Distirbutional robustness, regularizing variance, and adversaries

John Duchi
Based on joint work with Hongseok Namkoong and Aman Sinha

Stanford University

November 2017

Motivation

We do not want machine-learned systems to fail
when they get in the real world

Challenge one: Curly fries

WIRED Technology | Science | Culture | Video | Reviews | Magazine

Liking curly fries on Facebook reveals your high IQ

By PHILIPPA WARR

12 Mar 2013

What you Like on Facebook could reveal your race, age, IQ, sexuality and other personal data, even if you've set that information to "private".

Challenge one: Curly fries

WIRED Technology | Science | Culture | Video | Reviews | Magazine

Liking curly fries on Facebook reveals your high IQ

By PHILIPPA WARR
12 Mar 2013

What you Like on Facebook could reveal your race, age, IQ, sexuality and other personal data, even if you've set that information to "private".

Who doesn't like curly fries?

Challenge two: changes in environment

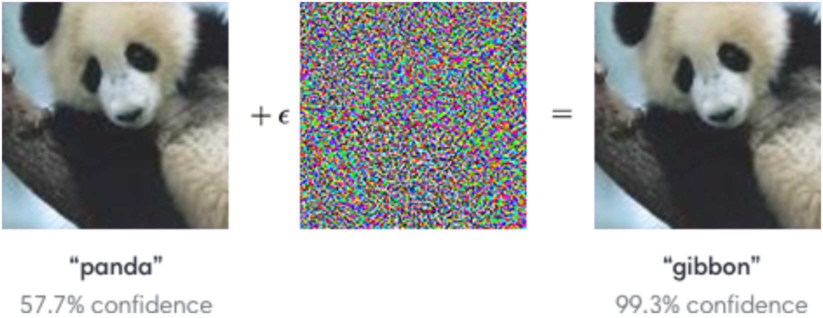
Learning to drive in California

Challenge two: changes in environment

Learning to drive in California

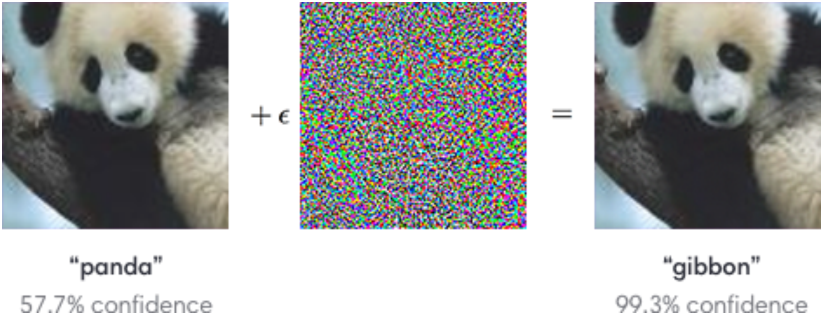
Driving in Ann Arbor

Challenge three: adversaries



[Goodfellow et al. 15]

Challenge three: adversaries



[Goodfellow et al. 15]

Paraphrased Quote:

We could put a transparent film on a stop sign, essentially imperceptible to a human, and a computer would see the stop sign as air (Dan Boneh)

Stochastic optimization problems

$$\text{minimize } R(\theta) := \mathbb{E}_{P_0}[\ell(\theta; Z)] = \int \ell(\theta; z) dP_0(z)$$

$$\text{subject to } \theta \in \Theta.$$

Empirical risk minimization: Often, solve

$$\widehat{\theta}_n = \operatorname*{argmin}_{\theta \in \Theta} \widehat{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; Z_i)$$

Stochastic optimization problems

$$\begin{aligned}&\text{minimize } R(\theta) := \mathbb{E}_{P_0}[\ell(\theta; Z)] = \int \ell(\theta; z) dP_0(z)\\ &\text{subject to } \theta \in \Theta.\end{aligned}$$

- Data/randomness is Z
- Loss function $\theta \mapsto \ell(\theta; z)$
- Parameter space Θ is a nonempty closed (convex) set

Empirical risk minimization: Often, solve

$$\widehat{\theta}_n = \operatorname*{argmin}_{\theta \in \Theta} \widehat{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; Z_i)$$

Stochastic optimization problems

$$\begin{aligned} &\text{minimize } R(\theta) := \mathbb{E}_{P_0}[\ell(\theta; Z)] = \int \ell(\theta; z) dP_0(z) \\ &\text{subject to } \theta \in \Theta. \end{aligned}$$

- Data/randomness is Z
- Loss function $\theta \mapsto \ell(\theta; z)$
- Parameter space Θ is a nonempty closed (convex) set
- Observe data $Z_i \stackrel{\text{iid}}{\sim} P_0$, $i = 1, \ldots, n$

Empirical risk minimization: Often, solve

$$\widehat{\theta}_n = \operatorname*{argmin}_{\theta \in \Theta} \widehat{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; Z_i)$$

Distributional robustness

$$R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; Z)]$$

Distributional robustness

$$R(\theta, \mathcal{P}) := \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(\theta; Z)]$$

Distributional robustness

$$R(\theta, \mathcal{P}) := \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(\theta; Z)]$$

- Uncertainty set $\mathcal{P}$ is set of "possible" distributions/worlds
- Different choices of uncertainty yield different behaviors
- Some sample-based uncertainty sets $\mathcal{P}$ certify future performance

Distributional robustness

$$R(\theta, \mathcal{P}) := \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(\theta; Z)]$$

- Uncertainty set $\mathcal{P}$ is set of “possible” distributions/worlds
- Different choices of uncertainty yield different behaviors
- Some sample-based uncertainty sets $\mathcal{P}$ certify future performance
- Much work in optimization literature: [Delage & Ye 10, Ben-Tal et al. 13, Bertsimas et al. 14, Lam & Zhou 15, Gotoh et al. 15]

Distributional robustness

$$R(\theta, \mathcal{P}) := \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(\theta; Z)]$$

- Uncertainty set $\mathcal{P}$ is set of "possible" distributions/worlds
- Different choices of uncertainty yield different behaviors
- Some sample-based uncertainty sets $\mathcal{P}$ certify future performance
- Much work in optimization literature: [Delage & Ye 10, Ben-Tal et al. 13, Bertsimas et al. 14, Lam & Zhou 15, Gotoh et al. 15]

Rest of this talk: Two vignettes showing some aspects of this approach

Vignette one: regularization by variance

Vignette one: regularization by variance

- Any learning algorithm has *bias* (approximation error) and *variance* (estimation error)

Vignette one: regularization by variance

- Any learning algorithm has *bias* (approximation error) and *variance* (estimation error)
- From empirical Bernstein's inequality, with probability $1 - \delta$

$$R(\theta) \le \underbrace{\widehat{R}_n(\theta)}_{\text{bias}} + \underbrace{\sqrt{\frac{2\mathrm{Var}_{\widehat{P}_n}\left(\ell(\theta; X)\right)}{n}}}_{\text{variance}} + \frac{C \log \frac{1}{\delta}}{n}$$

Vignette one: regularization by variance

- Any learning algorithm has *bias* (approximation error) and *variance* (estimation error)
- From empirical Bernstein's inequality, with probability $1 - \delta$

$$R(\theta) \le \underbrace{\widehat{R}_n(\theta)}_{\text{bias}} + \underbrace{\sqrt{\frac{2\mathrm{Var}_{\widehat{P}_n}\left(\ell(\theta; X)\right)}{n}}}_{\text{variance}} + \frac{C \log \frac{1}{\delta}}{n}$$

Goal: Trade between these automatically and optimally by solving

$$\widehat{\theta}^{\mathrm{var}} \in \operatorname*{argmin}_{\theta \in \Theta} \left\{ \widehat{R}_n(\theta) + \sqrt{\frac{2\mathrm{Var}_{\widehat{P}_n}\left(\ell(\theta; X)\right)}{n}} \right\}.$$

Optimizing for bias and variance

Good idea: Directly minimize bias + variance, certify optimality!

Optimizing for bias and variance

Good idea: Directly minimize bias + variance, certify optimality!
Minor issue: variance is wildly non-convex

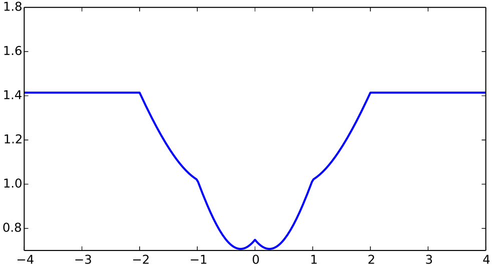


Figure: Variance of $\ell(\theta, X) = |\theta - X|$

Robust ERM

Goal:

$$\underset{\theta \in \Theta}{\text{minimize}} \; R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)]$$

Robust ERM

Goal:

$$\underset{\theta \in \Theta}{\text{minimize}} \;\; R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)]$$

Solve empirical risk minimization problem

$$\underset{\theta \in \Theta}{\text{minimize}} \;\; \sum_{i=1}^{n} \frac{1}{n} \ell(\theta; X_i)$$

Robust ERM

Goal:

$$\underset{\theta \in \Theta}{\text{minimize}} \;\; R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)]$$

Solve sample average optimization problem

$$\underset{\theta \in \Theta}{\text{minimize}} \;\; \sum_{i=1}^{n} \frac{1}{n} \ell(\theta; X_i)$$

Robust ERM

Goal:

$$\underset{\theta \in \Theta}{\text{minimize}} \;\; R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)]$$

Instead, solve *distributionally robust optimization (RO) problem*

$$\underset{\theta \in \Theta}{\text{minimize}} \;\; \sup_{p \in \mathcal{P}_{n,\rho}} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$

where $\mathcal{P}_{n,\rho}$ is some appropriately chosen set of vectors

Robust ERM

Goal:

$$\underset{\theta \in \Theta}{\text{minimize}} \; R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)]$$

Instead, solve *distributionally robust optimization (RO) problem*

$$\underset{\theta \in \Theta}{\text{minimize}} \; \sup_{p \in \mathcal{P}_{n,\rho}} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$

where $\mathcal{P}_{n,\rho}$ is some appropriately chosen set of vectors

This bit of talk: Give a principled statistical approach to choosing $\mathcal{P}_{n,\rho}$ and give stochastic optimality certificates for RO.

Empirical likelihood and robustness

Idea: Optimize over *uncertainty set* of possible distributions,

$$\mathcal{P}_{n,\rho} := \left\{ \text{Distributions } P \text{ such that } D(P\|\widehat{P}_n) \leq \frac{\rho}{n} \right\}$$

for some $\rho > 0$, where $D(P\|Q) = \int (p/q - 1)^2 q$

Empirical likelihood and robustness

Idea: Optimize over *uncertainty set* of possible distributions,

$$\mathcal{P}_{n,\rho} := \left\{ \text{Distributions } P \text{ such that } D(P \| \widehat{P}_n) \le \frac{\rho}{n} \right\}$$

for some $\rho > 0$, where $D(P\|Q) = \int (p/q - 1)^2 q$

Define (and optimize) *empirical likelihood upper confidence bound*

$$R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P \in \mathcal{P}_{n,\rho}} \mathbb{E}_P[\ell(\theta, X)] = \max_{p \in \mathcal{P}_{n,\rho}} \sum_{i=1}^n p_i \ell(\theta, X_i)$$

Empirical likelihood and robustness

Idea: Optimize over *uncertainty set* of possible distributions,

$$\mathcal{P}_{n,\rho} := \left\{ \text{Distributions } P \text{ such that } D(P \| \widehat{P}_n) \leq \frac{\rho}{n} \right\}$$

for some $\rho > 0$, where $D(P\|Q) = \int (p/q - 1)^2 q$

Define (and optimize) *empirical likelihood upper confidence bound*

$$R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P \in \mathcal{P}_{n,\rho}} \mathbb{E}_P[\ell(\theta, X)] = \max_{p \in \mathcal{P}_{n,\rho}} \sum_{i=1}^n p_i \ell(\theta, X_i)$$

Nice properties:

- Convex optimization problem
- Efficient solution methods [D. & Namkoong NIPS 16]

Robust Optimization = Variance Regularization

Theorem (D. & Namkoong)

Assume that ℓ is bounded over the space of decision vectors θ. Then

$$R_n(\theta; \mathcal{P}_{n,\rho}) = \widehat{R}_n(\theta) + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}} + O(\rho/n).$$

Robust Optimization = Variance Regularization

Theorem (D. & Namkoong)

Assume that ℓ is bounded over the space of decision vectors θ. Then

$$R_n(\theta; \mathcal{P}_{n,\rho}) = \widehat{R}_n(\theta) + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}} + O(\rho/n).$$

Choose $\widehat{\theta}^{\mathrm{rob}}$ to minimize robust empirical risk

$$R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P \in \mathcal{P}_{n,\rho}} \mathbb{E}_P[\ell(\theta, X)] = \max_{p \in \mathcal{P}_{n,\rho}} \sum_{i=1}^n p_i \ell(\theta, X_i).$$

Optimal bias variance tradeoff

Choose $\widehat{\theta}^{\mathrm{rob}}$ to minimize robust empirical risk

$$R_n(\widehat{\theta}^{\mathrm{rob}}, \mathcal{P}_{n,\rho}) = \min_{\theta \in \Theta} \max_{P \ll \widehat{P}_n} \left\{ \mathbb{E}_P[\ell(\theta; X)] : D_{\chi^2}\left(P \| \widehat{P}_n\right) \leq \frac{\rho}{n} \right\}.$$

Optimal bias variance tradeoff

Choose $\widehat{\theta}^{\mathrm{rob}}$ to minimize robust empirical risk

$$R_n(\widehat{\theta}^{\mathrm{rob}}, \mathcal{P}_{n,\rho}) = \min_{\theta \in \Theta} \max_{P \ll \widehat{P}_n} \left\{ \mathbb{E}_P[\ell(\theta; X)] : D_{\chi^2}\left(P \| \widehat{P}_n\right) \leq \frac{\rho}{n} \right\}.$$

Assume that $\Theta \subset \mathbb{R}^d$ compact with radius R and $\ell(\theta; X)$ is M-Lipschitz.

Optimal bias variance tradeoff

Choose $\widehat{\theta}^{\rm rob}$ to minimize robust empirical risk

$$R_n(\widehat{\theta}^{\rm rob}, \mathcal{P}_{n,\rho}) = \min_{\theta \in \Theta} \max_{P \ll \widehat{P}_n} \left\{ \mathbb{E}_P[\ell(\theta; X)] : D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n} \right\}.$$

Assume that $\Theta \subset \mathbb{R}^d$ compact with radius R and $\ell(\theta; X)$ is M-Lipschitz.

Theorem (D. & Namkoong 17)

Let $\rho = \log \frac{1}{\delta} + d \log n$. Then with probability at least $1 - \delta$,

$$R(\widehat{\theta}^{\rm rob}) \le \underbrace{R_n(\widehat{\theta}^{\rm rob}, \mathcal{P}_{n,\rho})}_{\text{optimality certificate}} + \frac{cMR}{n}\rho$$

$$\le \underbrace{\min_{\theta \in \Theta} \left\{ R(\theta) + 2\sqrt{\frac{2\rho \mathrm{Var}(\ell(\theta, \xi))}{n}} \right\}}_{\text{optimal tradeoff}} + \frac{cMR}{n}\rho$$

for some universal constant $c > 0$.

Experiment: Reuters Corpus (multi-label)

Problem: Classify documents as a **subset** of the 4 categories:

$$\Big\{\text{Corporate, Economics, Government, Markets}\Big\}$$

Experiment: Reuters Corpus (multi-label)

Problem: Classify documents as a **subset** of the 4 categories:

$$\Big\{\text{Corporate, Economics, Government, Markets}\Big\}$$

- Data: pairs $x \in \mathbb{R}^d$ represents document, $y \in \{-1, 1\}^4$ where $y_j = 1$ indicating x belongs j-th category.

Experiment: Reuters Corpus (multi-label)

Problem: Classify documents as a **subset** of the 4 categories:

$$\Big\{\text{Corporate, Economics, Government, Markets}\Big\}$$

- Data: pairs $x \in \mathbb{R}^d$ represents document, $y \in \{-1, 1\}^4$ where $y_j = 1$ indicating x belongs j-th category.
- Loss $\ell(\theta_j, (x, y)) = \log(1 + e^{-yx^\top \theta_j})$ for each $j = 1, \dots, 4$ and $\Theta = \big\{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq 1000\big\}$.

Experiment: Reuters Corpus (multi-label)

Problem: Classify documents as a **subset** of the 4 categories:

$$\Big\{\text{Corporate, Economics, Government, Markets}\Big\}$$

- Data: pairs $x \in \mathbb{R}^d$ represents document, $y \in \{-1, 1\}^4$ where $y_j = 1$ indicating x belongs j-th category.
- Loss $\ell(\theta_j, (x, y)) = \log(1 + e^{-yx^\top \theta_j})$ for each $j = 1, \ldots, 4$ and $\Theta = \big\{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq 1000\big\}$.
- $d = 47,236$, $n = 804,414$. 10-fold cross-validation.

Experiment: Reuters Corpus (multi-label)

Problem: Classify documents as a **subset** of the 4 categories:

$$\Big\{\text{Corporate, Economics, Government, Markets}\Big\}$$

- Data: pairs $x \in \mathbb{R}^d$ represents document, $y \in \{-1, 1\}^4$ where $y_j = 1$ indicating x belongs j-th category.
- Loss $\ell(\theta_j, (x, y)) = \log(1 + e^{-yx^\top \theta_j})$ for each $j = 1, \ldots, 4$ and $\Theta = \left\{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq 1000\right\}$.
- $d = 47,236$, $n = 804,414$. 10-fold cross-validation.

Table: Reuters Number of Examples

Corporate	Economics	Government	Markets
381,327	119,920	239,267	204,820

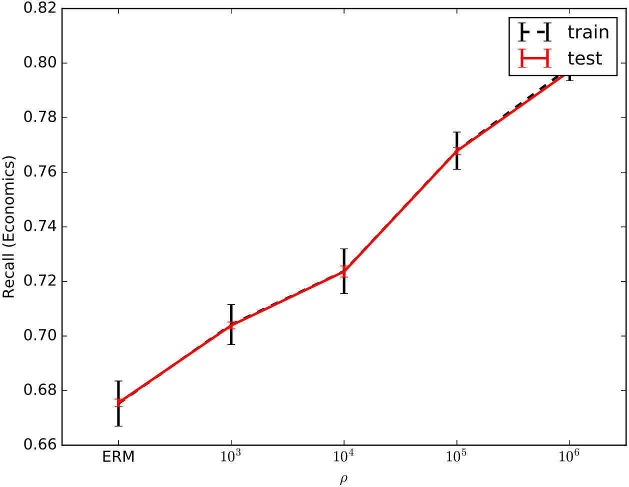
Experiment: Reuters Corpus (multi-label)

Table: Reuters Corpus (%)

	Precision		Recall		Corporate		Economics	
ρ	train	test	train	test	train	test	train	test
erm	92.72	92.7	90.97	90.96	90.2	90.25	67.53	67.56
10000	94.17	94.16	93.46	93.44	92.65	92.71	76.79	76.78

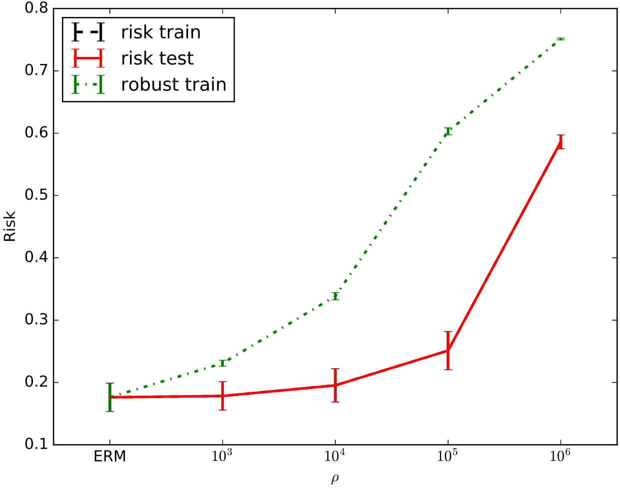
Experiment: Reuters Corpus (multi-label)

Figure: Recall on rare category (Economics)



Experiment: Reuters Corpus (multi-label)

Figure: Average logistic risk and confidence bound



Vignette two: Wasserstein robustness

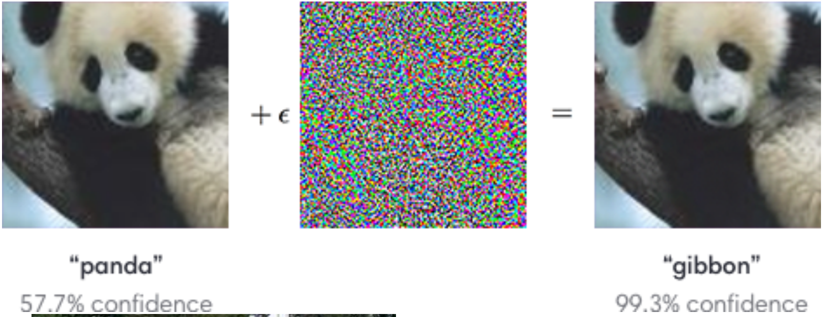
We do not want machine-learned systems to fail
when they get in the real world

Vignette two: Wasserstein robustness

We do not want machine-learned systems to fail
when they get in the real world

It is irresponsible to release systems into the world whose robustness
we do not understand

Challenges



A type of robustess

Robust optimization: instead of ℓ, look at robust loss

$$\ell_\epsilon(\theta; z) := \sup_{\|\Delta\| \le \epsilon} \ell(\theta; z + \Delta)$$

A type of robustess

Robust optimization: instead of ℓ, look at robust loss

$$\ell_\epsilon(\theta; z) := \sup_{\|\Delta\| \le \epsilon} \ell(\theta; z + \Delta)$$

- Adversarial attacks and defenses with heuristics and more advanced ideas [Goodfellow et al. 15, Jia and Liang 17, Papernot et al. 16, Madry et al. 17]

A type of robustess

Robust optimization: instead of ℓ, look at robust loss

$$\ell_\epsilon(\theta; z) := \sup_{\|\Delta\| \le \epsilon} \ell(\theta; z + \Delta)$$

- Adversarial attacks and defenses with heuristics and more advanced ideas [Goodfellow et al. 15, Jia and Liang 17, Papernot et al. 16, Madry et al. 17]

Minor issue: Usually this is NP-hard
Further issue: In neural network,

$$f_\theta(x) = \theta_1^T \sigma_{\mathrm{relu}}(\theta_2^T \sigma_{\mathrm{relu}}(\cdots))$$

and is is NP-hard to compute $\sup_\Delta \ell(f_\theta(x + \Delta))$

Distributional robustness

Question: How can we figure out how to "change" distribution right way to get robustness?

Distributional robustness

Question: How can we figure out how to "change" distribution right way to get robustness?

Let $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ be some cost function, and define *Wasserstein distance*

$$\begin{aligned} W_c(P, Q) &:= \inf_M \int c(z_1, z_2) dM(z_1, z_2) \\ &= \sup_f \left\{ \int f(z)(dP(z) - dQ(z)) \mid f(x) - f(z) \leq c(x, z) \right\} \end{aligned}$$

where M has P and Q as its marginal distributions

Wasserstein robustness

Look at distributionally robust risk

$$R(\theta, \mathcal{P}) := \sup_{P} \{\mathbb{E}_P[\ell(\theta; Z)] \mid P \in \mathcal{P}\}$$

Wasserstein robustness

Look at distributionally robust risk defined for $\rho \geq 0$

$$R(\theta, \rho) := \sup_{P} \left\{ \mathbb{E}_P[\ell(\theta; Z)] \ \text{ s.t. } W_c(P, P_0) \leq \rho \right\}$$

Wasserstein robustness

Look at distributionally robust risk defined for $\rho \geq 0$

$$R(\theta, \rho) := \sup_{P} \{\mathbb{E}_{P}[\ell(\theta; Z)] \;\text{ s.t. } W_c(P, P_0) \leq \rho\}$$

- Allows *changing support* to harder distributions
- Studied in robust optimization literature [Shafieezadeh-Abadeh et al. 15, Esfahani & Kuhn 15, Blanchet and Murthy 16]

Minor issue: Often still NP-hard

A first idea

(Simple) insight: If $\ell(\theta, z)$ is smooth in θ and z, then life gets a bit easier

A first idea

(Simple) insight: If $\ell(\theta, z)$ is smooth in θ and z, then life gets a bit easier

The function

$$\ell_\lambda(\theta; z) := \sup_{\Delta} \left\{ \ell(\theta; z + \Delta) - \frac{\lambda}{2} \|\Delta\|_2^2 \right\}$$

is efficient to compute (and differentiable, etc.) for *large enough* λ

Duality and robustness

Theorem (D., Namkoong, Sinha)

Let P_0 be any distribution on $\mathcal{Z}$ and $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ be any function. Then

$$\sup_{W_c(P,P_0)\le\rho} \mathbb{E}_P[\ell(\theta;Z)] = \inf_{\lambda\ge 0} \left\{ \int \sup_{z'} \left\{ \ell(\theta;z') - \lambda c(z',z) \right\} dP_0(z) + \lambda\rho \right\}$$

$$= \inf_{\lambda\ge 0} \left\{ \mathbb{E}_{P_0}\left[\ell_\lambda(\theta;Z)\right] + \lambda\rho \right\}.$$

Duality and robustness

Theorem (D., Namkoong, Sinha)

Let P_0 be any distribution on $\mathcal{Z}$ and $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ be any function. Then

$$\sup_{W_c(P,P_0)\le\rho} \mathbb{E}_P[\ell(\theta; Z)] = \inf_{\lambda\ge 0} \left\{ \int \sup_{z'} \left\{ \ell(\theta; z') - \lambda c(z', z) \right\} dP_0(z) + \lambda\rho \right\}$$
$$= \inf_{\lambda\ge 0} \left\{ \mathbb{E}_{P_0}\left[\ell_\lambda(\theta; Z)\right] + \lambda\rho \right\}.$$

Idea: Ignore that infimum, pick a large enough λ, and "solve"

$$\operatorname*{minimize}_{\theta} \ \mathbb{E}_{P_0}\left[\ell_\lambda(\theta; Z)\right]$$

Stochastic gradient algorithm

$$\underset{\theta}{\text{minimize}} \; \mathbb{E}_{P_0}[\ell_\lambda(\theta; Z)] = \mathbb{E}_{P_0}\left[\sup_{\Delta}\left\{\ell(\theta; Z + \Delta) - \frac{\lambda}{2}\|\Delta\|_2^2\right\}\right]$$

Repeat:

1. Draw $Z_k \stackrel{\text{iid}}{\sim} P$
2. Compute (approximate) maximizer

$$\widehat{Z}_k \approx \underset{z}{\operatorname{argmax}}\left\{\ell(\theta; z) - \frac{\lambda}{2}\|z - Z_k\|_2^2\right\}$$

3. Update

$$\theta_{k+1} := \theta_k - \alpha_k \nabla_\theta \ell(\theta_k; \widehat{Z}_k)$$

where α_k is a stepsize

Stochastic gradient algorithm

$$\underset{\theta}{\text{minimize}}\ \mathbb{E}_{P_0}[\ell_\lambda(\theta; Z)] = \mathbb{E}_{P_0}\left[\sup_{\Delta}\left\{\ell(\theta; Z + \Delta) - \frac{\lambda}{2}\|\Delta\|_2^2\right\}\right]$$

Repeat:

1. Draw $Z_k \overset{\text{iid}}{\sim} P$
2. Compute (approximate) maximizer

$$\widehat{Z}_k \approx \underset{z}{\text{argmax}}\left\{\ell(\theta; z) - \frac{\lambda}{2}\|z - Z_k\|_2^2\right\}$$

3. Update

$$\theta_{k+1} := \theta_k - \alpha_k \nabla_\theta \ell(\theta_k; \widehat{Z}_k)$$

where α_k is a stepsize

Theorem(ish): This converges with all the typical convergence properties

A certificate of robustness

A desiderata: We would like to certify that any learned θ has robustness properties

A certificate of robustness

A desiderata: We would like to certify that any learned θ has robustness properties

Theorem (D., Namkoong, Sinha 17)

With high probability, for all $\theta \in \Theta$ and uniformly in ρ,

$$\frac{1}{n}\sum_{i=1}^{n}\sup_{\Delta}\left\{\ell(\theta; Z_i + \Delta) - \frac{\lambda}{2}\|\Delta\|_2^2\right\} + \lambda\rho$$

$$\geq \sup_{P : W(P, P_0) \leq \rho}\left\{\mathbb{E}_P\left[\ell(\theta; Z)\right]\right\} - \frac{O(1)}{\sqrt{n}}$$

A certificate of robustness

A desiderata: We would like to certify that any learned θ has robustness properties

Theorem (D., Namkoong, Sinha 17)

With high probability, for all $\theta \in \Theta$

$$\frac{1}{n}\sum_{i=1}^{n}\sup_{\Delta}\left\{\ell(\theta; Z_i + \Delta) - \frac{\lambda}{2}\left\|\Delta\right\|_2^2\right\} + \lambda\widehat{W}(\theta) \geq \sup_{P:W(P,P_0)\leq\widehat{W}(\theta)}\left\{\mathbb{E}_P\left[\ell(\theta; Z)\right]\right\} - \frac{O(1)}{\sqrt{n}}$$

Empirical estimate: get an approximate divergence

$$\widehat{W}(\theta) := \frac{1}{2n}\sum_{i=1}^{n}\left\|\widehat{Z}_i(\theta) - Z_i(\theta)\right\|_2^2$$

where $\widehat{Z}_i = \mathrm{argmax}_z\{\ell(\theta; z) - \frac{\lambda}{2}\left\|z - Z_i\right\|_2^2\}$

Digging into neural networks

- Typically predict with

$$f_\theta(x) = \theta_1^\top \sigma_{\mathrm{relu}}(\theta_2^\top \sigma_{\mathrm{relu}}(\cdots))$$

where

$$\sigma_{\mathrm{relu}}(t) = \min\{1, (t)_+\}$$

Digging into neural networks

- Typically predict with

$$f_\theta(x) = \theta_1^\top \sigma_{\rm relu}(\theta_2^\top \sigma_{\rm relu}(\cdots))$$

where

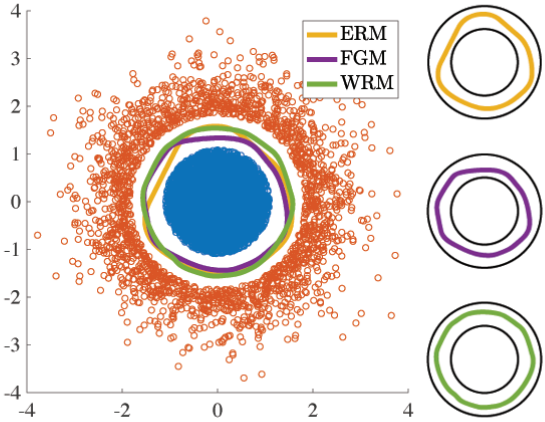
$$\sigma_{\rm relu}(t) = \min\{1, (t)_+\}$$

- Replace $\sigma_{\rm relu}$ with

$$\sigma_{\rm smooth}(t) = \begin{cases} \frac{(t)_+}{2\epsilon}^2 & \text{if } t \le \epsilon \\ t + \frac{\epsilon}{2} & \text{if } \epsilon \le t \le 1 - \epsilon \\ -\frac{(1-t)_+^2}{2\epsilon} + 1 & \text{if } t \ge t - \epsilon \end{cases}$$

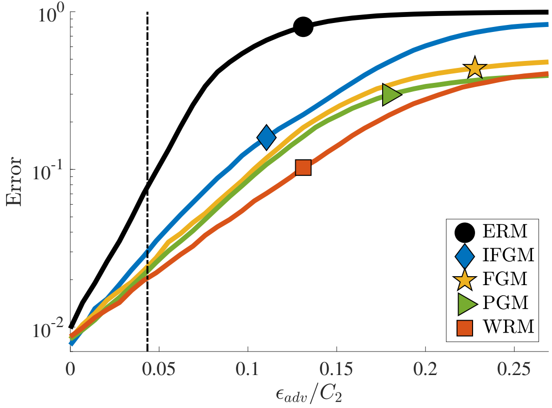
Simple Visualization

$$y = \mathrm{sign}(\|x\|_2 - \sqrt{2})$$



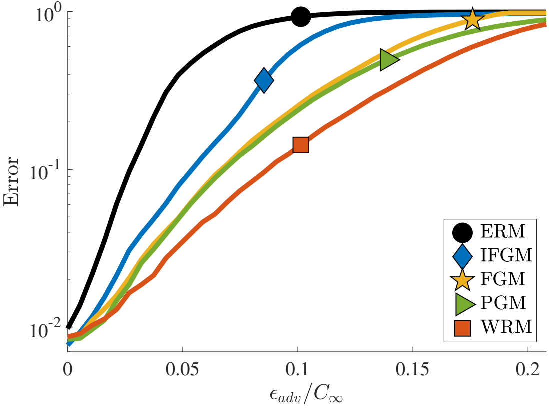
Experimental results: adversarial classification

- MNIST dataset with 3 convolutional layers, fully connected softmax top layer

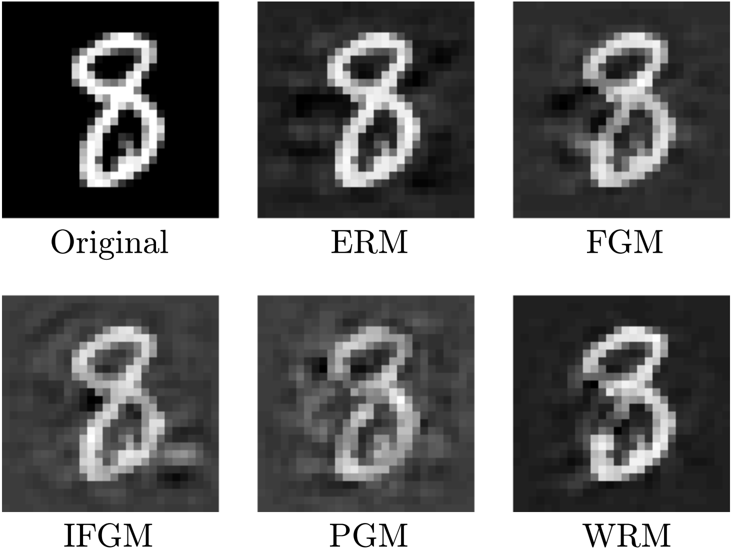


Experimental results: adversarial classification

- MNIST dataset with 3 convolutional layers, fully connected softmax top layer



Reading tea leaves



Reinforcement learning?

References

- H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences.

  In *Advances in Neural Information Processing Systems 29*, 2016

- H. Namkoong and J. C. Duchi. Variance regularization with convex objectives.

  In *Advances in Neural Information Processing Systems 30*, 2017

- A. Sinha, H. Namkoong, and J. C. Duchi. Certifiable distributional robustness with principled adversarial training.

  arXiv:1710.10571 [stat.ML], 2017