CS229M/STATS214: Machine Learning Theory

Lecturer: Tengyu Ma
Scribe: Roshni Sahoo and Sarah Wu

Lecture # 14
Mar 3rd, 2021

14.1 Review and overview

In the last lecture, we introduced the concept of *algorithmic regularization* (or *implicit regularization*) and demonstrated it for the problem of overparametrized linear regression. We also began applying it to a non-convex case where we assume that the ground-truth model is sparse. Consider the following quadratic parametrization:

$$y = \langle \beta^\star \odot \beta^\star, x \rangle, \tag{14.1}$$

where $\beta^\star$ is an r-sparse vector. We write $\beta^\star = \mathbf{1}_S$ where S is the set of indices on which $\beta^\star$ takes nonzero value. Since $\beta^\star$ is r-sparse, we have that $|S| \le r$. Our goal is to show that with $\Omega(r^2)$ samples, gradient descent on $\widehat{L}(\beta)$ starting from small initialization can converge to $\beta^\star$. This is is expressed more formally by the following theorem.

Theorem 14.1. *Let c be a sufficiently large universal constant. Suppose $n \ge cr^2 \log^2(d)$ and $\alpha \le 1/d^c$, then when $\dfrac{\log(d/\alpha)}{\eta} \lesssim T \lesssim \dfrac{1}{\eta\sqrt{d\alpha}}$, we have*

$$\left\|\beta^T \odot \beta^T - \beta^\star \odot \beta^\star\right\|_2^2 \le O\big(\alpha\sqrt{d}\big). \tag{14.2}$$

The main intuition for proving Theorem 14.1 is to establish strong uniform convergence on $S_r = \{\beta : \|\beta\|_0 \le O(r)\}$, i.e. gradient descent on the population loss $\widehat{L}$ is sufficiently "similar" to gradient descent on the empirical loss $\widehat{L}(\beta)$.

As a warm-up, we began to prove the following result on convergence for gradient descent on $L(\beta)$:

Theorem 14.2. *For sufficiently small η, gradient descent on $L(\beta)$ converges to $\beta^\star$ in $\Theta\Big(\frac{\log(1/(\epsilon\alpha))}{\eta}\Big)$ iterations with ϵ-error in ℓ_2-distance.*

In this lecture, we will wrap up the proof of Theorem 14.2 and then prove Theorem 14.1 in the case of $r = 1$.

14.2 Warm-up: Gradient descent on population loss

The analysis of Theorem 14.2 can be broken into two cases depending on where coordinate i of $\beta^\star$ is nonzero or not, i.e. $i \in S$ or $i \notin S$. Refer to the previous lecture for the proof in the case of $i \in S$. Here, we will prove the following lemma for $i \notin S$, then explain informally how the lemma implies Theorem 14.2.

Lemma 14.3. *For all $i \notin S$, when $t \le 1/(10\eta\alpha^2)$ we have that $\beta_i^t \le 2\alpha$.*

Proof. For a coordinate $i \notin S$, the gradient descent update for this problem becomes

$$\begin{aligned}
\beta_i^{t+1} &= \left[\beta^t - \eta(\beta^t \odot \beta^t - \beta^\star \odot \beta^\star) \odot \beta^t\right]_i && (14.3)\\
&= \beta_i^t - \eta(\beta_i^t \cdot \beta_i^t) \cdot \beta_i^t && \text{(since } \beta_i^\star = 0 \ \forall i \notin S) \quad (14.4)\\
&= \beta_i^t - \eta(\beta_i^t)^3. && (14.5)
\end{aligned}$$

Since our initialization β^0 was small, the update to these coordinates will be even smaller because $(\beta_i^t)^3$ is small. We can prove the desired claim using strong induction. Suppose $\beta_i^s \le 2\alpha$ for all $s \le t$ and $i \notin S$, and that $t+1 \le 1/(10\eta\alpha^2)$. Then, for all $s \le t$,

$$\begin{aligned}
\beta_i^{s+1} &= (1 - \eta(\beta_i^s)^2)\beta_i^s && (14.6)\\
&\le (1 + \eta(\beta_i^s)^2)\beta_i^s && (14.7)\\
&\le (1 + 4\eta\alpha^2)\beta_i^s. && \text{(since } \beta_i^s \le 2\alpha) \quad (14.8)
\end{aligned}$$

With strong induction, we can repeatedly apply this gradient update starting from $t = 0$ to obtain

$$\begin{aligned}
\beta_i^{t+1} &\le \beta_0 \cdot (1 + 4\eta\alpha^2)^t && (14.9)\\
&\le \beta_0 (1 + 4\eta\alpha^2)^{\frac{1}{10\eta\alpha^2}} && (14.10)\\
&\le \beta_0 \exp\left(\frac{4\eta\alpha^2}{10\eta\alpha^2}\right) && (14.11)\\
&= \beta_0 \cdot e^{2/5} && (14.12)\\
&\le 2\alpha, && (14.13)
\end{aligned}$$

which completes the inductive proof of the claim.

□

Informally, what we observe here is that when $i \notin S$, β_i stays small and will take many iterations before it even gets to 2α, which is close to 0 since α is chosen to be small. Last lecture, we saw that when $i \in S$, β_i converges to 1. Putting these two together, we can be sure that gradient descent on $L(\beta)$ converges to $\beta^\star$.

14.3 Main result: Gradient descent on empirical loss

The main part of this lecture focuses on analyzing gradient descent on the empirical risk $\widehat{L}(\beta)$. Analyzing the empirical risk is more complicated than analyzing the population risk, so we focus on the case when $\beta^\star$ is 1-sparse, i.e. $r = 1$. (When $r > 1$, the main idea is the same but requires some more advanced analysis techniques.) We restate the case of $r = 1$ in the following theorem.

Theorem 14.4. *Suppose $\eta \ge \widetilde{\Omega}(1)$. Then, gradient descent on $\widehat{L}(\beta)$ with $t = \Theta\left(\frac{\alpha \log(1/\delta)}{\eta}\right)$ steps satisfies*

$$\left\|\beta^t \odot \beta^t - \beta^\star \odot \beta^\star\right\|_2^2 \le \widetilde{O}\left(\frac{1}{\sqrt{n}}\right). \tag{14.14}$$

Remark 14.5. Note that Theorem 14.4 is a slightly weaker version of Theorem 14.1 for $r = 1$, since the bound on the RHS depends on the number of examples and not the initialization α. In Theorem 14.1, we could take α as small as we like to drive the bound to zero; we cannot do this for Theorem 14.4.

In the proof of this theorem, we will make use of the *restricted isometry property* (r, δ)-RIP (refer to last lecture) as well as the following lemma.

Lemma 14.6. *Suppose $x^{(1)}, x^{(2)}, \dots x^{(n)}$ satisfy the (r, δ)-RIP condition. Then, $\forall v, w$ such that $\|v\|_0 \leq r$ and $\|w\|_0 \leq r$, we have that*

$$\left| \frac{1}{n} \sum_{i=1}^{n} \langle x^{(i)}, v \rangle \langle x^{(i)}, w \rangle - \langle v, w \rangle \right| = \left| v^T \left(\frac{1}{n} \sum_{i=1}^{n} x^{(i)} (x^{(i)})^\top \right) w - \langle v, w \rangle \right| \tag{14.15}$$

$$\leq 4\delta \|v\|_2 \cdot \|w\|_2 . \tag{14.16}$$

Corollary 14.7. Taking $w = e_1, \dots, e_d$ in Lemma 14.6, we can conclude that

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \langle x^{(i)}, v \rangle x^{(i)} - v \right\|_\infty = \left\| \left(\frac{1}{n} \sum_{i=1}^{n} x^{(i)} (x^{(i)})^\top \right) v - v \right\|_\infty \tag{14.17}$$

$$\leq 4\delta \|v\|_2 . \tag{14.18}$$

Specifically, in the Gaussian case when $x^{(1)} \dots x^{(n)} \overset{\text{iid}}{\sim} \mathcal{N}(0, I_{d \times d})$ and when $n \geq \widetilde{\Omega}(r/\delta^2)$, with high probability the data satisfy the (r, δ)-RIP condition. It follows that when $r = 1$ and $\delta = \widetilde{O}(1/\sqrt{n})$, the data are $(1, \delta)$-RIP. With this, we proceed to prove Theorem 14.4 over the next few subsections.

14.3.1 Computing the gradient update $\nabla \widehat{L}(\beta)$

WLOG, assume that $\beta^\star = e_1$. We can decompose the gradient descent iterate β^t as

$$\beta^t = r_t \cdot e_1 + \zeta_t, \tag{14.19}$$

where $\zeta_t \perp e_1$. The idea is to prove convergence to $\beta^\star$ by showing that (i) $r_t \to 1$ as $t \to \infty$, and (ii) $\|\zeta_t\|_\infty \leq O(\alpha)$ for $t \leq \widetilde{O}(1/\eta)$. In other words, the *signal* r_t converges quickly to 1 while the *noise* ζ_t remains small for some number of initial iterations. One may be concerned that it is possible for the noise to amplify after many iterations, but we will not have to worry about this scenario if we can guarantee that β^t converges to $\beta^\star$ first.

We can compute the gradient of $\widehat{L}(\beta^t)$ as follows. Since $y^{(i)} = \langle \beta^\star \odot \beta^\star, x^{(i)} \rangle$ and $\beta^t = r_t e_1 + \zeta_t =$

$r_t\beta^\star + \zeta_t$,

$$\nabla \widehat{L}(\beta^t) = \frac{1}{n}\sum_{i=1}^{n}(\langle \beta^t \odot \beta^t, x^{(i)}\rangle - y^{(i)})x^{(i)} \odot \beta^t \tag{14.20}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\langle \beta^t \odot \beta^t - \beta^\star \odot \beta^\star, x^{(i)}\rangle)x^{(i)} \odot \beta^t \tag{14.21}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\langle r_t^2\beta^\star \odot \beta^\star + \zeta_t \odot \zeta_t - \beta^\star \odot \beta^\star, x^{(i)}\rangle x^{(i)} \odot \beta^t \tag{14.22}$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left\langle \big(r_t^2 - 1\big)\beta^\star \odot \beta^\star + \zeta_t \odot \zeta_t, x^{(i)}\right\rangle x^{(i)}}_{m_t} \odot \beta^t. \tag{14.23}$$

To simplify the analysis, we can rearrange some of the terms that are part of the gradient. Define m_t such that $\nabla \widehat{L}(\beta^t) = m_t \odot \beta^t$. Also, let $X = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}(x^{(i)})^\top$. Then,

$$m_t = \frac{1}{n}\sum_{i=1}^{n}\left\langle \big(r_t^2 - 1\big)\beta^\star \odot \beta^\star + \zeta_t \odot \zeta_t,\ x^{(i)}\right\rangle x^{(i)} \tag{14.24}$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} x^{(i)}\big(x^{(i)}\big)^\top\right)\big(r_t^2 - 1\big)\cdot(\beta^\star \odot \beta^\star) + \left(\frac{1}{n}\sum_{i=1}^{n} x^{(i)}(x^{(i)})^\top\right)(\zeta_t \odot \zeta_t) \tag{14.25}$$

$$= \underbrace{X\big(r_t^2 - 1\big)\cdot\big(\beta^\star \odot \beta^\star\big)}_{\text{part of } u_t} + \underbrace{X\big(\zeta_t \odot \zeta_t\big)}_{v_t}. \tag{14.26}$$

Now, define $u_t := (r_t^2 - 1)(\beta^\star \odot \beta^\star) - X(r_t^2 - 1)(\beta_* \odot \beta_*)$ and $v_t := X\big(\beta_t \odot \beta_t\big)$. Then we can rewrite the gradient as

$$\nabla \widehat{L}(\beta^t) = m_t \odot \beta^t = [(r_t^2 - 1)\beta^\star \odot \beta^\star - u_t + v_t] \odot \beta_t. \tag{14.27}$$

Our goal is to show that both u_t and v_t are small, so that $\nabla \widehat{L}(\beta^t)$ is close to its population version $\nabla L(\beta^t)$. Observe that X appears in both u_t and v_t. This matrix is challenging to deal with mathematically because it does not have full rank (because $n < d$). Instead, we rely on the RIP condition to reason about the behavior of X: the idea is that X behaves like the identity for sparse vector multiplication. Applying Corollary 14.7, we can bound $\|u_t\|_\infty$ as

$$\|u_t\|_\infty \le 4\delta \left\|(r_t^2 - 1)\beta^\star \odot \beta^\star\right\|_2 \le 4\delta||\beta^\star \odot \beta^\star||_2 \le 4\delta. \tag{14.28}$$

(In the second inequality, we assume that $|r_t| < 1$. We can do this because r_t starts out at α which is small; if $r_t \ge 1$, then we are already in the regime where gradient descent has converged.) We can bound $\|v_t\|_\infty$ in a similar manner: since Corollary 14.7 implies $\|v_t - \zeta_t \odot \zeta_t\|_\infty \le 4\delta \|\zeta_t \odot \zeta_t\|_2$,

$$\|v_t\|_\infty \le \|\zeta_t \odot \zeta_t\|_\infty + 4\delta \|\zeta_t \odot \zeta_t\|_2 \qquad \text{(by the triangle inequality)} \tag{14.29}$$

$$\le \|\zeta_t\|_\infty^2 + 4\delta \|\zeta_t \odot \zeta_t\|_1 \qquad \text{(since } \zeta_t \text{ very small)} \tag{14.30}$$

$$= \|\zeta_t\|_\infty^2 + 4\delta \|\zeta_t\|_2^2. \tag{14.31}$$

Note that the size of v_t depends on the size of the noise ζ_t. Thus, by bounding ζ_t (e.g. with a small initialization), we can ensure that v_t is also small. (Ensuring bounds on u_t is more difficult because it depends only on δ.) In the next subsection, we analyze the growth of ζ_t and r_t.

14.3.2 Dynamics analysis of ζ_t

First, we analyze the dynamics of the noise ζ_t, which we want to ensure does not grow too fast.

Lemma 14.8. *For all $t \le 1/(c\eta\delta)$ with sufficiently large constant c, we have*

$$\|\zeta_t\|_\infty \le 2\alpha, \qquad \|\zeta_t\|_2^2 \le \frac{1}{2}, \qquad \text{and} \quad \|\zeta_{t+1}\|_\infty \le \big(1 + O(\eta\delta)\big)\,\|\zeta_t\|_\infty . \tag{14.32}$$

Note that this result is weaker than what we were able to show for the population gradient (exponential growth with a small fixed rate), but we will ultimately show that the growth of the signal will be even faster.

Proof. Recall that the empirical gradient (14.27) is $\nabla \hat{L}(\beta) = \big[(r_t^2 - 1)\beta^\star \odot \beta^\star - u_t + v_t\big] \odot \beta^t$. Hence, the gradient update to β^t is

$$\begin{aligned}
\beta^{t+1} &= \beta^t - \eta\left[\left(r_t^2 - 1\right)\beta^\star \odot \beta^\star - u_t + v_t\right] \odot \beta^t && (14.33)\\
&= \underbrace{\beta^t - \eta\left(r_t^2 - 1\right)\beta^\star \odot \beta^\star \odot \beta^t}_{\text{GD update for population loss}} - \eta\left(-u_t + v_t\right) \odot \beta^t. && (14.34)
\end{aligned}$$

Recall that ζ_{t+1} is simply β^{t+1} except for the first coordinate (where it has a zero instead of r_{t+1}), i.e. ζ_{t+1} is the projection of β^{t+1} onto the subspace orthogonal to e_1. Hence,

$$\begin{aligned}
\zeta_{t+1} &= \left(I - e_1 e_1^\top\right)\beta^{t+1} && (14.35)\\
&= \left(I - e_1 e_1^\top\right)\cdot \beta^t - \eta\left(I - e_1 e_1^\top\right)(v_t - u_t) \odot \beta^t \quad \text{(by (14.34), second term = 0)} && (14.36)\\
&= \zeta_t - \eta\left[\left(I - e_1 e_1^T\right)(v_t - u_t) \odot \left(I - e_1 e_1^T\right)\beta^t\right] \qquad \text{(by distribution law for } \odot\text{)} && (14.37)\\
&= \zeta_t - \eta \underbrace{\left[\left(I - e_1 e_1^T\right)(v_t - u_t)\right]}_{\rho_t} \odot \zeta_t. && (14.38)
\end{aligned}$$

If we define ρ_t such that $\zeta_{t+1} = \zeta_t - \eta\rho_t \odot \zeta_t$, then the growth of ζ_t is dictated by the size of ρ_t. We can bound this as

$$\|\zeta_{t+1}\|_\infty \le (1 + \eta\,\|\rho_t\|_\infty)\,\|\zeta_t\|_\infty . \tag{14.39}$$

Now, we will prove the lemma by using strong induction on t. Suppose that the first two pieces of (14.32) hold for all iterations up to t. We can show that

$$\begin{aligned}
\|\rho_t\|_\infty &\le \|u_t\|_\infty + \|v_t\|_\infty && (14.40)\\
&\le 4\delta + \|\zeta_t\|_\infty^2 + 4\delta\,\|\zeta_t\|_2^2 \qquad \text{(by (14.28) and (14.31))} && (14.41)\\
&\le 4\delta + (2\alpha)^2 + 4\delta \cdot \frac{1}{2} \qquad \text{(by the inductive hypothesis)} && (14.42)\\
&\le 8\delta, && (14.43)
\end{aligned}$$

where the last step holds because we can take α to be arbitrarily small (e.g. $\alpha \le \text{poly}(1/n) \le O(\delta)$). Plugging this into (14.39), we have

$$\|\zeta_{t+1}\|_\infty \le (1 + 8\eta\delta) \|\zeta_t\|_\infty = \big(1 + O(\eta\delta)\big) \|\zeta_t\|_\infty , \tag{14.44}$$

which proves the third piece of the lemma. Using this piece, we can show that

$$\|\zeta_{t+1}\|_\infty \le (1 + 8\eta\delta)^{t+1} \|\zeta_0\|_\infty \le \big(1 + 8\eta\delta\big)^{1/(c\eta\delta)} \cdot \alpha \le 2\alpha \tag{14.45}$$

for a sufficiently large constant c, which proves the second piece. Finally, we show that

$$\|\zeta_{t+1}\|_2^2 \le \big(1 + 8\eta\delta\big)^{t+1} \|\zeta_0\|_2^2 \le \big(1 + 8\eta\delta\big)^{1/(c\eta\delta)} \cdot \alpha\sqrt{d} \le \frac{1}{2}, \tag{14.46}$$

if $\alpha \le \frac{1}{n^{O(1)}}$, which proves the first piece.

□

14.3.3 Dynamics analysis of r_t

Next, we analyze the dynamics of the signal r_t, which we want to show converges to 1.

Lemma 14.9. *For all $t \le 1/(c\eta\delta)$ with sufficiently large constant c, we have that*

$$r_{t+1} = \big(1 + \eta\big(1 - r_t^2\big)\big) r_t + O\big(\eta\delta\big) r_t.$$

Note that the first term on the RHS is r_{t+1} during gradient descent on the population loss, and the second term captures the error.

Proof. Recall that the gradient descent update from the empirical gradient (14.27) is

$$\beta^{t+1} = \beta^t - \eta\big[\big(r_t^2 - 1\big)\beta^\star \odot \beta^\star - u_t + v_t\big] \odot \beta_t. \tag{14.47}$$

We have that

$$\begin{aligned}
r_{t+1} &= \big\langle \beta^{t+1}, e_1 \big\rangle && (14.48)\\
&= \big\langle \beta^t, e_1 \big\rangle - \eta\big(r_t^2 - 1\big)\big\langle \beta^t, e_1 \big\rangle - \eta\big\langle v_t - u_t, e_1 \big\rangle\big\langle \beta^t, e_1 \big\rangle && (14.49)\\
&= r_t - \eta\big(r_t^2 - 1\big) r_t - \eta\big\langle v_t - u_t, e_1 \big\rangle r_t && (14.50)\\
&= \Big(1 + \eta\big(1 - r_t^2\big)\Big) r_t + \eta\big\langle u_t - v_t, e_1 \big\rangle r_t && (14.51)
\end{aligned}$$

so all we need to do is bound the second term as follows:

$$\begin{aligned}
|\eta\langle v_t - u_t, e_1\rangle r_t| &\le \eta \cdot r_t \|v_t - u_t\|_\infty && (14.52)\\
&\le \eta \cdot r_t \cdot 8\delta \qquad \text{(by (14.43))} && (14.53)\\
&= O(\eta\delta) \cdot r_t. && (14.54)
\end{aligned}$$

□

14.3.4 Putting it all together

Finally, we return to the proof of Theorem 14.4. By Lemma 14.9, we know that as long as $r_t \le 1/2$ it will grow exponentially fast, since

$$r_{t+1} \ge \Big(1 + \eta\big(1 - r_t^2\big) - O(\eta\delta)\Big) \cdot r_t \ge \left(1 + \frac{\eta}{2}\right) \cdot r_t. \tag{14.55}$$

This implies that at some $t_0 = O\Big(\frac{\log(1/\alpha)}{\eta}\Big)$, we'll observe $r_{t_0} > 1/2$ for the first time. Consider what happens after this point.

- When $1/2 < r_t \le 1$, we have that

$$\begin{aligned} 1 - r_{t+1} &\le 1 - r_t - \eta\big(1 - r_t^2\big) r_t + O(\eta\delta) \cdot r_t && (14.56)\\ &\le 1 - r_t - \frac{\eta\big(1 - r_t^2\big)}{2} + O(\eta\delta) && (14.57)\\ &\le 1 - r_t - \frac{\eta\big(1 - r_t\big)}{2} + O(\eta\delta) && (14.58)\\ &= \left(1 - \frac{\eta}{2}\right)(1 - r_t) + O(\eta\delta). && (14.59) \end{aligned}$$

  Thus, we can achieve $1 - r_{t+1} \le 2 \cdot \frac{O(n\delta)}{\eta/2} = O(\delta)$ in $\Theta\Big(\frac{\log(1/\delta)}{\eta}\Big)$ iterations.

- When $r_t > 1$, we can show in a similar manner that

$$r_{t+1} - 1 \le (1 - \eta)(r_t - 1) + O(\eta\delta) \le O(\delta), \tag{14.60}$$

  implying that r_t remains very close to 1 after the same order of iterations.

This completes the proof of Theorem 14.4, bounding the number of iterations needed for gradient descent on the empirical loss to converge to β^*. □