Benign Overfitting with Interpolating Linear Classifier without Subgaussianity

Ichiro Hashimoto

Department of Statistical Sciences, University of Toronto

UNIVERSITY OF TORONTO

Motivation

- Practical success of deep neural networks has provoked theoretically surprising phenomena in statistics. One of these phenomena, that has spurred intense theoretical research, is "benign overfitting": deep neural networks seem to generalize well in over-parametrized regime even though the networks show a perfect fit to noisy training data.
- It is now known that benign overfitting also occurs in various classical statistical models.
- For binary linear classification, previous works have proven that benign overfitting can occur while assuming that data are generated from subgaussian mixtures and the results were limited to specific regimes.

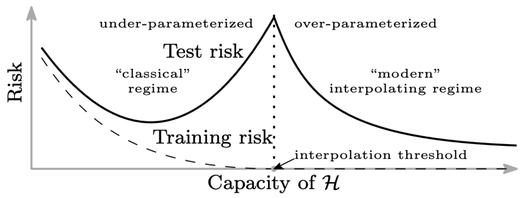


Figure 1. Risk curve: Classical regime vs. Over-parametrized regime[1]

non-Subgaussian Mixture Model

We consider binary linear classification, where the p-dimensional feature vectors $x_i, i = 1, \ldots, n$ are generated from a mixture of two distributions with mean μ and $-\mu$ with the same covariance matrix Σ.

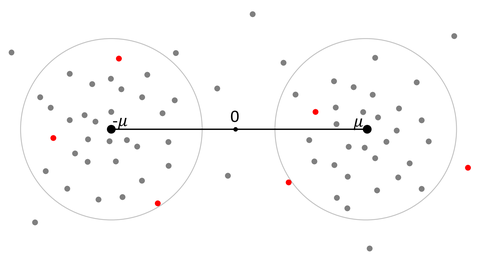


Figure 2. non-Subgaussian Mixture

In precise, the model on the feature vector x is induced by the relation

$$x \;=\; y\mu + z,$$

where we assume:

- $y \in \{-1, 1\}$ is a random variable satisfying $P(y = 1) = P(y = -1) = \frac{1}{2}$,
- $z = \Sigma^{1/2}\xi$, where $\Sigma \in \mathbb{R}^{p\times p}$ is positive definite and $\xi \in \mathbb{R}^p$ has independent entries $\xi_{(k)}, k = 1, \ldots, p$ that have mean zero and unit variance,
- the rth moments of the entries of ξ is bounded by K for some $2 < r \leq 4$.

Noise is introduced to the label y by flipping the sign of y with probability η. We call y a clean label and $\tilde{y}$ a noisy label.

Draw n samples $\{(x_i, \tilde{y}_i), i = 1, \ldots, n\}$ randomly from the distribution of $(x, \tilde{y})$.

Max Margin Classifier and Implicit Bias

We consider the maximum margin classifier $\hat{w}$, i.e. the solution to the hard-margin support vector machine:

$$\hat{w} \;=\; \arg\min \|w\|^2, \qquad \text{subject to } \langle w, \tilde{y}_i x_i\rangle \geq 1 \text{ for all } i = 1, 2, \ldots, n.$$

Our analysis on the maximum margin classifier is motivated by implicit bias induced by gradient descent on the logistic loss:

$$w_{t+1} = w_t - \eta\nabla_w L(w_t), \quad w_0 = 0, \quad t = 0, 1, 2, \ldots, \tag{1}$$

where $L(w)$ is defined by

$$L(w) \;:=\; \frac{1}{n}\sum_{i=1}^{n} \log\{1 + \exp(-\langle w, \tilde{y}_i x_i\rangle)\}.$$

([4]) showed that, when the dataset is linearly separable ($\exists w \in \mathbb{R}^p$ such that $\langle w, y_i x_i\rangle > 0$ for all i), linear classifier optimized by gradient descent (1) on the logistic loss with sufficiently small step size η converges in direction to the maximum margin classifier, that is

$$\lim_{t\to\infty} \frac{w_t}{\|w_t\|} \;=\; \frac{\hat{w}}{\|\hat{w}\|}.$$

Theorem 1 (Weak Signal Regime)

For any $\delta \in (0, 1/3]$, suppose $\|\mu\|^2 \geq C\dfrac{\|\Sigma^{1/2}\mu\|}{\sqrt{\delta}}$ and

$$\mathrm{Tr}(\Sigma) \geq C\max\left\{n\|\mu\|^2, \frac{n^{2/r+1/2}p^{2/r-1/2}\|\Sigma\|_F}{\delta^{2/r}}, \frac{n^{4/r+1/2}\|\Sigma\|_F}{\delta^{2/r}}, \frac{n^{3/2}\|\Sigma^{1/2}\mu\|}{\sqrt{\delta}}\right\}$$

for a sufficiently large constant $C > 1$.

Then, for sufficiently large n, with probability at least $1 - 3\delta$, there exists some constant $\tilde{c}$ such that

$$\mathbb{P}_{(x,\tilde{y})}\left(\langle\hat{w}_N, \tilde{y}x\rangle < 0\right) \leq \eta + \frac{\tilde{c}_1}{(1-2\eta)^2}\cdot\frac{\|\Sigma\|\mathrm{Tr}(\Sigma)}{n\|\mu\|^4}.$$

In particular, benign overfitting is guaranteed if $\mathrm{Tr}(\Sigma) = o(n\|\mu\|^4/\|\Sigma\|)$.

Theorem 2 (Strong Signal Regime)

For any $\delta \in (0, 1/3]$, suppose either of the following conditions is met for a sufficiently large constant $C > 1$:

- $\|\mu\|^2 \geq C\dfrac{\|\Sigma^{1/2}\mu\|}{\sqrt{\delta}}$ and

  $$C\max\left\{\frac{n^{2/r+1/2}p^{2/r-1/2}\|\Sigma\|_F}{\delta^{2/r}}, \frac{n^{4/r+1/2}\|\Sigma\|_F}{\delta^{2/r}}, \frac{n^{3/2}\|\Sigma^{1/2}\mu\|}{\sqrt{\delta}}\right\} \leq \mathrm{Tr}(\Sigma) \ll n\|\mu\|^2$$

  , or
- $\|\mu\| \geq C\sqrt{\mathrm{Tr}(\Sigma)}$ and

  $$\mathrm{Tr}(\Sigma) \geq C\max\left\{\frac{n^{2/r+1/2}p^{2/r-1/2}\|\Sigma\|_F}{\delta^{2/r}}, \frac{n^{4/r+1/2}\|\Sigma\|_F}{\delta^{2/r}}, \frac{n^{3/2}\|\Sigma^{1/2}\mu\|^2}{\delta\|\mu\|^2}\right\}.$$

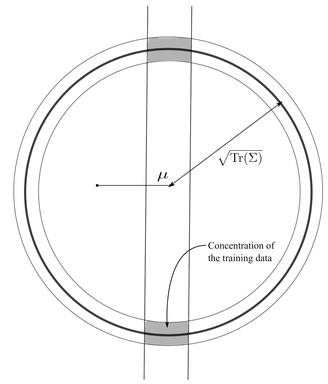
Then, for sufficiently large n, with probability at leas t $1 - \delta$, there exists some constant c such that

$$\mathbb{P}_{(x,\tilde{y})}\left(\langle\hat{w}_N, \tilde{y}x\rangle < 0\right) \leq \eta + \frac{c}{(1-2\eta)^2}\cdot\frac{n\|\Sigma\|}{\mathrm{Tr}(\Sigma)}.$$

In particular, benign overfitting is guaranteed if $\mathrm{Tr}(\Sigma) = \omega(n\|\Sigma\|)$.

Geometry behind Over-parametrized Regime

Concentration of Training Data



Phase Transition in Direction of Max Margin Classifier

$$\frac{\hat{w}}{\|\hat{w}\|^2} \approx \begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\tilde{y}_i x_i & \text{if } \|\mu\| \ll \sqrt{\dfrac{\mathrm{Tr}(\Sigma)}{n}} \\ \dfrac{1}{2}\left\{\dfrac{1}{(1-\eta)n}\displaystyle\sum_{i:\text{clean}}\tilde{y}_i x_i + \dfrac{1}{\eta n}\sum_{i:\text{noisy}}\tilde{y}_i x_i\right\} & \text{if } \|\mu\| \gg \sqrt{\dfrac{\mathrm{Tr}(\Sigma)}{n}}. \end{cases}$$

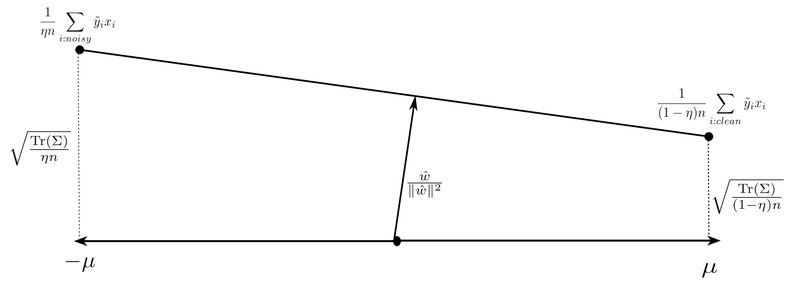


Figure 3. Strong Signal Regime

References

[1] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal.
Reconciling modern machine-learning practice and the classical bias–variance trade-off.
Proceedings of the National Academy of Sciences, 116(32):15849–15854, 2019.

[2] Yuan Cao, Quanquan Gu, and Mikhail Belkin.
Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures, 2022.

[3] Niladri S. Chatterji and Philip M. Long.
Finite-sample analysis of interpolating linear classifiers in the overparameterized regime.
Journal of Machine Learning Research, 22(129):1–30, 2021.

[4] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro.
The implicit bias of gradient descent on separable data, 2022.

[5] Ke Wang and Christos Thrampoulidis.
Binary classification of gaussian mixtures: Abundance of support vectors, benign overfitting, and regularization.
SIAM Journal on Mathematics of Data Science, 4(1):260–284, 2022.

[6] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals.
Understanding deep learning requires rethinking generalization.
In International Conference on Learning Representations, 2017.

Github: https://ichiro-h.github.io/

ichiro.hashimoto@mail.utoronto.ca