Variance-Reduced Min-Max Methods for Accelerated Distributionally Robust Machine Learning

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

September 2, 2024

Abstract

As machine learning systems are increasingly deployed in dynamic environments, there is a critical need for models that can maintain performance under distributional shifts. Traditional empirical risk minimization (ERM) approaches often falter when the training and test data distributions differ, making them vulnerable to adversarial attacks and other forms of distributional changes. To address this, distributionally robust supervised learning (DRSL) seeks to optimize models that are resilient to such shifts by framing training as a min-max problem between the model and an adversary. In this paper, we focus on DRSL with Wasserstein metrics, which offer a more flexible and theoretically sound approach to handling distributional shifts compared to traditional f-divergences. We propose novel stochastic gradient-based algorithms designed to efficiently solve large-scale Wasserstein DRSL problems. Specifically, we reformulate the problem as a smooth, convex-concave min-max optimization task and introduce two algorithms: *Stochastic Extragradient with Variance Reduction (SEVR)* and *Stochastic Proximal Point with Random Reshuffling (SPPRR)*. Both algorithms leverage the smoothness and finite-sum structure of the problem to achieve faster convergence rates than existing methods. Our theoretical analysis demonstrates the efficiency of these algorithms, and we provide complexity bounds that highlight their scalability. The results indicate that these methods are well-suited for practical applications where robustness to distributional shifts is essential.

Keywords: Distributionally Robust Learning; Wasserstein Distance; Stochastic Optimization; Min-Max Optimization; Variance Reduction

1 Introduction

The deployment of machine learning systems in real-world environments necessitates models that can adapt to or remain robust against distributional changes. Conventional supervised learning methods, particularly those based on Empirical Risk Minimization (ERM), typically assume that the training and test data are drawn from the same distribution. However, this assumption often fails in practice due to factors like selection bias, nonstationarity in the environment, or adversarial attacks. Such distributional shifts can significantly degrade the performance of models trained under the ERM paradigm.

To address these challenges, Distributionally Robust Supervised Learning (DRSL) has emerged as a powerful framework for developing models that are resilient to adversarial distributional shifts. DRSL approaches typically formulate the learning problem as a zero-sum game between the model and an adversary that perturbs the training data within a certain metric ball, such as the f-divergence or the Wasserstein distance. Among these, Wasserstein-based DRSL has gained prominence due to its ability to handle more general forms of perturbations, including those where the supports of the distributions do not overlap.

Despite the theoretical appeal of DRSL, its practical application is often hindered by the computational complexity of solving the associated min-max optimization problems. Existing approaches either rely on deterministic algorithms that scale poorly with data size or employ stochastic methods that do not fully exploit the problem's structure, leading to suboptimal convergence rates.

In this paper, we propose new stochastic algorithms designed to efficiently solve large-scale DRSL problems using Wasserstein metrics. Our approach avoids the non-smooth reformulations common in previous work by directly tackling the problem as a smooth, convex-concave min-max optimization task. We introduce two algorithms: *Stochastic Extragradient with Variance Reduction (SEVR)* and *Stochastic Proximal Point with Random Reshuffling (SPPRR)*. These algorithms leverage variance reduction techniques and random reshuffling to achieve faster convergence rates, making them more suitable for practical implementation on large datasets.

Backgrounds. With machine learning systems increasingly being deployed in real-world settings, there is an urgent need for machine learning approaches that can adapt to — or are robust to—changes in the environment. Despite this, the dominant paradigm for supervised learning [Hastie et al.(2009)] remains that of empirical risk minimization (ERM) [Vapnik(2013)], wherein a model is trained by minimizing a loss over a fixed set of training data. A key assumption underlying this approach is that the training data is from the same distribution as the test data — i.e., the distribution of the data does not change between training time and deployment.

Such assumptions are known to rarely hold in practice. Indeed, the distribution may change due to selection bias, nonstationarity in the environment [Quionero-Candela et al.(2009)], or even adversarial perturbations [Szegedy et al.(2013), Madry et al.(2018)], leaving machine learning models trained through ERM particularly susceptible to adversarial attacks [Szegedy et al.(2013), Carlini and Wagner(2017 or to degraded performance from distribution shifts.

Distributionally robust supervised learning (DRSL) seeks to address this issue by explicitly optimizing for solutions that are robust to adversarial distribution shifts. Note that DRSL is not a new terminology in the context of machine learning but has been frequently used in the existing works (see [Hu et al.(2018)] for an example). Common approaches to DRSL formulate training as a zero-sum game between the machine learning model and an adversary that perturbs the training data within a ball in either an f-divergence [Bagnell(2005), Ben-Tal et al.(2013), Namkoong and Duchi(2016), Namkoong and Duchi(2017), Hu et al.(2018)] or the Wasserstein distance [Gao et al.(2017), Sinha et al.(2018), Blanchet and Murthy(2019), Blanchet et al.(2019), Shafieezadeh-Abade For an overview of distributionally robust optimization and relevant applications, we refer the reader to a recent survey [Rahimian and Mehrotra(2019)].

Despite its appealing premise, DRSL suffers in practice from a relative dearth of algorithms for solving large constrained min-max optimization problems. Indeed, existing approaches to DRSL reformulate the problem as a non-smooth convex optimization problem [Wiesemann et al.(2014), Abadeh et al.(2015)], and then apply optimization algorithms which are deterministic or require solving complex sub-problems which quickly become computationally intractable as the number of data points increases [Li et al.(2019), Li et al.(2020), Liu et al.(2017), Lee and Mehrotra(2015), Luo and Mehrotra(2019)].

Recent years have seen an emerging interest in developing stochastic algorithms for solving large-scale DRSL. In particular, [Namkoong and Duchi(2016)] proposed stochastic algorithms for solving convex reformulations of DRSL with f-divergences. For Wasserstein DRSL, [Blanchet et al.(2018)] avoided the non-smooth reformulation and proved locally strong convexity of the dual problem under certain conditions, where standard stochastic gradient descent algorithms can be applied. In practice, Wasserstein distances have been found to be more favorable than f-divergences since they can be used to compare distributions with disjoint supports. Indeed since f divergences can only be used to compare distributions on the same support, the end result of DRSL with f-divergences is a classifier that it robust to reweighting of empirical distribution. Wasserstein DRSL, on the other hand, provides robustness to an adversary who perturbs the data in more general ways.

This paper focuses on providing stochastic gradient-based algorithms for DRSL with Wasserstein metrics (known as Wasserstein DRSL or WDRSL) which can handle large datasets common in supervised learning, while being provably faster than existing approaches. Key to our approach is that we refrain from reformulating the problem as a convex minimization problem, and instead reformulate it as a smooth, finite-sum, convex-concave min-max optimization problem which can be efficiently solved with specially constructed variance-reduced stochastic extragradient algorithms.

1.1 Contributions

We develop algorithms for solving WDRSL problems with generalized linear losses—a class of problems that includes WDRSL with logistic losses. The primary contributions of this paper are as follows: (i) We reformulate Wasserstein DRSL with generalized linear losses as a smooth, convex-concave min-max optimization problem, which allows for the application of efficient stochastic gradient-based algorithms. (ii) We introduce two novel algorithms, SEVR and SPPRR, that exploit the smoothness and finite-sum structure of the problem, leading to faster convergence rates compared to existing methods. (iii) We provide a thorough complexity analysis of both algorithms, demonstrating their theoretical advantages and practical efficiency.

In details, our contributions can be summarized as follows.

(i) We reformulate WDRSL with generalized linear losses as a constrained, smooth, convex-concave min-max optimization problem where the constraint sets for the maximization and minimization problems are the ℓ_∞-ball and the second-order cone, respectively. We then define the notion of optimality for the min-max formulation (cf. Definition 5), and show that it is consistent with the standard one (Definition 4).

(ii) We develop two simple stochastic algorithms for solving WDRSL, which we refer to as *Stochastic Extragradient with Variance Reduction* (SEVR) and *Stochastic Proximal Point with Random Reshuffling* (SPPRR). By a careful complexity analysis, we prove that these two algorithms are able to exploit the smoothness and finite-sum structure of our reformulation to converge faster than the existing off-the-shelf solvers and many deterministic first-order algorithmic frameworks for solving WDRSL considered in this paper.

Organization. The paper is organized as follows. In Section 2 we place our work in the context of recent advances in stochastic min-max optimization. In Section 3, we provide the basic setup for the WDRSL problems. In Section 4, we prove that the WDRSL problems can be reformulated as a smooth, convex-concave and finite-sum min-max optimization problem and specify the optimality criterion. In Section 5 and 6, we propose and analyze the SEVR and SPPRR algorithms for solving WDRSL problems and show that both algorithms achieve fast finite-time rates. Finally, we conclude this paper in Section 7.

Notation. We use bold lower-case letters (e.g., $\mathbf{x}$) to denote vectors, upper-case letters (e.g., X) to denote matrices, and calligraphic upper case letters (e.g., $\mathcal{X}$) to denote sets. The notion $[n]$ refers to $\{1, 2, \ldots, n\}$ for some integer $n > 0$. The notions $\mathbb{E}[\cdot \mid \xi]$ and $\mathbb{E}[\cdot]$ refer to the expectation conditioned on the random variable ξ and over all the randomness. We let $a \wedge b$ denote $\min\{a, b\}$.

For a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, we let $\nabla f(\mathbf{x})$ denote the gradient of f at $\mathbf{x}$. For a vector $\mathbf{x} \in \mathbb{R}^d$, we denote $\|\mathbf{x}\|$ as its ℓ_2-norm and $\langle \mathbf{x}, \mathbf{y} \rangle = x^\top y$ as the inner product between two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. For a constraint set $\mathcal{X} \subseteq \mathbb{R}^d$, we let $D_{\mathcal{X}}$ denote its diameter, where $D_{\mathcal{X}} = \max_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}} \|\mathbf{x} - \mathbf{x}'\|$ and let $\mathcal{P}_{\mathcal{X}}$ denote the orthogonal projection onto this set. Lastly, given the accuracy $\epsilon > 0$, the notation $a = O(b(\epsilon))$ stands for the upper bound $a \leq C \cdot b(\epsilon)$ where a constant $C > 0$ is independent of ϵ. Similarly, $a = \tilde{O}(b(\epsilon))$ indicates the previous inequality may depend on the logarithmic function of ϵ, and where a constant $C > 0$ is also independent of ϵ.

2 Related Work

Our work comes amid a surge of interest in gradient-based algorithms for a large class of emerging min-max optimization problems [Daskalakis et al.(2018), Mazumdar et al.(2020), Mokhtari et al.(2020a)]. Despite this activity, existing stochastic-gradient algorithms[1]—e.g., stochastic mirror-prox and its variants [Nemirovski et al.(2009), Juditsky et al.(2011), Chen et al.(2014), Chen et al.(2017)], single-call stochastic extragradient [Hsieh et al.(2019)], and epoch-wise stochastic gradient descent ascent [Yan et al.(2020)]—do not fully exploit the underlying structure of WDRSL and suffer from slow rates of convergence. In particular, current stochastic min-max optimization algorithms for solving finite-sum convex-concave min-max optimization problems, e.g., [Balamurugan and Bach(2016), Shi et al.(2017), Iusem et al.(2017), Du and Hu(2019), Chavdarova et al.(2019), Cui and Shanbhag(2019), Yang et al.(2020), Kotsalis et al.(2020)] often require large batches of data at each iteration or only achieve convergence guarantees if the objective function is strongly convex-concave. The objective function in WDRSL, however, is only convex-concave, motivating us to further tailor simple variance-reduced stochastic gradient algorithms, such as, e.g., [Balamurugan and Bach(2016), Chavdarova et al.(2019)] to effectively take advantage of the smoothness and finite-sum structure of WDRSL problems. In particular, our algorithms make use of two schemes commonly used in optimization to handle minimization problems that can be decomposed as finite sums — variance reduction and random reshuffling of the data.

Variance reduction (VR). Variance-reduced (VR) algorithms were originally proposed for solving finite-sum minimization problems [Johnson and Zhang(2013), Allen-Zhu and Yuan(2016), Reddi et al.(2016), Allen-Zhu and Hazan(2016), Fang et al.(2018), Zhou et al.(2018)] and have recently been extended to finite-sum min-max optimization and variational inequality (VI) problems [Balamurugan and Bach(2016), Chavdarova et al.(2019), Carmon et al.(2019), Alacaoglu and Malitsky(2021) The stochastic variance-reduced gradient (SVRG) algorithm [Balamurugan and Bach(2016), Chavdarova et al.(201 in particular, is the focus of our analysis since remains it computationally tractable in large problems. Previous works have shown that it achieves linear convergence rates in strongly convex-concave finite-sum problems, but its convergence in convex-concave [Xie et al.(2020)] problems remains unknown—an issue that we address in Section 5 by introducing a provably efficient algorithm which we call stochastic extragradient with variance reduction (SEVR). Recently, the concurrent work has appeared [Alacaoglu and Malitsky(2021)] that provides a new complexity bound for stochastic variance-reduced algorithms for solving finite-sum monotone VIs and thus convex-concave min-max optimization problems. Our algorithms are different from theirs and the theoretical results in two papers complement each other.

[1] For brevity, we focus on stochastic algorithms and leave deterministic algorithms, e.g., [Korpelevich(1976), Nemirovski(2004), Nesterov(2007), Chambolle and Pock(2011), Mazumdar et al.(2019), Liang and Stokes(2019), Thekumparampil et al.(2019), Mokhtari et al.(2020b), Lin et al.(2020)] out of our discussion.

Random reshuffling (RR). Algorithms that employ random reshuffling (RR) of the data are increasingly popular among practitioners in ERM problems, but, until recently, have proven challenging to analyze theoretically. Indeed, RR algorithms converge faster than the stochastic gradient algorithms on many problems [Bottou(2009), Recht and Ré(2013)], though the benefit comes at the cost of a significant complication to its theoretical analysis—the gradient estimators are now biased. Recent work in the optimization literature has established convergence rates for RR algorithms [Shamir(2016), Gürbüzbalaban et al.(2019), Haochen and Sra(2019), Nagaraj et al.(2019), Rajput et al.(2020), Safran and Shamir(2020), Nguyen et al.(2020), Mishchenko et al.(2020)]. In particular, they have been shown to converge faster than stochastic gradient descent on quadratic or strongly convex objectives [Gürbüzbalaban et al.(2019), Haochen and Sra(2019), Nguyen et al.(2020), Mishchenko et al.(2020)] and more recently for general smooth and convex objectives [Nagaraj et al.(2019)]. However, to date, random reshuffling has not been proposed for min-max optimization problems.

In this work, we introduce a min-max optimization algorithm that uses random reshuffling and provides performance guarantees. We show theoretically that the algorithm retains the impressive performance of RR methods when compared to other stochastic schemes in min-max optimization.

3 Preliminaries

In this section, we present a setup for the Wasserstein distributionally robust supervised learning (WDRSL) problem which covers a wide range of use cases.

WDRSL. In the statistical learning literature, it is common to assume that the pairs of a data sample and its label are independent, identically distributed, and drawn from some distribution $\mathbb{P}$ supported on $\mathbb{R}^d \times \{-1, 1\}$. In a *generalized linear problem* the parameter $\beta \in \mathbb{R}^d$ can be estimated by solving the following stochastic optimization problem:

$$\inf_{\beta \in \mathbb{R}^d} \left\{ \int_{\mathbb{R}^d \times \{-1,1\}} \left(\Psi(\langle \mathbf{x}, \beta \rangle) - y \langle \mathbf{x}, \beta \rangle \right) \mathbb{P}(d(\mathbf{x}, y)) \right\} \tag{1}$$

where $\Psi : \mathbb{R} \to \mathbb{R}$ is a smooth and nonlinear function, $y \in \{-1, 1\}$ denotes the response variable, $\mathbf{x} \in \mathbb{R}^d$ denotes the predictor (or covariate), and the expectation is with respect to the distribution $\mathbb{P}$. Problem (1) is referred to a generalized linear problem in the canonical form, and covers a wide range of application problems; see [Hardin and Hilbe(2007)] and [Dobson and Barnett(2018)] for the details.

Distributionally robust optimization, see, e.g., [Wiesemann et al.(2014)], seeks to minimize the worst-case expectation of the generalized linear loss function over an ambiguity set $\mathcal{P}$, defined as a neighborhood of the distribution $\mathbb{P}$. The resulting problem has the form:

$$\inf_{\beta \in \mathbb{R}^d} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E} \left[\Psi(\langle \mathbf{x}, \beta \rangle) - y \langle \mathbf{x}, \beta \rangle \right] \tag{2}$$

The set $\mathcal{P}$ can be thought of as the set of distributions in which an adversary could choose the data distribution, and it is common for it to be defined through moment constraints [Delage and Ye(2010)] or with metrics between probability distributions [Namkoong and Duchi(2016), Namkoong and Duchi(2017), Gao et al.(2017), Esfahani and Kuhn(2018), Blanchet and Murthy(2019), Blanchet et al.(2019)].

In practice, since we only have access to samples from $\mathbb{P}$, the ambiguity set is often defined as a neighborhood of the empirical distribution $\widehat{\mathbb{P}}_n$, with the most popular characterization making use of the Wasserstein distance between probability distributions [Abadeh et al.(2015), Gao et al.(2017), Sinha et al.(2018), Blanchet and Murthy(2019), Blanchet et al.(2019), Shafieezadeh-Abadeh et al.(2019)] due to computational tractability and theoretical properties [Esfahani and Kuhn(2018)].

Definition 1 (Wasserstein distance)**.** *Let μ and ν be two probability distributions supported on $\mathcal{Z} = \mathbb{R}^d \times \{-1, 1\}$ and let $\Pi(\mu, \nu)$ denote the set of all couplings (joint distributions) between μ and ν. The Wasserstein distance between μ and ν is defined by*

$$\mathcal{W}(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{Z}\times\mathcal{Z}} c(\mathbf{z}, \mathbf{z}')\pi(d\mathbf{z}, d\mathbf{z}')$$

where $\mathbf{z} = (\mathbf{x}, y) \in \mathbb{R}^d \times \{-1, 1\}$ and $c(\cdot, \cdot)$ is a well-defined metric on $\mathbb{R}^d \times \{-1, 1\}$.

This results in the *Wasserstein distributionally robust generalized linear problem*, which takes the form of Eq. (3), where the ambiguity set is defined as the δ-ball in the Wasserstein distance centered at the empirical distribution $\widehat{\mathbb{P}}_n$, $\mathcal{P} = \mathbb{B}_\delta(\widehat{\mathbb{P}}_n) = \{\mathbb{P} \mid \mathcal{W}(\mathbb{P}, \widehat{\mathbb{P}}_n) \leq \delta\}$:

$$\inf_{\beta\in\mathbb{R}^d} \sup_{\mathbb{P}\in\mathbb{B}_\delta(\widehat{\mathbb{P}}_n)} \mathbb{E}^{\mathbb{P}} \left[\Psi(\langle \mathbf{x}, \beta\rangle) - y\langle \mathbf{x}, \beta\rangle\right] \tag{3}$$

We remark that if $\delta = 0$ (no perturbation), the problem in Eq. (3) is equivalent to solving the classical empirical risk minimization problem.

4 Main Results: WDRSL as Structured Min-max Optimization

In the following three sections, we begin by deriving a new structured min-max optimization formulation of WDRSL with generalized linear losses. We then develop two simple stochastic iterative algorithms that exploit this structure to provably achieve fast finite-time rates. We defer proofs to the supplementary materials.

Before stating our main results, we recall some basic definitions for smooth functions.

Definition 2. *A function f is L-Lipschitz if for $\forall \mathbf{x}, \mathbf{x}'$: $|f(\mathbf{x}) - f(\mathbf{x}')| \leq L\|\mathbf{x} - \mathbf{x}'\|$.*

Definition 3. *A function f is ℓ-smooth if for $\forall \mathbf{x}, \mathbf{x}'$: $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\| \leq \ell\|\mathbf{x} - \mathbf{x}'\|$.*

We also require a notion of ϵ-optimality for the WDRSL problem described in Eq. (3).

Definition 4. *A point $\widehat{\beta}$ is an ϵ-optimal solution ($\epsilon \geq 0$) of Eq.* (3) *if*

$$\sup_{\mathbb{P}\in\mathbb{B}_\delta(\widehat{\mathbb{P}}_n)} \mathbb{E}^{\mathbb{P}} \left[\Psi(\langle \mathbf{x}, \widehat{\beta}\rangle) - y\langle \mathbf{x}, \widehat{\beta}\rangle\right] - \inf_{\beta\in\mathbb{R}^d} \sup_{\mathbb{P}\in\mathbb{B}_\delta(\widehat{\mathbb{P}}_n)} \mathbb{E}^{\mathbb{P}} \left[\Psi(\langle \mathbf{x}, \beta\rangle) - y\langle \mathbf{x}, \beta\rangle\right] \leq \epsilon$$

If $\epsilon = 0$, then $\widehat{\beta}$ is an optimal solution of WDRSL.

Definition 4 applies, inter alia, to the solutions to WDRSL with logistic loss functions [Abadeh et al.(2015)], though in this paper, we consider a more general setting and make the following assumption throughout.

Assumption 1. *The function Ψ is convex, L-Lipschitz and ℓ-smooth for some $L, \ell > 0$ and each data point $(\mathbf{x}_i, y_i)$ satisfies $\|\mathbf{x}_i\| \leq 1$ for all $i \in [n]$.*

Remark. Assumption 1 is mild and covers the WDRSL with logistic and generalized logistic functions [Stukel(1988), Aljarrah et al.(2020)]. In particular, **the second condition is only assumed for the simplicity and can be relaxed to $\|\mathbf{x}_i\| \leq G$ for all $i \in [n]$.** The Lipschitz objective function is necessary and is standard for analyzing stochastic algorithms either without access to full gradients [Nemirovski et al.(2009)] or with random reshuffling [Nagaraj et al.(2019)].

Given these definitions, we now show that the WDRSL problem defined in Eq. (3) is equivalent to a structured min-max optimization problem under Assumption 1 by [Abadeh et al.(2015)]. We provide the proofs in our work for completeness; see Section 4.1 for the details on the min-max reformulation. We then provide the explicit form of our min-max optimization reformulation of the WDRSL problem[2].

Min-max reformulation. It is worth mentioning that the linear model $\Psi(\langle \mathbf{x}, \cdot\rangle)$ is generally assumed in the context of distributionally robust optimization. Under Assumption 1 and for an empirical distribution $\widehat{\mathbb{P}}_n = (1/n)\sum_{i=1}^n \delta_{(\widehat{\mathbf{x}}_i, \widehat{y}_i)}$, the WDRSL problem in Eq. (3) is equivalent to the following structured min-max optimization model:

$$
\begin{aligned}
\min_{(\lambda,\beta)\in\mathbb{R}^d\times\mathbb{R}} \max_{\gamma\in\mathbb{R}^n} \quad & \left\{ \lambda(\delta-\kappa) + \frac{1}{n}\sum_{i=1}^n \Psi(\langle \widehat{\mathbf{x}}_i, \beta\rangle) + \frac{1}{n}\sum_{i=1}^n \gamma_i \left(\widehat{y}_i\langle \widehat{\mathbf{x}}_i, \beta\rangle - \lambda\kappa\right) \right\} \\
\text{s.t.} \quad & \|\beta\| \leq \lambda/(L+1), \ \|\gamma\|_\infty \leq 1
\end{aligned} \tag{4}
$$

where $\kappa > 0$ is a positive constant associated with the metric $c(\mathbf{z}, \mathbf{z}') = \|\mathbf{x} - \mathbf{x}'\| + \kappa|y - y'|$, for points $\mathbf{z} = (\mathbf{x}, y)$ and $\mathbf{z}' = (\mathbf{x}', y')$ in $\mathbb{R}^d$.

Remark. We remark that the min-max model in Eq. (4) has nice properties: (i) It enjoys a finite-sum structure, (ii) The objective function is Lipschitz, smooth, convex in $(\lambda, \beta) \in \mathbb{R} \times \mathbb{R}^d$ and concave in $\gamma \in \mathbb{R}^n$, and (iii) The constraint sets are in the form of second-order cones and the ℓ_∞-ball, such that the orthogonal projections can be computed analytically. For the general finite-sum convex-concave optimization, [Alacaoglu and Malitsky(2021)] provided a generic algorithm with finite-time convergence guarantee. Compared to their algorithm, our algorithms achieve better performance when n is huge and $\epsilon > 0$ is medium.

Remark. We assume that the existence of an optimal saddle point $\mathbf{u}^\star = (\lambda^\star, \beta^\star, \gamma^\star) \in \Lambda \times \Gamma$ in Eq. (4) and do not need to assume that this saddle point $\mathbf{u}^\star \in \Lambda \times \Gamma$ is unique. Nonetheless, the objective function in any optimal saddle point is the same due to Sion's min-max theorem [Sion(1958)].

Given our formulation we can define an ϵ-optimal solution for solving the WDRSL in Eq. (4) and show that it is consistent with the optimality condition in Definition 4. To do so, we denote $\mathbf{u} = (\lambda, \beta, \gamma)$ and let $L(\mathbf{u})$ be the objective function of the smooth min-max optimization model in Eq. (4) with an optimal saddle point $\mathbf{u}^\star = (\lambda^\star, \beta^\star, \gamma^\star) \in \Lambda \times \Gamma$. We define the operator $F : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^n$ as:

$$
F(\mathbf{u}) = \begin{pmatrix} \nabla_{(\lambda,\beta)} L(\mathbf{u}) \\ -\nabla_\gamma L(\mathbf{u}) \end{pmatrix} \tag{5}
$$

[2] We refer to [Gao et al.(2017)] and [Shafieezadeh-Abadeh et al.(2019)] for more general loss functions in WDRSL.

where:

$$\nabla_{(\lambda,\beta)} L(\mathbf{u}) = \begin{pmatrix} \delta - \kappa \left(1 + \frac{1}{n}\sum_{i=1}^n \gamma_i\right) \\ \frac{1}{n}\sum_{i=1}^n \Psi'(\langle \widehat{\mathbf{x}}_i, \beta\rangle)\widehat{\mathbf{x}}_i + \frac{1}{n}\sum_{i=1}^n \gamma_i \widehat{y}_i \widehat{\mathbf{x}}_i \end{pmatrix},$$

$$\nabla_{\gamma} L(\mathbf{u}) = \begin{pmatrix} \frac{1}{n}(\widehat{y}_1 \langle \widehat{\mathbf{x}}_1, \beta\rangle - \lambda\kappa) \\ \vdots \\ \frac{1}{n}(\widehat{y}_n \langle \widehat{\mathbf{x}}_n, \beta\rangle - \lambda\kappa) \end{pmatrix}$$

Finally, we write the constraint sets as $\Lambda = \{(\lambda, \beta) \in \mathbb{R} \times \mathbb{R}^d \mid \|\beta\| \leq \lambda/(L+1)\}$ and $\Gamma = \{\gamma \in \mathbb{R}^n \mid \|\gamma\|_\infty \leq 1\}$. Given this notation, we define the ϵ-optimal solution to the min-max formulation of the WDRSL problem.

Definition 5. *A point $\widehat{\mathbf{u}} = (\widehat{\lambda}, \widehat{\beta}, \widehat{\gamma})$ is an ϵ-optimal saddle-point solution of the WDRSL in Eq.* (4) *if $\widehat{\mathbf{u}} \in \Lambda \times \Gamma$ and $\Delta(\widehat{\mathbf{u}}) = L(\widehat{\lambda}, \widehat{\beta}, \gamma^\star) - L(\lambda^\star, \beta^\star, \widehat{\gamma}) \leq \epsilon$.*

Note that Definition 5 measures the optimality via appeal to the duality gap which is different from the objective function gap in optimization. Before introducing our algorithms for solving this problem, we first show that the two definitions of optimality, Definition 5 and Definition 4, are equivalent.

Theorem 1. *For sufficiently small $\epsilon > 0$, a point $\widehat{\beta}$ is an ϵ-optimal solution of the WDRSL in Eq.* (3) *if a point $\widehat{\mathbf{u}} = (\widehat{\lambda}, \widehat{\beta}, \widehat{\gamma})$ is an ϵ-optimal saddle-point solution of the WDRSL in Eq.* (4) *for some $\widehat{\lambda}$ and $\widehat{\gamma}$. Conversely, there exist $\widehat{\lambda}$ and $\widehat{\gamma}$ such that a point $\widehat{\mathbf{u}} = (\widehat{\lambda}, \widehat{\beta}, \widehat{\gamma})$ is an ϵ-optimal saddle-point solution of the WDRSL in Eq.* (4) *if a point $\widehat{\beta}$ is an an ϵ-optimal solution of the WDRSL in Eq.* (3)*.*

Theorem 1 shows that we can solve the original WDRSL problem posed in Eq. (3) by applying min-max optimization algorithms to the min-max reformulation described in Eq. (4). We remark that this result does not hold in general, and is a consequence of special structure of the WDRSL problem in Eq. (4). Most importantly, the gap function $\Delta(\widehat{\mathbf{u}})$ allows for a finite-time analysis of stochastic min-max optimization algorithms with variance reduction and random reshuffling.

Remark. It is important to remark the difference between the desired accuracy $\epsilon > 0$ and the perturbation level $\delta > 0$. Indeed, we use ϵ to characterize the optimality of the optimization problems. Furthermore, we can increase δ for $\mathcal{W}(\mathbb{P}, \widehat{\mathbb{P}}_n) \leq \delta$ defined in Eq. (3) in order to improve model robustness, which will not affect the computation of an ϵ-optimal saddle point solution of Eq. (4) when ϵ is sufficiently small. Given an ϵ-optimal saddle point solution, we can convert it to an ϵ-optimal solution of Eq. (3).

4.1 Proof of Theorem 1

We first present the formal statement of the min-max reformulation of WDRSL problems in Proposition 1.

Proposition 1 (Min-max Reformulation of WDRSL Problems)**.** *Under Assumption 1 and for an empirical distribution $\widehat{\mathbb{P}}_n = (1/n)\sum_{i=1}^n \delta_{(\widehat{\mathbf{x}}_i, \widehat{y}_i)}$, the WDRSL problem in Eq.* (3) *is equivalent to the following structured min-max optimization model:*

$$\begin{aligned} \min_{(\lambda,\beta)\in\mathbb{R}^d\times\mathbb{R}} \max_{\gamma\in\mathbb{R}^n} \quad & \left\{\lambda(\delta - \kappa) + \frac{1}{n}\sum_{i=1}^n \Psi(\langle \widehat{\mathbf{x}}_i, \beta\rangle) + \frac{1}{n}\sum_{i=1}^n \gamma_i \left(\widehat{y}_i \langle \widehat{\mathbf{x}}_i, \beta\rangle - \lambda\kappa\right)\right\} \\ \text{s.t.} \quad & \|\beta\| \leq \lambda/(L+1), \ \|\gamma\|_\infty \leq 1 \end{aligned} \tag{6}$$

where $\kappa > 0$ is a positive constant associated with the metric $c(\mathbf{z}, \mathbf{z}') = \|\mathbf{x} - \mathbf{x}'\| + \kappa|y - y'|$, for points $\mathbf{z} = (\mathbf{x}, y)$ and $\mathbf{z}' = (\mathbf{x}', y')$ in $\mathbb{R}^d$.

Before presenting the proof of the proposition, we provide a key technical lemma which is a straightforward generalization of [Abadeh et al.(2015), Lemma 1]. The proof is based on a simple yet nontrivial generalization and we present all the details for the sake of completeness.

Lemma 1. *Under Assumption 1 and let $(\widehat{\mathbf{x}}, \widehat{y}) \in \mathbb{R}^d \times \{-1, 1\}$ be a given pair of data sample. Then, for every $\lambda > 0$, we have*

$$\sup_{\mathbf{x} \in \mathbb{R}^d} \Psi(\langle \mathbf{x}, \beta \rangle) - \widehat{y}\langle \mathbf{x}, \beta \rangle - \lambda \|\widehat{\mathbf{x}} - \mathbf{x}\| = \begin{cases} \Psi(\langle \widehat{\mathbf{x}}, \beta \rangle) - \widehat{y}\langle \widehat{\mathbf{x}}, \beta \rangle, & \text{if } \|\beta\| \leq \lambda/(L+1) \\ -\infty, & \text{otherwise} \end{cases}$$

Proof of Proposition 1. The first part of the proof is the same as that of [Abadeh et al.(2015), Theorem 1] and we provide the details for the sake of completeness. Let us denote $\mathbf{z} = (\mathbf{x}, y) \in \mathcal{Z} = \mathbb{R}^d \times \{0, 1\}$ and use the shorthand $h_\beta(\mathbf{z}) = \Psi(\langle \mathbf{x}, \beta \rangle) - y\langle \mathbf{x}, \beta \rangle$. By the definition of the Wasserstein distance and the metric $c(\cdot, \cdot)$, we have

$$\sup_{\mathbb{P} \in \mathbb{B}_\delta(\widehat{\mathbb{P}}_n)} \mathbb{E}^{\mathbb{P}}\left[\Psi(\langle \mathbf{x}, \beta \rangle) - y\langle \mathbf{x}, \beta \rangle\right] = \sup_{\mathbb{P} \in \mathbb{B}_\delta(\widehat{\mathbb{P}}_n)} \int_{\mathcal{Z}} h_\beta(\mathbf{z}) \, \mathbb{P}(d\mathbf{z}) = \left\{ \begin{array}{cl} \sup\limits_{\pi \in \Pi(\mathbb{P}, \widehat{\mathbb{P}}_n)} & \int_{\mathcal{Z}} h_\beta(\mathbf{z}) \, \pi(d\mathbf{z}, \mathcal{Z}) \\ \text{s.t.} & \int_{\mathcal{Z} \times \mathcal{Z}} \|\mathbf{z} - \mathbf{z}'\| \, \pi(d\mathbf{z}, d\mathbf{z}') \leq \delta \end{array} \right.$$

Since the marginal distribution $\widehat{\mathbb{P}}_n$ of $\mathbf{z}'$ is discrete, the coupling π is completely determined by the conditional distribution $\mathbb{P}^i$ of $\mathbf{z}$ given $\mathbf{z}' = \widehat{\mathbf{z}}_i = (\widehat{\mathbf{x}}_i, \widehat{y}_i)$ for all $i \in [n]$. Thus, we have

$$\pi(d\mathbf{z}, d\mathbf{z}') = \frac{1}{n} \sum_{i=1}^n \delta_{(\widehat{\mathbf{x}}_i, \widehat{y}_i)}(d\mathbf{z}') \mathbb{P}^i(d\mathbf{z})$$

Putting these pieces together yields that

$$\sup_{\mathbb{P} \in \mathbb{B}_\delta(\widehat{\mathbb{P}}_n)} \mathbb{E}^{\mathbb{P}}\left[\Psi(\langle \mathbf{x}, \beta \rangle) - y\langle \mathbf{x}, \beta \rangle\right] = \left\{ \begin{array}{cl} \sup\limits_{\mathbb{P}^i} & \frac{1}{n} \sum_{i=1}^n \int_{\mathcal{Z}} h_\beta(\mathbf{z}) \, \mathbb{P}^i(d\mathbf{z}) \\ \text{s.t.} & \frac{1}{n} \sum_{i=1}^n \int_{\mathcal{Z}} \|\mathbf{z} - \widehat{\mathbf{z}}_i\| \, \mathbb{P}^i(d\mathbf{z}) \leq \delta \\ & \int_{\mathcal{Z}} \mathbb{P}^i(d\mathbf{z}) = 1 \end{array} \right.$$

By replacing $\mathbf{z}$ with $(\mathbf{x}, y)$ and decomposing each distribution $\mathbb{P}^i$ into unnormalized measures $\mathbb{P}^i_{\pm 1}(d\mathbf{x}) = \mathbb{P}^i(d\mathbf{x}, \{y = \pm 1\})$ supported on $\mathbb{R}^d$, the right-hand expression simplifies to

$$\left\{ \begin{array}{cl} \sup\limits_{\mathbb{P}^i_{\pm 1}} & \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}^d} h_\beta(\mathbf{x}, +1) \, \mathbb{P}^i_{+1}(d\mathbf{x}) + \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}^d} h_\beta(\mathbf{x}, -1) \, \mathbb{P}^i_{-1}(d\mathbf{x}) \\ \text{s.t.} & \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}^d} \|(\mathbf{x}, +1) - (\widehat{\mathbf{x}}_i, \widehat{y}_i)\| \, \mathbb{P}^i_{+1}(d\mathbf{x}) + \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}^d} \|(\mathbf{x}, -1) - (\widehat{\mathbf{x}}_i, \widehat{y}_i)\| \, \mathbb{P}^i_{-1}(d\mathbf{x}) \leq \delta \\ & \int_{\mathbb{R}^d} \mathbb{P}^i_{+1}(d\mathbf{x}) + \int_{\mathbb{R}^d} \mathbb{P}^i_{-1}(d\mathbf{x}) = 1 \end{array} \right.$$

For the inequality constraint, we further split the sum over $i \in [n]$ into two groups: $\{\widehat{y}_i = +1\}$ and

$\{\widehat{y}_i = -1\}$. Then we have

$$\begin{aligned}
\delta &\geq \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}^d} \|(\mathbf{x},+1)-(\widehat{\mathbf{x}}_i,\widehat{y}_i)\| \, \mathbb{P}^i_{+1}(d\mathbf{x}) + \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}^d} \|(\mathbf{x},-1)-(\widehat{\mathbf{x}}_i,\widehat{y}_i)\| \, \mathbb{P}^i_{-1}(d\mathbf{x}) \\
&= \frac{1}{n}\int_{\mathbb{R}^d}\sum_{\widehat{y}_i=+1}\left[\|\mathbf{x}-\widehat{\mathbf{x}}_i\| \, \mathbb{P}^i_{+1}(d\mathbf{x}) + \|\mathbf{x}-\widehat{\mathbf{x}}_i\| \, \mathbb{P}^i_{-1}(d\mathbf{x}) + 2\kappa\mathbb{P}^i_{-1}(d\mathbf{x})\right] \\
&\quad + \frac{1}{n}\int_{\mathbb{R}^d}\sum_{\widehat{y}_i=-1}\left[\|\mathbf{x}-\widehat{\mathbf{x}}_i\| \, \mathbb{P}^i_{-1}(d\mathbf{x}) + \|\mathbf{x}-\widehat{\mathbf{x}}_i\| \, \mathbb{P}^i_{+1}(d\mathbf{x}) + 2\kappa\mathbb{P}^i_{+1}(d\mathbf{x})\right] \\
&= \frac{2\kappa}{n}\int_{\mathbb{R}^d}\sum_{\widehat{y}_i=+1}\mathbb{P}^i_{-1}(d\mathbf{x}) + \sum_{\widehat{y}_i=-1}\mathbb{P}^i_{+1}(d\mathbf{x}) + \frac{1}{n}\int_{\mathbb{R}^d}\sum_{i=1}^{n}\|\mathbf{x}-\widehat{\mathbf{x}}_i\|\left(\mathbb{P}^i_{-1}(d\mathbf{x}) + \mathbb{P}^i_{+1}(d\mathbf{x})\right)
\end{aligned}$$

To this end, we conclude that

$$\begin{aligned}
&\sup_{\mathbb{P}\in\mathbb{B}_\delta(\widehat{\mathbb{P}}_n)} \mathbb{E}^{\mathbb{P}}\left[\Psi(\langle\mathbf{x},\beta\rangle) - y\langle\mathbf{x},\beta\rangle\right] \\
= &\left\{\begin{array}{ll}
\sup\limits_{\mathbb{P}^i_{\pm 1}} & \frac{1}{n}\sum_{i=1}^n\int_{\mathbb{R}^d} h_\beta(\mathbf{x},+1)\,\mathbb{P}^i_{+1}(d\mathbf{x}) + \frac{1}{n}\sum_{i=1}^n\int_{\mathbb{R}^d} h_\beta(\mathbf{x},-1)\,\mathbb{P}^i_{-1}(d\mathbf{x}) \\
\text{s.t.} & \frac{2\kappa}{n}\int_{\mathbb{R}^d}\sum_{\widehat{y}_i=+1}\mathbb{P}^i_{-1}(d\mathbf{x}) + \sum_{\widehat{y}_i=-1}\mathbb{P}^i_{+1}(d\mathbf{x}) + \frac{1}{n}\int_{\mathbb{R}^d}\sum_{i=1}^n\|\mathbf{x}-\widehat{\mathbf{x}}_i\|(\mathbb{P}^i_{-1}(d\mathbf{x}) + \mathbb{P}^i_{+1}(d\mathbf{x})) \leq \delta \\
& \int_{\mathbb{R}^d}\mathbb{P}^i_{+1}(d\mathbf{x}) + \int_{\mathbb{R}^d}\mathbb{P}^i_{-1}(d\mathbf{x}) = 1
\end{array}\right.
\end{aligned}$$

The above infinite-dimensional optimization problem over the measures $\mathbb{P}^i_{\pm 1}$ admits the semi-infinite dual and strong duality holds for any $\delta > 0$ due to [Shapiro(2001), Proposition 3.4]. That is to say, the following statement holds true,

$$\sup_{\mathbb{P}\in\mathbb{B}_\delta(\widehat{\mathbb{P}}_n)} \mathbb{E}^{\mathbb{P}}\left[\Psi(\langle\mathbf{x},\beta\rangle) - y\langle\mathbf{x},\beta\rangle\right] = \left\{\begin{array}{lll}
\inf\limits_{\lambda,s_i} & \lambda\delta + \frac{1}{n}\sum_{i=1}^n s_i & \\
\text{s.t.} & \sup_{\mathbf{x}\in\mathbb{R}^d} h_\beta(\mathbf{x},+1) - \lambda\|\widehat{\mathbf{x}}_i - \mathbf{x}\| - \lambda\kappa(1-\widehat{y}_i) \leq s_i, & \forall i\in[n] \\
& \sup_{\mathbf{x}\in\mathbb{R}^d} h_\beta(\mathbf{x},-1) - \lambda\|\widehat{\mathbf{x}}_i - \mathbf{x}\| - \lambda\kappa(1+\widehat{y}_i) \leq s_i, & \forall i\in[n] \\
& \lambda \geq 0 &
\end{array}\right.$$

Recall that $h_\beta(\mathbf{x}, y) = \Psi(\langle\mathbf{x},\beta\rangle) - y\langle\mathbf{x},\beta\rangle$. Then Lemma 1 with $\widehat{y} = \pm 1$ implies that

$$\sup_{\mathbf{x}\in\mathbb{R}^d} h_\beta(\mathbf{x},\pm 1) - \lambda\|\widehat{\mathbf{x}}_i - \mathbf{x}\| = h_\beta(\widehat{\mathbf{x}}_i,\pm 1) = \left\{\begin{array}{ll}
\Psi(\langle\widehat{\mathbf{x}}_i,\beta\rangle) \mp \langle\widehat{\mathbf{x}}_i,\beta\rangle, & \text{if } \|\beta\| \leq \frac{\lambda}{L+1} \\
+\infty & \text{otherwise}
\end{array}\right.$$

Putting these pieces together yields that

$$\inf_{\beta\in\mathbb{R}^d}\sup_{\mathbb{P}\in\mathbb{B}_\delta(\widehat{\mathbb{P}}_n)} \mathbb{E}^{\mathbb{P}}\left[\Psi(\langle\mathbf{x},\beta\rangle) - y\langle\mathbf{x},\beta\rangle\right] = \left\{\begin{array}{lll}
\min\limits_{\beta,\lambda,s_i} & \lambda\delta + \frac{1}{n}\sum_{i=1}^n s_i & \\
\text{s.t.} & \Psi(\langle\widehat{\mathbf{x}}_i,\beta\rangle) - \widehat{y}_i\langle\widehat{\mathbf{x}}_i,\beta\rangle \leq s_i, & \forall i\in[n] \\
& \Psi(\langle\widehat{\mathbf{x}}_i,\beta\rangle) + \widehat{y}_i\langle\widehat{\mathbf{x}}_i,\beta\rangle - 2\lambda\kappa \leq s_i, & \forall i\in[n] \\
& \|\beta\| \leq \frac{\lambda}{L+1} &
\end{array}\right. \tag{7}$$

This is a generalization of the optimization problem in [Abadeh et al.(2015), Eq. (7)]. It is equivalent to an unconstrained convex but nonsmooth optimization over the second-order cone as follows,

$$\begin{array}{ll}
\min\limits_{(\lambda,\beta)\in\mathbb{R}\times\mathbb{R}^d} & \lambda\delta + \frac{1}{n}\sum\limits_{i=1}^n \Psi(\langle\widehat{\mathbf{x}}_i,\beta\rangle) - \widehat{y}_i\langle\widehat{\mathbf{x}}_i,\beta\rangle + \frac{1}{n}\sum\limits_{i=1}^n \max\left\{0, 2\widehat{y}_i\langle\widehat{\mathbf{x}}_i,\beta\rangle - 2\lambda\kappa\right\} \\
\text{s.t.} & \|\beta\| \leq \lambda/(L+1)
\end{array} \tag{8}$$

Given a simple transformation given by

$$\max\{0,a\} = \frac{1}{2}(a+|a|) = \frac{1}{2}\left(a + \max_{|b|\leq 1} ab\right)$$

we have

$$\max\left\{0, 2\widehat{y}_i\langle \widehat{\mathbf{x}}_i, \beta\rangle - 2\lambda\kappa\right\} = \widehat{y}_i\langle \widehat{\mathbf{x}}_i, \beta\rangle - \lambda\kappa + \max_{|\gamma_i|\leq 1} \gamma_i\left(\widehat{y}_i\langle \widehat{\mathbf{x}}_i, \beta\rangle - \lambda\kappa\right) \tag{9}$$

Therefore, we conclude that the WDRSL problem in Eq. (3) is equivalent to the structured minimax optimization model in Eq. (4). This completes the proof. □

Proof of Theorem 1. We first prove that, for sufficiently small $\epsilon > 0$, a point $\widehat{\beta}$ is an ϵ-optimal solution of the WDRSL in Eq. (3) if a point $\widehat{\mathbf{u}} = (\widehat{\lambda}, \widehat{\beta}, \widehat{\gamma})$ is an ϵ-optimal saddle-point solution of the WDRSL in Eq. (4).

Without loss of generality, let $\epsilon \to 0$, we have $\widehat{\mathbf{u}} \to \mathbf{u}^\star$. Note that the term $\widehat{y}_i\langle \widehat{\mathbf{x}}_i, \beta\rangle - \lambda\kappa$ is a continuous function of (λ, β). Thus, we have

$$\widehat{y}_i\langle \widehat{\mathbf{x}}_i, \widehat{\beta}\rangle - \widehat{\lambda}\kappa \;\to\; \widehat{y}_i\langle \widehat{\mathbf{x}}_i, \beta^\star\rangle - \lambda^\star\kappa$$

This implies that, for sufficiently small $\epsilon > 0$, we have

$$\mathrm{sign}(\widehat{y}_i\langle \widehat{\mathbf{x}}_i, \widehat{\beta}\rangle - \widehat{\lambda}\kappa) \;=\; \mathrm{sign}\left(\widehat{y}_i\langle \widehat{\mathbf{x}}_i, \beta^\star\rangle - \lambda^\star\kappa\right) \tag{10}$$

Recall that the function $L(\mathbf{u})$ is defined by

$$L(\mathbf{u}) = \lambda(\delta - \kappa) + \frac{1}{n}\sum_{i=1}^n \Psi(\langle \widehat{\mathbf{x}}_i, \beta\rangle) + \frac{1}{n}\sum_{i=1}^n \gamma_i\left(\widehat{y}_i\langle \widehat{\mathbf{x}}_i, \beta\rangle - \lambda\kappa\right)$$

This implies that $\gamma^\star(\lambda, \beta)$ with the i-th entry $\gamma_i^\star(\lambda, \beta) = \mathrm{sign}(\widehat{y}_i\langle \widehat{\mathbf{x}}_i, \beta\rangle - \lambda\kappa)$ for $\forall i \in [n]$ is an unique solution that maximizes the concave function $L(\lambda, \beta, \cdot)$. Putting these pieces together with Eq. (10) implies that $\gamma^\star(\widehat{\lambda}, \widehat{\beta}) = \gamma^\star(\lambda^\star, \beta^\star) = \gamma^\star$. Therefore, we have

$$\Delta(\widehat{\mathbf{u}}) \leq \epsilon \implies L(\widehat{\lambda}, \widehat{\beta}, \gamma^\star(\widehat{\lambda}, \widehat{\beta})) - L(\lambda^\star, \beta^\star, \gamma^\star) \leq \epsilon \tag{11}$$

After some simple calculations (cf. Eq. (8) and Eq. (9)), we have $\|\widehat{\beta}\| \leq \widehat{\lambda}/(L+1)$ and

$$\begin{aligned}
& \widehat{\lambda}\delta + \frac{1}{n}\sum_{i=1}^n \Psi(\langle \widehat{\mathbf{x}}_i, \widehat{\beta}\rangle) - \widehat{y}_i\langle \widehat{\mathbf{x}}_i, \widehat{\beta}\rangle + \frac{1}{n}\sum_{i=1}^n \max\left\{0, 2\widehat{y}_i\langle \widehat{\mathbf{x}}_i, \widehat{\beta}\rangle - 2\widehat{\lambda}\kappa\right\} \\
\leq \quad & \lambda^\star\delta + \frac{1}{n}\sum_{i=1}^n \Psi(\langle \widehat{\mathbf{x}}_i, \beta^\star\rangle) - \widehat{y}_i\langle \widehat{\mathbf{x}}_i, \beta^\star\rangle + \frac{1}{n}\sum_{i=1}^n \max\left\{0, 2\widehat{y}_i\langle \widehat{\mathbf{x}}_i, \beta^\star\rangle - 2\lambda^\star\kappa\right\} + \epsilon
\end{aligned}$$

The primal-dual reformulation (cf. Eq. (7)) further implies that

$$\sup_{\mathbb{P}\in\mathbb{B}_\delta(\widehat{\mathbb{P}}_n)} \mathbb{E}^{\mathbb{P}}\left[\Psi(\langle \mathbf{x}, \widehat{\beta}\rangle) - y\langle \mathbf{x}, \widehat{\beta}\rangle\right] \;\leq\; \widehat{\lambda}\delta + \frac{1}{n}\sum_{i=1}^n \Psi(\langle \widehat{\mathbf{x}}_i, \widehat{\beta}\rangle) - \widehat{y}_i\langle \widehat{\mathbf{x}}_i, \widehat{\beta}\rangle + \frac{1}{n}\sum_{i=1}^n \max\left\{0, 2\widehat{y}_i\langle \widehat{\mathbf{x}}_i, \widehat{\beta}\rangle - 2\widehat{\lambda}\kappa\right\}$$

and

$$\inf_{\beta\in\mathbb{R}^d} \sup_{\mathbb{P}\in\mathbb{B}_\delta(\widehat{\mathbb{P}}_n)} \mathbb{E}^{\mathbb{P}}\left[\Psi(\langle \mathbf{x}, \beta\rangle) - y\langle \mathbf{x}, \beta\rangle\right] \;=\; \lambda^\star\delta + \frac{1}{n}\sum_{i=1}^n \Psi(\langle \widehat{\mathbf{x}}_i, \beta^\star\rangle) - \widehat{y}_i\langle \widehat{\mathbf{x}}_i, \beta^\star\rangle + \frac{1}{n}\sum_{i=1}^n \max\left\{0, 2\widehat{y}_i\langle \widehat{\mathbf{x}}_i, \beta^\star\rangle - 2\lambda^\star\kappa\right\}$$

Putting these pieces together yields that

$$\sup_{\mathbb{P}\in\mathbb{B}_\delta(\widehat{\mathbb{P}}_n)} \mathbb{E}^{\mathbb{P}}\left[\Psi(\langle \mathbf{x}, \widehat{\beta}\rangle) - y\langle \mathbf{x}, \widehat{\beta}\rangle\right] - \inf_{\beta\in\mathbb{R}^d} \sup_{\mathbb{P}\in\mathbb{B}_\delta(\widehat{\mathbb{P}}_n)} \mathbb{E}^{\mathbb{P}}\left[\Psi(\langle \mathbf{x}, \beta\rangle) - y\langle \mathbf{x}, \beta\rangle\right] \leq \epsilon$$

which implies the desired result.

Then we prove that, there exists $\widehat{\lambda}$ and $\widehat{\gamma}$ such that a point $\widehat{\mathbf{u}} = (\widehat{\lambda}, \widehat{\beta}, \widehat{\gamma})$ is an ϵ-optimal saddle-point solution of the WDRSL in Eq. (4) if a point $\widehat{\beta}$ is an an ϵ-optimal solution of the WDRSL in Eq. (3). Indeed, there exists $\widehat{\lambda} > 0$ such that

$$\sup_{\mathbb{P}\in\mathbb{B}_\delta(\widehat{\mathbb{P}}_n)} \mathbb{E}^{\mathbb{P}}\left[\Psi(\langle \mathbf{x}, \widehat{\beta}\rangle) - y\langle \mathbf{x}, \widehat{\beta}\rangle\right] = \widehat{\lambda}\delta + \frac{1}{n}\sum_{i=1}^n \Psi(\langle \widehat{\mathbf{x}}_i, \widehat{\beta}\rangle) - \widehat{y}_i\langle \widehat{\mathbf{x}}_i, \widehat{\beta}\rangle + \frac{1}{n}\sum_{i=1}^n \max\left\{0, 2\widehat{y}_i\langle \widehat{\mathbf{x}}_i, \widehat{\beta}\rangle - 2\widehat{\lambda}\kappa\right\}$$

In addition, we let $\widehat{\gamma} = \gamma^\star$. Then we have

$$\Delta(\widehat{\mathbf{u}}) = L(\widehat{\lambda}, \widehat{\beta}, \gamma^\star) - L(\lambda^\star, \beta^\star, \widehat{\gamma}) = L(\widehat{\lambda}, \widehat{\beta}, \gamma^\star) - L(\lambda^\star, \beta^\star, \gamma^\star) \leq L(\widehat{\lambda}, \widehat{\beta}, \gamma^\star(\widehat{\lambda}, \widehat{\beta})) - L(\lambda^\star, \beta^\star, \gamma^\star(\lambda^\star, \beta^\star))$$

Using the previous calculations, we have

$$\begin{aligned}
L(\widehat{\lambda}, \widehat{\beta}, \gamma^\star(\widehat{\lambda}, \widehat{\beta})) &= \widehat{\lambda}\delta + \frac{1}{n}\sum_{i=1}^n \Psi(\langle \widehat{\mathbf{x}}_i, \widehat{\beta}\rangle) - \widehat{y}_i\langle \widehat{\mathbf{x}}_i, \widehat{\beta}\rangle + \frac{1}{n}\sum_{i=1}^n \max\left\{0, 2\widehat{y}_i\langle \widehat{\mathbf{x}}_i, \widehat{\beta}\rangle - 2\widehat{\lambda}\kappa\right\} \\
&= \sup_{\mathbb{P}\in\mathbb{B}_\delta(\widehat{\mathbb{P}}_n)} \mathbb{E}^{\mathbb{P}}\left[\Psi(\langle \mathbf{x}, \widehat{\beta}\rangle) - y\langle \mathbf{x}, \widehat{\beta}\rangle\right] \\
L(\lambda^\star, \beta^\star, \gamma^\star(\lambda^\star, \beta^\star)) &= \lambda^\star\delta + \frac{1}{n}\sum_{i=1}^n \Psi(\langle \widehat{\mathbf{x}}_i, \beta^\star\rangle) - \widehat{y}_i\langle \widehat{\mathbf{x}}_i, \beta^\star\rangle + \frac{1}{n}\sum_{i=1}^n \max\left\{0, 2\widehat{y}_i\langle \widehat{\mathbf{x}}_i, \beta^\star\rangle - 2\lambda^\star\kappa\right\} \\
&= \inf_{\beta\in\mathbb{R}^d} \sup_{\mathbb{P}\in\mathbb{B}_\delta(\widehat{\mathbb{P}}_n)} \mathbb{E}^{\mathbb{P}}\left[\Psi(\langle \mathbf{x}, \beta\rangle) - y\langle \mathbf{x}, \beta\rangle\right]
\end{aligned}$$

Putting these pieces together yields that

$$\Delta(\widehat{\mathbf{u}}) \leq \sup_{\mathbb{P}\in\mathbb{B}_\delta(\widehat{\mathbb{P}}_n)} \mathbb{E}^{\mathbb{P}}\left[\Psi(\langle \mathbf{x}, \widehat{\beta}\rangle) - y\langle \mathbf{x}, \widehat{\beta}\rangle\right] - \inf_{\beta\in\mathbb{R}^d} \sup_{\mathbb{P}\in\mathbb{B}_\delta(\widehat{\mathbb{P}}_n)} \mathbb{E}^{\mathbb{P}}\left[\Psi(\langle \mathbf{x}, \beta\rangle) - y\langle \mathbf{x}, \beta\rangle\right] \leq \epsilon$$

which implies the desired result. □

5 Main Results: Extragradient Meets Variance Reduction

We present a simple iterative algorithm, *Stochastic Extragradient with Variance Reduction* (SEVR), for solving the min-max reformulation of the WDRSL problem, the pseudo-code of which is presented in Algorithm 1. For simplicity, we denote the component operator $F_i : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^n \rightarrow \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^n$ by

$$F_i(\mathbf{u}) = \begin{pmatrix} \delta - \kappa\,(1 + \gamma_i) \\ \Psi'(\langle \widehat{\mathbf{x}}_i, \beta\rangle)\widehat{\mathbf{x}}_i + \gamma_i\widehat{y}_i\widehat{\mathbf{x}}_i \\ 0 \\ \vdots \\ -(\widehat{y}_i\langle \widehat{\mathbf{x}}_i, \beta\rangle - \lambda\kappa) \\ \vdots \\ 0 \end{pmatrix} \tag{12}$$

Algorithm 1 Stochastic Extragradient with Variance Reduction (SEVR)

Input: $\tilde{\mathbf{u}}^0 = \mathbf{u}_0^0 \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^n$, epoch length k_0, step size $\eta > 0$ and the number of epochs S
Initialization: $l \leftarrow 0$ and $T \leftarrow k_0 2^S - k_0$
for $s = 0, 1, \ldots, S-1$ **do**
  Compute and store $F(\tilde{\mathbf{u}}^s)$
  $k_s \leftarrow k_0 2^s$
  for $t = 0, 1, \ldots, k_s - 1$ **do**
    Set $l \leftarrow l + 1$ and $\eta_{t+1}^s \leftarrow \frac{\eta\sqrt{T}}{\sqrt{2T-l}}$
    Uniformly sample with replacement two indices $i_t, j_t \in [n]$. Note that i_t and j_t are independent and identically distributed, and update $\mathbf{u}_{t+1}^s$ by Eq. (13) and Eq. (14)
  end for
  $\tilde{\mathbf{u}}^{s+1} \leftarrow \frac{1}{k_s} \sum_{t=1}^{k_s} \mathbf{u}_t^s$
  $\mathbf{u}_0^{s+1} \leftarrow \mathbf{u}_{k_s}^s$
end for
Output: $\tilde{\mathbf{u}}^S$

In SEVR, we start each epoch s of the SEVR algorithm with a reference point $\tilde{\mathbf{u}}^s$ and compute $F(\tilde{\mathbf{u}}^s)$ (cf. Eq. (5)). We then proceed with k_s projected extragradient updates of the form:

$$\begin{aligned} \mathbf{g}_t^s &= F(\tilde{\mathbf{u}}^s) + (F_{i_t}(\mathbf{u}_t^s) - F_{i_t}(\tilde{\mathbf{u}}^s)) \\ \bar{\mathbf{u}}_{t+1}^s &= \mathcal{P}_{\Lambda\times\Gamma}(\mathbf{u}_t^s - \eta_{t+1}^s \mathbf{g}_t^s) \qquad (13) \\ \bar{\mathbf{g}}_t^s &= F(\tilde{\mathbf{u}}^s) + (F_{j_t}(\bar{\mathbf{u}}_{t+1}^s) - F_{j_t}(\tilde{\mathbf{u}}^s)) \\ \mathbf{u}_{t+1}^s &= \mathcal{P}_{\Lambda\times\Gamma}(\mathbf{u}_t^s - \eta_{t+1}^s \bar{\mathbf{g}}_t^s) \qquad (14) \end{aligned}$$

and set the next reference point $\tilde{\mathbf{u}}^{s+1}$ to be the average of the iterates in the epoch: $(1/k_s) \sum_{t=1}^{k_s} \mathbf{u}_t^s$.

SEVR is inspired by the SVRG++ algorithm [Allen-Zhu and Yuan(2016)] to min-max optimization problems. It is related to the variance-reduced extragradient algorithms for saddle point problems (SVRG) proposed by [Balamurugan and Bach(2016)] and the stochastic variance-reduced extragradient (SVRE) algorithm proposed by [Chavdarova et al.(2019)]. A key difference between our approach, SEVR, and both SVRG and SVRE however, is that in SVRG and SVRE, the number of inner loops is a constant or is distributed according to a geometric random variable, while the proposed SEVR algorithm performs exponentially more inner loops as the iterates approach the optimal set. This yields improved complexity bounds by reducing the number of times the full gradient is computed. Further, both SVRG and SVRE have only been analyzed in the context of strongly convex-concave min-max problems while in the following theorem we provide complexity bounds for SEVR under the weaker assumption of convex-concave functions.

Theorem 2. *Let $L(\mathbf{u})$ be the objective function for the smooth min-max optimization problem in Eq. (4) with an optimal saddle point $\mathbf{u}^\star = (\lambda^\star, \beta^\star, \gamma^\star) \in \Lambda \times \Gamma$. Under the conditions of Assumption 1, assume that the initial vector $\tilde{\mathbf{u}}^0 = \mathbf{u}_0^0$ satisfies $\|\tilde{\mathbf{u}}^0 - \mathbf{u}^\star\| \leq D_\mathbf{u}$ and $\Delta(\tilde{\mathbf{u}}^0) \leq D_L$ for parameters $D_\mathbf{u}, D_L \geq 0$, and let $\epsilon \in (0,1)$ denote a desired accuracy. Then for a number of* *epochs S, initial epoch length k_0, and step size η, satisfying:*

$$S = O\left(\log_2\left(\frac{D_L}{\epsilon}\right)\right), \quad k_0 = \frac{D_{\mathbf{u}}^2}{\eta D_L}$$
$$\eta = O\left(\min\left\{\frac{1}{\ell+\kappa+1}, \frac{\epsilon}{(\ell+\kappa+1)^2 D_{\mathbf{u}}^2}, \frac{D_{\mathbf{u}}^2}{D_L}\right\}\right)$$

the iterates generated by the SEVR algorithm (cf. Algorithm 1) satisfy $\mathbb{E}[\Delta(\tilde{\mathbf{u}}^S)] \leq \epsilon$. Furthermore, the total number of gradient evaluations N satisfies:

$$N = O\left(n\log\left(\frac{1}{\epsilon}\right) + \frac{D_L}{\epsilon} + \frac{(\ell+\kappa+1)^2 D_{\mathbf{u}}^4}{\epsilon^2}\right)$$

The proof of Theorem 2 can be found in Section 5.1. Through Theorem 2, we immediately observe the benefits of our approach over deterministic algorithms for WDRSL that achieve complexity bounds of $O(n/\epsilon)$ [Liu et al.(2017), Zhang et al.(2019), Li et al.(2019)]. Indeed, our results show that SEVR achieves a better dependence on the number of samples n, in that the total number of gradient evaluations to find an ϵ-optimal stationary point is $O(n\log(1/\epsilon))$.[3] We also remark that SEVR improves over general-purpose interior-point methods [Lee and Mehrotra(2015), Luo and Mehrotra(2019)] which can be used to solve convex reformulations of WDRSL. Such methods have per-iteration costs at least quadratic in the problem dimension d while the per-iteration complexity of SEVR is only linear in d.

Remark. We remark that in contrast to many stochastic gradient algorithms the SEVR algorithm benefits from a sequence of adaptive yet *non-decaying* step sizes $\{\eta_t^s\}_{s\geq 0, t\geq 0}$, which is important in practice when a very large number of iterations is commonly desired.

Remark. Algorithm 1 is not a straightforward extension of [Allen-Zhu and Yuan(2016)]; indeed, it is an extragradient-based algorithm whereas the algorithm in [Allen-Zhu and Yuan(2016)] is an variant of gradient descent (GD). The gradient descent ascent, which extends GD to min-max optimization, might diverge or converges to limit circles in convex-concave setting. To analyze our algorithms, we construct a different potential function based on the duality gap, which is unnecessary in [Allen-Zhu and Yuan(2016)].

5.1 Proof of Theorem 2

In this section, we provide the detailed proofs for Theorem 2. Our derivation is based on a nontrivial combination of the analysis in [Allen-Zhu and Hazan(2016)] and [Chavdarova et al.(2019)]. For the simplicity, we denote $\mathbf{u} = (\lambda, \beta, \gamma)$. The convex-concave function $L : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$ and its component function $L_i : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$ are defined as follows:

$$L(\mathbf{u}) = \lambda(\delta - \kappa) + \frac{1}{n}\sum_{i=1}^n \Psi(\langle \widehat{\mathbf{x}}_i, \beta\rangle) + \frac{1}{n}\sum_{i=1}^n \gamma_i\left(\widehat{y}_i\langle \widehat{\mathbf{x}}_i, \beta\rangle - \lambda\kappa\right)$$

and

$$L_i(\mathbf{u}) = \lambda(\delta - \kappa) + \Psi(\langle \widehat{\mathbf{x}}_i, \beta\rangle) + \gamma_i\left(\widehat{y}_i\langle \widehat{\mathbf{x}}_i, \beta\rangle - \lambda\kappa\right)$$

[3] Recent work, [Alacaoglu and Malitsky(2021)], proved rates for a different stochastic variance-reduced algorithm under similar assumptions and derived a bound of $O(n + \sqrt{n}/\epsilon)$ which has a worse dependence on ϵ and n than that in Theorem 2 (when n is sufficiently large).

By the definition of F and F_i in Eq. (5) and Eq. (12), they can be expressed as follows:

$$F(\mathbf{u}) = \begin{pmatrix} \nabla_{(\lambda,\beta)} L(\mathbf{u}) \\ -\nabla_{\gamma} L(\mathbf{u}) \end{pmatrix}, \quad F_i(\mathbf{u}) = \begin{pmatrix} \nabla_{(\lambda,\beta)} L_i(\mathbf{u}) \\ -\nabla_{\gamma} L_i(\mathbf{u}) \end{pmatrix}$$

We also denote the constraint sets by

$$\begin{aligned} \Lambda &= \{(\lambda, \beta) \in \mathbb{R} \times \mathbb{R}^d \mid \|\beta\| \leq \lambda/(L+1)\} \\ \Gamma &= \{\gamma \in \mathbb{R}^n \mid \|\gamma\|_\infty \leq 1\} \end{aligned}$$

For the reference, we rewrite each iteration of the SEVR algorithm as follows:

$$\begin{aligned} \mathbf{g}_t^s &= F(\tilde{\mathbf{u}}^s) + (F_{i_t}(\mathbf{u}_t^s) - F_{i_t}(\tilde{\mathbf{u}}^s)), & \bar{\mathbf{u}}_{t+1}^s &= \mathcal{P}_{\Lambda\times\Gamma}(\mathbf{u}_t^s - \eta_{t+1}^s \mathbf{g}_t^s) \\ \bar{\mathbf{g}}_t^s &= F(\tilde{\mathbf{u}}^s) + (F_{j_t}(\bar{\mathbf{u}}_{t+1}^s) - F_{j_t}(\tilde{\mathbf{u}}^s)), & \mathbf{u}_{t+1}^s &= \mathcal{P}_{\Lambda\times\Gamma}(\mathbf{u}_t^s - \eta_{t+1}^s \bar{\mathbf{g}}_t^s) \end{aligned}$$

Finally, we denote $(\lambda^\star, \beta^\star, \gamma^\star)$ as an optimal saddle point of the smooth minimax optimization model in Eq. (4) and denote the duality gap function $\Delta(\mathbf{u})$ by

$$\Delta(\mathbf{u}) = L(\lambda, \beta, \gamma^\star) - L(\lambda^\star, \beta^\star, \gamma)$$

It is worth noting that the function $\Delta(\mathbf{u})$ is convex in $\mathbf{u}$ since L is a convex-concave function.

Technical Lemmas Our first lemma is to provide an upper bound for the variance of the gradient estimators $\mathbf{g}_t^s$ and $\bar{\mathbf{g}}_t^s$.

Lemma 2. *Under Assumption 1 and let $\mathbf{u}^\star = (\lambda^\star, \beta^\star, \gamma^\star) \in \Lambda \times \Gamma$ be an optimal saddle point of the smooth minimax optimization model in Eq. (4). Then, the following statement holds true,*

$$\begin{aligned} \mathbb{E}[\|\mathbf{g}_t^s - F(\mathbf{u}_t^s)\|^2 \mid \mathbf{u}_t^s, \tilde{\mathbf{u}}^s] &\leq 2(\ell+\kappa+1)^2 \left(\|\mathbf{u}_t^s - \mathbf{u}^\star\|^2 + \|\tilde{\mathbf{u}}^s - \mathbf{u}^\star\|^2\right) \\ \mathbb{E}[\|\bar{\mathbf{g}}_t^s - F(\bar{\mathbf{u}}_{t+1}^s)\|^2 \mid \bar{\mathbf{u}}_{t+1}^s, \tilde{\mathbf{u}}^s] &\leq 16(\ell+\kappa+1)\left(\Delta(\bar{\mathbf{u}}_{t+1}^s) + \Delta(\tilde{\mathbf{u}}^s)\right) \\ &\quad +2(\ell+\kappa+1)^2 \left(\|\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}^\star\|^2 + \|\tilde{\mathbf{u}}^s - \mathbf{u}^\star\|^2\right) \end{aligned}$$

Then, we provide a descent lemma for the iterates generated by the SEVR algorithm.

Lemma 3. *Under Assumption 1 and let $\mathbf{u}^\star = (\lambda^\star, \beta^\star, \gamma^\star) \in \Lambda \times \Gamma$ be an optimal saddle point of the smooth minimax optimization model in Eq. (4). Then, the following statement holds true,*

$$\begin{aligned} & \left(1 - 24\eta_{t+1}^s(\ell+\kappa+1)\right)\mathbb{E}[\Delta(\bar{\mathbf{u}}_{t+1}^s) \mid \mathbf{u}_t^s, \tilde{\mathbf{u}}^s] \\ \leq\ & \frac{1}{2\eta_{t+1}^s}\left(\|\mathbf{u}_t^s - \mathbf{u}^\star\|^2 - \mathbb{E}[\|\mathbf{u}_{t+1}^s - \mathbf{u}^\star\|^2 \mid \mathbf{u}_t^s, \tilde{\mathbf{u}}^s]\right) + 24\eta_{t+1}^s(\ell+\kappa+1)\Delta(\tilde{\mathbf{u}}^s) \\ & - \left(\frac{1}{2\eta_{t+1}^s} - \frac{3\eta_{t+1}^s(\ell+\kappa+1)^2}{2}\right)\mathbb{E}[\|\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}_t^s\|^2 \mid \mathbf{u}_t^s, \tilde{\mathbf{u}}^s] + 6\eta_{t+1}^s(\ell+\kappa+1)^2\|\tilde{\mathbf{u}}^s - \mathbf{u}^\star\|^2 \\ & +3\eta_{t+1}^s(\ell+\kappa+1)^2\|\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}^\star\|^2 + 3\eta_{t+1}^s(\ell+\kappa+1)^2\|\mathbf{u}_t^s - \mathbf{u}^\star\|^2 \end{aligned}$$

Now we are ready to prove a key technical lemma which is crucial to our subsequent analysis.

Lemma 4. *Under Assumption 1 and let $\mathbf{u}^\star = (\lambda^\star, \beta^\star, \gamma^\star) \in \Lambda \times \Gamma$ be an optimal saddle point of the smooth minimax optimization model in Eq. (4). If $k_0 \geq 1$, $0 < \eta \leq \frac{1}{100(\ell+\kappa+1)}$, and $\frac{1}{2\sqrt{2}\eta T} \geq 100\eta(\ell+\kappa+1)^2$, the following statement holds true,*

$$\mathbb{E}[\Delta(\tilde{\mathbf{u}}^S)] \leq \frac{1}{2^S}\left(\Delta(\tilde{\mathbf{u}}^0) + 18\eta(\ell+\kappa+1)^2\|\tilde{\mathbf{u}}^0 - \mathbf{u}^\star\|^2 + \frac{3\|\mathbf{u}_0^0 - \mathbf{u}^\star\|^2}{2k_s\eta_0^s}\right)$$

We are ready to present the

Proof of Theorem 2. Recall that the given parameter choices are

$$\begin{aligned} S &= 1 + \left\lfloor \log_2 \left(\frac{10 D_L}{\epsilon} \right) \right\rfloor \\ \eta &= \min\left\{ \frac{1}{100(\ell+\kappa+1)}, \frac{\epsilon}{2000\sqrt{2}(\ell+\kappa+1)^2 D_{\mathbf{u}}^2}, \frac{D_{\mathbf{u}}^2}{D_L} \right\} \\ k_0 &= \frac{D_{\mathbf{u}}^2}{\eta D_L} \ \geq \ 1 \end{aligned}$$

Then, we have

$$\frac{1}{2\sqrt{2}\eta T} \geq \frac{1}{2\sqrt{2}\eta k_0 \cdot 2^S} = \frac{1}{2\sqrt{2}\eta k_0} \frac{\epsilon}{10 D_L} = \frac{\epsilon}{20\sqrt{2} D_{\mathbf{u}}^2}$$

This together with

$$0 < \eta \leq \frac{\epsilon}{2000\sqrt{2}(\ell+\kappa+1)^2 D_{\mathbf{u}}^2}$$

yields that

$$\frac{1}{2\sqrt{2}\eta T} \geq \frac{\epsilon}{20\sqrt{2} D_{\mathbf{u}}^2} \geq 100\eta(\ell+\kappa+1)^2$$

Thus, Lemma 4 holds true and we have

$$\mathbb{E}[\Delta(\tilde{\mathbf{u}}^S)] \leq \frac{1}{2^S}\left(\Delta(\tilde{\mathbf{u}}^0) + 18\eta(\ell+\kappa+1)^2\|\tilde{\mathbf{u}}^0 - \mathbf{u}^\star\|^2 + \frac{3\|\mathbf{u}_0^0 - \mathbf{u}^\star\|^2}{\eta k_0 \sqrt{2}} \right)$$

Now we bound the three terms on the right hand side of the above inequality. Indeed, we have

$$\begin{aligned} \frac{\Delta(\tilde{\mathbf{u}}^0)}{2^S} &\leq \left(\frac{\Delta(\tilde{\mathbf{u}}^0)}{10 D_L} \right)\epsilon \ \leq \ \frac{\epsilon}{3} \\ \frac{18\eta(\ell+\kappa+1)^2\|\tilde{\mathbf{u}}^0 - \mathbf{u}^\star\|^2}{2^S} &\leq 18\eta(\ell+\kappa+1)^2\|\tilde{\mathbf{u}}^0 - \mathbf{u}^\star\|^2 \ \leq \ \left(\frac{18\|\tilde{\mathbf{u}}^0 - \mathbf{u}^\star\|^2}{400\sqrt{2} D_{\mathbf{u}}^2} \right)\epsilon \ \leq \ \frac{\epsilon}{3} \\ \frac{1}{2^S}\left(\frac{3\|\mathbf{u}_0^0 - \mathbf{u}^\star\|^2}{\eta k_0 \sqrt{2}} \right) &\leq \frac{\epsilon}{10 D_L} \frac{3\|\mathbf{u}_0^0 - \mathbf{u}^\star\|^2}{\sqrt{2}} \frac{D_L}{D_{\mathbf{u}}^2} \ \leq \ \left(\frac{3\|\mathbf{u}_0^0 - \mathbf{u}^\star\|^2}{10\sqrt{2} D_{\mathbf{u}}^2} \right)\epsilon \ \leq \ \frac{\epsilon}{3} \end{aligned}$$

This completes the proof. □

6 Main Results: Proximal Point Meets Random Reshuffling

The second algorithm we propose, *Stochastic Proximal Point with Random Reshuffling* (SPPRR), uses the idea of random reshuffling. In each epoch s of the SPPRR algorithm, we sample a random permutation $\{\sigma_0, \sigma_1, \ldots, \sigma_{n-1}\}$ of the set $[n]$ and proceed with n inexact proximal point updates of the form: $\mathbf{u}_{t+1}^s \approx \mathcal{P}_{\Lambda\times\Gamma}(\mathbf{u}_t^s - \eta F_{\sigma_t^s}(\mathbf{u}_{t+1}^s))$, where $\eta > 0$ is a step size and $\sigma = \sigma_t^s$. We then set $\mathbf{u}_0^{s+1} = \mathbf{u}_n^s$ and repeat the process for a total of S epochs. Note that a new permutation/shuffling is only generated at the beginning of each epoch; see Algorithm 2.

To compute the inexact proximal point updates, we first note that the exact proximal point update is equivalent to the solution of the fixed-point problem $\mathbf{u} = T(\mathbf{u})$, where the operator $T(\mathbf{u})$ is defined as $T(\mathbf{u}) := \mathcal{P}_{\Lambda\times\Gamma}(\mathbf{u}_t^s - \eta F_{\sigma_t}(\mathbf{u}))$. If $\eta \leq 1/(2(\ell+\kappa+1))$, it is easy to verify that T is

Algorithm 2 Stochastic Proximal Point with Random Reshuffling (SPPRR)

Input: $\mathbf{u}_0^0 \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^n$, step size $\eta > 0$ and the number of epochs S
Initialization:
for $s = 0, 1, \ldots, S-1$ **do**
  $\sigma^s \leftarrow$ a random permutation of the set $[n]$
  for $t = 0, 1, \ldots, n-1$ **do**
    $\mathbf{u}_{0,t}^s \leftarrow \mathbf{u}_t^s$
    for $i = 0, 1, \ldots, M-1$ **do**
      Update the iterate $\mathbf{u}_{i+1,t}^s$ by Eq. (15)
    end for
    $\mathbf{u}_{t+1}^s \leftarrow \mathbf{u}_{M,t}^s$
  end for
  $\mathbf{u}_0^{s+1} \leftarrow \mathbf{u}_n^s$
end for
Output: $\tilde{\mathbf{u}}^S = \frac{1}{nS} \sum_{s=0}^{S-1} \sum_{t=1}^{n} \mathbf{u}_t^s$

a contraction. Thus, if we let $\mathbf{u}_{0,t}^s = \mathbf{u}_t^s$, we can define our inexact proximal point update to be the result of M applications of the operator T to $\mathbf{u}_{0,t}^s$, which takes the following iterative form for $i = 1, ..., M$:

$$\mathbf{u}_{i+1,t}^s \;=\; \mathcal{P}_{\Lambda\times\Gamma}(\mathbf{u}_t^s - \eta F_{\sigma_t^s}(\mathbf{u}_{i,t}^s)) \tag{15}$$

We set $\mathbf{u}_{t+1}^s = \mathbf{u}_{M,t}^s$ and proceed to the next inexact proximal point update. In our theoretical treatment we set $M = O(\log(1/\epsilon))$ to guarantee that $\mathbf{u}_{i+1,t}^s$ is sufficiently close to the fixed point, while in practice it suffices to have $M = 2$ to observe fast convergence.

To our knowledge, the SPPRR algorithm is the first stochastic first-order algorithm for solving the min-max optimization problem with random reshuffling. The following theorem shows that the complexity bound of our algorithm is independent of the number of samples n up to a logarithmic factor.

Theorem 3. *Let $L(\mathbf{u})$ be the objective function of the smooth min-max optimization problem in Eq. (4) with an optimal saddle point $\mathbf{u}^\star = (\lambda^\star, \beta^\star, \gamma^\star) \in \Lambda \times \Gamma$, and define $D_{\mathbf{u}} = \|\mathbf{u}_0^0 - \mathbf{u}^\star\|$. Under the conditions of Assumption 1, if each F_i is bounded over $\Lambda \times \Gamma$ for all $i \in [n]$, then for a desired accuracy level $\epsilon \in (0,1)$, number of epochs S, step size η, and number of fixed point-iterations M that satisfy:*

$$S = \, O\left(\frac{1}{n}\left(\frac{(\ell+\kappa+1)D_{\mathbf{u}}^2}{\epsilon} + \frac{G^2 D_{\mathbf{u}}^2}{\epsilon^2}\right)\right)$$
$$\eta = \, \min\left\{\frac{1}{2(\ell+\kappa+1)}, \frac{\epsilon}{4G^2}\right\}, \; M \,=\, 1 + \lfloor \log_2(10nS) \rfloor$$

the iterates generated by the SPPRR algorithm satisfy $\mathbb{E}[\Delta(\tilde{\mathbf{u}}^S)] \leq \epsilon$. Further, the total number of gradient evaluations, N satisfies:

$$N = O\left(\left(\frac{(\ell+\kappa+1)D_{\mathbf{u}}^2}{\epsilon} + \frac{G^2 D_{\mathbf{u}}^2}{\epsilon^2}\right)\log\left(\frac{1}{\epsilon}\right)\right)$$

Our results extend convergence guarantees for random reshuffling for SGD on smooth, bounded functions [Nagaraj et al.(2019)] to smooth convex-concave min-max optimization regime. Note that

N does not depend on n; indeed, the random reshuffling algorithms can be viewed as stochastic gradient-based algorithms without replacement and the complexity bound is independent of n. Such result was known for minimization and our work extends it to min-max optimization.

Remark. We remark that the boundedness of the component functions is a standard assumption for proving the convergence of stochastic gradient algorithms with random reshuffling in minimization problems [Shamir(2016), Haochen and Sra(2019), Nagaraj et al.(2019), Rajput et al.(2020), Safran and Shamir(2020), Nguyen et al.(2020), Mishchenko et al.(2020)]. In the min-max regime, this translates to requiring the boundedness of each F_i over the constraint set $\Lambda \times \Gamma$.

Remark. It is worth emphasizing the importance of random reshuffling-based algorithms in the min-max optimization; indeed, while the random reshuffling-based algorithms generally do not improve the convergence rate in convex optimization, they are extremely efficient in practice. This motivates us to consider their extension to min-max optimization.

Remark. Algorithm 2 is not a straightforward extension of [Nagaraj et al.(2019)]; indeed, it is a proximal point-based algorithm whereas the algorithm in [Nagaraj et al.(2019)] is an variant of gradient descent (GD). The necessity of fixed-point sub-problem comes from the generality of min-max optimization. More specifically, [Nagaraj et al.(2019)] used the existence of *convex objective functions* for proving their Lemma 2 and 4 which can not be extended to convex-concave min-max optimization problems in a obvious way. We resolve this issue by changing the subproblem to the proximal subproblem and solving the resulting fixed-point problem.

6.1 Proof of Theorem 3

In this section, we provide the detailed proofs for Theorem 3. Our derivation extends the analysis in [Nagaraj et al.(2019)] from convex optimization to convex-concave minimax optimization. For the simplicity, we also denote $\mathbf{u} = (\lambda, \beta, \gamma)$ and the functions L and L_i by

$$\begin{aligned} L(\mathbf{u}) &= \lambda(\delta - \kappa) + \frac{1}{n}\sum_{i=1}^{n} \Psi(\langle \widehat{\mathbf{x}}_i, \beta \rangle) + \frac{1}{n}\sum_{i=1}^{n} \gamma_i \left(\widehat{y}_i \langle \widehat{\mathbf{x}}_i, \beta \rangle - \lambda\kappa \right) \\ L_i(\mathbf{u}) &= \lambda(\delta - \kappa) + \Psi(\langle \widehat{\mathbf{x}}_i, \beta \rangle) + \gamma_i \left(\widehat{y}_i \langle \widehat{\mathbf{x}}_i, \beta \rangle - \lambda\kappa \right) \end{aligned}$$

We also have

$$F(\mathbf{u}) = \begin{pmatrix} \nabla_{(\lambda,\beta)} L(\mathbf{u}) \\ -\nabla_\gamma L(\mathbf{u}) \end{pmatrix}, \quad F_i(\mathbf{u}) = \begin{pmatrix} \nabla_{(\lambda,\beta)} L_i(\mathbf{u}) \\ -\nabla_\gamma L_i(\mathbf{u}) \end{pmatrix}$$

and

$$\begin{aligned} \Lambda &= \{(\lambda, \beta) \in \mathbb{R} \times \mathbb{R}^d \mid \|\beta\| \leq \lambda/(L+1)\} \\ \Gamma &= \{\gamma \in \mathbb{R}^n \mid \|\gamma\|_\infty \leq 1\} \end{aligned}$$

For the reference, we rewrite each iteration of the SPPRR algorithm as follows:

$$\mathbf{u}_{i+1,t}^s \;=\; \mathcal{P}_{\Lambda \times \Gamma}(\mathbf{u}_t^s - \eta F_{\sigma_t^s}(\mathbf{u}_{i,t}^s))$$

Finally, we denote $(\lambda^\star, \beta^\star, \gamma^\star)$ as an optimal saddle point of the smooth minimax optimization model in Eq. (4) and recall that the duality gap function $\Delta(\mathbf{u})$ is defined by

$$\Delta(\mathbf{u}) = L(\lambda, \beta, \gamma^\star) - L(\lambda^\star, \beta^\star, \gamma)$$

It is worth noting that the function $\Delta(\mathbf{u})$ is convex in $\mathbf{u}$ since L is a convex-concave function.

Technical Lemmas Before presenting our technical lemmas, we define the exchangeable pair and some other key notations which are first introduced by [Nagaraj et al.(2019)] for convex optimization and become a standard machinery for analyzing stochastic algorithm with random reshuffling.

Suppose that we run the SPPRR algorithm for s epochs using the random permutation $\sigma^0, \sigma^1, \ldots, \sigma^{s-1}$ and obtain $\mathbf{u}_0^s = \mathbf{u}_n^{s-1}$. If $s = 0$, we start with the initial point $\mathbf{u}_0^0$. The exchange pair is defined by two different s-th epoch iterates $\{\mathbf{u}_t(\sigma^s)\}_{1 \leq t \leq n}$ and $\{\mathbf{u}_t(\tilde{\sigma}^s)\}_{1 \leq t \leq n}$ obtained by running the s-th epoch with independent uniform permutations σ^s and $\tilde{\sigma}^s$. Then, it is obvious that $\{\mathbf{u}_t(\sigma^s)\}_{1 \leq t \leq n}$ and $\{\mathbf{u}_t(\tilde{\sigma}^s)\}_{1 \leq t \leq n}$ are independent and identically distributed. This further implies that

$$\mathbb{E}[L_{\sigma_{t-1}^s}(\lambda_t(\tilde{\sigma}^s), \beta_t(\tilde{\sigma}^s), \gamma^\star) - L_{\sigma_{t-1}^s}(\lambda^\star, \beta^\star, \gamma_t(\tilde{\sigma}^s))] = \mathbb{E}[\Delta(\mathbf{u}_t(\tilde{\sigma}^s))] = \mathbb{E}[\Delta(\mathbf{u}_t(\sigma^s))] = \mathbb{E}[\Delta(\mathbf{u}_t^s)] \tag{16}$$

Let $\mathcal{D}_{t,s}$ be the distribution of the iterate $\mathbf{u}_t^s$ under the random shuffling σ^s and $\mathcal{D}_{t,s}^{(r)}$ be the distribution of the iterate $\mathbf{u}_t^s$ under the random shuffling σ^s conditioned on the event $\{\sigma_{t-1}^s = r\}$. This is different from the event $\{\sigma_{t+1}^s = r\}$ used in [Nagaraj et al.(2019)] since we analyze inexact proximal point update instead of projected gradient update. Nonetheless, our proof techniques are also based on the Kantorovich duality and the 1-Wasserstein and 2-Wasserstein distances between $\mathcal{D}_{t,s}$ and $\mathcal{D}_{t,s}^{(r)}$.

Definition 6 (1-Wasserstein and 2-Wasserstein distance)**.** *Suppose that μ and ν be two probability distributions over $\mathbb{R}^N$ such that $\mathbb{E}_\mu[\|X\|^2] < +\infty$ and $\mathbb{E}_\nu[\|Y\|^2] < +\infty$. Let $X \sim \mu$ and $Y \sim \nu$ be random vectors defined on a common measure space (i.e., they are coupled). The 1-Wasserstein and 2-Wasserstein distances between μ and ν are defined by*

$$\begin{aligned} \mathcal{W}_1(\mu, \nu) &= \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_\pi[\|X - Y\|] \\ \mathcal{W}_2(\mu, \nu) &= \inf_{\pi \in \Pi(\mu,\nu)} \sqrt{\mathbb{E}_\pi[\|X - Y\|^2]} \end{aligned}$$

where $\pi \in \Pi(\mu, \nu)$ denotes a coupling (or a joint distribution) over (X, Y) with marginals μ and ν.

By the above definition and Jensen's inequality, we have $\mathcal{W}_1(\mu, \nu) \leq \mathcal{W}_2(\mu, \nu)$. The following lemma summarizes a key characterization of 1-Wasserstein distance [Villani(2008), Theorem 5.10].

Lemma 5 (Kantorovich Duality [Villani(2008)])**.** *Suppose that μ and ν be two probability distributions over $\mathbb{R}^N$ such that $\mathbb{E}_\mu[\|X\|^2] < +\infty$ and $\mathbb{E}_\nu[\|Y\|^2] < +\infty$. Then, we have*

$$\mathcal{W}_1(\mu, \nu) = \sup_{g \text{ is 1-Lipschitz}} \mathbb{E}_\mu[g(X)] - \mathbb{E}_\nu[g(Y)]$$

The second lemma is to provide an upper bound for the difference between the unbiased gap $\mathbb{E}[\Delta(\mathbf{u}_t^s)]$ and the gap estimator $\mathbb{E}[L_{\sigma_{t-1}^s}(\lambda_t^s, \beta_t^s, \gamma^\star) - L(\lambda^\star, \beta^\star, \gamma_t^s)]$ using the 2-Wasserstein distance.

Lemma 6. *Let $\mathbf{u}^\star = (\lambda^\star, \beta^\star, \gamma^\star) \in \Lambda \times \Gamma$ be an optimal saddle point of the smooth minimax optimization model in Eq. (4). Under Assumption 1 and let F_i be bounded over $\Lambda \times \Gamma$ for all $i \in [n]$. Then, the following statement holds true,*

$$\left|\mathbb{E}[\Delta(\mathbf{u}_t^s)] - \mathbb{E}[L_{\sigma_{t-1}^s}(\lambda_t^s, \beta_t^s, \gamma^\star) - L_{\sigma_{t-1}^s}(\lambda^\star, \beta^\star, \gamma_t^s)]\right| \leq \frac{G}{n}\sum_{r=1}^{n} \mathcal{W}_2(\mathcal{D}_{t,s}, \mathcal{D}_{t,s}^{(r)})$$

Recall that the inexact oracle returns a point $\mathbf{u}_{M,t}^s$ that approximates the fixed-point problem $\mathbf{u} = \mathcal{P}_{\Lambda\times\Gamma}(\mathbf{u}_t^s - \eta F_{\sigma_t^s}(\mathbf{u}))$. For the simplicity of notation, we denote the fixed point by $\bar{\mathbf{u}}_t^s \in \Lambda \times \Gamma$ given the point $\mathbf{u}_t^s$. The following lemma provides an upper bound based on the coupling characterization of 2-Wasserstein distance under the certain error criterion.

Lemma 7. *Let $\mathbf{u}^\star = (\lambda^\star, \beta^\star, \gamma^\star) \in \Lambda \times \Gamma$ be an optimal saddle point of the smooth minimax optimization model in Eq. (4). Under Assumption 1 and let F_i be bounded over $\Lambda \times \Gamma$ for all $i \in [n]$. Suppose that $\epsilon \in (0,1)$ denotes the desired tolerance. Given any fixed random permutation σ_t^s, the following error criterion holds true,*

$$\|\mathbf{u}_t^s - \bar{\mathbf{u}}_{t-1}^s\| = \|\mathbf{u}_{M,t-1}^s - \bar{\mathbf{u}}_{t-1}^s\| \ \leq \ \frac{D_{\mathbf{u}}}{10(nS)^{3/2}}$$

Then, the following statement holds true,

$$\left|\mathbb{E}[\Delta(\mathbf{u}_t^s)] - \mathbb{E}[L_{\sigma_{t-1}^s}(\lambda_t^s, \beta_t^s, \gamma^\star) - L_{\sigma_{t-1}^s}(\lambda^\star, \beta^\star, \gamma_t^s)]\right| \ \leq \ 2\eta G^2 + \frac{D_{\mathbf{u}}}{5\sqrt{nS}}$$

Finally, we provide a upper bound on the term $\|\mathbf{u}_t^s - \mathbf{u}^\star\|$ and $\|\bar{\mathbf{u}}_{t-1}^s - \mathbf{u}^\star\|$ for $\forall t \in [n]$ and $\forall s \in \{0, 1, \ldots, S-1\}$. This bound is universal and depends on η, G, n, S, $D_{\mathbf{u}}$ and ϵ

Lemma 8. *Let $\mathbf{u}^\star = (\lambda^\star, \beta^\star, \gamma^\star) \in \Lambda \times \Gamma$ be an optimal saddle point of the smooth minimax optimization model in Eq. (4). Under Assumption 1 and let F_i be bounded over $\Lambda \times \Gamma$ for all $i \in [n]$. Suppose that the initial vector $\mathbf{u}_0^0$ satisfies that $\|\mathbf{u}_0^0 - \mathbf{u}^\star\| \leq D_{\mathbf{u}}$ for parameters $D_{\mathbf{u}} \geq 0$ and $\epsilon \in (0,1)$ denotes the desired tolerance. Given any fixed random permutation σ_t^s, the following error criterion holds true,*

$$\|\mathbf{u}_t^s - \bar{\mathbf{u}}_{t-1}^s\| = \|\mathbf{u}_{M,t-1}^s - \bar{\mathbf{u}}_{t-1}^s\| \ \leq \ \frac{D_{\mathbf{u}}}{10(nS)^{3/2}}$$

Then, the following statement holds true,

$$\max\left\{\|\bar{\mathbf{u}}_{t-1}^s - \mathbf{u}^\star\|, \|\mathbf{u}_t^s - \mathbf{u}^\star\|\right\} \ \leq \ 2D_{\mathbf{u}} + \eta G n S$$

We are ready to proceed with the

Proof of Theorem 3. Recall that the iterate $\bar{\mathbf{u}}_t^s$ is defined as the fixed point of the operator $T(\mathbf{u}) := \mathcal{P}_{\Lambda\times\Gamma}(\mathbf{u}_t^s - \eta F_{\sigma_t^s}(\mathbf{u}))$. By using the triangle inequality, we have

$$\|\mathbf{u}_t^s - \mathbf{u}^\star\|^2 \ \leq \ \|\mathbf{u}_t^s - \bar{\mathbf{u}}_{t-1}^s\|^2 + 2\|\mathbf{u}_t^s - \bar{\mathbf{u}}_{t-1}^s\|\|\bar{\mathbf{u}}_{t-1}^s - \mathbf{u}^\star\| + \|\bar{\mathbf{u}}_{t-1}^s - \mathbf{u}^\star\|^2$$

We claim that the following error criterion holds true,

$$\|\mathbf{u}_t^s - \bar{\mathbf{u}}_{t-1}^s\| = \|\mathbf{u}_{M,t-1}^s - \bar{\mathbf{u}}_{t-1}^s\| \ \leq \ \frac{D_{\mathbf{u}}}{10(nS)^{3/2}} \tag{17}$$

By Lemma 8, we have $\|\bar{\mathbf{u}}_{t-1}^s - \mathbf{u}^\star\| \leq 2D_{\mathbf{u}} + \eta G n S$. Putting these pieces together yields that

$$\|\mathbf{u}_t^s - \mathbf{u}^\star\|^2 \ \leq \ \|\bar{\mathbf{u}}_{t-1}^s - \mathbf{u}^\star\|^2 + \frac{2D_{\mathbf{u}}^2 + \eta G D_{\mathbf{u}} n S}{5(nS)^{3/2}} + \frac{D_{\mathbf{u}}^2}{100(nS)^3} \tag{18}$$

By definition of $\bar{\mathbf{u}}_t^s$, we have

$$\|\bar{\mathbf{u}}_{t-1}^s - \mathbf{u}^\star\|^2 \ \leq \ \|\mathbf{u}_{t-1}^s - \mathbf{u}^\star\|^2 - 2\eta\langle F_{\sigma_{t-1}^s}(\bar{\mathbf{u}}_{t-1}^s), \bar{\mathbf{u}}_{t-1}^s - \mathbf{u}^\star\rangle.$$

By the definition of F_i, we have

$$\|\bar{\mathbf{u}}_{t-1}^s - \mathbf{u}^\star\|^2 \;\leq\; \|\mathbf{u}_{t-1}^s - \mathbf{u}^\star\|^2 - 2\eta\left(L_{\sigma_{t-1}^s}(\bar{\lambda}_{t-1}^s, \bar{\beta}_{t-1}^s, \gamma^\star) - L_{\sigma_{t-1}^s}(\lambda^\star, \beta^\star, \bar{\gamma}_{t-1}^s)\right) \tag{19}$$

Since F_i is bounded by G for all $i \in [n]$, the function L is G-Lipschitz over $\Lambda \times \Gamma$. Thus, we have

$$\begin{aligned} & \left|(L_{\sigma_{t-1}^s}(\bar{\lambda}_{t-1}^s, \bar{\beta}_{t-1}^s, \gamma^\star) - L_{\sigma_{t-1}^s}(\lambda^\star, \beta^\star, \bar{\gamma}_{t-1}^s)) - (L_{\sigma_{t-1}^s}(\lambda_t^s, \beta_t^s, \gamma^\star) - L_{\sigma_{t-1}^s}(\lambda^\star, \beta^\star, \gamma_t^s))\right| \\ \leq \;\; & G\|\mathbf{u}_t^s - \bar{\mathbf{u}}_{t-1}^s\| \overset{\text{Eq. (17)}}{\leq} \frac{GD_{\mathbf{u}}}{10(nS)^{3/2}} \end{aligned} \tag{20}$$

Combining Eq. (18), Eq. (19) and Eq. (20) yields that

$$\begin{aligned} \|\mathbf{u}_t^s - \mathbf{u}^\star\|^2 \;\; & \leq \;\; \|\mathbf{u}_{t-1}^s - \mathbf{u}^\star\|^2 - 2\eta\left(L_{\sigma_{t-1}^s}(\lambda_t^s, \beta_t^s, \gamma^\star) - L_{\sigma_{t-1}^s}(\lambda^\star, \beta^\star, \gamma_t^s) - \frac{GD_{\mathbf{u}}}{10(nS)^{3/2}}\right) \\ & \quad + \frac{2D_{\mathbf{u}}^2 + \eta GD_{\mathbf{u}} nS}{5(nS)^{3/2}} + \frac{D_{\mathbf{u}}^2}{100(nS)^3} \end{aligned}$$

Taking the expectation of both sides and using Lemma 7, we have

$$\begin{aligned} \mathbb{E}[\|\mathbf{u}_t^s - \mathbf{u}^\star\|^2] \;\; & \leq \;\; \mathbb{E}[\|\mathbf{u}_{t-1}^s - \mathbf{u}^\star\|^2] - 2\eta\left(\mathbb{E}[\Delta(\mathbf{u}_t^s)] - 2\eta G^2 - \frac{GD_{\mathbf{u}}}{5\sqrt{nS}} - \frac{GD_{\mathbf{u}}}{10(nS)^{3/2}}\right) \\ & \quad + \frac{2D_{\mathbf{u}}^2 + \eta GD_{\mathbf{u}} nS}{5(nS)^{3/2}} + \frac{D_{\mathbf{u}}^2}{100(nS)^3} \end{aligned}$$

Summing the above inequality up over $t = 1, 2, \ldots, n$ and $s = 0, 1, \ldots, S-1$ and using the fact that $nS \geq 1$, we conclude that

$$\begin{aligned} \frac{1}{nS}\sum_{s=0}^{S-1}\sum_{t=1}^{n}\mathbb{E}[\Delta(\mathbf{u}_t^s)] \;\; & \leq \;\; \frac{\|\mathbf{u}_0^0 - \mathbf{u}^\star\|^2}{2\eta nS} + 2\eta G^2 + \frac{GD_{\mathbf{u}}}{5\sqrt{nS}} + \frac{GD_{\mathbf{u}}}{10(nS)^{3/2}} + \frac{D_{\mathbf{u}}^2}{5\eta(nS)^{3/2}} + \frac{GD_{\mathbf{u}}}{10\sqrt{nS}} + \frac{D_{\mathbf{u}}^2}{200\eta(nS)^3} \\ & \leq \;\; \frac{D_{\mathbf{u}}^2}{2\eta nS} + 2\eta G^2 + \frac{2GD_{\mathbf{u}}}{5\sqrt{nS}} + \frac{D_{\mathbf{u}}^2}{4\eta(nS)^{3/2}} \\ & \leq \;\; \frac{3D_{\mathbf{u}}^2}{4\eta nS} + 2\eta G^2 + \frac{2GD_{\mathbf{u}}}{5\sqrt{nS}} \end{aligned}$$

By the convexity of $\Delta(\cdot)$ and the definition of $\tilde{\mathbf{u}}^S$, we have

$$\mathbb{E}[\Delta(\tilde{\mathbf{u}}^S)] \;\leq\; \frac{1}{nS}\sum_{s=0}^{S-1}\sum_{t=1}^{n}\mathbb{E}[\Delta(\mathbf{u}_t^s)]$$

By the definition of η, we have

$$\begin{aligned} \frac{3D_{\mathbf{u}}^2}{4\eta nS} + 2\eta G^2 \;\; & \leq \;\; \frac{3D_{\mathbf{u}}^2}{4nS}\max\left\{2(\ell + \kappa + 1), \frac{4G^2}{\epsilon}\right\} + \frac{\epsilon}{2} \\ & \leq \;\; \frac{3D_{\mathbf{u}}^2}{4nS}\left(2(\ell + \kappa + 1) + \frac{4G^2}{\epsilon}\right) + \frac{\epsilon}{2} \\ & = \;\; \frac{2(\ell + \kappa + 1)D_{\mathbf{u}}^2}{nS} + \frac{3G^2 D_{\mathbf{u}}^2}{\epsilon nS} + \frac{\epsilon}{2} \end{aligned}$$

Putting these pieces together yields that

$$\mathbb{E}[\Delta(\tilde{\mathbf{u}}^S)] \;\leq\; \frac{2(\ell + \kappa + 1)D_{\mathbf{u}}^2}{nS} + + \frac{3G^2 D_{\mathbf{u}}^2}{\epsilon nS} + \frac{\epsilon}{2} + \frac{2GD_{\mathbf{u}}}{5\sqrt{nS}}$$

By the definition of S, we have $\mathbb{E}[\Delta(\tilde{\mathbf{u}}^S)] \leq \epsilon$. Then it remains to prove the claim for Eq. (17).

Proof of Eq. (17)**.** From the scheme of the SPPRR algorithm, we have $\mathbf{u}_{0,t-1}^s = \mathbf{u}_{t-1}^s$ and $\bar{\mathbf{u}}_{t-1}^s = \mathcal{P}_{\Lambda\times\Gamma}(\mathbf{u}_{t-1}^s - \eta F_{\sigma_{t-1}^s}(\bar{\mathbf{u}}_{t-1}^s))$. Thus, we have

$$\|\mathbf{u}_{0,t-1}^s - \bar{\mathbf{u}}_{t-1}^s\| \ \leq \ \|\mathbf{u}_{t-1}^s - (\mathbf{u}_{t-1}^s - \eta F_{\sigma_{t-1}^s}(\bar{\mathbf{u}}_{t-1}^s))\| \ \leq \ \eta G \ \leq \ \frac{\epsilon}{4G}$$

Since $0 < \eta < \frac{1}{2(\ell+\kappa+1)}$ and the operator F_i is $(\ell+\kappa+1)$-Lipschitz, we have

$$\|\mathcal{P}_{\Lambda\times\Gamma}(\mathbf{u}_{t-1}^s - \eta F_{\sigma_{t-1}^s}(\mathbf{u})) - \mathcal{P}_{\Lambda\times\Gamma}(\mathbf{u}_{t-1}^s - \eta F_{\sigma_{t-1}^s}(\mathbf{v}))\| \ \leq \ \eta\|F_{\sigma_{t-1}^s}(\mathbf{u}) - F_{\sigma_{t-1}^s}(\mathbf{v})\| \ \leq \ \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|$$

This implies that the operator $T(\cdot) = \mathcal{P}_{\Lambda\times\Gamma}(\mathbf{u}_{t-1}^s - \eta F_{\sigma_{t-1}^s}(\cdot))$ is a $\frac{1}{2}$-contraction. Then, the fixed-point iteration scheme implies that

$$\|\mathbf{u}_{M,t-1}^s - \bar{\mathbf{u}}_{t-1}^s\| \ \leq \ \frac{\|\mathbf{u}_{0,t-1}^s - \bar{\mathbf{u}}_{t-1}^s\|}{2^M} \ \leq \ \frac{\epsilon}{2^{M+2}G}$$

By the definition of M, we have $2^M \geq 10nS$. This together with the definition of S yields that

$$\|\mathbf{u}_{M,t-1}^s - \bar{\mathbf{u}}_{t-1}^s\| \ \leq \ \frac{\epsilon}{40GnS} \ \leq \ \frac{D_{\mathbf{u}}}{10(nS)^{3/2}}$$

which implies Eq. (17). □

7 Conclusions

In this work, we revisited Wasserstein DRSL through the lens of min-max optimization and demonstrated that WDRSL with generalized linear functions can be reformulated as constrained min-max optimization problems. We proposed two simple and efficient stochastic extra-gradient algorithms that leverage variance reduction and random reshuffling to exploit the smoothness and finite-sum structure of the reformulation. Our theoretical analysis guarantees that these algorithms converge to an ϵ-optimal solution at fast rates, significantly reducing computational overhead compared to standard approaches based on convex reformulations.. Our findings indicate that min-max formulations of robust optimization problems offer a promising and scalable alternative to traditional approaches, with potential applications beyond DRSL.

References

[Abadeh et al.(2015)] S. Abadeh, P. Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *NeurIPS*, pages 1576–1584, 2015.

[Alacaoglu and Malitsky(2021)] A. Alacaoglu and Y. Malitsky. Stochastic variance reduction for variational inequality methods. *ArXiv Preprint: 2102.08352*, 2021.

[Aljarrah et al.(2020)] M. A. Aljarrah, F. Famoye, and C. Lee. Generalized logistic distribution and its regression model. *Journal of Statistical Distributions and Applications*, 7(1):1–21, 2020.

[Allen-Zhu and Hazan(2016)] Z. Allen-Zhu and E. Hazan. Variance reduction for faster non-convex optimization. In *ICML*, pages 699–707. PMLR, 2016.

[Allen-Zhu and Yuan(2016)] Z. Allen-Zhu and Y. Yuan. Improved SVRG for non-strongly-convex or sum-of-non-convex objectives. In *ICML*, pages 1080–1089. PMLR, 2016.

[Bagnell(2005)] J. A. Bagnell. Robust supervised learning. In *AAAI*, volume 2, pages 714–719, 2005.

[Balamurugan and Bach(2016)] P. Balamurugan and F. Bach. Stochastic variance reduction methods for saddle-point problems. In *NeurIPS*, pages 1416–1424, 2016.

[Ben-Tal et al.(2013)] A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[Bertsekas(2009)] D. P. Bertsekas. *Convex Optimization Theory*. Athena Scientific Belmont, 2009.

[Blanchet and Murthy(2019)] J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[Blanchet et al.(2018)] J. Blanchet, K. Murthy, and F. Zhang. Optimal transport based distributionally robust optimization: Structural properties and iterative schemes. *ArXiv Preprint: 1810.02403*, 2018.

[Blanchet et al.(2019)] J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

[Bottou(2009)] L. Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. In *Proceedings of the Symposium on Learning and Data Science, Paris*, 2009.

[Carlini and Wagner(2017)] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.

[Carmon et al.(2019)] Y. Carmon, Y. Jin, A. Sidford, and K. Tian. Variance reduction for matrix games. In *NeurIPS*, page 11377–11388, 2019.

[Chambolle and Pock(2011)] 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(1):120–145, 2011.

[Chang and Lin(2011)] C-C. Chang and C-J. Lin. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

[Chavdarova et al.(2019)] T. Chavdarova, G. Gidel, F. Fleuret, and S. Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. In *NeurIPS*, pages 393–403, 2019.

[Chen et al.(2014)] Y. Chen, G. Lan, and Y. Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, 2014.

[Chen et al.(2017)] Y. Chen, G. Lan, and Y. Ouyang. Accelerated schemes for a class of variational inequalities. *Mathematical Programming*, 165(1):113–149, 2017.

[Cui and Shanbhag(2019)] S. Cui and U. V. Shanbhag. On the analysis of variance-reduced and randomized projection variants of single projection schemes for monotone stochastic variational inequality problems. *ArXiv Preprint: 1904.11076*, 2019.

[Daskalakis et al.(2018)] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training GANs with optimism. In *ICLR*, 2018. URL `https://openreview.net/forum?id=SJJySbbAZ`.

[Delage and Ye(2010)] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[Dobson and Barnett(2018)] A. J. Dobson and A. G. Barnett. *An Introduction to Generalized Linear Models*. CRC Press, 2018.

[Du and Hu(2019)] S. S. Du and W. Hu. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. In *AISTATS*, pages 196–205. PMLR, 2019.

[Esfahani and Kuhn(2018)] P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

[Fang et al.(2018)] C. Fang, C. J. Li, Z. Lin, and T. Zhang. Near-optimal non-convex optimization via stochastic path integrated differential estimator. *NeurIPS*, 31:689, 2018.

[Gao et al.(2017)] R. Gao, X. Chen, and A. J. Kleywegt. Distributional robustness and regularization in statistical learning. *ArXiv Preprint: 1712.06050*, 2017.

[Gürbüzbalaban et al.(2019)] M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, pages 1–36, 2019.

[Haochen and Sra(2019)] J. Haochen and S. Sra. Random shuffling beats SGD after finite epochs. In *ICML*, pages 2624–2633. PMLR, 2019.

[Hardin and Hilbe(2007)] J. W. Hardin and J. M. Hilbe. *Generalized Linear Models and Extensions*. Stata Press, 2007.

[Hastie et al.(2009)] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2009.

[Hsieh et al.(2019)] Y. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS*, pages 6938–6948, 2019.

[Hu et al.(2018)] W. Hu, G. Niu, I. Sato, and M. Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *ICML*, pages 2029–2037. PMLR, 2018.

[Iusem et al.(2017)] A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2):686–724, 2017.

[Johnson and Zhang(2013)] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NeurIPS*, pages 315–323, 2013.

[Juditsky et al.(2011)] A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[Korpelevich(1976)] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[Kotsalis et al.(2020)] G. Kotsalis, G. Lan, and T. Li. Simple and optimal methods for stochastic variational inequalities, i: operator extrapolation. *ArXiv Preprint: 2011.02987*, 2020.

[Lee and Mehrotra(2015)] C. Lee and S. Mehrotra. A distributionally-robust approach for finding support vector machines. *Manuscript, available at http://www. optimization-online. org/DB_HTML/2015/06/4965. html*, 2015.

[Li et al.(2019)] J. Li, S. Huang, and A. M-C. So. A first-order algorithmic framework for distributionally robust logistic regression. In *NeurIPS*, pages 3937–3947, 2019.

[Li et al.(2020)] J. Li, C. Chen, and A. M-C. So. Fast epigraphical projection-based incremental algorithms for Wasserstein distributionally robust support vector machine. In *NeurIPS*, page To appear, 2020.

[Liang and Stokes(2019)] T. Liang and J. Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *AISTATS*, pages 907–915. PMLR, 2019.

[Lin et al.(2020)] T. Lin, C. Jin, and M. I. Jordan. Near-optimal algorithms for minimax optimization. In *COLT*, pages 2738–2779. PMLR, 2020.

[Liu et al.(2017)] Y. Liu, X. Yuan, S. Zeng, and J. Zhang. Primal-dual hybrid gradient method for distributionally robust optimization problems. *Operations Research Letters*, 45(6):625–630, 2017.

[Luo and Mehrotra(2019)] F. Luo and S. Mehrotra. Decomposition algorithm for distributionally robust optimization using Wasserstein metric with an application to a class of regression models. *European Journal of Operational Research*, 278(1):20–35, 2019.

[Madry et al.(2018)] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[Mazumdar et al.(2020)] E. Mazumdar, L. J. Ratliff, and S. S. Sastry. On gradient-based learning in continuous games. *SIAM Journal on Mathematics of Data Science*, 2(1):103–131, 2020. doi: 10.1137/18M1231298.

[Mazumdar et al.(2019)] E. V. Mazumdar, M. I. Jordan, and S. S. Sastry. On finding local Nash equilibria (and only local Nash equilibria) in zero-sum games. *ArXiv Preprint: 1901.00838*, 2019.

[Mishchenko et al.(2020)] K. Mishchenko, A. Khaled, and P. Richtárik. Random reshuffling: Simple analysis with vast improvements. In *NeurIPS*, volume To appear, 2020.

[Mokhtari et al.(2020a)] A. Mokhtari, A. Ozdaglar, and S. Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *AISTATS*, pages 1497–1507. PMLR, 2020a.

[Mokhtari et al.(2020b)] A. Mokhtari, A. E. Ozdaglar, and S. Pattathil. Convergence rate of o(1/k) for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020b.

[Nagaraj et al.(2019)] D. Nagaraj, P. Jain, and P. Netrapalli. SGD without replacement: Sharper rates for general smooth convex functions. In *ICML*, pages 4703–4711. PMLR, 2019.

[Namkoong and Duchi(2016)] H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *NeurIPS*, pages 2216–2224, 2016.

[Namkoong and Duchi(2017)] H. Namkoong and J. C. Duchi. Variance-based regularization with convex objectives. In *NeurIPS*, pages 2971–2980, 2017.

[Nemirovski(2004)] A. Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[Nemirovski et al.(2009)] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[Nesterov(2007)] Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.

[Nguyen et al.(2020)] L. M. Nguyen, Q. Tran-Dinh, D. T. Phan, P. H. Nguyen, and M. van Dijk. A unified convergence analysis for shuffling-type gradient methods. *ArXiv Preprint: 2002.08246*, 2020.

[Quionero-Candela et al.(2009)] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

[Rahimian and Mehrotra(2019)] H. Rahimian and S. Mehrotra. Distributionally robust optimization: A review. *ArXiv Preprint: 1908.05659*, 2019.

[Rajput et al.(2020)] S. Rajput, A. Gupta, and D. Papailiopoulos. Closing the convergence gap of SGD without replacement. In *ICML*, pages 7964–7973. PMLR, 2020.

[Recht and Ré(2013)] B. Recht and C. Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.

[Reddi et al.(2016)] S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola. Stochastic variance reduction for nonconvex optimization. In *ICML*, pages 314–323. PMLR, 2016.

[Rockafellar(2015)] R. T. Rockafellar. *Convex Analysis*, volume 36. Princeton University Press, 2015.

[Safran and Shamir(2020)] I. Safran and O. Shamir. How good is SGD with random shuffling? In *COLT*, pages 3250–3284, 2020.

[Shafieezadeh-Abadeh et al.(2019)] S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

[Shamir(2016)] O. Shamir. Without-replacement sampling for stochastic gradient methods. In *NeurIPS*, volume 29, pages 46–54, 2016.

[Shapiro(2001)] A. Shapiro. On duality theory of conic linear problems. In *Semi-Infinite Programming*, pages 135–165. Springer, 2001.

[Shi et al.(2017)] Z. Shi, X. Zhang, and Y. Yu. Bregman divergence for stochastic variance reduction: saddle-point and adversarial prediction. In *NeurIPS*, pages 6031–6041, 2017.

[Sinha et al.(2018)] A. Sinha, H. Namkoong, and J. Duchi. Certifiable distributional robustness with principled adversarial training. In *ICLR*, 2018. URL `https://openreview.net/forum?id=Hk6kPgZA-`.

[Sion(1958)] M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

[Stukel(1988)] T. A. Stukel. Generalized logistic models. *Journal of the American Statistical Association*, 83(402):426–431, 1988.

[Szegedy et al.(2013)] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *ArXiv Preprint: 1312.6199*, 2013.

[Thekumparampil et al.(2019)] K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh. Efficient algorithms for smooth minimax optimization. In *NeurIPS*, pages 12680–12691, 2019.

[Vapnik(2013)] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Science & Business Media, 2013.

[Villani(2008)] C. Villani. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

[Wiesemann et al.(2014)] W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

[Xie et al.(2020)] G. Xie, L. Luo, Y. Lian, and Z. Zhang. Lower complexity bounds for finite-sum convex-concave minimax optimization problems. In *ICML*, pages 10504–10513. PMLR, 2020.

[Yan et al.(2020)] Y. Yan, Y. Xu, Q. Lin, W. Liu, and T. Yang. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. In *NeurIPS*, page To appear, 2020.

[Yang et al.(2020)] J. Yang, S. Zhang, N. Kiyavash, and N. He. A catalyst framework for minimax optimization. In *NeurIPS*, page To appear, 2020.

[Zhang et al.(2019)] Z. Zhang, S. Ahmed, and G. Lan. Efficient algorithms for distributionally robust stochastic optimization with discrete scenario support. *ArXiv Preprint: 1909.11216*, 2019.

[Zhou et al.(2018)] D. Zhou, P. Xu, and Q. Gu. Stochastic nested variance reduction for nonconvex optimization. *NeurIPS*, 31:3921–3932, 2018.

A Proof of Auxiliary Lemmas

A.1 Proof of Lemma 1

Proof of Lemma 1. Note that $\Psi(\langle \mathbf{x}, \beta \rangle) - \widehat{y}\langle \mathbf{x}, \beta \rangle - \lambda \|\widehat{\mathbf{x}} - \mathbf{x}\|$ is the difference of two different convex functions in $\mathbf{x}$. In order to maximize this function, we define its convex part as $h_\beta(\mathbf{x}) = \Psi(\langle \mathbf{x}, \beta \rangle) - \widehat{y}\langle \mathbf{x}, \beta \rangle$ and reformulate it using the conjugate function.

We first consider one-dimensional function $f(t) := \Psi(t) - \widehat{y}t$. Since Ψ is convex, the conjugate function of f is well defined and expressed by $f^\star(\theta) = \sup_{t \in \mathbb{R}}\{\theta t - \Psi(t) + \widehat{y}t\}$. Then, we claim that $f^\star(\theta) = +\infty$ when $|\theta| > L + 1$. Since Ψ is L-Lipschitz (cf. Assumption 1), we have

$$|\Psi(t) - \Psi(0)| = \left| \int_0^t \Psi'(s)\, ds \right| \le L|t| \implies \Psi(t) \le \Psi(0) + L|t|$$

This implies that $f^*(\theta) \ge \sup_{t \in \mathbb{R}}\{\theta t - L|t| - \Psi(0) + \widehat{y}t\}$. Since $\widehat{y} \in \{-1, 1\}$, we have

$$\theta t - L|t| - \Psi(0) + \widehat{y}t \longrightarrow \begin{cases} +\infty \text{ as } t \to +\infty, & \text{if } \theta > L + 1 \\ +\infty \text{ as } t \to -\infty, & \text{if } \theta < -L - 1 \end{cases}$$

Putting these pieces together yields the desired result. Moreover, the conjugate of $h_\beta(\mathbf{x}) = \Psi(\langle \mathbf{x}, \beta \rangle) - \widehat{y}\langle \mathbf{x}, \beta \rangle$ is therefore given by

$$h_\beta^\star(\xi) = \begin{cases} \inf\limits_{|\theta| \le L+1} f^\star(\theta), & \text{if } \xi = \theta\beta \\ +\infty, & \text{otherwise} \end{cases}$$

Since $h_\beta(\mathbf{x})$ is convex and continuous, it is equivalent to its bi-conjugate function [Rockafellar(2015)]. This implies the statement on the representation of the function h_β as follows,

$$h_\beta(\mathbf{x}) = \sup_{\xi \in \mathbb{R}^d} \langle \xi, \mathbf{x} \rangle - h_\beta^\star(\xi) = \sup_{|\theta| \le L+1} \langle \theta\beta, \mathbf{x} \rangle - f^\star(\theta) \tag{21}$$

The remaining steps are the same as that in [Abadeh et al.(2015), Lemma 1]. More specifically, Eq. (21) implies that h_β can be represented as the upper envelope of infinitely many linear functions. Using this representation, we have

$$\begin{aligned} \sup_{\mathbf{x} \in \mathbb{R}^d} \Psi(\langle \mathbf{x}, \beta \rangle) - \widehat{y}\langle \mathbf{x}, \beta \rangle - \lambda \|\widehat{\mathbf{x}} - \mathbf{x}\| &= \sup_{\mathbf{x} \in \mathbb{R}^d} \sup_{|\theta| \le L+1} \langle \theta\beta, \mathbf{x} \rangle - f^\star(\theta) - \lambda \|\widehat{\mathbf{x}} - \mathbf{x}\| \\ &= \sup_{\mathbf{x} \in \mathbb{R}^d} \sup_{|\theta| \le L+1} \langle \theta\beta, \mathbf{x} \rangle - f^\star(\theta) - \sup_{\|q\| \le \lambda} \langle q, \widehat{\mathbf{x}} - \mathbf{x} \rangle \\ &= \sup_{\mathbf{x} \in \mathbb{R}^d} \sup_{|\theta| \le L+1} \inf_{\|q\| \le \lambda} \langle \theta\beta, \mathbf{x} \rangle - f^\star(\theta) - \langle q, \widehat{\mathbf{x}} - \mathbf{x} \rangle \end{aligned}$$

By [Bertsekas(2009), Proposition 5.5.4], we have

$$\begin{aligned} \sup_{\mathbf{x} \in \mathbb{R}^d} \{\Psi(\langle \mathbf{x}, \beta \rangle) - \widehat{y}\langle \mathbf{x}, \beta \rangle - \lambda \|\widehat{\mathbf{x}} - \mathbf{x}\|\} &= \sup_{|\theta| \le L+1} \inf_{\|q\| \le \lambda} \sup_{\mathbf{x} \in \mathbb{R}^d} \langle \theta\beta, \mathbf{x} \rangle - f^\star(\theta) - \langle q, \widehat{\mathbf{x}} - \mathbf{x} \rangle \\ &= \sup_{|\theta| \le L+1} \inf_{\|q\| \le \lambda} \sup_{\mathbf{x} \in \mathbb{R}^d} \langle \theta\beta + q, \mathbf{x} \rangle - f^\star(\theta) - \langle q, \widehat{\mathbf{x}} \rangle \end{aligned}$$

Explicitly evaluating the maximization over $\mathbf{x} \in \mathbb{R}^d$, we have

$$\begin{aligned} \sup_{\mathbf{x} \in \mathbb{R}^d} \{\Psi(\langle \mathbf{x}, \beta \rangle) - \widehat{y}\langle \mathbf{x}, \beta \rangle - \lambda \|\widehat{\mathbf{x}} - \mathbf{x}\|\} &= \begin{cases} \sup\limits_{|\theta| \le L+1} \inf\limits_{\|q\| \le \lambda} & -f^\star(\theta) - \langle q, \widehat{\mathbf{x}} \rangle \\ \quad \text{s.t.} & \theta\beta + q = 0. \end{cases} \\ &= \begin{cases} \sup\limits_{|\theta| \le L+1} -f^\star(\theta) + \langle \theta\beta, \widehat{\mathbf{x}} \rangle, & \text{if } \sup\limits_{|\theta| \le L+1} \|\theta\beta\| \le \lambda \\ +\infty, & \text{otherwise.} \end{cases} \end{aligned}$$

In addition, we have

$$\begin{aligned}
\sup_{|\theta|\leq L+1} -f^\star(\theta) + \langle \theta\beta, \widehat{\mathbf{x}}\rangle &= \Psi(\langle \widehat{\mathbf{x}}, \beta\rangle) - \widehat{y}\langle \widehat{\mathbf{x}}, \beta\rangle \\
\sup_{|\theta|\leq L+1} \|\theta\beta\| \leq \lambda &\Leftrightarrow \|\beta\| \leq \frac{\lambda}{L+1}
\end{aligned}$$

Putting these pieces together yields the desired claim. □

A.2 Proof of Lemma 2

Proof of Lemma 2. Since the indices i_t and j_t are both uniformly sampled from the set $[n]$, the gradient estimator $\mathbf{g}_t^s$ and $\bar{\mathbf{g}}_t^s$ are both unbiased, meaning that

$$\mathbb{E}[\mathbf{g}_t^s \mid \mathbf{u}_t^s, \tilde{\mathbf{u}}^s] = F(\mathbf{u}_t^s), \qquad \mathbb{E}[\bar{\mathbf{g}}_t^s \mid \bar{\mathbf{u}}_{t+1}^s, \tilde{\mathbf{u}}^s] = F(\bar{\mathbf{u}}_{t+1}^s)$$

Since $\mathbb{E}[\|\xi - \mathbb{E}[\xi]\|^2] \leq \mathbb{E}[\|\xi\|^2]$ holds true for any random variable $\xi \in \mathbb{R}^d$, we have

$$\begin{aligned}
\mathbb{E}[\|\mathbf{g}_t^s - F(\mathbf{u}_t^s)\|^2 \mid \mathbf{u}_t^s, \tilde{\mathbf{u}}^s] &= \mathbb{E}[\|(F_{i_t}(\mathbf{u}_t^s) - F_{i_t}(\tilde{\mathbf{u}}^s)) - (F(\mathbf{u}_t^s) - F(\tilde{\mathbf{u}}^s))\|^2 \mid \mathbf{u}_t^s, \tilde{\mathbf{u}}^s] \\
&\leq \mathbb{E}[\|F_{i_t}(\mathbf{u}_t^s) - F_{i_t}(\tilde{\mathbf{u}}^s)\|^2 \mid \mathbf{u}_t^s, \tilde{\mathbf{u}}^s]
\end{aligned}$$

Note that Assumption 1 implies that Ψ' is ℓ-Lipschitz and $\|\mathbf{x}_i\| \leq 1$ for all $i \in [n]$. By the definition of the operator F_i, it is easy to verify that $\|F_i(\mathbf{u}) - F_i(\mathbf{u}')\| \leq (\ell + \kappa + 1)\|\mathbf{u} - \mathbf{u}'\|$. Thus, we have

$$\mathbb{E}[\|\mathbf{g}_t^s - F(\mathbf{u}_t^s)\|^2 \mid \mathbf{u}_t^s, \tilde{\mathbf{u}}^s] \leq (\ell + \kappa + 1)^2\|\mathbf{u}_t^s - \tilde{\mathbf{u}}^s\|^2$$

In addition, it follows from the Cauchy-Schwarz inequality that $\|\mathbf{u}_t^s - \tilde{\mathbf{u}}^s\|^2 \leq 2\|\mathbf{u}_t^s - \mathbf{u}^\star\|^2 + 2\|\tilde{\mathbf{u}}^s - \mathbf{u}^\star\|^2$. Putting these pieces together yields the first desired inequality. The remaining step is to prove the second desired inequality. By the similar argument, we have

$$\begin{aligned}
\mathbb{E}[\|\bar{\mathbf{g}}_t^s - F(\bar{\mathbf{u}}_{t+1}^s)\|^2 \mid \bar{\mathbf{u}}_{t+1}^s, \tilde{\mathbf{u}}^s] &= \mathbb{E}[\|(F_{j_t}(\bar{\mathbf{u}}_{t+1}^s) - F_{j_t}(\tilde{\mathbf{u}}^s)) - (F(\bar{\mathbf{u}}_{t+1}^s) - F(\tilde{\mathbf{u}}^s))\|^2 \mid \bar{\mathbf{u}}_{t+1}^s, \tilde{\mathbf{u}}^s] \\
&\leq \mathbb{E}[\|F_{j_t}(\bar{\mathbf{u}}_{t+1}^s) - F_{j_t}(\tilde{\mathbf{u}}^s)\|^2 \mid \bar{\mathbf{u}}_{t+1}^s, \tilde{\mathbf{u}}^s] \\
&\leq 2\mathbb{E}[\|F_{j_t}(\bar{\mathbf{u}}_{t+1}^s) - F_{j_t}(\mathbf{u}^\star)\|^2 \mid \bar{\mathbf{u}}_{t+1}^s, \tilde{\mathbf{u}}^s] + 2\mathbb{E}[\|F_{j_t}(\tilde{\mathbf{u}}^s) - F_{i_t}(\mathbf{u}^\star)\|^2 \mid \bar{\mathbf{u}}_{t+1}^s, \tilde{\mathbf{u}}^s] \quad (22)
\end{aligned}$$

Again, by the definition of the functions L_i and F_i, it is clear that L_i is $(\ell + \kappa)$-smooth and thus $\|F_i(\mathbf{u}) - F_i(\mathbf{u}')\| \leq (\ell + \kappa + 1)\|\mathbf{u} - \mathbf{u}'\|$. We define a regularized version of L_i by φ_i over the constraint set $\Lambda \times \Gamma$ as follows,

$$\varphi_i(\mathbf{u}) = L_i(\mathbf{u}) - \nabla L_i(\mathbf{u}^\star)^\top(\mathbf{u} - \mathbf{u}^\star) + \frac{\ell + \kappa + 1}{2}\left(\|\lambda - \lambda^\star\|^2 + \|\beta - \beta^\star\|^2\right) - \frac{\ell + \kappa + 1}{2}\|\gamma - \gamma^\star\|^2$$

This function φ_i is strongly convex-concave with the module $(\ell + \kappa + 1)/2$ and $2(\ell + \kappa + 1)$-smooth, and the unique optimal saddle point is $\mathbf{u}^\star$. This implies that

$$\begin{aligned}
\|\nabla\varphi_i(\mathbf{u})\|^2 &= \left\|\begin{pmatrix}\nabla_{(\lambda,\beta)}\varphi_i(\mathbf{u}) \\ -\nabla_\gamma\varphi_i(\mathbf{u})\end{pmatrix} - \begin{pmatrix}\nabla_{(\lambda,\beta)}\varphi_i(\mathbf{u}^\star) \\ -\nabla_\gamma\varphi_i(\mathbf{u}^\star)\end{pmatrix}\right\|^2 \leq 4(\ell + \kappa + 1)^2\|\mathbf{u} - \mathbf{u}^\star\|^2 \\
&\leq 8(\ell + \kappa + 1)\left(\varphi_i(\lambda, \beta, \gamma^\star) - \varphi_i(\lambda^\star, \beta^\star, \gamma)\right)
\end{aligned}$$

By the definition of φ_i and F_i, we have

$$\varphi_i(\lambda, \beta, \gamma^\star) - \varphi_i(\lambda^\star, \beta^\star, \gamma) = L_i(\lambda, \beta, \gamma^\star) - L_i(\lambda^\star, \beta^\star, \gamma) - F_i(\mathbf{u}^\star)^\top(\mathbf{u} - \mathbf{u}^\star) + \frac{\ell + \kappa + 1}{2}\|\mathbf{u} - \mathbf{u}^\star\|^2$$

Furthermore, we have

$$\nabla\varphi_i(\mathbf{u}) = \nabla L_i(\mathbf{u}) - \nabla L_i(\mathbf{u}^\star) + (\ell+\kappa+1)\begin{pmatrix} \lambda-\lambda^\star \\ \beta-\beta^\star \\ -(\gamma-\gamma^\star)\end{pmatrix}$$

By the Cauchy-Schwarz inequality, we have

$$\|F_i(\mathbf{u}) - F_i(\mathbf{u}^\star)\|^2 = \|\nabla L_i(\mathbf{u}) - \nabla L_i(\mathbf{u}^\star)\|^2 \leq 2\|\nabla\varphi_i(\mathbf{u})\|^2 + 2(\ell+\kappa+1)^2\|\mathbf{u}-\mathbf{u}^\star\|^2$$

Putting these pieces together with the fact that j_t is uniformly sampled from the set $[n]$ yields that

$$\begin{aligned}
& \mathbb{E}[\|F_{j_t}(\bar{\mathbf{u}}_{t+1}^s) - F_{j_t}(\mathbf{u}^\star)\|^2 \mid \bar{\mathbf{u}}_{t+1}^s, \tilde{\mathbf{u}}^s] \\
\leq \;& 16(\ell+\kappa+1)\mathbb{E}\left[L_{j_t}(\bar{\lambda}_{t+1}^s, \bar{\beta}_{t+1}^s, \gamma^\star) - L_{j_t}(\lambda^\star, \beta^\star, \bar{\gamma}_{t+1}^s) - F_{j_t}(\mathbf{u}^\star)^\top(\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}^\star) \mid \bar{\mathbf{u}}_{t+1}^s, \tilde{\mathbf{u}}^s\right] \\
& +2(\ell+\kappa+1)^2\|\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}^\star\|^2 \\
= \;& 16(\ell+\kappa+1)\left(\Delta(\bar{\mathbf{u}}_{t+1}^s) - F(\mathbf{u}^\star)^\top(\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}^\star)\right) + 2(\ell+\kappa+1)^2\|\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}^\star\|^2
\end{aligned}$$

Since $\mathbf{u}^\star$ is an optimal saddle point of the smooth minimax optimization model in Eq. (4), we have $F(\mathbf{u}^\star)^\top(\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}^\star) \geq 0$. Therefore, we conclude that

$$\mathbb{E}[\|F_{j_t}(\bar{\mathbf{u}}_{t+1}^s) - F_{j_t}(\mathbf{u}^\star)\|^2 \mid \bar{\mathbf{u}}_{t+1}^s, \tilde{\mathbf{u}}^s] \leq 16(\ell+\kappa+1)\Delta(\bar{\mathbf{u}}_{t+1}^s) + 2(\ell+\kappa+1)^2\|\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}^\star\|^2 \tag{23}$$

Similarly, we have

$$\mathbb{E}[\|F_{j_t}(\tilde{\mathbf{u}}^s) - F_{j_t}(\mathbf{u}^\star)\|^2 \mid \mathbf{u}_t^s, \tilde{\mathbf{u}}^s] \leq 16(\ell+\kappa+1)\Delta(\tilde{\mathbf{u}}^s) + 2(\ell+\kappa+1)^2\|\tilde{\mathbf{u}}^s - \mathbf{u}^\star\|^2 \tag{24}$$

Plugging Eq. (23) and Eq. (24) into Eq. (22) yields that

$$\mathbb{E}[\|\bar{\mathbf{g}}_t^s - F(\bar{\mathbf{u}}_{t+1}^s)\|^2 \mid \bar{\mathbf{u}}_{t+1}^s, \tilde{\mathbf{u}}^s] \;\leq\; 16(\ell+\kappa+1)\left(\Delta(\bar{\mathbf{u}}_{t+1}^s) + \Delta(\tilde{\mathbf{u}}^s)\right) + 2(\ell+\kappa+1)^2\left(\|\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}^\star\|^2 + \|\tilde{\mathbf{u}}^s - \mathbf{u}^\star\|^2\right)$$

This completes the proof. □

A.3 Proof of Lemma 3

Proof of Lemma 3. By the update formula for the iterates $\bar{\mathbf{u}}_{t+1}^s$ and $\mathbf{u}_{t+1}^s$, we have

$$\begin{aligned}
0 &\leq (\mathbf{u} - \bar{\mathbf{u}}_{t+1}^s)^\top\left(\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}_t^s + \eta_{t+1}^s\mathbf{g}_t^s\right), \quad \forall \mathbf{u} \in \Lambda\times\Gamma \\
0 &\leq (\mathbf{u} - \mathbf{u}_{t+1}^s)^\top\left(\mathbf{u}_{t+1}^s - \mathbf{u}_t^s + \eta_{t+1}^s\bar{\mathbf{g}}_t^s\right), \quad \forall \mathbf{u} \in \Lambda\times\Gamma
\end{aligned}$$

Letting $\mathbf{u} = \mathbf{u}_{t+1}^s$ in the first inequality and $\mathbf{u} = \mathbf{u}^\star$ in the second inequality and rearranging the resulting inequalities yields that

$$\begin{aligned}
0 &\leq \frac{1}{2\eta_{t+1}^s}\left(\|\mathbf{u}_t^s - \mathbf{u}_{t+1}^s\|^2 - \|\mathbf{u}_{t+1}^s - \bar{\mathbf{u}}_{t+1}^s\|^2 - \|\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}_t^s\|^2\right) + (\mathbf{u}_{t+1}^s - \bar{\mathbf{u}}_{t+1}^s)^\top\mathbf{g}_t^s \\
0 &\leq \frac{1}{2\eta_{t+1}^s}\left(\|\mathbf{u}_t^s - \mathbf{u}^\star\|^2 - \|\mathbf{u}_{t+1}^s - \mathbf{u}^\star\|^2 - \|\mathbf{u}_{t+1}^s - \mathbf{u}_t^s\|^2\right) + (\mathbf{u}^\star - \mathbf{u}_{t+1}^s)^\top\bar{\mathbf{g}}_t^s
\end{aligned}$$

Summing up the above two inequalities yields that

$$
\begin{aligned}
0 \ \leq \ & \frac{1}{2\eta_{t+1}^s}\left(\|\mathbf{u}_t^s-\mathbf{u}^\star\|^2-\|\mathbf{u}_{t+1}^s-\mathbf{u}^\star\|^2-\|\mathbf{u}_{t+1}^s-\bar{\mathbf{u}}_{t+1}^s\|^2-\|\bar{\mathbf{u}}_{t+1}^s-\mathbf{u}_t^s\|^2\right) \\
& +(\mathbf{u}_{t+1}^s-\bar{\mathbf{u}}_{t+1}^s)^\top\mathbf{g}_t^s+(\mathbf{u}^\star-\mathbf{u}_{t+1}^s)^\top\bar{\mathbf{g}}_t^s \\
= \ & \frac{1}{2\eta_{t+1}^s}\left(\|\mathbf{u}_t^s-\mathbf{u}^\star\|^2-\|\mathbf{u}_{t+1}^s-\mathbf{u}^\star\|^2-\|\mathbf{u}_{t+1}^s-\bar{\mathbf{u}}_{t+1}^s\|^2-\|\bar{\mathbf{u}}_{t+1}^s-\mathbf{u}_t^s\|^2\right) \\
& +(\mathbf{u}_{t+1}^s-\bar{\mathbf{u}}_{t+1}^s)^\top\left(\mathbf{g}_t^s-\bar{\mathbf{g}}_t^s\right)+(\mathbf{u}^\star-\bar{\mathbf{u}}_{t+1}^s)^\top\bar{\mathbf{g}}_t^s
\end{aligned}
\tag{25}
$$

Using the Young's inequality, we have

$$
(\mathbf{u}_{t+1}^s-\bar{\mathbf{u}}_{t+1}^s)^\top\left(\mathbf{g}_t^s-\bar{\mathbf{g}}_t^s\right) \ \leq \ \frac{\|\mathbf{u}_{t+1}^s-\bar{\mathbf{u}}_{t+1}^s\|^2}{2\eta_{t+1}^s}+\frac{\eta_{t+1}^s\|\mathbf{g}_t^s-\bar{\mathbf{g}}_t^s\|^2}{2}
\tag{26}
$$

Using the Cauchy-Schwarz inequality and the fact that F is $(\ell+\kappa+1)$-Lipschitz, we have

$$
\begin{aligned}
\|\mathbf{g}_t^s-\bar{\mathbf{g}}_t^s\|^2 \ & \leq \ 3\|\mathbf{g}_t^s-F(\mathbf{u}_t^s)\|^2+3\|F(\mathbf{u}_t^s)-F(\bar{\mathbf{u}}_{t+1}^s)\|^2+3\|\bar{\mathbf{g}}_t^s-F(\bar{\mathbf{u}}_{t+1}^s)\|^2 \\
& \leq \ 3\|\mathbf{g}_t^s-F(\mathbf{u}_t^s)\|^2+3\|\bar{\mathbf{g}}_t^s-F(\bar{\mathbf{u}}_{t+1}^s)\|^2+3(\ell+\kappa+1)^2\|\bar{\mathbf{u}}_{t+1}^s-\mathbf{u}_t^s\|^2
\end{aligned}
\tag{27}
$$

Combining Eq. (25), Eq. (26) and Eq. (27) yields that

$$
\begin{aligned}
(\bar{\mathbf{u}}_{t+1}^s-\mathbf{u}^\star)^\top\bar{\mathbf{g}}_t^s \ \leq \ & \frac{1}{2\eta_{t+1}^s}\left(\|\mathbf{u}_t^s-\mathbf{u}^\star\|^2-\|\mathbf{u}_{t+1}^s-\mathbf{u}^\star\|^2\right)-\left(\frac{1}{2\eta_{t+1}^s}-\frac{3\eta_{t+1}^s(\ell+\kappa+1)^2}{2}\right)\|\bar{\mathbf{u}}_{t+1}^s-\mathbf{u}_t^s\|^2 \\
& +\frac{3\eta_{t+1}^s}{2}\left(\|\mathbf{g}_t^s-F(\mathbf{u}_t^s)\|^2+\|\bar{\mathbf{g}}_t^s-F(\bar{\mathbf{u}}_{t+1}^s)\|^2\right)
\end{aligned}
$$

Taking the expectation of both sides of the above inequalities conditioned on $(\bar{\mathbf{u}}_{t+1}^s, \tilde{\mathbf{u}}^s)$, we have

$$
\begin{aligned}
& (\bar{\mathbf{u}}_{t+1}^s-\mathbf{u}^\star)^\top F(\bar{\mathbf{u}}_{t+1}^s) \\
\leq \ & \frac{1}{2\eta_{t+1}^s}\left(\|\mathbf{u}_t^s-\mathbf{u}^\star\|^2-\mathbb{E}[\|\mathbf{u}_{t+1}^s-\mathbf{u}^\star\|^2 \mid \bar{\mathbf{u}}_{t+1}^s,\tilde{\mathbf{u}}^s]\right)-\left(\frac{1}{2\eta_{t+1}^s}-\frac{3\eta_{t+1}^s(\ell+\kappa+1)^2}{2}\right)\|\bar{\mathbf{u}}_{t+1}^s-\mathbf{u}_t^s\|^2 \\
& +\frac{3\eta_{t+1}^s}{2}\left(\|\mathbf{g}_t^s-F(\mathbf{u}_t^s)\|^2+\mathbb{E}[\|\bar{\mathbf{g}}_t^s-F(\bar{\mathbf{u}}_{t+1}^s)\|^2 \mid \bar{\mathbf{u}}_{t+1}^s,\tilde{\mathbf{u}}^s]\right)
\end{aligned}
$$

Plugging the second inequality of Lemma 2 into the above inequality, we have

$$
\begin{aligned}
& (\bar{\mathbf{u}}_{t+1}^s-\mathbf{u}^\star)^\top F(\bar{\mathbf{u}}_{t+1}^s) \\
\leq \ & \frac{1}{2\eta_{t+1}^s}\left(\|\mathbf{u}_t^s-\mathbf{u}^\star\|^2-\mathbb{E}[\|\mathbf{u}_{t+1}^s-\mathbf{u}^\star\|^2 \mid \bar{\mathbf{u}}_{t+1}^s,\tilde{\mathbf{u}}^s]\right)-\left(\frac{1}{2\eta_{t+1}^s}-\frac{3\eta_{t+1}^s(\ell+\kappa+1)^2}{2}\right)\|\bar{\mathbf{u}}_{t+1}^s-\mathbf{u}_t^s\|^2 \\
& +\frac{3\eta_{t+1}^s}{2}\|\mathbf{g}_t^s-F(\mathbf{u}_t^s)\|^2+24\eta_{t+1}^s(\ell+\kappa+1)\left(\Delta(\bar{\mathbf{u}}_{t+1}^s)+\Delta(\tilde{\mathbf{u}}^s)\right)+3\eta_{t+1}^s(\ell+\kappa+1)^2\|\tilde{\mathbf{u}}^s-\mathbf{u}^\star\|^2 \\
& +3\eta_{t+1}^s(\ell+\kappa+1)^2\|\bar{\mathbf{u}}_{t+1}^s-\mathbf{u}^\star\|^2
\end{aligned}
$$

By the definition of F and the fact that L is a convex-concave function, we have

$$
(\bar{\mathbf{u}}_{t+1}^s-\mathbf{u}^\star)^\top F(\bar{\mathbf{u}}_{t+1}^s) \ \geq \ \Delta(\bar{\mathbf{u}}_{t+1}^s)
$$

Putting these pieces together yields that

$$\begin{aligned}
& \left(1 - 24\eta_{t+1}^s(\ell+\kappa+1)\right)\Delta(\bar{\mathbf{u}}_{t+1}^s) \qquad (28)\\
\leq \;& \frac{1}{2\eta_{t+1}^s}\left(\|\mathbf{u}_t^s - \mathbf{u}^\star\|^2 - \mathbb{E}[\|\mathbf{u}_{t+1}^s - \mathbf{u}^\star\|^2 \mid \bar{\mathbf{u}}_{t+1}^s, \tilde{\mathbf{u}}^s]\right) + \frac{3\eta_{t+1}^s}{2}\|\mathbf{g}_t^s - F(\mathbf{u}_t^s)\|^2 + 24\eta_{t+1}^s(\ell+\kappa+1)\Delta(\tilde{\mathbf{u}}^s) \\
& - \left(\frac{1}{2\eta_{t+1}^s} - \frac{3\eta_{t+1}^s(\ell+\kappa+1)^2}{2}\right)\|\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}_t^s\|^2 + 3\eta_{t+1}^s(\ell+\kappa+1)^2\|\tilde{\mathbf{u}}^s - \mathbf{u}^\star\|^2 \\
& + 3\eta_{t+1}^s(\ell+\kappa+1)^2\|\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}^\star\|^2
\end{aligned}$$

Taking the expectation of both sides of Eq. (28) conditioned on $(\mathbf{u}_t^s, \tilde{\mathbf{u}}^s)$ together with the tower property and the first inequality of Lemma 2, we have

$$\begin{aligned}
& \left(1 - 24\eta_{t+1}^s(\ell+\kappa+1)\right)\mathbb{E}[\Delta(\bar{\mathbf{u}}_{t+1}^s) \mid \mathbf{u}_t^s, \tilde{\mathbf{u}}^s] \\
\leq \;& \frac{1}{2\eta_{t+1}^s}\left(\|\mathbf{u}_t^s - \mathbf{u}^\star\|^2 - \mathbb{E}[\|\mathbf{u}_{t+1}^s - \mathbf{u}^\star\|^2 \mid \mathbf{u}_t^s, \tilde{\mathbf{u}}^s]\right) + 24\eta_{t+1}^s(\ell+\kappa+1)\Delta(\tilde{\mathbf{u}}^s) \\
& - \left(\frac{1}{2\eta_{t+1}^s} - \frac{3\eta_{t+1}^s(\ell+\kappa+1)^2}{2}\right)\mathbb{E}[\|\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}_t^s\|^2 \mid \mathbf{u}_t^s, \tilde{\mathbf{u}}^s] + 6\eta_{t+1}^s(\ell+\kappa+1)^2\|\tilde{\mathbf{u}}^s - \mathbf{u}^\star\|^2 \\
& + 3\eta_{t+1}^s(\ell+\kappa+1)^2\mathbb{E}[\|\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}^\star\|^2 \mid \mathbf{u}_t^s, \tilde{\mathbf{u}}^s] + 3\eta_{t+1}^s(\ell+\kappa+1)^2\|\mathbf{u}_t^s - \mathbf{u}^\star\|^2
\end{aligned}$$

This completes the proof. □

A.4 Proof of Lemma 4

Proof of Lemma 4. By the Cauchy-Schwarz inequality, we have

$$\|\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}^\star\|^2 \;\leq\; 2\|\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}_t^s\|^2 + 2\|\mathbf{u}_t^s - \mathbf{u}^\star\|^2$$

Rearranging the inequality in Lemma 3 with the above inequality and $0 < \eta_{t+1}^s \leq \eta$, we have

$$\begin{aligned}
& \left(1 - 24\eta(\ell+\kappa+1)\right)\mathbb{E}[\Delta(\bar{\mathbf{u}}_{t+1}^s) \mid \mathbf{u}_t^s, \tilde{\mathbf{u}}^s] \qquad (29)\\
\leq \;& \frac{1}{2\eta_{t+1}^s}\left(\|\mathbf{u}_t^s - \mathbf{u}^\star\|^2 - \mathbb{E}[\|\mathbf{u}_{t+1}^s - \mathbf{u}^\star\|^2 \mid \mathbf{u}_t^s, \tilde{\mathbf{u}}^s]\right) + 24\eta(\ell+\kappa+1)\Delta(\tilde{\mathbf{u}}^s) \\
& - \left(\frac{1}{2\eta} - \frac{63\eta(\ell+\kappa+1)^2}{2}\right)\mathbb{E}[\|\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}_t^s\|^2 \mid \mathbf{u}_t^s, \tilde{\mathbf{u}}^s] + 6\eta(\ell+\kappa+1)^2\|\tilde{\mathbf{u}}^s - \mathbf{u}^\star\|^2 \\
& - 12\eta(\ell+\kappa+1)^2\mathbb{E}[\|\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}^\star\|^2 \mid \mathbf{u}_t^s, \tilde{\mathbf{u}}^s] + 33\eta(\ell+\kappa+1)^2\|\mathbf{u}_t^s - \mathbf{u}^\star\|^2
\end{aligned}$$

Since $\eta \leq \frac{1}{100(\ell+\kappa+1)}$, we have

$$\begin{aligned}
1 - 24\eta(\ell+\kappa+1) & \geq \frac{2}{3} \\
24\eta(\ell+\kappa+1) & \leq \frac{1}{3} \\
\frac{1}{2\eta} - \frac{63\eta(\ell+\kappa+1)^2}{2} & \geq 0
\end{aligned}$$

Putting these pieces together with Eq. (29) and taking the expectation on all the randomness yields that

$$\begin{aligned}
\mathbb{E}[\Delta(\bar{\mathbf{u}}_{t+1}^s)] \;\leq\; & \frac{3}{4\eta_{t+1}^s}\left(\mathbb{E}[\|\mathbf{u}_t^s - \mathbf{u}^\star\|^2] - \mathbb{E}[\|\mathbf{u}_{t+1}^s - \mathbf{u}^\star\|^2]\right) + \frac{\mathbb{E}[\Delta(\tilde{\mathbf{u}}^s)]}{2} + 50\eta(\ell+\kappa+1)^2\mathbb{E}[\|\mathbf{u}_t^s - \mathbf{u}^\star\|^2] \\
& + 9\eta(\ell+\kappa+1)^2\mathbb{E}[\|\tilde{\mathbf{u}}^s - \mathbf{u}^\star\|^2] - 18\eta(\ell+\kappa+1)^2\mathbb{E}[\|\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}^\star\|^2]
\end{aligned}$$

Note that, by the definition of η_t^s, we have

$$\frac{1}{\eta_t^s} - \frac{1}{\eta_{t+1}^s} \geq \frac{1}{2\eta\sqrt{T}\sqrt{2T}} = \frac{1}{2\sqrt{2}\eta T} \geq 100\eta(\ell+\kappa+1)^2$$

Therefore, we conclude that

$$\begin{aligned} & \mathbb{E}[\Delta(\bar{\mathbf{u}}_{t+1}^s)] + 18\eta(\ell+\kappa+1)^2\mathbb{E}[\|\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}^\star\|^2] \\ \leq \; & \frac{3}{4}\left(\frac{\mathbb{E}[\|\mathbf{u}_t^s - \mathbf{u}^\star\|^2]}{\eta_t^s} - \frac{\mathbb{E}[\|\mathbf{u}_{t+1}^s - \mathbf{u}^\star\|^2]}{\eta_{t+1}^s}\right) + \frac{1}{2}\left(\mathbb{E}[\Delta(\tilde{\mathbf{u}}^s)] + 18\eta(\ell+\kappa+1)^2\mathbb{E}[\|\tilde{\mathbf{u}}^s - \mathbf{u}^\star\|^2]\right) \end{aligned} \tag{30}$$

Summing Eq. (30) up over $t = 0, 1, 2, \ldots, k_s - 1$ and dividing both sides by k_s, we have

$$\begin{aligned} & \mathbb{E}\left[\sum_{t=0}^{k_s-1} \frac{\Delta(\bar{\mathbf{u}}_{t+1}^s) + 18\eta(\ell+\kappa+1)^2\|\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}^\star\|^2}{k_s}\right] \\ \leq \; & \frac{3}{4}\left(\frac{\mathbb{E}[\|\mathbf{u}_0^s - \mathbf{u}^\star\|^2]}{k_s\eta_0^s} - \frac{\mathbb{E}[\|\mathbf{u}_{k_s}^s - \mathbf{u}^\star\|^2]}{k_s\eta_{k_s}^s}\right) + \frac{1}{2}\left(\mathbb{E}[\Delta(\tilde{\mathbf{u}}^s)] + 18\eta(\ell+\kappa+1)^2\mathbb{E}[\|\tilde{\mathbf{u}}^s - \mathbf{u}^\star\|^2]\right) \end{aligned}$$

Note that $\Delta(\mathbf{u})$ and $\|\mathbf{u} - \mathbf{u}^\star\|^2$ are both convex in $\mathbf{u}$. Since $\tilde{\mathbf{u}}^{s+1} = (1/k_s)\sum_{t=1}^{k_s} \bar{\mathbf{u}}_t^s$, we have

$$\Delta(\tilde{\mathbf{u}}^{s+1}) + 18\eta(\ell+\kappa+1)^2\|\tilde{\mathbf{u}}^{s+1} - \mathbf{u}^\star\|^2 \; \leq \; \sum_{t=0}^{k_s-1} \frac{\Delta(\bar{\mathbf{u}}_{t+1}^s) + 18\eta(\ell+\kappa+1)^2\|\bar{\mathbf{u}}_{t+1}^s - \mathbf{u}^\star\|^2}{k_s}$$

In addition, we have $\mathbf{u}_{k_s}^s = \mathbf{u}_0^{s+1}$, $\eta_{k_s}^s = \eta_0^{s+1}$ and $k_{s+1} = 2k_s$. Putting these pieces together yields that

$$\begin{aligned} & \mathbb{E}\left[\Delta(\tilde{\mathbf{u}}^{s+1}) + 18\eta(\ell+\kappa+1)^2\|\tilde{\mathbf{u}}^{s+1} - \mathbf{u}^\star\|^2 + \frac{3\|\mathbf{u}_0^{s+1} - \mathbf{u}^\star\|^2}{2k_{s+1}\eta_0^{s+1}}\right] \\ \leq \; & \frac{1}{2}\left(\mathbb{E}\left[\Delta(\tilde{\mathbf{u}}^s) + 18\eta(\ell+\kappa+1)^2\|\tilde{\mathbf{u}}^s - \mathbf{u}^\star\|^2 + \frac{3\|\mathbf{u}_0^s - \mathbf{u}^\star\|^2}{2k_s\eta_0^s}\right]\right) \end{aligned}$$

After telescoping for $s = 0, 1, 2, \ldots, S-1$ and using $k_0 \geq 1$, we have

$$\mathbb{E}[\Delta(\tilde{\mathbf{u}}^S)] \leq \frac{1}{2^S}\left(\Delta(\tilde{\mathbf{u}}^0) + 18\eta(\ell+\kappa+1)^2\|\tilde{\mathbf{u}}^0 - \mathbf{u}^\star\|^2 + \frac{3\|\mathbf{u}_0^0 - \mathbf{u}^\star\|^2}{\eta k_0\sqrt{2}}\right)$$

This completes the proof. □

A.5 Proof of Lemma 6

Proof of Lemma 6. By the definition of the exchangeable pair, we derive from Eq. (16) that

$$\begin{aligned} & \left|\mathbb{E}[\Delta(\mathbf{u}_t^s)] - \mathbb{E}[L_{\sigma_{t-1}^s}(\lambda_t^s, \beta_t^s, \gamma^\star) - L_{\sigma_{t-1}^s}(\lambda^\star, \beta^\star, \gamma_t^s)]\right| \\ = \; & \left|\mathbb{E}[L_{\sigma_{t-1}^s}(\lambda_t(\tilde{\sigma}^s), \beta_t(\tilde{\sigma}^s), \gamma^\star) - L_{\sigma_{t-1}^s}(\lambda^\star, \beta^\star, \gamma_t(\tilde{\sigma}^s))] - \mathbb{E}[L_{\sigma_{t-1}^s}(\lambda_t^s, \beta_t^s, \gamma^\star) - L_{\sigma_{t-1}^s}(\lambda^\star, \beta^\star, \gamma_t^s)]\right| \end{aligned}$$

By using the conditional expectation on $\{\sigma^s_{t-1} = r\}$ and the fact that σ^s and $\tilde{\sigma}^s$ are independent, we have

$$\mathbb{E}[L_{\sigma^s_{t-1}}(\lambda_t(\tilde{\sigma}^s), \beta_t(\tilde{\sigma}^s), \gamma^\star) - L_{\sigma^s_{t-1}}(\lambda^\star, \beta^\star, \gamma_t(\tilde{\sigma}^s))] = \frac{1}{n}\sum_{r=1}^n \mathbb{E}[L_r(\lambda_t(\tilde{\sigma}^s), \beta_t(\tilde{\sigma}^s), \gamma^\star) - L_r(\lambda^\star, \beta^\star, \gamma_t(\tilde{\sigma}^s))]$$

$$\mathbb{E}[L_{\sigma^s_{t-1}}(\lambda^s_t, \beta^s_t, \gamma^\star) - L_{\sigma^s_{t-1}}(\lambda^\star, \beta^\star, \gamma^s_t)] = \frac{1}{n}\sum_{r=1}^n \mathbb{E}[L_r(\lambda^s_t, \beta^s_t, \gamma^\star) - L_r(\lambda^\star, \beta^\star, \gamma^s_t) \mid \sigma^s_{t-1} = r]$$

Putting these pieces together with the triangle inequality yields that

$$\begin{aligned}
&\left|\mathbb{E}[\Delta(\mathbf{u}^s_t)] - \mathbb{E}[L_{\sigma^s_{t-1}}(\lambda^s_t, \beta^s_t, \gamma^\star) - L_{\sigma^s_{t-1}}(\lambda^\star, \beta^\star, \gamma^s_t)]\right| \\
\leq\ & \frac{1}{n}\sum_{r=1}^n \left|\mathbb{E}[L_r(\lambda_t(\tilde{\sigma}^s), \beta_t(\tilde{\sigma}^s), \gamma^\star) - L_r(\lambda^\star, \beta^\star, \gamma_t(\tilde{\sigma}^s))] - \mathbb{E}[L_r(\lambda^s_t, \beta^s_t, \gamma^\star) - L_r(\lambda^\star, \beta^\star, \gamma^s_t) \mid \sigma^s_{t-1} = r]\right| \\
\leq\ & \frac{1}{n}\sum_{r=1}^n \sup_{g \text{ is } G\text{-Lipschitz}} \left(\mathbb{E}[g(\lambda_t(\tilde{\sigma}^s), \beta_t(\tilde{\sigma}^s), \gamma_t(\tilde{\sigma}^s))] - \mathbb{E}[g(\lambda^s_t, \beta^s_t, \gamma^s_t) \mid \sigma^s_{t-1} = r]\right) \\
=\ & \frac{1}{n}\sum_{r=1}^n G \cdot \mathcal{W}_1(\mathcal{D}_{t,s}, \mathcal{D}^{(r)}_{t,s}),
\end{aligned}$$

where G is the Lipschitz parameter. This together with the fact that $\mathcal{W}_1(\mu, \nu) \leq \mathcal{W}_2(\mu, \nu)$ yields the desired result. $\square$

A.6 Proof of Lemma 7

Proof of Lemma 7. By Lemma 6, it suffices to upper bound the term $\mathcal{W}_2(\mathcal{D}_{t,s}, \mathcal{D}^{(r)}_{t,s})$. While the coupling used here has been established in [Nagaraj et al.(2019)], we hope to remark that their argument can not be directly applied since the iterates generated by the SPPRR algorithm does not necessarily satisfy [Nagaraj et al.(2019), Lemma 2]. Nonetheless, we prove that this issue can be fixed by controlling the error if $\|\mathbf{u}^s_{M,t} - \bar{\mathbf{u}}^s_t\| \leq \frac{\epsilon}{10GnS}$. To facilitate the readers, we divide our proof into three steps.

Coupling construction. By the definition of 2-Wasserstein distance, the tight upper bound is based on the construction of a particular coupling between $\mathcal{D}_{t,s}$ and $\mathcal{D}^{(r)}_{t,s}$. We use the one established in [Nagaraj et al.(2019)] and present the details as follows.

Let $\mathcal{R}_n$ be the set of all random permutations over the set $[n]$. For $i, j \in [n]$, we define the exchange function $E_{i,j} : \mathcal{R}_n \to \mathcal{R}_n$. That is to say, for any $\sigma \in \mathcal{R}_n$, the permutation $E_{i,j}(\sigma)$ stands for a new one where i-th and j-th entries of σ are exchangeable and it keeps everything else the same. We construct the operator $\omega_{r,t} : \mathcal{R}_n \to \mathcal{R}_n$ as follows:

$$\omega_{r,t}(\sigma) = \begin{cases} \sigma, & \text{if } \sigma_{t-1} = r \\ E_{t-1,j}(\sigma), & \text{if } \sigma_j = r \text{ and } j \neq t-1 \end{cases}$$

Intuitively, $\omega_{r,t}$ performs a single swap such that the $(t-1)$-th position of the permutation is r. Clearly, if σ^s is a random permutation at uniform, $\omega_{r,t}(\sigma^s)$ has the same distribution as $\sigma^s \mid \sigma^s_{t-1} = r$. Based on this construction, we have $\mathbf{u}_t(\sigma^s) \sim \mathcal{D}_{t,s}$ and $\mathbf{u}_t(\omega_{r,t}(\sigma^s)) \sim \mathcal{D}^{(r)}_{t,s}$. This gives a coupling between $\mathcal{D}_{t,s}$ and $\mathcal{D}^{(r)}_{t,s}$. Therefore, we conclude that

$$\mathcal{W}_2(\mathcal{D}_{t,s}, \mathcal{D}^{(r)}_{t,s}) \leq \sqrt{\mathbb{E}[\|\mathbf{u}_t(\sigma^s) - \mathbf{u}_i(\omega_{r,t}(\sigma^s))\|^2]} \tag{31}$$

Coupling upper bound. Let σ^s and $\tilde{\sigma}^s$ be two random permutations of the set $[n]$, we denote $\mathbf{u}_t(\sigma^s)$ and $\mathbf{u}_t(\tilde{\sigma}^s)$ by $\mathbf{v}_t$ and $\mathbf{w}_t$ respectively. It is clear that $\|\mathbf{v}_0 - \mathbf{w}_0\| = 0$ almost surely. Let $j \leq t$, we consider two cases. For the first case, we suppose that $\sigma_j^s = r \neq \tilde{r} = \tilde{\sigma}_j^s$. Then, due to the error criterion, we have

$$\|\mathbf{v}_j - \mathbf{w}_j\| \ \leq \ \|\mathbf{v}_j - \bar{\mathbf{v}}_{j-1}\| + \|\bar{\mathbf{v}}_{j-1} - \bar{\mathbf{w}}_{j-1}\| + \|\bar{\mathbf{w}}_{j-1} - \mathbf{w}_j\| \ \leq \ \|\bar{\mathbf{v}}_{j-1} - \bar{\mathbf{w}}_{j-1}\| + \frac{D_{\mathbf{u}}}{5(nS)^{3/2}}$$

where $\bar{\mathbf{v}}_{j-1}$ and $\bar{\mathbf{w}}_{j-1}$ are defined by

$$\begin{aligned} \bar{\mathbf{v}}_{j-1} &= \mathcal{P}_{\Lambda\times\Gamma}(\mathbf{v}_{j-1} - \eta F_r(\bar{\mathbf{v}}_{j-1})) \\ \bar{\mathbf{w}}_{j-1} &= \mathcal{P}_{\Lambda\times\Gamma}(\mathbf{w}_{j-1} - \eta F_{\tilde{r}}(\bar{\mathbf{w}}_{j-1})) \end{aligned}$$

Therefore, we have

$$\begin{aligned} \|\mathbf{v}_j - \mathbf{w}_j\| &\leq \|(\mathbf{v}_{j-1} - \eta F_r(\bar{\mathbf{v}}_{j-1})) - (\mathbf{w}_{j-1} - \eta F_{\tilde{r}}(\bar{\mathbf{w}}_{j-1}))\| + \frac{D_{\mathbf{u}}}{5(nS)^{3/2}} \\ &\leq \|\mathbf{v}_{j-1} - \mathbf{w}_{j-1}\| + 2\eta G + \frac{D_{\mathbf{u}}}{5(nS)^{3/2}} \end{aligned} \tag{32}$$

For the second case, we suppose that $\sigma_j^s = r = \tilde{\sigma}_j^s$. Then, by the same argument, we have

$$\|\mathbf{v}_j - \mathbf{w}_j\| \ \leq \ \|\bar{\mathbf{v}}_{j-1} - \bar{\mathbf{w}}_{j-1}\| + \frac{D_{\mathbf{u}}}{5(nS)^{3/2}}$$

where $\bar{\mathbf{v}}_{j-1}$ and $\bar{\mathbf{w}}_{j-1}$ are defined by

$$\begin{aligned} \bar{\mathbf{v}}_{j-1} &= \mathcal{P}_{\Lambda\times\Gamma}(\mathbf{v}_{j-1} - \eta F_r(\bar{\mathbf{v}}_{j-1})) \\ \bar{\mathbf{w}}_{j-1} &= \mathcal{P}_{\Lambda\times\Gamma}(\mathbf{w}_{j-1} - \eta F_r(\bar{\mathbf{w}}_{j-1})) \end{aligned}$$

Since F_r is monotone, it is easy to verify that $\|\bar{\mathbf{v}}_{j-1} - \bar{\mathbf{w}}_{j-1}\| \leq \|\mathbf{v}_{j-1} - \mathbf{w}_{j-1}\|$. Therefore, we have

$$\|\mathbf{v}_j - \mathbf{w}_j\| \ \leq \ \|\mathbf{v}_{j-1} - \mathbf{w}_{j-1}\| + \frac{D_{\mathbf{u}}}{5(nS)^{3/2}} \tag{33}$$

Combining Eq. (32) and Eq. (33) yields that

$$\|\mathbf{u}_t(\sigma^s) - \mathbf{u}_t(\tilde{\sigma}^s)\| \ \leq \ 2\eta G \cdot |\{j \leq t : \sigma_j^s \neq \tilde{\sigma}_j^s\}| + \frac{tD_{\mathbf{u}}}{5(nS)^{3/2}} \tag{34}$$

Main proof. By the definition of $\omega_{r,t}(\cdot)$, we have $|\{j \leq t : \sigma_j^s \neq [\omega_{r,t}(\sigma^s)]_j^s\}| \leq 1$. Plugging this into Eq. (34) and using the fact that $0 \leq t < n$ and $S \geq 1$ yields that

$$\|\mathbf{u}_t(\sigma^s) - \mathbf{u}_i(\omega_{r,t}(\sigma^s))\| \ \leq \ 2\eta G + \frac{D_{\mathbf{u}}}{5\sqrt{nS}}$$

Plugging the above inequality into Eq. (31) yields that

$$\mathcal{W}_2(\mathcal{D}_{t,s}, \mathcal{D}_{t,s}^{(r)}) \ \leq \ 2\eta G + \frac{D_{\mathbf{u}}}{5\sqrt{nS}}$$

This together with Lemma 6 yields the desired result. $\square$

A.7 Proof of Lemma 8

Proof of Lemma 8. By the definition, we have $\|\mathbf{u}_0^0 - \mathbf{u}^\star\| \leq D_{\mathbf{u}}$. Then we consider $\|\mathbf{u}_1^0 - \mathbf{u}^\star\|$. Indeed, regardless of any random permutation, we derive from the error criterion that

$$\|\mathbf{u}_1^0 - \mathbf{u}^\star\| \;\leq\; \|\mathbf{u}_1^0 - \bar{\mathbf{u}}_0^0\| + \|\bar{\mathbf{u}}_0^0 - \mathbf{u}^\star\| \;\leq\; \|\bar{\mathbf{u}}_0^0 - \mathbf{u}^\star\| + \frac{D_{\mathbf{u}}}{10(nS)^{3/2}}$$

By the definition of $\bar{\mathbf{u}}_0^0$ and the boundness of F_i for $\forall i \in [n]$, we have

$$\|\bar{\mathbf{u}}_0^0 - \mathbf{u}^\star\| = \|\mathcal{P}_{\Lambda\times\Gamma}(\mathbf{u}_0^0 - \eta F_{\sigma_0^0}(\bar{\mathbf{u}}_0^0)) - \mathbf{u}^\star\| \;\leq\; \|\mathbf{u}_0^0 - \mathbf{u}^\star\| + \eta G$$

Putting these pieces together yields that

$$\|\mathbf{u}_1^0 - \mathbf{u}^\star\| \;\leq\; \|\mathbf{u}_0^0 - \mathbf{u}^\star\| + \eta G + \frac{D_{\mathbf{u}}}{10(nS)^{3/2}} \;\leq\; D_{\mathbf{u}} + \eta G + \frac{D_{\mathbf{u}}}{10(nS)^{3/2}}$$

Repeating the above argument, we have

$$\begin{aligned}
\|\mathbf{u}_t^s - \mathbf{u}^\star\| \;&\leq\; \|\mathbf{u}_{t-1}^s - \mathbf{u}^\star\| + \eta G + \frac{D_{\mathbf{u}}}{10(nS)^{3/2}} \;\leq\; \|\mathbf{u}_0^s - \mathbf{u}^\star\| + t\left(\eta G + \frac{D_{\mathbf{u}}}{10(nS)^{3/2}}\right) \\
&\leq\; \|\mathbf{u}_0^0 - \mathbf{u}^\star\| + (ns+t)\left(\eta G + \frac{D_{\mathbf{u}}}{10(nS)^{3/2}}\right),
\end{aligned}$$

and

$$\|\bar{\mathbf{u}}_{t-1}^s - \mathbf{u}^\star\| \;\leq\; \|\mathbf{u}_{t-1}^s - \mathbf{u}^\star\| + \eta G \;\leq\; \|\mathbf{u}_0^0 - \mathbf{u}^\star\| + (ns+t)\left(\eta G + \frac{D_{\mathbf{u}}}{10(nS)^{3/2}}\right)$$

Since $1 \leq t \leq n$ and $0 \leq s \leq S-1$, we have $ns + t \leq nS$. This together with the previous two inequalities, $\|\mathbf{u}_0^0 - \mathbf{u}^\star\| \leq D_{\mathbf{u}}$ and $nS \geq 1$ yields the desired result. $\square$