Distirbutional robustness, regularizing variance, and adversaries

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

Stanford University

November 2017
We do not want machine-learned systems to fail when they get in the real world
Challenge one: Curly fries

Liking curly fries on Facebook reveals your high IQ

By PHILIPPA WARR
18 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

Liking curly fries on Facebook reveals your high IQ

By PHILIPPA WARR
18 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

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)

[Goodfellow et al. 15]
Challenge three: adversaries

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)

[Goodfellow et al. 15]
Stochastic optimization problems

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

subject to \( \theta \in \Theta \).

**Empirical risk minimization:** Often, solve

\[
\hat{\theta}_n = \arg\min_{\theta \in \Theta} \hat{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; Z_i)
\]
Stochastic optimization problems

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

subject to \( \theta \in \Theta \).

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

**Empirical risk minimization:** Often, solve

\[
\hat{\theta}_n = \arg\min_{\theta \in \Theta} \hat{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; Z_i)
\]
Stochastic optimization problems

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

subject to $\theta \in \Theta$.

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

**Empirical risk minimization:** Often, solve

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \hat{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; Z_i)$$
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

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]
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

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

From empirical Bernstein's inequality, with probability $1 - \delta$

$$R(\theta) \leq \hat{R}_n(\theta) + \sqrt{\frac{2 \text{Var}}{\hat{P}_n(\ell(\theta; X))}} \cdot n + C \log \frac{1}{\delta}$$

Goal: Trade between these automatically and optimally by solving

$$\hat{\theta}_{\text{var}} \in \arg\min_{\theta \in \Theta} \left( \hat{R}_n(\theta) + \sqrt{\frac{2 \text{Var}}{\hat{P}_n(\ell(\theta; X))}} \cdot n \right).$$
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 $1 - \delta$

$$R(\theta) \leq \hat{R}_n(\theta) + \sqrt{\frac{2\text{Var}_{\hat{P}_n}(\ell(\theta; X))}{n}} + \frac{C \log \frac{1}{\delta}}{n}$$
Vignette one: regularization by variance

- Any learning algorithm has \textit{bias} (approximation error) and \textit{variance} (estimation error).
- From empirical Bernstein’s inequality, with probability $1 - \delta$

$$R(\theta) \leq \hat{R}_n(\theta) + \sqrt{2\text{Var}_{\hat{P}_n}(\ell(\theta; X))} \frac{n}{n} + C \log \frac{1}{\delta} \frac{1}{n}$$

\textbf{Goal:} Trade between these automatically and optimally by solving

$$\hat{\theta}^{\text{var}} \in \arg\min_{\theta \in \Theta} \left\{ \hat{R}_n(\theta) + \sqrt{2\text{Var}_{\hat{P}_n}(\ell(\theta; X))} \frac{n}{n} \right\}.$$
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:

\[
\min_{\theta \in \Theta} R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)]
\]
Robust ERM

Goal:

\[
\minimize_{\theta \in \Theta} R(\theta) = \mathbb{E}_{P_0}\left[\ell(\theta; X)\right]
\]

Solve empirical risk minimization problem

\[
\minimize_{\theta \in \Theta} \sum_{i=1}^{n} \frac{1}{n} \ell(\theta; X_i)
\]
Robust ERM

Goal:

\[
\min_{\theta \in \Theta} \quad R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)]
\]

Solve sample average optimization problem

\[
\min_{\theta \in \Theta} \quad \sum_{i=1}^{n} \frac{1}{n} \ell(\theta; X_i)
\]
Robust ERM

Goal:

\[
\min_{\theta \in \Theta} \quad R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)]
\]

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

\[
\min_{\theta \in \Theta} \quad \sup_{p \in \mathcal{P}_{n,\rho}} \sum_{i=1}^{n} p_i \ell(\theta; X_i)
\]

where \( \mathcal{P}_{n,\rho} \) is some appropriately chosen set of vectors
Robust ERM

**Goal:**

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

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

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

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,

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

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,

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

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

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

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

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

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

for some \( \rho > 0 \), where \( D(P\|Q) = \int \frac{p}{q} - 1 \right)^2 q \)

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

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

**Nice properties:**

- Convex optimization problem
- Efficient solution methods [D. & Namkoong NIPS 16]
Robust Optimization = Variance Regularization

Theorem (D. & Namkoong)

Assume that $\ell$ is bounded over the space of decision vectors $\theta$. Then

$$R_n(\theta; P_{n,\rho}) = \hat{R}_n(\theta) + \sqrt{\frac{2\rho \text{Var}_{\hat{P}_n}(\ell(\theta; X))}{n}} + O(\rho/n).$$
Robust Optimization = Variance Regularization

Theorem (D. & Namkoong)

Assume that $\ell$ is bounded over the space of decision vectors $\theta$. Then

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

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

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

Choose $\hat{\theta}^{\text{rob}}$ to minimize robust empirical risk

$$R_n(\hat{\theta}^{\text{rob}}, P_n, \rho) = \min_{\theta \in \Theta} \max_{P \ll \hat{P}_n} \left\{ \mathbb{E}_P[\ell(\theta; X)] : D_{\chi^2}(P \| \hat{P}_n) \leq \frac{\rho}{n} \right\}.$$
Optimal bias variance tradeoff

Choose $\hat{\theta}^{\text{rob}}$ to minimize robust empirical risk

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

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

Choose $\hat{\theta}^{\text{rob}}$ to minimize robust empirical risk

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

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

Theorem (D. & Namkoong 17)

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

$$R(\hat{\theta}^{\text{rob}}) \leq R_n(\hat{\theta}^{\text{rob}}, \mathcal{P}_n, \rho) + \frac{cMR}{n} \rho$$

optimality certificate

$$\leq \min_{\theta \in \Theta} \left\{ R(\theta) + 2\sqrt{\frac{2\rho \text{Var}(\ell(\theta, \xi))}{n}} \right\} + \frac{cMR}{n} \rho$$

optimal tradeoff

for some universal constant $c > 0$. 
Experiment: Reuters Corpus (multi-label)

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

\{Corporate, Economics, Government, Markets\}
Experiment: Reuters Corpus (multi-label)

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

\[
\left\{ \text{Corporate, Economics, Government, Markets} \right\}
\]

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

\( d = 47, 236, n = 804, 414 \cdot 10 \)

10-fold cross-validation.

**Table: Reuters Number of Examples**

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corporate</td>
<td>381,327</td>
</tr>
<tr>
<td>Economics</td>
<td>119,920</td>
</tr>
<tr>
<td>Government</td>
<td>239,267</td>
</tr>
<tr>
<td>Markets</td>
<td>204,820</td>
</tr>
</tbody>
</table>
Problem: Classify documents as a subset of the 4 categories:

\[
\left\{ \text{Corporate, Economics, Government, Markets} \right\}
\]

- 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 \( \Theta = \{ \theta \in \mathbb{R}^d : \|\theta\|_1 \leq 1000 \} \).
**Problem:** Classify documents as a **subset** of the 4 categories:

\[
\{\text{Corporate, Economics, Government, Markets}\}
\]

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

- **Loss** \(\ell(\theta_j, (x, y)) = \log(1 + e^{-yx^\top\theta_j})\) for each \(j = 1, \ldots, 4\) and \(\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq 1000\}\).

- \(d = 47,236\), \(n = 804,414\). 10-fold cross-validation.
**Experiment: Reuters Corpus (multi-label)**

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

\[
\{ \text{Corporate, Economics, Government, Markets} \}
\]

- **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 \( \Theta = \{ \theta \in \mathbb{R}^d : \|\theta\|_1 \leq 1000 \} \).

- \( d = 47,236, n = 804,414. \) 10-fold cross-validation.

<table>
<thead>
<tr>
<th>Corporate</th>
<th>Economics</th>
<th>Government</th>
<th>Markets</th>
</tr>
</thead>
<tbody>
<tr>
<td>381,327</td>
<td>119,920</td>
<td>239,267</td>
<td>204,820</td>
</tr>
</tbody>
</table>
## Table: Reuters Corpus (%)

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>Precision train</th>
<th>Precision test</th>
<th>Recall train</th>
<th>Recall test</th>
<th>Corporate train</th>
<th>Corporate test</th>
<th>Economics train</th>
<th>Economics test</th>
</tr>
</thead>
<tbody>
<tr>
<td>erm</td>
<td>92.72</td>
<td>92.7</td>
<td>90.97</td>
<td>90.96</td>
<td>90.2</td>
<td>90.25</td>
<td>67.53</td>
<td>67.56</td>
</tr>
<tr>
<td>10000</td>
<td>94.17</td>
<td>94.16</td>
<td>93.46</td>
<td>93.44</td>
<td>92.65</td>
<td>92.71</td>
<td>76.79</td>
<td>76.78</td>
</tr>
</tbody>
</table>
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

“panda”
57.7% confidence

+ ε

= “gibbon”
99.3% confidence
A type of robustess

**Robust optimization:** instead of $\ell$, look at robust loss

$$\ell_\epsilon(\theta; z) := \sup_{\|\Delta\| \leq \epsilon} \ell(\theta; z + \Delta)$$
A type of robustness

**Robust optimization:** instead of $\ell$, look at robust loss

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

- 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, $f_\theta(x) = \theta^T_1 \sigma_{\text{relu}}(\theta^T_2 \sigma_{\text{relu}}(\cdots))$ and is is NP-hard to compute $\sup_{\Delta} \ell(f_\theta(x + \Delta))$
A type of robustness

**Robust optimization:** instead of $\ell$, look at robust loss

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

- 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,

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

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*

$$W_c(P, Q) := \inf_M \int c(z_1, z_2) dM(z_1, z_2)$$

$$= \sup_f \left\{ \int f(z)(dP(z) - dQ(z)) \mid f(x) - f(z) \leq c(x, z) \right\}$$

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 $\theta$ and $z$, then life gets a bit easier
A first idea

(Simple) insight: If $\ell(\theta, z)$ is smooth in $\theta$ 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* $\lambda$
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

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

$$= \inf_{\lambda \geq 0} \{ \mathbb{E}_{P_0} [\ell_\lambda(\theta; Z)] + \lambda \rho \} .$$
Duality and robustness

**Theorem (D., Namkoong, Sinha)**

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

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

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

**Idea:** Ignore that infimum, pick a large enough $\lambda$, and “solve”

$$
\minimize_{\theta} \mathbb{E}_{P_0}[\ell_{\lambda}(\theta; Z)]
$$
Stochastic gradient algorithm

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

Repeat:

1. Draw \( Z_k \overset{iid}{\sim} P \)

2. Compute (approximate) maximizer

\[
\hat{Z}_k \approx \arg\max_z \left\{\ell(\theta; z) - \frac{\lambda}{2} \|z - Z_k\|^2\right\}
\]

3. Update

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

where \( \alpha_k \) is a stepsize
Stochastic gradient algorithm

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

Repeat:

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

\[
\hat{Z}_k \approx \arg\max_{z} \left\{ \ell(\theta; z) - \frac{\lambda}{2} \| z - Z_k \|_2^2 \right\}
\]
3. Update

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

where \( \alpha_k \) 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 $\theta$ has robustness properties.
A certificate of robustness

A desiderata: We would like to certify that any learned $\theta$ has robustness properties

Theorem (D., Namkoong, Sinha 17)

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

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

A desiderata: We would like to certify that any learned $\theta$ has robustness properties

Theorem (D., Namkoong, Sinha 17)

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

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

Empirical estimate: get an approximate divergence

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

where $\hat{Z}_i = \arg\max_z \{ \ell(\theta; z) - \frac{\lambda}{2} \| z - Z_i \|_2^2 \}$
Digging into neural networks

- Typically predict with

\[ f_\theta(x) = \theta_1^T \sigma_{\text{relu}}(\theta_2^T \sigma_{\text{relu}}(\cdots)) \]

where

\[ \sigma_{\text{relu}}(t) = \min\{1, (t)_+\} \]
Digging into neural networks

- Typically predict with

\[ f_\theta(x) = \theta_1^\top \sigma_{\text{relu}}(\theta_2^\top \sigma_{\text{relu}}(\cdots)) \]

where

\[ \sigma_{\text{relu}}(t) = \min\{1, (t)_+\} \]

- Replace \( \sigma_{\text{relu}} \) with

\[
\sigma_{\text{smooth}}(t) = \begin{cases} 
(t)_+^2 
& \text{if } t \leq \epsilon \\
\frac{t + \epsilon}{2} 
& \text{if } \epsilon \leq t \leq 1 - \epsilon \\
-\frac{(1-t)^2}{2\epsilon} + 1 
& \text{if } t \geq t - \epsilon
\end{cases}
\]
Simple Visualization

\[ y = \text{sign}(\|x\|_2 - \sqrt{2}) \]
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

Original  ERM  FGM

IFGM  PGM  WRM
Reinforcement learning?
