Large-scale Non-convex Stochastic Constrained Distributionally Robust Optimization

Qi Zhang QZHANG48@BUFFALO.EDU
University at Buffalo
Yi Zhou YI.ZHOU@UTAH.EDU
University of Utah
Ashley Prater-Bennette ASHLEY.PRATER-BENNETTE@US.AF.MIL
Air Force Research Laboratory
Lixin Shen LSHEN03@SYR.EDU
Syracuse University
Shaofeng Zou SZOU3@BUFFALO.EDU
University at Buffalo

Abstract

Distributionally robust optimization (DRO) is a powerful framework for training robust models against data distribution shifts. This paper focuses on constrained DRO, which has an explicit characterization of the robustness level. Existing studies on constrained DRO mostly focus on convex loss function, and exclude the practical and challenging case with non-convex loss function, e.g., neural network. This paper develops a stochastic algorithm and its performance analysis for non-convex constrained DRO. The computational complexity of our stochastic algorithm at each iteration is independent of the overall dataset size, and thus is suitable for large-scale applications. We focus on the general Cressie-Read family divergence defined uncertainty set which includes χ^2-divergences as a special case. We prove that our algorithm finds an ϵ-stationary point with an improved computational complexity than existing methods. Our method also applies to the smoothed conditional value at risk (CVaR) DRO.

1. Introduction

Machine learning algorithms typically employ the approach of Empirical Risk Minimization (ERM), which minimizes the expected loss under the empirical distribution P_0 of the training dataset and assumes that test samples are generated from the same distribution. However, in practice, there usually exists a mismatch between the training and testing distributions due to various reasons, for example, task domains differences [3, 8]; minority group samples [15, 17] and adversarial attacks [14, 25]. Such a mismatch may lead to a significant performance degradation.

This challenge spurred noteworthy efforts on developing a framework of Distributionally Robust Optimization (DRO) e.g., [2, 31, 34]. In DRO, one seeks to optimize the expected loss under the worst-case distribution in an uncertainty set of distributions. Specifically, DRO aims to solve the following problem:

$$\inf_{x} \sup_{Q\sim\mathcal{U}(P_0)} \mathbb{E}_{S\sim Q}\, \ell(x;S), \tag{1}$$

where $\mathcal{U}(P_0)$ is an uncertainty set of distributions centered at P_0, P_0 is the empirical distribution of the training dataset, ℓ is the loss function, and x is the optimization variable. For example, the uncertainty set can be defined as

$$\mathcal{U}(P_0) := \{Q : D(Q\|P_0) \leq \rho\}, \tag{2}$$

where D is some distance-like metric, e.g., Kullback-Leibler (KL) divergence and χ^2 divergence, and ρ is the uncertainty level. In practice, for ease of implementation and analysis, a relaxed formulation of eq. (1), which is referred to as the penalized DRO, is usually solved [20, 23, 28, 35]:

$$\inf_x \sup_Q \mathbb{E}_{S\sim Q}\, \ell(x; S) - \lambda D(Q\|P_0), \tag{3}$$

where $\lambda > 0$ is a fixed hyperparameter that needs to be chosen manually. In contrast to constrained DRO in eq. (1), a regularization term is added to the objective function to keep the distribution Q and the distribution P_0 close, and the hyperparameter λ is manually chosen beforehand to control the tradeoff with minimizing the loss. From a Lagrangian perspective, the dual problems of these two formulations in eq. (1) and eq. (3) are similar. But for the penalized DRO, the Lagrangian multiplier λ is chosen beforehand, whereas in the constrained DRO, λ needs to be optimized, and thus the problem is more challenging. Compared with the penalized DRO setting, the constrained DRO problem in eq. (1) requires that the distribution Q to be strictly in the uncertainty set. Therefore, the obtained solution from the constrained DRO is minimax optimal for distributions in the uncertainty set, whereas it is hard to get such a guarantee for the penalized DRO relaxation.

In this paper, we focus on the challenging constrained DRO problem in eq. (1). In particular, we study the practical non-convex loss functions and focus on the general Cressie-Read family divergence defined uncertainty set [9, 20], which includes, e.g., χ^2 divergence, as a special case (see Section 2 for more details). We also investigate the smoothed conditional value at risk (CVaR) DRO problem. More importantly, we focus on the practical yet challenging large-scale scenario, where P_0 is the empirical distribution of N samples and N is very large. Our contributions can be summarized as below:

- In this paper, we generalize the analysis of the subsampling bias in [23] to the general Cressie-Read family. We further develop a Frank-Wolfe update on the dual variables in order to bound the gap between the objective and its optimal value given the optimization variable x and the biased estimate.

- The dual form of constrained DRO is neither smooth nor Lipschitz, making the convergence analysis difficult. We design an approximation of the original problem, and show that it is smooth and Lipschitz. The approximation error can be made arbitrarily small so that the solution to the approximation is still a good solution to the original. We then prove the strong duality of the approximated problem. Moreover, our strong duality holds for any φ-divergence DRO problem.

- We design a novel algorithm to solve the approximated problem and prove it converges to a stationary point of the constrained DRO problem with computational complexity at each iteration being independent of the training dataset size. Our proposed algorithm converges to a stationary point faster than existing methods [12].

Related work: φ-divergence DRO problems [1, 6] were widely studied, for example, CVaR in [7, 33, 36, 37], χ^2-divergence in [13, 16, 23], KL-divergence in [18, 28, 29] and Sinkhorn distance

[39]. However, the above studies are for some specific divergence function and can not be extended directly to the general Cressie-Read divergence family. The general φ-divergence DRO problem was studied in [20] but their method is for the penalized formulation and does not generalize to the constrained DRO. The general φ-divergence constrained DRO problem was studied in [9, 10, 26]. However, the above studies are for convex loss functions. To the best of our knowledge, our work is the first paper on large-scale non-convex constrained DRO with the general Cressie-Read divergence family. We note that the KL DRO was studied in [29], which however needs an exponential computational complexity. We achieve a polynomial computational complexity for the Cressie-Read divergence family.

2. Preliminaries and Problem Model

2.1. Notations

Let s be a sample in $\mathbb{S}$ and P_0 be the distribution on the points $\{s_i\}_{i=1}^N$, where N is the size of the support. Denote by $\Delta^n := \{\mathbf{p} \in \mathbb{R}^n | \sum_{i=1}^n p_i = 1, p_i \geq 0\}$ the n-dimensional probability simplex. Denote by $x \in \mathbb{R}^d$ the optimization variable. We denote by $\mathbb{1}_{\mathbb{X}}(x)$ the indicator function, where $\mathbb{1}_{\mathbb{X}}(x) = 0$ if $x \in \mathbb{X}$, otherwise $\mathbb{1}_{\mathbb{X}}(x) = \infty$. Let $\ell : \mathbb{R}^d \times \mathbb{S} \to \mathbb{R}$ be a non-convex loss function. Let $\|\cdot\|$ be the Euclidean norm and $(t)_+ := \max\{t, 0\}$ be the positive part of $t \in \mathbb{R}$. Denote ∇_x by the gradient to x. For a function $f : \mathbb{R}^d \to \mathbb{R}$, a point $x \in \mathbb{R}^d$ is said to be an ϵ-optimal solution if $|f(x) - f(x^*)| \leq \epsilon$, where $f(x^*)$ is the optimal value of f. If the function f is differentiable, a point $x \in \mathbb{R}^d$ is said to be first-order ϵ-stationary if $\|\nabla f(x)\| \leq \epsilon$.

2.2. Assumptions

In this paper, we take the following standard assumptions that are commonly used in the DRO literature [9, 23, 28, 29, 36, 39]:

- The non-convex loss function is bounded: $0 \leq \ell(x; s) \leq B$ for some $B > 0, \forall x \in \mathbb{R}^d, s \in \mathbb{S}$.
- The non-convex loss function is G-Lipschitz such that $|\ell(x_1; s) - \ell(x_2; s)| \leq G\|x_1 - x_2\|$ and L-smooth such that $\|\nabla_x \ell(x_1; s) - \nabla_x \ell(x_2; s)\| \leq L\|x_1 - x_2\|$ for any $x_1, x_2 \in \mathbb{R}^d$ and $s \in \mathbb{S}$.

2.3. DRO objective and its dual form

DRO problems shown in eq. (1) under different uncertainty sets are fundamentally different. Consider the uncertainty set defined by φ-divergence $D_\varphi(Q\|P_0)$, which is one of the most common choices in the literature and can be written as $D_\varphi(Q\|P_0) := \int \varphi\left(\frac{dQ}{dP_0}\right) dP_0$, where φ is a non-negative convex function such that $\varphi(1) = 0$ and $\varphi(t) = +\infty$ for ant $t < 0$. Then let the uncertainty set $\mathcal{U}(P_0) := \{Q : D_\varphi(Q\|P_0) \leq \rho\}$ where $\rho > 0$ is the radius of the uncertainty set.

In this paper, we study the general Cressie-Read family of φ-divergence [5, 38], where $\varphi_k(t) := \frac{t^k - kt + k - 1}{k(k-1)}$, $k \in (-\infty, +\infty) \setminus \{0, 1\}$. Let $k_* = \frac{k}{k-1}$. This family includes as special cases χ^2-divergence ($k = 2$) and KL divergence ($k \to 1$). When $k > 2$, the conjugate function of $\varphi_k(t)$ (which will be introduced later) is not smooth, thus the problem becomes hard to solve even in the penalized formulation [20]. In this paper, we focus on $k \in (1, 2]$ ($k_* \in [2, \infty)$).

Solving (1) directly is challenging due to the sup over Q. In [26], a finite-dimensional vector $\mathbf{q}$ was used to parameterize the distributions in the uncertainty set since $Q \ll P_0$ for φ-divergence. Then the DRO problem becomes a convex concave min-max problem. This method can be extended to the case with non-convex loss function by applying the algorithms for non-convex concave min-max problems [24, 30, 40]. However, the dimension of distribution in the uncertainty set is equal to the number of training samples. Thus, the computational complexity at each iteration is linear in the sample size and is prohibitive in large-scale applications.

To obtain a complexity independent of the sample size, one alternative is to use its dual. By duality, we can show that the DRO objective (1) can be equivalently written as [23, 34] $\inf_x \inf_{\lambda\geq 0,\tilde{\eta}\in\mathbb{R}}$ $\mathbb{E}_{S\sim P_0}\left[\lambda\varphi_k^*\left(\frac{\ell(x;S)-\tilde{\eta}}{\lambda}\right)+\lambda\rho+\tilde{\eta}\right]$, where $\varphi_k^*(t')=\sup_t\{t't-\varphi_k(t)\}$ is the conjugate function of $\varphi_k(t')$. In this way, the optimization problem under an unknown distribution is rewritten into one under a known distribution. The subsampling method can then be used, which leads to a complexity independent of the sample size (which will be introduced later). For the Cressie-Read family in (2.3), the corresponding objective can be written as

$$\inf_x \inf_{\lambda\geq 0,\eta\in\mathbb{R}} F(x;\lambda;\eta)=\mathbb{E}_{S\sim P_0}\Big[f(x;\lambda;\eta;S)\Big], \tag{4}$$

where $f(x;\lambda;\eta;s)=\frac{(k-1)^{k_*}}{k}(\ell(x;S)-\eta)_+^{k_*}\lambda^{1-k_*}+\lambda\left(\rho+\frac{1}{k(k-1)}\right)+\eta$. Therefore, we reformulate the DRO problem as one to minimize an objective function under a known distribution, where subsampling method could be used to reduce the complexity.

3. Analysis of Constrained DRO

3.1. Smooth and Lipschitz approximation

For $\lambda\in[0,+\infty),\eta\in\mathbb{R}$, the objective function $F(x;\lambda;\eta)$ is neither smooth nor Lipschitz. Thus it is difficult to implement gradient-based algorithms. In the following, we will construct an approximation of the original problem so that the objective function $F(x;\lambda;\eta)$ becomes smooth and Lipschitz by constraining both λ and η in some bounded intervals.

Since the loss function is bounded such that $0\leq\ell\leq B$, we can show that there exists an upper bound $\bar{\lambda}$ which only depends on k,ρ and B such that the optimal value $\lambda^*\leq\bar{\lambda}$. In this paper, we do not assume that $\lambda^*\geq\lambda_0>0$ as in [39]. Instead, we consider an approximation with $\lambda\in[\lambda_0,\bar{\lambda}]$, and show that the difference between the orignial and the approximation can be bounded. We can show corresponding optimal $\eta^*\in[-\bar{\eta},B]$, where $\bar{\eta}=\bar{\lambda}\left(\frac{k}{(k-1)^{k_*}k_*}\right)^{\frac{1}{k_*-1}}$. The proof can be found in Appendix B.

We show that the difference between the original and the approximation can be bounded in the following lemma.

Lemma 1 $\forall x\in\mathbb{R}^d, 0\leq\lambda_0\leq\bar{\lambda}$, $\left|\inf_{\lambda\in[\lambda_0,\bar{\lambda}],\eta\in[-\bar{\eta},B]}F(x;\lambda;\eta)-\inf_{\lambda\geq 0,\eta\in\mathbb{R}}F(x;\lambda;\eta)\right|\leq 2\lambda_0\rho$.

The proof can be found in Appendix C. Lemma 1 demonstrates that the non-smooth objective function can be approximated by a smooth objective function. A smaller λ_0 makes the gap smaller but the function "less smooth".

3.2. Convexity and smoothness on parameters

The advantage of our approximated problem is that the function is smooth in all x, λ, and η. Moreover, We find that the objective function is convex in λ and η though the loss function is non-convex in x in the Lemma 3. The proof can be found in Appendix D.

4. Mini-batch Algorithm

Existing constrained stochastic algorithm for general non-convex functions [12] can be used to solve the approximated problem directly. However, their method optimizes $y = (x; \lambda; \eta)$ as a whole. It can be seen that the objective function is non-convex in y and the computation complexity to get the ϵ- stationary point is $\mathcal{O}(\epsilon^{-3k_*-5})$.

In Lemma 3, we show that $F(x; z)$ is L_z-smooth in z and L_x-smooth in x. Moreover, $L_z \sim \mathcal{O}(\lambda_0^{-k_*-1})$, which is much larger then L_x when λ_0 is small, since $L_x \sim \mathcal{O}(\lambda_0^{-k_*+1})$. If we optimize all the parameters together, we need to implement non-convex algorithms to optimize a smooth function with a large smooth constant, which is not computationally efficient. However, if we optimize x and z separately, though $L_z > L_x$ which requires more resources to optimize z, the convexity in z makes it faster to converge to the optimal value of z.

This motivates us to consider a stronger convergence criterion. Instead of finding the ϵ- stationary point for $F(y)$, we can find (x, λ, η) such that $|\nabla_x F(x; \lambda; \eta)| \leq \epsilon, |F(x; \lambda; \eta) - \inf_{\lambda' \geq 0; \eta'} F(x; \lambda'; \eta')| \leq \epsilon$. We then provide our Stochastic gradient and Frank-Wolfe DRO algorithm (SFK-DRO), which optimizes x and z separately (see Algorithm 1). Define $D = \max_{z_1, z_2 \in \mathcal{M}} \|z_1 - z_2\|$, $\sigma = \frac{(k-1)^{k_*}}{k} k_* (B + \bar{\eta})^{k_*-1} G \lambda_0^{1-k_*}$, $\Delta = F(x_1; z_1) - \inf_{x, z \in \mathcal{M}} F(x; z)$ and C is a constant such that $C \geq D L_z$. The convergence rate is then provided in the following theorem.

Theorem 2 *With a suitable mini-batch size $n_x \sim \mathcal{O}(\lambda_0^{-2k_*+4} \epsilon^{-2}), n_z \sim \mathcal{O}(\epsilon^{-k_*})$ and $\alpha = \frac{1}{C}, \lambda_0 = \frac{\epsilon}{8\rho}$, for any small $\epsilon > 0$ such that $\frac{D L_z}{L_x} \sim \mathcal{O}(\epsilon^{-2}) \geq 2$ and $\frac{g}{C} \sim \mathcal{O}(\epsilon) \leq 1$, at most $T = 16 C \Delta \epsilon^{-2} \sim \mathcal{O}(\lambda_0^{-k_*-1} \epsilon^{-2})$ iterations are needed to guarantee a stationary point $(x_{t'+1}; z_{t'})$ in expectation: $\mathbb{E}\|\nabla_x F(x_{t'+1}; z_{t'})\| \leq \epsilon, \mathbb{E}\left[\left|F(x_{t'+1}; z_{t'}) - \inf_{\lambda \geq 0; \eta \in \mathbb{R}} F(x_{t'+1}; \lambda; \eta)\right|\right] \leq \epsilon$.*

The proof of Theorem 2 can be found in E and relies on the following lemma for our subsampling method. When we optimize z, an estimator $f_z(x, z) = \sum_{j=1}^{n_z} \frac{f(x; z; s_j)}{n_z}$ is build to estimate $F(x; z) = \mathbb{E}_{S \sim P_0} \left[f(x; z; S) \right]$. Though the estimator is unbiased, in our Frank-Wolfe update process [11, 19, 22] we need to estimate $\min F(x; z)$ via $\mathbb{E} \min f_z(x; z)$. In lemma 4, we show that this gap can be bounded by a decreasing function of the sample batch n_z and is independent of the total number of samples and can be found in Appendix F.

5. Smoothed CVaR

Our algorithm can also solve other DRO problems efficiently, for example, the Smoothed CVaR proposed in [20]. The CVaR DRO is an important φ-divergence DRO problem, where $\varphi(t) = \mathbb{1}_{[0, \frac{1}{\mu})}$ if $0 \leq t < \frac{1}{\mu}$, and $0 < \mu < 1$ is some constant. The dual of CVaR is non-differentiable, which is undesirable from an optimization viewpoint. To solve this problem, [20] proposed a new divergence function, which can be seen as a smoothed version of the CVaR. Their experiment results show the optimization of smoothed CVaR is much easier. However, [20]'s method only works for the penalized formulation of DRO. We show that our method can solve the constrained smoothed CVaR and and the complexity to get the ϵ-stationary point is $\mathcal{O}(\epsilon^{-7})$. The detailed proof can be found in Appendix G.

6. Numerical Results

In this section, we verify our theoretical results in solving an imbalanced classification problem. In the experiment, we consider a non-convex loss function and k is set to be 2 for the Cressie-Read family. We will show that 1) to optimize the same dual objective function, our proposed

Class	0	1	2	3	4	5	6	7	8	9
EMR	77.64	86.19	69.33	54.03	51.53	47.05	87.66	85.35	87.12	83.15
SFK-DRO	76.11	84.71	66.18	**54.95**	**58.65**	49.36	89.06	84.03	88.41	83.09
PAN-DRO	74.92	85.62	65.72	52.69	55.83	**49.50**	88.85	84.06	88.68	81.29

Table 1: Test Accuracy of each class for imbalanced CIFAR 10.

algorithm converges faster than the general Proximal Gradient Descent(PGD) algorithm [12]; 2) The performance proposed algorithm for the constrained DRO problem outperforms or is close to the performance of the penalized DRO with respect to the worst classes. Both of them outperform the baseline. The details can be found in Appendix H.

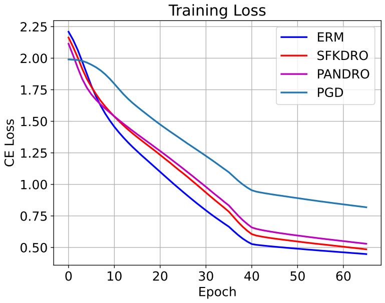

Figure 1: Training curve of classification task.

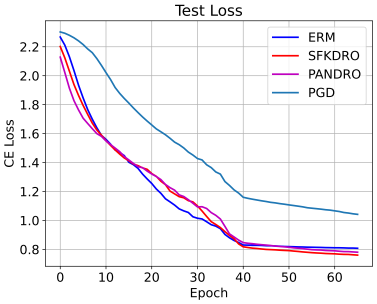

Figure 2: Test curve of classification task.

Results. In Figure 1, 2, we plot the value of the CE loss using different algorithms through the training process. It can be seen that to optimize the same dual objective function with the same learning rate, the PGD algorithm converges slower than our proposed DRO algorithms, which matches our theoretical results. Moreover, compared with ERM, the DRO algorithms have higher training losses but lower test losses, which demonstrates they are robust. We also provide the test accuracy of trained models in Table 1. It can be shown that for class $4, 5, 6$, the accuracies are the lowest due to the limited samples. For these classes, the performance of our SFK-DRO algorithm for the constrained DRO is better or close to the performance of PAN-DRO for the penalized DRO. Both DRO algorithms outperform the vanilla ERM algorithm.

7. Conclusion

In this paper, we developed the first stochastic algorithm for large-scale non-convex stochastic constrained DRO problems in the literature with theoretical convergence and complexity guarantee. We developed a smooth and Lipschitz approximation with bounded approximation error to the original problem. Compared with existing algorithms, the proposed algorithm has an improved convergence rate. The computational complexity at each iteration is independent of the size of the training dataset, and thus our algorithm is applicable to large scale applications. Our results hold for a general family of Cressie-Read divergences. It is of future interest to generalize our results to other distance-like metric, e.g., KL divergence, Wasserstain distance.

References

[1] Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966.

[2] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[3] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128, 2006.

[4] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: rebalanced mixup. In *Proceedings of Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 95–110. Springer, 2020.

[5] Noel Cressie and Timothy RC Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 46(3):440–464, 1984.

[6] Imre Csiszár. On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

[7] Sebastian Curi, Kfir Y Levy, Stefanie Jegelka, and Andreas Krause. Adaptive sampling for stochastic risk-averse learning. In *Proceedings of Advances in Neural Information Processing Systems*, volume 33, pages 1036–1047, 2020.

[8] Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126, 2006.

[9] John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

[10] John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3): 946–969, 2021.

[11] Marguerite Frank, Philip Wolfe, et al. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[12] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155 (1-2):267–305, 2016.

[13] Soumyadip Ghosh, Mark Squillante, and Ebisa Wollega. Efficient stochastic gradient descent for distributionally robust learning. *arXiv preprint arXiv:1805.08728*, 2018.

[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[15] Patrick J Grother, Patrick J Grother, P Jonathon Phillips, and George W Quinn. *Report on the evaluation of 2D still-image face recognition algorithms*. US Department of Commerce, National Institute of Standards and Technology, 2011.

[16] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

[17] Dirk Hovy and Anders Søgaard. Tagging performance correlates with author age. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)*, pages 483–488, 2015.

[18] Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 1(2):9, 2013.

[19] Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of International conference on machine learning*, pages 427–435. PMLR, 2013.

[20] Jikai Jin, Bohang Zhang, Haiyang Wang, and Liwei Wang. Non-convex distributionally robust optimization: Non-asymptotic analysis. In *Proceedings of Advances in Neural Information Processing Systems*, volume 34, pages 2771–2782, 2021.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of Advances in neural information processing systems*, volume 25, 2012.

[22] Simon Lacoste-Julien. Convergence rate of frank-wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

[23] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In *Proceedings of Advances in Neural Information Processing Systems*, volume 33, pages 8847–8860, 2020.

[24] Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *Proceedings of International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020.

[25] 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.

[26] Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Proceedings of Advances in neural information processing systems*, volume 29, 2016.

[27] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[28] Qi Qi, Zhishuai Guo, Yi Xu, Rong Jin, and Tianbao Yang. An online method for a class of distributionally robust optimization with non-convex objectives. In *Proceedings of Advances in Neural Information Processing Systems*, volume 34, pages 10067–10080, 2021.

[29] Qi Qi, Jiameng Lyu, Er Wei Bai, Tianbao Yang, et al. Stochastic constrained dro with a complexity independent of sample size. *arXiv preprint arXiv:2210.05740*, 2022.

[30] Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, 37(3):1087–1121, 2022.

[31] Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

[32] R Tyrrell Rockafellar and Roger JB Wets. *Variational Analysis*. Springer, 1998.

[33] R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

[34] Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017.

[35] Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

[36] Tasuku Soma and Yuichi Yoshida. Statistical learning with conditional value at risk. *arXiv preprint arXiv:2002.05826*, 2020.

[37] Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the cvar via sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[38] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[39] Jie Wang, Rui Gao, and Yao Xie. Sinkhorn distributionally robust optimization. *arXiv preprint arXiv:2109.11926*, 2021.

[40] Zi Xu, Huiling Zhang, Yang Xu, and Guanghui Lan. A unified single-loop alternating gradient projection algorithm for nonconvex–concave and convex–nonconcave minimax problems. *Mathematical Programming*, pages 1–72, 2023.

Appendix A. SFK-DRO algorithm

Algorithm 1 SFK-DRO

Input: Iteration number K, initial point (x_1, z_1), sample numbers n_x, n_z, step size α, and one constant C

```
Let $t = 1$
while $t \leq K$ do
    randomly select $n_x$ samples and compute $\nabla_x f_x(x_t, z_t) = \sum_{i=1}^{n_x} \frac{\nabla_x f(x_t; z_t; s_i)}{n_x}$.
    $x_{t+1} = x_t - \alpha \nabla_x f_x(x_t, z_t)$
    randomly select $n_z$ samples and compute $\nabla_z f_z(x_{t+1}, z_t) = \sum_{j=1}^{n_z} \frac{\nabla_z f(x_{t+1}; z_t; s_j)}{n_z}$
    $e_t$= $\arg\min_{e \in \mathcal{M}} \langle e, \nabla_z f_z(x_{t+1}; z_t) \rangle$
    $d_t = e_t - z_t$
    $g_t = \langle d_t, -\nabla_z f_z(x_{t+1}; z_t) \rangle$
    $\gamma_t = \min \left\{ \frac{g_k}{C}, 1 \right\}$
    $z_{t+1} = z_t + \gamma_t d_t$
    $t = t + 1$
end while
```

$t' = \arg\min_t \|\nabla_x f_x(x_t; z_t)\|^2 + g_t^2$

Output: $(x_{t'+1}, z_{t'})$

Appendix B. Bounds on the parameters

Proof Firstly, we show for bounded loss function ℓ, the optimal value λ^* has an upper bound. If $\lambda^* = 0$, then absolutely it has an upper bound. Otherwise, denote by $\lambda^*(\eta)$ for the optimal value of λ given η and $\lambda^* = \lambda^*(\eta^*)$. We then have $\nabla_\lambda F(x; \lambda^*(\eta); \eta) = 0$ since $F(x; \lambda; \eta)$ is convex in λ and η (which will be shown in). Denote by $\omega = (k(k-1)\rho + 1)^{\frac{1}{k_*}}$. It then follows that

$$\lambda^*(\eta) = (k-1)\omega^{-1} \mathbb{E}_{S \sim P_0} \left[(\ell(x; S) - \eta)_+^{k_*} \right]^{\frac{1}{k_*}}. \tag{5}$$

If $\eta^* \geq 0$, then $\lambda^*(\eta^*) \leq (k-1)\omega^{-1} B$.
If $\eta^* < 0$, combine (4) and (5), the objective changes into

$$\inf_{x; \eta \in \mathbb{R}} \bar{F}(x; \eta) = \omega \left(\mathbb{E}_{S \sim P_0} (\ell(x; S) - \eta)_+^{k_*} \right)^{1/k_*} + \eta.$$

For the optimal value η^*, we have $\nabla_\eta \bar{F}(x; \eta^*) = 0$. It follows that

$$\begin{aligned} &\nabla_\eta \bar{F}(x; \eta^*) \\ = &- \omega \left(\mathbb{E}_{S \sim P_0} (\ell(x; S) - \eta^*)_+^{k_*} \right)^{\frac{1-k_*}{k_*}} \\ &\times \left(\mathbb{E}_{S \sim P_0} (\ell(x; S) - \eta^*)_+^{k_*-1} \right) + 1 = 0. \end{aligned}$$

Therefore, we have

$$\begin{aligned}\frac{1}{\omega} &= \frac{\mathbb{E}_{S\sim P_0}(\ell(x;S)-\eta^*)_+^{k_*-1}}{\left[\mathbb{E}_{S\sim P_0}(\ell(x;S)-\eta^*)_+^{k_*}\right]^{1-\frac{1}{k_*}}} \\ &\geq \frac{|\eta^*|^{k_*-1}}{(B+|\eta^*|)^{k_*-1}},\end{aligned}$$

where the last inequality is due to the fact that $\ell(x;S)$ is bounded. Since $\omega > 1$, we have that

$$|\eta^*| \leq \frac{(\frac{1}{\omega})^{\frac{1}{k_*-1}}B}{1-(\frac{1}{\omega})^{\frac{1}{k_*-1}}}.$$

Thus, from (5) we have that

$$\lambda^* \leq (k-1)\omega^{-1}\left(1+\frac{(\frac{1}{\omega})^{\frac{1}{k_*-1}}}{1-(\frac{1}{\omega})^{\frac{1}{k_*-1}}}\right)B = \bar{\lambda}, \tag{6}$$

where $\bar{\lambda}$ only depands on the parameter k and the upper bound on the loss function B. In addition, for any fixed λ, for the optimal value $\eta^*(\lambda)$, we have $\nabla_\eta F(x;\lambda;\eta^*(\lambda)) = 0$. Thus, we have

$$\mathbb{E}_{S\sim P_0}\left[(\ell(x;S)-\eta^*(\lambda))_+^{k_*-1}\right] = \lambda^{k_*-1}\frac{k}{(k-1)^{k_*}k_*}.$$

Since $\lambda \in [\lambda_0, \bar{\lambda}]$, we have $\eta^* \in [-\bar{\eta}, B]$, where $\bar{\eta} = \bar{\lambda}\left(\frac{k}{(k-1)^{k_*}k_*}\right)^{\frac{1}{k_*-1}}$. This completes the proof. ■

Appendix C. Proof of Lemma 1

Proof We consider the following question first:

$$\sup_{D_{\varphi_k}(Q\|P_0)\leq\rho} \mathbb{E}_{S\sim Q}\,\ell(x;S) - \lambda_0 D_{\varphi_k}(Q\|P_0).$$

Suppose $\zeta(s) = \frac{dQ(s)}{dP_0(s)}$, then the question can be written as

$$\begin{aligned}&\sup_{\zeta\succeq 0}\int \ell(x;S)\zeta - \lambda_0\varphi_k(\zeta)dP_0 \\ &s.t. \int \varphi_k(\zeta)dP_0 \leq \rho, \int \zeta dP_0 = 1.\end{aligned}$$

The Lagrangian of the above problem can be written as

$$\begin{aligned}\mathcal{L}(x;\zeta;\lambda;\eta) = &\int \ell(x;S)\zeta - (\lambda_0+\lambda)\varphi_k(\zeta) - \eta\zeta dP_0 \\ &+ \lambda\rho + \eta.\end{aligned}$$

The problem is equivalent to

$$\sup_{\zeta \succeq 0} \inf_{\lambda \geq 0, \eta} \mathcal{L}(x; \zeta; \lambda; \eta).$$

Since the Slater condition holds, we can exchange the positions of $\sup$ and $\inf$ thus getting the dual form

$$\inf_{\lambda \geq 0, \eta} \lambda\rho + \eta + \sup_{\zeta \succeq 0} \int \ell(x; S)\zeta - (\lambda_0 + \lambda)\varphi_k(\zeta) - \eta\zeta dP_0.$$

Since the maximum operation can be moved inside the integral (Theorem 14.60 of [32]), we have that

$$\begin{aligned}
&\sup_{\zeta \succeq 0} \int \ell(x; S)\zeta - (\lambda_0 + \lambda)\varphi_k(\zeta) - \eta\zeta dP_0 \\
&= \int \sup_{\zeta \succeq 0} \ell(x; S)\zeta - (\lambda_0 + \lambda)\varphi_k(\zeta) - \eta\zeta dP_0 \\
&= \int \sup_{\zeta \succeq 0} \zeta[\ell(x; S) - \eta] - (\lambda_0 + \lambda)\varphi_k(\zeta) dP_0 \\
&= \int ((\lambda_0 + \lambda)\varphi_k)^*(\ell(x; S) - \eta) dP_0.
\end{aligned}$$

For each $\lambda > 0$ we get $(\lambda\varphi(\ell))^* = \lambda\varphi^*(\frac{\ell}{\lambda})$. Therefore, the objective function changes into

$$\inf_{\lambda \geq 0, \eta} \lambda\rho + \eta + (\lambda_0 + \lambda)\mathbb{E}_{S \sim P_0}\varphi_k^*\left(\frac{\ell(x; S) - \eta}{\lambda_0 + \lambda}\right).$$

We then have

$$\begin{aligned}
&\inf_{\lambda \in [\lambda_0, \bar{\lambda}], \eta \in [-\bar{\eta}, B]} F(x; \lambda; \eta) - \lambda_0\rho \\
&= \inf_{\lambda \geq \lambda_0, \tilde{\eta} \in \mathbb{R}} \mathbb{E}_{S \sim P_0}\left[\lambda\varphi_k^*(\frac{\ell(x; S) - \tilde{\eta}}{\lambda}) + (\lambda - \lambda_0)\rho + \tilde{\eta}\right] \\
&= \sup_{D_{\varphi_k}(Q \| P_0) \leq \rho} \mathbb{E}_{S \sim Q}\, \ell(x; S) - \lambda_0 D_{\varphi_k}(Q \| P_0), \qquad (7)
\end{aligned}$$

where the first equality is due to the definition of F, $\lambda^* \in [\lambda_0, \bar{\lambda}], \eta^* \in [-\bar{\eta}, B]$, and the second equality is due to the strong duality we provide above. Moreover, we have

$$\begin{aligned}
&\sup_{D_{\varphi_k}(Q \| P_0) \leq \rho} \mathbb{E}_{S \sim Q}\, \ell(x; S) \\
&- \sup_{D_{\varphi_k}(Q \| P_0) \leq \rho} \mathbb{E}_{S \sim Q}\, \ell(x; S) - \lambda_0 D_{\varphi_k}(Q \| P_0) \\
&\leq \lambda_0\rho. \qquad (8)
\end{aligned}$$

Combining (4),(7) and (8), we complete the proof. ■

Appendix D. Lemma 3 and its proof

Lemma 3 *Define* $z = (\lambda, \eta) \in \mathcal{M}$*, where* $\mathcal{M} = \{(\lambda, \eta) : \lambda \in [\lambda_0, \bar{\lambda}], \eta \in [-\bar{\eta}, B]\}$*. Then* $\forall x \in \mathbb{R}^d, z \in \mathcal{M}$*, the objective function* $F(x; z)$ *is convex and* L_z*-smooth in* z*, where* $L_z = \frac{1}{\lambda_0} + \frac{2(B+\bar{\eta})}{\lambda_0^2} + \frac{(B+\bar{\eta})^2}{2\lambda_0^3}$ *if* $k_* = 2$ *and* $L_z = \frac{(k-1)^{k_*}}{k} k_*(k_* - 1)\left(\frac{(B+\bar{\eta})^{k_*}}{\lambda_0^{k_*+1}} + \frac{(B+\bar{\eta})^{k_*-2}}{\lambda_0^{k_*-1}}\right)$ *if* $k_* > 2$*.*

Moreover, the objective function $F(x; z)$ *is* L_x*-smooth in* x*, where* $L_x = \frac{(k-1)^{k_*}}{k} k_* \lambda_0^{1-k_*}(B + \bar{\eta})^{k_*-2}((k_* - 1)G^2 + (B + \bar{\eta})L)$*.*

Proof From (4) we only need to prove $\phi(x; z) = \mathbb{E}_{S \sim P_0}\left[(\ell(x; S) - \eta)_+^{k_*} \lambda^{1-k_*}\right]$ is convex and smooth in z and smooth in x.

Firstly, we have

$$\nabla_\lambda \phi(x; z) = (1 - k_*)\mathbb{E}_{S \sim P_0}\left[(\ell(x; S) - \eta)_+^{k_*} \lambda^{-k_*}\right]$$

and

$$\nabla_\eta \phi(x; z) = -k_* \mathbb{E}_{S \sim P_0}\left[(\ell(x; S) - \eta)_+^{k_*-1} \lambda^{1-k_*}\right].$$

If $k_* = 2$, the problem becomes a χ^2-DRP problem and $\nabla_\eta \phi(z)$ is not differentiable when $\ell(x; S) - \eta = 0$. For any $z_1 = (\lambda_1; \eta_1), z_2 = (\lambda_2; \eta_2)$ where $\lambda_1, \lambda_2 \in [\lambda_0, \bar{\lambda}]$, we have that for any fixed $s \in \$$ and $a \in [0, 1]$

$$\begin{aligned}
&2\lambda_1\lambda_2(\ell(x; s) - \eta_1)_+(\ell(x; s) - \eta_2)_+ \\
&\leq \lambda_1^2(\ell(x; s) - \eta_2)_+^2 + \lambda_2^2(\ell(x; s) - \eta_1)_+^2.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\lambda_1\lambda_2\big(a^2(\ell(x; s) - \eta_1)_+^2 + (1 - a)^2(\ell(x; s) - \eta_2)_+^2 \\
&\quad + 2a(1 - a)(\ell(x; s) - \eta_1)_+(\ell(x; s) - \eta_2)_+\big) \\
\leq&(a(1 - a)\lambda_1^2 + (1 - a)^2\lambda_1\lambda_2)(\ell(x; s) - \eta_2)_+^2 \\
&+ (a(1 - a)\lambda_2^2 + a^2\lambda_1\lambda_2)(\ell(x; s) - \eta_1)_+^2.
\end{aligned} \tag{9}$$

In addition, we have

$$\begin{aligned}
&(\ell(x; s) - a\eta_1 + (1 - a)\eta_2)_+^2 \\
\leq&a^2(\ell(x; s) - \eta_1)_+^2 + (1 - a)^2(\ell(x; s) - \eta_2)_+^2 \\
&+ 2a(1 - a)(\ell(x; s) - \eta_1)_+(\ell(x; s) - \eta_2)_+.
\end{aligned} \tag{10}$$

Combine (9) and (10), we can get

$$\begin{aligned}
&\frac{1}{a\lambda_1 + (1 - a)\lambda_2}(\ell(x; s) - a\eta_1 - (1 - a)\eta_2)_+^2 \\
\leq&\left(\frac{a}{\lambda_1}(\ell(x; s) - \eta_1)_+^2 + \frac{1 - a}{\lambda_2}(\ell(x; s) - \eta_2)_+^2\right).
\end{aligned} \tag{11}$$

Take expectations for both sides, we have

$$\phi\left(x; az_1 + (1-a)z_2\right) \leq a\phi(x, z_1) + (1-a)\phi(x, z_2),$$

which demonstrates both $\phi(x;z)$ and $F(x;z)$ is convex in z. We then show F is smooth in z. We have that

$$\begin{aligned}
&\|\nabla_z\phi(x;z_1) - \nabla_z\phi(x;z_2))\| \\
=&\left|2\mathbb{E}_{s\sim P_0}\left[\frac{(\ell(x;s)-\eta_1)_+}{\lambda_1} - \frac{(\ell(x;s)-\eta_2)_+}{\lambda_2}\right]\right| \\
&+\left|\mathbb{E}_{s\sim P_0}\left[\frac{(\ell(x;s)-\eta_1)_+^2}{\lambda_1^2} - \frac{(\ell(x;s)-\eta_2)_+^2}{\lambda_2^2}\right]\right| \\
\leq&\left|2\mathbb{E}_{s\sim P_0}\left[\frac{(\ell(x;s)-\eta_1)_+}{\lambda_1} - \frac{(\ell(x;s)-\eta_2)_+}{\lambda_1}\right]\right| \\
&+\left|2\mathbb{E}_{s\sim P_0}\left[\frac{(\ell(x;s)-\eta_2)_+}{\lambda_1} - \frac{(\ell(x;s)-\eta_2)_+}{\lambda_2}\right]\right| \\
&+\left|\mathbb{E}_{s\sim P_0}\left[\frac{(\ell(x;s)-\eta_1)_+^2}{\lambda_1^2} - \frac{(\ell(x;s)-\eta_2)_+^2}{\lambda_1^2}\right]\right| \\
&+\left|\mathbb{E}_{s\sim P_0}\left[\frac{(\ell(x;s)-\eta_2)_+^2}{\lambda_1^2} - \frac{(\ell(x;s)-\eta_2)_+^2}{\lambda_2^2}\right]\right| \\
\leq&\frac{2|\eta_1-\eta_2|}{\lambda_0} + \frac{2(B+\bar{\eta})}{\lambda_0^2}|\lambda_1-\lambda_2| + \frac{2(B+\bar{\eta})}{\lambda_0^2}|\eta_1-\eta_2| \\
&+\frac{(B+\bar{\eta})^2}{\lambda_0^3}|\lambda_1-\lambda_2| \\
\leq&\left(\frac{2}{\lambda_0} + \frac{4(B+\bar{\eta})}{\lambda_0^2} + \frac{(B+\bar{\eta})^2}{\lambda_0^3}\right)|z_1-z_2|.
\end{aligned}$$

Therefore, $\phi(x;z)$ is $\frac{2}{\lambda_0} + \frac{4(B+\bar{\eta})}{\lambda_0^2} + \frac{(B+\bar{\eta})^2}{\lambda_0^3}$-smooth and $F(x,z)$ is $\frac{1}{\lambda_0} + \frac{2(B+\bar{\eta})}{\lambda_0^2} + \frac{(B+\bar{\eta})^2}{2\lambda_0^3}$-smooth in z.

If $k_* > 2$, $\nabla_\eta\phi(z)$ is differentiable. We can get the Hessian matrix of ϕ with respect to z as:

$$H = \begin{bmatrix}
k_*(k_*-1)\mathbb{E}_{S\sim P_0}\left[(\ell(x;S)-\eta)_+^{k_*}\lambda^{-k_*-1}\right], \\
k_*(k_*-1)\mathbb{E}_{S\sim P_0}\left[(\ell(x;S)-\eta)_+^{k_*-1}\lambda^{-k_*}\right]; \\
k_*(k_*-1)\mathbb{E}_{S\sim P_0}\left[(\ell(x;S)-\eta)_+^{k_*-1}\lambda^{-k_*}\right], \\
k_*(k_*-1)\mathbb{E}_{S\sim P_0}\left[(\ell(x;S)-\eta)_+^{k_*-2}\lambda^{1-k_*}\right]
\end{bmatrix}.$$

Suppose a_1, a_2 are the eigenvalues of H. We have

$$\begin{aligned}
a_1 + a_2 = tr(H) = k_*(k_*-1)\mathbb{E}_{S\sim P_0}\Big[& \\
&(\ell(x;S)-\eta)_+^{k_*}\lambda^{-k_*-1} + (\ell(x;S)-\eta)_+^{k_*-2}\lambda^{1-k_*}\Big] \\
&\geq 0
\end{aligned}$$

and

$$\begin{aligned}a_1a_2 = det(H) = &k_*^2(k_*-1)^2\lambda^{-2k_*}\\ &\times\Big(\mathbb{E}_{S\sim P_0}(\ell(x;S)-\eta)_+^{k_*}\mathbb{E}_{S\sim P_0}(\ell(x;S)-\eta)_+^{k_*-2}\\ &-\Big(E_{S\sim P_0}(\ell(x;S)-\eta)_+^{k_*-1}\Big)^2\Big)\geq 0.\end{aligned}$$

Thus H is semi-positive definite which demonstrates ϕ is convex in z. Moreover, the smooth constant should be the largest eigenvalue. Therefore we get

$$L_z = \frac{(k-1)^{k_*}}{k}k_*(k_*-1)\left(\frac{(B+\bar{\eta})^{k_*}}{\lambda_0^{k_*+1}}+\frac{(B+\bar{\eta})^{k_*-2}}{\lambda_0^{k_*-1}}\right)$$

Now we prove the objective is L_x-smooth in x. Firstly, we have

$$\nabla_x\phi(x;z) = k_*\lambda^{1-k_*}\mathbb{E}_{S\sim P_0}\left[(\ell(x;S)-\eta)_+^{k_*-1}\nabla_x\ell\right].$$

For any x_1, x_2 we have that

$$\begin{aligned}&\|\nabla_x\phi(x_1;z)-\nabla_x\phi(x_2;z))\|\\ \leq &k_*\lambda^{1-k_*}\big\|\mathbb{E}_{S\sim P_0}\big[(\ell(x_1;S)-\eta)_+^{k_*-1}\nabla_x\ell(x_1;S)\\ &-(\ell(x_2;S)-\eta)_+^{k_*-1}\nabla_x\ell(x_1;S)\big]\big\|\\ &+k_*\lambda^{1-k_*}\big\|\mathbb{E}_{S\sim P_0}\big[(\ell(x_2;S)-\eta)_+^{k_*-1}\\ &\times(\nabla_x\ell(x_2;S)-\nabla_x\ell(x_1;S))\big]\big\|.\end{aligned}$$

Since we have

$$\begin{aligned}&\|(\ell(x_1;S)-\eta)_+^{k_*-1}-(\ell(x_2;S)-\eta)_+^{k_*-1}\|\\ \leq&(B+\bar{\eta})^{k_*-2}\|\ell(x_1;S)-\ell(x_2;S)\|\\ \leq&(k_*-1)(B+\bar{\eta})^{k_*-2}G\|x_1-x_2\|,\end{aligned}$$

where the first inequality is because both $\ell(x;S)$ and η are bounded. And the second inequality is due to the fact that ℓ is smooth. Thus we have that

$$\begin{aligned}&\|(\ell(x_1;S)-\eta)_+^{k_*-1}-(\ell(x_2;S)-\eta)_+^{k_*-1}\|\\ \leq&k_*\lambda^{1-k_*}(B+\bar{\eta})^{k_*-2}(k_*-1)G^2\|x_1-x_2\|\\ &+k_*\lambda^{1-k_*}(B+\bar{\eta})^{k_*-1}L\|x_1-x_2\|.\end{aligned}$$

Therefore, $\phi(x;z)$ is $k_*\lambda^{1-k_*}(B+\bar{\eta})^{k_*-2}((k_*-1)G^2+(B+\bar{\eta})L)$-smooth and $F(x,z)$ is $\frac{(k-1)^{k_*}}{k}k_*\lambda^{1-k_*}(B+\bar{\eta})^{k_*-2}((k_*-1)G^2+(B+\bar{\eta})L)$-smooth in x.

■

Appendix E. Proof of Theorem 2

Proof For the update of x, we have that

$$x_{t+1} = x_t - \alpha \nabla_x f_x(x_t; z_t).$$

Since $F(x, z)$ is L_x-smooth in x, we have

$$\begin{aligned}
&F(x_{t+1}; z_t) \\
\leq &F(x_t; z_t) + \langle \nabla_x F(x_t; z_t), x_{t+1} - x_t \rangle \\
&+ \frac{L_x}{2} \|x_{t+1} - x_t\|^2 \\
= &F(x_t; z_t) - \alpha \nabla_x f_x(x_t; z_t) \top \nabla_x F(x_t; z_t) \\
&+ \frac{\alpha^2 L_x}{2} \|\nabla_x f_x(x_t; z_t)\|^2 \\
= &F(x_t; z_t) - \alpha \nabla_x f_x(x_t; z_t) \top \nabla_x F(x_t; z_t) \\
&+ \frac{\alpha^2 L_x}{2} \|\nabla_x f_x(x_t; z_t) - \nabla_x F_x(x_t; z_t) + \nabla_x F_x(x_t; z_t)\|^2 \\
\leq &F(x_t; z_t) - \alpha \nabla_x f_x(x_t; z_t) \top \nabla_x F(x_t; z_t) \\
&+ \alpha^2 L_x \|\nabla_x f_x(x_t; z_t) - \nabla_x F_x(x_t; z_t)\|^2 \\
&+ \alpha^2 L_x \|\nabla_x F_x(x_t; z_t)\|^2.
\end{aligned} \tag{12}$$

Given x_t and z_t, take the expectation for both sides of (24), we have that

$$\begin{aligned}
&\mathbb{E}[F(x_{t+1}; z_t) | x_t, z_t] \\
\leq &F(x_t; z_t) - \alpha \|\nabla_x F_x(x_t; z_t)\|^2 \\
&+ \alpha^2 L_x \mathbb{E}[\|\nabla_x f_x(x_t; z_t) - \nabla_x F_x(x_t; z_t)\|^2 | x_t, z_t] \\
&+ \alpha^2 L_x \|\nabla_x F_x(x_t; z_t)\|^2.
\end{aligned} \tag{13}$$

If $\alpha \leq \frac{1}{2L_x}$ we have that $\alpha^2 L_x \leq \frac{\alpha}{2}$ and

$$\begin{aligned}
\frac{\alpha}{2} \|\nabla_x F_x(x_t; z_t)\|^2 \leq &F(x_t; z_t) - \mathbb{E}[F(x_{t+1}; z_t) | x_t, z_t] \\
&+ \alpha^2 L_x \frac{\sigma^2}{n_x},
\end{aligned}$$

where the inequality is because $\nabla_x f(x; z; s) \leq \frac{(k-1)^{k_*}}{k} k_* (B + \bar{\eta})^{k_* - 1} G \lambda_0^{1-k_*} = \sigma$ is bounded. After that, we take expectations for both sides and we have

$$\begin{aligned}
\frac{\alpha}{2} \mathbb{E}\left[\|\nabla_x F_x(x_t; z_t)\|^2\right] \leq &\mathbb{E}[F(x_t; z_t)] - \mathbb{E}[F(x_{t+1}; z_t)] \\
&+ \alpha^2 L_x \frac{\sigma^2}{n_x}.
\end{aligned} \tag{14}$$

For the update of z and $\forall \gamma \in [0,1]$, we get an affine invariant version of the standard descent Lemma ((1.2.5) in [27])

$$\begin{aligned} & f_z(x_{t+1}; z_{t+1}) \\ \leq & f_z(x_{t+1}; z_t) + \gamma \langle \nabla_z f_z(x_{t+1}; z_t), d_t \rangle + \frac{\gamma^2}{2} C, \end{aligned}$$

where $C \geq DL_z$. In our algorithm we have $\gamma_t = \min\left\{\frac{g_t}{C}, 1\right\}$ and

$$\begin{aligned} \frac{g_t}{C} \leq & \frac{D\|\nabla_z f_z(x_{t+1}; z_t)\|}{DL_z} \\ \leq & \frac{(\rho + \frac{1}{k(k-1)}) + \frac{(k-1)^{k_*}(k_*-1)}{k}\lambda_0^{-k_*}}{L_z}. \end{aligned}$$

Since $L_z \sim \mathcal{O}(\lambda_0^{-k_*-1})$ thus for small λ_0 we have $\frac{g_t}{C} \leq 1$. Consequently, we can assume $\gamma = \frac{g_t}{C}$ and we have

$$\begin{aligned} & f_z(x_{t+1}; z_{t+1}) \\ \leq & f_z(x_{t+1}; z_t) - \frac{g_t}{C} g_t + \frac{(\frac{g_t}{C})^2}{2} C. \end{aligned} \tag{15}$$

Since $f_z(x; z)$ is convex in z, we have that $g_t \geq f_z(x_{t+1}; z_t) - \min_{z \in \mathcal{M}} f_z(x_{t+1}; z)$. Take expectations for both sides of (15), we have that

$$\mathbb{E}\left[\frac{g_t^2}{2C}\right] \leq \mathbb{E}[F(x_{t+1}; z_t)] - \mathbb{E}[F(x_{t+1}; z_{t+1})]. \tag{16}$$

By recursively adding (14) and (16), we have that

$$\begin{aligned} & \frac{1}{T}\sum_{t=1}^{T} \frac{\alpha}{2}\mathbb{E}\left[\|\nabla_x F_x(x_t; z_t)\|^2\right] + \mathbb{E}\left[\frac{g_t^2}{2C}\right] \\ \leq & \frac{F(x_1; z_1) - \mathbb{E}[F(x_{T+1}; z_{T+1})]}{T} + \alpha^2 L_x \frac{\sigma^2}{n_x}. \end{aligned}$$

Since $DL_z \sim \mathcal{O}(\lambda_0^{-k_*-1})$ and $L_x \sim \mathcal{O}(\lambda_0^{-k_*+1})$, we can find λ_0 small enough such that $C \geq DL_z \geq 2L_x$. Set $\alpha = \frac{1}{C}$, we then have that

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla_x F_x(x_t; z_t)\|^2\right] + \mathbb{E}\left[g_t^2\right] \leq \frac{2C\Delta}{T} + \frac{L_x \sigma^2}{n_x C}.$$

From Jensen's inequality, we have that

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla_x F_x(x_t; z_t)\||\right]^2 + \mathbb{E}\left[g_t\right]^2 \leq \frac{2C\Delta}{T} + \frac{L_x \sigma^2}{n_x C}.$$

When we set $T = 16C\Delta\epsilon^{-2} \sim \mathcal{O}(\lambda_0^{-k_*-1}\epsilon^{-2}), n_x = \frac{8L_x\sigma^2}{C\epsilon^2}$, for some $t \in [1, T]$ we have

$$\mathbb{E}\left[\|\nabla_x F_x(x_t; z_t)\|\right] \leq \frac{\epsilon}{2}$$

and

$$\mathbb{E}[g_t] \leq \frac{\epsilon}{2}$$

Since $F(x; z)$ is L_x-smooth in x, we have that

$$\begin{aligned}\|\nabla_x F_x(x_{t+1}; z_t) - \nabla_x F_x(x_t; z_t)\| &\leq L_x\|x_{t+1} - x_t\| \\ &\leq L_x\alpha\|\nabla_x f_x(x_t; z_t)\|.\end{aligned}$$

In addition,

$$\begin{aligned}&\mathbb{E}[\|\nabla_x f_x(x_t; z_t)\|^2] \\ \leq& 2\mathbb{E}[\|\nabla_x f_x(x_t; z_t) - \nabla_x F_x(x_t; z_t)\|^2 | x_t, z_t] \\ &+ 2\mathbb{E}[\|\nabla_x F_x(x_t; z_t)\|^2] \\ \leq& \frac{\epsilon^2}{2} + \frac{C\epsilon^2}{4L_x}\end{aligned}$$

Thus

$$\mathbb{E}\left[L_x\alpha\|\nabla_x f_x(x_t; z_t)\|\right] \leq \frac{L_x}{C}\sqrt{\frac{\epsilon^2}{2} + \frac{C\epsilon^2}{4L_x}} \leq \frac{\epsilon}{2}$$

since $C \geq 2L_x$. Therefore, we have

$$\mathbb{E}[\|\nabla_x F_x(x_{t+1}; z_t)\|] \leq \epsilon.$$

In addition, we have

$$g_t \geq f_z(x_{t+1}; z_t) - \min_{z \in \mathcal{M}} f_z(x_{t+1}; z).$$

Given x_{t+1} and z_t, take expectations for both sides and we have

$$\mathbb{E}[g_t | x_{t+1}, z_t] \geq F(x_{t+1}; z_t) - \mathbb{E}\left[\min_{z \in \mathcal{M}} f_z(x_{t+1}; z)\right].$$

By Lemma 4 we can get a $n_z \sim \mathcal{O}(\epsilon^{-k_*})$ such that $3B\sqrt{1 + k(k-1)\rho}\sqrt{\frac{4+\log(n_z)}{4n_z}} < \frac{\epsilon}{4}$ if $k_* = 2$ or $3B(1 + k(k-1)\rho)^{\frac{1}{k}}\left(\frac{1}{n_z} + \frac{1}{2^{k_*-1}(k_*-2)n_z}\right)^{\frac{1}{k_*}} < \frac{\epsilon}{4}$ if $k_* > 2$, we have

$$\left|\inf_{z \in \mathcal{M}}\left[F(x_{t+1}; z)\right] - \mathbb{E}\left[\inf_{z \in \mathcal{M}} f_z(x_{t+1}; z)\right]\right| \leq \frac{\epsilon}{4}.$$

By Lemma 1, when $\lambda_0 = \frac{\epsilon}{8\rho}$, we have

$$\left|\inf_{\lambda \in [\lambda_0, \bar{\lambda}], \eta \in [-\bar{\eta}, B]} F(x; \lambda; \eta) - \inf_{\lambda \geq 0, \eta \in \mathbb{R}} F(x; \lambda; \eta)\right| \leq \frac{\epsilon}{4}.$$

Thus we have

$$F(x_{t+1}; z_t) - \inf_{\lambda \geq 0, \eta \in \mathbb{R}} F(x; \lambda; \eta) \leq \epsilon. \tag{17}$$

which completes the proof. ■

Appendix F. Lemma 4 and its proof

Lemma 4 *For any bounded loss ℓ, $|\inf_{z\in\mathcal{M}}[F(x_{t+1};z)] - \mathbb{E}[\inf_{z\in\mathcal{M}} f_z(x_{t+1};z)]| \leq \mathcal{O}({n_z}^{-\frac{1}{k_*}})$.*

The (20) of [23] provides an inverse-cdf formulation of the DRO problem. By implementing the inverse-cdf formulation, the (42) and remark 1 of [23] show that

$$\begin{aligned}
&\left|\min_{z\in\mathcal{M}}[F(x;z)] - \mathbb{E}\left[\min_{z\in\mathcal{M}} f_z(x;z)\right]\right| \\
&\leq \int_0^1 (r(\beta)-r(1))(\beta\cdot h(\beta))' d\beta \\
&\leq \|r\|_k \|(\beta\cdot h(\beta))'\|_{k_*}, \qquad (18)
\end{aligned}$$

where $r \in \mathcal{R} := \{r : [0,1] \to \mathbb{R}_+ | \int_0^1 r(\beta)d\beta = 1 \text{ and } \int_0^1 \varphi_k(r(\beta))d\beta \leq \rho\}$, $h = 3B\min\left\{\sqrt{\frac{1}{\beta n_z}}, 1\right\}$ and the second inequality is due to the Hölder's inequality. Note this inequality holds for any fixed x, no matter whether the loss function is convex or not.
Since $\int_0^1 \varphi_k(r(\beta))d\beta \leq \rho$, we have that

$$\|r\|_k^k \leq 1 + k(k-1)\rho. \qquad (19)$$

Moreover, we have that

$$\begin{aligned}
&\|(\beta\cdot h(\beta))'\|_{k_*}^{k_*} \\
&= \int_0^{\frac{1}{n_z}} (3B)^{k_*} d\beta + \int_{\frac{1}{n_z}}^1 (\frac{3B}{2})^{k_*}\sqrt{\frac{1}{(\beta n_z)^{k_*}}} d\beta. \qquad (20)
\end{aligned}$$

For $k_* = 2$, we have

$$\|(\beta\cdot h(\beta))'\|_2^2 = (3B)^2 \frac{4+\log(n_z)}{4n_z} \qquad (21)$$

and if $k_* > 2$, we have that

$$\begin{aligned}
\|(\beta\cdot h(\beta))'\|_{k_*}^{k_*} &\leq (3B)^{k_*}\left(\frac{1}{n_z} + \frac{1}{2^{k_*}}\frac{2}{k_*-2}\frac{n^{0.5k_*-1}-1}{n^{0.5k_*}}\right) \\
&\leq (3B)^{k_*}\left(1 + \frac{1}{2^{k_*-1}(k_*-2)}\right)\frac{1}{n_z}. \qquad (22)
\end{aligned}$$

Combine (18),(19),(21) and (22), we can get the lemma and complete the proof.

Appendix G. Proof of smoothed CVaR

The divergence function of smoothed CVaR is

$$\varphi_s(t) = \begin{cases} t\log(t) + \frac{1-\mu t}{\mu}\log(\frac{1-\mu t}{1-\mu}), & t \in [0, \frac{1}{\mu}); \\ +\infty, & \text{otherwise.} \end{cases} \qquad (23)$$

The corresponding conjugate function is

$$\varphi_s^*(t) = \frac{1}{\mu}\log(1-\mu+\mu\exp(t)). \tag{24}$$

The objective function is then written as

$$\begin{aligned}&\inf_x \inf_{\lambda\geq 0,\eta\in\mathbb{R}} F_s(x;\lambda;\eta)\\ =&\mathbb{E}_{S\sim P_0}\left[\lambda\varphi_s^*(\frac{\ell(x;S)-\eta}{\lambda})+\lambda\rho+\eta\right].\end{aligned} \tag{25}$$

We will show that there exist upper bounds for the optimal values λ^* and η^* later. There exists a $\bar{\lambda}>0$ only depends on μ, B and ρ such that $\lambda^*\in[0,\bar{\lambda}]$ and $\eta^*\in[0,B]$.

This objective function is non-smooth when $\lambda\to 0$. Therefore, we take a similar approach as the one in Section 3.1 to approximate the original problem with $\lambda\in[\lambda_0,\bar{\lambda}]$. We bound the difference in the following lemma.

Lemma 5 $\forall x\in\mathbb{R}^d, \lambda_0\geq 0$,

$$\left|\inf_{\lambda\in[\lambda_0,\bar{\lambda}],\eta\in[0,B]} F_s(x;\lambda;\eta)-\inf_{\lambda\geq 0,\eta\in\mathbb{R}} F_s(x;\lambda;\eta)\right|\leq 2\lambda_0\rho.$$

The proof is similar to Lemma 1 thus is omitted here.

In addition, we will show that $F_s(x;z)$ is L_z'-smooth and convex in z, where $L_z'\sim\mathcal{O}(\lambda_0^{-3})$ if $\lambda\in[\lambda_0,\bar{\lambda}]$. Also it is easy to get $F_s(x;z)$ is L_x'-smooth in x, where $L_x'\sim\mathcal{O}(\lambda_0^{-2})$. Similar to eq. (42) and Remark 1 in [23], we can prove that

$$|\min_{z\in\mathcal{M}}[F_s(x_{t+1};z)]-\mathbb{E}\left[\min_{z\in\mathcal{M}} f_s(x_{t+1};z)\right]|\sim\mathcal{O}(n_s^{-0.5}). \tag{26}$$

We then use Algorithm 1 directly and the complexity to get the ϵ-stationary point is $\mathcal{O}(\epsilon^{-7})$.
Proof Bounded parameters: We have

$$\nabla_t\varphi_s^*(t)=\frac{1}{\mu}\frac{\mu\exp(t)}{1-\mu+\mu\exp(t)}\leq\frac{1}{\mu}$$

and

$$\nabla_t^2\varphi_s^*(t)=\frac{1}{\mu}\frac{\mu(1-\mu)\exp(t)}{(1-\mu+\mu\exp(t))^2}\leq\frac{1}{4\mu}$$

$$\nabla_\eta F_s(x;\lambda;\eta)=1-\mathbb{E}_{S\sim P_0}\varphi_s^{*\prime}(\frac{\ell(x;S)-\eta}{\lambda}).$$

$$\nabla_\eta^2 F_s(x;\lambda;\eta)=\frac{1}{\lambda}\mathbb{E}_{S\sim P_0}\varphi_s^{*\prime\prime}(\frac{\ell(x;S)-\eta}{\lambda}).$$

If $\eta>B$, then $\nabla_\eta F_s(x;\lambda;\eta)<0$. If $\eta<0$, then $\nabla_\eta F_s(x;\lambda;\eta)>0$. Thus $\eta^*\in[0,B]$.

$$\begin{aligned}\nabla_\lambda F_s(x;\lambda;\eta) =& \rho + \mathbb{E}_{S\sim P_0}\Big[\varphi_s^*(\frac{\ell(x;S)-\eta}{\lambda}) \\ & - \varphi_s^{*\prime}(\frac{\ell(x;S)-\eta}{\lambda})\frac{\ell(x;S)-\eta}{\lambda}\Big].\end{aligned}$$

$$\nabla_\lambda^2 F_s(x;\lambda;\eta) = \frac{1}{\lambda^3}\mathbb{E}_{S\sim P_0}\left[\varphi_s^{*\prime\prime}(\frac{\ell(x;S)-\eta}{\lambda})(\ell(x;S)-\eta)^2\right].$$

The second-order demonstrates that the $F_s(x;\lambda;\eta)$ is convex in λ. We then show that λ^* has an upper bound. If $\nabla_\lambda F_s(x;\lambda;\eta) \geq 0$ when $\lambda \to 0$, then $\lambda^* = 0$. If $\nabla_\lambda F_s(x;\lambda;\eta) < 0$ when $\lambda \to 0$, since both $\varphi_s^*(t), \varphi_s^{*\prime}(t)$ are increaing with t, we have that

$$\nabla_\lambda F_s(x;\lambda;\eta) > \rho + \varphi_s^*(\frac{-B}{\lambda}) - \frac{B}{\mu\lambda}. \tag{27}$$

Since we know $g(\lambda) = \rho + \varphi_s^*(\frac{-B}{\lambda}) - \frac{B}{\mu\lambda}$ is increasing with λ and $g(\lambda) < 0$ when $\lambda \to 0$, $g(\lambda) = \rho > 0$ when $\lambda \to \infty$. Thus we can find $\bar\lambda > 0$ that $g(\bar\lambda) = 0$. Moreover, the value of $\bar\lambda$ only depends on μ, B and ρ.

Smoothness: fix x, the Hessian matrix of $F_s(x;z)$ with respect to z as:

$$H_s = \begin{bmatrix} \frac{1}{\lambda}\mathbb{E}_{S\sim P_0}\varphi_s^{*\prime\prime}(\frac{\ell(x;S)-\eta}{\lambda}), \\ \frac{1}{\lambda^2}\mathbb{E}_{S\sim P_0}\left[\varphi_s^{*\prime\prime}(\frac{\ell(x;S)-\eta}{\lambda})(\ell(x;S)-\eta)\right]; \\ \frac{1}{\lambda^2}\mathbb{E}_{S\sim P_0}\left[\varphi_s^{*\prime\prime}(\frac{\ell(x;S)-\eta}{\lambda})(\ell(x;S)-\eta)\right], \\ \frac{1}{\lambda^3}\mathbb{E}_{S\sim P_0}\left[\varphi_s^{*\prime\prime}(\frac{\ell(x;S)-\eta}{\lambda})(\ell(x;S)-\eta)^2\right] \end{bmatrix}.$$

Suppose a_3, a_4 are the eigenvalues of H_s. We have $a_3 + a_4 > 0$ and $a_3 a_4 \geq 0$. And the function is L_z'-smooth and convex in z, where $L_z' \sim \mathcal{O}(\lambda_0^{-3})$ if $\lambda \in [\lambda_0, \bar\lambda]$. Also it is easy to get $F_s(x;\lambda;\eta)$ is L_x'-smooth in x, where $L_x' \sim \mathcal{O}(\lambda_0^{-2})$.

Bounded gap: denote $f_s(x,z) = \sum_{i=1}^{n_s}\frac{f(x;z;s)}{n_s}$, in order to use algorithm 1 directly, we need to estimate $\min F(x;z)$ via $\mathbb{E}\min f_z(x;z)$ and bound the gap. From the (42) and remark 1 of [23], we have that

$$\begin{aligned}&\left|\min_{z\in\mathcal{M}}[F_s(x_{t+1};z)] - \mathbb{E}\left[\min_{z\in\mathcal{M}} f_s(x_{t+1};z)\right]\right| \\ &\leq \int_0^1 (r(\beta)-r(1))(\beta\cdot h(\beta))' d\beta \end{aligned}\tag{28}$$

where $r \in \mathcal{R} := \{r : [0,1] \to \mathbb{R}_+ | \int_0^1 r(\beta)d\beta = 1$ and $\int_0^1 \varphi_s(r(\beta))d\beta \leq \rho\}$, $h = 3B\min\left\{\sqrt{\frac{1}{\beta n_s}}, 1\right\}$ Since $\int_0^1 \varphi_s(r(\beta))d\beta \leq \rho\}$, we have that $r(\beta) \leq \frac{1}{\mu}$ for any Moreover, we have that

$$\begin{aligned}\int_0^1 (\beta\cdot h(\beta))' d\beta =& \int_0^{\frac{1}{n_s}} 3B d\beta + 3B\int_{\frac{1}{n_s}}^1 \sqrt{\frac{1}{\beta n_s}} d\beta \\ =& \frac{3B}{n_s} + 6B(\sqrt{\frac{1}{n_s}} - \frac{1}{n_s})\end{aligned}$$

Thus the gap $\left|\min_{z\in\mathcal{M}}\left[F_s(x_{t+1};z)\right] - \mathbb{E}\left[\min_{z\in\mathcal{M}} f_s(x_{t+1};z)\right]\right| \sim \mathcal{O}(n_s^{-0.5})$. We then can use algorithm 1 directly and the complexity to get the ϵ-stationary point is $\mathcal{O}(\epsilon^{-7})$. ■

Appendix H. Experiments

In this section, we provide the details of our experiments.

Tasks. We conduct experiments on the imbalanced CIFAR-10 dataset, following the experimental setting in [4, 20]. The original CIFAR-10 test dataset consists of 10 classes, where each of the classes has 5000 images. We randomly select training samples from the original set for each class with the following sampling ratio: $\{0.804, 0.543, 0.997, 0.593, 0.390, 0.285, 0.959, 0.806, 0.967, 0.660\}$. We keep the test dataset unchanged.

Models. We learn the standard Alexnet model in [21] with the standard cross-entropy (CE) loss. For the comparison of convergence rate, we optimize the same dual objective with the PGD algorithm in [12]. To compare robustness, we optimize the ERM via vanilla SGD. In addition, we propose an algorithm PAN-DRO, which fixes λ and only optimizes η and the neural network. Thus it gets the solution for the penalized DRO problem.

Training details. We set $\lambda_1 = 1, \eta_1 = 0$, $\lambda_0 = 0.1, -\bar{\eta} = -10$, and the upper bounds $\bar{\lambda} = 10, B = 10$. To achieve a faster optimization rate, we set the learning rate $\alpha = 0.01$ before the first 40 epochs and $\alpha = 0.001$ after. The mini-batch size is chosen to be 128. All of the results are moving averaged by a window with size 5. The simulations are repeated by 4 times.

Appendix I. Complexity of PGD

I.1. Cressie-Read family

In the PGD algorithm [12], $y = (x;\lambda;\eta)$ is optimized as a whole. From Lemma 3, we know $F(y) = F(x;z)$ is L_z-smooth in z and L_x-smooth in x. Moreover, it is not hard to show that $\nabla_x F(x;z)$ is L_{xz}-Lipschitz in z and $\nabla_z F(x;z)$ is L_{xz}-Lipschitz in x, where $L_{zx}, L_{xz} \sim \mathcal{O}(\lambda_0^{-k_*})$. Therefore, we have

$$\begin{aligned}
\|\nabla_y F(y_1) - \nabla_y F(y_2)\| =& \|\nabla_x F(x_1;z_1) - \nabla_x F(x_2;z_2)\| + \|\nabla_z F(x_1;z_1) - \nabla_z F(x_2;z_2)\| \\
\leq& \|\nabla_x F(x_1;z_1) - \nabla_x F(x_1;z_2)\| + \|\nabla_z F(x_1;z_1) - \nabla_z F(x_1;z_2)\| \\
&+ \|\nabla_x F(x_1;z_2) - \nabla_x F(x_2;z_2)\| + \|\nabla_z F(x_1;z_2) - \nabla_z F(x_2;z_2)\| \\
\leq& L_{xz}\|z_1 - z_2\| + L_z\|z_1 - z_2\| + L_{zx}\|x_1 - x_2\| + L_x\|x_1 - x_2\| \\
\leq& (L_x + L_z + L_{xz} + L_{zx})\|y_1 - y_2\|.
\end{aligned}$$

Thus, $F(y)$ is L_y-smooth in y, where $L_y = L_x + L_z + L_{xz} + L_{zx} \sim \mathcal{O}(\lambda_0^{-k_*-1})$. According to Corollary 3 in [12] and $\lambda_0 \sim \mathcal{O}(\epsilon)$, we can get the ϵ- stationary point with the number of iterations $T \sim \mathcal{O}(\lambda_0^{-k_*-1}\epsilon^{-2})$ and batch size $n_p \sim \mathcal{O}(\lambda^{-2k_*}\epsilon^{-2})$. Thus, the total complexity is $\mathcal{O}(\epsilon^{-3k_*-5})$.

I.2. Smoothed CVaR

Similar to Cressie-Read family, we can show that $F_s(y)$ is L'_y-smooth in y, where $L'_y \sim \mathcal{O}(\lambda_0^{-3})$. According to Corollary 3 in [12], we can get the ϵ- stationary point with the number of iterations $T \sim \mathcal{O}(\lambda_0^{-3}\epsilon^{-2})$ and batch size $n_p \sim \mathcal{O}(\lambda^{-2}\epsilon^{-2})$. Thus, the total complexity is $\mathcal{O}(\epsilon^{-9})$.