Tikhonov Regularization is Optimal Transport Robust under Martingale Constraints

Anonymous Author(s)
Affiliation
Address
email

Abstract

Distributionally robust optimization has been shown to offer a principled way to regularize learning models. In this paper, we find that Tikhonov regularization is distributionally robust in an optimal transport sense (i.e., if an adversary chooses distributions in a suitable optimal transport neighborhood of the empirical measure), provided that suitable martingale constraints are also imposed. Further, we introduce a relaxation of the martingale constraints which not only provides a unified viewpoint to a class of existing robust methods but also leads to new regularization tools. To realize these novel tools, tractable computational algorithms are proposed. As a byproduct, the strong duality theorem proved in this paper can be potentially applied to other problems of independent interest.

1 Introduction

Regularization is an important tool in machine learning which is used in, for instance, reducing overfitting [20]. Recently, ideas from distributionally robust optimization (DRO) have led to a fresh viewpoint on regularization precisely in connection to overfitting; see, e.g., [7, 2, 17, 10, 26, 21, 22, 5] and the references therein.

In these references it is shown that many standard regularization-based estimators arise as the solution of a min-max game in which one wishes to minimize a loss over a class of parameters against an adversary that maximizes the out-of-distribution impact of any given parameter choice, that is, the adversary perturbs the empirical distribution in a certain way. The choice of adversarial distributions or perturbations in DRO is often non-parametric thus providing reassurance that the decision is reasonably robust to a wide range of out-of-distribution perturbations. For example, one such non-parametric choice is given by employing optimal transport costs [28] to construct a so-called distributional uncertainty set (e.g. a Wasserstein ball around the empirical distribution) for the adversary to choose. [24] shows that optimal transport-based DRO (OT-DRO) is closely related to adversarial robustness in the sense of steepest gradient loss contamination. This can be further explained by OT-DRO's hidden connection with generalized Lipschitz regularization [7]. Thus, understanding if a well-known regularization technique is actually distributionally robust and in what sense, allows us to understand its out-of-distribution benefits and potentially introduce improvements.

In this paper, we introduce a novel set of regularization techniques which incorporate martingale constraints into the OT-DRO framework. Our starting point is the conventional OT-DRO formulation. The conventional OT-DRO formulation can generally be interpreted as perturbing each data point in such a way that the average size perturbation is less than a given budget. In addition to this conventional formulation, we will impose a *martingale constraint* in the joint distribution of the empirical data and the resulting adversarially perturbed data.

Why do we believe that the martingale constraint makes sense as a regularization technique? It turns out that two random variables X and $\bar{X}$ form a martingale in the sense that $\mathbb{E}[\bar{X}|X] = X$ if and only

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

if the distribution of $\bar{X}$ dominates X in convex order [27]. In this sense, the adversary $\bar{X}$ will have higher dispersion in non-parametric sense than the observed data X but in a suitably constrained way so that the average locations are preserved. This novel OT-DRO constrained regularization, we believe, is helpful to potentially combat conservative solutions, see [13]. Moreover, by allowing a small amount of violation in the martingale property, we can control the regularization properties of this constraint, thus obtaining a natural interpolation towards the conventional OT-DRO formulation and potentially improved regularization performance. We point out that related optimal transport problems with martingale constraints have been studied in robust mathematical finance [1, 8].

Consider, for example, the linear regression setting with the exact martingale constraints, which means that for any given observed data point, the conditional expectation of the additive perturbation under the worst-case joint distribution equals zero. Surprisingly, we show that the resulting martingale DRO model is *exactly* equivalent to the ridge regression [15] with the Tikhonov regularization. To the best of our knowledge, this paper is the first work to interpret the Tikhonov regularization from a DRO perspective showing that it is distributionally robust in a precise non-parametric sense. In stark contrast, it is well-known that the conventional OT-based DRO model (*without* the martingale constraint) is identical to the regularized square-root regression problem [2]. Therefore, introducing an additional power in norm regularization (i.e., converting square-root regression to Tikhonov regularization) can be translated into adding martingale constraints in the adversarial perturbations thus reducing the adversarial power. A natural question that arises here is whether we can interpolate between the conventional DRO model and the Tikhonov regularization, and further improve them.

We will provide a comprehensive and positive answer to the above question in this paper. The key idea here is to relax the equality constraint on the conditional expectation of the adversarial violation and thus allow a small perturbation of the martingale property to gain more flexibility of the uncertainty set. This idea leads to another novel model, termed the perturbed martingale DRO in the sequel. Intuitively, if the relaxation is sufficiently loose, the perturbed martingale DRO model will reduce to the conventional DRO, which is formally equivalent to setting an infinite amount of possible violations for the martingale constraint. By contrast, if no violation is allowed, the perturbed martingale DRO will automatically reduce to the exact counterpart — Tikhonov regularization. As a result, we are able to introduce a new class of regularizers via the interpolation between the conventional DRO model and the Tikhonov regularization.

Furthermore, such insightful interpolation also works for a broad class of nonlinear learning models. Inspired by our extensive exploration of linear regression, the developed martingale DRO model can also provide a new principled adversarial training procedure for deep neural networks. Extensive experiments are conducted to demonstrate the effectiveness of the proposed perturbed martingale DRO model for both linear regression and deep neural network training under the adversarial setting.

We summarize our main contributions as below:

- We reveal a new hidden connection in this paper, that is, Tikhonov regularization is optimal transport robust when exact martingale constraints (i.e., convex order between adversary and empirical data) are imposed.
- Upon this surprising finding, we develop a new perturbed martingale DRO model, which not only provides a unified viewpoint of existing regularization techniques, but also leads to a new class of robust regularizers.
- We introduce an easy-to-implement computational approach to realize the model benefits in practice, for both linear regression and neural network training under the adversarial setting.
- As a byproduct, the strong duality theorem proved in this paper, acting as the main technical tool, can be potentially applied to other problems of independent interest.

2 Preliminaries

Let us introduce some basic definitions and concepts preparing for the later analysis.

Definition 2.1 (Optimal transport costs and the Wasserstein distance [18, 28])**.** *Suppose that* $c(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to [0,\infty]$ *is a lower semi-continuous cost function such that* $c(X,X) = 0$ *for every* $X \in \mathbb{R}^d$*. The optimal transport cost between two distributions* $\mathbb{Q}$ *and* $\mathbb{P}$ *supported on* $\mathbb{R}^d$ *is defined as*

$$D(\mathbb{Q},\mathbb{P}) \triangleq \min_{\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{X})} \left\{ \mathbb{E}_\pi \left[c(\bar{X}, X) \right] : P_1\pi = \mathbb{Q},\ P_2\pi = \mathbb{P} \right\}.$$

Here, $\mathcal{P}(\mathcal{X} \times \mathcal{X})$ is the set of joint probability distribution π of $(\bar{X}, X)$ supported on $\mathcal{X} \times \mathcal{X}$ while $P_1\pi$ and $P_2\pi$ respectively refer to the marginals of $\bar{X}$ and X under the joint distribution π.

If $c(X, \bar{X}) = \|\bar{X} - X\|$ is any given norm on $\mathbb{R}^d$, then D recovers the Wasserstein distance [28]. In this paper, we are interested in a flexible family of functions for the computational tractability, so called the Mahalanobis cost functions in the form of $c(\bar{X}, X) = (X - \bar{X})^\top M(X - \bar{X})$, where M is a d-by-d positive definite matrix.

Next, we consider the conventional OT-DRO problem:

$$\min_{\beta} \mathbb{L}_\beta(\widehat{\mathbb{P}}, \rho), \quad \text{where} \quad \mathbb{L}_\beta(\widehat{\mathbb{P}}, \rho) \triangleq \left\{ \begin{array}{ll} \sup\limits_{\pi} & \mathbb{E}_\pi[\ell(f_\beta(\bar{X}))] \\ \text{s.t.} & \pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) \\ & \mathbb{E}_\pi\left[c(\bar{X}, X)\right] \le \rho, \ P_2\pi = \widehat{\mathbb{P}}, \end{array} \right. \tag{2.1}$$

where $\widehat{\mathbb{P}} \triangleq \frac{1}{N}\sum_{i=1}^N \delta_{X_i}$ is the empirical distribution. Using Definition 2.1, we have $\mathbb{L}_\beta(\widehat{\mathbb{P}}, \rho) = \max_{\mathbb{Q}: D(\mathbb{Q}, \widehat{\mathbb{P}}) \le \rho} \mathbb{E}_{\mathbb{Q}}[\ell(f_\beta(\bar{X}))]$, which is the worst-case expected loss under all possible distributions around the empirical measure $\widehat{\mathbb{P}}$ at most ρ with respect to the OT distance. It is well-known that under appropriate assumptions, the DRO problem (2.1) is equivalent to the regularized square-root regression problem.

Proposition 2.2 ([2, Proposition 2.]). *Suppose that (i) the loss function $\ell(\cdot)$ is a convex quadratic function, i.e., $\nabla^2\ell(\cdot) = \gamma > 0$, where γ is a constant, (ii) the feature mapping $f_\beta(\bar{X}) = \beta^\top \bar{X}$ is linear, and (iii) the ground cost c is the squared Euclidean norm on $\mathcal{X} = \mathbb{R}^d$. Then*

$$\mathbb{L}_\beta(\widehat{\mathbb{P}}, \rho) = \left(\sqrt{\mathbb{E}_{\widehat{\mathbb{P}}}[\ell(f_\beta(X))]} + \sqrt{\rho}\|\beta\|_2\right)^2.$$

We also present the strong duality result for a general class of optimal transport based DRO models with martingale constraints. This result serves as our main technical tool for reformulating the DRO models, and it can also be applied to other semi-infinity structured DRO models, which could be of independent interests. We consider the primal problem

$$\begin{array}{ll} \sup\limits_{\pi} & \int_{\mathcal{X}\times\mathcal{X}} f(\bar{X})\,\mathrm{d}\pi \\ \text{s.t.} & \pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) \\ & \int_{\mathcal{X}\times\mathcal{X}} c(\bar{X}, X)\,\mathrm{d}\pi \le \rho, \ \ P_2\pi = \widehat{\mathbb{P}} \\ & \mathbb{E}_\pi[\bar{X}|X] = X \ \ \widehat{\mathbb{P}}\text{-a.s.} \end{array} \tag{Primal}$$

and its associated dual form

$$\inf_{\substack{\lambda \in \mathbb{R}_+ \\ \alpha_i \in \mathbb{R}^d \ \forall i}} \lambda\rho + \sum_{i=1}^N \alpha_i^\top X_i + \frac{1}{N}\sum_{i=1}^N \sup_{\bar{X}} \left[f(\bar{X}) - \alpha_i^\top \bar{X} - \lambda c(\bar{X}, X_i)\right]. \tag{Dual}$$

Here, $f : \mathcal{X} \to \mathbb{R}$ is upper semi-continuous and L^1-integrable. The next theorem states the strong duality result linking these two problems.

Theorem 2.3 (Strong duality). *Let $\widehat{\mathbb{P}} \triangleq \frac{1}{N}\sum_{i\in[N]} \delta_{X_i}$ be the reference measure. Suppose that (i) every sample point is in the interior of the cone generated by $\mathcal{X}$, i.e., $X_i \in \mathrm{int}(\mathrm{cone}(\mathcal{X}))\ \forall i \in [N]$, and (ii) the ambiguity radius $\rho > 0$. Then the strong duality holds, i.e.,* Val(Primal) = Val(Dual).

Remark 2.4. *At the heart of our analysis tools is the abstract semi-infinite duality theory for conic linear program [23, Proposition 3.4]. Notably, it will be tricky and subtle to reformulate our problem into the standard form and further carefully check the general Slater condition. Moreover, we indeed fill the technical gap in [16, Theorem 4.2].*

Notation. We use $\mathbb{S}^d_{++}$ to denote the set of d-by-d positive definite matrices and $\|X\|_M \triangleq \sqrt{X^\top M X}$ for any $X \in \mathbb{R}^d, M \in \mathbb{S}^d_{++}$; if M is the identity matrix, then we omit M and write $\|X\| \triangleq \sqrt{X^\top X}$. We use δ_X to denote the Dirac measure at X and let $\widehat{\mathbb{P}} \triangleq \frac{1}{N}\sum_{i=1}^N \delta_{X_i}$ be the empirical measure constructed from sample $\{X_1, \ldots, X_N\}$. We use $\mathbb{E}_P$ to denote the integration over P: $\mathbb{E}_P[f(X)] = \int_{\mathcal{X}} f(X)\mathrm{d}P$. Specifically, for $(\bar{X}, X)$ following joint distribution π, $\mathbb{E}_\pi[c(\bar{X}, X)] =$

$\int_{\mathcal{X}\times\mathcal{X}} c(\bar{X}, X)\mathrm{d}\pi$, $\mathbb{E}_\pi[f(\bar{X})] = \int_{\mathcal{X}\times\mathcal{X}} f(\bar{X})\mathrm{d}\pi$. We use $\ell(\cdot)$ to denote the loss function applied to the parametrized feature mapping f_β. Let $\nabla\ell(\cdot), \nabla^2\ell(\cdot)$ be the first and second order derivative of $\ell(\cdot)$ respectively. Notably, we use $\mathbb{L}_\beta(\widehat{\mathbb{P}}, \rho)$ to denote the objective function of the conventional DRO model (2.1); we use $L_\beta(\widehat{\mathbb{P}}, \rho)$ to refer to the exact martingale DRO model (3.1); we use $\mathcal{L}_\beta(\widehat{\mathbb{P}}, \rho, \epsilon)$ to refer to the perturbed martingale DRO model (3.2).

3 Tractable Reformulations

In this section, we introduce an optimal transport-based DRO model with the exact martingale constraint at first. That is, on top of the vanilla DRO model [3], we add an additional martingale equality constraint on its coupling. It is surprisingly interesting to find out that the resulting DRO approach is equivalent to empirical risk minimization with Tikhonov regularization. Naturally, we can relax the equality constraint and thus allow a small violation of the martingale property to enrich the uncertainty set. Formally, by sending violation size to infinity in the martingale constraint, our relaxation allows to interpolate between the conventional DRO formulation (i.e. with no martingale constraints) and Tikhonov regularization (which involves exact martingale constraints). Therefore, this relaxation further leads to a new class of regularizers in a principled way, which improves upon Tikhonov regularization as we show in our experiments.

Assumption 3.1. *The following assumptions hold throughout.*

(i) The ground cost $c(\cdot,\cdot)$ is the Mahalanobis cost with the weighting matrix $M \in \mathbb{S}^d_{++}$,

(ii) The domain $\mathcal{X}$ is unconstrained, i.e., $\mathcal{X} = \mathbb{R}^d$.

3.1 Optimal Transport-based DRO with Martingale Constraints

To start with, we investigate the exact martingale DRO problem:

$$\min_\beta L_\beta(\widehat{\mathbb{P}}, \rho), \qquad \text{where} \quad L_\beta(\widehat{\mathbb{P}}, \rho) \triangleq \left\{ \begin{array}{ll} \sup\limits_{\pi} & \mathbb{E}_\pi[\ell(f_\beta(\bar{X}))] \\ \text{s.t.} & \pi \in \mathcal{P}(\mathcal{X}\times\mathcal{X}) \\ & \mathbb{E}_\pi\left[c(\bar{X}, X)\right] \le \rho, \ P_2\pi = \widehat{\mathbb{P}} \\ & \mathbb{E}_\pi[\bar{X}|X] = X \quad \widehat{\mathbb{P}}\text{-a.s.}, \end{array} \right. \tag{3.1}$$

and $\rho \ge 0$ is the radius of uncertainty set centered at $\widehat{\mathbb{P}}$. Note that because $\widehat{\mathbb{P}}$ is the empirical measure, the martingale constraint implies that the *conditional* expected value of the perturbation obtained by modifying each observed data point equals the observed data point itself. The quantity $L_\beta(\widehat{\mathbb{P}}, \rho)$ is referred to as the worst-case expected loss of the model parameter β under the martingale DRO model. It is easy to see that $L_\beta(\widehat{\mathbb{P}}, \rho) \le \mathbb{L}_\beta(\widehat{\mathbb{P}}, \rho)$, where $\mathbb{L}_\beta$ is defined as in (2.1). This is because the adversary in (3.1) has a smaller feasible set, and thus is less powerful than the adversary in (2.1). Hence, the martingale DRO solution for problem (3.1) is considered to be *less* conservative than the conventional DRO solution for problem (2.1).

The next result asserts that the martingale DRO problem coincides with the Tikhonov regularization problem under similar conditions of Proposition 2.2.

Proposition 3.2 (Tikhonov equivalence)**.** *Suppose that (i) the loss function $\ell(\cdot)$ is a convex quadratic function, i.e., $\nabla^2\ell(\cdot) = \gamma > 0$, and (ii) the feature mapping $f_\beta(\bar{X}) = \beta^\top \bar{X}$ is linear. Then we have*

$$L_\beta(\widehat{\mathbb{P}}, \rho) = \mathbb{E}_{\widehat{\mathbb{P}}}[\ell(\beta^\top X)] + \frac{\gamma\rho}{2}\|\beta\|^2_{M^{-1}}.$$

If the Mahalanobis matrix M is the identity matrix, the martingale DRO model (3.1) recovers the Tikhonov regularization problem.

Proof of Proposition 3.2. By a change of the variable, let $\Delta = \bar{X} - X$ and we have

$$\begin{aligned} \sup_{\substack{\mathbb{E}_\pi[\|\Delta\|_M^2]\le\rho \\ \mathbb{E}_\pi[\Delta|X]=0}} \mathbb{E}_\pi\left[\ell(\beta^\top(X+\Delta))\right] &= \sup_{\substack{\mathbb{E}_\pi[\|\Delta\|_M^2]\le\rho \\ \mathbb{E}_\pi[\Delta|X]=0}} \mathbb{E}_\pi\left[\ell(\beta^\top X) + \nabla\ell(\beta^\top X)\beta^\top\Delta + \frac{\gamma}{2}\|\beta^\top\Delta\|^2\right] \\ &= \mathbb{E}_{\widehat{\mathbb{P}}}\left[\ell(\beta^\top X)\right] + \sup_{\substack{\mathbb{E}_\pi[\|\Delta\|_M^2]\le\rho \\ \mathbb{E}_\pi[\Delta|X]=0}} \mathbb{E}_\pi\left[\frac{\gamma}{2}\|\beta^\top\Delta\|^2\right] \end{aligned}$$

$$= \mathbb{E}_{\widehat{\mathbb{P}}}\left[\ell(\beta^\top X)\right] + \frac{\gamma\rho}{2}\|\beta\|_{M^{-1}}^2.$$

The last equality follows from the general Hölder's inequality. □

Example 3.3 (Linear regression)**.** *Let $X^\top \triangleq (Y, Z^\top) \in \mathbb{R}^d$ and $\beta^\top \triangleq (1, -b^\top) \in \mathbb{R}^d$, we have $\beta^\top X = Y - b^\top Z$. For any $Q \in \mathbb{S}_{++}^{d-1}$, we take $M = \mathrm{diag}(+\infty, Q)$, which implies that we do not allow transport of the response Y, then the problem (3.1) with $\gamma = 2$ becomes*

$$\min_b \left\{\mathbb{E}_{\widehat{\mathbb{P}}}\left[(Y - b^\top Z)^2\right] + \rho\|b\|_{Q^{-1}}^2\right\}.$$

In Appendix C.1, we give another instructive proof based on the strong duality result in Section 2. For general convex loss functions, we have the following certificate of robustness that provides upper and lower bound for the worst-case DRO loss in (3.1).

Corollary 3.4 (General convex loss functions)**.** *Suppose that (i) the loss function $\ell(\cdot)$ is μ-strongly convex and C-smooth, that is, $\ell(\theta) \geq \ell(\theta') + \nabla\ell(\theta')(\theta' - \theta) + \frac{\mu}{2}(\theta' - \theta)^2$ and $\ell(\theta) \leq \ell(\theta') + \nabla\ell(\theta')(\theta' - \theta) + \frac{C}{2}(\theta' - \theta)^2$ hold for all $\theta, \theta' \in \mathbb{R}$, and (ii) the feature mapping $f_\beta(\bar{X}) = \beta^\top \bar{X}$ is linear. Then we have for any $\rho \geq 0$,*

$$\mathbb{E}_{\widehat{\mathbb{P}}}[\ell(f_\beta(X))] + \frac{\mu\rho}{2}\|\beta\|_{M^{-1}}^2 \leq L_\beta(\widehat{\mathbb{P}}, \rho) \leq \mathbb{E}_{\widehat{\mathbb{P}}}[\ell(f_\beta(X))] + \frac{C\rho}{2}\|\beta\|_{M^{-1}}^2.$$

Example 3.5 (Logistic regression)**.** *Let $X = YZ \in \mathbb{R}^d$, where $Z \in \mathbb{R}^d, Y \in \{\pm 1\}$, and $\ell(t) = \log(1 + \exp(-t))$, where $\ell(\cdot)$ satisfies Assumption (i) in Corollary 3.4 with $C = \frac{1}{4}$. Then we have*

$$L_\beta(\widehat{\mathbb{P}}, \rho) \leq \mathbb{E}_{\widehat{\mathbb{P}}}[\log(1 + \exp(-Y\beta^\top Z))] + \frac{\rho}{8}\|\beta\|_{M^{-1}}^2.$$

3.2 Optimal Transport-based DRO with Perturbed Martingale Constraints

Now we turn to the relaxation of the martingale constraint to improve upon Tikhonov regularization and gain more flexibility. We consider the perturbed martingale coupling based DRO model (perturbed martingale DRO):

$$\min_\beta \mathcal{L}_\beta(\widehat{\mathbb{P}}, \rho, \epsilon), \quad \text{where } \mathcal{L}_\beta(\widehat{\mathbb{P}}, \rho, \epsilon) \triangleq \begin{cases} \sup\limits_\pi & \mathbb{E}_\pi[\ell(f_\beta(\bar{X}))] \\ \text{s.t.} & \pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) \\ & \mathbb{E}_\pi\left[c(\bar{X}, X)\right] \leq \rho, \ P_2\pi = \widehat{\mathbb{P}} \\ & \|\mathbb{E}_\pi[\bar{X}|X] - X\|_M \leq \epsilon \quad \widehat{\mathbb{P}}\text{-a.s.} \end{cases} \tag{3.2}$$

The parameter ϵ controls the allowed violations of the martingale constraint for the adversary. It is trivial that if we set $\epsilon = 0$, then we obtain $\mathcal{L}_\beta(\widehat{\mathbb{P}}, \rho, 0) = L_\beta(\widehat{\mathbb{P}}, \rho)$ and we recover the exact martingale DRO model (3.1). If we set $\epsilon = +\infty$ then the martingale constraint becomes ineffective, thus we have $\mathcal{L}_\beta(\widehat{\mathbb{P}}, \rho, +\infty) = \mathbb{L}_\beta(\widehat{\mathbb{P}}, \rho)$ and model (3.2) collapses to the DRO formulation (2.1). We thus can think of ϵ as an interpolating parameter connecting two extremes: the conventional DRO model (2.1) (at $\epsilon = +\infty$) and the exact martingale DRO model (3.1) (at $\epsilon = 0$).

However, the resulting optimization problem (3.2) constitutes an infinite-dimensional optimization problem over probability distributions and thus appears to be computationally intractable. To overcome this issue, we leverage Theorem 2.3 and prove that the problem (3.2) is actually equivalent to a finite-dimensional problem. To begin with, we present one crucial proposition that can be applied to more general settings.

Proposition 3.6 (General loss functions and feature mappings)**.** *Suppose that the loss function $\ell(\cdot)$ and the feature mapping $f_\beta(\cdot)$ are upper semi-continuous. Then, for any $\rho > 0$ and $\epsilon > 0$, the perturbed martingale DRO model (3.2) admits*

$$\mathcal{L}_\beta(\widehat{\mathbb{P}}, \rho, \epsilon) = \inf_{\lambda \geq 0, \alpha} \lambda\rho + \frac{\epsilon}{N}\sum_{i=1}^N \|\alpha_i\|_{M^{-1}} + \frac{1}{N}\sum_{i=1}^N \sup_{\Delta_i}\left[\ell(f_\beta(X_i + \Delta_i)) - \alpha_i^\top \Delta_i - \lambda\|\Delta_i\|_M^2\right]. \tag{3.3}$$

Sketch of proof. The key step is to decouple (3.2) as a two-layer optimization problem:

$$\begin{aligned} \mathcal{L}_\beta(\widehat{\mathbb{P}}, \rho, \epsilon) = \sup_{\|\eta_i\|_M \leq \epsilon \ \forall i} \quad & \sup_\pi \quad \int_{\mathcal{X}} \ell(f_\beta(\bar{X}))\mathrm{d}\pi \\ & \text{s.t} \quad \pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) \\ & \quad \int_{\mathcal{X}\times\mathcal{X}} c(\bar{X}, X)\mathrm{d}\pi \leq \rho, \ P_2\pi = \widehat{\mathbb{P}} \\ & \quad \int_{\mathcal{X}\times\mathcal{X}} \mathbb{I}_{X_i}(X) \cdot \bar{X}\mathrm{d}\pi = \tfrac{1}{N}(X_i + \eta_i) \quad \forall i \in [N]. \end{aligned} \tag{3.4}$$

Then we invoke Theorem 2.3 for the inner maximization problem over π and apply the Sion's minimax theorem [25] for the outer maximization over η. The desired result is obtained. $\square$

Due to the specific structure of the quadratic cost and the linear feature mapping, the dual variable λ admits a closed-form representation and $\alpha \in \mathbb{R}^{N\times d}$ in (3.3) can be reduced to N parallel one-dimensional optimization problems. Upon this observation, we thus conduct an instructive and intuitive reformulation in Theorem 3.7 for linear regression with proof detailed in Appendix C.2.

Theorem 3.7 (Linear regression)**.** *Suppose that (i) the loss function $\ell(\cdot)$ is a convex quadratic function, i.e., $\nabla^2\ell(\cdot) = \gamma > 0$ and (ii) the feature mapping $f_\beta(\bar{X}) = \beta^\top \bar{X}$ is linear. Then, the perturbed martingale DRO model (3.2) admits:*

$$\mathcal{L}_\beta(\widehat{\mathbb{P}}, \rho, \epsilon) = \mathbb{E}_{\widehat{\mathbb{P}}}[\ell(f_\beta(X))] + \frac{\rho\|\beta\|_{M^{-1}}^2}{2} + \boxed{R(\beta)}, \tag{3.5}$$

where the additional regularizer is defined as $R(\beta) \triangleq \|\beta\|_{M^{-1}} \min_{s\in\mathbb{R}^N} \left(\frac{\epsilon}{N}\|s\|_1 + \sqrt{\frac{\rho}{N}}\|G_\beta - s\|_2\right)$ and $G_\beta = (\nabla\ell(\beta^\top X_1), \ldots, \nabla\ell(\beta^\top X_N))^\top \in \mathbb{R}^N$.

Obviously, compared with the exact martingale constraint investigated in Proposition 3.2, the perturbed constraint we consider here involves an additional term $R(\beta)$. Fortunately, armed with Lemma D.3, we are able to shed light on its intuitive interpretation based on the quantitative relationship between modeling parameters ϵ and ρ. First, we discuss two extreme cases — there are interesting hidden connections between (3.2) and other existing regularization techniques.

Remark 3.8. *Intuitively, if ϵ is relatively small, the optimal $s^\star$ is zero which means that all perturbed martingale constraints will be active. Precisely, if $\epsilon^2 \le \rho$, the additional regularization term satisfies $R(\beta) = \epsilon\mathbb{E}_{\widehat{\mathbb{P}}}[\|\nabla_X \ell(f_\beta(X))\|_{M^{-1}}]$, which is so-called Jacobian or input gradient regularizer in the literature when M is an identity matrix. Recently, it has received much intention owing to its ability to improve adversarial robustness [29, 9]. Conversely, if ρ is relatively small, the optimal $s^\star$ is equal to G_β, implying that none of perturbed martingale constraints is active. In fact, if $\epsilon^2 \ge N\rho$, we have $R(\beta) = \sqrt{\rho\mathbb{E}_{\widehat{\mathbb{P}}}[\|\nabla_X \ell(f_\beta(X))\|_{M^{-1}}^2]}$. Then, we can easily figure out that $\mathcal{L}_\beta(\widehat{\mathbb{P}}, \rho, \epsilon)$ can be reduced to the conventional OT-based DRO model in Proposition 2.2.*

Then for the middle case, it is natural to infer that only part of constraints will be active. In Lemma D.3, we justify this conjecture rigorously. We refer the reader to Appendix D for the details. As such, the proposed martingale DRO model takes the first step bridging the input gradient regularization and regularized square-root regression problem in a unified framework. On the other hand, it also opens up an exciting brand new avenue of robustified regularizers. In the next section, we validate its effectiveness for the adversarial training task.

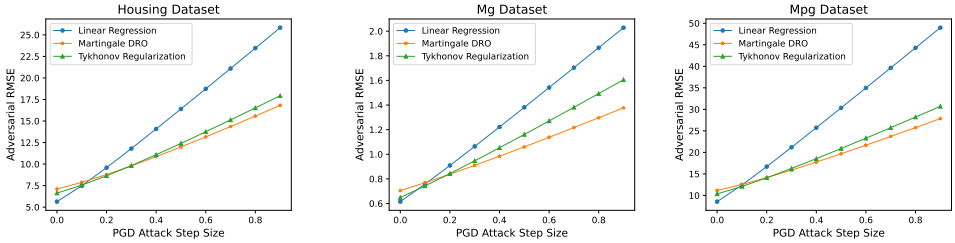
4 Optimization Algorithms

To take advantage of the proposed perturbed martingale DRO model, a natural question here is whether we can address problem (3.2) in a tractable manner. In this section, we answer the above question in the affirmative by developing two different computational paradigms for linear regression and deep neural network respectively.

4.1 Subgradient Method for Linear Regression

Unfortunately, the resulting formulation we conducted in Theorem 3.7 (i.e., (3.5)) for linear regression is potentially intractable for optimizing (β, s) jointly. One obvious computational challenge here is that the overall problem is not necessarily convex in (β, s), even if the original problem is convex over β. This is because the minima of convex functions is not convex. However, under a mild assumption, it is possible to give a reliable computational routine for solving the resulting problem over β directly.

To start with, a key observation is $\mathcal{L}_\beta(\widehat{\mathbb{P}}, \rho, \epsilon)$ is convex over β. This is essentially from the fact that the pointwise supremum of a class of convex functions is still convex. To check the details, $\mathcal{L}_\beta(\widehat{\mathbb{P}}, \rho, \epsilon)$ is originally defined in (3.2) and $\mathbb{E}_\pi[\ell(f_\beta(\bar{X}))]$ is convex over β for all π. Thus, a natural yet simple algorithm is subgradient method. The main difficulty is to obtain the correct subgradient oracle. Since the first two terms are smooth and strongly convex, we know that $R(\beta)$ is weakly convex and thus subdifferentially regular. All the subdifferential concepts are coincide. As such, we may simply use the Clarke subdifferential [6] in the sequel. Moreover, the sum rule is hold for



(a) Housing (b) Mg (c) Mpg

Figure 1: Compare the proposed martingale DRO with two standard benchmarks (linear regression and ridge regression) on three real world datasets — Housing, Mg and Mpg. The martingale DRO performs better than competing methods under large PGD step size.

computing the subgradient of $\mathcal{L}_\beta(\widehat{\mathbb{P}}, \rho, \epsilon)$ due to the weakly convexity. We have

$$\partial_\beta \mathcal{L}_\beta(\widehat{\mathbb{P}}, \rho, \epsilon) = \nabla_\beta \mathbb{E}_{\widehat{\mathbb{P}}}[\ell(f_\beta(X))] + \nabla_\beta \frac{\rho \|\beta\|_{M^{-1}}^2}{2} + \partial_\beta R(\beta).$$

The remaining question is how to compute the Clarke subdifferential of $R(\beta)$. When $\epsilon^2 \leq \rho$ and $\epsilon^2 \geq N\rho$, $R(\beta)$ will enjoy the convex composite structure. Thus, we can get the correct subgradient by invoking the chain rule developed in [19, Theorem 10.6] directly. The more subtle and tricky case is the middle one — $\rho < \epsilon^2 < N\rho$. Without of loss generality, we assume that $M = I$ for simplicity. Based on [6, Theorem 2.3.9], we have

$$\partial_\beta R(\beta) \subseteq \text{Conv}\left\{ \partial_\beta \left(\frac{\epsilon}{N} \sum_{i=1}^N \|s_i^\star \beta\|_1 + \sqrt{\frac{\rho}{N} \sum_{i=1}^N \|(\nabla \ell(\beta^\top X_i) - s_i^\star)\beta\|_2^2} \right),\ s^\star \in \mathcal{S}(\beta) \right\},$$

where $\mathcal{S}(\beta)$ is the optimal solution set and $\text{Conv}\{\cdot\}$ denotes the convex hull. If we assume the inclusion here is tight, then the vanilla subgradient method will converge to the optimal solution with the rate $\mathcal{O}(1/\sqrt{K})$ [4]. Empirically, we find out that the resulting subgradient method works well and the violated case will never happen.

4.2 A New Principled Adversarial Training Procedure for Deep Learning

In this subsection, we develop inexact stochastic gradient-type methods for (3.3) and thus are able to realized the benefits of the proposed DRO model in adversarial learning tasks. As the nonconvexity of $f_\beta(\cdot)$, the inner maximization problem over Δ_i will be no longer tractable. Therefore, we leverage the methodology proposed in [24] to gain the computational efficiency, that is, regarding the dual variable as a modeling parameter. To proceed, we make the same smoothness assumption in [24, Assumption B] (i.e., see Assumption E.1 in appendix for details).

Lemma 4.1 (Convex-concave minimax theorem without compactness)**.** *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and level bounded and $g : \mathbb{R}^m \to \mathbb{R}$ is strongly convex. Then we have*

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^m} \ f(x) + x^\top A y - g(y) = \max_{y \in \mathbb{R}^m} \min_{x \in \mathbb{R}^n} \ f(x) + x^\top A y - g(y).$$

Leveraging Lemma 4.1 and the smoothness assumption, (3.3) leads to a simple and instructive form:

$$\min_\beta \frac{1}{N} \sum_{i=1}^N \max_{\|\Delta_i\|_M \leq \epsilon} \left[\ell(f_\beta(X_i + \Delta_i)) - \lambda \|\Delta_i\|_M^2 \right]. \tag{4.1}$$

In contrast to the vanilla DRO model studied in [24], (4.1) further constrains the perturbation into a Euclidean ball with the correlation information M. Moreover, we can observe that the magnitude of ϵ decides how many martingale constraints will be active, which also perfectly matches our theoretical results and interpretations established for linear regression, see Theorem 3.7 and Remark 3.8.

From a computational viewpoint, if λ is large enough (see Lemma E.2 for details), the inner maximization problem is strongly concave and thus the outer minimization problem over β will be smooth. This motivates Algorithm 1, an inexact stochastic gradient method for Problem (4.1). The convergence guarantee is provided in [24, Theorem 2]. It is worthwhile mentioning that the resulting

new principled adversarial training is extremely easy to implement by only adding three lines of *Pytorch* code based on [24]. We refer the interested readers to Appendix F for details.

Algorithm 1: Martingale Distributionally Robust Optimization with Adversarial Training

Input : Sampling distribution $\widehat{\mathbb{P}}$, stepsize sequence $\{t_k\}_{k=0}^{K-1}$;
for $k = 0, 1, 2, \cdots, K-1$ **do**

Sample $X^k \sim \widehat{\mathbb{P}}$ and find an η-approximate maximizer $\hat{\Delta}^k$ satisfying

$$\|\hat{\Delta}^k - \Delta_k^\star\| \leq \eta, \quad \text{where} \quad \Delta_k^\star = \arg\max_{\|\Delta\|_M \leq \epsilon} \left\{ \ell(f_{\beta^k}(X^k + \Delta)) - \lambda \|\Delta\|_M^2 \right\}.$$

Set $\beta^{k+1} \leftarrow \beta^k - t_k \nabla_\beta \ell(f_{\beta^k}(X^k + \hat{\Delta}^k))$.

5 Numerical Results

In this section, we validate the effectiveness of our methods (referred to as *martingale DRO*) on both linear regression and deep neural networks under the adversarial setting. All simulations are implemented using Python 3.8 on: (1) a computer running Windows 10 with a 2.80GHz, Intel(R) Core(TM) i7-1165G7 processor and 16 GB of RAM, and (2) Google Colab with NVIDIA Tesla P100 GPU and 16 GB of RAM. As for the adversarial setting, we consider three types of attack, the detailed definitions of which are collected in Appendix F.

5.1 Linear Regression

To start with, we demonstrate the effectiveness of the proposed martingale DRO model (3.5) with the quadratic loss function and linear feature mapping, i.e., $\ell(f_\beta(X)) = \frac{1}{2}(Y - b^\top Z)^2$ with $X^\top \triangleq (Y, Z^\top)$ and $\beta^\top \triangleq (1, -b^\top)$, where Z is the feature vector and Y is the target variable. In this experiment, we test our method on three LIBSVM regression real world datasets [1]. More specifically, we randomly select 60% of the data to train the models and the rest as our test data. To showcase the effectiveness of martingale DRO model under adversarial setting, we apply one-step projected-gradient method (PGM) attack [14] on test data and report the performance in terms of the the root-mean-square error (RMSE) on adversarial test data. All numerical results with different step sizes for PGM attack are collected in Figure 1. As we mentioned in Proposition 3.2, since the exact martinagle DRO model (i.e., $\epsilon = 0$) is equivalent to Tikhonov regularization, we choose the same hyperparameter $\rho = 0.08$ for ridge regression and martingale DRO mode for fair comparison. We can easily observe that the martingale DRO can beat other two benchmarks over three real world dataset consistently when the step size for PGD attack is relatively large. This result also corroborates our theoretical intuition — the additional regularization $R(\beta)$ can further improve the adversarial robustness.

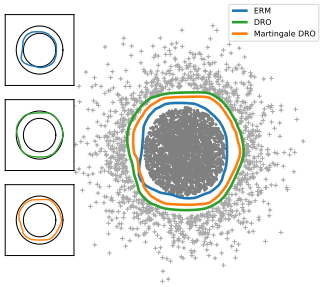


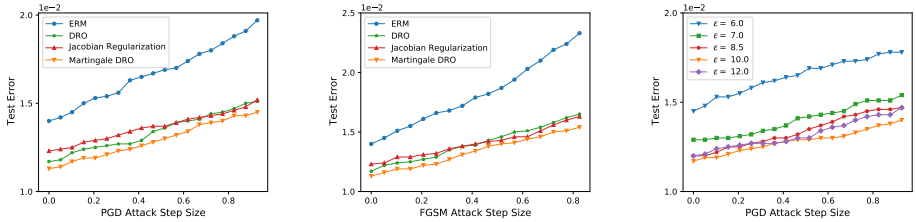
Figure 2: Synthetic data. DRO becomes too conservative, while martingale DRO is less conservative and leaves bigger margins from the sample points than competing methods.

5.2 Deep Neural Network for Adversarial Training

To begin with, we generate the synthetic training data $\{(Y_i, Z_i)\}_{i \in \mathcal{I}}$ with a wide margin as follows: generate i.i.d. $Z_j \sim N(0, I_2)$, where $Z \in \mathbb{R}^2$, I_2 is the identity matrix in $\mathbb{R}^2$; set $\mathcal{I} = \{j : \|Z_j\|_2 \notin (\sqrt{2}/\eta, \eta\sqrt{2})\}$, where $\eta = 1.6$; let $Y_i = \text{sign}(\|Z_i\|_2 - \sqrt{2})$, $\forall i \in \mathcal{I}$. We train a neural network with 3 hidden layers of size 4, 3 and 2 and ELU activations between layers. We compare our approach (cf. martingale DRO) with ERM and the conventional DRO approach developed in [24]. More details about the experiment setup are collected in Appendix F.

Experimental results on synthetic data are shown in Figure 2. Test data are shown in darkgray and gray with different shapes, which are generated by the previous procedure with a smaller margin ($\eta = 1.2$). Classification boundaries are shown in blue, green, and orange for ERM, DRO, and martingale DRO respectively, as well as with the true class boundaries of the test data. Intuitively, the boundary generated by ERM is too close to the true inner boundary since the majority of points are of darkgray class, while the DRO approach pushes the classification boundary outwards. However, as

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression.html



(a) PGD Attack (b) FGSM Attack (c) Sensitivity to ϵ

Figure 3: Compare the proposed martingale DRO with ERM and DRO on the MNIST datasets under PGD and FGSM attack; compare the proposed martingale DRO with different values of ϵ.

(a) Original (b) ERM (c) Jacobian Regularization (d) DRO (e) Martingale DRO

Figure 4: The largest DRO perturbation such that each model makes correct prediction.

illustrated in Figure 2, the DRO approach suffers from over-conservativeness and becomes entangled with the boundary of the outer gray class. In contrast, our martingale DRO boundary lands in between the two extremes and it leaves bigger margin. As shown in the previous sections, decreasing non-zero ϵ's push the perturbed martingale constraints towards the exact martingale constraints and force the classification boundary increasingly inward. More results are in Appendix F.

Then, we validate our method on the MNIST dataset [12]. For the classifier, we train a neural network equipped with 8×8, 6×6, 5×5 convolutional filter layers and ELU activations followed by a fully connected layer and softmax output. To show the robustness of our method, we test the performance of four methods (ERM, DRO, Jacobian regularization [11] and martingale DRO) under the PGD and FGSM attacks (Definition F.1) with test error defined to be: $1 -$ *classification accuracy*.

In Figures 3a and 3b, our martingale DRO model outperforms the other methods and still provide robustness under the ∞-norm FGSM attacks. In Figure 3c, we show the performance of our model with different ϵ's to show the sensitivity to parameter ϵ. As expected, when ϵ is relatively small, the model is not flexible enough and shows large test error. Alternatively, as ϵ becomes large, our model will behave like the original DRO model since the ϵ-constraint is almost invalid in this case.

Figure 4 visualizes the different levels of robustness for the four methods. For each test data point, we perturb the image using the DRO attacks (Definition F.2) with decaying level of perturbation and respectively record the first perturbed images that each model correctly classifies. In Figure 4, the original label is 6 and all methods output the correct prediction, whereas in the adversarial example that the DRO model predicts 6, the correct classification seems unreasonable to human eyes (see Appendix F for more examples). This observation shows an insight that the original DRO model is too conservative in predicting and our model puts more constraints on the perturbation when training thus providing a model that is more consistent to human eyes.

6 Closing Remarks

In this paper, we find that the OT-based DRO model is equivalent to Tikhonov regularization when exact martingale constraints are imposed. Upon this interesting hidden connection, we introduce a new model called the perturbed martingale DRO model, which not only provides a unified viewpoint to several common robust methods but also leads to new regularization tools. Empirically, we validate the effectiveness of our model in addressing the conservativeness issue for the conventional DRO model. From the statistical perspective, how to optimally select the size of uncertainty regions and the perturbation size of the martingale constraint, is a natural problem to be further explored.

References

[1] Mathias Beiglböck, Pierre Henry-Labordere, and Friedrich Penkner. Model-independent bounds for option prices—a mass transport approach. *Finance and Stochastics*, 17(3):477–501, 2013.

[2] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

[3] Jose Blanchet, Karthyek Murthy, and Fan Zhang. Optimal transport-based distributionally robust optimization: Structural properties and iterative schemes. *Mathematics of Operations Research*, 2021.

[4] Stephen Boyd, Lin Xiao, and Almir Mutapcic. Subgradient methods. *lecture notes of EE392o, Stanford University, Autumn Quarter*, 2004:2004–2005, 2003.

[5] Ruidi Chen and Ioannis C Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13), 2018.

[6] Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

[7] Zac Cranko, Zhan Shi, Xinhua Zhang, Richard Nock, and Simon Kornblith. Generalised Lipschitz regularisation equals distributional robustness. In *International Conference on Machine Learning*, pages 2178–2188. PMLR, 2021.

[8] Yan Dolinsky and H Mete Soner. Martingale optimal transport and robust hedging in continuous time. *Probability Theory and Related Fields*, 160(1):391–427, 2014.

[9] Chris Finlay and Adam M Oberman. Scaleable input gradient regularization for adversarial robustness. *arXiv preprint arXiv:1905.11468*, 2019.

[10] Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *arXiv preprint arXiv:1712.06050*, 2017.

[11] Judy Hoffman, Daniel A Roberts, and Sho Yaida. Robust learning with jacobian regularization. *arXiv preprint arXiv:1908.02729*, 2019.

[12] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[13] Jiashuo Liu, Zheyan Shen, Peng Cui, Linjun Zhou, Kun Kuang, and Bo Li. Distributionally robust learning with stable adversarial training. *arXiv preprint arXiv:2106.15791*, 2021.

[14] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[15] Gary C McDonald. Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):93–100, 2009.

[16] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

[17] Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. *Advances in Neural Information Processing Systems*, 30, 2017.

[18] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[19] R Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

[20] Stephan R Sain. *The Nature of Statistical Learning Theory*. Taylor & Francis, 1996.

[21] Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

[22] Soroosh Shafieezadeh Abadeh, Peyman M Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. *Advances in Neural Information Processing Systems*, 28, 2015.

[23] Alexander Shapiro. On duality theory of conic linear problems. In *Semi-infinite programming*, pages 135–165. Springer, 2001.

[24] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

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

[26] Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[27] Volker Strassen. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2):423–439, 1965.

[28] Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

[29] Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See section 3
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]