Minimax Theory and its Applications
Volume 12 (2027), No. 1, 59 to 63

Distributionally Robust Minimax Learning Through Wasserstein Ambiguity Sets and Variational Duality

Dr. Priya Vij[1], Ashu Nayak[2]

[1]Assistant Professor, Kalinga University, Naya Raipur, Chhattisgarh, India.
ku.priyavij@kalingauniversity.ac.in, https://orcid.org/0009-0005-4629-3413
[2]Assistant Professor, Kalinga University, Naya Raipur, Chhattisgarh, India.
ku.ashunayak@kalingauniversity.ac.in, https://orcid.org/0009-0002-8371-7324

Abstract:
A learner trained on a finite sample is, in some sense, betting that tomorrow will look much like today. This paper asks what happens when that bet is hedged: the learner plays a minimax game against an adversary who is free to move the empirical distribution anywhere in a Wasserstein ball of radius ε. We write the adversarial problem in its variational (Lagrangian) form and show how the dual degenerates, for Lipschitz losses, to the empirical risk plus a constant times ε; the constant for logistic regression is the dual norm of the weight vector, so the robust solution amounts to a standard norm-regularised learner, with the penalty fixed rather than tuned by cross validation. A primal variant confirms the dual value to within 0.001 on a sample of 200. A 30-trial simulation with 50 training points shows that the robust model with ε=0.1 reduces test log-loss from 0.839 to 0.637 and the loss under a 0.5-norm perturbation from 1.456 to 0.882, without affecting accuracy to within 0.2 points. The same setup illustrates the usual problem with robustness: at ε=0.2 accuracy falls by 2 to 4.5 points.

Keywords*:* minimax learning; distributionally robust optimization; Wasserstein distance; Lagrangian duality; logistic regression; regularization
MSC 2020: 90C47, 90C15, 62H30

1. Introduction

Start from what you know. We have N labelled examples, a parametric model and a loss. Empirical risk minimisation is assuming the empirical distribution P̂N is the real truth. When N is small, or when deployment data drifts away from the empirical distribution that assumption is the weak link in the chain. The minimax perspective replaces it with a weaker one: the real world is close to P̂N, and we should plan for the worst-case distribution in that small neighbourhood. Shapiro (2017) frames this idea within stochastic programming, and Rahimian and Mehrotra (2022) survey the countless choices for how the neighbourhood is defined. And you are asking how do we decide "close". If moment-based ambiguity sets such as those considered by Delage and Ye (2010) fix the first two moments and neglect the rest, ϕ-divergence balls will do only reweighting of the observed atoms, and hence will never assign probability to a point that is not in the sample. The Wasserstein metric is very different. It allows for moving probability along the geometry of the feature space, at a cost proportional to the distance traveled (Mohajerin Esfahani and Kuhn, 2018), and, under very mild conditions, it admits finite-sample guarantees and tractable convex reformulations (Gao and Kleywegt, 2023; Blanchet and Murthy, 2019).
We have a modest, pragmatic goal. We reintroduce the minimax learning problem, give a quick derivation of its variational dual, and read off what the dual says about familiar models. Three conclusions follow.

1. **A variational dual for the learner's game.** The supremum of the adversary over an infinite dimensional ball reduces to a one-dimensional infimum over a scalar multiplier λ, to which are coupled N independent inner maximisations (Theorem 1).
2. **Regularisation is an inevitable consequence, not a choice.** For losses with a finite growth rate, the dual reduces to empirical risk plus ε times that rate (for the logistic loss, it's the dual norm of the weights, see Corollary 1). The regularisation parameter is the size of the ambiguity set.
3. **A quick number crunch. We verify the dual value with an explicit primal perturbation.** We measure the impact of ε on clean and perturbed test loss-demonstrating where it begins to cause harm. Section 2 introduces the game. Section 3 presents the duality and minimax interchange. Section 4 introduces the first-order method. Section 5 shows our experiments, and the last section 6 discusses the limitations.

2. The learner–adversary game

Let $\Xi = \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} \subseteq \mathbb{R}^d$ a feature space and $\mathcal{Y}$ a finite label set. Samples $\xi_i = (x_i, y_i)$, $i = 1, \dots, N$, define the empirical measure $\hat{P}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_i}$. A model is indexed by $\theta \in \Theta \subseteq \mathbb{R}^m$ and incurs loss $\ell(\theta; \xi)$.
Transport cost. We use a cost that charges for moving features but forbids relabelling:

$$c\big((x,y),(x',y')\big) = \| x - x' \| + \infty \cdot \mathbf{1}\{y \neq y'\}.$$

The norm $\|\cdot\|$ is arbitrary and $\|\cdot\|_*$ denotes its dual. The type-1 Wasserstein distance (Villani, 2009) between P and Q is the cheapest expected cost over couplings π with marginals P and Q:

$$W(P,Q) = \inf_{\pi \in \Pi(P,Q)} \int_{\Xi \times \Xi} c\,(\xi,\xi')\,\pi(d\xi, d\xi').$$

Ambiguity set and game. For a radius $\varepsilon \geq 0$ let $\mathbb{B}_\varepsilon = \{Q : W\big(Q, \hat{P}_N\big) \leq \varepsilon\}$. The learner and the adversary play

$$J_\varepsilon = \min_{\theta \in \Theta} \sup_{Q \in \mathbb{B}_\varepsilon} \mathbb{E}_Q[\ell(\theta;\xi)]. \qquad (1)$$

The learner moves first and the adversary responds. When $\varepsilon = 0$ the game reduces to empirical risk minimisation. As ε grows the adversary gains room. In the limit only the loss's behaviour at infinity matters.

Standing assumptions. (A1) Θ is convex and compact. (A2) $\theta \mapsto \ell(\theta;\xi)$ is convex for each ξ. (A3) $x \mapsto \ell(\theta;x,y)$ is upper semicontinuous, and its growth rate

$$\kappa(\theta) = \sup_{y \in \mathcal{Y}} \limsup_{\|x'-x\| \to \infty} \frac{\ell(\theta;x',y) - \ell(\theta;x,y)}{\| x' - x \|}$$

is finite.

The value of κ does not depend on the base point x when ℓ is Lipschitz in x. Assumption (A3) is the condition under which the worst-case risk stays finite (Gao & Kleywegt, 2023).

3. Variational duality

The inner supremum in (1) is taken over a family of probability measures. The dual is taken over a single scalar value. We explain this and its implications below.

3.1. From measures to a multiplier

Fix θ. Any $Q \in \mathbb{B}_\varepsilon$ can be written as the second marginal of a coupling π whose first marginal is $\hat{P}_N$. Equivalently, Q is a mixture $\frac{1}{N}\sum_i Q_i$, where Q_i is the conditional law of the destination of atom ξ_i. Attaching a multiplier $\lambda \geq 0$ to the budget constraint gives the Lagrangian

$$\mathcal{L}(Q,\lambda) = \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N} \int_{\Xi} [\ell(\theta;\xi') - \lambda\, c(\xi_i,\xi')]\; Q_i(d\xi').$$

For every feasible Q and every $\lambda \geq 0$, the bracket integrates to at least $\mathbb{E}_Q[\ell] - \lambda\varepsilon$. Maximising each Q_i separately, the best choice is a point mass at a maximiser of the bracket. This gives weak duality at once. Strong duality is the substantive part, and it is the content of the next theorem.

Theorem 1 (variational dual). Under (A3), for every $\theta \in \Theta$ and $\varepsilon > 0$,

$$\sup_{Q \in \mathbb{B}_\varepsilon} \mathbb{E}_Q[\ell(\theta;\xi)] = \min_{\lambda \geq 0} \left\{ \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N} \sup_{x' \in \mathcal{X}} [\ell(\theta;x',y_i) - \lambda \| x' - x_i \|] \right\}. \qquad (2)$$

Proof sketch. The inequality "$\leq$" is the weak-duality argument above. The labels cannot move because the cost is infinite for $y \neq y'$. For "$\geq$" we invoke the general strong-duality theorem for Wasserstein balls around a nominal measure. Gao and Kleywegt (2023, Thm. 1) require only upper semicontinuity of the loss and a finite growth rate. Blanchet and Murthy (2019) give an equivalent statement for lower semicontinuous costs. Our cost is lower semicontinuous, and its infinite part only restricts the support of each Q_i to the fibre $\{y = y_i\}$. The minimum over λ is attained because the objective is convex and coercive in λ on $[\kappa(\theta), \infty)$. ▫

The right-hand side of (2) is the "variational" form. Each atom solves its own small maximisation, which is a local adversarial perturbation priced at λ per unit distance. The scalar λ sets that price globally so that, on average, the adversary spends exactly ε.

3.2. Collapse for Lipschitz losses

Corollary 1. Let $\mathcal{X} = \mathbb{R}^d$. Suppose $x \mapsto \ell(\theta;x,y)$ is $L(\theta)$-Lipschitz for every y, and its growth rate satisfies $\kappa(\theta) = L(\theta)$. Then

$$\sup_{Q \in \mathbb{B}_\varepsilon} \mathbb{E}_Q[\ell] = \frac{1}{N}\sum_{i=1}^{N} \ell\,(\theta;\xi_i) + \varepsilon\, L(\theta). \qquad (3)$$

Proof. If $\lambda < \kappa(\theta)$, the definition of κ lets x' run to infinity with the bracket unbounded, so the inner supremum is $+\infty$. If $\lambda \geq L(\theta)$, the Lipschitz bound gives $\ell(\theta;x',y_i) - \lambda \| x' - x_i \| \leq \ell(\theta;\xi_i)$, with equality at $x' = x_i$. Because $\kappa = L$, the feasible λ form exactly $[L(\theta), \infty)$. On that interval the objective of (2) is $\lambda\varepsilon$ plus the empirical risk, and it is minimised at $\lambda = L(\theta)$. ▫

Example (logistic regression). Take $\ell(\theta,b;x,y) = \log\Big(1 + \exp\big(-y(\theta^\top x + b)\big)\Big)$ with $y \in \{-1,+1\}$. The gradient in x equals $-y\,\sigma(\cdot)\,\theta$, where σ is the logistic function, so the Lipschitz constant is $\| \theta \|_*$ (the dual norm). Along the ray $x' = x - tyv$, with v a unit vector satisfying $\theta^\top v = \| \theta \|_*$, the slope tends to $\| \theta \|_*$ as $t \to \infty$. Hence $\kappa = L = \| \theta \|_*$, and (3) reads

$$J_\varepsilon = \min_{\theta,b} \frac{1}{N}\sum_{i=1}^{N} \log\left(1 + e^{-y_i(\theta^\top x_i + b)}\right) + \varepsilon \parallel \theta \parallel_*. \qquad (4)$$

The intercept b is not penalised because moving features cannot change it. The Euclidean case gives a group-lasso-type penalty $\varepsilon \parallel \theta \parallel_2$. The ℓ_∞ transport norm gives $\varepsilon \parallel \theta \parallel_1$. This is the known regularisation interpretation of Wasserstein-robust classifiers (Kuhn et al., 2019). Here it falls out of (2) in two lines.

3.3. Exchanging min and sup

Proposition 1. Under (A1)–(A2), and if $\theta \mapsto \ell(\theta;\xi)$ is continuous,

$$\min_{\theta\in\Theta} \sup_{Q\in\mathbb{B}_\varepsilon} \mathbb{E}_Q[\ell] = \sup_{Q\in\mathbb{B}_\varepsilon} \min_{\theta\in\Theta} \mathbb{E}_Q[\ell].$$

Proof. The ball $\mathbb{B}_\varepsilon$ is convex, since W is jointly convex. The map $Q \mapsto \mathbb{E}_Q[\ell]$ is linear, hence concave. The map $\theta \mapsto \mathbb{E}_Q[\ell]$ is convex and lower semicontinuous on the compact convex set Θ. Sion's theorem (1958) needs compactness on one side only, so it applies. ▫

The right side gives a statistical interpretation. A robust learner is an optimal Bayes learner against a least favourable distribution, or against a sequence of maximizing distributions when the supremum is not achieved. Thus, the minimax approach cannot only be viewed as conservative. It is the optimal approach to a particular but pessimistic world.

4. A first-order scheme

Formula (2) turns the game (1) into a joint minimisation over (θ, λ) with an inner maximisation per sample:

$$\min_{\theta\in\Theta,\ \lambda\geq 0} G(\theta,\lambda) = \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N} \phi_i(\theta,\lambda), \qquad \phi_i(\theta,\lambda) = \sup_{x'}[\ell(\theta;x',y_i) - \lambda \parallel x' - x_i \parallel].$$

Each ϕ_i is a supremum of functions that are jointly convex in (θ, λ), so G is convex. By Danskin's rule, a subgradient of ϕ_i is obtained by evaluating the partial gradients of the bracket at an inner maximiser $\hat{x}_i$. The general scheme is a stochastic subgradient method.

Algorithm 1 (dual stochastic subgradient). Input: ε, step sizes γ_k, batch size B, inner steps T.

1. Sample a mini-batch I_k of size B from $\{1, \dots, N\}$.
2. For each $i \in I_k$, start at $x' = x_i$ and run T steps of projected gradient ascent on $x' \mapsto \ell(\theta_k; x', y_i) - \lambda_k \parallel x' - x_i \parallel$. Call the result $\hat{x}_i$.
3. Set $g_\theta = \frac{1}{B}\sum_{i\in I_k} \nabla_\theta \ell(\theta_k; \hat{x}_i, y_i)$ and $g_\lambda = \varepsilon - \frac{1}{B}\sum_{i\in I_k} \parallel \hat{x}_i - x_i \parallel$.
4. Update $\theta_{k+1} = \Pi_\Theta(\theta_k - \gamma_k g_\theta)$ and $\lambda_{k+1} = \max\{0, \lambda_k - \gamma_k g_\lambda\}$.
5. Return the averaged iterate $(\bar{\theta}_K, \bar{\lambda}_K)$.

The λ-update has a plain meaning. If the adversary spent more than its budget on average, the price λ of transport goes up. If it spent less, the price goes down.

Convergence. Suppose the inner maximisations are solved exactly, $\Theta \times [0, \lambda_{\max}]$ has diameter D, and stochastic subgradients are bounded in second moment by M^2. With $\gamma_k = D/\left(M\sqrt{K}\right)$, the averaged iterate satisfies $\mathbb{E}\left[G\left(\bar{\theta}_K, \bar{\lambda}_K\right)\right] - \min G \leq DM/\sqrt{K}$. This is the robust stochastic approximation bound of Nemirovski et al. (2009). It needs no smoothness, which suits the kink that the norm term introduces at $\theta = 0$. For non-concave inner problems the bound holds with an additive error equal to the inner optimality gap.

The Lipschitz case needs no inner loop. Under Corollary 1, λ is eliminated in closed form ($\lambda^* = L(\theta)$), and Algorithm 1 reduces to (sub)gradient descent on (4). In Section 5 we use full-batch gradient descent on (4) with the Euclidean norm. We replace $\parallel \theta \parallel_2$ by the smooth surrogate $(\parallel \theta \parallel_2^2 + \tau)^{1/2}$ with $\tau = 10^{-8}$, use step size $0.2/(1 + k/200)^{1/2}$, and run 2,500 iterations from $\theta = 0$, $b = 0$.

5. Numerical illustration

These experiments are deliberately small-scale. They are designed to determine two things. One is if the dual value (3) is obtained using a real distribution of the ball, and the other is how ε balances between efficiency and robustness.

Data. The feature dimension is $d = 10$ and features are drawn as $x \sim \mathcal{N}(0, I)$. A ground-truth vector w is drawn at random and scaled to $\parallel w \parallel_2 = 2$. Labels follow $P(y = 1 \mid x) = \sigma(w^\top x)$, and 10% of labels are then flipped at random. The transport norm is Euclidean. All runs use NumPy with seed 7.

5.1. Is the dual bound attained?

We fit (4) on $N = 200$ points with $\varepsilon = 0.05$. The dual value is 0.61538. We then build two explicit members of the ball. In the first, every point is moved a distance ε in the direction $-y_i\theta/\parallel \theta \parallel_2$. Its expected loss is 0.58887. In the second, the whole budget goes to a single atom, which is moved a distance $N\varepsilon = 10$ in the same direction. We take the best of the 200 choices of atom, giving an expected loss of 0.61442.

The concentrated perturbation comes within 0.001 of the dual value, and the spread-out one falls well short. This is exactly what the proof of Corollary 1 predicts. The supremum is approached by sending a vanishing mass far

away, where the logistic loss is almost linear with slope $\| \theta \|_2$. The supremum is not attained by any finite move, which is why Proposition 1 speaks of maximising sequences.

5.2. Effect of the radius

For each of 30 trials we draw a training set of size $N \in \{50,200\}$ and a fresh test set of 5,000 points. We fit (4) for six values of ε. We report three quantities: test log-loss on clean data; test log-loss after each test point is shifted by 0.5 in the direction $-y\theta/\| \theta \|_2$, the worst ℓ_2 perturbation of that size for a linear score; and clean accuracy. Entries are means over 30 trials, with standard errors in parentheses.

Table 1: Test performance against Wasserstein radius ε (30 trials).

N	ε	Clean log-loss	Perturbed log-loss	Accuracy (%)
50	0 (ERM)	0.839 (0.036)	1.456 (0.079)	65.0
50	0.01	0.781 (0.027)	1.322 (0.060)	65.0
50	0.02	0.740 (0.021)	1.220 (0.047)	65.1
50	0.05	0.672 (0.010)	1.031 (0.025)	65.1
50	0.10	0.637 (0.006)	0.882 (0.013)	64.8
50	0.20	0.644 (0.004)	0.759 (0.006)	62.9
200	0 (ERM)	0.595 (0.003)	0.931 (0.012)	69.7
200	0.01	0.591 (0.003)	0.904 (0.010)	69.7
200	0.02	0.588 (0.003)	0.880 (0.009)	69.7
200	0.05	0.586 (0.002)	0.825 (0.007)	69.6
200	0.10	0.597 (0.003)	0.766 (0.004)	69.1
200	0.20	0.649 (0.004)	0.712 (0.002)	65.2

Three patterns stand out.

- **The benefits are greatest for small samples.** When N=50, the strongest model at ε=0.10 improves clean log-loss by 24% (0.839 to 0.637) and perturbed log-loss by 39% (1.456 to 0.882), at a cost of 0.2 points of accuracy. ERM without regularization on 50 points in 10 dimensions has an overconfident predictions, and the adversary's toll on θ_2 removes this.
- **The clean-loss optimum shifts to the left as N increases.** The optimal clean log-loss for N=200 is at ε=0.05, in comparison to ε=0.10 for N=50. This is consistent with the finite-sample guarantees in (Mohajerin Esfahani & Kuhn, 2018), where a certain radius that guarantees the likelihood level holds improves as N increases.
- **Perturbed loss is non-increasing, and Accuracy is not.** Loss under perturbation decreases monotonically in ε for both sample sizes, since that is what the game is protecting. For ε=0.20 the model is sufficiently reduced that accuracy drops by 2.1 points (N=50), and 4.5 points (N=200). The minimax learner protects the probability estimates, not the decision boundary, and pushes ε too far at the expense of accuracy.

So the radius is a real design parameter with an obvious interpretation: the transport budget the learner is willing to give up. It must be selected from the perturbation scale the application must endure, or by cross-validation in absence of that knowledge.

6. Concluding remarks

The Wasserstein minimax learner turns out to be less exotic than its definition: its variational dual (2) attributes only a single scalar coefficient to transport and decomposes the adversary into N local tasks. For any loss whose Lipschitz constant matches the rate of growth, the dual reduces to a standard regularised empirical risk (3). In the case of logistic regression with the norm penalty with weight epsilon, Sion's theorem implies the solution has a Bayesian interpretation as response to a least-favourable distribution.

Be explicit about the constraints. Corollary 1 requires unbounded features, with box-constrains features the adversary cannot attain the linear regime, and the dual must be in Algorithm 1. Our experiments are on a single synthetic model and a linear classifier. For non-Lipschitz losses such as squared loss, the type-2 distance is necessary and the penalty changes (Kuhn et al., 2010). We did not relax labels, but finite relabeling costs would introduce a margin shriveling term. Two extensions are quite straightforward. The first extension is to learn a data-dependent rule for ε based on the data-concentration results in (Mohajerin Esfahani & Kuhn, 2018). The second is to analyse Algorithm 1 for deep models with the inner maximisation being non-concave and solutions being available only approximately.

Data and code availability. This experiment includes synthetic data only. The code that reproduces the numbers in Section 5 is available from the author.

Conflict of interest. The author has no conflicts of interest to declare.

References

1. Blanchet, J., & Murthy, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, *44*(2), 565-600. https://doi.org/10.1287/moor.2018.0936

2. Delage, E., & Ye, Y. (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, *58*(3), 595-612. https://doi.org/10.1287/opre.1090.0741
3. Gao, R., & Kleywegt, A. (2023). Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, *48*(2), 603-655. https://doi.org/10.1287/moor.2022.1275
4. Kuhn, D., Esfahani, P. M., Nguyen, V. A., & Shafieezadeh-Abadeh, S. (2019). Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics* (pp. 130-166). Informs. https://doi.org/10.1287/educ.2019.0198
5. Mohajerin Esfahani, P., & Kuhn, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, *171*(1), 115-166. https://doi.org/10.1007/s10107-017-1172-1
6. Velan, I. (2026). Multi-Objective Optimization Of Urban Stormwater Infrastructure Considering Flood Mitigation, Lifecycle Cost, And Environmental Performance. *Journal of Smart Infrastructure and Environmental Sustainability*, *3*(3), 48-55.
7. Moti Ranjan Tandi. (2026). Resonance and Jump Phenomena in Microelectromechanical Systems with Cubic Stiffness Effects. *Applied Nonlinearity in Science and Technology*, *2*(3), 11–15.
8. Shapiro, A. (2017). Distributionally robust stochastic programming. *SIAM Journal on Optimization*, *27*(4), 2258-2275. https://doi.org/10.1137/16M1058297
9. Maidanov, K., & Fratlin, H. (2026). Mathematical Optimization Framework for Coordinated Power Flow and Voltage Regulation in Renewable Distribution Networks. *Journal of Applied Mathematical Models in Engineering*, 55-63.
10. Villani, C. (2009). *Optimal transport: old and new* (Vol. 338, p. 23). Berlin: springer. https://doi.org/10.1007/978-3-540-71050-9
11. Dr. M. Nalini. (2026). Soliton Interaction Analysis in Dispersive Plasma Media with Variable Density Gradients. *Applied Nonlinearity in Science and Technology*, *2*(3), 31–35.
12. Ahmad, H., Tariq, M., Cagin, T., Afzal, W., Hincal, E., Bayram, M., ... & Khan, I. (2026). Fractional inequalities for h-superquadratic stochastic processes. *Results in Nonlinear Analysis*, *9*(3), 80-105.
13. Dineshkumar, P. (2026). An Integrated Neurophysiological Framework For Investigating Cognitive Flexibility Through Behavioral Performance And Event-Related Potentials. *Advances in Cognitive and Neural Studies*, *2*(3), 38-45.
14. Surendar, A. (2026). A Systems-Based Framework For Evaluating Circular Resource Flows In Sustainable Urban Infrastructure Development. *Journal of Smart Infrastructure and Environmental Sustainability*, *3*(3), 56-63.
15. Rajan.C. (2025). Numerical Investigation of Acoustic Metamaterials Using Advanced Finite and Boundary Element Methods. *Advanced Computational Acoustics Engineering*, *3*(2), 25-30.
16. Eliana Marbel Rivas Téllez. (2026). Coupled Thermo-Mechanical Finite Element Analysis of Additively Manufactured Lattice Structures under Cyclic Loading Conditions. *Advances in Mechanical Engineering and Applications*, *2*(3), 25-32.
17. Marc Happi. (2026). Technology Incubation Programs and Women-Led Innovation. *Journal of Women, Innovation, and Technological Empowerment*, *2*(3), 33-39.