The Price of Implicit Bias in Robust ML

Nikolaos Tsilivis
NYU

Natalie Frank
University of Washington

Nati Srebro
TTIC

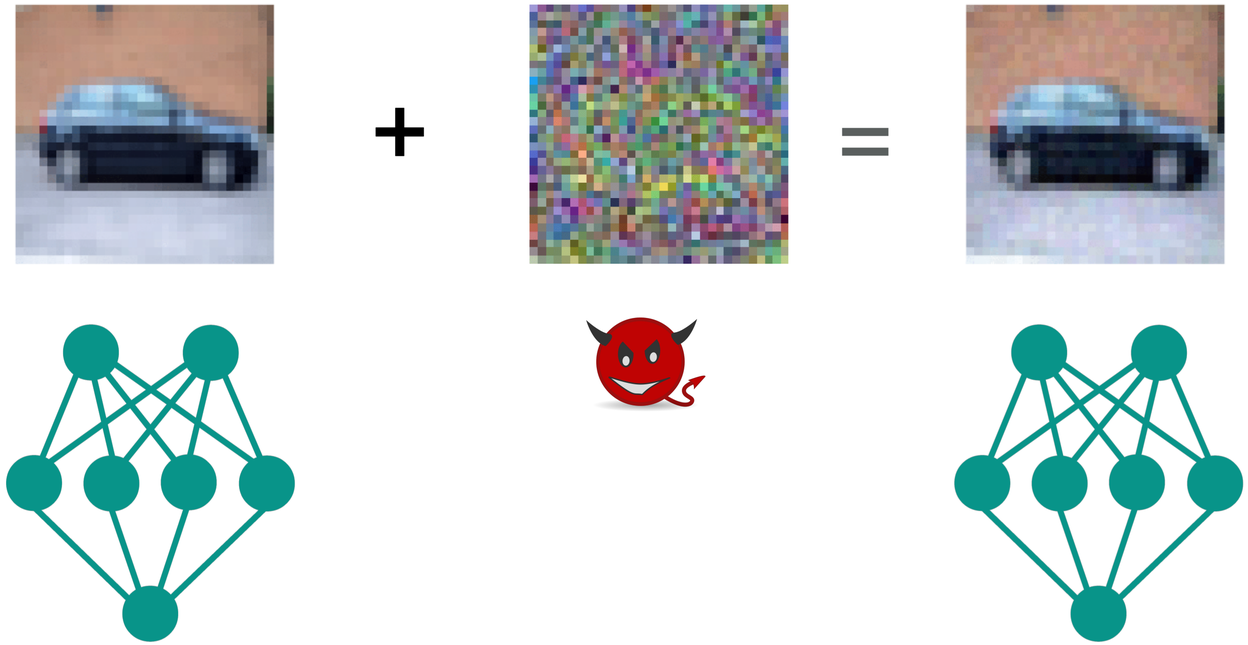
Julia Kempe
NYU & FAIR, Meta

ROBUST SUPERVISED LEARNING

Examples {(sample, label)} → Do well in (sample_new, label_new)

Test time robustness : Examples {(sample, label)} → Do well in *corrupted* *(sample_new', label_new')*

ADVERSARIAL ATTACKS?



'Car'

'Airplane'

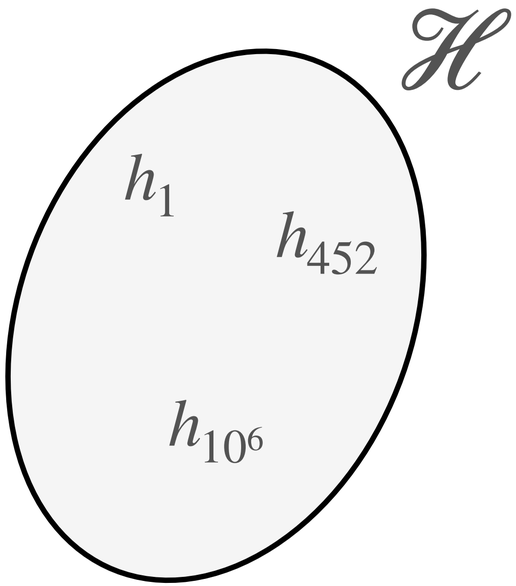
Szegedy et al. Intriguing properties of neural networks. ICLR 2014

Test of Time Award Runner Up, ICLR 2024

ROBUST EMPIRICAL RISK MINIMIZATION

Class of hypotheses $\mathscr{H}$, dataset $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$, perturbation level ϵ, threat model of ℓ_p balls:

$$min_{h \in \mathscr{H}} \frac{1}{m} \sum_{i=1}^{m} \max_{\|\mathbf{x}_i' - \mathbf{x}_i\|_p \leq \epsilon} \mathbf{1}\left[h(\mathbf{x}_i') \neq y_i\right]$$

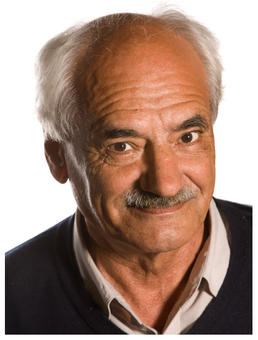


What is different in Robust ERM???

AGENDA

- Understand what is happening in linear models
    - (Robust) Generalization of linear models
    - ⇅
    - (Robust) Optimization of linear models
- Suggest what is happening in general

A PRIMER IN LEARNING THEORY (1/2)



Vapnik & Chernovenkis, 1971

Setup: Learn a concept $c^\star$ using data $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ from distribution D using class of hypotheses $\mathcal{H}$. For $h_{ERM} \in \arg\min_{h \in \mathcal{H}} L_S(h)$,

it holds with high probability:

$$L_D(h_{\text{ERM}}) \leq \inf_{h \in \mathcal{H}} L_D(h) + \sqrt{\frac{\text{capacity}(\mathcal{H})}{m}}$$

Approximation error + Estimation error

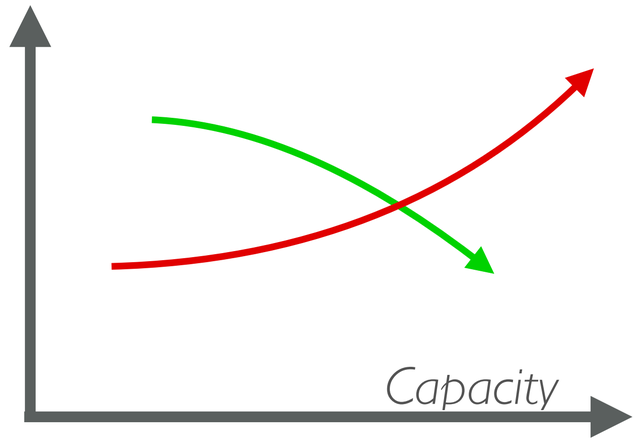
expected error

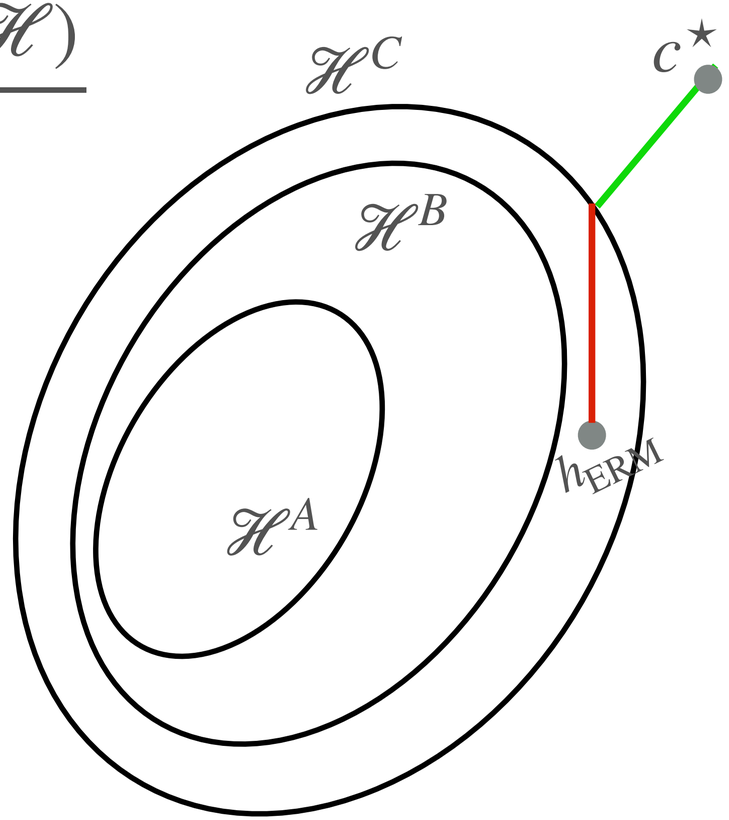
$$L_D(h) = \mathbb{E}_{(\mathbf{x},y)\sim D}\left[l(h(\mathbf{x}), y)\right]$$

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} l(h(\mathbf{x}_i), y_i)$$

empirical error

Tradeoff between fitness and complexity





A PRIMER IN LEARNING THEORY (2/2)

Setup: Data $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ from distribution D and hypothesis class $\mathscr{H}$.

Measure of Complexity: VC dimension

Measure of Complexity: **Rademacher complexity**

expected error

$$L_D(h) = \mathbb{E}_{(\mathbf{x},y)\sim D}\left[l(h(\mathbf{x}), y)\right]$$

empirical error

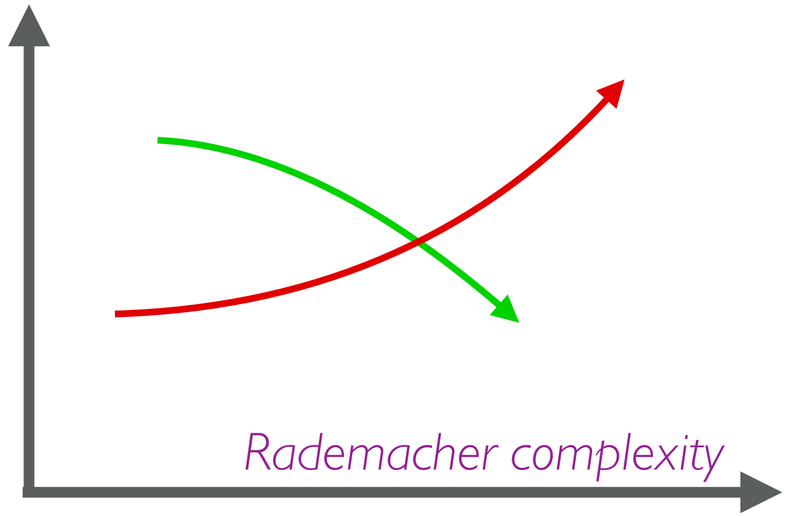
$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} l(h(\mathbf{x}_i), y_i)$$

$$L_D(h_{\text{ERM}}) \leq \inf_{h\in\mathscr{H}} L_D(h) + \text{Rad}(\mathscr{H})$$

Approximation error + Estimation error

$$\text{Rad}(\mathscr{H}) = \mathbb{E}_{\sigma_1,\ldots,\sigma_m}\left[\sup_{h\in\mathscr{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(\mathbf{x}_i)\right],$$

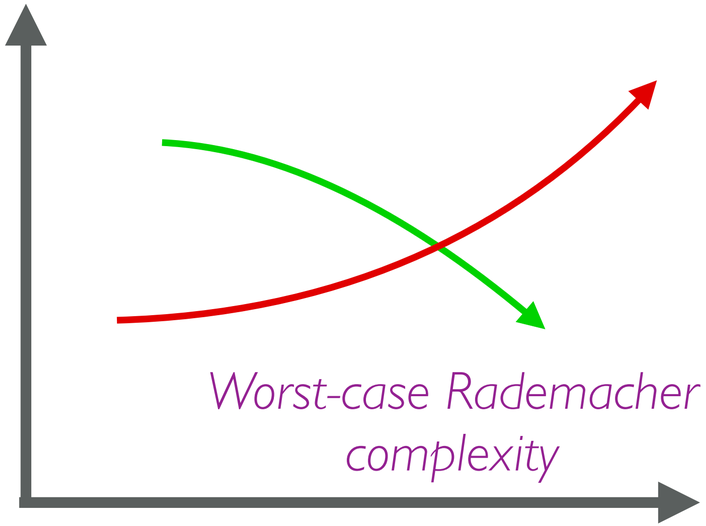
where $\sigma_i \sim Rad$ for all $i = 1,\ldots,m$.



OUR LEARNING SETUP

Worst-case Rademacher complexity

$$\mathrm{Rad}(\tilde{\mathscr{H}}) = \mathbb{E}_{\sigma_1,\ldots,\sigma_m}\left[\sup_{\|\mathbf{w}\|_r \le 1} \frac{1}{m}\sum_{i=1}^{m} \min_{\|\mathbf{x}_i' - \mathbf{x}_i\|_p \le \epsilon} \sigma_i \langle \mathbf{w}, \mathbf{x}_i' \rangle\right]$$



WORST-CASE RADEMACHER COMPLEXITY

Theorem [AFM 2020]

$$\mathrm{Rad}(\tilde{\mathscr{H}}) \leq \mathrm{Rad}(\mathscr{H}) + \frac{\epsilon \max(d^{1-\frac{1}{p}-\frac{1}{r}}, 1)}{2\sqrt{m}}$$

$$\mathrm{Rad}(\tilde{\mathscr{H}}) \geq \max\left(\mathrm{Rad}(\mathscr{H}), \frac{\epsilon \max(d^{1-\frac{1}{p}-\frac{1}{r}}, 1)}{2\sqrt{2m}}\right)$$

★ dimension dependent term, unless $1 - \frac{1}{p} - \frac{1}{r} \leq 0$

★ suggests *dual* regularization $r = \frac{p}{p-1}$

Setup: Hypothesis class $\mathscr{H}_r = \{\mathbf{x} \to \langle \mathbf{w}, \mathbf{x} \rangle : \|\mathbf{w}\|_r \leq 1\}$, dataset $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$, ℓ_p perturbations of size ϵ.

worst-case Rademacher complexity

$$\mathrm{Rad}(\tilde{\mathscr{H}}) = \mathbb{E}_{\sigma_1,\ldots,\sigma_m}\left[\sup_{\|\mathbf{w}\|_r \leq 1} \frac{1}{m} \sum_{i=1}^{m} \min_{\|\mathbf{x}_i' - \mathbf{x}_i\|_p \leq \epsilon} \sigma_i \langle \mathbf{w}, \mathbf{x}_i' \rangle\right]$$

$$\mathrm{Rad}(\mathscr{H}) = \mathbb{E}_{\sigma_1,\ldots,\sigma_m}\left[\sup_{\|\mathbf{w}\|_r \leq 1} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \langle \mathbf{w}, \mathbf{x}_i \rangle\right]$$

Rademacher complexity

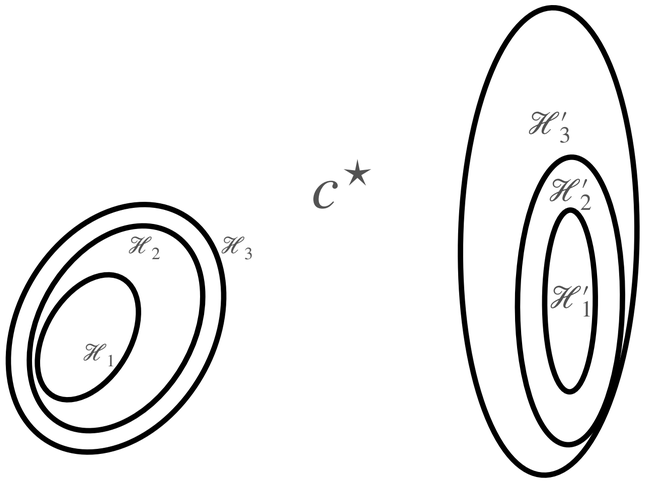
[AFM 2020] Awasthi et al. Adversarial Learning Guarantees for Linear Hypotheses and Neural Networks.

INTERIM SUMMARY

- Type of perturbation ℓ_p can influence "optimal" hypothesis class.

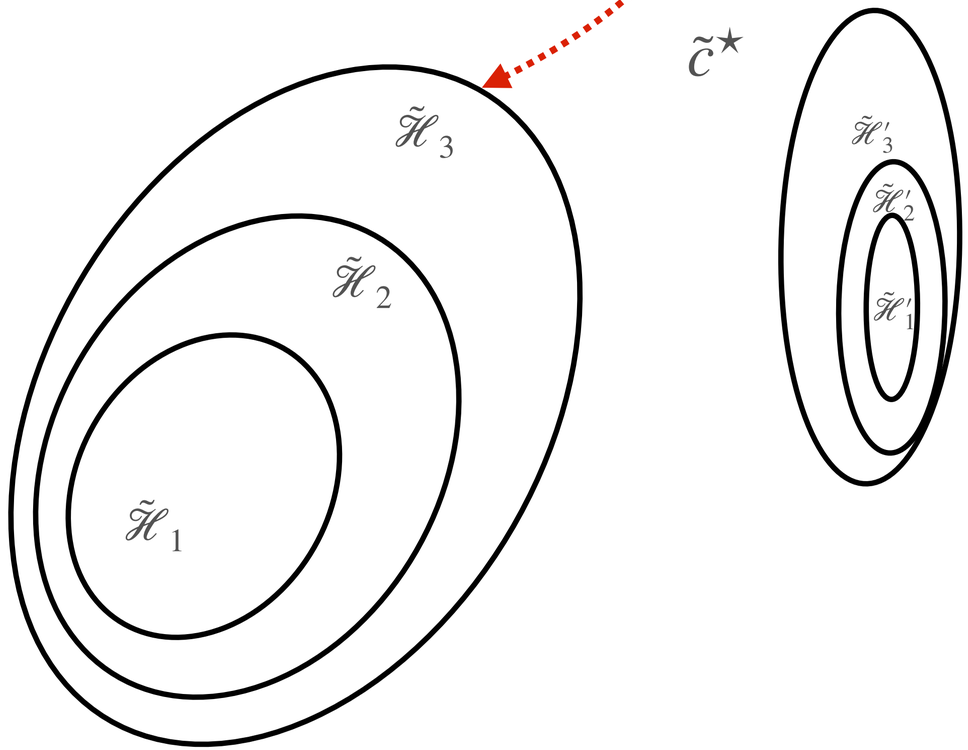
- Model selection becomes more important.

Standard ML



larger classes

Robust ML



IMPLICIT BIAS OF OPTIMIZATION

In supervised learning, we do not get to choose regularization; the regularization chooses us (through optimization)

Optimization algorithm, architecture, loss function, step size, initialization implicitly perform model selection.

IMPLICIT BIAS IN ROBUST ERM: ALGORITHM (1/2)

Unregularized loss $\tilde{L}_S(\mathbf{w}) = \sum_{i=1}^{m} \max_{\|\mathbf{x}_i' - \mathbf{x}_i\|_p \le \epsilon} e^{-y_i \langle \mathbf{w}, \mathbf{x}_i' \rangle}$.

Algorithm: *Steepest Descent* w.r.t. $\|\cdot\|_r$ norm

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \eta \Delta \mathbf{w}_t,$$

where $\Delta \mathbf{w}_t \in \arg \min_{\|\mathbf{u}\|_r \le \|\nabla L_S(\mathbf{w}_t)\|_{r\star}} \langle \mathbf{u}, \nabla L_S(\mathbf{w}_t) \rangle$.

Theorem

For any *separable* dataset and any initialization $\mathbf{w}_0$, for $\eta \to 0^+$, it holds:

$$\lim_{t\to\infty} \min_{i\in[m]} \min_{\|\mathbf{x}_i' - \mathbf{x}_i\|_p \le \epsilon} \frac{y_i \langle \mathbf{w}_t, \mathbf{x}_i' \rangle}{\|\mathbf{w}_t\|_r} = \max_{\mathbf{w}\neq 0} \min_{i\in[m]} \min_{\|\mathbf{x}_i' - \mathbf{x}_i\|_p \le \epsilon} \frac{y_i \langle \mathbf{w}, \mathbf{x}_i' \rangle}{\|\mathbf{w}\|_r}.$$

[**T.**, Frank, Srebro, Kempe. The Price of Implicit Bias in Adversarially Robust Generalization. NeurIPS 2024]

IMPLICIT BIAS IN ROBUST ERM: ALGORITHM (2/2)

Theorem

For any *separable* dataset and any initialization $\mathbf{w}_0$, for $\eta \to 0^+$, it holds:

$$\lim_{t\to\infty} \min_{i\in[m]} \min_{\|\mathbf{x}_i'-\mathbf{x}_i\|_p\le\epsilon} \frac{y_i\langle \mathbf{w}_t, \mathbf{x}_i'\rangle}{\|\mathbf{w}_t\|_r} = \max_{\mathbf{w}\neq 0} \min_{i\in[m]} \min_{\|\mathbf{x}_i'-\mathbf{x}_i\|_p\le\epsilon} \frac{y_i\langle \mathbf{w}, \mathbf{x}_i'\rangle}{\|\mathbf{w}\|_r}\,.$$

Algorithm: *Steepest Descent* w.r.t. $\|\cdot\|_r$ norm

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \eta\Delta\mathbf{w}_t,$$

$$\text{where } \Delta\mathbf{w}_t \in \arg\min_{\|\mathbf{u}\|_r\le\|\nabla L_S(\mathbf{w}_t)\|_{r\star}} \langle \mathbf{u}, \nabla L_S(\mathbf{w}_t)\rangle\,.$$

★ Equivalent to $\min_{\mathbf{w}} \|\mathbf{w}\|_r$ s.t. $\min_{\|\mathbf{x}_i'-\mathbf{x}_i\|_p\le\epsilon} y_i\langle \mathbf{w}, \mathbf{x}_i'\rangle \ge 1, \forall i \in [m]\,.$

★ For $r = 2$, robust ERM with GD $\to \|\mathbf{w}\|_2$ regularization (regardless of p).

★ Steepest descent for $r = p^\star = \dfrac{p}{p-1} \to \|\mathbf{w}\|_{p^\star}$ regularization.

[**T.**, Frank, Srebro, Kempe. The Price of Implicit Bias in Adversarially Robust Generalization. NeurIPS 2024]

IMPLICIT BIAS IN ROBUST ERM: ARCHITECTURE

Reparameterized linear model $f_{\text{diag}}(\mathbf{x}; \mathbf{w}) = \langle u_+^2 - u_-^2, \mathbf{x} \rangle, w = [\mathbf{u}_+, \mathbf{u}_-]$.

Unregularized loss $\tilde{L}_S(\mathbf{w}) = \sum_{i=1}^{m} \max_{\|\mathbf{x}_i' - \mathbf{x}_i\|_p \leq \epsilon} e^{-y_i f_{\text{diag}}(\mathbf{x}_i'; \mathbf{w})}$.

Proposition (Informal)

Consider GD over $f_{\text{diag}}(\,\cdot\,; \mathbf{w})$ and assume $\tilde{L}_s(\mathbf{w}) \to 0$, then the implicit bias of $\mathbf{w}$ corresponds to:

$$\min_{\mathbf{w}} \|\mathbf{w}\|_1 \text{ s.t. } \min_{\|\mathbf{x}_i' - \mathbf{x}_i\|_p \leq \epsilon} y_i \langle \mathbf{w}, \mathbf{x}_i' \rangle \geq 1, \forall i \in [m] .$$

★ Same algorithm, different architecture->different implicit bias.

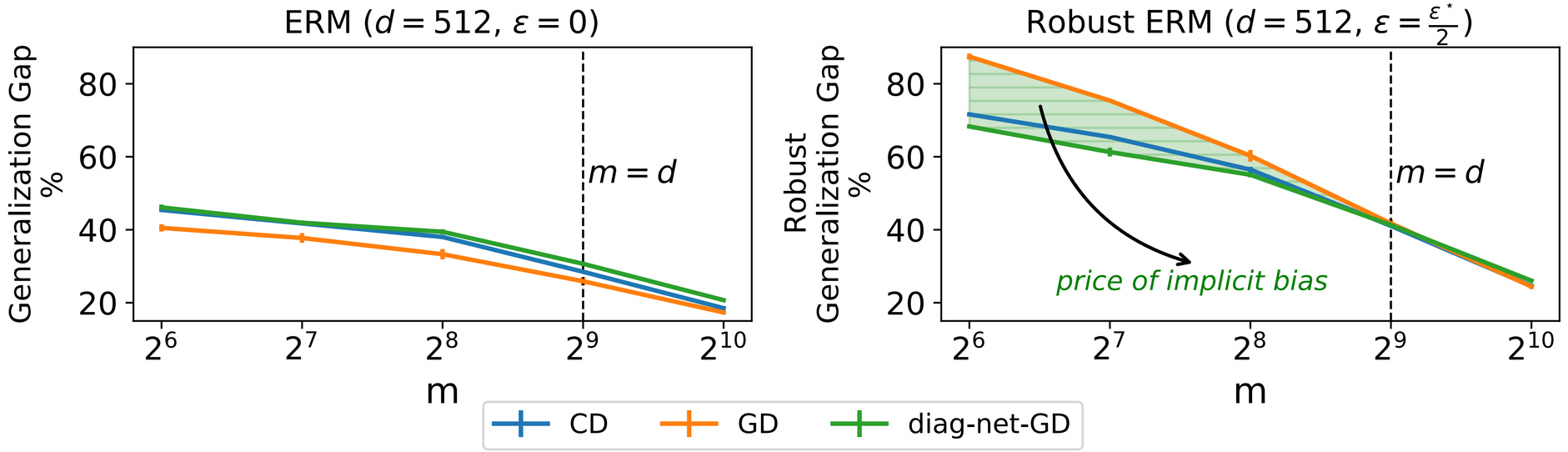
★ Favorable bias for $p = \infty$.

[**T.**, Frank, Srebro, Kempe. The Price of Implicit Bias in Adversarially Robust Generalization. NeurIPS 2024]

PRICE OF IMPLICIT BIAS

Coordinate Descent - CD:
Change only one coordinate per update (largest absolute gradient) - Steepest descent wrt ℓ_1 norm

- $\mathbf{x} \sim \mathcal{N}(0,I), \mathbf{w}^\star \sim \mathcal{N}(0,I), y = \mathrm{sgn}(\langle \mathbf{w}^\star, \mathbf{x} \rangle)$. Defend against ℓ_∞ perturbations.
- Find largest possible perturbation $\epsilon^\star$ and use $\epsilon < \epsilon^\star$.
- Compare GD on linear networks vs CD on linear networks vs GD on diagonal network.

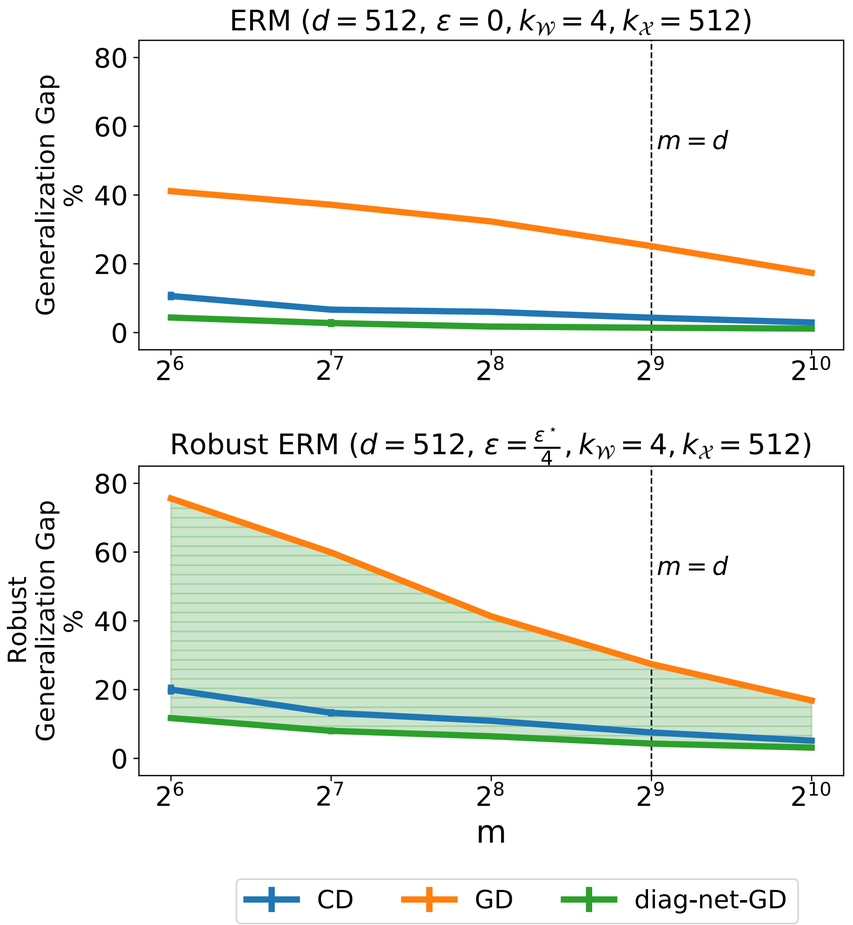


✅ *Type of perturbation ℓ_p can influence "optimal" hypothesis class.*

[**T.**, Frank, Srebro, Kempe. The Price of Implicit Bias in Adversarially Robust Generalization. NeurIPS 2024]

PRICE OF IMPLICIT BIAS

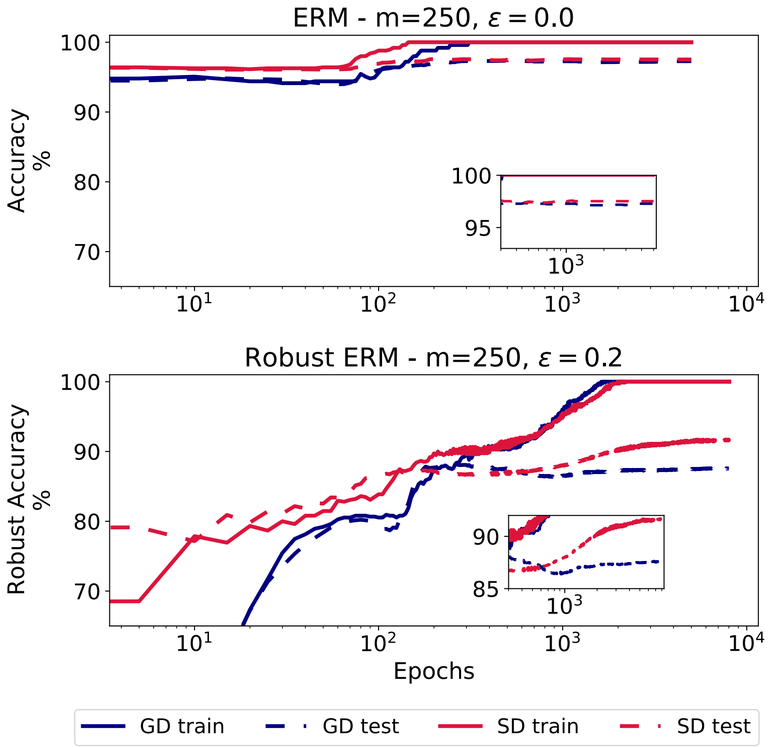
Same experiment with sparse teacher $\mathbf{w}^\star$

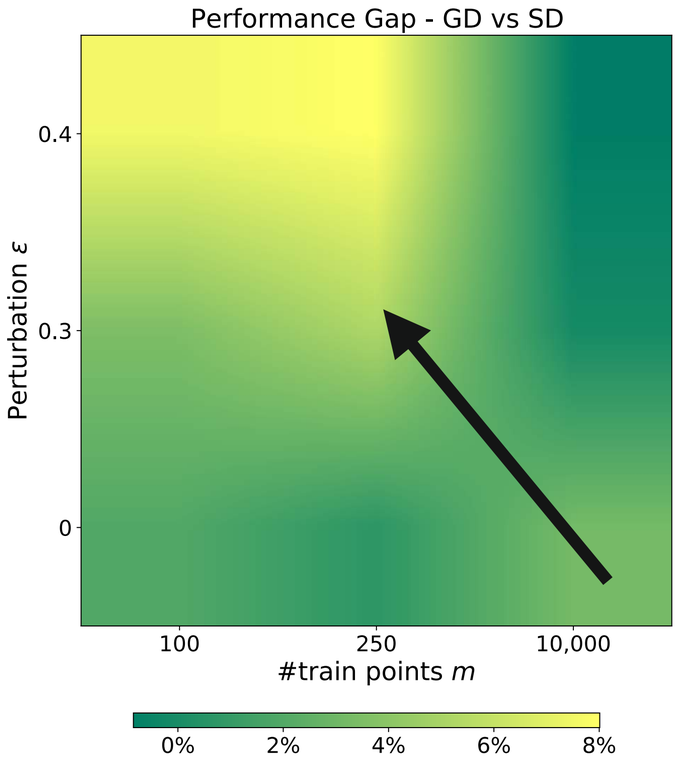


✅ *Model selection becomes more important.*

[**T.**, Frank, Srebro, Kempe. The Price of Implicit Bias in Adversarially Robust Generalization. NeurIPS 2024]

PRICE OF IMPLICIT BIAS IN NEURAL NETWORKS





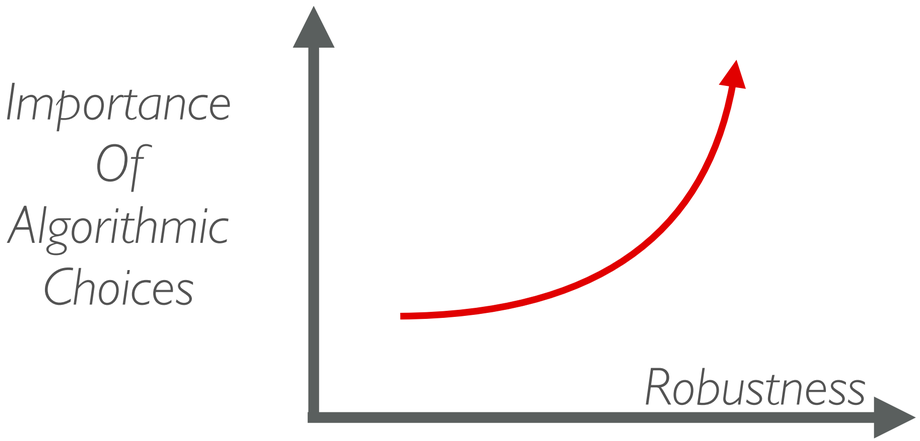
Sign Gradient Descent

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \operatorname{sgn}\left(\nabla L(\mathbf{w}_t)\right)$$

[**T.**, Vardi, Kempe. Flavors of Margin: Implicit Bias of Steepest Descent in Homogeneous Neural Networks]

[**T.**, Frank, Srebro, Kempe. The Price of Implicit Bias in Adversarially Robust Generalization. NeurIPS 2024]

CONCLUSION



Robust training as a testbed for theories for deep learning