Modern surprises in classical machine learning

Vidya Muthukumar
Assistant Professor, Schools of ECE and ISyE, Georgia Tech
Thanks to my wonderful collaborators!

Mikhail (Misha) Belkin

Daniel Hsu

Adhyyan Narang

Anant Sahai

Vignesh Subramanian

Ji Xu
High-level motivation: Success of **overparameterized** neural networks

Mysterious empirical success of *heavily overparameterized* neural networks...

Accuracy of models on CIFAR10 test dataset*  
(50,000 training points)

<table>
<thead>
<tr>
<th>model</th>
<th># params</th>
<th>train accuracy</th>
<th>test accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception</td>
<td>1,649,402</td>
<td>100.0</td>
<td>89.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>89.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>86.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>85.75</td>
</tr>
<tr>
<td>(fitting random labels)</td>
<td>100.0</td>
<td>100.0</td>
<td>9.78</td>
</tr>
<tr>
<td>Inception w/o BatchNorm</td>
<td>1,649,402</td>
<td>100.0</td>
<td>83.00</td>
</tr>
<tr>
<td>(fitting random labels)</td>
<td>100.0</td>
<td>100.0</td>
<td>10.12</td>
</tr>
<tr>
<td>Alexnet</td>
<td>1,387,786</td>
<td>99.90</td>
<td>81.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>99.82</td>
<td>79.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0</td>
<td>77.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0</td>
<td>76.07</td>
</tr>
<tr>
<td>(fitting random labels)</td>
<td>99.82</td>
<td>100.0</td>
<td>9.86</td>
</tr>
</tbody>
</table>

*Zhang et al (2017): “Understanding deep learning requires rethinking generalization”.*
High-level motivation: Success of overparameterized neural networks

Mysterious empirical success of heavily overparameterized neural networks...

Accuracy of models on CIFAR10 test dataset*
(50,000 training points)

<table>
<thead>
<tr>
<th>model</th>
<th># params</th>
<th>train accuracy</th>
<th>test accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception</td>
<td>1,649,402</td>
<td>100.0</td>
<td>89.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0</td>
<td>89.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0</td>
<td>86.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0</td>
<td>85.75</td>
</tr>
<tr>
<td>(fitting random labels)</td>
<td>100.0</td>
<td>9.78</td>
<td></td>
</tr>
<tr>
<td>Inception w/o BatchNorm</td>
<td>1,649,402</td>
<td>100.0</td>
<td>83.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0</td>
<td>82.00</td>
</tr>
<tr>
<td>(fitting random labels)</td>
<td>100.0</td>
<td>10.12</td>
<td></td>
</tr>
<tr>
<td>Alexnet</td>
<td>1,387,786</td>
<td>99.90</td>
<td>81.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>99.82</td>
<td>79.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0</td>
<td>77.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0</td>
<td>76.07</td>
</tr>
<tr>
<td>(fitting random labels)</td>
<td>99.82</td>
<td>9.86</td>
<td></td>
</tr>
</tbody>
</table>

*Zhang et al (2017): “Understanding deep learning requires rethinking generalization”.

zero training error, yet...
High-level motivation: Success of overparameterized neural networks

Mysterious empirical success of *heavily overparameterized* neural networks...

Accuracy of models on CIFAR10 test dataset*

*(50,000 training points)*

<table>
<thead>
<tr>
<th>model</th>
<th># params</th>
<th>train accuracy</th>
<th>test accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception</td>
<td>1,649,402</td>
<td>100.0</td>
<td>89.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0</td>
<td>89.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0</td>
<td>86.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0</td>
<td>85.75</td>
</tr>
<tr>
<td>(fitting random labels)</td>
<td>100.0</td>
<td>9.78</td>
<td></td>
</tr>
<tr>
<td>Inception w/o BatchNorm</td>
<td>1,649,402</td>
<td>100.0</td>
<td>83.00</td>
</tr>
<tr>
<td>(fitting random labels)</td>
<td>100.0</td>
<td>10.12</td>
<td></td>
</tr>
<tr>
<td>Alexnet</td>
<td>1,387,786</td>
<td>99.90</td>
<td>81.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>99.82</td>
<td>79.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0</td>
<td>77.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100.0</td>
<td>76.07</td>
</tr>
<tr>
<td>(fitting random labels)</td>
<td>99.82</td>
<td>9.86</td>
<td></td>
</tr>
</tbody>
</table>

*Zhang et al (2017): “Understanding deep learning requires rethinking generalization”.*
High-level motivation: Success of overparameterized neural networks

Mysterious empirical success of *heavily overparameterized* neural networks...

Accuracy of models on CIFAR10 test dataset* (50,000 training points)

<table>
<thead>
<tr>
<th>model</th>
<th># params</th>
<th>train accuracy</th>
<th>test accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception</td>
<td>1,649,402</td>
<td>100.0</td>
<td>89.05</td>
</tr>
<tr>
<td>BatchNorm w/o labels</td>
<td>1,649,402</td>
<td>100.0</td>
<td>83.00</td>
</tr>
<tr>
<td>(fitting random labels)</td>
<td>100.0</td>
<td>9.78</td>
<td></td>
</tr>
<tr>
<td>Inception w/o BatchNorm</td>
<td>1,649,402</td>
<td>100.0</td>
<td>82.00</td>
</tr>
<tr>
<td>(fitting random labels)</td>
<td>100.0</td>
<td>10.12</td>
<td></td>
</tr>
<tr>
<td>Alexnet</td>
<td>1,387,786</td>
<td>99.90</td>
<td>81.22</td>
</tr>
<tr>
<td>(fitting random labels)</td>
<td>99.82</td>
<td>9.86</td>
<td></td>
</tr>
</tbody>
</table>

Type of solution found (linear models/last layer of neural net): Gradient descent converges to solutions with *min-2-norm bias*

Optimization landscape: Overparameterization helps us easier achieve zero training error, yet...

Good classification accuracy

*Zhang et al (2017): “Understanding deep learning requires rethinking generalization”.*
Low-level motivation: Success of interpolation in linear models

Belkin, Hsu, Ma and Mandal (PNAS 2019):
Minimum-2-norm-interpolations of noisy data seem to empirically generalize better, as we overparameterize more.

Experiments on MNIST image dataset with random features, 2-layer (wide) neural networks.
Low-level motivation: Success of interpolation in linear models

Belkin, Hsu, Ma and Mandal (PNAS 2019): Minimum-2-norm-interpolations of noisy data seem to empirically generalize better, as we overparameterize more.

Experiments on MNIST image dataset with random features, 2-layer (wide) neural networks.

These phenomena also observed in local/nonparametric methods, boosting and random forests.
Questions we will answer for overparameterized linear model

- Regression (review): What happens when we interpolate noise?

- Classification: When are solutions that interpolate labels consistent?

- Classification: When is the hard-margin support-vector-machine (SVM) consistent?
Questions we will answer for overparameterized linear model

- Regression (review): What happens when we **interpolate** noise?

- Classification: When are **solutions that interpolate labels** consistent?

- Classification: When is the hard-margin **support-vector-machine (SVM)** consistent?
Review: Overparameterized linear regression

\[ Y = \phi(X)^T \beta^* + W \]

true parameter (signal)

output

features, dimension = \( d \)

noise, variance = \( \sigma^2 \)

\[ \mathbb{E}[\phi(X)] = 0, \quad \mathbb{E}[\phi(X)\phi(X)^T] = \Sigma \]
Review: Overparameterized linear regression

$$Y = \phi(X)^\top \beta^* + W$$

true parameter (signal)

output features, dimension = $d$

oise, variance = $\sigma^2$

$$\mathbb{E}[\phi(X)] = 0, \quad \mathbb{E}[\phi(X)\phi(X)^\top] = \Sigma$$

**Types of features considered:**

1. independent (sub)-Gaussian features

   $$\phi(X) := X, \Sigma = \text{diag}(\Lambda)$$

2. “Lifted” feature maps

   $$X \in [-1, 1], \text{first } d \text{ eigenfunctions}$$
   (e.g. Fourier, Legendre polynomials)
**Review: Overparameterized linear regression**

\[ Y = \phi(X)^\top \beta^* + W \]

true parameter (signal)

\[ \mathbb{E}[\phi(X)] = 0, \quad \mathbb{E}[\phi(X)\phi(X)^\top] = \Sigma \]

(output) samples

\[ A \beta = Y \] has infinitely many interpolating solutions

**Types of features considered:**

1. independent (sub)-Gaussian features

\[ \phi(X) := X, \Sigma = \text{diag}(\Lambda) \]

2. "Lifted" feature maps

\[ X \in [-1, 1], \] first \( d \) eigenfunctions

(e.g. Fourier, Legendre polynomials)
Review: Overparameterized linear regression

\[ Y = \phi(X)^\top \beta^* + W \]

true parameter (signal)

\[ \mathbb{E}[\phi(X)] = 0, \quad \mathbb{E}[\phi(X)\phi(X)^\top] = \Sigma \]

Features, dimension = \( d \)

noise, variance = \( \sigma^2 \)

(output) samples

(no. of features) \( d \gg n \) (no. of samples)

\[ A \beta = Y \]

has infinitely many interpolating solutions

Types of features considered:

1. independent (sub)-Gaussian features

\[ \phi(X) := X, \Sigma = \text{diag}(\Lambda) \]

2. “Lifted” feature maps

\[ X \in [-1, 1], \text{ first } d \text{ eigenfunctions} \]

(e.g. Fourier, Legendre polynomials)

Our focus today:

The minimum-2-norm interpolator

\[ \hat{\beta} := \arg \min ||\beta||_2 \]

subject to

\[ A \beta = Y. \]

equivalent to ridge regularization,

with \( \lambda \to 0 \)
Higher-dimensional models can interpolate noise in a harmless manner.
Higher-dimensional models can interpolate noise in a harmless manner.

E.g. orthonormal polynomial features \((n = 9)\)

Figure credits: Kailas Vodrahalli
Higher-dimensional models can interpolate noise in a harmless manner. E.g. orthonormal polynomial features \( n = 9 \).
Higher-dimensional models can interpolate noise in a harmless manner

E.g. orthonormal polynomial features \( (n = 9) \)

Figure credits: Kailas Vodrahalli
Harmless interpolation of noise in linear regression

Higher-dimensional models can interpolate noise in a harmless manner

E.g. orthonormal polynomial features \((n = 9)\)

Result: w.h.p. on training samples, min-2-norm interpolation achieves

\[
\text{Test MSE on pure noise} \asymp \sigma^2 \frac{n}{d}
\]

Phenomenon concurrently discovered in 2019 by

Belkin, Hsu and Xu; Bartlett, Long, Lugosi and Tsigler; Hastie, Montanari, Rosset and Tibshirani; M., Vodrahahalli, Subramanian and Sahai
Harmless interpolation of noise: in a picture

**Toy model:** regularly sampled 1-D data, **orthonormal** Fourier features, pure sinusoid “noise”

\[ (\Sigma = I_d) \]
Harmless interpolation of noise: in a picture

**Toy model:** regularly sampled 1-D data, *orthonormal* Fourier features, pure sinusoid “noise”

\[ \Sigma = I_d \]
Harmless interpolation of noise: in a picture

Toy model: regularly sampled 1-D data, **orthonormal** Fourier features, pure sinusoid “noise”

\( \sum = I_d \)
Harmless interpolation of noise: in a picture

**Toy model:** regularly sampled 1-D data, **orthonormal** Fourier features, pure sinusoid “noise” 

\[ \Sigma = I_d \]

\[ f(x) \]

- **High-frequency alias**
- **Zero-fit (ideal)**
- **Harmless interpolator:** combines aliases
- **Harmful interpolation**

\(~ d/n aliases — noise absorbed across aliases by min-2-norm interpolation\)

in real life: random data, random features, proofs use random matrix theory
A reminder: Why don’t we use $\ell_2$-regularization?

**Toy model:** regularly sampled 1-D data, **orthonormal** Fourier features, pure sinusoid signal ($\Sigma = I_d$)

- High-frequency alias
- $\sim d/n$ aliases — signal “bleeds” from min-2-norm interpolation

Examples of this galore in signal processing (waveform recovery), **statistical machine learning** (why LASSO is better than ridge)
A sensible model for $\ell_2$: implicit feature prioritization

$$\Sigma = \text{diag}(\Lambda) =$$

Bilevel covariance: $(n, d, s, R)$

$$\text{Weight on feature } (\lambda_j)$$

$$\text{Feature index (j)}$$

$$\text{Ratio} = R \gg 1$$

Will always interpolate noise harmlessly

(no. of prioritized features) $s \ll n$

$d \gg n$
A sensible model for $\ell_2$: implicit feature prioritization

$\Sigma = \text{diag}(\Lambda) = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_s \end{bmatrix}$

Bilevel covariance: $(n, d, s, R)$

- Will always interpolate
- Noise harmlessly

Example #1: “Weak features” model
(special case $s = 1$)

$$a_j(X) = f(X) + Z_j \quad \text{where} \quad Z_j \sim \mathcal{N}(0, \nu^2), \quad j = 1, \ldots, d.$$ 

Representative example for double descent when

$$Y = f(X) + W$$

inspired by spiked covariance ensembles and the study of minimum-Hilbert-norm interpolation in function space
A sensible model for $\ell_2$: implicit feature prioritization

Example #1: “Weak features” model
(special case $s = 1$)

$$a_j(X) = f(X) + Z_j \text{ where } Z_j \sim \mathcal{N}(0, \nu^2), \quad j = 1, \ldots, d.$$  

Representative example for double descent when

$$Y = f(X) + W$$

Example #2: Linear discriminant analysis
(special case $s = 1$)

$$a(X) = Y\mu + Z \text{ where}$$

$$Z \sim \mathcal{N}(0, \nu^2 I_d) \text{ and } Y \sim \text{Unif}[-1, 1]$$

$$\text{Ratio (R)} = \frac{\|\mu\|_2^2}{\nu^2}$$

Will always interpolate
noise harmlessly

*inspired by spiked covariance ensembles and the study of minimum-Hilbert-norm interpolation in function space*
Questions we will answer for overparameterized linear model

- Regression (review): What happens when we **interpolate** noise?

- Classification: When is the **minimum-2-norm interpolation of labels** consistent?

- Classification: When is the hard-margin **support-vector-machine (SVM)** consistent?
Overparameterized linear classification

\[ Y = \begin{cases} 
\text{sign}(\phi(X)\beta^*) & \text{w.p. } 1 - \nu \\
-\text{sign}(\phi(X)\beta^*) & \text{w.p. } \nu. 
\end{cases} \]

\[ \mathbb{E}[\phi(X)] = 0, \quad \mathbb{E}[\phi(X)\phi(X)^\top] = \Sigma \]

(no. of features) \( d \gg n \) (no. of samples)
Overparameterized linear classification

\[ Y = \begin{cases} \text{sign}(\phi(X)^\top \beta^*) \text{ w.p. } 1 - \nu \\ -\text{sign}(\phi(X)^\top \beta^*) \text{ w.p. } \nu. \end{cases} \]

\[ \mathbb{E}[\phi(X)] = 0, \quad \mathbb{E}[\phi(X)\phi(X)^\top] = \Sigma \]

Minimum-2-norm interpolation

\[ \hat{\beta} = \arg \min ||\beta||_2 \]

subject to

\[ A \hat{\beta} \approx Y \]

(label noise)

(only at training time)

(features)

(dimension = \(d\))

(no. of features) \(d \gg n\) (no. of samples)
Classification easier than regression

Historical evidence of classification easier than regression: low-dimensional settings, non-asymptotics (Devroye et al, Koltchinskii et al)

Regimes in the bilevel covariance model \( (n,d,s,R) \)
Classification easier than regression

Regression is consistent (Bartlett, Long, Lugosi & Tsigler)

Regimes in the bilevel covariance model (n,d,s,R)

Historical evidence of classification easier than regression: low-dimensional settings, non-asymptotics (Devroye et al, Koltchinskii et al)
Classification easier than regression

Historical evidence of classification easier than regression: low-dimensional settings, non-asymptotics (Devroye et al, Koltchinskii et al)

Classification is consistent, but regression is not
(M., Narang, Subramanian, Belkin, Hsu and Sahai)

Regression is consistent
(Bartlett, Long, Lugosi & Tsigler)

Regimes in the bilevel covariance model (n,d,s,R)
Classification easier than regression

Historical evidence of classification easier than regression: low-dimensional settings, non-asymptotics (Devroye et al, Koltchinskii et al)

Classification error $\frac{1}{n}$, Regression error $\frac{1}{d}$

Classification is consistent, but regression is not (M., Narang, Subramanian, Belkin, Hsu and Sahai)

Regression is consistent (Bartlett, Long, Lugosi & Tsigler)

Neither regression nor classification consistent (M., Narang, Subramanian, Belkin, Hsu and Sahai)

Regimes in the bilevel covariance model $(n,d,s,R)$
A sensible model: implicit feature prioritization

Can be shown that this assumption is necessary for consistency of min-2-norm interpolation (e.g. Tsigler and Bartlett, 2020).

\[ \Sigma = \text{diag}(\Lambda) = \]

Bilevel ensemble: \((n, d, s, R)\)

Will always interpolate noise harmlessly

\[ \text{Feature index (j)} \]

\[ s \ll n \]

\[ \text{Weight on feature } (\lambda_j) \]

\[ 10^{-6} \quad 10^{-5} \quad 10^{-4} \quad 10^{-3} \quad 10^{-2} \quad 10^{-1} \quad 10^0 \]

\[ 10^0 \quad 10^1 \quad 10^2 \quad 10^3 \quad 10^4 \]

\[ d \gg n \]

\[ \text{Ratio } = R \gg 1 \]
A sensible model: implicit **feature prioritization**

Can be shown that this assumption is necessary for consistency of min-2-norm interpolation (e.g. Tsigler and Bartlett, 2020)

\[
\Sigma = \text{diag}(\Lambda) =
\]

Bilevel ensemble: \((n, d, s, R)\)

\[
\text{Ratio} = R \gg 1
\]

Will **always** interpolate noise harmlessly

**Assumption: “known sparsity”**

\[
\beta_j^* \begin{cases} 
\neq 0 & \text{if } 1 \leq j \leq s \\
= 0 & \text{otherwise.}
\end{cases}
\]

w.l.o.g. can consider **1-sparsity**

Can be shown that this assumption is **necessary** for consistency of min-2-norm interpolation (e.g. Tsigler and Bartlett, 2020)
A sensible model: implicit **feature prioritization**

Can be shown that this assumption is **necessary** for consistency of min-2-norm interpolation (e.g. Tsigler and Bartlett, 2020)

\[ \Sigma = \text{diag}(\Lambda) = \]

\[ \frac{R}{1} \]

\[ \text{(no. of prioritized features) } s \ll n \]

\[ \text{Feature index (j)} \]

\[ \text{Weight on feature } (\lambda_j) \]

\[ \text{Ratio } = R \gg 1 \]

**Will always interpolate noise harmlessly**

**Assumption: “known sparsity”**

\[ \beta_j^* \begin{cases} 
\neq 0 & \text{if } 1 \leq j \leq s \\
= 0 & \text{otherwise.}
\end{cases} \]

w.l.o.g. can consider **1-sparsity**

Can be shown that this assumption is **necessary** for consistency of min-2-norm interpolation (e.g. Tsigler and Bartlett, 2020)
Recall: regression error

$$\Sigma = \text{diag}(\Lambda) =$$

**Theorem** (Bartlett, Long, Lugosi and Tsigler): For regression with the \textbf{min-2-norm interpolation},

$$\text{Test MSE } \approx \left( \frac{d-s}{d-s+nR} \right)^2$$

$$\rightarrow 0 \text{ as } n \rightarrow \infty \text{ iff } R \gg \frac{d}{n}.$$
Classification error: A sharp characterization

\[ \Sigma = \text{diag}(\Lambda) = \]

Bilevel covariance: \((n, d, s, R)\)

\[ \text{Weight on feature } (\lambda_j) \]

\[ \frac{\text{Ratio}}{R \gg 1} \]

(no. of prioritized features)

Will always interpolate noise harmlessly

**Theorem (M., Narang, Subramanian, Belkin, Hsu and Sahai):** For classification with the min-2-norm interpolation,

\[ \text{Test 0-1 loss} \approx \frac{1}{2} - \tan^{-1} \left( \frac{R}{\sqrt{(d-s)/n}} \right) \]

\[ \rightarrow 0 \text{ as } n \rightarrow \infty \text{ iff } R \gg \sqrt{\frac{d}{n}} \]

Classfn. error = \(1/2 - \tan^{-1}(SU/CN)\)

Survival SU

Contamination variance CN (like error from interpolation of noise)

Benign-ness of 0-1 loss described in Friedman, “On Bias, Variance, 0-1 loss, and the curse of dimensionality”
# Summary: Implications for \textit{consistency}

<table>
<thead>
<tr>
<th>Ratio (R)</th>
<th>$\gg d/n$</th>
<th>$\gg \sqrt{d/n}$, $\ll d/n$</th>
<th>$\ll \sqrt{d/n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymptotic classification test error, interpolation</td>
<td>0</td>
<td>0</td>
<td>$1/2$</td>
</tr>
</tbody>
</table>
Questions we will answer for overparameterized linear model

- Regression (review): What happens when we interpolate noise?
- Classification: When are solutions that interpolate labels consistent?
- Classification: When is the hard-margin support-vector-machine (SVM) consistent?
Overparameterized linear classification

\[ Y = \begin{cases} 
\text{sign}(\phi(X)^\top \beta^*) \text{ w.p. } 1 - \nu \\
-\text{sign}(\phi(X)^\top \beta^*) \text{ w.p. } \nu.
\end{cases} \]

\[ \mathbb{E}[\phi(X)] = 0, \quad \mathbb{E}[\phi(X)\phi(X)^\top] = \Sigma \]

(no. of features) \( d \gg n \) (no. of samples)
Overparameterized linear classification

\[ Y = \begin{cases} 
\text{sign}(\phi(X)^\top \beta^*) & \text{w.p. } 1 - \nu \\
-\text{sign}(\phi(X)^\top \beta^*) & \text{w.p. } \nu. 
\end{cases} \]

\[ \mathbb{E}[\phi(X)] = 0, \quad \mathbb{E}[\phi(X)\phi(X)^\top] = \Sigma \]

Hard-margin linear SVM

\[ \hat{\beta} = \text{arg min} ||\beta||_2 \]

subject to

\[ Y_i \cdot \phi(X_i)^\top \beta \geq 1, \quad i = 1, \ldots, n. \]

• Feasible solutions include **min-2-norm interpolation of binary labels**
• Outcome of GD on **last layer of neural network** achieving **zero training error**
SVM = interpolation with high probability

\[ \Sigma = \text{diag}(\lambda), \quad \lambda_j = \frac{1}{\sqrt{j}} \]
SVM = interpolation with high probability

Theorem (Hsu, M. and Xu): **hard-margin SVM = minimum-l2-interpolation** w.h.p. under sufficient overparameterization:

\[
\frac{||\lambda||_2^2}{||\lambda||_2} \gg n \quad \text{and} \quad \frac{||\lambda||_1}{||\lambda||_{\infty}} \gg n \log n
\]

\[
\Sigma = \text{diag}(\lambda), \quad \lambda_j = \frac{1}{\sqrt{j}}
\]
SVM = interpolation with high probability

\[ \Sigma = \text{diag}(\lambda), \lambda_j = \frac{1}{\sqrt{j}} \]

**Theorem** (Hsu, M. and Xu): **hard-margin SVM = minimum-l2-interpolation** w.h.p. under sufficient overparameterization:

\[
\frac{||\lambda||_2^2}{||\lambda||_2} \gg n \quad \text{and} \quad \frac{||\lambda||_1}{||\lambda||_\infty} \gg n \log n
\]

Implies equivalence in the bilevel covariance model if \( R \ll d/n \)
Implications for **consistency**

<table>
<thead>
<tr>
<th>Ratio (R)</th>
<th>$\gg d/n$</th>
<th>$\gg \sqrt{d/n}, \ll d/n$</th>
<th>$\ll \sqrt{d/n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymptotic classification test error, interpolation</td>
<td>0</td>
<td>0</td>
<td>1/2</td>
</tr>
<tr>
<td>Asymptotic classification test error, SVM</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Implications for consistency

<table>
<thead>
<tr>
<th>Ratio (R)</th>
<th>$d/n$</th>
<th>$\sqrt{d/n}$, $d/n$</th>
<th>$\sqrt{d/n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymptotic classification test error, interpolation</td>
<td>0</td>
<td>0</td>
<td>1/2</td>
</tr>
<tr>
<td>Asymptotic classification test error, SVM</td>
<td>0 (margin-based/sampling compression bounds)</td>
<td>0 (equivalence of SVM and interpolation)</td>
<td>1/2 (equivalence of SVM and interpolation)</td>
</tr>
</tbody>
</table>

Montanari, Ruan, Sohn and Yan, Chatterji and Long also analyze the SVM directly.
Proof technique and a connection to harmless interpolation
Proof technique and a connection to harmless interpolation

\((\mathbf{A}_{-i}, \mathbf{Y}_{-i})\) denote data matrix and output vector without the \(i^{th}\) training sample.
(\mathbf{A}_{-i}, \mathbf{Y}_{-i}) \text{ denote data matrix and output vector without the } i^{th} \text{ training sample }

\textbf{Dual witness technique: } we show that everything is a support vector iff 
\[ Y_i \mathbf{Y}_{-i}^T (\mathbf{A}_{-i} \mathbf{A}_{-i}^T)^{-1} \mathbf{A}_{-i} \phi(X_i) < 1 \text{ for all } i = 1, \ldots, n. \]
Proof technique and a connection to harmless interpolation

\[(A_{-i}, Y_{-i})\] denote data matrix and output vector without the \(i^{th}\) training sample

**Dual witness technique:** we show that everything is a support vector iff

\[Y_i Y_i^T (A_{-i} A_{-i}^T)^{-1} A_{-i} \phi(X_i) < 1 \text{ for all } i = 1, \ldots, n.\]

Can compare this with the excess test MSE arising from

**min-2-norm interpolation on noise:**

\[
\text{Test MSE} = \mathbb{E}_{X_{\text{new}}} \left[ (W^T (A A^T)^{-1} A \phi(X_{\text{new}}))^2 \right]
\]
Proof technique and a connection to harmless interpolation

\[(A_{-i}, Y_{-i})\] denote data matrix and output vector without the \(i^{th}\) training sample

**Dual witness technique:** we show that everything is a support vector iff

\[Y_i Y_{-i}^T (A_{-i} A_{-i}^T)^{-1} A_{-i} \phi(X_i) < 1 \text{ for all } i = 1, \ldots, n.\]

Can compare this with the excess test MSE arising from min-2-norm interpolation on noise:

\[
\text{Test MSE} = \mathbb{E}_{X_{\text{new}}} \left[ (W^T (A A^T)^{-1} A \phi(X_{\text{new}}))^2 \right]
\]

High-level takeaway: if all training points become support vectors, interpolation of noise will be harmless
Does the loss function matter?

- Classification asymptotically more benign than regression in high-dimensional regimes

- Classification can work well despite ultra-high-dimensionality, poor signal recovery and interpolation

- The SVM interpolates binary labels in ultra-high-dimensional settings

- Different loss functions at training time can yield similar solutions

And all of these insights connected to harmless interpolation!
Current research directions

• Corresponding analyses for **minimum-l1-norm interpolation**

• Corresponding analyses for **logistic test error**

• Characterizing **adversarial error**
A full theory of interpolation: old and new

minimum-norm interpolators

Deep nets interpolate data

Recent analyses of
kernel ridge regression,
linear regression

('14-'17)

('18-'20)
A full theory of interpolation: old and new

- **Kernel smoothers**
  - '68
    - Nadaraya-Watson kernel smoothing with singular “Hilbert” kernel interpolates data
      (Shepard)
  - '98
    - Hilbert kernel shown to be consistent
      (Devroye, Györfi and Krzyżak)
  - Deep nets interpolate data
  - '18
    - Hilbert kernel shown to be minimax-optimal
      (Belkin, Rakhlin and Tsybakov)
  - '18-'20
    - Recent analyses of kernel ridge regression, linear regression

- **Minimum-norm interpolators**
A full theory of interpolation: old and new

Kernel smoothers
Nearest-neighbors and variants
Minimum-norm interpolators

1-nearest-neighbor achieves twice the Bayes risk
(Cover and Hart)

Nadaraya-Watson kernel smoothing with singular “Hilbert” kernel interpolates data (Shepard)

Hilbert kernel shown to be consistent (Devroye, Györfi and Krzyżak)

Deep nets interpolate data

Hilbert kernel shown to be minimax-optimal (Belkin, Rakhlin and Tsybakov)

Recent analyses of kernel ridge regression, linear regression

Simplicial interpolation exponentially better than 1-NN on high-dimensional data (Belkin, Hsu and Mitra)

Deep nets interpolate data
A full theory of interpolation: old and new

- Kernel smoothers
- Nearest-neighbors and variants
- Minimum-norm interpolators

1-nearest-neighbor achieves twice the Bayes risk (Cover and Hart)

Nadaraya-Watson kernel smoothing with singular “Hilbert” kernel interpolates data (Shepard)

Hilbert kernel shown to be consistent (Devroye, Györfi and Krzyżak)

Hilbert kernel shown to be minimax-optimal (Belkin, Rakhlin and Tsybakov)

Deep nets interpolate data: '18 '18 '18-'20 ('14-'17)

Simplicial interpolation exponentially better than 1-NN on high-dimensional data (Belkin, Hsu and Mitra)

Connections between analyses? Implications for neural networks?