Can Uniform Convergence Explain Interpolation Learning?

D.J. Sutherland
TTI-Chicago → UBC

based on arXiv:2006.05942 (NeurIPS 2020), with:

Lijia Zhou
U Chicago

Nati Srebro
TTI-Chicago

Penn State Statistics Seminar, October 8 2020

Supervised learning

- Given i.i.d. samples $\mathbf{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$
  - features/covariates $\mathbf{x}_i \in \mathbb{R}^d$, labels/targets $y_i \in \mathbb{R}$

Supervised learning

- Given i.i.d. samples $\mathbf{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$
  - features/covariates $\mathbf{x}_i \in \mathbb{R}^d$, labels/targets $y_i \in \mathbb{R}$
- Want f such that $f(\mathbf{x}) \approx y$ for new samples from $\mathcal{D}$

Supervised learning

- Given i.i.d. samples $\mathbf{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$
  - features/covariates $\mathbf{x}_i \in \mathbb{R}^d$, labels/targets $y_i \in \mathbb{R}$
- Want f such that $f(\mathbf{x}) \approx y$ for new samples from $\mathcal{D}$:

$$f^* = \mathbf{argmin}\left[L_{\mathcal{D}}(f) := \mathop{\mathbb{E}}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}} L(f(\mathbf{x}), \mathrm{y})\right]$$

Supervised learning

- Given i.i.d. samples $\mathbf{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$
  - features/covariates $\mathbf{x}_i \in \mathbb{R}^d$, labels/targets $y_i \in \mathbb{R}$
- Want f such that $f(\mathbf{x}) \approx y$ for new samples from $\mathcal{D}$:

$$f^* = \mathbf{argmin} \left[L_{\mathcal{D}}(f) := \mathop{\mathbb{E}}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}} L(f(\mathbf{x}), \mathrm{y}) \right]$$

  - e.g. squared loss: $L(\hat{y}, y) = (\hat{y} - y)^2$

Supervised learning

- Given i.i.d. samples $\mathbf{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$
  - features/covariates $\mathbf{x}_i \in \mathbb{R}^d$, labels/targets $y_i \in \mathbb{R}$
- Want f such that $f(\mathbf{x}) \approx y$ for new samples from $\mathcal{D}$:

$$f^* = \mathbf{argmin} \left[L_{\mathcal{D}}(f) := \mathop{\mathbb{E}}_{(\mathbf{x},\mathrm{y})\sim\mathcal{D}} L(f(\mathbf{x}), \mathrm{y}) \right]$$

  - e.g. squared loss: $L(\hat{y}, y) = (\hat{y} - y)^2$
- Standard approaches based on empirical risk minimization:

$$\hat{f} \approx \mathbf{argmin} \left[L_{\mathbf{S}}(f) := \frac{1}{n} \sum_{i=1}^n L(f(\mathbf{x}_i), y_i) \right]$$

Statistical learning theory

We have lots of bounds like: with probability $\geq 1 - \delta$,

$$\sup_{f \in \mathcal{F}} |L_{\mathcal{D}}(f) - L_{\mathbf{S}}(f)| \leq \sqrt{\frac{C_{\mathcal{F},\delta}}{n}}$$

Statistical learning theory

We have lots of bounds like: with probability $\geq 1 - \delta$,

$$\sup_{f \in \mathcal{F}} |L_{\mathcal{D}}(f) - L_{\mathbf{S}}(f)| \leq \sqrt{\frac{C_{\mathcal{F},\delta}}{n}}$$

$C_{\mathcal{F},\delta}$ could be from VC dimension, covering number, RKHS norm, Rademacher complexity, fat-shattering dimension, ...

Statistical learning theory

We have lots of bounds like: with probability $\geq 1 - \delta$,

$$\sup_{f \in \mathcal{F}} |L_{\mathcal{D}}(f) - L_{\mathbf{S}}(f)| \leq \sqrt{\frac{C_{\mathcal{F},\delta}}{n}}$$

Statistical learning theory

We have lots of bounds like: with probability $\geq 1 - \delta$,

$$\sup_{f \in \mathcal{F}} |L_{\mathcal{D}}(f) - L_{\mathbf{S}}(f)| \leq \sqrt{\frac{C_{\mathcal{F},\delta}}{n}}$$

Then for large n, $L_{\mathcal{D}}(f) \approx L_{\mathbf{S}}(f)$, so $\hat{f} \approx f^*$

Statistical learning theory

We have lots of bounds like: with probability $\geq 1 - \delta$,

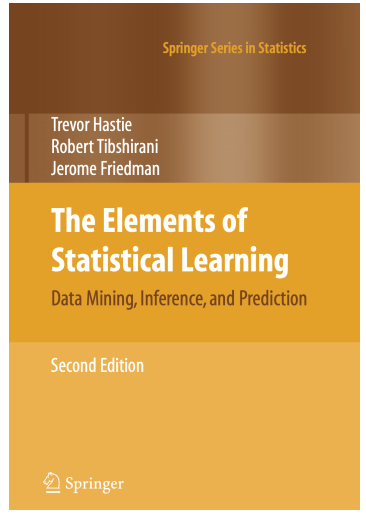
$$\sup_{f \in \mathcal{F}} |L_{\mathcal{D}}(f) - L_{\mathbf{S}}(f)| \leq \sqrt{\frac{C_{\mathcal{F},\delta}}{n}}$$

Then for large n, $L_{\mathcal{D}}(f) \approx L_{\mathbf{S}}(f)$, so $\hat{f} \approx f^*$

$$L_{\mathcal{D}}(\hat{f}) \leq L_{\mathbf{S}}(\hat{f}) + \sup_{f \in \mathcal{F}} |L_{\mathcal{D}}(f) - L_{\mathbf{S}}(f)|$$

Interpolation learning

Classical wisdom: “a model with zero training error is overfit to the training data and will typically generalize poorly”



Interpolation learning

Classical wisdom: “a model with zero training error is overfit to the training data and will typically generalize poorly”

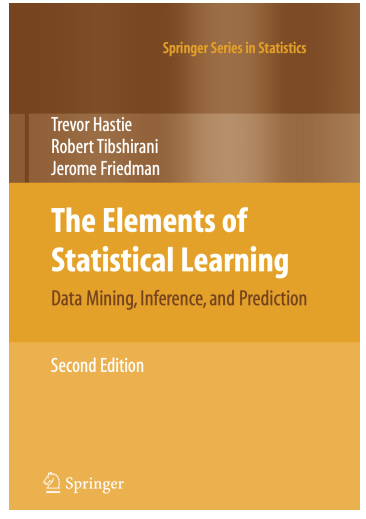


Table 1: The training and test accuracy (in percentage) of various models on the CIFAR10 dataset.

model	# params	random crop	weight decay	train accuracy	test accuracy
Inception	1,649,402	yes	yes	100.0	89.05
		yes	no	100.0	89.31
		no	yes	100.0	86.03
		no	no	100.0	85.75

Zhang et al., “Rethinking generalization”, ICLR 2017 $L_{\mathbf{S}}(\hat{f}) = 0; L_{\mathcal{D}}(\hat{f}) \approx 11\%$

Interpolation learning

Classical wisdom: “a model with zero training error is overfit to the training data and will typically generalize poorly”

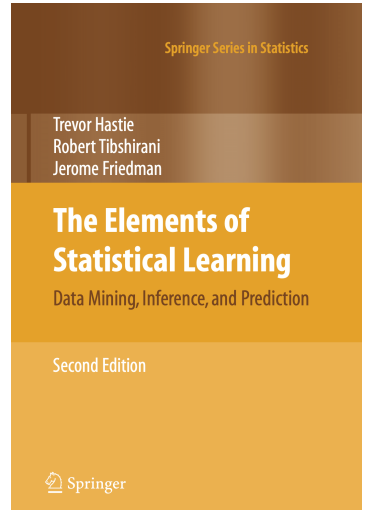


Table 1: The training and test accuracy (in percentage) of various models on the CIFAR10 dataset.

model	# params	random crop	weight decay	train accuracy	test accuracy
Inception	1,649,402	yes	yes	100.0	89.05
		yes	no	100.0	89.31
		no	yes	100.0	86.03
		no	no	100.0	85.75

Zhang et al., “Rethinking generalization”, ICLR 2017 $L_{\mathbf{S}}(\hat{f}) = 0; L_{\mathcal{D}}(\hat{f}) \approx 11\%$

We'll call a model with $L_{\mathbf{S}}(f) = 0$ an *interpolating* predictor

Interpolation learning

Classical wisdom: “a model with zero training error is overfit to the training data and will typically generalize poorly”

(when $L_{\mathcal{D}}(f^*) > 0$)

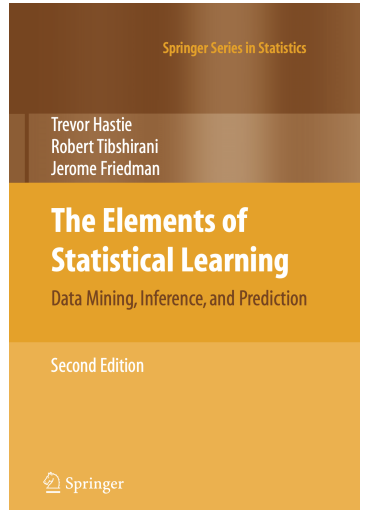


Table 1: The training and test accuracy (in percentage) of various models on the CIFAR10 dataset.

model	# params	random crop	weight decay	train accuracy	test accuracy
Inception	1,649,402	yes	yes	100.0	89.05
		yes	no	100.0	89.31
		no	yes	100.0	86.03
		no	no	100.0	85.75

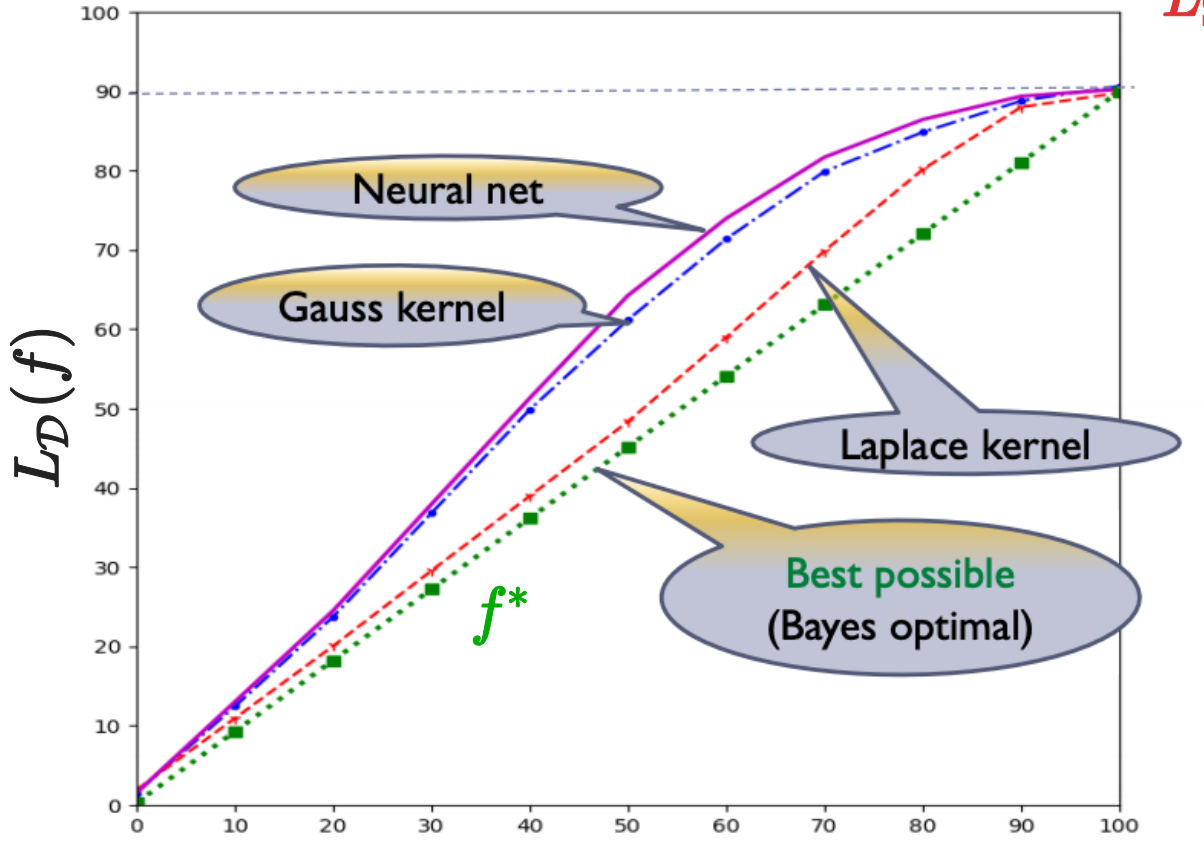
Zhang et al., “Rethinking generalization”, ICLR 2017 $L_{\mathbf{S}}(\hat{f}) = 0; L_{\mathcal{D}}(\hat{f}) \approx 11\%$

We'll call a model with $L_{\mathbf{S}}(f) = 0$ an *interpolating* predictor

Interpolation does not overfit even for very noisy data

All methods (except Bayes optimal) have zero training *square* loss.

$$L_S(\hat{f}) = 0$$



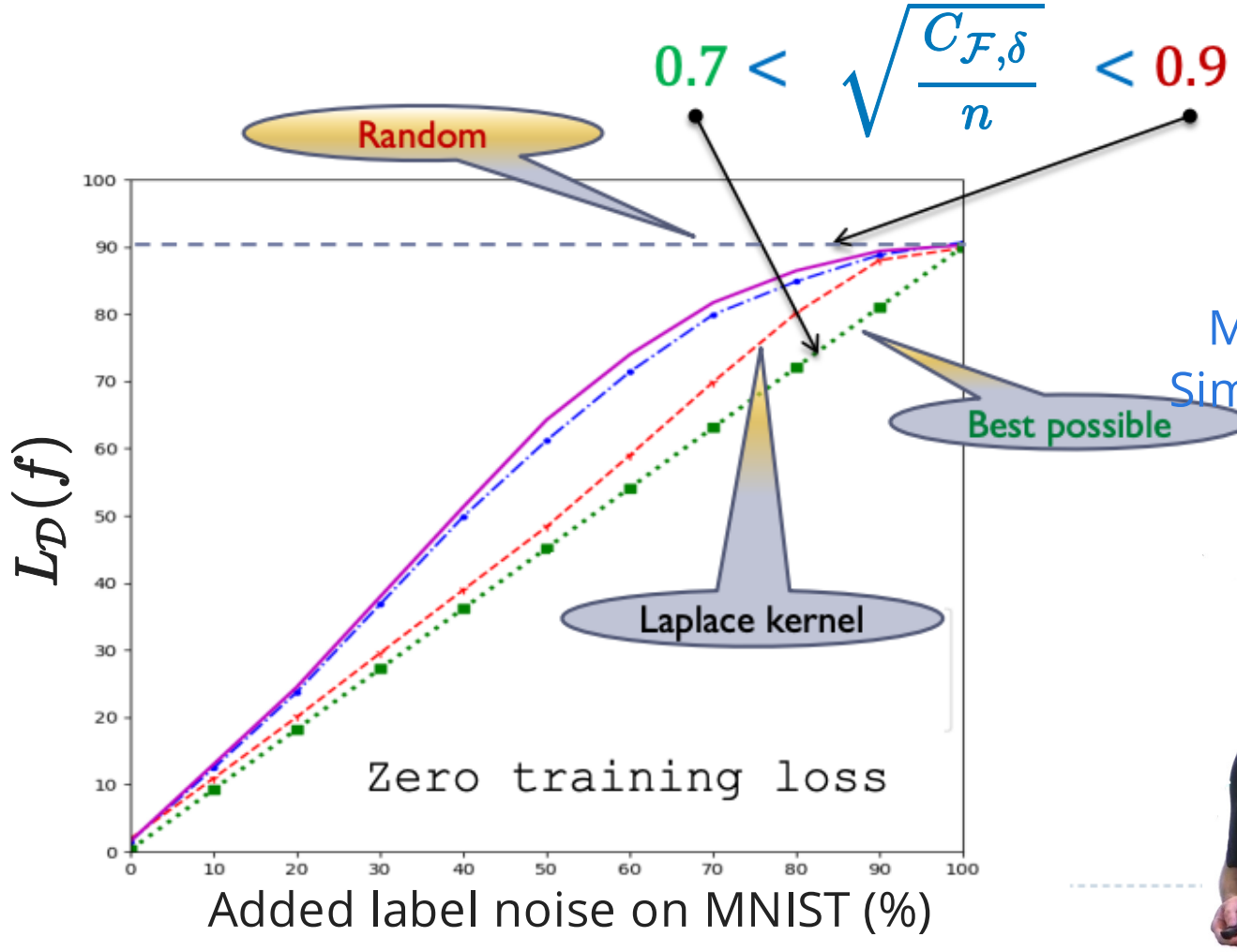
Misha Belkin
Simons Institute
July 2019

Belkin/Ma/Mandal, ICML 2018

Bounds?

$$L_{\mathcal{D}}(\hat{f}) \leq \underbrace{L_{\mathbf{S}}(\hat{f})}_{0} + \sqrt{\frac{C_{\mathcal{F},\delta}}{n}}$$

What kind of generalization bound could work here?



Misha Belkin
Simons Institute
July 2019

Belkin/Ma/Mandal, ICML 2018

Not a question of improving bounds

correct

$$0.7 < \sqrt{\frac{C_{\mathcal{F},\delta}}{n}} < 0.9 \qquad n \to \infty$$

nontrivial

Misha Belkin
Simons Institute
July 2019

There are no bounds like this and no reason they should exist.

A constant factor of 2 invalidates the bound!

Generalization theory for interpolation?

What theoretical analyses do we have?

- VC-dimension/Rademacher complexity/covering/margin bounds.
  - Cannot deal with interpolated classifiers when Bayes risk is non-zero.
  - Generalization gap cannot be bound when empirical risk is zero.
- Regularization-type analyses (Tikhonov, early stopping/SGD, etc.)
  - Diverge as $\lambda \to 0$ for fixed n.
- Algorithmic stability.
  - Does not apply when empirical risk is zero, expected risk nonzero.

$$L_{\mathcal{D}}(\hat{f}) \leq L_{\mathrm{S}}(\hat{f}) + \text{bound}$$

WYSIWYG bounds:

$= 0$

training loss

$\approx$

expected loss

Misha Belkin
Simons Institute
July 2019

- Classical smoothing methods (i.e., Nadaraya-Watson).
  - Most classical analyses do not support interpolation.
  - But 1-NN! (Also Hilbert regression Scheme, [Devroye, et al. 98])

Oracle bounds

expected loss

$\approx$

optimal loss

$$L_{\mathcal{D}}(\hat{f}) \leq L_{\mathcal{D}}(f^*) + \text{bound}$$

Generalization theory for interpolation?

What theoretical analyses do we have?

Lots of recent theoretical work on interpolation.

[Belkin+ NeurIPS 2018], [Belkin+ AISTATS 2018], [Belkin+ 2019], [Hastie+ 2019],

[Muthukumar+ JSAIT 2020], [Bartlett+ PNAS 2020], [Liang+ COLT 2020], [Montanari+ 2019], many more...

None* bound $\sup_{f \in \mathcal{F}} |L_{\mathcal{D}}(f) - L_{\mathbf{S}}(f)|$.

Is it possible to find such a bound?

Can uniform convergence explain interpolation learning?

- But 1-NN! (Also Hilbert regression Scheme, [Devroye, et al. 98])

$\approx$ optimal loss

$$L_{\mathcal{D}}(\hat{f}) \leq L_{\mathcal{D}}(f^*) + \text{bound}$$

Generalization theory for interpolation?

What theoretical analyses do we have?

Lots of recent theoretical work on interpolation.

[Belkin+ NeurIPS 2018], [Belkin+ AISTATS 2018], [Belkin+ 2019], [Hastie+ 2019],

[Muthukumar+ JSAIT 2020], [Bartlett+ PNAS 2020], [Liang+ COLT 2020], [Montanari+ 2019], many more...

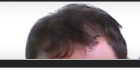
None* bound $\sup_{f \in \mathcal{F}} |L_{\mathcal{D}}(f) - L_{\mathbf{S}}(f)|$.

Is it possible to find such a bound?

Can uniform convergence explain interpolation learning?

But 1-NN! (Also Hilbert regression Scheme, [Devroye, et al. 98])

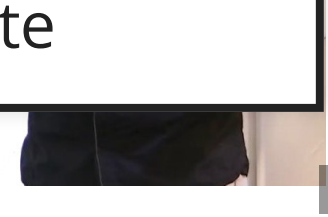
≈ optimal loss



*One exception-ish [Negrea/Dziugaite/Roy, ICML 2020]:

relates $\hat{f}$ to a surrogate predictor,

shows uniform convergence for the surrogate



A more specific version of the question

We're only going to worry about *consistency*:

$$\mathbb{E}[L_{\mathcal{D}}(\hat{f}) - L_{\mathcal{D}}(f^*)] \to 0$$

A more specific version of the question

We're only going to worry about *consistency*:

$$\mathbb{E}[L_{\mathcal{D}}(\hat{f}) - L_{\mathcal{D}}(f^*)] \to 0$$

...in a *non-realizable* setting: $L_{\mathcal{D}}(f^*) > 0$

A more specific version of the question

We're only going to worry about *consistency*:

$$\mathbb{E}[L_{\mathcal{D}}(\hat{f}) - L_{\mathcal{D}}(f^*)] \to 0$$

...in a *non-realizable* setting: $L_{\mathcal{D}}(f^*) > 0$

Is it possible to show consistency of an interpolator with

$$L_{\mathcal{D}}(\hat{f}) \leq \underbrace{L_{\mathbf{S}}(\hat{f})}_{0} + \sup_{f \in \mathcal{F}} |L_{\mathcal{D}}(f) - L_{\mathbf{S}}(f)|?$$

A more specific version of the question

We're only going to worry about *consistency*:

$$\mathbb{E}[L_{\mathcal{D}}(\hat{f}) - L_{\mathcal{D}}(f^*)] \to 0$$

...in a *non-realizable* setting: $L_{\mathcal{D}}(f^*) > 0$

Is it possible to show consistency of an interpolator with

$$L_{\mathcal{D}}(\hat{f}) \leq \underbrace{L_{\mathbf{S}}(\hat{f})}_{0} + \sup_{f \in \mathcal{F}} |L_{\mathcal{D}}(f) - L_{\mathbf{S}}(f)|?$$

This requires tight constants!

Our testbed problem

	"signal", d_S	"junk", $d_J \to \infty$
$\mathbf{x}$	$\mathbf{x}_S \sim \mathcal{N}\left(\mathbf{0}_{d_S}, \mathbf{I}_{d_S}\right)$	$\mathbf{x}_J \sim \mathcal{N}\left(\mathbf{0}_{d_J}, \frac{\lambda_n}{d_J}\mathbf{I}_{d_J}\right)$
$\mathbf{w}^*$	$\mathbf{w}_S^*$	$\mathbf{0}$

$$y = \underbrace{\langle \mathbf{x}, \mathbf{w}^* \rangle}_{\langle \mathbf{x}_S, \mathbf{w}_S{}^* \rangle} + \mathcal{N}(0, \sigma^2)$$

λ_n controls scale of junk: $\mathbb{E}\|\mathbf{x}_J\|^2 = \lambda_n$

Linear regression: $L(y, \hat{y}) = (y - \hat{y})^2$

Our testbed problem

	"signal", d_S	"junk", $d_J \to \infty$
$\mathbf{x}$	$\mathbf{x}_S \sim \mathcal{N}\left(\mathbf{0}_{d_S}, \mathbf{I}_{d_S}\right)$	$\mathbf{x}_J \sim \mathcal{N}\left(\mathbf{0}_{d_J}, \frac{\lambda_n}{d_J}\mathbf{I}_{d_J}\right)$
$\mathbf{w}^*$	$\mathbf{w}_S^*$	$\mathbf{0}$

$$y = \underbrace{\langle \mathbf{x}, \mathbf{w}^* \rangle}_{\langle \mathbf{x}_S, \mathbf{w}_S{}^* \rangle} + \mathcal{N}(0, \sigma^2)$$

λ_n controls scale of junk: $\mathbb{E}\|\mathbf{x}_J\|^2 = \lambda_n$

Linear regression: $L(y, \hat{y}) = (y - \hat{y})^2$

Min-norm interpolator: $\hat{\mathbf{w}}_{MN} = \underset{\mathbf{X}\mathbf{w}=\mathbf{y}}{\mathbf{argmin}} \|\mathbf{w}\|_2 = \mathbf{X}^\dagger \mathbf{y}$

Consistency of $\hat{\mathbf{w}}_{MN}$

As $d_J \to \infty$, $\hat{\mathbf{w}}_{MN}$ approaches ridge regression on the signal:

$$\langle \hat{\mathbf{w}}_{MN}, \mathbf{x} \rangle \overset{d_J \to \infty}{\longrightarrow} \langle \hat{\mathbf{w}}_{\lambda_n}, \mathbf{x}_S \rangle \text{ for almost all } \mathbf{x}$$

$$\hat{\mathbf{w}}_{\lambda_n} = \underset{\mathbf{w}_S \in \mathbb{R}^{d_S}}{\operatorname{argmin}} \left\| \mathbf{X}_S \mathbf{w}_S{}^2 - \mathbf{y} \right\|^2 + \lambda_n \left\| \mathbf{w}_S \right\|^2$$

Consistency of $\hat{\mathbf{w}}_{MN}$

As $d_J \to \infty$, $\hat{\mathbf{w}}_{MN}$ approaches ridge regression on the signal:

$$\langle \hat{\mathbf{w}}_{MN}, \mathbf{x} \rangle \overset{d_J \to \infty}{\longrightarrow} \langle \hat{\mathbf{w}}_{\lambda_n}, \mathbf{x}_S \rangle \text{ for almost all } \mathbf{x}$$

$$\hat{\mathbf{w}}_{\lambda_n} = \underset{\mathbf{w}_S \in \mathbb{R}^{d_S}}{\operatorname{argmin}} \left\| \mathbf{X}_S \mathbf{w}_S{}^2 - \mathbf{y} \right\|^2 + \lambda_n \left\| \mathbf{w}_S \right\|^2$$

If $\lambda_n = o(n)$, $\hat{\mathbf{w}}_{\lambda_n}$ is consistent: $L_{\mathcal{D}}(\hat{\mathbf{w}}_{\lambda_n}) \overset{n \to \infty}{\longrightarrow} \sigma^2$

Consistency of $\hat{\mathbf{w}}_{MN}$

As $d_J \to \infty$, $\hat{\mathbf{w}}_{MN}$ approaches ridge regression on the signal:

$$\langle \hat{\mathbf{w}}_{MN}, \mathbf{x} \rangle \overset{d_J \to \infty}{\longrightarrow} \langle \hat{\mathbf{w}}_{\lambda_n}, \mathbf{x}_S \rangle \text{ for almost all } \mathbf{x}$$

$$\hat{\mathbf{w}}_{\lambda_n} = \underset{\mathbf{w}_S \in \mathbb{R}^{d_S}}{\operatorname{argmin}} \left\| \mathbf{X}_S \mathbf{w}_S{}^2 - \mathbf{y} \right\|^2 + \lambda_n \left\| \mathbf{w}_S \right\|^2$$

If $\lambda_n = o(n)$, $\hat{\mathbf{w}}_{\lambda_n}$ is consistent: $L_{\mathcal{D}}(\hat{\mathbf{w}}_{\lambda_n}) \overset{n \to \infty}{\longrightarrow} \sigma^2$

$\hat{\mathbf{w}}_{MN}$ is consistent when d_S fixed, $d_J \to \infty$, $\lambda_n = o(n)$.

Could we have shown that with uniform convergence?

Consistency of $\hat{\mathbf{w}}_{MN}$

As $d_J \to \infty$, $\hat{\mathbf{w}}_{MN}$ approaches ridge regression on the signal:

$$\langle \hat{\mathbf{w}}_{MN}, \mathbf{x} \rangle \overset{d_J \to \infty}{\longrightarrow} \langle \hat{\mathbf{w}}_{\lambda_n}, \mathbf{x}_S \rangle \text{ for almost all } \mathbf{x}$$

$$\hat{\mathbf{w}}_{\lambda_n} = \underset{\mathbf{w}_S \in \mathbb{R}^{d_S}}{\operatorname{argmin}} \left\| \mathbf{X}_S \mathbf{w}_S{}^2 - \mathbf{y} \right\|^2 + \lambda_n \left\| \mathbf{w}_S \right\|^2$$

If $\lambda_n = o(n)$, $\hat{\mathbf{w}}_{\lambda_n}$ is consistent: $L_{\mathcal{D}}(\hat{\mathbf{w}}_{\lambda_n}) \overset{n \to \infty}{\longrightarrow} \sigma^2$

$\hat{\mathbf{w}}_{MN}$ is consistent when d_S fixed, $d_J \to \infty$, $\lambda_n = o(n)$.

Could we have shown that with uniform convergence?

No uniform convergence on norm balls

Theorem: If $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\|\le\|\hat{\mathbf{w}}_{MN}\|} \left|L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})\right|\right] = \infty.$$

No uniform convergence on norm balls

Theorem: If $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\|\leq\|\hat{\mathbf{w}}_{MN}\|} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})| \right] = \infty.$$

No uniform convergence on norm balls

Theorem: If $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\|\le\|\hat{\mathbf{w}}_{MN}\|} \left|L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})\right|\right] = \infty.$$

No uniform convergence on norm balls

Theorem: If $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J \to \infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\| \le \|\hat{\mathbf{w}}_{MN}\|} \left| L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \right| \right] = \infty.$$

Proof idea:

No uniform convergence on norm balls

Theorem: If $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J \to \infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\| \le \|\hat{\mathbf{w}}_{MN}\|} \left| L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \right| \right] = \infty.$$

Proof idea:

$$L_{\mathcal{D}}(\mathbf{w}) = (\mathbf{w} - \mathbf{w}^*)^{\mathsf{T}} \boldsymbol{\Sigma} (\mathbf{w} - \mathbf{w}^*) + L_{\mathcal{D}}(\mathbf{w}^*)$$

No uniform convergence on norm balls

Theorem: If $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J \to \infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\| \le \|\hat{\mathbf{w}}_{MN}\|} \left| L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \right| \right] = \infty.$$

Proof idea:

$$L_{\mathcal{D}}(\mathbf{w}) = (\mathbf{w} - \mathbf{w}^*)^{\mathsf{T}} \boldsymbol{\Sigma} (\mathbf{w} - \mathbf{w}^*) + L_{\mathcal{D}}(\mathbf{w}^*)$$

$$L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) = (\mathbf{w} - \mathbf{w}^*)(\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}})(\mathbf{w} - \mathbf{w}^*) + \left(L_{\mathcal{D}}(\mathbf{w}^*) - L_{\mathbf{S}}(\mathbf{w}^*)\right) - \text{cross term}$$

No uniform convergence on norm balls

Theorem: If $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\|\le\|\hat{\mathbf{w}}_{MN}\|} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|\right] = \infty.$$

Proof idea:

$$L_{\mathcal{D}}(\mathbf{w}) = (\mathbf{w}-\mathbf{w}^*)^{\mathsf{T}}\boldsymbol{\Sigma}(\mathbf{w}-\mathbf{w}^*) + L_{\mathcal{D}}(\mathbf{w}^*)$$

$$L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) = (\mathbf{w}-\mathbf{w}^*)(\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}})(\mathbf{w}-\mathbf{w}^*) + (L_{\mathcal{D}}(\mathbf{w}^*) - L_{\mathbf{S}}(\mathbf{w}^*)) - \text{cross term}$$

$$\sup[\ldots] \ge \|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}\| \cdot (\|\hat{\mathbf{w}}_{MN}\| - \|\mathbf{w}^*\|)^2 + o(1)$$

No uniform convergence on norm balls

Theorem: If $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\|\le\|\hat{\mathbf{w}}_{MN}\|} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|\right] = \infty.$$

Proof idea:

$$L_{\mathcal{D}}(\mathbf{w}) = (\mathbf{w}-\mathbf{w}^*)^{\mathsf{T}}\boldsymbol{\Sigma}(\mathbf{w}-\mathbf{w}^*) + L_{\mathcal{D}}(\mathbf{w}^*)$$

$$L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) = (\mathbf{w}-\mathbf{w}^*)(\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}})(\mathbf{w}-\mathbf{w}^*) + (L_{\mathcal{D}}(\mathbf{w}^*) - L_{\mathbf{S}}(\mathbf{w}^*)) - \text{cross term}$$

$$\sup[\ldots] \ge \|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}\| \cdot (\|\hat{\mathbf{w}}_{MN}\| - \|\mathbf{w}^*\|)^2 + o(1)$$

$$\Theta\left(\frac{n}{\lambda_n}\right)$$

No uniform convergence on norm balls

Theorem: If $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\|\le\|\hat{\mathbf{w}}_{MN}\|} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|\right] = \infty.$$

Proof idea:

$$L_{\mathcal{D}}(\mathbf{w}) = (\mathbf{w}-\mathbf{w}^*)^{\mathsf{T}}\boldsymbol{\Sigma}(\mathbf{w}-\mathbf{w}^*) + L_{\mathcal{D}}(\mathbf{w}^*)$$

$$L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) = (\mathbf{w}-\mathbf{w}^*)(\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}})(\mathbf{w}-\mathbf{w}^*) + (L_{\mathcal{D}}(\mathbf{w}^*) - L_{\mathbf{S}}(\mathbf{w}^*)) - \text{cross term}$$

$$\sup[\ldots] \ge \|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}\| \cdot (\|\hat{\mathbf{w}}_{MN}\| - \|\mathbf{w}^*\|)^2 + o(1)$$

$$\Theta\left(\sqrt{\tfrac{\lambda_n}{n}}\right) \qquad \Theta\left(\tfrac{n}{\lambda_n}\right)$$

Koltchinskii/Lounici, Bernoulli 2017

No uniform convergence on norm balls

Theorem: If $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J \to \infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\| \le \|\hat{\mathbf{w}}_{MN}\|} \left| L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \right| \right] = \infty.$$

Proof idea:

$$L_{\mathcal{D}}(\mathbf{w}) = (\mathbf{w} - \mathbf{w}^*)^{\mathsf{T}} \boldsymbol{\Sigma} (\mathbf{w} - \mathbf{w}^*) + L_{\mathcal{D}}(\mathbf{w}^*)$$

$$L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) = (\mathbf{w} - \mathbf{w}^*)(\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}})(\mathbf{w} - \mathbf{w}^*) + \left(L_{\mathcal{D}}(\mathbf{w}^*) - L_{\mathbf{S}}(\mathbf{w}^*)\right) - \text{cross term}$$

$$\sup[\ldots] \ge \|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\| \cdot \left(\|\hat{\mathbf{w}}_{MN}\| - \|\mathbf{w}^*\|\right)^2 + o(1) \to \infty$$

$$\Theta\left(\sqrt{\frac{\lambda_n}{n}}\right) \qquad \Theta\left(\frac{n}{\lambda_n}\right)$$

Koltchinskii/Lounici, Bernoulli 2017

A more refined uniform convergence analysis?

$\{\mathbf{w} : \|\mathbf{w}\| \leq B\}$ is no good.

A more refined uniform convergence analysis?

$\{\mathbf{w} : \|\mathbf{w}\| \leq B\}$ is no good. Maybe $\{\mathbf{w} : A \leq \|\mathbf{w}\| \leq B\}$?

A more refined uniform convergence analysis?

$\{\mathbf{w} : \|\mathbf{w}\| \leq B\}$ is no good. Maybe $\{\mathbf{w} : A \leq \|\mathbf{w}\| \leq B\}$?

Uniform convergence may be unable to explain generalization in deep learning

---

Vaishnavh Nagarajan
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA
vaishnavh@cs.cmu.edu

J. Zico Kolter
Department of Computer Science
Carnegie Mellon University &
Bosch Center for Artificial Intelligence
Pittsburgh, PA
zkolter@cs.cmu.edu

A more refined uniform convergence analysis?

Theorem (à la [Nagarajan/Kolter, NeurIPS 2019]):

For each $\delta \in (0, \frac{1}{2})$, let $\mathbf{Pr}\left(\mathbf{S} \in \mathcal{S}_{n,\delta}\right) \geq 1 - \delta$,

A more refined uniform convergence analysis?

Theorem (à la [Nagarajan/Kolter, NeurIPS 2019]):

For each $\delta \in (0, \frac{1}{2})$, let $\mathbf{Pr}\left(\mathbf{S} \in \mathcal{S}_{n,\delta}\right) \geq 1 - \delta$,

$\hat{\mathbf{w}}$ a *natural* consistent interpolator,

Natural interpolators: $\hat{\mathbf{w}}_S$ doesn't change if $\mathbf{X}_J$ flips to $-\mathbf{X}_J$. Examples:

$$\hat{\mathbf{w}}_{MN}, \qquad \underset{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}}{\mathbf{argmin}} \|\mathbf{w}\|_1, \qquad \underset{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}}{\mathbf{argmin}} \|\mathbf{w} - \mathbf{w}^*\|_2,$$

$\underset{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}}{\mathbf{argmin}} f_S(\mathbf{w}_S) + f_J(\mathbf{w}_J)$ with each f convex, $f_J(-\mathbf{w}_J) = f_J(\mathbf{w}_J)$

A more refined uniform convergence analysis?

Theorem (à la [Nagarajan/Kolter, NeurIPS 2019]):

For each $\delta \in (0, \frac{1}{2})$, let $\mathbf{Pr}\left(\mathbf{S} \in \mathcal{S}_{n,\delta}\right) \geq 1 - \delta$,

$\hat{\mathbf{w}}$ a *natural* consistent interpolator,

and $\mathcal{W}_{n,\delta} = \{\hat{\mathbf{w}}(\mathbf{S}) : \mathbf{S} \in \mathcal{S}_{n,\delta}\}$.

Natural interpolators: $\hat{\mathbf{w}}_S$ doesn't change if $\mathbf{X}_J$ flips to $-\mathbf{X}_J$. Examples:

$$\hat{\mathbf{w}}_{MN}, \qquad \operatorname*{argmin}_{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}} \|\mathbf{w}\|_1, \qquad \operatorname*{argmin}_{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}} \|\mathbf{w} - \mathbf{w}^*\|_2,$$

$\operatorname*{argmin}_{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}} f_S(\mathbf{w}_S) + f_J(\mathbf{w}_J)$ with each f convex, $f_J(-\mathbf{w}_J) = f_J(\mathbf{w}_J)$

A more refined uniform convergence analysis?

Theorem (à la [Nagarajan/Kolter, NeurIPS 2019]):

For each $\delta \in (0, \frac{1}{2})$, let $\mathbf{Pr}\left(\mathbf{S} \in \mathcal{S}_{n,\delta}\right) \geq 1 - \delta$,

$\hat{\mathbf{w}}$ a *natural* consistent interpolator,

and $\mathcal{W}_{n,\delta} = \{\hat{\mathbf{w}}(\mathbf{S}) : \mathbf{S} \in \mathcal{S}_{n,\delta}\}$. Then, almost surely,

Natural interpolators: $\hat{\mathbf{w}}_S$ doesn't change if $\mathbf{X}_J$ flips to $-\mathbf{X}_J$. Examples:

$$\hat{\mathbf{w}}_{MN}, \qquad \operatorname*{argmin}_{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}} \|\mathbf{w}\|_1, \qquad \operatorname*{argmin}_{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}} \|\mathbf{w} - \mathbf{w}^*\|_2,$$

$$\operatorname*{argmin}_{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}} f_S(\mathbf{w}_S) + f_J(\mathbf{w}_J) \text{ with each } f \text{ convex, } f_J(-\mathbf{w}_J) = f_J(\mathbf{w}_J)$$

A more refined uniform convergence analysis?

Theorem (à la [Nagarajan/Kolter, NeurIPS 2019]):

For each $\delta \in (0, \frac{1}{2})$, let $\mathbf{Pr}\left(\mathbf{S} \in \mathcal{S}_{n,\delta}\right) \geq 1 - \delta$,

$\hat{\mathbf{w}}$ a *natural* consistent interpolator,

and $\mathcal{W}_{n,\delta} = \{\hat{\mathbf{w}}(\mathbf{S}) : \mathbf{S} \in \mathcal{S}_{n,\delta}\}$. Then, almost surely,

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \sup_{\mathbf{S}\in\mathcal{S}_{n,\delta}} \sup_{\mathbf{w}\in\mathcal{W}_{n,\delta}} \left|L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})\right| \geq 3\sigma^2.$$

([Negrea/Dziugaite/Roy, ICML 2020] had a very similar result for $\hat{\mathbf{w}}_{MN}$)

Natural interpolators: $\hat{\mathbf{w}}_S$ doesn't change if $\mathbf{X}_J$ flips to $-\mathbf{X}_J$. Examples:

$$\hat{\mathbf{w}}_{MN}, \qquad \operatorname*{argmin}_{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}} \|\mathbf{w}\|_1, \qquad \operatorname*{argmin}_{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}} \|\mathbf{w} - \mathbf{w}^*\|_2,$$

$$\operatorname*{argmin}_{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}} f_S(\mathbf{w}_S) + f_J(\mathbf{w}_J) \text{ with each } f \text{ convex, } f_J(-\mathbf{w}_J) = f_J(\mathbf{w}_J)$$

A more refined uniform convergence analysis?

Theorem (à la [Nagarajan/Kolter, NeurIPS 2019]):

For each $\delta \in (0, \frac{1}{2})$, let $\mathbf{Pr}\left(\mathbf{S} \in \mathcal{S}_{n,\delta}\right) \geq 1 - \delta$,

$\hat{\mathbf{w}}$ a *natural* consistent interpolator,

and $\mathcal{W}_{n,\delta} = \{\hat{\mathbf{w}}(\mathbf{S}) : \mathbf{S} \in \mathcal{S}_{n,\delta}\}$. Then, almost surely,

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \sup_{\mathbf{S}\in\mathcal{S}_{n,\delta}} \sup_{\mathbf{w}\in\mathcal{W}_{n,\delta}} \left|L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})\right| \geq 3\sigma^2.$$

Proof shows that for most $\mathbf{S}$,

there's a typical predictor $\mathbf{w}$ (in $\mathcal{W}_{n,\delta}$)

that's good on most inputs ($L_{\mathcal{D}}(\mathbf{w}) \to \sigma^2$),

but very bad on *specifically* $\mathbf{S}$ ($L_{\mathbf{S}}(\mathbf{w}) \to 4\sigma^2$)

So, what are we left with?

So, what are we left with?

- Convergence of surrogates [Negrea/Dziugaite/Roy, ICML 2020]?

So, what are we left with?

- Convergence of surrogates [Negrea/Dziugaite/Roy, ICML 2020]?
  - Nice, but not really the same thing...

So, what are we left with?

- Convergence of surrogates [Negrea/Dziugaite/Roy, ICML 2020]?
  - Nice, but not really the same thing...
- Give up?

So, what are we left with?

- Convergence of surrogates [Negrea/Dziugaite/Roy, ICML 2020]?
  - Nice, but not really the same thing…
- Give up?
- Or…

One-sided uniform convergence?

We don't really care about small $L_{\mathcal{D}}$, big $L_{\mathbf{S}}$....
Could we bound $\mathbf{sup}\, L_{\mathcal{D}} - L_{\mathbf{S}}$ instead of $\mathbf{sup}\left|L_{\mathcal{D}} - L_{\mathbf{S}}\right|$?

One-sided uniform convergence?

We don't really care about small $L_{\mathcal{D}}$, big $L_{\mathbf{S}}$ ....
Could we bound $\sup L_{\mathcal{D}} - L_{\mathbf{S}}$ instead of $\sup \left| L_{\mathcal{D}} - L_{\mathbf{S}} \right|$?

- Existing uniform convergence proofs are “really” about $\left| L_{\mathcal{D}} - L_{\mathbf{S}} \right|$ [Nagarajan/Kolter, NeurIPS 2019]

One-sided uniform convergence?

We don't really care about small $L_{\mathcal{D}}$, big $L_{\mathbf{S}}$ ....
Could we bound $\sup L_{\mathcal{D}} - L_{\mathbf{S}}$ instead of $\sup \left| L_{\mathcal{D}} - L_{\mathbf{S}} \right|$?

- Existing uniform convergence proofs are “really” about $\left| L_{\mathcal{D}} - L_{\mathbf{S}} \right|$ [Nagarajan/Kolter, NeurIPS 2019]
- Strongly expect still ∞ for norm balls in our testbed
  - $\lambda_{\max}\left(\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\right)$ instead of $\left\| \boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}} \right\|$

One-sided uniform convergence?

We don't really care about small $L_{\mathcal{D}}$, big $L_{\mathbf{S}}$....
Could we bound $\sup L_{\mathcal{D}} - L_{\mathbf{S}}$ instead of $\sup \left| L_{\mathcal{D}} - L_{\mathbf{S}} \right|$?

- Existing uniform convergence proofs are "really" about $\left| L_{\mathcal{D}} - L_{\mathbf{S}} \right|$ [Nagarajan/Kolter, NeurIPS 2019]
- Strongly expect still ∞ for norm balls in our testbed
  - $\lambda_{\max}\left(\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\right)$ instead of $\left\| \boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}} \right\|$
- Not possible to show $\sup_{f \in \mathcal{F}} L_{\mathcal{D}} - L_{\mathbf{S}}$ is big for *all* $\mathcal{F}$
  - If $\hat{f}$ consistent and $\inf_f L_{\mathbf{S}}(f) \geq 0$, use $\mathcal{F} = \{f : L_{\mathcal{D}}(f) \leq L_{\mathcal{D}}(f^*) + \epsilon_{n,\delta}\}$

A broader view of uniform convergence

So far, used $L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \leq \sup_{\|\mathbf{w}\| \leq B} \left| L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \right|$

A broader view of uniform convergence

So far, used $L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \leq \sup_{\|\mathbf{w}\| \leq B} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|$

But *we only care about interpolators*. How about

$$\sup_{\|\mathbf{w}\| \leq B,\ L_{\mathbf{S}}(\mathbf{w}) = 0} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|?$$

A broader view of uniform convergence

So far, used $L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \leq \sup_{\|\mathbf{w}\| \leq B} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|$

But *we only care about interpolators*. How about

$$\sup_{\|\mathbf{w}\| \leq B,\ L_{\mathbf{S}}(\mathbf{w}) = 0} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|?$$

A broader view of uniform convergence

So far, used $L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \leq \sup_{\|\mathbf{w}\| \leq B} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|$

But *we only care about interpolators*. How about

$$\sup_{\|\mathbf{w}\| \leq B,\ L_{\mathbf{S}}(\mathbf{w}) = 0} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|?$$

Is this “uniform convergence”?

A broader view of uniform convergence

So far, used $L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \leq \sup_{\|\mathbf{w}\| \leq B} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|$

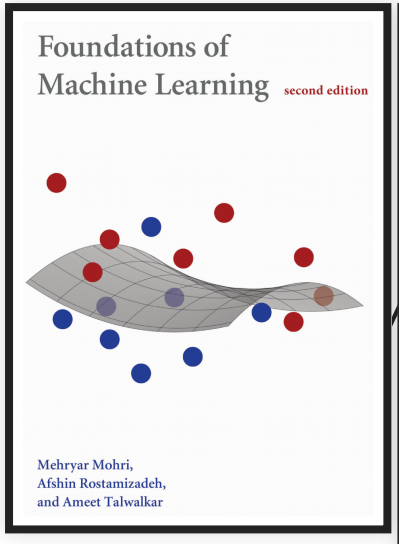
But *we only care about interpolators*. How about

$$\sup_{\|\mathbf{w}\| \leq B,\ L_{\mathbf{S}}(\mathbf{w}) = 0} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|?$$

Is this “uniform convergence”?

It's the standard notion for realizable ($L_{\mathcal{D}}(w^*) = 0$) analyses...

A broader view of uniform convergence



It's the stand

In the example of axis-aligned rectangles that we examined, the hypothesis h_S returned by the algorithm was always *consistent*, that is, it admitted no error on the training sample S. In this section, we present a general sample complexity bound, or equivalently, a generalization bound, for consistent hypotheses, in the case where the cardinality $|H|$ of the hypothesis set is finite. Since we consider consistent hypotheses, we will assume that the target concept c is in H.

Theorem 2.1 **Learning bounds — finite H, consistent case**
Let H be a finite set of functions mapping from $\mathcal{X}$ to $\mathcal{Y}$. Let $\mathcal{A}$ be an algorithm that for any target concept $c \in H$ and i.i.d. sample S returns a consistent hypothesis h_S: $\widehat{R}(h_S) = 0$. Then, for any $\epsilon, \delta > 0$, the inequality $\Pr_{S \sim D^m}[R(h_S) \leq \epsilon] \geq 1 - \delta$ holds if

$$m \geq \frac{1}{\epsilon}\left(\log |H| + \log \frac{1}{\delta}\right). \tag{2.8}$$

This sample complexity result admits the following equivalent statement as a generalization bound: for any $\epsilon, \delta > 0$, with probability at least $1 - \delta$,

$$R(h_S) \leq \frac{1}{m}\left(\log |H| + \log \frac{1}{\delta}\right). \tag{2.9}$$

Proof Fix $\epsilon > 0$. We do not know which consistent hypothesis $h_S \in H$ is selected by the algorithm $\mathcal{A}$. This hypothesis further depends on the training sample S. Therefore, we need to give a *uniform convergence bound*, that is, a bound that holds for the set of all consistent hypotheses, which a fortiori includes h_S. Thus,

A broader view of uniform convergence

So far, used $L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \leq \sup_{\|\mathbf{w}\| \leq B} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|$

But *we only care about interpolators*. How about

$$\sup_{\|\mathbf{w}\| \leq B,\ L_{\mathbf{S}}(\mathbf{w}) = 0} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|?$$

Is this “uniform convergence”?

It's the standard notion for realizable ($L_{\mathcal{D}}(w^*) = 0$) analyses...

A broader view of uniform convergence

So far, used $L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \leq \sup_{\|\mathbf{w}\| \leq B} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|$

But *we only care about interpolators*. How about

$$\sup_{\|\mathbf{w}\| \leq B,\ L_{\mathbf{S}}(\mathbf{w}) = 0} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|?$$

Is this “uniform convergence”?

It's the standard notion for realizable ($L_{\mathcal{D}}(w^*) = 0$) analyses...

Are there analyses like this for $L_{\mathcal{D}}(w^*) > 0$?

Optimistic rates

Applying [Srebro/Sridharan/Tewari 2010]: for all $\|\mathbf{w}\| \leq B$,

ξ_n: high-prob bound on $\max_{i=1,\ldots,n} \|\mathbf{x}_i\|^2$

$$L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \leq \tilde{\mathcal{O}}_P\left(\frac{B^2\xi_n}{n} + \sqrt{L_{\mathbf{S}}(\mathbf{w})\frac{B^2\xi_n}{n}}\right)$$

Optimistic rates

Applying [Srebro/Sridharan/Tewari 2010]: for all $\|\mathbf{w}\| \leq B$,

ξ_n: high-prob bound on $\max_{i=1,\ldots,n} \|\mathbf{x}_i\|^2$

$$L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \leq \tilde{\mathcal{O}}_P\left(\frac{B^2\xi_n}{n} + \sqrt{L_{\mathbf{S}}(\mathbf{w})\frac{B^2\xi_n}{n}}\right)$$

$$\sup_{\|\mathbf{w}\|\leq B,\ L_{\mathbf{S}}(\mathbf{w})=0} L_{\mathcal{D}}(\mathbf{w}) \leq c\frac{B^2\xi_n}{n} + o_P(1)$$

Optimistic rates

Applying [Srebro/Sridharan/Tewari 2010]: for all $\|\mathbf{w}\| \le B$,

ξ_n: high-prob bound on $\max_{i=1,\ldots,n} \|\mathbf{x}_i\|^2$

$$L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \le \tilde{\mathcal{O}}_P\left(\frac{B^2\xi_n}{n} + \sqrt{L_{\mathbf{S}}(\mathbf{w})\frac{B^2\xi_n}{n}}\right)$$

$$\sup_{\|\mathbf{w}\|\le B,\ L_{\mathbf{S}}(\mathbf{w})=0} L_{\mathcal{D}}(\mathbf{w}) \le c\frac{B^2\xi_n}{n} + o_P(1)$$

if $1 \ll \lambda_n \ll n$, $B = \|\hat{\mathbf{w}}_{MN}\|$, $\quad \to c\, L_{\mathcal{D}}(\mathbf{w}^*)$

Optimistic rates

Applying [Srebro/Sridharan/Tewari 2010]: for all $\|\mathbf{w}\| \leq B$,

ξ_n: high-prob bound on $\max_{i=1,\ldots,n} \|\mathbf{x}_i\|^2$

$$L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \leq \tilde{\mathcal{O}}_P\left(\frac{B^2\xi_n}{n} + \sqrt{L_{\mathbf{S}}(\mathbf{w})\frac{B^2\xi_n}{n}}\right)$$

$$c \leq 200,000 \log^3(n)$$

$$\sup_{\|\mathbf{w}\| \leq B,\ L_{\mathbf{S}}(\mathbf{w})=0} L_{\mathcal{D}}(\mathbf{w}) \leq c\frac{B^2\xi_n}{n} + o_P(1)$$

if $1 \ll \lambda_n \ll n$, $B = \|\hat{\mathbf{w}}_{MN}\|$, $\quad \to c\, L_{\mathcal{D}}(\mathbf{w}^*)$

Optimistic rates

Applying [Srebro/Sridharan/Tewari 2010]: for all $\|\mathbf{w}\| \leq B$,

ξ_n: high-prob bound on $\max_{i=1,\ldots,n} \|\mathbf{x}_i\|^2$

$$L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \leq \tilde{\mathcal{O}}_P\left(\frac{B^2\xi_n}{n} + \sqrt{L_{\mathbf{S}}(\mathbf{w})\frac{B^2\xi_n}{n}}\right)$$

$$c \leq 200,000\ \log^3(n)$$

$$\sup_{\|\mathbf{w}\|\leq B,\ L_{\mathbf{S}}(\mathbf{w})=0} L_{\mathcal{D}}(\mathbf{w}) \leq c\frac{B^2\xi_n}{n} + o_P(1)$$

if $1 \ll \lambda_n \ll n$, $B = \|\hat{\mathbf{w}}_{MN}\|$, $\quad \to c\, L_{\mathcal{D}}(\mathbf{w}^*)$

But if we suppose $c = 1$, would get a novel prediction:

$$\sup_{\|\mathbf{w}\|\leq\alpha\|\hat{\mathbf{w}}_{MN}\|,\ L_{\mathbf{S}}(\mathbf{w})=0} L_{\mathcal{D}}(\mathbf{w}) \leq \alpha^2\left[\sigma^2 + o_P(1)\right]$$

Main result

Theorem: If $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \mathbb{E}\left[\sup_{\substack{\|\mathbf{w}\|\le\alpha\|\hat{\mathbf{w}}_{MN}\| \\ L_{\mathbf{S}}(\mathbf{w})=0}} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})| \right] = \alpha^2 \, L_{\mathcal{D}}(\mathbf{w}^*)$$

Main result

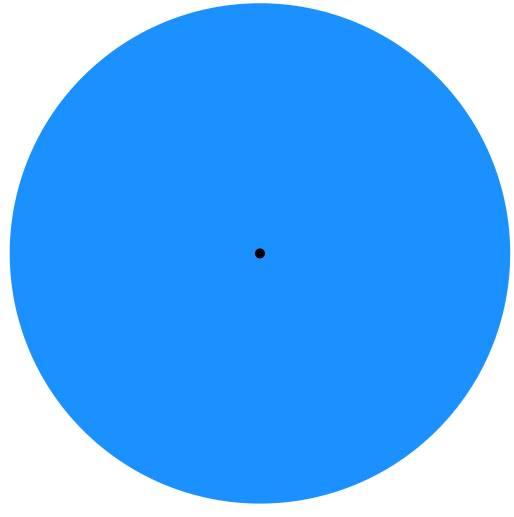
Theorem: If $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \mathbb{E}\left[\sup_{\substack{\|\mathbf{w}\|\le\alpha\|\hat{\mathbf{w}}_{MN}\| \\ L_{\mathbf{S}}(\mathbf{w})=0}} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})| \right] = \alpha^2 \, L_{\mathcal{D}}(\mathbf{w}^*)$$

- Confirms speculation based on $c = 1$ assumption
- Shows consistency with uniform convergence (of interpolators)
- New result for error of not-quite-minimal-norm interpolators
  - Norm $\|\hat{\mathbf{w}}_{MN}\| + \mathbf{const}$ is asympotically consistent
  - Norm $1.1\|\hat{\mathbf{w}}_{MN}\|$ is at worst $1.21\, L_{\mathcal{D}}(\mathbf{w}^*)$

What does $\{\mathbf{w} : \|\mathbf{w}\| \leq B,\ L_{\mathbf{S}}(\mathbf{w}) = 0\}$ look like?

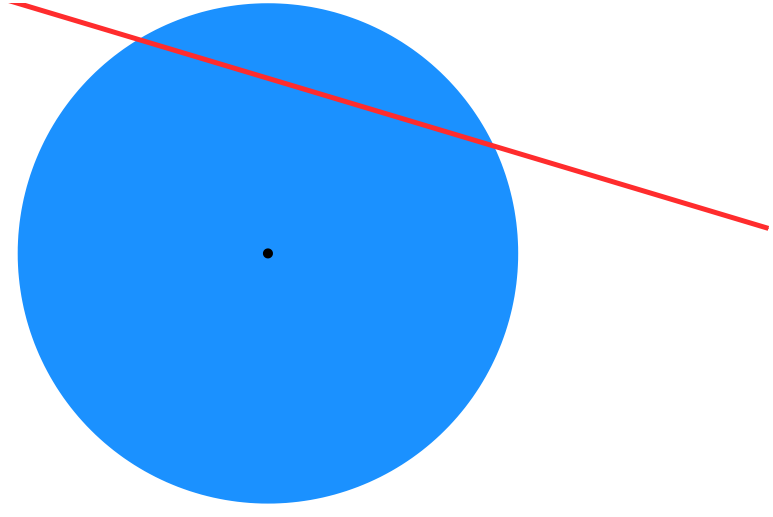
What does $\{\mathbf{w} : \|\mathbf{w}\| \leq B,\ L_{\mathbf{S}}(\mathbf{w}) = 0\}$ look like?



Intersection of d-ball

What does $\{\mathbf{w} : \|\mathbf{w}\| \leq B,\ L_{\mathbf{S}}(\mathbf{w}) = 0\}$ look like?

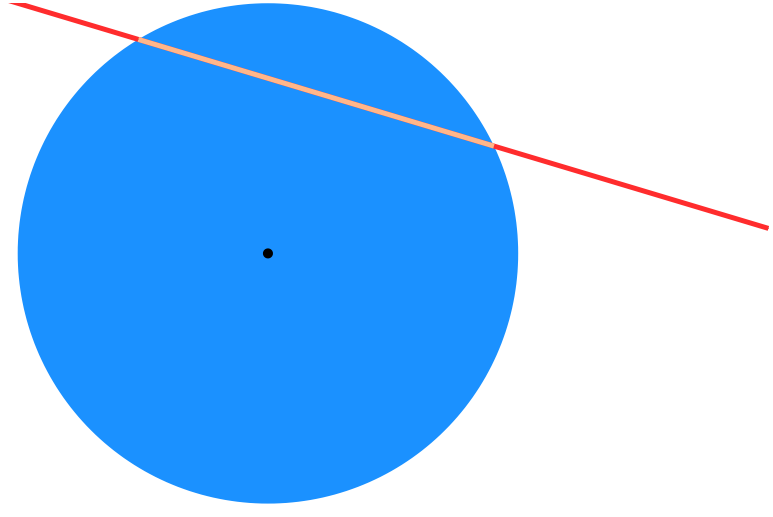
$\{\mathbf{w} : L_{\mathbf{S}}(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 = 0\}$ is the plane $\mathbf{X}\mathbf{w} = \mathbf{y}$



Intersection of d-ball with $(d-n)$-hyperplane:

What does $\{\mathbf{w} : \|\mathbf{w}\| \leq B,\ L_{\mathbf{S}}(\mathbf{w}) = 0\}$ look like?

$\{\mathbf{w} : L_{\mathbf{S}}(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 = 0\}$ is the plane $\mathbf{X}\mathbf{w} = \mathbf{y}$

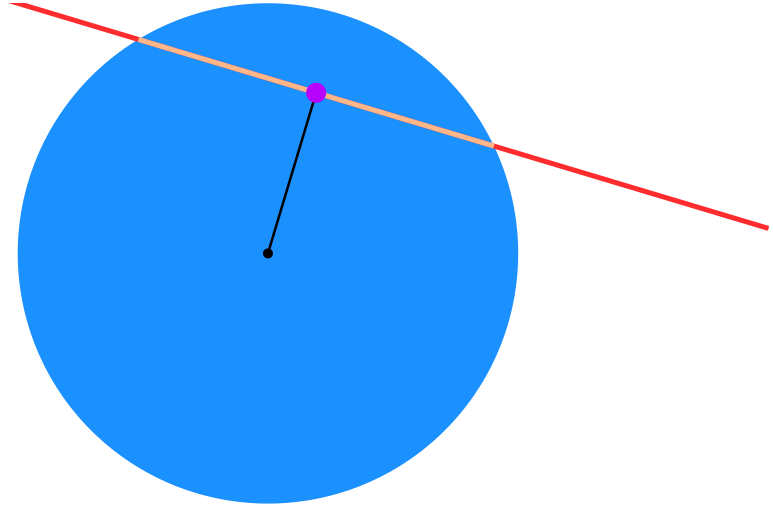


Intersection of d-ball with $(d-n)$-hyperplane:

$(d-n)$-ball

What does $\{\mathbf{w} : \|\mathbf{w}\| \leq B,\ L_{\mathbf{S}}(\mathbf{w}) = 0\}$ look like?

$\{\mathbf{w} : L_{\mathbf{S}}(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 = 0\}$ is the plane $\mathbf{X}\mathbf{w} = \mathbf{y}$

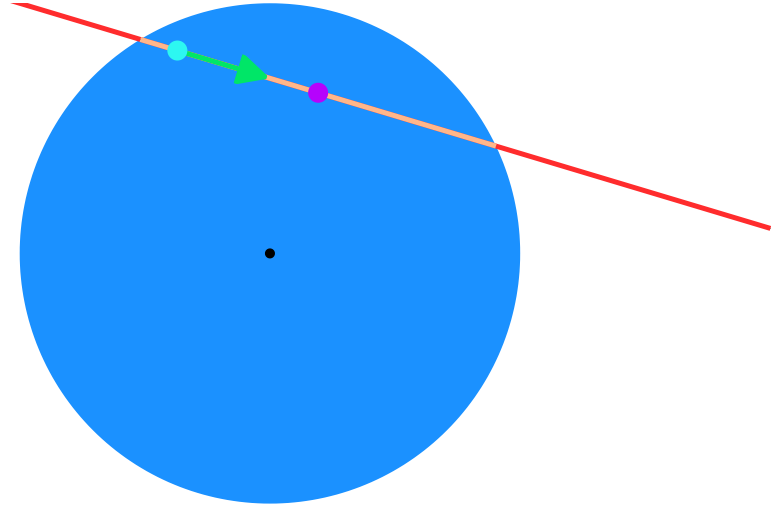


Intersection of d-ball with $(d-n)$-hyperplane:

$(d-n)$-ball centered at $\hat{\mathbf{w}}_{MN}$

What does $\{\mathbf{w} : \|\mathbf{w}\| \leq B,\ L_{\mathbf{S}}(\mathbf{w}) = 0\}$ look like?

$\{\mathbf{w} : L_{\mathbf{S}}(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 = 0\}$ is the plane $\mathbf{X}\mathbf{w} = \mathbf{y}$



Intersection of d-ball with $(d-n)$-hyperplane:
$(d-n)$-ball centered at $\hat{\mathbf{w}}_{MN}$

Can write as $\{\hat{\mathbf{w}} + \mathbf{F}\mathbf{z} : \mathbf{z} \in \mathbb{R}^{d-n},\ \|\hat{\mathbf{w}} + \mathbf{F}\mathbf{z}\| \leq B\}$
where $\hat{\mathbf{w}}$ is *any* interpolator, $\mathbf{F}$ is basis for $\mathbf{ker}(\mathbf{X})$

Decomposition via strong duality

Can change variables in $\sup_{\mathbf{w}:\|\mathbf{w}\| \leq B,\, L_{\mathbf{S}}(\mathbf{w})=0} L_{\mathcal{D}}(\mathbf{w})$ to

$$L_{\mathcal{D}}(\mathbf{w}^*) + \sup_{\mathbf{z}:\|\hat{\mathbf{w}}+\mathbf{F}\mathbf{z}\|^2 \leq B^2} (\hat{\mathbf{w}} + \mathbf{F}\mathbf{z} - w^*)^{\mathsf{T}} \boldsymbol{\Sigma} (\hat{\mathbf{w}} + \mathbf{F}\mathbf{z} - w^*)$$

Decomposition via strong duality

Can change variables in $\sup_{\mathbf{w}:\|\mathbf{w}\|\leq B,\, L_{\mathbf{S}}(\mathbf{w})=0} L_{\mathcal{D}}(\mathbf{w})$ to

$$L_{\mathcal{D}}(\mathbf{w}^*) + \sup_{\mathbf{z}:\|\hat{\mathbf{w}}+\mathbf{F}\mathbf{z}\|^2\leq B^2} (\hat{\mathbf{w}} + \mathbf{F}\mathbf{z} - w^*)^{\mathsf{T}} \boldsymbol{\Sigma} (\hat{\mathbf{w}} + \mathbf{F}\mathbf{z} - w^*)$$

Quadratic program, one quadratic constraint: strong duality

Decomposition via strong duality

Can change variables in $\sup_{\mathbf{w}:\|\mathbf{w}\|\leq B,\, L_{\mathbf{S}}(\mathbf{w})=0} L_{\mathcal{D}}(\mathbf{w})$ to

$$L_{\mathcal{D}}(\mathbf{w}^*) + \sup_{\mathbf{z}:\|\hat{\mathbf{w}}+\mathbf{F}\mathbf{z}\|^2\leq B^2} (\hat{\mathbf{w}} + \mathbf{F}\mathbf{z} - w^*)^{\mathsf{T}} \boldsymbol{\Sigma} (\hat{\mathbf{w}} + \mathbf{F}\mathbf{z} - w^*)$$

Quadratic program, one quadratic constraint: strong duality

Exactly equivalent to problem in *one* scalar variable:

$$L_{\mathcal{D}}(\hat{\mathbf{w}}) + \inf_{\mu > \|\mathbf{F}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{F}\|} \left\|\mathbf{F}^{\mathsf{T}}[\mu\hat{\mathbf{w}} - \Sigma(\hat{\mathbf{w}} - \mathbf{w}^*)]\right\|_{(\mu\mathbf{I}_{p-n} - \mathbf{F}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{F})^{-1}} + \mu(B^2 - \|\hat{\mathbf{w}}\|^2)$$

Decomposition via strong duality

Can change variables in $\sup_{\mathbf{w}:\|\mathbf{w}\| \leq B,\, L_{\mathbf{S}}(\mathbf{w})=0} L_{\mathcal{D}}(\mathbf{w})$ to

$$L_{\mathcal{D}}(\mathbf{w}^*) + \sup_{\mathbf{z}:\|\hat{\mathbf{w}}+\mathbf{F}\mathbf{z}\|^2 \leq B^2} (\hat{\mathbf{w}} + \mathbf{F}\mathbf{z} - w^*)^{\mathsf{T}} \boldsymbol{\Sigma} (\hat{\mathbf{w}} + \mathbf{F}\mathbf{z} - w^*)$$

Quadratic program, one quadratic constraint: strong duality

Exactly equivalent to problem in *one* scalar variable:

$$L_{\mathcal{D}}(\hat{\mathbf{w}}) + \inf_{\mu > \|\mathbf{F}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{F}\|} \left\| \mathbf{F}^{\mathsf{T}}[\mu\hat{\mathbf{w}} - \boldsymbol{\Sigma}(\hat{\mathbf{w}} - \mathbf{w}^*)] \right\|_{(\mu \mathbf{I}_{p-n} - \mathbf{F}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{F})^{-1}} + \mu(B^2 - \|\hat{\mathbf{w}}\|^2)$$

Can analyze this for different choices of $\hat{\mathbf{w}}$...

The minimal-risk interpolator

$$\hat{\mathbf{w}}_{MR} = \underset{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}}{\operatorname{argmin}} L_{\mathcal{D}}(\mathbf{w})$$

The minimal-risk interpolator

$$\begin{aligned}\hat{\mathbf{w}}_{MR} &= \operatorname*{argmin}_{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}} L_{\mathcal{D}}(\mathbf{w}) \\ &= \mathbf{w}^* + \Sigma^{-1}\mathbf{X}^\mathsf{T}(\mathbf{X}\Sigma^{-1}\mathbf{X}^\mathsf{T})^{-1}(Y - X\mathbf{w}^*)\end{aligned}$$

The minimal-risk interpolator

$$\begin{aligned}\hat{\mathbf{w}}_{MR} &= \operatorname*{argmin}_{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}} L_{\mathcal{D}}(\mathbf{w}) \\ &= \mathbf{w}^* + \Sigma^{-1}\mathbf{X}^\mathsf{T}(\mathbf{X}\Sigma^{-1}\mathbf{X}^\mathsf{T})^{-1}(Y - X\mathbf{w}^*)\end{aligned}$$

In Gaussian least squares generally, have that

$$\mathbb{E}\, L_{\mathcal{D}}(\hat{\mathbf{w}}_{MR}) = \frac{d-1}{d-1-n} L_{\mathcal{D}}(\mathbf{w}^*)$$

so $\hat{\mathbf{w}}_{MR}$ is consistent iff $n = o(d)$.

The minimal-risk interpolator

$$\begin{aligned}\hat{\mathbf{w}}_{MR} &= \operatorname*{argmin}_{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}} L_{\mathcal{D}}(\mathbf{w}) \\ &= \mathbf{w}^* + \Sigma^{-1}\mathbf{X}^\mathsf{T}(\mathbf{X}\Sigma^{-1}\mathbf{X}^\mathsf{T})^{-1}(Y - X\mathbf{w}^*)\end{aligned}$$

In Gaussian least squares generally, have that

$$\mathbb{E}\, L_{\mathcal{D}}(\hat{\mathbf{w}}_{MR}) = \frac{d-1}{d-1-n} L_{\mathcal{D}}(\mathbf{w}^*)$$

so $\hat{\mathbf{w}}_{MR}$ is consistent iff $n = o(d)$.

Very useful for lower bounds! [Muthukumar+ JSAIT 2020]

Restricted eigenvalue under interpolation

$$\kappa_{\mathbf{X}}(\boldsymbol{\Sigma}) = \sup_{\|\mathbf{w}\|=1,\ \mathbf{X}\mathbf{w}=\mathbf{0}} \mathbf{w}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{w}$$

Roughly, “how much” of $\boldsymbol{\Sigma}$ is “missed” by $\mathbf{X}$

Consistency up to $\left\|\hat{\mathbf{w}}_{MR}\right\|$

Analyzing dual with $\hat{\mathbf{w}}_{MR}$,
get without **any** distributional assumptions that

$$\sup_{\substack{\|\mathbf{w}\| \le \|\hat{\mathbf{w}}_{MR}\| \\ L_{\mathbf{S}}(\mathbf{w})=0}} L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\hat{\mathbf{w}}_{MR}) + \overset{1 \le \beta \le 4}{\beta}\, \kappa_X(\Sigma) \left[\|\hat{\mathbf{w}}_{MR}\|^2 - \|\hat{\mathbf{w}}_{MN}\|^2\right]$$

In the generic results, $L_{\mathcal{D}}$ means $L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\mathbf{w}^*) + (\mathbf{w}-\mathbf{w}^*)^{\mathsf{T}}\Sigma(\mathbf{w}-\mathbf{w}^*)$ for some $\mathbf{w}^*$

Consistency up to $\left\|\hat{\mathbf{w}}_{MR}\right\|$

Analyzing dual with $\hat{\mathbf{w}}_{MR}$,
get without **any** distributional assumptions that

$$\sup_{\substack{\|\mathbf{w}\| \le \|\hat{\mathbf{w}}_{MR}\| \\ L_{\mathbf{S}}(\mathbf{w})=0}} L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\hat{\mathbf{w}}_{MR}) + \beta\, \kappa_X(\Sigma) \left[\|\hat{\mathbf{w}}_{MR}\|^2 - \|\hat{\mathbf{w}}_{MN}\|^2\right]$$

$1 \le \beta \le 4$

(amount of missed energy) · (available norm)

In the generic results, $L_{\mathcal{D}}$ means $L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\mathbf{w}^*) + (\mathbf{w}-\mathbf{w}^*)^{\mathsf{T}}\Sigma(\mathbf{w}-\mathbf{w}^*)$ for some $\mathbf{w}^*$

Consistency up to $\left\|\hat{\mathbf{w}}_{MR}\right\|$

Analyzing dual with $\hat{\mathbf{w}}_{MR}$,
get without **any** distributional assumptions that

$$\sup_{\substack{\|\mathbf{w}\| \le \|\hat{\mathbf{w}}_{MR}\| \\ L_{\mathbf{S}}(\mathbf{w})=0}} L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\hat{\mathbf{w}}_{MR}) + \beta\, \kappa_X(\Sigma) \left[\|\hat{\mathbf{w}}_{MR}\|^2 - \|\hat{\mathbf{w}}_{MN}\|^2\right]$$

$1 \le \beta \le 4$

(amount of missed energy) · (available norm)

If $\hat{\mathbf{w}}_{MR}$ consistent, everything smaller-norm also consistent iff β term $\to \mathbf{0}$

In the generic results, $L_{\mathcal{D}}$ means $L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\mathbf{w}^*) + (\mathbf{w}-\mathbf{w}^*)^{\mathsf{T}}\Sigma(\mathbf{w}-\mathbf{w}^*)$ for some $\mathbf{w}^*$

Consistency up to $\left\|\hat{\mathbf{w}}_{MR}\right\|$

Analyzing dual with $\hat{\mathbf{w}}_{MR}$,
get without **any** distributional assumptions that

$$\sup_{\substack{\|\mathbf{w}\| \le \|\hat{\mathbf{w}}_{MR}\| \\ L_{\mathbf{S}}(\mathbf{w})=0}} L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\hat{\mathbf{w}}_{MR}) + \beta\, \kappa_X(\Sigma)\left[\|\hat{\mathbf{w}}_{MR}\|^2 - \|\hat{\mathbf{w}}_{MN}\|^2\right]$$

$1 \le \beta \le 4$

(amount of missed energy) · (available norm)

If $\hat{\mathbf{w}}_{MR}$ consistent, everything smaller-norm also consistent iff β term $\to \mathbf{0}$

In our setting:

$\hat{\mathbf{w}}_{MR}$ is consistent, $L_{\mathcal{D}}(\hat{\mathbf{w}}_{MR}) \to L_{\mathcal{D}}(\mathbf{w}^*)$

In the generic results, $L_{\mathcal{D}}$ means $L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\mathbf{w}^*) + (\mathbf{w}-\mathbf{w}^*)^{\mathsf{T}}\Sigma(\mathbf{w}-\mathbf{w}^*)$ for some $\mathbf{w}^*$

Consistency up to $\left\|\hat{\mathbf{w}}_{MR}\right\|$

Analyzing dual with $\hat{\mathbf{w}}_{MR}$,
get without **any** distributional assumptions that

$$\sup_{\substack{\|\mathbf{w}\| \le \|\hat{\mathbf{w}}_{MR}\| \\ L_{\mathbf{S}}(\mathbf{w})=0}} L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\hat{\mathbf{w}}_{MR}) + \beta\, \kappa_X(\Sigma)\left[\|\hat{\mathbf{w}}_{MR}\|^2 - \|\hat{\mathbf{w}}_{MN}\|^2\right] \qquad 1 \le \beta \le 4$$

(amount of missed energy) · (available norm)

If $\hat{\mathbf{w}}_{MR}$ consistent, everything smaller-norm also consistent iff β term $\to 0$

In our setting:

$\hat{\mathbf{w}}_{MR}$ is consistent, $L_{\mathcal{D}}(\hat{\mathbf{w}}_{MR}) \to L_{\mathcal{D}}(\mathbf{w}^*)$

$$\kappa_X(\boldsymbol{\Sigma}) \approx \frac{\lambda_n}{n} \qquad \mathbb{E}\left[\|\hat{\mathbf{w}}_{MR}\|^2 - \|\hat{\mathbf{w}}_{MN}\|^2\right] = \frac{\sigma^2 d_S}{\lambda_n} + o(1)$$

In the generic results, $L_{\mathcal{D}}$ means $L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\mathbf{w}^*) + (\mathbf{w}-\mathbf{w}^*)^{\mathsf{T}}\Sigma(\mathbf{w}-\mathbf{w}^*)$ for some $\mathbf{w}^*$

Consistency up to $\left\|\hat{\mathbf{w}}_{MR}\right\|$

Analyzing dual with $\hat{\mathbf{w}}_{MR}$,
get without **any** distributional assumptions that

$$\sup_{\substack{\|\mathbf{w}\| \le \|\hat{\mathbf{w}}_{MR}\| \\ L_{\mathbf{S}}(\mathbf{w})=0}} L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\hat{\mathbf{w}}_{MR}) + \beta\, \kappa_X(\Sigma) \left[\|\hat{\mathbf{w}}_{MR}\|^2 - \|\hat{\mathbf{w}}_{MN}\|^2\right]$$

$1 \le \beta \le 4$

(amount of missed energy) · (available norm)

If $\hat{\mathbf{w}}_{MR}$ consistent, everything smaller-norm also consistent iff β term $\to 0$

In our setting:

$\hat{\mathbf{w}}_{MR}$ is consistent, $L_{\mathcal{D}}(\hat{\mathbf{w}}_{MR}) \to L_{\mathcal{D}}(\mathbf{w}^*)$

$$\kappa_X(\boldsymbol{\Sigma}) \approx \frac{\lambda_n}{n} \qquad \mathbb{E}\left[\|\hat{\mathbf{w}}_{MR}\|^2 - \|\hat{\mathbf{w}}_{MN}\|^2\right] = \frac{\sigma^2 d_S}{\lambda_n} + o(1)$$

Plugging in: $\mathbb{E} \sup_{\|\mathbf{w}\| \le \|\hat{\mathbf{w}}_{MR}\|,\, L_{\mathbf{S}}(\mathbf{w})=0} L_{\mathcal{D}}(\mathbf{w}) \to L_{\mathcal{D}}(\mathbf{w}^*)$

In the generic results, $L_{\mathcal{D}}$ means $L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\mathbf{w}^*) + (\mathbf{w}-\mathbf{w}^*)^{\mathsf{T}}\Sigma(\mathbf{w}-\mathbf{w}^*)$ for some $\mathbf{w}^*$

Error up to $\alpha \left\| \hat{\mathbf{w}}_{MN} \right\|$

Analyzing dual with $\hat{\mathbf{w}}_{MN}$ for $\hat{\mathbf{w}}$, $\alpha \geq 1$, get in general:

$$\sup_{\substack{\|\mathbf{w}\| \leq \alpha \|\hat{\mathbf{w}}_{MN}\| \\ L_{\mathbf{S}}(\mathbf{w}) = 0}} L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\hat{\mathbf{w}}_{MN}) + (\alpha^2 - 1)\, \kappa_X(\Sigma)\, \|\hat{\mathbf{w}}_{MN}\|^2 + R_n$$

$R_n \to 0$ if $\hat{\mathbf{w}}_{MN}$ is consistent

Error up to $\alpha\left\|\hat{\mathbf{w}}_{MN}\right\|$

Analyzing dual with $\hat{\mathbf{w}}_{MN}$ for $\hat{\mathbf{w}}$, $\alpha \geq 1$, get in general:

$$\sup_{\substack{\|\mathbf{w}\| \leq \alpha \|\hat{\mathbf{w}}_{MN}\| \\ L_{\mathbf{S}}(\mathbf{w}) = 0}} L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\hat{\mathbf{w}}_{MN}) + (\alpha^2 - 1)\,\kappa_X(\Sigma)\,\|\hat{\mathbf{w}}_{MN}\|^2 + R_n$$

$R_n \to 0$ if $\hat{\mathbf{w}}_{MN}$ is consistent

In our setting:

$\hat{\mathbf{w}}_{MN}$ is consistent, because $\left\|\hat{\mathbf{w}}_{MN}\right\| \leq \left\|\hat{\mathbf{w}}_{MR}\right\|$

Error up to $\alpha\left\|\hat{\mathbf{w}}_{MN}\right\|$

Analyzing dual with $\hat{\mathbf{w}}_{MN}$ for $\hat{\mathbf{w}}$, $\alpha \geq 1$, get in general:

$$\sup_{\substack{\|\mathbf{w}\| \leq \alpha \|\hat{\mathbf{w}}_{MN}\| \\ L_{\mathbf{S}}(\mathbf{w}) = 0}} L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\hat{\mathbf{w}}_{MN}) + (\alpha^2 - 1)\, \kappa_X(\Sigma)\, \|\hat{\mathbf{w}}_{MN}\|^2 + R_n$$

$R_n \to 0$ if $\hat{\mathbf{w}}_{MN}$ is consistent

In our setting:

$\hat{\mathbf{w}}_{MN}$ is consistent, because $\left\|\hat{\mathbf{w}}_{MN}\right\| \leq \left\|\hat{\mathbf{w}}_{MR}\right\|$

$$\mathbb{E}\, \kappa_{\mathbf{X}}(\boldsymbol{\Sigma})\, \|\hat{\mathbf{w}}_{MN}\|^2 \to \sigma^2 = L_{\mathcal{D}}(\mathbf{w}^*)$$

Error up to $\alpha \left\| \hat{\mathbf{w}}_{MN} \right\|$

Analyzing dual with $\hat{\mathbf{w}}_{MN}$ for $\hat{\mathbf{w}}$, $\alpha \geq 1$, get in general:

$$\sup_{\substack{\|\mathbf{w}\| \leq \alpha \|\hat{\mathbf{w}}_{MN}\| \\ L_{\mathbf{S}}(\mathbf{w}) = 0}} L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\hat{\mathbf{w}}_{MN}) + (\alpha^2 - 1)\, \kappa_X(\Sigma)\, \|\hat{\mathbf{w}}_{MN}\|^2 + R_n$$

$R_n \to 0$ if $\hat{\mathbf{w}}_{MN}$ is consistent

In our setting:

$\hat{\mathbf{w}}_{MN}$ is consistent, because $\left\| \hat{\mathbf{w}}_{MN} \right\| \leq \left\| \hat{\mathbf{w}}_{MR} \right\|$

$$\mathbb{E}\, \kappa_{\mathbf{X}}(\boldsymbol{\Sigma}) \left\| \hat{\mathbf{w}}_{MN} \right\|^2 \to \sigma^2 = L_{\mathcal{D}}(\mathbf{w}^*)$$

Plugging in: $\mathbb{E} \sup_{\|\mathbf{w}\| \leq \alpha \|\hat{\mathbf{w}}_{MN}\|,\, L_{\mathbf{S}}(\mathbf{w}) = 0} L_{\mathcal{D}}(\mathbf{w}) \to \alpha^2\, L_{\mathcal{D}}(\mathbf{w}^*)$

Error up to $\alpha \left\| \hat{\mathbf{w}}_{MN} \right\|$

Analyzing dual with $\hat{\mathbf{w}}_{MN}$ for $\hat{\mathbf{w}}$, $\alpha \geq 1$, get in general:

$$\sup_{\substack{\|\mathbf{w}\| \leq \alpha \|\hat{\mathbf{w}}_{MN}\| \\ L_{\mathbf{S}}(\mathbf{w}) = 0}} L_{\mathcal{D}}(\mathbf{w}) = L_{\mathcal{D}}(\hat{\mathbf{w}}_{MN}) + (\alpha^2 - 1)\, \kappa_X(\Sigma)\, \|\hat{\mathbf{w}}_{MN}\|^2 + R_n$$

$R_n \to 0$ if $\hat{\mathbf{w}}_{MN}$ is consistent

In our setting:

$\hat{\mathbf{w}}_{MN}$ is consistent, because $\left\| \hat{\mathbf{w}}_{MN} \right\| \leq \left\| \hat{\mathbf{w}}_{MR} \right\|$

$$\mathbb{E}\, \kappa_{\mathbf{X}}(\boldsymbol{\Sigma}) \left\| \hat{\mathbf{w}}_{MN} \right\|^2 \to \sigma^2 = L_{\mathcal{D}}(\mathbf{w}^*)$$

Plugging in: $\mathbb{E} \sup_{\|\mathbf{w}\| \leq \alpha \|\hat{\mathbf{w}}_{MN}\|,\, L_{\mathbf{S}}(\mathbf{w}) = 0} L_{\mathcal{D}}(\mathbf{w}) \to \alpha^2\, L_{\mathcal{D}}(\mathbf{w}^*)$

...and we're done!

On Uniform Convergence and Low-Norm Interpolation Learning

Zhou, Sutherland, and Srebro [NeurIPS 2020] [arXiv:2006.05942]

- "Regular" uniform convergence can't explain consistency of $\hat{\mathbf{w}}_{MN}$
  - Uniform convergence over norm ball can't show *any* learning
- An "interpolating" uniform convergence bound does
  - Shows low norm is sufficient for interpolation learning here
  - Predicts exact worst-case error as norm grows
- Optimistic/interpolating rates might be able to explain interpolation learning more broadly
  - Need to get the constants on leading terms *exactly* right!
- Analyzing generalization gap via duality may be broadly applicable
  - Can always get upper bounds via weak duality