Does any of this stuff even work?
Interpolation and the limits of uniform convergence

CPSC 532S: Modern Statistical Learning Theory

28 March 2022

cs.ubc.ca/~dsuth/532S/22/

Admin

- Reminder: A3 (with edited Q1) due **tonight**
- A4 (mostly on kernels) will be posted ASAP
  - Just trying not to have to replace questions...
- Final will be available for most of the finals period
  - Optional bonus questions, to boost your assignment grade

Admin

- Reminder: A3 (with edited Q1) due **tonight**
- A4 (mostly on kernels) will be posted ASAP
  - Just trying not to have to replace questions...
- Final will be available for most of the finals period
  - Optional bonus questions, to boost your assignment grade
- Course grade will be curved
  - Nobody who showed reasonable understanding will fail
  - (even for grad student definition of failing)

Admin

- Reminder: A3 (with edited Q1) due **tonight**
- A4 (mostly on kernels) will be posted ASAP
  - Just trying not to have to replace questions...
- Final will be available for most of the finals period
  - Optional bonus questions, to boost your assignment grade
- Course grade will be curved
  - Nobody who showed reasonable understanding will fail
  - (even for grad student definition of failing)
- Teaching evals available; due April 11th
  - But please read Mike Gelbart's Teaching evaluations: the good, the bad, and the ugly before doing any of them
  - Numerical scores used heavily despite systematic bias

Deep learning vs kernels

- We've seen some stabs at deep learning approximation, generalization, and optimization
- NTK models, all three: as width $\to \infty$, NNs “work”

Deep learning vs kernels

- We've seen some stabs at deep learning approximation, generalization, and optimization
- NTK models, all three: as width $\to \infty$, NNs “work”
- So…are NTK models (or some tweak) all we need?

Deep learning vs kernels

- We've seen some stabs at deep learning approximation, generalization, and optimization
- NTK models, all three: as width $\longrightarrow \infty$, NNs “work”
- So…are NTK models (or some tweak) all we need?
- Bunch of results saying **no**

On the Power and Limitations of Random Features for Understanding Neural Networks

Gilad Yehudai **Ohad Shamir**
Weizmann Institute of Science
{gilad.yehudai,ohad.shamir}@weizmann.ac.il

- Roughly: there is a $w^* \in \mathbb{R}^d$ with $\|w^*\| = d^2$, $b^* \in \mathbb{R}$ s.t.
  - if $\mathbb{E}_{x \sim \mathcal{N}(0,I)}[(f(x) - \mathrm{ReLU}(\langle w^*, x\rangle + b^*))^2] \leq \frac{1}{50}$,
  - then $\|f\|_{\mathrm{NTK}} \geq \exp(\Omega(d))$
- and if f's init is isotropic, true for any w^* with $\|w^*\| = d^2$
- But GD learns this (at linear rate) with $\mathrm{poly}(d)$ samples

Quantifying the Benefit of Using Differentiable Learning over Tangent Kernels

Eran Malach Hebrew University of Jerusalem eran.malach@mail.huji.ac.il
Pritish Kamath Toyota Technological Institute at Chicago pritish@ttic.edu
Emmanuel Abbe EPFL emmanuel.abbe@epfl.ch
Nathan Srebro Toyota Technological Institute at Chicago nati@ttic.edu
Collaboration on the Theoretical Foundations of Deep Learning (deepfoundations.ai)

<table>
<tr><th colspan="2"></th><th>NTK at same Initialization</th><th>NTK at alternate randomized Initialization</th><th>NTK of arbitrary model or even an arbitrary Kernel</th></tr>
<tr><td colspan="2">GD with unbiased initialization ($\forall_x f_{\theta_0}(x) = 0$) ensures small error</td><td>▸ NTK edge $\geq$ poly^{-1} (Thm. 1)
▸ NTK edge can be $<$ poly^{-1} while GD reaches 0 loss (Separation 1)</td><td colspan="2">Edge with any kernel can be $<$ poly^{-1} while GD reaches 0 loss (Separation 2)</td></tr>
<tr><td rowspan="2">GD with arbitrary init. ensures small error</td><td>Kernel (or alt init) can depend on input dist. $\mathcal{D}_{\mathcal{X}}$</td><td rowspan="2">NTK edge can be $= 0$ while GD reaches arb. low loss (Separation 3)</td><td>▸ NTK edge $\geq$ poly^{-1} (Thm. 2)
▸ NTK edge can be $<$ poly^{-1} while GD reaches 0 loss (Separation 2)</td><td>Edge can be $<$ poly^{-1} while GD reaches 0 loss (Separation 2)</td></tr>
<tr><td>Dist-indep kernels</td><td colspan="2">edge with any kernel can be $<$ exp^{-1} while GD reaches arb. low loss (Separation 4)</td></tr>
</table>

Okay, fine, NTKs aren't the (whole) answer.

Okay, fine, NTKs aren't the (whole) answer.

What if we assume approximation and optimization are fine?

Okay, fine, NTKs aren't the (whole) answer.

What if we assume approximation and optimization are fine?
Current generalization bounds empirically aren't tight enough,

Okay, fine, NTKs aren't the (whole) answer.

What if we assume approximation and optimization are fine?
Current generalization bounds empirically aren't tight enough,
but can we hope to prove a tighter one?

Remainder of today is roughly this talk I've given before:

Can Uniform Convergence Explain Interpolation Learning?

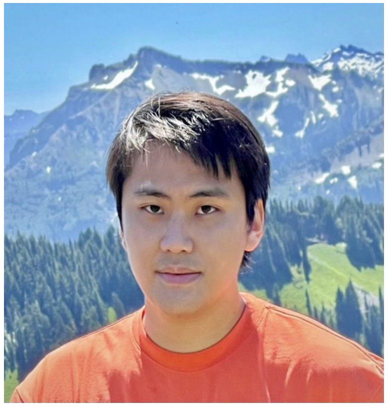
Danica J. Sutherland (she/her)
TTI-Chicago → UBC + Amii

based on [ZSS NeurIPS-20], [KZSS NeurIPS-21], [ZKSS 2021] with:

Lijia Zhou
UChicago

Frederic Koehler
MIT → Simons → Stanford

Nati Srebro
TTI-Chicago



Statistical learning theory

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

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

Statistical learning theory

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

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

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

Statistical learning theory

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

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

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

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

Statistical learning theory

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

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

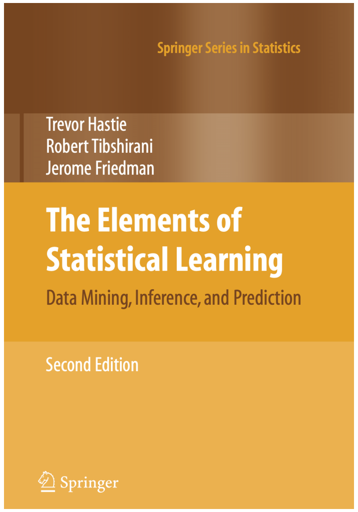
$C_{\mathcal{H},\delta}$ could be from Rademacher complexity, covering numbers, RKHS norm, VC dimension, fat-shattering dimension, ...

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

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

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”

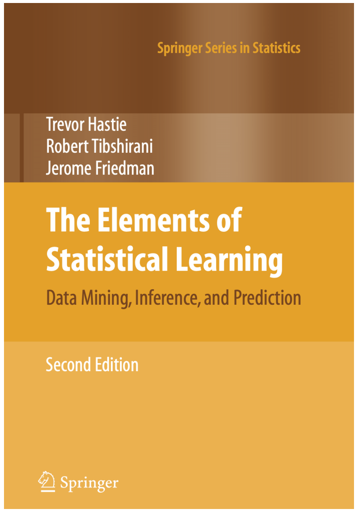


Table 1: The training and test accuracy (in percentage) of various models on the CIFAR10 dataset.

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

$$L_{\mathbf{S}}(\hat{h}) = 0; \, L_{\mathcal{D}}(\hat{h}) \approx 11\%$$

Interpolation learning

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

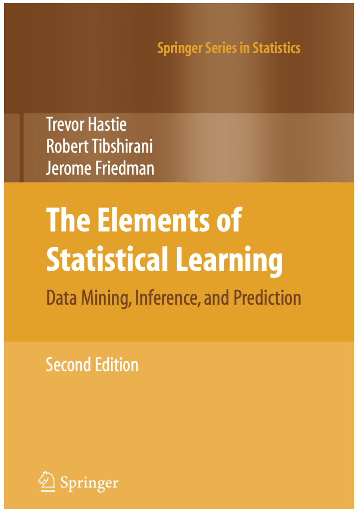


Table 1: The training and test accuracy (in percentage) of various models on the CIFAR10 dataset.

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 $L_{\mathbf{S}}(\hat{h}) = 0; L_{\mathcal{D}}(\hat{h}) \approx 11\%$

We'll call a model with $L_{\mathbf{S}}(h) = 0$ 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 $L_{\mathcal{D}}(h^*) > 0$)

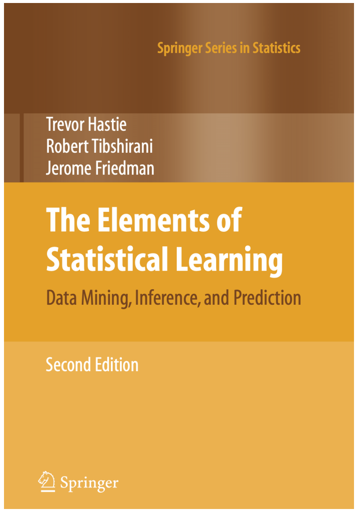


Table 1: The training and test accuracy (in percentage) of various models on the CIFAR10 dataset.

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

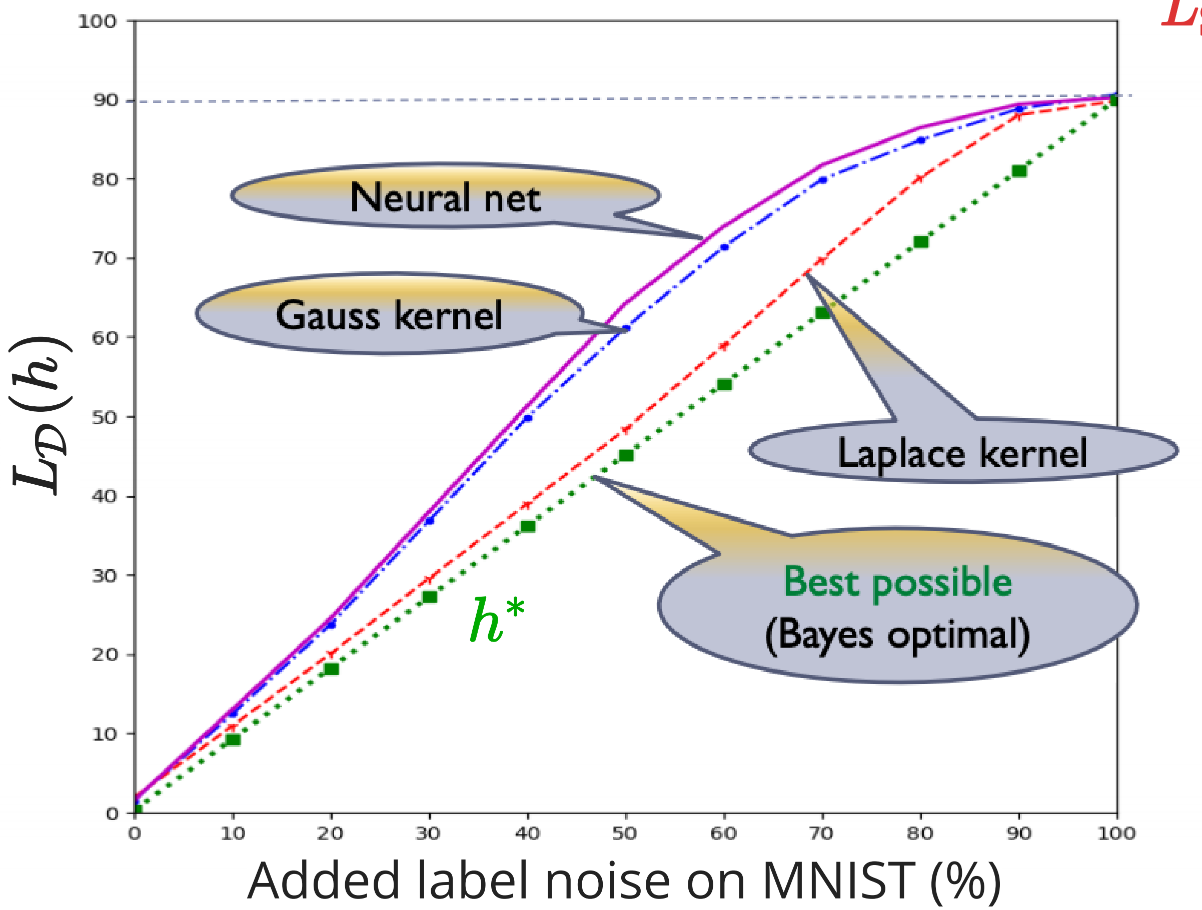
$$L_{\mathbf{S}}(\hat{h}) = 0;\ L_{\mathcal{D}}(\hat{h}) \approx 11\%$$

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

Interpolation does not overfit even for very noisy data

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

$$L_{\mathbf{S}}(\hat{h}) = 0$$



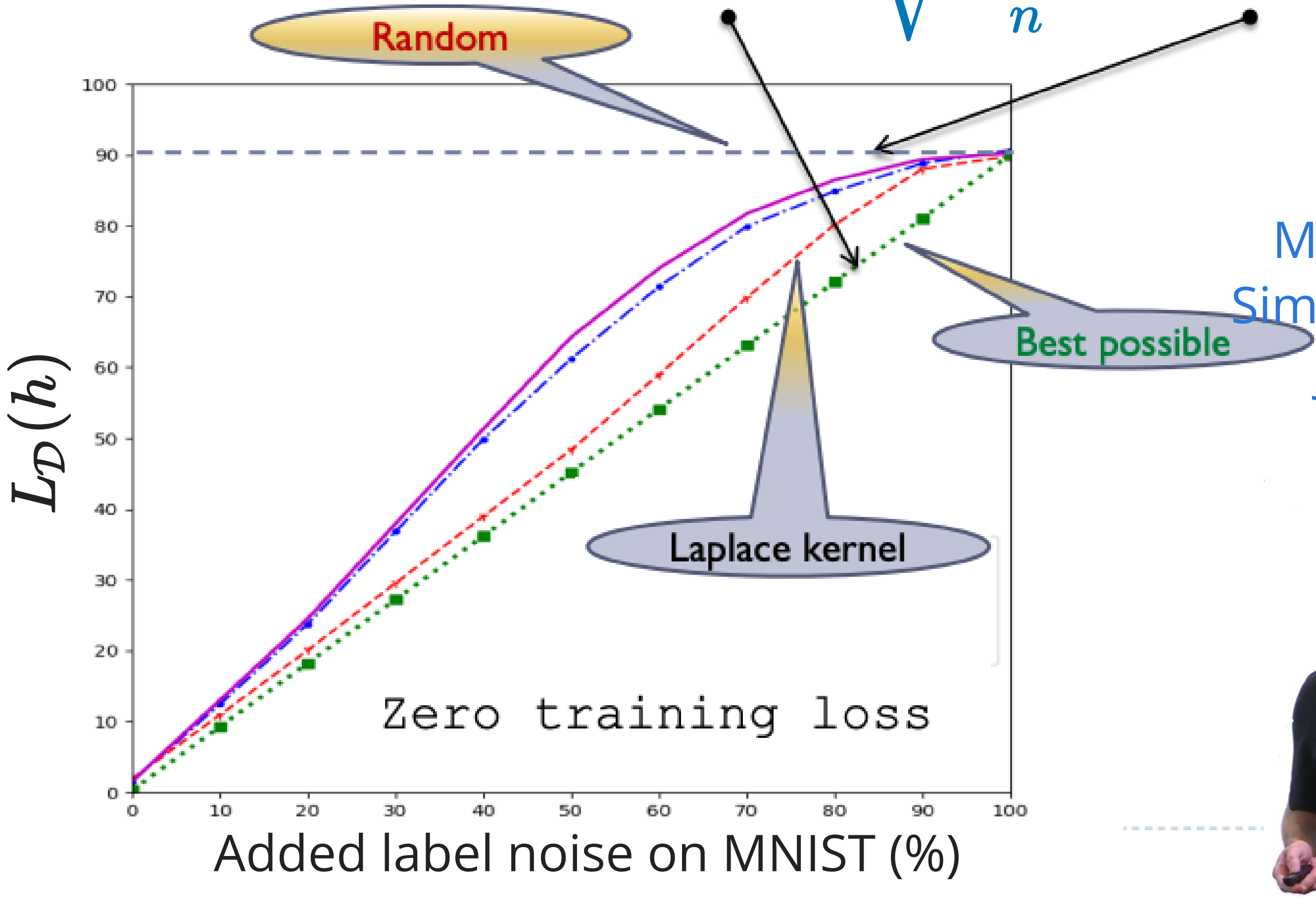
Misha Belkin
Simons Institute
July 2019

Belkin/Ma/Mandal, ICML 2018

Bounds? $L_{\mathcal{D}}(\hat{h}) \leq \underbrace{L_{\mathbf{S}}(\hat{h})}_{0} + \sqrt{\frac{C_{\mathcal{H},\delta}}{n}}$

What kind of generalization bound could work here?

$$0.7 < \sqrt{\frac{C_{\mathcal{H},\delta}}{n}} < 0.9$$



Misha Belkin
Simons Institute
July 2019

Belkin/Ma/Mandal, ICML 2018

Not a question of improving bounds

correct nontrivial

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

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.

$$L_{\mathcal{D}}(\hat{h}) \leq L_S(\hat{h}) + \text{bound}$$

WYSIWYG bounds:

training loss $= 0$

$\not\approx$

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

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

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 $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_{\mathbf{S}}(h)|$.

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

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

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…

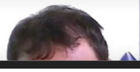
None* bound $\sup_{h \in \mathcal{H}} \left| L_{\mathcal{D}}(h) - L_{\mathbf{S}}(h) \right|$.

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 $\hat{h}$ to a surrogate predictor,

shows uniform convergence for the surrogate.

(Also, a few things since our first paper.)

A more specific version of the question

Today, we're mainly going to worry about *consistency*:

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

A more specific version of the question

Today, we're mainly going to worry about *consistency*:

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

...in a *noisy* setting: $L_{\mathcal{D}}(h^*) > 0$

A more specific version of the question

Today, we're mainly going to worry about *consistency*:

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

...in a *noisy* setting: $L_{\mathcal{D}}(h^*) > 0$

...for Gaussian linear regression:

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \Sigma) \quad y = \langle \mathbf{x}, w^* \rangle + \mathcal{N}(\mathbf{0}, \sigma^2) \quad L(y, \hat{y}) = (y - \hat{y})^2$$

A more specific version of the question

Today, we're mainly going to worry about *consistency*:

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

...in a *noisy* setting: $L_{\mathcal{D}}(h^*) > 0$

...for Gaussian linear regression:

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \Sigma) \quad y = \langle \mathbf{x}, w^* \rangle + \mathcal{N}(\mathbf{0}, \sigma^2) \quad L(y, \hat{y}) = (y - \hat{y})^2$$

Is it possible to show consistency of an interpolator with

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

A more specific version of the question

Today, we're mainly going to worry about *consistency*:

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

...in a *noisy* setting: $L_{\mathcal{D}}(h^*) > 0$

...for Gaussian linear regression:

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \Sigma) \quad y = \langle \mathbf{x}, w^* \rangle + \mathcal{N}(\mathbf{0}, \sigma^2) \quad L(y, \hat{y}) = (y - \hat{y})^2$$

Is it possible to show consistency of an interpolator with

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

This requires tight constants!

A testbed problem: “junk features”

	“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: $\mathbb{E}\|\mathbf{x}_J\|_2^2 = \lambda_n$

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

A testbed problem: "junk features"

	"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: $\mathbb{E}\|\mathbf{x}_J\|_2^2 = \lambda_n$

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

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

As $d_J \to \infty$, $\hat{\mathbf{w}}_{MN} = \mathbf{X}^\dagger \mathbf{y} \approx$ ridge regression on the signal

As $d_J \to \infty$, $\hat{\mathbf{w}}_{MN} = \mathbf{X}^\dagger \mathbf{y} \approx$ ridge regression on the signal

$$\hat{\mathbf{w}}_{\lambda_n} = \underset{\mathbf{w}_S \in \mathbb{R}^{d_S}}{\arg\min} \|\mathbf{X}_S \mathbf{w}_S{}^2 - \mathbf{y}\|^2 + \lambda_n \|\mathbf{w}_S\|^2$$

$$= \mathbf{X}_S^\top (\mathbf{X}_S \mathbf{X}_S^\top + \lambda_n I_n)^{-1} \mathbf{y}$$

As $d_J \to \infty$, $\hat{\mathbf{w}}_{MN} = \mathbf{X}^\dagger \mathbf{y} \approx$ ridge regression on the signal

$$\begin{aligned}\hat{\mathbf{w}}_{\lambda_n} &= \operatorname*{arg\,min}_{\mathbf{w}_S \in \mathbb{R}^{d_S}} \|\mathbf{X}_S \mathbf{w}_S{}^2 - \mathbf{y}\|^2 + \lambda_n \|\mathbf{w}_S\|^2 \\ &= \mathbf{X}_S^\mathsf{T} (\mathbf{X}_S \mathbf{X}_S^\mathsf{T} + \lambda_n I_n)^{-1} \mathbf{y}\end{aligned}$$

Designed setting so that $\mathbf{X}_J \mathbf{X}_J^\mathsf{T} \xrightarrow{a.s.} \lambda_n I_n$, so

As $d_J \to \infty$, $\hat{\mathbf{w}}_{MN} = \mathbf{X}^\dagger \mathbf{y} \approx$ ridge regression on the signal

$$\hat{\mathbf{w}}_{\lambda_n} = \underset{\mathbf{w}_S \in \mathbb{R}^{d_S}}{\arg\min} \|\mathbf{X}_S \mathbf{w}_S{}^2 - \mathbf{y}\|^2 + \lambda_n \|\mathbf{w}_S\|^2$$

$$= \mathbf{X}_S^\mathsf{T} (\mathbf{X}_S \mathbf{X}_S^\mathsf{T} + \lambda_n I_n)^{-1} \mathbf{y}$$

Designed setting so that $\mathbf{X}_J \mathbf{X}_J^\mathsf{T} \xrightarrow{a.s.} \lambda_n I_n$, so

- $(\hat{\mathbf{w}}_{MN})_S = \mathbf{X}_S^\mathsf{T} (\mathbf{X}_S \mathbf{X}_S^\mathsf{T} + \mathbf{X}_J \mathbf{X}_J^\mathsf{T})^{-1} \mathbf{y} \xrightarrow{a.s.} \hat{\mathbf{w}}_{\lambda_n}$

As $d_J \to \infty$, $\hat{\mathbf{w}}_{MN} = \mathbf{X}^{\dagger}\mathbf{y} \approx$ ridge regression on the signal

$$\hat{\mathbf{w}}_{\lambda_n} = \underset{\mathbf{w}_S \in \mathbb{R}^{d_S}}{\arg\min} \|\mathbf{X}_S \mathbf{w}_S{}^2 - \mathbf{y}\|^2 + \lambda_n \|\mathbf{w}_S\|^2$$

$$= \mathbf{X}_S^\top (\mathbf{X}_S \mathbf{X}_S^\top + \lambda_n I_n)^{-1} \mathbf{y}$$

Designed setting so that $\mathbf{X}_J \mathbf{X}_J^\top \xrightarrow{a.s.} \lambda_n I_n$, so

- $(\hat{\mathbf{w}}_{MN})_S = \mathbf{X}_S^\top (\mathbf{X}_S \mathbf{X}_S^\top + \mathbf{X}_J \mathbf{X}_J^\top)^{-1} \mathbf{y} \xrightarrow{a.s.} \hat{\mathbf{w}}_{\lambda_n}$
- $\langle (\hat{\mathbf{w}}_{MN})_J, \tilde{\mathbf{x}}_J \rangle = \mathbf{y}^\top (\mathbf{X}_S \mathbf{X}_S^\top + \mathbf{X}_J \mathbf{X}_J^\top)^{-1} \mathbf{X}_J^\top \tilde{\mathbf{x}}_J \xrightarrow{a.s.} 0$

As $d_J \to \infty$, $\hat{\mathbf{w}}_{MN} = \mathbf{X}^\dagger \mathbf{y} \approx$ ridge regression on the signal

$$\hat{\mathbf{w}}_{\lambda_n} = \operatorname*{arg\,min}_{\mathbf{w}_S \in \mathbb{R}^{d_S}} \|\mathbf{X}_S \mathbf{w}_S{}^2 - \mathbf{y}\|^2 + \lambda_n \|\mathbf{w}_S\|^2$$

$$= \mathbf{X}_S^\mathsf{T} (\mathbf{X}_S \mathbf{X}_S^\mathsf{T} + \lambda_n I_n)^{-1} \mathbf{y}$$

Designed setting so that $\mathbf{X}_J \mathbf{X}_J^\mathsf{T} \xrightarrow{a.s.} \lambda_n I_n$, so

- $(\hat{\mathbf{w}}_{MN})_S = \mathbf{X}_S^\mathsf{T} (\mathbf{X}_S \mathbf{X}_S^\mathsf{T} + \mathbf{X}_J \mathbf{X}_J^\mathsf{T})^{-1} \mathbf{y} \xrightarrow{a.s.} \hat{\mathbf{w}}_{\lambda_n}$
- $\langle (\hat{\mathbf{w}}_{MN})_J, \tilde{\mathbf{x}}_J \rangle = \mathbf{y}^\mathsf{T} (\mathbf{X}_S \mathbf{X}_S^\mathsf{T} + \mathbf{X}_J \mathbf{X}_J^\mathsf{T})^{-1} \mathbf{X}_J^\mathsf{T} \tilde{\mathbf{x}}_J \xrightarrow{a.s.} 0$

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

As $d_J \to \infty$, $\hat{\mathbf{w}}_{MN} = \mathbf{X}^\dagger \mathbf{y} \approx$ ridge regression on the signal

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

$$= \mathbf{X}_S^\top (\mathbf{X}_S \mathbf{X}_S^\top + \lambda_n I_n)^{-1} \mathbf{y}$$

Designed setting so that $\mathbf{X}_J \mathbf{X}_J^\top \xrightarrow{a.s.} \lambda_n I_n$, so

- $(\hat{\mathbf{w}}_{MN})_S = \mathbf{X}_S^\top (\mathbf{X}_S \mathbf{X}_S^\top + \mathbf{X}_J \mathbf{X}_J^\top)^{-1} \mathbf{y} \xrightarrow{a.s.} \hat{\mathbf{w}}_{\lambda_n}$
- $\langle (\hat{\mathbf{w}}_{MN})_J, \tilde{\mathbf{x}}_J \rangle = \mathbf{y}^\top (\mathbf{X}_S \mathbf{X}_S^\top + \mathbf{X}_J \mathbf{X}_J^\top)^{-1} \mathbf{X}_J^\top \tilde{\mathbf{x}}_J \xrightarrow{a.s.} 0$

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

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

As $d_J \to \infty$, $\hat{\mathbf{w}}_{MN} = \mathbf{X}^\dagger \mathbf{y} \approx$ ridge regression on the signal

$$\hat{\mathbf{w}}_{\lambda_n} = \operatorname*{arg\,min}_{\mathbf{w}_S \in \mathbb{R}^{d_S}} \|\mathbf{X}_S \mathbf{w}_S{}^2 - \mathbf{y}\|^2 + \lambda_n \|\mathbf{w}_S\|^2$$

$$= \mathbf{X}_S^\top (\mathbf{X}_S \mathbf{X}_S^\top + \lambda_n I_n)^{-1} \mathbf{y}$$

Designed setting so that $\mathbf{X}_J \mathbf{X}_J^\top \xrightarrow{a.s.} \lambda_n I_n$, so

- $(\hat{\mathbf{w}}_{MN})_S = \mathbf{X}_S^\top (\mathbf{X}_S \mathbf{X}_S^\top + \mathbf{X}_J \mathbf{X}_J^\top)^{-1} \mathbf{y} \xrightarrow{a.s.} \hat{\mathbf{w}}_{\lambda_n}$
- $\langle (\hat{\mathbf{w}}_{MN})_J, \tilde{\mathbf{x}}_J \rangle = \mathbf{y}^\top (\mathbf{X}_S \mathbf{X}_S^\top + \mathbf{X}_J \mathbf{X}_J^\top)^{-1} \mathbf{X}_J^\top \tilde{\mathbf{x}}_J \xrightarrow{a.s.} 0$

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

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

Could we have shown that with uniform convergence?

As $d_J \to \infty$, $\hat{\mathbf{w}}_{MN} = \mathbf{X}^\dagger \mathbf{y} \approx$ ridge regression on the signal

$$\hat{\mathbf{w}}_{\lambda_n} = \operatorname*{arg\,min}_{\mathbf{w}_S \in \mathbb{R}^{d_S}} \|\mathbf{X}_S \mathbf{w}_S{}^2 - \mathbf{y}\|^2 + \lambda_n \|\mathbf{w}_S\|^2$$

$$= \mathbf{X}_S^\top (\mathbf{X}_S \mathbf{X}_S^\top + \lambda_n I_n)^{-1} \mathbf{y}$$

Designed setting so that $\mathbf{X}_J \mathbf{X}_J^\top \xrightarrow{a.s.} \lambda_n I_n$, so

- $(\hat{\mathbf{w}}_{MN})_S = \mathbf{X}_S^\top (\mathbf{X}_S \mathbf{X}_S^\top + \mathbf{X}_J \mathbf{X}_J^\top)^{-1} \mathbf{y} \xrightarrow{a.s.} \hat{\mathbf{w}}_{\lambda_n}$
- $\langle (\hat{\mathbf{w}}_{MN})_J, \tilde{\mathbf{x}}_J \rangle = \mathbf{y}^\top (\mathbf{X}_S \mathbf{X}_S^\top + \mathbf{X}_J \mathbf{X}_J^\top)^{-1} \mathbf{X}_J^\top \tilde{\mathbf{x}}_J \xrightarrow{a.s.} 0$

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

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

Could we have shown that with uniform convergence?

A first attempt at uniform convergence

"Default" approach:

$$\mathbb{E} \sup_{\|\mathbf{w}\| \leq B_n} \left| L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \right|$$

A first attempt at uniform convergence

"Default" approach:

$$\mathbb{E} \sup_{\|\mathbf{w}\| \leq B_n} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|$$

A first attempt at uniform convergence

“Default” approach:

$$\mathbb{E} \sup_{\|\mathbf{w}\| \leq B_n} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})| \leq 2 \, \mathrm{Lip} \sqrt{\frac{B_n^2 \, \mathbb{E}\|\mathbf{x}\|^2}{n}}$$

A first attempt at uniform convergence

"Default" approach:

$$\mathbb{E} \sup_{\|\mathbf{w}\| \le B_n} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})| \le 2 \,\mathrm{Lip} \sqrt{\frac{B_n^2 \,\mathbb{E}\|\mathbf{x}\|^2}{n}}$$

With $B_n^2 = \mathbb{E}\left[\|\hat{\mathbf{w}}_{MN}\|^2\right]$, get

A first attempt at uniform convergence

"Default" approach: (assuming $\lambda_n \to \infty$)

$$\mathbb{E} \sup_{\|\mathbf{w}\| \le B_n} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})| \le 2\,\mathrm{Lip} \sqrt{\frac{B_n^2\, \mathbb{E}\|\mathbf{x}\|^2}{n}}$$

With $B_n^2 = \mathbb{E}\left[\|\hat{\mathbf{w}}_{MN}\|^2\right]$, get

A first attempt at uniform convergence

"Default" approach: (assuming $\lambda_n \to \infty$)

$$\mathbb{E} \sup_{\|\mathbf{w}\| \leq B_n} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})| \leq 2 \,\mathrm{Lip} \sqrt{\frac{B_n^2 \,\mathbb{E}\|\mathbf{x}\|^2}{n}}$$

With $B_n^2 = \mathbb{E}\left[\|\hat{\mathbf{w}}_{MN}\|^2\right]$, get $\sqrt{\frac{B_n^2 \,\mathbb{E}\|\mathbf{x}\|^2}{n}} \to \sigma$

A first attempt at uniform convergence

"Default" approach: (assuming $\lambda_n \to \infty$)

$$\mathbb{E} \sup_{\|\mathbf{w}\| \le B_n} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})| \le 2 \, \mathrm{Lip} \sqrt{\frac{B_n^2 \, \mathbb{E}\|\mathbf{x}\|^2}{n}}$$

With $B_n^2 = \mathbb{E}\left[\|\hat{\mathbf{w}}_{MN}\|^2\right]$, get $\sqrt{\frac{B_n^2 \, \mathbb{E}\|\mathbf{x}\|^2}{n}} \to \sigma$

Would need $\mathrm{Lip} \to \frac{1}{2}\sigma$…

A first attempt at uniform convergence

"Default" approach: (assuming $\lambda_n \to \infty$)

$$\mathbb{E} \sup_{\|\mathbf{w}\| \le B_n} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})| \le 2 \operatorname{Lip} \sqrt{\frac{B_n^2 \, \mathbb{E}\|\mathbf{x}\|^2}{n}}$$

With $B_n^2 = \mathbb{E}\left[\|\hat{\mathbf{w}}_{MN}\|^2\right]$, get $\sqrt{\frac{B_n^2 \, \mathbb{E}\|\mathbf{x}\|^2}{n}} \to \sigma$

Would need $\operatorname{Lip} \to \frac{1}{2}\sigma$…

but only have $\operatorname{Lip} \le \sup_{\|\mathbf{w}\| \le B_n} \max_i 2|\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i|$

A first attempt at uniform convergence

"Default" approach: (assuming $\lambda_n \to \infty$)

$$\mathbb{E} \sup_{\|\mathbf{w}\| \le B_n} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})| \le 2 \,\mathrm{Lip} \sqrt{\frac{B_n^2 \,\mathbb{E}\|\mathbf{x}\|^2}{n}}$$

With $B_n^2 = \mathbb{E}\left[\|\hat{\mathbf{w}}_{MN}\|^2\right]$, get $\sqrt{\frac{B_n^2 \,\mathbb{E}\|\mathbf{x}\|^2}{n}} \to \sigma$

Would need $\mathrm{Lip} \to \frac{1}{2}\sigma$...

but only have $\mathrm{Lip} \le \sup_{\|\mathbf{w}\| \le B_n} \max_i 2|\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i| \to \infty$

No uniform convergence on norm balls

Theorem: In junk features with $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\|_2 \le \|\hat{\mathbf{w}}_{MN}\|_2} \left| L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \right| \right] = \infty.$$

No uniform convergence on norm balls

Theorem: In junk features with $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\|_2 \le \|\hat{\mathbf{w}}_{MN}\|_2} \left| L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \right| \right] = \infty.$$

No uniform convergence on norm balls

Theorem: In junk features with $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\|_2 \le \|\hat{\mathbf{w}}_{MN}\|_2} \left| L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \right| \right] = \infty.$$

No uniform convergence on norm balls

Theorem: In junk features with $\lambda_n = o(n)$,

$$\lim_{n \to \infty} \lim_{d_J \to \infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\|_2 \leq \|\hat{\mathbf{w}}_{MN}\|_2} \left| L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \right| \right] = \infty.$$

Proof idea:

No uniform convergence on norm balls

<u>Theorem:</u> In junk features with $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\|_2 \le \|\hat{\mathbf{w}}_{MN}\|_2} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|\right] = \infty.$$

Proof idea:

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

No uniform convergence on norm balls

<u>Theorem:</u> In junk features with $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\|_2 \le \|\hat{\mathbf{w}}_{MN}\|_2} \left|L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})\right|\right] = \infty.$$

Proof idea:

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

$$\begin{aligned} L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) &= (\mathbf{w} - \mathbf{w}^*)(\mathbf{\Sigma} - \hat{\mathbf{\Sigma}})(\mathbf{w} - \mathbf{w}^*) \\ &\quad + (L_{\mathcal{D}}(\mathbf{w}^*) - L_{\mathbf{S}}(\mathbf{w}^*)) - \text{cross term} \end{aligned}$$

No uniform convergence on norm balls

Theorem: In junk features with $\lambda_n = o(n)$,

$$\lim_{n \to \infty} \lim_{d_J \to \infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\|_2 \le \|\hat{\mathbf{w}}_{MN}\|_2} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})| \right] = \infty.$$

Proof idea:

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

$$L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) = (\mathbf{w} - \mathbf{w}^*)(\mathbf{\Sigma} - \hat{\mathbf{\Sigma}})(\mathbf{w} - \mathbf{w}^*)$$

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

$$\sup[\ldots] \ge \|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_{op} \cdot (\|\hat{\mathbf{w}}_{MN}\|_2 - \|\mathbf{w}^*\|_2)^2 + o(1)$$

No uniform convergence on norm balls

<u>Theorem:</u> In junk features with $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\|_2 \le \|\hat{\mathbf{w}}_{MN}\|_2} \left| L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \right| \right] = \infty.$$

Proof idea:

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

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

$$\sup[\ldots] \ge \underbrace{\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_{op}}_{\Theta\left(\frac{n}{\lambda_n}\right)} \cdot (\|\hat{\mathbf{w}}_{MN}\|_2 - \|\mathbf{w}^*\|_2)^2 + o(1)$$

No uniform convergence on norm balls

Theorem: In junk features with $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\|_2 \le \|\hat{\mathbf{w}}_{MN}\|_2} \left| L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \right| \right] = \infty.$$

Proof idea:

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

$$\begin{aligned} L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) &= (\mathbf{w} - \mathbf{w}^*)(\mathbf{\Sigma} - \hat{\mathbf{\Sigma}})(\mathbf{w} - \mathbf{w}^*) \\ &\quad + (L_{\mathcal{D}}(\mathbf{w}^*) - L_{\mathbf{S}}(\mathbf{w}^*)) - \text{cross term} \end{aligned}$$

$$\sup[\ldots] \ge \underbrace{\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_{op}}_{\Theta\left(\sqrt{\frac{\lambda_n}{n}}\right)} \cdot \underbrace{(\|\hat{\mathbf{w}}_{MN}\|_2 - \|\mathbf{w}^*\|_2)^2}_{\Theta\left(\frac{n}{\lambda_n}\right)} + o(1)$$

Koltchinskii/Lounici, Bernoulli 2017

No uniform convergence on norm balls

Theorem: In junk features with $\lambda_n = o(n)$,

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \mathbb{E}\left[\sup_{\|\mathbf{w}\|_2 \le \|\hat{\mathbf{w}}_{MN}\|_2} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|\right] = \infty.$$

Proof idea:

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

$$\begin{aligned} L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) &= (\mathbf{w} - \mathbf{w}^*)(\mathbf{\Sigma} - \hat{\mathbf{\Sigma}})(\mathbf{w} - \mathbf{w}^*) \\ &\quad + (L_{\mathcal{D}}(\mathbf{w}^*) - L_{\mathbf{S}}(\mathbf{w}^*)) - \text{cross term} \end{aligned}$$

$$\sup[\ldots] \ge \underbrace{\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_{op}}_{\Theta\left(\sqrt{\frac{\lambda_n}{n}}\right)} \cdot \underbrace{(\|\hat{\mathbf{w}}_{MN}\|_2 - \|\mathbf{w}^*\|_2)^2}_{\Theta\left(\frac{n}{\lambda_n}\right)} + o(1) \to \infty$$

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

Vaishnav 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]):

In junk features, for each $\delta \in (0, \frac{1}{2})$, let $\Pr(\mathbf{S} \in \mathcal{S}_{n,\delta}) \geq 1 - \delta$,

A more refined uniform convergence analysis?

Theorem (à la [Nagarajan/Kolter, NeurIPS 2019]):
In junk features, for each $\delta \in (0, \frac{1}{2})$, let $\Pr(\mathbf{S} \in \mathcal{S}_{n,\delta}) \geq 1 - \delta$,
$\hat{\mathbf{w}}$ a *natural* consistent interpolator,

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

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

$\operatorname*{arg\,min}_{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}} f_S(\mathbf{w}_S) + f_J(\mathbf{w}_J)$ 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]):

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

$\hat{\mathbf{w}}$ 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:

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

$\operatorname*{arg\,min}_{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}} f_S(\mathbf{w}_S) + f_J(\mathbf{w}_J)$ 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]):
In junk features, for each $\delta \in (0, \frac{1}{2})$, let $\Pr\left(\mathbf{S} \in \mathcal{S}_{n,\delta}\right) \geq 1 - \delta$,
$\hat{\mathbf{w}}$ 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:

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

$\operatorname*{arg\,min}_{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}} f_S(\mathbf{w}_S) + f_J(\mathbf{w}_J)$ 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]):

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

$\hat{\mathbf{w}}$ 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:

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

$\operatorname*{arg\,min}_{\mathbf{w}:\mathbf{X}\mathbf{w}=\mathbf{y}} f_S(\mathbf{w}_S) + f_J(\mathbf{w}_J)$ 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]):

In junk features, for each $\delta \in (0, \frac{1}{2})$, let $\Pr(\mathbf{S} \in \mathcal{S}_{n,\delta}) \geq 1 - \delta$,

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

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

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

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* $\mathbf{S}$ ($L_{\mathbf{S}}(\mathbf{w}) \to 4\sigma^2$):

take $\hat{\mathbf{w}}$ with $\mathbf{X}_S$ the same, but $\mathbf{X}_J$ flipped to $-\mathbf{X}_J$

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 $\sup \left| L_{\mathcal{D}} - L_{\mathbf{S}} \right|$?

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 $\sup \lvert L_{\mathcal{D}} - L_{\mathbf{S}} \rvert$?

- Existing uniform convergence proofs are "really" about $\lvert L_{\mathcal{D}} - L_{\mathbf{S}} \rvert$ [Nagarajan/Kolter, NeurIPS 2019]
  - If you can bound $\mathfrak{R}$, can usually similarly bound $\mathfrak{R}'$

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]
  - If you can bound $\mathfrak{R}$, can usually similarly bound $\mathfrak{R}'$
- Strongly expect still ∞ for norm balls in our testbed
  - $\lambda_{\max}(\mathbf{\Sigma} - \hat{\mathbf{\Sigma}})$ instead of $\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_{op}$

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]
  - If you can bound $\mathfrak{R}$, can usually similarly bound $\mathfrak{R}'$
- Strongly expect still ∞ for norm balls in our testbed
  - $\lambda_{\max}(\mathbf{\Sigma} - \hat{\mathbf{\Sigma}})$ instead of $\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_{op}$
- Not possible to show $\sup_{h \in \mathcal{H}} L_{\mathcal{D}} - L_{\mathbf{S}}$ is big for *all* $\mathcal{H}$
  - If $\hat{h}$ consistent and $\inf_f L_{\mathbf{S}}(h) \geq 0$, use $\mathcal{H} = \{f : L_{\mathcal{D}}(h) \leq L_{\mathcal{D}}(h^*) + \epsilon_{n,\delta}\}$

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...
- Only do analyses based on e.g. exact form of $\hat{\mathbf{w}}_{MN}$?

So, what are we left with?

- Convergence of surrogates [Negrea/Dziugaite/Roy, ICML 2020]?
  - Nice, but not really the same thing...
- Only do analyses based on e.g. exact form of $\hat{\mathbf{w}}_{MN}$?
- We'd like to keep good things about uniform convergence:
  - Apply to more than just one specific predictor
  - Tell us more about “why” things generalize
  - Easier to apply without a nice closed form

So, what are we left with?

- Convergence of surrogates [Negrea/Dziugaite/Roy, ICML 2020]?
  - Nice, but not really the same thing…
- Only do analyses based on e.g. exact form of $\hat{\mathbf{w}}_{MN}$?
- We'd like to keep good things about uniform convergence:
  - Apply to more than just one specific predictor
  - Tell us more about “why” things generalize
  - Easier to apply without a nice closed form
- Or…

A broader view of uniform convergence

So far, used $L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \leq \sup_{\|\mathbf{w}\|_2 \leq B} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|$

A broader view of uniform convergence

So far, used $L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \leq \sup_{\|\mathbf{w}\|_2 \leq B} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|$

But *we only care about interpolators*. How about

$$\sup_{\|\mathbf{w}\|_2 \leq B,\ {\color{blue} L_{\mathbf{S}}(\mathbf{w})=0}} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|?$$

A broader view of uniform convergence

So far, used $L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \leq \sup_{\|\mathbf{w}\|_2 \leq B} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|$

But *we only care about interpolators*. How about

$$\sup_{\|\mathbf{w}\|_2 \leq B,\ L_{\mathbf{S}}(\mathbf{w})=0} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|?$$

Is this “uniform convergence”?

A broader view of uniform convergence

So far, used $L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \le \sup_{\|\mathbf{w}\|_2 \le B} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|$

But *we only care about interpolators*. How about

$$\sup_{\|\mathbf{w}\|_2 \le B,\; L_{\mathbf{S}}(\mathbf{w})=0} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|?$$

Is this “uniform convergence”?

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

A broader view of uniform convergence

S

Used at least since [Vapnik 1982] and [Valiant 1984]

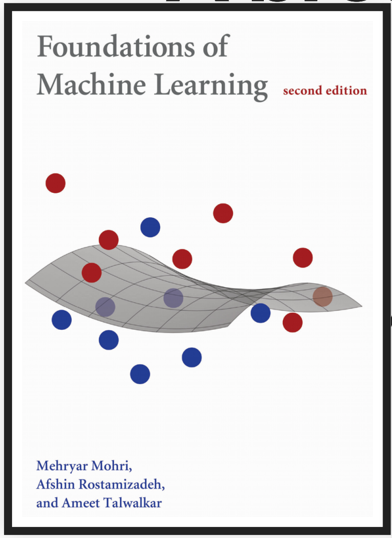
From [Devroye/Györfi/Lugosi 1996]:

PROOF. For $n\epsilon \leq 2$, the inequality is clearly true. So, we assume that $n\epsilon > 2$. First observe that since $\inf_{\phi \in \mathcal{C}} L(\phi) = 0$, $\widehat{L}_n(\phi_n^*) = 0$ with probability one. It is easily seen that

$$L(\phi_n^*) \leq \sup_{\phi : \widehat{L}_n(\phi)=0} |L(\phi) - \widehat{L}_n(\phi)|.$$

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 $\Pr_{S \sim \mathcal{D}^m}[R(h_S) \leq \epsilon] \geq 1 - \delta$ 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 $L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \leq \sup_{\|\mathbf{w}\|_2 \leq B} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|$

But *we only care about interpolators*. How about

$$\sup_{\|\mathbf{w}\|_2 \leq B,\ {\color{blue} L_{\mathbf{S}}(\mathbf{w})=0}} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|?$$

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 $L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \le \sup_{\|\mathbf{w}\|_2 \le B} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|$

But *we only care about interpolators*. How about

$$\sup_{\|\mathbf{w}\|_2 \le B,\ L_{\mathbf{S}}(\mathbf{w}) = 0} |L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w})|?$$

Is this “uniform convergence”?

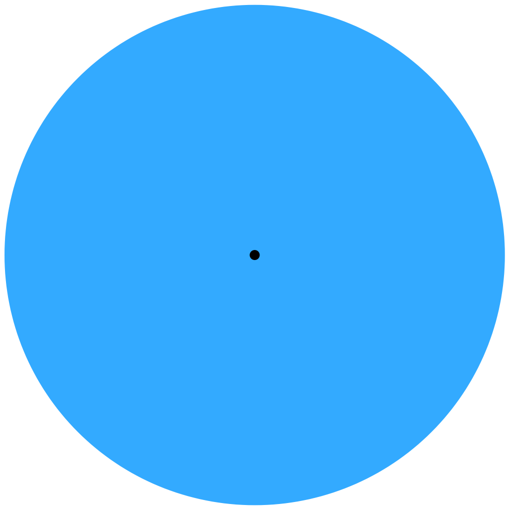
It's the standard notion for realizable ($L_{\mathcal{D}}(w^*) = 0$) analyses…

The interpolator ball in linear regression

What does $\{\mathbf{w} : \|\mathbf{w}\|_2 \leq B,\ L_{\mathbf{S}}(\mathbf{w}) = 0\}$ look like?

The interpolator ball in linear regression

What does $\{\mathbf{w} : \|\mathbf{w}\|_2 \leq B,\ L_{\mathbf{S}}(\mathbf{w}) = 0\}$ look like?

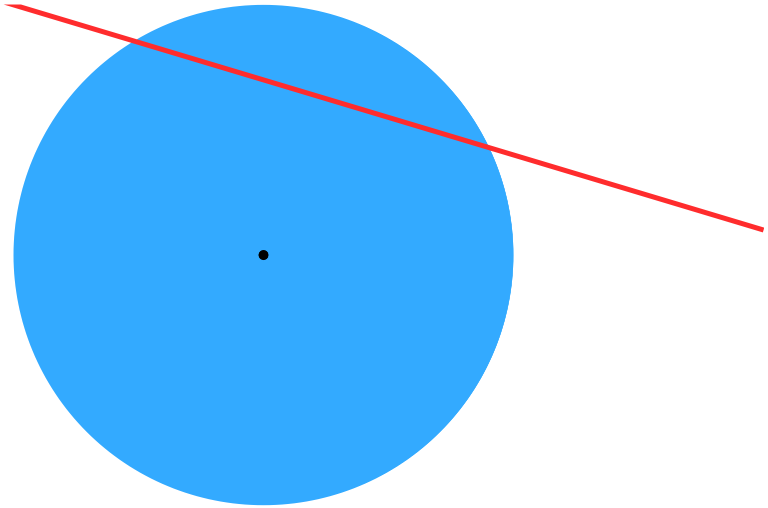


Intersection of d-ball

The interpolator ball in linear regression

What does $\{\mathbf{w} : \|\mathbf{w}\|_2 \leq B,\ L_{\mathbf{S}}(\mathbf{w}) = 0\}$ look like?

$\{\mathbf{w} : L_{\mathbf{S}}(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 = 0\}$ is the plane $\mathbf{X}\mathbf{w} = \mathbf{y}$

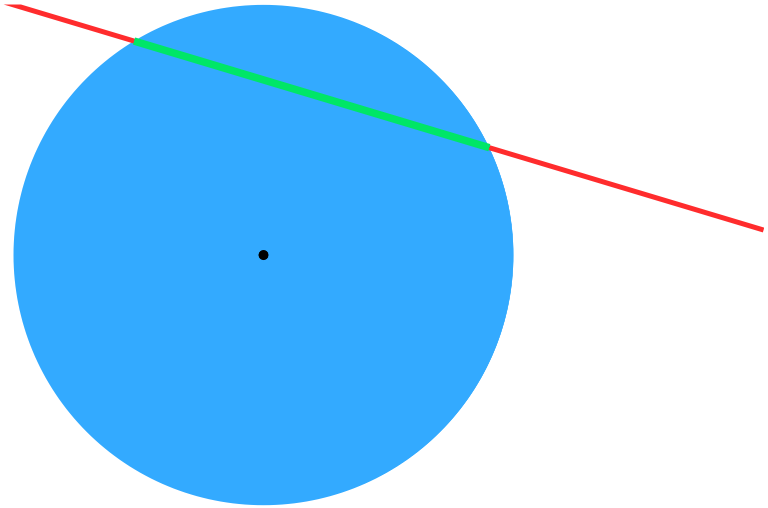


Intersection of d-ball with $(d-n)$-hyperplane:

The interpolator ball in linear regression

What does $\{\mathbf{w} : \|\mathbf{w}\|_2 \leq B,\ L_{\mathbf{S}}(\mathbf{w}) = 0\}$ look like?

$\{\mathbf{w} : L_{\mathbf{S}}(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 = 0\}$ is the plane $\mathbf{X}\mathbf{w} = \mathbf{y}$

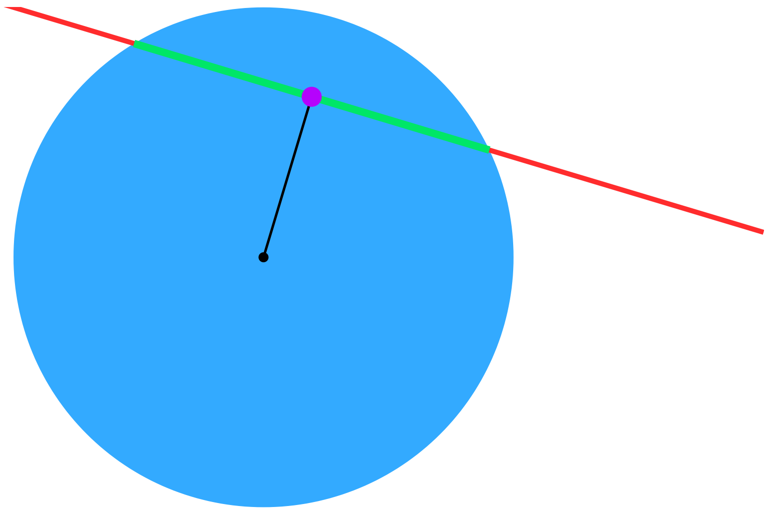


Intersection of d-ball with $(d-n)$-hyperplane:
$(d-n)$-ball

The interpolator ball in linear regression

What does $\{\mathbf{w} : \|\mathbf{w}\|_2 \leq B,\ L_{\mathbf{S}}(\mathbf{w}) = 0\}$ look like?

$\{\mathbf{w} : L_{\mathbf{S}}(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 = 0\}$ is the plane $\mathbf{X}\mathbf{w} = \mathbf{y}$



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

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

Optimistic rates

[Srebro/Sridharan/Tewari 2010] show:

$$L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \leq \tilde{\mathcal{O}}_P\left(\sqrt{L_{\mathbf{S}}(\mathbf{w})\,\bar{\mathfrak{R}}_n(\mathcal{H})^2} + \bar{\mathfrak{R}}_n(\mathcal{H})^2\right)$$

Optimistic rates

[Srebro/Sridharan/Tewari 2010] show:

$$L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \le \tilde{\mathcal{O}}_P\left(\sqrt{L_{\mathbf{S}}(\mathbf{w})\,\bar{\mathfrak{R}}_n(\mathcal{H})^2} + \bar{\mathfrak{R}}_n(\mathcal{H})^2\right)$$

ψ_n: high-prob bound on $\max_{i=1,\dots,n} \lVert \mathbf{x}_i \rVert_2^2$

$$\sup_{\lVert \mathbf{w} \rVert_2 \le B,\ L_{\mathbf{S}}(\mathbf{w})=0} L_{\mathcal{D}}(\mathbf{w}) \le c_n \frac{B^2 \psi_n}{n} + o_P(1)$$

Optimistic rates

[Srebro/Sridharan/Tewari 2010] show:

$$L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \le \tilde{\mathcal{O}}_P\left(\sqrt{L_{\mathbf{S}}(\mathbf{w})\, \bar{\mathfrak{R}}_n(\mathcal{H})^2} + \bar{\mathfrak{R}}_n(\mathcal{H})^2 \right)$$

ψ_n: high-prob bound on $\max_{i=1,\ldots,n} \|\mathbf{x}_i\|_2^2$

$$\sup_{\|\mathbf{w}\|_2 \le B,\, L_{\mathbf{S}}(\mathbf{w})=0} L_{\mathcal{D}}(\mathbf{w}) \le c_n \frac{B^2 \psi_n}{n} + o_P(1)$$

if $1 \ll \lambda_n \ll n$, $B = \|\hat{\mathbf{w}}_{MN}\|_2$, $\to c_n\, L_{\mathcal{D}}(\mathbf{w}^*)$

Optimistic rates

[Srebro/Sridharan/Tewari 2010] show:

$$L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \le \tilde{\mathcal{O}}_P\left(\sqrt{L_{\mathbf{S}}(\mathbf{w})\, \bar{\mathfrak{R}}_n(\mathcal{H})^2} + \bar{\mathfrak{R}}_n(\mathcal{H})^2\right)$$

ψ_n: high-prob bound on $\max_{i=1,\dots,n} \lVert \mathbf{x}_i \rVert_2^2$

$$\sup_{\lVert \mathbf{w} \rVert_2 \le B,\; L_{\mathbf{S}}(\mathbf{w})=0} L_{\mathcal{D}}(\mathbf{w}) \le c_n \frac{B^2 \psi_n}{n} + o_P(1)$$

if $1 \ll \lambda_n \ll n$, $B = \lVert \hat{\mathbf{w}}_{MN} \rVert_2$, $\quad \to c_n\, L_{\mathcal{D}}(\mathbf{w}^*)$

If this holds with $c_n \to 1$ (and $\mathfrak{R}_n$ instead of $\bar{\mathfrak{R}}_n$),
would explain consistency on junk features,
and predict that $B = \alpha \lVert \hat{\mathbf{w}}_{MN} \rVert_2$ gives $\alpha^2\, L_{\mathcal{D}}(\mathbf{w}^*)$

Optimistic rates

[Srebro/Sridharan/Tewari 2010] show:

$$L_{\mathcal{D}}(\mathbf{w}) - L_{\mathbf{S}}(\mathbf{w}) \le \tilde{\mathcal{O}}_P\left(\sqrt{L_{\mathbf{S}}(\mathbf{w})\,\bar{\mathfrak{R}}_n(\mathcal{H})^2} + \bar{\mathfrak{R}}_n(\mathcal{H})^2\right)$$

$c_n \le 200{,}000 \log^3(n)$ $\quad \psi_n$: high-prob bound on $\max_{i=1,\ldots,n} \|\mathbf{x}_i\|_2^2$

$$\sup_{\|\mathbf{w}\|_2 \le B,\; L_{\mathbf{S}}(\mathbf{w})=0} L_{\mathcal{D}}(\mathbf{w}) \le c_n \frac{B^2 \psi_n}{n} + o_P(1)$$

if $1 \ll \lambda_n \ll n$, $B = \|\hat{\mathbf{w}}_{MN}\|_2$, $\quad \to c_n\, L_{\mathcal{D}}(\mathbf{w}^*)$

If this holds with $c_n \to 1$ (and $\mathfrak{R}_n$ instead of $\bar{\mathfrak{R}}_n$),
would explain consistency on junk features,
and predict that $B = \alpha \|\hat{\mathbf{w}}_{MN}\|_2$ gives $\alpha^2\, L_{\mathcal{D}}(\mathbf{w}^*)$

Conjecture holds for Gaussian linear regression

For Gaussian linear regression, with general compact $\mathcal{H}$,
ignoring lower-order terms, we show w.h.p. that for all $w \in \mathcal{H}$,

$$L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w) \leq 2\sqrt{L_{\mathbf{S}}(w) \cdot \mathfrak{R}_n(\mathcal{H})^2} + \mathfrak{R}_n(\mathcal{H})^2$$

Conjecture holds for Gaussian linear regression

For Gaussian linear regression, with general compact $\mathcal{H}$, ignoring lower-order terms, we show w.h.p. that for all $w \in \mathcal{H}$,

$$L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w) \leq 2\sqrt{L_{\mathbf{S}}(w) \cdot \mathfrak{R}_n(\mathcal{H})^2} + \mathfrak{R}_n(\mathcal{H})^2$$

$$L_{\mathcal{D}}(w) \leq \left(\sqrt{L_{\mathbf{S}}(w)} + \mathfrak{R}_n(\mathcal{H})\right)^2$$

Conjecture holds for Gaussian linear regression

For Gaussian linear regression, with general compact $\mathcal{H}$, ignoring lower-order terms, we show w.h.p. that for all $w \in \mathcal{H}$,

$$L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w) \leq 2\sqrt{L_{\mathbf{S}}(w) \cdot \mathfrak{R}_n(\mathcal{H})^2} + \mathfrak{R}_n(\mathcal{H})^2$$

$$L_{\mathcal{D}}(w) \leq \left(\sqrt{L_{\mathbf{S}}(w)} + \mathfrak{R}_n(\mathcal{H})\right)^2$$

$$\sup_{w \in \mathcal{H}} \sqrt{L_{\mathcal{D}}(w)} - \sqrt{L_{\mathbf{S}}(w)} \leq \mathfrak{R}_n(\mathcal{H})$$

Conjecture holds for Gaussian linear regression

For Gaussian linear regression, with general compact $\mathcal{H}$, ignoring lower-order terms, we show w.h.p. that for all $w \in \mathcal{H}$,

$$L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w) \leq 2\sqrt{L_{\mathbf{S}}(w) \cdot \mathfrak{R}_n(\mathcal{H})^2} + \mathfrak{R}_n(\mathcal{H})^2$$

$$L_{\mathcal{D}}(w) \leq \left(\sqrt{L_{\mathbf{S}}(w)} + \mathfrak{R}_n(\mathcal{H})\right)^2$$

$$\sup_{w \in \mathcal{H}} \sqrt{L_{\mathcal{D}}(w)} - \sqrt{L_{\mathbf{S}}(w)} \leq \mathfrak{R}_n(\mathcal{H})$$

Proof *very specific* to Gaussian $\mathbf{x}$, pretty specific to linear models (but should work with sub-Gaussian noise) (extension beyond square loss is ongoing)

- Junk features setting: very stylized but “kind of like” deep learning
  - $\hat{\mathbf{w}}_{MN}$ is consistent, but usual uniform convergence can't show that
  - Uniform convergence over norm ball shows *nothing*

- Junk features setting: very stylized but "kind of like" deep learning
  - $\hat{\mathbf{w}}_{MN}$ is consistent, but usual uniform convergence can't show that
  - Uniform convergence over norm ball shows *nothing*
- Uniform convergence of interpolators does work
  - Together with new analysis of $\|\hat{\mathbf{w}}_{MN}\|$,
    ~matches previously known (nearly necessary) sufficient conditions
  - Shows low norm is sufficient for interpolation learning
  - Also apply to min-ℓ_1 interpolator [Wang/Donhauser/Yang AISTATS-22]
  - and two-layer random feature networks [Yang/Bai/Mei ICML-21]

- Junk features setting: very stylized but “kind of like” deep learning
  - $\hat{\mathbf{w}}_{MN}$ is consistent, but usual uniform convergence can't show that
  - Uniform convergence over norm ball shows *nothing*
- Uniform convergence of interpolators does work
  - Together with new analysis of $\|\hat{\mathbf{w}}_{MN}\|$,
    ~matches previously known (nearly necessary) sufficient conditions
  - Shows low norm is sufficient for interpolation learning
  - Also apply to min-ℓ_1 interpolator [Wang/Donhauser/Yang AISTATS-22]
  - and two-layer random feature networks [Yang/Bai/Mei ICML-21]
- Optimistic rates cover that theory, but also cover near-interpolators
  - Some non-square losses, but (so far) very specific to Gaussian data

- Junk features setting: very stylized but “kind of like” deep learning
  - $\hat{\mathbf{w}}_{MN}$ is consistent, but usual uniform convergence can't show that
  - Uniform convergence over norm ball shows *nothing*
- Uniform convergence of interpolators does work
  - Together with new analysis of $\|\hat{\mathbf{w}}_{MN}\|$,
    ~matches previously known (nearly necessary) sufficient conditions
  - Shows low norm is sufficient for interpolation learning
  - Also apply to min-ℓ_1 interpolator [Wang/Donhauser/Yang AISTATS-22]
  - and two-layer random feature networks [Yang/Bai/Mei ICML-21]
- Optimistic rates cover that theory, but also cover near-interpolators
  - Some non-square losses, but (so far) very specific to Gaussian data
- **Moving forward:**
  - “Plain” uniform convergence: maybe unlikely for realistic-ish NNs
  - Uniform convergence of interpolators / optimistic rates might work!
  - Or maybe $L_{\mathcal{D}}(\hat{h}) \leq L_{\mathcal{D}}(h^*) + \varepsilon$ type bounds…but unclear how

Backup slides