Distributed Inexact Newton-type Pursuit for Non-convex Sparse Learning

Bo Liu lb507@cs.rutgers.edu
Department of Computer Science, Rutgers Univeristy
Xiao-Tong Yuan xtyuan1980@gmail.com
BDAT Lab, Nanjing University of Information Science and Technology
Qingshan Liu qsliu@nuist.edu.cn
BDAT Lab, Nanjing University of Information Science and Technology
Dimitris N. Metaxas dnm@cs.rutgers.edu
Department of Computer Science, Rutgers University

Abstract

In this paper, we present a distributed greedy pursuit method for non-convex sparse learning under cardinality constraint. Given the training samples randomly partitioned across multiple machines, the proposed method alternates between local inexact optimization of a Newton-type approximation and centralized global results aggregation. Theoretical analysis shows that for a general class of objective functions with Lipschitze continues Hessian, the method converges linearly with contraction factor scales *inversely* with data size. Numerical results confirm the high communication efficiency of our method when applied to large-scale sparse learning tasks.

1 Introduction

Setup. We are interested in distributed computing methods for solving the following cardinality-constrained empirical risk minimization (ERM) problem:

$$\min_{w\in\mathbb{R}^p} F(w) = \frac{1}{m}\sum_{j=1}^{m} F_j(w) = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n} f(w; x_{ji}, y_{ji}), \quad \text{subject to } \|w\|_0 \leq k, \tag{1}$$

where f is a convex loss function, $\|w\|_0$ represents the number of non-zero entries in w, and we assume the training data $D = \{D_1, D_2, ..., D_m\}$ with $N = nm$ samples is evenly and randomly distributed over m machines; each machine j locally stores and accesses n training sample $D_j = \{x_{ji}, y_{ji}\}_{i=1}^n$. We refer to the above model as ℓ_0-ERM in this paper. Due to the presence of cardinality constraint, the problem is non-convex and NP-hard in general.

Related work. Iterative hard thresholding (IHT) methods have demonstrated superior scalability in solving (1) [1, 2]. It is known from [2] that if $F(w)$ is L-smooth and μ_s-strongly convex over s-sparse with some sparsity level $k = O\left(\frac{L^2}{\mu_s^2}\bar{k}\right)$, IHT-style methods reach the estimation error level $\|w^{(t)} - \bar{w}\| = \mathcal{O}\left(\sqrt{\bar{k}}\|\nabla F(\bar{w})\|_\infty/\mu_s\right)$ after $\mathcal{O}\left(\frac{L}{\mu_s}\log\left(\frac{\mu_s\|w^{(0)}-\bar{w}\|}{\sqrt{\bar{k}}\|\nabla F(\bar{w})\|_\infty}\right)\right)$ rounds of iteration. A distributed implementation of IHT was considered in [4] for compressive sensing. However, the linear dependence of the iteration complexity on the restricted condition number L/μ_s makes it inefficient in ill-conditioned settings.

Significant interest has recently been dedicated to designing distributed algorithms that have flexibility to adapt to the communication-computation tradeoffs [3, 7, 6]. There is a recent surge of developing Newton-type algorithm for distributed model learning [6, 5]. It was proved in [6] that if $F(w)$ is quadratic with condition number L/μ, the communication complexity (in high probability) to reach ϵ-precision is $\mathcal{O}\left(\frac{L^2}{\mu^2 n}\log(mp)\log\left(\frac{1}{\epsilon}\right)\right)$, which has an improved dependence on the condition number L/μ which could scale as large as $\mathcal{O}(\sqrt{mn})$ in regularized learning problems.

Open problem. Despite the attractiveness of distributed approximate/inexact Newton-type methods in classical regularized ERM learning, it still remains unclear whether this type of methods generalize equally well, both in theory and practice, to the non-convex ℓ_0-ERM model as defined in (1).

OPTML 2017: 10th NIPS Workshop on *Optimization for Machine Learning* (NIPS 2017).

Our contribution. In this paper, we give an affirmative answer to the above open question by developing a novel DANE-type distributed greedy pursuit method. We show for our method that the parameter estimation error bound $\|w^{(t)} - w^{(0)}\| = \mathcal{O}\left(\sqrt{k}\|\nabla F(\bar{w})\|_\infty/\mu_s\right)$ can be guaranteed in high probability after $\mathcal{O}\left(\frac{1}{1-\frac{L}{\mu_s}\sqrt{\frac{\log(mp)}{n}}}\log\left(\frac{\mu_s\|w^{(0)}-\bar{w}\|}{\sqrt{k}\|\nabla F(\bar{w})\|_\infty}\right)\right)$ rounds of communication. Comparing to the bound of distributed IHT [4], this bound has much improved dependence on restricted condition number when data size is sufficiently large. In sharp contrast, the required sample complexity in [7] for ℓ_1-regularized distributed learning is $n = \mathcal{O}\left(\frac{s^2L^2\log p}{\mu_s^2}\right)$ which is clearly inferior to ours.

Notation and definitions. We denote $\mathrm{H}_k(x)$ as a truncation operator which preserves the top k (in magnitude) entries of vector x and forces the remaining to be zero. The notation $\mathrm{supp}(x)$ represents the index set of nonzero entries of x. We conventionally define $\|x\|_\infty = \max_i |[x]_i|$ and define $x_{\min} = \min_{i\in\mathrm{supp}(x)} |[x]_i|$. For an index set S, we define $[x]_S$ and $[A]_{SS}$ as the restriction of x to S and the restriction of rows and columns of A to S, respectively. For an integer n, we abbreviate the set $\{1, ..., n\}$ to $[n]$. For any integer $s > 0$, we say $f(w)$ is restricted μ_s-strongly-convex and L_s-smooth if $\forall w, w'$ with $\|w - w'\|_0 \leq s$, $\frac{\mu_s}{2}\|w - w'\|^2 \leq f(w) - f(w') - \langle\nabla f(w'), w - w'\rangle \leq \frac{L_s}{2}\|w - w'\|^2$. Suppose that $f(w)$ is twice continuously differentiable. We say $f(w)$ has Restricted Lipschitz Hessian with constant $\beta_s \geq 0$ (or β_s-RLH) if $\forall w, w'$ with $\|w-w'\|_0 \leq s$, $\left\|\nabla^2_{SS}f(w) - \nabla^2_{SS}f(w')\right\| \leq \beta_s\|w - w'\|$, where we have used the abbreviation $\nabla^2_{SS}f := [\nabla^2 f]_{SS}$.

2 Distributed Inexact Newton-type Pursuit

2.1 Algorithm

The Distributed Inexact Newton-type PurSuit (DINPS) algorithm is outlined in Algorithm 1. Starting from an initial k-sparse approximation $w^{(0)}$, the procedure generates a sequence of intermediate k-sparse iterate $\{w^{(t)}\}_{t\geq 1}$ via distributed local sparse estimation and global synchronization among machines. More precisely, each iteration loop of DINPS can be decomposed into the following three consequent main steps:

Map-reduce gradient computation. In this first step, we evaluate the global gradient $\nabla F(w^{(t-1)}) = \frac{1}{m}\sum_{j=1}^m \nabla F_j(w^{(t-1)})$ at the current iterate via simple map-reduce averaging and distribute it to all machines for local computation.

Local inexact sparse approximation. In this step, each machine j constructs at the current iterate a local objective function as in (2), and then inexactly estimate a local k-sparse solution satisfying (3). This inexact sparse optimization step can be implemented using IHT-style algorithms which have been witnessed to offer fast and accurate solutions for centralized ℓ_0-estimation [8].

Centralized results aggregation. The master machine simply assigns to $w^{(t)}$ the first received k-sparse local output $w_{j_t}^{(t)}$ at the current round of iteration. Such an aggregation scheme is by nature asynchronous and hence robust to computation-power imbalance and communication delay.

The simplest way of initialization is to set $w^{(0)} = 0$, i.e., starting the iteration from scratch. Since the data samples are assumed to be evenly and randomly distributed on machines, another reasonable option of initialization is to minimize one of the local ℓ_0-ERM problems, say $w^{(0)} \approx \arg\min_{\|w\|_0\leq k} F_1(w)$, which is expected to be reasonably close to the global solution. A similar local initialization strategy was also considered for EDSL [7].

2.2 Main results

Deterministic result. The following is a deterministic result on the parameter estimation error bound of DINPS when the objective functions is twice differentiable with restricted Lipschitz Hessian.

Theorem 1. *Let $\bar{w}$ be a $\bar{k}$-sparse target vector with $\bar{k} \leq k$. Assume that each component $F_j(w)$ is μ_{3k}-strongly-convex and has β_{3k}-RLH. Let $\bar{H}_j = \nabla^2 F_j(\bar{w})$ and $\bar{H} = \frac{1}{m}\sum_{j=1}^m \bar{H}_j$. Assume that $\max_j \|\bar{H}_j - \eta\bar{H}\| \leq \theta\mu_{3k}/4$ for some $\theta \in (0, 1)$ and $\epsilon \leq \frac{k\eta^2\|\nabla F(\bar{w})\|_\infty^2}{2\mu_{3k}}$. Assume that $\|w^{(0)} - \bar{w}\| \leq \frac{\theta}{(1+\eta)\beta_{3k}}$ and $\|\nabla F(\bar{w})\|_\infty \leq \frac{\theta\mu_{3k}}{8.94\eta(1+\eta)\beta_{3k}\sqrt{k}}$. Set $\gamma = 0$, then $\|w^{(t)} - \bar{w}\| \leq \frac{5.47\eta\sqrt{k}\|\nabla F(\bar{w})\|_\infty}{(1-\theta)\mu_{3k}}$ after*

Algorithm 1: Distributed Inexact Newton-type PurSuit (DINPS)

Input : Loss functions $\{F_j(w)\}_{j=1}^m$ distributed over m different machines, sparsity level k, parameter $\gamma \geq 0$ and $\eta > 0$. Typically set $\gamma = 0$ and $\eta = 1$.

Initialization *Set* $w^{(0)} = 0$ *or estimate* $w^{(0)} \approx \arg\min_{\|w\|_0 \leq k} F_1(w)$.

for $t = 1, 2, ...$ **do**

/* **Map-reduce gradient evaluation** */

(**S1**) Compute $\nabla F(w^{(t-1)}) = \frac{1}{m}\sum_{j=1}^m \nabla F_j(w^{(t-1)})$ and broadcast it to all workers;

/* **Local sparse optimization:** */

(**S2**) **for** *all the workers* $j = 1, ..., m$ *in parallel* **do**

(i) Construct a local objective function:

$$P_j(w; w^{(t-1)} \mid \eta, \gamma) := \langle \eta \nabla F(w^{(t-1)}) - \nabla F_j(w^{(t-1)}), w \rangle + \frac{\gamma}{2}\|w - w^{(t-1)}\|^2 + F_j(w); \quad (2)$$

(ii) Estimate a k-sparse $w_j^{(t)}$ via approximately solving $\min_{\|w\|_0 \leq k} P_j(w; w^{(t-1)}, \eta, \gamma)$ up to sparsity level $\bar{k} \leq k$ and ϵ-precision, i.e., for any $\bar{k}$-sparse vector $\bar{w}$:

$$P_j(w_j^{(t)}; w^{(t-1)} \mid \eta, \gamma) \leq P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma) + \epsilon; \quad (3)$$

end

/* **Centralized results aggregation:** */

(**S3**) Set $w^{(t)} = w_{j_t}^{(t)}$ as the first received local solution;

end

Output : $w^{(t)}$.

$$t \geq \frac{1}{1-\theta} \log\left(\frac{\mu_{3k}\|w^{(0)} - \bar{w}\|}{\eta\sqrt{k}\|\nabla F(\bar{w})\|_\infty}\right)$$

rounds of iteration.

Proof sketch. The proof is rooted on the following key inequality:

$$\|w^{(t)} - \bar{w}\| \leq \frac{2\max_j \|\bar{H}_j - \eta\bar{H}\|}{\mu_{3k}}\|w^{(t-1)} - \bar{w}\| + \frac{(1+\eta)\beta_{3k}}{2}\|w^{(t-1)} - \bar{w}\|^2 + \frac{4.47\eta\sqrt{k}}{\mu_{3k}}\|\nabla F(\bar{w})\|_\infty.$$

Given this inequality, we can prove by induction that $\|w^{(t)} - \bar{w}\| \leq \frac{\theta}{(1+\eta)\beta_{3k}}$ holds for all $t \geq 0$, which then implies the recursive form $\|w^{(t)} - \bar{w}\| \leq \theta\|w^{(t-1)} - \bar{w}\| + \frac{4.47\eta\sqrt{k}}{\mu_{3k}}\|\nabla F(\bar{w})\|_\infty$.

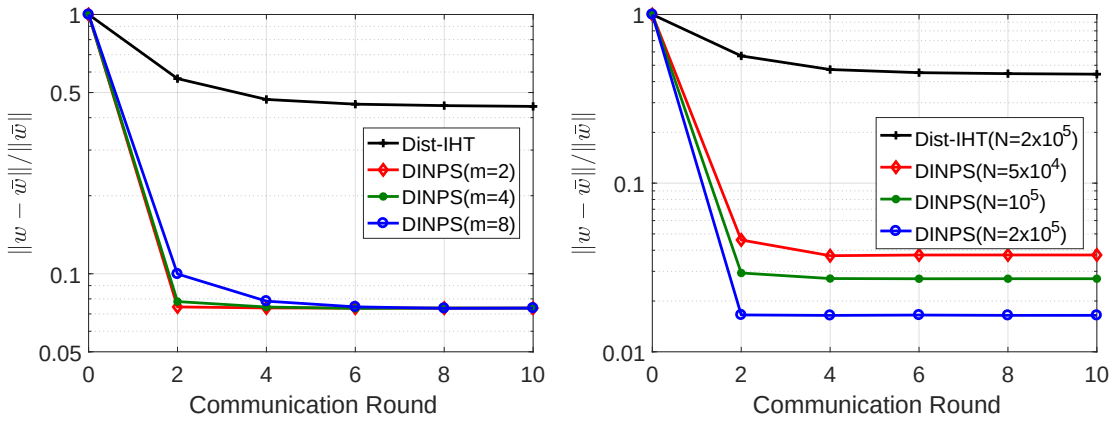
Remark. The main message conveyed by Theorem 1 is: Provided that $w^{(0)}$ is properly initialized and the estimation error level $\sqrt{k}\|\nabla F(\bar{w})\|_\infty$ is sufficiently small, DINPS for RLH objectives exhibits linear convergence behavior with contraction factor $\theta = \mathcal{O}(\max_j \|\bar{H}_j - \eta\bar{H}\|/\mu_{3k})$.

Stochastic result. We now turn to a stochastic setting where the samples are uniformly randomly distributed over m machines.

Corollary 1. *Let* $\bar{w}$ *be a* $\bar{k}$*-sparse target vector with* $\bar{k} \leq k$*. Assume that the samples are uniformly randomly distributed on* m *machines and each component* $F_j(w)$ *is* μ_{3k}*-strongly-convex and has* β_{3k}*-RLH. Assume* $\|\nabla^2 f(w^\top x_{ji}, y_{ji})\| \leq L$ *holds for all* $j \in [m]$ *and* $i \in [n]$*. Let* $\bar{H}_j = \nabla^2 F_j(\bar{w})$ *and* $\bar{H} = \frac{1}{m}\sum_{j=1}^m \bar{H}_j$*. Set* $\gamma = 0$ *and* $\eta = 1$*. Assume that* $\epsilon \leq \frac{k\|\nabla F(\bar{w})\|_\infty^2}{2\mu_{3k}}$*. Assume that* $\|w^{(0)} - \bar{w}\| \leq \frac{\theta}{2\beta_{3k}}$ *and* $\|\nabla F(\bar{w})\|_\infty \leq \frac{\theta\mu_{3k}}{17.88\eta\beta_{3k}\sqrt{k}}$*. For any* $\delta \in (0,1)$*, if* $n > \frac{512L^2\log(mp/\delta)}{\mu_{3k}^2}$*, then with probability at least* $1-\delta$*, Algorithm 1 will output solution* $w^{(t)}$ *satisfying* $\|w^{(t)} - \bar{w}\| \leq \frac{5.47\sqrt{k}\|\nabla F(\bar{w})\|_\infty}{(1-\theta)\mu_{3k}}$ *after*

$$t \geq \frac{1}{1-\theta} \log\left(\frac{\mu_{3k}\|w^{(0)} - \bar{w}\|}{\sqrt{k}\|\nabla F(\bar{w})\|_\infty}\right)$$

rounds of iteration with $\theta = \frac{L}{\mu_{3k}}\sqrt{\frac{512\log(mp/\delta)}{n}}$.



(a) F_{lr} with different m values. $N = 10^4$. (b) F_{lr} with different N values. $m = 10$.

Figure 1: Linear regression (denoted by F_{lr}) model training convergence, given varied number of machines (m) and training samples (N). The convergence of F_{lr} model training is evaluated by $\|w - \bar{w}\|/\|\bar{w}\|$.

Proof sketch. Based on [6, Lemma 2] we can show that $\max_j \|H_j - H\| \leq \theta\mu_{3k}/4$ holds with probability at least $1 - \delta$. The condition on n guarantees $\theta \in (0, 1)$.

Remark. Corollary 1 indicates that in the considered statistical setting, the contraction factor θ can be arbitrarily small given that the sample size $n = \mathcal{O}\left(\frac{L^2 \log(mp)}{\mu_{3k}^2}\right)$ is sufficiently large. This sample size complexity is clearly superior to the corresponding $n = \mathcal{O}\left(\frac{k^2 L^2 \log p}{\mu_{3k}^2}\right)$ complexity established in [7] for distributed ℓ_1-regularized sparse learning.

3 Experiment

We evaluate the empirical performance of DINPS on a set of simulated sparse linear regression tasks. A synthetic $N \times p$ design matrix is generated with each data sample x_i drawn from Gaussian distribution $\mathcal{N}(0, \Sigma)$ with $\Sigma_{j,k} = \begin{cases} 1 & \text{if } j = k \\ 1.5^{-\frac{|j-k|}{10}} & \text{otherwise} \end{cases}$. A sparse model parameter $\bar{w} \in \mathbb{R}^p$ is generated with the top $\bar{k}$ entries uniformly randomly valued in interval $(0, 0.1)$ and all the other entries set to be zero. The response variables $\{y_i\}_{i=1}^N$ are generated by $y_i = \bar{w}^\top x_i + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, 0.01)$. A distributed implementation of IHT (which we refer to as Dist-IHT) [4] is considered as a baseline algorithm for communication efficiency comparison. We set $p = 1000$. k value and parameter update step-sizes are chosen by grid search for optimal performance.

We first fix the training sample number to be $N = 10^4$ and vary the number of machines to be $m = 2, 4, 8$. The convergence curves of the considered algorithms in terms of round of communication are shown in Figure 1(a). Since Dist-IHT collects the gradient update from all workers before conducting gradient descent and hard-thresholding on master machine, its convergence behavior is irrelevant to the number of machines. In our DINPS algorithm, we adopt a more greedy parameter update strategy by solving a ℓ_0-constrained minimization problem on worker machines. As a result, it requires much fewer number of communication rounds between worker and master than Dist-IHT. We have also tested with the case when the number of machines is fixed as $m = 10$, and the number of training sample is varying to be $N = 5 \times 10^4$, 10^5 and 2×10^5. Figure 1(b) shows the convergence curves of the considered algorithms. Again we observe the superior communication efficiency of DINPS over Dist-IHT.

4 Conclusion

As a novel inexact Newton-type greedy pursuit method for distributed ℓ_0-ERM, DINPS has improved dependence on the restricted condition number than the prior distributed IHT methods.

References

[1] A. Beck and Y. C. Eldar. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, 2013.

[2] P. Jain, A. Tewari, and P. Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems*, 2014.

[3] M. I. Jordan, J. D. Lee, and Y. Yang. Communication-efficient distributed statistical inference. *arXiv preprint arXiv:1605.07689*, 2016.

[4] S. Patterson, Y. C. Eldar, and I. Keidar. Distributed compressed sensing for static and time-varying networks. *IEEE Transactions on Signal Processing*, 62(19):4931–4946, 2014.

[5] S. J. Reddi, J. Konečnỳ, P. Richtárik, B. Póczós, and A. Smola. Aide: Fast and communication efficient distributed optimization. *arXiv preprint arXiv:1608.06879*, 2016.

[6] O. Shamir, N. Srebro, and T. Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International Conference on Machine Learning*, 2014.

[7] J. Wang, M. Kolar, N. Srebro, and T. Zhang. Efficient distributed learning with sparsity. *International Conference on Machine Learning*, 2017.

[8] X.-T. Yuan, P. Li, and T. Zhang. Gradient hard thresholding pursuit for sparsity-constrained optimization. In *International Conference on Machine Learning*, 2014.

Abstract

This supplementary document contains the technical proofs of algorithm analysis in 10th NIPS Workshop on Optimization for Machine Learning submission entitled "Distributed Inexact Newton-type Pursuit for Non-convex Sparse Learning".

A A key lemma

The following lemma is key to our analysis.

Lemma 1. *Let $\bar{w}$ be a $\bar{k}$-sparse target vector with $\bar{k} \leq k$. Assume that each component $F_j(w)$ is μ_{3k}-strongly-convex and $\eta F(w) - F_j(w) - \frac{\gamma}{2}\|w\|^2$ has α_{3k}-RLG.*

$$\|w^{(t)} - \bar{w}\| \leq \frac{2\alpha_{3k}}{\gamma + \mu_{3k}}\|w^{(t-1)} - \bar{w}\| + \frac{3.47\eta\sqrt{k}}{\gamma + \mu_{3k}}\|\nabla F(\bar{w})\|_\infty + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}}.$$

Moreover, assume that each $F_j(w)$ has β_{3k}-RLH. Let $\bar{H}_j = \nabla^2 F_j(\bar{w})$ and $\bar{H} = \frac{1}{m}\sum_{j=1}^m \bar{H}_j$. Then

$$\begin{aligned}\|w^{(t)} - \bar{w}\| \leq & \frac{2(\gamma + \max_j \|\bar{H}_j - \eta\bar{H}\|)}{\gamma + \mu_{3k}}\|w^{(t-1)} - \bar{w}\| + \frac{(1+\eta)\beta_{3k}}{2}\|w^{(t-1)} - \bar{w}\|^2 + \frac{3.47\eta\sqrt{k}}{\gamma + \mu_{3k}}\|\nabla F(\bar{w})\|_\infty \\ & + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}}.\end{aligned}$$

Proof. Recall that $w^{(t)} = w^{(t)}_{j_t}$. By assumption we know that $F_{j_t}(w)$ is μ_{3k}-strongly-convex. Obviously, $P_{j_t}(w; w^{(t-1)} \mid \eta, \gamma)$ is $(\gamma + \mu_{3k})$-strongly-convex. Let $S^{(t)} = \text{supp}(w^{(t)})$ and $\bar{S} = \text{supp}(\bar{w})$. Consider $S = S^{(t)} \cup S^{(t-1)} \cup \bar{S}$. Then

$$\begin{aligned}& P_{j_t}(w^{(t)}; w^{(t-1)} \mid \eta, \gamma) \\ \geq & P_{j_t}(\bar{w}; w^{(t-1)} \mid \eta, \gamma) + \langle \nabla P(\bar{w}; w^{(t-1)} \mid \eta, \gamma), w^{(t)} - \bar{w}\rangle + \frac{\gamma + \mu_{3k}}{2}\|w^{(t)} - \bar{w}\|^2 \\ = & P_{j_t}(\bar{w}; w^{(t-1)} \mid \eta, \gamma) + \langle \nabla_S P(\bar{w}; w^{(t-1)} \mid \eta, \gamma), w^{(t)} - \bar{w}\rangle + \frac{\gamma + \mu_{3k}}{2}\|w^{(t)} - \bar{w}\|^2 \\ \overset{\xi_1}{\geq} & P_{j_t}(w^{(t)}; w^{(t-1)} \mid \eta, \gamma) - \epsilon - \|\nabla_S P(\bar{w}; w^{(t-1)} \mid \eta, \gamma)\|\|w^{(t)} - \bar{w}\| + \frac{\gamma + \mu_{3k}}{2}\|w^{(t)} - \bar{w}\|^2\end{aligned},$$

where "ξ_1" follows from the definition of $w^{(t)}$ as an approximate k-sparse minimizer of $P_{j_t}(w; w^{(t-1)} \mid \eta, \gamma)$ up to precision ϵ. By rearranging both sides of the above inequality with proper elementary calculation we get

$$\begin{aligned}& \|w^{(t)} - \bar{w}\| \\ \leq & \frac{2}{\gamma + \mu_{3k}}\|\nabla_S P_{j_t}(\bar{w}; w^{(t-1)} \mid \eta, \gamma)\| + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\ = & \frac{2}{\gamma + \mu_{3k}}\|\eta\nabla_S F(w^{(t-1)}) - \nabla_S F_{j_t}(w^{(t-1)}) + \gamma(\bar{w} - w^{(t-1)}) + \nabla_S F_{j_t}(\bar{w})\| + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\ = & \frac{2}{\gamma + \mu_{3k}}\|\eta\nabla_S F(w^{(t-1)}) - \eta\nabla_S F(\bar{w}) - (\nabla_S F_{j_t}(w^{(t-1)}) - \nabla_S F_{j_t}(\bar{w})) + \gamma(\bar{w} - w^{(t-1)}) + \eta\nabla_S F(\bar{w})\| \\ & + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\ \leq & \frac{2}{\gamma + \mu_{3k}}\left\|\left(\eta\nabla_S F(w^{(t-1)}) - \nabla_S F_{j_t}(w^{(t-1)}) - \gamma w^{(t-1)}\right) - (\eta\nabla_S F(\bar{w}) - \nabla_S F_{j_t}(\bar{w}) - \gamma\bar{w})\right\| \\ & + \frac{2\eta}{\gamma + \mu_{3k}}\|\nabla_S F(\bar{w})\| + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\ \overset{\zeta_1}{\leq} & \frac{2\alpha_{3k}}{\gamma + \mu_{3k}}\|w^{(t-1)} - \bar{w}\| + \frac{2\eta\sqrt{3k}}{\gamma + \mu_{3k}}\|\nabla F(\bar{w})\|_\infty + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}},\end{aligned}$$

where ζ_1 is according to the assumption that $\eta F(w) - F_{j_t}(w)$ has α_{3k}-RLG. This shows the validity of the first part.

Next we prove the second part. Similar to the above argument, we have

$$\begin{aligned}
\|w^{(t)} - \bar{w}\| \leq & \frac{2}{\gamma + \mu_{3k}} \|\nabla_S P_{j_t}(\bar{w}; w^{(t-1)} \mid \eta, \gamma)\| + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
\leq & \frac{2}{\gamma + \mu_{3k}} \left\| \gamma(w^{(t-1)} - \bar{w}) + \eta \nabla_S F(w^{(t-1)}) - \eta \nabla_S F(\bar{w}) - (\nabla_S F_{j_t}(w^{(t-1)}) - \nabla_S F_{j_t}(\bar{w})) \right\| \\
& + \frac{2\eta}{\gamma + \mu_{3k}} \|\nabla_S F(\bar{w})\| + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
\leq & \frac{2}{\gamma + \mu_{3k}} \left\| \gamma(w^{(t-1)} - \bar{w}) + \eta \nabla^2_{SS} F(\bar{w})(w^{(t-1)} - \bar{w}) - \nabla^2_{SS} F_{j_t}(\bar{w})(w^{(t-1)} - \bar{w}) \right\| \\
& + \frac{2\eta}{\gamma + \mu_{3k}} \|\nabla_S F(\bar{w})\| + \eta \left\| \nabla_S F(w^{(t-1)}) - \nabla_S F(\bar{w}) - \nabla^2_{SS} F(\bar{w})(w^{(t-1)} - \bar{w}) \right\| \\
& + \left\| \nabla_S F_{j_t}(w^{(t-1)}) - \nabla_S F_{j_t}(\bar{w}) - \nabla^2_{SS} F_{j_t}(\bar{w})(w^{(t-1)} - \bar{w}) \right\| + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
\leq & \frac{2}{\gamma + \mu_{3k}} \left(\gamma + \|\eta \nabla^2_{SS} F(\bar{w}) - \nabla^2_{SS} F_{j_t}(\bar{w})\| \right) \|w^{(t-1)} - \bar{w}\| + \frac{2\eta}{\gamma + \mu_{3k}} \|\nabla_S F(\bar{w})\| \\
& + \frac{(1+\eta)\beta_{3k}}{2} \|w^{(t-1)} - \bar{w}\|^2 + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
\leq & \frac{2(\gamma + \max_j \|\bar{H}_j - \eta \bar{H}\|)}{\gamma + \mu_{3k}} \|w^{(t-1)} - \bar{w}\| + \frac{(1+\eta)\beta_{3k}}{2} \|w^{(t-1)} - \bar{w}\|^2 \\
& + \frac{2\eta\sqrt{3k}}{\gamma + \mu_{3k}} \|\nabla F(\bar{w})\|_\infty + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}}.
\end{aligned}$$

This proves the second part. □

B Proof of Theorem 1

Proof. We first claim that $\|w^{(t)} - \bar{w}\| \leq \frac{\theta}{(1+\eta)\beta_{3k}}$ holds for all $t \geq 0$. This can be shown by induction. Obviously the claim holds for $t = 0$. Now suppose that $\|w^{(t-1)} - \bar{w}\| \leq \frac{\theta}{(1+\eta)\beta_{3k}}$ for some $t \geq 1$. Since $\gamma = 0$, according to Lemma 1 we have

$$\begin{aligned}
\|w^{(t)} - \bar{w}\| \leq & \frac{2 \max_j \|\bar{H}_j - \eta \bar{H}\|}{\mu_{3k}} \|w^{(t-1)} - \bar{w}\| + \frac{(1+\eta)\beta_{3k}}{2} \|w^{(t-1)} - \bar{w}\|^2 + \frac{3.47\eta\sqrt{k}}{\gamma + \mu_{3k}} \|\nabla F(\bar{w})\|_\infty \\
& + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
\overset{\zeta_1}{\leq} & \frac{\theta}{2} \|w^{(t-1)} - \bar{w}\| + \frac{(1+\eta)\beta_{3k}}{2} \|w^{(t-1)} - \bar{w}\|^2 + \frac{4.47\eta\sqrt{k}}{\mu_{3k}} \|\nabla F(\bar{w})\|_\infty \\
\leq & \theta \|w^{(t-1)} - \bar{w}\| + \frac{4.47\eta\sqrt{k}}{\mu_{3k}} \|\nabla F(\bar{w})\|_\infty \\
\overset{\zeta_2}{\leq} & \frac{\theta}{2(1+\eta)\beta_{3k}} + \frac{\theta}{2(1+\eta)\beta_{3k}} = \frac{\theta}{(1+\eta)\beta_{3k}},
\end{aligned}$$

where "ζ_1" follows from $\gamma = 0$ and the assumption on ϵ, and "ζ_2" follows from $\theta \leq 0.5$ and the condition of $\|\nabla F(\bar{w})\|_\infty \leq \frac{\theta \mu_{3k}}{8.94\eta(1+\eta)\beta_{3k}\sqrt{k}}$. Thus by induction $\|w^{(t)} - \bar{w}\| \leq \frac{\theta}{(1+\eta)\beta_{3k}}$ holds for all $t \geq 1$. Then it follows from the inequality "ζ_1" of the above we can see that for all $t \geq 0$,

$$\|w^{(t)} - \bar{w}\| \leq \theta \|w^{(t-1)} - \bar{w}\| + \frac{4.47\eta\sqrt{k}}{\mu_{3k}} \|\nabla F(\bar{w})\|_\infty.$$

By recursively applying the above inequality we get

$$\|w^{(t)} - \bar{w}\| \leq \theta^t \|w^{(0)} - \bar{w}\| + \frac{4.47\eta\sqrt{k}}{(1-\theta)(\mu_{3k})} \|\nabla F(\bar{w})\|_\infty$$

Based on the inequality $1 - x \leq \exp(-x)$ we need

$$t \geq \frac{1}{1-\theta} \log \left(\frac{(1-\theta)\mu_{3k} \|w^{(0)} - \bar{w}\|}{\eta\sqrt{k}\|\nabla F(\bar{w})\|_\infty} \right)$$

steps of iteration to achieve the precision of $\|w^{(t)} - \bar{w}\| \leq \frac{5.47\eta\sqrt{k}\|\nabla F(\bar{w})\|_\infty}{(1-\theta)\mu_{3k}}$. This proves the desired complexity bound. □

C Proof of Corollary 1

Lemma 2 which is based on a matrix concentration bound, indicates that the Hessian H_j is close to H when sample size is sufficiently large. The same result can be found in [6, Lemma 2].

Lemma 2. *Assume that $\|\nabla^2 f(w^\top x_{ji}, y_{ji})\| \leq L$ holds for all $j \in [m]$ and $i \in [n]$. Then for each j, with probability at least $1 - \delta$ over the samples,*

$$\max_j \|H_j - H\| \leq \sqrt{\frac{32L^2 \log(mp/\delta)}{n}},$$

where $H_j = \frac{1}{n}\sum_{i=1}^n \nabla^2 f(w^\top x_{ji}, y_{ji})$ and $H = \frac{1}{m}\sum_{j=1}^m H_j$.

From Lemma 2 we get that $\max_j \|H_j - H\| \leq \theta\mu_{3k}/4$ holds with probability at least $1 - \delta$. Since $n > \frac{512L^2 \log(mp/\delta)}{\mu_{3k}^2}$, we have $\theta \in (0, 1)$. By invoking Theorem 1 we get the desired result.