Noise Tolerance of Distributionally Robust Learning

Ramzi Dakhmouche [1] **Ivan Lunati** [2] **Hossein Gorji** [2]

Abstract

Given the importance of building robust machine learning models, considerable efforts have recently been put into developing training strategies that achieve robustness to outliers and adversarial attacks. Yet, a major aspect that remains an open problem is systematic robustness to global forms of noise such as those that come from measurements and quantization. Hence, we propose in this work an approach to train regression models from data with additive forms of noise, leveraging the Wasserstein distance as a loss function. Importantly, our approach is agnostic to the model structure, unlike the increasingly popular Wasserstein Distributionally Robust Learning paradigm (`WDRL`) which, we show, does not achieve improved robustness when the regression function is not convex or Lipschitz. We provide a theoretical analysis of the scaling of the regression functions in terms of the variance of the noise, for both formulations and show consistency of the proposed loss function. Lastly, we conclude with numerical experiments on physical PDE Benchmarks and electric grid data, demonstrating competitive performance.

1. Introduction

In real-world applications, collected data is often tainted with different forms of noise. Whether it is sensor noise in engineering systems or measurement uncertainty in biological experiments, such noise usually demands costly and time-consuming pre-processing steps, before meaningful results can be extracted using predictive machine learning algorithms. In order to streamline that process, different robust learning approaches have been proposed with a focus on robustness to outliers and adversarial attacks (Mohajerin Esfahani & Kuhn, 2018; Steinhardt et al., 2018; Bai et al., 2023; Levine & Feizi, 2020). Most of such strategies rely on augmenting the data with adversarial examples (Goodfellow et al., 2014; Madry et al., 2018) or designing suitable loss regularization techniques (Dong et al., 2020). However, for more global forms of noise, which are commonly encountered in practice, these approaches face both statistical and practical limitations. In the case of data augmentation, the limitations are inherent to its design, while adversarial regularization often targets bounded perturbations, thereby overlooking standard noise models that arise in real-world settings. In contrast, the increasingly popular paradigm of Wasserstein Distributionally Robust Learning (`WDRL`) (Mohajerin Esfahani & Kuhn, 2018; Shafieezadeh-Abadeh et al., 2019; Gao et al., 2024) represents a more general framework that allows for arbitrary perturbations, and is more theoretically appealing while leading to competitive performance. Yet, there seems to be a gap in the literature when it comes to robustness properties of `WDRL` with respect to global forms of noise, as pointed out by (Hu et al., 2020) for instance. In this work, we address this question in a regression setting from multiple perspectives:

1. We study the global robustness properties of the popular `WDRL` formulation, through a numerical analysis of its scaling in terms of the variance of the noise.
2. Notably, we show that `WDRL` may fail to improve the performance when the regression functions are neither Lipschitz nor convex.
3. To address this limitation, we propose a simple yet powerful robust learning approach that is agnostic to the structure of regression functions, enabling more expressive models. We further provide a theoretical analysis of its dependence on the variance of the noise.
4. We numerically demonstrate the performance of our proposed approach through various physical problem benchmarks and electric grid usage time series data.

2. Problem Setting

Consider the regression task of predicting response variables $y \in \mathcal{Y}$ from input features $x \in \mathcal{X}$. Given a class of regression functions $\{f_\theta,\, \theta \in \mathbb{R}^d\}$ and data samples $\{X_i, Y_i\}_{i \leq n}$ from an underlying distribution $\mu_{(X,Y)} \in \mathcal{P}_2(\mathcal{X} \times \mathcal{Y})$ with

[1]Institute of Mathematics, EPFL, Switzerland [2]Computational Engineering Lab, Empa, Switzerland. Correspondence to: Ramzi Dakhmouche <ramzi.dakhmouche@epfl.ch>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

$f_\theta : \mathcal{X} \to \mathcal{Y}$, the standard goal is to find a model $\theta \in \mathbb{R}^d$ that minimizes the empirical risk

$$\hat{\theta}_{\texttt{MSE}} \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \|Y_i - f_\theta(X_i)\|_2^2.$$

In this work, we focus on the setup where the response samples $(Y_i + \sigma\varepsilon_i)_{i \leq n}$ are tainted with independent and identically distributed noise with variance σ^2, with the objective of training deep learning models that are the least sensitive to the noise level σ.

3. Drawbacks of `WDRL`

In order to compute the objective function of `WDRL`, certain structural assumptions on the regression functions $f_\theta, \theta \in \mathbb{R}^d$ as well as the loss function ℓ, must be imposed. This is necessary since the original formulation involves solving an infinite dimensional optimization problem, which is generally intractable. For that matter, two main settings have been proposed:

(a) Assuming that the function $\ell_\theta : (x, y) \mapsto \ell(y, f_\theta(x))$ is a finite maximum of concave functions;

(b) Assuming that the function $\ell_\theta : (x, y) \mapsto \ell(y, f_\theta(x))$ is Lipschitz continuous.

In either cases, the objective function can be rewritten (Mohajerin Esfahani & Kuhn, 2018; Shafieezadeh Abadeh et al., 2015) under a tractable form[1] as follows

$$\begin{aligned} d_2((Y_i)_{i\leq n}, (f_\theta(X_i))_{i\leq n}) &:= \sup_{\substack{(X,Y)\sim\mu \\ \mathcal{W}_2(\mu,\hat{\mu})\leq\delta}} \mathbb{E}_\mu[\ell(Y, f_\theta(X))] \\ &= \inf_{\lambda\geq 0} \Bigg[\lambda\delta + \frac{1}{n}\sum_{i=1}^{n} \sup_{(\xi_1,\xi_2)\in\mathcal{X}\times\mathcal{Y}} \{\ell(\xi_1 - f_\theta(\xi_2)) \\ &\qquad -\lambda\|Y_i - \xi_1\|_2^2 - \lambda\|X_i - \xi_2\|_2^2\}\Bigg], \end{aligned}$$

where the optimal solutions $\lambda^\star(\theta)$ and $(\xi_1^\star(\theta), \xi_2^\star(\theta))$ are reached for all $\theta \in \mathbb{R}^d$. Yet, to satisfy (a) or (b) in a regression setting where the data distributions have unbounded domains, one typically needs to set $\ell(y, x) = |y - x|$ and to enforce structural properties of convexity or Lipschitzness on the neural network models, therefore, reducing their expressive power. A natural question that emerges is whether using the tractable expression of d_2 as a loss function, regardless of whether the equality holds, can improve the robustness of the neural network models. We provide a negative answer to this question by exploring the behavior of d_2 in training a convolutional neural operator (CNO) (Raonic et al., 2024) to solve the two-dimensional Navier-Stockes equation. In particular, we estimate the operator that maps the initial condition $(T = 0)$, represented as an image, to the final state $(T = 1)$. To this end, we train the `WDRL` regression model employing a stochastic descent- ascent algorithm, exploring the model behavior as the noise level increases. We use the hyperparameters optimized by the authors who proposed the CNO architecture (Raonic et al., 2023). We obtain the results shown in Figs. 1 and 2, for both Gaussian and heavy-tail noise distributions, respectively. For the latter case, we use the standard Cauchy distribution, where σ represents the scale parameter, as a Cauchy random variable does not have finite variance due to the heavy tails. We examine model performance via the mean absolute relative error (MAE). Our results indicate that under heavy-tail noise, `WDRL` training performs significantly worse than the standard `MSE` training. In the Gaussian noise setting, both lead to comparable results, without noticeable improvement from `WDRL`. This is in contrast with the novel regression approach introduced in the follow-up section, whose performance on this setting is demonstrated in Section 5.

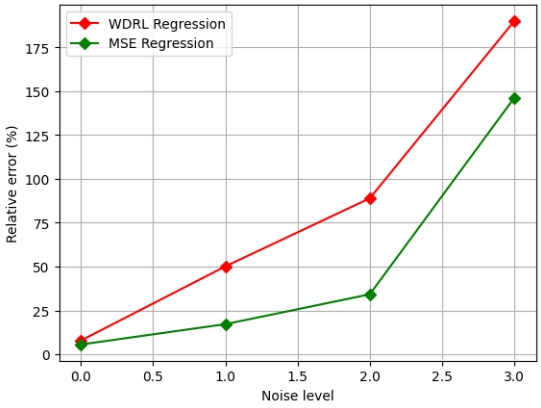

Figure 1. Test error evolution for Navier-Stockes operator learning with 30% corrupted training data with Cauchy noise

Remark. Note that this limitation of `WDRL` has not been raised so far, to the best of our knowledge, mainly because previous works focused on image classification where the data domains are bounded, honouring the Lipschitz property.

4. Wasserstein Batch Matching

The key idea behind our approach is to relax the strict matching between features X_i and their given responses $Y_i + \sigma\varepsilon_i$ for $i \in I_p$, where I_p denotes the index set of a training batch.

[1] (Blanchet & Murthy, 2019) propose a more general condition for the equality to hold, but leave the question of existence of optimizers, which is essential here, open.

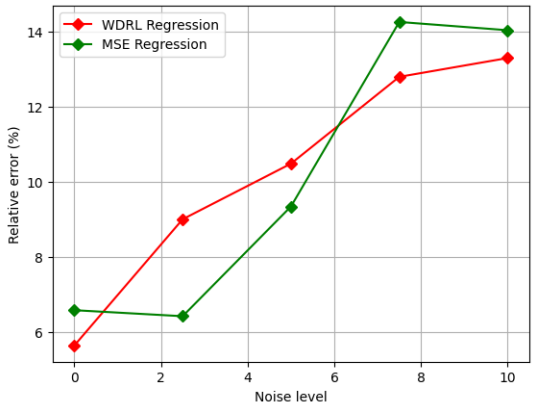

Figure 2. Test error evolution for Navier-Stockes operator learning with 30% corrupted training data with Gaussian noise

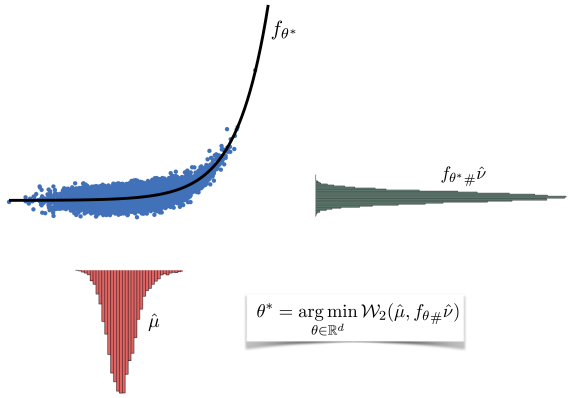

Figure 3. Wasserstein Batch Matching illustration. The regression through cloud of points in a batch (depicted by blue) is tackled by finding optimal map, depicted by black line, between distributions of $(X_i)_i$ and $(Y_i + \sigma\epsilon)_i$, shown by red and green histograms.

The motivation for this relaxation is that, in the presence of noise, the observed response $Y_i + \sigma\varepsilon_i$ already deviates from the true response Y_i. Consequently, if the batch elements are sufficiently close, allowing features to match responses without a fixed correspondence can lead to a more robust estimate while reducing the sensitivity of the loss function to noise.

4.1. Formulation & Consistency

The formal way to implement this idea is to compute the Wasserstein distance between the empirical distributions of the predictions $(f_\theta(X_i))_{i\in I_p}$ and the responses $(Y_i)_{i\in I_p}$, leading to Wasserstein Batch Matching (`WBM`) regression

$$\hat{\theta}_{\text{WBM}} \in \arg\min_{\theta\in\mathbb{R}^d} \sum_{p\geq 1} \mathcal{W}_2(m[(Y_i)_{i\in I_p}], m[(f_\theta(X_i))_{i\in I_p}]) \ ,$$

instead of the Mean Squared Error (`MSE`) regression. We illustrate the `WBM` idea in figure 3. For our setting of empirical distributions, note that the Wasserstein distance reduces to

$$\mathcal{W}_2(m[(Y_i)_{i\in I_p}], m[(f_\theta(X_i))_{i\in I_p}]) = \min_{P\in C} \left\langle P,\, M_{((Y_i)_{i\in I_p},(f_\theta(X_i))_{i\in I_p})} \right\rangle \ ,$$

where $M_{((Y_i)_{i\in I_p},(f_\theta(X_i))_{i\in I_p})} = \left(\|Y_i - f_\theta(X_j)\|_2^2 \right)_{i,j\in I_p}$ is the matrix of the pairwise norms between the predictions and the target values, and C the set of coupling matrices of dimension $\#I_p$. As a sanity check, we show in proposition (1) below that, asymptotically such a matching scheme recovers any continuously differentiable bandlimited function, among its co-monotonic functions from its samples, in the noise-free regime.

Proposition 4.1. ***(Consistency)*** *Let $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ be a continuously differentiable and integrable function with compactly supported Fourier transform and let $(f(x_{\phi(i)}))_{i\leq n}$ be its values sampled at ordered points $(x_i)_{i\leq n}$, where ϕ is an unknown permutation preserving the batch partition. Then, given a fixed batch size and an arbitrary amount of samples, f is completely characterized by minimizing*

$$\min_{g\in\mathcal{G}} \sum_{p\geq 1} \mathcal{W}_2(m[(f(x_i))_{i\in I_p}], m[(g(x_j))_{j\in I_p}]) \ ,$$

where $\{I_p, p \geq 1\}$ is the finite collection of batch index sets and $\mathcal{G}$ the set of continuously differentiable and integrable bandlimited functions that are co-monotonic with f, where we employ $m[(X_i)_{i\leq n}] = 1/n \sum_{i=1}^n \delta_{X_i}$, as the shorthand for empirical measure.

Two remarks are in order concerning the algorithmic aspects of the introduced approach.

Remark. (**Complexity**) From a computational complexity perspective, training with `WBM` involves solving a linear program at each training step, which costs $O(s)$, where $s = dim(\mathcal{Y})$ is the dimension of the response space $\mathcal{Y}$. However, this is independent of the strucutre of the regression functions. On the other hand, `WDRL` involves solving a minimax problem which is in $O(s^3)$ when the function ℓ_θ is convex-concave. However, in the absence of this structure, the problem can become arbitrarily hard.

Remark. (**Differentiability**) The proposed `WBM` loss is differentiable with respect to the regression parameters $\theta \in \mathbb{R}^d$ by the envelope theorem (Bonnans & Shapiro, 2013), which makes it well-suited for training deep learning models.

5. Numerical Results

We demonstrate the performance of WBM regression on two important practical problems: operator learning for PDEs and electric grid usage forecasting. Evaluation on test data is carried out using the mean absolute error (MAE) in all displayed figures, where we compare models trained with the MSE to those trained with WBM. We explore the robustness properties of WBM both to standard Gaussian and heavy-tail Cauchy noise. Heavy-tail noise is present in many real-world applications such as vibration sensors for intelligent monitoring, power consumption sensors, and LIDAR systems. It comes from transient events, sudden extreme changes such as short circuits, or atmospheric noise which exhibits heavy-tails. Additionally, we explore robustness to distribution shift properties, by training on noise-free data and testing on noisy data. Such a use-case is encountered in practice, when a model is developped based on cleaned data before being deployed on real data. We report the results for second practical problem in the appendix.

Learning of PDEs

PDEs model a wide range of physical and engineering problems and feature a rich set of dynamical processes that illustrate the performance of machine learning models on a wide range of practical regression problems. For that matter, we demonstrate the performance of WBM on a two of extensively used PDEs: the wave equation and the Navier-Stokes equation. More precisely, we consider the corresponding recently proposed benchmarks in (Raonic et al., 2024), where the task consists of learning operators mapping initial conditions ($T = 0$), represented as images, to the final state reached by the system, e.g., corresponding to ($T = 1$). The underlying images represent two-variable functions sampled at a given resolution. We train convolutional neural operators (Raonic et al., 2023), which have been proposed as featuring robustness properties notably to change in resolution. We compare models trained with the MSE loss to those trained with the WBM loss. We set the hyperparameters optimized in (Raonic et al., 2023), and keep the same for WBM training, except the batch size for which we explore different values. The convolutional neural operator architecture is based on mapping the sampled images back to function space using the Whittaker-Shannon interpolation formula (Raonic et al., 2023). We display the results in Figs. 6, 7. We note that, WBM regression consistently outperforms MSE regression. In particular, while both MSE and WDRL regressions indicate significant errors in Navier-Stokes operator learning subject to the Cauchy noise, as shown in Fig. 1, the introduced WBM regression demonstrates notable robustness.

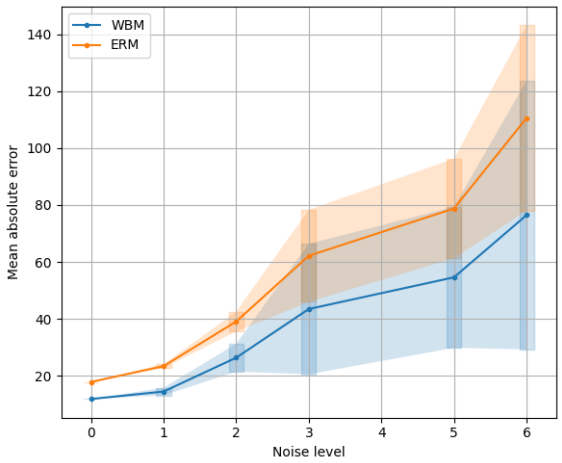

Figure 4. Test error evolution for Navier-Stockes operator learning - 30% corrupted test data with Cauchy noise

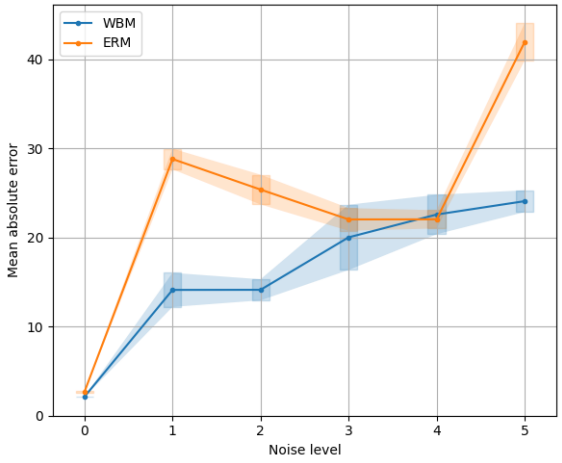

Figure 5. Test error evolution for wave equation operator learning - 30% corrupted training data with Gaussian noise

6. Discussion

In this paper, we considered the open problem of robustness to global forms of noise, for which we proposed a learning approach WBM, overcoming the drawbacks of WDRL. We investigated the scaling of the introduced regression along with other approaches with respect to noise levels, offering a theoretical justification for the gains achieved by WBM. Furthermore, we demonstrated the practical performance of WBM via several numerical experiments involving learning of physical PDE operators and electrical time series forecasting. We believe this work paves the way for robust learning methods that streamline the costly data pre-processing step, while advancing the development of reliable machine learning models.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Bai, X., He, G., Jiang, Y., and Obloj, J. Wasserstein distributional robustness of neural networks. *Advances in Neural Information Processing Systems*, 36, 2023.

Bartl, D., Drapeau, S., Obłój, J., and Wiesel, J. Sensitivity analysis of wasserstein distributionally robust optimization problems. *Proceedings of the Royal Society A*, 477 (2256):20210176, 2021.

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

Bonnans, J. F. and Shapiro, A. *Perturbation analysis of optimization problems*. Springer Science & Business Media, 2013.

Bui, A. T., Le, T., Tran, Q. H., Zhao, H., and Phung, D. A unified wasserstein distributional robustness framework for adversarial training. In *International Conference on Learning Representations*, 2022.

Chen, R. and Paschalidis, I. C. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19 (13):1–48, 2018.

Chen, S.-A., Li, C.-L., Arik, S. O., Yoder, N. C., and Pfister, T. Tsmixer: An all-mlp architecture for time series forecast-ing. *Transactions on Machine Learning Research*.

Dathathri, S., Dvijotham, K., Kurakin, A., Raghunathan, A., Uesato, J., Bunel, R. R., Shankar, S., Steinhardt, J., Goodfellow, I., Liang, P. S., et al. Enabling certification of verification-agnostic networks via memory-efficient semidefinite programming. *Advances in Neural Information Processing Systems*, 33:5318–5331, 2020.

Dong, Y., Deng, Z., Pang, T., Zhu, J., and Su, H. Adversarial distributional training for robust deep learning. *Advances in Neural Information Processing Systems*, 33: 8270–8283, 2020.

Elad, M., Kawar, B., and Vaksman, G. Image denoising: The deep learning revolution and beyond—a survey paper. *SIAM Journal on Imaging Sciences*, 16(3):1594–1654, 2023.

Föllmer, H. and Weber, S. The axiomatic approach to risk measures for capital determination. *Annual Review of Financial Economics*, 7(1):301–337, 2015.

Gao, R. and Kleywegt, A. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

Gao, R., Chen, X., and Kleywegt, A. J. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 72(3):1177–1191, 2024.

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

Hu, E. J., Swaminathan, A., Salman, H., and Yang, G. Improved image wasserstein attacks and defenses. *arXiv preprint arXiv:2004.12478*, 2020.

Ilbert, R., Odonnat, A., Feofanov, V., Virmaux, A., Paolo, G., Palpanas, T., and Redko, I. Samformer: Unlocking the potential of transformers in time series forecasting with sharpness-aware minimization and channel-wise attention. In *Forty-first International Conference on Machine Learning*, 2024.

Ito, K. and Xiong, K. Gaussian filters for nonlinear filtering problems. *IEEE transactions on automatic control*, 45 (5):910–927, 2000.

Jain, V. and Seung, S. Natural image denoising with convolutional networks. *Advances in neural information processing systems*, 21, 2008.

Krull, A., Buchholz, T.-O., and Jug, F. Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2129–2137, 2019.

Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., and Aila, T. Noise2noise: Learning image restoration without clean data. In *International Conference on Machine Learning*, pp. 2965–2974. PMLR, 2018.

Levine, A. and Feizi, S. Wasserstein smoothing: Certified robustness against wasserstein adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pp. 3938–3947. PMLR, 2020.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

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

Raghunathan, A., Steinhardt, J., and Liang, P. S. Semidefinite relaxations for certifying robustness to adversarial examples. *Advances in neural information processing systems*, 31, 2018.

Raonic, B., Molinaro, R., De Ryck, T., Rohner, T., Bartolucci, F., Alaifari, R., Mishra, S., and de Bézenac, E. Convolutional neural operators for robust and accurate learning of pdes. *Advances in Neural Information Processing Systems*, 36, 2023.

Raonic, B., Molinaro, R., De Ryck, T., Rohner, T., Bartolucci, F., Alaifari, R., Mishra, S., and de Bézenac, E. Convolutional neural operators for robust and accurate learning of pdes. *Advances in Neural Information Processing Systems*, 36, 2024.

Sardy, S., Tseng, P., and Bruce, A. Robust wavelet denoising. *IEEE transactions on signal processing*, 49(6): 1146–1152, 2001.

Shafieezadeh Abadeh, S., Mohajerin Esfahani, P. M., and Kuhn, D. Distributionally robust logistic regression. *Advances in neural information processing systems*, 28, 2015.

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

Staib, M. and Jegelka, S. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*, volume 3, pp. 4, 2017.

Steinhardt, J., Charikar, M., and Valiant, G. Resilience: A criterion for learning in the presence of arbitrary outliers. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, volume 94, pp. 45. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018.

Tjeng, V., Xiao, K. Y., and Tedrake, R. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*, 2019.

Wu, D., Xia, S.-T., and Wang, Y. Adversarial weight perturbation helps robust generalization. *Advances in neural information processing systems*, 33:2958–2969, 2020.

Xing, W. and Egiazarian, K. End-to-end learning for joint image demosaicing, denoising and super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3507–3516, 2021.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Conference*, volume 35, pp. 11106–11115. AAAI Press, 2021.

A. Proof of Proposition 4.1: Consistency

We assume without loss of generality that $\mathcal{X} = \mathcal{Y} = \mathbb{R}$. Let $f, g \in \mathcal{G}$. Given that g is continuously differentiable, it has bounded variations on every compact, that is for all $a, b \in \mathbb{R}$ such that $a < b$, we have

$$\sup_{pr \in Pr} \sum_{i=1}^{n_{pr}} |g(x_{i+1}) - g(x_i)| < +\infty \,,$$

where the supremum is taken over the set

$$\left\{ pr = \{x_0, \dots, x_{n_{pr}}\} \mid pr \text{ is a partition of } [a, b] \text{ satisfying } x_i \leq x_{i+1} \text{ for } 0 \leq i \leq n_{pr} - 1 \right\}$$

This implies that there exists a partition of the feature space into intervals of lengths $(\delta_n)_{n\in\mathbb{N}}$ such that g is monotonous on every interval. The same holds for f. Hence, we consider the partition formed by $I_i \cap J_j$ where $(I_i)_{i\in\mathbb{N}}$ and $(J_j)_{j\in\mathbb{N}}$ are the chosen partitions for f and g respectively. We denote by $(\delta_n)_{n\in\mathbb{N}}$ the new lengths. Since, f is bandlimited, let the support of its Fourier transform be included in $[-B, B]$ with $B > 0$. We can choose $(\delta_n)_n$ such that $\delta_n \leq \frac{1}{B}$ for all $n \in \mathbb{N}$. Furthermore, we can sample each interval a number of times equal to the prefixed batch size. Since, f satisfies

$$\begin{aligned} \min_{g\in\mathcal{G}} \sum_{p\geq 1} \mathcal{W}_2 \left(m[(f(x_i))_{i\in I_p}], m[(g(x_j))_{j\in I_p}] \right) &= \sum_{p\geq 1} \mathcal{W}_2(m[(f(x_i))_{i\in I_p}], m[(f(x_j))_{j\in I_p}]) \\ &= 0, \end{aligned}$$

we know that a minimizer $g \in \mathcal{G}$ of

$$g \mapsto \sum_{p\geq 1} \mathcal{W}_2 \left(m[(f(x_i))_{i\in I_p}], m[(g(x_j))_{j\in I_p}] \right)$$

must satisfy

$$\forall p \geq 1, \quad \mathcal{W}_2(m[(f(x_i))_{i\in I_p}], m[(g(x_j))_{j\in I_p}]) = 0 \,.$$

Furthermore, since f and g are co-monotonic, the Wasserstein matching recovers the true matching. Last, by the Shannon sampling theorem we conclude g is equal f.

B. Additional Numerical Results

Electric Grid Usage Forecasting
Predicting electric load is an important and timely problem, especially given the increasing share of renewable energy sources in all electrical grid networks. We employ the recently proposed state-of-the-art model `TSMixer` (Chen et al.), to forecast electric transformer usage from the popular ETDataset (Zhou et al., 2021; Ilbert et al., 2024). `TSMixer` is based on mixing operations via stacking multi-layer perceptions. We train the model with input sequence length of 336 and prediction sequence length of 96. We utilize the hyperparameters proposed by the authors, except the number of training epochs which we reduce to a single swap over the data. This is justified by the fact that we compare the model against itself trained with different loss functions. Hence, the comparison point can be chosen in a flexible way. We display the results in Figs. 6 and 7. In the former, the `WBM` regression outperforms `MSE`, whereas in the latter, their performance is similar, with `WBM` showing slight underperformance. This can be attributed to the minimal errors observed in both approaches, suggesting that the underlying model exhibits little sensitivity to noise in this setting.

Reproducibility. We provide a version of the code used for the numerical experiments in the following link: code. It is based on modifications of the publicly available code from (Raonic et al., 2023) and (Ilbert et al., 2024).

C. Related Works

Denoising and Filtering. Extensive research has been conducted on denoising and filtering techniques, ranging from Kalman filtering (Ito & Xiong, 2000) and wavelet denoising (Sardy et al., 2001) to deep learning based methods (Jain & Seung, 2008; Xing & Egiazarian, 2021; Krull et al., 2019; Lehtinen et al., 2018). For a comprehensive overview in the context of image data modalities, see (Elad et al., 2023). However, most of these approaches require low noise data,

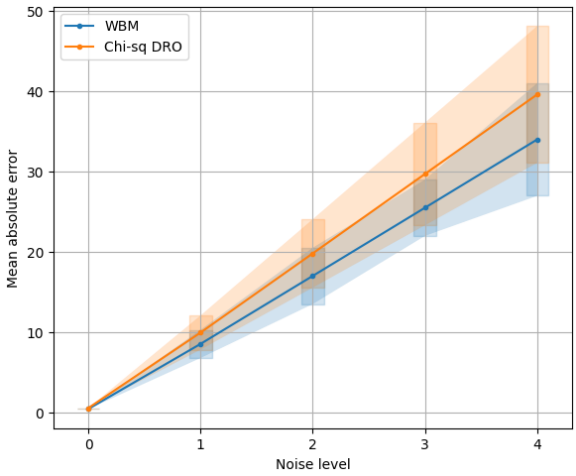

Figure 6. Test error evolution for electric time series forecasting - 30% corrupted test data with Cauchy noise

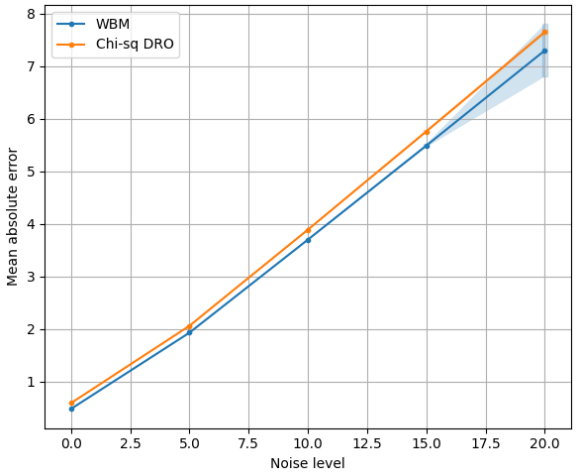

Figure 7. Test error evolution for electric time series forecasting - 30% corrupted test data with Gaussian noise

focus on Gaussian noise distributions or require an explicit noise model. Additionally, they introduce costly separate data pre-processing steps that must be performed prior to the modeling. In contrast, we propose an approach that directly trains competitive models from noisy data, eliminating the need for extensive pre-processing.

Adversarial Defense. Early works introduced techniques to augment the training data with adversarial examples (Goodfellow et al., 2014; Madry et al., 2018), leveraging the expressive power of neural networks to improve robustness. Building on this, several regularization techniques such as entropic regularization (Dong et al., 2020) and adversarial weight perturbation (Wu et al., 2020) have been proposed, further enhancing their performance. In parallel, *certified robustness* approaches have focused on quantifying the proportion of samples that remain robust to arbitrary perturbations within a given bound (Tjeng et al., 2019; Raghunathan et al., 2018; Dathathri et al., 2020). However, these techniques often lead to overly conservative models, which can degrade performance in the presence of global noise perturbations (Bai et al., 2023).

Distributionally Robust Optimization. It is concerned with minimizing the worst-case loss over a given set of distributions (Mohajerin Esfahani & Kuhn, 2018; Föllmer & Weber, 2015; Blanchet & Murthy, 2019), which is formally expressed as the minimax problem

$$\inf_{\theta\in\Theta} \sup_{Q\in\mathcal{P}} \mathbb{E}_Q\left[\ell_\theta(Z)\right]$$

where the supremum is taken over a suitably chosen class of distributions $Q \in \mathcal{P}$. Recent focus (Shafieezadeh Abadeh et al., 2015; Staib & Jegelka, 2017; Shafieezadeh-Abadeh et al., 2019; Chen & Paschalidis, 2018; Bartl et al., 2021; Gao et al., 2024) has been given to the formulation with Wasserstein ambiguity set $\mathcal{P} = B_\delta(P)$, which is the ball centered at the empirical distribution P with radius δ under the Wasserstein distance, leading to `WDRL`. See (Gao & Kleywegt, 2023) for a discussion on theoretical advantages of this choice. `WDRL` has demonstrated remarkable performance in out-of-sample linear regression (Mohajerin Esfahani & Kuhn, 2018) and classification (Shafieezadeh-Abadeh et al., 2019) tasks, as well as in defending against adversarial attacks (Bai et al., 2023; Bui et al., 2022) on neural networks. In contrast, we consider robustness to unbounded global forms of noise, which to the best of our knowledge, has not been investigated so far in the context of `WDRL`.