Topics in Deep Learning Theory (Spring 2025)

Lecture 12: GD converges to max-margin solutions

Instructor: Lei Wu Scribe: Lei Wu

The content of this lecture is a summary of the results in [Soudry et al., 2018, Gunasekar et al., 2018] (linear model) and [Ji and Telgarsky, 2018, Lyu and Li, 2019] (nonlinear model).

Notation. In this note, we use $\|\cdot\|$ to denote the ℓ_2 norm unless otherwise specified. We also use standard big-O notations: $O(\cdot), \Omega(\cdot)$ and $\Theta(\cdot)$.

We have studied how factors like model architectures, optimizers, hyperparameters impact the the implicit bias in training machine learning (ML) models. In this lecture, we further show that the loss function is also crucial for implicit bias.

Setup. Consider the binary classification problem. Let $\{(x_i, y_i)\}_{i=1}^n$ with $x_i \in \mathbb{R}^d$ and $y_i \in \{+1, -1\}$ be the training data. We make the following assumption:

Assumption 0.1.

- *(Linear separability) there exists a $w \in \mathbb{R}^d$ such that $w^\top x_i y_i \geq 0$.*
- $\max_i \|x_i\| \leq B$.

Under the above assumption, there exit many linear models that can perfectly classify all training data. However, for better generalization, the max-margin solution is often preferred.

Definition 0.2. Given a model $f : \mathbb{R}^d \mapsto \mathbb{R}$, the margin of f is defined as $\gamma(f) = \min_{i\in[n]} f(x_i)y_i$.

When $f(x) = f_w(x) := w^\top x$, we let $\gamma(w) = \gamma(f_w)$. The max-margin solution is given by

$$w^* = \operatorname*{argmax}_{\|w\|=1} \gamma(w) \qquad \gamma^* = \gamma(w^*).$$

For a classification problem, we would like to minimize the 0-1 loss. However, due to the non-smoothness and non-convexity of 0-1 loss, in practice, we turn to minimize some surrogate losses such as logistic loss, hinge loss, etc. Here, we consider the exponential loss

$$L(w) = \frac{1}{n}\sum_{i=1}^n e^{-y_i x_i^\top w}. \tag{1}$$

The gradient descent (GD) for minimizing $L(\cdot)$ is given by

$$w_{t+1} = w_t - \eta \nabla L(w_t).$$

We will show that w_t converges to the max-margin solution w^*, regardless the initialization.

1 Landscape Analysis

Before we delve into the dynamics analysis, we need some understanding of the landscape of $L(\cdot)$, which is very different from the landscape of regression.

Lemma 1.1 (Basic Properties)**.**

- *The minimizers of $L(\cdot)$ are unbounded.*
- *For any $w \in \mathbb{R}^d$ that perfectly classifies all data, we have $L(\lambda w) \to 0$ as $\lambda \to +\infty$.*

The proof is omited.

Lemma 1.2 (Gradient and Hessian)**.** *$v^\top \nabla L^2(w) v \le BL(w)\|v\|_2^2$ and $\gamma^* L(w) \le \|\nabla L(w)\|_2 \le BL(w)$.*

Proof. The gradient and Hessian is given by

$$\nabla L(w) = -\frac{1}{n}\sum_{i=1}^n e^{-y_i x_i^\top w} x_i y_i, \quad \nabla^2 L(w) = \frac{1}{n}\sum_{i=1}^n e^{-y_i x_i^\top w} x_i x_i^\top. \tag{2}$$

Note that

$$\|\nabla L(w)\| \ge \nabla L(w)^\top (-w^*) = \frac{1}{n}\sum_{i=1}^n e^{-y_i x_i^\top w} y_i x_i^\top w^* \ge \gamma^* L(w)$$

and

$$\|\nabla L(w)\| \le \frac{1}{n}\sum_{i=1}^n e^{-y_i x_i^\top w}\|x_i\| \le BL(w).$$

Additionally,

$$v^\top \nabla^2 L(w) v = \frac{1}{n}\sum_{i=1}^n e^{-w^\top x_i y_i}(v^\top x_i)^2 \le B\|v\|^2 L(w).$$

□

Remark 1.3. *This implies that 1) the sharpness (the largest eigenvalue of Hessian) decreases to zero as the loss converges to zero; 2) the gradient is well-controlled by the loss.*

2 Convergence Analysis

We make the following assumption on the learning rate:

Assumption 2.1. *Assume η to be sufficiently small such that in each step $L(w_{t+1}) \le L(w_t)$.*

Nearly all analysis of GD starts from the *descent inequality* :

$$\begin{aligned} L(w_{t+1}) &= L(w_t) - \eta\|\nabla L(w_t)\|^2 + \frac{\eta^2}{2}\inf_{\beta\in[0,1]} \nabla L(w_t)^\top \nabla^2 L(w_t - \beta\eta\nabla L(w_t))\nabla L(w_t) \\ &\le L(w_t) - \eta\|\nabla L(w_t)\|^2 + \frac{\eta^2 B}{2}\|\nabla L(w_t)\|^2 \inf_{\beta\in[0,1]} L(w_t - \beta\eta\nabla L(w_t)) \\ &\le L(w_t) - \eta\|\nabla L(w_t)\|^2 + \frac{\eta^2 B}{2}\|\nabla L(w_t)\|^2 L(w_t), \end{aligned} \tag{3}$$

where the last step use the assumption of η and the convexity of $L(\cdot)$. For notation simplicity, let $L_t = L(w_t)$ and $g_t = \|\nabla L(w_t)\|$.

2.1 Loss Convergence

Let us first examine the convergence of the loss. Note that although $L(\cdot)$ is convex, we cannot apply the standard results of convex optimization to conclude that $L(w_t) = O(1/t)$. Standard analysis of convex optimization assumes that the *minima are located in a compact domain*, whereas in our case, the minima are at infinity. In fact, the convergence of convex optimization can be arbitrarily slow if the minima are at infinity.

Lemma 2.2. $L(w_t) = \Theta(1/t)$.

Proof. By the descent inequality, when η is sufficiently small, we have

$$L(w_{t+1}) \leq L(w_t) - \frac{\eta}{2}\|\nabla L(w_t)\|^2 \leq L(w_t) - \frac{\gamma^*\eta}{2}L(w_t)^2.$$

This yield,

$$\frac{1}{L_{t+1}} \geq \frac{1}{L_t} + \frac{r^*\eta}{2 - r^*\eta L_t} \geq \frac{1}{L_t} + \frac{\gamma^*\eta}{2} \Rightarrow \frac{1}{L_t} \geq \frac{1}{L_0} + \frac{\gamma^*\eta t}{2} \tag{4}$$

which gives

$$L_t = O\left(\frac{1}{t}\right).$$

On the other hand, using $e^z \geq 1 + z$, we have

$$\begin{aligned} L(w_{t+1}) &= \frac{1}{n}\sum_{i=1}^{n} e^{-(w_t - \eta\nabla L(w_t))^\top x_i y_i} = \frac{1}{n}\sum_{i=1}^{n} e^{-w_t^\top x_i y_i} e^{\eta\nabla L(w_t)^\top x_i y_i} \\ &\geq \frac{1}{n}\sum_{i=1}^{n} e^{-w_t^\top x_i y_i}(1 + \eta\nabla L(w_t)^\top x_i y_i) \\ &= L(w_t) - \eta\|\nabla L(w_t)\|^2 \geq L(w_t) - \eta B^2 L(w_t), \end{aligned}$$

where the last step uses Lemma (1.2). Analogous to (4) , we can obtain

$$L(w_t) = \Omega\left(\frac{1}{t}\right).$$

□

Lemma 2.3. $\sum_{s=0}^{t}\|\nabla L(w_s)\|^2 \leq 2L(w_0)/\eta$ *and* $\sum_{s=0}^{t}\|\nabla L(w_s)\| = \Omega(\log t)$.

Proof. Due to $L(w_{t+1}) \leq L(w_t) - \frac{\eta}{2}\|\nabla L(w_t)\|^2$, we have

$$\sum_{s=0}^{t}\|\nabla L(w_s)\|^2 \leq \frac{2}{\eta}\left(L(w_0) - L(w_{t+1})\right) \leq 2L(w_0)/\eta.$$

Moreover,

$$\sum_{s=0}^{t}\|\nabla L(w_s)\| \geq \gamma^*\sum_{s=0}^{t} L(w_s) \geq C\gamma^*\sum_{s=0}^{t}\frac{1}{s} = C\gamma^*\log t.$$

□

2.2 The Margin Dynamics

Theorem 2.4. *The margin of GD solution satisfies* $|\gamma\left(\frac{w_t}{\|w_t\|}\right) - \gamma^*| = O(1/\log t)$.

Proof. By the descent inequality (3), we have

$$\begin{aligned} L_{t+1} &\le L_t\left(1 + \frac{B\eta^2 g_t^2}{2} - \frac{\eta g_t^2}{L_t}\right) \\ &\le L_t\left(1 + \frac{B\eta^2 g_t^2}{2} - \eta\gamma^* g_t\right) \\ &\le L_t \exp\left(\frac{B\eta^2 g_t^2}{2} - \eta\gamma^* g_t\right) \\ &\le L_0 \exp\left(\frac{B\eta^2}{2}\sum_{s=0}^{t} g_s^2 - \eta\gamma^*\sum_{s=0}^{t} g_s\right), \end{aligned}$$

where the second step uses Lemma (1.2). Noting that

$$L(w) = \frac{1}{n}\sum_{i=1}^{n} e^{-w^\top x_i y_i} \ge \frac{1}{n}\max_i e^{-w^\top x_i y_i} = \frac{1}{n}\exp(-\min_i w^\top x_i y_i), \tag{5}$$

we have

$$\min_i w_{t+1}^\top x_i y_i \ge -\log L(w_{t+1}) - \log n = \eta\gamma^*\sum_{s=0}^{t} g_s - \frac{B\eta^2}{2}\sum_{s=0}^{t} g_s^2 - \log n.$$

Additionally,

$$\|w_t + 1\| \le \|w_0\| + \eta\sum_{s=0}^{t}\|\nabla L(w_s)\| = \|w_0\| + \eta\sum_{s=0}^{t} r_s.$$

Thus the margin satisfies

$$\begin{aligned} \gamma\left(\frac{w_{t+1}}{\|w_{t+1}\|}\right) = \frac{\min_i w_{t+1}^\top x_i y_i}{\|w_{t+1}\|} &\ge \frac{\eta\gamma^*\sum_{s=0}^{t} r_s - \frac{B\eta^2}{2}\sum_{s=0}^{t} r_s^2 - \log n}{\|w_0\| + \eta\sum_{s=0}^{t} r_s} \\ &\ge \frac{C_1\eta\gamma^*\log t - 0.5B\eta L(w_0) - \log n}{\|w_0\| + \eta C_1 \log t} = \gamma^* - O\left(\frac{1}{\log t}\right), \end{aligned}$$

where we use Lemma (2.3).

□

Remark 2.5. *We can see that GD converges to the max-margin solution but the* ***margin's convergence is exponentially slow****.*

3 Extensions

3.1 Steepest Descent

We have studied GD, which can be viewed as steepest descent (GD) with respect to the ℓ_2 metric. In this subsection, we consider general steepest descent. Throughout this subsection, we

temperately use $\|\cdot\|$ to denote a generic norm in $\mathbb{R}^d$ and its dual norm is given by

$$\|u\|_* = \sup_{\|v\|\le 1} u^\top v.$$

SD with respect to the $\|\cdot\|$ norm is given by

$$w_{t+1} = \operatorname*{argmin}_{w\in\mathbb{R}^d} \left(L(w_t) - \nabla L(w_t)^\top (w - w_t) + \frac{\eta}{2}\|w - w_t\|^2 \right), \tag{6}$$

which can be further explicitly written as

$$w_{t+1} = w_t - \eta \|\nabla L(w_t)\|_* \delta_t \;\text{ with }\; \delta_t = \operatorname*{argmax}_{\|\delta\|\le 1} \delta^\top \nabla L(w_t).$$

Examples.

- When $\|\cdot\|$ is the ℓ_2 norm, SD recovers the standard GD.
- When $\|\cdot\|$ is the ℓ_1 norm, SD becomes

  $$w_{t+1} = w_t - \eta \partial_{j_t} L(w_t) e_{j_t}, \text{ with } j_t = \operatorname*{argmax}_j |\partial_i L(w_t)|,$$

  where $\nabla L(w) = (\partial_1 L(w), \ldots, \partial_d L(w))^\top \in \mathbb{R}^d$ and e_j is the one-hot vector. This is exactly the **`greedy`** `coordinate descent` as in each step, it select the coordinate whose gradient is largest to update.
- When $\|\cdot\|$ is ℓ_∞ norm, SD becomes `signGD`:

  $$w_{t+1} = w_t - \eta \|\nabla L(w_t)\|_1 \operatorname{sign}(\nabla L(w_t))$$

The implicit bias of SD for minimizing $L(\cdot)$ converges the max-margin solution with respect to the $\|\cdot\|$ solution:

Theorem 3.1. *For any norm $\|\cdot\|$, consider to minimize $L(\cdot)$ using SD with respect to $\|\cdot\|$. Suppose η to be sufficiently small. Then, SD converges to the following max-margin solution:*

$$\lim_{t\to\infty} \gamma\left(\frac{w_t}{\|w_t\|}\right) = \gamma^*, \quad \textit{with}\;\; \gamma^* = \operatorname*{argmax}_{\|w\|\le 1} \gamma(w).$$

The proof is analogous to that of GD and is left to homework.

Remark 3.2. *By the above theorem, the greedy coordinate descent will converge to max-margin solutions with ℓ_1 sparsity.*

3.2 Nonlinear Homogeneous Models

The above result can be also extended to nonlinear homogeneous models. Let $f(\cdot;\theta):\mathbb{R}^d \mapsto \mathbb{R}$ be a homogeneous model, i.e., $f(x;\lambda\theta) = \lambda^\alpha f(x;\theta)$ for any $\lambda \in \mathbb{R}_+$ and some positive constant α. Let

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n} e^{-y_i f(x_i;\theta)}.$$

Then, one can show that GD converges to the KKT point of the following max-margin problem

$$\min \ \|\theta\|_2^2 \tag{7}$$

$$s.t.\ y_i f(x_i;\theta) \geq 1. \tag{8}$$

Note that we can only show the convergence to KKT points. For general non-convex problem, we are unable to show GD converge to the max-margin solutions. For the proof and more details, we refer interested readers to [Lyu and Li, 2019].

References

[Gunasekar et al., 2018] Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. (2018). Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR.

[Ji and Telgarsky, 2018] Ji, Z. and Telgarsky, M. (2018). Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*.

[Lyu and Li, 2019] Lyu, K. and Li, J. (2019). Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*.

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