CS229M/STATS214: Machine Learning Theory

Lecturer: Tengyu Ma — Lecture # 11
Scribe: Andrew Wang — Feb 22nd, 2021

11.1 Review and overview

Last lecture, we discussed the role of optimization in understanding deep learning model generalization. In particular, gradient descent cannot always find the global minimum for a non-convex loss function. However, there are some special problems where (i) gradient descent can find a local minimum, and (ii) all local minima are also global minima. For these problems, gradient descent is able to find a global minimum. Based on these observations, we began carrying out a two step plan:

1. Characterize the set of functions for which gradient descent can find a global minimum, then
2. Identify machine learning problems that belong to this set.

We proved that principal component analysis (PCA) belongs in this set of problems (all local minima are global minima), and began proving a similar approximate claim for matrix completion. In this lecture, we finish this proof and begin discussing a new topic, the Neural Tangent Kernel (NTK), which will give further insights into local minima near a given initialization.

11.2 Matrix completion [GLM16]

Recall problem of rank-1 matrix completion problem. Let $M = zz^T$ be a rank-1 symmetric, positive semi-definite matrix for some $z \in \mathbb{R}^d$ such that $\|z\|_2 = 1$ and $\|z\|_\infty \le \mu/\sqrt{d}$. We observe random entries of M and must recover the remaining entries.

Definition 11.1. Suppose $M \in \mathbb{R}^{d\times d}$ and $\Omega \subseteq [d] \times [d]$. We define $P_\Omega(M)$ to be the matrix obtained by zeroing out every entry outside Ω.

Definition 11.2 (Matrix completion)**.** Suppose $M \in \mathbb{R}^{d\times d}$ and every entry of M is included in Ω with probability p. The *matrix completion task* is to recover M (with respect to some loss function) given the observation $P_\Omega(M)$.

In order to solve the matrix completion problem, one can search for x that locally minimizes the following objective:

$$f(x) = \frac{1}{2}\|P_\Omega(M - xx^T)\|_F^2 = \frac{1}{2}\sum_{(i,j)\in\Omega}(M_{ij} - x_i x_j)^2. \tag{11.1}$$

The goal is to prove that local minima of this objective function are close to a global minimum:

Theorem 11.3. *Assume $p = \dfrac{poly(\mu, \log d)}{d\epsilon^2}$. Then with high probability, all local minima of f are $O(\sqrt{\epsilon})$-close to $+z$ or $-z$ (the global minima of f).*

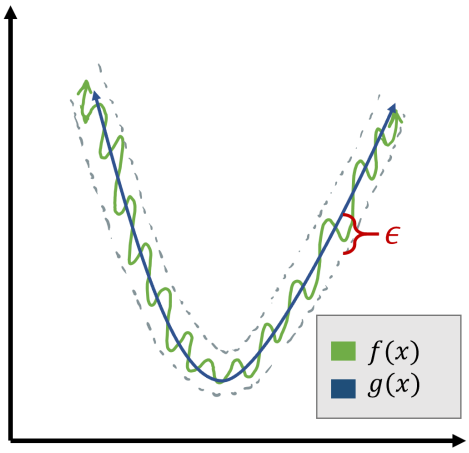

Figure 11.1: Even if $f(x)$ and $g(x)$ are no more than ϵ apart at any given x, the local minima of f may look dramatically different from the local minima of g.

Before presenting the proof, we make some observations that will guide the proof strategy.

Remark 11.4. $f(x)$ can be viewed as a sampled version of the PCA loss function $g(x) = \frac{1}{2}\|M - xx^T\|_F^2 = \frac{1}{2}\sum_{(i,j)\in[d]\times[d]}(M_{ij} - x_i x_j)^2$, in which we only observe a subset of the matrix entries. Thus, we would like to claim that $f(x) \approx g(x)$. However, matching the values of f and g is not sufficient to prove the theorem: even a small margin of error between f and g could lead to creation of many spurious local minima (see Figure 11.1 for an illustration). In order to ensure that the local minima of f look like the local minima of g, we will need further conditions like $\nabla f(x) \approx \nabla g(x)$ and $\nabla^2 f(x) \approx \nabla^2 g(x)$.

Remark 11.5. Key idea: concentration for scalars is easy. We can approximate a sum of scalars via a sample:

$$\sum_{(i,j)\in\Omega} T_{ij} \approx p \sum_{(i,j)\in[d]\times[d]} T_{ij}, \tag{11.2}$$

where we use $\approx$ to mean that

$$\left| \sum_{(i,j)\in\Omega} T_{ij} - p \sum_{(i,j)\in[d]\times[d]} T_{ij} \right| < \epsilon \tag{11.3}$$

with high probability. This suggests the strategy of casting the estimation of our desired quantities in the form of estimating a scalar sum via a sample. In particular, we note that for any matrices A and B,

$$\langle A, P_\Omega(B)\rangle = \sum_{(i,j)\in\Omega} A_{ij}B_{ij} \approx p\langle A, B\rangle. \tag{11.4}$$

To make use of this observation to understand the quantities of interest ($\nabla f(x)$ and $\nabla^2 f(x)$), we compute the bilinear and quadratic forms for $\nabla f(x)$ and $\nabla^2 f(x)$ respectively:

$$\langle v, \nabla f(x)\rangle = \langle v, P_\Omega(M - xx^T)x\rangle = \langle vx^T, P_\Omega(M - xx^T)\rangle, \tag{11.5}$$

where we have used the fact that $\langle A, BC\rangle = \langle AC^T, B\rangle$. Also note that vx^T is a rank-1 matrix and $M - xx^T$ is a rank-2 matrix.

$$\langle v, \nabla^2 f(x) v\rangle = \|P_\Omega(vx^T + xv^T)\|_F^2 - 2\langle P_\Omega(M - xx^T), vv^T\rangle \tag{11.6}$$

$$= \langle P_\Omega(vx^T + xv^T), vx^T + xv^T\rangle - 2\langle P_\Omega(M - xx^T), vv^T\rangle, \tag{11.7}$$

where we have used the fact that $\|P_\Omega(A)\|_F^2 = \langle P_\Omega(A), P_\Omega(A)\rangle = \langle P(\Omega(A), A\rangle$.

The key lemma that applies the scalar concentration to these matrix quantities is as follows:

Lemma 11.6. *Let $\epsilon > 0$, $p = \dfrac{poly(\mu, \log d)}{d\epsilon^2}$. Given that $A = uu^T, B = vv^T$ for some u, v satisfying $\|u\|_2 \le 1$, $\|v\|_2 \le 1$, $\|u\|_\infty \le \mu/\sqrt{d}$, $\|v\|_\infty \le \mu/\sqrt{d}$, we have $|\langle P_\Omega(A), B\rangle/p - \langle A, B\rangle| \le \epsilon$ w.h.p.*

If we can show that g has no bad local minima via a proof that only uses g via terms of the form $\langle v, \nabla g(x)\rangle$ and $\langle v, \nabla^2 g(x) v\rangle$, then by Lemma 11.6 this proof will automatically generalize to f by concentration.

Next, we prove some facts about g and show the analogous proofs for f that we will use in the proof of Theorem 11.3.

Lemma 11.7 (Connecting inner product and norm for g)**.** *If x satisfies $\nabla g(x) = 0$, then $\langle x, z\rangle^2 = \|x\|_2^4$.*

Proof.

$$\nabla g(x) = 0 \implies \langle x, \nabla g(x)\rangle = 0 \tag{11.8}$$

$$\implies \langle x, (zz^T - xx^T)x\rangle = 0 \qquad (\because \nabla g(x) = (M - xx^T)x) \tag{11.9}$$

$$\implies \langle x, z\rangle^2 = \|x\|_2^4. \tag{11.10}$$

□

Lemma 11.8 (Connecting inner product and norm for f)**.** *Suppose $\|x\|_\infty \le 2\mu/\sqrt{d}$. If x satisfies $\nabla f(x) = 0$, then $\langle x, z\rangle^2 \ge \|x\|_2^4 - \epsilon$ with high probability.*

Proof.

$$\nabla f(x) = 0 \implies \langle x, \nabla f(x)\rangle = 0 \tag{11.11}$$

$$\implies \langle x, \nabla g(x)\rangle \approx \langle x, \nabla f(x)\rangle/p \pm \epsilon \qquad \text{(by Lemma 11.6)} \tag{11.12}$$

$$\implies |\langle x, (zz^T - xx^T)x\rangle| \le \epsilon \qquad \text{w.h.p.} \tag{11.13}$$

$$\implies \langle x, z\rangle^2 \ge \|x\|_2^4 - \epsilon \qquad \text{w.h.p.} \tag{11.14}$$

□

Lemma 11.9 (Bound norm for g)**.** *If $\nabla^2 g(x) \succeq 0$, then $\|x\|_2^2 \ge 1/3$.*

Proof.

$$\nabla^2 g(x) \succeq 0 \implies \langle z, \nabla^2 g(x) z\rangle \ge 0 \tag{11.15}$$

$$\implies \|zx^T + xz^T\|_F^2 - 2z^T(zz^T - xx^T)z \ge 0 \tag{11.16}$$

$$\implies 1 \le \|x\|_2^2 + 2\langle x, z\rangle^2 \le \|x\|_2^2 + 2\|x\|_2^2 = 3\|x\|_2^2 \qquad \text{(by Cauchy-Schwarz)} \tag{11.17}$$

$$\implies \|x\|_2^2 \ge 1/3. \tag{11.18}$$

□

Lemma 11.10 (Bound norm for f)**.** *Suppose $\|x\|_\infty \le \mu/\sqrt{d}$. If $\nabla^2 f(x) \succeq 0$, then $\|x\|_2^2 \ge 1/3 - \epsilon/3$ with high probability.*

Proof.

$$\begin{aligned}
\nabla^2 f(x) \succeq 0 &\implies \langle z, \nabla^2 f(x) z\rangle \ge 0 && (11.19)\\
&\implies \langle z, \nabla^2 g(x) z\rangle \ge -\epsilon \qquad \text{w.h.p. (by Lemma 11.6)} && (11.20)\\
&\implies 3\|x\|_2^2 \ge 1 - \epsilon \qquad \text{w.h.p.} && (11.21)\\
&\implies \|x\|_2^2 \ge 1/3 - \epsilon/3 \qquad \text{w.h.p.} && (11.22)
\end{aligned}$$

□

Lemma 11.11 (g has no bad local minimum)**.** *All local minima of g are global minima.*

Proof.

$$\begin{aligned}
\nabla g(x) = 0 &\implies \langle z, \nabla g(x)\rangle = 0 && (11.23)\\
&\implies \langle z, (zz^T - xx^T)x\rangle = 0 && (11.24)\\
&\implies \langle x, z\rangle(1 - \|x\|_2^2) = 0. && (11.25)
\end{aligned}$$

Since $|\langle x, z\rangle| \ge 1/3 \ne 0$ (by Lemma 11.9), we must have $\|x\|_2^2 = 1$. But then Lemma 11.7 implies $\langle x, z\rangle^2 = \|x\|_2^4 = 1$, so $x = \pm z$ by Cauchy-Schwarz. □

We now prove Theorem 11.3, restated for convenience:

Theorem 11.12 (f has no bad local minimum)**.** *Assume $p = \dfrac{poly(\mu, \log d)}{d\epsilon^2}$. Then with high probability, all local minima of f are $O(\sqrt{\epsilon})$-close to $+z$ or $-z$.*

Proof. Observe that $\|x - z\|_2^2 = \|x\|_2^2 + \|z\|_2^2 - 2\langle x, z\rangle \le \|x\|_2^2 + 1 - 2\langle x, z\rangle$. Our goal is to show that this quantity is small with high probability, hence we need to bound $\|x\|_2^2$ and $\langle x, z\rangle$ w.h.p. Note that the following bounds in this proof are understood to hold w.h.p.

Let x be such that $\nabla f(x) = 0$. For $\epsilon \le 1/16$,

$$\begin{aligned}
\langle x, z\rangle^2 &\ge \|x\|_2^4 - \epsilon && \text{(by Lemma 11.8)} && (11.26)\\
&\ge (1/3 - \epsilon/3)^2 - \epsilon && \text{(by Lemma 11.10)} && (11.27)\\
&\ge 1/32. && && (11.28)
\end{aligned}$$

With this, we can get a bound on $\|x\|_2^2$:

$$\begin{aligned}
\nabla f(x) = 0 &\implies \langle x, \nabla f(x)\rangle = 0 && && (11.29)\\
&\implies |\langle z, \nabla g(x)\rangle| \le \epsilon && \text{(by Lemma 11.6)} && (11.30)\\
&\implies |\langle x, z\rangle| \cdot |1 - \|x\|_2^2| \le \epsilon && \text{(by dfn of } g) && (11.31)\\
&\implies |1 - \|x\|_2^2| \le 32\epsilon = O(\epsilon) && \text{(by (11.28))} && (11.32)\\
&\implies \|x\|_2^2 = 1 \pm O(\epsilon). && && (11.33)
\end{aligned}$$

Next, we bound $\langle x, z\rangle$:

$$\begin{aligned}
\langle x, z\rangle^2 &\geq \|x\|_2^4 - \epsilon && \text{(by Lemma 11.8)} && (11.34)\\
&\geq (1 - O(\epsilon))^2 - \epsilon && \text{(by (11.33))} && (11.35)\\
&= 1 - O(\epsilon). && && (11.36)
\end{aligned}$$

Finally, we put these quantities together to bound $\|x - z\|_2^2$. We have two cases:
Case 1: $\langle x, z\rangle \geq 1 - O(\epsilon)$. Then

$$\begin{aligned}
\|x - z\|_2^2 &= \|x\|_2^2 + \|z\|_2^2 - 2\langle x, z\rangle && (11.37)\\
&\leq \|x\|_2^2 + 1 - 2\langle x, z\rangle && (11.38)\\
&\leq 1 + O(\epsilon) + 1 - 2(1 - O(\epsilon)) && (11.39)\\
&\leq O(\epsilon). && (11.40)
\end{aligned}$$

Hence we conclude x is $O(\sqrt{\epsilon})$-close to z.
Case 2: $\langle x, z\rangle \leq -(1 - O(\epsilon))$. Then by an analogous argument, x is $O(\sqrt{\epsilon})$-close to $-z$. □

We have shown above that matrix completion of a rank-1 matrix has no spurious local minima. This proof strategy can be extended to handle higher-rank matrices and noisy matrices [GLM16]. The proof also demonstrates a generally useful proof strategy: often, reducing a hard problem to an easy problem results in solutions that do not give much insight into the original problem, because the proof techniques do not generalize. It can often be fruitful to seek a proof in the simplified problem that makes use of a restricted set of tools that could generalize to the harder problem. Here we limited ourselves to only using $\langle v, \nabla g(x)\rangle$ and $\langle v, \nabla^2 g(x) v\rangle$ in the easy case; these quantities could then be easily converted to analogous quantities in f via the concentration lemma (Lemma 11.6).

11.3 Other problems where all local minima are global minima

We have now demonstrated that two classes of machine learning problems, rank-1 PCA and rank-1 matrix completion, have no spurious local minima and are thus amenable to being solvable by gradient descent methods. We now outline some major classes of problems for which it is known that there are no spurious local minima.

- Principal component analysis (covered in previous lecture).
- Matrix completion (and other matrix factorization problems). On a related note, it has also been shown that linearized neural networks of the form $y = W_1 W_2 x$, where W_1 and W_2 are optimized separately, have no spurious local minima [BH89]. It should be noted that linearized neural networks are not very useful in practice since the advantage of optimizing W_1 and W_2 separately versus optimizing a single $W = W_1 W_2$ is not clear.
- Tensor decomposition. The problem is as follows:

$$\text{maximize} \quad \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d}\sum_{l=1}^{d} T_{ijkl} x_i x_j x_k x_l \quad \text{such that} \quad \|x\|_2 = 1. \tag{11.41}$$

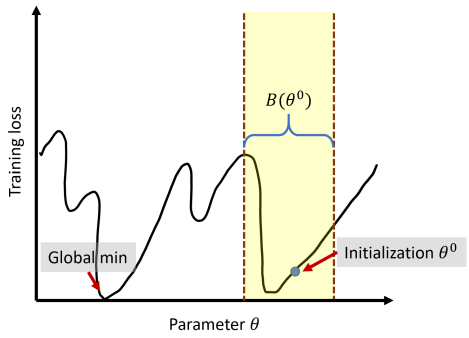

Figure 11.2: The training loss landscape around a given parameter initialization θ^0. We hope that the neighborhood around θ^0 contains a local minimum that is close to the global minimum.

Additionally, constraints are imposed on the tensor T to make the problem tractable. For example, one condition is that T must be a low-rank tensor with orthonormal components [GHJY15].

In general, the loss landscapes of neural networks (with nonlinearities) is currently not as well understood, though we now introduce the *neural tangent kernel* which allows us to make some characterizations of the loss near a given neural network initialization.

11.4 Neural tangent kernel (NTK) approach

We now move to a new topic of the *neural tangent kernel (NTK)*, which will give further insight into the loss landscape in the neighborhood of a given parameter initialization.

The key insight of the NTK approach is that if we take an appropriate random parameter initialization θ^0 (which we will choose later), we can identify a special neighborhood of θ^0, denoted $B(\theta^0)$, where "everything is nice". That is, the function is convex in $B(\theta^0)$, there is a global minimum in the $B(\theta^0)$, and the algorithm starting at θ^0 will converge to that global minimum. (See Figure 11.2 for an illustration.)

Take a random initialization $\theta = \theta^0$ and Taylor expand the loss around θ^0 w.r.t. θ:

$$f_\theta(x) = \underbrace{f_{\theta^0}(x) + \langle \nabla_\theta f_{\theta^0}(x), \theta - \theta^0 \rangle}_{g_\theta(x)} + O((\theta - \theta^0)^2). \tag{11.42}$$

In other words, we take the tangent plane to f_θ at x (a linear approximation). We observe that g_θ is an affine function of θ. Additionally, defining $\Delta\theta = \theta - \theta^0$, we see that the first term does not depend on $\Delta\theta$ while the second term $\langle \nabla_\theta f_{\theta^0}(x), \theta - \theta^0 \rangle$ is linear in $\Delta\theta$. (For convenience, we will sometimes choose to design θ^0 such that $f_{\theta^0}(x) = 0$ so that g_θ is linear in θ^0. However, the difference is not very important since $f_{\theta^0}(x)$ can simply be subsumed in the training labels y via $y' = y - f_{\theta^0}(x)$.)

Now we have that $y \approx \nabla_\theta f_{\theta^0}(x)^T \Delta\theta$. We can view $\phi(x) \triangleq \nabla_\theta f_{\theta^0}(x)$ as a feature map, i.e. we can rewrite the expression as $\phi(x)^T \Delta\theta$ where $\phi(x)$ is fixed (only depends on θ^0 and the architecture). This observation motivates the definition of the *neural tangent kernel*:

Definition 11.13 (Neural tangent kernel)**.** The *neural tangent kernel* K is defined as the function

$$K(x, x') = \langle \phi(x), \phi(x') \rangle = \langle \nabla f_{\theta^0}(x), \nabla f_{\theta^0}(x') \rangle. \tag{11.43}$$

In the next lecture, we will examine the optimization behavior for a neural network using the NTK approach. Suppose we fit $g_\theta(x)$ to y, i.e. we minimize the loss

$$\text{Loss} = \ell(\phi(x)^T \Delta\theta, y), \tag{11.44}$$

where $\phi(x)^T \Delta\theta$ is linear and the loss as a whole is convex. We will show that for a sufficiently wide neural network with proper initialization θ^0, optimizing $f_\theta(x)$ starting from θ^0 never leaves the neighborhood of θ^0, effectively behaving the same as optimizing $g_\theta(x)$. We will answer two questions:

1. Why does there exist a small neighborhood $B(\theta^0)$ such that there exists a global minimum in $B(\theta^0)$? (More surprising, involves proper design of θ^0.)
2. Does gradient descent on the original loss with respect to $f_\theta(x)$ stay in the neighborhood $B(\theta^0)$? (Less surprising, more technical.)

Bibliography

[BH89] Pierre Baldi and Kurt Hornik, *Neural networks and principal component analysis: Learning from examples without local minima*, Neural networks **2** (1989), no. 1, 53–58.

[GHJY15] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan, *Escaping from saddle points — online stochastic gradient for tensor decomposition*, Proceedings of The 28th Conference on Learning Theory (Paris, France) (Peter Grünwald, Elad Hazan, and Satyen Kale, eds.), Proceedings of Machine Learning Research, vol. 40, PMLR, 03–06 Jul 2015, pp. 797–842.

[GLM16] Rong Ge, Jason D Lee, and Tengyu Ma, *Matrix completion has no spurious local minimum*, Advances in Neural Information Processing Systems (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.