Lecture 10: Extensions to Classification

Geoff Pleiss

So far all of our results for linear models and neural networks have dealt with regression problems. What about classification—the setting neural networks are most commonly used for? We would like build classification analogs to the results we obtained for regression.

1. Gradient-based optimization biases overparameterized linear models to a "nice" global minimum.
2. We can characterize the asymptotic risk of these "nice" linear models in terms of the overparameterization ratio and the covariance spectrum.
3. Classification neural networks are well approximated by linear models, even throughout the course of optimization.

We will focus on (1) in this lecture.

---

1) Global Minima of Overparameterized Linear Classification

Consider the following linear binary classifier that assigns a class label $\{-1, 1\}$ to an input $\boldsymbol{x} \in \mathbb{R}^d$:

$$\hat{f}(\boldsymbol{x}) = \text{sign}\left(\hat{\boldsymbol{\theta}}^\top \boldsymbol{x}\right), \qquad \boldsymbol{\theta} \in \mathbb{R}^d. \tag{1}$$

We are given a training set $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ where $y_i \in \{-1, 1\}$ which is **linearly separable.** That is, there exists some $\boldsymbol{\theta}_s$ such that $y_i \boldsymbol{\theta}_s^\top \boldsymbol{x}_i > 0$ for all $i \in [1, n]$ (alternatively, $\text{sign}\left(\boldsymbol{\theta}_s^\top \boldsymbol{x}_i\right) = y_i$). Separability is guaranteed to occur when $d > n$ (i.e. in the overparameterized setting) and the $\boldsymbol{x}_i$ are not colinear.

Let $\boldsymbol{\Theta}_s$ be the set of all unit norm $\boldsymbol{\theta}$ that perfectly separate the training set:

$$\boldsymbol{\Theta}_s = \left\{\boldsymbol{\theta} \in \mathbb{R}^d \ : \ y_i \boldsymbol{\theta}^\top \boldsymbol{x}_i > 0 \text{ for all } i \in [1, n]\right\}. \tag{2}$$

A few notes are necessary here:

1. If $\boldsymbol{\theta}_s \in \boldsymbol{\Theta}_s$, then $\alpha\theta_s \in \boldsymbol{\Theta}_s$ for all $\alpha > 0$.
2. Let $\boldsymbol{\theta}_c$ be some vector in the nullspace of $\{\boldsymbol{x}_i\}_{i=1}^n$. Then $\boldsymbol{\theta}_s + \boldsymbol{\theta}_c \in \boldsymbol{\Theta}_s$.
3. With a simple analysis argument, let $\boldsymbol{\theta}_s \in \boldsymbol{\Theta}_s$, and let $\boldsymbol{\theta}_m \notin \boldsymbol{\Theta}_s$ be some other vector such that $y_i \boldsymbol{\theta}_m^\top \boldsymbol{x}_i \geq 0$ (i.e. there is some i with $y_i \boldsymbol{\theta}_m^\top \boldsymbol{x}_i = 0$). Then $\lambda\boldsymbol{\theta}_s + (1 - \lambda)\boldsymbol{\theta}_m \in \boldsymbol{\Theta}_s$ for all $\lambda \in (0, 1)$.

We have thus proven that there are an infinite number of vectors in $\boldsymbol{\Theta}_s$. Moreover, taking arguments (2) and (3), we have that the vectors in $\boldsymbol{\Theta}_s$ form a basis of $\mathbb{R}^d$! As with overparameterized linear regression, finding a classifier that achieves perfect training set accuracy is an underconstrained problem.

---

2) Gradient Descent Selects a "Nice" Global Minimum

For notational simplicity, we will use $\boldsymbol{z}_i = y_i \boldsymbol{x}_i$ from here on out.

In practice, we often select a $\hat{\boldsymbol{\theta}}$ by by minimizing a convex relaxation of the 0-1 loss function with gradient descent. Let's assume that we are minimizing the exponential loss function:[1]

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \exp\left(-y_i \boldsymbol{\theta}^\top \boldsymbol{x}_i\right). \tag{3}$$

It's fairly easy to characterize the global minima of this loss if our data are linearly separable.

- Assume $\boldsymbol{\theta} \in \boldsymbol{\Theta}_s$. Then $\boldsymbol{\theta}^\top \boldsymbol{z}_i > 0$ for all $i \in [1, n]$. Note that for all finite $\boldsymbol{\theta} \in \mathbb{R}^d$, we have that $\mathcal{L}(\boldsymbol{\theta}) > 0$ because $\exp(\boldsymbol{\theta}^\top \boldsymbol{z}_i)$ is strictly positive. However,

  $$\lim_{\alpha \to \infty} \mathcal{L}(\alpha \boldsymbol{\theta}) = \lim_{\alpha \to \infty} \sum_{i=1}^{n} \exp(-\alpha \boldsymbol{\theta}^\top \boldsymbol{z}_i) = 0.$$

  In other words $\lim_{\alpha \to \infty} \alpha \boldsymbol{\theta}$ is a global minimum for any $\boldsymbol{s} \in \boldsymbol{\Theta}_s$.

- Assume $\boldsymbol{\theta} \notin \boldsymbol{\Theta}_s$. Then there is some $\boldsymbol{\theta}^\top \boldsymbol{z}_i < 0$. The portion of this loss will be greater than 1 and cannot be brought down by scaling. Therefore, $\boldsymbol{\theta}$ is not a global minimum.

Gradient descent will converges to a global minimum of this convex optimization problem. We know that the limiting solution will have infinite norm because of the scaling argument above, so we are more interested in the direction of the limiting solution. Let $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}/\|\boldsymbol{\theta}\|$, where $\boldsymbol{\theta}$ is the limiting solution of gradient descent. As was the case with regression, maybe $\hat{\boldsymbol{\theta}}$ will converge to some unique "nice" global minimum, even though there is nothing explicit in the optimization problem forcing it to a particular global minimum...

2.1 The Intuitive Limiting Behaviour of Gradient Descent

Let $\boldsymbol{\theta}(t)$ be the parameter vector after t steps of gradient descent. As with our NTK optimization analysis, let's assume that gradient descent takes infinitesimally small steps, so that $\boldsymbol{\theta}(t)$ is governed by the differential equation:

$$\dot{\boldsymbol{\theta}}(t) = -\nabla \mathcal{L}(\boldsymbol{\theta}(t)) = \exp\left(-\boldsymbol{z}_i^\top \boldsymbol{\theta}(t)\right) \boldsymbol{z}_i. \tag{4}$$

Assume that we have run gradient descent for long enough so that $\boldsymbol{\theta}(t) \in \boldsymbol{\Theta}_s$—i.e. we have achieved perfect training set accuracy. At this point the amount that each data point contributes to the gradient becomes exponentially smaller as $-\boldsymbol{z}_i^\top \boldsymbol{\theta}(t)$ grows.

We will continue drive down the loss further by having the magnitude of $\boldsymbol{\theta}(t)$ grow. However, the data points that contribute the most to the gradient are those for which $-\boldsymbol{z}_i^\top \boldsymbol{\theta}(t)$ is least negative. Denoting this set of **support vectors** as $\mathbb{S}(t)$, i.e.:

$$\mathbb{S}(t) := \left\{ \boldsymbol{z}_i \,:\, |\boldsymbol{z}_i^{(t)} \boldsymbol{\theta}(t)| = \min_{i \in [1,n]} \left| \boldsymbol{z}_i^\top \boldsymbol{\theta}(t) \right| \right\}$$

the gradient will asymptotically be dominated by these points:

$$\dot{\boldsymbol{\theta}}(t) \approx \sum_{j \in \mathbb{S}(t)} \exp\left(-\boldsymbol{z}_j^\top \boldsymbol{\theta}(t)\right) \boldsymbol{z}_j,$$

[1] The results we will prove hold for other loss function, like the more common logistic loss function. However, the result is most straightforward with the exponential loss. Other loss functions have similar proofs but require more tedious bookkeeping.

and so $\boldsymbol{\theta}(t)$ is only growing in the directions spanned by $\boldsymbol{z}_j$. As $t \to \infty$ and $\|\boldsymbol{\theta}(t)\| \to \infty$, the normalized $\hat{\boldsymbol{\theta}}(t)/\|\boldsymbol{\theta}(t)\|$ will be a linear combination of these support vectors:

$$\hat{\boldsymbol{\theta}}(t) \to \sum_{i=1}^{n} \alpha_i \boldsymbol{z}_i, \qquad \begin{cases} \alpha_i > 0 & \text{if } \boldsymbol{z}_i \in \lim_{t\to\infty} \mathbb{S}(t) \\ \alpha_i = 0 & \text{if } \boldsymbol{z}_i \in \lim_{t\to\infty} \mathbb{S}(t) \end{cases} \tag{5}$$

Minimum norm interpretation. Discerning eyes will recognize Eq. (5) as the dual form of the support vector machine (SVM) classifier, which is (colinear) to the solution to the (primal) optimization problem:

$$\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_2^2 \quad \text{subject to} \quad \boldsymbol{\theta}^\top \boldsymbol{z}_i \geq 1 \quad \text{for all } i \in [1, n]. \tag{6}$$

The SVM solution is known as the **maximum margin classifier**, as it maximizes the minimum distance between all training points and the separating hyperplane. From Eq. (6) we see a close connection between the classification and regression gradient descent solutions: both are minimum norm solutions within some constrained optimization problem around data interpolation.

2.2 Proof

Let $\hat{\boldsymbol{\theta}}$ be the solution to Eq. (6). We will decompose $\boldsymbol{\theta}(t)$ into three arbitrary terms:

$$\boldsymbol{\theta}(t) = \log(t)\hat{\boldsymbol{\theta}} + \boldsymbol{w} + \boldsymbol{r}(t), \tag{7}$$

where $\boldsymbol{r}(t)$ is some residual term, and $\boldsymbol{w}$ is a vector used to construct the α_n coefficients in Eq. (5):

$$\boldsymbol{w} \,:\, \exp\left(-\boldsymbol{w}^\top \boldsymbol{z}_i\right) = \alpha_i \quad \Rightarrow \quad \hat{\boldsymbol{\theta}} = \sum_{\boldsymbol{z}_j \in \mathbb{S}} \exp\left(-\boldsymbol{w}^\top \boldsymbol{z}_j\right) \boldsymbol{z}_j. \tag{8}$$

(Note that such a $\boldsymbol{w} \in \mathbb{R}^d$ exists because there are fewer than n support vectors and $d > n$.) Our goal is to show that $\boldsymbol{r}(t)$ is bounded for all $t \in [0, \infty)$, and thus the $\log(t)\hat{\boldsymbol{\theta}}$ term dominates $\boldsymbol{\theta}(t)$ as $t \to \infty$.

Writing $\boldsymbol{r}(t)$ and its derivative as:

$$\begin{aligned} \boldsymbol{r}(t) &= \boldsymbol{\theta}(t) - \boldsymbol{w} - \log(t)\hat{\boldsymbol{\theta}} \\ &= \boldsymbol{\theta}(t) - \boldsymbol{w} - \log(t) \sum_{\boldsymbol{z}_j \in \mathbb{S}} \exp\left(\boldsymbol{w}^\top \boldsymbol{z}_j\right) \boldsymbol{z}_j \end{aligned}$$

we can write the differential equation governing $\|\boldsymbol{r}(t)\|_2^2$:

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\|\boldsymbol{r}(t)\|_2^2 &= \dot{\boldsymbol{r}}(t)^\top \boldsymbol{r}(t) \\
&= \left(-\frac{1}{t}\hat{\boldsymbol{\theta}} + \dot{\boldsymbol{\theta}}(t)\right)^\top \boldsymbol{r}(t) \\
&= \left(-\frac{1}{t}\underbrace{\hat{\boldsymbol{\theta}}}_{\sum_{\boldsymbol{z}_i\in\mathbb{S}}\exp(\boldsymbol{w}^\top\boldsymbol{z}_i)\boldsymbol{z}_i} + \sum_{i=1}^n \exp\left(-\boldsymbol{z}_i^\top \underbrace{\boldsymbol{\theta}(t)}_{\log(t)\hat{\boldsymbol{\theta}}+\boldsymbol{w}+\boldsymbol{r}(t)}\right)\boldsymbol{z}_i\right)^\top \boldsymbol{r}(t) \qquad \text{(plugging in Eq. (4))} \\
&= -\sum_{\boldsymbol{z}_i\in\mathbb{S}} \underbrace{\frac{1}{t}\exp(-\boldsymbol{w}^\top\boldsymbol{z}_i)\boldsymbol{z}_i^\top\boldsymbol{r}(t)}_{\beta_i} + \sum_{i=1}^n \underbrace{\exp\left(-\log(t)\boldsymbol{z}_i^\top\hat{\boldsymbol{\theta}} - \boldsymbol{z}_i^\top\boldsymbol{w} - \boldsymbol{z}_i^\top\boldsymbol{r}(t)\right)\boldsymbol{z}_i^\top\boldsymbol{r}(t)}_{\gamma_i} \\
&\qquad\qquad \text{(distributing in } \boldsymbol{r}(t)\text{)} \\
&= -\frac{1}{t}\sum_{\boldsymbol{z}_i\in\mathbb{S}} \underbrace{\exp(-\boldsymbol{w}^\top\boldsymbol{z}_i)\boldsymbol{z}_i^\top\boldsymbol{r}(t)}_{\beta_i} + \sum_{i=1}^n \underbrace{\exp\left(-\log(t)\boldsymbol{z}_i^\top\hat{\boldsymbol{\theta}} - \boldsymbol{z}_i^\top\boldsymbol{w} - \boldsymbol{z}_i^\top\boldsymbol{r}(t)\right)\boldsymbol{z}_i^\top\boldsymbol{r}(t)}_{\gamma_i} \\
&\qquad\qquad \text{(distributing in } \boldsymbol{r}(t)\text{)}
\end{aligned}$$

Unfortunately, this is a complicated differential equation involving $\boldsymbol{r}(t)$ terms. Let's see if we can bound out of the equation.

We'll start by rearrange terms in γ_i. Defining $u_i = \boldsymbol{w}^\top\boldsymbol{z}_i$ and $v_i(t) = \boldsymbol{z}_i^\top\boldsymbol{r}(t)$, we have:

$$\begin{aligned}
\gamma_i &= \exp\left(-\log(t)\boldsymbol{z}_i^\top\hat{\boldsymbol{\theta}}\right)\exp\left(-\boldsymbol{z}_i^\top\boldsymbol{w}\right)\exp\left(-\boldsymbol{z}_i^\top\boldsymbol{r}(t)\right)\boldsymbol{z}_i^\top\boldsymbol{r}(t) \\
&= \frac{1}{t^{\boldsymbol{z}_i^\top\hat{\boldsymbol{\theta}}}}\exp\left(-u_i\right)\exp\left(-v_i(t)\right)v_i(t)
\end{aligned}$$

and using the same terms for β_i, we have that

$$\beta_i = \frac{1}{t}\exp\left(-u_i\right)v_i(t).$$

If $\boldsymbol{z}_i \in \mathbb{S}$, then it is a support vector and so $\boldsymbol{z}_i^\top\boldsymbol{\theta}_i = 1$. Therefore, for these terms we have that

$$\gamma_i - \beta_i = \frac{1}{t}\exp(-u_i)\underbrace{v_i(t)\left(\exp\left(-v_i(t)\right)-1\right)}_{\leq 0}.$$

If $\boldsymbol{z}_i \notin \mathbb{S}$, then $\boldsymbol{z}_i^\top\hat{\boldsymbol{\theta}} > 1$. Recognizing that $\exp(-v_i(t))v_i(t) \leq 1$, we have that

$$\gamma_i \leq \frac{1}{t^c}\exp(-u_i),$$

where $c = \min_{\boldsymbol{z}_i\notin\mathbb{S}}\boldsymbol{z}_i^\top\hat{\boldsymbol{\theta}} > 1$.

All together, we have bounded away any $r(t)$ term out of our differential equation governing $\|\boldsymbol{r}(t)\|_2^2$. Its derivative is upper bounded by the sum of negative terms (yay!) coming from the support vectors as well as the sum of slowly-growing positive terms (yay!) coming from the non-support vectors. More mathematically,

$$\frac{d}{dt}\|\boldsymbol{r}(t)\|_2^2 = O\left(\frac{1}{t^c}\right),$$

which integrates to some finite constant. Thus, $\boldsymbol{r}(t)$ is bounded for all $t \in [0, \infty)$, and so $\boldsymbol{\theta}(t)$ is dominated by the $\hat{\boldsymbol{\theta}}$ term as $t \to \infty$.

---

3) Extensions

This maximum margin convergence has been extended to numerous other scenarios. Some well-known results:

- Soudry et al. [2018] extend this proof idea to non-exponential losses as well as to gradient descent with non-infinitessimal step sizes.
- Nacson et al. [2019] extend this proof idea to non-exponential losses as well as to gradient descent with non-infinitessimal step sizes.
- Lyu and Li [2020] extend this proof idea to neural networks with homogeneous activation functions.
- Ji and Telgarsky [2020] extend this proof idea to neural networks with non-homogeneous activation functions.

Analyzing the generalization of maximum margin classifiers is quite a bit more difficult than for linear regression. It is still a very active area of research, though many researchers have developed margin-based generalization bounds [e.g. Bartlett et al., 2017, Cao et al., 2021].

References

P. L. Bartlett, D. J. Foster, and M. J. Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

Y. Cao, Q. Gu, and M. Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. *Advances in Neural Information Processing Systems*, 34:8407–8418, 2021.

Z. Ji and M. Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.

K. Lyu and J. Li. Gradient descent maximizes the margin of homogeneous neural networks. 2020.

M. S. Nacson, N. Srebro, and D. Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *International Conference on Artificial Intelligence and Statistics*, pages 3051–3059, 2019.

D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.