Fundamental tradeoffs in distributionally adversarial training

Mohammad Mehrabi

Department of Data Sciences and Operations
University of Southern California

joint with Adel Javanmard (USC),
Ryan A. Rossi(Adobe Research),
Anup B. Rao (Adobe Reseach),
Tung Mai (Adobe Reseach)

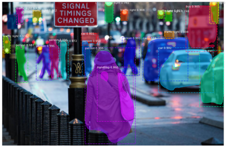
ICML 2021

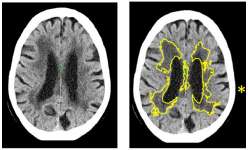
Modern data-driven algorithms

- Promising performance in dozens of safety-critical applications.

Modern data-driven algorithms

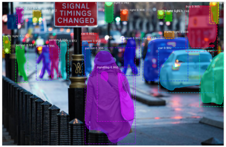
- Promising performance in dozens of safety-critical applications.

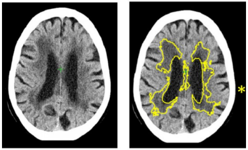




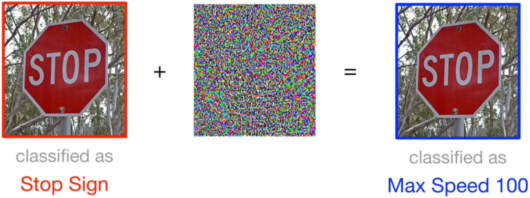
Modern data-driven algorithms

- Promising performance in dozens of safety-critical applications.



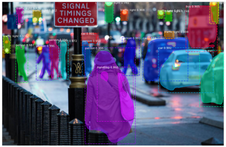


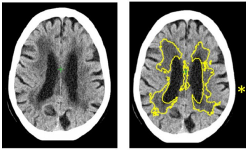
- Vulnerable to small discrepancies between training and test populations:



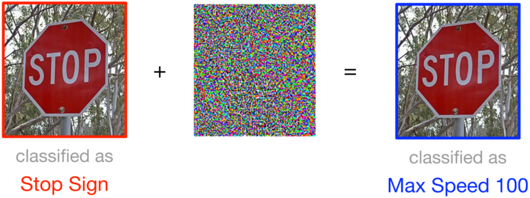
Modern data-driven algorithms

- Promising performance in dozens of safety-critical applications.





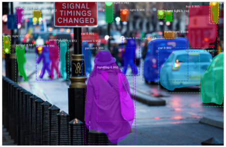
- Vulnerable to small discrepancies between training and test populations:

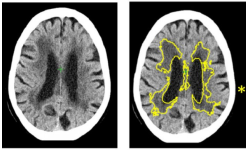


- Adversarial training is an effective technique to improve robustness

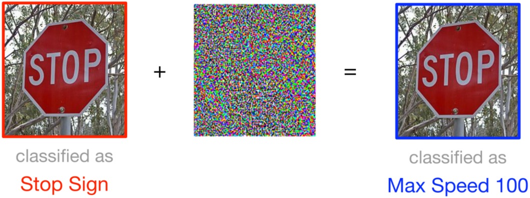
Modern data-driven algorithms

- Promising performance in dozens of safety-critical applications.






- Vulnerable to small discrepancies between training and test populations:



- Adversarial training is an effective technique to improve robustness
- Adversarial training degrades the model accuracy on benign test inputs

Classic supervised learning setup

- Data $\{z_i = (x_i, y_i)\}_{i=1:n} \overset{\text{iid}}{\sim} \mathbb{P}_z(\mathcal{Z})$ on metric space $\mathcal{Z}$ and norm $d(.,.)$
- Parametric loss $\ell(\theta; z = (x, y))$.

Classic supervised learning setup

- Data $\{z_i = (x_i, y_i)\}_{i=1:n} \overset{\text{iid}}{\sim} \mathbb{P}_z(\mathcal{Z})$ on metric space $\mathcal{Z}$ and norm $d(.,.)$
- Parametric loss $\ell(\theta; z = (x, y))$.
- Assess model θ performance:
  Standard Risk: SR(θ)=$\mathbb{E}_{z=(x,y)\sim P_z}[\ell(\theta; z)]$
  Expected loss on a new test data point from training population P_z

Classic supervised learning setup

- Data $\{z_i = (x_i, y_i)\}_{i=1:n} \overset{\text{iid}}{\sim} \mathbb{P}_z(\mathcal{Z})$ on metric space $\mathcal{Z}$ and norm $d(.,.)$
- Parametric loss $\ell(\theta; z = (x, y))$.
- Assess model θ performance:
  Standard Risk: $\text{SR}(\theta) = \mathbb{E}_{z=(x,y)\sim P_z}[\ell(\theta; z)]$
  Expected loss on a new test data point from <u>training population</u> P_z

Model performance when there is a <u>distributional shift</u> $\Rightarrow$ **Adversarial Risk**

Adversarial setup: distributional shift

Game between learner and adversary

Adversarial setup: distributional shift

Game between learner and adversary

Learner:

- Access to data generated iid from P_z
- Pick model θ (with empircal risk minimization, etc.)

Adversarial setup: distributional shift

Game between learner and adversary

Learner:

- Access to data generated iid from P_z
- Pick model θ (with empircal risk minimization, etc.)

Adversary:

- Access to the training distribution P_z and model θ
- Pick distribution of test data from an ε-neighborhood of P_z

Popular choice for an ε-neighborhood of P_z is Wasserstein ball: $\mathcal{U}_\varepsilon(P_z)$.

Adversarial setup: distributional shift

Game between learner and adversary

Learner:

- Access to data generated iid from P_z
- Pick model θ (with empircal risk minimization, etc.)

Adversary:

- Access to the training distribution P_z and model θ
- Pick distribution of test data from an ε-neighborhood of P_z

Popular choice for an ε-neighborhood of P_z is Wasserstein ball: $\mathcal{U}_\varepsilon(P_z)$.
Adversarial risk: $\mathrm{AR}(\theta) = \sup_{Q \in \mathcal{U}_\varepsilon(P_z)} \mathbb{E}_{z=(x,y)\sim Q}[\ell(\theta; z)]$

Main results

Fundamental question:

With unlimited number of training points and computational power:

Is there a model which is optimal in both standard and adversarial risks?

Is there a fundamental tradeoff between standard and adversarial risks?

Main results

Fundamental question:

With unlimited number of training points and computational power:

Is there a model which is optimal in both standard and adversarial risks?

Is there a fundamental tradeoff between standard and adversarial risks?

Main results:

- For three classes of statistical learning problems, indeed a tradeoff between standard and adversarial risk is manifested:
  i) Linear regression
  ii) Binary classification under a Gaussian mixtures model
  ii) The problem of learning an unknown function over a high-dimensional sphere using random features model

Main results

Fundamental question:

With unlimited number of training points and computational power:

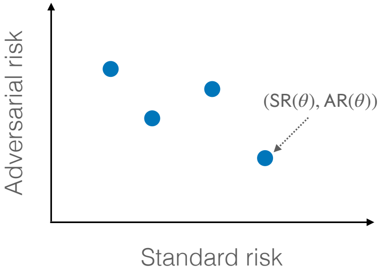
Is there a model which is optimal in both standard and adversarial risks?

Is there a fundamental tradeoff between standard and adversarial risks?

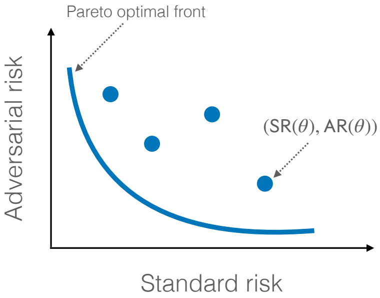
Main results:

- For three classes of statistical learning problems, indeed a tradeoff between standard and adversarial risk is manifested:
  i) Linear regression
  ii) Binary classification under a Gaussian mixtures model
  ii) The problem of learning an unknown function over a high-dimensional sphere using random features model
- Characterize such tradeoffs + effect of a variety of factors on them: problem dimension, adversary's power, complexity of the model class (e.g number of neurons)

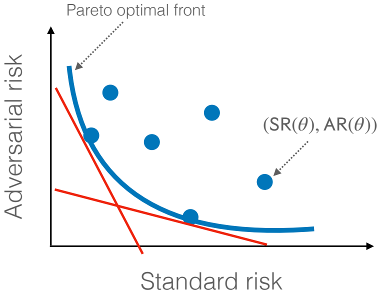
Pareto-optimal curve: characterization



Pareto-optimal curve: characterization

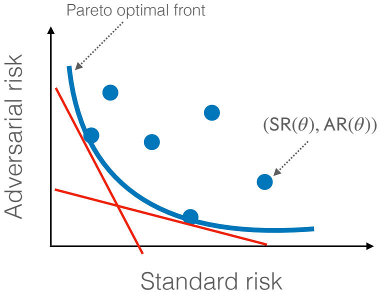


Pareto-optimal curve: characterization



$$\widehat{\theta}_\lambda = \arg\min_\theta \{\lambda \mathsf{SR}(\theta) + \mathsf{AR}(\theta)\}$$

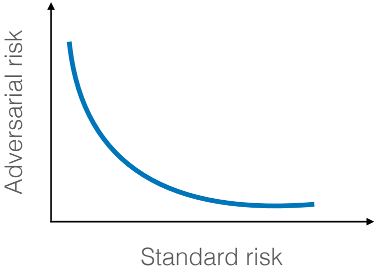
Pareto-optimal curve: characterization



$$\widehat{\theta}_\lambda = \arg\min_\theta \{\lambda \mathsf{SR}(\theta) + \mathsf{AR}(\theta)\}$$

Pareto optimal front: $\left\{ \left(\mathsf{SR}(\widehat{\theta}_\lambda), \mathsf{AR}(\widehat{\theta}_\lambda) \right), \lambda \geq 0 \right\}$

Pareto-optimal curve: characterization



$$\widehat{\theta}_\lambda = \arg\min_\theta \{\lambda \mathsf{SR}(\theta) + \mathsf{AR}(\theta)\}$$

Pareto optimal front: $\left\{ \left(\mathsf{SR}(\widehat{\theta}_\lambda), \mathsf{AR}(\widehat{\theta}_\lambda) \right), \lambda \geq 0 \right\}$

Adversarial risk: characterization

Adversarial Risk: $$\mathsf{AR}(\theta) = \sup_{Q \in \mathcal{U}_\varepsilon(P_Z)} E_{z=(x,y)\sim Q}[\ell(\theta; z)]$$

(Wasserstein ball) $\mathcal{U}_\varepsilon(P) = \{Q : W(Q, P) \le \varepsilon\}\,,$

(Wasserstein distance) $W(Q, P) = \inf_{\pi \in \mathsf{Cpl}(Q,P)} \left(\mathbb{E}_{(z_1,z_2)\sim\pi}[d^2(z_1, z_2)]\right)^{1/2}\,,$

(Metric on data points) $d(z, z') = ||x - x'||_{\ell_r} + \infty \cdot \mathbb{I}_{\{y \neq y'\}}$

Adversarial risk: characterization

Adversarial Risk: $\mathsf{AR}(\theta) = \sup_{Q \in \mathcal{U}_\varepsilon(P_Z)} E_{z=(x,y)\sim Q}[\ell(\theta; z)]$

$$\text{(Wasserstein ball)} \quad \mathcal{U}_\varepsilon(P) = \{Q : W(Q,P) \le \varepsilon\},$$

$$\text{(Wasserstein distance)} \quad W(Q,P) = \inf_{\pi \in \mathsf{Cpl}(Q,P)} \left(\mathbb{E}_{(z_1,z_2)\sim\pi}[d^2(z_1,z_2)]\right)^{1/2},$$

$$\text{(Metric on data points)} \quad d(z,z') = ||x - x'||_{\ell_r} + \infty \cdot \mathbb{I}_{\{y \neq y'\}}$$

Adversarial Risk dual problem:

$$\min_{\gamma \ge 0} \left\{ \gamma\varepsilon^2 + \mathbb{E}_{P_z}\Big[\underbrace{\Phi_\gamma(\theta; z)}_{\text{robust surrogate for } \ell(\theta;z)} \Big] \right\}$$

Adversarial risk: characterization

Adversarial Risk: $\mathsf{AR}(\theta) = \sup_{Q \in \mathcal{U}_\varepsilon(P_Z)} E_{z=(x,y)\sim Q}[\ell(\theta; z)]$

$$\text{(Wasserstein ball)} \quad \mathcal{U}_\varepsilon(P) = \{Q : W(Q,P) \le \varepsilon\},$$

$$\text{(Wasserstein distance)} \quad W(Q,P) = \inf_{\pi \in \mathsf{Cpl}(Q,P)} \left(\mathbb{E}_{(z_1,z_2)\sim\pi}[d^2(z_1,z_2)]\right)^{1/2},$$

$$\text{(Metric on data points)} \quad d(z,z') = ||x - x'||_{\ell_r} + \infty \cdot \mathbb{I}_{\{y \neq y'\}}$$

Adversarial Risk dual problem:

$$\min_{\gamma \ge 0} \left\{ \gamma\varepsilon^2 + \mathbb{E}_{P_z}\Big[\underbrace{\Phi_\gamma(\theta; z)}_{\text{robust surrogate for } \ell(\theta;z)} \Big] \right\}$$

Robust surrogate:

$$\Phi_\gamma(\theta; z_0) = \sup_{z \in \mathcal{Z}} \left\{ \ell(\theta; z) - \gamma \cdot d^2(z, z_0) \right\}$$

Adversarial risk: characterization

Adversarial Risk: $\text{AR}(\theta) = \sup_{Q \in \mathcal{U}_\varepsilon(P_Z)} E_{z=(x,y)\sim Q}[\ell(\theta; z)]$

$$\text{(Wasserstein ball)} \quad \mathcal{U}_\varepsilon(P) = \{Q : W(Q,P) \le \varepsilon\},$$

$$\text{(Wasserstein distance)} \quad W(Q,P) = \inf_{\pi \in \mathsf{Cpl}(Q,P)} \left(\mathbb{E}_{(z_1,z_2)\sim\pi}[d^2(z_1,z_2)]\right)^{1/2},$$

$$\text{(Metric on data points)} \quad d(z,z') = ||x - x'||_{\ell_r} + \infty \cdot \mathbb{I}_{\{y \neq y'\}}$$

Adversarial Risk dual problem:

$$\min_{\gamma \ge 0} \left\{ \gamma\varepsilon^2 + \mathbb{E}_{P_z}\Big[\underbrace{\Phi_\gamma(\theta; z)}_{\text{robust surrogate for } \ell(\theta;z)} \Big] \right\}$$

Robust surrogate:

$$\Phi_\gamma(\theta; z_0) = \sup_{z \in \mathcal{Z}} \left\{ \ell(\theta; z) - \gamma \cdot d^2(z, z_0) \right\}$$

Strong duality holds for Polish space $\mathcal{Z}$.

Pareto-optimal tradeoff: linear regression

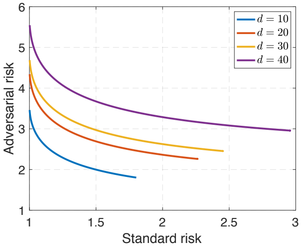
$$y = x^{\mathsf{T}}\theta_0 + \mathsf{N}(0,1)\,, \quad x \sim \mathsf{N}(0, \Sigma_{\mathbf{d}}), \quad \Sigma_{ij} = \rho^{|i-j|}$$

(square loss) $\ell(\theta; (x,y)) = (y - x^{\mathsf{T}}\theta)^2$

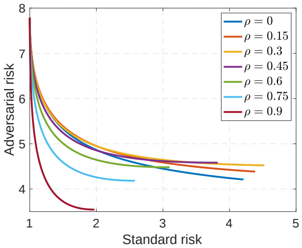
Pareto-optimal tradeoff: linear regression

$$y = x^\mathsf{T}\theta_0 + \mathsf{N}(0,1)\,, \quad x \sim \mathsf{N}(0, \Sigma_{\mathbf{d}}), \quad \Sigma_{ij} = \rho^{|i-j|}$$

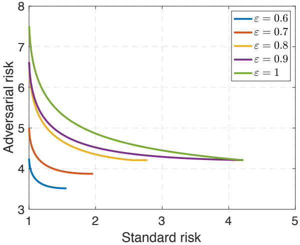
$$\text{(square loss) } \ell(\theta; (x,y)) = (y - x^\mathsf{T}\theta)^2$$



(a) Pareto optimal curve for several feature dimensions d.



(b) Pareto optimal curve for several feature dependency values ρ.



(c) Pareto optimal curve for several adversary's manipulative power ε .

Pareto-optimal tradeoff: binary classification

$$y \in \{+1, -1\}, \quad x \sim \mathsf{N}\left(y\mu, \Sigma_d\right), \quad \Sigma_{ij} = \rho^{|i-j|}$$

(linear classifiers) $\quad \ell(\theta; (x, y)) = \mathbb{I}\{yx^{\mathsf{T}}\theta \le 0\}$

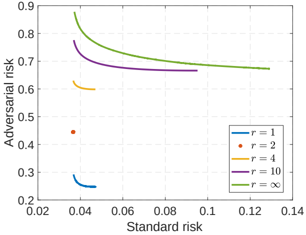
(metric on samples) $d(z, z') = ||x - x||_{\ell_r} + \infty \cdot \mathbb{I}\{y \neq y'\}$

Pareto-optimal tradeoff: binary classification

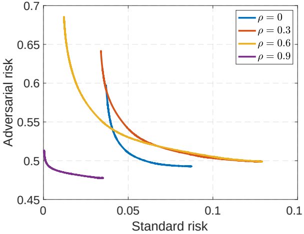
$$y \in \{+1, -1\}, \quad x \sim \mathsf{N}\left(y\mu, \Sigma_d\right), \quad \Sigma_{ij} = \rho^{|i-j|}$$

(linear classifiers) $\ell(\theta; (x, y)) = \mathbb{I}\{y x^{\mathsf{T}} \theta \leq 0\}$

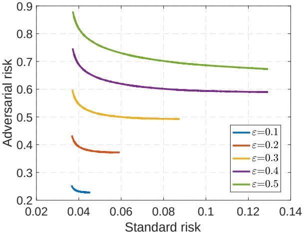
(metric on samples) $d(z, z') = ||x - x||_{\ell_r} + \infty \cdot \mathbb{I}\{y \neq y'\}$



(a) Pareto optimal curve for several ℓ_r norms on feature space.



(b) Pareto optimal curve for several feature dependency values (ρ) .



(c) Pareto optimal curve for several adversary's manipulative power ε.

Pareto-optimal tradeoff: learning non-linear functions

$$x \sim \mathsf{Unif}\left(\mathbb{S}^{d-1}(\sqrt{d})\right) \ ,$$

$$f(x) = \beta_0 + \beta_1^{\mathsf{T}} x + \underbrace{\frac{\beta_2}{d}\left(x^{\mathsf{T}} G x - \mathsf{tr}(G)\right)}_{\text{quadratic with } G \overset{\text{iid}}{\sim} \mathsf{N}(0,1)} + \mathsf{N}(0, \sigma^2)$$

(random features model) $\left\{ f(x, \theta, U) = \theta^T \sigma(Ux), U \in \mathbb{R}^{N \times d}, \theta \in \mathbb{R}^N \right\}$, rows of $U \overset{iid}{\sim} \mathbb{S}^{d-1}(1)$

Pareto-optimal tradeoff: learning non-linear functions

$$x \sim \text{Unif}\left(\mathbb{S}^{d-1}(\sqrt{d})\right),$$

$$f(x) = \beta_0 + \beta_1^\mathsf{T} x + \underbrace{\frac{\beta_2}{d}\left(x^\mathsf{T} G x - \text{tr}(G)\right)}_{\text{quadratic with } G \overset{\text{iid}}{\sim} \mathsf{N}(0,1)} + \mathsf{N}(0, \sigma^2)$$

(random features model) $\left\{f(x,\theta,U) = \theta^T \sigma(Ux), U \in \mathbb{R}^{N\times d}, \theta \in \mathbb{R}^N\right\}$, rows of $U \overset{iid}{\sim} \mathbb{S}^{d-1}(1)$