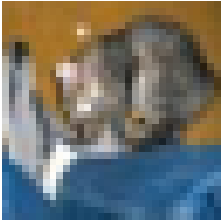
Principled Learning Method for Wasserstein Distributionally Robust Optimization with Local Perturbations

Yongchan Kwon [1] Wonyoung Kim [2]
Joong-Ho Won [2] Myunghee Cho Paik [2]

[1]Department of Biomedical Data Science, Stanford University

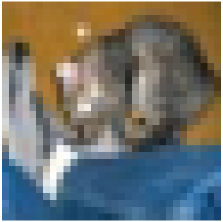
[2]Department of Statistics, Seoul National University

Contact: yckwon@stanford.edu



CIFAR-10: 94.1 % $\to$?? %
CIFAR-100: 74.4 % $\to$?? %

Motivation: state-of-the-art models are not robust




CIFAR-10: 94.1% → 73.0 % (**21.1** % drop)
CIFAR-100: 74.4% → 31.6 % (**42.8** % drop)

Overviews

- In this paper, we study Wasserstein distributionally robust optimization (WDRO) to make models robust.
- We develop a **principled and tractable** statistical inference method for WDRO.
- We formally present a locally perturbed data distribution and provide WDRO inference when **data are locally perturbed**.

Statistical learning problems

- Many statistical learning problems can be expressed by an optimization problem as follows:

$$\inf_{h \in \mathcal{H}} R(\mathbb{P}_{\mathrm{data}}, h) := \inf_{h \in \mathcal{H}} \int_{\mathcal{Z}} h(\zeta) d\mathbb{P}_{\mathrm{data}}(\zeta).$$

- Given observations $z_1, \ldots, z_n \sim \mathbb{P}_{\mathrm{data}}$ and the empirical distribution $\mathbb{P}_n := n^{-1} \sum_{i=1}^{n} \delta_{z_i}$, the empirical risk minimization (ERM) can be represented as

$$\inf_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} h(z_i). \tag{1}$$

- A solution of (1) asymptotically minimizes the true risk, but it performs poorly when **the test data distribution is different** from $\mathbb{P}_{\mathrm{data}}$.

Wasserstein distributionally robust optimization (WDRO)

- WDRO is the problem of learning a model minimizes the **worst-case risk** over the Wasserstein ball:

$$\inf_{h\in\mathcal{H}} \underbrace{\sup_{\mathbb{Q}\in\mathfrak{M}_{\alpha_n,p}(\mathbb{P}_n)} R(\mathbb{Q},h)}_{\text{worst-case risk}},$$

where $\mathfrak{M}_{\alpha_n,p}(\mathbb{P}_n)$ is the Wasserstein ball, a set of probability measures whose p-Wasserstein metric from $\mathbb{P}_n$ is less than $\alpha_n > 0$.

Illustration of WDRO

In ERM,

$$\inf_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} h(z_i)$$

In WDRO,

$$\inf_{h\in\mathcal{H}} \underbrace{\sup_{\mathbb{Q}\in\mathfrak{M}_{\alpha_n,p}(\mathbb{P}_n)} R(\mathbb{Q},h)}_{\text{worst-case risk}}$$

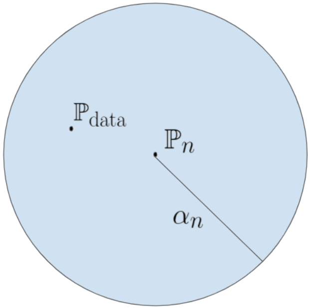


Figure: Illustration of Wasserstein ball $\mathfrak{M}_{\alpha_n,p}(\mathbb{P}_n)$.

▷ By the design of the local worst-case risk, a solution to WDRO can avoid overfitting to $\mathbb{P}_n$ and learn a robust model.

Main challenges in WDRO

WDRO is a powerful framework to train robust models! However, there are challenges.

1. Exact computation of the worst-case risk is **intractable** except for few simple settings.
   - it is difficult to find the inner supremum of the risk over the Wasserstein ball whose cardinality is infinity.
2. Even though we solve WDRO, we do not know any **theoretical properties** of a solution (e.g. risk consistency).

→ **We solve these two problems in this paper!**

Asymptotic equivalence between WDRO and penalty-based methods

Let $R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, h) := \sup_{\mathbb{Q}\in\mathfrak{M}_{\alpha_n,p}(\mathbb{P}_n)} R(\mathbb{Q}, h)$ and (α_n) be a vanishing sequence. In the following, we show that the worst-case risk can be approximated.

Theorem 1 (Informal; Approximation to local worst-case risk)

Let $\mathcal{Z}$ be an open and bounded subset of $\mathbb{R}^d$. For $k \in (0, 1]$, assume that a gradient of loss $\nabla_z h(z)$ is k-Hölder continuous and $\mathbb{E}_{\mathrm{data}}(\|\nabla_z h\|_*)$ is bounded below by some constant. Then for $p \in (1 + k, \infty)$, the following holds.

$$\left| R(\mathbb{P}_n, h) + \alpha_n \|\nabla_z h\|_{\mathbb{P}_n,p^*} - R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_n, h) \right| = O_p(\alpha_n^{1+k}).$$

Gao et al. (2017, Theorem 2) obtained a similar result when $\mathcal{Z} = \mathbb{R}^d$, yet our boundedness assumption on $\mathcal{Z}$ is reasonable in a sense that real computers store data in a finite number of states. Also, Theorem 1 is **sharper**.

Vanishing excess worst-case risk

Based on Theorem 1, for a vanishing sequence (α_n), we propose to minimize the following surrogate objective:

$$R^{\mathrm{prop}}_{\alpha_n,p}(\mathbb{P}_n, h) := R(\mathbb{P}_n, h) + \alpha_n \|\nabla_z h\|_{\mathbb{P}_n,p^*} \, . \tag{2}$$

Let $\hat{h}^{\mathrm{prop}}_{\alpha_n,p} = \mathrm{argmin}_{h\in\mathcal{H}} R^{\mathrm{prop}}_{\alpha_n,p}(\mathbb{P}_n, h)$.

Theorem 2 (Informal; Excess worst-case risk bound)

With the assumptions in Theorem 1, suppose $\mathcal{H}$ is uniformly bounded. Then, for $p \in (1+k, \infty)$, the following holds.

$$R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_{\mathrm{data}}, \hat{h}^{\mathrm{prop}}_{\alpha_n,p}) - \inf_{h\in\mathcal{H}} R^{\mathrm{worst}}_{\alpha_n,p}(\mathbb{P}_{\mathrm{data}}, h) = O_p\left(\frac{\mathfrak{C}(\mathcal{H}) \vee \alpha_n^{1-p}}{\sqrt{n}} \vee \log(n)\alpha_n^{1+k}\right),$$

where $\mathfrak{C}(\mathcal{H})$ is the Dudley's entropy integral.

Compared to Lee and Raginsky (2018), this form has the additional term $\log(n)\alpha_n^{1+k}$, which can be thought as a payoff for the approximation.

WDRO with locally perturbed data

Definition 3 (Locally perturbed data distribution)

For a dataset $\mathcal{Z}_n = \{z_1, \ldots, z_n\}$ and $\beta \geq 0$, we say $\mathbb{P}'_n$ is a β-locally perturbed data distribution if there exists a set $\{z'_1, \ldots, z'_n\}$ such that $\mathbb{P}'_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{z'_i}$ and z'_i can be expressed as

$$z'_i = z_i + e_i,$$

for $\|e_i\| \leq \beta$ and $i \in [n]$.

▷ Examples include denoising autoencoder (Vincent et al., 2010), Mixup (Zhang et al., 2017), and adversarial training (Goodfellow et al., 2014).

Extends the previous results

Theorem 4 (Informal; Parallel to Theorem 1)

Let (β_n) be a vanishing sequence and $\mathbb{P}'_n$ be a β_n-locally perturbed data distribution. With the assumptions in Theorem 1 and for $p \in (1+k, \infty)$, the following holds.

$$\left| R(\mathbb{P}'_n, h) + \alpha_n \|\nabla_z h\|_{\mathbb{P}'_n, p^*} - R^{\mathrm{worst}}_{\alpha_n, p}(\mathbb{P}_n, h) \right| = O_p(\alpha_n^{1+k} \vee \beta_n).$$

- Theorem 4 extends Theorem 1 to the cases when data are locally perturbed. The cost of perturbation is an additional error $O(\beta_n)$, which is negligible when $\beta_n \leq O(\alpha_n^{1+k})$.
- A similar extension for Theorem 2 is provided in the paper.

Numerical Experiments

- We conduct numerical experiments to demonstrate robustness of the proposed method using image classification datasets.
- We compare the following four methods:
  - Empirical risk minimization (ERM)
  - Proposed method (WDRO)
  - Empirical risk minimization with the Mixup (MIXUP)
  - Proposed method with the Mixup (WDRO+MIX)
- We use CIFAR-10 and CIFAR-100 datasets and train models using clean images.

Numerical Experiments: Accuracy comparison

Table: Accuracy comparison of the four methods using the clean and noisy test datasets with various training sample sizes. Average and standard deviation are denoted by 'average±standard deviation'.

Sample size	Clean ERM	Clean WDRO	Clean MIXUP	Clean WDRO+MIX	1% salt and pepper noise ERM	1% salt and pepper noise WDRO	1% salt and pepper noise MIXUP	1% salt and pepper noise WDRO+MIX
CIFAR-10								
2500	77.3 ± 0.8	77.1 ± 0.7	**81.4 ± 0.5**	**80.8 ± 0.7**	69.8 ± 1.8	71.9 ± 0.9	72.7 ± 1.6	**74.8 ± 0.9**
5000	83.3 ± 0.4	83.0 ± 0.3	**86.7 ± 0.2**	**85.6 ± 0.3**	75.2 ± 1.4	77.4 ± 0.5	76.4 ± 1.7	**79.6 ± 0.9**
25000	92.2 ± 0.2	91.4 ± 0.1	**93.3 ± 0.1**	92.4 ± 0.1	83.3 ± 0.8	**85.8 ± 0.5**	82.1 ± 1.7	**86.2 ± 0.3**
50000	94.1 ± 0.1	93.1 ± 0.1	**94.8 ± 0.2**	93.5 ± 0.2	84.1 ± 1.0	**87.4 ± 0.5**	82.5 ± 1.3	**87.3 ± 0.5**
CIFAR-100								
2500	33.8 ± 1.0	34.6 ± 1.7	**38.9 ± 0.6**	**39.4 ± 0.2**	29.2 ± 0.2	30.4 ± 1.2	33.2 ± 1.1	**35.0 ± 0.5**
5000	45.2 ± 0.9	43.7 ± 0.7	**49.9 ± 0.2**	**49.5 ± 0.4**	37.0 ± 0.8	38.1 ± 1.1	39.4 ± 1.3	**42.3 ± 0.7**
25000	67.8 ± 0.2	66.6 ± 0.3	**69.3 ± 0.3**	68.2 ± 0.3	51.0 ± 1.9	**56.5 ± 0.8**	49.6 ± 1.0	**55.8 ± 0.4**
50000	74.4 ± 0.2	73.5 ± 0.3	**75.2 ± 0.2**	73.8 ± 0.3	51.9 ± 1.3	**62.1 ± 0.5**	50.0 ± 3.0	60.6 ± 0.7

▷ In most cases, the proposed methods (WDRO, WDRO+MIX) show significantly better performance when test data are noisy.

Numerical Experiments: Accuracy comparison by noise intensity

Table: The comparison of the accuracy reduction on various salt and pepper noise intensities.

Probability of noisy pixels	ERM	WDRO	MIXUP	WDRO+MIX
CIFAR-10				
1%	10.1 ± 0.9	**5.7 ± 0.4**	12.4 ± 1.2	**6.2 ± 0.4**
2%	21.1 ± 1.9	**13.2 ± 0.5**	24.3 ± 1.4	**12.7 ± 0.8**
4%	39.7 ± 2.9	**32.9 ± 2.5**	43.5 ± 1.8	**30.9 ± 2.0**
CIFAR-100				
1%	22.5 ± 1.3	**11.4 ± 0.4**	25.2 ± 2.5	13.2 ± 0.7
2%	42.8 ± 2.3	**26.5 ± 1.0**	45.9 ± 3.4	29.7 ± 0.7
4%	61.7 ± 1.4	**50.0 ± 0.9**	63.9 ± 2.0	53.5 ± 0.9

Numerical Experiments: Gradient norm

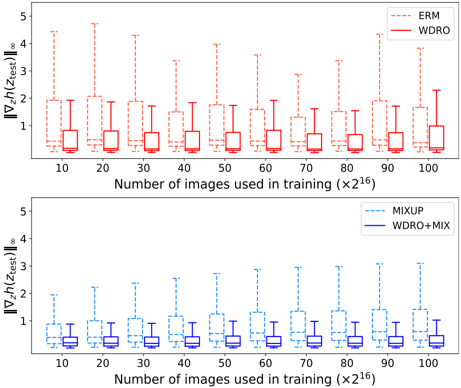


Figure: The box plots of the ℓ_∞-norm of the gradients when the number of images used in training increases from 10×2^{16} to 100×2^{16}.

Conclusion

- We develop a **principled and tractable** statistical inference method for WDRO.
- We formally present a locally perturbed data distribution and develop WDRO inference **when data are locally perturbed**.
- For more details, ArXiv & Github links:
  `https://arxiv.org/abs/2006.03333`
  `https://github.com/ykwon0407/wdro_local_perturbation`

References I

Gao, R., Chen, X., and Kleywegt, A. J. (2017). Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Lee, J. and Raginsky, M. (2018). Minimax statistical learning with Wasserstein distances. In *Advances in Neural Information Processing Systems*, pages 2687–2696.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.