DECOMPOSITION ALGORITHMS FOR DISTRIBUTIONALLY ROBUST OPTIMIZATION USING WASSERSTEIN METRIC*

FENGQIAO LUO† AND SANJAY MEHROTRA‡

Abstract. We study distributionally robust optimization (DRO) problems where the ambiguity set is defined using the Wasserstein metric. We show that this class of DRO problems can be reformulated as semi-infinite programs. We give an exchange method to solve the reformulated problem for the general nonlinear model, and a central cutting-surface method for the convex case, assuming that we have a separation oracle. We used a distributionally robust generalization of the logistic regression model to test our algorithm. Numerical experiments on the distributionally robust logistic regression models show that the number of oracle calls are typically $20 \sim 50$ to achieve 5-digit precision. The solution found by the model is generally better in its ability to predict with a smaller standard error.

Key words. distributionally robust optimization, Wasserstein metric, semi-infinite programming, cutting-surface algorithms, exchange method

AMS subject classifications. 90C34, 90C15, 90C25, 90C26, 65D05, 90C31, 90C05, 90C90

1. Introduction.

We consider a distributionally robust optimization problem defined using the Wasserstein metric:

$$\text{(WRO)} \qquad \inf_{\theta\in\Theta} \sup_{P\in\mathcal{P}} \mathbb{E}_P[h(\theta,\xi)],$$

where θ are the decision variables, and ξ are the model parameters. We assume that $h(\theta,\xi)$ is a continuous, but possibly a non-convex function in θ. Additional assumptions are made as results are developed in Sections 3 and 4. Let ξ be a random vector defined on a measurable space $(\Xi,\mathcal{F})$, where Ξ is the support in $\mathbb{R}^n$, and $\mathcal{F}$ is a σ-algebra that contains all singletons in Ξ, i.e., $\{\xi\}\in\mathcal{F},\ \ \forall\xi\in\Xi$. We let the distribution $\mathcal{P}$ followed by ξ be unknown, but assume that it belongs to an ambiguity set $\mathcal{P}$. Let $\mathcal{M}(\Xi,\mathcal{F})$ be the set of all probability distributions on $(\Xi,\mathcal{F})$, and define $\mathcal{P}$ by using the L_1-Wasserstein metric:

$$\mathcal{P} := \big\{P\in\mathcal{M}(\Xi,\mathcal{F}) \;\big|\; \mathcal{W}(P,P_0)\le r_0\big\}, \tag{1}$$

where P_0 is an empirical distribution satisfying: $P_0(\xi^i)=1/m,\ i\in[m]$, and $\{\xi^i\in\Xi : i\in[m]\}$ are the observed samples of ξ, and $r_0>0$ is a scalar. The L_1-Wasserstein metric is defined as [23]:

$$\mathcal{W}(P_1,P_2) := \inf_{K\in\mathcal{S}(P_1,P_2)} \int_{\Xi\times\Xi} d\big(s_1,s_2\big)K(ds_1\times ds_2), \tag{2}$$

where $\mathcal{S}(P_1,P_2) := \big\{K\in\mathcal{M}(\Xi\times\Xi,\mathcal{F}\times\mathcal{F}) :\ K(A\times\Xi)=P_1(A),\ K(\Xi\times A)=P_2(A),\ \forall A\in\mathcal{F}\big\}$ is the set of all joint probability distributions whose marginal distributions are P_1 and P_2, and $d(\cdot,\cdot)$ is a metric defined on Ξ, which is measurable in

*Submitted to the editors DATE.

Funding: This research was partially supported by NSF grants CMMI-1362003 and CMMI-1100868.

†Department of Industrial Engineering and Management Science, Northwestern University, Evanston, IL, 60208 (fengqiaoluo2014@u.northwestern.edu).

‡Department of Industrial Engineering and Management Science, Northwestern University, Evanston, IL, 60208 (mehrotra@iems.northwestern.edu).

$\mathcal{F} \times \mathcal{F}$. The set Ξ is partitioned as: $\Xi = \cup_{i=1}^{m+1} \Xi^i$, where $\Xi^i = \{\xi^i\}$ for $i \in [m]$ and $\Xi^{m+1} = \Xi \setminus (\cup_{i\in[m]} \Xi^i)$. We call (WRO) a Wasserstein-robust optimization problem from hereafter.

Wasserstein metric has been used to study the convergence of an empirical distribution from the i.i.d. samples to the true distribution. Specifically, Barrio et al. [1] showed that under the Wasserstein metric the empirical distribution converges to the true distribution almost surely as the number of samples go to infinity. Fournier and Guillin [22] have further shown that if the true distribution P_{true} has a light tail, then the probability $\Pr\{W(P_0, P_{\text{true}}) \geq r_0\}$ can be bounded from above by a function that decays exponentially with the sample size and r_0.

The (WRO) modeling framework allows to protect against ambiguity in the distribution when arriving at a decision. It also allows to perform sensitivity analysis with respect to the empirical distribution. This is useful particularly when the sample size m is small.

Contributions and organization of this paper. This paper makes the following contributions:

1. It is common in robust optimization to dualize the inner problem to develop a reformulation of the original model [2, 4, 6, 7]. The definition of Wasserstein metric in (2) uses a semi-infinite number of equality constraints, therefore its direct use is not suitable for dualization of the inner problem. We prove in Section 3 that the inner problem in (WRO) is equivalent to a conic linear optimization problem. We show that this conic program can be dualized with no duality gap, thus obtaining a semi-infinite programming reformulation of (WRO). The constraint sets in this semi-infinite program decompose in the observed samples. The decomposition allows for an independent separation problem for each observed sample ξ^i for algorithms that use cuts to solve the reformulated semi-infinite program. The results in Section 3 only assume that $h(\theta, \xi)$ is a continuous and bounded function in θ, and Ξ is a compact set (See Assumption 3.1). These results, for example, are applicable for problems involving mixed-integer variables used in the definition of Ξ.
2. We adapt the exchange algorithm and the central cutting-surface algorithm from [29] and [43] in Sections 4.1 and 4.2 for the general and convex cases, respectively. We show finite convergence of the exchange method to a solution with a desirable accuracy. The linear rate of convergence for the cutting surface algorithm presented here exploits the structure of the semi-infinite program. Specifically, a global linear rate of convergence is proved.
3. In Section 5.2 we present results on the computational performance of the central cutting-surface algorithm for solving WRLR problems. We find that the number of oracle calls are typically $20 \sim 50$, and the number of cuts added to the model are typically $3 \sim 10$ times the number of training samples. The solution time is $\lesssim 100$ times that of solving the ordinary logistic regression model.

   In Section 5.3, we present performance results on the quality of model obtained by the Wasserstein-robust logistic regression (WRLR) approach and compare it with the performance of the ordinary logistic regression (LR). Our motivation is to study a setting in which the number of available samples is small, and robustness is used to understand and possibly improve the quality of the trained model. This is typically the case at an early stage of a study (e.g., in healthcare) when limited data is available due to data collection ex-

penses. We use m to be $\{50, 75, 100, 150\}$ in the numerical experiments to test the performance of the Wasserstein-robust logistic regression model (See Section 5). Eleven data sets from UCI Machine Learning Repository are used. We use area under the receiver operator characteristic curve (AUC) to evaluate the performance of the models [69]. We find that the Wasserstein-robust logistic regression WRLR model has a significantly better out of sample performance than logistic-regression model (with $\alpha = 0.05$) in 24 (55%) cases. The predictive performance of WRLR is worse in 7 (16%) cases ($\alpha = 0.05$), and for the remaining 13 (29%) cases the difference is not statistically significant. The WRLR models also have smaller standard error when compared to logistic regression, suggesting that the model is more robust.

2. Literature Review. We now provide a literature review of prior work on distributionally robust optimization (DRO) and semi-infinite programming. These topics are relevant because the (WRO) problem is reformulated as a semi-infinite program, and a separation oracle is needed in the algorithms.

2.1. Distributionally robust optimization. Distributional robust optimization is a generalization of robust optimization (RO) [10, 24, 64], where an ambiguity set is used to model problem data distribution. The use of an ambiguity set in distributionally robust optimization overcomes the weakness of traditional robust optimization framework, as the RO model is often considered to be very conservative. Even when a deterministic model is a convex optimization problem, its DR-counterpart is NP-hard in most cases [2]. Therefore, literature on RO and DRO either makes assumptions on the function form of the objective and constraints with respect to uncertain parameters to ensure the convexity of the model [5, 11], or purposes convex reformulations to approximate original problems [24, 62].

Studies of DRO focus on ways of defining the ambiguity set, reformulations of these models into computationally tractable problems, probability guarantee of the constraint satisfaction by the true distribution, and applications. We briefly review these aspects and the literature below.

2.1.1. The ambiguity set. A common approach to describe ambiguity sets is to use moments of the distribution followed by the random parameters [9, 12, 18, 19, 43, 52, 56, 57]. Bertsimas and Popescu [11] discuss properties of probability distributions satisfying such constraints. It is also possible to use a *statistical distance* as a way of measuring the difference of two probability distributions. The statistical distances used in the DRO models are Wasserstein metric [21, 44, 49, 50, 54, 63], ϕ-divergence, χ^2-distance, Kullback-Leibler divergence [3, 13, 32, 41, 61, 67], and the Prokhorov metric [20].

2.1.2. Reformulation of DRO models. Shapiro and Kleywegt [57] show that under mild regularity conditions, the DRO problem with a deterministic set of constraints is equivalent to a stochastic programming problem based on a probability distribution that is a linear combination of distributions in the ambiguity set of the original DRO problem. It means that solving a DRO problem in this case is equivalent to solving a stochastic programming problem. Goh and Sim [24] investigate a two-stage DRO model whose objective has a linear structure, but can be used to express piece-wise linear utility functions and CVaR constraints. They show that by restricting the recourse variables to be affine mappings of uncertain parameters, this two-stage DRO model can be reformulated as a minimax linear programming problem. Delage and Ye [18] show that DRO problems whose ambiguity sets are defined

by the first and second moment inequalities are polynomial-time solvable under the assumption that the objective is convex in the decision variables and concave in uncertain parameters. They also provide semidefinite formulations for the data-driven problems. As an often used objective, the least-square loss function is convex in both decision variables and model parameters, hence violating the assumption in [18]. To overcome this obstacle, Mehrotra and Zhang [44] give conic and semidefinite programming reformulations of DR-least-squares problems with ambiguity sets defined using the first two moments, Wasserstein metric, and bounds on the probability density functions, respectively. Mehrotra and Papp [43] use the central cutting-surface algorithm they developed to solve DRO models where the ambiguity set is specified using arbitrarily many generalized moments. Wiesemann et al. [62] propose a framework for modeling and solving distributionally robust convex optimization problems, in which the ambiguity set is conically representable and constraint functions are piecewise affine in both decision variables and random parameters. They show that the reformulated problem is polynomial-time solvable under a strict nesting condition of the confidence sets. Esfahan and Kuhn [21] show that using the conic duality theory, the data-driven DRO problem with a Wasserstein-metric ambiguity set can be reduced to finitely many tractable convex optimization problems, if the loss function can be expressed as the point-wise maximum of finitely many concave functions in the uncertain parameter. However, this assumption on the loss function is violated by many statistical learning models, such as the logistic regression model considered in the computational section of this paper.

Many DRO problems involve robust chance constraints, which are often nonconvex. Jiang and Guan [33] study DR-chance constraints defined by the ϕ-divergences. They show that these constraints are equivalent to ordinary chance constraints based on some nominal probability measure. Hanasusanto et al. [26] show that if the ambiguity set in the robust chance constraint is defined by moments and satisfies a nested condition, the worst-case probability is an optimal solution of a conic optimization problem. For a recent review on tractable reformulations of robust chance constraints, see [51].

2.1.3. The probability bound. Studies on the probability that the true distribution is contained in the ambiguity set are related to the ambiguous chance constrained programming [14, 26, 70]. Erdoğan and Iyengar [20] show that the sampling of robust constraints is a good approximation for the DR-chance-constraint problem with a high probability. Delage and Ye [18] use the size of the ellipsoid confidence region using the second moment to satisfy a given level of probability guarantee. Esfahan and Kuhn [21] give a result on the out-of-sample performance guarantee of the solution to the data-driven DRO problem with a Wasserstein-metric ambiguity set. Calafiore and Ghaoui [14] show that chance constraints of linear inequalities with respect to a radial distribution (i.e., Gaussian distribution and uniform distribution on ellipsoidal support) can be converted explicitly into convex second-order cone constraints. Additionally, they show that distributionally robust chance constraints of linear inequalities under a few important distribution families (distributions with known mean and covariance, radially symmetric non-increasing distributions, etc.) can be guaranteed by some deterministic convex constraints. Bertsimas et al. [10] propose a novel scheme of constructing uncertainty sets for data-driven robust optimization problems using hypothesis tests. The resulting model is computationally tractable and has application insights regarding using statistical estimate for chance constraint violation. Ben-Tal et al. [3] show that the robust counterpart of linear

optimization problems with uncertainty set defined by ϕ-divergence are tractable for most choices of ϕ. Constructing confidence sets using ϕ-divergence is also studied in [41, 67].

2.1.4. Applications of DRO Models. Modeling and solutions of the distributionally robust counterparts of deterministic optimization models are investigated in various areas in operations research, including but not limit to: inventory management [68], scheduling and logistics [31, 37], and risk management [39, 45]. In statistical learning, Lee and Mehrotra [36] study distributionally robust linear support vector machine (DR-SVM) models with a Wasserstein-metric ambiguity set. They find that the (DR-SVM) model can be reformulated as a semi-infinite program, in which the master problems are convex quadratic programs and separation problems are linear programs. They also find that (DR-SVM) models have improved generalization capabilities than ordinary (SVM) models. Shaeezadeh-Abadeh et al. [54] investigate a distributionally robust logistic regression model using a Wasserstein-metric ambiguity set. For the model in [54], the uncertainty set of the attribute vector is assumed to be the entire $\mathbb{R}^n$, and the authors show that the semi-infinite constraints in the dual problem are equivalent to a single constraint obtained using the conjugate function. The assumption in [54] is not practical in settings such as healthcare, where certain physiological variables (e.g., heart rate, blood pressure) must necessarily be bounded. In the framework of this paper, the uncertainty set is assumed to be compact. Additionally, feature variables can be integer-valued.

2.2. Theory and numerical methods for semi-infinite programming. The study of distributional robust optimization models considered in this paper benefit from the known literature on semi-infinite programming. Semi-infinite programming (SIP) problems are optimization problems with constraints induced by a continuous parameter. The study of SIP is initialized by the work of Haar [25] and Charnes [15, 16, 17] focusing on linear-SIP problems. Later, the first and second order optimality conditions of general SIP were given in [27, 28, 30, 46, 47, 58]. For reviews of the theory and methods for SIP, see [29, 40, 53].

Numerical methods for solving convex SIP problems include primal methods [60], dual methods [29], penalty methods [38, 66], smooth approximation and projection methods [65], and cutting-plane methods [34]. Primal methods are based on searching for feasible descent directions, while dual methods are based on finding a solution of the system of KKT optimality conditions. In a penalty method, constraints are penalized in the objective and the penalty term is an integral of the constraint function over the continuous parameter. In a smooth approximation and projection method, infinitely many functions are replaced by a integral entropy function as an approximation. The SIP is solved by the smoothing projected gradient method. In a cutting-plane algorithm, a typical iteration involves solving a master problem with finitely many constraints and adding violated constraints obtained from solving a separation problem. Mehrotra and Papp [43] developed a central-cutting-surface algorithm with a linear rate of convergence for solving convex SIP problems. They demonstrate that adding cutting surfaces, as compared to cutting planes, can be computationally effective for problems in high dimension and ensure greater stability in the algorithm's performance. The central-cutting surface method is related to the exchange method [29], however it uses a centrality parameter in the algorithm.

3. Reformulation of the Wasserstein-robust Optimization Problem. In this section we show that (WRO) is equivalent to a semi-infinite program. This semi-

infinite program decomposes in scenarios. We note that the definition of $\mathcal{W}(P, P_0) \leq r_0$ involves infinitely many equality constraints of the form $K(A \times \Omega) = P,\ \forall A \in \mathcal{F}$. Since this form is not suitable for dualization of the inner problem, Theorem 3.6 below reformulates the inner problem in (WRO) as a conic linear program with finitely many constraints. Proposition 3.3 provides an intermediate result, and shows the continuity of the objective function of the inner problem in (WRO) with respect to the probability measure. Lemma 3.5 is a technical result that generalizes similar convergent sequence existence results for the finite dimensional sets to a set defined by the Wasserstein metric. We make the following general assumption throughout this paper. Additional assumptions are made at appropriate places as results are developed. Note that results in this section make no assumption on the metric $d(\cdot, \cdot)$ used in defining (2).

ASSUMPTION 3.1. *We assume that $r_0 > 0$, $\forall \theta \in \Theta$ in (2). The feasible region $\Theta \subseteq \mathbb{R}^n$ and the domain Ξ are compact. The function $h(\cdot, \cdot)$ is bounded on $\Theta \times \Xi$, and for every $\theta \in \Theta$, there exists $C(\theta) > 0$ such that the function $h(\theta, \cdot)$ satisfies $|h(\theta, s_1) - h(\theta, s_2)| \leq C(\theta) d(s_1, s_2),\ \forall s_1, s_2 \in \Xi$.*

DEFINITION 3.2. *Let $(\mathcal{X}, d_{\mathcal{X}})$ be a metric space formed by defining a metric d_X on the topological space $\mathcal{X}$. A function $f : \mathcal{X} \longmapsto \mathbb{R}$ is continuous in $(\mathcal{X}, d_{\mathcal{X}})$ if for a given $x \in \mathcal{X}$, and $\varepsilon > 0$, $\exists \delta > 0$ such that $\forall x' \in \mathcal{X}$, $d_{\mathcal{X}}(x', x) < \delta$ we have $|f(x') - f(x)| < \varepsilon$.*

Lemma 3.5 below shows that the interior of the Wasserstein ball has a sequence of distributions that converge to a chosen point on the boundary of the ball. We need the following two results, proved in Appendix A, in the proof of Lemma 3.5.

PROPOSITION 3.3. *Let Assumption 3.1 hold, then the function $f(\theta, \cdot) : \mathcal{M}(\Xi, \mathcal{F}) \longmapsto \mathbb{R}$ defined by $f(\theta, P) := \mathbb{E}_P[h(\theta, \xi)]$ is continuous in $(\mathcal{M}(\Xi, \mathcal{F}), \mathcal{W})$.*

PROPOSITION 3.4. *For any probability measure $P \in \mathcal{M}(\Omega, \mathcal{F})$, there exists a $K \in \mathcal{M}(\Xi \times \Xi, \mathcal{F} \times \mathcal{F})$ such that $K(A \times A) = P(A)$ and $K(\Xi \times A) = K(A \times \Xi) = P(A)$ for any $A \in \mathcal{F}$. For such a joint probability measure K, we have*

$$\int_{(s_1 \times s_2) \in \Xi \times \Xi} d\big(s_1, s_2\big) K(ds_1 \times ds_2) = 0. \tag{3}$$

LEMMA 3.5. *Let Assumption 3.1 hold, and $\mathcal{P}$ be defined as in (1). Let $\mathcal{P}' := \{P \in \mathcal{M}(\Xi, \mathcal{F}) :\ \mathcal{W}(P, P_0) < r_0\}$ be the interior of $\mathcal{P}$. Then for any $P \in \mathcal{P}$, there exists a sequence $\{P^n\}_{n=1}^{\infty} \subseteq \mathcal{P}'$ such that $\lim_{n \to \infty} \mathcal{W}(P^n, P) = 0$.*

Proof. For any given $P \in \mathcal{P}$ and for any $\varepsilon > 0$, by the definition of the Wasserstein metric, there exists $K^\varepsilon \in \mathcal{M}(\Xi \times \Xi, \mathcal{F} \times \mathcal{F})$ such that $K^\varepsilon(\Xi \times A) = P_0(A)$, $K^\varepsilon(A \times \Xi) = P(A)$, $\forall A \in \mathcal{F}$, and

$$\int_{(s_1 \times s_2) \in \Xi \times \Xi} d(s_1, s_2)\ K^\varepsilon(ds_1 \times ds_2) \leq \mathcal{W}\big(P_0, P\big) + \varepsilon. \tag{4}$$

Define $K_0 \in \mathcal{M}(\Xi \times \Xi, \mathcal{F} \times \mathcal{F})$ such that $K_0(A \times B) = P_0(A \cap B),\ \forall A, B \in \mathcal{F}$. By Proposition 3.4, we have

$$\int_{(s_1 \times s_2) \in \Xi \times \Xi} d(s_1, s_2) K_0(ds_1 \times ds_2) = 0. \tag{5}$$

Now define $\{P^n\}_{n=1}^{\infty} \subseteq \mathcal{M}(\Xi, \mathcal{F})$ as: $P^n := \lambda_n P + (1 - \lambda_n) P_0$ with $\lambda_n \in (0, 1)$ and $\lambda_n \to 1$. Define $K_n^\varepsilon := \lambda_n K^\varepsilon + (1 - \lambda_n) K_0$ as a probability measure in $\mathcal{M}(\Xi \times$

$\Xi, \mathcal{F} \times \mathcal{F})$. It is straightforward to verify that $K_n^\varepsilon(\Xi \times A) = P(A)$ and $K_n^\varepsilon(A \times \Xi) = P^n(A), \forall A \in \mathcal{F}$, using their definitions, which means the joint measure K_n^ε satisfies marginal conditions with respect to P and P^n. First, we need to verify that $\{P^n\}_{n=1}^\infty \subseteq \mathcal{P}'$. To see this, we have

$$\begin{aligned}
&\mathcal{W}(P^n, P_0) \le \int_{(s_1\times s_2)\in\Xi\times\Xi} d(s_1,s_2)\, K_n^\varepsilon(ds_1 \times ds_2) \\
&= \lambda_n \int_{\Xi\times\Xi} d(s_1,s_2)\, K^\varepsilon(ds_1 \times ds_2) + (1-\lambda_n) \int_{\Xi\times\Xi} d(s_1,s_2)\, K_0(ds_1 \times ds_2) \\
&\le \lambda_n \big[\mathcal{W}(P, P_0) + \varepsilon\big]. \qquad \text{using (4-5)} \qquad (6)
\end{aligned}$$

Since ε can be chosen arbitrarily, we set

$$\varepsilon = \min\left\{1,\ \frac{1}{2}\left(\frac{1}{\lambda_n} - 1\right)\mathcal{W}(P, P_0)\right\}. \qquad (7)$$

Substituting (7) into (6) yields $\mathcal{W}(P^n, P_0) \le (1/2 + \lambda_n/2)\mathcal{W}(P, P_0) < r_0$, hence $\{P^n\}_{n=1}^\infty \subseteq \mathcal{P}'$.

It remains to verify that $\lim_{n\to\infty} \mathcal{W}(P^n, P) = 0$. To see this, define $K \in \mathcal{M}(\Xi \times \Xi, \mathcal{F} \times \mathcal{F})$ such that $K(A \times B) = P(A \cap B)$, $\forall A, B \in \mathcal{F}$. By Proposition 3.4, we have

$$\int_{(s_1\times s_2)\in\Xi\times\Xi} d(s_1,s_2) K(ds_1 \times ds_2) = 0. \qquad (8)$$

Let $\widetilde{K}_n^\varepsilon := \lambda_n K + (1-\lambda_n)K^\varepsilon$ be a joint probability measure in $\mathcal{M}(\Xi \times \Xi, \mathcal{F} \times \mathcal{F})$. Then we have $\widetilde{K}^\varepsilon(\Xi \times A) = P(A)$ and $\widetilde{K}^\varepsilon(A \times \Xi) = P^n(A)$, $\forall A \in \mathcal{F}$, which means that $\widetilde{K}^\varepsilon$ satisfies marginal conditions with respect to P and P^n. It follows that

$$\begin{aligned}
&\mathcal{W}(P^n, P) \le \int_{(s_1\times s_2)\in\Xi\times\Xi} d(s_1,s_2)\, \widetilde{K}_n^\varepsilon(ds_1 \times ds_2) \\
&= \lambda_n \int_{\Xi\times\Xi} d(s_1,s_2)\, K(ds_1 \times ds_2) + (1-\lambda_n) \int_{\Xi\times\Xi} d(s_1,s_2)\, K^\varepsilon(ds_1 \times ds_2) \qquad (9) \\
&\le (1-\lambda_n)\big[\mathcal{W}(P, P_0) + \varepsilon\big]. \qquad \text{using (4),(8)}
\end{aligned}$$

Since $\lambda_n \to 1$, we have $\mathcal{W}(P^n, P) \to 0$ as $n \to \infty$. ∎

THEOREM 3.6. *Let Assumption 3.1 hold, and $\mathcal{P}$ be defined as in (1). For a given θ, the inner problem $\sup_{P\in\mathcal{P}} \mathbb{E}_P[h(\theta,\xi)]$ has a finite optimal value, and it is equivalent to the following conic linear program* (CLP):

$$\text{(CLP)} \quad \begin{array}{lll}
\sup\limits_{\mu} & \displaystyle\int_{s\in\Xi} h(\theta,s)\mu(ds \times \Xi) & \\
\text{s.t.} & \mu(\Xi \times \Xi^i) = 1/m, & i \in [m] \\
& \mu(\Xi \times \Xi^{m+1}) = 0, & \\
& \displaystyle\sum_{i\in[m]} \int_{s\in\Xi} d(s,s^i)\mu(ds \times \Xi^i) \le r_0, & \\
& \mu \succeq 0, &
\end{array}$$

where $\mu \succeq 0$ denotes that μ is a positive measure.

Proof. The inner problem in (WRO) can be written as:

$$\begin{array}{ll}\sup\limits_{P\in\mathcal{M}(\Xi,\mathcal{F})} & \int_{s\in\Xi} h(\theta,s)P(ds) \\ \text{s.t.} & \mathcal{W}(P,P_0)\le r_0.\end{array} \tag{10}$$

Let val(eqn#) denote the optimal value of a problem given by (eqn#). By Assumption 3.1 Ξ is compact, and $h(\theta,\cdot)$ is bounded in Ξ. Hence, the objective $\mathbb{E}_P[h(\theta,\xi)]$ is finite, and therefore val(10) is finite. We first show that val(10)=val(CLP). For this purpose we consider the following auxiliary problem:

$$\begin{array}{ll}\sup\limits_{P\in\mathcal{M}(\Xi,\mathcal{F})} & \int_{s\in\Xi} h(\theta,s)\ P(ds) \\ \text{s.t.} & \mathcal{W}(P,P_0)< r_0,\end{array} \tag{11}$$

whose feasible set is the interior of $\mathcal{P}$. We will prove that val(10) $\ge$ val(CLP) $\ge$ val(11), and val(10) = val(11), and hence val(10)=val(CLP).

We now show that val(CLP) $\ge$ val(11). Let $\widetilde{P}$ be a feasible solution of (11). Since we have $\mathcal{W}(\widetilde{P},P_0)<r_0$, by the definition of the Wasserstein metric in (2), there exists a $\widetilde{K}\in\mathcal{S}(\widetilde{P},P_0)$ satisfying:

$$\int_{\Xi\times\Xi} d(s_1,s_2)\ \widetilde{K}(ds_1\times ds_2)=\sum_{i\in[m]}\int_{s\in\Xi} d(s,s^i)\ \widetilde{K}(ds\times\Xi^i)\le r_0.$$

Therefore, $\widetilde{K}$ is a feasible solution of (CLP) with the objective value $\int_{s\in\Xi} h(\theta,s)\widetilde{K}(ds\times\Xi)$. Now observe that $\int_{s\in\Xi} h(\theta,s)\ \widetilde{K}(ds\times\Xi)=\int_{s\in\Xi} h(\theta,s)\ \widetilde{P}(ds)$, because $\widetilde{K}\in\mathcal{S}(\widetilde{P},P_0)$. Consequently, for any sequence $\{\widetilde{P}_k\}_1^\infty$ such that $\int_{s\in\Xi} h(\theta,s)\widetilde{P}_k(ds)\to$ val(11), there exist $\{\widetilde{K}_k\}_1^\infty$ satisfying $\int_{s\in\Xi} h(\theta,s)\ \widetilde{K}_k(ds\times\Xi)=\int_{s\in\Xi} h(\theta,s)\widetilde{P}_k(ds)$, hence $\int_{s\in\Xi} h(\theta,s)\ \widetilde{K}_k(ds\times\Xi)\to$ val(11). It follows that val(CLP) $\ge$ val(11).

Next we show that val(10) $\ge$ val(CLP). Suppose $\widehat{K}$ is a feasible solution of (CLP). Let $\widehat{P}\in\mathcal{M}(\Xi,\mathcal{F})$ be the marginal distribution of $\widehat{K}$ such that $\widehat{P}(A):=\widehat{K}(A\times\Xi)$, $\forall A\in\mathcal{F}$. Due to the constraints of (CLP), $\widehat{P}$ satisfies $\mathcal{W}(\widehat{P},P_0)\le r_0$; and hence $\widehat{P}$ is a feasible solution of (10). Because $\widehat{K}\in\mathcal{S}(\widehat{P},P_0)$, we have $\int_{s\in\Xi} h(\theta,s)\ \widehat{P}(ds)=\int_{s\in\Xi} h(\theta,s)\ \widehat{K}(ds\times\Xi)$. Consequently, for any sequence $\{\widehat{K}_k\}_1^\infty$ such that $\int_{s\in\Xi} h(\theta,s)\widehat{K}(ds\times\Xi)\to$ val(CLP), there exist a sequence $\{\widehat{P}_k\}_1^\infty$ satisfying $\int_{s\in\Xi} h(\theta,s)\ \widehat{P}_k(ds)=\int_{s\in\Xi} h(\theta,s)\ \widehat{K}_k(ds\times\Xi)$. It follows that $\int_{s\in\Xi} h(\theta,s)\ \widehat{P}_k(ds)\to$ val(CLP), and hence val(10) $\ge$ val(CLP).

We now show that val(10) = val(11). Since val(10) is finite, there exists a sequence of probability measures $\{P_m\}_{m=1}^\infty\subseteq\mathcal{P}$ such that

$$\lim_{m\to\infty}\int_{s\in\Xi} h(\theta,s)\ P_m(ds)=\text{val(10)}, \tag{12}$$

where $\mathcal{P}:=\{P\in\mathcal{M}(\Xi,\mathcal{F}):\mathcal{W}(P,P_0)\le r_0\}$. Let $\mathcal{P}':=\{P\in\mathcal{M}(\Xi,\mathcal{F}):\mathcal{W}(P,P_0)<r_0\}$ be the feasible set of (11). By Lemma 3.5, for any P_m ($m\ge1$), there exists a sequence $\{P_m^n\}_{n=1}^\infty\subseteq\mathcal{P}'$ such that $\lim_{n\to\infty}\mathcal{W}(P_m^n,P_m)=0$. Let $f(\theta,P):=\mathbb{E}_P[h(\theta,\xi)]$, then we have that val(10) $\ge$ val(11) $\ge f(\theta,P_m^n)$, $\forall m,n\ge1$. Moreover, we also have

that

$$\begin{aligned}
&\lim_{m\to\infty}\lim_{n\to\infty} |f(\theta, P_m^n) - \text{val}(10)| \\
&\leq \lim_{m\to\infty}\lim_{n\to\infty} \Big(|f(\theta, P_m^n) - f(\theta, P_m)| + |f(\theta, P_m) - \text{val}(10)| \Big) \\
&= 0, \qquad \text{[using Proposition 3.3 and (12)]}
\end{aligned}$$

which implies that val(11)=val(10). ∎

Note that the infimum used in (CLP) is over the joint measure μ. Moreover, in (CLP) both the objective and constraints are in linear form of the measure μ, which is a 'new decision variable' in the inner problem. Theorem 3.7 provides a dual of (CLP) as a semi-infinite linear program in the linear space of signed measures using conic programming duality. This theorem shows that the dual problem has a formulation which decomposes in ξ^i, $i \in [m]$.

THEOREM 3.7. *Let Assumption 3.1 hold. The dual of* (CLP) *can be written as the following semi-infinite program:*

$$\text{(CLP-D)} \qquad \begin{array}{ll} \min\limits_{v\in\mathbb{R}^{m+1}} & \dfrac{1}{m}\displaystyle\sum_{i=1}^{m} v_i + r_0 \cdot v_{m+1} \\ \text{s.t.} & h(\theta, s) - v_i - v_{m+1} \cdot d(s, \xi^i) \leq 0, \quad s \in \Xi, i \in [m] \\ & v_1, \ldots, v_m \in \mathbb{R}, \quad v_{m+1} \geq 0. \end{array}$$

Furthermore, strong duality holds, i.e., val(CLP)=val(CLP-D). *Additionally, for $r_0 > 0$ the optimum solution of (CLP-D) can be bounded by the following polytope:*

$$\begin{aligned} \mathcal{H} := \Big\{ v \in \mathbb{R}^{m+1} : & C_1 \leq v_i \leq (m+1)C_2 - mC_1, \ \textit{for } i \in [m], \\ & 0 \leq v_{m+1} \leq (C_2 - C_1)/r_0 \Big\}, \end{aligned} \tag{13}$$

where C_1 and C_2 are lower and upper bounds of $h(\cdot,\cdot)$ on $\Theta \times \Xi$, respectively.

Proof. We note that (CLP) can be rewritten as:

$$\begin{array}{ll} \sup\limits_{\mu} & \displaystyle\int_{(s_1\times s_2)\in\Xi\times\Xi} h(\theta, s_1)\ \mu(ds_1 \times ds_2) \\ \text{s.t.} & \displaystyle\int_{(s_1\times s_2)\in\Xi\times\Xi} \mathbf{1}_{\Xi^i}(s_2)\ \mu(s_1 \times s_2) = 1/m, \qquad i \in [m] \\ & \displaystyle\int_{(s_1\times s_2)\in\Xi\times\Xi} \mathbf{1}_{\Xi^{m+1}}(s_2)\ \mu(s_1 \times s_2) = 0, \\ & \displaystyle\int_{(s_1\times s_2)\in\Xi\times\Xi} \Big[\sum_{i\in[m]} d(s_1, s_2) \cdot \mathbf{1}_{\Xi^i}(s_2) \Big]\ \mu(ds_1 \times ds_2) \leq r_0, \\ & \mu \succeq 0, \end{array} \tag{14}$$

where $\mathbf{1}_{\Xi^i}(s)$ is an indicator function. Note that the second constraint in (14) is defined on the set $\Xi^{m+1} = \Xi \setminus (\cup_{i\in[m]}\Xi^i)$. For a given θ, we define functions $\{\psi_j\}_{j=0}^{m+2}$ as follows:

$$\psi_j(s_1, s_2) = \begin{cases} h(\theta, s_1) & j = 0 \\ \mathbf{1}_{\Xi^j}(s_2) & j \in [m+1] \\ \sum_{i=1}^{m} d(s_1, s_2) \cdot \mathbf{1}_{\Xi^i}(s_2) & j = m+2. \end{cases} \tag{15}$$

Clearly $\{\psi_j\}_{j=0}^{m+2}$ are bounded $\mathcal{F}\times\mathcal{F}$-measurable functions on $\Xi\times\Xi$. It follows that $\{\psi_j\}_{j=0}^{m+2}$ are μ-integrable for any $\mu\succeq 0$. We will first put (14) in a standard form of conic linear program. Let $\mathcal{X}$ be the linear space of *finite signed measures*, and $\mathcal{X}^+ := \{\mu\in\mathcal{X}:\ \mu\succeq 0\}$ be the set of non-negative measures which is a convex cone in $\mathcal{X}$. Let $\mathcal{X}'$ be the set of functions that are μ-integrable for all $\mu\in\mathcal{X}^+$, i.e., $\mathcal{X}' := \{f:\Xi\times\Xi\to\mathbb{R}\mid f\in L^1(\Xi\times\Xi,\mu),\ \forall\mu\in\mathcal{X}^+\}$. Define $\langle\mu,f\rangle := \int_{(s_1\times s_2)\in\Xi\times\Xi} f(s_1,s_2)\mu(ds_1\times ds_2),\ \forall\mu\in\mathcal{X}, f\in\mathcal{X}'$. Define the linear operator $\mathcal{A}:\mathcal{X}\to\mathbb{R}^{m+2}$ as

$$\mathcal{A}\mu := \big[\langle\mu,\psi_1\rangle,\ldots,\langle\mu,\psi_{m+2}\rangle\big]^T. \tag{16}$$

Define a vector b as $b = \big[\underbrace{1/m,\ldots,1/m}_{m},0,r_0\big]^T\in\mathbb{R}^{m+2}$, and a convex cone $\mathcal{K} := \mathbf{0}^{m+1}\times(-\infty,0]$ in $\mathbb{R}^{m+2}$. Using the above notations, (14) can be rewritten as:

$$\begin{aligned}\sup_{\mu}\quad & \langle\mu,\psi_0\rangle\\ \text{s.t.}\quad & \mathcal{A}\mu - b\in\mathcal{K}\\ & \mu\in\mathcal{X}^+.\end{aligned} \tag{CLP1}$$

Using Lemma A.1(a) in Appendix A from [55], we have the following dual of (CLP1):

$$\begin{aligned}\inf_{w}\quad & -b^T\cdot w\\ \text{s.t.}\quad & \mathcal{A}^*w+\psi_0\in-(\mathcal{X}^+)^*\\ & w\in\mathcal{K}^*,\end{aligned} \tag{17}$$

where $\mathcal{K}^*=\mathbb{R}^{m+1}\times(-\infty,0]$ is the dual cone of $\mathcal{K}$, and $-(\mathcal{X}^+)^* = \{f\in\mathcal{X}':\ \langle\mu,f\rangle\le 0,\forall\mu\succeq 0\}$ is the polar cone of $\mathcal{X}^+$. The adjoint operator $\mathcal{A}^*$ acts on $\mathbb{R}^{m+2}$ as: $\mathcal{A}^*v=\sum_{i=1}^{m+2}v_i\psi_j,\ \forall v\in\mathbb{R}^{m+2}$. Therefore, the dual problem (17) can be expressed as follows:

$$\begin{aligned}\min_{w}\quad & -\frac{1}{m}\sum_{i=1}^{m}w_i - r_0\cdot w_{m+2}\\ \text{s.t.}\quad & \langle\mu,\psi_0\rangle+\sum_{i=1}^{m+2}w_i\langle\mu,\psi_i\rangle\le 0,\quad\forall\mu\in\mathcal{X}^+,\\ & w_{m+2}\le 0.\end{aligned} \tag{18}$$

We have

$$\begin{aligned}\sum_{i=1}^{m+2}w_i\langle\mu,\psi_i\rangle+\langle\mu,\psi_0\rangle = &\sum_{i=1}^{m}\int_{\Xi\times\Xi^i}\big[w_i+w_{m+2}\cdot d(s_1,\xi^i)+h(\theta,s_1)\big]\mu(ds_1\times ds_2)\\ &+\int_{\Xi\times\Xi^{m+1}}\big[w_{m+1}+h(\theta,s_1)\big]\mu(ds_1\times ds_2).\end{aligned} \tag{19}$$

Since $\mathcal{F}$ contains all singletons in Ξ, all integrands on the RHS of (19) must be non-positive to ensure the constraint $\sum_{i=1}^{m+2}w_i\langle\mu,\psi_i\rangle+\langle\mu,\psi_0\rangle\le 0,\ \forall\mu\in\mathcal{X}^+$. Otherwise, we concentrate the measure μ on the points $(\hat{s}\times\xi^i)$, $i\in[m]$, or the set $(\hat{s},\Xi^{m+1})$ at which one of the integrands is positive to achieve a contradiction. In other words, $\sum_{i=1}^{m+2}w_i\langle\mu,\psi_i\rangle+\langle\mu,\psi_0\rangle\le 0,\ \forall\mu\in\mathcal{X}^+$ are equivalent to the following constraints:

$$w_i + w_{m+2}\cdot d(s,\xi^i)+h(\theta,s)\le 0,\quad\forall s\in\Xi\quad i\in[m], \tag{20}$$

$$w_{m+1}+h(\theta,s)\le 0,\quad\forall s\in\Xi. \tag{21}$$

Since w_{m+1} does not appear in the objective function of (18), and constraint (21) can be satisfied by choosing w_{m+1} sufficiently small, we can remove w_{m+1} and constraint (21) from the dual problem. By rewriting variables w_i as $-v_i$ for $i \in [m]$ and w_{m+2} as $-v_{m+1}$, the dual problem (18) can be rewritten as (CLP-D).

We now show that strong duality holds. We only need to verify that val(CLP) is finite and show that Sol(CLP-D) is nonempty and bounded [55] (see Lemma A.1(b), in Appendix A). First, since the objective in (CLP) is an expectation of a bounded function, val(CLP) is finite. Second, we note that the feasible set of (CLP-D), denoted as Fsb(CLP-D), is a closed convex subset in $\mathbb{R}^{m+1}$. Since $C_1 \le h \le C_2$ on $\Theta \times \Xi$, $v = \big[\underbrace{C_2, \ldots, C_2}_{m}, 0\big]^T$ is a feasible solution of (CLP-D), we have val(CLP-D) $\le C_2$. Moreover, substituting $s = \xi^i$ in the i^{th} constraint of (CLP-D), we have $v_i \ge h(\theta, \xi^i) \ge C_1,\ i \in [m]$, which means that C_1 is a lower bound on v_i, $i \in [m]$. To show that Sol(CLP-D) is bounded, consider any $v^* \in$ Sol(CLP-D). Since val(CLP-D) $\le C_2$, we have $\frac{1}{m}\sum_{i=1}^m v_i^* + r_0 \cdot v_{m+1}^* \le C_2$. It follows that

$$
\begin{aligned}
&v_{m+1}^* \le \frac{1}{r_0}\Big(C_2 - \frac{1}{m}\sum_{i=1}^m v_i^*\Big) \le \frac{1}{r_0}(C_2 - C_1),\\
&v_i^* \le m(C_2 + r_0 \cdot v_{m+1}^*) - \sum_{j=1}^{i-1} v_j^* - \sum_{j=i+1}^{m} v_j^* \le (m+1)C_2 - mC_1, \quad i \in [m],
\end{aligned} \tag{22}
$$

which provide upper bounds on v_i^*, $i \in [m+1]$. Therefore, Sol(CLP-D) is bounded by the following compact set:

$$
\begin{aligned}
\mathcal{H} := \Big\{ v \in \mathbb{R}^{m+1} : {} & C_1 \le v_i \le (m+1)C_2 - mC_1, \text{ for } i \in [m],\\
& 0 \le v_{m+1} \le (C_2 - C_1)/r_0 \Big\}.
\end{aligned} \tag{23}
$$

Now Fsb(CLP-D)$\cap\mathcal{H}$ is a non-empty compact set in $\mathbb{R}^{m+1}$, and Sol(CLP-D) is a subset of Fsb(CLP-D) $\cap\ \mathcal{H}$. By Weierstrass Theorem [8], there exists an optimal solution to (CLP-D). Therefore, Sol(CLP-D) is nonempty, bounded, and attains its optimum at a solution in $\mathcal{H}$. ∎

COROLLARY 3.8. *Let Assumption 3.1 hold. The Wasserstein-robust optimization problem* (WRO) *is equivalent to the following semi-infinite program:*

$$
\begin{aligned}
\min_{\theta, v} \quad & \frac{1}{m}\sum_{i=1}^m v_i + r_0 \cdot v_{m+1}\\
\text{s.t.} \quad & h(\theta, s) - v_i - v_{m+1} \cdot d(s, \xi^i) \le 0, \quad s \in \Xi, i \in [m]\\
& \theta \in \Theta, v \in \mathcal{H}.
\end{aligned} \tag{WRO-D}
$$

Proof. From Theorem 3.7, the inner problem of (WRO) is reformulated as a minimization problem with semi-infinite constraints. We can now combine (CLP-D) with the outer problem of (WRO), and have an equivalent combined formulation as (WRO-D). Since from Theorem 3.7 the optimal solution of the inner problem is bounded in the polytope $\mathcal{H}$, these additional constraints are added to (WRO-D). ∎

We note that the constraints in (WRO-D) decompose in the scenarios ξ^i, $i \in [m]$; and for a given s, $d(s, \xi^i)$ is a constant.

4. Algorithms for the (WRO) Refomulation. Corollary 3.8 shows that the Wasserstein-robust optimization problem (WRO) is equivalent to (WRO-D), which is a semi-infinite program. Any algorithm for solving a general semi-infinite program can now be applied to solve (WRO-D). For a general continuous function $h(\theta, s)$ in θ we present a modification of the exchange algorithm [29] in Section 4.1, which ensures ε-optimality after solving a finite number of finitely constrained master problems. In the special case where $h(\theta, s)$ is a convex function of θ, (WRO-D) is a convex semi-infinite program. We adapt the cutting surface algorithm in [43] for (WRO-D) and use its structure to achieve a global linear rate of convergence.

Let $x = [\theta, v]$ be the decision variables, and define the following functions:

$$f(x) := \frac{1}{m}\sum_{i=1}^{m} v_i + r_0 \cdot v_{m+1}, \tag{24}$$
$$g_i(x, s) := h(\theta, s) - v_i - v_{m+1} \cdot d(s, \xi^i), \quad i \in [m].$$

The problem (WRO-D) can be rewritten as:

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{s.t.} & g_i(x, s) \le 0, \quad s \in \Xi, i \in [m] \\ & x \in X, \end{array} \tag{SIP}$$

where $X = \Theta \times \mathcal{H}$. Problem (SIP) is a semi-infinite program. An approach to obtain a solution of such problems is to solve relaxation problems (master problems) with a finite number of constraints, and add a violated constraint obtained from solving a separation problem (defined for $\forall\ i \in [m]$) to tighten the formulation iteratively. In particular, the separation problem of (SIP) for identifying a violated constraint at the solution $(\widetilde{x}, \widetilde{v})$ of the current master problem can be written as follows:

$$\max_{s \in \Xi} g_i(\widetilde{x}, s) = h(\widetilde{\theta}, s) - \widetilde{v}_i - \widetilde{v}_{m+1} \cdot d(s, \xi^i), \quad \text{for } i \in [m]. \tag{Sep-i}$$

The inequality generated from solving (Sep-i) is called a feasibility cut.

For clarity of notations, we consider the following general form of a semi-infinite program:

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{s.t.} & g(x, t) \le 0, \quad t \in T, \\ & x \in X, \end{array} \tag{gen-SIP}$$

where $X \subseteq \mathbb{R}^{k_1}$ and $T \subseteq \mathbb{R}^{k_2} \times \mathbb{Z}^{k_3}$, allowing that T may be defined as a mixed-integer set.

4.1. A modified exchange algorithm for (WRO) model. We now assume that we have an oracle to solve the master and separation problems of (gen-SIP). The modified exchange algorithm given in Algorithm 1 allows an ε-optimal solution to (gen-SIP), when compared with the original exchange algorithm. Theorem 4.2 shows that our modified exchange method finds a solution of a desired accuracy in finitely many iterations.

DEFINITION 4.1. *A point $z_0 \in Z$ is an ε-feasible solution of* (gen-SIP) *if* $\max\limits_{t \in T} g(z_0, t) \le \varepsilon$. *A point $z_0 \in Z$ is an ε-optimal solution of* (gen-SIP) *if z_0 is ε-feasible and $f(z_0) \le$* val(gen-SIP).

THEOREM 4.2. *Let $Z \times T$ be compact, and $g(z,t)$ be continuous on $Z \times T$. Suppose we have an oracle that generates an optimal solution of the problem $\min_{z \in Z} \{f(z) : g(z,t) \le 0,\ t \in T'\}$ for any finite set $T' \subseteq T$, and an oracle that generates an ε-optimal solution of the problem $\max_{t \in T} g(z,t)$ for any $z \in Z$ and $\varepsilon > 0$. Then Algorithm 1 returns an ε-optimal solution of (gen-SIP) in finitely many iterations.*

Proof. Since $g(z,t)$ is continuous on $Z \times T$, and $Z \times T$ is compact, it follows that $g(z,t)$ is uniformly continuous on $Z \times T$. Therefore, there exists an $\alpha > 0$ such that

$$|g(z',t') - g(z,t)| \le \frac{\varepsilon}{2}, \quad \text{if } \|z' - z\| + \|t' - t\| \le \alpha. \tag{25}$$

First, we prove by contradiction that the algorithm terminates in finitely many iterations. Suppose the algorithm generates infinite sequences $\mathcal{Z} = \{z_k\}_0^\infty$ and $\mathcal{T} = \{t_k\}_0^\infty$ without terminating. We claim that $t_{k+1} \notin \cup_{i=1}^k B(t_i, \alpha)$ for every k, where $B(t_i, \alpha)$ is the closed ball of center t_i and radius α in $\mathbb{R}^{k_2+k_3}$. If not, there exists t_k such that $t_k \in B(t_i, \alpha)$ for some $i < k$. Using (25) and $t_i \in T_{z-1}$, we have $g(z_{k-1}, t_k) \le g(z_{k-1}, t_i) + \frac{\varepsilon}{2} \le \frac{\varepsilon}{2}$, indicating that the termination criteria is satisfied. Therefore, the above claim holds. Now consider the compact set: $\mathcal{B} = \cup_{\{t \in T\}} B(t, \alpha)$. Since $B(t_k, \alpha) \subseteq \mathcal{B}$ for every k, and by the claim we have $B(t_i, \alpha) \cap B(t_j, \alpha) = \emptyset$ for every $t_i, t_j \in \mathcal{T}$ with $i \ne j$, it follows that $\sum_{i=0}^\infty \mathrm{vol}(B(t_i, \alpha)) \le \mathrm{vol}(\mathcal{B})$, which leads to a contradiction.

We now prove that once the algorithm terminates, it returns an ε-solution. Suppose it terminates at the end of the nth iteration. Then z_n is an optimal solution of the problem $\min_{z \in Z} \{f(z) : g(z,t) \le 0,\ t \in T_n\}$. It follows that $f(z_n) \le \mathrm{val}(T_n) \le \mathrm{val}$(gen-SIP), where $\mathrm{val}(T_n)$ is the optimal value of the problem (gen-SIP) with constraint index set T_n. By the separation oracle and termination criteria, it also holds that $\max_{t \in T} g(z_k, t) \le g(z_n, t_{n+1}) + \frac{\varepsilon}{2} \le \varepsilon$. Hence, z_n is an ε-optimal solution of (gen-SIP). ∎

Algorithm 1 A modified exchange algorithm to solve (gen-SIP).

Prerequisites: Two oracles specified in Theorem 4.2.
Output: An ε-optimal solution of (gen-SIP).
Step 1 Set $T_0 \leftarrow \emptyset$, $k \leftarrow 0$.
Step 2 Determine an optimal solution z_k of the problem $\min_{z \in Z} \{f(z) : \text{s.t. } g(z,t) \le 0,\ t \in T_k\}$.
Step 3 Determine a $\frac{\varepsilon}{2}$-optimal solution t_{k+1} of the problem $\max_{t \in T} g(z_k, t)$. If $g(z_k, t_{k+1}) \le \frac{\varepsilon}{2}$, stop and return z_k; otherwise let $T_{k+1} \leftarrow T_k \cup \{t_{k+1}\}$, $k \leftarrow k+1$ and go to Step 2

4.2. A central cutting-surface algorithm for the convex case. In this section we make the following additional assumption.

ASSUMPTION 4.3. *The feasible region Θ is convex, and the function $h(\cdot, s)$ is convex for every $s \in \Xi$. Furthermore, there exists a parameter B satisfying the following condition:*

$$B > \|\eta\|, \quad \forall\, \eta \in \partial_\theta h(\theta, s)\ \ \forall\, \theta \in \Theta,\ s \in \Xi, \tag{26}$$

where $\partial_\theta h(\theta, s)$ is the sub differential set of $h(\theta, s)$ at θ.

Since the master problem of (SIP) is a convex optimization problem, we assume it can be solved efficiently up to optimality. We now present a central cutting-surface algorithm to solve (SIP). A pseudo-code for this algorithm is given in Algorithm 2. The algorithm is initialized with a solution $x^{(0)} = [\theta^0, \mathbf{0}_{m+1}]$, where θ^0 may be taken as a solution of the empirical deterministic optimization problem:

$$\text{(EDO)} \qquad \min_{\theta \in \Theta} \frac{1}{m} \sum_{i=1}^{m} h(\theta, \xi^i).$$

At the kth iteration of this algorithm (Step 1) a master problem with a centering argumentation is solved. This problem is defined by a set of constraints (with the index set $Q^{(k-1)} \subseteq \Xi$) inherited from the $(k-1)$th iteration. The master problem at the kth iteration is formulated as follows:

$$\text{(Master)} \qquad \begin{array}{ll} \max\limits_{x,w,t} & w \\ \text{s.t.} & t + w \le M^{(k-1)}, \\ & f(x) - t + B \cdot w \le 0, \\ & g_i(x,s) + B \cdot w \le 0, \quad \forall i \in [m], \;\; s \in Q^{(k-1)}, \\ & x \in X. \end{array}$$

The (Master) problem is a convex optimization problem. For clarity, we drop the index i in (Master) in Algorithm 2. We assume that there is an oracle to find an ε-optimal solution to (Sep-i) for any $\varepsilon > 0$. The algorithm terminates if no feasibility cut is found. Otherwise, the newly found feasibility cut is added to the working set (Step 4). At the end of each iteration, we may optionally drop certain constraints that are not binding at the current solution of the master problem (Step 6).

Algorithm 2 A central cutting-surface algorithm from [43] to solve (SIP).

Prerequisites: Assumptions 3.1 and 4.3 hold. There exists an oracle to find an ε-optimal solution to (Sep-i).
Input: A strict upper bound U of the objective function h, a centering parameter $B > 0$ which satisfies (26), a tolerance error ε, an arbitrary $\alpha > 1$ specifying how aggressively cuts are dropped.
Output: An ε-optimal solution to (SIP).
Step 1 (Initialization). Set $k \leftarrow 1$, $M^{(0)} \leftarrow U$, $v^{(0)} \leftarrow \mathbf{0}_{m+1}$, $x^{(0)} \leftarrow [\theta^0, \mathbf{0_{m+1}}]$, $\widetilde{x}^{(0)} \leftarrow x^{(0)}$, where θ^0 is a solution to (EDO). Let $Q^{(0)} \leftarrow \{s^{(0)}\}$, where $s^{(0)} = \xi$.
Step 2 (Solve a master problem). Determine the optimal solution $(x^{(k)}, w^{(k)})$ to (Master).
Step 3 (Optimal soluton?). If $w^{(k)} = 0$, stop and return $\widetilde{x}^{(k-1)}$.
Step 4 (Feasible solution?). Find an ε-optimal solution denoted by $s^{(k)}$ to (Sep-i) using the oracle. If $g(x^{(k)}, s^{(k)}) > 0$ and go to Step 5, otherwise go to Step 6.
Step 5 (Add a cut). Set $Q^{(k)} \leftarrow Q^{(k-1)} \cup \{s^{(k)}\}$, $\widetilde{x}^{(k)} \leftarrow \widetilde{x}^{(k-1)}$ and $M^{(k)} \leftarrow M^{(k-1)}$. Go to Step 6.
Step 6 (Update best know ε-feasible solution). Set $Q^{(k)} \leftarrow Q^{(k-1)}$, $\widetilde{x}^{(k)} \leftarrow x^{(k)}$ $M^{(k)} \leftarrow f(x^{(k)})$.
Step 7 (Optionally drop cuts). Let $D = \left\{ s^{(l)} \in Q^{(k)} : \; l \in \{0\} \cup [k] \;\middle|\; w^{(l)} \ge \alpha w^{(k)} \text{ and } g(x^{(k)}, s^{(l)}) + B \cdot w^{(k)} < 0 \right\}$ and set $Q^{(k)} \leftarrow Q^{(k)} \setminus D$. Set $k \leftarrow k + 1$ and go to Step 2.

The convergence of Algorithm 2 is given by Theorem 4.4. This theorem is a refinement of the linear rate of convergence result proved in Theorem 8 of [43] (See Appendix A) for the central cutting surface algorithm when specialized to (WRO-D).

THEOREM 4.4. *Convergence of the central cutting-surface algorithm.*

(a) *Algorithm 2 either finds an ε-optimal solution of (SIP) in finitely many iterations or generates $\{\widetilde{x}^{(k)}\}_{k=1}^{\infty}$ that each accumulated point is an ε-optimal solution of (SIP).*

(b) *Algorithm 2 converges linearly in objective function value between consecutive feasibility cuts. The rate of convergence satisfies:*

$$\rho \leq 1 - \frac{1}{2B'+1}, \tag{27}$$

where $B' := \max\{\sqrt{r_0^2+1+(1/m)},\ \sqrt{B^2+L^2+1}\}$ and $L = \max\{d(s,\xi^i) : s \in \Xi,\ i \in [m]\}$.

Proof. Note that (SIP) is equivalent to the following reformulation:

$$\begin{array}{ll} \min\limits_{x,t} & t \\ \text{s.t.} & f(x) - t \leq 0, \\ & g_i(x,s) \leq 0, \quad \forall s \in \Xi,\ \forall i \in [m] \\ & x \in X. \end{array} \tag{28}$$

Treating t as x_0 in (38), we now verify that (28) satisfies Assumption A.2 in Appendix A. Since $\Theta \times \Xi$ is bounded and $h(\cdot,\cdot)$ is continuous, $\exists C_1, C_2$ such that $h(\cdot,\cdot) \in [C_1, C_2]$ on $\Theta \times \Xi$, and the objective value of (WRO) is in $[C_1, C_2]$. Also, since there is no duality gap, we can set the dual objective $C_1 \leq f(x) \leq C_2$ for all $x \in X$, and set $t \in [C_1, C_2]$, without affecting the optimal solution and the optimal value. Hence, Assumption A.2 (1) is satisfied. To verify Assumption A.2 (2), we note that for any $\eta > 0$, we can verify that $[\overline{t}, \overline{x}]$ is a Slater point of (28), where $\overline{t} = C_2 + 2\eta$ and $\overline{x} = [\theta^0, (C_2+\eta)\mathbf{1_m}, 0]$ ($\mathbf{1_m}$ is the m-dimensional vector with all entries being 1). Now we show that Assumption A.2 (3) is also satisfied. Now let us take a subgradient of $f(x) - t$ and $g_i(x,s)$ ($\forall i \in [m]$) with respect to the decision variables $[t, x]$, and use Assumption 4.3, we can set the centerality parameter B' to be $B' := \max\{\sqrt{r_0^2+1+(1/m)},\ \sqrt{B^2+L^2+1}\}$, where $L = \max\{d(s,\xi^i) :\ s \in \Xi,\ i \in [m]\}$. The oracle is assumed to be given based on the prerequisites of Algorithm 2, hence Assumption A.2 (4) is satisfied.

We now apply Theorem A.3 on the semi-infinite program (28) and the corresponding master problem (Master). Part (a) directly follows from Theorem A.3 (a)-(c). By Theorem A.3 (d), for $k \geq \widehat{k}$ iterations, where $w^{(\widehat{k})} < \eta/B$, we have

$$\rho^{(k)} \leq 1 - \frac{\eta - B'w^{(k)}}{\eta + B'(\overline{t} - f^*)}, \tag{29}$$

where f^* is the optimal value of (28). Since (29) holds for every $\eta > 0$, we can select an η to maximize $\frac{\eta - B'w^{(k)}}{\eta + B'(\overline{t}-f^*)}$, hence minimizing the upper bound of $\rho^{(k)}$ in Theorem A.3. Let $F(\eta) := \frac{\eta - B'w^{(k)}}{\eta + B'(\overline{t}-f^*)}$, and substitute $\overline{t} = C_2 + 2\eta$ in $F(\eta)$, we have

$$F(\eta) = \frac{\eta - B'w^{(k)}}{(2B'+1)\eta + B'(C_2 - f^*)}, \quad F'(\eta) = \frac{B'w^{(k)}(2B'+1) + B'(C_2 - f^*)}{[(2B'+1)\eta + B'(C_2 - f^*)]^2}. \tag{30}$$

Since $w^{(k)} > 0$, $f^* \leq C_2$, we have $F'(\eta) > 0$ for all $\eta > 0$. Therefore, the maximum value of $F(\cdot)$ is:

$$\max_{\eta>0} F(\eta) = F(\infty) = \frac{1}{2B'+1}. \tag{31}$$

It follows that the uniform rate of convergence satisfies: $\rho \leq 1 - \frac{1}{2B'+1}$. ∎

4.3. Computational tractability of the separation problem. We now discuss the computational difficulty of solving (Sep-i), which depends on the function form of $h(\theta, s)$ in s for a given θ, the metric d, and the uncertainty set Ξ. Since $\widetilde{v}_{m+1} \geq 0$ and in most applications $d(\cdot,\cdot)$ is chosen to be a vector norm, the term $-\widetilde{v}_i - \widetilde{v}_{m+1} d(s, \xi^i)$ in (Sep-i) is concave in s. Therefore, in the case where $h(\theta, s)$ is concave in s, and Ξ is a convex set, (Sep-i) becomes a convex optimization problem. For a very general case where $h(\theta,\cdot)$ and $d(\xi^i,\cdot)$ are continuously differentiable over the compact (not necessarily convex) uncertainty set Ξ, drawing a sufficiently many independent uniform samples $s^t \in \Xi$ and verifying if objective value of (Sep-i) is greater than ε can either identify a violated constraint (not necessarily the most violated constraint), or conclude that the solution of the current master problem is ε-optimal with high probability (See [43] Section 5.2). Furthermore, for cases where $h(\theta,\cdot)$ is a polynomial function, $d(\cdot,\cdot)$ is an Euclidean norm, and Ξ is specified by polynomial inequalities (e.g., an ellipsoid), (Sep-i) falls in to the category of polynomial optimization. The global optimal solution in such cases can be obtained by solving a sequence of SDP relaxations (a primal approach) [35] or a sequence of SOS relaxations (a dual approach) [48].

For some important models from statistical learning such as linear regression, linear support vector machine, logistic regression, etc., the loss function h has the form $h(\theta_0 + \theta^T x, y)$, where x is the feature vector and y is the response value. For this case, (Sep-i) can be solved efficiently using a branch-and-bound scheme based on piece-wise linear approximations of $h(\theta_0 + \theta^T x, y)$. For more details about this approach, see [42].

5. Numerical Experiments. We investigate the following Wasserstein-robust logistic regression (WRLR) model as a numerical example to illustrate our algorithmic ideas and the performance of (WRO):

$$\min_{[\theta_0,\theta]\in\Theta} \max_{P\in\mathcal{P}} \mathbb{E}_P\Big[\log\big(1+\exp[-y(\theta_0+\theta^T x)]\big)\Big], \tag{WRLR}$$

where $x \in \mathbb{R}^n$ is the feature vector, and $y \in \{0,1\}$ is the label. The semi-infinite reformulation of (WRLR) is written as follows:

$$\begin{aligned} \min_{\theta_0,\theta,v}\ & \frac{1}{m}\sum_{i=1}^{m} v_i + r_0 \cdot v_{m+1} \\ \text{s.t.}\ & \log\big(1+\exp\big[-y(\theta_0+\theta^T x)\big]\big) - v_i - v_{m+1}\cdot d(s,\xi^i) \leq 0,\ s\in\Xi, i\in[m], \\ & [\theta_0,\theta]\in\Theta, v\in\mathcal{H}, \end{aligned} \tag{32}$$

where $s = [x, y]$, and $\xi^i = [x^i, y^i]$ is the ith ($i \in [m]$) observed sample. We assume that only the feature vector x has uncertainty but not the label y. Therefore, the uncertainty set Ξ can be written as $\Xi = \Xi_0 \cup \Xi_1$, where Ξ_0, Ξ_1 are the uncertainty sets for feature vectors with the label 0, 1 respectively. The uncertainty set Ξ_0 is defined as

Table 1: Summary of data sets from UCI Machine Learning Repository

Data set	Area	No. Attrib.	No. Observ.
BA	Finance	4	1372
VC	Health care	6	310
PID	Health care	8	768
BCW	Health care	9	699
ST-H	Health care	13	270
EES	Health care	14	14980
SPT-H	Health care	22	267
ION	Aerospace	34	351
SPTF-H	Health care	44	267
SPAM	Computer	57	4601
CB	Aerospace	60	208

an n-dimensional hyper-rectangle such that each dimension corresponds to an interval $[\overline{x}_j \pm \sigma_j]$ for the feature x_j, where $\overline{x}_j$ is the sample mean of observations with label 0, and σ_j is the sample standard deviation. The uncertainty set Ξ_1 is defined similarly. We used the l_1 norm to define the metric d on Ξ_0 (Ξ_1), i.e., $d(x, x') = \|x - x'\|_1$.

5.1. Numerical setup. We present computational effort in solving the semi-infinite dual problem (WRO-D) of (WRLR) using the cutting-surface algorithm (See Section 4.2). We also present the out of sample predictive performance of the (WRLR) model compared with the ordinary logistic regression model (LR). The algorithm for solving (WRO-D) were implemented in C++, and the computational tests were performed on an Intel Xeon CPU with 4 GB of RAM. The cutting-surface algorithm for (WRO-D) consists of iteratively solving the master problem (Master) and the separation problem (Sep-i). The convex master problem (Master) is solved using Ipopt 3.12.4 [59] which implements a primal-dual interior point method. The separation problem (Sep-i) is solved using the branch-and-bound scheme based on sequentially piece-wise linear approximating $h_\theta(u) := \log(1 + e^{-u})$, and each convex optimization subproblem is solved using CPLEX 12.6.3. We used 11 data sets (those with less than 60 features) from the UCI machine learning repository in our computational testing, which are: Banknote authentication (BA), Vertebral column (VC), Pima Indians diabetes (PID), Breast cancer Wisconsin (BCW), Statlog heart (ST-H), EEG eye state (EES), SPECT heart (SPT-H), Ionosphere (ION), SPECTF heart (SPTF-H), Spambase (SPAM), Connectionist bench (CB). A summary of these datasets is given in Table 1. We now describe the data generation for our test problems. We chose m ($m = 50, 75, 100, 150$) observations from each of the UCI data sets. We kept the class labels of the chosen observations unchanged.

5.2. Computational effort in solving WRLR. The computational effort in solving the semi-infinite dual problem (WRO-D) of (WRLR) for 11 data sets is given in Table 2 for different choices of m. The numbers are averaged over the 100 experiments. Columns 4-8 give the number of outer iterations in Algorithm 2, total number of constraints (cuts) added to the master problem at termination, the CPU time for solving a problem instance, and the percentage of time spend in solving the master and separation problems, respectively. We also provide the CPU time for solving (LR) in Column 3. The results show that the number of calls to the master problem

is approximately $4 \sim 40$ when solving (WRLR). Over these calls approximately $15m \sim 50m$ cutting surfaces are added. In other words, approximately $15m \sim 50m$ artificial samples are identified. For data sets with more features, the program spends a larger fraction of time on solving master problems since their scale becomes larger. The computational time of solving (LR) models is less than 1 second for data sets with feature size less than 20, and for data sets with feature size between $30 \sim 60$ the computational time is less than 20 seconds. The average time of solving (WRLR) models is $\lesssim 100$ times that of solving the (LR) models.

5.3. Predictive performance of the WRLR model. We now compare the predictive performance of the (WRLR) model with the (LR) model. For each data set, we randomly select m samples out of all the observed samples to train both models. For each combination of data set and the training sample size m, we performed 100 experiments. Both trained (LR) and (WRLR) models are used to predict the remaining observations from the data set, and the corresponding AUC values (area under the ROC curve) are recorded. AUC value is the most popular metric used for evaluating the performance of a model used in medical literature. In each experiment, training samples are selected randomly and independently. The mean AUC values ($\overline{\text{AUC}}$) over 100 experiments and the standard errors are listed in Table 2. The p-values in this table are based on the hypothesis test: $H_0 : \overline{\text{AUC}}^{\text{OOS}}_{\text{WRLR}} \leq \overline{\text{AUC}}^{\text{OOS}}_{\text{LR}}$ versus $H_1 : \overline{\text{AUC}}^{\text{OOS}}_{\text{WRLR}} > \overline{\text{AUC}}^{\text{OOS}}_{\text{LR}}$, where $\overline{\text{AUC}}^{\text{OOS}}_{\text{WRLR}}$ and $\overline{\text{AUC}}^{\text{OOS}}_{\text{LR}}$ denote the out of sample mean AUC values corresponding to (WRLR) and (LR), respectively. Statistically, the smaller the p-value, the more likely H_1 is true.

To train the (WRLR) model, one needs to specify the radius r_0 of the Wasserstein ball. One way to determine this radius is to use the concentration inequality $\Pr\big(\mathcal{W}(P_{\text{true}}, P_0) \leq r_0\big) \geq 1 - \gamma$. The theoretical bounds in [22] are of limited value. Instead we used six candidate empirical Wasserstein radii: $\{0, 0.01, 0.05, 0.1, 0.5, 1\}$[1] and a cross-validation approach to select the best r_0 from these. Specifically, we used the following 4-fold cross-validation approach: we divided m training samples into 4 subsets and used any three of them to train WRLR with every candidate value of r_0 and tested the model on the remaining subset. We finally picked the r_0 value corresponding to the best mean AUC value over 4 folds. Once this r_0 is selected, it is used for out of sample testing on the remaining observations that is not part of the chosen m samples.

The comparison shows that the quanity ($\overline{\text{AUC}}^{\text{OOS}}_{\text{WRLR}} - \overline{\text{AUC}}^{\text{OOS}}_{\text{LR}}$) ranges from -.0462 to .1005 and the relative difference ranges from -.3791 to .7122, in the studied cases. We observe that $\overline{\text{AUC}}^{\text{OOS}}_{\text{WRLR}}$ is greater than $\overline{\text{AUC}}^{\text{OOS}}_{\text{LR}}$ in 34 (77%) cases out of all 44 cases. The standard errors associated with (WRLR) are smaller than that of (LR) in 31 (72%) cases. This suggests that not only the distributionally-robust model is better, its performance is also more stable. It is seen from the p-values at the significance level $\alpha = 0.05$, (WRLR) outperformances (LR) in 24 (55%) cases which are from seven data sets: BA, BCW, ST-H, SPT-H, ION, SPTF-H and SPAM. For 7 (16%) cases which are from data sets: VC, PID and EES, $\overline{\text{AUC}}^{\text{OOS}}_{\text{WRLR}}$ is significantly smaller than $\overline{\text{AUC}}^{\text{OOS}}_{\text{LR}}$, indicating (WRLR) is not as good as (LR) for these three data sets. For the remaining 13 (29%) cases that are from data sets: PID, ST-H, EES, SPT-H, SPAM and CB, the difference in mean AUC is not statistically significant.

[1] Note that with $r_0 = 0$, the (WRLR) reduces to the (LR) model.

Table 2: Computational performance of solving WRLR and predictive performance of WRLR compared with LR for 11 data sets. Listed values are average of 100 experiments.

		Computational Performance						Predictive Performance						
		LR	WRLR					LR		WRLR				
Data set	m	CPU [sec]	No. main iters.[1]	No. cuts	CPU [sec]	Master[2](%)	Sep.[3](%)	Mean AUC	S.E.	Mean AUC	S.E.	Diff.	Rel. Diff.[4]	p-value
BA	50	0.022	3.8	66.9	1.21	13.74	86.26	.9985	.0002	.9991	.0001	.0006	.4202	.0062
	75	0.025	4.3	90.8	0.86	17.96	82.04	.9985	.0001	.9994	.0000	.0009	.5775	.0000
	100	0.032	3.9	116.7	1.83	13.89	86.11	.9993	.0000	.9995	.0000	.0002	.2724	.0001
	150	0.042	4.6	157.5	2.30	14.42	85.58	.9996	.0000	.9997	.0000	.0000	.1179	.0000
VC	50	0.021	7.2	195	1.65	30.51	69.49	.8782	.0025	.8320	.0034	-.0462	-.3791	1.0000
	75	0.032	7.1	271.3	1.82	29.15	70.85	.8867	.0022	.8441	.0029	-.0426	-.3755	1.0000
	100	0.047	6.6	323.3	3.73	32.44	67.56	.8887	.0023	.8504	.0025	-.0383	-.3442	1.0000
	150	0.064	6.5	541.5	5.19	30.63	69.37	.8951	.0023	.8601	.0029	-.0350	-.3337	1.0000
PID	50	0.032	7.5	203.7	4.19	22.20	77.80	.7542	.0058	.7564	.0037	.0022	.0089	.3756
	75	0.044	6.2	244	5.15	19.78	80.22	.8009	.0020	.8000	.0017	-.0009	-.0043	.6281
	100	0.05	7.5	240.4	3.85	13.34	86.66	.7880	.0025	.7840	.0019	-.0041	-.0192	.9034
	150	0.087	6	403.9	9.73	19.01	80.99	.8189	.0011	.8084	.0020	-.0105	-.0577	1.0000
BCW	50	0.051	8.6	251.8	6.19	31.44	68.56	.9716	.0020	.9916	.0004	.0200	.7040	.0000
	75	0.076	9.4	284	7.88	27.48	72.52	.9773	.0016	.9886	.0010	.0112	.4954	.0000
	100	0.097	8.9	501.4	12.84	34.28	65.72	.9790	.0025	.9940	.0001	.0150	.7122	.0000
	150	0.124	9.6	786.1	27.31	41.79	58.21	.9889	.0007	.9945	.0001	.0056	.5049	.0000
ST-H	50	0.069	11.7	242.7	12.18	30.56	69.44	.8317	.0035	.8808	.0029	.0490	.2914	.0000
	75	0.159	8.5	321.9	13.02	27.58	72.42	.8504	.0036	.8903	.0018	.0398	.2664	.0000
	100	0.147	9.6	381.5	10.50	24.15	75.85	.8945	.0022	.9064	.0011	.0120	.1133	.0000
	150	0.193	8.1	472.5	21.14	29.02	70.98	.8986	.0017	.8990	.0017	.0004	.0042	.4319
EES	50	0.041	15.7	251.9	7.80	37.72	62.28	.5874	.0037	.5902	.0035	.0029	.0070	.2877
	75	0.079	12.1	435	11.11	37.37	62.63	.6095	.0038	.5996	.0041	-.0099	-.0252	.9602
	100	0.322	11.9	577.8	17.78	36.55	63.45	.6221	.0033	.6175	.0032	-.0047	-.0124	.8459
	150	0.383	8.5	737.8	31.56	41.78	58.22	.6245	.0023	.6139	.0026	-.0106	-.0282	.9987
SPT-H	50	0.241	21.5	938.8	38.87	88.91	11.09	.8126	.0016	.8176	.0018	.0050	.0269	.0203
	75	0.343	24.5	1031.2	53.43	83.40	16.60	.8221	.0021	.8311	.0030	.0089	.0501	.0078
	100	0.54	19.4	1122.5	63.08	83.44	16.56	.8311	.0022	.8370	.0037	.0058	.0346	.0866
	150	0.987	13.4	1384	49.70	77.91	22.09	.8301	.0029	.8567	.0033	.0267	.1569	.0000
ION	50	0.912	34.7	740.6	173.37	63.56	36.44	.8429	.0024	.8708	.0021	.0279	.1775	.0000

Continued on the next page

Table 2 – continued from previous page

		Computational Performance						Predictive Performance						
		LR	WRLR					LR		WRLR				
Data set	m	CPU [sec]	No. main iters.[1]	No. cuts	CPU [sec]	Master[2](%)	Sep.[3](%)	Mean AUC	S.E.	Mean AUC	S.E.	Diff.	Rel. Diff.[4]	p-value
	75	2.415	28.2	1208.5	229.49	62.95	37.05	.8582	.0031	.8919	.0018	.0338	.2381	.0000
	100	1.51	23.1	1502.5	441.58	68.51	31.49	.8606	.0029	.8967	.0018	.0360	.2584	.0000
	150	2.774	17.8	1718.3	386.63	61.30	38.70	.8715	.0024	.9006	.0021	.0291	.2264	.0000
SPTF-H	50	2.144	38.9	438.1	62.68	70.38	29.62	.6818	.0048	.7517	.0056	.0700	.2198	.0000
	75	3.718	31.6	791.4	112.66	74.41	25.59	.7255	.0045	.8260	.0031	.1005	.3661	.0000
	100	3.453	29	983.7	144.66	75.50	24.50	.7638	.0033	.8368	.0027	.0730	.3091	.0000
	150	5.694	23.7	1327.6	221.44	77.87	22.13	.8142	.0039	.8632	.0037	.0490	.2638	.0000
SPAM	50	3.655	50.5	538.2	266.40	66.10	33.90	.8969	.0012	.8984	.0010	.0015	.0142	.1796
	75	4.76499	46.9	918.8	422.79	60.15	39.85	.9219	.0014	.9249	.0019	.0030	.0386	.0998
	100	11.065	40.5	1368	409.94	65.66	34.34	.9164	.0024	.9331	.0008	.0166	.1992	.0000
	150	15.76	36.8	1989.7	1414.10	75.31	24.69	.9346	.0010	.9387	.0007	.0041	.0630	.0005
CB	50	5.59802	44.6	420.4	120.13	42.30	57.70	.8081	.0043	.8136	.0033	.0055	.0286	.1549
	75	9.23099	45.7	601.2	228.41	53.83	46.17	.8105	.0027	.8120	.0016	.0015	.0080	.3138
	100	10.67	43.7	763.5	203.12	51.79	48.21	.8218	.0037	.8228	.0043	.0010	.0056	.4298
	150	19.887	39.4	1362	570.95	60.77	39.23	.8405	.0061	.8461	.0062	.0056	.0352	.2597

[1] Number of outer iterations described in Algorithm 2.
[2] Fraction of time spent on solving master problems described in Algorithm 2.
[3] Fraction of time spent on solving separation problems described in Algorithm 2.
[4] Defined as $(\overline{\text{AUC}}_{\text{WRLR}} - \overline{\text{AUC}}_{\text{LR}})/(1 - \overline{\text{AUC}}_{\text{LR}})$.

649 **6. Concluding Remarks.** The computational results presented in this paper
650 used a Wasserstein-robust framework for the logistic regression model, where all the
651 decision variables are continuous, and the feasible set is assumed to be convex. The
652 modified exchange algorithm of this paper is applicable to a broad class of nonlin-
653 ear optimization problems. These algorithms can be implemented further within a
654 branch-and-bound framework. Since the decomposition framework for solving the
655 dual of (WRO) is suitable for possibly mixed-integer decision variables and model pa-
656 rameters, this framework can also be adapted to model and solve (WRO) application
657 problems such as those arising in scheduling, logistics and supply chain management.

658 **Acknowledgement.** We are grateful to Changhyeok Lee and Liwei Zeng for a
659 valuable discussion. We also acknowledge NSF grants CMMI-1362003 and CMMI-
660 1100868 that were used to support this research.

661 REFERENCES

662 [1] E. D. Barrio, E. Gine, and C. Matran, *Central limit theorems for the wasserstein distance*
663 *between the empirical and the true distributions*, The Annals of Probability, 27 (1999),
664 pp. 1009–1071.
665 [2] A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski, *Robust Optimization*, Princeton Series in
666 Applied Mathematics, Princeton University Press, August 2009.
667 [3] A. Ben-Tal, D. Hertog, A. D. Waegenaere, B. Melenberg, and G. Rennen, *Robust*
668 *solutions of optimization problems affected by uncertain probabilities*, Management Science,
669 59 (2013), pp. 341–357.
670 [4] A. Ben-Tal and A. Nemirovski, *Robust convex optimization*, Mathematics of Operations
671 Research, 23 (1998), pp. 769–805.
672 [5] A. Ben-Tal and A. Nemirovski, *Robust solutions to uncertain linear programs*, Operations
673 Research Letters, 25 (1999), pp. 1–13.
674 [6] A. Ben-Tal and A. Nemirovski, *Robust optimization - methodology and applications*, Math-
675 ematical Programming, 92 (2002), pp. 453–480.
676 [7] A. Ben-Tal, A. Nemirovski, and C. Roos, *Robust solutions of uncertain quadratic and conic-*
677 *quadratic problems*, SIAM J. OPTIM., 13 (2002), pp. 535–560.
678 [8] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex Analysis and Optimization*, Athena
679 Scientific Optimization and Computation, Athena Scientific, April 2003.
680 [9] D. Bertsimas, X. V. Doan, K. Natarajan, and C. Teo, *Models for minimax stochastic linear*
681 *optimization problems with risk aversion*, Math. Oper. Res., 35 (2010), pp. 580–602.
682 [10] D. Bertsimas, V. Gupta, and N. Kallus, *Data-driven robust optimization*, 2013. https:
683 //arxiv.org/pdf/1401.0212v2.pdf.
684 [11] D. Bertsimas and I. Popescu, *Optimal inequalities in probability theory: A convex optimiza-*
685 *tion approach*, SIAM J. OPTM., 15 (2005), pp. 780–804.
686 [12] J. R. Birge and R. J.-B. Wets, *Computing bounds for stochastic programming problems*
687 *by means of a generalized moment problem*, Mathematics of Operations Research, 12(1)
688 (1987), pp. 149–162.
689 [13] G. C. Calafiore, *Ambiguous risk measures and optimal robust portfolios*, SIAM J. OPTIM.,
690 18 (2007), pp. 853–877.
691 [14] G. C. Calafiore and L. E. Ghaoui, *On distributionally robust chance-constrained linear*
692 *programs*, J. Optim. Theory Appl., 130 (2006), pp. 1–22.
693 [15] A. Charnes, W. W. Cooper, and K. O. Kortanek, *Duality, haar programs and finite se-*
694 *quence spaces*, Proceedings of the National Academy of Science, 48 (1962), pp. 783–786.
695 [16] A. Charnes, W. W. Cooper, and K. O. Kortanek, *Duality in semi-infinite programs and*
696 *some works of haar and carathéodory*, Management Science, 9 (1963), pp. 209–228.
697 [17] A. Charnes, W. W. Cooper, and K. O. Kortanek, *On the theory of semi-infinite program-*
698 *ming and a generalization of the kuhn-tucker saddle point theorem for arbitrary convex*
699 *functions*, Naval Research Logistics Quarterly, 16 (1969), pp. 41–51.
700 [18] E. Delage and Y. Ye, *Distributional robust optimization under moment uncertainty with*
701 *application to data-driven problems*, Operations Research, 58 (2010).
702 [19] J. Dupačová, *The minimax approach to stochastic programming and an illustrative applica-*
703 *tion*, Stochastics, 20 (1987), pp. 73–88.
704 [20] E. Erdoğan and G. Iyengar, *Ambiguous chance constrained problems and robust optimiza-*

tion, Mathematical Programming, 107 (2006), pp. 37–61.
[21] P. M. Esfahan and D. Kuhn, *Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulatioins*, 2015. https://arxiv.org/pdf/1505.05116.pdf.
[22] N. Fournier and A. Guillin, *On the rate of convergence in wasserstein distance of the empirical measure*, Probab. Theory Relat. Fields, 162 (2015), pp. 707–738.
[23] C. R. Givens and R. M. Shortt, *A class of wasserstein metrics for probability distributions*, Michigan Math. J., 31 (1984), pp. 231–240.
[24] J. Goh and M. Sim, *Distributionally robust optimization and its tractable approximations*, Oper. Res., 58 (2010), pp. 902–917.
[25] A. Haar, *Über linear ungleichungen*, Acta Mathematica Szeged, 2 (1924).
[26] G. Hanasusanto, V. Roitch, D. Kuhn, and W. Wiesemann, *A distributionally robust perspective on uncertainty quantification and chance constrained programming*, Mathematical Programming, 151 (2015), pp. 35–62.
[27] R. Hettich and H. T. Jongen, *On first and second order conditions for local optima for optimization problems in finite dimensions*, Methods Oper. Res., 23 (1977), pp. 82–97.
[28] R. Hettich and H. T. Jongen, *Semi-infinite programming: conditions of optimality and applications*, in Optimization Techniques, Lecture Notes in Control and Inform. Sci., J. Stoer, ed., Springer-Verlag, Berlin, Heidelberg, New York, 1977, ch. 7, pp. 82–97.
[29] R. Hettich and K. O. Kortanek, *Semi-infinite programming: theory, methods, and applications*, SIAM REVIEW, 35 (1993), pp. 380–429.
[30] R. Hettich and G. Still, *Second order optimality conditions for generalized semi-infinite programming problems*, Optimization, 34 (1995), pp. 195–211.
[31] P. Jaillet, J. Qi, and M. Sim, *Routing optimization under uncertainty*, Oper. Res., 64 (2016), pp. 186–200.
[32] R. Jiang and Y. Guan, *Risk-averse two-stage stochastic program with distributional ambiguity*, 2015. https://www.optimization-online.org/DB_FILE/2015/05/4908.pdf.
[33] R. Jiang and Y. Guan, *Data-driven chance constrained stochastic program*, Mathematical Programming, 158 (2016), pp. 291–327.
[34] K. O. Kortanek and H. No, *A central cutting plane algorithm for convex semi-infinite programming problems*, SIAM J. OPTM, 3 (1993), pp. 901–918.
[35] J. B. Lasserre, *Global optimization with polynomials and the problem of moments*, SIAM J. OPTIM., 11 (2001), pp. 796–817.
[36] C. Lee and S. Mehrotra, *A Distributionally-robust Approach for Finding Support Vector Machines*, 2015. eprints for the optimization community, http://www.optimization-online.org/DB_HTML/2015/06/4965.html.
[37] S. Liao, C. Delft, and J. P. Vial, *Distributionally robust workforce scheduling in call centers with uncertain arrival rates*, Optim. Methods Softw., 28 (2013), pp. 501–522.
[38] Q. Lin, R. Loxton, K. L. Teo, Y. H. Wu, and C. J. Yu, *A new exact penalty method for semi-infinite programming problems*, J. Comp. Appl. Math., 261 (2014), pp. 271–286.
[39] D. Z. Long and J. Qi, *Distributionally robust discrete optimization with entropic value-at-risk*, Oper. Res. Lett., 42 (2014), pp. 532–538.
[40] M. López and G. Still, *Semi-infinite programming*, European Journal of Operational Research, 180 (2007), pp. 491–518.
[41] D. K. Love and G. Bayraksan, *Phi-divergence constrained ambiguous stochastic programs for data-driven optimization*, 2016. http://www.optimization-online.org/DB_HTML/2016/03/5350.html.
[42] F. Q. Luo, C. Lee, and S. Mehrotra, *Application of distributionally robust optimization with wasserstein metric on logistic regression and support vector machine*, tech. report., Northwestern University, 2017.
[43] S. Mehrotra and D. Papp, *A cutting surface algorithm for semi-infinite convex programming with an application to moment robust optimization*, SIAM J. OPTM, 24 (2015), pp. 1670–1697.
[44] S. Mehrotra and H. Zhang, *Models and algorithms for distributionally robust least squares problems*, Mathematical Programming, 146 (2014), pp. 123–141.
[45] K. Natarajan, D. Pachamanova, and M. Sim, *Incorporating asymmetric distributional information in robust value-at-risk optimization*, Manag. Sci., 54 (2008), pp. 573–585.
[46] G. Nuernberger, *Global unicity in optimization and approximation*, Z. Angew. Math. Mech., 65 (1985), pp. T319–T321.
[47] G. Nuernberger, *Global unicity in semi-infinite optimization*, Numer. Funct. Anal. Optic., 8 (1985), pp. 173–191.
[48] P. A. Parrilo, *Semidefinite programming relaxations for semialgebraic problems*, Math. Pro-

gram. Ser. B, 96 (2003), pp. 293–320.
[49] G. C. Pflug, A. Pichler, and D. Wozabal, *Then 1/n investment strategy is optimal under high model ambiguity*, J. Bank. Financ., 36 (2012), pp. 410–417.
[50] G. Plug and D. Wozabal, *Ambiguity in portfolio selection*, Quantitative Finance, 7 (2007), pp. 435–442.
[51] K. Postek, D. Hertog, and B. Melenberg, *Computationally tractable counterparts of distributionally robust constraints on risk measures*, SIAM REV., 58 (2016), pp. 603–650.
[52] A. Prékopa, *Stochastic Programming*, vol. 324 of Mathematics and Its Applications, Springer Netherlands, 1995.
[53] R. Reemtsen and S. Görner, *Numerical methods for semi-infinite programming: A survey*, in Semi-infinite Programming, Nonconvex Optimization and Its Applications, R. Reemtsen and J. J. Rückmann, eds., Kluwer Boston, Boston, 1998, ch. 25, pp. 195–275.
[54] S. Shafieezadeh-Abadeh, P. M. Esfahani, and D. Kuhn, *Distributional robust logistic regression*. arXiv:1509.09259, 2015.
[55] A. Shapiro, *On duality theory of conic linear problems*, in Semi-Infinite Programming, vol. 57 of Nonconvex Optimization and Its Applications, Springer US, 2001, pp. 135–165.
[56] A. Shapiro and S. Ahmed, *On a class of minimax stochastic programs*, SIAM J. OPTIM., 14 (2004), pp. 1237–1249.
[57] A. Shapiro and A. Kleywegt, *Minimax analysis of stochastic problems*, Optimization Methods and Software, 17 (2002), pp. 523–542.
[58] G. Still, *Generalized semi-infinite programming: theory and methods*, European Journal of Operational Research, 119 (1999), pp. 301–313.
[59] A. Wächter and L. T. Biegler, *On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming*, Mathematical Programming, 106 (2006), pp. 25–57.
[60] S. Wang and Y. Yuan, *Feasible method for semi-infinite programs*, SIAM J. OPTM, 25 (2015), pp. 2537–2560.
[61] Z. Wang, P. W. Glynn, and Y. Ye, *Likelihood robust optimization for data-driven problems*, Comput. Manag. Sci., 13 (2016), pp. 241–261.
[62] W. Wiesemann, D. Kuhn, and M. Sim, *Distributional robust convex optimization*, Operations Research, 62 (2015), pp. 1358–1376.
[63] D. Wozabal, *A framework for optimization under ambiguity*, Annals of Operations Research, 193 (2012), pp. 21–47.
[64] H. Xu, C. Caramanis, and S. Mannor, *A distributional interpretation of robust optimization*, Mathematics of Operations Research, 37 (2012), pp. 95–110.
[65] M. Xu, S. Y. Wu, and J. J. Ye, *Solving semi-infinite programs by smoothing projected gradient method*, Comp. Optim. Appl., 59 (2014), pp. 591–616.
[66] X. Q. Yang, Z. Y. Chen, and J. C. Zhou, *Optimality conditions for semi-infinite and generalized semi-infinite programs via l_p exact penalty functions*, 2015. http://www.mypolyuweb.hk/~mayangxq/2015YCZ.pdf.
[67] İ. Yanikoğlu and D. Hertog, *Safe approximations of ambiguous chance constraints using historical data*, INFORMS J. Comput., 25 (2013), pp. 666–681.
[68] Y. Zhang, Z. M. Shen, and S. Song, *Distributionally robust optimization of two-stage lot-sizing problems*, Prod. Oper. Manag., 25 (2016), pp. 2116–2131.
[69] M. H. Zweig and G. Campbell, *Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine*, Clin Chem., 39 (1993), pp. 561–577.
[70] S. Zymler, D. Kuhn, and B. Rustem, *Distributional robust joint chance constraints with second-order moment information*, Math. Programming, 137 (2013), pp. 167–198.

Appendix A. Proofs and supplements for Sections 3 and 4.

Proof of Proposition 3.3.

Proof. By the definition of Wasserstein metric in (2), if two probability measures P_1 and P_2 satisfying $\mathcal{W}(P_1, P_2) < \varepsilon$, $\varepsilon > 0$, then there exists a $K \in \mathcal{M}(\Xi \times \Xi, \mathcal{F} \times \mathcal{F})$ such that

$$K(A \times \Xi) = P_1(A), \quad K(\Xi \times A) = P_2(A), \quad \forall A \in \mathcal{F} \tag{33}$$

$$\int_{(s_1 \times s_2)\Xi \times \Xi} d\big(s_1, s_2\big) K(ds_1 \times ds_2) \leq \varepsilon. \tag{34}$$

Therefore, we have

$$
\begin{aligned}
|f(P_1) - f(P_2)| &= \left| \int_{s_1 \in \Xi} h(\theta, s_1) P_1(ds_1) - \int_{s_2 \in \Xi} h(\theta, s_2) P_2(ds_2) \right| \\
&= \left| \int_{s_1 \in \Xi} h(\theta, s_1) K(ds_1 \times \Xi) - \int_{s_2 \in \Xi} h(\theta, s_2) K(\Xi \times ds_2) \right| \\
&= \left| \int_{(s_1 \times s_2) \in \Xi \times \Xi} h(\theta, s_1) K(ds_1 \times ds_2) - \int_{(s_1 \times s_2) \in \Xi \times \Xi} h(\theta, s_2) K(ds_1 \times ds_2) \right| \qquad (35) \\
&\leq \int_{(s_1 \times s_2) \in \Xi \times \Xi} |h(\theta, s_1) - h(\theta, s_2)| K(ds_1 \times ds_2) \\
&\leq C(\theta) \int_{(s_1 \times s_2) \in \Xi \times \Xi} d(s_1, s_2) K(ds_1 \times ds_2) \\
&\leq C(\theta) \varepsilon,
\end{aligned}
$$

which shows that f is continuous in $(\mathcal{M}(\Xi, \mathcal{F}), \mathcal{W})$. ∎

Proof of Proposition 3.4.

Proof. We divide the proof into the following two steps: (a) Show that such a joint probability measure K exists. Define a probability measure $K \in \mathcal{M}(\Xi \times \Xi, \mathcal{F} \times \mathcal{F})$ such that $K(A \times B) := P(A \cap B)$ for all $A, B \in \mathcal{F}$, and $K\left(\cup_{i=1}^{\infty} A_i \times B_i\right) = \sum_{i=1}^{\infty} K(A_i \times B_i)$ for all countable collections $\left\{A_i \times B_i\right\}_{i=1}^{\infty}$ of pairwise disjoint sets in $\mathcal{F} \times \mathcal{F}$. It is straightforward to verify that such K is as desired.

(b) Show that $\int_{(s_1 \times s_2) \in \Xi \times \Xi} d\big(s_1, s_2\big) K(ds_1 \times ds_2) = 0$. We prove by contradiction. Assume $\int_{(s_1 \times s_2) \in \Xi \times \Xi} d\big(s_1, s_2\big) K(ds_1 \times ds_2) > \varepsilon$, for some $\varepsilon > 0$, then we have $K(A) > 0$, where $A := \{(s_1 \times s_2) \in \Xi \times \Xi : \ d(s_1, s_2) > 0\}$. Let $A_n := \{(s_1 \times s_2) \in \Xi \times \Xi : d(s_1, s_2) \geq 1/n\}$. Since $A = \cup_{n=1}^{\infty} A_n$, it follows that $\exists\ m$ such that $K(A_m) > 0$. Since $A_m \in \mathcal{F} \times \mathcal{F}$, it can be expressed as: $A_m = \cup_{i=1}^{\infty} S_i^1 \times S_i^2$, where $S_i^1, S_i^2 \in \mathcal{F}$. This implies that $\exists\ i$ such that $K(S_i^1 \times S_i^2) > 0$. By the definition of K, we have $K(S_i^1 \times S_i^2) = P(S_i^1 \cap S_i^2) > 0$, hence $S_i^1 \cap S_i^2$ is nonempty, implying that $\exists\ s \in S_i^1 \cap S_i^2$ and hence $(s \times s) \in S_i^1 \times S_i^2 \subseteq A_m$. However, since $d(\cdot, \cdot)$ is a metric, we have $d(s, s) = 0$ which contradicts to $d(s, s) \geq 1/m$. ∎

Duality theorem of conic linear programming.

THEOREM A.1 (Proposition 2.8(iii) from [55]). *Let V and W be linear spaces (over real numbers), with V' and W' being their dual space respectively. Also let $C \subseteq V$ and $K \subseteq W$ be convex cones. Let $A : V \to W$ be a linear mapping and $A^* : W' \to V'$ be its adjoint mapping. Consider the conic linear optimization problem of the following form:*

$$
\begin{aligned}
&(36a) \qquad (\mathrm{P}) \quad \min_{v \in C} \quad \langle c, v \rangle \\
&(36b) \qquad \text{s.t.} \quad Av + b \in K.
\end{aligned}
$$

Then the dual of (P) *can be formulated as*

$$
\begin{aligned}
&(37a) \qquad (\mathrm{D}) \quad \max_{w^* \in -K^*} \quad \langle w^*, b \rangle \\
&(37b) \qquad \text{s.t.} \quad A^* w^* + c \in C^*,
\end{aligned}
$$

where C^* and K^* are the dual cones of C and K respectively.

(a) *The weak duality holds, e.g.,* $\mathrm{val(P)} \geq \mathrm{val(D)}$.
(b) *If* $\mathrm{val(P)}$ *is finite,* Y *is a finite dimensional Banach space, and the optimal solution set* $\mathrm{Sol(D)}$ *of* (D) *is nonempty and bounded, then* $\mathrm{val(P)}=\mathrm{val(D)}$.

Convergence of the central cutting surface algorithm for convex semi-infinite programs. We summarize the convergence of the central cutting surface algorithm from [43] for solving a general semi-infinite convex optimization problem of the form:

$$\begin{aligned} &\text{minimize} && x_0 \\ &\text{subject to} && g(x,t) \leq 0 \quad \forall t \in T, \\ &&& x \in X, \end{aligned} \tag{38}$$

where the decision variable is x whose first coordinate (and also the objective) is denoted by x_0. Assume the following conditions are satisfied (following the notation from [43]):

ASSUMPTION A.2. (1) *The set* $X \subseteq \mathbb{R}^n$ *is compact and convex.*
(2) *There exists a Slater point* $\overline{x}$ *and* $\eta > 0$ *satisfying* $\overline{x} \in X$ *and* $g(\overline{x}, t) \leq -\eta$ *for every* $t \in T$.
(3) *The function* $f(\cdot)$ *and* $g(\cdot, t)$ *are convex and sub-differentiable for every* $t \in T$*; moreover, these sub-differentials are uniformly bounded: there exists a* $B > 0$ *such that for every* $x \in X$ *and* $t \in T$*, every subgradient* $d \in \partial_x g(x,t)$ *satisfies* $\|d\| \leq B$.
(4) *For every point* $x \in X$ *that is not* ε*-feasible, there exists an oracle that can find in finite time a* $t \in T$ *satisfying* $g(x,t) > 0$.

The master problem of (38) at the kth iteration has the following form:

$$\begin{aligned} &\text{maximize} && w \\ &\text{subject to} && x_0 + w \leq M^{(k-1)}, \\ &&& g(x,t) + Bw \leq 0 \quad \forall t \in Q^{(k-1)}, \\ &&& x \in X. \end{aligned} \tag{39}$$

THEOREM A.3. ([43] Theorems 2-4 and Theorem 8) *Let* $(x^{(k)}, w^{(k)})$ *be the solution to the master problem at the* k*th iteration, and* $\widetilde{x}(k)$ *be the best know* ε*-feasible solution at the end of the* k*th iteration. The following statements hold:*
(a) *If Algorithm* 1 *terminates in the* k*th iteration, then* $\widetilde{x}^{(k-1)}$ *is an* ε*-optimal solution to* (38).
(b) *If Algorithm* 1 *does not terminate, then there exists an index* $\widehat{k}$ *such that the sequence* $\{\widetilde{x}^{(\widehat{k}+i)}\}_{i=1}^{\infty}$ *consists entirely of* ε*-feasible solutions.*
(c) *If Algorithm* 1 *does not terminate, then the sequence* $\{\widetilde{x}^{(k)}\}_{i=1}^{\infty}$ *has an accumulation point, and each accumulation point is an* ε*-optimal solution to* (38).
(d) *Algorithm* 1 *converges linearly in objective function value between consecutive feasible cuts after the first* $\hat{k}$ *iterations, where* $\widehat{k}$ *satisfies* $w^{(\widehat{k})} < \eta/B$*. Denote by* x^* *an optimal solution of* (38)*, and let* $\rho^{(k)} = \frac{\widetilde{x}_0^{(k)} - x_0^*}{\widetilde{x}_0^{(k-1)} - x_0^*}$*, then we have*

$$\rho^{(k)} \leq 1 - \frac{\eta - Bw^{(k)}}{\eta + B(\overline{x}_0 - x_0^*)} \tag{40}$$