Classification vs Regression in Overparameterized Regimes: Does the Loss Function Matter?

Adhyyan Narang

University of Washington

August 8, 2024
Collaborators

Joint work with:

Vidya Muthukumar  Vignesh Subramanian  Misha Belkin  Daniel Hsu  Anant Sahai
**Empirical Observation**

| model           | # params | train accuracy | test accuracy |
|-----------------|----------|----------------|---------------|
| Inception       | 1,649,402| 100.0          | 89.05         |
| (fitting random labels) | 100.0   | 89.31          | 86.03         |
|                 | 100.0   | 85.75          |               |
| Inception w/o BatchNorm | 1,649,402| 100.0          | 83.00         |
| (fitting random labels) | 100.0   | 82.00          | 10.12         |
| Alexnet         | 1,387,786| 99.90          | 81.22         |
| (fitting random labels) | 99.82  | 79.66          | 77.36         |
|                 | 100.0   | 76.07          |               |
| MLP 3x512       | 1,735,178| 100.0          | 53.35         |
| (fitting random labels) | 100.0 | 52.39          |               |
| MLP 1x512       | 1,209,866| 99.80          | 50.39         |
| (fitting random labels) | 99.34 | 50.51          |               |

- From Zhang et.al "Understanding Deep Learning Requires Rethinking Generalization (2016)".
- CIFAR 10 (50,000 train examples)
- Benign overfitting happens for classification too
Regression Very Quick Recap

Minimum 2-norm interpolator

\[ \hat{\alpha}_{\text{MNI}} = \min_{\alpha \in \mathbb{R}^d} \| \alpha \| \]

s.t \( X_i^\top \alpha = Y_i \) for all \( i = 1 \ldots n \)

This admits the closed form expression:

\[ \hat{\alpha}_{\text{MNI}} = A_{\text{train}}^\dagger Y_{\text{train}}. \]

Analysis of MSE Risk

\[ \mathcal{E}_{\text{test}}(\hat{\alpha}) \]
\[ = \mathbb{E} \left[ (\langle X, \alpha^* \rangle + \epsilon - \langle X, \hat{\alpha} \rangle)^2 \right] \]
\[ = \mathbb{E} \left[ (\langle X, \hat{\alpha} - \alpha^* \rangle)^2 \right] + \mathbb{E}[\epsilon^2] \]
\[ = \mathbb{E} \left[ (\hat{\alpha} - \alpha^*)^\top X X^\top (\hat{\alpha} - \alpha^*) \right] + \sigma^2 \]
\[ = (\hat{\alpha} - \alpha^*)^\top \Sigma (\hat{\alpha} - \alpha^*) + \sigma^2 \]
\[ = \| \Sigma^{1/2}(\hat{\alpha} - \alpha^*) \|^2_2 - \sigma^2 \]
\[ = \| \Sigma^{1/2}(\hat{\alpha} - \alpha^*) \|^2_2. \]
Analyzing classification is more challenging

### Minimum 2-norm interpolator

$$\hat{\alpha}_{\text{MNI}} = \min_{\alpha \in \mathbb{R}^d} \|\alpha\| \quad \text{s.t.} \quad X_i^\top \alpha = Y_i \quad \text{for all} \quad i = 1 \ldots n$$

This admits the closed form expression:

$$\hat{\alpha}_{\text{MNI}} = A_{\text{train}}^\dagger Y_{\text{train}}.$$

### Support Vector Machine

$$\hat{\alpha}_{\text{SVM}} = \min_{\alpha \in \mathbb{R}^d} \|\alpha\| \quad \text{s.t.} \quad Y_i X_i^\top \alpha \geq 1 \quad \text{for all} \quad i = 1, \ldots, n.$$

Now, the solution is not in closed form anymore, and the risk does not admit an easy form.
1. Setup

2. Proliferation of Support Vectors

3. Benign overfitting: Classification v/s Regression
**Data Model**

**Gaussian Features** \( X_i \sim \mathcal{N}(0, \Sigma) \)

Denote by \( \Lambda = [\lambda_1 \ldots \lambda_n] \) the spectrum of \( \Sigma \)

**Labels**

\[
Z_i = \langle X_i, \alpha^* \rangle \quad \text{and}
\]

\[
Y_i = \begin{cases} 
\text{sgn}(Z_i) \quad \text{with probability} \quad (1 - \nu^*) \\
-\text{sgn}(Z_i) \quad \text{with probability} \quad \nu^*.
\end{cases}
\]
Interpolating Estimators, Risk

**Interpolators**

\[
\hat{\alpha}_{\text{binary}} = \min_{\alpha \in \mathbb{R}^d} \|\alpha\| \quad \text{s.t} \quad X_i^\top \alpha = Y_i
\]

\[
\hat{\alpha}_{\text{real}} = \min_{\alpha \in \mathbb{R}^d} \|\alpha\| \quad \text{s.t} \quad X_i^\top \alpha = Z_i
\]

\[
\hat{\alpha}_{\text{SVM}} = \min_{\alpha \in \mathbb{R}^d} \|\alpha\| \quad \text{s.t} \quad Y_i X_i^\top \alpha \geq 1
\]

Third = First when all constraints are tight.

**Regression Risk**

\[
\mathcal{R}(\hat{\alpha}) = \mathbb{E}[\langle X, \alpha^* - \hat{\alpha} \rangle^2]
\]

**Classification Risk**

\[
\mathcal{C}(\hat{\alpha}) = \mathbb{P}[\text{sgn}(\langle X, \hat{\alpha} \rangle) \neq \text{sgn}(\langle X, \alpha^* \rangle)]
\]
1. Setup

2. Proliferation of Support Vectors

3. Benign overfitting: Classification v/s Regression
Curious Empirical Observation

- Fix $n = 32$ and $\Sigma = I$
Theoretical Result

**Theorem**

If \( \Sigma = I_d \) and \( d > n \log(n) + n - 1 \), then for any fixed \( Y_{\text{train}} \in \{-1, 1\}^n \), we have with probability \((1 - \frac{2}{n})\)

\[
\hat{\alpha}_{\text{binary}} = \hat{\alpha}_{\text{SVM}}
\]
Curious Empirical Observation 2

- Fix \( n = 519 \), \( d = 12167 \) and vary \( \Lambda \).
- As “effective overparameterization” is increased, the fraction of support vectors increases.
Theoretical Result

**Theorem**

If $\Sigma$ satisfies

$$\frac{\|\Lambda\|_1}{\|\Lambda\|_2} \geq n\sqrt{\log(n)} \quad \text{and} \quad \frac{\|\Lambda\|_1}{\|\Lambda\|_\infty} \geq n\sqrt{n \log(n)}$$

then simultaneously for all $Y_{\text{train}} \in \{-1, 1\}^n$, we have with probability $(1 - \frac{2}{n})$

$$\hat{\alpha}_{\text{binary}} = \hat{\alpha}_{\text{SVM}}$$

- Note that $d \geq \left(\frac{\|\Lambda\|_1}{\|\Lambda\|_2}\right)^2 \geq \frac{\|\Lambda\|_1}{\|\Lambda\|_\infty}$.
- In the isotropic setting, these are all equal.
- So these ratios measure how far we are from isotropic.
Equivalence of Loss Functions

The outcome of training loss functions in the linear model (separable data)

- **Logistic loss**
  - Gradient Descent
  - (Ji and Telgarsky, 2019)
  - (Soudry et al., 2018)

- **Squared loss**
  - Gradient Descent, initialized at 0
  - (Engl et al., 1996)

---

Hard-margin SVM

Equivalent with sufficient overparameterization

Min $\ell_2$-norm interpolation
Proof technique

• By complementary slackness, the $i$th point is a support vector when the $i$th dual constraint is strictly feasible.
• Dual condition is expressed cleanly, and goes through when Gram matrix is close to diagonal.
• This happens in high dimensions whp

Intuition

• In the small $d$ or highly anisotropic case, a lot of weight is placed on small features.
• So you would probably overshoot the constraint.
• But when you have many features to use, you have more “fine-grained control” and is cheaper to be tight.
"On the proliferation of support vectors in high dimensions" Hsu, Muthukumar, Xu (2020): Sharpens the second theorem here, and provides a converse result

"Support vector machines and linear regression coincide with very high-dimensional features.” Ardeshir, Sanford, Hsu (2021): Show that above paper is tight

"Benign overfitting in binary classification of gaussian mixtures” Wang, Thrampoulidis (2021): Show the same for Gaussian Mixture Models

"Benign overfitting in multiclass classification: All roads lead to interpolation.” Wang, Muthukumar, Thrampoulidis (2021): Multiclass extension
Assumption (1-sparse) For some unknown $t \in \{1 \ldots s\}$, assume that $\alpha^* = e_t$
Survival and Contamination

Survival (Signal Recovery)

\[ SU(\hat{\alpha}) = \frac{\hat{\alpha}_t}{\alpha^*_t} \]

Contamination (False discovery of features)

\[ B = \sum_{j \neq t} \hat{\alpha}_j X_j \]
\[ CN(\hat{\alpha}) = \sqrt{\mathbb{E}[B^2]} \]

Then,

\[ R(\hat{\alpha}) = (1 - SU(\hat{\alpha}))^2 + CN(\hat{\alpha})^2 \]

And,

\[ C(\hat{\alpha}) = 1 - \tan^{-1} \left( \frac{SU(\hat{\alpha})}{CN(\hat{\alpha})} \right) \]
Results

Theorem (Bartlett, Long, Lugosi and Tsigler)

\[ R(\hat{\alpha}_{\text{real}}) \approx \left( \frac{d - s}{d - s + nR} \right)^2 \]

Taking the limit,

\[ \to 0 \text{ as } n \to \infty \text{ if and only if } R \gg \frac{d}{n}. \]

Theorem (Present Work)

\[ C(\hat{\alpha}_{\text{binary}}) \approx \frac{1}{2} - \tan^{-1} \left( \frac{R}{\sqrt{(d - s)/n}} \right) \]

\[ \to 0 \text{ as } n \to \infty \text{ if and only if } R \gg \sqrt{\frac{d}{n}}. \]
## Separating Regime

| Ratio (R)      | $\gg \frac{d}{n}$ | $\gg \sqrt{\frac{d}{n}}$, $\ll \frac{d}{n}$ | $\ll \sqrt{\frac{d}{n}}$ |
|---------------|--------------------|--------------------------------------------|-----------------|
| Classification| 0                  | 0                                          | $\frac{1}{2}$  |
| Regression    | 0                  | 1                                          | 1               |

**Note:**
- Benign overfitting does not always happen – it depends on the quality of features and the razor.
- The second and third column co-incide with the regime where support vectors proliferate.
Summary

• With high enough effective overparameterization, support vectors proliferate.
• This paves the way to analyze the SVM by looking at the 2-norm interpolator.
• Identify clear separating regimes between regression and classification.

Since then:
• Community: Extend to multiclass, kernels, mixture models.
• My work: The same phenomena that lead to benign overfitting cause adversarial examples! Would be happy to give a talk on this at some point.