OPTIMAL METHODS FOR CONVEX RISK AVERSE DISTRIBUTED OPTIMIZATION

GUANGHUI LAN† AND ZHE ZHANG†

Abstract. This paper studies the communication complexity of convex risk averse optimization over a network. The problem generalizes the well-studied risk-neutral finite-sum distributed optimization problem and its importance stems from the need to handle risk in an uncertain environment. For algorithms in the literature, there exists a gap in communication complexities for solving risk-averse and risk-neutral problems. We propose two distributed algorithms, namely the distributed risk averse optimization (DRAO) method and the distributed risk averse optimization with sliding (DRAO-S) method, to close the gap. Specifically, the DRAO method achieves the optimal communication complexity by assuming a certain saddle point subproblem can be easily solved in the server node. The DRAO-S method removes the strong assumption by introducing a novel saddle point sliding subroutine which only requires the projection over the ambiguity set $P$. We observe that the number of $P$-projections performed by DRAO-S is optimal. Moreover, we develop matching lower complexity bounds to show that communication complexities of both DRAO and DRAO-S are not improvable. Numerical experiments are conducted to demonstrate the encouraging empirical performance of the DRAO-S method.

Keywords: risk-averse optimization, distributed optimization, first-order algorithm, convex optimization, lower complexity.

AMS 2000 subject classification: 90C25, 90C15, 68W15, 49M27, 49M29

1. Introduction. In this paper, we consider the following risk-averse optimization problem over a star-shape (worker-server) communication network [7]:

$$\min_{x \in X} \{f(x) := \max_{p \in P} \sum_{i=1}^{m} p_i f_i(x) - \rho^*(p) + u(x)\}, \tag{1.1}$$

where $P \subseteq \Delta^m := \{ p \in \mathbb{R}^m \mid \sum_{i=1}^{m} p_i = 1, p_i \geq 0 \}$ and $X \subseteq \mathbb{R}^n$ and $\Pi_i \subseteq \mathbb{R}^{m_i}$ are closed and convex. Functions $f_i(x)$, $u(x)$, and $\rho^*(p)$ are proper closed and convex. We assume the scenario (or local) cost function $f_i$ to be only available to worker node $i$ and focus on the situation where $f_i$'s are either all smooth or all structured non-smooth. We use the following generic representation for both types of $f_i$'s:

$$f_i(x) = \max_{\pi_i \in \Pi_i} \langle A_i x, \pi_i \rangle - f_i^*(\pi_i),$$

where $\Pi_i$ is a closed convex set and $f_i^*$ is a proper, closed and convex function. Specifically, if $f_i$ is smooth, then $A_i$ is the identity matrix $I \in \mathbb{R}^{n \times n}$, $f_i^*$ is the Fenchel conjugate to $f_i$, and $\Pi_i = \text{dom}(f_i^*)$ [7]. If $f_i$ is structured non-smooth [28], then $A_i \in \mathbb{R}^{m_i \times n}$ is a linear operator, $\Pi_i$ is bounded, and the $f_i^*$-prox mapping can be solved efficiently [10]. This type of structured non-smooth function has found a wide range of applications, including total variation regularization in image processing [31], low-rank tensor [12, 37], overlapped group lasso [21, 35], and graph regularization [49, 35]. Additionally, we assume the (strongly) convex regularization term $u(x)$ and $\rho^*, P$ are both available to the server node.

Observe that if the ambiguity set $P$ consists of only a single probability mass function $\bar{p}$, then (1.1) is called risk neutral and it can be written as a finite sum problem (see Chapter 5 of [16]):

$$\min_{x \in X} \sum_{i=1}^{m} \bar{p}_i f_i(x) + u(x). \tag{1.2}$$

However, if the costs among workers are imbalanced (different importance, limited availability of data, etc.), taking an average over the costs across workers might be meaningless or operationally wrong. In such cases, non-trivial $\rho^*$ and $P$ in (1.1) generalizes risk-neutral optimization to risk-averse optimization and distributionally robust optimization (DRO). Specifically, if $f := (f_1, \ldots, f_m)$ denotes the scenario costs and
\( \rho \) is a convex risk measure, it can be formulated as (1.1) using Fenchel conjugates (see Definition 6.4 and Theorem 6.5 of [34]):

\[
\rho(f) := \text{arg max}_{p \in P} \langle p, f \rangle - \rho^*(p). \tag{1.3}
\]

For example, if we denote the (reference) probability mass function by \( \bar{p} \), some widely used risk measures and their conjugates are given as follows.

- **Mean semideviation of order \( r \):**
  \[
  \rho(f) = \sum_{i=1}^{m} \bar{p}_i f_i + c(\sum_{i=1}^{m} \bar{p}_i |f_i - \mathbb{E}[f_i]|_r/r) = \max_{p \in P} \rho(p, f),
  \]
  where the ambiguity set \( P := \{ p \in \Delta_m^n : 3 \bar{c}_i \geq 0 \text{ s.t. } p_i = \bar{p}_i(1 + \bar{c}_i - (\bar{c}_i, \bar{p})_s) \text{, } \|\bar{c}_i\|_s \leq c \}, c \in [0, 1] \) and \( ||\cdot||_s \) is the conjugate norm to \( ||\cdot||_r \), i.e., \( 1/s + 1/r = 1 \).

- **Entropic risk:**
  \[
  \rho(f) = \tau^{-1} \log \sum_{i=1}^{m} \bar{p}_i \exp(\tau f_i) = \max_{p \in \Delta_m^n} \langle p, f \rangle - \tau^{-1} \sum_{i=1}^{m} p_i \log(p_i/\bar{p}_i). \tag{1.4}
  \]

- **Distributionally robust objective:**
  \[
  \rho(f) := \sup_{p \in P} \langle f, p \rangle \text{ for some uncertainty set } P.
  \]

The incorporation of all the above risk measures makes our problem (1.1) more challenging than the finite sum problem (1.2). We note that (1.1) also covers a popular risk measure CV@R with \( \rho(f) = \max_{p \in \Delta_m^n} \langle p, f \rangle - \rho^*(p) \). The risk measure admits a finite-sum reformulation, \( \rho(f) = \inf_{t} \sum_{i=1}^{m} \bar{p}_i \{ |f_i - t|_+/(\alpha + t) \} \), but the function \( \tilde{f}_i(x, t) := [f_i(x) - t]_+/\alpha + t \) is nonsmooth with a very large Lipschitz-continuity constant, even if the original \( f_i \) is smooth. In contrast, our conjugate formulation avoids the situation.

As alluded earlier, we assume that the communication network has a star topology with a computationally powerful server node connected directly to many worker nodes. During a communication round, all the worker nodes send their local information to the central server and the server node broadcasts solutions to all worker nodes. This type of distributed optimization framework is very popular in machine learning, such as federated learning [11], where the data are held privately in each worker (device) and the central server learns a global model by communicating with the workers. Since communication in a network tends to be slower than computation inside a single node by orders of magnitude and less communications implies a better protection of privacy, one of the main goals of this paper is to study the system’s communication complexity, i.e., the number of communication rounds required to find a point \( \tilde{x} \in X \) s.t. \( f(\tilde{x}) - f(x^*) \leq \epsilon \), where \( x^* \) denotes an optimal solution of (1.1).

Risk averse optimization problems of form (1.1) have a wide range of applications in portfolio selection [22], renewable energy [23], power security [10], telecommunication [19] and climate change planning [30]. As a concrete example, the massive multiple-input multiple-output (MIMO) system in 5G communication network consists of multiple active antennas and terminal devices [30] [19]. The active antennas at the base station should be configured to ensure stable connections for all the terminals in its service area. Such an objective can be formulated as (1.1) with \( f_i \) being the negative data speed at the \( i \)th terminal device and \( (P, \rho^*) \) being the conjugates of the mean-semideviation risk measure. To gather information for the downlink and uplink channels, the base station needs to communicate with terminal devices. In a highly mobile environment, it is crucial to find an optimal antenna configuration within only a small number of communication rounds. A second example is motivated by climate change. The state government may wish to invest in infrastructure to prepare for climate change. Each scenario cost function \( f_i \) may denote the long-term economic cost estimated by a certain climate model and a certain impact model [30]. In order to avoid the downside risk, \( (P, \rho^*) \) could be chosen as the conjugate to the some risk measures mentioned above, say the entropic risk measure. Because these models involve large amounts of data and costly simulation runs, we might need to store \( f_i \)'s on separate computing nodes and use a communication network to find the optimal policy.

Our formulation is also applicable for the computationally demanding distributionally robust optimization (DRO). DRO provides a powerful framework for either learning from limited data [13] or data-driven decision making [2] [38]. Under the assumption of a finite scenario support \( \Xi = [\xi_1, \ldots, \xi_m] \), we could use \( f_i(x) := f(x, \xi_i) \) to denote the cost under scenario \( \xi_i \) and choose \( \rho \) to be the risk measure induced by the corresponding probability uncertainty set [51]. Implemented on a distributed communication network with the evaluation of \( f(x, \xi_i) \)'s performed in parallel on different machines, a small number of communication rounds is essential for fast computation.

Additionally, the risk-averse formulation in (1.1) could also be useful for federated learning between organizations, i.e., the cross-silo federated learning [11]. Cross-silo federated learning has found applications in finance risk prediction in reinsurance [39], drug discovery [5], electronic health record mining [6] and smart
If the workers represent demographically partitioned organizations or geographically partitioned data centers, we could choose $\rho$ to be the mean-semideviation risk measure to ensure that the trained model offers consistent performance for the different populations. Moreover, the risk measures may also provide incentives for competing organizations to cooperate. For example, consider the operations of competing airlines, we can choose $f_i(x)$ to be the expected negative improvement in profit of the $i^{th}$ airline and choose $\rho(f(x)) := \max_{i \in [m]} f_i(x)$ to guarantee benefits for every participant. In both cases, a smaller number of communication rounds implies a better protection of privacy.

Despite the importance of problem (1.1), however, the study of its communication complexity and the development of efficient algorithms are rather limited. Since (1.1) can be viewed as a trilinear saddle point problem, we can potentially apply some recently developed first-order algorithms (e.g. 40 41) for solving it. However, these methods are designed without special consideration for communication burden. Specifically, the most related algorithm is the sequential dual (SD) method, which was first proposed in [40] for the implementation of the SD method has communication complexities of $O(D_1D_Xo/\sqrt{\epsilon} + D_PD_1MA_D/Xo/\epsilon)$ and $O(D_1D_Xo/\epsilon + D_PD_1MA_D/Xo/\epsilon)$ for the smooth and the structured non-smooth problems, respectively. Here $L_f$, $D_1$, $MA_P$, and $D_Xo$ correspond to the the overall smoothness constant, the dual radius, the operator norm of $A_1$, the radius of $P$, and the distance to optimal solution (see Tables 1.1 and 1.2, and Section 3 for their precise definitions). On the other hand, for the risk-neutral problem (1.2) with $P := \{\bar{p}\}$, direct distributed implementations of the Nesterov accelerated gradient method [26] and the primal-dual algorithm 33 can achieve communication complexities of $O(\sqrt{L_1D_Xo}/\sqrt{\epsilon})$ and $O(D_1MA_D/Xo/\epsilon)$ for the smooth and the structured non-smooth problems, respectively, which were shown to be tight (see, e.g., 33). Clearly, there exists a significant gap in communication complexities, especially for smooth problems where the $O(1/\epsilon)$ communication complexity for the risk-averse setting is an order of magnitude larger than the $O(1/\sqrt{\epsilon})$ bound for the risk-neutral setting. Therefore we pose the following research question:

| Can we solve the risk averse problem over a star-shape network with the same communication complexity as the finite-sum problem? |

This paper intends to provide a positive answer for this question through the following three steps.

First, we propose a conceptual distributed risk-averse optimization (DRAO) method. It is inspired by works of Nesterov (Section 2.3.1 of 27) and Lan [14] on composite optimization of the form $\min_{x} \rho(f(x))$ for a smooth vector function $f$. While Nesterov 27 considers the problem with $\rho(f(x)) = \max_{i=1,...,m} f_i(x)$, Lan 14 generalizes $\rho$ to any monotone convex function. They are able to achieve an $O(1/\sqrt{\epsilon})$ oracle complexity of $f$ by incorporating the following inner-linearization prox-mapping into the accelerated gradient descent (AGD) method or a bundle variant into the accelerated prox-level (APL) method:

$$x_t \leftarrow \arg\min_{x \in X} \rho(f_i(x_t^t) + \langle \nabla f_i(x_t^t), x - x_t^t \rangle, ..., f_m(x_t^t) + \langle \nabla f_m(x_t^t), x - x_t^t \rangle) + \frac{\rho}{2} \|x - x_t^t\|^2 \tag{1.4}$$

Such an update is a simplified version of (1.1) with $f_i(x_t^t) + \langle \nabla f_i(x_t^t), x - x_t^t \rangle$ denoting some (iterative) linearization of $f_i$ at $x_t^t$ and $\frac{\rho}{2} \|x - x_t^{t-1}\|^2$ being the proximal term. Similarly, we modify the SD method by combining the $p$ and $x$-prox updates into a single $(x,p)$-prox update given by

$$x_t \leftarrow \arg\min_{x \in X} \max_{p \in B} \sum_{i=1}^{m} p_i \left[ (x, A_i \pi_i^t) - f_i^*(\pi_i^t) \right] - \rho^*(p) + u(x) + \frac{\rho}{2} \|x - x_t^{t-1}\|^2 \tag{1.5}$$

where $(x, A_i \pi_i^t) - f_i^*(\pi_i^t)$ also represents some iterative linearization of $f_i$ specified by the dual variable $\pi_i^t$. In fact, rewriting $\rho$ in its primal form [1.3] shows that (1.5) is equivalent to

$$x_t \leftarrow \arg\min_{x \in X} \rho \left( (x, A_1 \pi_1^t) - f_1^*(\pi_1^t), ..., (x, A_m \pi_m^t) - f_m^*(\pi_m^t) \right) + u(x) + \frac{\rho}{2} \|x - x_t^{t-1}\|^2 \tag{1.3}$$

$$x_t \leftarrow \arg\min_{x \in X} \rho \left( (x, A_1 \pi_1^t) - f_1^*(\pi_1^t), ..., (x, A_m \pi_m^t) - f_m^*(\pi_m^t) \right) + u(x) + \frac{\rho}{2} \|x - x_t^{t-1}\|^2 \tag{1.3}$$

$$x_t \leftarrow \arg\min_{x \in X} \rho \left( (x, A_1 \pi_1^t) - f_1^*(\pi_1^t), ..., (x, A_m \pi_m^t) - f_m^*(\pi_m^t) \right) + u(x) + \frac{\rho}{2} \|x - x_t^{t-1}\|^2 \tag{1.3}$$
which matches \( \pi_t^i \) if \( \pi_t^i \) is selected to be \( \nabla f_i(x^t) \) for smooth \( f_i \)'s. Such a modification of the SD method leads to the DRAO method. As shown in Table 1.1, it achieves the optimal first-order (FO) oracle complexities for \( f \) (or \( \Pi \)-projection complexities) for both the smooth and the structured non-smooth problems. Since \((\rho^*, P)\) is available to the server, the above \((x, p)\)-prox update can be performed entirely on the server. So the method achieves the desired communication complexities shown in Table 1.1. However, this approach requires \( \rho \) to be simple so that (1.5) can be efficiently solved. This assumption might be too strong in practice. For example, if \( m \) is large, the \((x, p)\)-prox update in (1.5) with either the mean-semideviation \( \rho \) or the Kantorvich ambiguity set \( P \) can be computationally challenging.

### Table 1.1: Communication Complexity and FO Oracle Complexity of \( f \) for DRAO and DRAO-S

|                        | Convex (\( \alpha = 0 \)) | Strongly convex (\( \alpha > 0 \)) |
|------------------------|-----------------------------|----------------------------------|
| Smooth                 | \( O(\sqrt{L_f}||x^0 - x^*||/\sqrt{\epsilon}) \) | \( O(\sqrt{L_f}/\alpha \log(1/\epsilon)) \) |
| Structured Non-smooth  | \( O(M_A D_{\Pi} ||x^0 - x^*||/\epsilon) \) | \( O(M_A D_{\Pi}/\sqrt{\epsilon}) \) |

|                        | Convex (\( \alpha = 0 \)) | Strongly convex (\( \alpha > 0 \)) |
|------------------------|-----------------------------|----------------------------------|
| Smooth                 | \( O(D_P M ||x^0 - x^*||/\epsilon) \) | \( O((L_f/\alpha)^{1/4} M_D P/\alpha \sqrt{\epsilon}) \) |
| Structured Non-smooth  | \( O(D_P M_{\Pi} ||x^0 - x^*||/\epsilon) \) | \( O(D_P M_{\Pi}/\sqrt{\epsilon}) \) |

\( M_A = \max_{\in \Xi} \|A_i\|, \quad D_{\Pi} = \max_{\in \Xi} \|\pi_i - \bar{\pi}_i\| \).

\( D_P \) denotes \( P \)-'s radius. \( M \) denotes the operator norm of \( \|\nabla f_1(x), \ldots, \nabla f_m(x)\| \) over some bounded ball around \( x^* \) and \( M_{\Pi} \) denotes the operator norm of \( \|A_i \pi_1, \ldots, A_m \pi_m\| \) over the whole feasible region \( \Pi \).

Number of P-projections required to generate an \( \epsilon \)-close solution, i.e., \( \|x^N - x^*\|^2 \leq \epsilon \).

Second, we overcome the restrictive assumption of \( \rho \) being simple in DRAO by developing a saddle point sliding (SPS) subroutine. It replaces the \((x, p)\)-prox update step in the DRAO method by performing only a finite number of \( P \)-projections and \( X \)-projections to solve the saddle point subproblem associated with \( \pi^t \) in (1.5) inexactely. The new method, called distributed risk averse optimization with sliding (DRAO-S), maintains the same communication complexities as DRAO. DRAO-S is also computationally efficient. Since each inner iteration of the sliding subroutine requires one \( P \)-projection and one \( X \)-projection, the optimal numbers of these projections are optimal in most cases. As shown in Table 1.2, they match the lower bounds for solving a single \((x, p)\) bi-linear saddle point problem, i.e., (1.5) with a fixed \( \pi^t \) and \( \eta_t = 0 \). Such a result is similar to that of the gradient sliding (GS) method \([15]\) for solving an additive composite problem.

\[
\min_{x \in X} f(x) + g(x). \tag{1.6}
\]

The GS method can achieve both optimal \( f \)-oracle and optimal \( g \)-oracle complexities. However, our nested composite problem appears to be more challenging. This is because for a fixed \( x \), the optimal dual variables \( p \) and \( \pi \) in (1.1) are dependent, while the optimal dual variables \( \pi f \) and \( \pi g \) associated with the saddle point reformulation of (1.6) through conjugate duality are independent. In fact, (1.6) can always be rewritten as a nested composite problem (see the discussion in Example 3 of (14)). Additionally, the SPS subroutine in the DRAO-S method is initialized differently from the usual sliding subroutines in \([15]\) and \([17]\). Such a modification simplifies both the outer loop algorithm and the convergence analysis. So it may allow easier applications of the sliding technique to other problem settings. Furthermore, an interesting feature of the DRAO-S method is that its inner loop, the SPS subroutine, can adjust dynamically to the varying levels of difficulty, characterized by \( \|\pi^t\| \), of the saddle point subproblem (1.5). This allows us to remove the assumption of the smooth \( f_i \)'s being Lipschitz continuous, which is required by the SD method in (11), but may not hold if the domain \( X \) is unbounded.

Third, we show that the communication complexities of both DRAO and DRAO-S are not improvable by constructing lower complexity bounds. Previous developments are restricted to a trivial \( P \) and the smooth

---

\( ^2 \)Except for the strongly convex smooth problem which is worse off by a factor of \((L_f/\alpha)^{1/4}\).
problem \cite{32}. We propose a more general computation model which includes both the $f_r$-gradient oracle and the $f^*_i$-prox mapping oracle, and introduce a different set of problem parameters appropriate for the risk-averse problem. They allow us to develop, for a non-trivial $P$ and for both the smooth and the structured non-smooth problems, new lower complexity bounds matching the upper communication complexity bounds possessed by DRAO and DRAO-S.

The rest of the paper is organized as follows. Preliminary Section 2 reviews a gap function in \cite{40} which will guide the algorithm design. Then Section 3 and Section 4 propose and analyze the DRAO and DRAO-S methods, respectively. Section 5 provides lower communication complexity bounds and Section 6 provides some encouraging numerical results. Finally, some concluding remarks are made in Section 7.

1.1. Notation & Assumptions. The following assumptions and notations will be used throughout the paper.

- The set of optimal solutions to (1.1), $X^*$, is nonempty, and $x^*$ is an arbitrary optimal solution. $f_*$ denotes the optimal objective, $f(x^*)$. $R_0$ denotes an estimate of the distance from the initial point to $x^*$, i.e., $R_0 \geq \|x^0 - x^*\|$.
- $D_P$ denotes the radius of $P$, i.e., $D_P := \max_{p,\tilde{p}\in P} \sqrt{2U(p,\tilde{p})}$ if a Bregman distance $U$ is used.
- $f : \mathbb{R}^n \to \mathbb{R}^m$ denotes a vector of scenario cost functions, $[f_1; \ldots; f_m]$, and $\nabla f(x) : \mathbb{R}^n \to \mathbb{R}^{m \times n}$ denotes the Jacobian matrix function.
- We refer to the following computation as either a prox mapping or a projection:
  \[
  \hat{w} \leftarrow \arg\min_{w \in \mathbb{W}} (g(w) + h(w) + \tau V(w; \hat{w}),
  \]
  where the vector $g$ represents some “descent direction” (the gradient for example), and $h(w)$ is a simple convex function \cite{16}. Moreover, $V$ denotes the Bregman distance, $\hat{w}$ is a prox center, and $\tau$ is a stepsize parameter. Together they ensure the output $\hat{w}$ is close to $\hat{w}$. In particular, we call it an $x$, a $\pi$, or a $p$-prox mapping (an $X$, a $\Pi$ or a $P$-projection) if $W = X$ and $h \equiv 0$, $W = \Pi$ and $h = f^*_i$, or $W = P$ and $h = \rho^* \equiv 0$, respectively. Sometimes, the term prox update also is used to emphasize that the prox mapping is performed to update $w^t := \hat{w}$ from $w := \hat{w}^{-1}$.

2. Preliminary: $Q$-gap function. In this section, we introduce a gap function \cite{40} which will guide our algorithmic development throughout the rest of the paper. For notation convenience, we denote $\pi \equiv (\pi_1, \ldots, \pi_m)$ and $\Pi \equiv \Pi_1 \times \Pi_2 \times \ldots \times \Pi_m$ so that (1.1) can be written as

\[
\min_{x \in X} \max_{p \in P} \max_{\pi \in \Pi} \{L(x; p, \pi) := \sum_{i=1}^m p_i \langle A_i x, \pi_i \rangle - f^*_i(\pi_i) - \rho^*(p) + u(x)\}. \tag{2.1}
\]

The following duality result holds (see Proposition 2.1 of \cite{40}) for the preceding trilinear saddle point problem.

**Lemma 2.1.** Let $f$ and $\mathcal{L}$ be defined in (1.1) and (2.1), then the following statements hold for all $x \in X$.

a) **Weak Duality:** \(f(x) \geq \mathcal{L}(x; p, \pi)\) for all $p \in P$, $\pi \in \Pi$.

b) **Strong Duality:** \(f(x) = \mathcal{L}(x; \hat{p}, \hat{\pi})\) for any $\hat{\pi}_i \in \arg\max_{\pi_i \in \Pi} \langle A_i x, \pi_i \rangle - f^*_i(\pi_i)$, $i = 1, \ldots, m$, and any $\hat{p} \in \arg\max_{p \in P} \sum_{i=1}^m p_i f_i(x)$.

We measure the quality of a feasible solution $z = (x, p, \pi)$ by a gap function $Q$ associated with some feasible reference point $\hat{z} := (\hat{x}; \hat{p}, \hat{\pi})$:

\[
Q(z; \hat{z}) := \mathcal{L}(x; \hat{p}, \hat{\pi}) - \mathcal{L}(x; p, \pi). \tag{2.2}
\]

The following result provides a bound on the function optimality gap from above by $Q$.

**Lemma 2.2.** Let $Q$ be defined in (2.2), then

\[
f(x) - f(x^*) \leq \max_{\hat{p} \in P, \hat{\pi} \in \Pi} Q((x; p, \pi); (x^*; \hat{p}, \hat{\pi})). \tag{2.3}
\]

Moreover, the optimal solution $x^*$ of (1.1), together with some $\pi^*_i \in \arg\max_{\pi_i \in \Pi} \langle A_i x^*, \pi_i \rangle - f^*_i(\pi_i)$, $i = 1, \ldots, m$, and some $p^* \in \arg\max_{p \in P} \sum_{i=1}^m p_i f_i(x^*)$ forms a saddle point $z^* := (x^*; p^*; \pi^*)$ of (2.1), i.e.,

\[
Q(z; z^*) \geq 0, \ \forall z \equiv (x; p, \pi) \in X \times P \times \Pi. \tag{2.4}
\]
Proof. Let \( \hat{p} \) and \( \hat{\pi} \) be defined in Lemma 2.1. By Lemma 2.1 we have \( f(x) - f(x^*) \leq \mathcal{L}(x, \hat{p}, \hat{\pi}) - \mathcal{L}(x^*, p, \pi) = Q((x; p, \pi); (x^*; \hat{p}, \hat{\pi})) \), from which (2.3) follows immediately. Next, the first order optimality condition of (1.1) implies that there exist some \( \pi^* \in \arg \max_{x \in \Pi} \langle A_i x, \pi_i \rangle - f_i^*(\pi_i) \) such that \( \sum_{i=1}^m p_i^* A_i^\top \pi_i^* + g^*, x - x^* \geq 0 \) for any \( x \in X \). This observation together with the definition of \( \mathcal{L} \) in (2.1) then imply that

\[
\mathcal{L}(x; p^*, \pi^*) \geq \mathcal{L}(x^*; p^*, \pi^*), \forall x \in X.
\]

Moreover, due to our choice of \( (p^*, \pi^*) \), Lemma 2.1 also implies that

\[
f(x^*) = \mathcal{L}(x^*; p^*, \pi^*) \geq \mathcal{L}(x^*; p, \pi), \forall (p, \pi) \in P \times \Pi.
\]

Then (2.4) follows from combining the preceding two inequalities. \( \square \)

In view of Lemma 2.2 we can use \( Q \) to guide our search for an \( \epsilon \)-optimal solution. In particular, we decompose \( Q \) into three component gap functions given by

\[
Q(z; \hat{z}) = Q_x(z; \hat{z}) + Q_p(z; \hat{z}) + Q_\pi(z; \hat{z})
\]

with

\[
\begin{align*}
Q_\pi(z; \hat{z}) &:= \mathcal{L}(\hat{x}; \hat{p}, \hat{\pi}) - \mathcal{L}(\hat{x}; \hat{p}, \hat{\pi}) = \sum_{i=1}^m \hat{p}_i \left( \langle A_i \hat{x}, \hat{\pi}_i \rangle - f_i^*(\hat{\pi}_i) \right). \\
Q_p(z; \hat{z}) &:= \mathcal{L}(z; \hat{p}, \hat{\pi}) - \mathcal{L}(z; \hat{p}, \hat{\pi}) = \sum_{i=1}^m (\hat{p}_i - \hat{p}_i) \left( \langle A_i \hat{x}, \pi_i \rangle - f_i^*(\pi_i) \right). \\
Q_x(z; \hat{z}) &:= \mathcal{L}(\hat{x}; \hat{p}, \hat{\pi}) - \mathcal{L}(\hat{x}; \hat{p}, \hat{\pi}) = \sum_{i=1}^m \hat{p}_i A_i^\top \hat{\pi}_i, \hat{x} - \hat{x}) + u(\hat{x}) - u(\hat{x}).
\end{align*}
\]

3. Upper Bounds for Communication Complexity. In this section, we propose the distributed risk averse optimization (DRAO) method to provide upper bounds on communication complexity. The algorithm and its convergence properties are presented in Subsection 3.1 and its convergence analysis is presented in Subsection 3.2.

3.1. DRAO method. The DRAO method is designed for solving the min-max-max trilinear saddle point problem in (2.1). It is inspired by two algorithms for optimizing nested composite problems. First, the sequential dual (SD) algorithm, proposed in [10], [11], achieves the optimal dependence of iteration complexity on every layer of nested function by performing sequential proximal updates to their dual variables. DRAO employs similar sequential proximal updates for \( \pi, p \) and \( x \) to obtain optimal dependence of iteration complexity on \( f \). Second, the accelerated prox-level (APL) algorithm, proposed in [14], can reduce the number of outer iterations further by solving a more complicated proximal sub-problem (1.4). Here, DRAO combines the separate \( p \) and \( x \) proximal updates into a single \( (x, p) \) prox update step in the server node to save communication.

Algorithm 1 describes a generic DRAO method which will be later specialized for solving the smooth and the structured nonsmooth problems. As shown in Algorithm 1 the server first sends an extrapolated point \( \hat{x} \) to the workers for them to perform dual proximal updates in Line 3. The only goal is to reduce the component gap function \( Q_{\pi_i} \) (c.f. (2.3)). Here we intentionally leave the prox-function \( V_i \) in an abstract form because its selection and the resulting implementation of the proximal update will depend on the smoothness properties of \( f_i \). Next, the server collects the newly generated \( A_i \pi_i^* \) in Line 4 to solve the \( (x, p) \) prox update problem in Line 5 so as to reduce both \( Q_x \) and \( Q_p \).
Observe that in the generic DRAO algorithm, we assume subproblems in Line 3 and 5 to be solved exactly by the workers and server, respectively. In fact, Line 3 reduces to local gradient evaluations in the smooth case, while requiring a prox mapping for the structured nonsmooth case. Line 5 requires us to solve a structured bilinear saddle point problem. We will discuss in more details how to solve these problems approximately in next section, while focusing on the communication complexity now.

First, we consider the smooth problem where all $A_i$’s are identity matrices and all $f_i$’s are smooth such that $\|\nabla f_i(x_1) - \nabla f_i(x_2)\| \leq L_i \|x_1 - x_2\|$, $\forall x_1, x_2 \in \mathbb{R}^n$. Since the Fenchel conjugate to a smooth convex function is strongly convex [8], a natural choice of the prox-function $V_i$ in DRAO would be the Bregman distance function generated by $f_i^*$ given by

$$W_{f_i^*}(\pi_i; \tilde\pi_i) := f_i^*(\pi_i) - f_i^*(\tilde\pi_i) - \langle (f_i^*)'(\tilde\pi_i), \pi_i - \tilde\pi_i \rangle.$$  \hfill eq:dual_bregman

It has been shown in [18] [16] [11] that the $\pi_i$ proximal update in Line 3 of Algorithm 1 is equivalent to a gradient evaluation. Specifically, with $\bar{x} = x_0$ and $\pi_0 = \nabla f_i(\bar{x})$, Line 3 reduces to the following steps:

$$x^t \leftarrow (\bar{x} + \tau_t x^{t-1})/(1 + \tau_t),$$  \hfill eq:smooth_alg1

$$\pi_i^t \leftarrow \nabla f_i(x^t),$$  \hfill eq:smooth_alg2

$$f_i^*(\pi_i^t) \leftarrow \langle \bar{x}, \pi_i^t \rangle - f_i(x^t).$$  \hfill eq:smooth_alg3

Plugging $f_i^*(\pi_i^t)$ defined in \[eq:smooth_alg3\] into Line 5 of Algorithm 1, we can completely remove any information about the conjugate function $f_i^*$. Therefore, DRAO is a purely primal algorithm for the smooth problem.

To discuss the convergence properties of DRAO, we need to properly define some Lipschitz smoothness constants. For a given $p \in P$, let us denote $f_p(x) := \sum_{i=1}^m p_i f_i(x)$. Clearly $f_p$ is a smooth convex function with Lipschitz continuous gradients, i.e., $\|\nabla f_p(x_1) - \nabla f_p(x_2)\| \leq L_p \|x_1 - x_2\|, \forall x_1, x_2 \in X$. Moreover, $L_p \leq \sum_{i=1}^m p_i L_i$. We define an aggregate smoothness constant $L_f$ to characterize the overall smoothness property of the risk averse problem (2.1):

$$L_f = \max_{p \in P} L_p.$$  \hfill eq:smo_cst

Observe that in the risk neutral case where $P = \{(1/m, \ldots, 1/m)\}$, $L_f$ is the global smoothness constant of $f$ [33], which is upper bounded by $\frac{1}{m} \sum_{i=1}^m L_i$. In the robust case when $P = \Delta^+_m$, $L_f = \max_i L_i$.

Theorem 3.1 and 3.2 below show the convergence rates of the DRAO method applied to the aforementioned smooth problems, for non-strongly convex $u(x)$ and strongly convex $u(x)$ respectively. Their proofs are given in Section 3.2.

**Theorem 3.1.** Let $L_f$ be defined in (3.5). If $\{x^t\}_{t=1}^T$ are generated by the DRAO method applied to smooth problems with

$$\omega_t = t, \ \theta_t = (t - 1)/t, \ \tau_t = (t - 1)/2, \ \eta_t = 2L_f/t.$$
Then for a reference point \( \hat{z} := (\hat{x}; \hat{p}, \hat{\pi}) \) in which \( \hat{\pi} = \nabla f_i(\hat{x}) \) for some \( \hat{x} \in X \), we have
\[
\sum_{t=1}^{N} \omega_t Q(z^t; \hat{z}) + L_f \| x^N - \hat{x} \|^2 \leq L_f \| x^0 - \hat{x} \|^2. \tag{3.6}
\]
In particular, the ergodic solution \( \bar{x}^N \) satisfies
\[
f(\bar{x}^N) - f(x^*) \leq 2L_f \| x^0 - x^* \|^2 / N(N + 1), \forall N \geq 1. \tag{3.7}
\]

**Theorem 3.2.** Let \( L_f \) be defined in (3.5). Assume, in addition, that \( u(x) \) is \( \alpha \)-strongly convex for some \( \alpha > 0 \). Let \( \kappa := L_f / \alpha \) denote the condition number. If \( \{x^t\}_{t=1}^{N} \) are generated by the DRAO method applied to smooth problems with
\[
\theta_t = \theta := \frac{\sqrt{\kappa + 1} - 1}{\sqrt{\kappa + 1}}, \quad \omega_t = \omega := (\frac{1}{\theta})^t - 1, \quad \tau_t = \tau := \frac{\sqrt{\kappa + 1} - 1}{2}, \quad \eta_t = \eta := \frac{\alpha}{\sqrt{\kappa + 1}}, \tag{3.8}
\]
Then for a reference point \( \hat{z} := (\hat{x}; \hat{p}, \hat{\pi}) \) in which \( \hat{\pi} = \nabla f_i(\hat{x}) \) for some \( \hat{x} \in X \), we have
\[
\sum_{t=1}^{N} \omega_t Q(z^t; \hat{z}) + \frac{\alpha(\sqrt{\kappa + 1} - 1)}{4\theta^2} \| x^N - \hat{x} \|^2 \leq \frac{\alpha}{\kappa + 1}(\| x^0 - \hat{x} \|^2 + L_f \| x^0 - x^* \| ). \tag{3.9}
\]
In particular, the last iterate \( x^N \) converges geometrically:
\[
\| x^N - x^* \|^2 \leq \theta^N (1 + \kappa) \| x^0 - x^* \|^2, \forall N \geq 1. \tag{3.10}
\]

We make two remarks regarding the above convergence results. First, selecting the saddle point \( z^* \) defined in Lemma 2.2 as \( \hat{z} \), Theorem 3.1 (c.f. (3.6)) and 3.2 (c.f. (3.9)) imply that all generated iterates, \( \{x^t\}_{t=1}^{N} \), are inside some ball around \( x^* \):
\[
\| x^t - x^* \| \leq (1 + L_f / \alpha) \| x^0 - x^* \| \quad \text{if } \alpha > 0.
\]
This shows that the search space for \( x^t \) is essentially bounded. Such a property will become useful when we solve saddle point subproblem in Line 5 of DRAO approximately in the next section. Second, Theorem 3.1 and 3.2 imply, respectively, \( O(L_f \| x^0 - x^* \|^2 / \sqrt{\alpha}) \) and \( O(\sqrt{L_f / \alpha} \log(1/\epsilon)) \) communication complexities to find \( \epsilon \)-optimal solutions. It is interesting note that with \( L_f \) defined in (3.5), these results are valid even if \( P \) is larger than the probability simplex, i.e., \( \Delta_m^P \subsetneq P \subset R_m^P \). We will show later in Section 5 that these communication complexity bounds are not improvable in general.

Next, let us consider the structured non-smooth problem. Because \( f_i^* \) may not be strongly convex, the Bregman distance function \( W_{f_i^*} \) (c.f. (3.1)) is no longer suitable for \( \pi_t \) prox update. Instead we choose \( \forall_{i} \pi_t := \frac{1}{2} \| \pi_t - \hat{\pi}_t \|^2 \), so that the \( \pi_t \) proximal update is given by:
\[
\pi_t \leftarrow \text{arg max}_{\pi_t} (A_t x^t, \pi_t) - f_i^*(\pi_t) - \frac{\eta}{2} \| \pi_t - \pi_t^{-1} \|^2. \tag{3.11}
\]

Theorem 3.3 below states the convergence properties of Algorithm 1 applied to the structured nonsmooth problem and its proof is provided in Section 3.2. We need to define the maximum linear operator norm \( M_A \) and the maximum dual radius \( D_{\Pi} \) as
\[
M_A := \max_{i \in [m]} \| A_i \|_{2,2}, \quad D_{\Pi} := \max_{i \in [m]} \max_{\pi_t, \pi_t^{-1}} \| \pi_t - \hat{\pi}_t \|. \tag{3.12}
\]
Note that \( M_A D_{\Pi} \) provides an estimate of the Lipschitz continuity constant of \( \sum_{i} p_i f_i(x) \).

**Theorem 3.3.** Let a structured non-smooth risk-averse problem (1.1) be given. Let \( M_A \) and \( D_{\Pi} \) be defined above in (3.12) and let \( R_0 \geq \| x^0 - x^* \| \).

1. If \( \alpha = 0 \) and the stepsizes satisfy
\[
\omega_t = 1, \quad \theta_t = 1, \quad \lambda_t = M_A D_{\Pi} / R_0, \quad \tau_t = M_A R_0 / D_{\Pi},
\]
then the following convergence rate holds for the solution $\bar{x}^N$ returned by the DRAO algorithm
\[ f(\bar{x}^N) - f(x^*) \leq M_A D^0 R_0 / N. \]  
\tag{3.13} \]  

b) If $\alpha > 0$ and the stepsizes satisfy
\[ \omega_t = t, \ \theta_t = (t - 1) / t, \ \eta_t = t\alpha / 3, \ \tau_t = 3M_2^2 / t\alpha, \]
then the following convergence rate holds for the solution $\bar{x}^N$ returned by the DRAO algorithm
\[ f(\bar{x}^N) - f(x^*) \leq \left( \alpha \| x^0 - x^* \|^2 / 3 + 3M_2^2 D^2_\Pi / \alpha \right) / N(N + 1). \]  
\tag{3.14} \]  

The preceding theorem gives us $O(R_0 D^1_\Pi M_A / \epsilon)$ and $O(M_A D^1_\Pi / \sqrt{\epsilon N})$ communication complexities for solving the structured nonsmooth problem under the non-strongly convex and the strongly convex settings, respectively. These complexity bounds are worse than those of the smooth problem by an order of magnitude. Hence, it is interesting to observe that the smoothness properties of the scenario cost functions have significant impact on communication complexity, even under the assumption that the workers are equipped with the capability to solve the $\pi_i$ proximal update in (3.11).

3.2. Convergence analysis. Our main goal in this subsection is to establish the convergence rates associated with the DRAO method stated in Theorems 3.1, 3.2, and 3.3.

We will first show some general convergence properties about the generic DRAO method in Algorithm 1. Notice that this result holds regardless of the strong convexity of $f_i^*$ (i.e., $\mu = 0$ is allowed in (3.15)) and the strong convexity of $u$ (i.e., $\alpha = 0$ is allowed), and thus can be applied to both smooth and nonsmooth problems under either convex or strongly convex settings.

**Proposition 3.4.** Let $\{z^t \equiv (x^t; p^t, \pi^t)\}_{t=1}^N$ be generated by Algorithm 1 for some $p^t \in \arg\max_{p \in P} \sum_{i=1}^m p_i(\langle x^t, \pi^t \rangle - f_i^*(\pi^t))$. Let $\hat{z} := (\hat{x}; \hat{p}, \hat{\pi}) \in X \times P \times \Pi$ be a given reference point (c.f. (2.2)). Let $\mu$ be a non-negative constant satisfying
\[ f_i^*(\pi_i) - f_i^*(\hat{\pi}_i) - \langle g_i^*(\pi_i), \pi_i - \hat{\pi}_i \rangle \geq \mu V_i(\pi_i; \hat{\pi}_i), \ \forall \pi_i, \hat{\pi}_i \in \Pi, \forall i \in [m]. \]  
\tag{3.15} \]  

If there exists a positive constant $q$ satisfying
\[ \sum_{i=1}^m p_i V_i(\pi_i^1; \pi_i^{t-1}) \geq \frac{1}{2q} \left\| \sum_{i=1}^m p_i A_i^\top (\pi_i^t - \pi_i^{t-1}) \right\|^2, \ \forall t \geq 2, \ \forall p \in P, \]  
\tag{3.16} \]  
and the stepsizes satisfy the following conditions for all $t \geq 2$:
\[ \omega_{t-1} = \omega_t \theta_t, \]  
\[ \eta_{t-1} \tau_t \geq \theta_t q, (\tau_N + \mu) \eta_N \geq q, \]  
\[ \omega_t \eta_t \leq \omega_{t-1} (\eta_{t-1} + \alpha), \omega_t \tau_t \leq \omega_{t-1} (\tau_{t-1} + \alpha), \]
then the next bound is valid for all $\hat{z} := (\hat{x}; \hat{p}, \hat{\pi}) \in X \times P \times \Pi$ and $N \geq 1$:
\[ \sum_{i=1}^N \omega_i Q_i(z^t; \hat{z}) + \omega_{N-1} (\eta_N + \alpha) \left\| x^N - \hat{x} \right\|^2 \leq \frac{\omega_p^N}{2} \left\| x^0 - \hat{x} \right\|^2 + \omega_1 \tau_1 \sum_{i=1}^m p_i V_i(\hat{\pi}_i; \pi_i^0). \]  
\tag{3.17} \]  

**Proof.** Let $Q$, $Q_p$ and $Q_x$ be defined in (2.5). We begin by analyzing the convergence of $Q$. It follows from the definition of $\tilde{x}^t$ that
\[ \langle \tilde{\pi}_i - \pi_i^t, A_i(\tilde{x}^t - x^t) \rangle = -\langle \tilde{\pi}_i - \pi_i^{t-1}, A_i(x^t - x^{t-1}) \rangle + \theta_t \langle \tilde{\pi}_i - \pi_i^{t-1}, A_i(x^{t-1} - x^{t-2}) \rangle \]
\[ + \omega_t \tau_t \sum_{i=1}^m p_i V_i(\hat{\pi}_i; \pi_i^t). \]  
\[ \tag{3.18} \]  
\[ \text{We assume the strong convexity modulus $\alpha$ to be small such that the $M_2^2 D^2_\Pi / \alpha$ term dominates in (3.14).} \]
Since $\partial u$ gradient evaluation point $x$

So, combining (3.21) and (3.20), taking a $p$

The optimality condition for the dual update in Line 3 of Algorithm 1 (see Lemma 3.1 of [16]) implies

$$\sum_{i=1}^{N} \omega_i \langle \pi_i - \tilde{\pi}_i, A_i x^t \rangle + f_i^*(\pi_i) - f_i^*(\tilde{\pi}_i) + \langle \pi_i - \tilde{\pi}_i, A_i (\hat{x}^t - x^t) \rangle$$

$$\leq \tau V_i(\tilde{\pi}_i; \pi_i^{-1}) - (\tau_i + \mu) V_i(\tilde{\pi}_i; \pi_i) - \tau_i V_i(\pi_i; \pi_i^{-1})$$

So, combining the above two relations, taking the $\omega_t$ weighted sum of the resulting inequalities and using the conditions that $\omega_{t-1} = \omega_t \theta_t$ and $\omega_t \tau_t \leq \omega_{t-1}(\tau_{t-1} + \mu)$, we obtain

$$\sum_{i=1}^{N} \omega_i \langle \pi_i - \tilde{\pi}_i, A_i x^t \rangle + f_i^*(\pi_i) - f_i^*(\tilde{\pi}_i)$$

$$\leq - (\omega_N(\tau_N + \mu) V_i(\tilde{\pi}_i; \pi_i) - \omega_N(\tau_i - \pi_i N, A_i (x_N - x_i N - 1)))$$

$$- \sum_{i=1}^{N} [(\omega_i \tau_i V_i(\pi_i; \pi_i^{-1}) + \omega_{t-1}(\sum_{i=1}^{N} \tilde{\pi}_i A_i^T(\pi_i; \pi_i^{-1}), x_N - x_i N^{-1})) + \omega_1 \tau_1 V_i(\tilde{\pi}_i; \pi_i)]$$

A $\hat{p}_i$-weighted sum of the above inequality leads to the desired $Q_\pi$ convergence bound given by

$$\sum_{i=1}^{N} \omega_i Q_\pi(z_i, \hat{z})$$

$$\leq - (\omega_N(\tau_N + \mu) \sum_{i=1}^{N} \hat{p}_i V_i(\tilde{\pi}_i; \pi_i) - \omega_N(\sum_{i=1}^{N} \hat{p}_i A_i^T(\pi_i; \pi_i), x_N - x_i N^{-1}))$$

$$- \sum_{i=1}^{N} [(\omega_i \tau_i \sum_{i=1}^{m} \hat{p}_i V_i(\pi_i; \pi_i^{-1}) + \omega_{t-1}(\sum_{i=1}^{N} \tilde{\pi}_i A_i^T(\pi_i; \pi_i^{-1}), x_N - x_i N^{-1})) + \omega_1 \tau_1 \sum_{i=1}^{m} \hat{p}_i V_i(\tilde{\pi}_i; \pi_i)]$$

Next, we consider $x^t$ and $p^t$ generated by Line 5 in Algorithm 1. Denote the proximal sub-problem by $F(x; \pi^t) := \max_{p \in P} \sum_{i=1}^{m} p_i [\langle x, v_i \rangle - f_i^*(\pi_i)] - \rho^*(p) + u(x) + \frac{\theta}{2} \left\| x - x_i^{t-1} \right\|^2$. Since $x^t \in \arg \min_{x \in X} F(x; \pi^t)$, the first order necessity condition implies the existence of some maximizer $p^t$ and some subgradient $u'(x^t) \in \partial u(x^t)$ such that

$$\sum_{i=1}^{m} p_i u_i^t + \eta(x^t - x_i^{t-1}) + u'(x^t) \in -\partial \delta_X(x^t)$$

$$\Rightarrow \langle \sum_{i=1}^{m} p_i A_i \pi_i, x^t - x_i \rangle + \langle \eta(x^t - x_i^{t-1}) + u'(x^t), x_i^{t-1} \rangle \leq 0.$$ 

Since $u(x^t) + \alpha \left\| x^t - \hat{x} \right\|^2 / 2 - u(\hat{x}) \leq \langle u'(x^t), x^t - \hat{x} \rangle$, due to the $\alpha$-strong convexity of $u$, and $\frac{\theta}{2} \left\| x^t - x_i^{t-1} \right\|^2 - \frac{\eta}{2} \left\| x_i^{t-1} - \hat{x} \right\|^2 = \langle \eta_t(x^t - x_i^{t-1}), x_i^{t-1} - \hat{x} \rangle$, we get

$$Q_x(z_i, \hat{z}) + \frac{\eta}{2} \left\| x_i^{t-1} - \hat{x} \right\|^2 \leq \frac{\eta}{2} \left\| x^t - x_i^{t-1} \right\|^2 - \frac{\eta}{2} \left\| x^t - x_i^{t-1} \right\|^2.$$ 

Additionally, being a maximizer, $p^t$ satisfies

$$p^t \in \arg \max_{p \in P} \sum_{i=1}^{m} p_i [\langle x, A_i \pi_i \rangle - f_i^*(\pi_i)] - \rho^*(p)$$

$$\Rightarrow \sum_{i=1}^{m} \hat{p}_i - \rho^*(p) \langle \pi_i, A_i \pi_i \rangle - f_i^*(\pi_i) + \rho^*(p) - \rho^*(\hat{p}) \leq 0 \Rightarrow Q_p(z_i, \hat{z}) \leq 0.$$ 

So, combining (3.21) and (3.20), taking a $\omega_t$-weighted sum of the resulting inequality and using $\omega_t \eta_t \leq \omega_{t-1}(\eta_t - \alpha)$, we obtain

$$\sum_{i=1}^{N} \omega_t (Q_x(z_i, \hat{z}) + Q_p(z_i, \hat{z})) + \frac{\omega_t}{2} \left\| x^N - \hat{x} \right\|^2 \leq \omega_t \left\| x^0 - \hat{x} \right\|^2 - \sum_{i=1}^{N} \omega_t \left\| x^t - x_i^{t-1} \right\|^2.$$ 

Then utilizing (3.16) and the Young’s inequality, (3.18) follows immediately by adding (3.22) to (3.19).

We now specialize the results obtained in Proposition 3.4 for the smooth problems. Observe that the gradient evaluation point $x^t$ is common for all workers. This allows us to easily characterize the strong convexity modulus of the aggregate prox-penalty function $\sum_{i=1}^{m} \hat{p}_i W_i f_i(x)$ (c.f. (3.16)) in the next lemma.

**Lemma 3.5.** Let $\hat{p} \in P$ be given. If $\pi_i = \nabla f_i(x)$ and $\bar{\pi}_i = \nabla f_i(\bar{x})$ for some $x$ and $\bar{x}$, then

$$\sum_{i=1}^{m} \hat{p}_i W_i f_i(\pi_i; \bar{\pi}_i) \geq \frac{1}{2L_f} \left\| \sum_{i=1}^{m} \hat{p}_i A_i^T(\pi_i; \bar{\pi}_i) \right\|^2.$$ 

(3.23)
Proof. Let \( f_\bar{x}(x) := (\sum_{i=1}^{m} \hat{p}_i f_i(x)) \). Then by the definition of \( L_f \) in (3.5), \( f_\bar{x} \) is \( L_f \)-smooth and its conjugate \((f_\bar{x})^*\) is \( 1/L_f \)-strongly convex. Next, we relate \( D(f_\bar{x})^* \) to \( \sum_{i=1}^{m} \hat{p}_i W_{f_i}^* \) to calculate its strong convexity modulus. Since \( \sum_{i=1}^{m} \hat{p}_i \pi_i \in \nabla f_\bar{x}(x), \sum_{i=1}^{m} \hat{p}_i \pi_i \in \nabla f_\bar{x}(x) \), we have by Fenchel duality that
\[
\sum_{i=1}^{m} \hat{p}_i (\pi_i - \bar{\pi}_i, \nabla f_\bar{x}^* (\bar{\pi}_i)) = \sum_{i=1}^{m} \hat{p}_i (\pi_i - \bar{\pi}_i, \bar{x}) = \langle \sum_{i=1}^{m} \hat{p}_i (\pi_i - \bar{\pi}_i), \nabla f_\bar{x}^* (\sum_{i=1}^{m} \hat{p}_i \pi_i) \rangle,
\]
and similarly, \( \sum_{i=1}^{m} \hat{p}_i f_i^* (\pi_i) = (f_\bar{x})^* (\sum_{i=1}^{m} \hat{p}_i \pi_i) \). Thus
\[
\sum_{i=1}^{m} \hat{p}_i W_{f_i}^* (\pi_i; \bar{\pi}_i) = (f_\bar{x})^* (\sum_{i=1}^{m} \hat{p}_i \pi_i) - (f_\bar{x})^* (\sum_{i=1}^{m} \hat{p}_i \pi_i), \nabla f_\bar{x}^* (\sum_{i=1}^{m} \hat{p}_i \pi_i)
= D(f_\bar{x})^* (\sum_{i=1}^{m} \hat{p}_i \pi_i; \sum_{i=1}^{m} \hat{p}_i \pi_i) \geq \frac{1}{2L_f} \| \sum_{i=1}^{m} \hat{p}_i (\pi_i - \bar{\pi}_i) \|^2 = \frac{1}{2L_f} \| \sum_{i=1}^{m} \hat{p}_i \bar{A}_i^* (\pi_i - \bar{\pi}_i) \|^2,
\]
where the last inequality follows from \( A_i^* = I \) in smooth problems.

We are now ready to prove Theorems 3.1 and 3.2.

Proof of Theorem 3.1. We apply Proposition 3.4 to obtain the convergence result in (3.7). Since \( f_i^* \) is 1-strongly convex with respect to \( W_{f_i}^* \), \( \mu = 1 \) satisfies condition (3.15). Since \( \pi_i^* = \nabla f_i(x_i^*) \) and \( \bar{\pi}_i = \nabla f_i(\bar{x}) \) for some \( \bar{x}, q = L_f \) satisfies condition (3.16) (c.f. Lemma 3.5). Moreover, since stepsizes proposed in Theorem 3.1 verifies (3.17), all the requirements in Proposition 3.4 are met. Thus Proposition 3.4 leads to (3.6), i.e.,
\[
\sum_{t=1}^{N} \omega_t L(x_t^t: \hat{\pi}, \bar{\pi}) - \sum_{t=1}^{N} \omega_t L(x_t^*: p_t, \pi^*) + L f \| x_N - x^* \|^2 \leq L f \| x_0 - x^* \|^2.
\]
In particular, with \( \hat{\pi}_N^* = \nabla f_N(x_N^*) \) and \( \hat{\rho}_N^* \in \arg \max_{\rho \in P} \sum_{i=1}^{m} \hat{p}_i f_i(\hat{x}_N^*) \) such that \( f(\hat{x}_N^*) = L(\hat{x}_N^*; \hat{\rho}_N^*, \hat{\pi}_N^*) \) (see Lemma 2.1), we have
\[
\sum_{t=1}^{N} \omega_t L(x_t^t: \hat{\pi}_N^*, \bar{\pi}_N) - \sum_{t=1}^{N} \omega_t L(x_t^*: p_t, \pi^*) \leq L f \| x_N - x^* \|^2.
\]
Because \( L(\cdot; \hat{\rho}_N^*, \hat{\pi}_N^*) \) is linear for \( x \), the first term is equal to
\[
\sum_{t=1}^{N} \omega_t L(x_t^*: p_t, \pi^*) = \frac{N(N+1)}{2} L(\hat{x}_N^*; \hat{\rho}_N^*, \hat{\pi}_N^*) = \frac{N(N+1)}{2} f(\hat{x}_N^*).
\]
Due to the weak duality in Lemma 2.1, the second term is upper bounded by
\[
\sum_{t=1}^{N} \omega_t f(x_t^*) \leq \sum_{t=1}^{N} \omega_t f(x_t^*) = \frac{N(N+1)}{2} f(x^*).
\]
Then the desired inequality in (3.7) follows immediately.

Proof of Theorem 3.2. Similar to the preceding proof, the proposed stepsizes (c.f. 3.8), together with \( \mu = 1 \) and \( q = L_f \), verify the requirements in Proposition 3.4 thus
\[
\sum_{t=1}^{N} \omega_t (L(x_t^*: \hat{\pi}, \bar{\pi}) - L(x_t^*: p_t, \pi^*)) + \omega_N (\eta + \alpha) \| x_N - x^* \|^2 \leq \omega_1 (\eta \| x_0 - x^* \|^2 + \tau \sum_{i=1}^{m} \hat{p}_i W_{f_i}^* (\hat{\pi}_i; \pi_i^*)).
\]
Using the relation of conjugate Bregman distance functions and the identity \( \nabla (\sum_{i=1}^{m} \hat{p}_i f_i(x)) = \sum_{i=1}^{m} \hat{p}_i \nabla f_i(x) \), the last term can be upper bounded by
\[
\sum_{i=1}^{m} \hat{p}_i W_{f_i}^* (\hat{\pi}_i; \pi_i^*) = \sum_{i=1}^{m} \hat{p}_i (D_f(x_0^*; \bar{\pi})) = (\sum_{i=1}^{m} \hat{p}_i f_i(x^*)) - (\sum_{i=1}^{m} \hat{p}_i f_i(x) - \langle \nabla (\sum_{i=1}^{m} \hat{p}_i f_i(x), x_0^* - \bar{x} \rangle \leq \frac{L_f}{2} \| x_0^* - \bar{x} \|^2.
\]
Thus the \( Q \) convergence bound (c.f. 3.9) follows immediately:
\[
\sum_{t=1}^{N} \omega_t Q(x_t^t; \hat{x}) + \frac{\omega_1 (\sqrt{\frac{3m+1}{4m}} - 1)}{\omega_1} \| x_N - \hat{x} \|^2 \leq \frac{\omega_1 (\sqrt{\frac{3m+1}{4m}} - 1)}{\omega_1} \| x_0^* - \hat{x} \|^2 + L f \| x^* - \hat{x} \|^2.
\]
Additionally, setting the preceding $\hat{z}$ to the saddle point $z^*$ defined in Lemma 2.2 (c.f. (2.4)) such that $Q(z^t; z^*) \geq 0 \forall t$ and dividing both sides by $\alpha(\sqrt{N+1}-1)/4\delta^n$, the geometric convergence of $x^N$ to $x^*$ (c.f. (3.10)) can be deduced.

Now we move on to present convergence proofs for structured non-smooth problems.

**Proof of Theorem 3.3** First, we consider part a) with a non-strongly convex $u(x)$. The result is also a consequence of Proposition 3.4. Since $V_i(\pi_i; \hat{\pi}_i) := \frac{1}{2} \|P_i - \hat{\pi}_i\|^2 \geq \frac{1}{2\pi_n} \|A_i^T (\pi_i - \hat{\pi}_i)\|^2$, the condition (3.16) is satisfied with $q = M^{2}_{x}$ due to the Jensen’s inequality. Since $f^*_i$ is convex, the condition (3.16) is satisfied with $\mu = 0$. Additionally, since the chosen stepizes in Theorem 3.3 satisfy the condition (3.17), all the requirements for Proposition 3.4 are met. Thus for any feasible $\hat{z}$, we have

$$\sum_{t=1}^{N} \omega_t Q(z^t; \hat{z}) + \frac{\omega_t \delta^2}{2} \|x^N - \hat{x}\|^2 \leq \frac{\omega_t \delta^2}{2} \|x^0 - \hat{x}\|^2 + \frac{\omega_t \delta^2}{2} \sum_{i=1}^{m} \hat{p}_i \|\hat{\pi}_i - \pi_i\|^2.$$  

Let $\hat{\pi}_N \in \arg\max_{\pi_i \in \Pi} \langle \pi_i, A_i \hat{x}^N \rangle - f_i^*(\pi_i)$ and $\hat{p}_N \in \arg\max_{p \in P} \sum_{i=1}^{m} p_i f_i(\hat{x}^N)$ such that $f(\hat{x}^N) = \mathcal{L}(\hat{x}^N; \hat{p}_N, \hat{\pi}_N)$ (c.f. Lemma 2.1). Setting $\hat{z}$ to $z^N := (x^*; \hat{p}_N, \hat{\pi}_N)$ leads to

$$\sum_{t=1}^{N} \omega_t Q(z^t; \hat{z}) \leq \frac{\omega_t \delta^2}{2} R_0^2 + \frac{\omega_t \delta^2}{2} D_R = R_0 M_A D_R.$$  

Then the resulting convergence bound in (3.13) can be deduced using the fact $(\sum_{t=1}^{N} \omega_t) f(\hat{x}^N) - f(x^*) \leq \sum_{t=1}^{N} \omega_t Q(z^t; \hat{z}).$

As for part b), the derivation is the same except for a different set of stepsizes selected to take advantage of the $\alpha$-strong convexity of $u(x)$.

### 4. The DRAO-S method

The practical application of the DRAO method is limited by the requirement of computing exact solutions to the following saddle point problem in Line 5 of Algorithm 1

$$x^t \leftarrow \arg\min_{x \in \mathcal{X}} \max_{p \in P} \sum_{i=1}^{m} p_i [\langle x, v^t_i \rangle - f_i^*(\pi^t_i)] - \rho^*(p) + u(x) + \frac{\eta_t}{2} \|x - x^t\|^2 \quad \text{(4.1)}$$

We relax this requirement by only assuming the ability to efficiently compute the $p$-prox mapping defined in (1.7). In the DRO setting, the efficient implementations for risk measures induced by several probability uncertainty sets are described in [19]. For the entropic risk measure, if we select prox-function to be $U(p; \tilde{p}) := \sum_{i=1}^{m} p_i \log(p_i/\tilde{p}_i)$, the computation amounts to only a softmax evaluation. For the mean semi-deviation risk of order two, the computation can be implemented as a quadratically constrained quadratic program (QCQP).

In this section, we design a novel saddle point sliding (SPS) subroutine to solve (4.1) inexactly and call the resulting method **distributed risk-averse optimization with sliding (DRAO-S)**. We show that the communication complexity of DRAO-S will be in the same order of magnitude as DRAO. Moreover, the total number of $P$-projections required by DRAO-S will be mostly optimal, in the sense that it is equivalent to the optimal one required for solving problem (1.1) with linear local cost functions $f_i$’s.

**4.1. The Algorithm and Convergence Results.** The SPS subroutine for solving (4.1) inexactly is presented in Algorithm 2. It is closely related to the classic primal dual (PD) algorithm (see 3 [16]) for solving a structured bi-linear saddle point problem given by

$$\min_{y \in \mathcal{Y}} \max_{p \in P} \langle v^t, y, p \rangle - \rho^*(p),$$

where the matrix $v^t$ is obtained from stacking $v^t_1, v^t_2, \ldots, v^t_m$ from Line 3 of Algorithm 1 horizontally. In iteration $s$, the subroutine computes an extrapolated prediction of $S_i \in P$ in Line 2, a $y$-prox update in Line 3, and then a $p$-prox update in Line 4. The $p$-prox update utilizes a general Bregman distance function $U$, since different choices of $U$ may take advantage of different geometry of $P$. At the end of $S_t$ iterations, SPS returns a weighted ergodic average of $\{y^s\}$ as an approximate solution of problem (4.1).
The DRAO-S method is obtained by making the following two modifications to DRAO. First, we supply an additional pair of initial points \((p^0, y^0) \in P \times X\). For simplicity we set \(y^0 = x^0\) and \(p^{-1} = p^0\). Second, we replace the definition of \(x^t\) in Line 5 of DRAO with the output solution of the SPS subroutine according to:

\[
(x^t, y^t, p^t, p^t) = SPS(x^{t-1}, y^{t-1}, p^{t-1}, p^{t-1}, \{v^t_i\}, \{v^{t-1}_i\} | \eta_t, \{\delta_i\}, \{\gamma_i\}, \{\beta_i\}, \{q_s\}, S_t; \forall t \geq 1. \tag{4.2}
\]

Due to its similarity to the DRAO method, we call the outer loop of the DRAO-S method (Algorithm 1) with modification (4.2) the outer DRAO loop and say a phase of the DRAO-S method happens if \(t\) is increased by 1. Accordingly, we call the inner loop of the DRAO-S method (Algorithm 2) the inner SPS loop and say an (inner) iteration happens if \(t\) is incremented by 1. Intuitively, if the inner iteration limits \(S_t\) are large enough, \(x^t\) obtained from the inner SPS loop will be a good approximate solution of (4.1). However, the \(P\)-projection complexity, i.e., the total number of inner iterations given by \(T_t\), might be too large. In addition, observe the computation burdens of the server subproblem and the worker subproblem during each communication round are different: the server requires several rounds of \(P\) and \(X\)-projections for the SPS subroutine, while the worker needs just one \(\pi_e\)-prox mapping. However, since the server is often more powerful, e.g. the cloud-edge system, this asymmetry may have limited impact on the overall computation performance of the system.

We note here that the DRAO-S method is related to the Primal Dual Sliding (PDS) method in [17], of which the sliding subroutine is also a primal dual type algorithm. However, since we are dealing with a nested trilinear saddle point problem, rather than the sum of two bilinear saddle point problems, the DRAO-S method differs from the PDS method in two important ways. First, the linear operator \(v^t\) for the saddle point subproblem (c.f. (4.1)) changes in every phase. To coordinate consecutive inner SPS loops, we construct a special momentum term using both the current \(v^t\) and the previous \(v^{t-1}\) when transitioning into a new phase, i.e., Line 2 of Algorithm 2. This construction is inspired by the novel momentum term in the SD method [40]. Second, as opposed to the single initialization in both the PDS method and the GS method, the \(y^t\) prox update in Line 3 of Algorithm 2 utilizes two distinct initialization points, the ergodic average \(x^{t-1}\) and the last iterate obtained in the last phase, \(y^0\). Even though both are approximate solutions of (4.1), \(y^t\) is used only in the inner SPS loop while the ergodic average \(x^t\) is used in the outer loop. As will be discussed in the next subsection, this decoupling appears to significantly simplify, as compared to other sliding algorithms [13, 17], the convergence analysis and the selection of stepsizes.

Now, let us first consider the smooth problem. The stepsizes associated with the inner SPS loop need to adapt dynamically in two aspects. First, the inner iteration limit \(S_t\) needs to be an increasing function of \(t\) to maintain the same communication complexity as the DRAO method. Second, as a primal dual type algorithm, the inner SPS loop stepsizes, \(\gamma_t\), \(\delta_t\) and \(\beta_t\), need to satisfy a certain condition related to the operator norm of \(v^t\), i.e., \(\gamma_t \beta_t \geq \delta_t \|v^t\|^2\), to ensure convergence. Specifically, if the Bregman distance function \(U\) in Line 4 of Algorithm 2 is 1-strongly convex with respect to \(\|\cdot\|_U\), the operator norm of interest

---

**Algorithm 2 Saddle Point Sliding (SPS) Subroutine**

**Input:** Initial points \(x^{t-1}, y^0 \in X, p^0, p^{-1} \in P\), and gradients \(\{v^t_i\}, \{v^{t-1}_i\}\). Non-negative stepsizes \(\eta_t, \{\delta_i\}, \{\gamma_i\}\) and \(\{\beta_i\}\), averaging weights \(\{q_s\}\), and iteration number \(S_t\).

1. for \(s = 1, 2, 3, \ldots S_t\) do
2. \[\bar{v}^s \leftarrow \begin{cases} \sum_{i=1}^m p_i v^t_i + \delta_i \sum_{i=1}^m (p^0_i - p^{-1}_i)v^{t-1}_i & \text{if } s = 1, \\ \sum_{i=1}^m p_i v^{t-1}_i + \delta_i \sum_{i=1}^m (p^{t-1}_i - p^{-2}_i)v^t_i & \text{if } s \geq 2. \end{cases} \]
3. \[y^s \leftarrow \arg \min_{y \in X} (y, \bar{v}^s) + u(y) + \frac{\eta_s}{2} \|y - y^{s-1}\|^2 + \frac{\eta_s}{2} \|y - x^{t-1}\|^2. \]
4. \[P_t \leftarrow \arg \max_{p \in P} \sum_{i=1}^m p_i (\langle v^t_i, y^s \rangle - f_i^* (\pi_i)) - p^* (p) - \eta_s U (p; p^{s-1}). \]
5. end for
6. return \(x^t := \sum_{s=1}^{S_t} q_s y^s/\sum_{s=1}^{S_t} q_s, y^t := y^{S_t}, \bar{p}^t := \sum_{s=1}^{S_t} q_s p^s/\sum_{s=1}^{S_t} q_s, p^t := p^{S_t}\) and 

\[y^t = p^{S_t-1}. \]
is given by

$$M_t := \|v^t\|_{2, U^*} := \max_{\|p\| \leq 1, \|y\| \leq 1} \sum_{l=1}^{m} p_l(v_l^t)^\top y.$$ 

Under the smooth setting, however, a prior bound on $M_t$ may not exist if $X$ is unbounded. So we choose $\gamma_s^t$ and $\beta_s^t$ to adjust dynamically to $M_t$ in each phase. In particular, the next theorem presents the stepsize choice and the convergence result under the non-strongly convex setting.

**Theorem 4.1.** Let a smooth risk-averse problem (c.f. (1.1)) be given. Let $M_t$ and $L_f$ be defined in (4.3) and (3.5), and let $R_0$ and $D_P$ be defined in Subsection 1.1. If $\{x^t\}_{t=1}^N$ are generated by the DRAO-S method (4.2) with the following stepsizes:

$$\omega_t = t, \quad \theta_t = (t-1)/(t), \quad \tau_t = (t-1)/2, \quad \eta_t = 2L_f/t,$$

$$\Delta > 0, \quad S_t = [t\Delta M_t], \quad \tilde{M}_t = \frac{s_t t}{\Delta}, \quad \beta_s^t = \beta^t = \frac{D_P M_t}{R_0}, \quad \gamma_s^t = \gamma^t = \frac{R_0 M_t}{D_P}, \quad \text{stp:sps_sm}$$

then the solution $\bar{x}^N$ returned by the outer DRAO loop satisfies

$$f(\bar{x}^N) - f(x^*) \leq \frac{2L_f^2R_0^2}{N(N+1)} + \frac{2D_P R_0 N}{N(N+1)\Delta}, \forall N \geq 1,$$

and there exists an uniform upper bound $\tilde{M}$ for $M_t$, i.e., $\tilde{M} \geq M_t, \forall t \geq 1$. In addition, if $\Delta = D_P / (L_f R_0)$, the convergence bound can be simplified to

$$f(\bar{x}^N) - f(x^*) \leq \frac{4L_f^2R_0^2}{N(N+1)}, \forall N \geq 1.$$  

A few remarks are in place regarding the above result. First, the stepsizes in the outer DRAO loop are exactly the same as that of Theorem 3.1. Since each phase requires only two rounds of communication, the DRAO-S method has a communication complexity of $O(\sqrt{L_f R_0} / \sqrt{\epsilon})$. Second, $M_t$ defined in (4.4) is the smallest upper bound of $M_t$ needed to make the inner iteration limit $S_t$ an integer. The factor $\Delta$ in (4.4) plays the role of a conversion factor between the $P$-projection complexity and the communication complexity ($\Pi$ projection complexity). A communication complexity of the order $O(1 / \sqrt{\epsilon})$ can be maintained for any $\Delta > 0$ and the specific choice in (4.4) is needed only for optimal constant dependence. Third, since the number of phases needed to find an $\epsilon$-optimal solution is bounded by $N_\epsilon := 2L_f R_0 / \sqrt{\epsilon}$, the total number of $P$-projections is given by

$$\sum_{t=1}^{N} \lfloor t\Delta M_t \rfloor \leq \Delta \tilde{M} N_\epsilon^2 + N_\epsilon = O(D_P \tilde{M} R_0 / \epsilon).$$

Fourth, both the inner SPS loop stepsizes, $\beta_s^t$, $\gamma_s^t$ and $\delta_s^t$, and the inner iteration limit $S_t$ adjust dynamically to the varying operator norm $M_t$ characterizing the difficulty of the saddle point problem (4.1) of each phase. Specifically, when the saddle point problem is easy, i.e., $M_t$ is small, $\gamma_s^t$, $\beta_s^t$, and $S_t$ become small so that a small number of inner iterations is performed, and vice versa. Thus, when most $M_t$’s are significantly smaller than the upper bound $\tilde{M}$, the total number of $P$-projections can be much smaller than $O(D_P \tilde{M} R_0 / \epsilon)$. Such a saving is possible because $S_t$ can compensate for the changing $\beta_s^t$ and $\gamma_s^t$ such that the effective proximal penalty parameter, $\omega_t\gamma_s^t$ and $\omega_t\beta_s^t$, remains constant across phases. In contrast, it is difficult for single loop primal dual type algorithms, such as the SD method (11), to adjust dynamically to the varying operator norm of $v^t$ in each iteration.

The following theorem presents the stepsize choice and the convergence result under the strongly convex setting.

**Theorem 4.2.** Let a smooth problem $f$ (c.f. (1.1)) with $\alpha > 0$ be given. Let $R_0$ and $D_P$ be defined in Subsection 1.1. Let the smoothness constant $L_f$ be defined in (3.5) such that $\kappa := L_f / \alpha$ denotes the condition...
number. If \( \{x^t\}_{t=1}^N \) are generated by the DRAO-S method \((4.2)\) with the following stepsize:

\[
\theta_t = \theta := \frac{\sqrt{8\kappa t - 1}}{\sqrt{8\kappa t + 1}}, \quad \omega_t = (\frac{1}{\theta})^{t-1}, \quad \tau_t = \tau := \frac{\sqrt{8\kappa t - 1}}{2}, \quad \eta_t = \eta := \frac{\alpha(\sqrt{8\kappa t + 1}-1)}{4},
\]

\[
\beta_s = \alpha(s+1), \quad \gamma_s = \frac{2\bar{s}^2(s+1)}{\omega(s+2)}, \quad \delta_t = (s-1)/s,
\]

\[
q_t^s = s, \quad S_t = [(2\omega_1\Delta)^{1/2}M_t], \quad \Delta > 0,
\]

then the last iterate \( x^N \) converges geometrically:

\[
\|x^N - x^*\|^2 \leq \theta^N \left( (1 + 2\kappa)\|x^0 - x^*\|^2 + \frac{4\beta_s^5}{\alpha\eta}\right), \forall N \geq 1,
\]

and there exist an \( \bar{M} \) such that \( M_t \leq \bar{M} \forall t \geq 1 \). If \( \Delta = 2D_P^2/\eta L_f R_0^2 \) and \( \kappa \geq 1 \), the overall P-projection complexity is \( O(\frac{1}{4}MD_P \Delta + \sqrt{\kappa} \log(\frac{1}{\epsilon})) \).

Notice that the strong convexity modulus \( \alpha \) is split into two to accelerate both the outer DRAO loop and the inner SPS loop. For the outer DRAO loop, the proposed stepsize is the same as that of Theorem 3.2 if \( \alpha/2 \) is viewed as the strong convexity modulus, hence the DRAO-S method maintains the same order of communication complexity, i.e., \( O(\sqrt{\kappa} \log(1/\epsilon)) \). For the inner SPS loop, the stepsizes are similar to that of the accelerated primal dual method (see [16]) with an \( \alpha/2 \) strong convexity modulus. Moreover, similar to Theorem 4.1, the stepsize \( \gamma^t_s \) and the inner iteration limit \( S_t \) adjust dynamically to the varying operator norm \( M_t \) in each phase. It is also worth noting that the constant dependence of the \( P \)-projection complexity on \( \kappa \) in Theorem 4.2 is larger than the optimal by a factor of \( \kappa^{1/4} \). This complexity can be further improved to \( O(\overline{MD_P}/\alpha \sqrt{\kappa} + \sqrt{\kappa} \log(1/\epsilon)) \) if some stepsize choice similar to Theorem 6 of [42] is utilized.

Next, let us move on to the structured non-smooth problem. Since II is assumed be bounded, the following uniform upper bound, the strong convexity modulus \( \bar{M} \) (c.f. (4.3)) is useful for convergence analysis,

\[
\tilde{M} = \max_{\pi \in \Pi} \{ \|[A_1^T \pi_1; \ldots; A_m^T \pi_m]\|_2, U_* := \max_{\pi \in \Pi} \max_{\|y\|_2 \leq 1, \|\| \leq 1} \sum_{i=1}^m p_t(A_i^T \pi_t, y) \}.
\]

Specifically, the stepsize choices and the convergence properties of the DRAO-S method, applied to both non-strongly and strongly convex settings, are presented in the next theorem.

**Theorem 4.3.** Let a structured non-smooth problem \( f \) \((1.1)\) be given. Let \( D_P, M_A, \) and \( M_t \) be defined in \((3.12)\) and \((4.9)\), and let \( D_P \) and \( R_0 \) be defined in Subsection I.1.

a) If \( \alpha = 0 \) and the stepsizes are given by

\[
\omega_t = 1, \quad \theta_t = 1, \quad \eta_t = M_A D_P / 2R_0, \quad \tau_t = M_A R_0 / 2D_P, \quad \alpha > 0, \quad S_t = [M_A \Delta], \quad \Delta > 0, \quad \tilde{M}_t = \frac{S_t}{\Delta}, \quad \beta_s = \beta := \frac{D_P M_t}{R_0}, \quad \gamma_s = \gamma = \frac{R_0 \tilde{M}_t}{D_P},
\]

\[
q_t^s = 1, \quad \delta_t^s = \begin{cases} \tilde{M}_t / \tilde{M}_{t-1} & \text{if } t \geq 2, \\ 1 & \text{if } t = 1 \end{cases}, \quad \text{and } \delta_t^s = 1 \forall s \geq 2,
\]

then the solution \( \tilde{x}^N \) returned by the DRAO-S method satisfies

\[
f(\tilde{x}^N) - f(x^*) \leq \frac{M_A D_P R_0}{N} + \frac{D_P R_0}{\Delta N}, \forall N \geq 1.
\]

In particular, if \( \Delta = \frac{D_P}{M_A D_P} \), the convergence bound can be simplified to

\[
f(\tilde{x}^N) - f(x^*) \leq \frac{2M_A D_P R_0}{N} \forall N \geq 1.
\]

b) If \( \alpha > 0 \) and the stepsizes are given by

\[
\omega_t = t, \quad \theta_t = (t-1)/t, \quad \eta_t = ta/6, \quad \tau_t = 6/ta,
\]

\[
\Delta > 0, \quad S_t = [\Delta \tilde{M}_t^2], \quad \Delta > 0, \quad \gamma_s = \gamma^t := 4\tilde{M}_t^2 / \alpha t,
\]

\[
\beta_s = \begin{cases} \frac{2}{t} & \text{if } s = 1 \\ \frac{2}{t} & \forall s \geq 2, \end{cases}, \quad \delta_t^s = \begin{cases} \frac{t-1}{t} & \text{if } s = 1 \\ \frac{1}{t} & \forall s \geq 2, \end{cases}, \quad q_t^s = 1,
\]

\[
\Delta > 0, \quad S_t = [(2\omega_1\Delta)^{1/2}M_t], \quad \Delta > 0, \quad \gamma_s = \gamma^t := 4\tilde{M}_t^2 / \alpha t,
\]

\[
\beta_s = \begin{cases} \frac{2}{t} & \text{if } s = 1 \\ \frac{2}{t} & \forall s \geq 2, \end{cases}, \quad \delta_t^s = \begin{cases} \frac{t-1}{t} & \text{if } s = 1 \\ \frac{1}{t} & \forall s \geq 2, \end{cases}, \quad q_t^s = 1,
\]
then the solution \(\bar{x} \) returned by the DRAO-S method satisfies
\[
f(\bar{x}^N) - f(x^*) \leq \frac{1}{N(N+1)}(\frac{1}{\alpha} R_0^2 + \frac{6M_2^2 D_0^2}{\alpha} + \frac{1M_2^2 D_0^3}{\alpha\Delta}).
\]
\[eq:sps-f-ns-qtr\]
In particular, if \(\Delta = D_0^2/(M_2^2 D_0^2 + \alpha^2 R_0^2/6)\), the convergence bound can be simplified to
\[
f(\bar{x}^N) - f(x^*) \leq \frac{1}{N(N+1)}(\alpha R_0^2 + \frac{10M_2^2 D_0^2}{\alpha}).
\]
\[eq:sps-f-ns-str1\]

A few remarks are in place. First, observe our choice of inner iteration limit \(S_t\) adjust dynamically to the operator norm \(M_t\) for the non-strongly convex case \(1.10\), but not for the strongly convex case \(4.13\). This shortcoming is an artifact of the order of prox updates in Algorithm 2. If a \(p\)-prox update, utilizing a \(y\)-momentum prediction, is performed before the \(y\)-prox update, \(S_t\) can be chosen to be \([\Delta M_2^2]\) to achieve the same effect. Moreover, since two communication rounds is required for each phase, the preceding result implies an \(\mathcal{O}(1/\epsilon)\) (\(\mathcal{O}(1/\sqrt{\epsilon})\)) communication complexity when \(\alpha = 0\) (resp. \(\alpha > 0\)). Since the inner iteration limit \(S_t\) is bounded, it also implies an \(\mathcal{O}(1/\epsilon)\) (resp. \(\mathcal{O}(1/\sqrt{\epsilon})\)) \(P\)-projection complexity. In particular, with the specific choices of \(\Delta\) shown above, the DRAO-S method can achieve the optimal constant dependence on problem parameters, that is, \(\mathcal{O}(R_0 M_2 D_0^2/\epsilon)\) communication and \(\mathcal{O}(M_2 D_0^2/\sqrt{\epsilon})\) \(P\)-projection complexities when \(\alpha = 0\), and \(\mathcal{O}(M_2 D_0^2/\sqrt{\epsilon})\) communication and \(\mathcal{O}(M_2 D_0^2/\sqrt{\epsilon})\) \(P\)-projection complexities when \(\alpha > 0\).

### 4.2. Convergence Analysis

Our goal in this subsection is to establish the convergence rates of the DRAO-S method stated in Theorem 1.1 and 1.3. First, we present a recursive bound to characterize the convergence property of each inner SPS loop under both the non-strongly convex (\(\alpha = 0\)) and the strongly convex (\(\alpha > 0\)) settings.

**Proposition 4.4.** Let \(M_t := \|v^t\|_{2, U^t}\) and \(M_t - 1 := \|v^{t-1}\|_{2, U^t}\). If the SPS stepsizes in Algorithm 2 satisfy:
\[
\delta_t = a_t \delta_t - 1, \quad \beta_t \gamma_t - 1 \geq \delta_t M_t^2, \forall s \geq 2,
\]
\[req:SPS_phase\]
\[
\gamma_t \geq \frac{1}{\delta_t + \gamma_t - 1} \geq \delta_t M_t^2, \forall s \geq 2,
\]
then the generated iterates, \((\{x^t, p^t\})\) of the following relation for all \(x \in X, \ p \in P\) and \(S \geq 1\):
\[
\{\sum_{s=1}^{S} q_s [\mathcal{L}(x^t; p, \pi^t) - \mathcal{L}(x^t; \tilde{p}^s, \pi^t)] + \delta_t q_t (y^t - x, \sum_{i=1}^{m} v_i^{t-1}(p_i^t - p_i^{t-1}) - q_s (y^t - x, \sum_{i=1}^{m} v_i^{t-1}(p_i^t - p_i^{t-1})) + \frac{q_s}{2} [\eta_t\|x^{t-1}\|_2^2 + \eta_t \|x^{t-1}\|_2^2 - \eta_t \|x^{t-1}\|_2^2]) \leq q_t \gamma_t U(p; p^0) - q_S \gamma_t U(p; p^S) + \frac{1}{2}\|p^S - p^{S-1}\|_2^2 + \frac{q_t \gamma_t M_t^2}{2\eta_t} \|p^0 - p^{S-1}\|_2^2 \leq q_t \beta_t \gamma_t \|y^t - x\|_2^2 - \|x^{t-1}\|_2^2.
\]
\[eq:SPS_phase_bound\]

**Proof.** Let an \(x \in X\) and a \(p \in P\) be given. First, consider the convergence of \(y^t\). Since \(u(y) + \eta_t \|y - x\|_2^2/2\) has a strong convexity modulus of \(\alpha + \eta_t\), the \(y\)-proximal update in Line 3 of Algorithm 2 leads a three point inequality (see Lemma 3.1 of [16]):
\[
\langle y^t - x, \hat{v}^t \rangle + u(y^t) - u(x) + \frac{1}{2}(\beta_t + \alpha + \eta_t) \|y^t - y^{t-1}\|_2^2 + \beta_t \|y^t - y^{t-1}\|_2^2 - \beta_t \|y^{t-1} - x\|_2^2
\]
\[eq:p2_0_x\]
\[
+ \frac{\eta_t}{2} (\|y^t - x\|_2^2 - \|x - x^{t-1}\|_2^2) \leq 0.
\]
Equivalently, we have
\[
\langle y^t - x, \hat{v}^t \rangle + u(y^t) - u(x) + \frac{1}{2}(\beta_t + \alpha + \eta_t) \|y^t - y^{t-1}\|_2^2 + \beta_t \|y^t - y^{t-1}\|_2^2 - \beta_t \|y^{t-1} - x\|_2^2
\]
\[eq:p2_0_x\]
\[
+ \frac{\eta_t}{2} (\|y^t - x\|_2^2 - \|x - x^{t-1}\|_2^2) \leq 0.
\]
In particular, the definition of $\bar{v}^*$ in Line 2 of Algorithm 2 implies

$$
\langle y^s - x, \bar{v}^* \rangle = (y^s - x, \sum_{i=1}^{m} p_i^s v_i^s) - (y^s - x, \sum_{i=1}^{m} (p_i^s - p_i^{-1}) v_i^s) + \delta_s (y^s - y^{s-1}, \sum_{i=1}^{m} (p_i^s - p_i^{-2}) v_i^s), \forall s \geq 2,
$$

$$
\langle y^1 - x, \bar{v}^1 \rangle = (y^1 - x, \sum_{i=1}^{m} p_1^1 v_1^1) - (y^1 - x, \sum_{i=1}^{m} (p_1^1 - p_i^1) v_i^1) + \delta_1 (y^1 - y^0, \sum_{i=1}^{m} (p_1^0 - p_i^{-1}) v_i^{-1}).
$$

So, substituting them into (4.18), summing up the resulting inequality with weight $q_s$, noting the step-sizes conditions in (4.16), and utilizing Young's inequality, we get

$$
\sum_{s=1}^{S} q_s \left( \mathcal{L}(y^s; p^s, \pi^s) - \mathcal{L}(x; p^s, \pi^s) + \frac{1}{2} \eta \| y^s - x^{t-1} \|^2 + (\eta + \alpha/2) \| y^s - x \|^2 - \eta \| x - x^{t-1} \|^2 \right) + \frac{q_s^2 \gamma_{t+1}}{2} \| p^s - p^{s-1} \|^2_U
\leq \sum_{s=2}^{\infty} \frac{q_{s-1} \omega_{s-1}}{2} \| p^{s-1} - p^{s-2} \|^2_U + \frac{q_s^2 \gamma_{t+1}}{2} \| p^0 - p^{-1} \|^2_U
- \frac{1}{2} \gamma_s (\beta_S + \alpha/2) \| y^s - x \|^2 - q_1 \beta_1 \| y^0 - x \|^2.
$$

Next, consider the convergence of $p^*$. The $p$-proximal update in Line 4 of Algorithm 2 implies

$$
\mathcal{L}(y^s; p^*, \pi^s) - \mathcal{L}(y^s; p^s, \pi^s) + \gamma_s U(p; p^s) + U(p^s; p^{s-1}) - U(p; p^{s-1}) \leq 0.
$$

Observing the strong convexity of $U$ with respect to $\| \cdot \|_U$ and the stepsize conditions in (4.16), the $q_s$ weighted sum satisfies

$$
\sum_{s=1}^{S} q_s \left( \mathcal{L}(y^s; p^s, \pi^s) - \mathcal{L}(y^s; p^s, \pi^s) + \frac{1}{2} \eta \| y^s - x^{t-1} \|^2 + (\eta + \alpha/2) \| y^s - x \|^2 - \eta \| x - x^{t-1} \|^2 \right)
+ q_s \delta_1 (y^0 - x, \sum_{i=1}^{m} v_i^{-1} (p_i^0 - p_i^{-1}) - q_s (y^s - x, \sum_{i=1}^{m} v_i^s (p_i^s - p_i^{s-1})))
\leq q_1 \gamma_1 U(p; p^0) - q_s \gamma_s U(p; p^s) + \frac{1}{2} \| p^s - p^{s-1} \|^2_U + \frac{q_s^2 \gamma_{t+1}}{2} \| p^0 - p^{-1} \|^2_U
- \frac{1}{2} \gamma_s (\beta_S + \alpha/2) \| y^s - x \|^2 - q_1 \beta_1 \| y^0 - x \|^2.
$$

Moreover, since $\mathcal{L}(y^s; p^s, \pi^s)$, $\| y^s - x^{t-1} \|^2$ and $\| y^s - x \|^2$ are convex with respect to $y^s$ and $\mathcal{L}(x; p^s, \pi^s)$ is linear with respect to $p^s$, the desired convergence bound in (4.17) can be derived using the Jensen’s inequality.

Notice that an $\omega_t$-weighted sum of the terms related to the outer DRAO loop, $x^t$ and $\bar{p}^t$, in (4.17) from the preceding proposition is the same as the $Q_x$ and $Q_p$ convergence bound in the proof of Proposition 3.4 (c.f. (3.22)). A convergence bound of $Q$ in the DRAO-S method can be deduced by plugging it into the proof of Proposition 3.4, i.e., the analysis for the $Q$ convergence in the DRAO method.

**Proposition 4.5.** Let $z^t := \{ x^t, \bar{p}^t, \pi^t \}$ be generated by the DRAO-S method with the outer DRAO loop stepsize satisfying (3.15), (3.16) and (3.17), and the inner SPS loop stepsize satisfying (4.16). Let $\tilde{\omega}^t := \omega_t / \sum_{s=1}^{S} q_s$ denote the effective summation weight for the inner SPS loops and let $M_t$ be defined in (4.3). Let the following inter-phase stepsize requirements for inner SPS loop hold for $t \geq 2$:

$$
\begin{align*}
\tilde{\omega}_1 \delta_1^2 M_{t-1}^2 &\leq \tilde{\omega}_1^{-1} \beta_1 q_{s_{t-1}}^{-1} q_{s_{t-1}}^{-1} \gamma_{t-1} \gamma_{t-1}, \quad M_{N}^2 \leq \gamma_{s_N}^N (\beta_{s_N}^N + \alpha/2), \\
\tilde{\omega}_t \delta_t^2 q_t^1 &\leq \tilde{\omega}_t^{-1} q_{s_{t-1}}^{-1} q_{s_{t-1}}^{-1} + \alpha/2, \quad \tilde{\omega}_t \delta_t^2 q_t^1 \leq \tilde{\omega}_t^{-1} q_{s_{t-1}}^{-1} \gamma_{t-1}, \gamma_{t-1} \gamma_{t-1}, \\
\tilde{\omega}_t \delta_t^2 q_t^1 &\leq \tilde{\omega}_t^{-1} q_{s_{t-1}}^{-1}.
\end{align*}
$$

*Except that $\alpha/2$, instead of $\alpha$, is regarded as the strong convexity modulus for the outer DRAO loop.*
Then the following $Q$-convergence bound holds for any reference point $z := (x^*, p, \pi)$ and for all $N \geq 1$

\[
\sum_{t=1}^{N} \omega_t Q(z^t, z) + \left[ \omega_N (\eta_N + \alpha/2) \right] \|x^N - x^*\|^2 / 2 
\leq \omega_1 \frac{1}{\tau_1} \sum_{i=1}^{m} p_i W_i^2 (\pi_i; \pi_0^*) + \left( \omega_1 \eta_1 + \hat{\omega} q_1 \beta_1 \right) \|x^0 - x^*\|^2 / 2 + \hat{\omega} q_1 \gamma_1 D_p^2 / 2. \tag{4.21}
\]

The above bound also holds if the last condition in (4.20) is replaced by \[
\gamma_{S_t}^{(\beta_{S_t} + \alpha/2)} \geq M^2, \quad \beta_1 = 0, \quad \text{and} \quad \delta_t^i = 0 \quad \forall t \geq 1. \tag{4.22}
\]

**Proof.** As pointed out before the statement of the proposition, dividing both sides of (4.17) by $\sum_{s=1}^{S_t} q_s$, taking its $\omega$-weighted sum, and noting the choice of initialization points in (4.2) and the telescope convergence bound induced by the stepsize requirements (4.20), we get a convergence bound of $Q_x$ and $Q_p$ given by

\[
\sum_{t=1}^{N} \omega_t \left[ Q(x^t; z) + Q_p(z^t; z) \right] + \sum_{t=1}^{N} \frac{\omega}{\sqrt{2}} \eta_t \|x^t - x^{t-1}\|^2 + \omega_N (\eta_N + \alpha/2) \|x^N - x^*\|^2 
\leq \hat{\omega} q_1 \gamma_1 U(p; p^0) + \beta_t \|y^0 - x^0\|^2 / 2 + \omega_1 \frac{1}{\tau} \|x^0 - x^*\|^2. \tag{4.23}
\]

The preceding bound is almost the same as its counterpart in Proposition 3.4, i.e., (4.22). Moreover, since the generation of $\pi^t$ in the outer DRAO loop of the DRAO-S method is also the same as that of the DRAO method, the $Q_x$ convergence bound in Proposition 3.4 (c.f. (3.19)) is also valid. The desired $Q$ convergence bound in (4.21) then follows from combining (3.19) with (4.23), and noting $U(p; p^0) \leq D_p^2 / 2$ and $y^0 = x^0$.

In addition, if the alternative stepsize requirement (4.22) is satisfied, (4.17) can be simplified further to

\[
\sum_{s=1}^{N} q_s (\mathcal{L}(x^t; p, \pi^t) - \mathcal{L}(x; p^0, \pi^0)) + \frac{1}{\sqrt{2}} \eta t \|x^t - x^{t-1}\|^2 + (\eta t + \alpha/2) \|x^t - x\|^2 - \eta t \|x^{t-1} - x\|^2 
\leq q_1 \gamma_1 U(p; p^0) - q_{S_t} \gamma_{S_t} U(p; p^0). \tag{4.24}
\]

Then a similar argument would lead to the $Q$ convergence bound in (4.21) as well.\hfill \blacksquare

The next convergence proofs of the DRAO-S method for the smooth problem, i.e., Theorem 4.1 and Theorem 4.2 are direct applications of Proposition 4.5.

**Proof of Theorem 4.1** It is easy to verify that the stepsize choice in (4.4) satisfies the requirements in Proposition 4.5, thus the following convergence bound is valid for any reference point $z := (x^*, p, \pi)$ and for all $N \geq 1$:

\[
\sum_{t=1}^{N} \omega_t Q(z^t, z) + L_f \|x^N - x^*\|^2 \leq L_f R_0^2 + D_P R_0 / \Delta \leq 2 L_f R_0^2. \tag{4.25}
\]

Let $\hat{p}^N = \nabla f(x^N)$ and $\hat{p}^N \in \arg\max_{p \in P} \sum_{i=1}^{m} p_i f_i(x^N)$ such that $f(x^N) = \mathcal{L}(x^N; \hat{p}^N, \hat{p}^N)$ (see Lemma 2.1), then the desired convergence result in (4.5) can be deduced by choosing the reference point to be $(x^*, \hat{p}^N, \hat{p}^N)$. Next we show the boundedness of $M_t$. Since $\pi^t = \nabla f(x^t)$ (c.f. (3.3)), $f$ is smooth and $x^t$ is a convex combination of $x^0$ and $\{x^t\}$, the boundedness of $\|\pi^t\|_{\mathcal{U}}$ is a consequence of the boundedness of $x^t$. Setting the reference point to the saddle point $(x^*, p^*, \pi^*)$ (c.f. Lemma 2.2), we get from (4.24)

\[
L_f \|x^N - x^*\|^2 \leq 2 L_f R_0^2 \quad \forall N \geq 2.
\]

So $x^t$’s are restricted to be a bounded ball around $x^*$. Due to $\theta_l \leq 1$, the extrapolated $x^t$’s are also restricted to a bounded ball around $x^*$. Thus, the boundedness of $\{M_t\}_{t=1}^{\infty}$ follows.\hfill \blacksquare

**Proof of Theorem 4.2** Again, we can verify that the stepsize choice in (4.7) satisfies the alternative requirements in Proposition 4.5 (c.f. (4.22)). Setting the reference point $z$ to the saddle point $(x^*, p^*, \pi^*)$ (c.f. Lemma 2.2), we get from (4.20) the desired geometric convergence of $x^N$ (4.8), i.e.,

\[
\|x^N - x^*\|^2 \leq \theta^N (1 + 2\kappa) \|x^0 - x^*\|^2 + \frac{4D_P^2}{\eta_0^2}, \quad \forall N \geq 1. \tag{4.25}
\]
Observe that the above convergence bound also implies the boundedness of \( \{x^t\} \). Thus the existence of an uniform bound for \( M_t \) follows from an argument similar to that of the proof of Theorem 4.1.

Next, we establish an upper bound on the total of inner iterations when \( \Delta := 2D_P^2/\eta L_f \|x^0 - x^*\|^2 \) and \( \kappa \) is large. The specific choice of \( \Delta \) allows us to simplify (4.25) further to

\[
\|x^N - x^*\|^2 \leq \theta^N (5\kappa R)^2.
\]

Let \( N_\epsilon \) denotes the least number of phases required to satisfy \( \|x^N - x^*\|^2 \leq \epsilon \). Clearly, \( N_\epsilon = O(\sqrt{\kappa} \log(1/\epsilon)) \).

Specifically, since \( \kappa \geq 1 \) implies \( 1/\theta \leq 2 \), we have \( (1/\theta)^N \leq 10\kappa R_0/\epsilon \).

As for the inner iteration number, a bound for \( S_t \) is provided by the stepsizes requirement (4.7). Since \( M_t \leq M \) and \( S_t \leq 1 + \sqrt{2}\omega_2M \), the total number is upper bounded by

\[
\sum_{t=1}^{N_\epsilon} S_t \leq N_\epsilon + \sum_{t=1}^{N_\epsilon} (\sqrt{\frac{1}{\theta}} + 1)\sqrt{\Delta M} = N_\epsilon + \frac{1}{\sqrt{1/\theta - 1}} \sqrt{\Delta M} \leq N_\epsilon + \frac{16\sqrt{L_fR_0}}{\alpha \sqrt{\epsilon}} \leq N_\epsilon + \frac{64\kappa^{1/4}MD_P}{\alpha \sqrt{\epsilon}},
\]

where the second last inequality follows from observing \( \sqrt{1 + 1} - 1 \geq 1/4 \) for \( l \leq 1 \) such that

\[
\frac{1}{\sqrt{1/\theta - 1}} \sqrt{\Delta M} \leq 2(\sqrt{8\kappa} + 1) \leq 4\kappa^{1/4}.
\]

Thus the number of inner iterations, and hence the \( P \) projection complexity, are upper bounded by \( O(\frac{\kappa^{1/4}MD_P}{\alpha \sqrt{\epsilon}} + \sqrt{\kappa} \log(\frac{1}{\epsilon})) \).

**Proof of Theorem 4.3**

The proof is similar to that of Theorem 3.3. Let us first consider the non-strongly convex case. Since \( M_{AH} \geq M_t, \forall t \), the stepsizes choice in (4.10) satisfies all the requirements in Proposition 4.4 (c.f. (4.20)). So substituting the stepsizes choice into (4.21), we obtain the following convergence bound of the \( Q \) gap function for any reference point \( z := (x^*, p, \pi) \):

\[
\sum_{t=1}^{N} Q(z^t; z) + \frac{\theta}{2} \|x^N - x\|^2 + \frac{3}{2} U(z; p^N) \leq \left( \frac{3}{2} + \frac{\kappa}{2\kappa} \right) \|x^0 - x\|^2 + \frac{3}{2} U(z; p^0) + \tau \sum_{t=1}^{m} p_t V_t(\pi; \pi^0).
\]

The desired function value convergence bound (4.11) follows immediately by selecting the reference point to be \((x^*, p^N, \pi^N)\), where \( \pi^N = \nabla f_t(\bar{x}^N) \) and \( p^N \in \arg\max_{p \in P} \sum_{t=1}^{m} p_t f_t(\bar{x}^N) \).

The convergence bound (4.14) for the strongly case also follows from substituting the stepsizes choice in (4.13) into (4.21).

**5. Lower Communication Complexities.** In this section, we establish theoretical lower bounds on distributed risk averse optimization to show the communication complexities of both DRAO and DRAO-S are not improvable. Specifically, we propose a distributed prox mapping (DPM) computing environment consisted of the following elements and construct lower bounds for all algorithms that can be implemented on the environment.

- **Local memory:** the server node has a finite local memory \( M_s \) and each worker node has a finite primal and a finite dual memory, \( M_t \) and \( M_t^s \), respectively. In the beginning, the local memories contain only the trivial vector 0, i.e.,

\[
M_{i,0} = M_{s,0} := \{0\}, \ M_{t,0}^s := \{0\} \ \forall i \in [m].
\]

In one communication round, these local memories can be updated by both local computation and server-worker communication:

\[
M_{s,t+1} := M_{s,t}^p \cup M_{s,t}^{\text{comm}}, \ M_{i,t+1} := M_{i,t}^p \cup M_{i,t}^{\text{comm}}, \ M_{t,0}^s := M_{t,0}^{\text{CP}}, \ \forall i \in [m],
\]

where \( M_{s,t}^p \) and \( M_{t,0}^{\text{CP}} \) represent iterates found via local computation, and \( M_{s,t}^{\text{comm}} \) and \( M_{i,t}^{\text{comm}} \) denote the vector(s) communicated to the server and the \( i \)-th worker node.
• **Server-worker communication:** in one communication round, each worker can send one vector from its local primal memory and receive one vector from the server’s memory:

\[ \mathcal{M}_{i,t}^{\text{comm}} \in \text{span}(\mathcal{M}_{s,t-1}), \quad \mathcal{M}_{s,t}^{\text{comm}} := \{ y_i \in \text{span}(\mathcal{M}_{i,t-1}), i \in [m] \}. \]

• **Local computations:** between communication rounds, each worker can query its dual prox mapping oracle and the \( A_i \)-multiplication oracle for \( L \geq 0 \) times.

\[ \mathcal{M}_{i,t}^{cp,0} := \mathcal{M}_{i,t}^{cp,L}, \quad \mathcal{M}_{i,t}^{\pi,cp} := \mathcal{M}_{i,t}^{cp,0} \quad \text{where} \quad \mathcal{M}_{i,t}^{cp,0} := \mathcal{M}_{i,t-1}, \mathcal{M}_{i,t}^{\pi,0} := \mathcal{M}_{i,t-1}. \]

For \( l = 1, 2, 3, \ldots, L \):

\[ \mathcal{M}_{i,t}^{cp,l} := \mathcal{M}_{i,t}^{cp,l-1} \cup \{ A_i \pi_i, A_i \bar{x} \}, \quad \mathcal{M}_{i,t}^{\pi,l} := \mathcal{M}_{i,t}^{\pi,l-1} \cup \{ \pi_i, A_i \bar{x} \}, \quad \bar{x} \in \text{span}(\mathcal{M}_{i,t}^{cp,l-1}), \]

\[ \pi_i \in \text{span}(\mathcal{M}_{i,t}^{\pi,l-1}), \quad \pi_i^{**} \in \arg \max \langle A_i \bar{x}, \pi_i \rangle - f_i^{*}(\pi_i) - \frac{\tau}{2} \| \pi_i - \pi_i^{**} \|^2, \tau \geq 0. \]

The server node can query its \( u(x) \) prox mapping oracle for \( L \geq 0 \) times.

\[ \mathcal{M}_{s,t}^{cp} := \mathcal{M}_{s,t}^{cp,L} \quad \text{where} \quad \mathcal{M}_{s,t}^{cp,0} := \mathcal{M}_{s,t-1}. \]

For \( l = 1, 2, 3, \ldots, L \):

\[ \mathcal{M}_{s,t}^{cp,l} := \mathcal{M}_{s,t}^{cp,l-1} \cup \{ x_i \}, \quad \text{where} \quad x_i := \arg \min_{x \in X} u(x) + \frac{\tau}{2} \| x - \bar{x} \|^2, \bar{x} \in \text{span}(\mathcal{M}_{s,t}^{cp,l-1}). \]

• **Output solution:** the output solution \( x^t \) comes from local primal memories,

\[ x^t \in \text{span}((U_{i \in [m]} \mathcal{M}_{i,t}) \cup \mathcal{M}_{s,t}), t \geq 1. \]

Essentially, the only hard requirement of the DPM environment is that only one vector can be sent and received by each worker during one communication round. The computations supported by the DPM environment are quite strong in several aspects. First, it allows gradient evaluation of \( f_i \) since it is equivalent to the \( \pi_i \)-prox mapping (c.f. (5.1)) with \( \tau = 0 \), i.e.,

\[ \bar{x}_i = \nabla f_i(\bar{x}) \Leftrightarrow \bar{x}_i \in \arg \max_{\pi_i \in \Pi_i} \langle \pi_i, \bar{x} \rangle - f_i^{*}(\pi_i). \]

Second, it allows a possibly large number of local computation steps to be performed between communications. This assumption of generous computing resource at each node helps us to focus on the communication bottleneck. Third, it allows the freedom to make an arbitrary selection from the span of the local memory for communication, computation, and outputting solutions. For example, it might appear that the DRAO method is not implementable because Line 5 of Algorithm 1 requires an \( A_i \)-multiplication oracle.

However, if we let \( (x^t, \bar{p}^t) \) be an optimal pair of saddle point solutions in the \( \langle x, p \rangle \)-prox mapping step (c.f. (4.1)), the output \( x^t \) can be written alternatively as

\[ x^t \leftarrow \arg \max_{x \in X} \eta_i \| x - \bar{x} \|^2 / 2 + u(x), \]

where \( \bar{x} := x^{t-1} - \sum_{i=1}^{m} p_i^t v_i^t / \eta_i \) and \( x \in \text{span}(\mathcal{M}_{s,t}) \). Moreover, the computation and communication of \( f_i^{*}(\pi_i^t) \)'s are unnecessary because they are only used for generating \( \bar{p}^t \). Since all other steps are directly supported, the DRAO method can be implemented on the DPM environment. Indeed, our setup implies that the desired \( p \) can be obtained from any oracle when selecting \( x \) (from the span of local memory of the server). This renders all communication and computation related to \( p \) unnecessary. So the DRAO-S method, and, more generally, any distributed algorithm consisting of the \( x \)-prox mapping, the \( \pi \)-prox mapping, and some \( p \) update can be implemented on the DPM environment. For simplicity, we will call any algorithm implemented in the DPM environment a DPM algorithm for the rest of this section.

Now we present some hard instances, inspired by [23, 27, 33], for all DPM algorithms. We first describe a network topology and a general result which will be used in all our constructions. As shown in Figure 5.1, the problem has only two workers, node 1 and node 2. Let \( K_i \) denote the subspace with non-zero entries only in the first \( i \) coordinates, \( K_i := \{ x \in \mathbb{R}^n : x_j = 0 \forall j > i \} \). We will construct \( f_1 \) and \( f_2 \) such that the
iterate $x^t$ generated in $t$ communication rounds will be restricted to a certain $K_i$. Towards that end, we call a hard problem odd-even preserving if the memories generated by any DPM algorithm satisfies

$$\mathcal{M}_{1,0} \cup \mathcal{M}_{2,0} \cup \mathcal{M}_{s,0} \subset K_2,$$

$$\mathcal{M}_{1,t−1} \subset K_i \iff \begin{cases} 
\mathcal{M}_{1,t}^{cp} \subset K_i & i \geq 2 \text{ even } \\
\mathcal{M}_{1,t}^{ip} \subset K_{i+1} & i \geq 2 \text{ odd }
\end{cases},$$

$$\mathcal{M}_{s,t−1} \subset K_i \iff \begin{cases} 
\mathcal{M}_{s,t}^{cp} \subset K_i & i \geq 2 \text{ even } \\
\mathcal{M}_{s,t}^{ip} \subset K_{i+1} & i \geq 2 \text{ odd }
\end{cases}.$$ (5.2)

This property stipulates that the progresses on the reachable subspace $K_i$ are possible only on node 1 or 2 depending on if $i$ is odd or even, so that a large number of communication rounds between node 1 and 2 are necessary for a non-trivial solution. Importantly, the next lemma characterizes the progress of a DPM algorithm when applied to this kind of hard problem.

**Lemma 5.1.** If the odd-even preserving property (5.2) holds, the output solution $x^t$ generated by a DPM algorithm after $t$ communication rounds satisfy $x^t \subset K_{\lfloor t/2 \rfloor + 2}$.

**Proof.** Let $\mathcal{M}_t := \text{span}(\mathcal{M}_{1,t} \cup \mathcal{M}_{2,t} \cup \mathcal{M}_{s,t})$, and let $t(i) := \min \{t \geq 0 : \exists y \in M_i, j \geq i \text{ s.t. } y_j \neq 0\}$ denote the first time a vector with a non-zero $j$ is obtained after one communication round. Next, the odd-even preserving property again implies $\mathcal{M}_{2,t(i)} \subset K_i$ and $\mathcal{M}_{s,t(i)} \subset K_{i+1}$, and $\mathcal{M}_{2,t(i)} \subset K_i$ and $\mathcal{M}_{s,t(i)} \subset K_{i+1}$ after one communication round. Therefore, we have $t(i+1) > t(i) + 1$, i.e., $t(i+1) \geq \min \{t(2) \geq 0\}$, the largest non-zero index in $\mathcal{M}_t$ satisfies $t \geq t(i) \geq t(2) + 2i - 4 \geq 2i - 4$, thus $i \leq \lfloor t/2 \rfloor + 2$.

We are now ready to provide lower bounds under different problem settings. The next two results establish tight lower bounds for the smooth problem with a non-strongly convex $u(x)$ and a strongly convex $u(x)$, respectively.

**Theorem 5.2.** Let $L_f > 0$, $R_0 \geq 1$ and $\epsilon > 0$ be given. For a sufficiently large problem dimension, i.e., $n \geq 2\sqrt{L_f R_0/\epsilon}$, there exists a smooth hard problem of form (1.1) with an aggregate smoothness constant $L_f$ (c.f. (3.5)), $\| x^0 - x^* \| \leq R_0$ such that any DPM algorithm takes at least $\Omega(\sqrt{L_f R_0/\epsilon})$ communication rounds to find an $\epsilon$-optimal solution.

**Proof.** Consider the following hard problem parameterized by $\beta \geq 0, \alpha \geq 0$ and $k \geq 4$,

$$f(x) := \max_{p \in \Delta_2^k} p_1 f_1(x) + p_2 f_2(x) + u(x) \text{ with } X = \mathbb{R}^{2k+1}, u(x) = 0,$$

$$f_1(x) := \beta [2 \sum_{i=1}^k (x(2i-1) - x_{2i})^2 + x_1^2 + x_{2k+1}^2 - 2\gamma x_1],$$

$$f_2(x) := \beta [2 \sum_{i=1}^k (x(2i - 2i+1) - x_{2i})^2 + x_1^2 + x_{2k+1}^2 - 2\gamma x_1].$$

Its aggregate smoothness constant $\bar{L}_f$ (c.f. (3.5)) satisfies $\bar{L}_f \leq 6\beta$, and its optimal solution $(x^*, p^*)$ satisfies

$$p^* = \left[ \frac{1}{2}, \frac{1}{2} \right], \quad x^*_1 = \gamma (1 - \frac{1}{2k+2}) \forall i \leq 2k + 1, \text{ s.t. } \| x^0 - x^* \| \leq \sqrt{\gamma k + 1} \text{ and } f_* = -\beta \gamma^2 [1 - \frac{1}{2k+2}].$$

Their optimality can be verified with the first order conditions:

$$0 = \nabla f_1(x^*) + \frac{1}{2} f_2(x^*)$$

and $[1/2, 1/2] \in \arg \max_{p \in \Delta_2^k} p_1 f_1(x^*) + p_2 f_2(x^*)$.

The even-odd preserving property also holds for (5.3). To see this, consider the worker node $f_1$. Let an even $i \geq 2$ be given and assume $\mathcal{M}_{1,t−1} \subset K_i$, i.e., $\mathcal{M}_{1,t}^{ip} \subset K_i$. Because $A_i = I$, the update rule in (5.1)
imply that $\mathcal{M}_{1,t}^{\tau,0} \subset \mathcal{K}_i$. We show $\mathcal{M}_{1,t}^{\tau,L} \cup \mathcal{M}_{1,t}^{\tau,L} \subset \mathcal{K}_i$ for all $l \geq 0$ by induction. Clearly, the statement holds for $l = 0$. If $\mathcal{M}_{1,t}^{\tau,L-1} \cup \mathcal{M}_{1,t}^{\tau,L-1} \subset \mathcal{K}_i$, $\bar{x}$ and $\bar{\pi}_i$ chosen in (5.1) must be in $\mathcal{K}_i$. As for the $\sigma_1$-prox mapping, if $\sigma = 0, \pi_1^{t+1} := \nabla f_1(\bar{x}) \in \mathcal{K}_i$. If $\sigma > 0$, Lemma 8.1 in the appendix allows us to write $\pi_1^{t+1}$ as

$$\pi_1^{t+1} = \pi_1 + \frac{1}{\tau}(\bar{x} - y),$$

where $y \leftarrow \arg \min_y f_1(y) + \frac{1}{2\tau} \|\bar{x} + \tau \bar{\pi}_1 - y\|^2$.

In particular, $y \in \mathcal{K}_i$ because

$$y = \arg \min_x \beta (2 \sum_{j=1}^{l+1} (x_{2j-1} - x_{2j})^2 + x_1^2 - 2\gamma x_1) + \frac{1}{2\tau} \sum_{j=1}^{l+1} \|\bar{x}_j + \tau \bar{\pi}_1 x_j - x_j\|^2 + \beta x_{2k+1} + \frac{1}{2\tau} \sum_{j=i+1} x_j^2.$$

So $\pi_1^{t+1} \in \mathcal{K}_i$ also holds for $\sigma > 0$. Thus the principle of induction implies that $\mathcal{M}_{1,t}^{\tau,L} \cup \mathcal{M}_{1,t}^{\tau,L} \subset \mathcal{K}_i$, $\forall L \geq 0$, i.e., $\mathcal{M}_{1,t}^{\tau,L} \subset \mathcal{K}_i$. In addition, when $i \geq 2$ odd and $\mathcal{M}_{1,t-1} \subset \mathcal{K}_i$, we have $\mathcal{M}_{1,t-1} \subset \mathcal{K}_{i+1}$. Since $i + 1$ is even, the preceding result implies that $\mathcal{M}_{1,t}^{\tau,L} \subset \mathcal{K}_{i+1}$. A similar result can also be derived for the worker $2$ for both even and odd $i \geq 2$. Therefore, problem (5.3) satisfies the even-odd preserving property.

Applying Lemma 5.1, the output solution $x^k$ from any DPM algorithm in $k$ communication rounds must satisfy $x^k \in \mathcal{K}_k$. In particular, let $f := (f_1 + f_2)/2$ denote a lower bound for $f$. Then $f(x^k) \geq \min_{x \in \mathcal{K}_k} f(x) \geq \min f(x) = -\beta^2 \gamma^2 (1 - \frac{1}{\tau^2}).$

Now we set the parameters in (5.3) to obtain the desired lower bound. If $\epsilon \geq L f R_0^2 / 256$, $\Omega(\sqrt{L f R_0 / \epsilon}) = \Omega(1)$, so the lower bound clearly hold. Otherwise, we set $\beta := L f / 6$, $\gamma := R_0 / \sqrt{k + 1}$ and $k := \lfloor \sqrt{L f R_0 / 4 \sqrt{\epsilon}} \rfloor$ such that (5.3) is $L_f$-smooth (c.f. (3.5)) with $\|x^0 - x^k\| \leq R_0$ and $k \geq 4$. A solution $x^k$ generated by any DPM algorithm in $k$ communication rounds satisfy $f(x^k) - f^* \geq \frac{\gamma^2 \beta}{2k+2} \geq \epsilon$. Thus they imply the desired $\Omega(\sqrt{L f R_0 / \epsilon})$ lower communication complexity bound when the problem dimension is $2\sqrt{L f R_0 / 4 \epsilon} + 1$.

We remark here that the above risk-averse lower bound is the same as the risk-neutral lower bound of $\Omega(\sqrt{L f R_0 / \epsilon})$, developed in [33], if $P$ is singleton set of the empirical distribution, $P = \{\bar{p} := (1/m, ..., 1/m)\}$, and $L_{f,\bar{p}}$ denotes the aggregate smoothness constant over $\bar{p}$. But, other than the intuition that the risk-averse problem should be harder than the risk-neutral problem, the later bound offers little guarantee. In fact, our risk-averse lower bound can be larger than the risk-neutral lower bound because the aggregate smoothness constant $L_f$ (c.f. (3.3)) defined over a non-trivial $P$ can be significantly larger than $L_{f,\bar{p}}$. For example, consider an expanded version of (5.3) constructed by adding $(m - 2)$ additional workers with constant local cost functions, $f_i(x) \equiv C$ for some $C < f^*$, and by setting $P$ to the $m$-dimensional simplex $\Delta_+^m$. The same argument as above will lead to the same lower bound of $\Omega(\sqrt{L f R_0 / \epsilon})$ for the expanded problem. However, because the smoothness constants of $\{f_i\}_{i=3}^m$ are zero, we have $L_{f,\bar{p}} \leq 2L_{f} / m << L_{f}$.

**Theorem 5.3.** Let $L_f > 8\alpha > 0$ and $\epsilon > 0$ be given. There exists an infinite-dimensional smooth problem of form (1.1) with an aggregate smoothness constant $L_f$ (c.f. (3.5)) and a strong convexity modulus $\alpha$ such that any DPM algorithm requires at least $\Omega(\sqrt{L_f / \alpha \log(1/\epsilon)})$ communication rounds to find an $\epsilon$-close solution, i.e., $x$ such that $\|x - x^*\|^2 \leq \epsilon$.

**Proof.** Again we prove the result by construction. Consider the following infinite dimensional problem

---

#We ignore the problem parameter $R_0$ inside the log.
parameterized by $\beta > 2\alpha$

$$f(x) := \max_{p \in \Delta^+_K}p_1f_1(x) + p_2f_2(x) + u(x) \text{ where } X := \mathbb{R}^\infty$$

$$u(x) := \frac{\beta}{2} \|x\|^2, f_1(x) := \frac{\beta - \alpha}{4}[x^\top A_1 x - 2x_1], f_2(x) := \frac{\beta - \alpha}{4}[x^\top A_2 x - 2x_1],$$

$$A_1 := \begin{bmatrix} \frac{\beta - \alpha}{4+\beta} & -1 & 1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \\ \vdots & \vdots & \vdots \end{bmatrix}, A_2 := \begin{bmatrix} \frac{\beta - \alpha}{4+\beta} & -1 & 1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \\ \vdots & \vdots & \vdots \end{bmatrix}.$$  \hspace{1cm} smst_l:prob (5.4)

Clearly, the aggregate smoothness constant (c.f. (3.5)) of the problem is bounded by $\beta - \alpha$ and its strong convexity modulus is $\alpha$. The optimal solutions are given by $p^* = (1/2, 1/2)$ and $x^*$, with $x^*_i = (1 - \sqrt{\gamma})^i \forall i \geq 1$, since they satisfy the first order optimality conditions:

$$\nabla (\frac{1}{2} f_1 + \frac{1}{2} f_2 + u)(x^*) = 0 \text{ and } (\frac{1}{2}, \frac{1}{2}) \in \arg \max_{p \in \Delta^+_K} p_1 f_1(x^*) + p_2 f_2(x^*).$$

Moreover, similar to Theorem 5.2, the alternating block diagonal structure of $A_1$ and $A_2$ implies the even-odd preserving property, so $x^k$ generated by any DPM algorithm in $k \geq 4$ communications rounds satisfy $x^k \in K_k$, i.e.,

$$\|x^k - x^*\|^2 \geq \sum_{i=k+1}^{\infty}(x^*_i)^2 = \left(1 - \frac{\sqrt{\gamma}}{1 + \sqrt{\gamma}}\right)^{2k}\left(1 - (1 - \frac{\sqrt{\gamma}}{1 + \sqrt{\gamma}})^2\right) = \left(1 - \frac{\sqrt{\gamma}}{1 + \sqrt{\gamma}}\right)^{2k} R_0^2$$

where $R_0^2 := \left(1 - \frac{\sqrt{\gamma}}{1 + \sqrt{\gamma}}\right)^2/(1 - (1 - \frac{\sqrt{\gamma}}{1 + \sqrt{gamma}})^2).$ Thus it takes at least $\Omega((\beta/2)(1/\log(1/\epsilon)))$ communication rounds to obtain an $x$ with $\|x - x^*\|^2 \leq \epsilon$.

Now we derive the lower bound for different $\epsilon$’s. If $\epsilon \geq (1 - 2 \sqrt{1/\kappa})^8 R_0^2$, then $\Omega(\sqrt{L_f/\alpha \log(1/\epsilon)}) = \Omega(1)$, so the desired lower bound holds. Otherwise, we can set $\beta := L_f + \alpha$ such that the hard problem (5.4) is $L_f$-smooth, and the desired lower communication bound of $\Omega(\sqrt{L_f/\alpha \log(1/\epsilon)})$ follows from (5.5).

We remark here that a finite dimensional hard problem can also be obtained by modifying (5.4) according to [18]. Next, we move on to consider the structured non-smooth problem.

**THEOREM 5.4.** Let $M_A > 0, D_{\Pi} > 0, R_0 \geq 1$ and $\epsilon > 0$ be given. When the problem dimension $n$ is sufficiently large (specified below), there exists a structured non-smooth problem $f$ of form (1.1) with $M_A \geq \max_{i \in [m]} \|A_i\|, D_{\Pi} \geq \max_{i \in [m]} \max_{x, z \in \Pi_i} \|\pi_i - \tilde{\pi}_i\|$ (c.f. (3.12)) and $R_0 \geq \|x^0 - x^*\|$ such that the following communication lower bounds hold.

a) When $u(x)$ is convex and $n > 2[D_{\Pi} M_A R_0/96]\epsilon$, any DPM algorithm requires $\Omega(M_A D_{\Pi} R_0/\epsilon)$ communication rounds to find an $\epsilon$-optimal solution.

b) When $u(x)$ is $\alpha > 0$ strongly convex and $n > 2[D_{\Pi} M_A/48 \sqrt{\alpha} \epsilon$, any DPM algorithm requires $\Omega(M_A D_{\Pi}/\sqrt{\alpha} \epsilon)$ communication rounds to find an $\epsilon$-optimal solution.

**Proof.** We consider the following hard problem parameterized by $k \geq 4$, $\alpha$, $\gamma_A$ and $\gamma_\pi$:

$$f(x) := \max_{p \in \Delta^+_K} p_1 f_1(x) + p_2 f_2(x) \text{ with } X = \mathbb{R}^{2k+1}, u(x) = \frac{\alpha}{2} \|x\|^2$$

$$f_1(x) := \gamma_A \gamma_\pi \left[2 \sum_{i=1}^k |x_{2i-1} - x_{2i}| - \left(\frac{\alpha}{2} + \frac{1}{\beta}\right) x_1\right],$$

$$f_2(x) := \gamma_A \gamma_\pi \left[2 \sum_{i=1}^k |x_{2i} - x_{2i+1}| - \left(\frac{\alpha}{2} + \frac{1}{\beta}\right) x_1\right].$$

In particular, the scenario cost functions $f_1$ and $f_2$ are specified in the structured maximization form (c.f. 23).
(1.1)), \( f_i(x) := \max_{x_i \in \Pi} \langle A_i x, \pi_i \rangle - f^*_i(\pi) \) with

\[
A_1 := \begin{bmatrix} -\left(\frac{1}{2} + \frac{1}{k} \right) & -1 & \ldots & -1 \\ 1 & -1 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & -1 & \ldots & -1 \end{bmatrix}, \quad \Pi_1 := \gamma_\pi \{1\} \times [-2, 2]^k \subset \mathbb{R}^{k+1}, \quad f^*_i(\pi_1) \equiv 0,
\]

\[
A_2 := \begin{bmatrix} -\left(\frac{1}{2} + \frac{1}{k} \right) & 1 & -1 & \ldots & -1 \\ 1 & -1 & \ldots & 1 \end{bmatrix}, \quad \Pi_2 := \gamma_\pi \{1\} \times [-2, 2]^k \subset \mathbb{R}^{k+1}, \quad f^*_2(\pi_2) \equiv 0.
\]

Clearly, \( \max_{x \in [2]} \| A_i \|_{2,2} \leq 2\gamma_A \) and \( \max_{x \in [m]} \max_{\pi_i, \pi_i \in \Pi} \| \pi_i - \pi_i \| \leq 5\sqrt{k}\gamma_\pi \). Furthermore, \( p^* := [1/2, 1/2] \) and \( x^* := \frac{7\gamma_A^2}{2k\alpha} \) form the optimal solution since they satisfy the first order optimality conditions given by:

\[
0 \in \partial \left( \frac{1}{2}f_1 + \frac{1}{2}f_2 + u(x) \right)(x^*) = 0 \quad \text{and} \quad \left( \frac{1}{2}, \frac{1}{2} \right) \in \arg \max_{p \in \Delta_2^+} p f_1(x^*) + p f_2(x^*)
\]

So \( f_* = f(x^*) = -\left( \frac{7\gamma_A^2}{2k\alpha} + \frac{7\gamma_A^2}{4k\alpha} \right), \quad \| x^* - x^0 \|^2 = \frac{7\gamma_A^2}{2k\alpha}(2k + 4) \leq \frac{7\gamma_A^2}{k\alpha^2} \).

Now, let us verify (5.6) satisfies the even-odd preserving property. Consider the worker \( f_1 \). Let an even \( i \geq 2 \) be given and assume \( \mathcal{M}_{1,t-1} \subset \mathcal{K}_i, \) i.e., \( \mathcal{M}^{\pi} \subset \mathcal{K}_i \). Let \( \mathcal{S}_j \) denote a (dual) subspace with non-zero entries only in the first \( j \) coordinates, \( \mathcal{S}_j := \{ \pi \in \mathbb{R}^{K+1} : \pi_l = 0 \quad \forall \ l > j \} \). Since the \( 2i-1 \)th and the \( 2i \)th coordinates of \( \gamma_A \) are non-zero if the \( i+1 \) th coordinate of \( \gamma_A \) is non-zero for any \( i \geq 2 \), the update rule (5.1) in the DPM environment implies that \( \mathcal{M}^{\pi,0} \subset \mathcal{S}_{i/2+1} \). We show \( \mathcal{M}^{\pi} \subset \mathcal{K}_i \) and \( \mathcal{M}^{\pi} \subset \mathcal{S}_{i/2+1} \) for all \( l \geq 0 \) by induction. Clearly the statement holds for \( l = 0 \). Moreover, if \( \mathcal{M}^{\pi} \subset \mathcal{K}_i \) and \( \mathcal{M}^{\pi} \subset \mathcal{S}_{i/2+1} \), \( A_1 \bar{x} \) and \( \bar{\pi} \) must be in \( \mathcal{S}_{i/2+1} \), so the dual proximal in (5.1) can be written as:

\[
\bar{\pi}^{\pi} := \max_{x \in \Pi, \sum_{j=1}^{i/2+1} (A_1 x) \bar{\pi}_{1,j} - \tau/2(\bar{\pi}_{1,j} - \bar{\pi}_{1,j})^2 - \tau/2 \sum_{j=i/2+2}^{i+k+1} (\bar{\pi}_{1,j})^2.
\]

This leads us to \( \bar{\pi}^{\pi} \subset \mathcal{S}_{i/2+1} \) and \( A_1 \bar{\pi}^{\pi} \subset \mathcal{K}_i \), i.e., \( \mathcal{M}^{\pi} \subset \mathcal{K}_i \). In addition, when \( i \geq 2 \) is odd and \( \mathcal{M}^{\pi} \subset \mathcal{S}_{i/2+1} \), we have \( \mathcal{M}^{\pi} \subset \mathcal{K}_i \). Since \( i + 1 \) is even, the preceding result implies that \( \mathcal{M}^{\pi} \subset \mathcal{K}_i \).

The property (5.2) for the worker \( f_2 \) for both even and odd \( i \)'s can also be deduced in a similar way. Therefore we have shown that the even-odd preserving property holds for the hard problem (5.6).

Next, applying Lemma 5.1 the solution \( x^k \) returned by any DPO algorithm in \( k \geq 4 \) communication rounds satisfy \( x^k \in \mathcal{K}_k \). We provide a lower bound of \( f \) on \( \mathcal{K}_k \). Let \( f := \frac{1}{2}(f_1 + f_2 + u(x)) \) denote a uniform lower bound for \( f \) given by:

\[
\tilde{f}(x) := \gamma_\pi \gamma_A \sum_{i=1}^{2k} |x_i - x_i^1| - (1 + \frac{1}{k})x_i^1 + \frac{1}{k} \| x \|^2.
\]

In order to find the minimum of \( \tilde{f} \) on \( \mathcal{K}_k \), observe that arranging \( \{ x_i \}_{i=1}^k \) in a decreasing order decreases \( \tilde{f} \). Moreover, if \( x_k < 0 \), setting all negative coordinates to zero decreases \( \tilde{f} \), so we can focus on \( x_1 \geq x_2 \ldots \geq x_k \geq x_{k+1} = \ldots = x_{2k+1} = 0 \).

\[
\min_{x \in \mathcal{K}_k} \tilde{f}(x) = \min_{x_k \in \mathcal{K}_k} -\frac{2\gamma_A}{k} x_k + \frac{\alpha}{k} x_k^2 \geq -\frac{2\gamma_A}{k}\frac{x_k^2}{k\alpha}.
\]

Thus, \( f(x^k) - f_* \geq \min_{x \in \mathcal{K}_k} \tilde{f}(x) - f_* \geq \frac{7\gamma_A^2}{4k\alpha} \).

Finally, we choose appropriate problem parameters to establish the lower bounds. If \( \epsilon \geq D_\Omega(M_A M_{H_1} R_0 / \epsilon), \quad \Omega(M_A M_{H_1} R_0 / \epsilon) = \Omega(1) \), so the lower bound in a) clearly holds. Otherwise, setting \( k := \lceil D_\Omega(M_A M_{H_1} R_0 / 96\epsilon) \rceil, \quad n := 2k + 1, \quad \gamma_\pi := \frac{D_\Omega}{5\sqrt{\epsilon}}, \quad \gamma_A := \frac{M_A}{2}, \) and \( \alpha := \gamma_\pi \gamma_A / R_0 \sqrt{\epsilon} \), the parameters of (5.6) satisfy \( \max_{x \in [2]} \| A_1 x \|_{2,2} \leq M_A, \quad \max_{x \in [m]} \max_{x_i, x_i \in \Pi} \| \pi_i - \pi_i \| \leq D_\Omega, \quad 4 \leq k, \) and \( \| x^0 - x^* \| \leq R_0 \). Since the minimum optimality gap attainable in \( k = \Omega(M_A M_{H_1} R_0 / \epsilon) \) communication rounds is lower bounded by \( \epsilon \), the result in a) follows.

Now consider \( u(x) \) being \( \alpha \)-strongly convex for a fixed \( \alpha > 0 \). If \( \epsilon \geq D_\Omega M_A^2 / 40000\alpha \), \( \Omega(M_A \sqrt{D_\Omega / \alpha}) = \Omega(1) \), so the lower bound in b) clearly holds. Otherwise, setting \( k := \lceil D_\Omega M_A / 48\sqrt{\alpha} \rceil, \quad n := \lceil D_\Omega M_A / 48\sqrt{\alpha} \rceil, \quad 2k + 1 \)
2k + 1, \( \gamma \in \Delta := D_{\Pi}/5\sqrt{k} \), and \( \gamma_A := M_A/2 \), the parameters of (5.6) satisfy \( \max_{i \in [m]} \| A_i \|_{2,2} \leq M_A \), \( \max_{i \in [m]} \max_{i, \bar{i} \in \Pi} \| \pi_i - \bar{\pi} \| / D_{\Pi} \leq 4 \leq k \), and \( \| x^0 - x^* \| \leq R_0 \). Since the minimum optimality gap attainable in \( k = \Omega(M_A D_{\Pi}/\sqrt{\alpha \epsilon}) \) communication rounds is lower bounded by \( \epsilon \), the result in b) follows.

6. Numerical Experiments. In this section, we present a few numerical experiments to verify the theoretical convergence properties of the proposed DRAO-S method.

6.1. Implementation Details. The numerical experiments are implemented in MATLAB 2021b and are tested on an Alienware Desktop with a 4.20 GHz Intel Core i7 processor and 16 GB of 2400MHz DDR4 memory. The stepsize of the DRAO-S method is chosen according to Theorem 4.1, 4.2 and 4.3. The implementation details of the proximal mappings are deferred to the Appendix. Parameter tuning is used to achieve better empirical performance. The DRAO-S method is first tested on a few trial stepsizes, each running for only 20 phases. Next, the one achieving the lowest objective value during the trials is selected to achieve better empirical performance. The DRAO-S method are chosen according to Theorem 4.1, 4.2 and 4.3. The \( \| x^0 - x^* \| \) communication rounds is lower bounded by \( \epsilon \), the result in b) follows.

### Table 1

| Parameter | Choices | Conservative Estimate |
|-----------|---------|-----------------------|
| \( L_f \) | \{ \( L_f \), 0.3\( L_f \) \} | \( L_f := \max_{i \in [m]} \| H_i^T H_i \| \) |
| \( M_t \) | \{ \( M_t \), 0.3\( M_t \) \} | \( M_t := \| \| \nabla f_1(x) \| , \ldots, \| \nabla f_m(x) \| \| \) |

So there are four sets of trial stepsizes. For the structured non-smooth two-stage stochastic program, the parameters \( M_{\Pi} \) and \( M_A \) used for the calculations in (4.10) and (4.13) are given by

| Parameter | Choices | Conservative Estimate |
|-----------|---------|-----------------------|
| \( M_A \) | \{ \( M_A \), 0.3\( M_A \), 0.1\( M_A \) \} | \( M_A := \max_{i \in [m]} \| T_i \| \) |
| \( M_{\Pi} \) | \{ \( M_{\Pi} \), 0.3\( M_{\Pi} \), 0.1\( M_{\Pi} \) \} | \( M_{\Pi} := \| [ T_1^e ; \ldots ; T_m^e ] \| \) |

So there are nine sets of trial stepsizes.

Fig. 6.1: Convergence of DRAO-S for a Randomly Generated Robust Linear Regression Problem

6.2. Risk Averse Linear Regression Problem. For the smooth case, the following risk-averse linear regression problem of the form (6.1) is considered:

\[
\text{CV@R}_\alpha(f_1(x), \ldots, f_m(x)) + \frac{\alpha}{2} \| x \|^2 \text{ with } f_i(x) := \frac{1}{2} \| H_i x - b_i \|^2, X := \mathbb{R}^n. \tag{6.1}
\]

Here \( f_i \) denotes the loss function associated with the \( i^{th} \) dataset. Such a problem is motivated by the need for a single model to work well in most cases, either for fairness or risk aversion. For example, the state education department might wish to build a model to help teachers to identify students who need extra help.
| #Scenarios | Opt. Gap | 10% Risk | 5% Risk | 1.25% Risk |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| Non-strongly Convex $\alpha = 0$ | | | | |
| 20 | 10% | 6 | 16 | 8 | 50 | 8 | 50 |
| | 1% | 31 | 529 | 38 | 1236 | 38 | 1236 |
| | 0.1% | 76 | 2858 | 85 | 6282 | 85 | 6282 |
| 50 | 10% | 3 | 4 | 6 | 16 | 7 | 40 |
| | 1% | 16 | 128 | 20 | 159 | 38 | 1192 |
| | 0.1% | 63 | 1935 | 67 | 2206 | 68 | 4245 |
| 200 | 10% | 3 | 4 | 3 | 4 | 5 | 14 |
| | 1% | 6 | 19 | 12 | 72 | 23 | 260 |
| | 0.1% | 30 | 516 | 51 | 1336 | 65 | 2113 |

| Strongly Convex Condition Number $\alpha = 10$ | | | | |
| 20 | 1=3 | 32 | 665 | 39 | 1254 | 39 | 1254 |
| | 1=4 | 43 | 1926 | 44 | 2139 | 44 | 2139 |
| | 1=5 | 48 | 2080 | 49 | 3603 | 49 | 3603 |
| 50 | 1=3 | 19 | 159 | 29 | 485 | 39 | 1306 |
| | 1=4 | 36 | 854 | 38 | 1132 | 44 | 2051 |
| | 1=5 | 41 | 1558 | 43 | 2061 | 49 | 3754 |
| 200 | 1=3 | 16 | 54 | 14 | 69 | 29 | 454 |
| | 1=4 | 28 | 205 | 32 | 437 | 42 | 1777 |
| | 1=5 | 40 | 660 | 44 | 1357 | 46 | 2934 |

Table 6.1: Communications Rounds and P-Projects Required by DRAO-S for Linear Regression under a CV@R Risk

$(H_i, b_i)$ could represent the data collected in the $i^{th}$ county and the CV@R risk measure could be used to ensure fairness among counties.

In our experiments, we set $n = 40$, and generate matrices $H_i \in \mathbb{R}^{40 \times 200}$, and $b_i \in \mathbb{R}^{40}$ randomly. We generate an estimate of $f_*$ by running the bundle level method 20 to an extremely high degree of accuracy. We record the average number of communication rounds and P-projects steps needed, over five randomly generated instances, to achieve the desired relative optimality gap, i.e., $(f(x') - f_*)/f_* \leq \epsilon$ under different settings. In particular, the DRAO-S method is tested on problems with different levels of risk and different numbers of computing nodes to understand how the communication and the P-projection complexities depend on $D_P$ and $m$ in practice. The results are presented in Table 6.1. As for the number of computing nodes $m$, both the number of P-projections and the number of communication rounds scale well with it. In fact, they seem to decrease slightly when $m$ increases. As for $D_P$, recall that a lower risk level corresponds to a larger ambiguity set $P$ and hence a larger radius $D_P$ (c.f. Subsection 1.1). Therefore, both the number of P-projections and the number of communication rounds increase with $D_P$, but the number of communication rounds seems to have a weaker dependence on it. Additionally, typical convergence curves of the DRAO-S method are plotted in Figure 6.1 and they seem to verify the theoretical convergence guarantees. When $\alpha = 0$, Table 6.1 and the convergence curve in Figure 6.1a illustrate a communication complexity and a P-projection complexity on the order of $\mathcal{O}(1/\sqrt{\epsilon})$ and $\mathcal{O}(1/\epsilon)$, respectively. When $\alpha > 0$, the convergence curves in Figure 6.1b and Table 6.1 illustrate a communication complexity and a P-projection complexity on the order of $\mathcal{O}(\log(1/\epsilon))$ and $\mathcal{O}(1/\sqrt{\epsilon})$, respectively. Thus, the DRAO-S method can find highly accurate solutions within a small number of communication rounds.

### 6.3. Risk Averse Two-Stage Stochastic Programming

For the structured non-smooth case, we compare the DRAO-S method with the SD method 40 using the same risk-averse two-stage stochastic linear programming problem from 40:

$$
\min_{x \in \mathbb{R}^n} \ c^T x + \text{CV@R}_\delta (g_1(x), \ldots, g_m(x)) + \frac{\alpha}{2} \|x\|^2,
\text{ s.t. } 0 \leq x_j \leq U \ \forall j \in [n],
\quad g_i(x) := \min_{y_i \in \mathbb{R}} \{y_i^T e_i, \ s.t. \ R y_i \geq d_i - T_i x\}. $$

The problem models the capacity expansion decision of an electricity company. Being the sole provider for electricity, the company has to meet all demand profiles $\{d_i\}$ using a combination of installed capacity, with an availability factor of $T_i$, and electricity purchased from outside the grid, at a unit cost of $e_i$. Being risk averse, the company intends to find a decision that keeps the total cost low for roughly $(1 - \delta)$ of all possible scenarios.
In our experiments, we set $n = 40$ and $l = 20$, generate $T_i \in \mathbb{R}_{20 \times 40}$, $c_i \in \mathbb{R}_{20}$, $d_i \in \mathbb{R}_{20}$ and $c \in \mathbb{R}_{40}$ randomly, and choose $R := I_{20, 20}$ to be the simple complete recourse matrix. We record down the average number of communication rounds required to achieve the desired relative optimality gaps for both methods in Table 6.2. Clearly, DRAO-S enjoys significant savings compared to the SD method. The number of communications rounds required by DRAO-S is also less sensitive to the risk level and $D_P$. Moreover, typical convergence curves are plotted in Figure 6.2a and 6.2b. They seem to verify the theoretical communication complexities of DRAO-S on the orders of $O(1/\epsilon)$ and $O(1/\sqrt{\epsilon})$, respectively, for the non-strongly convex and the strongly convex problems.

### 6.4. Risk Measure induced by the $\chi^2$ Ambiguity Set.

Next, we test these algorithms on a more complicated quadratically constrained set $P$. Given a radius parameter $r$, the modified $\chi^2$ probability uncertainty set respect to the empirical probability $[1/m, \ldots, 1/m]$ is given by

$$P_r = \{p \in \mathbb{R}_{m}^m : \sum_{i=1}^m p_i = 1, \|p - [1/m, \ldots, 1/m]\|^2 \leq r\}.$$ 

Inspired by the $\chi^2$ test, $P_r$ is useful in DRO [4]. We conduct our experiments with the induced risk-measure $\rho(g) = \max_{p \in P_r} \langle p, g \rangle$ on both the linear regression problem (6.1) and the two-stage stochastic program (6.2). The average number of communication rounds required to reach the desired sub-optimality for various levels of $r$ are recorded in Table 6.3 and 6.4. Since a larger $r$ implies a larger $P$, the results are consistent with our findings under the CV@R setting.

### 7. Conclusion.

This paper proposes the problem of distributed risk averse optimization. A conceptual DRAO method and a more practical DRAO-S method are proposed. Both of them are able to solve the
Optimality of their communication complexities are established with matching lower bounds. And preliminary numerical experiments seem to indicate promising empirical performance for DRAO-S.

In future work, we will attempt to extend our proposed methods to the more general cross-device federated learning setting \([11]\) where \(f_i\)'s are accessible only via a stochastic first order oracle and the communication network is unreliable. We will also attempt to study the extension to more complicated risk measures for which \(p\)-prox mappings are prohibitively expensive and only gradient evaluations are possible.

### References

1. A. Beck, *First-order methods in optimization*, vol. 25, SIAM, 2017.
2. D. Bertsimas and A. Thiele, *Robust and data-driven optimization: modern decision making under uncertainty*, in Models, methods, and applications for innovative decision making, INFORMS, 2006, pp. 95–122.
3. A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, Journal of mathematical imaging and vision, 40 (2011), pp. 120–145.
4. Y. Chen, H. Sun, and H. Xu, *Decomposition and discrete approximation methods for solving two-stage distributionally robust optimization problems*, Computational Optimization and Applications, 76 (2021), pp. 205–238.
5. E. CORDIS, *Machine learning ledger orchestration for drug discovery*, [https://featurecloud.eu/about/our-vision/](https://featurecloud.eu/about/our-vision/). Retrieved: July. 2022.

---

#### Table 6.3: Communications Rounds and \(P\)-Projections Required by DRAO-S for Linear Regression under a modified \(\chi^2\) Risk Measure

| #Scenarios | Opt. Gap | \(r = 0.05\) | \(r = 0.1\) | \(r = 0.2\) |
|-------------|----------|-------------|-------------|-------------|
|             | Non-strongly Convex \(\alpha = 0\) | #Comm | \(P\)-proj | #Comm | \(P\)-proj | #Comm | \(P\)-proj |
| 20          | 10%      | 3 4        | 3 4        | 3 4       |             |
|             | 1%       | 14 93     | 17 134     | 20 148    |             |
|             | 0.1%     | 28 274    | 40 944     | 73 2710   |             |
| 50          | 10%      | 3 4        | 3 4        | 3 4       |             |
|             | 1%       | 8 33      | 15 83      | 15 184    |             |
|             | 0.1%     | 21 216    | 30 524     | 69 2369   |             |
| 200         | 10%      | 3 4        | 3 4        | 3 4       |             |
|             | 1%       | 22 22     | 14 95      | 19 104    |             |
|             | 0.1%     | 21 207    | 38 472     | 79 2426   |             |

#### Table 6.4: Communication Rounds Required by Two-Stage Stochastic Program under a modified \(\chi^2\) Risk Measure

| #Scenarios | Opt. Gap | \(r = 0.05\) | \(r = 0.1\) | \(r = 0.2\) |
|-------------|----------|-------------|-------------|-------------|
|             | Non-strongly Convex \(\alpha = 0\) | SD | DRAO-S | SD | DRAO-S | SD | DRAO-S |
| 20          | 10%      | 88 41      | 125 62     | 190 54     |             |
|             | 1%       | 703 202    | 1066 330   | 2032 345   |             |
| 50          | 10%      | 246 46     | 319 57     | 388 75     |             |
|             | 1%       | 1543 313   | 2146 377   | 2838 409   |             |
| 200         | 10%      | 194 16     | 273 32     | 332 45     |             |
|             | 1%       | 1747 270   | 2818 315   | 3191 335   |             |

---

risk-averse problem with the same communication complexities as those for solving the risk-neutral problem. The optimality of their communication complexities are established with matching lower bounds. And preliminary numerical experiments seem to indicate promising empirical performance for DRAO-S.
8. Appendix. **Lemma 8.1.** Let \( f_i : \mathbb{R}^n \to \mathbb{R} \) be a proper convex closed function and \( f_i^* \) be its Fenchel conjugate, then the following are equivalent for all \( y \in X, \pi_i \in \mathbb{R}^n, \tau > 0 \):

\[
\begin{align*}
\pi_i^t & \leftarrow \text{arg max}_{\pi_i} (\bar{y}, \pi_i) - f_i^*(\pi_i) - \frac{\tau}{2} \| \pi_i - \bar{\pi}_i \|^2, \\
\pi_i^t & \leftarrow \bar{\pi}_i + \frac{1}{\tau} (y - \bar{y}), \text{ where } y \leftarrow \text{arg min}_{\pi_i} f_i(y) + \frac{1}{2\tau} \| y + \tau \bar{\pi}_i - y \|^2.
\end{align*}
\]

**Proof.** Let us fix an \( i \in [m] \) and let \( \bar{u} := \bar{y} + \tau \bar{\pi}_i \). Consider a Moreau envelop of \( f_i \) given by \( g(u) = (f_i \circ \frac{1}{2\tau} \| \cdot \|^2)(u) := \inf_y f_i(y) + \frac{1}{2\tau} \| u - y \|^2 \). Since \( f_i \) is convex, \( g \) is convex and smooth over \( \mathbb{R}^n \), thus \( \partial g(\bar{u}) \) is non-empty and unique.

Next, define \( \bar{g}(u) := f_i(y) + \frac{1}{2\tau} \| u - y \|^2 \) such that \( g(u) \leq \bar{g}(u) \). Since \( y := \inf_y f_i(y) + \frac{1}{2\tau} \| \bar{u} - y \|^2 \) in (8.2) implies \( g(\bar{u}) = \bar{g}(\bar{u}) \), the subgradient of \( g \) at \( \bar{u} \) must be a subgradient of \( \bar{g} \), a dominating function, at \( \bar{u} \), i.e.,

\[ \partial g(\bar{u}) \subset \partial \bar{g}(\bar{u}) = \{ \pi_i^t := \frac{1}{\tau} (\bar{u} - y) \} \]

Therefore \( \pi_i^t = \nabla g(\bar{u}) \). Using infinitesimal convolution identity (c.f. Theorem 4.16 in [1]) \( (g)^*(\pi_i) = (f_i \circ \frac{1}{2\tau} \| \cdot \|^2)^*(\pi_i) = f_i^*(\pi_i) + \frac{\tau}{2} \| \pi_i \|^2 \forall \pi_i \), the equivalence between maximization and sub-gradient evaluation, and the fact \( u := \bar{y} + \tau \bar{\pi}_i \), we get

\[ \pi_i^t \in \partial g(\bar{u}) = \partial (f_i \circ \frac{1}{2\tau} \| \cdot \|^2)(\bar{u}) \]

\[ \pi_i^t \in \partial g(\bar{u}) \]

8.1. Efficient Implementations for Proximal Mappings. Since \( X \) is either a box or \( \mathbb{R}^n \), the \( x \)-prox mappings are implemented with closed-form solutions. The \( \pi \)-prox mappings also admit closed-form solutions. For the linear regression problem in (6.1), the equivalent primal gradient computation amounts to a matrix-vector multiplication. For the two-stage stochastic program in (6.2), since the simple form solutions. For the linear regression problem in (6.1), the equivalent primal gradient computation.

\[ p^t = \text{arg max}_{p \geq 0} \langle p, g \rangle - \frac{1}{2} \| p - p^{t-1} \|^2 \iff p^t = \min_{\lambda \in \mathbb{R}} \text{arg max}_{p \geq 0} \langle p, g \rangle - \frac{1}{2} \| p - p^{t-1} \|^2 + \lambda \sum_{i=1}^m p_i - 1 \]

**Proof:**

For a fixed \( \lambda \), the inner solution \( p(\lambda) \) can be computed via a component-wise vector thresholding and the optimal \( \lambda^t \) is characterized by the root condition \( \sum_{i=1}^m p_i(\lambda^t) - 1 = 0 \). Since \( p(\lambda) \) is a monotonically non-decreasing function of \( \lambda \), an accurate approximation to \( \lambda^t \) and hence \( p^t \) can be found by a binary search on \( \lambda \). Next, when \( \rho \) is the risk measure induced by the \( \chi^2 \) ambiguity set, we can dualize the \( \chi^2 \) constraint to express the \( p^t \)-prox mapping equivalently as follows.

\[ p^t = \text{arg max}_{p \geq 0} \langle p, g \rangle - \frac{1}{2} \| p - p^{t-1} \|^2 \iff p^t = \min_{u \in \mathbb{R}} \text{arg max}_{p \geq 0} \langle p, g \rangle - \frac{1}{2} \| p - p^{t-1} \|^2 + u \| [1/m, \ldots, 1/m] \|^2 - r \]

**Proof:**

For a fixed \( u \), the inner solution \( p(u) \) above can be solved similarly to (8.3). A sufficient optimality condition for \( u^t \) is the KKT condition, i.e. either \( u^t = 0 \), or \( u^t > 0 \) and \( \| p(u^t) - [1/m, \ldots, 1/m] \|^2 - r = 0 \). So an accurate \( u^t \) and hence \( p^t \) can be found by a binary search on \( u \).