Journal of Machine Learning Research 21 (2020) 1-51 Submitted 9/19; Revised 2/20; Published 8/20

On Convergence of Distributed Approximate Newton Methods: Globalization, Sharper Bounds and Beyond

Xiao-Tong Yuan XTYUAN1980@GMAIL.COM
Cognitive Computing Lab
Baidu Research
Beijing 100193, China

Ping Li PINGLI98@GMAIL.COM
Cognitive Computing Lab
Baidu Research
Bellevue, WA 98004, USA

Editor: Sathiya Keerthi

Abstract

The DANE algorithm is an approximate Newton method popularly used for communication-efficient distributed machine learning. Reasons for the interest in DANE include scalability and efficiency. Convergence of DANE, however, can be tricky; its appealing convergence rate is only rigorous for quadratic objective function, and for more general convex functions the known results are no stronger than those of the classic first-order methods. To remedy these drawbacks, we propose in this article some new alternatives of DANE which are more suitable for analysis. We first introduce a simple variant of DANE equipped with backtracking line search, for which global asymptotic convergence and sharper local non-asymptotic convergence guarantees can be proved for both quadratic and non-quadratic strongly convex functions. Then we propose a heavy-ball method to accelerate the convergence of DANE, showing that the near-tight local rate of convergence can be established for strongly convex functions, and with proper modification of the algorithm about the same result applies globally to linear prediction models. Numerical evidence is provided to confirm the theoretical and practical advantages of our methods.

Keywords: Communication-efficient distributed learning, Approximate Newton method, Global convergence, Heavy-Ball acceleration.

1. Introduction

Distributed learning is a promising tool for alleviating the pressure of ever-increasing data and/or model scale in modern machine learning systems. In this article, we study the distributed optimization algorithms for solving the following empirical risk minimization (ERM) problem:

$$\min_{w\in\mathbb{R}^p} F(w) := \frac{1}{N}\sum_{i=1}^{N} f(w; x_i, y_i), \tag{1}$$

where $\{x_i, y_i\}_{i=1}^N$ is a training sample of size N, $f(w; x_i, y_i)$ is a loss function evaluated at the data point (x_i, y_i) which is smooth and convex in w. Such a finite-sum formulation

encapsulates a large body of statistical learning problems including least square regression, logistic regression and support vector machines, to name a few. We assume without loss of generality that the training data $\mathcal{D} = \{D_1, ..., D_m\}$ with $N = mn$ samples is evenly and randomly distributed over m different machines; each machine j locally stores and accesses n training samples $D_j = \{x_{ji}, y_{ji}\}_{i=1}^n$. Let us denote $F_j(w) := \frac{1}{n}\sum_{i=1}^n f(w; x_{ji}, y_{ji})$ the local empirical risk evaluated on D_j. The global objective is then to minimize the average of these local empirical risk functions

$$\min_{w\in\mathbb{R}^p} F(w) = \frac{1}{m}\sum_{j=1}^m F_j(w). \tag{2}$$

Recently, significant interest has been dedicated to designing distributed algorithms and systems that have flexibility to adapt to the communication-computation tradeoffs, e.g., for parameter estimation (Jaggi et al., 2014; Shamir et al., 2014) and statistical inference (Wang et al., 2017a; Jordan et al., 2018). A common spirit of these communication-efficient methods is trying to quickly optimize the objective value (or estimation accuracy) to certain precision using a minimal number of inter-machine communication rounds.

In this article we revisit the Distributed Approximate NEwton (DANE) algorithm proposed by Shamir et al. (2014) for solving (2), which is now one of the most popular second-order methods for communication-efficient distributed machine learning. We analyze its convergence behavior, expose problems and issues, and propose alternative algorithms more suitable for the task. We contribute to derive some new results, insights and algorithms, using a unified and more elementary framework of Lyapunov analysis.

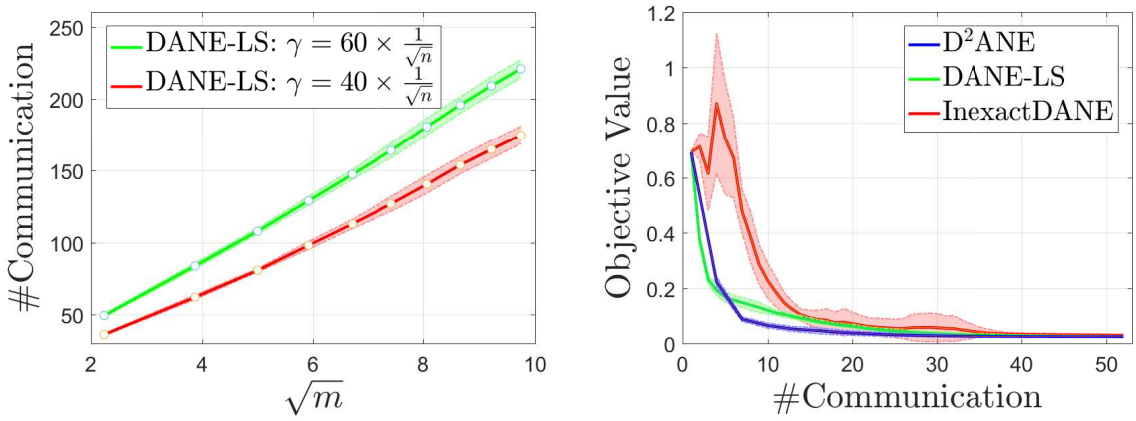
1.1 Review of the DANE Method

For the distributed ERM problem (2), the iteration (communication) complexity of first-order distributed approaches including (accelerated) gradient descent and ADMM (alternating direction method of multipliers) (Boyd et al., 2011) tend to suffer from the unsatisfactory polynomial dependence on condition number. To tackle this problem, Shamir et al. (2014) proposed the DANE method that takes advantage of the stochastic nature of problem: the i.i.d. data samples $\{x_i, y_i\}$ are uniformly distributed and each local subproblem should be close to the global problem when data size becomes sufficiently large. At the t-th iteration loop of DANE, in parallel each individual *worker* machine j optimizes a local subproblem $w_j^{(t)} = \arg\min_w P_j^{(t-1)}(w)$ in which

$$P_j^{(t-1)}(w) := \langle \eta\nabla F(w^{(t-1)}) - \nabla F_j(w^{(t-1)}), w\rangle + \frac{\gamma}{2}\|w - w^{(t-1)}\|^2 + F_j(w). \tag{3}$$

Then the *master* machine computes and broadcasts the averaged model $w^{(t)} = \frac{1}{m}\sum_{j=1}^m w_j^{(t)}$ and its full gradient $\nabla F(w^{(t)}) = \frac{1}{m}\sum_{j=1}^m \nabla F_j(w^{(t)})$ in a map-reduce fashion.

The construction of the local objective (3) is inspired by the idea of leveraging the global first-order information and local higher-order information for local processing. If $F(w)$ is quadratic with condition number $\kappa = L/\mu$ (see Table 2 for notation), then the communication complexity (with high probability) of DANE to reach ϵ-precision was shown to be $\mathcal{O}\left(\kappa^2 n^{-1}\log(1/\epsilon)\right)$ which has a much improved dependency on the condition number κ



(a) Quadratic loss: communication complexity

(b) Logistic loss: global convergence

Figure 1: (a) The number of communication rounds (y-axis) versus number of machines (x-axis) curves of DANE on a synthetic ridge regression task ($N = 2000$, $p = 200$). Here we set $\mu = \mathcal{O}(1/\sqrt{mn})$, $\gamma = \mathcal{O}(1/\sqrt{n})$ and precision $\epsilon = 10^{-5}$. Roughly speaking, the communication complexity scales linearly with respect to $\sqrt{m}$. (b) Illustration of the convergence behavior of D^{2}ANE, DANE-LS and InexactDane on a synthetic logistic regression task ($N = 1000, p = 200, m = 4$) with $\gamma = \mathcal{O}(1/\sqrt{n})$. Each experiment is randomly replicated 10 times.

that could scale as large as $\mathcal{O}(\sqrt{mn})$ in statistical learning problems. The InexactDane (Reddi et al., 2016) method is an inexact implementation of DANE that allows the local sub-problem to be solved inexactly but still possess the above improved communication complexity bounds for quadratic problems. By applying Nesterov's acceleration technique, AIDE (Reddi et al., 2016) and MP-DANE (Wang et al., 2017b) further reduce the communication complexity to $\mathcal{O}\left(\sqrt{\kappa} n^{-1/4} \log(1/\epsilon)\right)$ in the quadratic case, which is nearly tight in view of the lower bound established by Arjevani and Shamir (2015).

On top of the attractive communication-efficiency, another appealing aspect of DANE lies in its versatility in implementation. This is because by nature DANE is an algorithm-agnostic optimization framework, in the sense that the local subproblems can be solved by applying virtually any algorithms designed for the global problem. From the perspective of implementation, this enables fast adaptation of the available single-machine algorithm code to distributed software platform. This contrasts DANE from those algorithm-specific methods such as DiSCO (Zhang and Xiao, 2015) (rooted from damped Newton method) and DSVRG (Shamir, 2016; Lee et al., 2017) (rooted from SVRG). What's more, DANE does not require to access a second-order oracle for its execution, nor does it restrict to any specific problem structure such as the linear prediction models focused by DSCOVR (Xiao et al., 2019) and GIANT (Wang et al., 2018).

Open issues and motivation. Despite the above-mentioned advantages of DANE and its variants, this family of algorithms still exhibits several issues regarding convergence that remain open for exploration, which are raised below.

- Question 1. *Is the convergence bound of plain DANE tight even for quadratic problems?* The communication complexity of plain (exact or inexact) DANE is known to be $\mathcal{O}\left(\kappa^2 n^{-1}\log(1/\epsilon)\right)$ for stochastic quadratic problems (Shamir et al., 2014; Reddi et al., 2016). Since for outer-loop communication DANE only needs to access a first-order oracle of the global problem, we have strong reason to conjecture that the factor on condition number matching this mechanism should be as sharp as $\kappa n^{-1/2}$, even without any momentum acceleration. As visualized in Figure 1(a) for a ridge regression example with $\kappa = \mathcal{O}(\sqrt{mn})$, it is roughly the case that the number of communication rounds scales linearly with respect to $\sqrt{m}$. This leaves a potential theoretical gap between m and $\sqrt{m}$ to be closed.

- Question 2. *Can the strong guarantees of DANE be extended to non-quadratic problems?* The strong communication complexity bounds of DANE-type methods, with or without acceleration, are so far only rigorous for quadratic problems (Shamir et al., 2014; Reddi et al., 2016; Wang et al., 2017b). For more generic convex loss functions, the related bounds are obtained under $\gamma = \mathcal{O}(L)$ which are as slow as those of the ordinary first-order methods and thus are less informative for theoretical justification of performance. It is not clear if DANE-type methods can be guaranteed to converge in the regime $\gamma \ll L$ of interest. In Figure 1(b), we plot the convergence curves of InexactDane under $\gamma = \mathcal{O}(Ln^{-1/2})$ on a synthetic logistic regression task, from which we can observe that apparent zigzag effect occurs in the early stage of communication. Therefore, a natural question to ask is whether the desirable strong guarantees of DANE can be extended to a wider problem spectrum beyond ridge regression.

The primary goal of this work is to answer Question 1 and Question 2 so as to gain deeper understanding of the convergence behavior of DANE in theory and practice.

1.2 Overview of Our Contribution

We address the above questions regarding the convergence of DANE and make progress towards fully understanding DANE both for quadratic and non-quadratic convex functions. To achieve this goal, we propose two new alternatives which are more suitable for convergence analysis as well as for algorithm acceleration. We first propose the *DANE-LS* algorithm as a slight modification of DANE equipped with backtracking line search. The motivation of introducing the line search step is to ensure global asymptotic convergence and facilitate local non-asymptotic analysis for non-quadratic convex problems, which is key to answering Question 2. As another notable difference, DANE-LS only requires the master machine (say F_1) to solve its local subproblem to obtain the next iterate, while the worker machines (say $F_j, j = 2, ..., m$) wait. Such a modification turns out to be beneficial for improving the convergence analysis for quadratic loss, which answers Question 1.

We then show that DANE can be readily accelerated via applying the heavy-ball acceleration technique (Polyak, 1964; Qian, 1999). To this end, we modify the iteration of DANE by adding a small momentum term $\beta(w^{(t-1)} - w^{(t-2)})$ for some $\beta > 0$ to the current iterate

	Method	Quadratic Problem	Non-quadratic Problem
Without momentum acceleration	DANE	$\mathcal{O}\left(\frac{\kappa^2}{n}\log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left(\kappa\log\left(\frac{1}{\epsilon}\right)\right)$
	InexactDane	$\mathcal{O}\left(\frac{\kappa^2}{n}\log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left(\kappa\log\left(\frac{1}{\epsilon}\right)\right)$
	DANE-LS (**ours**)	$\mathcal{O}\left(\frac{\kappa}{\sqrt{n}}\log\left(\frac{1}{\epsilon}\right)\right)$	Globally convergent with local rate $\mathcal{O}\left(\frac{p^{1/2}\kappa}{\sqrt{n}}\log\left(\frac{1}{\epsilon}\right)\right)$
With momentum acceleration	AIDE	$\mathcal{O}\left(\frac{\sqrt{\kappa}}{n^{1/4}}\log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left(\sqrt{\kappa}\log\left(\frac{1}{\epsilon}\right)\right)$
	MP-DANE	$\mathcal{O}\left(\frac{\sqrt{\kappa}}{n^{1/4}}\log\left(\frac{1}{\epsilon}\right)\right)$	✗
	DANE-HB (**ours**)	$\mathcal{O}\left(\frac{\sqrt{\kappa}}{n^{1/4}}\log\left(\frac{1}{\epsilon}\right)\right)$	Local rate: $\mathcal{O}\left(\frac{p^{1/4}\sqrt{\kappa}}{n^{1/4}}\log\left(\frac{1}{\epsilon}\right)\right)$
	D^{2}ANE (**ours**)	$\mathcal{O}\left(\frac{\sqrt{\kappa}}{n^{1/4}}\log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left(\frac{\sqrt{\kappa}}{n^{1/4}}\log^2\left(\frac{1}{\epsilon}\right)\right)$

Table 1: Comparison of communication complexity bounds of different DANE-type methods without (top panel) or with (bottom panel) momentum acceleration. All the bounds for quadratic problem and our results for non-quadratic problem hold with high probability over the random draw of local i.i.d. data. The other results are deterministic. The x-mark "✗" indicates that the related result was not available in the original reference of method.

$w^{(t)}$. We call this alternative as *DANE-HB*. For quadratic problems, we prove that such a simple momentum strategy boosts the communication complexity of DANE to match those of AIDE and MP-DANE but with more elementary analysis. As a perhaps more interesting contribution, DANE-HB can also be shown to have about the same near-optimal bound for strongly convex and twice differentiable objectives in a vicinity of the minimizer, which has not been covered by the previous analysis. For the widely used linear prediction model with smooth and convex losses, we further develop *D^{2}ANE* as a nested Newton-type extension of DANE-HB which can be shown to converge globally at a near-optimal rate.

Highlight of results. Table 1 summarizes our main results on communication complexity of DANE-LS and DANE-HB/D^{2}ANE in stochastic setting and compares them with prior DANE-type methods. These results are divided into two groups respectively for quadratic and non-quadratic strongly convex problems. We use the big $\mathcal{O}$ notation to hide the logarithmic factors involving quantities other than ϵ. As highlighted in the colored cells of Table 1, we contribute several new theoretical insights into DANE, as elaborated below.

- The bound highlighted in *light red* shade gives an affirmative answer to Question 1. That is, in the quadratic case, DANE-LS attains a tighter communication complexity bound $\mathcal{O}\left(\kappa n^{-1/2}\log(1/\epsilon)\right)$ than the already known $\mathcal{O}\left(\kappa^2 n^{-1}\log(1/\epsilon)\right)$ bound for DANE. Such an improvement is achieved by applying a minimal modification of algorithm with model averaging removed on the master machine (note that the line search option of DANE-LS is not activated for quadratic problems). This implies that even without any momentum acceleration, DANE actually can converge faster than already recognized in theory.

- The result highlighted in *light blue* shade answers Question 2 affirmatively. To be more specific, blessed by the backtracking line search, DANE-LS with arbitrary values of $\gamma > 0$ can be proved to converge globally to the unique minimizer when the objective function is strongly convex and twice differentiable. In Figure 1(b) we illustrate the global convergence of DANE-LS when applied to a synthetic logistic regression task. The benefit of line search to DANE-type methods has also been numerically observed in Wang et al. (2018), but without theoretical justification. In the late stage of iteration when the iterate is sufficiently close to the minimizer, the communication complexity of DANE-LS is upper bounded by $\mathcal{O}\left(p^{1/2}\kappa n^{-1/2}\log(1/\epsilon)\right)$. Here the additional factor $p^{1/2}$ arises from invoking uniform concentration analysis to the spectral norm $\|\nabla^2 F_1 - \nabla^2 F\|$ over a bounded domain of interest.

- From the third column of Table 1 we can see that DANE-HB matches AIDE and MP-DANE in communication complexity for quadratic problem. For non-quadratic smooth and strongly convex loss functions, the results highlighted in *light brown* shade shows that DANE-HB possesses an $\mathcal{O}\left(p^{1/4}\sqrt{\kappa}n^{-1/4}\log(1/\epsilon)\right)$ communication complexity bound in a local area around the minimizer. Specially for linear prediction models, by integrating DANE-HB into an inexact Newton-type quadratic approximation framework, we can show that an improved near-tight bound $\mathcal{O}\left(\sqrt{\kappa}n^{-1/4}\log^2(1/\epsilon)\right)$ holds globally for D^{2}ANE, hence answers Question 2 when algorithm acceleration is considered. In contrast, the global convergence bound is as slow as $\mathcal{O}\left(\sqrt{\kappa}\log\left(1/\epsilon\right)\right)$ for AIDE, while for MP-DANE such a bound is not available. See Figure 1(b) for an illustration of the convergence behavior of D^{2}ANE and Table 3 for additional comparison with some other relevant distributed learning methods.

1.3 Other Related Work

Driven by the ever-increasing demand on scaling up machine learning models in modern distributed computing environment, a vast body of distributed optimization algorithms has been developed in several relevant lines of research. A substantial number of these works, including the DANE-type algorithms we work on in this article, focus on communication-efficient distributed learning which is preferable when the network has severely limited bandwidth and high latency (Jaggi et al., 2014; Richtárik and Takáč, 2016; Lee et al., 2017; Jordan et al., 2018; Chen et al., 2020). For a class of self-concordant empirical risk functions, Zhang and Xiao (2015) proposed DiSCO as a distributed inexact damped Newton method in which the Newton step is optimized via a preconditioned conjugate gradient procedure. For quadratic problems, DiSCO attains a near-tight communication complexity bound $\mathcal{O}\left(\sqrt{\kappa}n^{-1/4}\log(1/\epsilon)\right)$ which was soon after matched by AIDE. The SPAG method proposed by Hendrikx et al. (2020) is a preconditioned accelerated first-order algorithm that achieves near-tight rate of convergence for distributed optimization. For high-dimensional sparse estimation, EDSL (Wang et al., 2017a) and DINPS (Liu et al., 2019) respectively extend the idea of DANE to solving distributed ℓ_1-ERM and ℓ_0-ERM problems, obtaining analogous improvement in communication efficiency.

For large-scale convex learning with linear models, CoCoA (Jaggi et al., 2014) and CoCoA+ (Ma et al., 2015; Smith et al., 2018) were developed inside the framework of block coordinate descent/ascent to perform expensive local computations with the aim of reduc-

ing the overall communications across the network. In the same problem setting, Xiao et al. (2019) proposed DSCOVR as a family of randomized primal-dual block coordinate algorithms for asynchronous distributed optimization with a roughly $\mathcal{O}(m\log(1/\epsilon))$ communication complexity bound. Also for linear prediction models, the GIANT method (Wang et al., 2018) improves over DANE in communication complexity bound under the condition that sample size should be sufficiently larger than feature size.

With additional memory and preprocessing at each machine, Lee et al. (2017) showed that SVRG (Johnson and Zhang, 2013) can be adapted for distributed optimization to attain $\mathcal{O}(1)$ communication complexity, and nearly linear speed-up in first-order oracle computation complexity can be achieved in the regime where sample size dominates condition number. Specifically for linear models, a more efficient implementation of distributed SVRG method was proposed and analyzed by Shamir (2016) under the without replacement sampling strategy. By combining DSVRG with minibatch passive-aggressive updates, the MP-DSVRG method (Wang et al., 2017b) was shown to have provably better tradeoff between communication and memory efficiency for quadratic loss functions. The equivalence between a distributed implementation of SVRG and InexactDane has been revealed in the framework of Federated SVRG (Konečnỳ et al., 2016) for distributed machine learning with extremely large number of nodes. Last but not least, the well designed distributed learning platforms such as MapReduce (Dean and Ghemawat, 2008), Apache Spark (Zaharia et al., 2016), Petuum (Xing et al., 2015), Parameter Server (Li et al., 2014), "AI-Box" (GPU parameter servers for commercial Ads CTR models) (Zhao et al., 2019, 2020), etc., have significantly facilitated the system implementation of distributed optimization algorithms.

1.4 Organization and Notation

Paper organization. The rest of this article is organized as follows: In Section 2, we introduce DANE-LS as a new alternative of DANE with backtracking line search and analyze its convergence rate for quadratic and non-quadratic convex problems. In Section 3, we propose DANE-HB to accelerate DANE using heavy-ball approach, along with a variant specifically designed for linear prediction with convex losses. The numerical evaluation results are presented and discussed in Section 4. Finally, we conclude this article in Section 5. All the technical proofs of theoretical results are deferred to the appendix section.

Notation. The key quantities and notations that commonly used in our analysis are summarized in Table 2. In stochastic setting, unless otherwise stated, we use the big $\mathcal{O}$ notation that hides inside the logarithmic factors involving quantities other than ϵ.

2. Globalization of DANE with Sharper Analysis

In this section, we provide a global and sharper analysis of the plain version of DANE method without applying any momentum acceleration. The analysis is actually conducted on a modified version of DANE augmented with backtracking line search, while only a master machine is allocated to do local computation in an inexact manner. Such simple modifications turn out to be beneficial for the global asymptotic and local non-asymptotic analysis of DANE.

Notation	Definition
m	number of worker machines
n	number of training samples distributed on each individual worker machine
$N = mn$	total number of training samples
p	number of features
$F(w)$	global risk function
$F_1(w)$	local risk function on the master machine
L	Lipschitz smoothness modulus of the gradient vector $\nabla F(w)$
ν	Lipschitz smoothness modulus of the Hessian matrix $\nabla^2 F(w)$
μ	strong convexity modulus of $F(w)$
$\kappa = L/\mu$	condition number of $F(w)$
β	momentum strength coefficient for heavy-ball acceleration
ϵ	sub-optimality of the global problem
ε	sub-optimality of the local subproblem
γ	regularization strength of the local subproblem
δ	tail probability bound in stochastic setting
$[N]$	abbreviation of the index set $\{1, ..., N\}$
$\|x\| = \sqrt{x^\top x}$	Euclidean norm of a vector x
$\lambda_{\max}(A)$	the largest eigenvalue of a matrix A
$\lambda_{\min}(A)$	the smallest eigenvalue of a matrix A
$A \succeq B$	$A - B$ is symmetric, positive semi-definite
$A \succ B$	$A - B$ is symmetric, positive definite
$\|A\|$	spectral norm of matrix A
$\rho(A)$	spectral radius of A, i.e., its largest (in modulus) eigenvalue

Table 2: Table of notation.

2.1 Leveraging Backtracking Line Search

Since DANE is essentially an approximated second-order method, it is natural to consider introducing an additional line search operation to hopefully guarantee global convergence while preserving the appealing local rate of convergence. In practice, the numerical evidence in the work of Wang et al. (2018) has already demonstrated that backtracking line search is beneficial for improving the convergence performance of DANE-type methods, although without any theoretical support. In view of these, we propose the DANE-LS (DANE with Line Search) method which is outlined in Algorithm 1. The notable differences between DANE-LS and the vanilla DANE are summarized in below:

- For non-quadratic problems, two optional backtracking line search steps (as highlighted in gray shade) are conducted on the master machine. The Option-I needs to evaluate the global objective value and hence requires additional communication cost. By only accessing the locally available information, the Option-II is free of evaluating the global objective value but at the price of introducing an additional hyper-parameter ν which quantifies the smoothness of Hessian.

Algorithm 1: DANE with backtracking Line Search: `DANE-LS`(γ, ρ, ν)

Input : Parameters $\gamma, \nu > 0$, $\rho \in (0, 1/3]$.
Output: $w^{(t)}$.
Initialization *Set* $w^{(0)} = 0$ *or* $w^{(0)} \approx \arg\min_w F_1(w)$.
for $t = 1, 2, ...$ **do**
`/* Global computation on the master machine associated with` $F_1(w)$ `*/`
Compute $\nabla F(w^{(t-1)}) = \frac{1}{m}\sum_{j=1}^m \nabla F_j(w^{(t-1)})$;
Estimate $\tilde{w}^{(t)}$ such that $\|\nabla P^{(t-1)}(\tilde{w}^{(t)})\| \le \varepsilon_t$, where

$$P^{(t-1)}(w) := \langle \nabla F(w^{(t-1)}) - \nabla F_1(w^{(t-1)}), w\rangle + \frac{\gamma}{2}\|w - w^{(t-1)}\|^2 + F_1(w); \quad (4)$$

if *The objective function* F *is non-quadratic* **then**
`/* Backtracking line search */`
Update $w^{(t)} = (1-\eta_t)w^{(t-1)} + \eta_t \tilde{w}^{(t)}$ with proper $\eta_t \in (0, 1]$ which satisfies either of the following *sufficient descent* condition for the provided ρ:

(**Option-I**) `/* Line-search with global value evaluation. */`

$$F(w^{(t)}) \le F(w^{(t-1)}) - \psi(\tilde{w}^{(t)}, w^{(t-1)}), \quad (5)$$

where

$$\psi(\tilde{w}^{(t)}, w^{(t-1)}) := \eta_t\rho\langle \nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)}\rangle - \eta_t\varepsilon_t\|\tilde{w}^{(t)} - w^{(t-1)}\|;$$

(**Option-II**) `/* Line-search without global value evaluation. */`

$$\langle \nabla F(w^{(t-1)}), w^{(t)} - w^{(t-1)}\rangle + (w^{(t)} - w^{(t-1)})^\top(\nabla^2 F_1(w^{(t-1)}) + \gamma I)(w^{(t)} - w^{(t-1)}) + \frac{\nu}{6}\|w^{(t)} - w^{(t-1)}\|^3 \le -\psi(\tilde{w}^{(t)}, w^{(t-1)}).$$

end
else
$w^{(t)} = \tilde{w}^{(t)}$;
end
`/* Local gradient evaluation on worker machines */`
For each machine j, compute $\nabla F_j(w^{(t)})$ and send it to the master machine;
end

- As another notable difference, only a master machine is in charge of solving a local subproblem associated with $F_1(w)$ to obtain the next iterate, during which time the other worker machines stay idle. Such a master-slave architecture has been widely adopted and investigated in many distributed machine learning and statistical inference approaches (Shamir, 2016; Lee et al., 2017; Wang et al., 2017a; Jordan et al., 2018). Allowing only master to do the heavy lifting is certainly more energy saving and less sensitive to network latency.

As the consequence of the above modifications, DANE-LS can be shown to improve over DANE not only for non-quadratic convex objectives (see Section 2.3) but also for the well-studied quadratic case (see Section 2.2). Moreover, the master-slave computing architecture eases the extension of analysis to the heavy-ball acceleration presented in the next section. It is noteworthy that the local subproblem is allowed to be solved inexactly with sub-optimality $\|\nabla P^{(t-1)}(\tilde{w}^{(t)})\| \leq \varepsilon_t$. Such a local sub-optimality condition is computationally more tractable for verification than those of InexactDane and AIDE with unknown local minimizers involved, and hence is more practical from the perspective of algorithm implementation.

2.2 Sharper Bounds for Quadratic Functions

We start by analyzing DANE-LS in a simple yet informative regime where the loss functions are quadratic. In this setting, the line search options will not be activated throughout the algorithm execution. Our analysis assumes the conditions of strong convexity and Lipschitz smoothness which are conventionally used in analyzing distributed optimization algorithms.

Definition 1 (Strong Convexity/Smoothness) *A differentiable function g is μ-strongly-convex and L-smooth if $\forall w, w'$,*

$$\frac{\mu}{2}\|w - w'\|^2 \leq g(w) - g(w') - \langle \nabla g(w'), w - w' \rangle \leq \frac{L}{2}\|w - w'\|^2.$$

The ratio value $\kappa = L/\mu$ is referred to as the condition number. We further introduce the concept of Lipschitz continuous Hessian which characterizes the smoothness of gradient.

Definition 2 (Lipschitz Continuous Hessian) *We say a twice continuously differentiable function g has Lipschitz continuous Hessian with constant $\nu \geq 0$ (ν-LH) if $\forall w, w'$,*

$$\left\|\nabla^2 g(w) - \nabla^2 g(w')\right\| \leq \nu \|w - w'\|.$$

Let $w^* = \arg\min_w F(w)$ denote the global minimizer of F. The following theorem is our main result on the convergence rate of DANE-LS for quadratic functions in terms of parameter estimation error.

Theorem 3 (Convergence rate of DANE-LS for quadratic loss) *Assume that the loss function is quadratic. Let H and H_1 be the Hessian matrices of the global objective F and local objective F_1, respectively. Assume that $\mu I \preceq H \preceq LI$. Given precision $\epsilon > 0$, if $\|H_1 - H\| \leq \gamma$ and $\varepsilon_t \leq \frac{\mu^2 \|\nabla F(w^{(t-1)})\|}{2(\mu+2\gamma)L}$, then Algorithm 1 will output $w^{(t)}$ satisfying $\|w^{(t)} - w^*\| \leq \epsilon$ after*

$$t \geq \frac{2(\mu + 2\gamma)}{\mu} \log\left(\frac{\sqrt{\kappa}\|w^{(0)} - w^*\|}{\epsilon}\right)$$

rounds of iterations.

As a comparison, the communication complexity bounds established for DANE (Shamir et al., 2014, Lemma 1) and InexactDane (Reddi et al., 2016, Corollary 1) are both of the order $\mathcal{O}\left(\gamma^2/\mu^2 \log(1/\epsilon)\right)$, which are inferior to the $\mathcal{O}\left(\gamma/\mu \log(1/\epsilon)\right)$ bound established in Theorem 3, as long as $\gamma/\mu > 1$. After a careful inspection of the technical proofs in Shamir et al. (2014); Reddi et al. (2016), we notice that the looseness of the former bounds essentially comes from the reduce step conducted by master machine for aggregating models from local workers, and such an issue is seemingly difficult to be remedied inside the original architecture of DANE. In this paper, with the proposed modifications, the tighter bound in Theorem 3 can be attained based on a fairly elementary analysis. This answers Question 1 affirmatively.

We further derive the following result as an implication of Theorem 3 to the stochastic setting where the samples are uniformly randomly distributed over machines.

Corollary 4 *Assume the conditions in Theorem 3 hold and $\|\nabla^2 f(w; x_i, y_i)\| \leq L$ for all $i \in [N]$. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the samples drawn to construct F_1, Algorithm 1 with $\gamma = L\sqrt{\frac{32\log(p/\delta)}{n}}$ will output $w^{(t)}$ satisfying $\|w^{(t)} - w^*\| \leq \epsilon$ after*

$$t \geq \left(2 + 4\kappa\sqrt{\frac{32\log(p/\delta)}{n}}\right) \log\left(\frac{\sqrt{\kappa}\|w^{(0)} - w^*\|}{\epsilon}\right)$$

rounds of iterations.

Remark 5 *In statistical learning problems, the condition number κ could scale as large as $\mathcal{O}(\sqrt{mn})$ for optimal generalization (Shalev-Shwartz et al., 2009). If this is the case, then Corollary 4 implies an $\mathcal{O}\left(\sqrt{m}\log(1/\epsilon)\right)$ communication complexity bound for stochastic quadratic problems, which contrasts itself from the $\mathcal{O}\left(m\log(1/\epsilon)\right)$ bound previously known for vanilla DANE and InexactDane methods. Notice, such improvement is of particular interest in the regime of federated machine learning where the number of computing nodes m could be huge (Konečnỳ et al., 2016; McMahan et al., 2017).*

We comment that in the quadratic case, DANE-LS shares an identical spirit of preconditioning to the distributed preconditioned conjugate gradient (DPCG) method developed for computing the inexact Newton step of DiSCO (Zhang and Xiao, 2015). Actually, based on the similarity between local and global Hessian matrices, F_1 essentially serves as a preconditioner which is effective in significantly reducing the condition number of local objective

$P^{(t)}$ when local data is sufficiently correlated to the global one. As we will show shortly in the next subsection that such a preconditioning effect of F_1 is also beneficial for improving the communication efficiency of DANE-LS for non-quadratic strongly convex problems. From the viewpoint of implementation, in contrast to DPCG that is implemented based on preconditioned conjugate gradient method, the local preconditioned subproblems in our method can be more flexibly optimized via a wider spectrum of algorithms including the stochastic variance reduction methods to gain better computational efficiency.

2.3 Global Analysis for Strongly Convex Functions

We now move to consider the more general regime in which the objective function is strongly convex and twice differentiable with Lipschitz continuous Hessian. First, we show in the following key lemma that the proposed global and local backtracking line search steps are always feasible under natural conditions.

Lemma 6 (Feasibility of line search) *Assume that F is L-smooth and F_1 is μ-strongly convex. For any given $\rho \in (0,1)$,*

(a) if the length of search satisfies

$$0 < \eta_t \leq \min\left\{1, \frac{2(\gamma+\mu)(1-\rho)}{L}\right\},$$

then the global backtracking line search (Option-I) is feasible, i.e.,

$$F(w^{(t)}) \leq F(w^{(t-1)}) - \psi(\tilde{w}^{(t)}, w^{(t-1)}),$$

where $\psi(\tilde{w}^{(t)}, w^{(t-1)}) := \eta_t\rho\langle\nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)}\rangle - \eta_t\varepsilon_t\|\tilde{w}^{(t)} - w^{(t-1)}\|$.

(b) Moreover, assume that $F_1(w)$ has ν-LH and $\exists D > 0$ such that $\|\tilde{w}^{(t)} - w^{(t-1)}\| \leq D$ for all $t \geq 0$. If

$$\eta_t \leq \min\left\{1, \frac{-(3\nu D + 6(\gamma+\mu)) + \sqrt{(3\nu D + 6(\gamma+\mu))^2 + 96(1-\rho)\nu D(\gamma+\mu)}}{4\nu D}\right\},$$

then the local backtracking line search (Option-II) is feasible, i.e.,

$$\langle\nabla F(w^{(t-1)}), w^{(t)} - w^{(t-1)}\rangle + \frac{1}{2}(w^{(t)} - w^{(t-1)})^\top(\nabla^2 F_1(w^{(t-1)}) + \gamma I)(w^{(t)} - w^{(t-1)}) + \frac{\nu}{6}\|w^{(t)} - w^{(t-1)}\|^3 \leq -\psi(\tilde{w}^{(t)}, w^{(t-1)}).$$

Remark 7 *The bound D in the part (b) of Lemma 6 is reasonable if we focus on an ℓ_2-norm bounded domain of interest Ω such that $D = \max_{w,w'\in\Omega}\|w - w'\|$. The result also implies that if the Option-I is carried out under Armijo rule for global line search, then the additional rounds of communication needed for global objective evaluation is roughly of the order $\mathcal{O}\left(\log\left(\frac{L}{(\gamma+\mu)(1-\rho)}\right)\right)$.*

The following theorem is our main result on the global convergence of DANE-LS.

Theorem 8 (Global convergence of DANE-LS) *Assume that $F(w)$ and $F_1(w)$ are L-smooth, μ-strongly-convex and have ν-LH. Suppose that $\varepsilon_t \leq \frac{\rho(\mu+\gamma)}{2(L+\gamma)+\rho(\mu+\gamma)}\|\nabla F(w^{(t-1)})\|$.*

(a) Then the objective value sequence $\{F(w^{(t)})\}$ generated by Algorithm 1 with the global line search step (Option-I) converges and the difference norm sequence $\{\|\tilde{w}^{(t)}-w^{(t-1)}\|\}$ converges to zero.

(b) Assume in addition that $\sup_w \|\nabla^2 F_1(w)-\nabla^2 F(w)\| \leq \gamma$ and $\|\tilde{w}^{(t)}-w^{(t-1)}\|$ is bounded from above for all $t \geq 0$. Then the objective value sequence $\{F(w^{(t)})\}$ generated by Algorithm 1 with the local line search step (Option-II) converges and the difference norm sequence $\{\|\tilde{w}^{(t)} - w^{(t-1)}\|\}$ converges to zero.

Remark 9 *Theorem 8 suggests a natural way of controlling the termination of Algorithm 1 by monitoring either the objective value progress $|F(w^{(t)}) - F(w^{(t-1)})|$ or the estimation vector difference $\|\tilde{w}^{(t)} - w^{(t-1)}\|$.*

Local non-asymptotic convergence. We further study the local convergence behavior of DANE-LS. The starting point is to show, via the following lemma, that the unit length eventually satisfies the sufficient descent condition in (5).

Lemma 10 (Acceptability of unit length for line search) *Assume that the conditions in Theorem 8 hold. Then for any fixed $\rho \in (0, 1/3]$, the unit length $\eta_t = 1$ guarantees the sufficient descent condition (5) provided that t is sufficiently large.*

The following lemma establishes the local convergence rate of Algorithm 1 when $\eta_t \equiv 1$, i.e., when the unit length is always accepted by the backtracking line search.

Lemma 11 (Local convergence rate of DANE-LS) *Assume that F and F_1 are L-smooth, μ-strongly-convex and have ν-LH. Assume that $\sup_w \|\nabla^2 F_1(w) - \nabla^2 F(w)\| \leq \gamma$. Let $\tau = \left\lceil \frac{\mu+2\gamma}{2\mu} \log(4\kappa) \right\rceil$. Suppose that $\varepsilon_t \leq \min\left\{(\gamma+\mu)^2, \frac{\|\nabla F(w^{(t-1)})\|^2}{L^2}\right\}$ and $\max_{0\leq i\leq \tau-1} \|w^{(i)} - w^*\| \leq \frac{(\gamma+\mu)}{4(6\nu+1)\sqrt{\kappa}\tau}$. Then for any $\epsilon > 0$, Algorithm 1 with $\eta_t \equiv 1$ will attain estimation error $\|w^{(t)} - w^*\| \leq \epsilon$ after*

$$t \geq 4\tau \log\left(\frac{\gamma+\mu}{4(6\nu+1)\sqrt{\kappa}\tau\nu\epsilon}\right)$$

rounds of iterations.

Remark 12 *Lemma 11 essentially shows that up to the logarithmic factors on κ and τ, the local communication complexity of DANE-LS is upper bounded by $\mathcal{O}(\gamma/\mu \log(1/\epsilon))$, which matches the bound for the quadratic function.*

We are now ready to present our main result on the local non-asymptotic convergence of DANE-LS for strongly convex functions.

Theorem 13 (Non-asymptotic convergence of DANE-LS) *Assume that F and F_1 are μ-strongly-convex, L-smooth and have ν-LH. Assume that $\sup_w \|\nabla^2 F_1(w) - \nabla^2 F(w)\| \leq \gamma$. Suppose that $\rho \in (0, 1/3]$ and*

$$\varepsilon_t \leq \min\left\{(\gamma+\mu)^2, \frac{\|\nabla F(w^{(t-1)})\|^2}{L^2}, \frac{\rho(\mu+\gamma)}{2(L+\gamma)+\rho(\mu+\gamma)}\|\nabla F(w^{(t-1)})\|\right\}.$$

Let $\tau = \left\lceil \frac{\mu+2\gamma}{2\mu}\log(4\kappa)\right\rceil$. Then there exists a time stamp t_0, which is invariant to ϵ, such that Algorithm 1 will output solution $w^{(t)}$ satisfying $\|w^{(t)} - w^*\| \leq \epsilon$ after

$$t \geq t_0 + 4\tau \log\left(\frac{\gamma+\mu}{4(6\nu+1)\sqrt{\kappa}\tau}\left(\frac{1}{\epsilon}\right)\right)$$

rounds of iterations.

Remark 14 *Theorem 13 reveals that DANE-LS converges globally towards w^* and in a local area around w^* it enjoys a linear rate of convergence with complexity $\mathcal{O}(\gamma/\mu \log(1/\epsilon))$. We comment on the choice of γ in the stochastic setting. For the considered L-smooth objective functions, standard uniform concentration theory (see, e.g., Zhang and Xiao, 2015; Mei et al., 2018) suggests that the concentration bound $\sup_w \|\nabla^2 F_1(w) - \nabla^2 F(w)\| \leq \gamma = \mathcal{O}\left(L\sqrt{p/n}\right)$ holds with high probability over a bounded domain of interest. Then with such a choice of γ the local communication complexity of DANE-LS is bounded as $\mathcal{O}\left(p^{1/2}\kappa n^{-1/2}\log(1/\epsilon)\right)$ with high probability, which shows the benefit of statistical correlation of local problems for global optimization when $n \gg p$. This result partially answers Question 2 as raised in Section 1.1.*

3. Heavy-Ball Acceleration of DANE

We further introduce a simple yet effective momentum acceleration method for DANE based on the classic heavy-ball approach (Polyak, 1964), which has long been acknowledged to work favorably for accelerating first-order optimization methods (Ghadimi et al., 2015; Wilson et al., 2016; Loizou and Richtárik, 2017; Zhou et al., 2018; Chee and Li, 2020).

3.1 The DANE-HB Algorithm

As outlined in Algorithm 2, the proposed DANE-HB method shares an almost identical centralized computing architecture to DANE-LS. The main difference is that for local sub-problem optimization in the master machine, we first estimate $\tilde{w}^{(t)} \approx \arg\min_w P^{(t-1)}(w)$, and then compute $w^{(t)} = \tilde{w}^{(t)} + \beta(w^{(t-1)} - w^{(t-2)})$ as a linear combination of $\tilde{w}^{(t)}$ and the previous two iterates, where $\beta > 0$ is the momentum strength coefficient. It is optional to implement the backtracking line search steps (like in Algorithm 1) which work well in our numerical practice to obtain global convergence, although there is no theoretical guarantee that the difference vector $w^{(t)} - w^{(t-1)}$ should point to a descent direction. Regarding initialization, the simplest way is to set $w^{(-1)} = w^{(0)} = 0$, i.e., starting the iteration from scratch. Since $F_1(w)$ is expected to be close to $F(w)$ in stochastic setting, another reasonable option of initialization is to set $w^{(-1)} = w^{(0)} \approx \arg\min_w F_1(w)$ which is also expected to be close to the global solution w^*.

Algorithm 2: DANE with Heavy-Ball acceleration: `DANE-HB`(γ, β)

Input : Parameters $\gamma, \beta > 0$.
Output: $w^{(t)}$.
Initialization *Set* $w^{(0)} = 0$ *or* $w^{(0)} \approx \arg\min_w F_1(w)$. *Let* $w^{(-1)} = w^{(0)}$.
for $t = 1, 2, \ldots$ **do**
  `/* Global computation on master machine */`
  Compute $\nabla F(w^{(t-1)}) = \frac{1}{m}\sum_{j=1}^m \nabla F_j(w^{(t-1)})$;
  Estimate $\tilde{w}^{(t)}$ such that $\|\nabla P^{(t-1)}(\tilde{w}^{(t)})\| \le \varepsilon_t$, where $P^{(t-1)}$ is defined by (4);
  Compute $w^{(t)} = \tilde{w}^{(t)} + \beta(w^{(t-1)} - w^{(t-2)})$;
  (Optionally) Conduct backtracking line search.
  `/* Local gradient evaluation on worker machines */`
  For each machine j, compute $\nabla F_j(w^{(t)})$ and send it to the master machine;
end

3.2 Convergence Results for Quadratic Functions

The following result shows that the heavy-ball acceleration strategy can improve the communication efficiency of DANE for quadratic problems.

Theorem 15 (Convergence rate of DANE-HB for quadratic function) *Assume that the loss function is quadratic. Let H and H_1 be the Hessian matrices of the global objective F and local objective F_1, respectively. Assume that $\mu I \preceq H \preceq LI$. Set $\beta = \left(1 - \sqrt{\frac{\mu}{\mu+2\gamma}}\right)^2$ and $\varepsilon_t = \frac{\sqrt{2}(\mu+\gamma)\|\nabla F(w^{(0)})\|}{2L(t+1)^2}\left(1 - \frac{1}{2}\sqrt{\frac{\mu}{\mu+2\gamma}}\right)^{t+1}$. Given precision $\epsilon > 0$, if $\|H_1 - H\| \le \gamma$, then Algorithm 2 will output $w^{(t)}$ satisfying $\|w^{(t)} - w^*\| \le \epsilon$ after*

$$t \ge 2\sqrt{\frac{\mu+2\gamma}{\mu}} \log\left(\frac{2\sqrt{2}c\|w^{(0)} - w^*\|}{\epsilon}\right)$$

rounds of iterations, where c is a constant relying on $\sqrt{\mu/(\mu+2\gamma)}$.

The following corollary is the implication of Theorem 15 in stochastic setting.

Corollary 16 *Assume the conditions in Theorem 15 hold and $\|\nabla^2 f(w; x_i, y_i)\| \le L$ for all $i \in [N]$. Then for any $\delta \in (0,1)$, with probability at least $1 - \delta$ over the random samples drawn to construct F_1, Algorithm 2 with $\gamma = L\sqrt{\frac{32\log(p/\delta)}{n}}$ will attain estimation error $\|w^{(t)} - w^*\| \le \epsilon$ after*

$$t \ge \mathcal{O}\left(\frac{\sqrt{\kappa}}{n^{1/4}} \log^{1/4}\left(\frac{p}{\delta}\right) \log\left(\frac{1}{\epsilon}\right)\right)$$

rounds of iterations.

Remark 17 *The result shows that in the quadratic case, DANE-HB matches the communication complexity lower bounds (up to logarithmic factors) proved by Arjevani and Shamir (2015). Similar guarantees for quadratic problems have also been proved for AIDE and MP-DANE based on the catalyst acceleration technique (Lin et al., 2015), and for DiSCO based on preconditioned conjugate gradient methods.*

3.3 Convergence Results for Strongly Convex Functions

We further study the performance of DANE-HB when applied to a broad class of strongly convex functions with Lipschitz continuous Hessian. In this general case, the following result shows that in a vicinity of the global minimizer, DANE-HB enjoys the same appealing rate of convergence as established for the ridge regression problems.

Theorem 18 (Local convergence rate of DANE-HB) *Assume that F and F_1 are L-smooth, μ-strongly-convex and has ν-LH. Assume that $\sup_w \|\nabla^2 F_1(w) - \nabla^2 F(w)\| \leq \gamma$. Choose $\beta = \left(1 - \sqrt{\mu/(\mu+2\gamma)}\right)^2$. Let $\tau = \left\lceil 2\sqrt{(\mu+2\gamma)/\mu}\log(2c)\right\rceil$ in which c is a constant dependent on $\sqrt{\mu/(\mu+2\gamma)}$. Assume that $\varepsilon_t \leq \min\left\{(\gamma+\mu)^2, \|\nabla F(w^{(t-1)})\|^2/L^2\right\}$. Given precision $\epsilon > 0$, if $\max_{-1\leq i\leq \tau-1}\|w^{(i)} - w^*\| \leq \frac{\gamma+\mu}{4(6\nu+1)\sqrt{2}c\tau}$, then Algorithm 2 will output $w^{(t)}$ satisfying $\|w^{(t)} - w^*\| \leq \epsilon$ after*

$$t \geq 4\tau \log\left(\frac{\gamma+\mu}{4(6\nu+1)c\tau}\left(\frac{1}{\epsilon}\right)\right)$$

rounds of iterations.

Remark 19 *It has been proved that AIDE converges at the rate of $\mathcal{O}\left(\sqrt{\kappa}\log(1/\epsilon)\right)$ for non-quadratic strongly convex functions with $\gamma = \mathcal{O}(L)$, and that result is global (Reddi et al., 2016, Theorem 6). In a local area around the global minimizer, we obtain the $\mathcal{O}\left(\sqrt{\gamma/\mu}\log(1/\epsilon)\right)$ rounds of communication bound in Theorem 18 for an arbitrary $\gamma > 0$ as long as the γ-related condition $\sup_w \|\nabla^2 F_1(w) - \nabla^2 F(w)\| \leq \gamma$ holds. Particularly, with the choice of $\gamma = \mathcal{O}\left(L\sqrt{p/n}\right)$ as suggested by Zhang and Xiao (2015, Lemma 5), the local communication complexity of DANE-HB scales as $\mathcal{O}\left(p^{1/4}\sqrt{\kappa}n^{-1/4}\log(1/\epsilon)\right)$ with high probability, which matches DiSCO and outperforms AIDE when $n \gg p$ in large-scale statistical learning problems.*

3.4 Extension for Convex Optimization with Linear Models

So far, DANE-HB has been shown to converge globally for the quadratic objective, whilst for non-quadratic problems it can merely be shown to converge in a vicinity of the global minimizer. In this section, we move to study a special class of convex learning problems with linear regression or prediction models. More specifically, we consider the loss function of the form

$$f(w; x_i, y_i) = l(w^\top x_i, y_i) + \frac{\mu}{2}\|w\|^2,$$

where $l(w^\top x_i; y_i)$ is a convex function that measures the linear regression/prediction loss of w at data point (x_i, y_i) and $\mu > 0$ controls the strength of ℓ_2-regularization. For example, the quadratic loss $l(w^\top x_i, y_i) = \frac{1}{2}(y_i - w^\top x_i)^2$ is used in least squares regression and the logistic loss $l(w^\top x_i, y_i) = \log\left(1 + \exp(-y_i w^\top x_i)\right)$ in logistic binary classification. Then we can re-express Problem (1)

$$\min_{w\in\mathbb{R}^p} F(w) = \tilde{F}(w) + \frac{\mu}{2}\|w\|^2,$$

where $\tilde{F}(w) := \frac{1}{N}\sum_{i=1}^{N} l(w^\top x_i, y_i)$. For such a special strongly convex problem, we propose a double-loop extension of DANE-HB and showcase that the proposed method enjoys a global near-optimal communication complexity bound.

3.4.1 The D^{2}ANE Algorithm

Algorithm 3: Distributed Doubly Approximate Newton: D^{2}ANE(γ, β, ℓ)

Input : Hyper-parameters $\gamma, \beta, \ell > 0$. Typically $\gamma = \mathcal{O}(1/\sqrt{n})$.
Output: $w^{(t)}$.
Initialization *Set* $w^{(0)} = 0$ *or* $w^{(0)} \approx \arg\min_w F_1(w)$.
for $t = 1, 2, \ldots$ **do**

(S1) Construct a quadratic approximation function to F at $w^{(t-1)}$ which is expressed as $Q^{(t-1)}(w) :=$

$$F(w^{(t-1)}) + \langle \nabla F(w^{(t-1)}), w - w^{(t-1)} \rangle + \frac{1}{2}(w - w^{(t-1)})^\top H (w - w^{(t-1)}), \quad (6)$$

where $H = \frac{\ell X X^\top}{N} + \mu I$.

(S2) Estimate $w^{(t)} =$ DANE-HB(γ, β) by applying DANE-HB (Algorithm 2) to $Q^{(t-1)}(w)$ with a warm-start initialization $w^{(t-1)}$ such that

$$Q^{(t-1)}(w^{(t)}) \le \min_w Q^{(t-1)}(w) + \epsilon_t.$$

end

D^{2}ANE (Distributed Doubly Approximate NEwton) is formally stated in Algorithm 3. The algorithm contains an outer-loop iteration for constructing an approximate Newton-type quadratic approximation to the global empirical risk, which is then optimized via an inner-loop DANE-HB method. More specifically, at each iterate $w^{(t-1)}$, we first construct in the step S1 a quadratic approximation function $Q^{(t-1)}(w)$ to the original problem around $w^{(t-1)}$ as expressed by (6). Then in the step S2 we apply DANE-HB as an inner-loop iterative procedure to (approximately) optimize $Q^{(t-1)}$ in a distributed fashion. Suppose that the loss function $l(\cdot, \cdot)$ is twice differentiable with respect to its first argument and $|l''(a, \cdot)| \le \ell$ for all a. Then we can verify that for any w, the Hessian matrix of F can be upper bounded as $\nabla^2 F(w) = \frac{1}{N}\sum_{i=1}^{N} l''(w^\top x_i, y_i) x_i x_i^\top + \mu I \preceq \frac{\ell X X^\top}{N} + \mu I = H$. This implies that $Q^{(t-1)}$ is an upper bound of the second-order Taylor expansion of F at $w^{(t-1)}$, which justifies our calling $Q^{(t-1)}$ as an approximate Newton-type quadratic approximation to F.

An alternative way for constructing the outer-loop quadratic approximation in Algorithm 3 is to replace H with the exact Hessian matrix $\nabla^2 F(w^{(t)})$ in $Q^{(t)}$. For DiSCO, such an exact Newton approximation step has been shown to work favorably for optimizing self-concordant functions via damped Newton method (Zhang and Xiao, 2015). While it is prospective to adapt D^{2}ANE to that framework with quadratic subproblems solved by DANE-HB rather than DPCG, we nevertheless still choose to work on the inexact Newton

step (6) with a fixed Hessian H, which turns out to imply stronger communication bound than its exact Newton counter part in terms of the dependence on feature dimension.

3.4.2 Convergence Analysis

Let X_1 denote the subset of data samples associated with F_1 that were stored on the master machine. The following is our main result on the convergence rate of D^{2}ANE for strongly convex learning with linear models.

Theorem 20 (Convergence of D^{2}ANE) *Assume that the univariate functions l_i are ℓ-smooth and σ-strongly convex. Assume without loss of generality that $\|x_i\| \leq 1$. Let $H = \frac{\ell}{N} XX^\top + \mu I$ and $H_1 = \frac{\ell}{n} X_1 X_1^\top + \mu I$. Choose $\beta = \left(1 - \sqrt{\frac{\mu}{\mu+2\gamma}}\right)^2$ and $\epsilon_t = \frac{\sigma}{2\ell} \exp\left\{-\frac{\sigma(t-1)}{2\ell}\right\}$. If $\|H_1 - H\| \leq \gamma$, then Algorithm 3 will output solution $w^{(t)}$ with sub-optimality $F(w^{(t)}) - F(w^*) \leq \epsilon$ after*

$$t \geq \frac{2\ell}{\sigma} \log\left(\frac{\max\left\{1, F(w^{(0)}) - F(w^*)\right\}}{\epsilon}\right)$$

rounds of outer-loop iterations and

$$\mathcal{O}\left(\frac{\ell}{\sigma}\sqrt{\frac{\gamma}{\mu}} \log^2\left(\frac{1}{\epsilon}\right)\right)$$

rounds of inner-loop iterations of DANE-HB.

Remark 21 *When the univariate function l_i is second-order differentiable, the condition of ℓ-smooth and σ-strongly convex is identical to $\sigma \leq l_i''(\cdot) \leq \ell$. For the quadratic loss function $l(w^\top x_i, y_i) = \frac{1}{2}(y_i - w^\top x_i)^2$, we have $\ell = \sigma = 1$. For the binary logistic loss $l(w^\top x_i, y_i) = \log\left(1 + \exp(-y_i w^\top x_i)\right)$, without loss of generality we assume $\|x_i\| \leq 1\ \forall i$ and the domain of interest[1] is bounded, i.e., $\|w\| \leq B$ for some $B > 0$. Then we can verify that $\ell = 1/4$ and $\sigma = \exp(B)/(1+\exp(B))^2$ which typically does not scale with feature dimension.*

The following is a stochastic variant of Theorem 20 in the setting where the samples are uniformly randomly distributed over machines.

Corollary 22 *Assume the conditions in Theorem 20 hold. Then for any $\delta \in (0,1)$, with probability at least $1 - \delta$ over the samples drawn to construct $F_1(w)$, Algorithm 3 with $\gamma = (\ell + \mu)\sqrt{\frac{32\log(p/\delta)}{n}}$ will attain sub-optimality $F(w^{(t)}) - F(w^*) \leq \epsilon$ after*

$$t = \mathcal{O}\left(\frac{\ell\sqrt{\kappa}}{\sigma n^{1/4}} \log^{1/4}\left(\frac{p}{\delta}\right) \log^2\left(\frac{1}{\epsilon}\right)\right).$$

rounds of inner-loop iterations of DANE-HB.

1. Concerning the domain of interest for D^{2}ANE, let us consider the initialization $w^{(0)} = \arg\min_w F_1(w) + \frac{\gamma}{2}\|w\|^2$ with $\gamma = \mathcal{O}(1/\sqrt{n})$. Then in view of the stability arguments (see, e.g. Zhang and Xiao, 2015, Lemma 5) we can verify that $\mathbb{E}[F(w^{(0)})] \leq F(w^*) + \mathcal{O}(1/\sqrt{n})$ holds under mild conditions. It can be seen from the proof of Theorem 20 that $F(w^{(t)}) \leq F(w^{(t-1)})$ for all $t \geq 1$, and thus $\mathcal{W} := \{w : F(w) \leq F(w^*) + \mathcal{O}(1/\sqrt{n})\}$ is a domain of interest which is expected to be well bounded around w^*, e.g., in view of the strong convexity of F such that $\mathcal{W} \subseteq \left\{w : \|w - w^*\| \leq \mu^{-1/2} n^{-1/4}\right\}$.

Method	Ridge regression	Logistic regression
GIANT	$\mathcal{O}\left(\log\left(\frac{\kappa}{\epsilon}\right)\right)$	✗
DSVRG	$\mathcal{O}\left(\frac{\kappa}{n}\log\left(\frac{1}{\epsilon}\right)+\frac{\kappa^2}{mn}\log^2\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left(\frac{\kappa}{n}\log\left(\frac{1}{\epsilon}\right)+\frac{\kappa^2}{mn}\log^2\left(\frac{1}{\epsilon}\right)\right)$
DiSCO	$\mathcal{O}\left(\frac{\sqrt{\kappa}}{n^{1/4}}\log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left(p^{1/4}\left(\frac{\sqrt{\kappa}}{n^{1/4}}\log\left(\frac{1}{\epsilon}\right)+\frac{\kappa^{3/2}}{n^{3/4}}\right)\right)$
DANE-HB (**ours**)	$\mathcal{O}\left(\frac{\sqrt{\kappa}}{n^{1/4}}\log\left(\frac{1}{\epsilon}\right)\right)$	Local rate: $\mathcal{O}\left(p^{1/4}\frac{\sqrt{\kappa}}{n^{1/4}}\log\left(\frac{1}{\epsilon}\right)\right)$
D^{2}ANE (**ours**)	$\mathcal{O}\left(\frac{\sqrt{\kappa}}{n^{1/4}}\log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left(\frac{\sqrt{\kappa}}{n^{1/4}}\log^2\left(\frac{1}{\epsilon}\right)\right)$

Table 3: Comparison of communication complexity for different distributed learning methods. The x-mark "✗" indicates that the related result was not explicitly reported in the corresponding reference.

Remark 23 *To our best knowledge, this is the first provable near-optimal communication complexity bound of DANE-type methods for non-quadratic loss functions.*

3.5 Comparison against Prior Methods

In Section 1.2 we have highlighted the advantages of our proposed algorithms over several prior DANE-type methods (see Table 1). In this subsection, to further compare our methods against other distributed learning algorithms beyond DANE, we list in Table 3 the amount of communication required by DANE-HB/D^{2}ANE and several representative sample-distributed learning algorithms for solving ridge regression and logistic regression problems. The amount of communication is measured by the number of vectors of size p transmitted among the networked machines. Here we do not count the communication cost spent for distributing data to machines which is required virtually by all the sample-distributed methods. The only exception is DSVRG which, in addition to data allocation, also requires to distribute a random subset of data in order to guarantee unbiased estimation of batch gradient for local optimization. In the following elaboration, we highlight the key messages taken from Table 3.

- *Results for ridge regression.* In this quadratic loss setting, GIANT (Wang et al., 2018) has logarithmic dependence on the condition number κ and hence is superior to the other methods that have polynomial bounds on κ. However, such an improvement of GIANT is only valid in the well-conditioned regime where the sample size N should be sufficiently larger than feature dimension p. In contrast, without needing to assume $N \gg p$, DiSCO (Zhang and Xiao, 2015), DANE-HB and D^{2}ANE require $\mathcal{O}\left(\sqrt{\kappa}n^{-1/4}\log(1/\epsilon)\right)$ rounds of communications with $\mathcal{O}(p)$ bits communicated per round. The amount of communication required by DSVRG (Lee et al., 2017) is $\mathcal{O}\left(\kappa n^{-1}\log(1/\epsilon)+\kappa^2(mn)^{-1}\log^2(1/\epsilon)\right)$ in which the additional term $\kappa^2(mn)^{-1}\log^2(1/\epsilon)$ arises from distributing a multi-set sampled with replacement from the entire data, and it certainly dominates the bound when $\kappa=\Omega(m)$. If this is the case, then DSVRG will be comparable or superior to DiSCO/DANE-HB/D^{2}ANE when $\kappa=\mathcal{O}(n^{1/2}m^{2/3})$, and otherwise the former will be inferior to the latter in communication efficiency.

- *Results for logistic regression.* For general smooth loss functions such as logistic loss, GIANT exhibits linear-quadratic local convergence behavior but without any communication complexity bound explicitly provided. The amount of communication required by DSVRG is still $\mathcal{O}\left(\kappa n^{-1}\log(1/\epsilon)+\kappa^2(mn)^{-1}\log^2(1/\epsilon)\right)$. For DiSCO, the communication complexity becomes $\mathcal{O}\left(p^{1/4}\left(\sqrt{\kappa}/n^{-1/4}\log(1/\epsilon)+\kappa^{3/2}n^{-3/4}\right)\right)$ in which the factor $p^{1/4}$ comes from the uniform concentration analysis of the time varying Hessian matrices. DANE-HB has a slightly improved $\mathcal{O}\left(p^{1/4}\left(\sqrt{\kappa}/n^{-1/4}\log(1/\epsilon)\right)\right)$ bound in a local area around the minimizer. These bounds are inferior to that of DSVRG in high dimensional settings. For D^{2}ANE, the bound is $\mathcal{O}\left(\sqrt{\kappa}n^{-1/4}\log^2(1/\epsilon)\right)$ which has no polynomial dependence on p thanks to the shared Hessian H among the approximate Newton approximation steps. Similar to the previous discussions for the quadratic case, given that $\kappa=\Omega(m)$, DSVRG will be comparable or superior to D^{2}ANE when $\kappa=\mathcal{O}(n^{1/2}m^{2/3})$, and otherwise D^{2}ANE performs better.

To summarize the above discussions, DANE-HB and D^{2}ANE are able to offer competitive or superior communication efficiency to the considered distributed learning algorithms in high-dimensional and ill-conditioned (e.g., $\kappa=\Omega(n^{1/2}m^{2/3})$) problem regimes.

4. Experiments

In this section, we present a numerical study for theory verification and algorithm evaluation. In the theory verification part, we conduct simulations on linear regression and binary logistic regression problems to verify the strong convergence guarantees established for DANE-LS, DANE-HB and D^{2}ANE. Then in the algorithm evaluation part, we run experiments on synthetic and real data binary logistic regression tasks to evaluate the numerical performance of these alternatives with comparison to several state-of-the-art distributed learning methods. We simulate the distributed environment on a single server powered by dual Intel(R) Xeon(R) E5-2630V4@2.2GHz CPU with multiple logic processors simulating multiple machines. All the considered methods are implemented in Matlab R2018b on Microsoft Windows 10. The local subproblems on the master machine are solved by an SVRG solver from SGDLibrary (Kasai, 2017), and the momentum coefficient β in DANE-HB is set according to Theorem 15. We replicate each experiment 10 times over random split of data and report the results in mean-value along with error bar. We initialize $w^{(0)}=0$ throughout our numerical study.

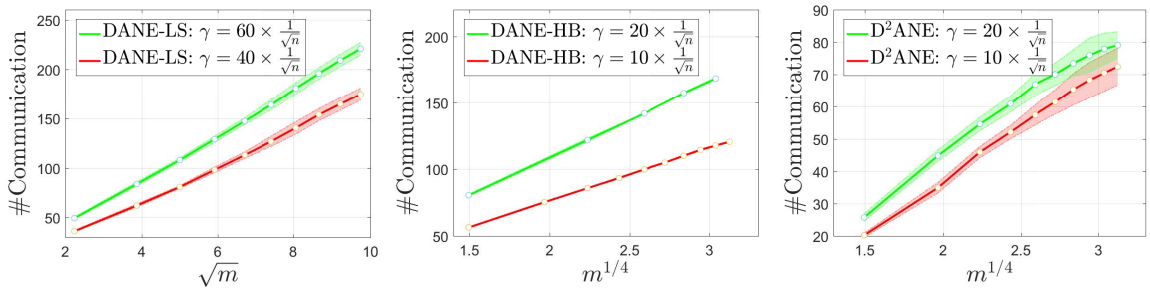
4.1 Theory Verification

The following experimental protocol is considered for theory verification study.

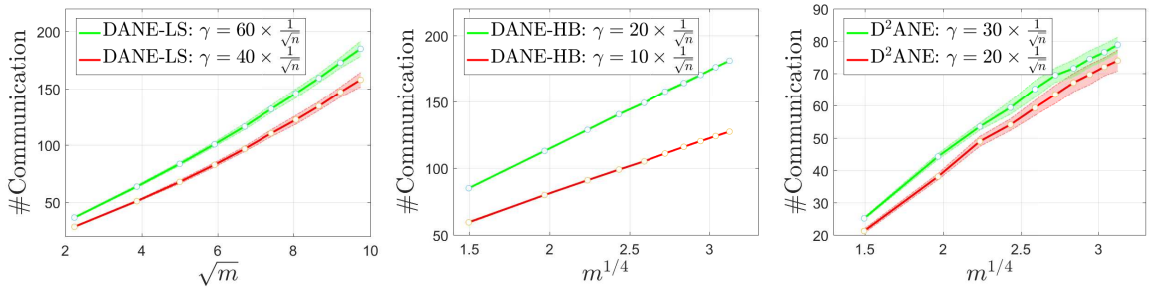
- To verify the bounds established in Theorem 3 for DANE-LS and in Theorem 15 for DANE-HB for quadratic problems, we consider the ridge regression model with loss function $f(w;x_i,y_i)=\frac{1}{2}(w^\top x_i-y_i)^2+\frac{\mu}{2}\|w\|^2$. The feature points $\{x_i\}_{i=1}^N$ are sampled from standard multivariate normal distribution. The responses $\{y_i\}_{i=1}^N$ are generated according to a linear model $y_i=\bar{w}^\top x_i+e_i$ with a random Gaussian vector $\bar{w}\in\mathbb{R}^p$ and random Gaussian noise $e_i\sim\mathcal{N}(0,\sigma^2)$.

- For D^{2}ANE, we verify its communication complexity bounds as in Theorem 20 by applying it to the binary logistic regression model with loss function $f(w; x_i, y_i) = \log\left(1 + \exp(-y_i w^\top x_i)\right) + \frac{\mu}{2}\|w\|^2$. We consider a simulation task in which each data feature x_i is sampled from standard multivariate normal distribution and its label $y_i \in \{-1, +1\}$ is determined according to the conditional probability $\mathbb{P}(y_i|x_i; \bar{w}) = \exp(2y_i \bar{w}^\top x)/(1 + \exp(2y_i \bar{w}^\top x_i))$ with a p-dimensional Gaussian vector $\bar{w}$.

For our simulation study, we test with feature dimensions $p \in \{200, 500\}$. We fix $N = 10p$, $\mu = 1/\sqrt{N}$, and study the impact of varying number of machines m and regularization $\gamma = \mathcal{O}(1/\sqrt{n})$ on the needed rounds of communication to reach sub-optimality $\epsilon = 10^{-6}$. We replicate the experiment 10 times over random split of data.



(a) $p = 200$



(b) $p = 500$

Figure 2: Theory verification: the number of communication rounds (y-axis) versus number of machines (x-axis) curves of DANE-LS (left panels) and DANE-HB (middle panels) on a synthetic ridge regression task, and of D^{2}ANE (right panels) on a synthetic logistic regression task.

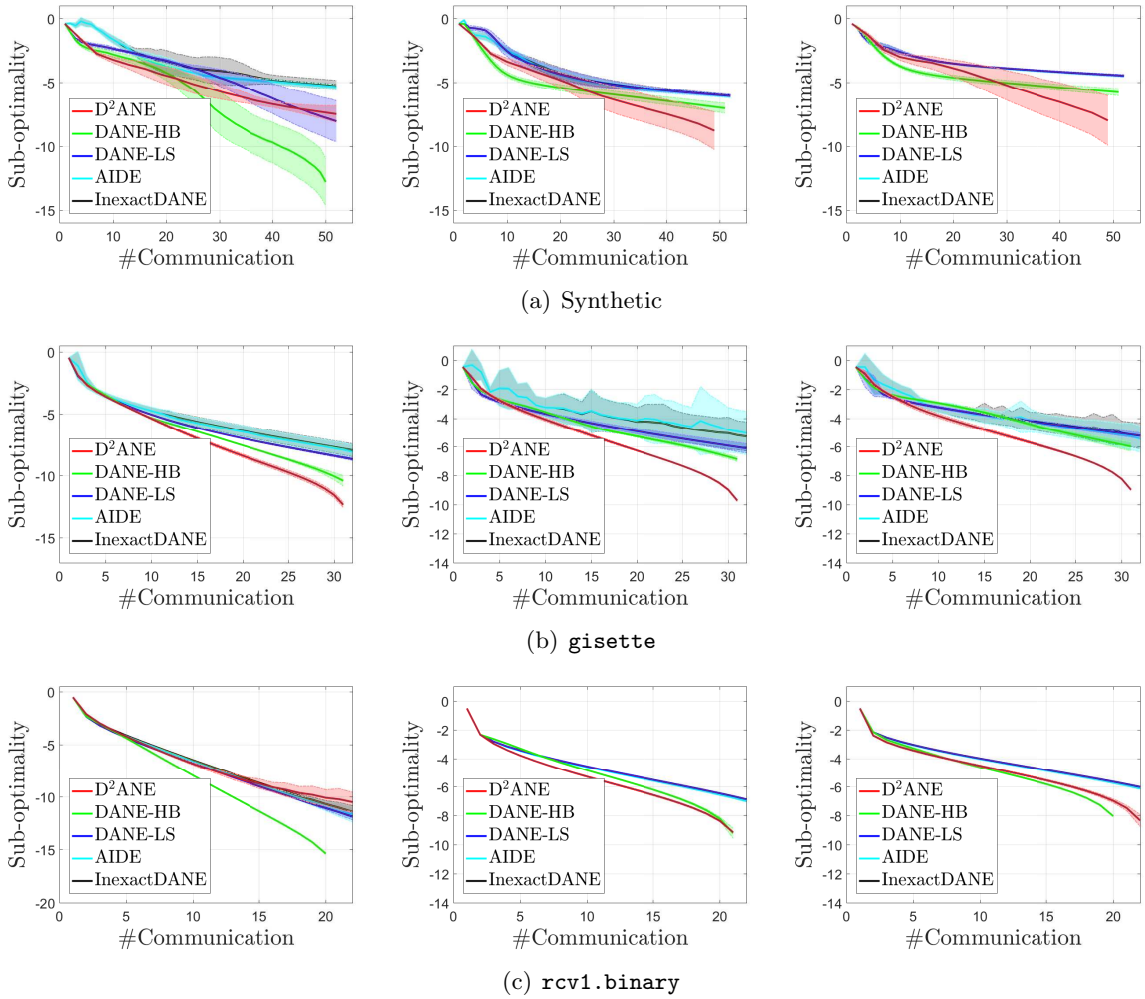
Figure 2 shows the evolving curves (error bar shaded in color) of the needed communication rounds as functions of number of machines achieved by DANE-LS (left panel), DANE-HB (middle panel) and D^{2}ANE (right panel)in the considered setting. Visually speaking, the number of communication rounds scales roughly linearly with respect to $\sqrt{m}$ for DANE-LS and to $m^{1/4}$ for DANE-HB and D^{2}ANE, under varying values of γ. We can also observe that smaller γ leads to fewer rounds of communication. These results are consistent with the theoretical predictions in Theorem 3, Theorem 15 and Theorem 20.

4.2 Algorithm Evaluation

We further compare the convergence performance of our methods with several representative communication-efficient distributed learning methods. For the sake of presentation clarity, we divide the numerical study into two categories using the DANE-type methods and other type of methods as baselines, respectively.

4.2.1 Comparison against DANE-type Methods

In this part, we carry out experiments to compare our methods with InexactDane and AIDE (Reddi et al., 2016), for binary logistic regression problems. We begin with a simulation study using the same data generation protocol as in the previous theory verification study, with $p = 200$, $N = 10p$, $\gamma = 40/\sqrt{n}$, $\mu = 1/\sqrt{N}$ and $m \in \{4, 16, 32\}$.



(a) Synthetic

(b) `gisette`

(c) `rcv1.binary`

Figure 3: Algorithm evaluation with comparison to DANE-type methods: the objective value sub-optimality evolving curves on synthetic and real-data logistic regression tasks with $m = 4$ (left panels), $m = 16$ (middle panels) and $m = 32$ (right panels).

Figure 3(a) shows the sub-optimality (in objective value) convergence curves (w.r.t. communication rounds) of the considered algorithms. From these curves we can see that DANE-LS, DANE-HB and D^{2}ANE are stable in convergence while InexactDane and AIDE exhibit zigzag effect in early iterations when $m = 4, 16$. The convergence instability of the plain DANE method has also been observed in the original work of Shamir et al. (2014). The stability of our proposed methods shows the benefit of line search (for DANE-LS and DANE-HB) and double-loop Newton approximation (for D^{2}ANE) for improving the convergence behavior of DANE-type methods. In terms of communication efficiency, we can see that: i) DANE-LS is superior or comparable to InexactDane and AIDE in optimizing the global objective value after the same rounds of communication; and ii) DANE-HB and D^{2}ANE converge considerably faster than other methods. These observations confirm the effectiveness of heavy-ball approach for accelerating the convergence of DANE.

Next, we evaluate the convergence performance of the considered algorithms on two real data sets `gisette` (Guyon et al., 2005) ($p = 5000$, $N = 6000$) and `rcv1.binary` (Lewis et al., 2004) ($p = 47236$, $N = 20242$). For each data set, we fix the regularization parameter $\mu = 10^{-5}$ and test with $m \in \{4, 16, 32\}$. The results are shown in the middle and bottom rows of Figure 3 from which we have the following observations:

- For `gisette`, Figure 3(b) shows that DANE-LS, DANE-HB and D^{2}ANE converge more stably than InexactDane and AIDE. In terms of communication efficiency, D^{2}ANE outperforms the other considered methods with a clear margin and DANE-HB is the runner-up. DANE-LS converges slightly faster than InexactDane and AIDE when $m = 4, 16$, while these three algorithms are comparable when $m = 32$.
- For `rcv1.binary`, Figure 3(c) shows that all the considered algorithms converge smoothly. In most cases, DANE-HB and D^{2}ANE outperform DANE-LS, InexactDane and AIDE which exhibit very close performance on this data.

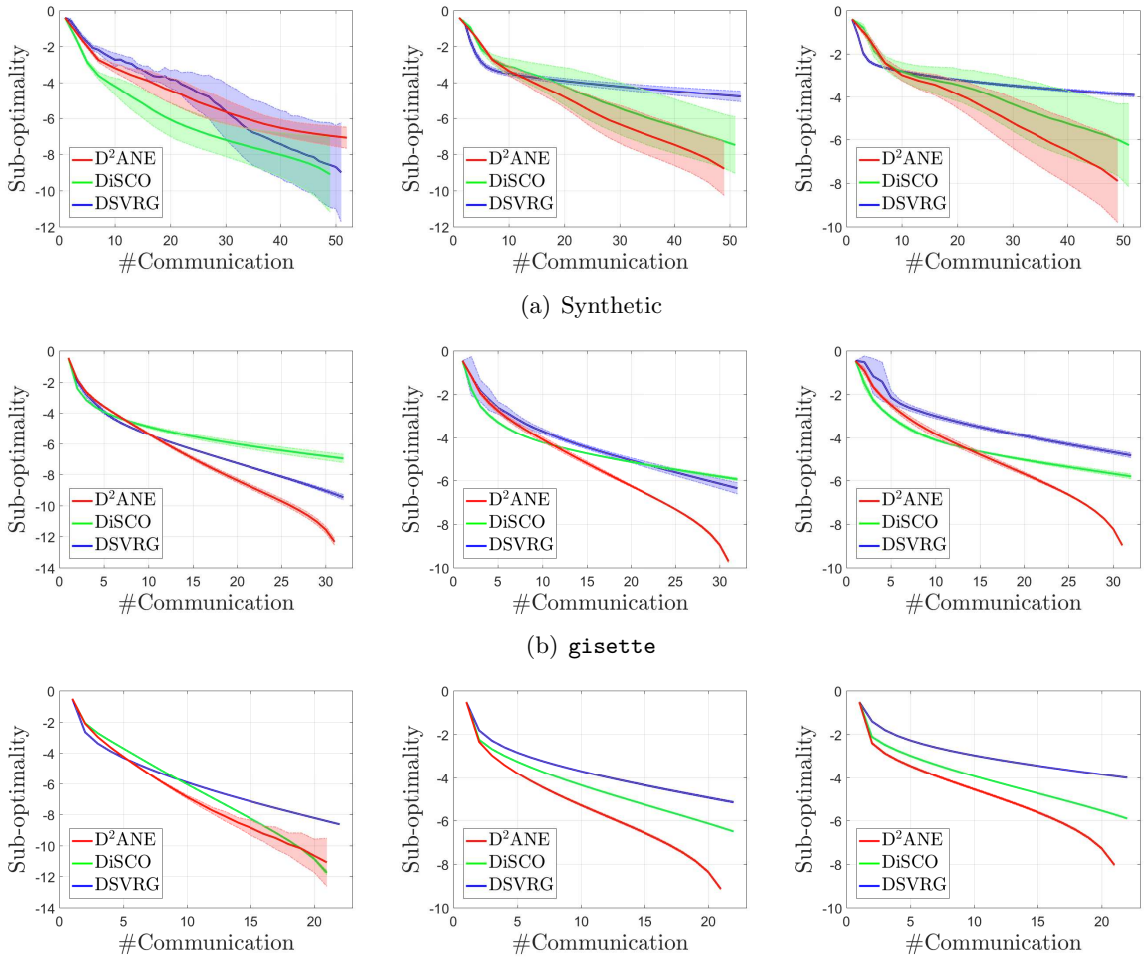
To summarize this set of experiments, our proposed algorithms are stabler than the prior DANE-type methods which matches the global convergence theory established for our algorithms. In many cases, DANE-HB and D^{2}ANE substantially outperform the other methods in communication efficiency.

4.2.2 Comparison Against the Methods beyond DANE

In this group of evaluation, we compare the performance of D^{2}ANE with DSVRG (Lee et al., 2017) and DiSCO (Zhang and Xiao, 2015) which are among others two representative first-order and second-order algorithms for communication-efficient distributed learning. The evaluation is conducted on the same data sets as used in the previous experiment, and the results are shown in Figure 4. Here we omit the results of DANE-LS and DANE-HB in order to avoid redundancy of presentation because in most cases these two methods are inferior or comparable to D^{2}ANE as previously shown.

Below we summarize the main observations that can be made from these results:

- Results on synthetic data: D^{2}ANE $\geq$ DiSCO $\geq$ DSVRG. As shown in Figure 4(a), DiSCO outperforms the other considered algorithms when relatively small $m = 4$ number of machines is used. For relatively large $m = 16, 32$, D^{2}ANE and DiSCO converge faster than DSVRG.



(a) Synthetic

(b) `gisette`

(c) `rcv1.binary`

Figure 4: Algorithm evaluation with comparison to the distributed optimization methods beyond DANE: the objective value sub-optimality evolving curves on synthetic and real logistic regression tasks with $m = 4$ (left panels), $m = 16$ (middle panels) and $m = 32$ (right panels).

- Results on `gisette`: D^{2}ANE $>$ DiSCO $\geq$ DSVRG. From the curves in Figure 4(b) we can see that D^{2}ANE outperforms DiSCO and DSVRG with a clear margin.

- Results on `rcv1.binary`: D^{2}ANE $>$ DiSCO $>$ DSVRG. Figure 4(c) shows that D^{2}ANE significantly outperforms DiSCO and DSVRG especially for relatively large $m = 16, 32$.

Overall, D^{2}ANE performs the best in communication efficiency among all the considered algorithms. DiSCO is found to be competitive or superior to DSVRG in many places.

5. Conclusions

In this article, we have made progress towards deeply understanding the mysterious convergence behavior of the popularly applied DANE method for communication-efficient distributed convex optimization. To this end, we propose two new alternatives, DANE-LS and DANE-HB, which are more suitable for global asymptotic and local non-asymptotic analysis, and also effective for momentum acceleration. The core messages conveyed by our study include:

(1) *The plain DANE method can actually converge faster than already known.* For quadratic problems, even without any momentum acceleration, DANE-LS attains a tighter communication complexity bound than what already revealed for plain DANE.

(2) *Line search is beneficial to DANE.* For non-quadratic strongly convex functions, blessed by the backtracking line search under Armijo rule, DANE-LS converges globally under a wider spectrum of γ than DANE, at an appealing local rate of convergence.

(3) *Heavy-ball acceleration is effective for DANE.* DANE-HB possesses a near-tight communication complexity bound for quadratic functions. Whilst for non-quadratic convex functions, DANE-HB exhibits an identical performance in the vicinity of minimizer. For convex optimization with linear models, we proposed the D^{2}ANE method as a double-loop approximate Newton extension of DANE-HB that has been shown to have global convergence with near-tight communication complexity bounds.

Numerical results support our theoretical findings and confirm that DANE-LS, DANE-HB and D^{2}ANE are safe and in many places more attractive alternatives to the prior DANE-type methods for both quadratic and non-quadratic convex optimization problems. We expect that the theory and algorithms developed in this article will fuel future investigation on distributed non-convex optimization problems such as distributed training of deep neural nets across multiple machines or GPUs. Also, we believe our improved DANE-type methods should have practical implications in large-scale federated optimization for privacy-preserving collaborative machine learning.

Acknowledgements

The authors sincerely thank the anonymous reviewers for providing constructive comments that led to a substantial improvement of this article. Xiao-Tong Yuan would also like to acknowledge the support from National Major Project of China for New Generation of AI under Grant No.2018AAA0100400 and Natural Science Foundation of China (NSFC) under Grant No.61876090 and No.61936005.

Appendix A. Some Auxiliary Lemmas

Here we introduce a set of auxiliary lemmas which will be used for proving the main results in the article. For the sake of readability, we defer the relevant proofs into Appendix D. The following elementary lemmas will be used frequently throughout our analysis.

Lemma 24 *Let A and B be two symmetric and positive definite matrices and $B \succeq \mu I$ for some $\mu > 0$. If $\|A - B\| \leq \gamma$, then $(A+\gamma I)^{-1}B$ is diagonalizable and*

$$\lambda_{\max}(A+\gamma I)^{-1}B \leq 1, \quad \lambda_{\min}((A+\gamma I)^{-1}B) \geq \frac{\mu}{\mu+2\gamma}.$$

Moreover, the following spectral norm bounds hold:

$$\|I - (A+\gamma I)^{-1/2}B(A+\gamma I)^{-1/2}\| \leq \frac{2\gamma}{\mu+2\gamma}\ ,\ \|I - B^{1/2}(A+\gamma I)^{-1}B^{1/2}\| \leq \frac{2\gamma}{\mu+2\gamma}.$$

Let us denote $\rho(A)$ the spectral radius of A, i.e., the largest (in modulus) eigenvalue of a square matrix A.

Lemma 25 *Let $A \in \mathbb{R}^{d\times d}$ be a square matrix with positive real eigenvalues such that $0 < \mu \leq \lambda_{\min}(A) \leq \lambda_{\max}(A) \leq L$. Assume that A is diagonalizable. Then*

$$\rho\left(\begin{bmatrix} (1+\beta)I - \eta A & -\beta I \\ I & 0 \end{bmatrix}\right) \leq \max\{|1-\sqrt{\eta\mu}|, |1-\sqrt{\eta L}|\},$$

where $\beta = \max\{|1-\sqrt{\eta\mu}|, |1-\sqrt{\eta L}|\}^2$.

An important relationship between the spectral norm $\|A\|$ and spectral radius $\rho(A)$ is given by the equality $\rho(A) = \lim_{t\to\infty}\|A^t\|^{1/t}$, which implies the following classic lemma.

Lemma 26 *For $\lim_{t\to\infty} A^t = 0$ it is necessary and sufficient that $\rho(A) < 1$ and for every $\delta > 0$ there exists a constant $c = c(\delta)$ such that*

$$\|A^t\| \leq c(\rho(A)+\delta)^t$$

for all positive integers t.

The following lemma is elementary and will be used in many places of our analysis.

Lemma 27 *Assume that function g has ν-LH. Then*

$$\left\|\Delta g(w, w')\right\| \leq \frac{\nu}{2}\|w - w'\|^2,$$

where $\Delta g(w, w') := \nabla g(w) - \nabla g(w') - \nabla^2 g(w')(w - w')$.

The next lemma, which is based on a matrix concentration bound (Tropp, 2012), shows that the Hessian of $F_1(w)$ is close to that of $F(w)$ when the sample size is sufficiently large. The same result appears in the work of Shamir et al. (2014).

Lemma 28 *Assume that $\|\nabla^2 f(w^\top x_i, y_i)\| \leq L$ holds for all $i \in [N]$. Let $H(w) = \nabla^2 F(w)$ and $H_1(w) = \nabla^2 F_1(w)$. Then for each fixed w, with probability at least $1-\delta$ over the samples drawn to construct $F_1(w)$, the following bound holds:*

$$\|H_1(w) - H(w)\| \leq \sqrt{\frac{32L^2\log(p/\delta)}{n}}.$$

Appendix B. Proofs for Section 2

We collect in this appendix section the technical proofs of the results in Section 2, including Theorems 3, Theorem 8 and Theorem 13, and their corollaries.

B.1 Proof of Theorem 3

In this appendix subsection, we prove Theorem 3 as restated in below.

Theorem 3 (Convergence rate of DANE-LS for quadratic loss) *Assume that the loss function is quadratic. Let H and H_1 be the Hessian matrices of the global objective F and local objective F_1, respectively. Assume that $\mu I \preceq H \preceq LI$. Given precision $\epsilon > 0$, if $\|H_1 - H\| \leq \gamma$ and $\varepsilon_t \leq \frac{\mu^2 \|\nabla F(w^{(t-1)})\|}{2(\mu+2\gamma)L}$, then Algorithm 1 will output $w^{(t)}$ satisfying $\|w^{(t)} - w^*\| \leq \epsilon$ after*

$$t \geq \frac{2(\mu+2\gamma)}{\mu} \log\left(\frac{\sqrt{\kappa}\|w^{(0)} - w^*\|}{\epsilon}\right)$$

rounds of iterations.

Proof Since the objective is quadratic, for any $w^{(t-1)}$ the optimal solution $w^* = \arg\min_w F(w)$ can always be expressed as

$$w^* = w^{(t-1)} - H^{-1}\nabla F(w^{(t-1)}).$$

Since $H_1^{(t)} \equiv H_1$ holds in the quadratic case, the gradient equation of $P^{(t-1)}$ at $w^{(t)}$ implies

$$w^{(t)} = w^{(t-1)} - (H_1+\gamma I)^{-1}\nabla F(w^{(t-1)}) + (H_1+\gamma I)^{-1}\nabla P^{(t-1)}(w^{(t)}).$$

Combining the above two equalities yields

$$w^{(t)} - w^* = (I - \eta(H_1+\gamma I)^{-1}H)(w^{(t-1)} - w^*) + (H_1+\gamma I)^{-1}\nabla P^{(t-1)}(w^{(t)}).$$

By multiplying $H^{1/2}$ on both sides of the above recurrent form we have

$$H^{1/2}(w^{(t)}-w^*) = (I-H^{1/2}(H_1+\gamma I)^{-1}H^{1/2})H^{1/2}(w^{(t-1)}-w^*)+H^{1/2}(H_1+\gamma I)^{-1}\nabla P^{(t-1)}(w^{(t)})$$

Let $u^{(t)} = H^{1/2}(w^{(t)} - w^*)$. Based on the basic inequality $\|Tx\| \leq \|T\|\|x\|$ we obtain

$$\begin{aligned}
&\|u^{(t)}\| \\
\leq&\|I - H^{1/2}(H_1+\gamma I)^{-1}H^{1/2}\|\|u^{(t-1)}\| + \|H^{1/2}(H_1+\gamma I)^{-1}H^{1/2}\|\|H^{-1/2}\nabla P^{(t-1)}(w^{(t)})\| \\
\overset{\zeta_1}{\leq}&\frac{2\gamma}{\mu+2\gamma}\|u^{(t-1)}\| + \frac{\varepsilon_t}{\sqrt{\mu}} \\
\overset{\zeta_2}{\leq}&\left(1 - \frac{\mu}{\mu+2\gamma}\right)\|u^{(t-1)}\| + \frac{\mu}{2(\mu+2\gamma)}\|u^{(t-1)}\| \\
=&\left(1 - \frac{\mu}{2(\mu+2\gamma)}\right)\|u^{(t-1)}\|,
\end{aligned}$$

where in the inequality "ζ_1" we have used Lemma 24 and $\|H^{1/2}(H_1+\gamma I)^{-1}H^{1/2}\| \leq 1$ which are valid in view of $\|H_1 - H\| \leq \gamma$ and $H \succeq \mu I$, "ζ_2" follows from the condition $\varepsilon_t \leq \frac{\mu^2\|\nabla F(w^{(t-1)})\|}{2(\mu+2\gamma)L}$ which implies $\frac{\varepsilon_t}{\sqrt{\mu}} \leq \frac{\mu\sqrt{\mu}\|w^{(t-1)}-w^*\|}{2(\mu+2\gamma)} \leq \frac{\mu\|u^{(t-1)}\|}{2(\mu+2\gamma)}$. The above inequality directly leads to

$$\begin{aligned}
&\|w^{(t)} - w^*\| \\
\leq &\frac{1}{\sqrt{\mu}}\|u^{(t)}\| \leq \frac{1}{\sqrt{\mu}}\left(1 - \frac{\mu}{2(\mu+2\gamma)}\right)^t \|u^{(0)}\| \\
\leq &\sqrt{\frac{L}{\mu}}\left(1 - \frac{\mu}{2(\mu+2\gamma)}\right)^t \|w^{(0)} - w^*\|.
\end{aligned}$$

Since $(1-x)^t \leq \exp\{-xt\}$, we can show from the above that $\|w^{(t)} - w^*\| \leq \epsilon$ is valid when

$$t \geq \frac{2(\mu+2\gamma)}{\mu}\log\left(\frac{\sqrt{L}\|w^{(0)} - w^*\|}{\sqrt{\mu}\epsilon}\right).$$

This concludes the proof. ■

Further, we prove Corollary 4 as restated in below.

Corollary 4 *Assume the conditions in Theorem 3 hold and $\|\nabla^2 f(w; x_i, y_i)\| \leq L$ for all $i \in [N]$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the samples drawn to construct F_1, Algorithm 1 with $\gamma = L\sqrt{\frac{32\log(p/\delta)}{n}}$ will output $w^{(t)}$ satisfying $\|w^{(t)} - w^*\| \leq \epsilon$ after*

$$t \geq \left(2 + 4\kappa\sqrt{\frac{32\log(p/\delta)}{n}}\right)\log\left(\frac{\sqrt{\kappa}\|w^{(0)} - w^*\|}{\epsilon}\right)$$

rounds of iterations.

Proof Since $H(w) \equiv H$ and $H_1(w) \equiv H_1$ in the quadratic case, we know from Lemma 28 that $\|H_1 - H\| \leq \gamma = L\sqrt{\frac{32\log(p/\delta)}{n}}$ holds with probability at least $1-\delta$. By invoking Theorem 3 we obtain the desired bound. ■

B.2 Proof of Theorem 8

We provide in this appendix subsection a detailed proof of Theorem 8 as restated below.

Theorem 8 (Global convergence of DANE-LS) *Assume that $F(w)$ and $F_1(w)$ are L-smooth, μ-strongly-convex and have ν-LH. Suppose that $\varepsilon_t \leq \frac{\rho(\mu+\gamma)}{2(L+\gamma)+\rho(\mu+\gamma)}\|\nabla F(w^{(t-1)})\|$.*

(a) Then the objective value sequence $\{F(w^{(t)})\}$ generated by Algorithm 1 with the global line search step (Option-I) converges and the difference norm sequence $\{\|\tilde{w}^{(t)} - w^{(t-1)}\|\}$ converges to zero.

(b) Assume in addition that $\sup_w \|\nabla^2 F_1(w) - \nabla^2 F(w)\| \leq \gamma$ *and* $\|\tilde{w}^{(t)} - w^{(t-1)}\|$ *is bounded from above for all* $t \geq 0$. *Then the objective value sequence* $\{F(w^{(t)})\}$ *generated by Algorithm 1 with the local line search step (Option-II) converges and the difference norm sequence* $\{\|\tilde{w}^{(t)} - w^{(t-1)}\|\}$ *converges to zero.*

As a key step, we first need to prove the following restated Lemma 6.

Lemma 6 (Feasibility of line search) *Assume that* F *is* L*-smooth and* F_1 *is* μ*-strongly convex. For any given* $\rho \in (0, 1)$,

(a) if the length of search satisfies

$$0 < \eta_t \leq \min\left\{1, \frac{2(\gamma+\mu)(1-\rho)}{L}\right\},$$

then the global backtracking line search (Option-I) is feasible, i.e.,

$$F(w^{(t)}) \leq F(w^{(t-1)}) - \psi(\tilde{w}^{(t)}, w^{(t-1)}),$$

where $\psi(\tilde{w}^{(t)}, w^{(t-1)}) := \eta_t \rho \langle \nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)} \rangle - \eta_t \varepsilon_t \|\tilde{w}^{(t)} - w^{(t-1)}\|$.

(b) Moreover, assume that $F_1(w)$ *has* ν*-LH and* $\exists D > 0$ *such that* $\|\tilde{w}^{(t)} - w^{(t-1)}\| \leq D$ *for all* $t \geq 0$. *If*

$$\eta_t \leq \min\left\{1, \frac{-(3\nu D + 6(\gamma+\mu)) + \sqrt{(3\nu D + 6(\gamma+\mu))^2 + 96(1-\rho)\nu D(\gamma+\mu)}}{4\nu D}\right\},$$

then the local backtracking line search (Option-II) is feasible, i.e.,

$$\langle \nabla F(w^{(t-1)}), w^{(t)} - w^{(t-1)} \rangle + \frac{1}{2}(w^{(t)} - w^{(t-1)})^\top (\nabla^2 F_1(w^{(t-1)}) + \gamma I)(w^{(t)} - w^{(t-1)}) + \frac{\nu}{6}\|w^{(t)} - w^{(t-1)}\|^3 \leq -\psi(\tilde{w}^{(t)}, w^{(t-1)}).$$

Proof Let us define

$$r^{(t)} = \nabla P^{(t-1)}(\tilde{w}^{(t)}) = \nabla F_1(\tilde{w}^{(t)}) + \nabla F(w^{(t-1)}) - \nabla F_1(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}). \quad \text{(A.1)}$$

From the definition of $\tilde{w}^{(t)}$ we have that $\|r^{(t)}\| \leq \varepsilon_t$. Since $F(w)$ is L-smooth, we have

$$\begin{aligned}
&F(w^{(t)}) \\
\leq& F(w^{(t-1)}) + \langle \nabla F(w^{(t-1)}), w^{(t)} - w^{(t-1)} \rangle + \frac{L}{2}\|w^{(t)} - w^{(t-1)}\|^2 \\
=& F(w^{(t-1)}) + \eta_t \langle \nabla F(w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)} \rangle + \frac{L\eta_t^2}{2}\|\tilde{w}^{(t)} - w^{(t-1)}\|^2 \\
\overset{\zeta_1}{\leq}& F(w^{(t-1)}) - \eta_t \langle \nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)} \rangle \\
&+ \eta_t \langle r^{(t)}, \tilde{w}^{(t)} - w^{(t-1)} \rangle + \frac{L\eta_t^2}{2}\|\tilde{w}^{(t)} - w^{(t-1)}\|^2 \\
\overset{\zeta_2}{\leq}& F(w^{(t-1)}) - \left(\eta_t - \frac{L\eta_t^2}{2(\gamma+\mu)}\right) \langle \nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)} \rangle \\
&+ \eta_t \varepsilon_t \|\tilde{w}^{(t)} - w^{(t-1)}\|,
\end{aligned}$$

where "ζ_1" follows from (A.1) and "ζ_2" is due to the μ-strong-convexity of F_1 which implies $\langle \nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)} \rangle \geq (\mu+\gamma)\|\tilde{w}^{(t)} - w^{(t-1)}\|^2$. To make a successful global line search, we simply require $-\left(\eta_t - \frac{L\eta_t^2}{2(\gamma+\mu)}\right) \leq -\eta_t \rho$, which obviously can be guaranteed by setting

$$0 < \eta_t \leq \min\left\{1, \frac{2(\gamma+\mu)(1-\rho)}{L}\right\}.$$

This prove the result in Part(a).

To prove the result in Part(b), we first note that the equality (A.1) is identical to

$$\nabla F(w^{(t-1)}) = -(\nabla^2 F_1(w^{(t-1)}) + \gamma I)(\tilde{w}^{(t)} - w^{(t-1)}) - \Delta F_1(\tilde{w}^{(t)}, w^{(t-1)}) + r^{(t)}. \tag{A.2}$$

Then based on the definition of $w^{(t)}$ we can derive that

$$\begin{aligned}
&\langle \nabla F(w^{(t-1)}), w^{(t)} - w^{(t-1)} \rangle + \frac{1}{2}(w^{(t)} - w^{(t-1)})^\top (\nabla^2 F_1(w^{(t-1)}) + \gamma I)(w^{(t)} - w^{(t-1)}) \\
&\quad + \frac{\nu}{6}\|w^{(t)} - w^{(t-1)}\|^3 \\
=&\eta_t \langle \nabla F(w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)} \rangle + \frac{\eta_t^2}{2}(\tilde{w}^{(t)} - w^{(t-1)})^\top (\nabla^2 F_1(w^{(t-1)}) + \gamma I)(\tilde{w}^{(t)} - w^{(t-1)}) \\
&\quad + \frac{\nu\eta_t^3}{6}\|\tilde{w}^{(t)} - w^{(t-1)}\|^3 \\
\overset{\zeta_1}{=}&\eta_t \langle \nabla F(w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)} \rangle - \frac{\eta_t^2}{2}\langle \nabla F(w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)} \rangle + \frac{\nu\eta_t^3}{6}\|\tilde{w}^{(t)} - w^{(t-1)}\|^3 \\
&\quad - \frac{\eta_t^2}{2}\langle \Delta F_1(\tilde{w}^{(t)}, w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)} \rangle + \frac{\eta_t^2}{2}\langle r^{(t)}, \tilde{w}^{(t)} - w^{(t-1)} \rangle \\
\overset{\zeta_2}{\leq}&\left(\eta_t - \frac{\eta_t^2}{2}\right)\langle \nabla F(w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)} \rangle + \left(\frac{\nu\eta_t^2}{4} + \frac{\nu\eta_t^3}{6}\right)\|\tilde{w}^{(t)} - w^{(t-1)}\|^3 \\
&\quad + \frac{\eta_t^2}{2}\langle r^{(t)}, \tilde{w}^{(t)} - w^{(t-1)} \rangle \\
\overset{\zeta_3}{=}&-\left(\eta_t - \frac{\eta_t^2}{2}\right)\langle \nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)} \rangle \\
&\quad + \left(\frac{\nu\eta_t^2}{4} + \frac{\nu\eta_t^3}{6}\right)\|\tilde{w}^{(t)} - w^{(t-1)}\|^3 + \eta_t\langle r^{(t)}, \tilde{w}^{(t)} - w^{(t-1)} \rangle \\
\leq&-\left(\eta_t - \frac{\eta_t^2}{2}\right)\langle \nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)} \rangle \\
&\quad + \left(\frac{\nu\eta_t^2}{4} + \frac{\nu\eta_t^3}{6}\right)D\|\tilde{w}^{(t)} - w^{(t-1)}\|^2 + \eta_t\varepsilon_t\|\tilde{w}^{(t)} - w^{(t-1)}\| \\
\overset{\zeta_4}{\leq}&\left(-\left(\eta_t - \frac{\eta_t^2}{2}\right) + \left(\frac{\nu\eta_t^2}{4} + \frac{\nu\eta_t^3}{6}\right)\frac{D}{\gamma+\mu}\right)\times \\
&\langle \nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)} \rangle + \eta_t\varepsilon_t\|\tilde{w}^{(t)} - w^{(t-1)}\|,
\end{aligned}$$

where "ζ_1" follows from (A.2), "ζ_2" uses $\|\Delta\tilde{F}(w^{(t-1)}, \tilde{w}^{(t)})\| \leq \frac{\nu}{2}\|\tilde{w}^{(t)} - w^{(t-1)}\|^2$, "$\zeta_3$" follows from (A.1) and "ζ_4" is due to the μ-strong-convexity of F_1 which implies $\langle \nabla F_1(\tilde{w}^{(t)}) -$

$\nabla F_1(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)}\rangle \geq (\mu+\gamma)\|\tilde{w}^{(t)} - w^{(t-1)}\|^2$. In order to make a successful line search, it suffices to set

$$-\left(\eta_t - \frac{\eta_t^2}{2}\right) + \left(\frac{\nu\eta_t^2}{4} + \frac{\nu\eta_t^3}{6}\right)\frac{D}{\gamma+\mu} \leq -\eta_t\rho$$

which indeed can be guaranteed by setting

$$0 < \eta_t \leq \min\left\{1, \frac{-(3\nu D + 6(\gamma+\mu)) + \sqrt{(3\nu D + 6(\gamma+\mu))^2 + 96(1-\rho)\nu D(\gamma+\mu)}}{4\nu D}\right\}.$$

The proof of Part(b) is concluded. ■

We also need the following lemma which bounds the values of $\|\nabla F(\tilde{w}^{(t)})\|$ and $\|\tilde{w}^{(t)} - w^*\|$ with respect to the distance $\|\tilde{w}^{(t)} - w^{(t-1)}\|$ and sub-optimality ε_t.

Lemma 29 *Assume that F and F_1 have Lipschitz continuous Hessian. If $\sup_w \|\nabla^2 F_1(w) - \nabla^2 F(w)\| \leq \gamma$, then at any time instant t it is true that*

$$\|\nabla F(\tilde{w}^{(t)})\| \leq 2\gamma\|\tilde{w}^{(t)} - w^{(t-1)}\| + \varepsilon_t, \quad \|\tilde{w}^{(t)} - w^*\| \leq \frac{2\gamma}{\mu}\|\tilde{w}^{(t)} - w^{(t-1)}\| + \frac{\varepsilon_t}{\mu}.$$

Proof From the local sub-optimality condition we have

$$\|\nabla P^{(t-1)}(\tilde{w}^{(t)})\| = \|\nabla F_1(\tilde{w}^{(t)}) + \nabla F(w^{(t-1)}) - \nabla F_1(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)})\| \leq \varepsilon_t.$$

Then

$$\begin{aligned}
&\|\nabla F(\tilde{w}^{(t)})\| \\
=&\|\nabla F(\tilde{w}^{(t)}) - \nabla P^{(t-1)}(\tilde{w}^{(t)}) + \nabla P^{(t-1)}(\tilde{w}^{(t)})\| \\
\leq&\|\nabla F(\tilde{w}^{(t)}) - \nabla F_1(\tilde{w}^{(t)}) - \nabla F(w^{(t-1)}) + \nabla F_1(w^{(t-1)}) - \gamma(\tilde{w}^{(t)} - w^{(t-1)})\| + \varepsilon_t \\
=&\|(\nabla^2(F - F_1)(w') + \gamma I)(\tilde{w}^{(t)} - w^{(t-1)})\| + \varepsilon_t \\
\leq&2\gamma\|\tilde{w}^{(t)} - w^{(t-1)}\| + \varepsilon_t,
\end{aligned}$$

where in the last inequality we have used $\sup_w \|\nabla^2 F(w) - \nabla^2 F_1(w)\| \leq \gamma$. This proves the first inequality. The second inequality follows readily from the strong convexity of F such that $\mu\|\tilde{w}^{(t)} - w^*\| \leq \|\nabla F(\tilde{w}^{(t)}) - \nabla F(w^*)\| = \|\nabla F(\tilde{w}^{(t)})\|$. ■

Now we are in the position to prove the main result in Theorem 8.

Proof [of Theorem 8] Part (a): We first prove the convergence of the objective value sequence. Based on (A.1), the smoothness of F_1 and the condition $\varepsilon_t \leq \frac{\rho(\mu+\gamma)}{2(L+\gamma)+\rho(\mu+\gamma)}\|\nabla F(w^{(t-1)})\|$ we can show that

$$\begin{aligned}
\varepsilon_t \geq&\|r_t\| \\
\geq&\|\nabla F(w^{(t-1)})\| - \|\nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)})\| \\
\geq&\left(\frac{2(L+\gamma)}{\rho(\mu+\gamma)} + 1\right)\varepsilon_t - (L+\gamma)\|\tilde{w}^{(t)} - w^{(t-1)}\|,
\end{aligned}$$

which then implies

$$\varepsilon_t \leq \frac{\rho(\gamma+\mu)}{2}\|\tilde{w}^{(t)} - w^{(t-1)}\|. \tag{A.3}$$

Since $F(w)$ is L-smooth and $F_1(w)$ is μ-strongly convex, from the first part of Lemma 6 we know that the global line search is feasible at each step of iteration and thus

$$\begin{aligned}
&F(w^{(t)}) \\
\leq& F(w^{(t-1)}) - \eta_t\rho\langle\nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)}\rangle \\
&+ \eta_t\varepsilon_t\|\tilde{w}^{(t)} - w^{(t-1)}\| \\
\overset{\zeta_1}{\leq}& F(w^{(t-1)}) - \eta_t\rho(\gamma+\mu)\|\tilde{w}^{(t)} - w^{(t-1)}\|^2 + \frac{\eta_t\rho(\gamma+\mu)}{2}\|\tilde{w}^{(t)} - w^{(t-1)}\|^2 \\
=& F(w^{(t-1)}) - \frac{\eta_t\rho(\gamma+\mu)}{2}\|\tilde{w}^{(t)} - w^{(t-1)}\|^2,
\end{aligned}$$

where in "ζ_1" we have used the bound (A.3). From Lemma 29 we know that $\|\tilde{w}^{(t)} - w^{(t-1)}\| \neq 0$ unless $\tilde{w}^{(t)}$ admits a global minimizer of F. Then based on the above inequality the sequence $\{F(w^{(t)})\}$ is decreasing. Since $F(w^{(t)}) \geq F(w^*) > -\infty$, it must hold that $\{F(w^{(t)})\}$ converges. Also from the above inequality we have

$$\eta_t\rho(\gamma+\mu)\|\tilde{w}^{(t)} - w^{(t-1)}\|^2 \leq 2(F(w^{(t-1)}) - F(w^{(t)})),$$

which implies $\|\tilde{w}^{(t)} - w^{(t-1)}\| \to 0$ as $t \to \infty$.

Proof of part(b): Since $F(w)$ has ν-smooth, we have

$$\begin{aligned}
&F(w^{(t)}) \\
\leq& F(w^{(t-1)}) + \langle\nabla F(w^{(t-1)}), w^{(t)} - w^{(t-1)}\rangle + \frac{1}{2}(w^{(t)} - w^{(t-1)})^\top\nabla^2 F(w^{(t-1)})(w^{(t)} - w^{(t-1)}) \\
&+ \frac{\nu}{6}\|w^{(t)} - w^{(t-1)}\|^3 \\
\overset{\zeta_1}{\leq}& F(w^{(t-1)}) + \langle\nabla F(w^{(t-1)}), w^{(t)} - w^{(t-1)}\rangle \\
&+ \frac{1}{2}(w^{(t)} - w^{(t-1)})^\top(\nabla^2 F_1(w^{(t-1)}) + \gamma I)(w^{(t)} - w^{(t-1)}) + \frac{\nu}{6}\|w^{(t)} - w^{(t-1)}\|^3 \\
\overset{\zeta_2}{\leq}& F(w^{(t-1)}) - \eta_t\rho\langle\nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)}\rangle \\
&+ \eta_t\varepsilon_t\|\tilde{w}^{(t)} - w^{(t-1)}\| \\
\overset{\zeta_3}{\leq}& F(w^{(t-1)}) - \eta_t\rho(\gamma+\mu)\|\tilde{w}^{(t)} - w^{(t-1)}\|^2 + \frac{\eta_t\rho(\gamma+\mu)}{2}\|\tilde{w}^{(t)} - w^{(t-1)}\|^2 \\
=& F(w^{(t-1)}) - \frac{\eta_t\rho(\gamma+\mu)}{2}\|\tilde{w}^{(t)} - w^{(t-1)}\|^2,
\end{aligned}$$

where "ζ_1" follows from $\|\nabla^2 F_1(w^{(t-1)}) - \nabla^2 F(w^{(t-1)})\| \leq \gamma$ such that $\nabla^2 F_1(w^{(t-1)}) - \nabla^2 F(w^{(t-1)}) + \gamma I \succeq 0$, in "$\zeta_2$" we have used the second part of Lemma 6, and "ζ_3" is due to the bound (A.3). Based on an identical argument to that of part(a) we can show that the sequence $\{F(w^{(t)})\}$ converges and $\|\tilde{w}^{(t)} - w^{(t-1)}\| \to 0$ as $t \to \infty$. This completes the proof. ■

B.3 Proof of Theorem 13

This appendix subsection is devoted to providing a detailed proof of Theorem 13 as restated in below.

Theorem 13 (Non-asymptotic convergence of DANE-LS) *Assume that F and F_1 are μ-strongly-convex, L-smooth and have ν-LH. Assume that $\sup_w \|\nabla^2 F_1(w) - \nabla^2 F(w)\| \leq \gamma$. Suppose that $\rho \in (0, 1/3]$ and*

$$\varepsilon_t \leq \min\left\{(\gamma+\mu)^2, \frac{\|\nabla F(w^{(t-1)})\|^2}{L^2}, \frac{\rho(\mu+\gamma)}{2(L+\gamma)+\rho(\mu+\gamma)}\|\nabla F(w^{(t-1)})\|\right\}.$$

Let $\tau = \left\lceil \frac{\mu+2\gamma}{2\mu}\log(4\kappa)\right\rceil$. Then there exists a time stamp t_0, which is invariant to ϵ, such that Algorithm 1 will output solution $w^{(t)}$ satisfying $\|w^{(t)} - w^*\| \leq \epsilon$ after

$$t \geq t_0 + 4\tau \log\left(\frac{\gamma+\mu}{4(6\nu+1)\sqrt{\kappa}\tau}\left(\frac{1}{\epsilon}\right)\right)$$

rounds of iterations.

To prove the theorem, we first need to prove the following restated Lemma 10.

Lemma 10 (Acceptability of unit length for line search) *Assume that the conditions in Theorem 8 hold. Then for any fixed $\rho \in (0, 1/3]$, the unit length $\eta_t = 1$ guarantees the sufficient descent condition (5) provided that t is sufficiently large.*

Proof Since $F(w)$ has ν-LH, it holds that

$$\begin{aligned}
&F(w^{(t)})\\
\leq&F(w^{(t-1)}) + \langle \nabla F(w^{(t-1)}), w^{(t)} - w^{(t-1)}\rangle + \frac{1}{2}(w^{(t)} - w^{(t-1)})^\top \nabla^2 F(w^{(t-1)})(w^{(t)} - w^{(t-1)})\\
&+ \frac{\nu}{6}\|w^{(t)} - w^{(t-1)}\|^3\\
\leq&F(w^{(t-1)}) + \langle \nabla F(w^{(t-1)}), w^{(t)} - w^{(t-1)}\rangle\\
&+ \frac{1}{2}(w^{(t)} - w^{(t-1)})^\top (\nabla^2 F_1(w^{(t-1)}) + \gamma I)(w^{(t)} - w^{(t-1)}) + \frac{\nu}{6}\|w^{(t)} - w^{(t-1)}\|^3,
\end{aligned}$$

where in the last inequality we have used the assumption $\|\nabla^2 F_1(w^{(t-1)}) - \nabla^2 F(w^{(t-1)})\| \leq \gamma$. Based on the above inequality, it suffices to prove

$$\begin{aligned}
&\langle \nabla F(w^{(t-1)}), w^{(t)} - w^{(t-1)}\rangle + \frac{1}{2}(w^{(t)} - w^{(t-1)})^\top (\nabla^2 F_1(w^{(t-1)}) + \gamma I)(w^{(t)} - w^{(t-1)})\\
&+ \frac{\nu}{6}\|w^{(t)} - w^{(t-1)}\|^3\\
\leq& -\frac{1}{3}\langle \nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)}\rangle + \varepsilon_t\|\tilde{w}^{(t)} - w^{(t-1)}\|.
\end{aligned}$$

To this end, by mimicking the arguments in the proof of Lemma 6 we can show that

$$\begin{aligned}
&\langle \nabla F(w^{(t-1)}), w^{(t)} - w^{(t-1)} \rangle + \frac{1}{2}(w^{(t)} - w^{(t-1)})^\top (\nabla^2 F_1(w^{(t-1)}) + \gamma I)(w^{(t)} - w^{(t-1)}) \\
&+ \frac{\nu}{6}\|w^{(t)} - w^{(t-1)}\|^3 \\
\leq & - \left(\eta_t - \frac{\eta_t^2}{2}\right) \langle \nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)} \rangle \\
&+ \left(\frac{\nu\eta_t^2}{4} + \frac{\nu\eta_t^3}{6}\right) \|\tilde{w}^{(t)} - w^{(t-1)}\|^3 + \eta_t \varepsilon_t \|\tilde{w}^{(t)} - w^{(t-1)}\| \\
\overset{\zeta_1}{\leq} & - \left(\eta_t - \frac{\eta_t^2}{2}\right) \langle \nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)} \rangle \\
&+ \frac{1}{\mu + \gamma}\left(\frac{\nu\eta_t^2}{4} + \frac{\nu\eta_t^3}{6}\right) \|\tilde{w}^{(t)} - w^{(t-1)}\| \langle \nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^{(t-1)}) \\
&+ \gamma(\tilde{w}^{(t)} - w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)} \rangle + \eta_t \varepsilon_t \|\tilde{w}^{(t)} - w^{(t-1)}\| \\
&\left(-\left(\eta_t - \frac{\eta_t^2}{2}\right) + \frac{5\nu\|\tilde{w}^{(t)} - w^{(t-1)}\|}{12(\gamma + \mu)}\right) \langle \nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^{(t-1)}) \\
&+ \gamma(\tilde{w}^{(t)} - w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)} \rangle + \eta_t \varepsilon_t \|\tilde{w}^{(t)} - w^{(t-1)}\|,
\end{aligned}$$

where "ζ_1" is due to $\langle \nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)} \rangle \geq (\mu + \gamma)\|\tilde{w}^{(t)} - w^{(t-1)}\|^2$ and in the last inequality we have used the fact $\eta_t \leq 1$. When t is sufficiently large, from Theorem 8 we know that $\|\tilde{w}^{(t)} - w^{(t-1)}\|$ will be sufficiently close to zero so that $\frac{5\nu\|\tilde{w}^{(t)} - w^{(t-1)}\|}{12(\gamma+\mu)} \leq \frac{1}{6}$. Then we consider $\eta_t = 1$ in the above inequality so that

$$\begin{aligned}
&\langle \nabla F(w^{(t-1)}), w^{(t)} - w^{(t-1)} \rangle + \frac{1}{2}(w^{(t)} - w^{(t-1)})^\top (\nabla^2 F_1(w^{(t-1)}) + \gamma I)(w^{(t)} - w^{(t-1)}) \\
&+ \frac{\nu}{6}\|w^{(t)} - w^{(t-1)}\|^3 \\
\leq & - \frac{1}{3}\langle \nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}), \tilde{w}^{(t)} - w^{(t-1)} \rangle + \varepsilon_t \|\tilde{w}^{(t)} - w^{(t-1)}\|,
\end{aligned}$$

which implies that unit length is acceptable for any $\rho \in (0, 1/3]$. ■

We also need the following restated Lemma 11 which establishes the local convergence rate of Algorithm 1 when $\eta_t \equiv 1$, i.e., the unit length is always accepted by the backtracking line search.

Lemma 11 (Local convergence rate of DANE-LS) *Assume that F and F_1 are L-smooth, μ-strongly-convex and have ν-LH. Assume that $\sup_w \|\nabla^2 F_1(w) - \nabla^2 F(w)\| \leq \gamma$. Let $\tau = \left\lceil \frac{\mu+2\gamma}{2\mu} \log(4\kappa) \right\rceil$. Suppose that $\varepsilon_t \leq \min\left\{(\gamma+\mu)^2, \frac{\|\nabla F(w^{(t-1)})\|^2}{L^2}\right\}$ and $\max_{0\leq i\leq \tau-1} \|w^{(i)} - w^*\| \leq \frac{(\gamma+\mu)}{4(6\nu+1)\sqrt{\kappa\tau}}$. Then for any $\epsilon > 0$, Algorithm 1 with $\eta_t \equiv 1$ will attain estimation error $\|w^{(t)} - w^*\| \leq \epsilon$ after*

$$t \geq 4\tau \log\left(\frac{\gamma + \mu}{4(6\nu + 1)\sqrt{\kappa}\tau\nu\epsilon}\right)$$

rounds of iterations.

Proof Since $\eta_t \equiv 1$, we always have $w^{(t)} = \tilde{w}^{(t)}$. In view of the first-order optimality condition $\nabla F(w^*) = 0$ we can derive

$$
\begin{aligned}
&\nabla P^{(t-1)}(w^{(t)}) \\
=&\nabla F_1(w^{(t)}) - \nabla F_1(w^{(t-1)}) + \nabla F(w^{(t-1)}) + \gamma(w^{(t)} - w^{(t-1)}) \\
=&\nabla F_1(w^{(t)}) - \nabla F_1(w^*) + \nabla F_1(w^*) - \nabla F_1(w^{(t-1)}) + \nabla F(w^{(t-1)}) - \nabla F(w^*) + \gamma(w^{(t)} - w^{(t-1)}) \\
=&\Delta F_1(w^{(t)}, w^*) + \nabla^2 F_1(w^*)(w^{(t)} - w^*) - \Delta F_1(w^{(t-1)}, w^*) - \nabla^2 F_1(w^*)(w^{(t-1)} - w^*) \\
&+ \Delta F(w^{(t-1)}, w^*) + \nabla^2 F(w^*)(w^{(t-1)} - w^*) + \gamma(w^{(t)} - w^{(t-1)}) \\
=&\Delta F_1(w^{(t)}, w^*) + (\nabla^2 F_1(w^*) + \gamma I)(w^{(t)} - w^*) - \Delta F_1(w^{(t-1)}, w^*) - (\nabla^2 F_1(w^*) + \gamma I)(w^{(t-1)} - w^*) \\
&+ \Delta F(w^{(t-1)}, w^*) + \nabla^2 F(w^*)(w^{(t-1)} - w^*).
\end{aligned}
$$

Multiplying $(\nabla^2 F_1(w^*)+\gamma I)^{-1}$ on both sides of the above with proper rearrangement yields

$$
\begin{aligned}
&w^{(t)} - w^* \\
=&\left(I - (\nabla^2 F_1(w^*) + \gamma I)^{-1}\nabla^2 F(w^*)\right)(w^{(t-1)} - w^*) \\
&+ (\nabla^2 F_1(w^*) + \gamma I)^{-1}\left(\Delta F_1(w^{(t-1)}, w^*) - \Delta F(w^{(t-1)}, w^*) - \Delta F_1(w^{(t)}, w^*) + \nabla P^{(t-1)}(w^{(t)})\right).
\end{aligned}
$$

Let $H^* = \nabla^2 F(w^*)$ and $H_1^* = \nabla^2 F_1(w^*)$. Similar to the previous analysis, we work on the three term recurrence in matrix form

$$
u^{(t)} = Au^{(t-1)} + r^{(t-1)} \tag{A.4}
$$

where $u^{(t)} := w^{(t)} - w^*$, $A := I - (H_1^* + \gamma I)^{-1}H^*$ and

$$
r^{(t-1)} := (H_1^*+\gamma I)^{-1}\left(\Delta F_1(w^{(t-1)}, w^*) - \Delta F(w^{(t-1)}, w^*) - \Delta F_1(w^{(t)}, w^*) + \nabla P^{(t-1)}(w^{(t)})\right).
$$

We next bound $\|r^{(t-1)}\|$ with respect to $\|u^{(t-1)}\|$ and the local optimization precision ε_t.

$$
\begin{aligned}
&\|r^{(t-1)}\| \leq \left\|(H_1^* + \gamma I)^{-1}\right\| \left\|\Delta F_1(w^{(t-1)}, w^*) - \Delta F(w^{(t-1)}, w^*) - \Delta F_1(w^{(t)}, w^*)\right\| \\
&+ \left\|(H_1^* + \gamma I)^{-1}\right\| \|\nabla P^{(t-1)}(w^{(t)})\| \\
\leq&\frac{\nu}{2(\gamma+\mu)}\|w^{(t)} - w^*\|^2 + \frac{\nu}{\gamma+\mu}\|w^{(t-1)} - w^*\|^2 + \frac{\varepsilon_t}{\gamma+\mu},
\end{aligned} \tag{A.5}
$$

where we have used $H_1^* = \nabla^2 F_1(w^*) \succeq \mu I$ and the Lipschitz Hessian assumption such that $\|\Delta F_1(w^{(t)}, w^*)\| \leq \frac{\nu}{2}\|w^{(t)}-w^*\|^2$, $\|\Delta F_1(w^{(t-1)}, w^*)\| \leq \frac{\nu}{2}\|w^{(t-1)}-w^*\|^2$ and $\|\Delta F(w^{(t-1)}, w^*)\| \leq \frac{\nu}{2}\|w^{(t-1)} - w^*\|^2$, and also by assumption $\|\nabla P^{(t-1)}(w^{(t)})\| \leq \varepsilon_t$. In the following step we bound $\|w^{(t)}-w^*\|$ with respect to $\|w^{(t-1)}-w^*\|$. Since $\tilde{F}(w)$ is μ-strongly-convex, $P^{(t-1)}(w)$

is naturally $(\gamma+\mu)$-strongly-convex. Therefore

$$\begin{aligned}
&\|w^{(t)}-w^*\| \\
\leq&\frac{1}{\gamma+\mu}\|\nabla P^{(t-1)}(w^{(t)})-\nabla P^{(t-1)}(w^*)\| \overset{\zeta_1}{\leq} \frac{1}{\gamma+\mu}\|\nabla P^{(t-1)}(w^*)\|+\frac{\varepsilon_t}{\gamma+\mu} \\
=&\frac{1}{\gamma+\mu}\|\nabla F(w^{(t-1)})-\nabla F_1(w^{(t-1)})+\gamma(w^*-w^{(t-1)})+\nabla F_1(w^*)\|+\frac{\varepsilon_t}{\gamma+\mu} \\
=&\frac{1}{\gamma+\mu}\left\|\Big(\nabla F(w^{(t-1)})-\nabla F_1(w^{(t-1)})\Big)-(\nabla F(w^*)-\nabla F_1(w^*))+\gamma(w^*-w^{(t-1)})\right\| \\
&+\frac{\varepsilon_t}{\gamma+\mu}\leq\frac{2\gamma}{\gamma+\mu}\|w^{(t-1)}-w^*\|+\frac{\varepsilon_t}{\gamma+\mu}\leq 2\|w^{(t-1)}-w^*\|+\frac{\varepsilon_t}{\gamma+\mu},
\end{aligned} \tag{A.6}$$

where "ζ_1" follows from the sub-optimality of $w^{(t)}=\tilde{w}^{(t)}$ with respect to $P^{(t-1)}$ and the last inequality is implied by the assumption $\|\nabla^2 F(w)-\nabla^2 F_1(w)\|\leq\gamma$ for all w over a bonded domain of interest. By combining (A.5) and (A.6), and using the basic inequality $(a+b)^2\leq 2a^2+2b^2$ we arrive at

$$\begin{aligned}
\|r^{(t-1)}\|\leq&\frac{4\nu}{\gamma+\mu}\|w^{(t-1)}-w^*\|^2+\frac{\nu\varepsilon_t^2}{(\gamma+\mu)^3}+\frac{\nu}{\gamma+\mu}\|w^{(t-1)}-w^*\|^2+\frac{\varepsilon_t}{\gamma+\mu} \\
\overset{\zeta_1}{\leq}&\frac{5\nu}{\gamma+\mu}\|u^{(t-1)}\|^2+\frac{(\nu+1)\varepsilon_t}{\gamma+\mu}\overset{\zeta_2}{\leq}\frac{6\nu+1}{\gamma+\mu}\|u^{(t-1)}\|^2,
\end{aligned}$$

where in the inequality "ζ_1" we have used the assumption on ε_t which implies $\varepsilon_t\leq(\gamma+\mu)^2$, and "$\zeta_2$" follows from $\varepsilon_t\leq\frac{\|\nabla F(w^{(t-1)})\|^2}{L^2}\leq\|w^{(t-1)}-w^*\|^2=\|u^{(t-1)}\|^2$. Since $\|H_1^*-H^*\|\leq\gamma$ and $H^*\succeq\mu I$, by applying Lemma 24 we obtain that

$$\begin{aligned}
\|A^t\|=&\|(I-(H_1^*+\gamma I)^{-1}H^*)^t\| \\
=&\left\|\Big((H^*)^{-1/2}(I-(H^*)^{1/2}(H_1^*+\gamma I)^{-1}(H^*)^{1/2})(H^*)^{1/2}\Big)^t\right\| \\
=&\left\|(H^*)^{-1/2}(I-(H^*)^{1/2}(H_1^*+\gamma I)^{-1}(H^*)^{1/2})^t(H^*)^{1/2}\right\| \\
\leq&\sqrt{\frac{L}{\mu}}\|I-(H^*)^{1/2}(H_1^*+\gamma I)^{-1}(H^*)^{1/2}\|^t\leq\sqrt{\frac{L}{\mu}}\left(1-\frac{\mu}{\mu+2\gamma}\right)^t.
\end{aligned} \tag{A.7}$$

In the following argument, to simplify notation, we abbreviate

$$c=\sqrt{\frac{L}{\mu}},\quad \vartheta=\frac{6\nu+1}{\gamma+\mu},\quad \rho=1-\frac{\mu}{\mu+2\gamma}$$

such that

$$\|A^t\|\leq c\rho^t,\quad \|r^{(t)}\|\leq\vartheta\|u^{(t)}\|^2.$$

Let us consider the following defined integer

$$\tau=\left\lceil\frac{\mu+2\gamma}{2\mu}\log\left(\frac{4L}{\mu}\right)\right\rceil$$

such that $\|A^\tau\| \leq \frac{1}{2}$. We now prove by induction that for any integer $k \geq 0$,

$$\max_{0\leq i\leq \tau-1} \|u^{(k\tau+i)}\| \leq \frac{1}{4c\tau\vartheta}\left(\frac{3}{4}\right)^k.$$

The assumption $\max_{0\leq i\leq\tau-1}\|w^{(i)}-w^*\| \leq \frac{1}{4c\tau\vartheta}$ guarantees that the bound is valid for the base case $k=0$, i.e., $\max_{0\leq i\leq\tau-1}\|u^{(i)}\| \leq \frac{1}{4c\tau\vartheta}$. Now assume that the desired bound holds for some $k \geq 0$. By recursively applying (A.4) we can show that

$$\begin{aligned}
&\|u^{((k+1)\tau)}\| \\
=&\left\|A^\tau u^{(k\tau)} + \sum_{i=0}^{\tau-1} A^i r^{(k\tau+\tau-1-i)}\right\| \\
\leq&\|A^\tau\|\|u^{(k\tau)}\| + \sum_{i=0}^{\tau-1}\|A^i\|\|r^{(k\tau+\tau-1-i)}\| \\
\overset{\zeta_1}{\leq}&\frac{1}{2}\|u^{(k\tau)}\| + \vartheta c\sum_{i=0}^{\tau-1}\|u^{(k\tau+\tau-1-i)}\|^2 \\
\overset{\zeta_2}{\leq}&\frac{1}{2}\left(\frac{3}{4}\right)^k\frac{1}{4c\tau\vartheta} + \frac{1}{4\tau}\sum_{i=0}^{\tau-1}\|u^{k\tau+i}\| \\
\overset{\zeta_3}{\leq}&\frac{1}{2}\left(\frac{3}{4}\right)^k\frac{1}{4c\tau\vartheta} + \frac{1}{4}\left(\frac{3}{4}\right)^k\frac{1}{4c\tau\vartheta} = \left(\frac{3}{4}\right)^{k+1}\frac{1}{4c\tau\vartheta},
\end{aligned}$$

where "ζ_1" is due to (A.7) which implies $\|A^i\| \leq c$ for all $i \geq 1$ and it also has used $\|r^{(t)}\| \leq \vartheta\|u^{(t)}\|^2$, "$\zeta_2$" and "$\zeta_3$" are based on the induction step and $\|u^{k\tau+i}\| \leq \frac{1}{4c\tau\vartheta}$ for all $0 \leq i \leq \tau-1$. By using the same argument as the above, we can show that $\|u^{((k+1)\tau+i)}\| \leq \frac{1}{4c\tau\vartheta}\left(\frac{3}{4}\right)^{k+1}$ for all $0 \leq i \leq \tau-1$. This proves that $\max_{0\leq i\leq\tau-1}\|u^{(k\tau+i)}\| \leq \frac{1}{4c\tau\vartheta}\left(\frac{3}{4}\right)^k$ holds for all $k \geq 0$. Specially for $i=0$ we have

$$\|w^{(k\tau)} - w^*\| = \|u^{k\tau}\| \leq \frac{1}{4c\tau\vartheta}\left(\frac{3}{4}\right)^k.$$

Therefore, we need $t \geq 4\tau\log\left(\frac{1}{4c\tau\vartheta\epsilon}\right)$ to guarantee the estimation bound $\|w^{(t)}-w^*\| \leq \epsilon$. This completes the proof. ■

We are now ready to prove the main theorem.

Proof [of Theorem 13] Under the given conditions, from Theorem 8 and Lemma 10 we know that for any prefixed $\rho \in (0, 1/3]$, there exists a sufficiently large t_0 such that for all $t \geq t_0$, the unit length $\eta_t = 1$ is acceptable while the following bound holds:

$$\|\tilde{w}^{(t)} - w^{(t-1)}\| \leq \frac{6\mu}{13\gamma+\mu}\left(\frac{\gamma+\mu}{4(6\nu+1)\sqrt{\kappa}\tau}\right). \tag{A.8}$$

Since $\varepsilon_t \leq \frac{\rho(\mu+\gamma)}{2(L+\gamma)+\rho(\mu+\gamma)}\|\nabla F(w^{(t-1)})\|$, the bound in (A.3) holds such that

$$\varepsilon_t \leq \frac{\rho(\gamma+\mu)}{2}\|\tilde{w}^{(t)} - w^{(t-1)}\| \leq \frac{\gamma+\mu}{6}\|\tilde{w}^{(t)} - w^{(t-1)}\|,$$

where we have used $\rho \leq 1/3$. Then based on Lemma 29 and (A.8), the following holds for all $t \geq t_0$,

$$\begin{aligned}\|w^{(t)} - w^*\| =& \|\tilde{w}^{(t)} - w^*\| \leq \frac{2\gamma}{\mu}\|\tilde{w}^{(t)} - w^{(t-1)}\| + \frac{\varepsilon_t}{\mu} \\ &\leq \left(\frac{2\gamma}{\mu} + \frac{\gamma+\mu}{6\mu}\right)\|\tilde{w}^{(t)} - w^{(t-1)}\| \leq \frac{\gamma+\mu}{4(6\nu+1)\sqrt{\kappa}\tau}.\end{aligned}$$

Since $\varepsilon_t \leq \min\left\{(\gamma+\mu)^2, \frac{\|\nabla F(w^{(t-1)})\|^2}{L^2}\right\}$, by invoking Lemma 11 we obtain $\|w^{(t_0+t_1)} - w^*\| \leq \epsilon$ after

$$t_1 \geq 4\tau \log\left(\frac{\gamma+\mu}{4(6\nu+1)\sqrt{\kappa}\tau}\left(\frac{1}{\epsilon}\right)\right),$$

where $\tau = \left\lceil \frac{\mu+2\gamma}{2\mu}\log(4\kappa)\right\rceil$. This concludes the proof. ■

Appendix C. Proofs for Section 3

We collect in this appendix section the technical proofs of the results in Section 3, including Theorems 15, Theorem 18, Theorem 13 and their corollaries.

C.1 Proof of Theorem 15

We now prove Theorem 15 which is following restated.

Theorem 15 (Convergence rate of DANE-HB for quadratic function) *Assume that the loss function is quadratic. Let H and H_1 be the Hessian matrices of the global objective F and local objective F_1, respectively. Assume that $\mu I \preceq H \preceq LI$. Set $\beta = \left(1 - \sqrt{\frac{\mu}{\mu+2\gamma}}\right)^2$ and $\varepsilon_t = \frac{\sqrt{2}(\mu+\gamma)\|\nabla F(w^{(0)})\|}{2L(t+1)^2}\left(1 - \frac{1}{2}\sqrt{\frac{\mu}{\mu+2\gamma}}\right)^{t+1}$. Given precision $\epsilon > 0$, if $\|H_1 - H\| \leq \gamma$, then Algorithm 2 will output $w^{(t)}$ satisfying $\|w^{(t)} - w^*\| \leq \epsilon$ after*

$$t \geq 2\sqrt{\frac{\mu+2\gamma}{\mu}}\log\left(\frac{2\sqrt{2}c\|w^{(0)} - w^*\|}{\epsilon}\right)$$

rounds of iterations, where c is a constant relying on $\sqrt{\mu/(\mu+2\gamma)}$.

Proof [of Theorem 15] Since the objective is quadratic, for any $w^{(t-1)}$ the optimal solution $w^* = \arg\min_w F(w)$ can always be expressed as

$$w^* = w^{(t-1)} - H^{-1}\nabla F(w^{(t-1)}). \tag{A.9}$$

Since $H_1^{(t)} \equiv H_1$ holds in the quadratic case, based the gradient equation of $P^{(t-1)}$ at $\tilde{w}^{(t)}$ we have

$$\tilde{w}^{(t)} = w^{(t-1)} - (H_1+\gamma I)^{-1}\nabla F(w^{(t-1)}) + (H_1+\gamma I)^{-1}\nabla P^{(t-1)}(\tilde{w}^{(t)}).$$

Then from the definition of $w^{(t)} = \tilde{w}^{(t)} + \beta(w^{(t-1)} - w^{(t-2)})$ we have

$$w^{(t)} = w^{(t-1)} - \eta(H_1 + \gamma I)^{-1}\nabla F(w^{(t-1)}) + \beta(w^{(t-1)} - w^{(t-2)}) + r^{(t-1)}, \tag{A.10}$$

where the residual term $r^{(t-1)}$ is given by

$$r^{(t-1)} = (H_1 + \gamma I)^{-1}\nabla P^{(t-1)}(\tilde{w}^{(t)}).$$

Plugging (A.9) into (A.10) yields

$$w^{(t)} - w^* = ((1+\beta)I - (H_1 + \gamma I)^{-1}H)(w^{(t-1)} - w^*) - \beta(w^{(t-2)} - w^*) + r^{(t-1)}.$$

Now let us study the three term recurrence in matrix form

$$\begin{aligned}
\begin{bmatrix} w^{(t)} - w^* \\ w^{(t-1)} - w^* \end{bmatrix} &= \begin{bmatrix} (1+\beta)I - (H_1 + \gamma I)^{-1}H & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} w^{(t-1)} - w^* \\ w^{(t-2)} - w^* \end{bmatrix} + r^{(t-1)} \\
&= \begin{bmatrix} (1+\beta)I - (H_1 + \gamma I)^{-1}H & -\beta I \\ I & 0 \end{bmatrix}^t \begin{bmatrix} w^{(0)} - w^* \\ w^{(-1)} - w^* \end{bmatrix} \\
&\quad + \sum_{\tau=0}^{t-1} \begin{bmatrix} (1+\beta)I - (H_1 + \gamma I)^{-1}H & -\beta I \\ I & 0 \end{bmatrix}^\tau r^{(t-1-\tau)}.
\end{aligned}$$

Let us abbreviate

$$u^{(t)} := \begin{bmatrix} w^{(t)} - w^* \\ w^{(t-1)} - w^* \end{bmatrix}, \quad A := \begin{bmatrix} (1+\beta)I - (H_1 + \gamma I)^{-1}H & -\beta I \\ I & 0 \end{bmatrix}.$$

It follows from the preceding recursion form and the bastic fact $\|Tx\| \le \|T\|\|x\|$ that

$$\|u^{(t)}\| \le \|A^t\|\|u^{(0)}\| + \sum_{\tau=0}^{t-1} \|A^\tau\|\|r^{(t-1-\tau)}\|. \tag{A.11}$$

Let us now temporarily assume that $\rho(A) < 1$ and consider $\delta = \frac{1-\rho(A)}{2}$. From Lemma 26 we know that there exists a constant $c = c(\delta)$ such that for all $t \ge 0$:

$$\|A^t\| \le c(\rho(A) + \delta)^t = c\left(\frac{1+\rho(A)}{2}\right)^t. \tag{A.12}$$

Next we show that $\rho(A) < 1$ is indeed the case under the conditions of the theorem. Since $\|H_1 - H\| \le \gamma$ and $H \succeq \mu I$, by invoking Lemma 24 we obtain that $(H_1 + \gamma I)^{-1}H$ is diagonalizable and

$$\frac{\mu}{\mu + 2\gamma} \le \lambda_{\min}((H_1 + \gamma I)^{-1}H) \le \lambda_{\max}((H_1 + \gamma I)^{-1}H) \le 1.$$

Given the setting of $\beta = \left(1 - \sqrt{\frac{\mu}{\mu+2\gamma}}\right)^2$, it is known from Lemma 25 (with $\eta = 1$) that

$$\rho(A) \le 1 - \sqrt{\frac{\mu}{\mu + 2\gamma}}.$$

Notice that $\|r^{(t)}\| \leq \frac{\varepsilon_t}{\mu+\gamma}$ holds for all t which follows immediately from $\|\nabla P^{(t-1)}(\tilde{w}^{(t)})\| \leq \varepsilon_t$ and $H_1 \succeq \mu I$. Then combining the above bound with (A.11) and (A.12) yields

$$\begin{aligned}
&\|w^{(t)} - w^*\| \\
\leq &\|u^{(t)}\| \leq c\left(1 - \frac{1}{2}\sqrt{\frac{\mu}{\mu+2\gamma}}\right)^t \|u^{(0)}\| + \frac{c}{\mu+\gamma}\sum_{\tau=0}^{t-1}\varepsilon_{t-1-\tau}\left(1 - \frac{1}{2}\sqrt{\frac{\mu}{\mu+2\gamma}}\right)^\tau \\
\overset{\zeta_1}{\leq} &\sqrt{2}c\left(1 - \frac{1}{2}\sqrt{\frac{\mu}{\mu+2\gamma}}\right)^t \|w^{(0)} - w^*\| + \frac{c\sqrt{2}}{2}\sum_{\tau=0}^{t-1}\frac{1}{(t-\tau)^2}\left(1 - \frac{1}{2}\sqrt{\frac{\mu}{\mu+2\gamma}}\right)^t \|w^{(0)} - w^*\| \\
\leq &2\sqrt{2}c\left(1 - \frac{1}{2}\sqrt{\frac{\mu}{\mu+2\gamma}}\right)^t \|w^{(0)} - w^*\|,
\end{aligned}$$

where in the inequality "ζ_1" we have used $w^{(0)} = w^{(-1)}$ and the condition

$$\begin{aligned}
\varepsilon_t \leq &\frac{\sqrt{2}(\mu+\gamma)\|\nabla F(w^{(0)})\|}{2L(t+1)^2}\left(1 - \frac{1}{2}\sqrt{\frac{\mu}{\mu+2\gamma}}\right)^{t+1} \\
\leq &\frac{\sqrt{2}(\mu+\gamma)\|w^{(0)} - w^*\|}{2(t+1)^2}\left(1 - \frac{1}{2}\sqrt{\frac{\mu}{\mu+2\gamma}}\right)^{t+1},
\end{aligned}$$

and in the last inequality we have used $\sum_{\tau=0}^{t-1}\frac{1}{(t-\tau)^2} \leq 1 + \int_1^\infty \frac{1}{x^2}dx \leq 2$. Since $1 - a \leq e^{-a}$, it follows directly from the preceding bound that $\|w^{(t)} - w^*\| \leq \epsilon$ is valid when

$$t \geq 2\sqrt{\frac{\mu+2\gamma}{\mu}}\log\left(\frac{2\sqrt{2}c\|w^{(0)} - w^*\|}{\epsilon}\right).$$

This concludes the proof. ■

C.2 Proof of Theorem 18

We now prove the following restated Theorem 18 about the local convergence rate of DANE-HB when applied to strongly convex loss functions with Lipschitz Continuous Hessian.

Theorem 18 (Local convergence rate of DANE-HB) *Assume that F and F_1 are L-smooth, μ-strongly-convex and has ν-LH. Assume that $\sup_w \|\nabla^2 F_1(w) - \nabla^2 F(w)\| \leq \gamma$. Choose $\beta = \left(1 - \sqrt{\mu/(\mu+2\gamma)}\right)^2$. Let $\tau = \left\lceil 2\sqrt{(\mu+2\gamma)/\mu}\log(2c)\right\rceil$ in which c is a constant dependent on $\sqrt{\mu/(\mu+2\gamma)}$. Assume that $\varepsilon_t \leq \min\left\{(\gamma+\mu)^2, \|\nabla F(w^{(t-1)})\|^2/L^2\right\}$. Given precision $\epsilon > 0$, if $\max_{-1\leq i\leq\tau-1}\|w^{(i)} - w^*\| \leq \frac{\gamma+\mu}{4(6\nu+1)\sqrt{2}c\tau}$, then Algorithm 2 will output $w^{(t)}$ satisfying $\|w^{(t)} - w^*\| \leq \epsilon$ after*

$$t \geq 4\tau\log\left(\frac{\gamma+\mu}{4(6\nu+1)c\tau}\left(\frac{1}{\epsilon}\right)\right)$$

rounds of iterations.

Proof [of Theorem 18] The proof mimics that of Lemma 11 with proper adaptation to the heave-ball momentum formulation. For the sake of completeness, here we provide the full details of proof. Since $\nabla F(w^*) = 0$, we can show that

$$\begin{aligned}
&\nabla P^{(t-1)}(\tilde{w}^{(t)}) \\
=&\nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^{(t-1)}) + \nabla F(w^{(t-1)}) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}) \\
=&\nabla F_1(\tilde{w}^{(t)}) - \nabla F_1(w^*) + \nabla F_1(w^*) - \nabla F_1(w^{(t-1)}) + \nabla F(w^{(t-1)}) - \nabla F(w^*) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}) \\
=&\Delta F_1(\tilde{w}^{(t)}, w^*) + \nabla^2 F_1(w^*)(\tilde{w}^{(t)} - w^*) - \Delta F_1(w^{(t-1)}, w^*) - \nabla^2 F_1(w^*)(w^{(t-1)} - w^*) \\
&+ \Delta F(w^{(t-1)}, w^*) + \nabla^2 F(w^*)(w^{(t-1)} - w^*) + \gamma(\tilde{w}^{(t)} - w^{(t-1)}) \\
=&\Delta F_1(\tilde{w}^{(t)}, w^*) + (\nabla^2 F_1(w^*) + \gamma I)(\tilde{w}^{(t)} - w^*) - \Delta F_1(w^{(t-1)}, w^*) \\
&- (\nabla^2 F_1(w^*) + \gamma I)(w^{(t-1)} - w^*) + \Delta F(w^{(t-1)}, w^*) + \nabla^2 F(w^*)(w^{(t-1)} - w^*).
\end{aligned}$$

Multiplying $(\nabla^2 F_1(w^*)+\gamma I)^{-1}$ on both sides of the above followed by proper rearrangement yields

$$\begin{aligned}
&\tilde{w}^{(t)} - w^* \\
=&\left(I - (\nabla^2 F_1(w^*) + \gamma I)^{-1}\nabla^2 F(w^*)\right)(w^{(t-1)} - w^*) \\
&+ (\nabla^2 F_1(w^*) + \gamma I)^{-1}\left(\nabla P^{(t-1)}(\tilde{w}^{(t)}) - \Delta F_1(\tilde{w}^{(t)}, w^*) + \Delta F_1(w^{(t-1)}, w^*) - \Delta F(w^{(t-1)}, w^*)\right) \\
=&\left(I - (\nabla^2 F_1(w^*) + \gamma I)^{-1}\nabla^2 F(w^*)\right)(w^{(t-1)} - w^*) \\
&+ (\nabla^2 F_1(w^*) + \gamma I)^{-1}\left(\Delta F_1(w^{(t-1)}, w^*) - \Delta F(w^{(t-1)}, w^*) - \Delta F_1(\tilde{w}^{(t)}, w^*) + \nabla P^{(t-1)}(\tilde{w}^{(t)})\right).
\end{aligned}$$

Recall the update $w^{(t)} = \tilde{w}^{(t)} + \beta(w^{(t-1)} - w^{(t-2)})$. It follows that

$$\begin{aligned}
&w^{(t)} - w^* \\
=&\left((1+\beta)I - (\nabla^2 F_1(w^*) + \gamma I)^{-1}\nabla^2 F(w^*)\right)(w^{(t-1)} - w^*) - \beta(w^{(t-2)} - w^*) \\
&+ (\nabla^2 F_1(w^*) + \gamma I)^{-1}\left(\Delta F_1(w^{(t-1)}, w^*) - \Delta F(w^{(t-1)}, w^*) - \Delta F_1(\tilde{w}^{(t)}, w^*) + \nabla P^{(t-1)}(\tilde{w}^{(t)})\right).
\end{aligned}$$

Let $H^* = \nabla^2 F(w^*)$ and $H_1^* = \nabla^2 F_1(w^*)$. Similar to the previous analysis, we work on the three term recurrence in matrix form

$$u^{(t)} = Au^{(t-1)} + r^{(t-1)} \tag{A.13}$$

where $u^{(t)} := \begin{bmatrix} w^{(t)} - w^* \\ w^{(t-1)} - w^* \end{bmatrix}$, $A := \begin{bmatrix} (1+\beta)I - (H_1^* + \gamma I)^{-1}H^* & -\beta I \\ I & 0 \end{bmatrix}$ and

$$r^{(t-1)} := \begin{bmatrix} (H_1^* + \gamma I)^{-1}\left(\Delta F_1(w^{(t-1)}, w^*) - \Delta F(w^{(t-1)}, w^*) - \Delta F_1(\tilde{w}^{(t)}, w^*) + \nabla P^{(t-1)}(\tilde{w}^{(t)})\right) \\ 0 \end{bmatrix}.$$

Provided that $\varepsilon_t \leq \min\left\{(\gamma+\mu)^2, \|\nabla F(w^{(t-1)})\|^2/L^2\right\}$, using about the same arguments as those in the proof of Lemma 11, we can bound $\|r^{(t-1)}\|$ with respect to $\|u^{(t-1)}\|$ as

$$\|r^{(t-1)}\| \leq \frac{6\nu + 1}{\gamma + \mu}\|u^{(t-1)}\|^2.$$

Since $\|H_1^* - H^*\| \leq \gamma$ and $H^* \succeq \mu I$, by applying Lemma 24 we obtain that $(H_1^* + \gamma I)^{-1} H^*$ is diagonalizable and

$$\frac{\mu}{\mu + 2\gamma} \leq \lambda_{\min}\left((H_1^* + \gamma I)^{-1} H^*\right) \leq \lambda_{\max}\left((H_1^* + \gamma I)^{-1} H^*\right) \leq 1.$$

Given $\beta = \left(1 - \sqrt{\frac{\mu}{\mu+2\gamma}}\right)^2$, it is true in view of Lemma 25 (with $\eta = 1$) that

$$\rho(A) \leq 1 - \sqrt{\frac{\mu}{\mu + 2\gamma}}.$$

Let $\delta = \frac{1-\rho(A)}{2}$. From Lemma 26 we know that there exists a constant $c = c(\delta)$ such that for all $t \geq 0$:

$$\|A^t\| \leq c(\rho(A) + \delta)^t = c\left(\frac{1 + \rho(A)}{2}\right)^t \leq c\left(1 - \frac{1}{2}\sqrt{\frac{\mu}{\mu + 2\gamma}}\right)^t. \tag{A.14}$$

Without loss of generality we assume $c \geq 1$. To simplify notation, we abbreviate $\vartheta = \frac{6\nu+1}{\gamma+\mu}$ in the following argument. Let us consider the following defined integer

$$\tau = \left\lceil 2\sqrt{\frac{\mu + 2\gamma}{\mu}} \log(2c) \right\rceil,$$

which ensures $\|A^\tau\| \leq \frac{1}{2}$. We now prove by induction that for any integer $k \geq 0$,

$$\max_{0 \leq i \leq \tau - 1} \|u^{(k\tau+i)}\| \leq \frac{1}{4c\tau\vartheta}\left(\frac{3}{4}\right)^k.$$

The assumption $\max_{-1 \leq i \leq \tau-1} \|w^{(i)} - w^*\| \leq \frac{1}{4\sqrt{2}c\tau\vartheta}$ guarantees that the bound is valid for the base case $k = 0$, i.e., $\max_{0 \leq i \leq \tau-1} \|u^{(i)}\| \leq \frac{1}{4c\tau\vartheta}$. Now assume that the above bound is valid for some $k \geq 0$. Then based on the recursive form (A.13) we can derive

$$\begin{aligned}
&\|u^{((k+1)\tau)}\| \\
=&\left\| A^\tau u^{(k\tau)} + \sum_{i=0}^{\tau-1} A^i r^{(k\tau+\tau-1-i)} \right\| \\
\leq&\|A^\tau\|\|u^{(k\tau)}\| + \sum_{i=0}^{\tau-1} \|A^i\|\|r^{(k\tau+\tau-1-i)}\| \\
\overset{\zeta_1}{\leq}&\frac{1}{2}\|u^{(k\tau)}\| + \vartheta c \sum_{i=0}^{\tau-1} \|u^{(k\tau+\tau-1-i)}\|^2 \\
\overset{\zeta_2}{\leq}&\frac{1}{2}\left(\frac{3}{4}\right)^k \frac{1}{4c\tau\vartheta} + \frac{1}{4\tau}\sum_{i=0}^{\tau-1} \|u^{k\tau+i}\| \\
\overset{\zeta_3}{\leq}&\frac{1}{2}\left(\frac{3}{4}\right)^k \frac{1}{4c\tau\vartheta} + \frac{1}{4}\left(\frac{3}{4}\right)^k \frac{1}{4c\tau\vartheta} = \left(\frac{3}{4}\right)^{k+1} \frac{1}{4c\tau\vartheta},
\end{aligned}$$

where "ζ_1" is due to (A.14) which implies $\|A^i\| \leq c$ for all $i \geq 1$ and it also has used $\|r^{(t)}\| \leq \vartheta \|u^{(t)}\|^2$, "$\zeta_2$" and "$\zeta_3$" are based on the induction step and $\|u^{k\tau+i}\| \leq \frac{1}{4c\tau\vartheta}$ for all $0 \leq i \leq \tau - 1$. Similarly, we can show that $\|u^{((k+1)\tau+i)}\| \leq \frac{1}{4c\tau\vartheta}\left(\frac{3}{4}\right)^{k+1}$ for all $0 \leq i \leq \tau - 1$. This proves that $\max_{0\leq i\leq \tau-1} \|u^{(k\tau+i)}\| \leq \frac{1}{4c\tau\vartheta}\left(\frac{3}{4}\right)^{k}$ holds for all $k \geq 0$. Particularly, we obtain

$$\|w^{(k\tau)} - w^*\| \leq \|u^{k\tau}\| \leq \frac{1}{4c\tau\vartheta}\left(\frac{3}{4}\right)^{k}.$$

Therefore, to reach $\|w^{(t)} - w^*\| \leq \epsilon$ we need $t \geq 4\tau \log\left(\frac{1}{4c\tau\vartheta\epsilon}\right)$. This completes the proof. ■

C.3 Proof of Theorem 20

In this subsection, we prove Theorem 20 as restated below.

Theorem 20 (Convergence of D^{2}ANE) *Assume that the univariate functions l_i are ℓ-smooth and σ-strongly convex. Assume without loss of generality that $\|x_i\| \leq 1$. Let $H = \frac{\ell}{N}XX^\top + \mu I$ and $H_1 = \frac{\ell}{n}X_1X_1^\top + \mu I$. Choose $\beta = \left(1 - \sqrt{\frac{\mu}{\mu+2\gamma}}\right)^2$ and $\epsilon_t = \frac{\sigma}{2\ell}\exp\left\{-\frac{\sigma(t-1)}{2\ell}\right\}$. If $\|H_1 - H\| \leq \gamma$, then Algorithm 3 will output solution $w^{(t)}$ with sub-optimality $F(w^{(t)}) - F(w^*) \leq \epsilon$ after*

$$t \geq \frac{2\ell}{\sigma}\log\left(\frac{\max\left\{1, F(w^{(0)}) - F(w^*)\right\}}{\epsilon}\right)$$

rounds of outer-loop iterations and

$$\mathcal{O}\left(\frac{\ell}{\sigma}\sqrt{\frac{\gamma}{\mu}}\log^2\left(\frac{1}{\epsilon}\right)\right)$$

rounds of inner-loop iterations of DANE-HB.

Proof We first analyze the outer-loop iteration complexity. As defined in Algorithm 3, at each time instance t the quadratic subproblem is optimized to certain ϵ_t-sub-optimality, i.e.,

$$Q^{(t-1)}(w^{(t)}) \leq \min_w Q^{(t-1)}(w) + \epsilon_t.$$

The value of ϵ_t will be specified shortly in the analysis to follow. Let us abbreviate $l_i(w^\top x_i) = l(w^\top x_i, y_i)$ with l_i being a univariate function. For any $\eta \in [0, 1]$, the smoothness

of l_i and the sub-optimality of $w^{(t)}$ imply

$$\begin{aligned}
&F(w^{(t)})\\
=&\tilde{F}(w^{(t)}) + \frac{\mu}{2}\|w^{(t)}\|^2 = \frac{1}{N}\sum_{i=1}^N l_i(x_i^\top w^{(t)}) + \frac{\mu}{2}\|w^{(t)}\|^2\\
\leq&\frac{1}{N}\sum_{i=1}^N \Big\{ l_i(x_i^\top w^{(t-1)}) + l_i'(x_i^\top w^{(t-1)}) x_i^\top (w^{(t)} - w^{(t-1)})\\
&+\frac{\ell}{2}(w^{(t)} - w^{(t-1)})^\top x_i x_i^\top (w^{(t)} - w^{(t-1)})\Big\} + \frac{\mu}{2}\|w^{(t)}\|^2\\
=&\tilde{F}(w^{(t-1)}) + \langle \nabla \tilde{F}(w^{(t-1)}), w^{(t)} - w^{(t-1)}\rangle + \frac{\ell}{2N}(w^{(t)} - w^{(t-1)})^\top XX^\top (w^{(t)} - w^{(t-1)})\\
&+\frac{\mu}{2}\|w^{(t)}\|^2\\
=&Q^{(t-1)}(w^{(t)}) \leq Q^{(t-1)}((1-\eta)w^{(t-1)} + \eta w^*) + \varepsilon_t\\
=&F(w^{(t-1)}) + \eta\langle \nabla F(w^{(t-1)}), w^* - w^{(t-1)}\rangle\\
&+\frac{\eta^2\ell}{2}(w^* - w^{(t-1)})^\top \left(\frac{XX^\top}{N} + \frac{\mu}{\ell} I\right)(w^* - w^{(t-1)}) + \varepsilon_t.
\end{aligned}$$

On the other side, from the strong-convexity of $l_i(\cdot)$ we can show that

$$\begin{aligned}
&F(w^*)\\
=&\frac{1}{N}\sum_{i=1}^N l_i(x_i^\top w^*) + \frac{\mu}{2}\|w^*\|^2\\
\geq&\frac{1}{N}\sum_{i=1}^N \Big\{ f_i(x_i^\top w^{(t-1)}) + f_i'(x_i^\top w^{(t-1)}) x_i^\top (w^* - w^{(t-1)})^\top + \frac{\sigma}{2}(w^* - w^{(t-1)})^\top x_i x_i^\top (w^* - w^{(t-1)})\Big\}\\
&+\frac{\mu}{2}\|w^{(t-1)}\|^2 + \mu\langle w^{(t-1)}, w^* - w^{(t-1)}\rangle + \frac{\mu}{2}\|w^* - w^{(t-1)}\|^2\\
=&F(w^{(t-1)}) + \langle \nabla F(w^{(t-1)}), w^* - w^{(t-1)}\rangle + \frac{\sigma}{2}(w^* - w^{(t-1)})^\top \left(\frac{XX^\top}{N} + \frac{\mu}{\sigma} I\right)(w^* - w^{(t-1)})\\
\geq&F(w^{(t-1)}) + \langle \nabla F(w^{(t-1)}), w^* - w^{(t-1)}\rangle + \frac{\sigma}{2}(w^* - w^{(t-1)})^\top \left(\frac{XX^\top}{N} + \frac{\mu}{\ell} I\right)(w^* - w^{(t-1)}),
\end{aligned}$$

where in the last inequality we have used the basic fact $\ell \geq \sigma$. By setting $\eta = \sigma/\ell \in (0, 1]$ and combining the above two inequalities we arrive at

$$F(w^{(t)}) - F(w^*) \leq \left(1 - \frac{\sigma}{\ell}\right)(F(w^{(t-1)}) - F(w^*)) + \epsilon_t.$$

Since l_i is ℓ-smooth and $\|x_i\| \leq 1$, we can verify that F is $(\ell + \mu)$-smooth. In view of the condition

$$\epsilon_t = \frac{\sigma}{2\ell}\exp\left\{-\frac{\sigma(t-1)}{2\ell}\right\}, \tag{A.15}$$

it can be straightforwardly shown by induction that

$$F(w^{(t)}) - F(w^*) \leq \exp\left\{-\frac{\sigma t}{2\ell}\right\} \max\left\{1, F(w^{(0)}) - F(w^*)\right\}.$$

Then for any desired precision $\epsilon > 0$, the sub-optimality $F(w^{(t)}) - F(w^*) \leq \epsilon$ holds provided that

$$t \geq T = \frac{2\ell}{\sigma} \log\left(\frac{\max\left\{1, F(w^{(0)}) - F(w^*)\right\}}{\epsilon}\right).$$

To prove the inner-loop iteration complexity, from Theorem 15 and (A.15) we know that the condition $Q^{(t-1)}(w^{(t)}) \leq \min_w Q^{(t-1)}(w) + \epsilon_t$ is valid when the inner loop is sufficiently executed with $\mathcal{O}\left(\sqrt{\frac{\gamma}{\mu}} \log\left(\frac{1}{\epsilon_t}\right)\right)$ rounds of iterations. Therefore, the overall inner-loop iteration complexity to attain $F(w^{(t)}) - F(w^*) \leq \epsilon$ is dominated by

$$\begin{aligned}
&\mathcal{O}\left(\sum_{t=1}^{T}\left\{\sqrt{\frac{\gamma}{\mu}}\left(\log\left(\frac{\ell}{\sigma}\right) + (t-1)\frac{\sigma}{\ell}\right)\right\}\right) \\
=&\mathcal{O}\left(\sqrt{\frac{\gamma}{\mu}}\left(T\log\left(\frac{\ell}{\sigma}\right) + T^2\frac{\sigma}{\ell}\right)\right) \leq \mathcal{O}\left(\frac{\ell}{\sigma}\sqrt{\frac{\gamma}{\mu}}\log^2\left(\frac{1}{\epsilon}\right)\right).
\end{aligned}$$

This proves the desired bound. ■

Appendix D. Proof of Auxiliary Lemmas

D.1 Proof of Lemma 24

Proof As both $A+\gamma I$ and B are symmetric and positive definite, we know that the eigenvalues of $(A+\gamma I)^{-1}B$ are positive real numbers and identical to those of $(A+\gamma I)^{-1/2}B(A+\gamma I)^{-1/2}$. Consider the following eigenvalue decomposition of $(A+\gamma I)^{-1/2}B(A+\gamma I)^{-1/2}$:

$$(A+\gamma I)^{-1/2}B(A+\gamma I)^{-1/2} = Q^\top \Lambda Q,$$

where $Q^\top Q = I$ and Λ is a diagonal matrix with eigenvalues as diagonal entries. Then

$$(A+\gamma I)^{-1}B = (A+\gamma I)^{-1/2}Q^\top \Lambda Q (A+\gamma I)^{1/2},$$

which is a diagonal eigenvalue decomposition, and hence $(A+\gamma I)^{-1}B$ is diagonalizable.

To prove the eigenvalue bounds of $(A+\gamma I)^{-1}B$, it suffices to prove the same bounds for $(A+\gamma I)^{-1/2}B(A+\gamma I)^{-1/2}$. Since $\|A-B\| \leq \gamma$, we have $B \preceq A+\gamma I$ which implies $(A+\gamma I)^{-1/2}B(A+\gamma I)^{-1/2} \preceq I$ and hence $\lambda_{\max}((A+\gamma I)^{-1/2}B(A+\gamma I)^{-1/2}) \leq 1$. Moreover, since $B \succeq \mu I$, it holds that $\frac{2\gamma}{\mu}B - \gamma I \succeq \gamma I \succeq A - B$. Then we obtain $(A+\gamma I)^{-1/2}B(A+\gamma I)^{-1/2} \succeq \frac{\mu}{\mu+2\gamma}I$ which implies $\lambda_{\min}((A+\gamma I)^{-1/2}B(A+\gamma I)^{-1/2}) \geq \frac{\mu}{\mu+2\gamma}$. Therefore it must hold

$$\|I - (A+\gamma I)^{-1/2}B(A+\gamma I)^{-1/2}\| \leq 1 - \frac{\mu}{\mu+2\gamma} = \frac{2\gamma}{\mu+2\gamma}.$$

Similarly, we can show that $\frac{\mu}{\mu+2\gamma}I \preceq B^{1/2}(A+\gamma I)^{-1}B^{1/2} \preceq I$, and thus $\|I - B^{1/2}(A+\gamma I)^{-1}B^{1/2}\| \leq \frac{2\gamma}{\mu+2\gamma}$. ■

D.2 Proof of Lemma 25

Proof Let $0 < \mu \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d \leq L$ be the eigenvalues of A and Λ be a diagonal matrix whose diagonal entries are $\{\lambda_i\}$ in a non-decreasing order. Since A is diagonalizable, it can be verified that the eigenvalues of the following two $2d \times 2d$ matrices coincide:

$$T_1 = \begin{bmatrix} (1+\beta)I - \eta A & -\beta I \\ I & 0 \end{bmatrix}, \quad T_2 = \begin{bmatrix} (1+\beta)I - \eta \Lambda & -\beta I \\ I & 0 \end{bmatrix}.$$

It is possible to permute the matrix T_2 to a block diagonal matrix with 2×2 blocks of the form

$$\begin{bmatrix} 1+\beta-\eta\lambda_i & -\beta \\ 1 & 0 \end{bmatrix}.$$

Therefore we have

$$\begin{aligned} &\rho\left(\begin{bmatrix} (1+\beta)I - \eta A & -\beta I \\ I & 0 \end{bmatrix}\right) \\ =&\rho\left(\begin{bmatrix} (1+\beta)I - \eta \Lambda & -\beta I \\ I & 0 \end{bmatrix}\right) = \max_{i\in[d]} \rho\left(\begin{bmatrix} 1+\beta-\eta\lambda_i & -\beta \\ 1 & 0 \end{bmatrix}\right). \end{aligned}$$

For each $i \in [d]$, the eigenvalues of the 2×2 block matrices are given by the roots of

$$\lambda^2 - (1+\beta-\eta\lambda_i)\lambda + \beta = 0.$$

Given that $\beta \geq |1 - \sqrt{\eta\lambda_i}|^2$, the roots of the above equation are imaginary and both have magnitude $\sqrt{\beta}$. Since $\beta = \max\{|1-\sqrt{\eta\mu}|^2, |1-\sqrt{\eta L}|^2\}$, the magnitude of each root is at most $\max\{|1-\sqrt{\eta\mu}|, |1-\sqrt{\eta L}|\}$. This proves the desired spectral radius bound. ■

Appendix E. Computational Complexity of DANE-HB and D^{2}ANE

In addition to the communication complexity analysis, here we further provide a computational complexity analysis for DANE-HB in order to gain better understanding of its overall computational efficiency. We first restrict our attention to the quadratic setting in which the global convergence of DANE-HB is guaranteed. At each communication round t, the master machine needs to solve the local subproblem $\tilde{w}^{(t)} \approx \arg\min_w P^{(t-1)}(w)$ to certain desired precision. Inspired by Federated SVRG (Konečnỳ et al., 2016) which essentially applies SVRG (Johnson and Zhang, 2013) to the local optimization of InexactDane , we specify that the local minimization of DANE-HB is solved by SVRG. Clearly such a specification of DANE-HB only needs to access the first-order information of the loss functions. Following Johnson and Zhang (2013); Zhang and Xiao (2017), we employ the incremental first-order oracle (IFO) complexity as the computational complexity metric for solving the finite-sum minimization problem (1).

Definition 30 *An IFO takes an index $i \in [N]$ and a point $(x_i, y_i) \in \{x_j, y_j\}_{j=1}^N$, and returns the pair $(f(w; x_i, y_i), \nabla f(w; x_i, y_i))$.*

As a consequence of Corollary 16, the following result summaries the computational complexity of DANE-HB for quadratic problems.

Corollary 31 (Computational complexity of DANE-HB for quadratic objective) *Assume the conditions in Corollary 16 hold and the local subproblems are solved by SVRG. Then for sufficiently small $\delta > 0$, with probability at least $1 - \delta$ over the random samples drawn to construct F_1, the IFO complexity of DANE-HB for attaining estimation error $\|w^{(t)} - w^*\| \leq \epsilon$ is bounded in expectation (w.r.t. stochastic gradient estimation) by*

$$\mathcal{O}\left(\sqrt{\kappa}\left(n^{3/4}+n^{1/4}\right)\log^{1/4}\left(\frac{p}{\delta}\right)\log^2\left(\frac{1}{\epsilon}\right)+\sqrt{\kappa}n^{3/4}\log^{1/4}\left(\frac{p}{\delta}\right)\log\left(\frac{1}{\epsilon}\right)\right),$$

Proof Recollect that $\gamma = L\sqrt{\frac{32\log(p/\delta)}{n}}$ in Corollary 16. From Corollary 16 we know that with probability at least $1-\delta$ over the random choice of F_1, $\|w^{(t)} - w^*\| \leq \epsilon$ after

$$T = \mathcal{O}\left(\frac{\sqrt{\kappa}}{n^{1/4}}\log^{1/4}\left(\frac{p}{\delta}\right)\log\left(\frac{1}{\epsilon}\right)\right)$$

rounds of outer-loop communication. In each round of outer-loop communication, each worker machine needs to compute the local batch gradient over n samples, and thus the outer-loop full gradient computation can be done in parallel with IFO complexity

$$\mathcal{O}\left(\sqrt{\kappa}n^{3/4}\log^{1/4}\left(\frac{p}{\delta}\right)\log\left(\frac{1}{\epsilon}\right)\right).$$

It is standard to know that the IFO complexity of the inner-loop SVRG computation can be bounded in expectation by

$$\mathcal{O}\left(\left(n+\frac{L+\gamma}{\gamma+\mu}\right)\log\left(\frac{1}{\varepsilon_t}\right)\right) \leq \mathcal{O}\left(\left(n+\sqrt{\frac{n}{\log(p/\delta)}}\right)\log\left(\frac{1}{\epsilon}\right)\right) \leq \mathcal{O}\left(\left(n+\sqrt{n}\right)\log\left(\frac{1}{\epsilon}\right)\right),$$

where we have used $\log(1/\varepsilon_t) \leq \mathcal{O}(\log(1/\epsilon))$ for all $t \leq T$ and $\log(p/\delta) \geq 1$ for sufficiently small tail bound δ. Combing the above inner-loop and outer-loop IFO bounds yields the following overall expectation (w.r.t. SVRG) computation complexity bound

$$\mathcal{O}\left(\sqrt{\kappa}\left(n^{3/4}+n^{1/4}\right)\log^{1/4}\left(\frac{p}{\delta}\right)\log^2\left(\frac{1}{\epsilon}\right)+\sqrt{\kappa}n^{3/4}\log^{1/4}\left(\frac{p}{\delta}\right)\log\left(\frac{1}{\epsilon}\right)\right),$$

which holds with probability at least $1-\delta$ over the randomness of F_1. ■

For an instance, let us consider the conventional regularized learning problems in which the condition number κ scales as large as $\mathcal{O}(\sqrt{N}) = \mathcal{O}(\sqrt{mn})$. In this case, the above result implies that with high probability over the random construction of F_1, the expected IFO complexity bound of DANE-HB with local SVRG optimization is dominated by

$$\mathcal{O}\left(\left(m^{1/4}n+m^{1/4}n^{1/2}\right)\log^2\left(\frac{1}{\epsilon}\right)+m^{1/4}n\log\left(\frac{1}{\epsilon}\right)\right).$$

For comparison, the expected IFO complexity bound of the classic single-machine SVRG is given by

$$\mathcal{O}\left(\left(mn+\sqrt{mn}\right)\log\left(\frac{1}{\epsilon}\right)\right).$$

Since the sample size mn dominates the condition number $\sqrt{mn}$ in this example, up to logarithm factors, DANE-HB is roughly $\times m^{3/4}$ cheaper than SVRG in computational cost for quadratic problems, which also matches the corresponding result of MP-DANE (Wang et al., 2017b).

Analogously, by combining Corollary 22 and Corollary 31 we can establish the following result on the overall IFO complexity bound of D^{2}ANE for linear models.

Corollary 32 (Computation complexity of D^{2}ANE) *Assume the conditions in Corollary 22 hold and the local subproblems are solved by SVRG. Then for sufficiently small $\delta>0$, with probability at least $1-\delta$ over the random samples drawn to construct F_1, the IFO complexity of D^{2}ANE for attaining sub-optimality $F(w^{(t)})-F(w^*)\leq\epsilon$ is bounded in expectation (w.r.t. stochastic gradient estimation) by*

$$\mathcal{O}\left(\frac{\ell\sqrt{\kappa}}{\sigma}\left(n^{3/4}+n^{1/4}\right)\log^{1/4}\left(\frac{p}{\delta}\right)\log^3\left(\frac{1}{\epsilon}\right)+\frac{\ell\sqrt{\kappa}}{\sigma}n^{3/4}\log^{1/4}\left(\frac{p}{\delta}\right)\log^2\left(\frac{1}{\epsilon}\right)\right).$$

When the condition number scales as $\kappa=\mathcal{O}(\sqrt{N})=\mathcal{O}(\sqrt{mn})$ in regularized statistical learning problems, Corollary 32 shows that with high probability over the random construction of F_1, the expected IFO complexity of D^{2}ANE implemented with SVRG (for local optimization) is upper bounded by

$$\mathcal{O}\left(\frac{\ell}{\sigma}\left(m^{1/4}n+m^{1/4}n^{1/2}\right)\log^3\left(\frac{1}{\epsilon}\right)+\frac{\ell}{\sigma}m^{1/4}n\log^2\left(\frac{1}{\epsilon}\right)\right).$$

This above bound indicates that up to logarithm factors, D^{2}ANE is roughly $\times m^{3/4}$ cheaper than SVRG in computational cost for strongly convex optimization with linear models.

References

Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1756–1764, 2015.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Jerry Chee and Ping Li. Understanding and detecting convergence for stochastic gradient descent with momentum. *arXiv preprint arXiv:2008.12224*, 2020.

Xiangyi Chen, Xiaoyun Li, and Ping Li. Toward communication efficient adaptive gradient method. In *ACM-IMS Foundations of Data Science Conference (FODS)*, 2020.

Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, January 2008.

Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European Control Conference (ECC)*, pages 310–315. IEEE, 2015.

Isabelle Guyon, Steve Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the nips 2003 feature selection challenge. In *Advances in Neural Information Processing Systems (NIPS)*, pages 545–552, 2005.

Hadrien Hendrikx, Lin Xiao, Sebastien Bubeck, Francis Bach, and Laurent Massoulie. Statistically preconditioned accelerated gradient method for distributed optimization. In *International Conference on Machine Learning (ICML)*, 2020.

Martin Jaggi, Virginia Smith, Martin Takác, Jonathan Terhorst, Sanjay Krishnan, Thomas Hofmann, and Michael I Jordan. Communication-efficient distributed dual coordinate ascent. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NIPS)*, pages 315–323, 2013.

Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, pages 1–14, 2018.

Hiroyuki Kasai. SGDLibrary: A MATLAB library for stochastic optimization algorithms. *Journal of Machine Learning Research*, 18:215–1, 2017.

Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

Jason D Lee, Qihang Lin, Tengyu Ma, and Tianbao Yang. Distributed stochastic variance reduced gradient methods by sampling extra data with replacement. *Journal of Machine Learning Research*, 18(1):4404–4446, 2017.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5(Apr):361–397, 2004.

Mu Li, David G Andersen, Alex J Smola, and Kai Yu. Communication efficient distributed machine learning with the parameter server. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3384–3392, 2015.

Bo Liu, Xiao-Tong Yuan, Lezi Wang, Qingshan Liu, Junzhou Huang, and Dimitris Metaxas. Distributed inexact newton-type pursuit for non-convex sparse learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Nicolas Loizou and Peter Richtárik. Linearly convergent stochastic heavy ball method for minimizing generalization error. *arXiv preprint arXiv:1710.10737*, 2017.

Chenxin Ma, Virginia Smith, Martin Jaggi, Michael Jordan, Peter Richtarik, and Martin Takac. Adding vs. averaging in distributed primal-dual optimization. In *International Conference on Machine Learning(ICML)*, pages 1973–1982, 2015.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1273–1282, 2017.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

B. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

Sashank J Reddi, Jakub Konečnỳ, Peter Richtárik, Barnabás Póczós, and Alex Smola. AIDE: Fast and communication efficient distributed optimization. *arXiv preprint arXiv:1608.06879*, 2016.

Peter Richtárik and Martin Takáč. Distributed coordinate descent method for learning with big data. *Journal of Machine Learning Research*, 17(1):2657–2681, 2016.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *Annual Conference on Learning Theory (COLT)*, 2009.

Ohad Shamir. Without-replacement sampling for stochastic gradient methods. In *Advances in Neural Information Processing Systems (NIPS)*, pages 46–54, 2016.

Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International Conference on Machine Learning (ICML)*, pages 1000–1008, 2014.

Virginia Smith, Simone Forte, Ma Chenxin, Martin Takáč, Michael I Jordan, and Martin Jaggi. Cocoa: A general framework for communication-efficient distributed optimization. *Journal of Machine Learning Research*, 18:230, 2018.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Jialei Wang, Mladen Kolar, Nathan Srebro, and Tong Zhang. Efficient distributed learning with sparsity. In *International Conference on Machine Learning (ICML)*, pages 3636–3645, 2017a.

Jialei Wang, Weiran Wang, and Nathan Srebro. Memory and communication efficient distributed stochastic optimization with minibatch prox. In *Annual Conference on Learning Theory (COLT)*, pages 1882–1919, 2017b.

Shusen Wang, Farbod Roosta-Khorasani, Peng Xu, and Michael W Mahoney. Giant: Globally improved approximate newton method for distributed optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2338–2348, 2018.

Ashia C Wilson, Benjamin Recht, and Michael I Jordan. A lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635*, 2016.

Lin Xiao, Adams Wei Yu, Qihang Lin, and Weizhu Chen. Dscovr: Randomized primal-dual block coordinate algorithms for asynchronous distributed optimization. *Journal of Machine Learning Research*, 20(43):1–58, 2019.

Eric P Xing, Qirong Ho, Wei Dai, Jin Kyu Kim, Jinliang Wei, Seunghak Lee, Xun Zheng, Pengtao Xie, Abhimanu Kumar, and Yaoliang Yu. Petuum: A new platform for distributed machine learning on big data. *IEEE Transactions on Big Data*, 1(2):49–67, 2015.

Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. Apache spark: A unified engine for big data processing. *Commun. ACM*, 59(11):56–65, October 2016. ISSN 0001-0782.

Yuchen Zhang and Lin Xiao. DiSCO: Distributed optimization for self-concordant empirical loss. In *International Conference on Machine Learning (ICML)*, pages 362–370, 2015.

Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *Journal of Machine Learning Research*, 18(1):2939–2980, 2017.

Weijie Zhao, Jingyuan Zhang, Deping Xie, Yulei Qian, Ronglai Jia, and Ping Li. Aibox: CTR prediction model training on a single node. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 319–328, 2019.

Weijie Zhao, Deping Xie, Ronglai Jia, Yulei Qian, Ruiquan Ding, Mingming Sun, and Ping Li. Distributed hierarchical GPU parameter server for massive scale deep learning ads systems. In *Proceedings of Machine Learning and Systems (MLSys)*, 2020.

Pan Zhou, Xiaotong Yuan, and Jiashi Feng. Efficient stochastic gradient hard thresholding. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1984–1993, 2018.