Manifold Identification for Ultimately Communication-Efficient Distributed Optimization

Yu-Sheng Li [*1] **Wei-Lin Chiang** [*1] **Ching-pei Lee** [2]

Abstract

This work proposes a progressive manifold identification approach for distributed optimization with sound theoretical justifications to greatly reduce both the rounds of communication and the bytes communicated per round for partly-smooth regularized problems such as the ℓ_1- and group-LASSO-regularized ones. Our two-stage method first uses an inexact proximal quasi-Newton method to iteratively identify a sequence of low-dimensional manifolds in which the final solution would lie, and restricts the model update within the current manifold to gradually lower the order of the per-round communication cost from the problem dimension to the dimension of the manifold that contains a solution and makes the problem within it smooth. After identifying this manifold, we take superlinear-convergent truncated semismooth Newton steps computed by preconditioned conjugate gradient to largely reduce the communication rounds by improving the convergence rate from the existing linear or sublinear ones to a superlinear rate. Experiments show that our method can be orders of magnitudes lower in the communication cost and an order of magnitude faster in the running time than the state of the art.

1. Introduction

Distributed computing that simultaneously utilizes multiple machines is essential for dealing with the huge volume of data faced nowadays. Therefore, distributed optimization has become an active research topic in machine learning in recent years. In distributed optimization, the additional computing power and storage from multiple machines can greatly reduce the time for computation, memory access, and disk I/O. On the other hand, expensive inter-machine communication is frequently needed to synchronize the current model and calculate the update steps, so the communication cost, which can be an order of magnitude more expensive than local computation, usually becomes the bottleneck.

It is clear that the overall communication cost is proportional to both (i) the number of communication rounds and (ii) the cost per round, so reducing either factor can lower the overall expense on communication. State-of-the-art communication-efficient distributed optimization methods such as (Zhang & Xiao, 2015; Zheng et al., 2017; Lee et al., 2017; Lee & Chang, 2019; Lee et al., 2019) mainly focused on improving the former, which is closely related to the iteration complexity or the convergence speed. On the other hand, the reduction of the latter is usually achieved by compression or coding techniques that communicate inexact information using fewer bytes (Aji & Heafield, 2017; Lin et al., 2017; Wen et al., 2017; Chen et al., 2018). These approaches can harm the convergence speed of the optimization process due to the introduced noise, so the overall communication cost might not be reduced because of the increased communication rounds, and relevant study is restricted to specific problem classes with limited theoretical support. This work aims at decreasing both factors simultaneously through a two-stage progressive manifold identification approach with thorough theoretical guarantees.

We consider a distributed setting of minimizing the following regularized problem using K machines.

$$\min_{\boldsymbol{w}\in\Re^d} \quad F(\boldsymbol{w}) \coloneqq \sum_{k=1}^{K} f_k(\boldsymbol{w}) + \Psi(\boldsymbol{w}), \tag{1}$$

where each f_k is differentiable and exclusively available on the kth machine, their sum is Lipschitz-continuously-differentiable (sometimes called smooth) but might not be convex, and the regularization term Ψ is convex, proper, and closed but possibly nonsmooth. We assume further that (1) has a solution set $\Omega \neq \emptyset$ and Ψ is partly smooth (Lewis, 2002) everywhere. Loosely speaking, a convex function

[*]Equal contribution [1]Department of Computer Science, National Taiwan University, Taipei, Taiwan [2]Department of Mathematics & Institute for Mathematical Sciences, National University of Singapore, Singapore. Correspondence to: Yu-Sheng Li <r07922087@csie.ntu.edu.tw>, Wei-Lin Chiang <r06922166@csie.ntu.edu.tw>, Ching-pei Lee <leechingpei@gmail.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

Ψ is partly smooth at a point $\boldsymbol{w}$ relative to a $\mathcal{C}^2$ manifold $\mathcal{M}$ if $\Psi|_{\mathcal{M}}$ is $\mathcal{C}^2$ around $\boldsymbol{w}$ and for directions leaving $\mathcal{M}$, the change of Ψ is sharp. A more technical definition is in Definition 3.5. We focus on the case in which the manifolds are translates of subspaces of $\Re^d$ such that for some given matrices A_i, for each point $\boldsymbol{w}$ there is an index set $I_{\boldsymbol{w}}$ such that the manifold

$$\mathcal{M}_{\boldsymbol{w}} := \{\bar{\boldsymbol{w}} \mid A_i\bar{\boldsymbol{w}} = A_i\boldsymbol{w} \quad \forall i \in I_{\boldsymbol{w}}\} \tag{2}$$

contains $\boldsymbol{w}$ and makes $\Psi|_{\mathcal{M}_{\boldsymbol{w}}}$ $\mathcal{C}^2$ around $\boldsymbol{w}$. For example, when $\Psi(\cdot) = \|\cdot\|_1$, each A_i is the unit vector of the ith coordinate and $I_{\boldsymbol{w}} = \{i \mid \boldsymbol{w}_i = 0\}$.

An important class of partly smooth functions is the sparsity-inducing penalty functions such that Ψ is block-separable and for each block, the only point of nonsmoothness is the origin. Examples in machine learning include the ℓ_1 and the group-LASSO regularizations that promote sparsity (Tibshirani, 1996; Meier et al., 2008). Partly smoothness and a mild regularity condition (see (19)) ensures that an optimal solution lies in a low-dimensional smooth manifold, and identifying which helps us lower the communication cost.

We utilize partly smoothness to propose a two-stage progressive manifold identification approach for (1) that is ultimately communication-efficient and call it MADPQN (Manifold-Aware Distributed Proximal Quasi-Newton). The first stage is a trust-region-like distributed inexact proximal quasi-Newton method that predicts on the fly the manifold $\mathcal{M}^*$ in which a final solution would lie, and utilizes the current prediction to gradually reduce the communication cost from the $O(d)$ in existing work to $O(|\mathcal{M}^*|)$. Empirically, $|\mathcal{M}^*| \ll d$ is often observed. This stage conducts progressive manifold selection such that at each iteration, we identify a sub-manifold of the current one and confine the next iterate to it to keep the communication cost low. After the manifold becomes fixed, the second stage applies a superlinear-convergent truncated semismooth Newton (SSN) method with preconditioned conjugate gradient (PCG) on the problem restricted to the manifold. When the manifold contains an optimal solution, in contrast to the existing sublinear or linear rates, the superlinear convergence of the SSN steps greatly cuts the number of communication rounds while maintaining the $O(|\mathcal{M}^*|)$ per-round communication cost, making MADPQN even more communication-efficient, especially for obtaining high-precision solutions.

Contributions. This paper is the first to utilize manifold identification in distributed optimization for improving communication efficiency systematically. Unlike existing work that only switches to smooth optimization after the final manifold is identified but does nothing prior to that, our progressive strategy utilizes the intermediate manifolds identified in the process to reduce the costs of both computation and communication throughout. Our two-stage approach accelerates the optimization, whether the final manifold is of very low dimension or not. For the former case, the first stage makes the communication cost per round much smaller; for the latter, we reach the final manifold swiftly and the fast-convergent second stage then greatly cuts the number of communication rounds. In theory, unlike existing approaches that focus on only one factor, MADPQN has both lower per-round communication cost and fewer communication rounds such that the asymptotic communication complexity to achieve an ϵ-accurate solution for (1) can be as low as $O(|\mathcal{M}^*| \log\log(1/\epsilon))$, in contrast to the $O(d\log(1/\epsilon))$ or $O(d/\epsilon)$ cost of existing approaches. Experiments show that MADPQN can be more orders of magnitudes more communication-efficient than the state of the art and the running time can be an order of magnitude faster.

Organization. This paper is organized as follows. In Section 2, we give details of the proposed MADPQN algorithm. Section 3 then provides theoretical supports, including its convergence rate and the ability for manifold identification. Related research is reviewed in Section 4. Experiments in Section 5 showcase the empirical performance of MADPQN. Section 6 then concludes this work. Due to the space limit, proofs, implementation details, and additional experiments are all put in the appendices. Based on this study, we have released a package for use at `http://www.github.com/leepei/madpqn/`.

2. An Ultimately Communication-efficient Distributed Optimization Algorithm

MADPQN has three main components–(1) inexact proximal quasi-Newton, (2) progressive manifold selection, and (3) local acceleration through smooth optimization. Our two-stage method uses the first two components at the first stage, and at the second stage we switch to the third component. We will first introduce each component respectively, and then combine them into an integrated algorithm.

2.1. Where Do Communication Costs Occur?

Communication is the usual bottleneck in distributed optimization. We discuss where the major communication takes place. In our description, we assume an *all-reduce* model, meaning that each machine serves simultaneously as a master and a worker, such that each machine has its own local data, but all other information is global and all machines conduct the same operations.

We see that the summation in (1) requires distributing $\boldsymbol{w}$ to all machines, which takes a round of communication with cost up to $O(d)$. Summing the components up then takes an additional $O(1)$ communication cost.

In most efficient algorithms, we need to compute the gradient of the smooth part $f := \sum f_k$, which is of the form

$$\nabla f(\boldsymbol{w}) = \sum_{k=1}^{K} \nabla f_k(\boldsymbol{w}). \tag{3}$$

Since $\boldsymbol{w}$ has been made available on all machines when evaluating $F(\boldsymbol{w})$, computing (3) through all-reduce takes a round of $O(d)$ communication. Hence, the major communication at each iteration occurs when synchronizing $\boldsymbol{w}$ on all machines and computing the gradient, and each communicates a d-dimensional vector. A goal of our method design is reducing the size of the vectors to be communicated.

2.2. Distributed Inexact Proximal Quasi-Newton

We give an overview of the inexact distributed proximal quasi-Newton method (DPLBFGS) by Lee et al. (2019) and our modifications. This is the major building block of the first stage of our algorithm. We adopt DPLBFGS because of its fast convergence as a second-order method that can effectively reduce the communication rounds, and solving its subproblem requires little communication.

At the tth iteration, given the current iterate $\boldsymbol{w}^t$, DPLBFGS solves the following subproblem approximately to obtain a tentative update direction $\boldsymbol{p}$:

$$\min_{\boldsymbol{p}\in\mathbb{R}^d} Q_{H_t}(\boldsymbol{p};\boldsymbol{w}^t) := \nabla f(\boldsymbol{w}^t)^\top \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^\top H_t \boldsymbol{p} + \Psi(\boldsymbol{w}^t+\boldsymbol{p}) - \Psi(\boldsymbol{w}^t), \tag{4}$$

where H_t is the LBFGS approximation of the Hessian that we will give details below.

After $\boldsymbol{p}$ is obtained, we use the trust-region variant in (Lee et al., 2019) to ensure sufficient function decrease, as this approach can identify a manifold $\mathcal{M}^*$ containing a solution $\boldsymbol{w}^* \in \Omega$ to (1) within finite iterations (see Section 3). Given any symmetric and positive definite matrix H_t for defining the subproblem (4) and parameters $\sigma_1 \in (0,1]$ and $\theta > 1$, we accept the direction $\boldsymbol{p}$ as the update direction $\boldsymbol{p}^t$ if

$$F\left(\boldsymbol{w}^t+\boldsymbol{p}\right) \le F\left(\boldsymbol{w}^t\right) + \sigma_1 Q_{H_t}\left(\boldsymbol{p};\boldsymbol{w}^t\right); \tag{5}$$

otherwise we repeatedly update H_t by

$$H_t \leftarrow \theta H_t \tag{6}$$

and resolve (4) defined by the new H_t until (5) holds. The iterate is then updated by $\boldsymbol{w}^{t+1} = \boldsymbol{w}^t + \boldsymbol{p}^t$. We will denote the initial choice for H_t as $\tilde{H}_t$ and the final one as $\hat{H}_t$.

DPLBFGS uses the LBFGS approximation of the (generalized) Hessian (Liu & Nocedal, 1989) as $\tilde{H}_t$. Using the compact representation (Byrd et al., 1994), given a prespecified integer $m > 0$, let $m(t) := \min(m, t)$, and define

$$\boldsymbol{s}_i := \boldsymbol{w}^{i+1} - \boldsymbol{w}^i, \quad \boldsymbol{y}_i := \nabla f(\boldsymbol{w}^{i+1}) - \nabla f(\boldsymbol{w}^i), \quad \forall i,$$

Algorithm 1: MADPQN: An ultimately communication-efficient two-stage manifold identification method for (1)

Given $\theta > 1, \sigma_1 \in (0,1), \tilde{\epsilon}, \delta > 0$, integers $m, S, T > 0$, a sequence $\{\epsilon_j\} \ge \epsilon > 0$, an initial point $\boldsymbol{w}^{0,0}$

for $j = 0, 1, 2, \ldots$ **do**
  $U_{j,0} \leftarrow \emptyset, Z_{j,0} \leftarrow \emptyset, \tilde{\mathcal{M}}_{j,-1} = \tilde{\mathcal{M}}_{j,0} = \Re^d$
  for $t = 0, 1, 2, \ldots$ **do**
    Compute $P_{\tilde{\mathcal{M}}_{j,t-1}}(\nabla f(\boldsymbol{w}^{j,t}))$ by (13)
    if $t > 0$ **then**
      Compute $P_{\tilde{\mathcal{M}}_{j,t-1}}(\boldsymbol{y}_{t-1})$, and $\gamma_{j,t}$ by (15)
      if (14) *holds for* $i = t-1$ **then**
        Update $U_{j,t}$ by (8) and update $\tilde{Z}_{j,t}$ using $\tilde{Z}_{j,t-1}$ and $\boldsymbol{s}_{j,t-1}^\top U_t$
      end
      Select $\mathcal{M}_{j,t} \subset \mathcal{M}_{j,t-1}$ with $\boldsymbol{w}^{j,t} \in \mathcal{M}_{\boldsymbol{w}^{j,t}} \subseteq \mathcal{M}_{j,t}$ using $P_{\tilde{\mathcal{M}}_{j,t-1}}(\nabla f(\boldsymbol{w}^{j,t}))$ and decide a basis matrix $B_{j,t}$ of $\tilde{\mathcal{M}}_{j,t}$
    end
    if $\mathcal{M}_{j,t}$ *unchanged for* S *consecutive iterations (ignoring* $\mathcal{M}_{\tilde{j},0}$ *for all* $\tilde{j}$*), last update is not SSN, and* $\nabla^2 F|_{\mathcal{M}_{\boldsymbol{w}^{j,t}}}$ *is positive definite enough* **then**
      Compute a truncated SSN step $\boldsymbol{p}$ for $F|_{\mathcal{M}_{\boldsymbol{w}^{j,t}}}$ by PCG and line search
    end
    if $\mathcal{M}_{j,t}$ *unchanged for* S *consecutive iterations and last update is SSN* **then**
      $H_{j,t} \leftarrow \gamma_{j,t} I$
    end
    while *True* **do**
      if $H_{j,t}$ *is diagonal* **then**
        Solve (12) exactly to obtain $\boldsymbol{p} = B_{j,t}\boldsymbol{t}$ and the objective Q
      end
      Solve (12) approximately using SpaRSA to get $\boldsymbol{p} = B_{j,t}\boldsymbol{t}$ and the objective Q
      if $F(\boldsymbol{w}^{j,t}+\boldsymbol{p}) \le F(\boldsymbol{w}^{j,t}) + \sigma_1 Q$ **then**
        break
      end
      $H_{j,t} \leftarrow \theta H_{j,t}$
    end
    $\boldsymbol{w}^{j,t+1} \leftarrow \boldsymbol{w}^{j,t} + \boldsymbol{p}$
    if $|Q| \le \epsilon_j$ *and* $t \ge T$ **then**
      $\boldsymbol{w}^{j+1,0} \leftarrow \boldsymbol{w}^{j,t+1}$
      break
    end
  end
end

our $\tilde{H}_t$ is

$$\tilde{H}_t = \gamma_t I - U_t Z_t^{-1} U_t^\top, \tag{7}$$

where γ_t is bounded in a positive interval for all t, and

$$U_t \coloneqq [\gamma_t S_t, Y_t], \quad Z_t \coloneqq \begin{bmatrix} \gamma_t S_t^\top S_t, & L_t \\ L_t^\top & -D_t \end{bmatrix}, \tag{8}$$
$$S_t \coloneqq \left[\boldsymbol{s}_{t-m(t)}, \boldsymbol{s}_{t-m(t)+1}, \ldots, \boldsymbol{s}_{t-1}\right],$$
$$Y_t \coloneqq \left[\boldsymbol{y}_{t-m(t)}, \boldsymbol{y}_{t-m(t)+1}, \ldots, \boldsymbol{y}_{t-1}\right],$$
$$D_t \coloneqq \operatorname{diag}\left(S_t^\top Y_t\right), (L_t)_{i,j} \coloneqq \begin{cases} \left(S_t^\top Y_t\right)_{i,j} & \text{if } i > j, \\ 0 & \text{else.} \end{cases}$$

When $t = 0$, $\tilde{H}_0 = \gamma_0 I$ and we solve (4) exactly by simply applying one proximal operation.

To construct and utilize $\tilde{H}_t$ efficiently, especially when d is large, we pick a partition $\mathcal{J}_1, \ldots, \mathcal{J}_K$ of $\{1, \ldots, d\}$ and store $(U_t)_{\mathcal{J}_k,:}$ on the kth machine; since m is usually small, all machines keep a copy of the whole Z_t matrix.

If f is not strongly convex, it is possible that (7) is not positive definite, making (4) ill-conditioned. We thus follow Li & Fukushima (2001) to take a predefined $\delta > 0$ and discard all pairs of $(\boldsymbol{s}_i, \boldsymbol{y}_i)$ that do not satisfy $\boldsymbol{s}_i^\top \boldsymbol{y}_i \geq \delta \boldsymbol{s}_i^\top \boldsymbol{s}_i$.

For solving (4), as $\tilde{H}_t$ is generally not diagonal, there is no easy closed-form solution. Thus DPLBFGS uses an iterative solver SpaRSA (Wright et al., 2009), a proximal-gradient method with fast empirical convergence, to get an approximate solution to this subproblem. By treating (4) as another regularized problem with the smooth part being quadratic, at each round SpaRSA only requires computing the gradient of the smooth part and at most a couple of proximal operations and function value evaluations. These are all relatively cheap in comparison to calculating ∇f in general, as $\tilde{H}_t$ does not couple with the original function that might involve operating on a huge amount of data. The gradient of the smooth part at a subproblem iterate $\tilde{\boldsymbol{p}}$ is computed through

$$\nabla f(\boldsymbol{w}^t) + \tilde{H}_t \tilde{\boldsymbol{p}} = \nabla f(\boldsymbol{w}^t) + \gamma_t \tilde{\boldsymbol{p}} - U_t(Z_t^{-1}(U_t^\top \tilde{\boldsymbol{p}})). \tag{9}$$

The computation of $U_t^\top \tilde{\boldsymbol{p}}$ is distributed, through

$$U_t^\top \tilde{\boldsymbol{p}} = \sum_{k=1}^{K} (U_t)_{\mathcal{J}_k,:}^\top \tilde{\boldsymbol{p}}_{\mathcal{J}_k} \tag{10}$$

and a round of $O(m)$ communication. The $\mathcal{J}_k$ part of the gradient (9) is then computed locally on the kth machine as Z_t is maintained on all machines. Thus the iterate $\tilde{\boldsymbol{p}}$ is also computed and stored in a distributed manner according to the partition $\{\mathcal{J}_k\}$, consistent with the storage of U_t.

2.3. Progressive Manifold Selection

The usual bottleneck of DPLBFGS is the computation of ∇f, which involves going through the whole dataset and communicating a vector of length $O(d)$, and the latter can be much more costly in practice. However, as the trust-region DPLBFGS approach is able to identify a manifold $\mathcal{M}^*_{\boldsymbol{w}^*} \ni \boldsymbol{w}^*$ within finite iterations, it makes sense to confine all subsequent iterates to $\mathcal{M}^*_{\boldsymbol{w}^*}$, reducing the amount of communication per round from $O(d)$ bytes to $O(|\mathcal{M}^*_{\boldsymbol{w}^*}|)$.

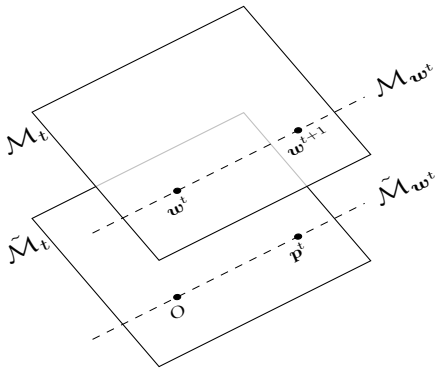


Figure 1. An example of updating within manifolds.

As the manifolds we consider in (2) are translates of subspaces of $\Re^d$, in the form

$$\mathcal{M}_{\boldsymbol{w}^t} = \boldsymbol{w}^t + \tilde{\mathcal{M}}_{\boldsymbol{w}^t} \tag{11}$$

for some subspace $\tilde{\mathcal{M}}_{\boldsymbol{w}^t}$, if we force $\boldsymbol{w}^i \in \mathcal{M}_{\boldsymbol{w}^t}$ for all $i > t$, we must have $\boldsymbol{p}^i \in \tilde{\mathcal{M}}_{\boldsymbol{w}^t}$ for all $i \geq t$. This is the central idea we use to reduce the communication and computation. At the tth iteration, with the currently manifold $\mathcal{M}_t$ selected by our algorithm that contains $\boldsymbol{w}^t$ and $\mathcal{M}_{\boldsymbol{w}^t}$ (the one making Ψ partly smooth around $\boldsymbol{w}^t$), we will select a sub-manifold $\mathcal{M}_{t+1} \subseteq \mathcal{M}_t$ and update $\boldsymbol{w}^{t+1}$ within $\mathcal{M}_{t+1}$. Equivalently, our algorithm selects a sequence of subspaces $\{\tilde{\mathcal{M}}_t\}$, and the update steps satisfy $\boldsymbol{p}^t \in \tilde{\mathcal{M}}_t$. Therefore, the subproblem (4) becomes $\min_{\boldsymbol{p} \in \tilde{\mathcal{M}}_t} Q_{H_t}(\boldsymbol{p}; \boldsymbol{w}^t)$. One thing to note is that $\mathcal{M}_t$ might not be the one that makes Ψ $\mathcal{C}^2$ at $\boldsymbol{w}^t$, which we denote by $\mathcal{M}_{\boldsymbol{w}^t}$, but we always have $\mathcal{M}_{\boldsymbol{w}^t} \subseteq \mathcal{M}_t$. We use a super-manifold of $\mathcal{M}_{\boldsymbol{w}^t}$ to allow flexibility because it is likely $\mathcal{M}_{\boldsymbol{w}^{t+1}} \nsubseteq \mathcal{M}_{\boldsymbol{w}^t}$, unless $\boldsymbol{w}^t$ is close to $\boldsymbol{w}^*$. An illustration is shown in Figure 1.

To really reduce the communication cost through the idea above, we need to conduct a change of basis first. As each $\tilde{\mathcal{M}}_t$ is a subspace of $\Re^d$, we can find an orthonormal basis $\{\boldsymbol{b}_{t,1}, \ldots, \boldsymbol{b}_{t,N(t)}\}$ of it. We denote $B_t \coloneqq [\boldsymbol{b}_{t,1}, \ldots, \boldsymbol{b}_{t,N(t)}]$ and approximately minimize the following equivalent problem

$$\begin{aligned} \min_{\boldsymbol{t} \in \Re^{N(t)}} \; & \nabla f(\boldsymbol{w}^t)^\top B_t \boldsymbol{t} + \frac{1}{2}(B_t \boldsymbol{t})^\top H_t B_t \boldsymbol{t} \\ & + \Psi(\boldsymbol{w}^t + B_t \boldsymbol{t}) - \Psi(\boldsymbol{w}^t), \end{aligned} \tag{12}$$

and only transmit the solution $\tilde{\boldsymbol{t}}$ to all machines for constructing $\boldsymbol{p} = B_t \tilde{\boldsymbol{t}}$. This reduces the communication cost for updating $\boldsymbol{w}^t$ from $O(d)$ to $O(N(t))$. When $\mathcal{M}_t = \mathcal{M}^*$,

we get $N(t) = O(|\mathcal{M}^*|)$ and the cost of this communication is greatly reduced. For the $\Psi(\boldsymbol{w}^t + B_t\boldsymbol{t})$ term in (12), the proximal operation seems harder to calculate, but since the basis is selected according to the structure of Ψ, this is not a problem. For the manifold selection, we use $\partial F|_{\mathcal{M}_t}(\boldsymbol{w}^{t+1})$ to decide how to shrink the current $\mathcal{M}_t$; see Appendix A.1 for more details.

Next, we reduce the communication cost for calculating the gradient. Let $P_{\tilde{\mathcal{M}}_t}$ be the Euclidean projection onto $\tilde{\mathcal{M}}_t$ and $\tilde{\mathcal{M}}_t^C$ be the complement space of $\tilde{\mathcal{M}}_t$. Because $\boldsymbol{v} = P_{\tilde{\mathcal{M}}_t}(\boldsymbol{v}) + P_{\tilde{\mathcal{M}}_t^C}(\boldsymbol{v})$ for any $\boldsymbol{v} \in \Re^d$, we always have

$$\nabla f(\boldsymbol{w}^t)^\top \boldsymbol{p} = P_{\tilde{\mathcal{M}}_t}(\nabla f(\boldsymbol{w}^t))^\top \boldsymbol{p}, \quad \forall \boldsymbol{p} \in \tilde{\mathcal{M}}_t.$$

Thus we only need $P_{\tilde{\mathcal{M}}_t}(\nabla f(\boldsymbol{w}^t))$ in (12) as $B_t\boldsymbol{t} \in \tilde{\mathcal{M}}_t$. On the other hand, we use $P_{\tilde{\mathcal{M}}_{t-1}}(\nabla f(\boldsymbol{w}^t))$ to select $\mathcal{M}_t$, and since $\tilde{\mathcal{M}}_t \subseteq \tilde{\mathcal{M}}_{t-1}$, we need to compute

$$P_{\tilde{\mathcal{M}}_{t-1}}(\nabla f(\boldsymbol{w}^t)) = B_{t-1}\sum_{i=1}^{K} B_{t-1}^\top \nabla f_k(\boldsymbol{w}^t). \tag{13}$$

The summation in (13) is where the communication occurs, and the bytes transmitted is now lowered to $O(N(t-1))$. Therefore, the communication cost per iteration has been reduced to $O(N(t-1) + N(t)) = O(N(t-1))$ as $N(t)$ is decreasing with respect to t. If B_j consists of columns of B_i for all $i \le j$, and the coordinates of the gradient are aligned with the columns in B_0 (as is the case when Ψ is a sparsity-inducing term like the ℓ_1 or $\ell_{2,1}$ norms), then $B_{t-1}^\top \nabla f_k(\boldsymbol{w}^t)$ is simply taking the coordinates of $\nabla f_k(\boldsymbol{w}^t)$ selected in B_{t-1}, so the computation is also cheaper than obtaining the whole $\nabla f(\boldsymbol{w}^t)$.

The remaining issue is to construct $\tilde{H}_t$ when we only have $P_{\tilde{\mathcal{M}}_{t-1}}(\nabla f(\boldsymbol{w}^t))$ and therefore $P_{\tilde{\mathcal{M}}_{t-1}}(\boldsymbol{y}_{t-1})$. Since $\tilde{H}_t$ appears in (12) in the form

$$B_t^\top \tilde{H}_t B_t = \gamma_t B_t^\top B_t - (B_t^\top U_t) Z_t^{-1} (B_t^\top U_t)^\top,$$

as long as we can construct Z_t correctly, $P_{\tilde{\mathcal{M}}_{t-1}}(\boldsymbol{y}_{t-1})$ is enough for obtaining $B_t^\top U_t$ and thus $B_t^\top \tilde{H}_t B_t$ as $\boldsymbol{s}_{t-1} = \boldsymbol{p}^{t-1} \in \tilde{\mathcal{M}}_{t-1}$. Note that because $\texttt{span}(B_{t+1}) = \tilde{\mathcal{M}}_{t+1} \subseteq \tilde{\mathcal{M}}_t = \texttt{span}(B_t)$, keeping $B_t^\top U_t$ alone is enough for updating $B_{t+1}^\top U_t$. To compute the correct Z_t with only $B_t^\top U_t$ available, the key is to reuse entries already appeared in Z_{t-1}. We thus only compute the newly appeared entries, which are $S_t^\top \boldsymbol{s}_{t-1}$ and $Y_t^\top \boldsymbol{s}_{t-1}$. As now we only compute $P_{\tilde{\mathcal{M}}_{t-1}}(\nabla f(\boldsymbol{w}^t))$, $P_{\tilde{\mathcal{M}}_{t-1}}(\boldsymbol{y}_{t-1})$ is used instead to update Z_t, and we call the matrix obtained this way $\tilde{Z}_t$. Section 3 will show that $\tilde{Z}_t = Z_t$, so our algorithm correctly maintains the LBFGS approximation and the corresponding fast convergence.

Our safeguard mechanism is also calculated using the components within the subspace only, as follows.

$$P_{\tilde{\mathcal{M}}_i}(\boldsymbol{s}_i)^\top P_{\tilde{\mathcal{M}}_i}(\boldsymbol{y}_i) \ge \delta P_{\tilde{\mathcal{M}}_i}(\boldsymbol{s}_i)^\top P_{\tilde{\mathcal{M}}_i}(\boldsymbol{s}_i). \tag{14}$$

For the choice of γ_t, we take a small $\tilde{\epsilon} > 0$ and set the following in our implementation.

$$\gamma_0 = \max\left\{\left|\frac{\nabla f(\boldsymbol{w}^0)^\top \nabla^2 f(\boldsymbol{w}^0) \nabla f(\boldsymbol{w}^0)}{\|\nabla f(\boldsymbol{w}^0)\|_2^2}\right|, \tilde{\epsilon}\right\},$$
$$\gamma_{t+1} = \begin{cases} \frac{P_{\tilde{\mathcal{M}}_t}(\boldsymbol{y}_t)^\top P_{\tilde{\mathcal{M}}_t}(\boldsymbol{y}_t)}{P_{\tilde{\mathcal{M}}_t}(\boldsymbol{y}_t)^\top P_{\tilde{\mathcal{M}}_t}(\boldsymbol{s}_t)} & \text{if (14)}, \\ \gamma_t & \text{otherwise}, \end{cases} \quad \forall t \ge 0. \tag{15}$$

2.4. Acceleration After Identifying the Active Set

The regularization Ψ we consider is partly smooth within $\mathcal{M}^*$ around a solution $\boldsymbol{w}^*$. Therefore, when $\mathcal{M}^*$ is correctly identified, the problem can be reduced to a smooth optimization problem by restricting the subsequent iterates within $\mathcal{M}^*$. We can therefore utilize smooth optimization algorithms that are both more efficient and faster in convergence. In particular, we use a truncated semismooth Newton (SSN) method that is superlinear-convergent when we are close to a solution (Qi & Sun, 1993), which is the case when $\mathcal{M}^*$ can be correctly identified. The idea of SSN is that since ∇f is Lipschitz continuous, f is twice-differentiable almost everywhere. Therefore, its generalized Hessian (denoted by $\nabla^2 f$) always exists (Hiriart-Urruty et al., 1984) and is hence used as the substitute of the real Hessian for computing the Newton step. When f is twice-differentiable, the algorithm reduces to the normal Newton method. We use preconditioned conjugate gradient (PCG) to approximately solve the SSN linear system and follow Hsia et al. (2018) to use the diagonal entries of $\nabla^2 f$ as the preconditioner. Backtracking line search is then applied to ensure sufficient objective decrease. To determine that we have found $\mathcal{M}^*$ to safely conduct SSN within the manifold, we assert $\mathcal{M}^* = \mathcal{M}_t$ when $\mathcal{M}_t$ remains unchanged for S consecutive iterations, for some predefined S.

To ensure convergence in case that $\nabla^2 f$ is not positive-definite, every time after an SSN step, we conduct a DPLBFGS step or a proximal gradient (PG) step (when we have entered the superlinear-convergent phase) on F restricted to $\mathcal{M}_t$. We also conduct manifold selection again after the DPLBFGS or PG step, and continue smooth optimization only when the manifold remains the same. More detailed description of our implementation is in Appendix A.2.

2.5. Safeguards

Our progressive manifold selection might sometimes be too aggressive and thus miss the correct manifold. To avoid this problem, we adopt a two-layer loop strategy. For any variable that we previously use t as its counter, now we use (j,t) to indicate that it is at the jth outer iteration and the tth inner iteration. At the beginning of the jth outer iteration, we discard $U_{j-1,t}$, $\tilde{Z}_{j-1,t}$ and the selected manifold to start afresh from the current iterate. The inner loop then conducts

the algorithm described above, until $|Q_{\hat{H}_{j,t}}(\boldsymbol{p}^{j,t};\boldsymbol{w}^{j,t})|$ is smaller than a predefined $\epsilon_j > 0$ and $t \geq T$ for some prespecified integer $T \geq 0$. We use an annealing strategy such that $\{\epsilon_j\}$ is a sequence decreasing to some $\epsilon > 0$. The overall algorithm is summarized in Algorithm 1.

3. Analysis

We first analyze the convergence of the proposed algorithm, and then show that $\mathcal{M}^*$ is identified in finite iterations. The latter justifies our approach of conducting progressive manifold identification on the fly and taking SSN steps after the manifold is fixed.

We remark that using the domain of $\Re^d$ in (1) is just for notational ease, and our algorithm and analysis extends directly to any Euclidean spaces.

3.1. Convergence

The following lemma shows that progressive manifold selection works correctly with the quasi-Newton approach.

Lemma 3.1. *For any $j = 0, 1, \ldots,$ under the progressive manifold identification scheme $\mathcal{M}_{j,t+1} \subseteq \mathcal{M}_{j,t}$ for all $t \geq 0$, $\tilde{Z}_{j,t} = Z_{j,t}$ always holds true. Moreover, $B_{j,t+1}^\top U_{j,t} = B_{j,t+1}^\top B_{j,t}(B_{j,t}^\top U_{j,t})$ for all i, t, so maintaining $B_{j,t}^\top U_{j,t}$ suffices for constructing (12).*

We now present the convergence rate of our algorithm, which provides an upper bound for the communication rounds. Parts of our analysis are extended from (Lee et al., 2019). We will use the following notation in this section.

$$Q_{H_{j,t}}^{j,t} := \min_{\boldsymbol{p}\in\tilde{\mathcal{M}}_{j,t}} \left(Q_{H_{j,t}}(\boldsymbol{p};\boldsymbol{w}^{j,t})\right).$$

Throughout the analysis, we assume that there is a fixed $\eta \in [0, 1)$ such that every time we solve (12) for $t > 0$, the solution $\boldsymbol{p}$ is η-approximate:

$$Q_{H_{j,t}}(\boldsymbol{p};\boldsymbol{w}^{j,t}) - Q_{H_{j,t}}^{j,t} \leq \eta(Q_{H_{j,t}}(\boldsymbol{0};\boldsymbol{w}^{j,t}) - Q_{H_{j,t}}^{j,t}), \tag{16}$$

and when $t = 0$, the solution is exact (as $H_{j,0} = \gamma_{j,0}I$). Condition (16) can easily be satisfied without the knowledge of $Q_{H_{j,t}}^{j,t}$, provided the eigenvalues of the quadratic term are bounded in a positive range and the subproblem solver is linear-convergent on strongly convex problems. See more discussions in (Lee & Wright, 2019; Lee et al., 2019). We thus assume (16) holds for all $t > 0$ with some fixed $\eta \in [0, 1)$, although its explicit value might be unknown. We first show that $\tilde{H}_{j,t}$ and $\hat{H}_{j,t}$ have bounded eigenvalues.

Lemma 3.2. *There are constants $\tilde{C}_1 \geq \tilde{C}_2 > 0$ such that $\tilde{C}_1 I \succeq \tilde{H}_{j,t} \succeq \tilde{C}_2 I$ for all j, t, and at every (j, t)th iteration, (6) is conducted at most a constant times before (5) is satisfied. Therefore, There are constants $C_1 \geq C_2 > 0$ such that $C_1 I \succeq \hat{H}_{j,t} \succeq \tilde{H}_{j,t} \succeq C_2 I$.*

We now show that the inner loops terminate finitely and give the convergence rate of the outer loop.

Lemma 3.3. *The jth inner loop ends in $o(1/\epsilon_j)$ steps. When f is convex and the quadratic growth condition*

$$g(\boldsymbol{w}) - \min_{\hat{\boldsymbol{w}}} g(\hat{\boldsymbol{w}}) \geq \min_{\boldsymbol{w}^*\in \mathit{argmin}\, g(\boldsymbol{w})} \mu\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2, \quad \forall \boldsymbol{w} \in \mathit{dom}(g), \tag{17}$$

for some fixed $\mu > 0$ is satisfied by $F|_{\mathcal{M}_{j,t}}$ for all j, t within the level set $\{\boldsymbol{w} \mid F|_{\mathcal{M}_{j,t}}(\boldsymbol{w}) \leq F|_{\mathcal{M}_{j,t}}(\boldsymbol{w}^{j,t})\}$, the inner loop ends in $O(\log(1/\epsilon_j))$ steps.

Theorem 3.4. *Consider (1), and fix $\epsilon > 0$. When f is convex, for reaching a point $\boldsymbol{w}$ with $F(\boldsymbol{w}) - F^* \leq \epsilon$, it takes $O(1/\epsilon)$ outer iterations if*

$$\max_{\boldsymbol{w}:F(\boldsymbol{w})\leq F(\boldsymbol{w}^{0,0})} \min_{\boldsymbol{w}^*\in\Omega} \|\boldsymbol{w} - \boldsymbol{w}^*\|_2 \tag{18}$$

is bounded, and $O(\log(1/\epsilon))$ if F satisfies (17). When f is nonconvex, it takes $o(1/\epsilon)$ outer iterations to have $|Q_{\hat{H}_{j,0}}(\boldsymbol{p}^{j,0};\boldsymbol{w}^{j,0})| \leq \epsilon$, and the first-order optimality $\boldsymbol{0} \in \partial F(\boldsymbol{w}^{j,0})$ holds if and only if $Q_{\hat{H}_{j,0}}(\boldsymbol{p}^{j,0};\boldsymbol{w}^{j,0}) = 0$.

3.2. Finite-time Manifold Identification

In what follows, we establish that the proposed algorithm can identify the correct manifold within finite time under the partly smoothness assumption (Lewis, 2002).

Definition 3.5 (Partly smooth functions)**.** *A convex function Ψ is partly smooth at $\boldsymbol{w}$ relative to a set $\mathcal{M}$ containing $\boldsymbol{w}$ if $\partial\Psi(\boldsymbol{w}) \neq \emptyset$ and:*

1. *Around $\boldsymbol{w}$, $\mathcal{M}$ is a $\mathcal{C}^2$-manifold and $\Psi|_{\mathcal{M}}$ is $\mathcal{C}^2$.*
2. *The affine span of $\partial\Psi(\boldsymbol{w})$ is a translate of the normal space to $\mathcal{M}$ at $\boldsymbol{w}$.*
3. *$\partial\Psi$ is continuous at $\boldsymbol{w}$ relative to $\mathcal{M}$.*

We first show that cluster points of the iterates are stationary.

Theorem 3.6. *All cluster points of $\{\boldsymbol{w}^{j,0}\}_{j=0}^\infty$ are critical. Moreover, if f is lower-bounded and Ψ is coercive, $\{\boldsymbol{w}^{j,0}\}_{j=0}^\infty$ has at least one cluster point.*

We now present the main result that $\mathcal{M}^*$ can be identified within finite iterations.

Theorem 3.7. *If a cluster point $\boldsymbol{w}^*$ of $\{\boldsymbol{w}^{j,0}\}$ satisfies*

$$-\nabla f(\boldsymbol{w}^*) \in \text{relint}\, \partial\Psi(\boldsymbol{w}^*), \tag{19}$$

where relint *denotes the relative interior, and Ψ is partly smooth relative to $\mathcal{M}^*$ around $\boldsymbol{w}^*$, then $\boldsymbol{w}^{j,1}$ identifies $\mathcal{M}^*$ in finite outer iterations. If moreover (12) is always solved exactly, then for j large enough and in the convergent subsequence, $\boldsymbol{w}^{j,t} \in \mathcal{M}^*$ for all $t > 0$.*

Remark 3.8. *Theorems 3.4 and 3.7 are applicable to the case in which F is nonconvex as long as Ψ is convex. Therefore, MADPQN is also applicable to nonconvex problems with a differentiable loss, including deep learning models.*

When the nondegenerate condition (19) fails, we are unable to identify $\mathcal{M}^*$. However, if Ψ is a mirror-stratifiable function (Fadili et al., 2018), we can still identify a superset manifold $\hat{\mathcal{M}}^*$ of $\mathcal{M}^*$, so MADPQN is still able to reduce the per-round communication cost to $O(|\hat{\mathcal{M}}^*|)$ at the first stage. However, it will not enter the second stage, so the overall communication cost is expected to be slightly higher, though still superior to existing approaches. See Appendix B.3 for the empirical effectiveness of the second stage.

3.3. Superlinear Convergence of the Second Stage

At the smooth optimization stage of MADPQN, eventually we alternate between a PG step and a truncated SSN step. If ∇f is semismooth (see Definition C.2), then we can obtain a two-step superlinear convergence of the iterates to the optimal solution when the truncated SSN steps are accurate enough. Due to the space limit, we leave the details of computing the truncated SSN steps and the superlinear convergence in Appendices A.2 and C.2.

4. Related Work

Manifold Identification. The major tool we use is manifold identification for partly smooth functions studied by Hare & Lewis (2004); Lewis (2002); Lewis & Zhang (2013). By applying their theory, many first-order algorithms are known to have the ability of manifold identification for partly smooth functions, including proximal-gradient-type methods (Nutini et al., 2017b; Liang et al., 2017a), the alternating direction method for multipliers (Liang et al., 2017b), regularized dual averaging (Lee & Wright, 2012), proximal coordinate descent (Wright, 2012; Nutini et al., 2017a), and variance reduction stochastic methods (Poon et al., 2018). However, these algorithms are usually not the choice for distributed optimization as non-stochastic first-order methods converge slowly in practice, and the frequent update of the iterate in stochastic methods incurs much more rounds of communication. Second-order algorithms tend to be superior for distributed optimization (see, *e.g.*, (Zhang & Xiao, 2015; Lee et al., 2019)), but current theory for their ability of manifold identification is restricted to the case in which (4) is solved exactly (Hare & Lewis, 2007; Hare, 2011; Sun et al., 2019). This work is the first to study and utilize such property in distributed optimization to improve communication efficiency. A key difference from all the works above is that instead of passively waiting until the final manifold is identified, our approach actively tries to find the manifold, and improves the computation and communication efficiency by utilizing from the beginning on the fact that the final solution lies in a lower-dimensional manifold.

Active Set Approaches. Another related idea is working-set or active-set methods for constrained problems in single-core optimization. These approaches force inequality constraints selected in the active-set to be at equality to save computation. Active-set methods can be extended to (1) when Ψ has an equivalent constrained form. For example, Yuan et al. (2012) used working set selection for ℓ_1-regularized logistic regression for proximal-Newton with coordinate descent as the subproblem solver, and Zhong et al. (2014) took a similar idea for ℓ_1-regularized M-estimators. These works showed that working-set heuristics can improve the empirical running time, but did not provide theoretical supports. Johnson & Guestrin (2015) considered a primal-dual approach for working set selection for ℓ_1-regularized problems with theoretical guarantees and nice practical performance, but it is hard to generalize the results to other regularizers. Our usage of progressive manifold identification is essentially a type of working set selection, but unlike existing works, MADPQN is fully supported by the theory of manifold identification and is applicable to all partly smooth regularizers, including but not limited to the ℓ_1 and group-LASSO norms. The stage of smooth optimization method is not present in the methods with working set heuristics either. Moreover, the purposes are totally different—these works use active sets to reduce their computation cost, while our progressive manifold identification is mainly for improving the communication efficiency.

Distributed Optimization for (1). For optimizing (1) in a distributed manner, although there are some heuristic approaches designed for specific regularizers that work well in practice, such as L-COMM (Chiang et al., 2018) and OWLQN (Andrew & Gao, 2007) for ℓ_1-regularized problems, the only method we are aware of that applies to general regularizers is the one proposed by Lee et al. (2019). They consider an inexact distributed proximal quasi-Newton method for (1) with theoretical convergence supports, and the first stage of MADPQN is extended from it. However, they did not utilize the partly smoothness of Ψ, so the communication cost per-round of their method is fixed to $O(d)$, while the communication cost per round of MADPQN is $O(N(t-1))$, which converges to $O(|\mathcal{M}^*|)$ rapidly. When $|\mathcal{M}^*| \ll d$, our method has a much lower communication cost per round (see Figure 3). Another difference between MADPQN and (Lee et al., 2019) is the truncated SSN steps, which greatly helps the convergence and reduces the number of communication rounds.

Distributed Optimization for Linear Models with Feature-wise Storage. For linear models such that $f_k(\boldsymbol{w}) = g_k(X_k^\top \boldsymbol{w})$ for some function g_k and some matrix X_k, instead of assuming each machine has exclusive access to an f_k, some works such as (Mahajan et al., 2017; Dünner et al., 2018) consider the scheme in which each node stores some rows of $X := [X_1, \ldots, X_K] \in \Re^{d\times n}$. These approaches communicate a vector of length $O(n)$ to synchronize $X^\top \boldsymbol{w}$ instead, so they might have a lower com-

munication cost when $n \ll d$. However, they do not work for nonlinear models. Even for linear models, $|\mathcal{M}^*| < n$ is often observed and these methods are unable to utilize the low-dimensional manifolds to reduce the communication cost. On the other hand, when $d \ll n$, their communication cost becomes prohibitively high, while MADPQN can handle both large d and large n.

5. Experiments

We conduct experiments to examine the empirical behavior of MADPQN on the Amazon cloud. For all experiments, we use 32 EC2 instances, each with one thread, connected through Open MPI. Convergence is compared through the relative difference to the optimal function value $(F(\boldsymbol{w}) - F^*)/F^*$, where F^* is the optimal objective value of (1), obtained approximately by solving the problem to a high precision.

We use the following fixed parameters for MADPQN throughout the experiments, and it works quite robustly, so there is no need to tune these parameters: $\theta = 2$, $m = 10$, $T = 2$, $S = 10$, $\sigma_1 = 10^{-4}$, $\tilde{\epsilon} = 10^{-10}$, $\delta = 10^{-10}$, $\epsilon = 10^{-14}$, $\epsilon_j = \max\{\epsilon, 10^{-4-3j}\}$.

Table 1. Data statistics for (20).

Data set	#instances (n)	#features (d)	$\lvert\mathcal{M}^*\rvert$
news20	19,996	1,355,191	506
webspam	350,000	16,609,143	793
avazu-site	25,832,830	999,962	11,867

Table 2. Data statistics for (21). Note that $d = \hat{d} \times c$.

Data set	n	$\hat{d}$	c	$\lvert\mathcal{M}^*\rvert$
sector.scale	6,412	55,197	105	94,920
rcv1-test.multiclass	518,571	47,236	53	386,529

5.1. Experiment on ℓ_1-regularized Problems

We first examine the performance of MADPQN on ℓ_1-regularized logistic regression, whose objective in (1) is

$$F(\boldsymbol{w}) = \|\boldsymbol{w}\|_1 + C\sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i \boldsymbol{w}^\top \boldsymbol{x}_i\right)\right), \quad (20)$$

where each $(\boldsymbol{x}_i, y_i) \in \Re^d \times \{-1, 1\}$ is an instance-label pair stored in one of the machines (so the sum over the data points on the kth machine forms f_k), and $C > 0$ is a given parameter to balance the two terms. For this problem, each A_i in (2) is the standard unit vector of the ith coordinate (so are the columns of $B_{j,t}$), $I_{\boldsymbol{w}} = \{i \mid \boldsymbol{w}_i = 0\}$, and $|\mathcal{M}^*|$ is the number of nonzero elements in the solution $\boldsymbol{w}^*$.

We consider public datasets listed in Table 1 (from http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets) and set $C = 1$ in all experiments. For news20, we use 32 t2.micro instances, while for the remaining two larger datasets, we use 32 r5n.large instances. We first examine the effectiveness of our progressive manifold selection. Table 1 shows that $|\mathcal{M}^*|$ can be much smaller than d, proving the effectiveness of utilizing manifold identification to reduce the communication cost per round. We further compare the time spent on communication of our method with and without progressive manifold selection in Figure 3, and see that the progressively selected manifold quickly approaches $\mathcal{M}^*$ and effectively keeps the cumulative communication cost low as the manifold dimension drops.

We then compare MADPQN with the following state-of-the-art methods for (20) implemented in C/C++.

- DPLBFGS (Lee et al., 2019)
- OWLQN (Andrew & Gao, 2007)
- L-COMM (Chiang et al., 2018)

For all of them, we use information from the past 10 iterations and set the step size shrinking parameter to 0.5 to have a fair comparison with MADPQN (equivalent to $m = 10$ and $\theta = 2$). Note that distributed first-order methods are not included in our comparison because they have been shown by Lee et al. (2019) (for proximal-gradient-type methods) and Dünner et al. (2018) (for stochastic methods) to be inferior to the methods we compare with. Comparison with the feature-wise approach by Dünner et al. (2018) and other additional experiments are in Appendix B.

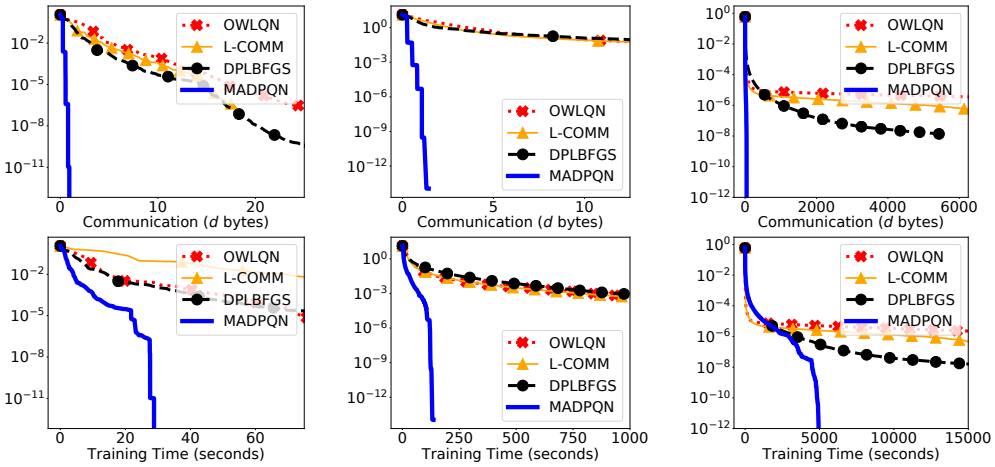
We compare the running time and the bytes communicated (divided by d) in Figure 2. We observe that the communication cost and the running time of MADPQN are both better than all other methods, although the difference in the running time is less significant. This is because that local computation still occupies a prominent portion of the total running time. Note that the manifold selection procedure in MADPQN takes less than 1% of the total running time so it is not the cause.

5.2. Experiment on Group-LASSO-regularized Problems

We next consider the group-LASSO-regularized multinomial logistic regression problem for multiclass classification:

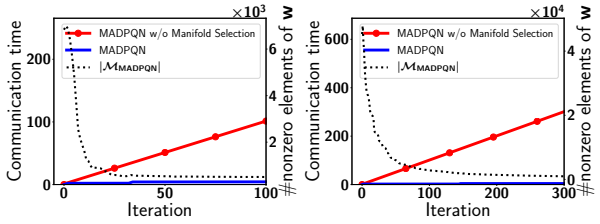
$$F(W) = \sum_{i=1}^{\hat{d}} \sqrt{\sum_{j=1}^{c} W_{i,j}^2} + C\sum_{i=1}^{n} l(W^\top \boldsymbol{x}_i; y_i), W \in \Re^{\hat{d} \times c}, \quad (21)$$

where $c > 0$ is the number of possible classes in the problem, each $(\boldsymbol{x}_i, y_i) \in \Re^{\hat{d}} \times \{1, \ldots, c\}$ is an instance-label



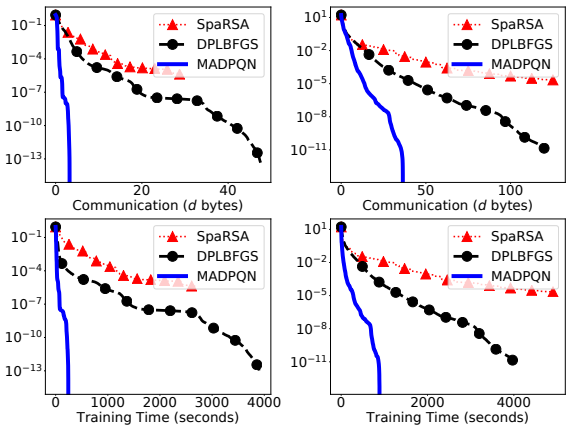
(a) news20 (b) webspam (c) avazu-site

Figure 2. Comparison of methods for (20). We present the number of bytes communicated divided by d (upper) and the training time (lower) vs. the relative difference to the optimal value.



(a) news20 (b) webspam

Figure 3. Iteration (x-axis) vs. total time spent on communication (solid lines, y-axis on the left) and the number of nonzero elements in the current $\boldsymbol{w}$ of MADPQN (dotted line, y-axis on the right).



(a) sector.scale (b) rcv1-test.multiclass

Figure 4. Comparison of methods for (21). We present the bytes communicated (upper) and the training time (lower) vs. the relative difference to the optimal value.

pair stored in one of the machines, and the loss term is defined as $l(\boldsymbol{z}; y) := -\log(\exp(z_y)/\sum_{i=1}^{c} \exp(z_i))$. For this problem, each A_i in (2) is the same as that for (20), $I_W = \{i \mid W_{i,j} = 0, \forall j\}$, and $|\mathcal{M}^*|$ is c times the number of rows in the solution W^* that contain at least one nonzero entry. Note that the problem dimension in this problem is $d = c \times \hat{d}$. For this problem, we compare MADPQN with DPLBFGS and SpaRSA (Wright et al., 2009) on the public datasets in Table 2, as OWLQN and L-COMM only apply to ℓ_1-regularized problems. The result of using 32 t2.micro instances is shown in Figure 4, and we see that MADPQN is significantly more efficient than existing approaches in both the communication cost and the running time.

6. Conclusions

In this work, we utilize manifold identification to greatly improve the communication efficiency and hence the overall running time of distributed optimization. Our algorithm cuts the bytes communicated per round through progressively selecting a sub-manifold where an optimal solution is likely to lie. After the correct manifold is identified, our algorithm further exploits partly smoothness of the problem to accelerate the convergence, and reduce the communication complexity by switching to a superlinear-convergent truncated semismooth Newton method. Experiments show that the overall communication cost and running time of our method are orders of magnitude lower than the state of the art. Future work includes the extension to nonlinear manifolds using Riemannian optimization, applying our algorithm to deep learning training with sparsity-inducing norms (Wen et al., 2016), and the analysis of manifold identification when the subproblems are solved inexactly.

Acknowledgement

This work was supported in part by the AWS Cloud Credits for Research program of Amazon Inc.

References

Aji, A. F. and Heafield, K. Sparse communication for distributed gradient descent. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 440–445, 2017.

Andrew, G. and Gao, J. Scalable training of L1-regularized log-linear models. In *Proceedings of the International Conference on Machine Learning*, 2007.

Byrd, R. H., Nocedal, J., and Schnabel, R. B. Representations of quasi-Newton matrices and their use in limited memory methods. *Mathematical Programming*, 63(1-3): 129–156, 1994.

Chen, T., Giannakis, G., Sun, T., and Yin, W. LAG: Lazily aggregated gradient for communication-efficient distributed learning. In *Advances in Neural Information Processing Systems*, pp. 5050–5060, 2018.

Chiang, W.-L., Li, Y.-S., Lee, C.-p., and Lin, C.-J. Limited-memory common-directions method for distributed l1-regularized linear classification. In *Proceedings of SIAM International Conference on Data Mining*, 2018.

Dünner, C., Lucchi, A., Gargiani, M., Bian, A., Hofmann, T., and Jaggi, M. A distributed second-order algorithm you can trust. In *Proceedings of the International Conference on Machine Learning*, 2018.

Fadili, J., Malick, J., and Peyré, G. Sensitivity analysis for mirror-stratifiable convex functions. *SIAM Journal on Optimization*, 28(4):2975–3000, 2018.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. LIBLINEAR: a library for large linear classification. *Journal of Machine Learning Research*, 9: 1871–1874, 2008.

Hare, W. Identifying active manifolds in regularization problems. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 261–271. Springer, 2011.

Hare, W. L. and Lewis, A. S. Identifying active constraints via partial smoothness and prox-regularity. *Journal of Convex Analysis*, 11(2):251–266, 2004.

Hare, W. L. and Lewis, A. S. Identifying active manifolds. *Algorithmic Operations Research*, 2(2):75–75, 2007.

Hiriart-Urruty, J.-B., Strodiot, J.-J., and Nguyen, V. H. Generalized Hessian matrix and second-order optimality conditions for problems with $C^{1,1}$ data. *Applied Mathematics & Optimization*, 11(1):43–56, 1984.

Hsia, C.-Y., Chiang, W.-L., and Lin, C.-J. Preconditioned conjugate gradient methods in truncated Newton frameworks for large-scale linear classification. In *Proceedings of the Asian Conference on Machine Learning*, 2018.

Johnson, T. and Guestrin, C. Blitz: A principled meta-algorithm for scaling sparse optimization. In *Proceedings of the International Conference on Machine Learning*, pp. 1171–1179, 2015.

Lee, C.-p. and Chang, K.-W. Distributed block-diagonal approximation methods for regularized empirical risk minimization. *Machine Learning*, 2019.

Lee, C.-p. and Wright, S. J. Inexact successive quadratic approximation for regularized optimization. *Computational Optimization and Applications*, 72:641–674, 2019.

Lee, C.-p., Wang, P.-W., Chen, W., and Lin, C.-J. Limited-memory common-directions method for distributed optimization and its application on empirical risk minimization. In *Proceedings of the SIAM International Conference on Data Mining*, 2017.

Lee, C.-p., Lim, C. H., and Wright, S. J. A distributed quasi-Newton algorithm for primal and dual regularized empirical risk minimization. Technical report, 2019. arXiv:1912.06508.

Lee, S. and Wright, S. J. Manifold identification in dual averaging for regularized stochastic online learning. *Journal of Machine Learning Research*, 13:1705–1744, 2012.

Lewis, A. S. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3):702–725, 2002.

Lewis, A. S. and Zhang, S. Partial smoothness, tilt stability, and generalized hessians. *SIAM Journal on Optimization*, 23(1):74–94, 2013.

Li, D.-H. and Fukushima, M. On the global convergence of the BFGS method for nonconvex unconstrained optimization problems. *SIAM Journal on Optimization*, 11 (4):1054–1064, 2001.

Liang, J., Fadili, J., and Peyré, G. Activity identification and local linear convergence of forward–backward-type methods. *SIAM Journal on Optimization*, 27(1):408–437, 2017a.

Liang, J., Fadili, J., and Peyré, G. Local convergence properties of douglas–rachford and alternating direction method of multipliers. *Journal of Optimization Theory and Applications*, 172(3):874–913, 2017b.

Lin, Y., Han, S., Mao, H., Wang, Y., and Dally, W. J. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *NIPS Workshop on ML Systems*, 2017.

Liu, D. C. and Nocedal, J. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

Mahajan, D., Keerthi, S. S., and Sundararajan, S. A distributed block coordinate descent method for training ℓ_1 regularized linear classifiers. *Journal of Machine Learning Research*, 18:1–35, 2017.

Meier, L., Van De Geer, S., and Bühlmann, P. The group LASSO for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70 (1):53–71, 2008.

Mifflin, R. Semismooth and semiconvex functions in constrained optimization. *SIAM Journal on Control and Optimization*, 15(6):959–972, 1977.

Nocedal, J. and Wright, S. *Numerical Optimization*. Springer, second edition, 2006.

Nutini, J., Laradji, I., and Schmidt, M. Let's make block coordinate descent go fast: Faster greedy rules, message-passing, active-set complexity, and superlinear convergence. Technical report, 2017a. arXiv:1712.08859.

Nutini, J., Schmidt, M., and Hare, W. "active-set complexity" of proximal gradient: How long does it take to find the sparsity pattern? In *NIPS Workshop on Optimization for Machine Learning*, 2017b.

Poon, C., Liang, J., and Schönlieb, C.-B. Local convergence properties of SAGA/prox-SVRG and acceleration. In *Proceedings of the International Conference on Machine Learning*, 2018.

Qi, L. and Sun, J. A nonsmooth version of Newton's method. *Mathematical programming*, 58(1-3):353–367, 1993.

Rockafellar, R. T. and Wets, R. J.-B. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Shi, W., Ling, Q., Wu, G., and Yin, W. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

Sun, Y., Jeong, H., Nutini, J., and Schmidt, M. Are we there yet? manifold identification of gradient-related proximal methods. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 1110–1119, 2019.

Tibshirani, R. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B*, 58:267–288, 1996.

Wen, W., Wu, C., Wang, Y., Chen, Y., and Li, H. Learning structured sparsity in deep neural networks. In *Advances in neural information processing systems*, pp. 2074–2082, 2016.

Wen, W., Xu, C., Yan, F., Wu, C., Wang, Y., Chen, Y., and Li, H. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in neural information processing systems*, pp. 1509–1519, 2017.

Wright, S. J. Accelerated block-coordinate relaxation for regularized optimization. *SIAM Journal on Optimization*, 22(1):159–186, 2012.

Wright, S. J., Nowak, R. D., and Figueiredo, M. A. T. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

Yuan, G.-X., Ho, C.-H., and Lin, C.-J. An improved GLMNET for L1-regularized logistic regression. *Journal of Machine Learning Research*, 13:1999–2030, 2012.

Zhang, Y. and Xiao, L. DiSCO: Distributed optimization for self-concordant empirical loss. In *International Conference on Machine Learning*, pp. 362–370, 2015.

Zheng, S., Wang, J., Xia, F., Xu, W., and Zhang, T. A general distributed dual coordinate optimization framework for regularized loss minimization. *Journal of Machine Learning Research*, 18:1–52, 2017.

Zhong, K., Yen, I. E.-H., Dhillon, I. S., and Ravikumar, P. K. Proximal quasi-newton for computationally intensive $l1$-regularized M-estimators. In *Advances in Neural Information Processing Systems*, 2014.

A. Implementation Details

In this appendix, we give more implementation details of our algorithm.

A.1. Manifold Selection

The manifold selection rule for $\mathcal{M}_{j,t}$ can directly affect the performance of our algorithm. If the selection is not aggressive enough, the dimension of the manifolds and therefore the per-iteration communication cost may remain large for many iterations. On the other hand, if the selection is too aggressive, it is possible that from very early on, $\mathcal{M}_{j,t}$ does not contain the correct manifold $\mathcal{M}^*$ as a submanifold anymore, and all subsequent inner iterations would give only limited objective improvement as no optimal solution is contained in the manifold we are confined to. Therefore, we want to have some rules that become more aggressive when we are certain that the current iterate is close to a solution. To achieve this, we use an adaptive strategy.

We consider the situation such that there is a known basis $\{\boldsymbol{b}^j_{t-1,i}\}_{i=1}^{N(t-1)}$ of the subspace $\tilde{\mathcal{M}}_{j,t-1}$ (see (11)) and the manifold selection procedure for $\tilde{\mathcal{M}}_{j,t}$ is simply picking vectors from this basis. Our first goal is to ensure that $\mathcal{M}_{\boldsymbol{w}^{j,t}}$ (the manifold that makes Ψ smooth around $\boldsymbol{w}^{j,t}$) is always included in $\mathcal{M}_{j,t}$ (the manifold we currently select), so that at least we can conduct smooth optimization within a subset of $\mathcal{M}_{j,t}$. Thus, our first step is to include the vectors that form a basis of $\tilde{\mathcal{M}}_{\boldsymbol{w}^{j,t}}$. We then examine the remaining vectors. The identification theorem (Theorem 3.7) shows that the requirement for manifold identification is $-\nabla f(\boldsymbol{w}^*) \in \operatorname{relint} \partial\Psi(\boldsymbol{w}^*)$. Therefore, for each $\boldsymbol{b}^j_{t-1,i}$ not selected in the first step, we examine the distance between $-\nabla_{\boldsymbol{b}^j_{t-1,i}} f(\boldsymbol{w}^{j,t})$ and the boundary of $\partial_{\boldsymbol{b}^j_{t-1,i}} \Psi(\boldsymbol{w}^{j,t})$, which we denote by $\operatorname{bdd}(\partial_{\boldsymbol{b}^j_{t-1,i}} \Psi(\boldsymbol{w}^{j,t}))$. First of all, if

$$-\nabla_{\boldsymbol{b}^j_{t-1,i}} f(\boldsymbol{w}^{j,t}) \notin \partial_{\boldsymbol{b}^j_{t-1,i}} \Psi(\boldsymbol{w}^{j,t}) \quad \Leftrightarrow \quad \mathbf{0} \notin \partial_{\boldsymbol{b}^j_{t-1,i}} F(\boldsymbol{w}^{j,t}), \tag{22}$$

it means that 0 is not in the directional derivative along this direction, so for sure we should still include this vector in our next basis. On the other hand, if (22) does not hold, we see how likely it would turn to hold in the future iterations. More specifically, we give a threshold $\xi_{j,t} > 0$ at the (j,t)th iteration, and if

$$\mathbf{0} \in \partial_{\boldsymbol{b}^j_{t-1,i}} F(\boldsymbol{w}^{j,t}), \quad \text{and} \quad \operatorname{dist}\left(-\nabla \boldsymbol{b}^j_{t-1,i} f(\boldsymbol{w}^{j,t}), \operatorname{bdd}(\partial_{\boldsymbol{b}^j_{t-1,i}} \Psi(\boldsymbol{w}^{j,t}))\right) < \xi_{j,t}, \tag{23}$$

we deem that (22) might turn to hold again when we get closer to a solution, meaning that $\tilde{\mathcal{M}}^*$ might also include it, and thus this vector $\boldsymbol{b}^j_{t-1,i}$ is also included in the basis for constructing $\tilde{\mathcal{M}}_{j,t}$. Now to decide the value of $\xi_{j,t}$, we want it to gradually decrease to 0 when we approach a critical point. Thus, we make it proportional to the subproblem objective $|Q_{\hat{H}_{j,t-1}}(\boldsymbol{p}^{j,t-1}; \boldsymbol{w}^{j,t-1})|$, which is an indicator for how much we can still improve the objective value within the current manifold as shown in Theorem 3.4. More specifically, we set

$$\xi_{j,t} = \xi_0 \min\left\{1, \frac{|Q_{\hat{H}_{j,t-1}}(\boldsymbol{p}^{j,t-1}; \boldsymbol{w}^{j,t-1})|}{|Q_{\hat{H}_{j,0}}(\boldsymbol{p}^{0,0}; \boldsymbol{w}^{0,0})|}\right\}$$

for some ξ_0. This way, it converges to 0 asymptotically as shown in Theorem 3.4, and thus $\mathcal{M}_{j,t}$ converges $\mathcal{M}^*_{j,t}$. In our implementation, we let

$$\xi_0 = d^{-1}.$$

To summarize, the vectors $\boldsymbol{b}^j_{t-1,i}$ in the basis of $\tilde{\mathcal{M}}_{j,t-1}$ that satisfy (22) or (23) and those that are contained in $\tilde{\mathcal{M}}_{j,t-1}$ are selected as the basis for $\tilde{\mathcal{M}}_{j,t}$ and therefore to construct the next manifold selected.

As mentioned in Section 2.3, we use $P_{\tilde{\mathcal{M}}_{j,t-1}}(\nabla f(\boldsymbol{w}^{j,t}))$ to decide the next subspace $\tilde{\mathcal{M}}_{j,t}$. Therefore, the gradient calculation has a slightly higher communication cost than updating the iterate.

A.2. Preconditioned Conjugate Gradient and Backtracking Line Search for Truncated Semismooth Newton Steps

When we are certain that the manifold does not change anymore, we start the smooth optimization stage of using truncated semismooth Newton steps. We will sometimes call them Newton steps or SSN steps for short. When an SSN step is computed, we first check the largest step size we can take without passing through a point of nonsmoothness of Ψ. This is the initial step size α_0 we take, and then we do a backtracking line search for ensuring sufficient function decrease starting

from $\min\{1, \alpha_0\}$. Assume that the current iterate is $\boldsymbol{w}$, the corresponding manifold that makes F smooth around $\boldsymbol{w}$ is $\mathcal{M}$, its parallel subspace is $\tilde{\mathcal{M}}$, and the SSN step is $\boldsymbol{p}$. Given parameters $\sigma, \beta \in (0, 1)$, we find the smallest nonnegative integer i such that

$$F\left(\boldsymbol{w} + \alpha_0 \beta^i \boldsymbol{p}\right) \leq F\left(\boldsymbol{w}\right) + \sigma \alpha_0 \beta^i \left(\nabla F|_{\mathcal{M}}(\boldsymbol{w})^\top \boldsymbol{p} + \frac{1}{2} \boldsymbol{p}^\top \nabla^2 F|_{\mathcal{M}}(\boldsymbol{w}) \boldsymbol{p}\right), \tag{24}$$

Note that since f is Lipschitz-continuously differentiable and $\Psi|_{\mathcal{M}}$ is $\mathcal{C}^2$ around $\boldsymbol{w}$, a generalized Hessian $\nabla^2 F|_{\mathcal{M}}(\boldsymbol{w})$ always exists (Hiriart-Urruty et al., 1984). When there are multiple possible generalized Hessians, we pick an arbitrary one. The final step size is then set to $\alpha_0 \beta^i$. If α_0 is too small, or this decrease condition cannot be satisfied even when $\alpha_0 \beta^i$ is already extremely small, we discard this smooth optimization step and turn back to proximal quasi-Newton steps.

The truncated semismooth Newton step is obtained by applying preconditioned conjugate gradient (PCG) to find an approximate solution $\boldsymbol{p} \in \tilde{\mathcal{M}}$ to the linear system

$$\nabla^2 F|_{\mathcal{M}}(\boldsymbol{w}) \boldsymbol{p} = -\nabla F|_{\mathcal{M}}(\boldsymbol{w}),$$

or equivalently, if B is an orthonormal basis of $\tilde{\mathcal{M}}$,

$$B^\top \nabla^2 F|_{\mathcal{M}}(\boldsymbol{w}) B \boldsymbol{t} = -B^\top \nabla F|_{\mathcal{M}}(\boldsymbol{w}), \quad \boldsymbol{p} = B\boldsymbol{t}.$$

More details of PCG can be found in, for instance, Nocedal & Wright (2006, Chapter 7). Note that we take the diagonal entries $B^\top \nabla^2 F|_{\mathcal{M}}(\boldsymbol{w}) B$ as the preconditioner in PCG.

We know that if $\boldsymbol{w}$ is close enough to a solution point $\boldsymbol{w}^* \in \Omega$, the generalized Hessian at $\boldsymbol{w}^*$ is positive definite, and $\boldsymbol{w}$ and $\boldsymbol{w}^*$ are in the same manifold, then unit step size and superlinear convergence can be expected. However, those conditions are not verifiable in practice, and it is possible that after taking many PCG iterations to compute a highly accurate truncated semismooth Newton step, we only get a very small step size, resulting in a limited objective decrease with disproportionately expensive computation. Therefore, before letting PCG run for lengthy iterations, we start from something with a much lower cost. When any of the following conditions holds, we terminate PCG when it reaches an iteration bound P that gradually increases from some integer P_0.

1. When there is no previous Newton step.
2. When $\mathcal{M}_{j,t}$, the current selected manifold, differs from $\mathcal{M}_{\boldsymbol{w}^{j,t}}$, the manifold that makes Ψ smooth around the current iterate $\boldsymbol{w}^{j,t}$. This situation indicates that the manifold that makes Ψ smooth is likely to change in later iterations.
3. When the last SSN step results in a step size smaller than 1.

If none of these conditions holds, we take a factor $\zeta > 1$ and set

$$P \leftarrow \min\{\zeta P, |\tilde{\mathcal{M}}_{\boldsymbol{w}^{j,t}}|\},$$

as in theory PCG finds the exact solution of the Newton linear system within $|\tilde{\mathcal{M}}_{\boldsymbol{w}^{j,t}}|$ steps. See, for example, Nocedal & Wright (2006, Chapter 7). In our implementation, we set $P_0 = 5$ and $\zeta = 10$.

We do not know in practice whether $H = B^\top \nabla^2 F|_{\mathcal{M}}(\boldsymbol{w}) B$ is positive definite or not as well. When it is not positive definite at $\boldsymbol{w}^*$, superlinear convergence does not happen. Even worse, PCG might not be able to obtain a solution for the linear system (as there might be no solution at all). Theoretically, we should check the smallest eigenvalue of H first and when it is too small, we should not try to compute a Newton step. However, computing the smallest eigenvalue can be too expensive. Alternatively, we use an early exit mechanism and a trust-region technique in PCG to avoid such cases. At the Tth PCG iteration, two vectors are generated: the current approximate solution $\boldsymbol{t}^{(T)}$ to the linear system and the direction $\boldsymbol{s}^{(T)}$ for updating $\boldsymbol{t}^{(T)}$. For some small threshold $\tau > 0$, we exit PCG if

$$\frac{\left(\boldsymbol{t}^{(T)}\right)^\top H \boldsymbol{t}^{(T)}}{\left(\boldsymbol{t}^{(T)}\right)^\top \boldsymbol{t}^{(T)}} \leq \tau, \quad \text{or} \quad \frac{\left(\boldsymbol{s}^{(T)}\right)^\top H \boldsymbol{s}^{(T)}}{\left(\boldsymbol{s}^{(T)}\right)^\top \boldsymbol{s}^{(T)}} \leq \tau,$$

which implies that indeed H has an eigenvalue no larger than τ. In our implementation, we set $\tau = 10^{-8}$.

The trust-region technique uses the fact (straightforward from the KKT conditions) that given any vector $\boldsymbol{a}$ and any symmetric matrix H, for any $\Delta > 0$, there exists a $\lambda \geq 0$ such that $H + \lambda I$ is positive definite and the solution to

$$\min_{\boldsymbol{t}:\|\boldsymbol{t}\|_2 \leq \Delta} \boldsymbol{a}^\top \boldsymbol{t} + \frac{1}{2}\boldsymbol{t}^\top H \boldsymbol{t} \tag{25}$$

is the same as the solution to

$$\min_{\boldsymbol{t}} \boldsymbol{a}^\top \boldsymbol{t} + \frac{1}{2}\boldsymbol{t}^\top (H + \lambda I)\boldsymbol{t}.$$

PCG solves the former problem without the trust region constraint, so incorporating it with a trust-region technique takes more observations. Classical analysis for PCG shows that $\|\boldsymbol{t}^{(T)}\|$ increases monotonically with T, and when H is not positive definite, it is not hard to see that $\|\boldsymbol{t}^{(T)}\|$ increases to infinity. Therefore, we place an upper bound Δ such that when $\|\boldsymbol{t}^{(T)}\| > \Delta$, we terminate PCG and use $B\boldsymbol{t}^{(T)}$ as the final truncated semismooth Newton step. Each time we set Δ to be large enough such that it does not change the solution unless the smallest eigenvalue of H is too small.

For the safeguards above, when PCG is terminated early, we still use the obtained approximate solution as a possible update direction and conduct the backtracking line search procedure first to see if we can get sufficient objective decrease, as line search is usually much cheaper than the computation of an update direction.

Except for the above safeguards, we terminate PCG when

$$\left\|\nabla^2 F|_{\mathcal{M}}(\boldsymbol{w})\boldsymbol{p} + \nabla F|_{\mathcal{M}}(\boldsymbol{w})\right\| \leq 0.1 \min\left\{1, \|\nabla F|_{\mathcal{M}}(\boldsymbol{w})\|^2\right\} \tag{26}$$

to obtain superlinear convergence; see Theorem C.3 for more details.

Assume that at the (j,t)th iteration a Newton step is taken. When PCG is not terminated early, and the final accepted step size is 1, at the $(j,t+1)$th iteration, we take the next update to be a simple proximal gradient step, with the initial step size being $\gamma_{j,t+1}$ defined in (15). That is, we obtain the update by

$$\boldsymbol{p}^{j,t+1} = \arg\min_{\boldsymbol{p}\in\bar{\mathcal{M}}_{j,t+1}} \nabla f(\boldsymbol{w}^{j,t+1})^\top \boldsymbol{p} + \frac{\gamma_{j,t+1}}{2}\boldsymbol{p}^\top \boldsymbol{p} + \Psi(\boldsymbol{w}^{j,t+1} + \boldsymbol{p}) \tag{27}$$

and iteratively enlarge $\gamma_{j,t+1}$ through (6) and resolve (27) until (5) holds. Otherwise, we continue the next step using the proximal quasi-Newton procedure. This is because that when PCG terminates normally and the step size is one, likely we are entering the superlinear convergence phase so the proximal step is just a safeguard and we do not wish to spend too much time on it, and thus we take a proximal gradient but not a much more expensive proximal quasi-Newton step.

A.3. Coordinate Random Shuffling

In Section 2.2, we described that the rows of U_t are stored in a distributed manner, according to a partition $\mathcal{J}_1, \ldots, \mathcal{J}_K$ of $\{1, \ldots, d\}$. To facilitate the computation of (10), which is the most expensive one in the subproblem solve, we should make each $\mathcal{J}_i$ of equal size so that each machine has an equal computational burden, assuming all machines have similar computational power. (When the computational environment is heterogeneous, the burden distribution should be proportional to the computational power of the machines.) In this sense, it is most straightforward to let each $\mathcal{J}_k$ be a set of consecutive numbers as this also improves memory locality for accessing other information in the memory. However, when manifold selection is involved, it is possible that the coordinates are ordered in a specific, non-random way such that consecutive coordinates are more likely to be correlated, so that they tend to be included in or excluded from the current manifold simultaneously. This can make the computation burden distribution unbalanced. To even out the computational burden, we apply coordinate random shuffling in assigning the partition. That is, we first randomly shuffle $\{1, \ldots, d\}$ and then let each $\mathcal{J}_k$ take consecutive elements of the shuffled set. This way, the computational burden on each machine becomes more balanced in expectation, both before and after manifold selection; experiments in Section B.4 verify this claim.

B. Additional Experiments and Setting Details

This appendix provides additional experiments and setting details for our and other algorithms. We first give implementation details of other solvers we compare with in Section B.1. In Section B.2, comparison on more datasets is conducted. Section B.3 examines the effectiveness of the two stages of MADPQN. Coordinate shuffling discussed in Section A.3 is then scrutinized in Section B.4. Finally, we compare MADPQN with a state-of-the-art feature-wise storage method (that works for convex and linear models only) in Section B.5.

Table 3. Additional data statistics.

Data set	#instances (n)	#features (d)	$\lvert\mathcal{M}^*\rvert$
rcv1-test	677,399	47,226	5,095
yahoo-japan	176,203	832,026	3,182
yahoo-korea	460,554	3,052,939	9,818
epsilon	400,000	2,000	1,463
url	2,396,130	3,231,961	25,399
KDD2010-b	19,264,097	29,890,096	2,007,370

B.1. Implementation Details of Other Solvers Compared

In Section 5, several methods are used to solve (20). To make a fair comparison, all methods are implemented in C++ and MPI. Here we give more details about the implementations and parameters used in these methods.

- OWLQN (Andrew & Gao, 2007) and L-COMM (Chiang et al., 2018): We use the implementation in the experimental codes of Chiang et al. (2018).[1] $m = 10$ and $\theta = 2$ are used for both methods.
- DPLBFGS (Lee et al., 2019): We use the experimental codes released by the authors. $m = 10$ and $\theta = 2$ are used.[2]

For MADPQN and DPLBFGS, we take the maximum SpaRSA steps for solving the subproblem to be 100.

B.2. Comparisons on More Datasets

In addition to the datasets listed in Table 1, we further conduct experiments on more datasets listed in Table 3.[3] We use 32 t2.micro instances for the first three smaller datasets, and 32 r5n.large instances for the rest in order to fit the whole dataset into the main memory. The results are presented in Figure 5. Similar to the observation in Section 5.1, our proposed method is always faster than others in terms of the communication cost. Because $\lvert\mathcal{M}^*\rvert \approx d$ for rcv1-test and epsilon (see Table 3), the difference between MADPQN and others on these two is less significant, but MADPQN is still the most communication-efficient. Regarding the training time, our method is generally the fastest to achieve nearly optimal solutions, with the only exception on rcv1-test, while other methods suffer from slower convergence especially on difficult problems such as KDD2010-b (and avazu-site in Figure 2). It is also worth noting that especially for rcv1-test and url, the difference in the training time between MADPQN and others is much smaller than that in their communication costs, indicating that the proportion of communication costs in the overall running time is smaller and the running time improvement from reducing the communication is thus limited. This also explains why although MADPQN is the most communication-efficient for rcv1-test, its running time is not the fastest. However, even under these settings that do not favor MADPQN, it is still competitive with the state of the art. Moreover, we expect to see more improvement if more machines are used so that the computation burden per machine is lighter and the communication cost becomes relatively significant. Overall speaking, our method is faster than others in most datasets and enjoy better convergence in the later stage of optimization. Even when the setting does not favor our method, MADPQN is still at least competitive, and in other settings, MADPQN outperforms the state of the art significantly.

B.3. Experiments on the Effectiveness of Progressive Manifold Selection and Smooth Optimization

In Sections 2.3, 2.4, A.1 and A.2, we have introduced the progressive manifold identification strategy and smooth optimization acceleration to reduce the communication costs and rounds. Now we use (20) to examine the effectiveness of these two techniques, respectively, by comparing MADPQN without manifold selection (and thus no smooth optimization as well), MADPQN without (semismooth) Newton steps, and MADPQN. We call the first variant MADPQN-noMS, and the second MADPQN-noN. The difference between MADPQN-noMS, and MADPQN-noN indicates the effectiveness of manifold selection, while the comparison between MADPQN-noN and MADPQN shows how the Newton steps accelerate the

[1] https://www.csie.ntu.edu.tw/~cjlin/papers/l-commdir-l1/

[2] https://github.com/leepei/dplbfgs

[3] yahoo-japan and yahoo-korea are two document binary classification datasets, and all others are downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets.

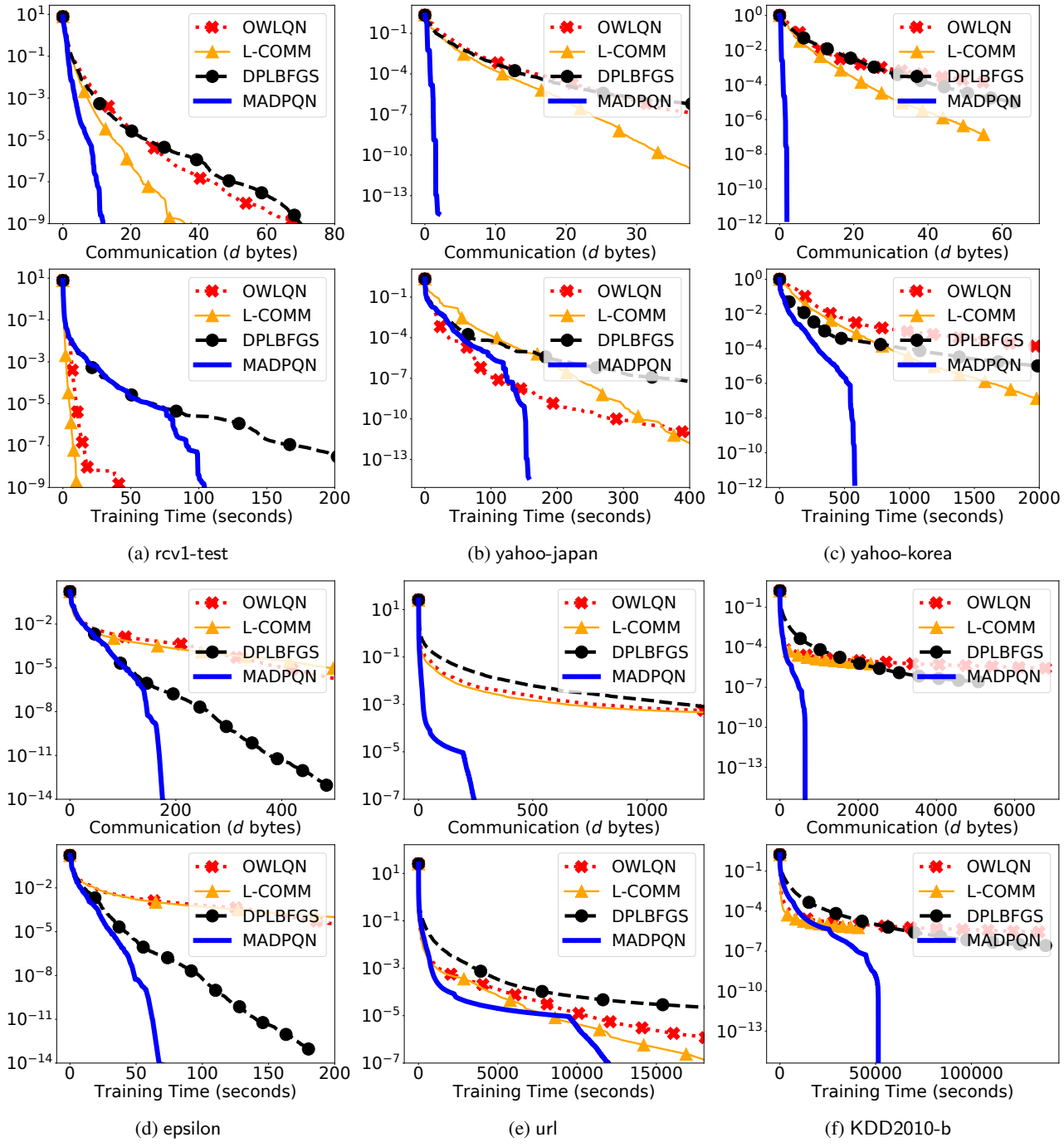
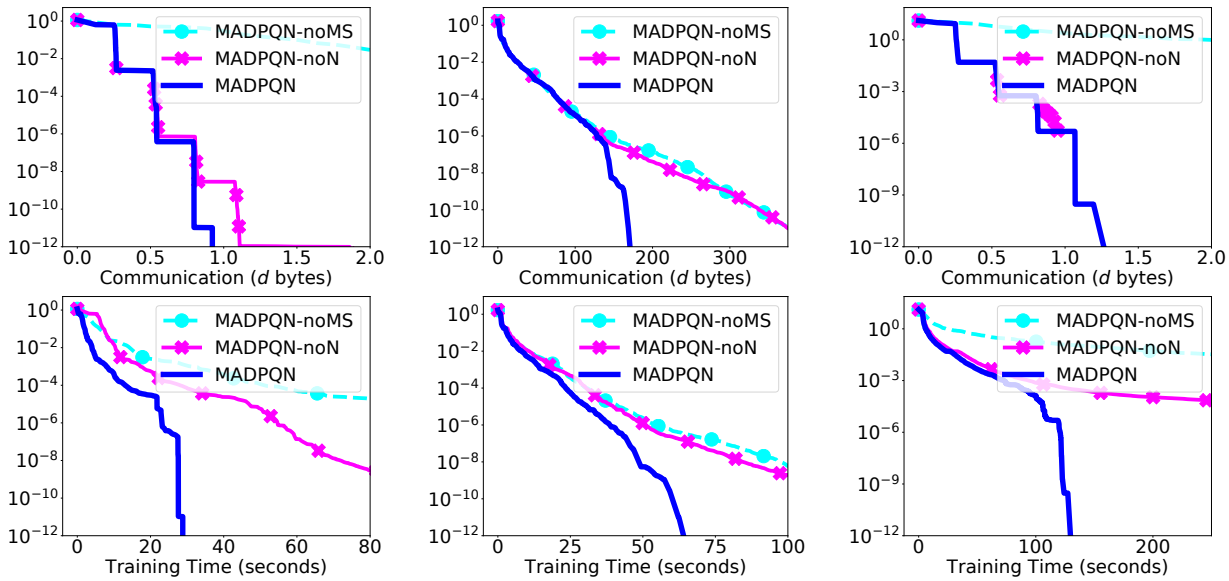


Figure 5. More comparisons of methods for (20). We present the number of bytes communicated (upper) and the training time (lower) vs. the relative difference to the optimal value.

optimization process. We present the results in Figure 6. For news20 and webspam, as their final solutions $\boldsymbol{w}^*$ are extremely sparse (see Table 1), meaning that $|\mathcal{M}^*|$ is small, our progressive shrinking strategy becomes very useful. Hence, compared with MADPQN-noMS, we observe significant reductions of the communication cost in MADPQN and MADPQN-noN. On the other hand, the final solution $\boldsymbol{w}^*$ of epsilon is relatively dense and the original problem dimension is also low. In such a case, we do not see much difference between MADPQN-noMS and MADPQN-noN because $|\mathcal{M}^*| \approx d$ and communication is likely not the bottleneck. However, after the smooth optimization stage is added in MADPQN, which



(a) news20 (b) epsilon (c) webspam

Figure 6. Effectiveness of progressive manifold selection and smooth optimization. We present the number of bytes communicated (upper) and the training time (lower) versus the difference to the optimal value. MADPQN-noMS indicates our method without the manifold selection and the semismooth Newton parts, and MADPQN-noN indicates our method without the semismooth Newton part.

Table 4. The minimum and maximum numbers of nonzero coordinates in $\boldsymbol{w}$ across machines after MADPQN identifies the correct manifold. For example, for news20 without random shuffling the coordinates, one machine contains no nonzero coordinate while another contains 392.

Data set	without shuffling		with shuffling	
	min	max	min	max
news20	0	392	9	25
webspam	0	172	9	36
yahoo-japan	0	2,114	78	117
yahoo-korea	0	8,549	277	336
rcv1-test	105	205	136	177
url	413	7,226	754	858
KDD2010-b	37,273	264,462	62,296	63,119
epsilon	33	52	30	53
avazu-site	325	407	338	410

starts to take Newton steps after identifying the correct manifold, the convergence is significantly improved in the later stage of optimization, showing the necessity of both stages.

B.4. Random Shuffling on the Coordinates

In Section 2.2, we mentioned that a partition $\mathcal{J}_1, \ldots, \mathcal{J}_K$ of $\{1, \ldots, d\}$ is used to indicate the coordinates handled by each machine, and for (20) and (21), this partition actually aligns directly with the basis of the manifolds. We notice that during the training, the numbers of nonzero coordinates could greatly vary on different machines, leading to an unbalanced workload that causes high synchronization costs between machines when random shuffling is not introduced (see Table 4 for the sizes of smallest and largest nonzero sets on (20)). To alleviate this issue, we apply coordinate shuffling described in Section A.3 to MADPQN (and also all other solvers as well for a fair comparison, although it is not quite useful for other approaches). To see its effectiveness, we present in Table 4 the sizes of nonzero sets after conducting a random shuffling of

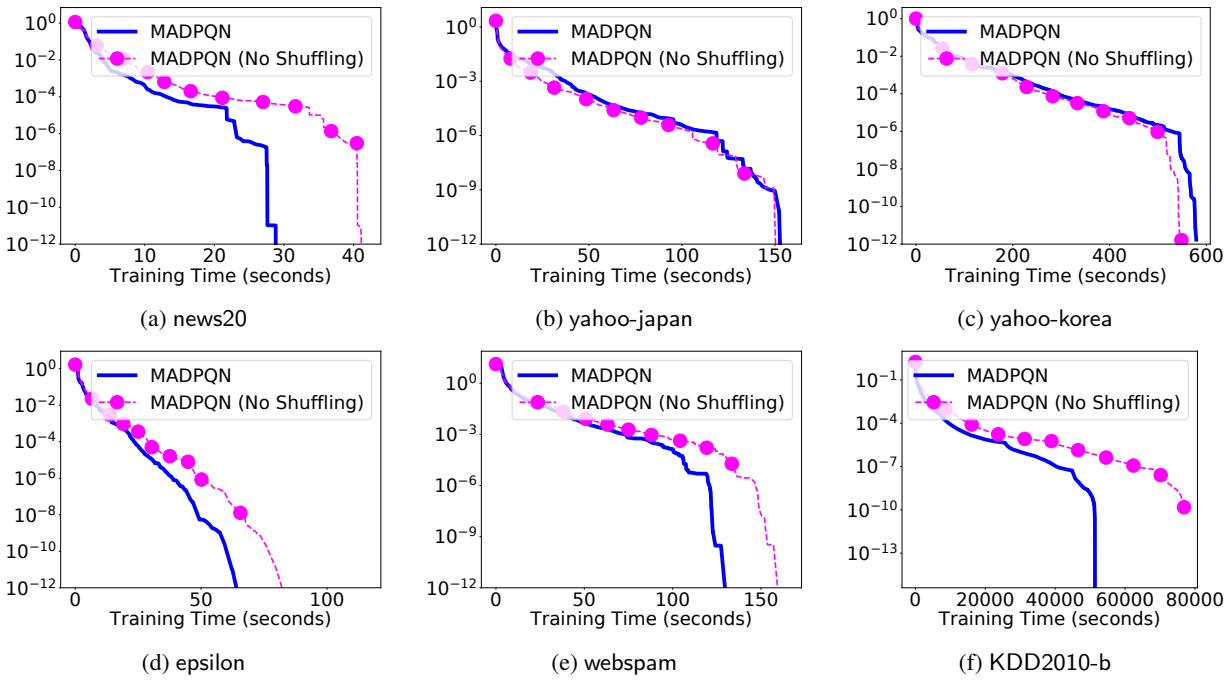


Figure 7. Comparisons of MADPQN with and without random shuffling the coordinates on (20). We present the training time versus the relative difference to the optimal value.

the coordinates for (20). Results show that the sizes become more balanced, making the workload distributed more evenly. We further conduct a running time comparison on (20) in Figure 7. We observe a clear gain of random shuffling in four of the six datasets presented since their synchronization costs are reduced, and for the remaining two datasets, the running time difference between with and without random shuffling is quite minor.

B.5. Comparison With the Feature-wise Scheme

We proceed on to compare with methods using feature-wise storage for linear models we reviewed in Section 4. We compare MADPQN and ADN (Dünner et al., 2018) as it is considered the state of the art for this type of methods. As the authors did not give details about the local sub-problem solver, we consider an efficient Newton-type random-permutation cyclic coordinate descent-based method (Yuan et al., 2012) implemented in the package LIBLINEAR (Fan et al., 2008) to solve the local problem and implement ADN in C/C++. Cyclic coordinate descent is known to converge Q-linearly so the requirement in (Dünner et al., 2018) is satisfied. We aggregate the local solutions after one epoch of cyclic coordinate descent to make the computation/communication ratio closer to our method for a fair comparison.

The comparison is shown in Figure 8. We notice that ADN has a communication cost close to MADPQN on datasets with $n \ll d$, including news20 and webspam, but performs poorly on other datasets with a larger n. Note that ADN is consistently slow in the running time for its heavier computation per iteration as it always processes all coordinates, so even on datasets that it has a low communication cost, the running time is not competitive. We also notice on all datasets that the convergence of ADN slows down significantly after reaching a certain medium precision.

C. More Analysis

We discuss more theoretical properties of our algorithm in this appendix. In particular, we discuss in detail the inner loop stopping condition and the superlinear convergence of the truncated semismooth Newton steps.

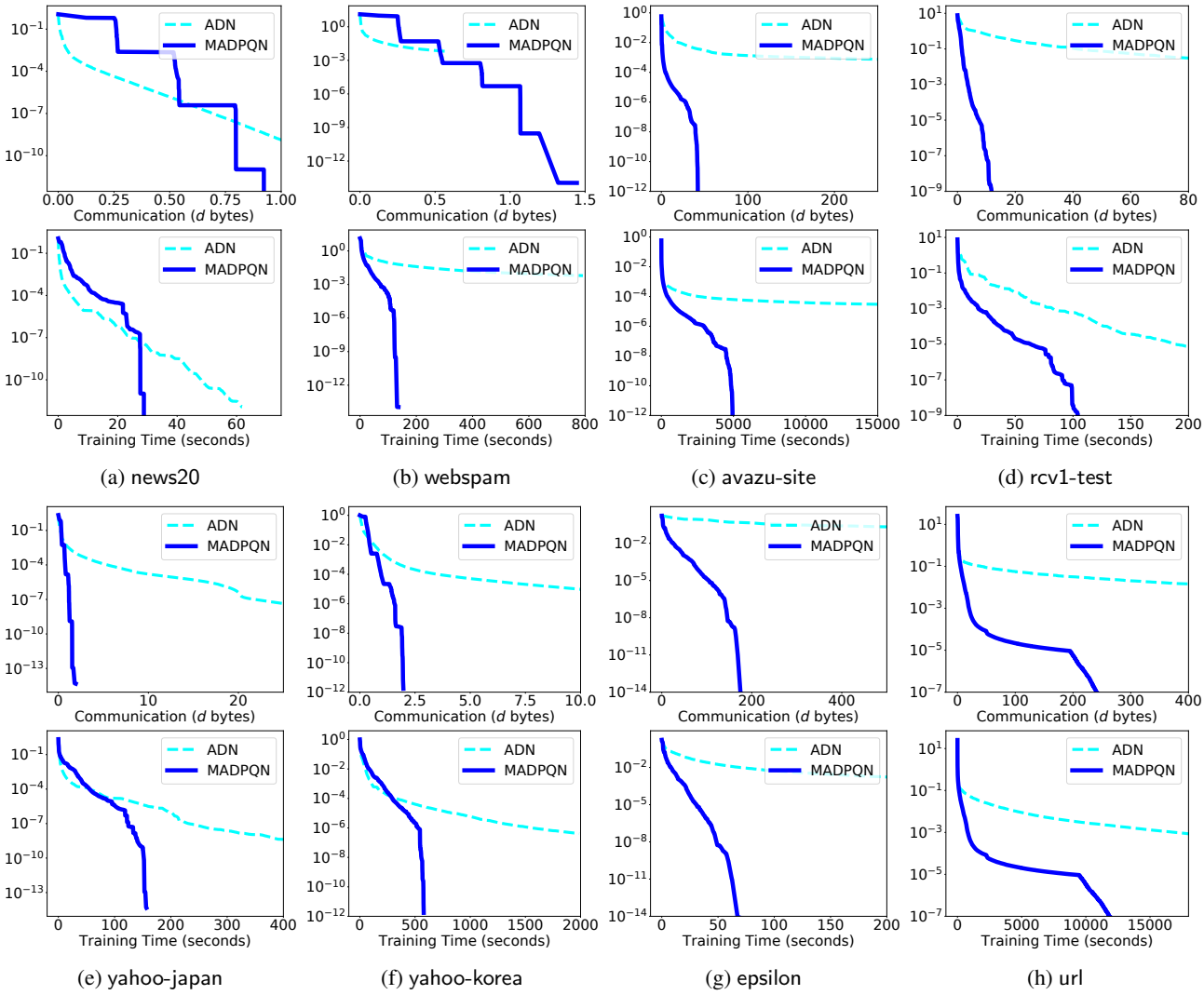


Figure 8. More comparisons of methods for (20). We present the number of bytes communicated (upper) and the training time (lower) vs. the relative difference to the optimal value.

C.1. Inner Loop Stopping Condition

In our implementation, we used the objective of the subproblem (12) as our stopping condition. The reason is that it provides an estimate of how far we are from the best solution we can get within the current manifold.

Lemma C.1. *If $\boldsymbol{p}^{j,t}$ is obtained through solving (12) approximately and (16) is satisfied, the first-order optimality condition restricted to $\mathcal{M}_{j,t}$, $\mathbf{0} \in \partial F|_{\mathcal{M}_{j,t}}(\boldsymbol{w}^{j,t})$, holds if and only if $|Q_{\hat{H}_{j,t}}(\boldsymbol{p}^{j,t}; \boldsymbol{w}^{j,t})| = 0$.*

When f is convex, (16) is always satisfied, and (18) is bounded, there are constants $R_1, R_2 > 0$ such that

$$\frac{\tilde{\Delta}_{j,t}}{\sigma_1} \geq |Q_{\hat{H}_{j,t}}(\boldsymbol{p}^{j,t}; \boldsymbol{w}^{j,t})| \geq \min\left\{R_1 \tilde{\Delta}_{j,t}^2, R_2 \tilde{\Delta}_{j,t}\right\}, \tag{28}$$

where

$$\tilde{\Delta}_{j,t} \coloneqq F(\boldsymbol{w}^{j,t}) - F_{j,t}^*, \quad F_{j,t}^* \coloneqq \min_{\boldsymbol{w} \in \mathcal{M}_{j,t}} F(\boldsymbol{w}).$$

When $F|_{\mathcal{M}_{j,t}}$ further satisfies (17) for some fixed $\mu > 0$ for all $\boldsymbol{w}$ within the level set $\{\boldsymbol{w} \mid F|_{\mathcal{M}_{j,t}}(\boldsymbol{w}) \leq F|_{\mathcal{M}_{j,t}}(\boldsymbol{w}^{j,t})\}$, the lower bound of (28) can be tighten to

$$|Q_{\hat{H}_{j,t}}(\boldsymbol{p}^{j,t}; \boldsymbol{w}^{j,t})| \geq R_3 \tilde{\Delta}_{j,t}$$

for some $R_3 > 0$.

C.2. Two-step Superlinear Convergence of the Newton Steps

We show that if ∇f is semismooth, we have already found the correct manifold $\mathcal{M}^*$, and when we are close enough to a local solution, we can get asymptotic superlinear convergence using the truncated semismooth Newton steps, even if between two Newton steps there is always one proximal gradient step.

Before starting our analysis, we give the definition of semismoothness (Mifflin, 1977) that is needed later.

Definition C.2 (Semismoothness)**.** *For a Lipschitz continuous function $G : \Re^m \to \Re^n$, for some integers $m, n \geq 0$, we say that G is semismooth at a point x if G is directionally differentiable at x and for any $V \in \partial G(x + \Delta x)$ with $\Delta x \to 0$,*

$$G(x + \Delta x) - G(x) - V\Delta x = o\left(\|\Delta x\|\right). \tag{29}$$

We notice that the class of semismooth functions is much broader than the class of differentiable functions, as the latter implies the former but not vice versa. We will assume that ∇f is semismooth to get superlinear convergence in our analysis.

We will assume that around a critical point $\boldsymbol{w}^*$ of the iterates, all the generalized Hessians H restricted to the manifold $\mathcal{M}^*$ have eigenvalues lower-bounded by a constant $r > 0$ for $\tilde{\mathcal{M}}^*$, in the sense

$$\boldsymbol{v}^\top H \boldsymbol{v} \geq r\|\boldsymbol{v}\|^2, \forall \boldsymbol{v} \in \tilde{\mathcal{M}}^* := \mathcal{M}^* - \boldsymbol{w}^*.$$

This combined with the nondegenerate condition $-\nabla f(\boldsymbol{w}^*) \in \operatorname{relint} \partial\Psi(\boldsymbol{w}^*)$ can be shown to be equivalent to the quadratic growth condition locally within the manifold (Wright, 2012). But we do not go along that direction further as it is not the focus of our analysis.

We now show that when the gradient is semismooth (as Ψ is $\mathcal{C}^2$ when restricted to the manifold, we just need ∇f to be semismooth) and when the generalized Hessians around $\boldsymbol{w}^*$ are all positive definite enough, the iterates of alternating between truncated semismooth Newton and proximal gradient steps converge to to $\boldsymbol{w}^*$ superlinearly in a two-step manner. We will assume that unit step size is already accepted. Under the positive definiteness assumption of the generalized Hessians around $\boldsymbol{w}^*$, standard argument for truncated Newton method indicates that unit step size is eventually accepted, and thus we omit discussion for this part. Parts of our proof are motivated by the proof of Wright (2012, Theorem 3.7).

Theorem C.3. *Assume that $\boldsymbol{w}^*$ is a cluster point of the iterates of our algorithm that is critical, Ψ is partly smooth around $\boldsymbol{w}^*$ with respect to the $\mathcal{C}^2$ manifold $\mathcal{M}^*$, $\nabla F|_{\mathcal{M}^*}$ is semismooth around $\boldsymbol{w}^*$, and there is $\delta_1 > 0$ and $r > 0$ such that for all $\boldsymbol{w} \in \mathcal{M}^*$ with $\|\boldsymbol{w} - \boldsymbol{w}^*\|_2 \leq \delta_1$, all the generalized Hessian*

$$\nabla^2 F|_{\mathcal{M}^*}(\boldsymbol{w}) \in \partial \nabla F|_{\mathcal{M}^*}(\boldsymbol{w})$$

satisfy

$$\boldsymbol{v}^\top \nabla^2 F|_{\mathcal{M}^*}(\boldsymbol{w})\boldsymbol{v} \geq r\|\boldsymbol{v}\|^2, \forall \boldsymbol{v} \in \tilde{\mathcal{M}}^*. \tag{30}$$

Then there exists another constant $\delta_2 > 0$ such that if:

- *The iterate $\boldsymbol{w}^{j,t}$ is already in $\mathcal{M}^*$ with $\|\boldsymbol{w}^{j,t} - \boldsymbol{w}^*\| \leq \delta_2$,*
- *We conduct one proximal gradient step (27) at the (j,t)th iteration of our algorithm, and*
- *The next update step at the $(j, t+1)$th iteration is a truncated semismooth Newton step satisfying (26) and unit step size satisfies (24) so that it is accepted,*

then we have that

$$\|\boldsymbol{w}^{j,t+2} - \boldsymbol{w}^*\| = o\left(\|\boldsymbol{w}^{j,t} - \boldsymbol{w}^*\|\right).$$

Namely, $\{\boldsymbol{w}^{j,2i}\}_i$ converges to $\boldsymbol{w}^*$ superlinearly.

D. Proofs

D.1. Proof of Lemma 3.1

Proof. We assume without loss of generality that (14) is always satisfied, as otherwise it just involves some index changes. We prove the first claim by induction. As no outer iteration is involved in this lemma, we omit the counter j for outer iterations.

The detailed update rule for $\tilde{Z}_t$ is:

$$\tilde{Z}_t := \begin{bmatrix} \gamma_t \tilde{Z}_t^{1,1} & \tilde{Z}_t^{2,1} \\ \left(\tilde{Z}_t^{2,1}\right)^\top & \tilde{Z}_t^{2,2} \end{bmatrix}, \tag{31}$$

with each block being of size $m(t) \times m(t)$ and defined as (note that $\tilde{Z}_t^{2,2}$ is a diagonal matrix)

$$\left(\tilde{Z}_t^{1,1}\right)_{i,j} = \begin{cases} \left(\tilde{Z}_{t-1}^{1,1}\right)_{i,j} & \text{if } m(t-1) < m \text{ and } i,j < m(t), \\ \left(\tilde{Z}_{t-1}^{1,1}\right)_{i+1,j+1} & \text{if } m(t-1) = m \text{ and } i,j < m, \\ P_{\tilde{\mathcal{M}}_{t-1}}\left(\boldsymbol{s}_{t-m(t)+i-1}\right)^\top P_{\tilde{\mathcal{M}}_{t-1}}\left(\boldsymbol{s}_{t-m(t)+j-1}\right) & \text{otherwise.} \end{cases} \tag{32a}$$

$$\left(\tilde{Z}_t^{2,2}\right)_{i,i} = \begin{cases} \left(\tilde{Z}_{t-1}^{2,2}\right)_{i,i} & \text{if } m(t-1) < m \text{ and } i < m(t), \\ \left(\tilde{Z}_{t-1}^{2,2}\right)_{i+1,i+1} & \text{if } m(t-1) = m \text{ and } i < m, \\ P_{\tilde{\mathcal{M}}_{t-1}}\left(\boldsymbol{s}_{t-1}\right)^\top P_{\tilde{\mathcal{M}}_{t-1}}\left(\boldsymbol{y}_{t-1}\right) & \text{otherwise.} \end{cases} \tag{32b}$$

$$\left(\tilde{Z}_t^{2,1}\right)_{i,j} = \begin{cases} \left(\tilde{Z}_{t-1}^{2,1}\right)_{i,j} & \text{if } m(t-1) < m \text{ and } i,j < m(t), \\ \left(\tilde{Z}_{t-1}^{2,1}\right)_{i+1,j+1} & \text{if } m(t-1) = m \text{ and } i,j < m, \\ P_{\tilde{\mathcal{M}}_{t-1}}\left(\boldsymbol{s}_{t-1}\right)^\top P_{\tilde{\mathcal{M}}_{t-1}}\left(\boldsymbol{y}_{t-m(t)+j-1}\right) & \text{if } m(t) = i > j, \\ 0 & \text{otherwise.} \end{cases} \tag{32c}$$

First, for $t = 1$ the two are the same because $\tilde{\mathcal{M}}_0 = \Re^d$, which implies $\boldsymbol{s}_0 = P_{\tilde{\mathcal{M}}_0}(\boldsymbol{s}_0)$ and $\boldsymbol{y}_0 = P_{\tilde{\mathcal{M}}_0}(\boldsymbol{y}_0)$, and there is no entry obtained from the previous iteration. Next, suppose that $\tilde{Z}_t = Z_t$ for $t = 1, 2, \ldots, T-1$. It is trivial to see that the update rules (32) without the projection part updates Z_t correctly by comparing (32) and (8). Therefore, we just need to check the correctness after the projection kicks in. For $t = T$, we see that because the previous update direction is obtained through solving (12) that confines the solution to $\tilde{\mathcal{M}}_{t-1}$, we have $\boldsymbol{s}_{T-1} = \boldsymbol{p}^{T-1} \in \tilde{\mathcal{M}}_{T-1}$. Therefore,

$$P_{\tilde{\mathcal{M}}_{T-1}^C}(\boldsymbol{s}_{T-1}) = \mathbf{0}.$$

As $\tilde{\mathcal{M}}_{T-1}$ is a subspace, for any vector $\boldsymbol{v}$, we always have

$$\boldsymbol{v} = P_{\tilde{\mathcal{M}}_{T-1}^C}(\boldsymbol{v}) + P_{\tilde{\mathcal{M}}_{T-1}}(\boldsymbol{v}),$$

so

$$\boldsymbol{s}_{T-1}^\top \boldsymbol{v} = P_{\tilde{\mathcal{M}}_{T-1}}\left(\boldsymbol{s}_{T-1}\right)^\top \boldsymbol{v} = P_{\tilde{\mathcal{M}}_{T-1}}\left(\boldsymbol{s}_{T-1}\right)^\top P_{\tilde{\mathcal{M}}_{T-1}}\left(\boldsymbol{v}\right).$$

This shows that the new entries in $\tilde{Z}_T$ also accord with the entries of Z_T at the same locations, as all new entries computed in (32) are inner products between $\boldsymbol{s}_{T-1}$ and some other vectors. Thus, by induction, we have proven that (8) and (32) give the same result, so $\tilde{Z}_T = Z_T$ as desired.

For the second result, we know that $\boldsymbol{b}_{t+1,i} \in \tilde{\mathcal{M}}_{t+1} \subseteq \tilde{\mathcal{M}}_t$ for all i, and therefore for any vector $\boldsymbol{v}$, we have that

$$\boldsymbol{b}_{t+1,i}^\top \boldsymbol{v} = \boldsymbol{b}_{t+1,i}^\top P_{\tilde{\mathcal{M}}_t}(\boldsymbol{v}). \tag{33}$$

Since B_t is an orthonormal basis of $\tilde{\mathcal{M}}_t$, we know that $B_t B_t^\top$ is the orthogonal projection operator onto $\tilde{\mathcal{M}}_t$, which completes our proof using (33). □

D.2. Proof of Lemma 3.2

Proof. In this proof, we denote the Lipschitz constant for ∇f by L. We first show that for the initial $\tilde{H}_{j,t}$, there are $\tilde{C}_1 \geq \tilde{C}_2 > 0$ such that

$$\tilde{C}_1 I \succeq \tilde{H}_{j,t} \succeq \tilde{C}_2 I, \quad \forall j, t \geq 0. \tag{34}$$

Consider the case $t = 0$. The lower bound $\gamma_{j,0} \geq \tilde{\epsilon}$ is trivial from (15). For the upper bound, the Lipschitz continuity of the gradient suggests that $\|\nabla^2 f(\boldsymbol{w})\| \leq L$ for any $\boldsymbol{w}$ at where f is twice-differentiable. Therefore, we have that $\gamma_{j,0} \leq L$ as the generalized Hessian $\nabla^2 f(\boldsymbol{w}^{j,0})$ is the limit of a sequence of real Hessians. Thus, (15) gives $\gamma_{j,0} \leq \max\{L, \tilde{\epsilon}\}$. The two results above combined then proves that

$$\tilde{\epsilon} \leq \gamma_{j,0} \leq \max\{L, \tilde{\epsilon}\}. \tag{35}$$

Now consider the case of $t > 0$. We assume without loss of generality that (14) is satisfied for all $t \geq 0$. Otherwise, we can just shift the indices in the proof since those pairs of $(\boldsymbol{s}_{j,t}, \boldsymbol{y}_{j,t})$ that failed (14) are discarded. We first show that $\gamma_{j,t}$ for $t > 0$ defined by (15) satisfies

$$\gamma_{j,t} \in \left[\delta, \frac{L^2}{\delta}\right], \quad \forall t > 0, \forall j. \tag{36}$$

For the upper bound, we have from (14) that

$$\left\|P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{y}_{j,t-1}\right)\right\| \geq \delta \left\|P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{s}_{j,t-1}\right)\right\|, \quad \forall t > 0, \forall j,$$

so

$$\gamma_{j,t} = \frac{\left\|P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{y}_{j,t-1}\right)\right\|^2}{P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{y}_{j,t-1}\right)^\top P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{s}_{j,t-1}\right)} \geq \frac{\left\|P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{y}_{j,t-1}\right)\right\|^2}{\left\|P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{y}_{j,t-1}\right)\right\| \left\|P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{s}_{j,t-1}\right)\right\|} = \frac{\left\|P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{y}_{j,t-1}\right)\right\|}{\left\|P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{s}_{j,t-1}\right)\right\|} \geq \delta.$$

For the lower bound, we first note from the Lipschitz continuity of ∇f that

$$\left\|P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{y}_{j,t-1}\right)\right\| \leq \|\boldsymbol{y}_{j,t-1}\| \leq L\|\boldsymbol{s}_{j,t-1}\| = L\left\|P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{s}_{j,t-1}\right)\right\|. \tag{37}$$

We thus get that

$$\gamma_{j,t} = \frac{\left\|P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{y}_{j,t-1}\right)\right\|^2}{P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{s}_{j,t-1}\right)^\top P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{s}_{j,t-1}\right)} \frac{P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{s}_{j,t-1}\right)^\top P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{s}_{j,t-1}\right)}{P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{y}_{j,t-1}\right)^\top P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{s}_{j,t-1}\right)} \overset{(14)}{\leq} \left(\frac{\left\|P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{y}_{j,t-1}\right)\right\|}{\left\|P_{\tilde{\mathcal{M}}_{j,t-1}}\left(\boldsymbol{s}_{j,t-1}\right)\right\|}\right)^2 \frac{1}{\delta} \overset{(37)}{\leq} \frac{L^2}{\delta}.$$

Thus, we have proven (36).

Next, we show that $\tilde{H}_{j,t}$ has bounded eigenvalues. Following Liu & Nocedal (1989), the construction of $\tilde{H}_{j,t}$ is equivalent to

$$\begin{aligned}
H_{j,t}^{(0)} &\coloneqq \gamma_{j,t} I, \\
H_{j,t}^{(i+1)} &\coloneqq H_{j,t}^{(i)} - \frac{H_{j,t}^{(i)} \boldsymbol{s}_{j,t-m(t)+i} \boldsymbol{s}_{j,t-m(t)+i}^\top H_{j,t}^{(i)}}{\boldsymbol{s}_{j,t-m(t)+i}^\top H_{j,t}^{(i)} \boldsymbol{s}_{t-m(t)+i}} + \frac{\boldsymbol{y}_{t-m(t)+i} \boldsymbol{y}_{t-m(t)+i}^\top}{\boldsymbol{y}_{t-m(t)+i}^\top \boldsymbol{s}_{t-m(t)+i}}, \quad i = 0, \ldots, m(t) - 1, \\
\tilde{H}_{j,t} &= H_{j,t}^{(m(t))}.
\end{aligned} \tag{38}$$

We start from another sequence of matrices $\{B_{j,t}^{(i)}\}_{i=0}^{m(t)}$ instead:

$$\begin{cases}
B_{j,t}^{(0)} & \coloneqq \frac{1}{\gamma_{j,t}} I, \\
B_{j,t}^{(i+1)} & \coloneqq V_{j,t-m(t)+i} B_{j,t}^{(i)} V_{j,t-m(t)+i}^\top + \rho_{j,t-m(t)+i} \boldsymbol{s}_{j,t-m(t)+i} \boldsymbol{s}_{j,t-m(t)+i}^\top,
\end{cases} \tag{39}$$

with

$$V_{j,i} \coloneqq \left(I - \rho_{j,i} \boldsymbol{s}_{j,i} \boldsymbol{y}_{j,i}^\top\right), \quad \rho_{j,i} \coloneqq \frac{1}{\boldsymbol{s}_{j,i}^\top \boldsymbol{y}_{j,i}}, i = 0, 1, \ldots. \tag{40}$$

It is obvious that $B_t^{(i)}$ are all symmetric. By an easy induction, we see that they are also positive semidefinite, as (14) ensures that $\rho_{j,i} > 0$ for all i. Now we show that $B_{j,t}^{(i)}$ are actually positive definite for all i through induction. Clearly,

since $\gamma_{j,t} > 0$, $B_{j,t}^{(0)} \succ 0$. Assume that $B_{j,t}^{(i)} \succ 0$ for $i = 0, 1, \ldots, i_0$, then for $i = i_0 + 1$, if there is a vector $\boldsymbol{v}$ such that $\boldsymbol{v}^\top B_{j,t}^{(i_0+1)} \boldsymbol{v} = 0$, we must have that

$$\begin{cases} \boldsymbol{v}^\top \boldsymbol{s}_{j,t-m(t)+i_0} & = 0, \\ \boldsymbol{v}^\top V_{j,t-m(t)+i_0} B_{j,t}^{(i_0)} V_{j,t-m(t)+i_0}^\top \boldsymbol{v} & = 0. \end{cases} \quad \Rightarrow \quad (\boldsymbol{v} - \boldsymbol{0})^\top B_{j,t}^{(i_0)} (\boldsymbol{v} - \boldsymbol{0}) = 0,$$

which implies that $\boldsymbol{v} = \boldsymbol{0}$. Therefore, $B_{j,t}^{(i_0+1)} \succ 0$.

Since all $B_{j,t}^{(i)}$ are positive definite, they are invertible. By the matrix inversion lemma, we can see that the inverse of $B_{j,t}^{(i)}$ is exactly $H_{j,t}^{(i)}$, so $H_{j,t}^{(i)}$ are all invertible and positive definite. We can therefore bound the trace of $H_{j,t}^{(i)}$ by

$$\text{trace}\left(H_{j,t}^{(i)}\right) \leq \text{trace}\left(H_{j,t}^{(0)}\right) + \sum_{k=t-m(t)}^{t-m(t)+i-1} \frac{\boldsymbol{y}_{j,k}^\top \boldsymbol{y}_{j,k}}{\boldsymbol{y}_{j,k}^\top \boldsymbol{s}_{j,k}} \leq \gamma_{j,t} d + \frac{iL^2}{\delta} \leq \frac{(i+1)L^2}{\delta} \leq \frac{(m+1)L^2}{\delta}, \quad \forall i, j \geq 0, \forall t > 0. \tag{41}$$

Therefore, from the positive semidefiniteness of $H_{j,t}^{(i)}$, (41) implies the existence of $\tilde{C}_1 > 0$ such that

$$H_{j,t}^{(i)} \preceq \tilde{C}_1 I, \quad i = 0, \ldots, m(t), \quad \forall t > 0, \forall j \geq 0.$$

Next, for its lower bound, from the formulation for (38) in (Liu & Nocedal, 1989), and the upper bound $\|H_{j,t}^{(i)}\| \leq \tilde{C}_1$, we have

$$\det\left(\tilde{H}_{j,t}\right) = \det\left(H_{j,t}^{(0)}\right) \prod_{k=t-m(t)}^{t-1} \frac{\boldsymbol{y}_{j,k}^\top \boldsymbol{s}_{j,k}}{\boldsymbol{s}_{j,k}^\top \boldsymbol{s}_{j,k}} \frac{\boldsymbol{s}_{j,k}^\top \boldsymbol{s}_{j,k}}{\boldsymbol{s}_{j,k}^\top H_{j,t}^{(k-t+m(t))} \boldsymbol{s}_{j,k}} \geq \gamma_{j,t}^d \left(\frac{\delta}{\tilde{C}_1}\right)^{m(t)} > 0.$$

We then take

$$c_2 = \delta^d \min_{i=0,1,\ldots,m} \left(\frac{\delta}{\tilde{C}_1}\right)^i$$

and see immediately from (36) that

$$\det\left(\tilde{H}_{j,t}\right) \geq c_2, \quad \forall t > 0, \forall j \geq 0.$$

We therefore get that

$$\lambda_{\min}\left(\tilde{H}_{j,t}\right) \geq \frac{\det\left(\tilde{H}_{j,t}\right)}{\lambda_{\max}\left(\tilde{H}_{j,t}\right)^{d-1}} \geq \frac{c_2}{\tilde{C}_1^{d-1}} =: \tilde{C}_2 > 0,$$

where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the smallest and the largest eigenvalues, respectively. The part of finite modification then follows from Lemma 4 of (Lee & Wright, 2019), and the final result is clear from (6) and that $\tilde{H}_t$ is positive definite. □

D.3. Proof of Lemma 3.3

Proof. In this proof, without loss of generality, we skip all the semismooth Newton steps, as it just involves at most a shift of the iteration counters because we alternate between a Newton step and a proximal quasi-Newton or proximal gradient step. Consider the acceptance criterion (5) for the update step. From (16) and that $Q_H(\boldsymbol{0}; \boldsymbol{w}) = 0$ for any H and any $\boldsymbol{w}$, we know that $Q_{\hat{H}_{j,t}}\left(\boldsymbol{p}^{j,t}; \boldsymbol{w}^{j,t}\right) \leq 0$ for all t. Therefore, we can see that for any $T \geq 0$ and any j,

$$\begin{aligned} \sum_{t=0}^{T} \sigma_1 \left|Q_{\hat{H}_{j,t}}\left(\boldsymbol{p}^{j,t}; \boldsymbol{w}^{j,t}\right)\right| &= -\sum_{t=0}^{T} \sigma_1 Q_{\hat{H}_{j,t}}\left(\boldsymbol{p}^{j,t}; \boldsymbol{w}^{j,}t\right) \\ &\leq \sum_{t=0}^{T} \left(F(\boldsymbol{w}^{j,t}) - F(\boldsymbol{w}^{j,t+1}\right) \\ &\leq F(\boldsymbol{w}^{j,0}) - F(\boldsymbol{w}^{j,T+1}) \\ &\leq F(\boldsymbol{w}^{j,0}) - F^*. \end{aligned}$$

By dividing both sides by $\sigma_1(T+1)$, we get that

$$c_{j,T} \coloneqq \min_{0\le t\le T} \left| Q_{\hat{H}_{j,t}}\left(\boldsymbol{p}^{j,t}; \boldsymbol{w}^{j,t}\right)\right| \le \frac{F(\boldsymbol{w}^{j,0}) - F^*}{\sigma_1(T+1)}.$$

As for any fixed j, $\{c_{j,T}\}$ is decreasing, nonnegative, and summable with respect to T, Proposition 3.4 and Theorem 3.5 of (Shi et al., 2015) show that $\{c_{j,T}\}_{T=0,1,\dots}$ decreases at a rate of $o(1/T)$, showing that within $o(1/\epsilon_j)$ inner iterations, $c_{j,T}$ will be smaller than ϵ_j. Note that this part does not depend on the eigenvalues of $\hat{H}_{j,t}$, and thus it holds regardless whether a proximal quasi-Newton step or a proximal gradient step is taken after an SSN step.

When (17) holds, again from the acceptance criterion, and let F_t^* be defined as

$$F_{j,t}^* \coloneqq \min_{\boldsymbol{w}\in\mathcal{M}_{j,t}} F(\boldsymbol{w}), \tag{42}$$

clearly we have

$$F(\boldsymbol{w}^{j,t}) - F_{j,t}^* \ge F(\boldsymbol{w}^{j,t}) - F(\boldsymbol{w}^{j,t+1}) \ge \sigma_1 \left| Q_{\hat{H}_{j,t}}(\boldsymbol{p}^{j,t}; \boldsymbol{w}^{j,t})\right|. \tag{43}$$

From Corollary 2 of (Lee & Wright, 2019), $F(\boldsymbol{w}^{j,t}) - F_{j,t}^*$ converges to 0 at a Q-linear rate as $\hat{H}_{j,t}$ has bounded eigenvalues for all j and t (Lemma 3.2 and note that if we take (27), $\gamma_{j,t}$ are also bounded), so it takes at most $O(\log(1/\epsilon_j))$ inner iterations to make $|Q_{\hat{H}_{j,t}}(\boldsymbol{p}^{j,t}; \boldsymbol{w}^{j,t})|$ smaller than ϵ_j. Notice that $\mathcal{M}_{j,t}$ might be shrinking, but since $F_{j,t}^*$ does not decrease when $\mathcal{M}_{j,t}$ shrinks (as the constraint for the optimization problem becomes more restrictive), and the Lipschitz constant only increases when confined to a subspace, the Q-linear rate at each iteration can only become faster but not slower. Therefore, the Q-linear convergence rate indeed holds, even if the selected manifold keeps shrinking. □

D.4. Proof of Theorem 3.4

Proof. First consider the case that f is convex. We will denote the upper bound for (18) by R, and define

$$F^* \coloneqq \min_{\boldsymbol{w}} F(\boldsymbol{w}), \quad \Delta_{j,t} \coloneqq F(\boldsymbol{w}^{j,t}) - F^*. \tag{44}$$

Clearly $0 \le R < \infty$ and $\Delta_{j,t} \ge 0$. We notice from $Q_H(\mathbf{0}; \boldsymbol{w}) = 0$ for any H and $\boldsymbol{w}$ and (16) that

$$Q_{\hat{H}_{j,t}}(\boldsymbol{p}^{j,t}; \boldsymbol{w}^{j,t}) \le 0.$$

This together with (5) shows that the function value is always decreasing. Now we consider only $\boldsymbol{w}^{j,0}$ and $\boldsymbol{w}^{j,1}$ for all j. We know that in this case $\hat{H}_{j,0} = \gamma_{j,0} I$, and from Lemma 3.2 we know that $\hat{H}_{j,0} = \hat{\gamma}_{j,0} I$ for some $\hat{\gamma}_{j,0}$ that is bounded in a range $[C_2, C_1]$ for some $C_1 \ge C_2 > 0$, for all $j \ge 0$. As we solve (4) exactly for $t = 0$, we have from the convexity of f and therefore F that

$$\begin{aligned}
& Q_{\hat{H}_{j,0}}(\boldsymbol{p}^{j,0}; \boldsymbol{w}^{j,0}) \\
=& \min_{\boldsymbol{p}} \nabla f\left(\boldsymbol{w}^{j,0}\right)^\top \boldsymbol{p} + \frac{\hat{\gamma}_{j,0}}{2}\boldsymbol{p}^\top\boldsymbol{p} + \Psi\left(\boldsymbol{w}^{j,0} + \boldsymbol{p}\right) - \Psi\left(\boldsymbol{w}^{j,0}\right) \\
\le& \min_{\boldsymbol{p}} f\left(\boldsymbol{w}^{j,0} + \boldsymbol{p}\right) + \Psi\left(\boldsymbol{w}^{j,0} + \boldsymbol{p}\right) + \frac{\hat{\gamma}_{j,0}}{2}\boldsymbol{p}^\top\boldsymbol{p} - F\left(\boldsymbol{w}^{j,0}\right) \\
\le& \min_{\lambda\in[0,1]} F\left(\boldsymbol{w}^{j,0} + \lambda\left(P_\Omega\left(\boldsymbol{w}^{j,0}\right) - \boldsymbol{w}^{j,0}\right)\right) + \frac{\hat{\gamma}_{j,0}\lambda^2}{2}\left(P_\Omega\left(\boldsymbol{w}^{j,0}\right) - \boldsymbol{w}^{j,0}\right)^\top\left(P_\Omega\left(\boldsymbol{w}^{j,0}\right) - \boldsymbol{w}^{j,0}\right) - F\left(\boldsymbol{w}^{j,0}\right) \\
\le& \min_{\lambda\in[0,1]} (1-\lambda) F\left(\boldsymbol{w}^{j,0}\right) + \lambda F^* + \frac{\hat{\gamma}_{j,0}\lambda^2}{2}\left(\boldsymbol{w}^{j,0} - P_\Omega\left(\boldsymbol{w}^{j,0}\right)\right)^\top\left(\boldsymbol{w}^{j,0} - P_\Omega\left(\boldsymbol{w}^{j,0}\right)\right) - F\left(\boldsymbol{w}^{j,0}\right) \\
\le& \min_{\lambda\in[0,1]} -\lambda\Delta_{j,0} + \frac{\hat{\gamma}_{j,0}\lambda^2}{2}\|\boldsymbol{w}^{j,0} - P_\Omega\left(\boldsymbol{w}^{j,0}\right)\|^2 \qquad (45) \\
\le& \min_{\lambda\in[0,1]} -\lambda\Delta_{j,0} + \frac{R^2\hat{\gamma}_{j,0}\lambda^2}{2} \\
\le& \min_{\lambda\in[0,1]} -\lambda\Delta_{j,0} + \frac{C_1 R^2\lambda^2}{2}. \qquad (46)
\end{aligned}$$

The minimizer of (46) is

$$\lambda = \min\left\{1, \frac{\Delta_{j,0}}{C_1 R^2}\right\},$$

and by using which in (46), we get

$$Q_{\hat{H}_{j,0}}(\boldsymbol{p}^{j,0}; \boldsymbol{w}^{j,0}) \leq -\frac{1}{2}\Delta_{j,0}$$

in the former case and

$$Q_{\hat{H}_{j,0}}(\boldsymbol{p}^{j,0}; \boldsymbol{w}^{j,0}) \leq -\frac{\Delta_{j,0}^2}{2C_1 R^2}$$

in the latter. We then get that

$$\Delta_{j+1,0} \leq \Delta_{j,1} \leq \Delta_{j,0} - \frac{\sigma_1}{2}\min\left\{\Delta_{j,0}, \frac{\Delta_{j,0}^2}{C_1 R^2}\right\} \leq \Delta_{j,0} - \frac{\sigma_1}{2}\min\left\{\Delta_{j,0}, \frac{\Delta_{j,0}^2}{C_1 R^2 + \Delta_{0,0}}\right\} \leq \Delta_{j,0} - \frac{\sigma_1}{2}\frac{\Delta_{j,0}^2}{C_1 R^2 + \Delta_{0,0}},$$

where in the last inequality we used the fact that $\Delta_{j,0}$ is decreasing with respect to j. Therefore, after dividing both sides by $\Delta_{j,0}\Delta_{j+1,0}$, we obtain

$$\Delta_{j,0}^{-1} \leq \Delta_{j+1,0}^{-1} - \frac{1}{C_1 R^2 + \Delta_{0,0}}.$$

By summing the inequality above from $j = 0$ to $j = T - 1$ and telescoping, we get

$$\Delta_{T,0}^{-1} \geq \Delta_{0,0}^{-1} + \frac{T}{C_1 R^2 + \Delta_{0,0}} \geq \frac{T}{C_1 R^2 + \Delta_{0,0}}.$$

This shows that $\Delta_{T,0}$ decreases at a rate of $O(1/T)$, and therefore it takes at most $O(1/\epsilon)$ such steps to reach an ϵ-accurate solution.

When F satisfies (17), we apply it to (45) and get

$$Q_{\hat{H}_{j,0}}(\boldsymbol{p}^{j,0}; \boldsymbol{w}^{j,0}) \leq \min_{\lambda\in[0,1]} -\lambda\Delta_{j,0} + \frac{\hat{\gamma}_{j,0}\lambda^2}{2\mu}\Delta_{j,0} \leq \min_{\lambda\in[0,1]} -\lambda\Delta_{j,0} + \frac{C_1\lambda^2}{2\mu}\Delta_{j,0} = \min_{\lambda\in[0,1]} \left(\frac{C_1\lambda^2}{2\mu} - \lambda\right)\Delta_{j,0}. \quad (47)$$

Let

$$C_3 \coloneqq \min_{\lambda\in[0,1]} \left(\frac{C_1\lambda^2}{2\mu} - \lambda\right),$$

then, with some simple calculation, clearly $C_3 \in [-1, 0)$. From (5), we see that

$$\Delta_{j+1,0} - \Delta_{j,0} \leq \Delta_{j,1} - \Delta_{j,0} \leq \sigma_1 Q_{\hat{H}_{j,0}}\left(\boldsymbol{p}^{j,0}; \boldsymbol{w}^{j,0}\right) \leq \sigma_1 C_3 \Delta_{j,0} \quad \Rightarrow \quad \Delta_{j+1,0} \leq (1 + \sigma_1 C_3)\,\Delta_{j,0},$$

showing that $\Delta_{j,0}$ decreases at a Q-linear rate (as $\sigma_1 \in (0, 1]$ and $C_3 \in [-1, 0)$), so it takes at most $O(\log(1/\epsilon))$ such proximal gradient steps to get an ϵ-accurate solution.

Finally, consider the case in which f is nonconvex. We use (5) to get

$$\sigma_1 \sum_{j=0}^{T} |Q_{\hat{H}_{j,0}}(\boldsymbol{p}^{j,0}; \boldsymbol{w}^{j,0})| \leq \sum_{j=0}^{T} (F(\boldsymbol{w}^{j,0}) - F(\boldsymbol{w}^{j,1})) \leq F(\boldsymbol{w}^{0,0}) - F^*. \quad (48)$$

Define

$$Q_k \coloneqq \min_{0\leq j\leq k} \left|Q_{\hat{H}_{j,0}}(\boldsymbol{p}^{j,0}; \boldsymbol{w}^{j,0})\right|,$$

and we get that Q_k is decreasing, nonnegative, and summable. Thus again by Proposition 3.4 and Theorem 3.5 of (Shi et al., 2015), Q_k decreases at a rate of $o(1/k)$, showing that it takes at most $o(1/\epsilon)$ outer iterations to decrease Q_k to below ϵ.

For the final claim, from Lemma 3.2 we know that $\hat{H}_{j,0} = \hat{\gamma}_{j,0} I$ for some $\hat{\gamma}_{j,0}$ that is bounded in a range $[C_2, C_1]$ for some $C_1 \geq C_2 > 0$. Because $Q_{\hat{\gamma}_{j,0}I}(\cdot\,; \boldsymbol{w}^{j,0})$ is strongly convex, $Q_{\hat{\gamma}_{j,0}I}(\boldsymbol{0}; \boldsymbol{w}^{j,0}) = 0$, and (16), we know that $Q_{\hat{\gamma}_{j,0}I}(\boldsymbol{p}^{j,0}; \boldsymbol{w}^{j,0}) = 0$ if and only if $\boldsymbol{p}^{j,0} = \boldsymbol{0}$. The optimality condition of the subproblem (4) is

$$\boldsymbol{0} \in \nabla f(\boldsymbol{w}^{j,0}) + \hat{\gamma}_{j,0}\boldsymbol{p}^{j,0} + \partial\Psi\left(\boldsymbol{w}^{j,0} + \boldsymbol{p}^{j,0}\right). \quad (49)$$

When $\boldsymbol{p}^{j,0} = \mathbf{0}$, (49) implies that $\mathbf{0} \in \partial F(\boldsymbol{w}^{j,0})$. On the other hand, we have that

$$\begin{aligned} Q_{\hat{\gamma}_{j,0} I}\left(\boldsymbol{p}^{j,0}; \boldsymbol{w}^{j,0}\right) &= \nabla f(\boldsymbol{w}^{j,0})^\top \boldsymbol{p}^{j,0} + \frac{\hat{\gamma}_{j,0}}{2}\left\|\boldsymbol{p}^{j,0}\right\|^2 + \Psi\left(\boldsymbol{w}^{j,0} + \boldsymbol{p}^{j,0}\right) - \Psi\left(\boldsymbol{w}^{j,0}\right) \\ &\geq \nabla f(\boldsymbol{w}^{j,0})^\top \boldsymbol{p}^{j,0} + \frac{\hat{\gamma}_{j,0}}{2}\left\|\boldsymbol{p}^{j,0}\right\|^2 + \boldsymbol{s}^\top \boldsymbol{p}^{j,0}, \quad \forall \boldsymbol{s} \in \partial\Psi\left(\boldsymbol{w}^{j,0}\right). \end{aligned} \tag{50}$$

When $\mathbf{0} \in \partial F(\boldsymbol{w}^{j,0})$ holds, we see that $-\nabla f(\boldsymbol{w}^{j,0}) \in \partial\Psi\left(\boldsymbol{w}^{j,0}\right)$, so (50) implies that

$$Q_{\hat{\gamma}_{j,0} I}\left(\boldsymbol{p}^{j,0}; \boldsymbol{w}^{j,0}\right) = \nabla f(\boldsymbol{w}^{j,0})^\top \boldsymbol{p}^{j,0} + \frac{\hat{\gamma}_{j,0}}{2}\left\|\boldsymbol{p}^{j,0}\right\|^2 + \Psi\left(\boldsymbol{w}^{j,0} + \boldsymbol{p}^{j,0}\right) - \Psi\left(\boldsymbol{w}^{j,0}\right) \geq \mathbf{0}^\top \boldsymbol{p}^{j,0} + \frac{\hat{\gamma}_{j,0}}{2}\left\|\boldsymbol{p}^{j,0}\right\|^2 \geq 0.$$

However, we have noticed above that $Q_{\hat{\gamma}_{j,0} I}\left(\boldsymbol{p}^{j,0}; \boldsymbol{w}^{j,0}\right) \leq 0$, so it must be 0, as desired.

□

D.5. Proof of Theorem 3.6

Proof. From (48), we see that $|Q_{\hat{H}_{j,0}}(\boldsymbol{p}^{j,0}; \boldsymbol{w}^{j,0})|$ converges to 0 as j approaches infinity. From Lemma 3.2 we know that $Q_{\hat{H}_{j,0}}(\cdot; \boldsymbol{w}^{j,0})$ is C_2-strongly convex for some $C_2 > 0$ for all j, so it satisfies (17) with $\mu = C_2/2$ and $\boldsymbol{p}^{j,0}$ is the only solution to (4) (we assumed that for $t = 0$ the solution is exact as it is just a simple proximal operation). These imply that $\boldsymbol{p}^{j,0}$ converges to $\mathbf{0}$ as well. On the other hand, from, for example, Proposition 8.7 of (Rockafellar & Wets, 2009), we know that the subdifferential $\partial\Psi$ is outer-semicontinuous, meaning that for any sequence of points $\hat{\boldsymbol{w}}^j$ converging to a point $\hat{\boldsymbol{w}}$, we have

$$\lim_{j\to\infty} \partial\Psi\left(\hat{\boldsymbol{w}}^j\right) \subseteq \partial\Psi(\hat{\boldsymbol{w}}). \tag{51}$$

Thus if there is a subsequence $\{\boldsymbol{w}^{j,0}\}_{j\in K}$ that converges to a cluster point $\boldsymbol{w}^*$, we have from the optimality condition of (4) that

$$-\left(\nabla f\left(\boldsymbol{w}^{j,0}\right) + \hat{\gamma}_{j,0}\boldsymbol{p}^{j,0}\right) \in \partial\Psi\left(\boldsymbol{w}^{j,0} + \boldsymbol{p}^{j,0}\right).$$

By taking limit on j over K on it, using (51), and that $\hat{\gamma}_{j,0} \in [C_2, C_1]$ for some $C_1 \geq C_2 > 0$ for all j from Lemma 3.2, we get that

$$-(\nabla f(\boldsymbol{w}^*) + \mathbf{0}) \in \partial\Psi(\boldsymbol{w}^* + \mathbf{0}) \quad \Rightarrow \quad \mathbf{0} \in \partial F(\boldsymbol{w}^*),$$

showing that $\boldsymbol{w}^*$ is a critical point and $\{\boldsymbol{w}^{j,1}\}_{j\in K}$ converges to $\boldsymbol{w}^*$ as well.

When f is lower-bounded and Ψ is coercive, the level sets of F are compact. Then from the Bolzano-Weierstrass Theorem, the sequence $\{\boldsymbol{w}^{j,0}\}$ must have a subsequence converging to a point $\boldsymbol{w}^*$, as the whole sequence lies in a compact set. □

D.6. Proof of Theorem 3.7

Proof. Let K denotes the set of the indices of the subsequence of $\{\boldsymbol{w}^{j,0}\}$ that converges to $\boldsymbol{w}^*$. From (49) and Lemma 3.2, we know that

$$\begin{aligned} \operatorname{dist}(-\nabla f(\boldsymbol{w}^{j,1}), \partial\Psi(\boldsymbol{w}^{j,1})) &\leq \| - \hat{\gamma}_{j,0}\boldsymbol{p}^{j,0} + (\nabla f(\boldsymbol{w}^{j,1}) - \nabla f(\boldsymbol{w}^{j,0}))\| \\ &= \| - \hat{\gamma}_{j,0}\boldsymbol{p}^{j,0} + (\nabla f(\boldsymbol{w}^{j,0} + \boldsymbol{p}^{j,0}) - \nabla f(\boldsymbol{w}^{j,0}))\| \\ &\leq C_1\|\boldsymbol{p}^{j,0}\| + \|\nabla f(\boldsymbol{w}^{j,0} + \boldsymbol{p}^{j,0}) - \nabla f(\boldsymbol{w}^{j,0})\| \\ &\leq C_1\|\boldsymbol{p}^{j,0}\| + L\|\boldsymbol{p}^{j,0}\| \\ &= (L + C_1)\|\boldsymbol{p}^{j,0}\| \to 0 \end{aligned}$$

for $j \in K$, where L is the Lipschitz constant of ∇f and the convergence is from the proof of Theorem 3.6. Therefore, all the conditions of Lewis & Zhang (2013, Theorem 4.10) are satisfied, showing that $\boldsymbol{w}^{j,1} = \boldsymbol{w}^{j,0} + \boldsymbol{p}^{j,0} \in \mathcal{M}^*$ holds for all $j \in K$ large enough, or equivalently when $\boldsymbol{w}^{j,0}$ is close enough to $\boldsymbol{w}^*$.

Next, we take into consideration the inner iterates. For the jth outer iteration, We denote by $T(j)$ the set for the indices of the inner iterations in which we did not use a truncated semismooth Newton step. We then construct another two sequences of vectors $\{\hat{\boldsymbol{w}}^{j,t}\}$ and $\{\hat{\boldsymbol{p}}^{j,t}\}$ for $j \in K$ and $t = 0, 1, \ldots,$ by $\hat{\boldsymbol{w}}^{j,0} \equiv \boldsymbol{w}^{j,0}$, $\hat{\boldsymbol{w}}^{j,1} \equiv \boldsymbol{w}^{j,1}$, $\hat{\boldsymbol{p}}^{j,0} \equiv \boldsymbol{p}^{j,0}$, and letting $\hat{\boldsymbol{w}}^{j,t+1}$ and

$\hat{\boldsymbol{p}}^{j,t+1}$ for $t+1 \in T(j)$ be the iterates and steps generated by solving (4) under $\Re^d$ instead of the selected manifold/subspace, using the previous $\hat{\boldsymbol{w}}^{j,t}$. For those $t+1 \notin T(j)$, we do the same truncated semismooth Newton step within the current smooth manifold in which $\hat{\boldsymbol{w}}^{j,t+1}$ lies. Then by using an argument similar to that in the proof of Lemma 3.3, we have

$$\sigma_1 \sum_{j\in K}\sum_{t\in T(j)} |Q_{\hat{H}_{j,t}}(\hat{\boldsymbol{p}}^{j,t};\hat{\boldsymbol{w}}^{j,t})| \le F(\boldsymbol{w}^{0,0}) - F^*,$$

showing that $\sum_{t\in T(j)} |Q_{\hat{H}_{j,t}}(\hat{\boldsymbol{p}}^{j,t};\hat{\boldsymbol{w}}^{j,t})|$ converges to 0 as well. This in turn implies that $\{\sum_{t\in T(j)} \|\hat{\boldsymbol{p}}^{j,t}\|^2\}_{j\in K}$ also converges to 0, as $Q_{\hat{H}_{j,t}}$ satisfies (17) with $\mu = C_2$ for some fixed $C_2 > 0$ for all j and t by Lemma 3.2, $Q_{\hat{H}_{j,t}}(\boldsymbol{0},\hat{\boldsymbol{w}}^{j,t};\hat{\boldsymbol{p}}^{j,t}) = 0$, and $\hat{\boldsymbol{p}}^{j,t}$ is the only exact solution (as assumed in the theorem statement).

For those $t \notin T(j)$, we know that there are only finite such t according to Lemma 3.3. We show that the corresponding update steps also converge to 0 as well, in the sense that $\{\sum_{t\notin T(j)} \|\hat{\boldsymbol{p}}^{j,t}\|^2\}_{j\in K}$ also approaches 0. We let the iterations $t \notin T(j)$ be denoted by $i_{j,k}, k = 0, 1, \ldots$, and denote the largest number of possible inner iterations by K_1, we then have that $k \le K_1$ and $i_{j,k} \le K_1$ for all j, k. For any given j and $k = 0$, we have $t \in T(j)$ for all $t < i_{j,0}$. As we know that $\hat{\boldsymbol{w}}^{j,0}$ converges to $\boldsymbol{w}^*$ and $\sum_{t<i_{j,0}} \|\hat{\boldsymbol{p}}^{j,t}\|^2$ converges to $\boldsymbol{0}$ as j approaches infinity, we see that

$$\hat{\boldsymbol{w}}^{j,i_{j,0}} = \hat{\boldsymbol{w}}^{j,0} + \sum_{t=0}^{i_{j,0}-1} \hat{\boldsymbol{p}}^{j,t} \to \boldsymbol{w}^* + \boldsymbol{0} = \boldsymbol{w}^*, \tag{52}$$

thus $\hat{\boldsymbol{w}}^{j,i_{j,0}}$ also converges to $\boldsymbol{w}^*$. Since these iterates all lie in a region close to $\boldsymbol{w}^*$, we can find a compact set $\mathcal{S}$ that encompasses them. As $\Psi|_{\mathcal{M}^*}$ is $\mathcal{C}^2$, its Hessian is actually bounded within $\mathcal{S}$. Moreover, the generalized Hessian of f has eigenvalues upper bounded by L, and thus the generalized Hessian of $F|_{\mathcal{M}^*}$ is upper-bounded within the region of interest. On the other hand, from the trust-region technique for the SSN steps in (25), the real matrix used for obtaining the SSN step is also lower-bounded. Because the gradient within the smooth manifold is continuous and the iterates are confined to a compact set, we know that the exact SSN step converges to $\boldsymbol{0}$ as well, and thus so does $\hat{\boldsymbol{p}}^{j,i_{j,0}}$ as implied by (26). Now assume $\hat{\boldsymbol{p}}^{j,i_{j,k}}$ converges to $\boldsymbol{0}$ for $k = 0, 1, \ldots, T$, then for $k = T+1$, if $i_{j,T+1}$ exists, we see that by reusing (52) with $i_{j,0}$ replaced by $i_{j,T}$, $\hat{\boldsymbol{w}}^{j,i_{j,T+1}}$ also converges to $\boldsymbol{w}^*$ (as the sum has no more than K_1 terms), so the step $\hat{\boldsymbol{p}}^{j,i_{j,k}}$ also converges to $\boldsymbol{0}$. This shows that the whole sequence of $\hat{\boldsymbol{w}}^{j,t}$ for $j \in K$ converges to $\boldsymbol{w}^*$.

Therefore, we have from the optimality condition of (4) that for $t \in T(j)$,

$$\begin{aligned}\operatorname{dist}(-\nabla f(\hat{\boldsymbol{w}}^{j,t+1}), \partial\Psi(\hat{\boldsymbol{w}}^{j,t+1})) &\le \|\hat{H}_{j,t}\hat{\boldsymbol{p}}^{j,t} + (\nabla f(\hat{\boldsymbol{w}}^{j,t+1}) - \nabla f(\hat{\boldsymbol{w}}^{j,t}))\| \\ &\le \|\hat{H}_{j,t}\|\|\hat{\boldsymbol{p}}^{j,t}\| + L\|\hat{\boldsymbol{p}}^{j,t}\| \\ &\le (C_1 + L)\|\hat{\boldsymbol{p}}^{j,t}\| \to 0,\end{aligned}$$

and again from Theorem 4.10 of (Lewis & Zhang, 2013), we know that after finite outer iterations, the sequence $\{\hat{\boldsymbol{w}}^{j,t+1}\}$ for $j \in K$ identifies and always stays within the correct manifold $\mathcal{M}^*$ (note that those $\hat{\boldsymbol{w}}^{j,0}$ are not included in this sequence). Since those semismooth Newton steps do not move the iterates away from its current manifold, they do not affect this result. We denote by $(\hat{j}, \hat{t})$ the first iteration from where on $\hat{\boldsymbol{w}}^{j,t}$ for $j \in K$ do not leave $\mathcal{M}^*$.

Finally, consider our original iterates and steps $\{\boldsymbol{w}^{j,t}\}$ and $\{\boldsymbol{p}^{j,t}\}$. Clearly, as our previous construction calibrates $\boldsymbol{w}^{j,1}$ and $\hat{\boldsymbol{w}}^{j,1}$, for any $j \in K$ satisfying $j > \hat{j}$, we know that $\boldsymbol{w}^{j,1}$ has already identified $\mathcal{M}^*$. Thus in our manifold selection strategy, we must have $\mathcal{M}^* \subseteq \mathcal{M}_{j,1}$. Then since the solution of (4) at the next iteration actually lies in $\tilde{\mathcal{M}}^*$, the subspace parallel to $\mathcal{M}^*$, we have that the respective solutions of (4) and (12) result in identical update steps. Therefore, we see that $\boldsymbol{p}^{j,t} = \hat{\boldsymbol{p}}^{j,t}$ for all $j \in K$ satisfying $j > \hat{j}$ and all $t \in T(j)$. By induction, the semismooth Newton steps also coincide. Therefore, for $j \in K$ and $j \ge \hat{j}$, $\boldsymbol{w}^{j,t} = \hat{\boldsymbol{w}}^{j,t}$ for all t. This suggests that the inner iterates stay within $\mathcal{M}^*$ as desired. □

D.7. Proof of Lemma C.1

Proof. Let the optimal solution to (12) be $\boldsymbol{t}^*_{j,t}$ and let $\boldsymbol{p}^*_{j,t} := B_{j,t}\boldsymbol{t}^*_{j,t}$, then it is clear from (16) and the strong convexity of the subproblem (from Lemma 3.2) that

$$|Q_{\hat{H}_{j,t}}(\boldsymbol{p}^{j,t};\boldsymbol{w}^{j,t})| = 0 \quad\Leftrightarrow\quad |Q_{\hat{H}_{j,t}}(\boldsymbol{p}^*_{j,t};\boldsymbol{w}^{j,t})| = 0 \quad\Leftrightarrow\quad \boldsymbol{p}^{j,t} = \boldsymbol{p}^*_{j,t} = \boldsymbol{0}. \tag{53}$$

The optimality condition of the subproblem (12) is

$$\mathbf{0} \in B_{j,t}^\top \left(\nabla f(\boldsymbol{w}^{j,t}) + \hat{H}_{j,t}\boldsymbol{p}_{j,t}^* + \partial \Psi \left(\boldsymbol{w}^{j,t} + \boldsymbol{p}_{j,t}^* \right) \right).$$

As $B_{j,t}$ is an orthonormal basis of $\tilde{\mathcal{M}}_{j,t}$, this holds true if and only if

$$\mathbf{0} \in P_{\tilde{\mathcal{M}}_{j,t}} \left(\nabla f(\boldsymbol{w}^{j,t}) + \hat{H}_{j,t}\boldsymbol{p}_{j,t}^* + \partial \Psi \left(\boldsymbol{w}^{j,t} + \boldsymbol{p}_{j,t}^* \right) \right). \tag{54}$$

When $|Q_{\hat{H}_{j,t}}(\boldsymbol{p}^{j,t}; \boldsymbol{w}^{j,t})| = 0$, (53) shows that $\boldsymbol{p}_{j,t}^* = \mathbf{0}$, and (54) implies that

$$\mathbf{0} \in P_{\tilde{\mathcal{M}}_{j,t}} \left(\nabla f(\boldsymbol{w}^{j,t}) + \partial \Psi \left(\boldsymbol{w}^{j,t} \right) \right) = \partial_{\tilde{\mathcal{M}}_{j,t}} F(\boldsymbol{w}^{j,t}) = \partial F|_{\mathcal{M}_{j,t}}(\boldsymbol{w}^{j,t}).$$

On the other hand, when $\mathbf{0} \in \partial F|_{\mathcal{M}_{j,t}}(\boldsymbol{w}^{j,t}) = \partial_{\tilde{\mathcal{M}}_{j,t}} F(\boldsymbol{w}^{j,t})$ holds, we have that

$$\begin{aligned} Q_{\hat{H}_{j,t}} \left(\boldsymbol{p}^{j,t}; \boldsymbol{w}^{j,t} \right) &= \nabla_{\tilde{\mathcal{M}}_{j,t}} f(\boldsymbol{w}^{j,t})^\top \boldsymbol{p}^{j,t} + \frac{1}{2} \left(\boldsymbol{p}^{j,t} \right)^\top \hat{H}_{j,t} \boldsymbol{p}^{j,t} + \Psi \left(\boldsymbol{w}^{j,t} + \boldsymbol{p}^{j,t} \right) - \Psi \left(\boldsymbol{w}^{j,t} \right) \\ &\geq \nabla_{\tilde{\mathcal{M}}_{j,t}} f(\boldsymbol{w}^{j,t})^\top \boldsymbol{p}^{j,t} + \frac{1}{2} \left(\boldsymbol{p}^{j,t} \right)^\top \hat{H}_{j,t} \boldsymbol{p}^{j,t} + \boldsymbol{s}^\top \boldsymbol{p}^{j,t}, \quad \forall \boldsymbol{s} \in \partial \Psi_{\tilde{\mathcal{M}}_{j,t}} \left(\boldsymbol{w}^{j,t} \right). \end{aligned} \tag{55}$$

Since $\mathbf{0} \in \partial_{\tilde{\mathcal{M}}_{j,t}} F(\boldsymbol{w}^{j,t})$, we see that $-\nabla_{\tilde{\mathcal{M}}_{j,t}} f(\boldsymbol{w}^{j,t}) \in \partial_{\tilde{\mathcal{M}}_{j,t}} \Psi \left(\boldsymbol{w}^{j,t} \right)$, so (55) and the positive definiteness of $\hat{H}_{j,t}$ imply that

$$\begin{aligned} 0 &\geq Q_{\hat{H}_{j,t}} \left(\boldsymbol{p}^{j,t}; \boldsymbol{w}^{j,t} \right) \\ &= \nabla_{\tilde{\mathcal{M}}_{j,t}} f(\boldsymbol{w}^{j,t})^\top \boldsymbol{p}^{j,t} + \frac{1}{2} \left(\boldsymbol{p}^{j,t} \right)^\top \hat{H}_{j,t} \boldsymbol{p}^{j,t} + \Psi \left(\boldsymbol{w}^{j,t} + \boldsymbol{p}^{j,t} \right) - \Psi \left(\boldsymbol{w}^{j,t} \right) \\ &\geq \mathbf{0}^\top \boldsymbol{p}^{j,t} + \frac{1}{2} \left(\boldsymbol{p}^{j,t} \right)^\top \hat{H}_{j,t} \boldsymbol{p}^{j,t} \geq 0, \end{aligned}$$

which shows that $Q_{\hat{H}_{j,t}}(\boldsymbol{p}^{j,t}; \boldsymbol{w}^{j,t}) = 0$.

Now consider the case that f is convex. The upper bound in (28) is straightforward from the acceptance condition (5) and that the largest possible function decrease is bounded by $\tilde{\Delta}_{j,t}$. For this part, convexity is actually not needed. We then consider the lower bound. We will denote the upper bound for (18) by R and we define

$$\Omega_{j,t} \coloneqq \left\{ \hat{\boldsymbol{w}}^* \mid \hat{\boldsymbol{w}}^* \in \mathcal{M}_{j,t}, F(\hat{\boldsymbol{w}}^*) = F_{j,t}^* \right\}$$

as the solution set for $F|_{\mathcal{M}_{j,t}}$. Clearly we have $0 \leq R < \infty$. From Lemma 3.2 we know that $\hat{H}_{j,t}$ has eigenvalues bounded in a range $[C_2, C_1]$ for some $C_1 \geq C_2 > 0$. We then get from (16) that

$$\begin{aligned} &\frac{1}{1-\eta} Q_{\hat{H}_{j,t}}(\boldsymbol{p}^{j,t}; \boldsymbol{w}^{j,t}) \\ &\leq Q_{\hat{H}_{j,t}}(\boldsymbol{p}_{j,t}^*; \boldsymbol{w}^{j,t}) \\ &= \min_{\boldsymbol{p}} \nabla f \left(\boldsymbol{w}^{j,t} \right)^\top \boldsymbol{p} + \frac{1}{2} \boldsymbol{p}^\top \hat{H}_{j,t} \boldsymbol{p} + \Psi \left(\boldsymbol{w}^{j,t} + \boldsymbol{p} \right) - \Psi \left(\boldsymbol{w}^{j,t} \right) \\ &\leq \min_{\boldsymbol{p}} f \left(\boldsymbol{w}^{j,t} + \boldsymbol{p} \right) + \Psi \left(\boldsymbol{w}^{j,t} + \boldsymbol{p} \right) + \frac{1}{2} \boldsymbol{p}^\top \hat{H}_{j,t} \boldsymbol{p} - F \left(\boldsymbol{w}^{j,t} \right) \\ &\leq \min_{\lambda \in [0,1]} F \left(\boldsymbol{w}^{j,t} + \lambda \left(P_{\Omega_{j,t}} \left(\boldsymbol{w}^{j,t} \right) - \boldsymbol{w}^{j,t} \right) \right) + \frac{\lambda^2}{2} \left(P_{\Omega_{j,t}} \left(\boldsymbol{w}^{j,t} \right) - \boldsymbol{w}^{j,t} \right)^\top \hat{H}_{j,t} \left(P_{\Omega_{j,t}} \left(\boldsymbol{w}^{j,t} \right) - \boldsymbol{w}^{j,t} \right) - F \left(\boldsymbol{w}^{j,t} \right) \\ &\leq \min_{\lambda \in [0,1]} (1-\lambda) F \left(\boldsymbol{w}^{j,t} \right) + \lambda F_{j,t}^* + \frac{\lambda^2}{2} \left(\boldsymbol{w}^{j,t} - P_{\Omega_{j,t}} \left(\boldsymbol{w}^{j,t} \right) \right)^\top \hat{H}_{j,t} \left(\boldsymbol{w}^{j,t} - P_{\Omega_{j,t}} \left(\boldsymbol{w}^{j,t} \right) \right) - F \left(\boldsymbol{w}^{j,t} \right) \\ &\leq \min_{\lambda \in [0,1]} -\lambda \tilde{\Delta}_{j,t} + \frac{\|\hat{H}_{j,t}\| \lambda^2}{2} \| \boldsymbol{w}^{j,t} - P_{\Omega_{j,t}} \left(\boldsymbol{w}^{j,t} \right) \|^2 \qquad (56) \\ &\leq \min_{\lambda \in [0,1]} -\lambda \tilde{\Delta}_{j,t} + \frac{C_1 R^2 \lambda^2}{2}. \qquad (57) \end{aligned}$$

The solution of (57) is

$$\lambda = \min\left\{1, \frac{\tilde{\Delta}_{j,t}}{C_1 R^2}\right\}.$$

When the former holds, it implies that

$$\frac{1}{1-\eta} Q_{\hat{H}_{j,t}}(\boldsymbol{p}^{j,t}; \boldsymbol{w}^{j,t}) \leq -\tilde{\Delta}_{j,t} + \frac{C_1 R^2}{2} \leq -\frac{\tilde{\Delta}_{j,t}}{2},$$

and when the latter holds, we get

$$\frac{1}{1-\eta} Q_{\hat{H}_{j,t}}(\boldsymbol{p}^{j,t}; \boldsymbol{w}^{j,t}) \leq -\frac{\tilde{\Delta}_{j,t}^2}{2C_1 R^2}.$$

These two combined then proves (28).

When $F|_{\tilde{\mathcal{M}}_{j,t}}$ further satisfies (17), (56) implies that

$$\frac{1}{1-\eta} Q_{\hat{H}_{j,t}}(\boldsymbol{p}^{j,t}; \boldsymbol{w}^{j,t}) \leq \min_{\lambda\in[0,1]} -\lambda\tilde{\Delta}_{j,t} + \frac{C_1\lambda^2\tilde{\Delta}_{j,t}}{\mu} = \min_{\lambda\in[0,1]} \left(\frac{C_1\lambda^2}{\mu} - \lambda\right)\tilde{\Delta}_{j,t},$$

and the final claim holds by taking

$$R_3 = -(1-\eta)\min_{\lambda\in[0,1]} \left(\frac{C_1\lambda^2}{\mu} - \lambda\right).$$

□

D.8. Proof of Theorem C.3

Proof. We first show that if $\boldsymbol{p}^{j,t}$ is a proximal gradient step (27) with the "step size" $\hat{\gamma}_{j,t} \in [C_2, C_1]$ for some $C_1 \geq C_2 > 0$, there exists $\tilde{L} > 0$ and $\delta_2 > 0$ such that

$$\|\boldsymbol{p}^{j,t}\| \leq \tilde{L}\|\boldsymbol{w}^{j,t} - \boldsymbol{w}^*\|, \forall \boldsymbol{w}^{j,t} \in \mathcal{M}^* \bigcap B_{\delta_2}(\boldsymbol{w}^*), \tag{58}$$

where $B_{\delta_2}(\boldsymbol{w}^*)$ is the open ℓ_2-norm ball centered at $\boldsymbol{w}^*$ with radius δ_2. Assume this claim is false, then there is a sequence $\{\boldsymbol{w}_i\}$ in $\mathcal{M}^*$ converging to $\boldsymbol{w}^*$ and the corresponding proximal gradient steps $\{\boldsymbol{p}_i\}$, with the step sizes being $\{\gamma_i\}$ and $\gamma_i \geq \gamma$ for some $\gamma > 0$, such that

$$\lim_{i\to\infty} \frac{\|\boldsymbol{w}_i - \boldsymbol{w}^*\|}{\|\boldsymbol{p}_i\|} = 0. \tag{59}$$

Note that the proximal gradient steps are obtained through

$$\boldsymbol{p}_i = \arg\min_{\boldsymbol{p}} \nabla f(\boldsymbol{w}_i)^\top \boldsymbol{p} + \frac{\gamma_i}{2}\boldsymbol{p}^\top\boldsymbol{p} + \Psi(\boldsymbol{w}_i + \boldsymbol{p}). \tag{60}$$

We know from Theorem 3.7 that $\boldsymbol{p}_i \in \tilde{\mathcal{M}}^*$ for i large enough, so we omitted the constraint of restricting the update within the current selected manifold as they eventually result in the same update step. Since $\boldsymbol{p}_i$ is the optimal solution to the strongly convex problem (60), we have that

$$\nabla f(\boldsymbol{w}_i)^\top \boldsymbol{p}_i + \frac{\gamma_i}{2}\boldsymbol{p}_i^\top\boldsymbol{p}_i + \Psi(\boldsymbol{w}_i + \boldsymbol{p}_i) \leq \nabla f(\boldsymbol{w}_i)^\top(\boldsymbol{w}^* - \boldsymbol{w}_i) + \frac{\gamma_i}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_i\|^2 + \Psi(\boldsymbol{w}^*). \tag{61}$$

We also have from the optimality condition $-\nabla f(\boldsymbol{w}^*) \in \partial\Psi(\boldsymbol{w}^*)$ that

$$\Psi(\boldsymbol{w}_i + \boldsymbol{p}_i) \geq \Psi(\boldsymbol{w}^*) - \nabla f(\boldsymbol{w}^*)^\top(\boldsymbol{w}_i + \boldsymbol{p}_i - \boldsymbol{w}^*).$$

Substituting the result above into (61), we get that

$$\frac{\gamma_i}{2}\boldsymbol{p}_i^\top\boldsymbol{p}_i \leq (\nabla f(\boldsymbol{w}^*) - \nabla f(\boldsymbol{w}_i))^\top(\boldsymbol{w}_i + \boldsymbol{p}_i - \boldsymbol{w}^*) + \frac{\gamma_i}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_i\|^2 \leq L\|\boldsymbol{w}_i - \boldsymbol{w}^*\|^2 + L\|\boldsymbol{w}_i - \boldsymbol{w}^*\|\|\boldsymbol{p}_i\| + \frac{\gamma_i}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_i\|^2.$$

After dividing both sides by $\gamma_i\|\boldsymbol{p}_i\|^2$ and letting i approach infinity, (59) indicates that the right-hand side approaches 0 as γ_i are lower-bounded, while the left-hand side is always $1/2$, giving a contradiction. Therefore, (58) holds true for some $\delta_2 > 0$ and some $\tilde{L} > 0$.

It is clear that since $\Psi|_{\mathcal{M}^*}$ is $\mathcal{C}^2$ around $\boldsymbol{w}^*$, its gradient is locally Lipschitz within $B_{\delta_2}(\boldsymbol{w}^*)$, and therefore $\nabla F|_{\mathcal{M}^*}$ is locally Lipschitz continuous with some constant $L_2 > 0$ within this range. Next, let H denote the generalized Hessian at $\boldsymbol{w}^{j,t+1}$ that we use for computing the truncated semismooth Newton step at the $(j, t+1)$ iteration and $\boldsymbol{p}^{j,t+1}$ denote this truncated semismooth Newton step, we have

$$
\begin{aligned}
&\|H\left(\boldsymbol{w}^{j,t+2} - \boldsymbol{w}^*\right)\| \\
=& \|H\left(\boldsymbol{w}^{j,t+1} + \boldsymbol{p}^{j,t+1} - \boldsymbol{w}^*\right)\| \\
=& \|H\boldsymbol{w}^{j,t+1} - \nabla F|_{\mathcal{M}^*}(\boldsymbol{w}^{j,t+1}) + \nabla F|_{\mathcal{M}^*}(\boldsymbol{w}^{j,t+1}) + H\boldsymbol{p}^{j,t+1} - H\boldsymbol{w}^*\| \\
\leq& \|\nabla F|_{\mathcal{M}^*}(\boldsymbol{w}^{j,t+1}) + H\boldsymbol{p}^{j,t+1}\| + \left\|\left(H\left(\boldsymbol{w}^{j,t+1} - \boldsymbol{w}^*\right) - \nabla F|_{\mathcal{M}^*}\left(\boldsymbol{w}^{j,t+1}\right) + \nabla F|_{\mathcal{M}^*}\left(\boldsymbol{w}^*\right)\right)\right\| \\
\leq& 0.1\|\nabla F|_{\mathcal{M}}(\boldsymbol{w}^{j,t+1})\|^2 + \left\|H\left(\boldsymbol{w}^{j,t+1} - \boldsymbol{w}^*\right) - \nabla F|_{\mathcal{M}^*}\left(\boldsymbol{w}^{j,t+1}\right) + \nabla F|_{\mathcal{M}^*}\left(\boldsymbol{w}^*\right)\right\| \\
\leq& 0.1L_2^2\|\boldsymbol{w}^{j,t+1} - \boldsymbol{w}^*\|^2 + o\left(\left\|\boldsymbol{w}^{j,t+1} - \boldsymbol{w}^*\right\|\right) = o\left(\left\|\boldsymbol{w}^{j,t+1} - \boldsymbol{w}^*\right\|\right).
\end{aligned}
$$

Notice that in the last inequality, we used the semismoothness of the gradient in (29) and the termination condition (26) of PCG. Finally, as $\boldsymbol{w}^{j,t+2}, \boldsymbol{w}^* \in \mathcal{M}^*$, we know that $\boldsymbol{w}^{j,t+2} - \boldsymbol{w}^* \in \bar{\mathcal{M}}^*$, and therefore from (30), we have that

$$
r\|\boldsymbol{w}^{j,t+2} - \boldsymbol{w}^*\| \leq \|H\left(\boldsymbol{w}^{j,t+2} - \boldsymbol{w}^*\right)\| \Rightarrow \|\boldsymbol{w}^{j,t+2} - \boldsymbol{w}^*\| = \frac{1}{r}o\left(\left\|\boldsymbol{w}^{j,t+1} - \boldsymbol{w}^*\right\|\right).
$$

Finally, from (58), we know that

$$
o\left(\left\|\boldsymbol{w}^{j,t+1} - \boldsymbol{w}^*\right\|\right) = o\left(\left\|\boldsymbol{w}^{j,t} - \boldsymbol{w}^*\right\|\right).
$$

Notice that when $\|\boldsymbol{w}^{j,t} - \boldsymbol{w}^*\|$ is small enough (modify δ_2 if necessary), $\|\boldsymbol{w}^{j,t+2} - \boldsymbol{w}^*\| \leq \delta_2$ can be guaranteed, and the superlinear convergence continues at the next step. Thus we have completed the proof. □