Compressive Sensing with a Multiple Convex Sets Domain

Hang Zhang*, Afshin Abdi*, and Faramarz Fekri*.
*School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA.

Abstract—In this paper, we study a general framework for compressive sensing assuming the existence of the prior knowledge that \( x^* \) belongs to the union of multiple convex sets, \( x^* \in \bigcup_i \mathcal{C}_i \). In fact, by proper choices of these convex sets in the above framework, the problem can be transformed to well known CS problems such as the phase retrieval, quantized compressive sensing, and model-based CS. First we analyze the impact of this prior knowledge on the minimum number of measurements \( M \) to guarantee the uniqueness of the solution. Then we formulate a universal objective function for signal recovery, which is both computationally inexpensive and flexible. Then an algorithm based on multiplicative weight update recovery, which is both computationally inexpensive and flexible. Finally, we investigate as to how we can improve the signal recovery by introducing regularizers into the objective function.

I. INTRODUCTION

In the traditional compressive sensing (CS) \([\mathbb{I} \), sparse signal \( x \) is reconstructed via

\[
\min_x \|x\|_1, \quad \text{s.t.} \quad y = Ax, \tag{1}
\]

where \( y \in \mathbb{R}^M \) denotes the measurement vector and \( A \in \mathbb{R}^{M \times N} \) is the measurement matrix.

In this paper, we assume the existence of extra prior knowledge that \( x \) lies in the union of some convex sets, \( x \in \bigcup_{i=1}^L \mathcal{C}_i \), where \( L \) denotes the number of constraint sets and \( \mathcal{C}_i \) is the \( i \)-th convex constraint set. Therefore, we now wish to solve

\[
\min_x \|x\|_1, \quad \text{s.t.} \quad y = Ax, \quad x \in \bigcup_{i=1}^L \mathcal{C}_i \tag{2}
\]

This ill-posed inverse problem (i.e., given measurement \( y \), solving for \( x \)) turns out to be a rather general form of CS. For example, setting \( \bigcup_{i=1}^L \mathcal{C}_i = \mathbb{R}^n \) simplifies our problem to the traditional CS problem. In the following, we will further show that by appropriate choices of these convex sets, \( \mathcal{C}_i \) can be transformed to the phase retrieval \([\mathbb{I} \), quantized compressive sensing \([\mathbb{A} \), or model-based CS \([\mathbb{5} \) problems.

A. Relation with other problems

a) Phase retrieval: Consider the noiseless phase retrieval problem in which the measurements are given by

\[
y_i = |\langle \alpha_i, x \rangle|^2, \quad 1 \leq i \leq l,
\]

where \( y_i \) is the \( i \)-th measurement and \( \alpha_i \) denotes the corresponding coefficients. Considering the first measurement, the constraint \( \sqrt{y_i} = |\langle \alpha_i, x \rangle| \) can be represented via \( x \in \mathcal{B}_+^{(1)} \bigcup \mathcal{B}_-^{(1)} \) where \( \mathcal{B}_+^{(1)} = \{x : \langle \alpha_1, x \rangle = +\sqrt{y_1} \} \) and \( \mathcal{B}_-^{(1)} = \{x : \langle \alpha_1, x \rangle = -\sqrt{y_1} \} \). Following these steps, the constraints \( \{y_i = |\langle \alpha_i, x \rangle|^2\}_{i=1}^l \) can be transformed to \( x \in \bigcap_{i=1}^l (\mathcal{B}_+^{(i)} \bigcup \mathcal{B}_-^{(i)}) = \bigcup_{j=1}^{l'} \mathcal{C}_j \), where \( \mathcal{C}_j \) denotes the corresponding \( j \)-th constraint set. Setting sensing matrix \( A = 0 \) will restore the phase retrieval to our setting.

b) Quantized compressive sensing: In this scenario, the measurements are quantized, i.e.,

\[
y_i = Q(\langle \alpha_i, x \rangle), \quad 1 \leq i \leq L,
\]

where \( Q(\cdot) \) is the quantizer. By defining \( \mathcal{C}_i := \{x : \langle \alpha_i, x \rangle \in Q^{-1}(y_i)\} \), the quantized CS can be easily transformed to \( \mathcal{B}_+^{(i)} \bigcup \mathcal{B}_-^{(i)} \).

c) Model-based compressive sensing: These lines of works \([\mathbb{5} \), \([\mathbb{6} \), \([\mathbb{7} \) are the most similar work to our model, where they consider

\[
y = Ax, \quad x \in \bigcup_{i=1}^L \mathcal{L}_i.
\]

Here, \( \mathcal{L}_i \) is assumed to be a linear space whereas the only assumption we make on the models is being a convex set. Hence, their model can be regarded as a special case of our problem.

In \([\mathbb{5} \), the author studied the minimum number of measurements \( M \) under different models, i.e., shape of \( \mathcal{L}_i \), and modified CoSaMP algorithms \([\mathbb{1} \) to reconstruct signal. In \([\mathbb{6} \), the authors expanded the signal onto different basis and transformed model-based CS to be block-sparse CS. In \([\mathbb{7} \), the author studied model-based CS with incomplete sensing matrix information and reformulated it as a matrix completion problem.

B. Our contribution:

a) Statistical Analysis: We analyze the minimum number of measurements to ensure uniqueness of the solution. We first show that the conditions for the uniqueness can be represented as \( \min_{u \in E} \|Au\|_2 > 0 \), for an appropriate set \( E \). Assuming the entries of the sensing matrix \( A \) are i.i.d. Gaussian, we relate the probability of uniqueness to the number of measurements, \( M \). Our results show that depending on the structure of \( \mathcal{C}_i \) ’s, the number of measurements can be reduced significantly.
b) Optimization Algorithm: We propose a novel formulation and the associated optimization algorithm to reconstruct the signal $\hat{x}$. First, note that existing algorithms on e.g., model-based CS are not applicable to our problem as they rely heavily on the structure of constraint sets. For example, a key idea in model-based CS is to consider expansion of $\hat{x}$ onto the basis of each $\mathcal{C}_i$ and then rephrase the constraint as the block sparsity on the representation of $\hat{x}$ on the union of bases. However, such an approach may add complicated constraints on the coefficients of $\hat{x}$ in the new basis, as the sets $\mathcal{C}_i$’s are not necessarily simple subspaces.

Note that although $\mathcal{C}_i$’s are assumed to be convex, their union $\bigcup_i \mathcal{C}_i$ is not necessarily a convex set, which makes the optimization problem hard to solve. By introducing an auxiliary variable, $p$, we convert the non-convex optimization problem to a biconvex problem. Using multiplicative weight update [8] from online learning theory [9], we design an algorithm with convergence speed of $O(T^{-1/2})$ to a local minimum. Further, we investigate improving the performance of the algorithm by incorporating appropriate regularization. Compared to the naive idea of solving $L$ simultaneous optimization problems

$$\min_x \|x\|_1, \quad \text{s.t.} \quad y = Ax, \quad x \in \mathcal{C}_i$$

and choosing the best solution out of $L$ results, our method is computationally less-expensive and more flexible.

Finally, we would like to mention that due to the space limit, here we only present the main results and the detailed proofs are provided in the extended version of this paper, available at [10].

II. System Model

Let $A \in \mathbb{R}^{M \times N}$ be the measurement matrix, and consider the setup

$$y = Ax^*, \quad \text{and} \quad x^* \in \bigcup_{i=1}^L \mathcal{C}_i,$$  \hspace{1cm} (3)

where $x^*$ is a $K$-sparse high-dimensional signal, $y \in \mathbb{R}^M$ is the measurement vector, and $\mathcal{C}_i \subset \mathbb{R}^N$, $i = 1, 2, \ldots, L$, is a convex set.

Due to the sparsity of $x^*$, we propose to reconstruct $x^*$ via

$$\hat{x} = \text{argmin}_x \|x\|_1, \quad \text{s.t.} \quad y = Ax, \quad x \in \bigcup_{i=1}^L \mathcal{C}_i,$$  \hspace{1cm} (4)

where $\hat{x}$ denotes the reconstructed signal. Let $d \triangleq \hat{x} - x^*$ be the deviation of the reconstructed signal $\hat{x}$ from the true signal $x^*$.

In the following, we will study the inverse problem in (4) from two perspectives: the statistical and the computational aspects.

III. Statistical Property

In this section, we will find the minimum number of measurements $M$ to $\hat{x} = x^*$, i.e., $d = 0$. 

**Definition.** The tangent cone $\mathcal{T}_x$ for $\|x\|_1$ is defined as

$$\mathcal{T}_x \triangleq \{e : \|x + te\|_1 \leq \|x\|_1, \exists t \geq 0\}.$$ 

Geometric interpretation of $\mathcal{T}_x$ is that it contains all directions that lead to smaller $\|\cdot\|_1$ originating from $x$. In the following analysis, we use $\mathcal{T}$ as a compact notation for $\mathcal{T}_x$.

**Definition.** The Gaussian width $\omega(\cdot)$ associated with set $U$ is defined as $\omega(U) \triangleq E \sup_{x \in U} (g(x), g \sim \mathcal{N}(0, I))$. [12], [13].

Define cone $\mathcal{C}_{i,j}$ as

$$\mathcal{C}_{i,j} \triangleq \{z \mid z = t(x_1 - x_2), \exists t > 0, x_1 \in \mathcal{C}_i, x_2 \in \mathcal{C}_j\},$$

which denotes the cone consisting of all vectors $z$ that are parallel with $x_1 - x_2$, $x_1 \in \mathcal{C}_i$, and $x_2 \in \mathcal{C}_j$. Then we define event $\mathcal{E}$ as

$$\mathcal{E} \triangleq \bigcup_{i,j} (\text{null}(A) \cap \mathcal{T} \cap \mathcal{C}_{i,j}) = \{0\}.$$ 

**Lemma 1.** If we have event $\mathcal{E}$ to be satisfied, then we can guarantee the correct recovery of $x$, i.e., $\hat{x} = x^*$.

**Proof.** We prove this lemma by showing $d = 0$. Assume that there exists non-zero $e \in \text{null}(A) \cap \mathcal{T} \cap (\bigcup_{i,j} \mathcal{C}_{i,j})$. We can show that signal $x^* + te$, where $t$ is some positive constant such that $\|x^* + te\|_1 \leq \|x^*\|_1$, satisfying constraints described by (3). This implies that $d = te \neq 0$ and the wrong recovery of $x^*$.

Since a direct computation of the probability of event $\mathcal{E}$ can be difficult, we analyze the following equivalent event,

$$\min_{x \in \mathcal{T} \cap (\bigcup_{i,j} \mathcal{C}_{i,j})} \|Ax\|_2 > 0.$$ 

For the simplicity of analysis, we assume that the entries $A_{i,j}$ of $A$ are i.i.d. normal $\mathcal{N}(0, 1)$. Using Gordon’s escape from mesh theorem [12], we obtain the following result that relates $\text{Pr}(\mathcal{E})$ with the number of measurements $M$. The proof can be found in [10].

**Theorem 2.** Let $a_M = E\|g\|_2$, where $g \in \mathcal{N}(0, I_{M \times M})$, and $\omega(\cdot)$ denotes the Gaussian width. Provided that $a_M \geq \omega(\mathcal{T})$ and $(1 - 2\epsilon)a_M \geq \omega(\mathcal{C}_{i,j})$ for $1 \leq i, j \leq L$ and $\epsilon > 0$, we have

$$\text{Pr}(\mathcal{E}) \geq 1 - \min \left\{ \begin{array}{l} \text{Pr} \left\{ \min_{u \in \mathcal{T} \setminus \{0\}} \|Au\|_2 > 0 \right\}, \\
\text{Pr} \left\{ \min_{u \in \bigcap_{i,j} \mathcal{C}_{i,j} \setminus \{0\}} \|Au\|_2 > 0 \right\} \end{array} \right\}.$$
where $\mathcal{P}_1$ and $\mathcal{P}_2$ are defined as
\[
\mathcal{P}_1 \leq \min \left(1, \exp \left(-\frac{(a_m - \omega(T))^2}{2}\right)\right)
\]
\[
\mathcal{P}_2 \leq \min \left(1, \frac{3}{2} \exp \left(-\frac{c^2 a_M^2}{2}\right) + \sum_{i<j} \exp \left(-\frac{(1-2\varepsilon)a_M - \omega(\mathcal{C}_{ij})^2}{2}\right)\right).
\]

Here $\mathcal{P}_1$ is associated with the descent cone of the optimization function, namely, $\|x\|_1$, while $\mathcal{P}_2$ is associated with the prior knowledge $x \in \bigcup_i \mathcal{C}_i$ [Theorem 2] implies that event $\mathcal{E}$ (uniqueness) holds with higher probability than the traditional CS due to the extra constraint $x \in \bigcup_i \mathcal{C}_i$. If we fix $\Pr(\mathcal{E})$, we can separately calculate the corresponding $M$ with and without the constraint $x \in \bigcup_i \mathcal{C}_i$. The difference $\Delta M$ would indicate the savings in the number of measurements due to the additional structure $x \in \bigcup_i \mathcal{C}_i$ over the traditional CS. Here we give a simple illustrative example.

Example 3. Consider the constraint set
\[
\mathcal{C}_i = \{(0, \cdots, 0, x_i, \cdots, x_{K+i}, 0, \cdots, 0)\},
\]
where $1 \leq i \leq N - K$. Here we assume $N = O(K^c)$, where $c > 1$ is constant. In the sequel we will show that [Theorem 2] gives us the order $M = O(K)$ to ensure solution uniqueness as $K$ approaches infinity, which gives us the same bound as shown in [5].

Setting $\varepsilon = 1/4$, we can bound $\mathcal{P}_2$ as
\[
\mathcal{P}_2 \leq \frac{3}{2} \exp \left(-\frac{a_M^2}{32}\right) + \frac{N^2}{2} \exp \left(-\frac{(a_M - 2a_M^2K)^2}{8}\right),
\]
provided $a_M \geq 2a_K$. With the relation $\frac{M}{\sqrt{M+1}} \leq a_M \leq \sqrt{M}$ [11], [12] and setting $M = 3K$, we have
\[
\mathcal{P}_2 \leq c_1 \exp(-c_2K) + c_3N^2 \exp(-c_4K),
\]
where $c_1, c_2, c_3, c_4 > 0$ are some positive constants. Since $N = O(K^c)$, we can see $\mathcal{P}_2$ shrinks to zero as $K$ approaches infinity, which implies the solution uniqueness.

Comparing with the traditional CS theory without prior knowledge $x \in \bigcup_i \mathcal{C}_i$, our bound reduces the number of measurements from $M = O(K \log N/K) = O(K \log K)$ to $M = O(K)$.

IV. COMPUTATIONAL ALGORITHM

Apart from the statistical property, another important aspect of [8] is to design an efficient algorithm. One naive idea is to consider and solve $L$ separate optimization problems
\[
\hat{x}^{(i)} = \arg \min_x \|x\|_1, \ y = Ax, \ x \in \mathcal{C}_i,
\]
and then selecting the best one, i.e., the sparsest reconstructed signal among all $\hat{x}^{(i)}$'s. However, this method has two drawbacks:

- It does not exploit the fact that $\|x\|_1$ and constraint $y = Ax$ remains the same for all $\mathcal{C}_i$ and performs similar updates for all $\mathcal{C}_i$ repeatedly, which lead to prohibitively high computational cost when $L$ is large.

- It is inflexible. For example, some prior knowledge of which $\mathcal{C}_i$ the true signal $x^*$ is more likely to reside might be available. The above method cannot incorporate such priors.

To overcome the above drawbacks, we (i) reformulate (4) to a more tractable objective function, and (ii) propose a computationally efficient algorithm to solve it. In the following, we assume that $x$ is bounded in the sense that for a constant $R$, $\|x\|_2 \leq R$.

A. Reformulation of the objective function

We introduce an auxiliary variable $p$ and rewrite the Lagrangian form in (4) as
\[
\min_{x} \min_{p \in \Delta_L} \sum_i p_i \left(\|x\|_1 + \hat{1}(x \in \mathcal{C}_i) + \frac{\lambda_1}{2}\|y - Ax\|^2_2 + \frac{\lambda_2}{2}\|x\|^2_2\right),
\]
where $\Delta_L$ is the simplex $\{p_i \geq 0, \sum_i p_i = 1\}$, $\hat{1}()$ is the truncated indicator function, which is 0 when its argument is true and is some large finite number $C$ otherwise, and $\lambda_1, \lambda_2 > 0$ are the Lagrange multipliers. The term $\|y - Ax\|^2_2$ is used to penalize for the constraint $y = Ax$ while $\|x\|^2_2$ corresponds to the energy constraint $\|x\|^2_2 \leq R$. It can be easily shown that solving (5) for large enough $C$ ensures $x \in \bigcup_i \mathcal{C}_i$.

Apart from the universality, our formulation is very flexible. We can easily adjust to the case that $x$ belongs to the intersection, i.e., $x \in \bigcap_i \mathcal{C}_i$ via modifying $\min_{p \in \Delta_L}$ in (5) to $\max_{p \in \Delta_L}$. Besides, to the best of our knowledge, this is the first time that such a formulation (5) is proposed.

In the following, we will focus on the computational methods. Note that the difficulties in solving (5) are due to two aspects:

- Optimization over $p$: Although classical methods to minimize over $p$ with fixed $x$, e.g., alternative minimization and ADMM [14], can calculate local minimum efficiently (due to the bi-convexity of (5)), they can be easily trapped in the local-minima. This is because some entries in $p$ can be set to zero and hence $x$ will be kept away from the corresponding set $\mathcal{C}_i$ thereafter. To handle this problem, we propose to use multiplicative weight update [8] and update $p$ with the relation $p^{(t+1)}_i \propto p^{(t)}_i e^{-\eta p^{(t)}_i f_i(x)}$, where $p^{(t)}$ denotes $p$'s value in the $t$th iteration. This update relation avoids the sudden change of $p^{(t)}$'s entries from non-zero to zero, which could have forced $x^{(t)}$ being trapped in a local minimum.

- Optimization over $x$: Due to the non-smoothness of $\hat{1}(x \in \mathcal{C}_i)$ and $\|x\|_1$ in (5) and the difficulties in calculating their sub-gradients, directly minimizing (5) would be computationally prohibitive. We propose to first approximate $\hat{1}(x \in \mathcal{C}_i)$ with a smooth function $h_i(x)$ and update $x^{(t)}$ with the relation (8) used in proximal gradient descent [15].
**Definition** ($L_g$-strongly smooth [15]). Function $g(\cdot) : \mathcal{X} \rightarrow \mathbb{R}$ is $L_g$-strongly smooth iff

$$g(y) \leq g(x) + \langle \nabla g(x), y - x \rangle + \frac{L_g}{2}\|x - y\|^2,$$

for all $x, y$ in the domain $\mathcal{X}$.

**B. Non-convex Proximal Multiplicative Weighting Algorithm**

Here we directly approximate the truncated indicator function $\mathbb{1}_{\{x \in \mathcal{C}_i\}}$ by $L_{h,i}$ strongly-smooth convex penalty functions $h_i(x)$, which may be different for different shapes of convex sets. For example, consider the convex set $\mathcal{C}_i$ in **Example 3**. We may define $h_i(x) = \sum_{j \in [i, i + K]} x_j$, where $[a, b]$ denotes the region from $a$ to $b$. While for the set $\{x : \langle a, x \rangle \leq b\}$, we may instead adopt the modified log-barrier function with a finite value. Then (5) can be rewritten as

$$\min_{p} \min_{x} \mathcal{L}(p, x) \triangleq \sum_{i=1}^{L} p_i f_i(x), \tag{6}$$

where $f_i(x)$ is defined as

$$f_i(x) \triangleq \|x\|_1 + h_i(x) + \frac{\lambda_1}{2}\|y - Ax\|^2 + \frac{\lambda_2}{2}\|x\|^2.$$

Hence, the optimization problem in (6) can be solved via Alg. [11]

**Algorithm 1 Non-convex Proximal Multiplicative Weighting Algorithm**

- **Initialization**: Initialize all variables with uniform weight $p_i^{(0)} = L^{-1}$ and $x^{(0)} = 0$.
- **Time $t$**: We update $p_i^{(t+1)}$ as

$$p_i^{(t+1)} \propto p_i^{(t)} e^{-\eta_i^{(t)} f_i(x^{(t)})}, \tag{7}$$

where $p_i^{(t)}$ denotes the $i$th element of $p^{(t)}$. Then we adopt proximal gradient descent for $x^{(t)}$ update

$$x^{(t+1)} = \text{prox}_{\eta_i^{(t)} \|x\|_1} \left[ x^{(t)} - \eta_i^{(t)} \sum_i p_i^{(t)} \left( \nabla_x h_i(x^{(t)}) + \lambda_1 A^T(Ax - y) + \lambda_2 x^{(t)} \right) \right], \tag{8}$$

where the proximal operator $\text{prox}_{\|x\|_1}(x)$ is defined as [10] $\arg\min_x \left( \|x\|_1 + 1/2\|x - x^{(t)}\|^2 \right)$.
- **Output**: Calculate the average value $ar{p} = (\sum_i p_i^{(t)}) / T$ and $\bar{x} = (\sum_i x^{(t)}) / T$. Then output $\bar{x}$ by projecting $\bar{x}$ onto the set of $\cup_i \mathcal{C}_i$.

**Lemma 4.** $h(x) \triangleq \sum_i p_i h_i(x) + \lambda_1\|y - Ax\|^2/2 + \lambda_2\|x\|^2/2$ is strongly-smooth with some positive constant denoted as $L_h$.

**Proof.** First, we can check that $\lambda_1\|y - Ax\|^2/2 + \lambda_2\|x\|^2/2$ is strongly-smooth. Denote the corresponding parameter as $L_{h,0}$. Meanwhile, due to the construction of $h_i(x)$, it is strongly-smooth for every $i$. Since $p_i$ is non-negative for every $i$, we can easily prove the following inequality

$$h(x_1) \geq h(x_2) + \langle \nabla h(x_2), x_1 - x_2 \rangle + \frac{L_h}{2}\|x_1 - x_2\|^2,$$

where $L_h$ is defined as $\min\{L_{h,i}, 0 \leq i \leq L\}$.

Then we have the following theorem. The proof can be found in [10].

**Theorem 5.** Let $\eta_i^{(t)} = \eta_i \leq L_h^{-1}$, and $\eta_p^{(t)} = R_f^{-1} \sqrt{2\log L/T}$, where $|f_i(.)| \leq R_f$, $\|x\|_2 \leq R$. Then we have

$$\frac{\min_p \sum_{i} \mathcal{L}(p, x^{(t)})}{T} \leq \frac{\sum_{i} \mathcal{L}(p^{(t)}, x^{(t)})}{T} + \frac{\sum_{i} \mathcal{L}(p^{(t)}, x^{(t)})}{T} \leq \frac{2R^2}{\eta_i T} + \frac{R_f \log L}{T}.$$

Due to the difficulties in analyzing the global optimum, in Theorem 5 we focus on analyzing the closeness between the average value $(\sum_i \mathcal{L}(p^{(t)}, x^{(t)}) / T)$ to its local minimum. The first term denotes the gap between average value $(\sum_i \mathcal{L}(p^{(t)}, x^{(t)}) / T)$ and the optimal value of $\mathcal{L}(p, x)$ with $x^{(t)}$ being fixed. Similarly, the second term represents the gap with $p^{(t)}$ being fixed. As $T \rightarrow \infty$, the sum of these two bounds approaches zero at the rate of $O(T^{-1/2})$.

Moreover note that setting $\eta_p^{(t)}$ requires the oracle knowledge of $T$, which is impractical. This artifacts can easily be fixed by the doubling trick [9] §2.3.1. In addition, we have proved the following theorem [10].

**Theorem 6.** Let $\eta_i^{(t)} \leq L_h^{-1}$, where $|f_i(.)| \leq R_f$. Then we have

$$\frac{1}{T} \sum_{t} \|x^{(t+1)} - x^{(t)}\|^2 \leq \frac{2\mathcal{L}(p^{(0)}, x^{(0)})}{L_h T} + \frac{4R^2 \sum_i \eta_i^{(t)}}{L_h T},$$

This theorem discusses the convergence speed with respect to the $x^{(t)}$ update. Due to the $O(T^{-1/2})$ of the first term on the right side of the above inequality, the best convergence rate we can obtain is $O(T^{-1})$, which is achievable by $\eta_p^{(t)} \propto t^{-2}$. However, using fixed learning rate $\eta_p$ as in Theorem 5 would result in the convergence rate of $O(T^{-1/2})$.

**C. Regularization for $p$**

Note that we can interpret $p_i$, the $i$th element of $p$ in (6) as the likelihood of $x^* \in \mathcal{C}_i$. Without any prior knowledge about which set $\mathcal{C}_i$ the true signal $x^*$ resides, variable $p$ is uniformly distributed among all possible distributions $\Delta_L$. When certain prior information is available, its distribution is skewed towards certain distributions, namely $q$.

Here we would like to exploit such prior knowledge and add regularizers to penalize $p$ towards $q$. Meanwhile our analysis
shows that such a regularizer may also contribute towards the performance improvement. In this paper, we adopt $\| \cdot \|_2$, and hence the modified function $\mathcal{L}(p, x)$ can be written as

$$\mathcal{L}(p, x) = \mathcal{L}(p) + \frac{\lambda_3}{2} \| p - q \|_2^2,$$

where $\lambda_3 > 0$ is a constant used to balance $\mathcal{L}(p, x)$ and $\frac{1}{2} \| p - q \|_2^2$. Based on different applications, other norms such as KL-divergence or $l_1$ norm can be used as the regularizer.

Then we substitute the update equation (7) as

$$p^{(t+1)} = p^{(t)} - \eta_p^{(t)} g^{(t)},$$

where $g^{(t)} = \nabla_{p^{(t)}} \mathcal{L}(p, x^{(t)}) = f(x^{(t)}) + \lambda_3 (p^{(t)} - q)$, and $f(x^{(t)})$ denotes the vector whose $i$th element is $f_i(x^{(t)})$. Similar as above, we obtain the following theorems.

**Theorem 7.** Provided that $\| g^{(t)} \|_2 \leq R_g$, by setting $\eta^{(t)} = \eta_w \leq L_h^{-1}$ and $\eta_p^{(t)} = \lambda_h^{-1}$ we conclude that

$$\frac{\min_p \sum_t \mathcal{L}(p, x^{(t)})}{T} \leq \frac{\sum_t \mathcal{L}(p^{(t)}, x^{(t)})}{T} - \frac{\sum_t \mathcal{L}(p^{(t)}, x^{(t)})}{T} \leq \frac{R_g^2 \log T}{2 \lambda_3 T} + \frac{R_g^2}{2 \eta_w T}.$$

Comparing with **Theorem 5** implies that the regularizers improve the optimal rate from $\mathcal{O}(T^{-1/2})$ to $\mathcal{O}(\log T/T)$. Therefore, our framework can exploit the prior information to improve the recovery performance whereas the naive method of iterative computation fails to achieve as such.

**Theorem 8.** Provided that $\| g^{(t)} \|_2 \leq R_g$, by setting $\eta^{(t)} = \eta_w \leq L_h^{-1}$ we conclude that

$$1 \leq T \sum_t \| x^{(t+1)} - x^{(t)} \|_2^2 \leq \frac{2 \mathcal{L}(p^{(0)}, x^{(0)})}{L_h T} + \frac{2R_g^2}{L_h T} \sum_t \left( \eta_p^{(t)} + \frac{\lambda_3 (\eta_p^{(t)})^2}{2} \right).$$

In this case, if we set $\eta_p^{(t)} = t^{-2}$, then $\frac{1}{T} \sum_t \| x^{(t+1)} - x^{(t)} \|_2^2$ would decrease at the rate of $\mathcal{O}(T^{-1})$, which is the same as **Theorem 6**.

**V. Conclusion**

In this paper, we studied the compressive sensing with a multiple convex-set domain. First we analyzed the impact of prior knowledge $x \in \bigcup_i \mathcal{C}_i$ on the minimum number of measurements $M$ to guarantee uniqueness of the solution. We gave an illustrative example and showed that significant savings in $M$ can be achieved. Then we formulated a universal objective function and develop an algorithm for the signal reconstruction. We show that in terms of the speed of convergence to local minimum, our proposed algorithm based on multiplicative weight update and proximal gradient descent can achieve the optimal rate of $\mathcal{O}(T^{-1/2})$. Further, in terms of $T^{-1} \| x^{(t+1)} - x^{(t)} \|_2^2$, the optimal speed increases to $\mathcal{O}(T^{-1})$. Moreover, provided that we have a prior knowledge about $p$, we show that we can improve the optimal recovery performance by $\| \cdot \|_2$ regularizers, and hence increasing the above convergence rate from $\mathcal{O}(T^{-1/2})$ to $\mathcal{O}(\log T/T)$. Detailed proofs can be found in [10].

**REFERENCES**

[1] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*, vol. 1, no. 3.

[2] Y. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*, ser. Compressed Sensing: Theory and Applications. Cambridge University Press, 2012. [Online]. Available: https://books.google.com/books?id=Gm3hICwN0YC

[3] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski, “Phase retrieval via matrix completion,” 2011.

[4] W. Dai, H. V. Pham, and O. Milenkovic, “Quantized compressive sensing,” arXiv preprint arXiv:0901.0749, 2009.

[5] 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. 2002–2010, 2010.

[6] M. F. Duarte and Y. C. Eldar, “Structured compressed sensing: From theory to applications,” IEEE Transactions on Signal Processing, vol. 59, no. 9, pp. 4053–4085, 2011.

[7] J. Silva, M. Chen, Y. C. Eldar, G. Sapiro, and L. Carin, “Blind compressed sensing over a structured union of subspaces,” arXiv preprint arXiv:1103.2469, 2011.

[8] S. Arora, E. Hazan, and S. Kale, “The multiplicative weights update method: a meta-algorithm and applications.” Theory of Computing, vol. 8, no. 1, pp. 121–164, 2012.

[9] S. Shalev-Shwartz et al., “Online learning and online convex optimization,” Foundations and Trends® in Machine Learning, vol. 4, no. 2, pp. 107–194, 2012.

[10] H. Zhang, A. Abdi, and F. Fekri, “Compressive sensing with a multiple convex sets domain,” arXiv preprint, 2019.

[11] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, “The convex geometry of linear inverse problems,” Foundations of Computational Mathematics, vol. 12, no. 6, pp. 805–849, 2012.

[12] Y. Gordon, “On milman’s inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$,” in Geometric Aspects of Functional Analysis. Springer, 1988, pp. 84–106.

[13] H. Zhang, A. Abdi, and F. Fekri, “Compressive sensing with energy constraint,” in IEEE Information Theory Workshop (ITW’17), Kaohsiung, Taiwan, Nov. 2017.

[14] S. Boyd, “Alternating direction method of multipliers.”

[15] A. Beck, *First-Order Methods in Optimization*, ser. MOS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics, 2017. [Online]. Available: https://books.google.com/books?id=wrkDwAAMAAJ

[16] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” SIAM journal on imaging sciences, vol. 2, no. 1, pp. 183–202, 2009.

[17] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, 2013. [Online]. Available: https://books.google.com/books?id=O3yoAQAACAAJ

[18] Q. Qian, S. Zhu, J. Tang, R. Jin, B. Sun, and H. Li, “Robust optimization over multiple domains,” arXiv preprint arXiv:1805.07588, 2018.
APPENDIX A

PROOF OF THEOREM 2

Proof. Note that for any-vector non-zero \(h \in \text{null}(A) \cap \mathcal{F} \cap \left(\bigcup_{i,j} \tilde{C}_{i,j}\right)\), we can always rescale to make it unit-norm. Hence we can rewrite the event \(\mathcal{E}\) as
\[
\mathcal{E} = \left\{ \text{null}(A) \cap S_2^{n-1} \cap \mathcal{F} \cap \left(\bigcup_{i,j} \tilde{C}_{i,j}\right) = \emptyset \right\}.
\]

For the conciseness of notation, we define \(\tilde{\mathcal{E}}\) to be \(\tilde{\mathcal{E}} = \bigcup_{i,j} \tilde{C}_{i,j}\). Then we upper-bound \(1 - \Pr(\mathcal{E})\) as
\[
1 - \Pr(\mathcal{E}) = \Pr\left(\text{null}(A) \cap S_2^{n-1} \cap \mathcal{F} \cap \tilde{\mathcal{E}} \neq \emptyset\right) \leq \min\left\{ \Pr\left(\text{null}(A) \cap S_2^{n-1} \cap \mathcal{F} \neq \emptyset\right), \Pr\left(\text{null}(A) \cap S_2^{n-1} \cap \tilde{\mathcal{E}} \neq \emptyset\right) \right\},
\]

where \((i)\) is because
\[
\left\{ \text{null}(A) \cap S_2^{n-1} \cap \mathcal{F} \cap \tilde{\mathcal{E}} \neq \emptyset\right\} \subseteq \left\{ \text{null}(A) \cap S_2^{n-1} \cap \mathcal{F} \neq \emptyset\right\},
\]
\[
\left\{ \text{null}(A) \cap S_2^{n-1} \cap \mathcal{F} \cap \tilde{\mathcal{E}} \neq \emptyset\right\} \subseteq \left\{ \text{null}(A) \cap S_2^{n-1} \cap \tilde{\mathcal{E}} \neq \emptyset\right\}.
\]

With Lemma 9 and Lemma 10, we can separately bound \(\mathcal{P}_1\) and \(\mathcal{P}_2\) and finish the proof.

Lemma 9. We have \(\mathcal{P}_1 \leq \min\left\{ 1, \exp\left(-\frac{(a_m - \omega(\mathcal{F}))^2}{2}\right)\right\}\), if \(a_m \geq \omega(\mathcal{F})\).

Proof. Note that we have
\[
\Pr\left(\text{null}(A) \cap S_2^{n-1} \cap \mathcal{F} \neq \emptyset\right) + \Pr\left(\text{null}(A) \cap S_2^{n-1} \cap \mathcal{F} = \emptyset\right) = 1.
\]

Then we lower-bound \(\mathcal{P}_1^c\) as
\[
\mathcal{P}_1^c = \Pr\left\{ \min_{u \in S_2^{n-1} \cap \mathcal{F}} \|Au\|_2 > 0 \right\} \geq 1 - \exp\left(-\frac{(a_m - \omega(\mathcal{F}))^2}{2}\right),
\]

provided \(a_m \geq \omega(\mathcal{F})\), where \((i)\) is because of Corollary 3.3 in [11], and \(\omega(\cdot)\) denotes the Gaussian width.

Lemma 10. If \((1 - 2\epsilon a_M) \geq \omega(\tilde{C}_{ij})\), we have
\[
\mathcal{P}_2 \leq \min\left\{ 1, \frac{3}{2} \exp\left(-\frac{\epsilon^2 a_M^2}{2}\right) + \sum_{i \leq j} \exp\left(-\frac{(1 - 2\epsilon)a_M - \omega(\tilde{C}_{ij})^2}{2}\right)\right\},
\]

Proof. Note that we have
\[
\Pr\{\text{null}(A) \cap S_2^{n-1} \cap \tilde{\mathcal{E}} = \emptyset\} = 1.
\]

Here we upper-bound \(\mathcal{P}_2\) via lower-bounding \(\mathcal{P}_2^c\). First we define \(\mathcal{P}_2^c(d)\) as
\[
\mathcal{P}_2^c(d) \triangleq \Pr\left\{ \min_{u \in \cup_{\mathcal{A}_j, j}, v \in \text{null}(A)} \|u - v\|_2 \geq d \right\}.
\]

Then we have \(\mathcal{P}_2 = \lim_{d \to 0} \mathcal{P}_2^c(d)\). The following proof trick is fundamentally the same as that are used in Theorem 4.1 in [12] but in a clear format by only keeping the necessary parts for this scenario. We only present it for the self-containing of this paper and do not claim any novelties.

We first define \(\mathcal{A}_{i,j} = S_2^{n-1} \cap \tilde{C}_{i,j}\) and two quantities \(Q_1\) and \(Q_2\) as
\[
Q_1 \triangleq \Pr\left\{ \min_{u \in \cup_{\mathcal{A}_{i,j}}} \|Au\|_2 \geq d(1 + \epsilon_1)E\|A\|_2\right\},
\]
\[
Q_2 \triangleq \Pr\left\{ \sum_{i \leq j, \mathcal{A}_{i,j}} \left( \frac{\sum_{\tilde{g}_{i,j}}}{2} \right)^{1/2} + \sum_{j} u_{j_2} h_{j_2} \geq d(1 + \epsilon_1)E\|A\|_2 + \epsilon_2 a_M \right\},
\]
where \( a_M = \mathbb{E}\|g\|_2 \), \( g \in \mathcal{N}(0, I_{M \times M}) \), and \( g_j, h_i \) are iid standard normal random variables \( \mathcal{N}(0,1) \). The following proof is divided into 3 parts.

**Step 1.** We prove that \( \mathcal{P}_2(d) + e^{-c_2^2 a_M^2 /2} \geq Q_1 \), which is done by

\[
Q_1 = \Pr \left\{ \min_{u \in \mathcal{F}_{i,j}} \|Au\|_2 \geq d(1 + \epsilon_1)\mathbb{E}\|A\|_2 \right\} \\
= \Pr \left\{ \min_{u \in \mathcal{F}_{i,j}, v \in \text{null}(A)} \|A(u - v)\|_2 \geq d(1 + \epsilon_1)\mathbb{E}\|A\|_2 \right\} \\
\leq \Pr \left\{ \min_{u \in \mathcal{F}_{i,j}, v \in \text{null}(A)} \|A\|_2 \|u - v\|_2 \geq d(1 + \epsilon_1)\mathbb{E}\|A\|_2 \right\} \\
\leq \Pr \left\{ \|A\|_2 \geq (1 + \epsilon_1)\mathbb{E}\|A\|_2 \right\} + \Pr \left\{ \min_{u \in \mathcal{F}_{i,j}, v \in \text{null}(A)} \|u - v\|_2 \geq d \right\} \\
\leq \exp \left( -\frac{c_2^2 (\mathbb{E}\|A\|_2)^2}{2} \right) + \mathcal{P}_2(d)
\]

where in \( (i) \) we use \( \|A\|_2 \|u - v\|_2 \geq \|A(u - v)\|_2 \), in \( (ii) \) we use the union bound for

\[
\left\{ \min_{u \in \mathcal{F}_{i,j}, v \in \text{null}(A)} \|A\|_2 \|u - v\|_2 \geq d(1 + \epsilon_1)\mathbb{E}\|A\|_2 \right\} \\
\subseteq \left\{ \|A\|_2 \geq (1 + \epsilon_1)\mathbb{E}\|A\|_2 \right\} \cup \left\{ \min_{u \in \mathcal{F}_{i,j}, v \in \text{null}(A)} \|u - v\|_2 \geq d \right\},
\]

in \( (iii) \) we use the Gaussian concentration inequality Lipschitz functions (Theorem 5.6 in [17]) for \( \|A\|_2 \), and in \( (iv) \) we use \( \|A\|_2 \geq \|Ae_1\|_2 = \| \sum_{i=1}^{M} A_{i,1} \|_2 \), where \( e_1 \) denotes the canonical basis.

**Step 2.** We prove that \( Q_1 + \frac{1}{2} e^{-c_2^2 a_M^2 /2} \geq Q_2 \), which is done by

\[
Q_1 + \frac{1}{2} e^{-c_2^2 a_M^2 /2} \geq Q_1 + \Pr \{ g \geq \epsilon_2 a_M \} \\
\geq \Pr \left\{ \min_{u \in \mathcal{F}_{i,j_1,j_1}} \|Au\|_2 + g \|u\|_2 \geq d(1 + \epsilon)\mathbb{E}\|A\|_2 + \epsilon_2 a_M \|u\|_2 \right\} \\
= \Pr \left\{ \bigcap_{i_1 < j_1} \bigcap_{u \in \mathcal{F}_{i_1,j_1}} \|Au\|_2 + g \|u\|_2 \geq d(1 + \epsilon)\mathbb{E}\|A\|_2 + \epsilon_2 a_M \|u\|_2 \right\} \\
\geq \Pr \left\{ \bigcap_{i_1 < j_1} \bigcap_{u \in \mathcal{F}_{i_1,j_1}} \left( \sum_{i_2=1}^{M} g_{i_2}^2 \right)^{\frac{1}{2}} \right\} + \sum_{j_2=1}^{N} e_{j_2} h_{j_2} \geq d(1 + \epsilon_1)\mathbb{E}\|A\|_2 + \epsilon_2 a_M \right\} \right\},
\]

where in \( (i) \) \( g \) is a RV satisfying standard normal distribution, in \( (ii) \) we use the union bound, and \( (iii) \) comes from Lemma 3.1 in [12] and \( \|u\|_2 = 1 \).
Step 3. We lower bound $Q_2$ by considering Then we bound $R$ as

$$1 - Q_2 = \Pr \left\{ \bigcup_{i_1 \leq j_1} \bigcup_{u \in \mathcal{S}_{i_1,j_1}} \left[ \left( \sum_{i_2=1}^{M} g_{i_2}^2 \right)^\dagger + \sum_{j_2=1}^{N} u_{j_2} h_{j_2} \leq d(1 + \epsilon_1)\|A\|_2 + \epsilon_2 a_M \right] \right\}$$

$$\leq \Pr \left\{ \left( \sum_{i_2=1}^{M} g_{i_2}^2 \right)^\dagger \leq (1 - \epsilon_2) a_M \right\} + \Pr \left\{ \bigcup_{i_1 \leq j_1} \bigcup_{u \in \mathcal{S}_{i_1,j_1}} \sum_{j_2=1}^{N} u_{j_2} h_{j_2} \leq d(1 + \epsilon_1)\|A\|_2 - (1 - 2\epsilon_2) a_M \right\}$$

$$\leq \Pr \left\{ \left( \sum_{i_2=1}^{M} g_{i_2}^2 \right)^\dagger - a_M \leq -\epsilon_2 a_M \right\} + \Pr \left\{ \bigcup_{i_1 \leq j_1} \bigcup_{u \in \mathcal{S}_{i_1,j_1}} \sum_{j_2=1}^{N} u_{j_2} h_{j_2} \leq d(1 + \epsilon_1)\|A\|_2 - (1 - 2\epsilon_2) a_M \right\}$$

$$\leq \exp \left( -\frac{\epsilon_2^2 a_M^2}{2} \right) + \Pr \left\{ \bigcup_{i_1 \leq j_1} \bigcup_{u \in \mathcal{S}_{i_1,j_1}} \sum_{j_2=1}^{N} u_{j_2} h_{j_2} \leq d(1 + \epsilon_1)\|A\|_2 - (1 - 2\epsilon_2) a_M \right\}$$

$$\leq \exp \left( -\frac{\epsilon_2^2 a_M^2}{2} \right) + \exp \left[ \frac{\left( 1 - 2\epsilon_2 \right) a_M - d(1 + \epsilon_1)\|A\|_2 - \omega(\tilde{e}_{ij})}{2} \right]^2$$

where in (i) we use $\mathbb{E} \left( \sum_{i_2=1}^{M} g_{i_2}^2 \right) = a_M$ and Gaussian concentration inequality in [17], in (ii) we use union-bound, in (iii) we define $\tilde{h} = -h$ and flip the sign by the symmetry of Gaussian variables, and in (iv) we use the definition of $\omega(\tilde{e}_{ij})$. Assuming $(1 - 2\epsilon_2) a_M \geq d(1 + \epsilon_1)\|A\|_2 + \omega(\tilde{e}_{ij})$, we finish the proof via the Gaussian concentration inequality [17].

Combining the above together and set $d(1 + \epsilon_1) \to 0$ while $\epsilon_1 \to \infty$, we conclude that

$$\mathcal{P}_2^\circ \geq 1 - \frac{3}{2} \exp \left( -\frac{\epsilon_2^2 a_M^2}{2} \right) - \sum_{i \leq j} \exp \left( -\frac{\left( 1 - 2\epsilon_2 \right) a_M - \omega(\tilde{e}_{ij})}{2} \right)^2$$

provided $(1 - 2\epsilon) a_M \geq \omega(\tilde{e}_{ij})$, and finish the proof.

\[ \]
Proof. Based on the definition of $p^*$, we note that $\sum_i \mathcal{F}_t^i$ is non-negative and prove the lower-bound. Then we prove its upper-bound.

Since the function is linear, optimal $p^*$ must be at the edge of $\Delta_L$ and we denote the non-zero entry as $i^*$. Hence, we could study it via the multiplicative weight algorithm analysis [8]. First we rewrite the update equation (7). Define $w^{(0)} = 1 \in \mathbb{R}^L$ and update $w^{(t+1)}$ as

$$ w_i^{(t+1)} = w_i^{(t)} \exp \left( -\eta_p f_i(x_i^{(t)}) \right), \quad p_i^{(t+1)} = \frac{w_i^{(t+1)}}{\sum_i w_i^{(t+1)}}, $$

where $(\cdot)_i$ denotes the $i$th element, and $p^{(t+1)}$ can be regarded as the normalized version of $w^{(t+1)}$.

First we define $\Psi_t$ as

$$ \Psi_t = \sum_{i=1}^L w_i^{(t)}. $$

Then we have $\Psi_0 = L$ while $\Psi_T \geq w_i^{(T)} = \exp \left( -\sum_{t=1}^T \eta_p f_i(x_i^{(t)}) \right) = \exp \left( -\eta_p \sum_{t=1}^T f_i(x_i^{(t)}) \right)$, where $\eta_p = \eta_p = \sqrt{2 \log L/T}$.

Then we study the division $\Psi_{t+1}/\Psi_t$ as

$$ \Psi_{t+1} = \sum_i w_i^{(t+1)} = \sum_i w_i^{(t)} \exp \left( -\eta_p f_i(x_i^{(t)}) \right) \leq \sum_i w_i^{(t)} \left( 1 - \eta_p f_i(x_i^{(t)}) + \frac{\eta_p^2 f_i^2(x_i^{(t)})}{2} \right) = \Psi_t \left( 1 - \eta_p \left\langle p^{(t)}, f(x^{(t)}) \right\rangle + \frac{\eta_p^2 R_t^2}{2} \right) \leq \Psi_t \exp \left( -\eta_p \left\langle p^{(t)}, f(x^{(t)}) \right\rangle + \frac{\eta_p^2 R_t^2}{2} \right) $$

where in (i) we use $e^{-x} \leq 1 + x + x^2/2$ for $x \geq 0$, and in (ii) we use $e^x \geq 1 + x$ for all $x \in \mathbb{R}$. Using the above relation iteratively, we conclude that

$$ \frac{\Psi_T}{\Psi_0} \leq \exp \left( -\eta_p \sum_t \left\langle p^{(t)}, f(x^{(t)}) \right\rangle + \frac{\eta_p^2 T R_t^2}{2} \right), $$

which gives us

$$ \log \Psi_T \leq \log L - \eta_p \sum_t \left\langle p^{(t)}, f(x^{(t)}) \right\rangle + \frac{T \eta_p^2 R_t^2}{2}. $$

With relation $\log \Psi_T \geq \log w_i^{(T)}$, we obtain

$$ \eta_p \sum_t \left( \left\langle p^{(t)}, f(x^{(t)}) \right\rangle - \left\langle e_{i^*}, f(x^{(t)}) \right\rangle \right) = \eta_p \sum_t \left( \left\langle p^{(t)}, f(x^{(t)}) \right\rangle - \left\langle p^*, f(x^{(t)}) \right\rangle \right) \leq \log L + \frac{T \eta_p^2 R_t^2}{2}, $$

where $e_{i^*}$ denotes the canonical basis, namely, has 1 in its $i^*$th entry and all others to be zero.

Lemma 12. Define $\mathcal{F}_t^i = \mathcal{L}(p^{(t)}, x^{(t)}) - \mathcal{L}(p^{(t)}, x^*)$, where $x^*$ is defined in (10), and set $\eta_w^{(t)} = \eta_w \leq L^{-1}$, then we have

$$ 0 \leq \sum_t \mathcal{F}_t^i \leq \frac{1}{2 \eta_w} \|x^{(0)} - x^*\|^2_2. $$

Proof. From the definition of $x^*$, we can prove the non-negativeness of $\sum_t \mathcal{F}_t^i$. Here we focus on upper-bounding $\sum_t \mathcal{F}_t^i$ by separately analyzing each term $\mathcal{F}_t^i$. For the conciseness of notation, we drop the time index $t$. Define $h(x)$ as

$$ h(x) = \sum_i p_i h_i(x) + \frac{\lambda_1 \|y - Ax\|^2_2}{2} + \frac{\lambda_2 \|x\|^2}{2}. $$
First, we rewrite the update equation as
\[
x^{(t+1)} = x^{(t)} - \left\{ \left( x^{(t)} - \text{prox}_{\eta_w \| \cdot \|_1} \left[ x^{(t)} - \eta_w \nabla h(x^{(t)}) \right] \right) \right\},
\]
which means that
\[
H(x^{(t)}) = \frac{x^{(t)} - \text{prox}_{\eta_w \| \cdot \|_1} \left[ x^{(t)} - \eta_w \nabla h(x^{(t)}) \right]}{\eta_w},
\]
where \( \text{prox}_{\eta_w \| \cdot \|_1} (x) \) is defined as [15]
\[
\text{prox}_{\eta_w \| \cdot \|_1} (x) = \arg\min_z \quad \eta_w \| z \|_1 + \frac{1}{2} \| x - z \|_2^2.
\]
Here we need one important property of \( H(x^{(t)}) \), that is widely in the analysis of proximal gradient descent (direct results of Theorem 6.39 in [13]) and states
\[
H(x^{(t)}) \in \nabla h(x^{(t)}) + \partial \| x^{(t+1)} \|_1.
\]
Here we consider the update relation \( f(x^{(t+1)}) - f(z) \) as
\[
\sum_t \mathcal{S}_2^t = \sum_i p_i \left[ f_i(x^{(t+1)}) - f_i(x^*) \right] = \| x^{(t+1)} \|_1 + h(x^{(t+1)}) - \| x^* \|_1 - h(x^*)
\]
\[
\leq (i) \quad \left( \partial \| x^{(t+1)} \|_1, x^{(t+1)} - x^* \right) + h(x^{(t)}) + \left\langle \nabla h(x^{(t)}), x^{(t+1)} - x^{(t)} \right\rangle + \frac{L_h}{2} \| x^{(t+1)} - x^{(t)} \|_2^2 - h(x^*)
\]
\[
\leq (ii) \quad \left( \partial \| x^{(t+1)} \|_1, x^{(t+1)} - x^* \right) + \left\langle \nabla h(x^{(t)}), x^{(t)} - x^* + x^{(t+1)} - x^{(t)} \right\rangle + \frac{L_h}{2} \| x^{(t+1)} - x^{(t)} \|_2^2
\]
\[
= (iii) \quad \left( \partial \| x^{(t+1)} \|_1 + \nabla h(x^{(t)}), x^{(t+1)} - x^* \right) + \frac{L_h}{2} \| x^{(t+1)} - x^{(t)} \|_2^2
\]
\[
\leq (iv) \quad \left( H(x^{(t)}), x^{(t+1)} - x^* \right) + \frac{L_h}{2} \| x^{(t+1)} - x^{(t)} \|_2^2
\]
\[
\leq (v) \quad \frac{1}{\eta_w} \left\langle x^{(t)} - x^{(t+1)}, x^{(t+1)} - x^* \right\rangle + \frac{L_h}{2} \| x^{(t+1)} - x^{(t)} \|_2^2
\]
\[
\leq (vi) \quad \frac{1}{\eta_w} \left\langle x^{(t)} - x^{(t+1)}, x^{(t+1)} - x^* \right\rangle + \frac{1}{2 \eta_w} \| x^{(t+1)} - x^{(t)} \|_2^2
\]
\[
= (vii) \quad \frac{1}{\eta_w} \left\langle x^{(t)} - x^{(t+1)}, x^{(t+1)} - x^* \right\rangle + \frac{1}{2 \eta_w} \| x^{(t+1)} - x^{(t)} \|_2^2
\]
\[
= \frac{1}{2 \eta_w} \| x^{(t)} - x^* \|_2^2 - \frac{1}{2 \eta_w} \| x^{(t+1)} - x^* \|_2^2,
\]
where in (i) we use \( \| x^* \|_1 \geq \| x^{(t+1)} \|_1 + \| \partial \| x^{(t+1)} \|_1 \|_1, x^* = x^{(t+1)} \) based on the definition of sub-gradients, and \( h(x^{(t+1)}) \leq h(x^{(t)}) + \| \nabla h(x^{(t)}), x^{(t+1)} \| + \frac{L_h}{2} \| x^{(t+1)} - x^{(t)} \|_2^2 \) from the \( L_h \) smoothness of \( h(\cdot) \), in (ii) we use \( h(x^*) \geq h(x^{(t)}) + \left\langle \nabla h(x^{(t)}), x^* - x^{(t)} \right\rangle \) since \( h(\cdot) \) is convex, in (iii) we use \( H(x^{(t)}) \in \nabla h(x^{(t)}) + \partial \| x^{(t+1)} \|_1 \), and in (iv) we use \( x^{(t+1)} = x^{(t)} - \eta_w H(x^{(t)}) \), and in (vii) we use \( \eta_w \leq L^{-1} \).

Hence, we finish the proof by
\[
\sum_t \mathcal{S}_2^t \leq \sum_t \left[ \frac{1}{2 \eta_w} \| x^{(t-1)} - x^* \|_2^2 - \frac{1}{2 \eta_w} \| x^{(t)} - x^* \|_2^2 \right]
\]
\[
= \frac{1}{2 \eta_w} \| x^{(0)} - x^* \|_2^2 - \frac{1}{2 \eta_w} \| x^{(T)} - x^* \|_2^2 \leq \frac{1}{2 \eta_w} \| x^{(0)} - x^* \|_2^2 \leq \frac{4 R^2}{2 \eta_w},
\]
where in (i) we use \( \| x^{(0)} - x^* \|_2 \leq \| x^* \|_2 + \| x^{(0)} \| \leq 2 R \).
**APPENDIX C**

**PROOF OF THEOREM 6**

**Proof.** First we define \( h(x) = \sum_t p_t h_t(x) + \lambda_1 \| y - Ax \|^2_2 + \lambda_2 \| x \|^2_2 / 2 \). Then we consider the term \( \mathcal{L}(p(t), x^{(t)}) - \mathcal{L}(p(t), x^{(t)}) \) and have

\[
\begin{align*}
\mathcal{L}(p(t), x^{(t)}) - \mathcal{L}(p(t), x^{(t)}) &= \| x^{(t)} \|_1 - \| x^{(t)} \|_1 + \sum_i p_i(t) \left( h_i(x^{(t)}) - h_i(x^{(t)}) \right) \\
&\leq \langle \partial \| x^{(t)} \|_1, x^{(t)} - x^{(t)} \rangle + \sum_i p_i(t) \left( \langle \nabla h_i(x^{(t)}), x^{(t)} - x^{(t)} \rangle + \frac{L_i}{2} \| x^{(t)} - x^{(t)} \|^2_2 \right) \\
&\leq \left( \frac{\partial \| x^{(t)} \|_1}{\eta^{(t)}} \langle x^{(t)} - x^{(t)} \rangle + \frac{L_i}{2} \| x^{(t)} - x^{(t)} \|^2_2 \right) \\
&\leq \langle \frac{\partial \| x^{(t)} \|_1}{\eta^{(t)}} \rangle \| x^{(t)} - x^{(t)} \|^2_2 \\
&\leq \frac{L_i}{2} \| x^{(t)} - x^{(t)} \|^2_2,
\end{align*}
\]

where in (i) we have \( \| x^{(t)} \|_1 \geq \| x^{(t)} \|_1 + \langle \partial \| x^{(t)} \|_1, x^{(t)} - x^{(t)} \rangle \) from the definition of sub-gradient [15], and \( h_i(x^{(t)}) \leq h_i(x^{(t)}) + \langle \nabla h_i(x^{(t)}), x^{(t)} - x^{(t)} \rangle + \frac{L_i}{2} \| x^{(t)} - x^{(t)} \|^2_2 \), in (ii) we use the property \( \partial \| x^{(t)} \|_1 + \sum_i p_i(t) \nabla h_i(x^{(t)}) \in H(x^{(t)}) \), in (iii) we use \( x^{(t)} = x^{(t)} - \eta^{(t)} H(x^{(t)}) \), and in (iv) we use \( \eta^{(t)} \leq L^{-1} \).

Adopting similar tricks as [18], we could upper-bound \( \| x^{(t)} - x^{(t)} \|^2_2 \) as

\[
\| x^{(t)} - x^{(t)} \|^2_2 \leq \frac{2}{L_i} \left( \mathcal{L}(p(t), x^{(t)}) - \mathcal{L}(p(t), x^{(t)}) \right)
\]

Then we separately discuss bound \( \mathcal{F}^2_t \) and \( \mathcal{F}^1_t \). Since most terms of \( \sum_t \mathcal{F}^1_t \) will be cancelled after summization, we focus the analysis on bounding \( \mathcal{F}^2_t \), which is

\[
\mathcal{F}^2_t = \left\langle p(t) - p(t), f(x^{(t)}) \right\rangle \leq \| p(t) - p(t) \|_1 \| f(x^{(t)}) \|_\infty^1.
\]

where \( f(x^{(t)}) \) denotes the vector whose \( i \)th element is \( f_i(x^{(t)}) \). Notice that we have

\[
\| p(t) - p(t) \|_1 \leq \left( i \right) \left( i \right) \| p(t) \| \leq \| p(t) - p(t), f(x^{(t)}) \| \leq \eta^{(t)} \| p(t) - p(t), f(x^{(t)}) \|_\infty \leq \eta^{(t)} R_f \| p(t) - p(t), f(x^{(t)}) \|_1,\]

which gives us \( \| p(t) - p(t) \|_1 \leq \eta^{(t)} R_f \), where (i) is because of Pinsker’s inequality (Theorem 4.19 in [17]) and (ii) is because of Lemma 13. To conclude, we have upper-bound \( \mathcal{F}^2_t \) as

\[
\mathcal{F}^2_t \leq \eta^{(t)} R_f^2.
\]

Then we finish the proof as

\[
\sum_t \| x^{(t)} - x^{(t)} \|^2_2 \leq \frac{2 \mathcal{L}(p(0), x(0)) - 2 \mathcal{L}(p(T+1), x(T+1))}{L_i} + \frac{4R_f^2 \sum_t \eta^{(t)}}{L_h},
\]

where (i) is because \( \mathcal{L}(p(T+1), x(T+1)) \geq 0 \).
Lemma 13. With Alg. [7] we have

\[ D_{KL} \left( p^{(t+1)} \| p^{(t)} \right) \leq \eta_p^{(t)} \left\langle p^{(t)} - p^{(t+1)}, f(x^{(t)}) \right\rangle, \]

where \( f(x^{(t)}) \) denotes the vector whose \( i \)-th element is \( f_i(x^{(t)}) \).

Proof. Here we have

\[ D_{KL} \left( p^{(t+1)} \| p^{(t)} \right) = \sum_i p_i^{(t+1)} \log \left( \frac{p_i^{(t+1)}}{p_i^{(t)}} \right) = \sum_i p_i^{(t)} \log \left( e^{-\eta_p^{(t)} f_i(x_i)} \right) = -\log (Z_t) - \eta_p^{(t)} \sum_i p_i^{(t+1)} f_i(x_i) \]

where \( Z_t = \prod_i p_i^{(t)} e^{-\eta_p f_i(x_i)} \).

Then we have

\[ \eta_p^{(t)} \left\langle p^{(t)} - p^{(t+1)}, f(x^{(t)}) \right\rangle = D_{KL} \left( p^{(t+1)} \| p^{(t)} \right) + \log \left( \sum_i p_i^{(t)} e^{-\eta_p^{(t)} f_i(x_i)} \right) + \eta_p^{(t)} \left\langle p^{(t)}, f(x^{(t)}) \right\rangle \]

\[ \geq D_{KL} \left( p^{(t+1)} \| p^{(t)} \right) + \log \left( \prod_i e^{-\eta_p^{(t)} p_i^{(t)} f_i(x_i)} \right) + \eta_p^{(t)} \left\langle p^{(t)}, f(x^{(t)}) \right\rangle \]

\[ = D_{KL} \left( p^{(t+1)} \| p^{(t)} \right) + \sum_i \log \left( e^{-\eta_p^{(t)} p_i^{(t)} f_i(x_i)} \right) + \eta_p^{(t)} \left\langle p^{(t)}, f(x^{(t)}) \right\rangle = D_{KL} \left( p^{(t+1)} \| p^{(t)} \right), \]

where in (i) we use \( \sum_i p_i x_i \geq \prod_i x_i^{p_i} \) such that \( \sum_i p_i = 1, p_i \geq 0 \).

\[ \square \]

APPENDIX D
PROOF OF THEOREM 7

Proof. Define \( p^* \) and \( x^* \) as

\[ p^* = \arg\min_p \sum_t \mathcal{L} \left( p, x^{(t)} \right), \quad x^* = \arg\min_x \sum_t \mathcal{L} \left( p^{(t)}, x \right), \] (10)

respectively First we define \( \mathcal{F}_1^t \) and \( \mathcal{F}_2^t \) as

\[ \mathcal{F}_1^t = \mathcal{L} \left( p^{(t)}, x^{(t)} \right) - \mathcal{L} \left( p^*, x^{(t)} \right); \]

\[ \mathcal{F}_2^t = \mathcal{L} \left( p^{(t)}, x^{(t)} \right) - \mathcal{L} \left( p^{(t)}, x^* \right), \]

respectively. Then our goal becomes bounding \( |\sum_t \mathcal{F}_1^t| + |\sum_t \mathcal{F}_2^t| \). For term \( |\sum_t \mathcal{F}_2^t| \), the analysis stays the same as Lemma 12 Here we focus on bounding \( \sum_t \mathcal{F}_1^t \), which proceeds as

\[ \sum_t \mathcal{F}_1^t = \sum_t \left( \mathcal{L} \left( p^{(t)}, x^{(t)} \right) - \mathcal{L} \left( p, x^{(t)} \right) \right) = \sum_t \left\{ -\left\langle \nabla_t \mathcal{L} \left( p^{(t)}, x^{(t)} \right), p - p^{(t)} \right\rangle - \frac{\lambda_3}{2} \left\| p^{(t)} - p \right\|_2^2 \right\} \]

\[ = \sum_t \left\{ \left\langle g^{(t)}, p^{(t)} - p \right\rangle - \frac{\lambda_3}{2} \left\| p^{(t)} - p \right\|_2^2 \right\} \]

Then we consider the distance \( \left\| p^{(t+1)} - p \right\|_2^2 \) which is

\[ \left\| p^{(t+1)} - p \right\|_2^2 = \left\| p^{(t)} - \eta_p g^{(t)} - \eta_p g^{(t)} \right\|_2^2 \leq \left\| p^{(t)} - \eta_p g^{(t)} - p \right\|_2 \]

\[ = \left\| p^{(t)} - p \right\|_2^2 + \eta_p^2 \left\| g^{(t)} \right\|_2^2 - 2\eta_p \left\langle g^{(t)}, p - p \right\rangle, \]

where in (i) we use the contraction property for projection, which gives us

\[ \left\langle g^{(t)}, p - p \right\rangle \leq \frac{\eta_p^2 \left\| g^{(t)} \right\|_2^2}{2} + \frac{\left\| p^{(t)} - p \right\|_2^2 - \left\| p^{(t+1)} - p \right\|_2^2}{2\eta_p} \]
By setting $\eta_p(t) = (\lambda t)^{-1}$, we have

$$\sum_t \mathcal{J}^t_1 \leq \frac{R_g^2 \log T}{2\lambda_3} + \frac{\lambda_3}{2} \sum_t (\|p(t) - p\|^2 - \|p(t+1) - p\|^2) - \frac{\lambda_3}{2} \sum_t \|p(t) - p\|^2$$

$$= \frac{R_g^2 \log T}{2\lambda_3} + \frac{\lambda_3}{2} \sum_t \|p(t) - p\|^2 - \frac{\lambda_3 (T + 1)}{2} \|p(T+1) - p\|^2 - \frac{\lambda_3}{2} \sum_t \|p(t) - p\|^2$$

$$= \frac{R_g^2 \log T}{2\lambda_3} - \frac{\lambda_3 (T + 1)}{2} \|p(T+1) - p\|^2 \leq \frac{R_g^2 \log T}{2\lambda_3}.$$ 

Hence, we have

$$\sum_t \mathcal{J}^t_1 + \mathcal{J}^t_2 \leq \frac{R_g^2 \log T}{2\lambda_3 T} + \frac{R_g^2}{2\eta_\ell T},$$

which completes the proof.

**APPENDIX E**

**PROOF OF THEOREM 8**

Proof. Following the same procedure in section C, we can bound

$$\|x^{(t+1)} - x^{(t)}\|^2 \leq \frac{2}{L_h} \left[ \mathcal{L}(p(t), x(t)) - \mathcal{L}(p(t), x(t+1)) \right].$$

Since $\mathcal{J}^t_1$ will cancel themselves after summarization, we focus on bounding $\mathcal{J}^t_2$. Then we have

$$\mathcal{L}(p(t), x(t)) = \mathcal{L}(p(t+1), x(t+1)) + \left( \nabla_p \mathcal{L}(p(t), x(t)) \right) + \frac{\lambda_3}{2} \|p(t+1) - p(t)\|^2$$

$$= \mathcal{L}(p(t+1), x(t+1)) + \left( f(x(t+1)) + \lambda_3 (p(t+1) - q), p(t) - p(t+1) \right) + \frac{\lambda_3}{2} \|p(t+1) - p(t)\|^2,$$

where in (i) we have

$$\nabla_p \mathcal{L}(p(t+1), x(t+1)) = f(x(t+1)) + \lambda_3 \left( p(t+1) - q \right).$$

Then we have

$$\mathcal{L}(p(t+1), x(t+1)) - \mathcal{L}(p(t), x(t)) = \left( g(t+1), p(t) - p(t+1) \right) + \frac{\lambda_3}{2} \|p(t+1) - p(t)\|^2$$

$$\leq \|g(t+1)\|_2 \|p(t+1) - p(t)\|_2 + \frac{\lambda_3}{2} \|p(t+1) - p(t)\|^2$$

$$\leq \|g(t+1)\|_2 \|\eta_p(t)g(t)\|_2 + \frac{\lambda_3}{2} \left( \eta_p(t) \right)^2 \leq R_g^2 \left( \eta_p(t) + \frac{\lambda_3}{2} \left( \eta_p(t) \right)^2 \right).$$

Hence, we conclude that

$$\sum_t \|x^{(t+1)} - x^{(t)}\|^2 \leq \frac{2 \mathcal{L}(p(0), x(0))}{L_h} + \frac{2R_g^2}{L_h} \sum_t \left( \eta_p(t) + \frac{\lambda_3}{2} \left( \eta_p(t) \right)^2 \right),$$

where $\|g(t)\|_2 \leq R_g.$