Precise Tradeoffs in [and asymptotics of] Adversarial Training for Linear Regression

Behrad Moniri Samar Hadou
University of Pennsylvania

STAT 972 Final Presentation

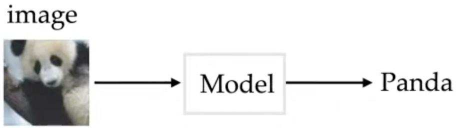
Precise Tradeoffs in Adversarial Training for Linear Regression

Adel Javanmard, Mahdi Soltanolkotabi, Hamed Hassani

Conference on Learning Theory (COLT), 2020.

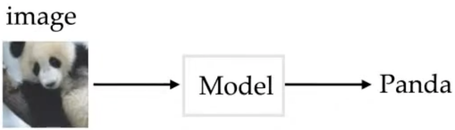
Motivation

- Modern Neural Networks are very good tools for prediction.



Motivation

- Modern Neural Networks are very good tools for prediction.



- Modern Neural Networks are not robust to adversarial attacks.

Notations

- Data: $(\mathbf{x}_i, y_i) \sim \mathbb{P}(\mathbb{R}^d, \mathbb{R})$
- Model: $f_{\boldsymbol{\theta}}(\cdot) : \mathbb{R}^d \to \mathbb{R}$
- Loss Function: $\ell(\boldsymbol{\theta}, \mathbf{x}, y) = (y - f_\theta(\mathbf{x}))^2$

Traditional Supervised learning

Traditional Supervised learning

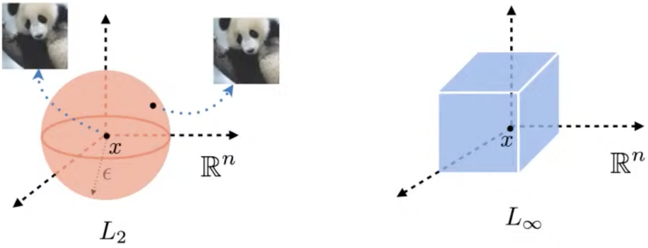
- Population Loss:

$$SR(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x},y}[\ell(\boldsymbol{\theta}, \mathbf{x}, y)]$$

- Empirical Risk Minimization:

$$\widehat{\boldsymbol{\theta}}_{ERM} = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \mathbf{x}_i, y_i)$$

$L_p,\ p \geq 1$: Simplest Possible Geometry



Robust supervised learning

- Adversarial Loss:

$$AR(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x},y}\left[\max_{||\boldsymbol{\delta}||_2 \leq \varepsilon} \ell(\boldsymbol{\theta}, \mathbf{x} + \boldsymbol{\delta}, y)\right]$$

- Adverasrial Training:

$$\widehat{\boldsymbol{\theta}^{\varepsilon}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \max_{||\boldsymbol{\delta}_i||_2 \leq \varepsilon} \ell(\boldsymbol{\theta}, \mathbf{x}_i + \boldsymbol{\delta}_i, y_i)$$

Observations

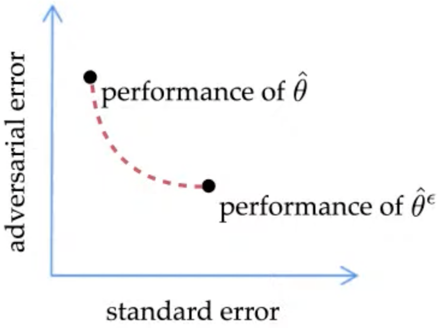
- $AR(\widehat{\boldsymbol{\theta}}_{ERM})$ is large and $AR(\widehat{\boldsymbol{\theta}^{\varepsilon}})$ is much smaller.

Observations

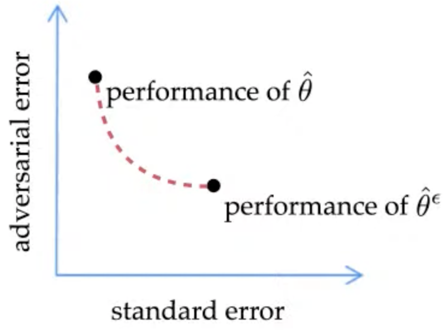
- $AR(\widehat{\boldsymbol{\theta}}_{ERM})$ is large and $AR(\widehat{\boldsymbol{\theta}^{\varepsilon}})$ is much smaller.
- $SR(\widehat{\boldsymbol{\theta}^{\varepsilon}})$ is larger than $SR(\widehat{\boldsymbol{\theta}}_{ERM})$.

Observations

- $AR(\widehat{\boldsymbol{\theta}}_{ERM})$ is large and $AR(\widehat{\boldsymbol{\theta}^{\varepsilon}})$ is much smaller.
- $SR(\widehat{\boldsymbol{\theta}^{\varepsilon}})$ is larger than $SR(\widehat{\boldsymbol{\theta}}_{ERM})$.



- $AR(\widehat{\boldsymbol{\theta}}_{ERM})$ is large and $AR(\widehat{\boldsymbol{\theta}^{\varepsilon}})$ is much smaller.
- $SR(\widehat{\boldsymbol{\theta}^{\varepsilon}})$ is larger than $SR(\widehat{\boldsymbol{\theta}}_{ERM})$.



Questions:

- Is there a fundamental tradeoff between SR and AR?
- How can we algorithmically achieve this tradeoff?

Linear Regression: Fundamental Tradeoffs

Gaussian Linear Regression

- We consider standard gaussian linear regression with

$$y_i = \langle \mathbf{x}_i, \boldsymbol{\theta}_0 \rangle + w_i \quad \text{where} \quad \mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p) \quad w_i \sim \mathcal{N}\left(0, \sigma_0^2\right)$$

  for $1 \leq i \leq n$.
- We also focus on training linear models of the form $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle$

Gaussian Linear Regression

- We consider standard gaussian linear regression with

$$y_i = \langle \mathbf{x}_i, \boldsymbol{\theta}_0 \rangle + w_i \quad \text{where} \quad \mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p) \quad w_i \sim \mathcal{N}(0, \sigma_0^2)$$

  for $1 \leq i \leq n$.
- We also focus on training linear models of the form $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle$
- We can write:

$$\mathrm{SR}(\widehat{\boldsymbol{\theta}}) := \mathbb{E}\left[(y - \langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}} \rangle)^2 \right]$$

Gaussian Linear Regression

- We consider standard gaussian linear regression with

$$y_i = \langle \mathbf{x}_i, \boldsymbol{\theta}_0 \rangle + w_i \quad \text{where} \quad \mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p) \quad w_i \sim \mathcal{N}(0, \sigma_0^2)$$

  for $1 \leq i \leq n$.
- We also focus on training linear models of the form $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle$
- We can write:

$$\mathrm{SR}(\widehat{\boldsymbol{\theta}}) := \mathbb{E}\left[(y - \langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}} \rangle)^2 \right] = \sigma_0^2 + \left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \right\|_{\ell_2}^2,$$

Gaussian Linear Regression

- We consider standard gaussian linear regression with

$$y_i = \langle \mathbf{x}_i, \boldsymbol{\theta}_0 \rangle + w_i \quad \text{where} \quad \mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p) \quad w_i \sim \mathcal{N}(0, \sigma_0^2)$$

  for $1 \leq i \leq n$.
- We also focus on training linear models of the form $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle$
- We can write:

$$\mathrm{SR}(\widehat{\boldsymbol{\theta}}) := \mathbb{E}\left[(y - \langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}} \rangle)^2\right] = \sigma_0^2 + \left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \right\|_{\ell_2}^2,$$

$$\mathrm{AR}(\widehat{\boldsymbol{\theta}}) := \mathbb{E}\left[\max_{\|\boldsymbol{\delta}\|_{\ell_2} \leq \varepsilon} (y - \langle \boldsymbol{x} + \boldsymbol{\delta}, \widehat{\boldsymbol{\theta}} \rangle)^2\right]$$

Gaussian Linear Regression

- We consider standard gaussian linear regression with

$$y_i = \langle \mathbf{x}_i, \boldsymbol{\theta}_0 \rangle + w_i \quad \text{where} \quad \mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p) \quad w_i \sim \mathcal{N}(0, \sigma_0^2)$$

  for $1 \leq i \leq n$.
- We also focus on training linear models of the form $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle$
- We can write:

$$\mathrm{SR}(\widehat{\boldsymbol{\theta}}) := \mathbb{E}\left[(y - \langle \boldsymbol{x}, \widehat{\boldsymbol{\theta}} \rangle)^2\right] = \sigma_0^2 + \left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \right\|_{\ell_2}^2,$$

$$\mathrm{AR}(\widehat{\boldsymbol{\theta}}) := \mathbb{E}\left[\max_{\|\boldsymbol{\delta}\|_{\ell_2} \leq \varepsilon} (y - \langle \boldsymbol{x} + \boldsymbol{\delta}, \widehat{\boldsymbol{\theta}} \rangle)^2\right]$$

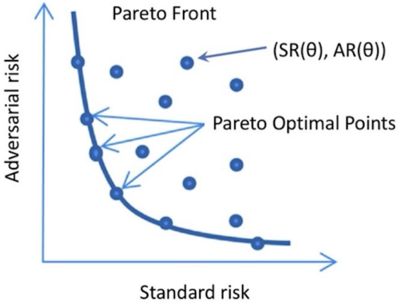
$$= \left(\sigma_0^2 + \left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \right\|_{\ell_2}^2 + \varepsilon^2 \|\widehat{\boldsymbol{\theta}}\|_{\ell_2}^2\right.$$

$$\left. + 2\sqrt{\frac{2}{\pi}} \varepsilon \, \|\widehat{\boldsymbol{\theta}}\|_{\ell_2} \left(\sigma_0^2 + \left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \right\|_{\ell_2}^2\right)^{1/2}\right.$$

Pareto Optimal

Pareto-optimal points are the intersection points of the region with the supporting lines:

$$\boldsymbol{\theta}^{\lambda} := \arg\min_{\boldsymbol{\theta}} \lambda\, SR(\boldsymbol{\theta}) + AR(\boldsymbol{\theta})$$



Pareto Optimal Curve

The solution $\boldsymbol{\theta}^\lambda$ is given by

$$\boldsymbol{\theta}^\lambda = \left(1 + \gamma_0^\lambda\right)^{-1} \boldsymbol{\theta}_0,$$

with γ_0^λ the fixed point of the following two equations:

$$\gamma_0^\lambda = \frac{\varepsilon_{\text{test}}^2 + \sqrt{\frac{2}{\pi}}\varepsilon_{\text{test}}\, A^\lambda}{1 + \lambda + \sqrt{\frac{2}{\pi}}\frac{\varepsilon_{\text{test}}}{A^\lambda}}$$

$$A^\lambda = \frac{1}{\|\boldsymbol{\theta}_0\|_{\ell_2}} \left(\left(1 + \gamma_0^\lambda\right)^2 \sigma_0^2 + \left(\gamma_0^\lambda\right)^2 \|\boldsymbol{\theta}_0\|_{\ell_2}^2\right)^{1/2}.$$

Linear Regression: Algorithmic Tradeoffs

Algorithmic Tradeoffs

- Consider a class of estimators $\left\{\widehat{\boldsymbol{\theta}^{\varepsilon}} : \varepsilon \geq 0\right\}$ constructed via the following saddle point problem:

$$\widehat{\boldsymbol{\theta}^{\varepsilon}} \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \max_{\|\boldsymbol{\delta}_i\|\leq\varepsilon} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \langle \boldsymbol{x}_i + \boldsymbol{\delta}_i, \boldsymbol{\theta}\rangle\right)^2$$

- Can one of these (adversarially trained) estimators achieve the optimal tradeoff?

Algorithmic Tradeoffs

- Consider a class of estimators $\left\{\widehat{\boldsymbol{\theta}^{\varepsilon}} : \varepsilon \geq 0\right\}$ constructed via the following saddle point problem:

$$\widehat{\boldsymbol{\theta}^{\varepsilon}} \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \max_{\|\boldsymbol{\delta}_i\|\leq\varepsilon} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \langle \boldsymbol{x}_i + \boldsymbol{\delta}_i, \boldsymbol{\theta}\rangle\right)^2$$

- Can one of these (adversarially trained) estimators achieve the optimal tradeoff?
- The answer is in the limit.

Algorithmic Tradeoffs

- Assume that $n \to \infty$, $d \to \infty$ and $n/d \to \delta$.

Algorithmic Tradeoffs

- Assume that $n \to \infty$, $d \to \infty$ and $n/d \to \delta$.
- Can we find an asymptotic expression for $AR(\widehat{\boldsymbol{\theta}^{\varepsilon}})$ and $SR(\widehat{\boldsymbol{\theta}^{\varepsilon}})$?

- Assume that $n \to \infty$, $d \to \infty$ and $n/d \to \delta$.
- Can we find an asymptotic expression for $AR(\widehat{\boldsymbol{\theta}^{\varepsilon}})$ and $SR(\widehat{\boldsymbol{\theta}^{\varepsilon}})$?
- Note that these expression can both be written in terms of only $\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|_{\ell_2}^2$ and $\left\|\widehat{\boldsymbol{\theta}}\right\|_{\ell_2}^2$.

- Assume that $n \to \infty$, $d \to \infty$ and $n/d \to \delta$.
- Can we find an asymptotic expression for $AR(\widehat{\boldsymbol{\theta}^{\varepsilon}})$ and $SR(\widehat{\boldsymbol{\theta}^{\varepsilon}})$?
- Note that these expression can both be written in terms of only $\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|_{\ell_2}^2$ and $\left\|\widehat{\boldsymbol{\theta}}\right\|_{\ell_2}^2$.
- To do this, we will use Convex Gaussian Minmax Theorem (CGMT).

- Standard linear regression has widely been studied in the proportional limit:

History

- Standard linear regression has widely been studied in the proportional limit:
  - The underparameterized regime:

    [1] Antonia M. Tulino and Sergio Verdu. Random matrix theory and wireless communications. Foundations and Trends in Communications and Information Theory, 2004.

- Standard linear regression has widely been studied in the proportional limit:
  - The underparameterized regime:

    [1] Antonia M. Tulino and Sergio Verdu. Random matrix theory and wireless communications. Foundations and Trends in Communications and Information Theory, 2004.
  - General case:

    [2] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: ridge regression and classification. Annals of Statistics, 2018.

    [3] Trevor Hastie, Andrea Montanari, Saharon Rosset, Ryan J. Tibshirani. Surprises in High-Dimensional Ridgeless Least Squares Interpolation, Annals of Statistics, 2022.

- Standard linear regression has widely been studied in the proportional limit:
  - The underparameterized regime:

    [1] Antonia M. Tulino and Sergio Verdu. Random matrix theory and wireless communications. Foundations and Trends in Communications and Information Theory, 2004.

  - General case:

    [2] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: ridge regression and classification. Annals of Statistics, 2018.

    [3] Trevor Hastie, Andrea Montanari, Saharon Rosset, Ryan J. Tibshirani. Surprises in High-Dimensional Ridgeless Least Squares Interpolation, Annals of Statistics, 2022.

They all use the Marchenko-Pastur limit. Here, we cannot use that because there is no closed form for the estimator.

Theorem (Convex Gaussian Min-Max Theorem (CGMT) – informal)

For $\mathbf{X}$ with i.i.d standard normal entries and $\psi(\cdot,\cdot)$ a convex-concave function, define

$$\Phi(\mathbf{X}) := \min_{\mathbf{z}} \max_{\mathbf{u}} \ \mathbf{u}^T \mathbf{X} \mathbf{z} + \psi(\mathbf{z}, \mathbf{u}) \quad (PO)$$

$$\phi(\mathbf{g}, \mathbf{h}) := \min_{\mathbf{z}} \max_{\mathbf{u}} \ \|\mathbf{z}\| \mathbf{g}^T \mathbf{u} + \|\mathbf{u}\| \mathbf{h}^T \mathbf{z} + \psi(\mathbf{z}, \mathbf{u}) \quad (AO)$$

We have $\Phi(\mathbf{X}) \approx \phi(\mathbf{g}, \mathbf{h})$, in which $\mathbf{g}, \mathbf{h}$ are standard Gaussian random vectors. Also the norms of the solutions for both optimization problems are equal.

[Thrampoulidis, Oymak, and Hassibi; 2016 & 2018]

Finding the asymptotic expressions for $AR(\widehat{\boldsymbol{\theta}^{\varepsilon}})$ and $SR(\widehat{\boldsymbol{\theta}^{\varepsilon}})$:

Finding the asymptotic expressions for $AR(\widehat{\boldsymbol{\theta}^{\varepsilon}})$ and $SR(\widehat{\boldsymbol{\theta}^{\varepsilon}})$:

- **Step 1**: Adversarial loss has a closed form:

$$\widehat{\boldsymbol{\theta}^{\varepsilon}} \in \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \max_{\|\boldsymbol{\delta}_i\| \leq \varepsilon} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \langle \boldsymbol{x}_i + \boldsymbol{\delta}_i, \boldsymbol{\theta} \rangle\right)^2$$

$$= \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{2n} \sum_{i=1}^{n} \left(|y_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle| + \varepsilon \|\boldsymbol{\theta}\|_{\ell_2}\right)^2$$

- **Step 2**: Write in the form of a Primary Optimization.

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2n}\sum_{i=1}^{n}\left(|y_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle| + \varepsilon\|\boldsymbol{\theta}\|_{\ell_2}\right)^2$$

- **Step 2**: Write in the form of a Primary Optimization.

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2n}\sum_{i=1}^{n}\left(|y_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle| + \varepsilon\|\boldsymbol{\theta}\|_{\ell_2}\right)^2$$

$$= \min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2n}\sum_{i=1}^{n}\left(|w_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta} - \boldsymbol{\theta}_0\rangle| + \varepsilon\|\boldsymbol{\theta}\|_{\ell_2}\right)^2$$

- **Step 2**: Write in the form of a Primary Optimization.

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2n}\sum_{i=1}^{n}\left(|y_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle| + \varepsilon\|\boldsymbol{\theta}\|_{\ell_2}\right)^2$$

$$= \min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2n}\sum_{i=1}^{n}\left(|w_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta} - \boldsymbol{\theta}_0\rangle| + \varepsilon\|\boldsymbol{\theta}\|_{\ell_2}\right)^2$$

$$= \min_{\boldsymbol{z}\in\mathbb{R}^d, \mathbf{v}\in\mathbb{R}^n} \frac{1}{2n}\sum_{i=1}^{n}\left(|\mathbf{v}_i| + \varepsilon\|\mathbf{z} + \boldsymbol{\theta}_{\mathbf{0}}\|_{\ell_2}\right)^2 \qquad \text{s.t. } \mathbf{v} = \mathbf{w} - \mathbf{X}\mathbf{z}$$

- **Step 2**: Write in the form of a Primary Optimization.

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2n}\sum_{i=1}^{n}\left(|y_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle| + \varepsilon\|\boldsymbol{\theta}\|_{\ell_2}\right)^2$$

$$= \min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2n}\sum_{i=1}^{n}\left(|w_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta} - \boldsymbol{\theta}_0\rangle| + \varepsilon\|\boldsymbol{\theta}\|_{\ell_2}\right)^2$$

$$= \min_{\boldsymbol{z}\in\mathbb{R}^d, \mathbf{v}\in\mathbb{R}^n} \frac{1}{2n}\sum_{i=1}^{n}\left(|\mathbf{v}_i| + \varepsilon\|\mathbf{z} + \boldsymbol{\theta}_{\mathbf{0}}\|_{\ell_2}\right)^2 \qquad \text{s.t. } \mathbf{v} = \mathbf{w} - \mathbf{X}\mathbf{z}$$

$$= \min_{\boldsymbol{z}\in\mathbb{R}^d, \mathbf{v}\in\mathbb{R}^n} \frac{1}{2n}\left(||\mathbf{v}||_{\ell_2}^2 + n\varepsilon^2||\mathbf{z} + \boldsymbol{\theta}_0||_{\ell_2}^2 + 2\varepsilon\|\mathbf{z} + \boldsymbol{\theta}_{\mathbf{0}}\|_{\ell_2}||\mathbf{v}||_{\ell_1}\right)$$

- **Step 2**: Write in the form of a Primary Optimization.

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2n}\sum_{i=1}^{n}\left(|y_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle| + \varepsilon\|\boldsymbol{\theta}\|_{\ell_2}\right)^2$$

$$= \min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2n}\sum_{i=1}^{n}\left(|w_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta} - \boldsymbol{\theta}_0\rangle| + \varepsilon\|\boldsymbol{\theta}\|_{\ell_2}\right)^2$$

$$= \min_{\boldsymbol{z}\in\mathbb{R}^d, \mathbf{v}\in\mathbb{R}^n} \frac{1}{2n}\sum_{i=1}^{n}\left(|\mathbf{v}_i| + \varepsilon\|\mathbf{z} + \boldsymbol{\theta_0}\|_{\ell_2}\right)^2 \qquad \text{s.t. } \mathbf{v} = \mathbf{w} - \mathbf{X}\mathbf{z}$$

$$= \min_{\boldsymbol{z}\in\mathbb{R}^d, \mathbf{v}\in\mathbb{R}^n} \frac{1}{2n}\left(||\mathbf{v}||_{\ell_2}^2 + n\varepsilon^2||\mathbf{z} + \boldsymbol{\theta}_0||_{\ell_2}^2 + 2\varepsilon\|\mathbf{z} + \boldsymbol{\theta_0}\|_{\ell_2}||\mathbf{v}||_{\ell_1}\right)$$

$$= \min_{\boldsymbol{z}\in\mathbb{R}^d, \mathbf{v}\in\mathbb{R}^n} \max_{\mathbf{u}\in\mathbb{R}^n} \frac{1}{2n}\left(||\mathbf{v}||_{\ell_2}^2 + n\varepsilon^2||\mathbf{z} + \boldsymbol{\theta}_0||_{\ell_2}^2 + 2\varepsilon\|\mathbf{z} + \boldsymbol{\theta_0}\|_{\ell_2}||\mathbf{v}||_{\ell_1}\right) + \frac{1}{2n}\mathbf{u}^\top(\mathbf{v} - \mathbf{w} + \mathbf{X}\mathbf{z})$$

- CGMT PO and AO forms:

$$\Phi(\mathbf{X}) := \min_{\mathbf{z}} \max_{\mathbf{u}} \; \mathbf{u}^T \mathbf{X} \mathbf{z} + \psi(\mathbf{z}, \mathbf{u}) \quad (PO)$$

$$\phi(\mathbf{g}, \mathbf{h}) := \min_{\mathbf{z}} \max_{\mathbf{u}} \; \|\mathbf{z}\| \mathbf{g}^T \mathbf{u} + \|\mathbf{u}\| \mathbf{h}^T \mathbf{z} + \psi(\mathbf{z}, \mathbf{u}) \quad (AO)$$

- CGMT PO and AO forms:

$$\Phi(\mathbf{X}) := \min_{\mathbf{z}} \max_{\mathbf{u}} \ \mathbf{u}^T \mathbf{X} \mathbf{z} + \psi(\mathbf{z}, \mathbf{u}) \quad (PO)$$

$$\phi(\mathbf{g}, \mathbf{h}) := \min_{\mathbf{z}} \max_{\mathbf{u}} \ \|\mathbf{z}\| \mathbf{g}^T \mathbf{u} + \|\mathbf{u}\| \mathbf{h}^T \mathbf{z} + \psi(\mathbf{z}, \mathbf{u}) \quad (AO)$$

- Primary Optimization:

$$\min_{\mathbf{z}\in\mathbb{R}^d, \mathbf{v}\in\mathbb{R}^n} \max_{\mathbf{u}\in\mathbb{R}^n} \quad \frac{1}{2n} \left(||\mathbf{v}||_{\ell_2}^2 + n\varepsilon^2 ||\mathbf{z} + \boldsymbol{\theta}_0||_{\ell_2}^2 + 2\varepsilon \|\mathbf{z} + \boldsymbol{\theta}_0\|_{\ell_2} ||\mathbf{v}||_{\ell_1} \right) + \frac{1}{2n} \mathbf{u}^\top (\mathbf{v} - \mathbf{w} + \mathbf{X}\mathbf{z})$$

Algorithmic Tradeoffs

- CGMT PO and AO forms:

$$\Phi(\mathbf{X}) := \min_{\mathbf{z}} \max_{\mathbf{u}} \; \mathbf{u}^T \mathbf{X} \mathbf{z} + \psi(\mathbf{z}, \mathbf{u}) \quad (PO)$$

$$\phi(\mathbf{g}, \mathbf{h}) := \min_{\mathbf{z}} \max_{\mathbf{u}} \; \|\mathbf{z}\| \mathbf{g}^T \mathbf{u} + \|\mathbf{u}\| \mathbf{h}^T \mathbf{z} + \psi(\mathbf{z}, \mathbf{u}) \quad (AO)$$

- Primary Optimization:

$$\min_{\boldsymbol{z} \in \mathbb{R}^d, \mathbf{v} \in \mathbb{R}^n} \max_{\mathbf{u} \in \mathbb{R}^n} \quad \frac{1}{2n} \left(||\mathbf{v}||_{\ell_2}^2 + n\varepsilon^2 ||\mathbf{z} + \boldsymbol{\theta}_0||_{\ell_2}^2 + 2\varepsilon \|\mathbf{z} + \boldsymbol{\theta}_\mathbf{0}\|_{\ell_2} ||\mathbf{v}||_{\ell_1} \right) + \frac{1}{2n} \mathbf{u}^\top (\mathbf{v} - \mathbf{w} + \mathbf{X}\mathbf{z})$$

- Hence, the Auxiliary Optimization is:

$$\min_{\boldsymbol{z} \in \mathbb{R}^d, \mathbf{v} \in \mathbb{R}^n} \max_{\mathbf{u} \in \mathbb{R}^n} \quad \frac{1}{2n} \left(\|\boldsymbol{z}\|_{\ell_2} \boldsymbol{g}^T \boldsymbol{u} + \|\boldsymbol{u}\|_{\ell_2} \boldsymbol{h}^T \boldsymbol{z} - \boldsymbol{u}^T \boldsymbol{\omega} + \boldsymbol{u}^T \boldsymbol{v} \right) + \frac{1}{2n} \left(||\mathbf{v}||_{\ell_2}^2 + n\varepsilon^2 ||\mathbf{z} + \boldsymbol{\theta}_0||_{\ell_2}^2 + 2\varepsilon \|\mathbf{z} + \boldsymbol{\theta}_\mathbf{0}\|_{\ell_2} ||\mathbf{v}||_{\ell_1} \right).$$

- **Step 3**: Study the Auxiliary Optimization

$$\min_{\boldsymbol{z}\in\mathbb{R}^d,\mathbf{v}\in\mathbb{R}^n}\max_{\mathbf{u}\in\mathbb{R}^n}\quad \frac{1}{2n}\left(\|\boldsymbol{z}\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{u}+\|\boldsymbol{u}\|_{\ell_2}\boldsymbol{h}^T\boldsymbol{z}-\boldsymbol{u}^T\omega+\boldsymbol{u}^T\boldsymbol{v}\right)$$
$$+\frac{1}{2n}\left(||\mathbf{v}||_{\ell_2}^2+n\varepsilon^2||\mathbf{z}+\boldsymbol{\theta}_0||_{\ell_2}^2+2\varepsilon\|\mathbf{z}+\boldsymbol{\theta}_\mathbf{0}\|_{\ell_2}||\mathbf{v}||_{\ell_1}\right).$$

- **Scalarization**: Starting with the maximization over $\mathbf{u}$, let $\mathbf{u}=\beta\tilde{\mathbf{u}}$.

$$\max_{\mathbf{u}\in\mathbb{R}^n}\quad \frac{1}{2n}\left(\|\boldsymbol{z}\|_{\ell_2}\boldsymbol{g}^T\boldsymbol{u}+\|\boldsymbol{u}\|_{\ell_2}\boldsymbol{h}^T\boldsymbol{z}-\boldsymbol{u}^T\omega+\boldsymbol{u}^T\boldsymbol{v}\right)$$
$$=\max_{\beta}\quad\frac{1}{2n}\left(\beta\mathbf{h}^T\mathbf{z}+\left\|\|\mathbf{z}\|_{\ell_2}\mathbf{g}-\mathbf{w}+\mathbf{v}\right\|_{\ell_2}\right).$$

- Repeat for the other variables $\mathbf{z}$ and $\mathbf{v}$.

Eventually, the AO is reduced to

$$\max_{0\le\beta\le K_\beta}\ \sup_{\gamma,\tau_h\ge 0}\ \min_{0\le\alpha\le K_\alpha}\ \min_{\tau_g\ge 0}\quad D\left(\alpha,\beta,\gamma,\tau_h,\tau_g\right),$$

with

$$D\left(\alpha,\beta,\gamma,\tau_h,\tau_g\right) =$$

$$\frac{\delta\beta}{2\left(\tau_g+\beta\right)}\left(\alpha^2+\sigma^2\right) - \frac{\alpha}{2\tau_h}\left(\gamma^2+\beta^2\right) + \gamma\sqrt{\frac{\alpha^2\beta^2}{\tau_h^2}+V^2} - \frac{\alpha\tau_h}{2} + \frac{\beta\tau_g}{2}$$

$$+\ \delta\mathbf{1}_{\left\{\gamma\left(\tau_g+\beta\right)>\sqrt{\frac{2}{\pi}}\delta\varepsilon\beta\sqrt{\alpha^2+\sigma^2}\right\}}\frac{\beta^2\left(\alpha^2+\sigma^2\right)}{2\tau_g\left(\tau_g+\beta\right)}\left(\operatorname{erf}\left(\frac{\tau_*}{\sqrt{2}}\right) - \frac{\gamma\left(\tau_g+\beta\right)}{\delta\varepsilon\beta\sqrt{\alpha^2+\sigma^2}}\tau_*\right)$$

and τ_* is the unique solution to

$$\frac{\gamma\left(\tau_g+\beta\right)}{\delta\varepsilon\beta\sqrt{\alpha^2+\sigma^2}} - \frac{\beta}{\tau_g}\tau - \tau\cdot\operatorname{erf}\left(\frac{\tau}{\sqrt{2}}\right) - \sqrt{\frac{2}{\pi}}e^{-\frac{\tau^2}{2}} = 0$$

- It holds in probability that

$$\lim_{n\to\infty} \frac{1}{d} \left\| \widehat{\boldsymbol{\theta}}^{\varepsilon} - \boldsymbol{\theta}_0 \right\|_{\ell_2}^2 = \alpha_*^2,$$

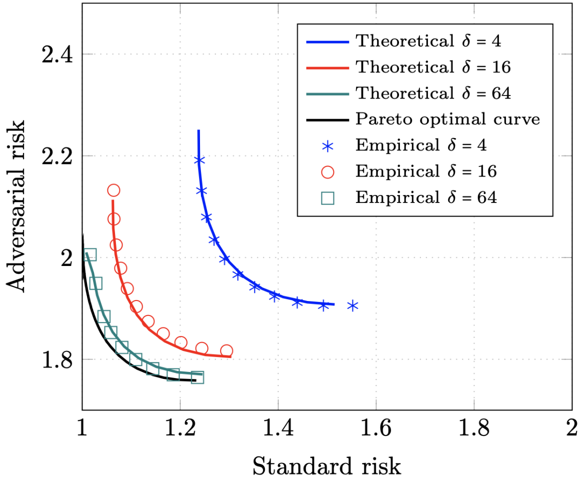
$$\lim_{n\to\infty} \frac{1}{\sqrt{d}} \left\| \widehat{\boldsymbol{\theta}}^{\varepsilon} \right\|_{\ell_2} = \frac{\beta_\star \tau_\star \sqrt{\alpha_*^2 + \sigma^2}}{\varepsilon \tau_{g*}}.$$

- Hence, the following also holds in probability

$$\lim_{n\to\infty} \mathrm{SR}\left(\hat{\theta}^{\varepsilon}\right) = \sigma^2 + \alpha_*^2,$$

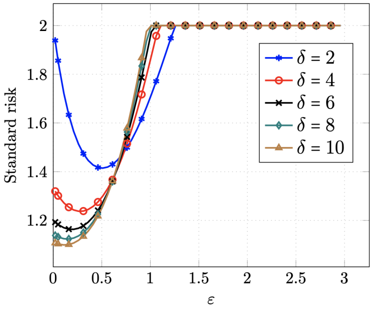
$$\lim_{n\to\infty} \mathrm{AR}\left(\widehat{\theta}^{\varepsilon}\right) = \left(\sigma^2 + \alpha_*^2 + \varepsilon^2 \left(\alpha_*^2 + \sigma^2\right) \left(\frac{\beta_* \tau_*}{\varepsilon \tau_{g*}}\right)^2\right) + 2\sqrt{\frac{2}{\pi} \frac{\varepsilon\ \beta_* \tau_*}{\varepsilon \tau_{g*}}} (\sigma^2 + \alpha_*^2).$$

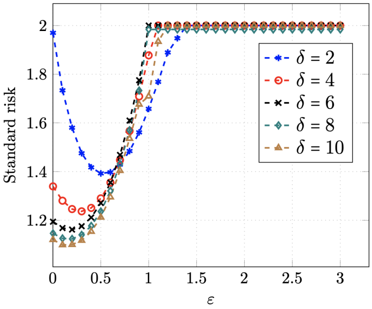
Algorithmic Tradeoffs



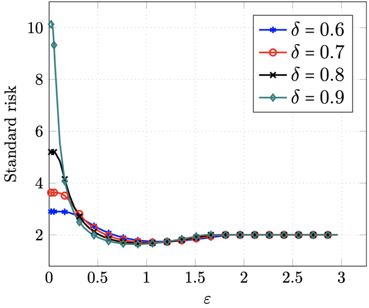
Role of Overparameterization

Overparameterized

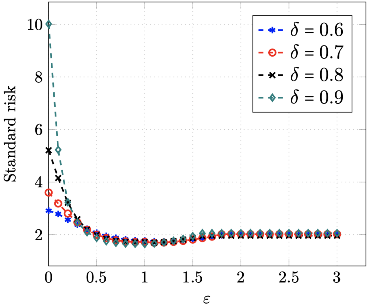




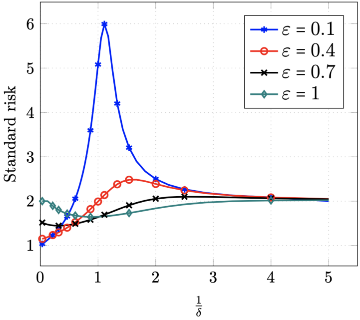
Underparameterized



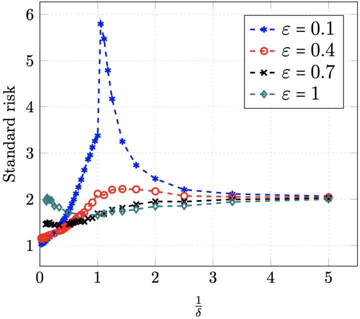
(a) Theoretical curves



(b) Empirical curves



(a) Theoretical curves



(b) Empirical curves

Interpolation threshold depends on ε.

What else can be done?

What else can be done?

- Adversarial training of random feature models: $y = \boldsymbol{\theta}^\top \sigma(W\boldsymbol{x}) + \epsilon$.
- $W \in R^{N\times d}$, $\boldsymbol{\theta} \in \mathbb{R}^d$, and we have n samples.
- $\psi_1 = N/n$ and $\psi_2 = n/d$.

What else can be done?

- Adversarial training of random feature models: $y = \boldsymbol{\theta}^\top \sigma(W\boldsymbol{x}) + \epsilon$.
- $W \in R^{N \times d}$, $\boldsymbol{\theta} \in \mathbb{R}^d$, and we have n samples.
- $\psi_1 = N/n$ and $\psi_2 = n/d$.

Hamed Hassani and Adel Javanmard. The curse of overparametrization in adversarial training: Precise analysis of robust generalization for random features regression, 2022.

What else can be done?

- Adversarial training of random feature models: $y = \boldsymbol{\theta}^\top \sigma(W\boldsymbol{x}) + \epsilon$.
- $W \in R^{N \times d}$, $\boldsymbol{\theta} \in \mathbb{R}^d$, and we have n samples.
- $\psi_1 = N/n$ and $\psi_2 = n/d$.

Hamed Hassani and Adel Javanmard. The curse of overparametrization in adversarial training: Precise analysis of robust generalization for random features regression, 2022.

- Idea (Gaussian Equivalence):

$$\begin{aligned}\sigma(W\boldsymbol{x}) &= \mu_0 \mathbf{1} + \mu_1 W\boldsymbol{x} + \mu_2 \sigma_\perp(W\boldsymbol{x}) \quad \mathbb{E}[W\boldsymbol{x}\sigma_\perp(\mathbf{W}\boldsymbol{x})^\top] = 0 \\ &= \mu_0 \mathbf{1} + \mu_1 W\boldsymbol{x} + \mathbf{u}\end{aligned}$$

What else can be done?

- Adversarial training of random feature models: $y = \boldsymbol{\theta}^\top \sigma(W\boldsymbol{x}) + \epsilon$.
- $W \in R^{N \times d}$, $\boldsymbol{\theta} \in \mathbb{R}^d$, and we have n samples.
- $\psi_1 = N/n$ and $\psi_2 = n/d$.

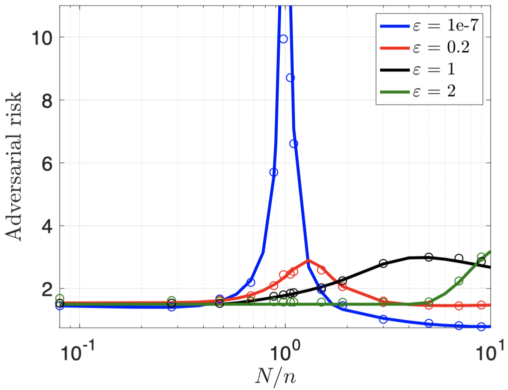
Hamed Hassani and Adel Javanmard. The curse of overparametrization in adversarial training: Precise analysis of robust generalization for random features regression, 2022.

- Idea (Gaussian Equivalence):

$$\begin{aligned}\sigma(W\boldsymbol{x}) &= \mu_0\mathbf{1} + \mu_1 W\boldsymbol{x} + \mu_2\sigma_\perp(W\boldsymbol{x}) \quad \mathbb{E}[W\boldsymbol{x}\sigma_\perp(\mathbf{W}\boldsymbol{x})^\top] = 0 \\ &= \mu_0\mathbf{1} + \mu_1 W\boldsymbol{x} + \mathbf{u}\end{aligned}$$

- Then, use CGMT for the linear regression that pops out.

Results for Random Features



Thanks!

Thank You!