# Compressive Sensing with a Multiple Convex Sets Domain 

Hang Zhang, Afshin Abdi, and Faramarz Fekri<br>School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA.


#### Abstract

In this paper, we study a general framework for compressive sensing assuming the existence of the prior knowledge that $\mathbf{x}^{\natural}$ belongs to the union of multiple convex sets, $\mathbf{x}^{\natural} \in \bigcup_{i} \mathcal{C}_{i}$. 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 and proximal gradient descent is proposed and analyzed for signal reconstruction. Finally, we investigate as to how we can improve the signal recovery by introducing regularizers into the objective function.


## 1 Introduction

In the traditional compressive sensing (CS) Eldar \& Kutyniok (2012); Foucart \& Rauhut (n.d.), sparse signal $\mathbf{x}$ is reconstructed via

$$
\begin{equation*}
\min _{\mathbf{x}}\|\mathbf{x}\|_{1}, \quad \text { s.t. } \quad \mathbf{y}=\mathbf{A x} \tag{1.1}
\end{equation*}
$$

where $\mathbf{y} \in \mathbb{R}^{M}$ denotes the measurement vector and $\mathbf{A} \in \mathbb{R}^{M \times N}$ is the measurement matrix.
In this paper, we assume the existence of extra prior knowledge that $\mathbf{x}$ lies in the union of some convex sets, $\mathbf{x} \in \bigcup_{i=1}^{L} \mathcal{C}_{i}$, 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

$$
\begin{equation*}
\min _{\mathbf{x}}\|\mathbf{x}\|_{1}, \quad \text { s.t. } \quad \mathbf{y}=\mathbf{A x}, \quad \mathbf{x} \in \bigcup_{i=1}^{L} \mathcal{C}_{i} \tag{1.2}
\end{equation*}
$$

This ill-posed inverse problem (i.e., given measurement $\mathbf{y}$, solving for $\mathbf{x}$ ) turns out to be a rather general form of CS. For example, setting $\cup \mathcal{C}_{i}=\mathbb{R}^{n}$ simplifies our problem to the traditional CS problem. In the following, we will further show that by appropriate choices of these convex sets, Eq. (1.2) can be transformed to the phase retrieval Candes et al. (2011); Chen \& Candes (2015), quantized compressive sensing Dai et al. (2009), or model-based CS Baraniuk et al. (2010) problems.

### 1.1 Relation with other problems

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

$$
\begin{equation*}
y_{i}=\left|\left\langle\mathbf{a}_{i}, \mathbf{x}\right\rangle\right|^{2}, \quad 1 \leq i \leq l, \tag{1.3}
\end{equation*}
$$

where $y_{i}$ is the $i$-th measurement and $\mathbf{a}_{i}$ denotes the corresponding coefficients. Considering the first measurement, the constraint $\sqrt{y_{1}}=\left|\left\langle\mathbf{a}_{1}, \mathbf{x}\right\rangle\right|$ can be represented via $\mathbf{x} \in \mathcal{B}_{+}^{(1)} \cup \mathcal{B}_{-}^{(1)}$ where $\mathcal{B}_{+}^{(1)}=\left\{\mathbf{x}:\left\langle\mathbf{a}_{1}, \mathbf{x}\right\rangle=\sqrt{y_{1}}\right\}$ and $\mathcal{B}_{-}^{(1)}=\left\{\mathbf{x}:\left\langle\mathbf{a}_{1}, \mathbf{x}\right\rangle=-\sqrt{y_{1}}\right\}$. Following these steps, the constraints $\left\{y_{i}=\left\langle\mathbf{a}_{i}, \mathbf{x}\right\rangle^{2}\right\}_{i=1}^{l}$ can be transformed to $\mathbf{x} \in \bigcap_{i}\left(\mathcal{B}_{+}^{(i)} \cup \mathcal{B}_{-}^{(i)}\right)=\bigcup_{j=1}^{2^{l}} \mathcal{C}_{j}$, for some appropriately defined $C_{j}$ 's given by the intersection of different $\mathcal{B}_{ \pm}^{(i)}$. Setting sensing matrix $\mathbf{A}=\mathbf{0}$ will restore the phase retrieval to our setting.

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

$$
\begin{equation*}
y_{i}=Q\left(\left\langle\mathbf{a}_{i}, \mathbf{x}\right\rangle\right), 1 \leq i \leq L, \tag{1.4}
\end{equation*}
$$

where $Q(\cdot)$ is the quantizer. Since $Q^{-1}(\cdot)$ is an interval on real line, $\mathcal{C}_{i}$ would be a convex set and the quantized CS can be easily transformed to Eq. (1.2).

Model-based compressive sensing These lines of works Baraniuk et al. (2010); Duarte \& Eldar (2011); Silva et al. (2011) are the most similar work to our model, where they consider

$$
\begin{equation*}
\mathbf{y}=\mathbf{A} \mathbf{x}^{\natural}, \quad \mathbf{x}^{\natural} \in \bigcup_{i} \mathcal{L}_{i} . \tag{1.5}
\end{equation*}
$$

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 Baraniuk et al. (2010), the author studied the minimum number of measurements $M$ under different models, i.e., shape of $\mathcal{L}_{i}$, and modified CoSaMP algorithms Foucart \& Rauhut (n.d.) to reconstruct signal. In Duarte \& Eldar (2011), the authors expanded the signal onto different basis and transformed model-based CS to be block-sparse CS. In Silva et al. (2011), the author studied modelbased CS with incomplete sensing matrix information and reformulated it as a matrix completion problem.

### 1.2 Our contribution:

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 $\min _{u \in E}\|\mathbf{A u}\|_{2}>0$, for an appropriate set $E$. Assuming the entries of the sensing matrix $\mathbf{A}$ 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.

Optimization Algorithm We propose a novel formulation and the associated optimization algorithm to reconstruct the signal $\mathbf{x}$. 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 $\mathbf{x}$ onto the basis of each $\mathcal{C}_{i}$ and then rephrase the constraint as the block sparsity on the representation of $\mathbf{x}$ on the union of bases. However, such an approach may add complicated constraints on the coefficients of $\mathbf{x}$ 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 Eq. (1.2) hard to solve. By introducing an auxiliary variable, $\mathbf{p}$, we convert the non-convex optimization problem to a biconvex problem. Using multiplicative weight update Arora et al. (2012) from online learning theory Shalev-Shwartz et al. (2012), we design an algorithm with convergence speed of $\mathcal{O}\left(T^{-1 / 2}\right)$ 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

$$
\begin{equation*}
\min _{\mathbf{x}}\|\mathbf{x}\|_{1} \text {, s.t. } \mathbf{y}=\mathbf{A x}, \quad \mathbf{x} \in \mathcal{C}_{i}, \tag{1.6}
\end{equation*}
$$

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

## 2 System Model

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

$$
\begin{equation*}
\mathbf{y}=\mathbf{A} \mathbf{x}^{\natural}, \quad \text { and } \quad \mathbf{x}^{\natural} \in \bigcup_{i=1}^{L} \mathcal{C}_{i}, \tag{2.1}
\end{equation*}
$$

where $\mathbf{x}^{\natural}$ is a $K$-sparse high-dimensional signal, $\mathbf{y} \in \mathbb{R}^{M}$ 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 $\mathbf{x}^{\natural}$, we propose to reconstruct $\mathbf{x}^{\natural}$ via

$$
\begin{equation*}
\widehat{\mathbf{x}}=\operatorname{argmin}_{\mathbf{x}}\|\mathbf{x}\|_{1}, \quad \text { s.t. } \mathbf{y}=\mathbf{A x}, \quad \mathbf{x} \in \bigcup_{i=1}^{L} \mathcal{C}_{i} \tag{2.2}
\end{equation*}
$$

where $\widehat{\mathbf{x}}$ denotes the reconstructed signal. Let $\mathbf{d} \triangleq \widehat{\mathbf{x}}-\mathbf{x}^{\natural}$ be the deviation of the reconstructed signal $\widehat{\boldsymbol{x}}$ from the true signal $\mathbf{x}^{\natural}$. In the following, we will study the inverse problem in Eq. (2.2) from two perspectives; the statistical and the computational aspects.

## 3 Statistical Property

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

Definition 1. The tangent cone $\mathcal{T}_{\mathbf{x}}$ for $\|\mathbf{x}\|_{1}$ is defined as Chandrasekaran et al. (2012)

$$
\begin{equation*}
\mathcal{T}_{\mathbf{x}} \triangleq\left\{\mathbf{e}:\|\mathbf{x}+t \mathbf{e}\|_{1} \leq\|\mathbf{x}\|_{1}, \exists t \geq 0\right\} \tag{3.1}
\end{equation*}
$$

Geometric interpretation of $\mathcal{T}_{\mathbf{x}}$ is that it contains all directions that lead to smaller $\|\cdot\|_{1}$ originating from $\mathbf{x}$. In the following analysis, we use $\mathcal{T}$ as a compact notation for $\mathcal{T}_{\mathbf{x}^{\natural}}$. Easily we can prove that $\mathbf{d} \in \mathcal{T}$.
Definition 2. The Gaussian width $\omega(\cdot)$ associated with set $U$ is defined as $\omega(U) \triangleq \mathbb{E} \sup _{\mathbf{x} \in U}\langle\mathbf{g}, \mathbf{x}\rangle, \mathbf{g} \sim$ $\mathcal{N}(\mathbf{0}, \mathbf{I})$, Gordon (1988).

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

$$
\begin{equation*}
\widetilde{\mathcal{C}}_{i, j} \triangleq\left\{\mathbf{z} \mid \mathbf{z}=t\left(\mathbf{x}_{1}-\mathbf{x}_{2}\right), \exists t>0, \mathbf{x}_{1} \in \mathcal{C}_{i}, \mathbf{x}_{2} \in \mathcal{C}_{j}\right\} \tag{3.2}
\end{equation*}
$$

which denotes the cone consisting of all vectors $\mathbf{z}$ that are parallel with $\mathbf{x}_{1}-\mathbf{x}_{2}, \mathbf{x}_{1} \in \mathcal{C}_{i}$, and $\mathbf{x}_{2} \in \mathcal{C}_{j}$. Then we define event $\mathcal{E}$ as

$$
\begin{equation*}
\mathcal{E} \triangleq\left\{\bigcup_{i, j}\left(\operatorname{null}(\mathbf{A}) \bigcap \mathcal{T} \cap \widetilde{\mathcal{C}}_{i, j}\right)=\{\mathbf{0}\}\right\} \tag{3.3}
\end{equation*}
$$

Lemma 1. We can guarantee the correct recovery of $\mathbf{x}$, i.e., $\widehat{\mathbf{x}}=\mathbf{x}^{\natural}$, iff we have event $\mathcal{E}$ to be satisfied.
Proof. This proof is fundamentally same as Chandrasekaran et al. (2012); Zhang et al. (Nov. 2017.). First we prove that $\mathcal{E}$ leads to $\widehat{\mathbf{x}} \neq \mathbf{x}^{\natural}$. Provided $\mathbf{d} \triangleq \widehat{\mathbf{x}}-\mathbf{x}^{\natural} \neq 0$, we then have a $\widehat{\mathbf{x}} \neq \mathbf{x}^{\natural}$ such that $\|\widehat{\mathbf{x}}\|_{1} \leq\left\|\mathbf{x}^{\natural}\right\|_{1}$. Setting $\mathbf{e} \| \mathbf{d}$, we hav a non-zero $\mathbf{e} \in \operatorname{null}(\mathbf{A}) \cap \mathcal{T} \cap\left(\cup_{i, j} \widetilde{\mathcal{C}}_{i, j}\right)$, which violates $\mathcal{E}$.
Then we prove that $\widehat{\mathbf{x}} \neq \mathbf{x}^{\natural}$ implies $\mathcal{E}$. Assume that there exists non-zero $\mathbf{e} \in \operatorname{null}(\mathbf{A}) \cap \mathcal{T} \cap\left(\cup_{i, j} \widetilde{\mathcal{C}}_{i, j}\right)$. We can show that signal $\mathbf{x}^{\natural}+t \mathbf{e}$, where $t$ is some positive constant such that $\left\|\mathbf{x}^{\natural}+t \mathbf{e}\right\|_{1} \leq\left\|\mathbf{x}^{\natural}\right\|_{1}$, satisfying constraints described by Eq. (2.1). This implies that $\mathbf{d}=t \mathbf{e} \neq \mathbf{0}$ and the wrong recovery of $\mathbf{x}^{\natural}$.

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

$$
\begin{equation*}
\min _{\mathbf{x} \in \mathcal{T} \cap\left(\cup_{i, j} \widetilde{\mathcal{c}}_{i j}\right)}\|\mathbf{A x}\|_{2}>0 . \tag{3.4}
\end{equation*}
$$

For the simplicity of analysis, we assume that the entries $A_{i, j}$ of $\mathbf{A}$ are i.i.d. normal $\mathcal{N}(0,1)$. Using Gordon's escape from mesh theorem Gordon (1988), we obtain the following result that relates $\operatorname{Pr}(\mathcal{E})$ with the number of measurements $M$.

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

$$
\begin{equation*}
\operatorname{Pr}(\mathcal{E}) \geq 1-(\underbrace{\operatorname{Pr}\left(\min _{\mathbf{u} \in \mathcal{T}^{c} \backslash\{\mathbf{0}\}}\|\mathbf{A} \mathbf{u}\|_{2}>0\right)}_{\mathcal{P}_{1}} \wedge \underbrace{\operatorname{Pr}\left(\min _{\boldsymbol{u} \in \cap \widetilde{\mathcal{C}}_{i} \backslash\{\mathbf{0}\}}\|\mathbf{A} \mathbf{u}\|_{2}>0\right)}_{\mathcal{P}_{2}}) \tag{3.5}
\end{equation*}
$$

where $a \wedge b$ denotes the minimum of $a$ and $b$, and $\mathcal{P}_{1}$ and $\mathcal{P}_{2}$ can be bounded as

$$
\begin{align*}
& \mathcal{P}_{1} \leq 1 \wedge \exp \left(-\frac{\left(a_{M}-\omega(\mathcal{T})\right)^{2}}{2}\right) \\
& \mathcal{P}_{2} \leq 1 \wedge \frac{3}{2} \exp \left(-\frac{\epsilon^{2} a_{M}^{2}}{2}\right)+\sum_{i \leq j} \exp \left(-\frac{\left((1-2 \epsilon) a_{M}-\omega\left(\widetilde{\mathcal{C}}_{i j}\right)\right)^{2}}{2}\right) \tag{3.6}
\end{align*}
$$

Thm. 1 links the probability of correct recovery of Eq. (2.2) with the number of measurements $M$, and the "size" of constraint set. Detailed explanation is given as the following. To ensure highprobability of $\mathcal{E}$, we would like to $\mathcal{P}_{1} \wedge \mathcal{P}_{2}$ to approach zero, which requires large value of $a_{M}$. Meanwhile, $a_{M}$ is a monotonically increasing function of the sensor number $M$. Hence, we can obtain the minimum sensor number $M$ requirement by unique recovery via investigating $a_{M}$.
Remark 1. Notice that $\mathcal{P}_{1}$ is associated with the descent cone $\mathcal{T}$ of the optimization function, namely, $\|\mathbf{x}\|_{1}$, while $\mathcal{P}_{2}$ is associated with the prior knowledge $\mathbf{x} \in \bigcup_{i} \mathcal{C}_{i}$. Thm. 1 implies that event $\mathcal{E}$ (uniqueness) holds with higher probability than the traditional CS due to the extra constraint $\mathbf{x} \in \bigcup_{i} \mathcal{C}_{i}$. If we fix $\operatorname{Pr}(\mathcal{E})$, we can separately calculate the corresponding $M$ with and without the constraint $\mathbf{x} \in \bigcup_{i} \mathcal{C}_{i}$. The difference $\Delta M$ would indicate the savings in the number of measurements due to the additional structure $\mathbf{x} \in \bigcup_{i} \mathcal{C}_{i}$ over the traditional CS.

One simple example is attached below to illustrate the improvement brought by Thm. 1.
Example 2. Consider the constraint set

$$
\begin{equation*}
\mathcal{C}_{i}=\left\{\left(0, \cdots, 0, x_{i}, \cdots, x_{K+i}, 0, \cdots, 0\right)\right\} \tag{3.7}
\end{equation*}
$$

where $1 \leq i \leq N-K$. We study the asymptotic behavior of Thm. 1 when $N$ is of order $\mathcal{O}\left(K^{c}\right)$, where $c>1$ is constant. In the sequel we will show that Thm. 1 gives us the order $M=\mathcal{O}(K)$ to ensure solution uniqueness as K approaches infinity, which gives us the same bound as shown in Baraniuk et al. (2010) and suggests the tightness of our result.
Setting $\epsilon=1 / 4$, we can bound $\mathcal{P}_{2}$ as

$$
\begin{equation*}
\mathcal{P}_{2} \leq \frac{3}{2} \exp \left(-\frac{a_{M}^{2}}{32}\right)+\frac{N^{2}}{2} \exp \left(-\frac{\left(a_{M}-2 a_{2 K}\right)^{2}}{8}\right) \tag{3.8}
\end{equation*}
$$

provided $a_{M} \geq 2 a_{K}$. With the relation $\frac{M}{\sqrt{M+1}} \leq a_{M} \leq \sqrt{M}$ Chandrasekaran et al. (2012); Gordon (1988) and setting $M=3 K$, we have

$$
\begin{equation*}
\mathcal{P}_{2} \leq c_{1} \exp \left(-c_{2} K\right)+c_{3} N^{2} \exp \left(-c_{4} K\right) \tag{3.9}
\end{equation*}
$$

where $c_{1}, c_{2}, c_{3}, c_{4}>0$ are some positive constants. Since $N=\mathcal{O}\left(K^{c}\right)$, 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 $M=\mathcal{O}(K \log N / K)=\mathcal{O}(K \log K)$ to $M=\mathcal{O}(K)$.

## 4 Computational Algorithm

Apart from the statistical property, another important aspect of Eq. (2.2) is to design an efficient algorithm. One naive idea is to consider and solve $L$ separate optimization problems

$$
\begin{equation*}
\widehat{\mathbf{x}}^{(i)}=\operatorname{argmin}_{\mathbf{x}}\|\mathbf{x}\|_{1}, \text { s.t. } \mathbf{y}=\mathbf{A} \mathbf{x}, \quad \mathbf{x} \in \mathcal{C}_{i}, \tag{4.1}
\end{equation*}
$$

and then selecting the best one, i.e., the sparsest reconstructed signal among all $\widehat{\mathbf{x}}^{(i)}$ 's. However, this method has two drawbacks:

- It requires solving $L$ separate optimization problems, which in many applications might be prohibitively large and difficult to handle, but the proposed method is based on one single optimization procedure.
- It is inflexible. For example, some prior knowledge of which $\mathcal{C}_{i}$ the true signal $\mathbf{x}^{\natural}$ is more likely to reside might be available. The above method cannot incorporate such priors.

To overcome the above drawbacks, we (i) reformulate Eq. (2.2) to a more tractable objective function, and (ii) propose a computationally efficient algorithm to solve it. In the following, we assume that $\mathbf{x}$ is bounded in the sense that for a constant $R,\|\mathbf{x}\|_{2} \leq R$.

### 4.1 Reformulation of the objective function

We introduce an auxiliary variable $\mathbf{p}$ and rewrite the Lagrangian form in Eq. (2.2) as

$$
\begin{equation*}
\min _{\mathbf{x}} \min _{\mathbf{p} \in \Delta_{L}} \sum_{i} p_{i}\left(\|\mathbf{x}\|_{1}+\widetilde{\mathbb{1}}\left(\mathbf{x} \in \mathcal{C}_{i}\right)+\frac{\lambda_{1}}{2}\|\mathbf{y}-\mathbf{A} \mathbf{x}\|_{2}^{2}+\frac{\lambda_{2}\|\mathbf{x}\|_{2}^{2}}{2}\right) \tag{4.2}
\end{equation*}
$$

where $\Delta_{L}$ is the simplex $\left\{p_{i} \geq 0, \sum_{i} p_{i}=1\right\}, \widetilde{1}(\cdot)$ 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 $\|\mathbf{y}-\mathbf{A x}\|_{2}^{2}$ is used to penalize for the constraint $\mathbf{y}=\mathbf{A x}$ while $\|\mathbf{x}\|_{2}^{2}$ corresponds to the energy constraint $\|\mathbf{x}\|_{2} \leq R$. It can be easily shown that solving Eq. (4.2) for large enough $C$ ensures $\mathbf{x} \in \bigcup_{i}^{L} \mathcal{C}_{i}$.

```
Algorithm 1 Non-convex Proximal Multiplicative Weighting Algorithm
    - Initialization: Initialize all variables with uniform weight \(p_{i}^{(0)}=L^{-1}\) and \(\mathbf{x}^{(0)}=\mathbf{0}\).
    - For time \(t=1\) to \(T\) : We update \(p_{i}^{(t+1)}\) and \(\mathbf{x}^{(t)}\) as
\[
\begin{align*}
& p_{i}^{(t+1)} \propto p_{i}^{(t)} e^{-\eta_{p}^{(t)} f_{i}\left(\mathbf{x}^{(t)}\right)}  \tag{4.3}\\
& \mathbf{x}^{(t+1)}=\operatorname{prox}_{\eta_{w}^{(t)}\|\cdot\|_{1}}\left[\mathbf{x}^{(t)}-\eta_{x}^{(t)} \sum_{i} p_{i}^{(t)}\left(\nabla_{x} h_{i}\left(\mathbf{x}^{(t)}\right)+\lambda_{1} \mathbf{A}^{\top}\left(\mathbf{A} \mathbf{x}^{(t)}-\mathbf{y}\right)+\lambda_{2} \mathbf{x}^{(t)}\right)\right],
\end{align*}
\]
```

where $p_{i}^{(t)}$ denotes the $i$ th element of $\mathbf{p}^{(t)}$, and the proximal operator prox $\|_{\|}(\mathbf{x})$ is defined as $\operatorname{argmin}_{\mathbf{z}}\left(\|\mathbf{z}\|_{1}+\frac{1}{2}\|\mathbf{z}-\mathbf{x}\|_{2}^{2}\right)$ Beck \& Teboulle (2009).

- Output: Calculate the average value $\overline{\mathbf{p}}=\frac{\sum_{t} \mathbf{p}^{(t)}}{T}$ and value $\overline{\mathbf{x}}=\frac{\sum_{t} \mathbf{x}^{(t)}}{T}$. Then output $\widehat{\mathbf{x}}$ by projecting $\overline{\mathbf{x}}$ onto the set of $\bigcup_{i} \mathcal{C}_{i}$.

Apart from the universality, our formulation has the following benefits:

- It is memory efficient. Compared with the naive idea that needs to store $L$ different $\widehat{\mathbf{x}}^{(i)}$, our method only needs to track one $\widehat{\mathbf{x}}$ and one redundant variable $\mathbf{p}$. This reduces the storing memory from $\mathcal{O}(N L)$ to $\mathcal{O}(N+L)$.
- It is very flexible. We can easily adjust to the case that $\mathbf{x}$ belongs to the intersection, i.e., $\mathbf{x} \in \bigcap_{i} \mathcal{C}_{i}$ via modifying $\min _{\mathbf{p} \in \Delta_{L}}$ in Eq. (4.2) to $\max _{\mathbf{p} \in \Delta_{L}}$.
Besides, to the best of our knowledge, this is the first time that such a formulation Eq. (4.2) is proposed. In the following, we will focus on the computational methods. Note that the difficulties in solving Eq. (4.2) are due to two aspects:
- Optimization over p: Although classical methods to minimize over $\mathbf{p}$ with fixed $\mathbf{x}$, e.g., alternative minimization and ADMM Boyd (n.d.), can calculate local minimum efficiently (due to the biconvexity of Eq. (4.2), they can be easily trapped in the local-minima. This is because some entries in $\mathbf{p}$ can be set to zero and hence $\mathbf{x}$ will be kept away from the corresponding set $\mathcal{C}_{i}$ thereafter. To handle this problem, we propose to use multiplicative weight update Arora et al. (2012) and update $\mathbf{p}$ with the relation $\mathbf{p}^{(t+1)} \propto \mathbf{p}^{(t)} e^{-\eta_{p}^{(t)} f_{i}(\mathbf{x})}$, where $\mathbf{p}^{(t)}$ denotes $\mathbf{p}^{\prime}$ s value in the $t$ th iteration. This update relation avoids the sudden change of $\mathbf{p}^{(t)}$ 's entries from non-zero to zero, which could have forced $\mathbf{x}^{(t)}$ being trapped in a local minimum.
- Optimization over $\mathbf{x}$ : Due to the non-smoothness of $\widetilde{\mathbb{1}}\left(\mathbf{x} \in \mathcal{C}_{i}\right)$ and $\|\mathbf{x}\|_{1}$ in Eq. (4.2) and the difficulties in calculating their sub-gradients, directly minimizing Eq. (4.2) would be computationally prohibitive. We propose to first approximate $\widetilde{\mathbb{1}}\left(\mathbf{x} \in \mathcal{C}_{i}\right)$ with a smooth function $h_{i}(\mathbf{x})$ and update $\mathbf{x}^{(t)}$ with the relation Eq. (4.4) used in proximal gradient descent Beck (2017).
Definition 3 ( $L_{g}$-strongly smooth Beck (2017)). Function $g(\cdot): \mathcal{X} \mapsto \mathbb{R}$ is $L_{g}$-strongly smooth iff

$$
\begin{equation*}
g(\mathbf{y}) \leq g(\mathbf{x})+\langle\nabla g(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle+\frac{L_{g}}{2}\|\mathbf{x}-\mathbf{y}\|_{2}^{2} \tag{4.4}
\end{equation*}
$$

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

### 4.2 Non-convex Proximal Multiplicative Weighting Algorithm

Here we directly approximate the truncated indicator function $\widetilde{\mathbb{1}}\left(\mathbf{x} \in \mathcal{C}_{i}\right)$ by $L_{h, i}$ strongly-smooth convex penalty functions $h_{i}(\mathbf{x})$, which may be different for different shapes of convex sets. For example, consider the convex set $\mathcal{C}_{i}$ in Example. 2. We may define $h_{i}(\mathbf{x})=\sum_{j \notin[i, i+K]}^{N} x_{j}^{2}$, where $[a, b]$ denotes the region from $a$ to $b$. While for the set $\{\mathbf{x}:\langle\mathbf{a}, \mathbf{x}\rangle \leq b\}$, we may instead adopt the modified log-barrier function with a finite value. Then Eq. (4.2) can be rewritten as

$$
\begin{equation*}
\min _{\mathbf{p}} \min _{\mathbf{x}} \mathcal{L}(\mathbf{p}, \mathbf{x}) \triangleq \sum_{i=1}^{L} p_{i} f_{i}(\mathbf{x}) \tag{4.5}
\end{equation*}
$$

where $f_{i}(\mathbf{x})$ is defined as

$$
\begin{equation*}
f_{i}(\mathbf{x}) \triangleq\|\mathbf{x}\|_{1}+h_{i}(\mathbf{x})+\frac{\lambda_{1}}{2}\|\mathbf{y}-\mathbf{A} \mathbf{x}\|_{2}^{2}+\frac{\lambda_{2}}{2}\|\mathbf{x}\|_{2}^{2} \tag{4.6}
\end{equation*}
$$

Hence, the optimization problem in (4.5) can be solved via Alg. 1
Lemma 2. $h(\mathbf{x}) \triangleq \sum_{i} p_{i} h_{i}(\mathbf{x})+\frac{\lambda_{1}\|\mathbf{y}-\mathbf{A x}\|_{2}^{2}}{2}+\frac{\lambda_{2}\|\mathbf{x}\|_{2}^{2}}{2}$ is strongly-smooth with some positive constant denoted as $L_{h}$.

Proof. First, we can check that $\frac{\lambda_{1}\|\mathbf{y}-\mathbf{A x}\|_{2}^{2}}{2}+\frac{\lambda_{2}\|\mathbf{x}\|_{2}^{2}}{2}$ is strongly-smooth. Denote the corresponding parameter as $L_{h, 0}$. Meanwhile, due to the construction of $h_{i}(\mathbf{x})$, it is strongly-smooth for every $i$. Since $p_{i}$ is non-negative for every $i$, we can easily prove the following inequality

$$
\begin{equation*}
h\left(\mathbf{x}_{1}\right) \geq h\left(\mathbf{x}_{2}\right)+\left\langle\nabla h\left(\mathbf{x}_{2}\right), \mathbf{x}_{1}-\mathbf{x}_{2}\right\rangle+\frac{L_{h}}{2}\left\|\mathbf{x}_{1}-\mathbf{x}_{2}\right\|_{2}^{2} \tag{4.7}
\end{equation*}
$$

where $L_{h}$ is defined as $\min \left(L_{h, i}\right), 0 \leq i \leq L$.
Then we have the following theorem.
Theorem 3. Let $\eta_{x}^{(t)}=\eta_{x} \leq L_{h}^{-1}$, and $\eta_{p}^{(t)}=R_{f}^{-1} \sqrt{2 \log L / T}$, where $\left|f_{i}(\cdot)\right| \leq R_{f},\|\mathbf{x}\|_{2} \leq R$. Then we have

$$
\begin{align*}
& \left|\frac{\min _{\mathbf{p}} \sum_{t} \mathcal{L}\left(\mathbf{p}, \mathbf{x}^{(t)}\right)}{T}-\frac{\sum_{t} \mathcal{L}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t)}\right)}{T}\right| \\
+ & \left|\frac{\sum_{t} \mathcal{L}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t)}\right)}{T}-\frac{\min _{\mathbf{x}} \sum_{t} \mathcal{L}\left(\mathbf{p}^{(t)}, \mathbf{x}\right)}{T}\right|  \tag{4.8}\\
\leq & \frac{2 R^{2}}{\eta_{x} T}+R_{f} \sqrt{\frac{\log L}{T}}
\end{align*}
$$

where $T$ denotes the number of iterations.
Due to the difficulties in analyzing the global optimum, in Theorem 3 we focus on analyzing the closeness between the average value $\frac{\Sigma_{\not} \mathcal{L}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t)}\right)}{T}$ to its local minimum. The first term denotes the gap between average value $\frac{\sum_{\mathcal{L}} \mathcal{L}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t)}\right)}{T}$ and the optimal value of $\mathcal{L}(\mathbf{p}, \mathbf{x})$ with $\mathbf{x}^{(t)}$ being fixed. Similarly, the second term represents the gap with $\mathbf{p}^{(t)}$ being fixed. As $T \rightarrow \infty$, the sum of these two bounds approaches to zero at the rate of $\mathcal{O}\left(T^{-1 / 2}\right)$.
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 (Shalev-Shwartz et al. , 2012, §2.3.1). In addition, we have proved the following theorem.
Theorem 4. Let $\eta_{w}^{(t)} \leq L_{h}^{-1}$, where $\left|f_{i}(\cdot)\right| \leq R_{f}$. Then we have

$$
\begin{equation*}
\frac{1}{T} \sum_{t}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2} \leq \frac{2 \mathcal{L}\left(\mathbf{p}^{(0)}, \mathbf{x}^{(0)}\right)}{L_{h} T}+\frac{4 R_{f}^{2} \sum_{t} \eta_{p}^{(t)}}{L_{h} T} . \tag{4.9}
\end{equation*}
$$

This theorem discusses the convergence speed with respect to the $\mathbf{x}^{(t)}$ update. Due to the $\mathcal{O}\left(T^{-1}\right)$ of the first term on the right side of the above inequality, the best convergence rate we can obtain is $\mathcal{O}\left(T^{-1}\right)$, which is achievable by $\eta_{p}^{(t)} \propto t^{-2}$. However, using fixed learning rate $\eta_{p}$ as in Thm. 3 would result in the convergence rate of $\mathcal{O}\left(T^{-1 / 2}\right)$.

### 4.3 Regularization for $\mathbf{p}$

Another drawback of the naive method is that they cannot exploit the prior knowledge. For example, if we know that the true $\mathbf{x}^{\natural}$ is most likely to reside in set $\mathcal{C}_{1}$. With the naive method, we cannot use this information but separately solve Eq. (2.2) for all $L$ sets. In the sequel, we will show that our formulation Eq. (4.2) can incorporate such prior knowledge by adding regularizers for $\mathbf{p}$, and bring certain performance improvement.
Note that we can interpret $p_{i}$, the $i$-th element of $\mathbf{p}$ in Eq. (4.5) as the likelihood of $\mathbf{x}^{\natural} \in \mathcal{C}_{i}$. Without any prior knowledge about which set $\mathcal{C}_{i}$ the true signal $\mathbf{x}^{\natural}$ resides, variable $\mathbf{p}$ is uniformly distributed among all possible distributions $\Delta_{L}$. When certain prior information is available, its distribution is skewed towards certain distributions, namely $\mathbf{q}$.
In this paper, we adopt $\|\cdot\|_{2}^{2}$ to regularize $\mathbf{p}$ towards $\mathbf{q}$ and write the modified function $\mathcal{L R}(\mathbf{p}, \mathbf{x})$ as

$$
\begin{equation*}
\mathcal{L R}(\mathbf{p}, \mathbf{x})=\mathcal{L}(\mathbf{p}, \mathbf{x})+\frac{\lambda_{3}}{2}\|\mathbf{p}-\mathbf{q}\|_{2}^{2} \tag{4.10}
\end{equation*}
$$

where $\lambda_{3}>0$ is a constant used to balance $\mathcal{L}(\mathbf{p}, \mathbf{x})$ and $\frac{1}{2}\|\mathbf{p}-\mathbf{q}\|_{2}^{2}$. 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 Eq. (4.3) as

$$
\begin{equation*}
\mathbf{p}^{(t+1)}=\mathbb{P}_{\Delta}\left(\mathbf{p}^{(t)}-\eta_{p}^{(t)} \mathbf{g}^{(t)}\right) \tag{4.11}
\end{equation*}
$$

where $\mathbf{g}^{(t)}=\nabla_{\mathbf{p}^{(t)}} \mathcal{L} \mathcal{R}\left(\mathbf{p}, \mathbf{x}^{(t)}\right)=\mathbf{f}\left(\mathbf{x}^{(t)}\right)+\lambda_{3}\left(\mathbf{p}^{(t)}-\mathbf{q}\right)$, and $\mathbf{f}\left(\mathbf{x}^{(t)}\right)$ denotes the vector whose $i$ th element is $f_{i}\left(\mathbf{x}^{(t)}\right)$. Similar as above, we obtain the following theorems.
Theorem 5. Provided that $\left\|\mathbf{g}^{(t)}\right\|_{2} \leq R_{g}$, by setting $\eta_{x}^{(t)}=\eta_{x} \leq L_{h}^{-1}$ and $\eta_{p}^{(t)}=(\lambda t)^{-1}$ we conclude that

$$
\begin{align*}
& \left|\frac{\min _{\mathbf{p}} \sum_{t} \mathcal{L} \mathcal{R}\left(\mathbf{p}, \mathbf{x}^{(t)}\right)}{T}-\frac{\sum_{t} \mathcal{L} \mathcal{R}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t)}\right)}{T}\right| \\
+ & \left|\frac{\sum_{t} \mathcal{L} \mathcal{R}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t)}\right)}{T}-\frac{\min _{\mathbf{x}} \sum_{t} \mathcal{L} \mathcal{R}\left(\mathbf{p}^{(t)}, \mathbf{x}\right)}{T}\right|  \tag{4.12}\\
\leq & \frac{R_{g}^{2} \log T}{2 \lambda_{3} T}+\frac{R^{2}}{2 \eta_{x} T^{\prime}}
\end{align*}
$$

Comparing with Thm. 3, Thm. 5 implies that the regularizers improve the optimal rate from $\mathcal{O}\left(T^{-1 / 2}\right)$ 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 6. Provided that $\left\|\mathbf{g}^{(t)}\right\|_{2} \leq R_{g}$, by setting $\eta_{w}^{(t)}=\eta_{x} \leq L_{h}^{-1}$ we conclude that

$$
\begin{equation*}
\frac{1}{T} \sum_{t}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2} \leq \frac{2 \mathcal{L} \mathcal{R}\left(\mathbf{p}^{(0)}, \mathbf{x}^{(0)}\right)}{L_{h} T}+\frac{2 R_{g}^{2}}{L_{h} T} \sum_{t}\left(\eta_{p}^{(t)}+\frac{\lambda_{3}\left(\eta_{p}^{(t)}\right)^{2}}{2}\right) \tag{4.13}
\end{equation*}
$$

In this case, if we set $\eta_{p}^{(t)}$ as $t^{-2}$, then $\frac{1}{T} \sum_{t}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2}$ would decrease at the rate of $\mathcal{O}\left(T^{-1}\right)$, which is the same as Thm. 4.

## 5 Conclusion

In this paper, we studied the compressive sensing with a multiple convex-set domain. First we analyzed the impact of prior knowledge $\mathbf{x} \in \bigcup_{i} \mathcal{C}_{i}$ 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 $\mathcal{O}\left(T^{-1 / 2}\right)$. Further, in terms of $T^{-1}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2}$, the optimal speed increases to $\mathcal{O}\left(T^{-1}\right)$. Moreover, provided that we have a prior knowledge about $\mathbf{p}$, we show that we can improve the optimal recovery performance by $\|\cdot\|_{2}^{2}$ regularizers, and hence increasing the above convergence rate from $\mathcal{O}\left(T^{-1 / 2}\right)$ to $\mathcal{O}\left(\frac{\log T}{T}\right)$.

## References

Arora, Sanjeev, Hazan, Elad, \& Kale, Satyen. 2012. The Multiplicative Weights Update Method: a Meta-Algorithm and Applications. Theory of Computing, 8(1), 121-164.

Baraniuk, Richard G, Cevher, Volkan, Duarte, Marco F, \& Hegde, Chinmay. 2010. Model-based compressive sensing. IEEE Transactions on Information Theory, 56(4), 1982-2001.

Веск, A. 2017. First-Order Methods in Optimization. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics.

Beck, Amir, \& Teboulle, Marc. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM journal on imaging sciences, 2(1), 183-202.

Boucheron, S., Lugosi, G., \& Massart, P. 2013. Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford University Press.

Boyd, Stephen. Alternating direction method of multipliers.
Candes, Emmanuel J, Eldar, Yonina C, Strohmer, Thomas, \& Voroninski, Vladislav. 2011. Phase Retrieval via Matrix Completion.

Chandrasekaran, Venkat, Recht, Benjamin, Parrilo, Pablo A, \& Willsky, Alan S. 2012. The convex geometry of linear inverse problems. Foundations of Computational mathematics, 12(6), 805-849.

Chen, Yuxin, \& Candes, Emmanuel. 2015. Solving random quadratic systems of equations is nearly as easy as solving linear systems. Pages 739-747 of: Advances in Neural Information Processing Systems.

Dai, Wei, Pham, Hoa Vinh, \& Milenkovic, Olgica. 2009. Quantized compressive sensing. arXiv preprint arXiv:0901.0749.

Duarte, Marco F, \& Eldar, Yonina C. 2011. Structured compressed sensing: From theory to applications. IEEE Transactions on Signal Processing, 59(9), 4053-4085.

Eldar, Y.C., \& Kutyniok, G. 2012. Compressed Sensing: Theory and Applications. Compressed Sensing: Theory and Applications. Cambridge University Press.

Foucart, Simon, \& Rauhut, Holger. A mathematical introduction to compressive sensing. Vol. 1.
Gordon, Yehoram. 1988. On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^{n}$. Pages 84-106 of: Geometric Aspects of Functional Analysis. Springer.
Qian, Qi, Zhu, Shenghuo, Tang, Jiasheng, Jin, Rong, Sun, Baigui, \& Li, Hao. 2018. Robust Optimization over Multiple Domains. arXiv preprint arXiv:1805.07588.
Shalev-Shwartz, Shai, et al. . 2012. Online learning and online convex optimization. Foundations and Trends $®$ ) in Machine Learning, 4(2), 107-194.
Silva, Jorge, Chen, Minhua, Eldar, Yonina C, Sapiro, Guillermo, \& Carin, Lawrence. 2011. Blind compressed sensing over a structured union of subspaces. arXiv preprint arXiv:1103.2469.

Zhang, H., Abdi, A., \& Fekri, F. Nov. 2017.. Compressive Sensing with Energy Constraint. In: IEEE Information Theory Workshop (ITW'17).

## Appendix A Proof of Theorem 1

Proof. Note that for any-vector non-zero $h \in \operatorname{null}(\mathbf{A}) \cap \mathcal{T} \cap\left(\bigcup_{i, j} \widetilde{\mathcal{C}}_{i, j}\right)$, we can always rescale to make it unit-norm. Hence we can rewrite the event $\mathcal{E}$ as

$$
\begin{equation*}
\mathcal{E}=\left\{\operatorname{null}(\mathbf{A}) \bigcap \mathrm{S}_{2}^{n-1} \bigcap \mathcal{T} \bigcap\left(\bigcup_{i, j} \widetilde{\mathcal{C}}_{i, j}\right)=\varnothing\right\} . \tag{A.1}
\end{equation*}
$$

For the conciseness of notation, we define $\widetilde{\mathcal{C}}$ to be $\widetilde{\mathcal{C}}=\bigcup_{i, j} \widetilde{\mathcal{C}}_{i, j}$. Then we upper-bound $1-\operatorname{Pr}(\mathcal{E})$ as

$$
\begin{align*}
& 1-\operatorname{Pr}(\mathcal{E})=\operatorname{Pr}\left(\operatorname{null}(\mathbf{A}) \bigcap S_{2}^{n-1} \bigcap \mathcal{T} \cap \widetilde{\mathcal{C}} \neq \varnothing\right) \\
& \underbrace{(i)}_{\mathcal{P}_{1}} \underbrace{\operatorname{Pr}\left(\operatorname{null}(\mathbf{A}) \bigcap S_{2}^{n-1} \bigcap \mathcal{T} \neq \varnothing\right)}_{\mathcal{P}_{2}} \wedge \underbrace{\operatorname{Pr}\left(\operatorname{null}(\mathbf{A}) \bigcap S_{2}^{n-1} \cap \widetilde{\mathcal{C}} \neq \varnothing\right)}, \tag{A.2}
\end{align*}
$$

where $(i)$ is because

$$
\begin{align*}
& \left\{\operatorname{null}(\mathbf{A}) \bigcap \mathrm{S}_{2}^{n-1} \cap \mathcal{T} \cap \widetilde{\mathcal{C}} \neq \varnothing\right\} \subseteq\left\{\operatorname{null}(\mathbf{A}) \bigcap \mathrm{S}_{2}^{n-1} \cap \mathcal{T} \neq \varnothing\right\}, \\
& \left\{\operatorname{null}(\mathbf{A}) \bigcap \mathrm{S}_{2}^{n-1} \cap \mathcal{T} \bigcap \widetilde{\mathcal{C}} \neq \varnothing\right\} \subseteq\left\{\operatorname{null}(\mathbf{A}) \bigcap \mathrm{S}_{2}^{n-1} \cap \widetilde{\mathcal{C}} \neq \varnothing\right\} \tag{A.3}
\end{align*} .
$$

With Lemma 3 and Lemma 4 , we can separately bound $\mathcal{P}_{1}$ and $\mathcal{P}_{2}$ and finish the proof.

Lemma 3. We have $\mathcal{P}_{1} \leq 1 \wedge \exp \left(-\frac{\left(a_{m}-\omega(\mathcal{T})\right)^{2}}{2}\right)$, if $a_{m} \geq \omega(\mathcal{T})$.
Proof. Note that we have

$$
\begin{equation*}
\underbrace{\operatorname{Pr}\left(\operatorname{null}(\mathbf{A}) \bigcap S_{2}^{n-1} \bigcap \mathcal{T} \neq \varnothing\right)}_{\mathcal{P}_{1}}+\underbrace{\operatorname{Pr}\left(\operatorname{null}(\mathbf{A}) \bigcap \mathrm{S}_{2}^{n-1} \cap \mathcal{T}=\varnothing\right)}_{\mathcal{P}_{1}^{c}}=1 \tag{A.4}
\end{equation*}
$$

Then we lower-bound $\mathcal{P}_{1}^{c}$ as

$$
\begin{equation*}
\mathcal{P}_{1}^{c}=\operatorname{Pr}\left(\min _{\mathbf{u} \in S_{2}^{n-1} \cap \mathcal{T}}\|\mathbf{A u}\|_{2}>0\right) \stackrel{(i)}{\geq} 1-\exp \left(-\frac{\left(a_{m}-\omega(\mathcal{T})\right)^{2}}{2}\right), \tag{A.5}
\end{equation*}
$$

provided $a_{m} \geq \omega(\mathcal{T})$, where ( $i$ ) is because of Corollary 3.3 in Chandrasekaran et al. (2012), and $\omega(\cdot)$ denotes the Gaussian width.

Lemma 4. If $\omega\left(\widetilde{\mathcal{C}}_{i j}\right) \leq 1-2 \epsilon a_{M}$, we have

$$
\begin{equation*}
\mathcal{P}_{2} \leq 1 \wedge \frac{3}{2} \exp \left(-\frac{\epsilon^{2} a_{M}^{2}}{2}\right)+\sum_{i \leq j} \exp \left(-\frac{\left((1-2 \epsilon) a_{M}-\omega\left(\widetilde{\mathcal{C}}_{i j}\right)\right)^{2}}{2}\right) \tag{A.6}
\end{equation*}
$$

Proof. Note that we have

$$
\begin{equation*}
\underbrace{\operatorname{Pr}\left(\operatorname{null}(\mathbf{A}) \bigcap S_{2}^{n-1} \bigcap \widetilde{\mathcal{C}} \neq \varnothing\right)}_{\mathcal{P}_{2}}+\underbrace{\operatorname{Pr}\left(\operatorname{null}(\mathbf{A}) \bigcap S_{2}^{n-1} \cap \widetilde{\mathcal{C}}=\varnothing\right)}_{\mathcal{P}_{2}^{c}}=1 \tag{A.7}
\end{equation*}
$$

Here we upper-bound $\mathcal{P}_{2}$ via lower-bounding $\mathcal{P}_{2}^{c}$. First we define $\mathcal{P}_{2}^{c}(d)$ as

$$
\begin{equation*}
\mathcal{P}_{2}^{c}(d) \triangleq \operatorname{Pr}\left(\min _{\mathbf{u} \in \cup \mathcal{S}_{i, j}, \mathbf{v} \in \operatorname{null}(\mathbf{A})}\|\mathbf{u}-\mathbf{v}\|_{2} \geq d\right) \tag{A.8}
\end{equation*}
$$

Then we have $\mathcal{P}_{2}^{c}=\lim _{d \rightarrow 0} \mathcal{P}_{2}^{c}(d)$. The following proof trick is fundamentally the same as that are used in Theorem 4.1 in Gordon (1988) 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 $\mathcal{S}_{i, j}=S_{2}^{n-1} \cap \widetilde{\mathcal{C}}_{i, j}$ and two quantities $Q_{1}$ and $Q_{2}$ as

$$
\begin{align*}
& Q_{1} \triangleq \operatorname{Pr}\left(\min _{\mathbf{u} \in \cup \mathcal{S}_{i, j}}\|\mathbf{A} \mathbf{u}\|_{2} \geq d\left(1+\epsilon_{1}\right) \mathbb{E}\|\mathbf{A}\|_{2}\right) \\
& Q_{2} \triangleq \operatorname{Pr}\left(\bigcap_{i_{1} \leq j_{1}} \bigcap_{\mathbf{u} \in \mathcal{S}_{i_{1}, j_{1}}}\left(\left(\sum_{i_{2}=1}^{M} g_{i_{2}}^{2}\right)^{1 / 2}+\sum_{j_{2}} u_{j_{2}} h_{j_{2}} \geq d\left(1+\epsilon_{1}\right) \mathbb{E}\|\mathbf{A}\|_{2}+\epsilon_{2} a_{M}\right)\right) \tag{A.9}
\end{align*}
$$

where $a_{M}=\mathbb{E}\|\mathbf{g}\|_{2}, \mathbf{g} \in \mathcal{N}\left(\mathbf{0}, \mathbf{I}_{M \times M}\right)$, and $g_{j}, h_{i}$ are iid standard normal random variables $\mathcal{N}(0,1)$. The following proof is divided into 3 parts.
Step I. We prove that $\mathcal{P}_{2}^{c}(d)+e^{-\epsilon_{1}^{2} a_{M}^{2} / 2} \geq Q_{1}$, which is done by

$$
\begin{align*}
& Q_{1}=\operatorname{Pr}\left(\min _{\mathbf{u} \in \cup \mathcal{S}_{i, j}}\|\mathbf{A u}\|_{2} \geq d\left(1+\epsilon_{1}\right) \mathbb{E}\|\mathbf{A}\|_{2}\right) \\
& =\operatorname{Pr}\left(\min _{\mathbf{u} \in \cup \mathcal{S}_{i, j}, \mathbf{v} \in \operatorname{null}(\mathbf{A})}\|\mathbf{A}(\mathbf{u}-\mathbf{v})\|_{2} \geq d\left(1+\epsilon_{1}\right) \mathbb{E}\|\mathbf{A}\|_{2}\right) \\
& \stackrel{\text { (i) }}{\leq} \operatorname{Pr}\left(\min _{\mathbf{u} \in \cup \mathcal{S}_{i, j,}, \mathbf{v} \in \operatorname{null}(\mathbf{A})}\|\mathbf{A}\|_{2}\|\mathbf{u}-\mathbf{v}\|_{2} \geq d\left(1+\epsilon_{1}\right) \mathbb{E}\|\mathbf{A}\|_{2}\right) \\
& \leq \operatorname{Pr}\left(\|\mathbf{A}\|_{2} \geq\left(1+\epsilon_{1}\right) \mathbb{E}\|\mathbf{A}\|_{2}\right)+\underbrace{\operatorname{Pr}\left(\min _{\mathbf{u} \in \cup \mathcal{S}_{i, j}, \mathbf{v} \in \operatorname{null}(\mathbf{A})}\|\mathbf{u}-\mathbf{v}\|_{2} \geq d\right)}  \tag{A.10}\\
& \leq \mathcal{P}_{2}^{c}(d) \\
& \leq \\
& \text { (iii) } \exp \left(-\frac{\epsilon_{1}^{2}\left(\mathbb{E}\|\mathbf{A}\|_{2}\right)^{2}}{2}\right)+\mathcal{P}_{2}^{c}(d) \\
& \text { (iv) } \exp \left(-\frac{\epsilon_{1}^{2} a_{M}^{2}}{2}\right)+\mathcal{P}_{2}^{c}(d),
\end{align*}
$$

where in $(i)$ we use $\|\mathbf{A}\|_{2}\|\mathbf{u}-\mathbf{v}\|_{2} \geq\|\mathbf{A}(\mathbf{u}-\mathbf{v})\|_{2}$, in (ii) we use the union bound for

$$
\begin{align*}
& \left\{\min _{\mathbf{u} \in \cup \mathcal{S}_{i, j}, \mathbf{v} \in \operatorname{null}(\mathbf{A})}\|\mathbf{A}\|_{2}\|\mathbf{u}-\mathbf{v}\|_{2} \geq d\left(1+\epsilon_{1}\right) \mathbb{E}\|\mathbf{A}\|_{2}\right\}  \tag{A.11}\\
\subseteq & \left\{\|\mathbf{A}\|_{2} \geq\left(1+\epsilon_{1}\right) \mathbb{E}\|\mathbf{A}\|_{2}\right\} \cup\left\{\begin{array}{l}
\left.\min _{\mathbf{u} \in \cup \mathcal{S}_{i, j}, \mathbf{v} \in \operatorname{null}(\mathbf{A})}\|\mathbf{u}-\mathbf{v}\|_{2} \geq d\right\}
\end{array}\right.
\end{align*}
$$

in (iii) we use the Gaussian concentration inequality Lipschitz functions (Theorem 5.6 in Boucheron et al. (2013)) for $\|\mathbf{A}\|_{2}$, and in $(v)$ we use $\|\mathbf{A}\|_{2} \geq\left\|\mathbf{A} \mathbf{e}_{1}\right\|_{2}=\left\|\sum_{i=1}^{M} A_{i, 1}\right\|_{2}$, where $\mathbf{e}_{1}$ denotes the canonical basis.
Step II. We prove that $Q_{1}+\frac{1}{2} e^{-\epsilon_{2}^{2} a_{M}^{2} / 2} \geq Q_{2}$, which is done by

$$
\begin{align*}
& \quad Q_{1}+\frac{1}{2} e^{-\epsilon^{2} a_{M}^{2} / 2} \stackrel{(i)}{\geq} Q_{1}+\operatorname{Pr}\left\{g \geq \epsilon_{2} a_{M}\right\} \\
& \stackrel{\text { (ii) }}{\geq} \operatorname{Pr}\left(\min _{\mathbf{u} \in \cup \mathcal{S}_{i_{1}, j_{1}}}\|\mathbf{A u}\|_{2}+g\|\mathbf{u}\|_{2} \geq d(1+\epsilon) \mathbb{E}\|\mathbf{A}\|_{2}+\epsilon_{2} a_{M}\|\mathbf{u}\|_{2}\right) \\
& =\operatorname{Pr}\left(\bigcap_{i_{1} \leq j_{1}} \bigcap_{\mathbf{u} \in \mathcal{S}_{i_{1}, j_{1}}}\|\mathbf{A} \mathbf{u}\|_{2}+g\|\mathbf{u}\|_{2} \geq d(1+\epsilon) \mathbb{E}\|\mathbf{A}\|_{2}+\epsilon_{2} a_{M}\|\mathbf{u}\|_{2}\right)  \tag{A.12}\\
& \stackrel{\text { (iii) }}{\geq} \operatorname{Pr}\left(\bigcap_{Q_{2}} \bigcap_{i_{1} \leq j_{1}} \bigcap_{u \in \mathcal{S}_{i_{1}, j_{1}}}\left(\sum_{i_{2}=1}^{M} g_{i_{2}}^{2}\right)^{\frac{1}{2}}+\sum_{j_{2}=1}^{N} u_{j_{2}} h_{j_{2}} \geq d\left(1+\epsilon_{1}\right) \mathbb{E}\|\mathbf{A}\|_{2}+\epsilon_{2} a_{M}\right)
\end{align*}
$$

where in $(i) \mathbf{g}$ is a RV satisfying standard normal distribution, in (ii) we use the union bound, and (iii) comes from Lemma 3.1 in Gordon (1988) and $\|\mathbf{u}\|_{2}=1$.

Step III. We lower bound $Q_{2}$ as

$$
\begin{align*}
& 1-Q_{2}=\operatorname{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)^{\frac{1}{2}}+\sum_{j_{2}=1}^{N} u_{j_{2}} h_{j_{2}} \leq d\left(1+\epsilon_{1}\right) \mathbb{E}\|\mathbf{A}\|_{2}+\epsilon_{2} a_{M}\right]\right) \\
& \leq \operatorname{Pr}\left(\left(\sum_{i_{2}=1}^{M} g_{i_{2}}^{2}\right)^{\frac{1}{2}} \leq\left(1-\epsilon_{2}\right) a_{M}\right)+\operatorname{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\left(1+\epsilon_{1}\right) \mathbb{E}\|\mathbf{A}\|_{2}-\left(1-2 \epsilon_{2}\right) a_{M}\right) \\
& \leq \operatorname{Pr}\left(\left(\sum_{i_{2}=1}^{M} g_{i_{2}}^{2}\right)^{\frac{1}{2}}-a_{M} \leq-\epsilon_{2} a_{M}\right)+\operatorname{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\left(1+\epsilon_{1}\right) \mathbb{E}\|\mathbf{A}\|_{2}-\left(1-2 \epsilon_{2}\right) a_{M}\right) \\
& \stackrel{(i)}{\leq} \exp \left(-\frac{\epsilon_{2}^{2} a_{M}^{2}}{2}\right)+\operatorname{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\left(1+\epsilon_{1}\right) \mathbb{E}\|\mathbf{A}\|_{2}-\left(1-2 \epsilon_{2}\right) a_{M}\right) \\
& \stackrel{(i i)}{\leq} \exp \left(-\frac{\epsilon_{2}^{2} a_{M}^{2}}{2}\right)+\sum_{i_{1} \leq j_{1}} \operatorname{Pr}\left(\bigcup_{u \in \mathcal{S}_{i_{1}, j_{1}}} \sum_{j_{2}=1}^{N} u_{j_{2}} h_{j_{2}} \leq d\left(1+\epsilon_{1}\right) \mathbb{E}\|\mathbf{A}\|_{2}-\left(1-2 \epsilon_{2}\right) a_{M}\right) \\
& \stackrel{(\text { iii }}{\leq} \exp \left(-\frac{\epsilon_{2}^{2} a_{M}^{2}}{2}\right)+\sum_{i_{1} \leq j_{1}} \operatorname{Pr}\left(\max _{u \in \mathcal{S}_{i_{1} j_{1}}} \sum_{j_{2}=1}^{N} u_{j_{2}} h_{j_{2}}^{\prime} \geq\left(1-2 \epsilon_{2}\right) a_{M}-d\left(1+\epsilon_{1}\right) \mathbb{E}\|\mathbf{A}\|_{2}\right) \\
& \stackrel{(i v)}{\leq} \exp \left(-\frac{\epsilon_{2}^{2} a_{M}^{2}}{2}\right)+\sum_{i_{1} \leq j_{1}} \exp \left(-\frac{\left(\left(1-2 \epsilon_{2}\right) a_{M}-d\left(1+\epsilon_{1}\right) \mathbb{E}\|\mathbf{A}\|_{2}-\omega\left(\widetilde{\mathcal{C}_{i j}}\right)\right)^{2}}{2}\right) \text {, } \tag{A.13}
\end{align*}
$$

where in (i) we use $\mathbb{E} \sqrt{\sum_{i_{2}=1}^{M} g_{i_{2}}^{2}}=a_{M}$ and Gaussian concentration inequality in Boucheron et al. (2013), in (ii) we use union-bound, in (iii) we define $h^{\prime}=-h$ and flip the sign by the symmetry of Gaussian variables, and in (iv) we use the definition of $\omega\left(\widetilde{\mathcal{C}}_{i j}\right)$. Assuming $\left(1-2 \epsilon_{2}\right) a_{M} \geq d(1+$ $\left.\epsilon_{1}\right) \mathbb{E}\|\mathbf{A}\|_{2}+\omega\left(\widetilde{\mathcal{C}}_{i j}\right)$, we finish the proof via the Gaussian concentration inequality Boucheron et al. (2013).

Combining the above together and set $d\left(1+\epsilon_{1}\right) \rightarrow 0$ while $\epsilon_{1} \rightarrow \infty$, we conclude that

$$
\begin{equation*}
\mathcal{P}_{2}^{c} \geq 1-\frac{3}{2} \exp \left(-\frac{\epsilon^{2} a_{M}^{2}}{2}\right)-\sum_{i \leq j} \exp \left(-\frac{\left((1-2 \epsilon) a_{M}-\omega\left(\widetilde{\mathcal{C}}_{i j}\right)\right)^{2}}{2}\right) \tag{A.14}
\end{equation*}
$$

provided $(1-2 \epsilon) a_{M} \geq \omega\left(\widetilde{\mathcal{C}}_{i j}\right)$, and finish the proof.

## Appendix B Proof of Theorem 3

Proof. Define $\mathbf{p}^{*}$ and $\mathbf{x}^{*}$ as

$$
\begin{equation*}
\mathbf{p}^{*}=\operatorname{argmin}_{\mathbf{p}} \sum_{t} \mathcal{L}\left(\mathbf{p}, \mathbf{x}^{(t)}\right), \quad \mathbf{x}^{*}=\operatorname{argmin}_{\mathbf{x}} \sum_{t} \mathcal{L}\left(\mathbf{p}^{(t)}, \mathbf{x}\right) . \tag{B.1}
\end{equation*}
$$

respectively First we define $\mathcal{T}_{1}^{t}$ and $\mathcal{T}_{2}^{t}$ as

$$
\begin{align*}
& \mathcal{T}_{1}^{t}=\mathcal{L}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t)}\right)-\mathcal{L}\left(\mathbf{p}^{*}, \mathbf{x}^{(t)}\right) \\
& \mathcal{T}_{2}^{t}=\mathcal{L}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t)}\right)-\mathcal{L}\left(\mathbf{p}^{(t)}, \mathbf{x}^{*}\right), \tag{B.2}
\end{align*}
$$

respectively. Then our goal becomes bounding $\left|\sum_{t} \mathcal{T}_{1}^{t}\right|+\left|\sum_{t} \mathcal{T}_{2}^{t}\right|$. With Lemma 5 and Lemma 6, we have finished the proof.

Lemma 5. Define $\mathcal{T}_{1}^{t}=\mathcal{L}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t)}\right)-\mathcal{L}\left(\mathbf{p}^{*}, \mathbf{x}^{(t)}\right)$, where $\mathbf{p}^{*}$ is defined in Eq. (B.1), then we have

$$
\begin{equation*}
0<\sum_{t} \mathcal{T}_{1}^{t} \leq R_{f} \sqrt{T \log L} \tag{B.3}
\end{equation*}
$$

when $\eta_{p}^{(t)}=R_{f} \sqrt{2 \log L / T}$.

Proof. Based on the definition of $\mathbf{p}^{*}$, we note that $\sum_{t} \mathcal{T}_{1}^{t}$ is non-negative and prove the lower-bound. Then we prove its upper-bound.

Since the function is linear, optimal $\mathbf{p}^{*}$ 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 Arora et al. (2012). First we rewrite the update equation (4.3). Define $\mathbf{w}^{(0)}=\mathbf{1} \in \mathbb{R}^{L}$ and update $\mathbf{w}^{(t+1)}$ as

$$
\begin{equation*}
w_{i}^{(t+1)}=w_{i}^{(t)} \exp \left(-\eta_{p}^{(t)} f_{i}\left(\mathbf{x}^{(t)}\right)\right), p_{i}^{(t+1)}=\frac{w_{i}^{(t+1)}}{\sum_{i} w_{i}^{(t+1)}} \tag{B.4}
\end{equation*}
$$

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

$$
\begin{equation*}
\Psi_{t}=\sum_{i=1}^{L} w_{i}^{(t)} \tag{B.5}
\end{equation*}
$$

Then we have $\Psi_{0}=L$ while $\Psi_{T} \geq w_{i^{*}}^{(T)}=\exp \left(-\sum_{t=1}^{T} \eta_{p}^{(t)} f_{i^{*}}\left(\mathbf{x}^{(t)}\right)\right)=\exp \left(-\eta_{p} \sum_{t=1}^{T} f_{i^{*}}\left(\mathbf{x}^{(t)}\right)\right)$, where $\eta_{p}^{(t)}=\eta_{p}=\sqrt{2 \log L / T}$.

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

$$
\begin{align*}
& \Psi_{t+1}=\sum_{i} w_{i}^{(t+1)}=\sum_{i} w_{i}^{t} \exp \left(-\eta_{p} f_{i}\left(\mathbf{x}^{(t)}\right)\right) \\
& \stackrel{(i)}{\leq} \sum_{i} w_{i}^{t}\left(1-\eta_{p} f_{i}\left(\mathbf{x}^{(t)}\right)+\frac{\eta_{p}^{2} f_{i}^{2}\left(\mathbf{x}^{(t)}\right)}{2}\right) \\
&= \Psi_{t}\left(1-\eta_{p}\left\langle\mathbf{p}^{(t)}, f\left(\mathbf{x}^{(t)}\right)\right\rangle+\frac{\eta_{p}^{2} R_{f}^{2}}{2}\right)  \tag{B.6}\\
& \stackrel{(i i)}{\leq} \Psi_{t} \exp \left(-\eta_{p}\left\langle\mathbf{p}^{(t)}, f\left(\mathbf{x}^{(t)}\right)\right\rangle+\frac{\eta_{p}^{2} R_{f}^{2}}{2}\right)
\end{align*}
$$

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

$$
\begin{equation*}
\frac{\Psi_{T}}{\Psi_{0}} \leq \exp \left(-\eta_{p} \sum_{t}\left\langle\mathbf{p}^{(t)}, f\left(\mathbf{x}^{(t)}\right)\right\rangle+\frac{\eta_{p}^{2} T R_{f}^{2}}{2}\right), \tag{B.7}
\end{equation*}
$$

which gives us

$$
\begin{equation*}
\log \Psi_{T} \leq \log L-\eta_{p} \sum_{t}\left\langle\mathbf{p}^{(t)}, f\left(\mathbf{x}^{(t)}\right)\right\rangle+\frac{T \eta_{p}^{2} R_{f}^{2}}{2} . \tag{B.8}
\end{equation*}
$$

With relation $\log \Psi_{T} \geq \log w_{i^{*}}^{(T)}$, we obtain

$$
\begin{equation*}
\eta_{p} \sum_{t}\left(\left\langle\mathbf{p}^{(t)}, \boldsymbol{f}\left(\mathbf{x}^{(t)}\right)\right\rangle-\left\langle\mathbf{e}_{i^{*},} \boldsymbol{f}\left(\mathbf{x}^{(t)}\right)\right\rangle\right)=\eta_{p} \sum_{t}\left(\left\langle\mathbf{p}^{(t)}, \boldsymbol{f}\left(\mathbf{x}^{(t)}\right)\right\rangle-\left\langle\mathbf{p}^{*}, f\left(\mathbf{x}^{(t)}\right)\right\rangle\right) \leq \log L+\frac{T \eta_{p}^{2} R_{f}^{2}}{2} \tag{B.9}
\end{equation*}
$$

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

Lemma 6. Define $\mathcal{T}_{2}^{t}=\mathcal{L}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t)}\right)-\mathcal{L}\left(\mathbf{p}^{(t)}, \mathbf{x}^{*}\right)$, where $\mathbf{x}^{*}$ is defined in Eq. (B.1), and set $\eta_{x}^{(t)}=\eta_{x} \leq$ $L_{h}^{-1}$, then we have

$$
\begin{equation*}
0 \leq \sum_{t} \mathcal{T}_{2}^{t} \leq \frac{1}{2 \eta_{x}}\left\|\mathbf{x}^{(0)}-\mathbf{x}^{*}\right\|_{2}^{2} . \tag{B.10}
\end{equation*}
$$

Proof. From the definition of $\mathbf{x}^{*}$, we can prove the non-negativeness of $\sum_{t} \mathcal{T}_{2}^{t}$. Here we focus on upper-bounding $\sum_{t} \mathcal{T}_{2}^{t}$ by separately analyzing each term $\mathcal{T}_{2}^{t}$. For the conciseness of notation, we drop the time index $t$. Define $h(\mathbf{x})$ as

$$
\begin{equation*}
h(\mathbf{x})=\sum_{i} p_{i} h_{i}(\mathbf{x})+\frac{\lambda_{1}\|\mathbf{y}-\boldsymbol{A} \boldsymbol{x}\|_{2}^{2}}{2}+\frac{\lambda_{2}\|\mathbf{x}\|_{2}^{2}}{2} . \tag{B.11}
\end{equation*}
$$

First, we rewrite the update equation as

$$
\begin{equation*}
\mathbf{x}^{(t+1)}=\mathbf{x}^{(t)}-\underbrace{\left(\mathbf{x}^{(t)}-\operatorname{prox}_{\eta_{x}\|\cdot\|_{1}}\left[\mathbf{x}^{(t)}-\eta_{w} \nabla h\left(\mathbf{x}^{(t)}\right)\right]\right)}_{\eta_{x} H\left(\mathbf{x}^{(t)}\right)}, \tag{B.12}
\end{equation*}
$$

which means that

$$
\begin{equation*}
H\left(\mathbf{x}^{(t)}\right)=\frac{\mathbf{x}^{(t)}-\operatorname{prox}_{\eta_{x}\|\cdot\|_{1}}\left(\mathbf{x}^{(t)}-\eta_{w} \nabla h\left(\mathbf{x}^{(t)}\right)\right)}{\eta_{x}} \tag{B.13}
\end{equation*}
$$

where $\operatorname{prox}_{\eta_{x}\|\cdot\|_{1}}(\mathbf{x})$ is defined as Beck (2017)

$$
\begin{equation*}
\operatorname{prox}_{\eta_{x}\|\cdot\|_{1}}(\mathbf{x})=\operatorname{argmin}_{\mathbf{z}} \eta_{x}\|\mathbf{z}\|_{1}+\frac{1}{2}\|\mathbf{x}-\mathbf{z}\|_{2}^{2} \tag{B.14}
\end{equation*}
$$

Here we need one important property of $H\left(\mathbf{x}^{(t)}\right)$, that is widely in the analysis of proximal gradient descent (direct results of Theorem 6.39 in Beck (2017)) and states

$$
\begin{equation*}
H\left(\mathbf{x}^{(t)}\right) \in \nabla h\left(\mathbf{x}^{(t)}\right)+\partial\left\|\mathbf{x}^{(t+1)}\right\|_{1} \tag{B.15}
\end{equation*}
$$

Here we consider the update relation $f\left(\mathbf{x}^{(t+1)}\right)-f(z)$ as

$$
\begin{align*}
& \mathcal{T}_{2}^{t+1}=\sum_{i} p_{i}^{(t)}\left[f_{i}\left(\mathbf{x}^{(t+1)}\right)-f_{i}\left(\mathbf{x}^{*}\right)\right]=\left\|\mathbf{x}^{(t+1)}\right\|_{1}+h\left(\mathbf{x}^{(t+1)}\right)-\left\|\mathbf{x}^{*}\right\|_{1}-h\left(\mathbf{x}^{*}\right) \\
& \stackrel{(i)}{\leq}\left\langle\partial\left\|\mathbf{x}^{(t+1)}\right\|_{1}, \mathbf{x}^{(t+1)}-\mathbf{x}^{*}\right\rangle+h\left(\mathbf{x}^{(t)}\right)+\left\langle\nabla h\left(\mathbf{x}^{(t)}\right), \mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\rangle+\frac{L_{h}}{2}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2}-h\left(\mathbf{x}^{*}\right) \\
& \stackrel{(i i)}{\leq}\left\langle\partial\left\|\mathbf{x}^{(t+1)}\right\|_{1}, \mathbf{x}^{(t+1)}-\mathbf{x}^{*}\right\rangle+\left\langle\nabla h\left(\mathbf{x}^{(t)}\right), \mathbf{x}^{(t)}-\mathbf{x}^{*}+\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\rangle+\frac{L_{h}}{2}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2} \\
& \stackrel{(i i i)}{=}\left\langle\partial\left\|\mathbf{x}^{(t+1)}\right\|_{1}+\nabla h\left(\mathbf{x}^{(t)}\right), \mathbf{x}^{(t+1)}-\mathbf{x}^{*}\right\rangle+\frac{L_{h}}{2}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2} \\
& \stackrel{(i v)}{\leq}\left\langle H\left(\mathbf{x}^{(t)}\right), \mathbf{x}^{(t+1)}-\mathbf{x}^{*}\right\rangle+\frac{L_{h}}{2}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2} \\
& \stackrel{(v)}{=} \frac{1}{\eta_{x}}\left\langle\mathbf{x}^{(t)}-\mathbf{x}^{(t+1)}, \mathbf{x}^{(t+1)}-\mathbf{x}^{*}\right\rangle+\frac{L_{h}}{2}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2} \\
& \stackrel{(v i)}{\leq} \frac{1}{\eta_{x}}\left\langle\mathbf{x}^{(t)}-\mathbf{x}^{(t+1)}, \mathbf{x}^{(t+1)}-\mathbf{x}^{*}\right\rangle+\frac{1}{2 \eta_{x}}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2} \\
& =\frac{1}{\eta_{x}}\left\langle\mathbf{x}^{(t)}-\mathbf{x}^{(t+1)}, \mathbf{x}^{(t+1)}-\mathbf{x}^{*}\right\rangle+\frac{1}{2 \eta_{x}}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{*}\right\|_{2}^{2}+\frac{1}{2 \eta_{x}}\left\|\mathbf{x}^{(t)}-\mathbf{x}^{*}\right\|_{2}^{2}+\frac{1}{\eta_{x}}\left\langle\mathbf{x}^{(t+1)}-\mathbf{x}^{*}, \mathbf{x}^{*}-\mathbf{x}^{(t)}\right\rangle \\
& =\frac{1}{2 \eta_{x}}\left\|\mathbf{x}^{(t)}-\mathbf{x}^{*}\right\|_{2}^{2}+\frac{1}{2 \eta_{x}}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{*}\right\|_{2}^{2}+\frac{1}{\eta_{x}}\left\langle\mathbf{x}^{(t+1)}-\mathbf{x}^{*}, \mathbf{x}^{*}-\mathbf{x}^{(t)}+\mathbf{x}^{(t)}-\mathbf{x}^{(t+1)}\right\rangle \\
& =\frac{1}{2 \eta_{x}}\left\|\mathbf{x}^{(t)}-\mathbf{x}^{*}\right\|_{2}^{2}-\frac{1}{2 \eta_{x}}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{*}\right\|_{2}^{2} \tag{B.16}
\end{align*}
$$

where in (i) we use $\left\|\mathbf{x}^{*}\right\|_{1} \geq\left\|\mathbf{x}^{(t+1)}\right\|_{1}+\left\langle\partial\left\|\mathbf{x}^{(t+1)}\right\|_{1}, \mathbf{x}^{*}-\mathbf{x}^{(t+1)}\right\rangle$ based on the definition of subgradients, and $h\left(\mathbf{x}^{t+1}\right) \leq h\left(\mathbf{x}^{(t)}\right)+\left\langle\nabla h\left(\mathbf{x}^{(t)}\right), \mathbf{x}^{(t+1)}\right\rangle+L_{h}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2} / 2$ from the $L_{h}$ smoothness of $h(\cdot)$, in (ii) we use $h\left(\mathbf{x}^{*}\right) \geq h\left(\mathbf{x}^{(t)}\right)+\left\langle\nabla h\left(\mathbf{x}^{(t)}\right), \mathbf{x}^{*}-\mathbf{x}^{(t)}\right\rangle$ since $h(\cdot)$ is convex, in (iii) we use $H\left(\mathbf{x}^{(t)}\right) \in \nabla h\left(\mathbf{x}^{(t)}\right)+\partial\left\|\mathbf{x}^{(t+1)}\right\|_{1}$, and in (iv) we use $\mathbf{x}^{(t+1)}=\mathbf{x}^{(t)}-\eta_{x} H\left(\mathbf{x}^{(t)}\right)$, and in (vi) we use $\eta_{x} \leq L^{-1}$.

Hence, we finishes the proof by

$$
\begin{align*}
& \sum_{t} \mathcal{T}_{2}^{t} \leq \sum_{t}\left[\frac{1}{2 \eta_{x}}\left\|\mathbf{x}^{(t-1)}-\mathbf{x}^{*}\right\|_{2}^{2}-\frac{1}{2 \eta_{x}}\left\|\mathbf{x}^{(t)}-\mathbf{x}^{*}\right\|_{2}^{2}\right]  \tag{B.17}\\
= & \frac{1}{2 \eta_{x}}\left\|\mathbf{x}^{(0)}-\mathbf{x}^{*}\right\|_{2}^{2}-\frac{1}{2 \eta_{x}}\left\|\mathbf{x}^{(T)}-\mathbf{x}^{*}\right\|_{2}^{2} \leq \frac{1}{2 \eta_{x}}\left\|\mathbf{x}^{(0)}-\mathbf{x}^{*}\right\|_{2}^{2} \stackrel{(i)}{\leq} \frac{4 R^{2}}{2 \eta_{x}},
\end{align*}
$$

where in (i) we use $\left\|\mathbf{x}^{(0)}-\mathbf{x}^{*}\right\|_{2} \leq\left\|\mathbf{x}^{*}\right\|_{2}+\left\|\mathbf{x}^{(0)}\right\| \leq 2 R$.

## Appendix C Proof of Theorem 4

Proof. First we define $h(\mathbf{x})=\sum_{i} p_{i} h_{i}(\mathbf{x})+\lambda_{1}\|\mathbf{y}-\boldsymbol{A} \boldsymbol{x}\|_{2}^{2} / 2+\lambda_{2}\|\mathbf{x}\|_{2}^{2} / 2$. Then we consider the term $\mathcal{L}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t+1)}\right)-\mathcal{L}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t)}\right)$ and have

$$
\begin{align*}
& \quad \mathcal{L}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t+1)}\right)-\mathcal{L}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t)}\right)=\left\|\mathbf{x}^{(t+1)}\right\|_{1}-\left\|\mathbf{x}^{(t)}\right\|_{1}+\sum_{i} p_{i}^{(t)}\left(h_{i}\left(\mathbf{x}^{(t+1)}\right)-h_{i}\left(\mathbf{x}^{(t)}\right)\right) \\
& \stackrel{(i)}{\leq}\left\langle\partial\left\|\mathbf{x}^{(t+1)}\right\|_{1}, \mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\rangle+\sum_{i} p_{i}^{(t)}\left[\left\langle\nabla h_{i}\left(\mathbf{x}^{(t)}\right), \mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\rangle+\frac{L_{h}}{2}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2}\right] \\
& \stackrel{(i i)}{=}\langle\underbrace{\left.\partial\left\|\mathbf{x}^{(t+1)}\right\|_{1}+\sum_{i} p_{i}^{(t)} \nabla h_{i}\left(\mathbf{x}^{(t)}\right), \mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\rangle+\frac{L_{h}}{2}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2}}_{H\left(\mathbf{x}^{(t)}\right)} \\
& \stackrel{(i i i)}{=}\left\langle\frac{\mathbf{x}^{(t)}-\mathbf{x}^{(t+1)}}{\eta_{w}^{(t)}}, \mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\rangle+\frac{L_{h}}{2}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2} \\
& = \\
& =\frac{1}{2}\left(L_{h}-\frac{2}{\eta_{w}^{(t)}}\right)\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2}  \tag{C.1}\\
& \text { (iv) } \\
& \leq \\
& \leq \\
& =\frac{L_{h}}{2}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2}
\end{align*}
$$

where in (i) we have $\left\|\mathbf{x}^{(t)}\right\|_{1} \geq\left\|\mathbf{x}^{(t+1)}\right\|_{1}+\left\langle\partial\left\|\mathbf{x}^{(t+1)}\right\|_{1}, \mathbf{x}^{(t)}-\mathbf{x}^{(t+1)}\right\rangle$ from the definition of subgradient Beck (2017), and $h_{i}\left(\mathbf{x}^{(t+1)}\right) \leq h_{i}\left(\mathbf{x}^{(t)}\right)+\left\langle\nabla h_{i}\left(\mathbf{x}^{(t)}\right), \mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\rangle+\frac{L_{h}}{2}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2}$, in (ii) we use the property $\partial\left\|\mathbf{x}^{(t+1)}\right\|_{1}+\sum_{i} p_{i}^{(t)} \nabla h_{i}\left(\mathbf{x}^{(t)}\right) \in H\left(\mathbf{x}^{(t)}\right)$, in (iii) we use $\mathbf{x}^{(t+1)}=\mathbf{x}^{(t)}-$ $\eta_{w}^{(t)} H\left(\mathbf{x}^{(t)}\right)$, and in $(i v)$ we use $\eta_{w}^{(t)} \leq L^{-1}$.

Adopting similar tricks as Qian et al. (2018), we could upper-bound $\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2}$ as

$$
\begin{align*}
& \left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2} \leq \frac{2}{L_{h}}\left[\mathcal{L}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t)}\right)-\mathcal{L}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t+1)}\right)\right] \\
= & \frac{2}{L_{h}}[\underbrace{\mathcal{L}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t)}\right)-\mathcal{L}\left(\mathbf{p}^{(t+1)}, \mathbf{x}^{(t+1)}\right)}_{\mathcal{T}_{1}^{t}}+\underbrace{\mathcal{L}\left(\mathbf{p}^{(t+1)}, \mathbf{x}^{(t+1)}\right)-\mathcal{L}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t+1)}\right)}_{\mathcal{T}_{2}^{t}}] . \tag{C.2}
\end{align*}
$$

Then we separately discuss bound $\mathcal{T}_{1}^{t}$ and $\mathcal{T}_{2}^{t}$. Since most terms of $\sum_{t} \mathcal{T}_{1}^{t}$ will be cancelled after summarization, we focus the analysis on bounding $\mathcal{T}_{2}^{t}$, which is

$$
\begin{equation*}
\mathcal{T}_{2}^{t}=\left\langle\mathbf{p}^{(t+1)}-\mathbf{p}^{(t)}, f\left(\mathbf{x}^{(t+1)}\right)\right\rangle \leq\left\|\mathbf{p}^{(t+1)}-\mathbf{p}^{(t)}\right\|_{1} \underbrace{\left\|f\left(\mathbf{x}^{(t+1)}\right)\right\|_{\infty}}_{\leq R_{f}} \tag{C.3}
\end{equation*}
$$

where $\boldsymbol{f}\left(\mathbf{x}^{(t+1)}\right)$ denotes the vector whose $i$ th element is $f_{i}\left(\mathbf{x}^{(t+1)}\right)$. Notice that we have

$$
\begin{align*}
& \left\|\mathbf{p}^{(t+1)}-\mathbf{p}^{(t)}\right\|_{1}^{2} \stackrel{(i)}{\leq} 2 D_{K L}\left(\mathbf{p}^{(t+1)} \| \mathbf{p}^{(t)}\right) \stackrel{(i i)}{\leq} 2 \eta_{p}^{(t)}\left\langle\mathbf{p}^{(t)}-\mathbf{p}^{(t+1)}, \boldsymbol{f}\left(\mathbf{x}^{(t)}\right)\right\rangle  \tag{C.4}\\
\leq & 2 \eta_{p}^{(t)}\left\|\mathbf{p}^{(t+1)}-\mathbf{p}^{(t)}\right\|_{1}\left\|\boldsymbol{f}\left(\mathbf{x}^{(t)}\right)\right\|_{\infty} \leq 2 \eta_{p}^{(t)} R_{f}\left\|\mathbf{p}^{(t+1)}-\mathbf{p}^{(t)}\right\|_{1}
\end{align*}
$$

which gives us $\left\|\mathbf{p}^{(t+1)}-\mathbf{p}^{(t)}\right\|_{1} \leq 2 \eta_{p}^{(t)} R_{f}$, where $(i)$ is because of Pinsker's inequality (Theorem 4.19 in Boucheron et al. (2013)) and (ii) is because of Lemma 7. To conclude, we have upper-bound $\mathcal{T}_{2}^{t}$ as

$$
\begin{equation*}
\mathcal{T}_{2}^{t} \leq 2 \eta_{p}^{(t)} R_{f}^{2} \tag{C.5}
\end{equation*}
$$

Then we finish the proof as

$$
\begin{align*}
& \sum_{t}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2} \leq \frac{2 \mathcal{L}\left(\mathbf{p}^{(0)}, \mathbf{x}^{(0)}\right)-2 \mathcal{L}\left(\mathbf{p}^{(T+1)}, \mathbf{x}^{(T+1)}\right)}{L_{h}}+\frac{4 R_{f}^{2} \sum_{t} \eta_{p}^{(t)}}{L_{h}}  \tag{C.6}\\
\stackrel{(i)}{\leq} & \frac{2 \mathcal{L}\left(\mathbf{p}^{(0)}, \mathbf{x}^{(0)}\right)}{L_{h}}+\frac{4 R_{f}^{2} \sum_{t} \eta_{p}^{(t)}}{L_{h}}
\end{align*}
$$

where $(i)$ is because $\mathcal{L}\left(\mathbf{p}^{(T+1)}, \mathbf{x}^{(T+1)}\right) \geq 0$.
Lemma 7. With Alg. 1, we have

$$
\begin{equation*}
D_{K L}\left(\mathbf{p}^{(t+1)} \| \mathbf{p}^{(t)}\right) \leq \eta_{p}^{(t)}\left\langle\mathbf{p}^{(t)}-\mathbf{p}^{(t+1)}, f\left(\mathbf{x}^{(t)}\right)\right\rangle \tag{C.7}
\end{equation*}
$$

where $\boldsymbol{f}\left(\mathbf{x}^{(t)}\right)$ denotes the vector whose ith element is $f_{i}\left(\mathbf{x}^{(t)}\right)$.

Proof. Here we have

$$
\begin{align*}
& D_{K L}\left(\mathbf{p}^{(t+1)} \| \mathbf{p}^{(t)}\right)=\sum_{i} p_{i}^{(t+1)} \log \left(\frac{p_{i}^{(t+1)}}{p_{i}^{(t)}}\right) \\
= & \sum_{i} p_{i}^{(t+1)} \log \frac{e^{-\eta_{p}^{(t)} f_{i}\left(x_{t}\right)}}{Z_{t}}=-\log \left(Z_{t}\right)-\eta_{p}^{(t)} \sum_{i} p_{i}^{(t+1)} f_{i}\left(\mathbf{x}^{(t)}\right)  \tag{C.8}\\
= & -\log \left(Z_{t}\right)-\eta_{p}^{(t)}\left\langle\mathbf{p}^{(t)}, f\left(\mathbf{x}^{(t)}\right)\right\rangle-\eta_{p}^{(t)}\left\langle\mathbf{p}^{(t+1)}-\mathbf{p}^{(t)}, f\left(\mathbf{x}^{(t)}\right)\right\rangle,
\end{align*}
$$

where $Z_{t} \triangleq \sum_{i} p_{i}^{(t)} e^{-\eta_{p} f_{i}\left(\mathbf{x}^{(t)}\right)}$.
Then we have

$$
\begin{align*}
& \eta_{p}^{(t)}\left\langle\mathbf{p}^{(t)}-\mathbf{p}^{(t+1)}, \boldsymbol{f}\left(\mathbf{x}^{(t)}\right)\right\rangle=D_{K L}\left(\mathbf{p}^{(t+1)} \| \mathbf{p}^{(t)}\right)+\log \left(\sum_{i} p_{i}^{(t)} e^{-\eta_{p}^{(t)} f_{i}\left(\mathbf{x}^{(t)}\right)}\right)+\eta_{p}^{(t)}\left\langle\mathbf{p}^{(t)}, f\left(\mathbf{x}_{t}\right)\right\rangle \\
& \stackrel{(i)}{\geq} D_{K L}\left(\mathbf{p}^{(t+1)} \| \mathbf{p}^{(t)}\right)+\log \left[\prod_{i} e^{-\eta_{p}^{(t)} p_{i}^{t} f_{i}\left(\mathbf{x}^{(t)}\right)}\right]+\eta_{p}^{(t)}\left\langle\mathbf{p}^{(t)}, f\left(\mathbf{x}^{(t)}\right)\right\rangle \\
&= D_{K L}\left(\mathbf{p}^{(t+1)} \| \mathbf{p}^{(t)}\right)+\underbrace{\sum_{i} \log \left(e^{-\eta_{p}^{(t)} p_{i}^{(t)} f_{i}\left(\mathbf{x}^{(t)}\right)}\right)+\eta_{p}^{(t)}\left\langle\mathbf{p}^{(t)}, f\left(\mathbf{x}_{t}\right)\right\rangle}_{0}=D_{K L}\left(\mathbf{p}^{(t+1)} \| \mathbf{p}^{(t)}\right), \tag{C.9}
\end{align*}
$$

where in (i) we use $\sum_{i} p_{i} x_{i} \geq \prod_{i} x_{i}^{p_{i}}$ such that $\sum_{i} p_{i}=1, p_{i} \geq 0$.

## Appendix D Proof of Theorem 5

Proof. Define $\mathbf{p}^{*}$ and $\mathbf{x}^{*}$ as

$$
\begin{equation*}
\mathbf{p}^{*}=\operatorname{argmin}_{\mathbf{p}} \sum_{t} \mathcal{L} \mathcal{R}\left(\mathbf{p}, \mathbf{x}^{(t)}\right), \quad \mathbf{x}^{*}=\operatorname{argmin}_{\mathbf{x}} \sum_{t} \mathcal{L} \mathcal{R}\left(\mathbf{p}^{(t)}, \mathbf{x}\right) \tag{D.1}
\end{equation*}
$$

respectively First we define $\mathcal{T}_{1}^{t}$ and $\mathcal{T}_{2}^{t}$ as

$$
\begin{align*}
& \mathcal{T}_{1}^{t}=\mathcal{L} \mathcal{R}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t)}\right)-\mathcal{L} \mathcal{R}\left(\mathbf{p}^{*}, \mathbf{x}^{(t)}\right) \\
& \mathcal{T}_{2}^{t}=\mathcal{L R}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t)}\right)-\mathcal{L} \mathcal{R}\left(\mathbf{p}^{(t)}, \mathbf{x}^{*}\right) \tag{D.2}
\end{align*}
$$

respectively. Then our goal becomes bounding $\left|\sum_{t} \mathcal{T}_{1}^{t}\right|+\left|\sum_{t} \mathcal{T}_{2}^{t}\right|$, For term $\left|\sum_{t} \mathcal{T}_{2}^{t}\right|$, the analysis stays the same as Lemma 6. Here we focus on bounding $\sum_{t} \mathcal{T}_{1}^{t}$, which proceeds as

$$
\begin{align*}
& \sum_{t} \mathcal{T}_{1}^{t}=\sum_{t}\left(\mathcal{L R}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t)}\right)-\mathcal{L R}\left(\mathbf{p}, \mathbf{x}^{(t)}\right)\right)=\sum_{t}\{-\langle\underbrace{\nabla_{t} \mathcal{L R}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t)}\right)}_{\mathbf{g}^{(t)}}, \mathbf{p}-\mathbf{p}^{(t)}\rangle-\frac{\lambda_{3}}{2}\left\|\mathbf{p}^{(t)}-\mathbf{p}\right\|_{2}^{2}\} \\
= & \sum_{t}\left\{\left\langle\mathbf{g}^{(t)}, \mathbf{p}^{(t)}-\mathbf{p}\right\rangle-\frac{\lambda_{3}}{2}\left\|\mathbf{p}^{(t)}-\mathbf{p}\right\|_{2}^{2}\right\} \tag{D.3}
\end{align*}
$$

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

$$
\begin{align*}
& \left\|\mathbf{p}^{(t+1)}-\mathbf{p}\right\|_{2}^{2}=\left\|\mathbb{P}_{\Delta}\left(\mathbf{p}^{(t)}-\eta_{p}^{t} \mathbf{s}^{(t)}\right)-\mathbf{p}\right\|_{2}^{2} \stackrel{(i)}{\leq}\left\|\mathbf{p}^{(t)}-\eta_{p}^{(t)} \mathbf{g}^{(t)}-\mathbf{p}\right\|_{2}^{2}  \tag{D.4}\\
= & \left\|\mathbf{p}^{(t)}-\mathbf{p}\right\|_{2}^{2}+\left(\eta_{p}^{(t)}\right)^{2}\left\|\mathbf{g}^{(t)}\right\|_{2}^{2}-2 \eta_{p}^{(t)}\left\langle\mathbf{g}^{(t)}, \mathbf{p}^{(t)}-\mathbf{p}\right\rangle,
\end{align*}
$$

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

$$
\begin{equation*}
\left\langle\mathbf{g}^{(t)}, \mathbf{p}^{(t)}-\mathbf{p}\right\rangle \leq \frac{\eta_{p}^{(t)}\left\|\mathbf{g}^{(t)}\right\|_{2}^{2}}{2}+\frac{\left\|\mathbf{p}^{(t)}-\mathbf{p}\right\|_{2}^{2}-\left\|\mathbf{p}^{(t+1)}-\mathbf{p}\right\|_{2}^{2}}{2 \eta_{p}^{(t)}} \tag{D.5}
\end{equation*}
$$

By setting $\eta_{p}^{(t)}=(\lambda t)^{-1}$, we have

$$
\begin{align*}
& \sum_{t} \mathcal{T}_{1}^{t} \leq \frac{R_{g}^{2} \log T}{2 \lambda_{3}}+\frac{\lambda_{3}}{2} \sum_{t} t\left(\left\|\mathbf{p}^{(t)}-\mathbf{p}\right\|_{2}^{2}-\left\|\mathbf{p}^{(t+1)}-\mathbf{p}\right\|_{2}^{2}\right)-\frac{\lambda_{3}}{2} \sum_{t}\left\|\mathbf{p}^{(t)}-\mathbf{p}\right\|_{2}^{2} \\
= & \frac{R_{g}^{2} \log T}{2 \lambda_{3}}+\frac{\lambda_{3}}{2} \sum_{t}\left\|\mathbf{p}^{(t)}-\mathbf{p}\right\|_{2}^{2}-\frac{\lambda_{3}(T+1)}{2}\left\|\mathbf{p}^{(T+1)}-\mathbf{p}\right\|_{2}^{2}-\frac{\lambda_{3}}{2} \sum_{t}\left\|\mathbf{p}^{(t)}-\mathbf{p}\right\|_{2}^{2}  \tag{D.6}\\
= & \frac{R_{g}^{2} \log T}{2 \lambda_{3}}-\frac{\lambda_{3}(T+1)}{2}\left\|\mathbf{p}^{(T+1)}-\mathbf{p}\right\|_{2}^{2} \leq \frac{R_{g}^{2} \log T}{2 \lambda_{3}} .
\end{align*}
$$

Hence, we have

$$
\begin{equation*}
\frac{\sum_{t} \mathcal{T}_{1}^{t}+\mathcal{T}_{2}^{t}}{T} \leq \frac{R_{g}^{2} \log T}{2 \lambda_{3} T}+\frac{R^{2}}{2 \eta_{x} T} \tag{D.7}
\end{equation*}
$$

which completes the proof.

## Appendix E Proof of Theorem 6

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

$$
\begin{align*}
& \left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2} \leq \frac{2}{L_{h}}\left[\mathcal{L R}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t)}\right)-\mathcal{L} \mathcal{R}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t+1)}\right)\right] \\
= & \frac{2}{L_{h}}[\underbrace{\mathcal{L R}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t)}\right)-\mathcal{L} \mathcal{R}\left(\mathbf{p}^{(t+1)}, \mathbf{x}^{(t+1)}\right)}_{\mathcal{T}_{1}^{t}}+\underbrace{\mathcal{L} \mathcal{R}\left(\mathbf{p}^{(t+1)}, \mathbf{x}^{(t+1)}\right)-\mathcal{L} \mathcal{R}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t+1)}\right)}_{\mathcal{T}_{2}^{t}}] . \tag{E.1}
\end{align*}
$$

Since $\mathcal{T}_{1}^{t}$ will cancel themselves after summarization, we focus on bounding $\mathcal{T}_{2}^{t}$. Then we have

$$
\begin{align*}
& \mathcal{L R}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t+1)}\right)=\mathcal{L} \mathcal{R}\left(\mathbf{p}^{(t+1)}, \mathbf{x}^{(t+1)}\right)+\left\langle\nabla_{\mathbf{p}} \mathcal{L \mathcal { R }}\left(\mathbf{p}^{(t+1)}, \mathbf{x}^{(t+1)}\right), \mathbf{p}^{(t)}-\mathbf{p}^{(t+1)}\right\rangle+\frac{\lambda_{3}}{2}\left\|\mathbf{p}^{(t+1)}-\mathbf{p}^{(t)}\right\|_{2}^{2} \\
& \stackrel{(i)}{=}  \tag{E.2}\\
& \mathcal{L R}\left(\mathbf{p}^{(t+1)}, \mathbf{x}^{(t+1)}\right)+\left\langle\boldsymbol{f}\left(\mathbf{x}^{(t+1)}\right)+\lambda_{3}\left(\mathbf{p}^{(t+1)}-q\right), \mathbf{p}^{(t)}-\mathbf{p}^{(t+1)}\right\rangle+\frac{\lambda_{3}}{2}\left\|\mathbf{p}^{(t+1)}-\mathbf{p}^{(t)}\right\|_{2}^{2}
\end{align*}
$$

where in (i) we have

$$
\begin{equation*}
\nabla_{\mathbf{p}} \mathcal{L} \mathcal{R}\left(\mathbf{p}^{(t+1)}, \mathbf{x}^{(t+1)}\right)=f\left(\mathbf{x}^{(t+1)}\right)+\lambda_{3}\left(\mathbf{p}^{(t+1)}-\boldsymbol{q}\right) \tag{E.3}
\end{equation*}
$$

Then we have

$$
\begin{align*}
& \mathcal{L R}\left(\mathbf{p}^{(t+1)}, \mathbf{x}^{(t+1)}\right)-\mathcal{L R}\left(\mathbf{p}^{(t)}, \mathbf{x}^{(t+1)}\right)=\left\langle\mathbf{g}^{(t+1)}, \mathbf{p}^{(t)}-\mathbf{p}^{(t+1)}\right\rangle+\frac{\lambda_{3}}{2}\left\|\mathbf{p}^{(t+1)}-\mathbf{p}^{(t)}\right\|_{2}^{2} \\
\leq & \left\|\mathbf{g}^{(t+1)}\right\|_{2}\left\|\mathbf{p}^{(t+1)}-\mathbf{p}^{(t)}\right\|_{2}+\frac{\lambda_{3}}{2}\left\|\mathbf{p}^{(t+1)}-\mathbf{p}^{(t)}\right\|_{2}^{2}  \tag{E.4}\\
\leq & \left\|\mathbf{g}^{(t+1)}\right\|_{2}\left\|\eta_{p}^{(t)} \mathbf{g}^{(t)}\right\|_{2}+\frac{\lambda_{3}\left(\eta_{p}^{(t)}\right)^{2}}{2}\left\|\mathbf{g}^{(t)}\right\|_{2}^{2} \leq R_{g}^{2}\left(\eta_{p}^{(t)}+\frac{\lambda_{3}\left(\eta_{p}^{(t)}\right)^{2}}{2}\right) .
\end{align*}
$$

Hence, we conclude that

$$
\begin{equation*}
\sum_{t}\left\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\right\|_{2}^{2} \leq \frac{2 \mathcal{L R}\left(\mathbf{p}^{(0)}, \mathbf{x}^{(0)}\right)}{L_{h}}+\frac{2 R_{g}^{2}}{L_{h}} \sum_{t}\left(\eta_{p}^{(t)}+\frac{\lambda_{3}\left(\eta_{p}^{(t)}\right)^{2}}{2}\right) \tag{E.5}
\end{equation*}
$$

where $\left\|\mathbf{g}^{(t)}\right\|_{2} \leq R_{g}$.

