximal operator given by

$$\mathrm{prox}_P(x) = \operatorname*{argmin}_{z\in\mathbb{R}^n} \left\{ P(z) + \frac{1}{2}\|z - x\|^2 \right\}.$$

Thus, by Lemma 2.4 of (Combettes & Wajs, 2005), R is Lipschitz continuous. Since ∇f is also Lipschitz continuous, we see that both y^+ and g^+ are Lipschitz continuous. This, together with Proposition 1, implies the existence of a constant $\rho > 0$ such that for all $x \in \mathbb{R}^n$ satisfying $d(x, \mathcal{X}) \leq \rho$,

$$\left\| \left(y^+(x), g^+(x)\right) - (\bar{y}, \bar{g}) \right\| \leq \delta. \tag{27}$$

Using the definition of R in (3), we have

$$\mathbf{0} \in \nabla f(x) + R(x) + \partial P(x + R(x)). \tag{28}$$

Hence, by (5) and (26), for all $x \in \mathbb{R}^n$,

$$x + R(x) \in \Sigma\left(y^+(x), g^+(x)\right).$$

This, together with (25) and (27), yields

$$d(x + R(x), \Sigma(\bar{y}, \bar{g})) \leq \theta \left\| \left(y^+(x), g^+(x)\right) - (\bar{y}, \bar{g}) \right\| \tag{29}$$

whenever $d(x, \mathcal{X}) \leq \rho$. Now, using the fact that $\nabla f(x) = A^T \nabla h(Ax)$ and $\bar{g} = A^T \nabla h(\bar{y})$, we bound

$$\begin{aligned} \left\| y^+(x) - \bar{y} \right\| &\leq \|Ax - \bar{y}\| + \|A\| \cdot \|R(x)\|, \\ \left\| g^+(x) - \bar{g} \right\| &\leq L\|A^T\| \cdot \|Ax - \bar{y}\| + \|R(x)\|, \end{aligned}$$

where $L > 0$ is the Lipschitz constant of ∇h. Thus, by letting $M = \max\left\{\|A\|, L\|A^T\|, 1\right\}$, we obtain from (29) that

$$d(x + R(x), \Sigma(\bar{y}, \bar{g})) \leq M\theta(\|Ax - \bar{y}\| + \|R(x)\|)$$

whenever $d(x, \mathcal{X}) \leq \rho$. In view of Proposition 2 and the inequality $d(x, \mathcal{X}) \leq d(x + R(x), \mathcal{X}) + \|R(x)\|$, this implies that

$$d(x, \mathcal{X}) \leq \kappa_0(\|Ax - \bar{y}\| + \|R(x)\|) \tag{30}$$

whenever $d(x, \mathcal{X}) \leq \rho$, where $\kappa_0 = \max\{M\theta, 1\}$. Upon squaring both sides of (30) and using the inequality $(a + b)^2 \leq 2(a^2 + b^2)$, which is valid for all $a, b \in \mathbb{R}$, we have

$$d^2(x, \mathcal{X}) \leq 2\kappa_0^2 \left(\|Ax - \bar{y}\|^2 + \|R(x)\|^2\right) \tag{31}$$

whenever $d(x, \mathcal{X}) \leq \rho$. Since h is strongly convex on any compact subset of $\mathbb{R}^m$, there exists a constant $\sigma > 0$ such that for all $x \in \mathbb{R}^n$ satisfying $d(x, \mathcal{X}) \leq \rho$,

$$\begin{aligned} \sigma\|Ax - \bar{y}\|^2 &\leq \langle \nabla h(Ax) - \nabla h(\bar{y}), Ax - \bar{y} \rangle \\ &= \langle \nabla f(x) - \bar{g}, x - \bar{x} \rangle, \end{aligned} \tag{32}$$

where $\bar{x}$ is the projection of x onto $\mathcal{X}$. Using the convexity of P, for any $u \in \partial P(x + R(x))$ and $v \in \partial P(\bar{x})$, we have

$$\langle u - v, x + R(x) - \bar{x} \rangle \geq 0. \tag{33}$$

Due to (28) and the optimality of $\bar{x}$, we can take $u = -\nabla f(x) - R(x)$ and $v = -\bar{g}$ in (33) to get

$$\langle \nabla f(x) - \bar{g}, x - \bar{x} \rangle + \|R(x)\|^2 \leq \langle \bar{g} - \nabla f(x) + \bar{x} - x, R(x) \rangle.$$

Since $\|R(x)\|^2 \geq 0$ and ∇f is Lipschitz continuous, by the Cauchy-Schwarz inequality, there exists a constant $\kappa_1 > 0$ such that

$$\langle \nabla f(x) - \bar{g}, x - \bar{x} \rangle \leq \kappa_1 \|x - \bar{x}\| \cdot \|R(x)\|.$$

Combining this with (31) and (32), we see that there exists a constant $\kappa_2 > 0$ such that for all $x \in \mathbb{R}^n$ satisfying $d(x, \mathcal{X}) \leq \rho$,

$$d^2(x, \mathcal{X}) \leq \kappa_2 \left(\|x - \bar{x}\| \cdot \|R(x)\| + \|R(x)\|^2\right).$$

Upon solving this quadratic inequality, we obtain a constant $\kappa > 0$ such that

$$d(x, \mathcal{X}) \leq \kappa \|R(x)\|$$

whenever $d(x, \mathcal{X}) \leq \rho$. This completes the proof. □

7.4. Proof of Lemma 2

Since the epigraph of P is polyhedral, it can be represented as

$$\text{epi}(P) = \{(z, w) \in \mathbb{R}^n \times \mathbb{R} \mid C_z z + C_w w \leq d\},$$

where $C_z \in \mathbb{R}^{l\times n}$ and $C_w, d \in \mathbb{R}^l$ for some $l \geq 1$. We claim that for any $x, g \in \mathbb{R}^n$, $-g \in \partial P(x)$ if and only if there exists a scalar $s \in \mathbb{R}$ such that (x, s) is an optimal solution to the following linear program:

$$\begin{array}{ll} \min & \langle g, z\rangle + w \\ \text{s.t.} & C_z z + C_w w \leq d, \\ & z \in \mathbb{R}^n,\ w \in \mathbb{R}. \end{array} \tag{34}$$

Indeed, if $-g \in \partial P(x)$, then by definition,

$$P(z) \geq P(x) - \langle g, z - x\rangle, \quad \forall z \in \text{dom}(P).$$

Upon rearranging, we have

$$P(x) + \langle g, x\rangle \leq P(z) + \langle g, z\rangle \leq w + \langle g, z\rangle, \quad \forall (z, w) \in \text{epi}(P).$$

This implies that $(x, P(x))$ is an optimal solution to (34). Conversely, if (x, s) is an optimal solution to (34), then $s = P(x)$ because otherwise $(x, P(x))$ is a feasible solution to (34) with lower objective value. Hence,

$$P(x) + \langle g, x\rangle \leq P(z) + \langle g, z\rangle, \quad \forall z \in \text{dom}(P),$$

which, by the definition of subgradient, implies that $-g \in \partial P(x)$. This establishes the claim. Now, using (5) and the optimality conditions of the linear program (34), we have

$$\Sigma(y, g) = \left\{x \mid (x, s, \gamma) \in \mathcal{S}(y, g) \ \text{ for some } s \in \mathbb{R}, \gamma \in \mathbb{R}^l\right\}, \tag{35}$$

where

$$\mathcal{S}(y, g) = \left\{ (z, w, \lambda) \left| \begin{array}{r} Az = y, \\ C_z^T\lambda + g = \mathbf{0}, \\ C_w^T\lambda + 1 = 0, \\ \lambda \geq \mathbf{0}, \\ C_z z + C_w w \leq d, \\ \langle \lambda, C_z z + C_w w - d\rangle = 0 \end{array} \right. \right\}.$$

The set-valued function $\mathcal{S}$ is a polyhedral multifunction because $\text{gph}(\mathcal{S})$, which is a subset of $\mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^l$, is a finite union of polyhedral convex sets. Moreover, we see from (35) that $\text{gph}(\Sigma)$ is the projection of $\text{gph}(\mathcal{S})$ onto $\mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^n$. Hence, $\text{gph}(\Sigma)$ is also a finite union of polyhedral convex sets, which implies that Σ is a polyhedral multifunction. □

7.5. Proof of Lemma 3

The bound on $d(x, \mathcal{P}_1)$ follows from the well-known Hoffman bound; see, e.g., Lemma 2.2 of (Luo & Tseng, 1992). To prove the bound on $d(x, \mathcal{P}_2)$, recall that by definition,

$$d(x, \mathcal{P}_2) = \min_{v\in\mathcal{P}_2} \|x - v\|.$$

Consider a fixed $x \in \mathbb{R}^n$ and $p \in [1, \infty]$. It is clear that $d(x, \mathcal{P}_2) = 0$ if $x_J = \mathbf{0}$. Hence, suppose that $x_J \neq \mathbf{0}$. Set

$$v_J = \begin{cases} -\dfrac{\|x_J\|_p}{\|d_J\|_p} \cdot d_J & \text{if } d_J \neq \mathbf{0}, \\ \mathbf{0} & \text{otherwise,} \end{cases}$$

and $v_{J^c} = x_{J^c}$, where $J^c = \{1, \ldots, n\} \setminus J$. Then, we have $v \in \mathcal{P}_2$. Moreover,

$$d(x, \mathcal{P}_2) \leq \|x - v\| = \|x_J - v_J\| = \begin{cases} \|x_J\|_p \cdot \left\| \dfrac{d_J}{\|d_J\|_p} + \dfrac{x_J}{\|x_J\|_p} \right\| & \text{if } d_J \neq \mathbf{0}, \\ \|x_J\| & \text{otherwise.} \end{cases}$$

Using the convention that $u/\|u\|_p = \mathbf{0}$ if $u = \mathbf{0}$, we can summarize the above results as

$$d(x, \mathcal{P}_2) \leq \|x_J\|_p \cdot \left\| \frac{d_J}{\|d_J\|_p} + \frac{x_J}{\|x_J\|_p} \right\|.$$

Since the above inequality holds for arbitrary $x \in \mathbb{R}^n$ and $p \in [1, \infty]$, the proof is completed. □

7.6. Proof of Proposition 3

Consider a fixed $p \in (1, \infty)$. For any $x \in \mathbb{R}^n$, we have

$$\partial \|x\|_p = \begin{cases} \dfrac{1}{d(x)} \left(\text{sgn}(x_1)|x_1|^{p-1}, \ldots, \text{sgn}(x_n)|x_n|^{p-1} \right) & \text{if } x \neq \mathbf{0}; \\ \{z \in \mathbb{R}^n \mid \|z\|_q \leq 1\} & \text{if } x = \mathbf{0}, \end{cases} \tag{36}$$

where $d(x) = (\sum_{i=1}^n |x_i|^p)^{\frac{1}{q}}$. From the above expression, we see that for any $z \in \partial \|x\|_p$,

$$\begin{aligned} \|z\|_q &= 1 \quad \text{if } x \neq \mathbf{0}; \\ \|z\|_q &\leq 1 \quad \text{if } x = \mathbf{0}. \end{aligned}$$

Hence, if $-g \in \omega \partial \|x\|_p$ for some $x \in \mathbb{R}^n$, then $\|g\|_q \leq \omega$. In particular, we have $\mathcal{S} = \emptyset$ when $\|g\|_q > \omega$ and $\mathcal{S} = \{\mathbf{0}\}$ when $\|g\|_q < \omega$. On the other hand, if $\|g\|_q = \omega$, then either $x = \mathbf{0}$ or

$$-g = \frac{\omega}{d(x)} \left(\text{sgn}(x_1)|x_1|^{p-1}, \ldots, \text{sgn}(x_n)|x_n|^{p-1} \right). \tag{37}$$

In either case, we have

$$x_i = -\text{sgn}(g_i) \left(\frac{|g_i|}{\omega} d(x) \right)^{\frac{1}{p-1}} = -\text{sgn}(g_i) \cdot \left(\frac{|g_i|}{\omega} \right)^{\frac{q}{p}} \cdot \|x\|_p, \quad i = 1, \ldots, n.$$

This shows that if $x_1, x_2 \neq \mathbf{0}$ and x_1, x_2 satisfies (37), then x_1 must be a positive multiple of x_2. Now, observe that $-v(g)$ satisfies (37). Hence, we conclude that

$$\mathcal{S} = \{x \in \mathbb{R}^n \mid x \text{ is a non-negative multiple of } -v(g)\} = \{x \in \mathbb{R}^n \mid x = a \cdot v(g),\ a \leq 0\}.$$

Lastly, if $p \in (1, 2]$, then $q/p \geq 1$. In this case, the function $t \mapsto \text{sgn}(t)|t|^{\frac{q}{p}}$ is continuously differentiable and hence locally Lipschitz. Thus, for any $t \in \mathbb{R}$, there exist constants $\nu > 0$ and $\delta > 0$ such that

$$\left| \text{sgn}(s)|s|^{\frac{q}{p}} - \text{sgn}(t)|t|^{\frac{q}{p}} \right| \leq \nu |s - t| \quad \text{whenever } |s - t| \leq \delta.$$

This implies (6). □

7.7. Proof of Lemma 5

Using (5) and (7), we can write

$$\Sigma(y, g) = \{x \in \mathbb{R}^n \mid Ax = y\} \cap C(g),$$

where

$$C(g) := \{x \in \mathbb{R}^n \mid -g \in \partial P(x)\} = \left\{ x \in \mathbb{R}^n \;\middle|\; \begin{array}{l} x_J = a_J \cdot v(g_J),\ a_J \leq 0,\ \forall J \in \mathcal{J}_1^g, \\ x_J = \mathbf{0},\ \forall J \in \mathcal{J}_2^g \end{array} \right\}.$$

Let $\mathcal{N}(A)$ be the nullspace of A. The following proposition provides a characterization of the boundedness of $\Sigma(y, g)$:

Proposition 5 *Suppose that $\Sigma(y,g)$ is non-empty. Then, $\Sigma(y,g)$ is bounded if and only if $\mathcal{N}(A)\cap C(g)=\{\mathbf{0}\}$.*

Proof Let $x\in\Sigma(y,g)$ be arbitrary. Suppose there exists a vector $d\in\mathbb{R}^n\setminus\{\mathbf{0}\}$ such that $d\in\mathcal{N}(A)\cap C(g)$. Since $C(g)$ is a convex cone, for any $s\geq 0$, we have $x+sd\in C(g)$. Moreover, we have $A(x+sd)=Ax=y$. It follows that $x+sd\in\Sigma(y,g)$ for all $s\geq 0$; *i.e.*, $\Sigma(y,g)$ is unbounded.

Conversely, suppose that $\Sigma(y,g)$ is unbounded. Since $\Sigma(y,g)$ is a non-empty closed convex set, by Theorem 8.4 of (Rockafellar, 1970), there exists a vector $d\in\mathbb{R}^n\setminus\{\mathbf{0}\}$ such that for any $x\in\Sigma(y,g)$ and $s\geq 0$, $x+sd\in\Sigma(y,g)$. This implies that $Ad=\mathbf{0}$ and $d\in C(g)$, which in turn implies that $d\in\mathcal{N}(A)\cap C(g)$. □

In view of Proposition 5, it suffices to show the existence of a constant $\delta>0$ such that whenever $(\tilde{y},\tilde{g})\in\mathbb{R}^m\times\mathbb{R}^n$ satisfies $\|\tilde{g}-g\|\leq\delta$ and $\Sigma(\tilde{y},\tilde{g})$ is non-empty, we have

$$\mathcal{N}(A)\cap C(\tilde{g})=\{\mathbf{0}\}.$$

Suppose to the contrary that the above does not hold. Then, there exist sequences $\{y^k\}_{k\geq 0}$, $\{g^k\}_{k\geq 0}$, and $\{d^k\}_{k\geq 0}$ such that $\Sigma(y^k,g^k)$ is non-empty and $\mathbf{0}\neq d^k\in\mathcal{N}(A)\cap C(g^k)$ for all $k\geq 0$, and that $g^k\to g$. Since both $\mathcal{N}(A)$ and $C(g^k)$ are cones, we have

$$\bar{d}^k:=\frac{d^k}{\|d^k\|}\in\mathcal{N}(A)\cap C(g^k).$$

Note that $\|\bar{d}^k\|=1$ for all $k\geq 0$. Thus, by passing to a subsequence if necessary, we may assume that $\bar{d}^k\to\bar{d}$ for some $\bar{d}\in\mathbb{R}^n\setminus\{\mathbf{0}\}$. Clearly, we have $\bar{d}\in\mathcal{N}(A)$. Moreover, by definition of $C(g^k)$ and the fact that $\bar{d}^k\in C(g^k)$, we have $-g^k\in\partial P(\bar{d}^k)$ for all $k\geq 0$. Since $g^k\to g$ and $\bar{d}^k\to\bar{d}$, Theorem 24.4 of (Rockafellar, 1970) implies that $-g\in\partial P(\bar{d})$, or equivalently, $\bar{d}\in C(g)$. It follows that $\mathbf{0}\neq\bar{d}\in\mathcal{N}(A)\cap C(g)$, which, together with Proposition 5, contradicts the boundedness of $\Sigma(y,g)$. This completes the proof of Lemma 5. □

7.8. Proof of Proposition 4

For simplicity and consistency, let $f(x)=\frac{1}{2}\|Ax-b\|^2$ and $P(x)=\|x\|_p$, where $p\in(2,\infty)$. We first show that $\bar{x}=(1,0)$ is an optimal solution to problem (20). Indeed, using (36), we have

$$\nabla f(\bar{x})=(-1,0),\quad \partial P(\bar{x})=(1,0).$$

Thus, $\mathbf{0}\in\nabla f(\bar{x})+\partial P(\bar{x})$, which implies the optimality of $\bar{x}$. Next, we show that $\bar{x}=(1,0)$ is the only optimal solution to problem (20); *i.e.*, $\mathcal{X}=\{\bar{x}\}$. Let $\tilde{x}\in\mathcal{X}$ be arbitrary. Since Ax is invariant over $\mathcal{X}$, we have

$$A\tilde{x}=A\bar{x}=1,$$

which implies that $\tilde{x}_1=1$. Moreover, since $\nabla f(x)$ is also invariant over $\mathcal{X}$, we have $\nabla f(\tilde{x})=\nabla f(\bar{x})=(-1,0)$. Now, the optimality of $\tilde{x}$ yields $(1,0)\in\partial P(\tilde{x})$. This, together with Proposition 3, implies that $\tilde{x}$ is a non-negative multiple of $(1,0)$. Since $\tilde{x}_1=1$, we conclude that $\tilde{x}=(1,0)=\bar{x}$, as desired. Finally, we have $v^*=f(\bar{x})+P(\bar{x})=3/2$. □

7.9. Proof of Lemma 6

By definition of $R(x^k)$, we have

$$\mathbf{0}\in\nabla f(x^k)+R(x^k)+\partial P\left(x^k+R(x^k)\right).$$

Adding x^k to both sides and rearranging, we get

$$x^k-\nabla f(x^k)\in x^k+R(x^k)+\partial P\left(x^k+R(x^k)\right),\tag{38}$$

which is a relationship of the form $u\in(I+\partial P)(z)$. Since ∂P is a maximal monotone operator (see, e.g., (Minty, 1964)), a result of Minty (Minty, 1962) states that given any $u\in\mathbb{R}^n$, there exists a unique vector $z=z(u)\in\mathbb{R}^n$ such that $u\in(I+\partial P)(z(u))$. Thus, it remains to show that $R(x^k)=(0,-\delta_k^{1/q})$ satisfies (38).

To begin, we use the definition of x^k and the fact that $\nabla f(x)=(x_1-2,0)$ to compute

$$x^k-\nabla f(x^k)=(2,x_2^k)=\left(2,\frac{2-(1-\delta_k)^{\frac{1}{q}}}{(1-\delta_k)^{\frac{1}{p}}}\cdot\delta_k^{\frac{1}{p}}+\delta_k^{\frac{1}{q}}\right).$$

Now, let $z^k = x^k + (0, -\delta_k^{1/q})$. Then,

$$z^k = \left(2 - (1-\delta_k)^{\frac{1}{q}}, \frac{2-(1-\delta_k)^{\frac{1}{q}}}{(1-\delta_k)^{\frac{1}{p}}} \cdot \delta_k^{\frac{1}{p}}\right) = \frac{2-(1-\delta_k)^{\frac{1}{q}}}{(1-\delta_k)^{\frac{1}{p}}} \left((1-\delta_k)^{\frac{1}{p}}, \delta_k^{\frac{1}{p}}\right).$$

Using (36), it can be verified that for $p \in (2, \infty)$,

$$\partial P(z^k) = \left((1-\delta_k)^{\frac{p-1}{p}}, \delta_k^{\frac{p-1}{p}}\right) = \left((1-\delta_k)^{\frac{1}{q}}, \delta_k^{\frac{1}{q}}\right).$$

It follows that

$$z^k + \partial P(z^k) = \left(2, \frac{2-(1-\delta_k)^{\frac{1}{q}}}{(1-\delta_k)^{\frac{1}{p}}} \cdot \delta_k^{\frac{1}{p}} + \delta_k^{\frac{1}{q}}\right) = x^k - \nabla f(x^k). \tag{39}$$

Upon comparing (38) and (39), we conclude that $R(x^k) = (0, -\delta_k^{1/q})$, as desired. □