Chapter 9

Implicit/Algorithmic Regularization Effect

One of the miracles of modern deep learning is the phenomenon of *algorithmic regularization* (also known as *implicit regularization* or *implicit bias*): although the loss landscape may contain infinitely many global minimizers, many of which do not generalize well, in practice our optimizer (e.g. SGD) tends to recover solutions with good generalization properties.

The focus of this chapter will be to illustrate algorithmic regularization in simple settings. In particular, we first show that gradient descent (with the right initialization) identifies the minimum norm interpolating solution in overparametrized linear regression. Next, we show that for a certain non-convex reparametrization of the linear regression task (where the data is generated from a sparse ground-truth model), gradient descent (again, suitably initialized) approximately recovers a sparse solution (with good generalization). Finally, we discuss algorithmic regularization in the "classification" setting, and how stochasticity can contribute to algorithmic regularization.

9.1 Implicit regularization effect of zero initialization in over-parametrized linear regression

We prove that gradient descent (initialized at the origin) converges to the minimum norm (interpolating solution) (assuming such a solution exists).

Let $X \triangleq \left[x^{(1)}, ..., x^{(n)}\right]^\top \in \mathbb{R}^{n \times d}$ denote our data matrix and $\vec{y} \triangleq \left[y^{(1)}, ..., y^{(n)}\right]^\top \in \mathbb{R}^n$ denote our label vector, where $n < d$. (Assume X is full rank.) Our goal is to find a weight vector β that minimizes our empirical loss function $\widehat{L}(\beta) \triangleq \frac{1}{2}\|\vec{y} - X\beta\|_2^2$.

As we are in the overparametrized setting with $n < d$ and X full rank, there exist infinitely many global minimizers that interpolate the data and hence achieve zero loss. In fact, the following lemma shows that the set of global minimizers forms a subspace.

Lemma 9.1. *Let X^+ denote the pseudoinverse[1] of X. Then β is a global minimizer if and only if $\beta = X^+\vec{y} + \zeta$ for some ζ such that $\zeta \perp x_1, ..., x_n$.*

Proof. For any $\beta \in \mathbb{R}^d$, we can decompose it as $\beta = X^+\vec{y} + \zeta$ for some $\zeta \in \mathbb{R}^d$. Since

$$X\beta = X(X^+\vec{y} + \zeta) = \vec{y} + X\zeta, \tag{9.1}$$

β is a global minimizer if and only if $X\zeta = 0$, which happens if and only if $\zeta \perp x_1, ..., x_n$.

□

[1] Since X is full rank, $XX^\top$ is invertible and so we have $X^+ = X^\top(XX^\top)^{-1}$. Note that $XX^+X = X$.

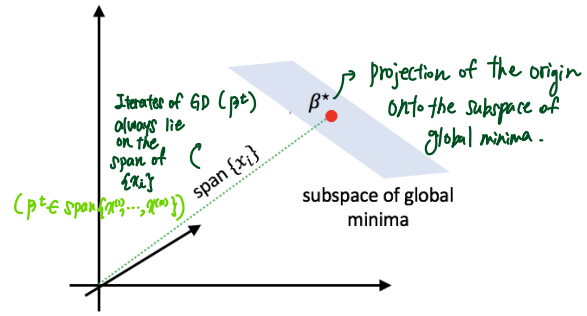


Figure 9.1: Visualization of proof intuition for Theorem 9.3. The solution $\beta^\star$ is the projection of the origin onto the subspace of global minima.

From Lemma 9.1, we can derive an explicit formula for the minimum norm interpolant $\beta^\star \triangleq \operatorname{argmin}_{\beta:\widehat{L}(\beta)=0} \|\beta\|_2$.

Corollary 9.2. $\beta^\star = X^+\vec{y}$.

Proof. Take any β such that $\widehat{L}(\beta) = 0$, and write $\beta = X^+\vec{y} + \zeta$. Then from the definition of X^+ and the fact that $X\zeta = 0$ (see the proof of Lemma 9.1), we have

$$\begin{aligned}
||\beta||_2^2 &= ||X^+\vec{y}||_2^2 + ||\zeta||_2^2 + 2\langle X^+\vec{y}, \zeta\rangle && (9.2)\\
&= ||X^+\vec{y}||_2^2 + ||\zeta||_2^2 + 2\langle X^\top(XX^\top)^{-1}\vec{y}, \zeta\rangle && (9.3)\\
&= ||X^+\vec{y}||_2^2 + ||\zeta||_2^2 + 2\langle (XX^\top)^{-1}y, X\zeta\rangle && (9.4)\\
&= ||X^+\vec{y}||_2^2 + ||\zeta||_2^2 && \text{(because } X\zeta = 0\text{)} \quad (9.5)\\
&\geq ||X^+\vec{y}||_2^2, && (9.6)
\end{aligned}$$

(with equality if and only if $\zeta = 0$.) □

Now, suppose we learn β using "gradient descent" with initialization β^0, where at iteration t we set $\beta^t = \beta^{t-1} - \eta\nabla\widehat{L}(\beta^{t-1})$ for some learning rate η. Since $\widehat{L}(\beta)$ is convex, we know from standard results in convex optimization that gradient descent will converge to a global minimizer for a suitably chosen learning rate η (in particular, taking η to be sufficiently small). Assuming $\beta^0 = 0$, we will in fact recover the minimum norm interpolating solution.

Theorem 9.3. *Suppose gradient descent on $\widehat{L}(\beta)$ with initialization $\beta^0 = 0$ converges to a solution $\hat{\beta}$ such that $\widehat{L}(\hat{\beta}) = 0$. Then $\hat{\beta} = \beta^\star$.*

The main idea of the proof is that the iterates of gradient descent always lie in the span of the $x^{(i)}$'s (see Figure 9.1 for an illustration).

Proof. We first show via induction that $\beta^t \in \text{span}\{x^{(1)}, \dots, x^{(n)}\}$ for all t. For the induction base case, note that $\beta^0 = 0 \in \text{span}\{x^{(1)}, \dots, x^{(n)}\}$. Now suppose $\beta^{t-1} \in \text{span}\{x^{(1)}, \dots, x^{(n)}\}$. Recall that $\beta^t = \beta^{t-1} - \eta\nabla\widehat{L}(\beta^{t-1})$. Since left-multiplying any vector by $X^\top$ amounts to taking a linear combination of the rows of X, it follows that $\eta\nabla\widehat{L}(\beta^{t-1}) = \eta X^\top(X\beta^{t-1} - \vec{y}) \in \text{span}\{x^{(1)}, \dots, x^{(n)}\}$, and so $\beta^t = \beta^{t-1} - \eta\nabla\widehat{L}(\beta^{t-1}) \in \text{span}\{x^{(1)}, \dots, x^{(n)}\}$. This proves the induction step.

Next, we show that $\hat{\beta} \in \text{span}\{x^{(1)}, \dots, x^{(n)}\}$ and $\widehat{L}(\hat{\beta}) = 0$ implies $\hat{\beta} = \beta^\star$. By definition, $\hat{\beta} \in \text{span}\{x^{(1)}, \dots, x^{(n)}\}$ implies $\hat{\beta} = X^\top v$ for some $v \in \mathbb{R}^n$. Since $\widehat{L}(\hat{\beta}) = 0$, we have $0 = X\hat{\beta} - \vec{y} = XX^\top v - \vec{y}$. This implies $v = (XX^\top)^{-1}y$, and so $\hat{\beta} = X^\top v = X^\top(XX^\top)^{-1}\vec{y} = X^+\vec{y} = \beta^\star$. □

9.2 Implicit regularization of small initialization in nonlinear models

We give another example of implicit regularization effect of small initialization in a non-convex version of the overparametrized linear regression task considered in the previous section. The results in this subsection are largely simplifications of the paper Li et al. [2017] which studies over-parameterized compressed sensing and two-layer neural nets with quadratic activation.

We assume $x^{(1)}, ..., x^{(n)} \overset{\text{iid}}{\sim} \mathcal{N}(0, I_{d\times d})$ and $y^{(i)} = f_{\beta^\star}(x^{(i)})$, where the ground truth vector $\beta^\star$ is r-sparse (i.e. $\|\beta^\star\|_0 = r$). For simplicity, we assume $\beta^\star_i = \mathbf{1}\{i \in S\}$ for some $S \subset [d]$ such that $|S| = r$. We again analyze the overparametrized setting, where this time $n \ll d$ but also $n \geq \Omega(r^2)$.

Our goal is to find a weight vector that minimizes our empirical loss function

$$\widehat{L}(\beta) \triangleq \frac{1}{4n}\sum_{i=1}^{n}\left(y^{(i)} - f_\beta(x^{(i)})\right)^2, \tag{9.7}$$

where $f_\beta(x) \triangleq \langle \beta \odot \beta, x\rangle$. The operation $\odot$ denotes the Hadamard product: for $u, v \in \mathbb{R}^d$, $u \odot v \in \mathbb{R}^d$ is defined by $(u \odot v)_i \triangleq u_i v_i$ for $i = 1, \ldots, d$.

9.2.1 Main results of algorithmic regularization

Note that while f_β is still linear over x, our loss is no longer convex over β. (To see this, suppose $\beta \neq 0$ is a global minimizer. Then we have $\widehat{L}(0) > \widehat{L}(\beta) = \widehat{L}(-\beta)$.) Thus, the effect of algorithmic regularization induced by gradient descent will be much different from the overparametrized linear regression setting.

In the previous setting of linear regression, solutions with low ℓ_2 norm are desirable as they tend to generalize well. In the present setting, we know our ground-truth parameter $\beta^\star$ is sparse. Thus, we want to learn a sparse solution $\hat{\beta}$, avoiding non-sparse solutions that may not generalize well. One approach to finding sparse solutions, called *lasso regression*, is to minimize the ℓ_1-regularized proxy loss

$$\sum_{i=1}^{n}\left(\langle \theta, x^{(i)}\rangle - y^{(i)}\right)^2 + \lambda\|\theta\|_1 \tag{9.8}$$

with respect to θ, where $\theta = \beta \odot \beta$. However, it turns out that we can equivalently learn a sparse solution by running gradient descent from a suitable initialization on the original *unregularized* loss.

To be specific, let $\beta^0 = \alpha\mathbf{1} \in \mathbb{R}^d$ be the initialization where α is a small positive number. The update rule of gradient descent algorithm is given by $\beta^{t+1} = \beta^t - \eta\nabla\widehat{L}(\beta^t)$. The next theorem shows that when $n = \widetilde{\Omega}(r^2)$, gradient descent on $\widehat{L}(\beta)$ converges to $\beta^\star$.

Theorem 9.4. *Let c be a sufficiently large universal constant. Suppose $n \geq cr^2\log^2(d)$ and $\alpha \leq 1/d^c$, then when $\dfrac{\log(d/\alpha)}{\eta} \lesssim T \lesssim \dfrac{1}{\eta\sqrt{d\alpha}}$, we have*

$$\left\|\beta^\top \odot \beta^\top - \beta^\star \odot \beta^\star\right\|_2^2 \leq O\left(\alpha\sqrt{d}\right). \tag{9.9}$$

(Here, T indexes the gradient descent steps.)

We make several remarks about Theorem 9.4 before presenting the proof.

Remark 9.5. In this problem we do not use $\beta^0 = 0$ as the initialization point because $\beta = 0$ is a critical point, that is, $\nabla\widehat{L}(0) = 0$. Note that the lower bound on T depends logarithimically on $1/\alpha$, so we can take α to be a small inverse polynomial on d and the lower bound won't change much. Also, the upper bound depends polynomially on $1/\alpha$ (which is considered very big when c is sufficiently large), so we do not need to use early stopping in a serious way.

Remark 9.6. Theorem 9.4 is a simplified version of Theorem 1.1 in [Li et al., 2018].

Remark 9.7. $\widehat{L}(\beta)$ has many global minima. To see this, observe that the number of parameters is d and the number of constraints to fit all the examples is $O(n)$ because there are only n examples. Recall that for overparameterized model we have $d \gg n$; consequently, there exists many global minima of $\widehat{L}(\beta)$.

Remark 9.8. $\beta^\star$ is the min-norm solution in this case. That is,

$$\beta^\star = \operatorname{argmin} \|\beta\|_2^2 \qquad \text{s.t. } \widehat{L}(\beta) = 0. \tag{9.10}$$

Informally, this is because we can view $\beta \odot \beta$ as a vector $\theta \in \mathbb{R}^d$, which leads to $\|\beta\|_2^2 = \|\theta\|_1$. Then in the θ space (and with a little abuse of notation), the optimization problem (9.10) becomes

$$\theta^\star = \operatorname{argmin} \|\theta\|_1 \qquad \text{s.t. } \widehat{L}(\theta) = 0, \tag{9.11}$$

which is a lasso regression, whose solution is sparse.

Remark 9.9. In this non-linear case and the linear case before, gradient descent with small initialization converges to minimum ℓ_2-norm solution. Similarly, in the NTK regime, gradient descent converges to a solution that is very close to the initialization. Therefore, it seems conceivable that GD generally prefers global minima nearest to the initialization. However, we do not have a general theorem for this phenomenon (and the instructor also believes that this is not universally true without other conditions).

9.2.2 Ground work for proof and the restricted isometry property

In this section we prepare the ground work for the proof of Theorem 9.4.

We start by showing several basic properties about $\widehat{L}(\beta)$. Note that for any fixed vector $v \in \mathbb{R}^d$ and $x \in \mathbb{R}^d$, when x is drawn from $\mathcal{N}(0, I)$, we have

$$\mathbb{E}\left[\langle x, v\rangle^2\right] = \mathbb{E}\left[v^\top x x^\top v\right] = v^\top \mathbb{E}\left[x x^\top\right] v = \|v\|_2^2. \tag{9.12}$$

It follows that

$$L(\beta) = \frac{1}{4} \mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,I)} \left[(y - \langle \beta \odot \beta, x\rangle^2\right] \quad \text{(9.1 그대로)} \tag{9.13}$$

$$= \frac{1}{4} \mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,I)} \left[\langle \beta^\star \odot \beta^\star - \beta \odot \beta, x\rangle^2\right] \quad \text{(by definition of } y) \tag{9.14}$$

$$= \frac{1}{4} \|\beta^\star \odot \beta^\star - \beta \odot \beta\|_2^2 \cdot \mathbb{E}\left[\langle x, v\rangle^2\right] = \|v\|_2^2 \quad \text{(by (9.12))} \tag{9.15}$$

Note that (9.15) is the metric that we use to characterize how close β is to the ground-truch parameter $\beta^\star$ (see (9.9)).

In the following lemma we show that $\widehat{L}(\beta) \approx L(\beta)$ by uniform convergence. Generally speaking, uniform convergence of the loss function for all β requires $n \geq \Omega(d)$ samples, so in our setting (where $n \ll d$) $\widehat{L}(\beta) \approx L(\beta)$ does not always hold. However, since we assume $\beta^\star$ is sparse, the analysis only requires uniform convergence for sparse vectors.

Lemma 9.10. *Assume $n \geq \widetilde{\Omega}(r^2)$. With high probability over the randomness in $x^{(1)}, \cdots, x^{(n)}$, $\forall v$ such that $\|v\|_0 \leq r$ we have*

$$(1-\delta)\|v\|_2^2 \leq \frac{1}{n}\sum_{i=1}^{n} \langle v, x^{(i)}\rangle^2 \leq (1+\delta)\|v\|_2^2. \tag{9.16}$$

Lemma 9.10 is a special case of Lemma 2.2 in [Li et al., 2018] so the proof is omitted here. We say the set $\{x^{(1)}, \cdots, x^{(n)}\}$ (or $X = [x^{(1)}, \cdots, x^{(n)}]$) satisfies (r, δ)-*RIP condition* (*restricted isometric property*) if (9.16) holds.

$\forall v$ s.t. $\|v\|_0 \le r$ (where $n \gtrsim \Omega(r^2)$),

$$(1-\delta)\|v\|_2^2 \le \frac{1}{n}\sum_{i=1}^{n}\langle v, x^{(i)}\rangle^2 \le (1+\delta)\|v\|_2^2. \tag{9.16}$$

By algebraic manipulation, (9.16) is equivalent to

$$(1-\delta)\|v\|_2^2 \le v^\top \left(\frac{1}{n}\sum_{i=1}^{n} x^{(i)}(x^{(i)})^\top\right) v \le (1+\delta)\|v\|_2^2. \tag{9.17}$$

In other words, from the point of view of a sparse vector v we have $\sum_{i=1}^n x^{(i)}(x^{(i)})^\top \approx I$. (Note however that $\sum_{i=1}^n x^{(i)}(x^{(i)})^\top$ is not close to $I_{d\times d}$ in other notions of closeness.) For example, $\sum_{i=1}^n x^{(i)}(x^{(i)})^\top$ is not close to $I_{d\times d}$ in spectral norm. Another way to see this is that $\sum_{i=1}^n x^{(i)}(x^{(i)})^\top$ is a $d \times d$ matrix but only has rank $n \ll d$.)

As a result, with the RIP condition we have $\widehat{L}(\beta) \approx L(\beta)$ if β is sparse. With more tools we can also get $\nabla\widehat{L}(\beta) \approx \nabla L(\beta)$. Let us define the set $S_r = \{\beta : \|\beta\|_0 \le O(r)\}$, the set where we have uniform convergence of $\widehat{L}$ and $\nabla\widehat{L}$. Informally, as long as we are in the set S_r, $\widehat{L}$ and $\nabla\widehat{L}$ have similar behavior to their population counterparts. (Note, on the other hand, that there exists a dense $\beta \notin S_r$ such that $\widehat{L}(\beta) = 0$ but $L(\beta) \gg 0$.)

The RIP condition also gives us the following lemma which will be needed for the proof of Theorem 9.4.

Lemma 9.11. *Suppose $x^{(1)}, x^{(2)}, \dots x^{(n)}$ satisfy the (r,δ)-RIP condition. Then, $\forall v, w$ such that $\|v\|_0 \le r$ and $\|w\|_0 \le r$, we have that*

$$\left|\frac{1}{n}\sum_{i=1}^{n}\langle x^{(i)}, v\rangle\langle x^{(i)}, w\rangle - \langle v, w\rangle\right| = \left|v^T\left(\frac{1}{n}\sum_{i=1}^{n} x^{(i)}(x^{(i)})^\top\right)w - \langle v, w\rangle\right| \tag{9.18}$$

$$\le 4\delta\,\|v\|_2 \cdot \|w\|_2 \tag{9.19}$$

Corollary 9.12. Taking $w = e_1, \dots, e_d$ in Lemma 9.11, we can conclude that

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\langle x^{(i)}, v\rangle x^{(i)} - v\right\|_\infty = \left\|\left(\frac{1}{n}\sum_{i=1}^{n} x^{(i)}(x^{(i)})^\top\right)v - v\right\|_\infty \tag{9.20}$$

$$\le 4\delta\,\|v\|_2\,. \tag{9.21}$$

9.2.3 Warm-up for analysis: Gradient descent on population loss

The main intuition for proving Theorem 9.4 is to leverage the uniform convergence when β belongs to the set S_r (see Figure 9.2). Note that the initialization β^0 is not exactly r-sparse, but taking α to be sufficiently small, β^0 is approximately 0-sparse. The proof is decomposed into the following steps:

1. Gradient descent on $L(\beta)$ converges to $\beta^\star$ without leaving S_r, and
2. Gradient descent on $\widehat{L}(\beta)$ is similar to gradient descent on $L(\beta)$ inside S_r.

Combining the two steps we can show that gradient descent on $\widehat{L}(\beta)$ does not leave S_r and converges to $\beta^\star$.

As a warm up, we prove the following theorem for gradient descent on $L(\beta)$.

Theorem 9.13. *For sufficiently small η, gradient descent on $L(\beta)$ converges to $\beta^\star$ in $\Theta\left(\frac{\log(1/(\epsilon\alpha))}{\eta}\right)$ iteration with ϵ-error in ℓ_2-distance.*

Proof. Since

$$\nabla L(\beta) = (\beta\odot\beta - \beta^\star\odot\beta^\star)\odot\beta, \tag{9.22}$$

the gradient descent step is

$$\beta^{t+1} = \beta^t - \eta(\beta^t\odot\beta^t - \beta^\star\odot\beta^\star)\odot\beta^t. \tag{9.23}$$

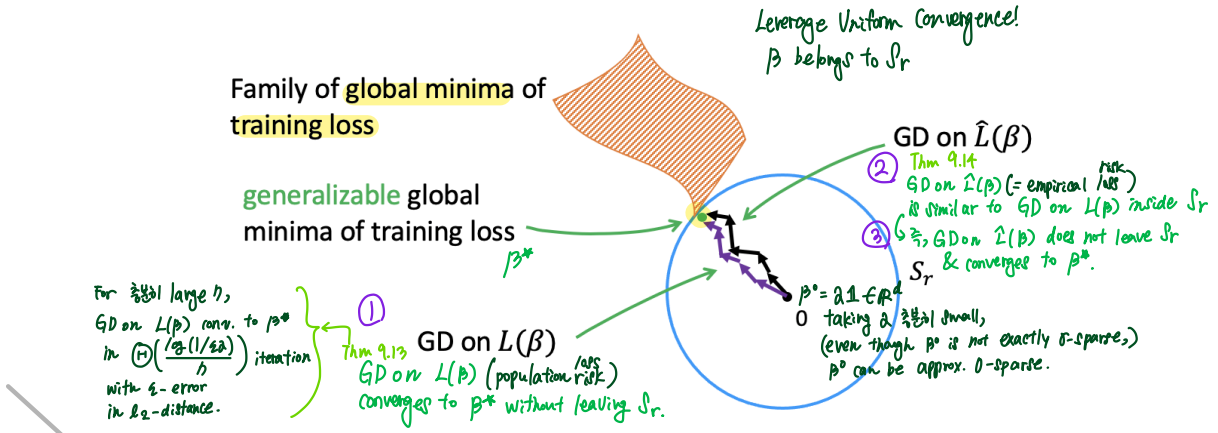


Figure 9.2: Visualization of proof intuition for Theorem 9.4.

Recall that $\beta^\star = \mathbf{1}\{i \in S\}$ and $\beta^0 = \alpha\mathbf{1}$, and the update rule above decouples across the coordinates of β^t. Thus, we only need to show that $|\beta_i^\star - \beta_i^t| \le \epsilon$ for the number of iterations stated in the Theorem.

Case 1: $i \in S$. For $i \in S$, the update rule for coordinate i is

$$\beta_i^{t+1} = \beta_i^t - \eta(\beta_i^t \cdot \beta_i^t - 1 \cdot 1) \cdot \beta_i^t \tag{9.24}$$

$$= \beta_i^t - \eta\left[(\beta_i^t)^2 - 1\right]\beta_i^t. \tag{9.25}$$

Consider the following two cases:

- If $\beta_i^t \le 1/2$, we have

$$\beta_i^{t+1} = \beta_i^t\left[1 + \eta\left(1 - (\beta_i^t)^2\right)\right] \tag{9.26}$$

$$\ge \beta_i^t\left(1 + \frac{3}{4}\eta\right). \tag{9.27}$$

Consequently, β_i^{t+1} grow exponentially, and it takes $\Theta\left(\frac{\log(1/\alpha)}{\eta}\right)$ iterations for β_i^t to grow from α to at least $1/2$.[2] This will bring us into the second case.

- if $\beta_i^t \ge 1/2$, we have

$$1 - \beta_i^{t+1} = 1 - \beta_i^t + \eta\left[(\beta_i^t)^2 - 1\right]\beta_i^t \tag{9.28}$$

$$= 1 - \beta_i^t - \eta\left(1 - \beta_i^t\right)\left(1 + \beta_i^t\right)\beta_i^t \tag{9.29}$$

$$\le \left(1 - \beta_i^t\right) - \eta\left(1 - \beta_i^t\right)\beta_i^t \qquad \text{(because } 1 + \beta_i^t \ge 1\text{)} \tag{9.30}$$

$$= \left(1 - \beta_i^t\right)\left(1 - \eta\beta_i^t\right) \tag{9.31}$$

$$\le \left(1 - \beta_i^t\right)(1 - \eta/2). \qquad \text{(because } \beta_i^t \ge 1/2\text{)} \tag{9.32}$$

Therefore it takes $\Theta\left(\frac{\log(1/\epsilon)}{\eta}\right)$ iterations to achieve $1 - \beta_i^t \le \epsilon$.

Case 2: $i \notin S$. For all $i \notin S$, we claim (informally) that it is sufficient to show that when $t \le 1/(10\eta\alpha^2)$, $\beta_i^t \le 2\alpha$. This is because 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.

[2] This is because $(1+\eta)^{1/\eta} \approx e$, so $(1+\eta)^{c/\eta} \approx e^c$.

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

$$\beta_i^{t+1} = \left[\beta^t - \eta(\beta^t \odot \beta^t - \beta^\star \odot \beta^\star) \odot \beta^t\right]_i \tag{9.33}$$

$$= \beta_i^t - \eta(\beta_i^t \cdot \beta_i^t) \cdot \beta_i^t \qquad (\text{since } \beta_i^\star = 0 \ \forall i \notin S) \tag{9.34}$$

$$= \beta_i^t - \eta(\beta_i^t)^3. \tag{9.35}$$

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 $s \le t$,

$$\beta_i^{s+1} = (1 - \eta(\beta_i^s)^2)\beta_i^s \tag{9.36}$$

$$\le (1 + \eta(\beta_i^s)^2)\beta_i^s \tag{9.37}$$

$$\le (1 + 4\eta\alpha^2)\beta_i^s. \qquad (\text{since } \beta_i^s \le 2\alpha) \tag{9.38}$$

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

$$\beta_i^{t+1} \le \beta_0 \cdot (1 + 4\eta\alpha^2)^t \tag{9.39}$$

$$\le \beta_0 (1 + 4\eta\alpha^2)^{\frac{1}{10\eta\alpha^2}} \tag{9.40}$$

$$\le \beta_0 \exp\left(\frac{4\eta\alpha^2}{10\eta\alpha^2}\right) \tag{9.41}$$

$$= \beta_0 \cdot e^{2/5} \tag{9.42}$$

$$\le 2\alpha, \tag{9.43}$$

which completes the inductive proof of the claim. □

9.2.4 Proof of main result: gradient descent on empirical loss

Analyzing gradient descent on the empirical risk $\widehat{L}(\beta)$ is more complicated than analyzing gradient descent on 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.)

Note that in our setup, i.e. when $x^{(1)} \dots x^{(n)} \overset{\text{iid}}{\sim} \mathcal{N}(0, I_{d\times d})$ and when $n \ge \widetilde{\Omega}(r/\delta^2)$, 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. This will allow us to use the lemmas involving the RIP condition for the proof.

We restate the case of $r = 1$ in the following theorem.

Theorem 9.14. *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*

$$\left\|\beta^t \odot \beta^t - \beta^\star \odot \beta^\star\right\|_2^2 \le \widetilde{O}\left(\frac{1}{\sqrt{n}}\right). \tag{9.44}$$

Remark 9.15. Note that Theorem 9.14 is a slightly weaker version of Theorem 9.4 for $r = 1$, since the bound on the RHS depends on the number of examples and not the initialization α. In Theorem 9.4, we could take α as small as we like to drive the bound to zero; we cannot do this for Theorem 9.14.

We proceed to prove Theorem 9.14 with the follow steps:

1. Computing the gradient update $\nabla\widehat{L}(\beta)$,
2. Dynamics analysis of noise ζ_t,
3. Dynamics analysis of signal r_t, and

4. Putting it all together.

(Step 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

$$\beta^t = r_t \cdot e_1 + \zeta_t, \tag{9.45}$$

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) $\|\zeta_t\|_\infty \le O(\alpha)$ for $t \le \widetilde{O}(1/\eta)$. 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 $\beta^t = r_t e_1 + \zeta_t = r_t \beta^\star + \zeta_t$,

$$\nabla \widehat{L}(\beta^t) = \frac{1}{n} \sum_{i=1}^n (\langle \beta^t \odot \beta^t, x^{(i)} \rangle - y^{(i)}) x^{(i)} \odot \beta^t \tag{9.46}$$

$$= \frac{1}{n} \sum_{i=1}^n (\langle \beta^t \odot \beta^t - \beta^\star \odot \beta^\star, x^{(i)} \rangle) x^{(i)} \odot \beta^t \tag{9.47}$$

$$= \frac{1}{n} \sum_{i=1}^n \langle r_t^2 \beta^\star \odot \beta^\star + \zeta_t \odot \zeta_t - \beta^\star \odot \beta^\star, x^{(i)} \rangle x^{(i)} \odot \beta^t \tag{9.48}$$

$$= \underbrace{\frac{1}{n} \sum_{i=1}^n \left\langle (r_t^2 - 1) \beta^\star \odot \beta^\star + \zeta_t \odot \zeta_t, x^{(i)} \right\rangle x^{(i)}}_{m_t} \odot \beta^t. \tag{9.49}$$

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,

$$m_t = \frac{1}{n} \sum_{i=1}^n \left\langle (r_t^2 - 1) \beta^\star \odot \beta^\star + \zeta_t \odot \zeta_t, x^{(i)} \right\rangle x^{(i)} \tag{9.50}$$

$$= \left(\frac{1}{n} \sum_{i=1}^n x^{(i)} (x^{(i)})^\top \right) (r_t^2 - 1) \cdot (\beta^\star \odot \beta^\star) + \left(\frac{1}{n} \sum_{i=1}^n x^{(i)} (x^{(i)})^\top \right) (\zeta_t \odot \zeta_t) \tag{9.51}$$

$$= \underbrace{X (r_t^2 - 1) \cdot (\beta^\star \odot \beta^\star)}_{\text{part of } u_t} + \underbrace{X (\zeta_t \odot \zeta_t)}_{v_t}. \tag{9.52}$$

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

$$\nabla \widehat{L}(\beta^t) = m_t \odot \beta^t = [(r_t^2 - 1)(\beta^\star \odot \beta^\star) - u_t + v_t] \odot \beta_t. \tag{9.53}$$

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 9.12, we can bound $\|u_t\|_\infty$ as

$$\|u_t\|_\infty \le 4\delta \left\| (r_t^2 - 1) \beta^\star \odot \beta^\star \right\|_2 \le 4\delta \|\beta^\star \odot \beta^\star\|_2 \le 4\delta. \tag{9.54}$$

(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 9.12 implies $\|v_t - \zeta_t \odot \zeta_t\|_\infty \le 4\delta \|\zeta_t \odot \zeta_t\|_2$,

$$\|v_t\|_\infty \le \|\zeta_t \odot \zeta_t\|_\infty + 4\delta \|\zeta_t \odot \zeta_t\|_2 \qquad \text{(by the triangle inequality)} \tag{9.55}$$

$$\le \|\zeta_t\|_\infty^2 + 4\delta \|\zeta_t \odot \zeta_t\|_2 \qquad \text{(since } \zeta_t \text{ very small)} \tag{9.56}$$

$$= \|\zeta_t\|_\infty^2 + 4\delta \|\zeta_t\|_2^2. \tag{9.57}$$

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 two subsections, we analyze the growth of ζ_t and r_t.

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

$$\|\zeta_t\|_\infty \le 2\alpha, \qquad \|\zeta_t\|_2^2 \le \frac{1}{2}, \qquad \text{and} \qquad \|\zeta_{t+1}\|_\infty \le \big(1 + O(\eta\delta)\big) \|\zeta_t\|_\infty. \tag{9.58}$$

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 (9.53) 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

$$\beta^{t+1} = \beta^t - \eta \big[(r_t^2 - 1)\,\beta^\star \odot \beta^\star - u_t + v_t\big] \odot \beta^t \tag{9.59}$$

$$= \underbrace{\beta^t - \eta\,(r_t^2 - 1)\,\beta^\star \odot \beta^\star \odot \beta^t}_{\text{GD update for population loss}} - \eta\,(-u_t + v_t) \odot \beta^t. \tag{9.60}$$

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,

$$\zeta_{t+1} = \big(I - e_1 e_1^\top\big)\,\beta^{t+1} \tag{9.61}$$

$$= \big(I - e_1 e_1^\top\big) \cdot \beta^t - \eta \big(I - e_1 e_1^\top\big)(v_t - u_t) \odot \beta^t. \qquad \text{(by (9.60), second term} = 0) \tag{9.62}$$

$$= \zeta_t - \eta \big[\big(I - e_1 e_1^T\big)(v_t - u_t) \odot \big(I - e_1 e_1^T\big)\beta^t\big] \qquad \text{(by distribution law for } \odot) \tag{9.63}$$

$$= \zeta_t - \eta \underbrace{\big[\big(I - e_1 e_1^T\big)(v_t - u_t)\big]}_{\rho_t} \odot \zeta_t. \tag{9.64}$$

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

$$\|\zeta_{t+1}\|_\infty \le (1 + \eta \|\rho_t\|_\infty) \|\zeta_t\|_\infty. \tag{9.65}$$

Now, we will prove the lemma by using strong induction on t. Suppose that the first two pieces of (9.58) hold for all iterations up to t. We can show that

$$\|\rho_t\|_\infty \le \|u_t\|_\infty + \|v_t\|_\infty \tag{9.66}$$

$$\le 4\delta + \|\zeta_t\|_\infty^2 + 4\delta \|\zeta_t\|_2^2 \qquad \text{(by (9.54) and (9.57))} \tag{9.67}$$

$$\le 4\delta + (2\alpha)^2 + 4\delta \cdot \frac{1}{2} \qquad \text{(by the inductive hypothesis)} \tag{9.68}$$

$$\le 8\delta, \tag{9.69}$$

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 (9.65), we have

$$\|\zeta_{t+1}\|_\infty \le (1 + 8\eta\delta) \|\zeta_t\|_\infty = \big(1 + O(\eta\delta)\big) \|\zeta_t\|_\infty, \tag{9.70}$$

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

$$\|\zeta_{t+1}\|_\infty \le (1+8\eta\delta)^{t+1}\|\zeta_0\|_\infty \le (1+8\eta\delta)^{1/(c\eta\delta)}\cdot\alpha \le 2\alpha \tag{9.71}$$

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

$$\|\zeta_{t+1}\|_2^2 \le (1+8\eta\delta)^{t+1}\|\zeta_0\|_2^2 \le (1+8\eta\delta)^{1/(c\eta\delta)}\cdot\alpha\sqrt{d} \le \frac{1}{2}, \tag{9.72}$$

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

□

(Step 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 9.17. *For all $t \le 1/(c\eta\delta)$ with sufficiently large constant c, we have that*

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

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 (9.53) is

$$\beta^{t+1} = \beta^t - \eta\big[(r_t^2-1)\beta^\star\odot\beta^\star - u_t + v_t\big]\odot\beta_t. \tag{9.73}$$

We have that

$$\begin{aligned} r_{t+1} &= \langle\beta^{t+1}, e_1\rangle && (9.74)\\ &= \langle\beta^t, e_1\rangle - \eta(r_t^2-1)\langle\beta^t, e_1\rangle - \eta\langle v_t-u_t, e_1\rangle\langle\beta^t, e_1\rangle && (9.75)\\ &= r_t - \eta(r_t^2-1)r_t - \eta\langle v_t-u_t, e_1\rangle r_t && (9.76)\\ &= \big(1+\eta(1-r_t^2)\big)r_t + \eta\langle u_t-v_t, e_1\rangle r_t && (9.77)\end{aligned}$$

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

$$\begin{aligned} |\eta\langle v_t-u_t, e_1\rangle r_t| &\le \eta\cdot r_t\,\|v_t-u_t\|_\infty && (9.78)\\ &\le \eta\cdot r_t\cdot 8\delta \qquad \text{(by (9.69))} && (9.79)\\ &= O(\eta\delta)\cdot r_t. && (9.80)\end{aligned}$$

□

Putting it all together Finally, we return to the proof of Theorem 9.14. By Lemma 9.17, we know that as long as $r_t \le 1/2$ it will grow exponentially fast, since

$$r_{t+1} \ge \Big(1+\eta(1-r_t^2)-O(\eta\delta)\Big)\cdot r_t \ge \Big(1+\frac{\eta}{2}\Big)\cdot r_t. \tag{9.81}$$

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

$$1 - r_{t+1} \le 1 - r_t - \eta(1 - r_t^2)(r_t) + O(\eta\delta) \cdot (r_t) \tag{9.82}$$

$$\le 1 - r_t - \frac{\eta(1 - r_t^2)}{2} + O(\eta\delta) \tag{9.83}$$

$$\le 1 - r_t - \frac{\eta(1 - r_t)}{2} + O(\eta\delta) \tag{9.84}$$

$$= \left(1 - \frac{\eta}{2}\right)(1 - r_t) + O(\eta\delta). \tag{9.85}$$

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

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

$$r_{t+1} - 1 \le (1 - \eta)(r_t - 1) + O(\eta\delta) \le O(\delta), \tag{9.86}$$

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

This completes the proof of Theorem 9.14, bounding the number of iterations needed for gradient descent on the empirical loss to converge to β^*. □

9.3 From small to large initialization: a precise characterization

We have previously discussed how certain initializations of gradient descent converge to minimum-norm solutions. In the sequel, we characterize the "effect of initialization" more precisely—we will show that in a variant of the settings in Section 9.2, we can precisely compute the corresponding regularizer (induced by any initialization). We will see that when the initialization is small, we obtain the bias towards minimum norm solution (in the parameter space used in optimization), whereas when the initialization is large, we are in the NTK regime (Section 8.4) where the implicit bias is towards the min norm solution under the NTK kernel. The materials in this subsection are simplifications of results in the recent paper Woodworth et al. [2020].

9.3.1 Preparation: gradient flow

To simplify the analysis, we will consider the concept of gradient flow, i.e. gradient descent with an infinitesimal learning rate. This allows us omit the "second order effect" from the learning rate and simplify the analysis.

We begin by recalling the gradient descent update formula. In our previous description of gradient descent, we indexed the updated parameters by $t = 1, 2, \ldots$. Anticipating our generalization to infinitesimal steps, we will index the updated parameters using parentheses (instead of subscripts). In particular, the standard gradient descent update given a loss function $L(w)$ is

$$w(t+1) = w(t) - \eta \nabla L(w(t)). \tag{9.87}$$

If we scale the time by η so that each update (by gradient descent) corresponds to a time step of size η (rather than size 1), the update becomes

$$w(t+\eta) = w(t) - \eta \nabla L(w(t)). \tag{9.88}$$

Taking $\eta \to 0$ yields a differential equation, which can be thought of as a continuous process rather than discrete updates:

$$w(t+dt) = w(t) - dt \cdot \nabla L(w(t)). \tag{9.89}$$

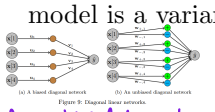
This can also be written as:

$$\dot{w}(t) = -\nabla L(w(t)) \quad \text{with} \quad \dot{w}(t) = \frac{\partial w(t)}{\partial t} \tag{9.90}$$

This allows us to ignore the η^2 term (alternatively the (dt^2) term), which will simplify some of the technical details that follow.

9.3.2 Characterizing the implicit bias of initialization

The results in this section are slight simplification of the recent paper by Woodworth et al. [2020]. The model is a variant of the one we introduced in (9.7). Recalling that $x^{\odot 2} = x \odot x$, let



$$f_w(x) = \left(w_+^{\odot 2} - w_-^{\odot 2}\right)^\top x. \tag{9.91}$$

where $w_+, w_- \in \mathbb{R}^d$. Let w denote the concatenation of the two parameter vectors, i.e. $= (w_+, w_-)$. In (9.7), we defined $f_\beta(x) = (\beta \odot \beta)^\top x$; this model can only represent positive linear combinations of x. By contrast, $f_w(x)$ can represent any linear model. Moreover, if we choose our initialization for w such that $w_+(0) = w_-(0)$, we obtain $f_{w(0)}(x) \equiv 0$ for all x. Similar to our analysis of the NTK, this initialization will simplify the subsequent derivations.

Next, we define the following loss function,

$$\widehat{L}(w) = \frac{1}{2}\sum_{i=1}^{n}\left(y^{(i)} - f_w(x^{(i)})\right)^2, \tag{9.92}$$

and consider the initialization

$$w_+(0) = w_-(0) = \alpha \cdot \vec{\mathbf{1}} \tag{9.93}$$

where $\vec{\mathbf{1}}$ denotes the all-ones vector. The analysis technique still applies to any general initializations as long as all the dimension are initialized to be non-zero, but the the initialization scale is the most important factor, and therefore we chose this simplification for the ease of exposition.

Note that every $w = (w_+, w_-)$ corresponds to a de facto linear function of x. We denote the resulting linear model as θ_w:

$$\theta_w = w_+^{\odot 2} - w_-^{\odot 2}. \tag{9.94}$$

Note that $\theta_w^\top x = f_w(x)$.

Let $w(\infty)$ denote the limit of the gradient flow, i.e.

$$w(\infty) = \lim_{t\to\infty} w(t). \tag{9.95}$$

Then, the converged linear model in the θ space is defined by $\theta_\alpha(\infty) = \theta_{w(\infty)}$—we are interested in understanding its properties. For simplicity, we will omit the ∞ index and refer to this quantity as θ_α. We assume throughout that the limit exists and all corresponding regularity conditions are met.

Let

$$X = \begin{bmatrix} x^{(1)\top} \\ \vdots \\ x^{(n)\top} \end{bmatrix} \in \mathbb{R}^{n\times d} \quad \text{and} \quad \vec{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}. \tag{9.96}$$

In the sequel, we formally state our result relating the complexity of the solution discovered by gradient flow to the size of the initialization.

Theorem 9.18 (Theorem 1 in Woodworth et al. [2020]). *For any $0 < \alpha < \infty$, assume that gradient flow with initialization $w_+(0) = w_-(0) = \alpha \cdot \vec{1}$ converges to a solution that fits the data exactly: $X\theta_\alpha = \vec{y}$.*[3] *Then, the solution satisfies the following notion of minimum complexity:*

$$\theta_\alpha = \operatorname{argmin}_\theta Q_\alpha(\theta) \tag{9.97}$$

$$\text{s.t.} \quad X\theta = \vec{y} \tag{9.98}$$

where

$$Q_\alpha(\theta) = \alpha^2 \cdot \sum_{i=1}^{n} q\left(\frac{\theta_i}{\alpha^2}\right) \tag{9.99}$$

and

$$q(z) = 2 - \sqrt{4 + z^2} + z \cdot \operatorname{arcsinh}\left(\frac{z}{2}\right) \tag{9.100}$$

In words, Theorem 9.18 claims that θ_α is the minimum complexity solution for the complexity measure Q_α.

Remark 9.19. In particular, when $\alpha \to \infty$ we have that

$$q(\theta_i/\alpha^2) \asymp \theta_i^2/\alpha^4 \tag{9.101}$$

and so

$$Q_\alpha(\theta) \asymp \frac{1}{\alpha^2} \|\theta\|_2^2 . \tag{9.102}$$

This means that if $\alpha \to \infty$ than the complexity measure Q_α is the ℓ_2-norm, $||\theta||_2$. If $\alpha \to 0$, then the complexity measure becomes

$$q\left(\frac{\theta_i}{\alpha^2}\right) \asymp \frac{|\theta_i|}{\alpha^2} \log\left(\frac{1}{\alpha^2}\right) \quad \text{(by Taylor expansion)} \tag{9.103}$$

and so,

$$Q_\alpha(\theta) \asymp \frac{\|\theta\|_1}{\alpha^2} \log\left(\frac{1}{\alpha^2}\right) \tag{9.104}$$

To summarize, for $\alpha \to \infty$, the constrained minimization problem we solve in (9.98) yields the minimum ℓ_2-norm solution of θ (i.e. the ℓ_4-norm for w). When $\alpha \to 0$, solving (9.98) yields the minimum ℓ_1-norm θ (which is the ℓ_2-norm for w). For $0 < \alpha < \infty$, we obtain some interpolation (of ℓ_1 and ℓ_2 regularization) of the optimum.

Remark 9.20. Note that when $\alpha \to 0$, the intuition is similar to what we had observed in previous analyses; in particular, the solution is the global minimum closest to the initialization. Note however, that when $\alpha \neq 0$, the solution discovered by gradient descent will not *exactly* correspond to the solution closest to the initialization.

Remark 9.21. When $\alpha \to \infty$, we claim that the model optimization is in the neural tangent kernel (NTK) regime. Recall that we had two parameters, (σ, β), that determined if we could treat the optimization problem as a kernel regression. Further recall that σ denotes the minimum singular value of Φ and β is the Lipschitzness of the gradient. Let us now compute σ and β for large α initializations of our model.

[3] This assumption can likely be proved to be true and thus not required. Here we still include the condition because the original paper Woodworth et al. [2020] assumed it.

For $w_-(0) = w_+(0) = \alpha\vec{1}$,

$$\nabla f_{w(0)}(x) = 2\begin{bmatrix} w_+(0)\cdot x \\ -w_-(0)\odot x \end{bmatrix} = 2\alpha\begin{bmatrix} x \\ -x \end{bmatrix} \tag{9.105}$$

by the chain rule. It is clear then that both σ and β linearly depend on α. This implies that

$$\frac{\beta}{\sigma^2} \to 0 \quad \text{as } \alpha \to \infty \tag{9.106}$$

since the denominator is $O(\alpha^2)$, while the numerator is $O(\alpha)$. In particular, the features used in this kernel method are:

$$\phi(x) = \nabla f_{w(0)}(x) = 2\alpha\begin{bmatrix} x \\ -x \end{bmatrix} \tag{9.107}$$

The neural tangent kernel perspective then gives an alternative proof of this complexity minimization result for $\alpha \to \infty$. In the NTK regime, the solution (to our convex problem) is always the minimum ℓ_2-norm solution for the feature matrix, which in this case equals $\begin{bmatrix} X \\ -X \end{bmatrix}$.

Note that practice tends not to follow the assumptions made here. Often, people either do not use large initializations or do not use infinitesimally small step sizes. But this is a good thing because we do not want to be in the NTK regime; being in the NTK regime implies that we are doing no different or better than just using a kernel method.

We can now prove Theorem 9.18, which is similar to the overparametrized linear regression proof of Theorem 9.3.

This proof follows in two steps:

1. We find an invariance maintained by the optimizer. In the overparametrized linear regression proof of Theorem 9.3, we required $\theta \in \text{span}\{x^{(i)}\}$. For this proof, we will use a slightly more complicated invariance.
2. We characterize the solution using this invariance. The invariance, which depends on α, will tell us which zero error solution the optimization converges to.

Note also that all of these conditions only depend upon the empirically observed samples. The invariance and minimum is not defined with respect to any population quantities.

Proof. Let

$$\tilde{X} = \begin{bmatrix} X & -X \end{bmatrix} \in \mathbb{R}^{n\times 2d} \quad \text{and} \quad w(t) = \begin{bmatrix} w_+(t) \\ w_-(t) \end{bmatrix} \in \mathbb{R}^{2d}. \tag{9.108}$$

Then, the model output on n data points can be described in matrix notation as follows:

$$\left(\tilde{X}w(t)^{\odot 2}\right) = \left(\begin{bmatrix} X & -X \end{bmatrix}\begin{bmatrix} w_+(t)^{\odot 2} \\ w_-(t)^{\odot 2} \end{bmatrix}\right) = \begin{bmatrix} f_{w(t)}(x^{(1)}) \\ \vdots \\ f_{w(t)}(x^{(n)}) \end{bmatrix} \in \mathbb{R}^n. \tag{9.109}$$

Given the loss function,

$$L(w(t)) = \frac{1}{2}\left\|\tilde{X}w(t)^{\odot 2} - \vec{y}\right\|_2^2, \tag{9.110}$$

the gradient of $w(t)$ can be computed as

$$\dot{w}(t) = -\nabla L(w(t)) \tag{9.111}$$

$$= -\nabla \left(\left\| \tilde{X} w(t)^{\odot 2} - \vec{y} \right\|_2^2 \right) \tag{9.112}$$

$$= \left(2\tilde{X}^\top r(t)\right) \odot w(t) \qquad \text{(chain rule)} \tag{9.113}$$

where $r(t) = \tilde{X} w(t)^{\odot 2} - \vec{y}$ denotes the residual vector. We see that the $\tilde{X}^\top r(t)$ term in (9.113) is reminiscent of linear regression for which it would correspond to the gradient, although the $\odot w(t)$ reminds us that this problem is indeed quadratic.

We cannot directly solve this differential equation, but we claim that

$$w(t) = w(0) \odot \exp\left(-2\tilde{X}^\top \int_0^t r(s) ds \right) \qquad \text{(exp is applied entry-wise)} \tag{9.114}$$

which is not quite a closed form solution of equation 9.113 since $r(s)$ is still a function of $w(t)$. To understand how we obtained this "solution," we consider a more abstract setting. Suppose that

$$\dot{u}(t) = v(t) u(t) \tag{9.115}$$

We can then "solve" this differential equation as follows. Rearranging, we observe that

$$\frac{\dot{u}(t)}{u(t)} = v(t) \tag{9.116}$$

$$\frac{d \log u(t)}{dt} = v(t) \quad \text{(chain rule)} \tag{9.117}$$

$$\log u(t) - \log u(0) = \int_0^t v(s) ds \quad \text{(integration)} \tag{9.118}$$

$$\frac{u(t)}{u(0)} = \exp\left(\int_0^t v(s) ds \right) \tag{9.119}$$

In our problem, $u \leftrightarrow w_i$ and $v \leftrightarrow (\tilde{X}^\top r(t))_i$.

We have characterized w, but we want to transform this to a characterization that involves θ. Recall that $w_+(0) = \alpha \vec{1}$ and $w_-(0) = \alpha \vec{1}$ so that $w(0) = \alpha \vec{1} \in \mathbb{R}^{2d}$. Additionally, we have that $\theta(t) = w_+(t)^{\odot 2} - w_-(t)^{\odot 2}$. We can now apply (9.114) to expand $w(t)$ and simplify.

Note that if we have $\tilde{X}^\top = \begin{bmatrix} X^\top \\ -X^\top \end{bmatrix} \in \mathbb{R}^{2n \times d}$, then for some vector v,

$$\left(\exp(-2\tilde{x}^\top v)\right)^{\odot 2} = \begin{bmatrix} \exp(-2X^\top v) \\ \exp(2X^\top v) \end{bmatrix}^{\odot 2} \tag{9.120}$$

$$= \begin{bmatrix} \exp(-4X^\top v) \\ \exp(4X^\top v) \end{bmatrix}. \tag{9.121}$$

Applying this result for $v = \int_0^t r(s) ds$, we obtain that:

$$\theta(t) = w_+(t)^{\odot 2} - w_-(t)^{\odot 2} \tag{9.122}$$

$$= \alpha^2 \left[\exp\left(-4X^\top \int_0^t r(s) ds \right) - \exp\left(4X^\top \int_0^t r(s) ds \right) \right] \tag{9.123}$$

$$= 2\alpha^2 \sinh\left(-4X^\top \int_0^t r(s) ds \right). \tag{9.124}$$

Letting $t \to \infty$, we have that

$$\theta_\alpha = 2\alpha^2 \sinh\left(-4X^\top \int_0^\infty r(s)ds\right). \tag{9.125}$$

Lastly, we also know

$$X\theta_\alpha = \vec{y} \tag{9.126}$$

since this is the assumption by the theorem (which should can be proven because the optimization should converge to a zero-error solution). We next show that (9.125) and (9.126) are also sufficient conditions for a solution to the constrained optimization problem given by (9.98). In particular, (9.125) and (9.126) correspond to the Karush-Kuhn-Tucker (or KKT) conditions of (9.98).

A KKT condition is an optimality condition for constrained optimization problems. While these conditions can have a variety of formulations and typically one can invoke some off-the-shelf theorems to use them, we can motivate the conditions we encountered by considering the following general optimization program:

$$\begin{aligned} &\text{argmin} && Q(\theta) && (9.127)\\ &\text{s.t.} && X\theta = \vec{y}. && (9.128) \end{aligned}$$

We say that θ satisfies the (first order) KKT conditions if

$$\begin{cases} \nabla Q(\theta) = X^\top \nu \text{ for some } \nu \in \mathbb{R}^n & (9.129)\\ X\theta = \vec{y} & (9.130) \end{cases}$$

More intuitively, we know that optimality implies that there are no first order local improvements that satisfy the constraint (up to first order). Then, consider a perturbation $\Delta\theta$. In order to satisfy the constraint, we must enforce the following:

$$\Delta\theta \perp \text{row-span}\{X\} \quad \text{so} \quad X\Delta\theta = 0 \tag{9.131}$$

So, if we look at $\theta + \Delta\theta$ satisfying the constraint, we can use a Taylor expansion to show that

$$Q(\theta + \Delta\theta) = Q(\theta) + \langle \Delta\theta, \nabla Q(\theta)\rangle \le Q(\theta) \tag{9.132}$$

because if $\langle \Delta\theta, \nabla Q(\theta)\rangle$ is positive it violates the optimality assumption. In fact, it is very easy to make the sign flip for $\langle \Delta\theta, \nabla Q(\theta)\rangle$ because you can flip $\Delta\theta$ to be the opposite direction. This means that

$$\forall\, \Delta\theta \perp \text{row-span}\{X\}, \quad \langle \Delta\theta, \nabla Q(\theta)\rangle = 0 \tag{9.133}$$

because if it is negative, you can equivalently flip it to be positive which violates optimality. This means that $Q(\theta) \subseteq \text{row-span}\{X\}$, or $Q(\theta) = X^\top \nu$ for some ν.

Returning to our problem, the KKT condition gives

$$\nabla Q(\theta) = X^\top \nu \tag{9.134}$$

and the invariance gives us

$$\theta_\alpha = 2\alpha^2 \sinh\left(-4X^\top \int_0^\infty r(s)ds\right) \quad (9.125) \tag{9.135}$$

$$= 2\alpha^2 \sinh\left(-4X^\top v'\right) \tag{9.136}$$

where we let $v' = \int_0^\infty r(s)ds$ for simplicity. Taking the gradient of Q gives

$$\nabla Q_\alpha(\theta) = \text{arcsinh}\left(\frac{1}{2\alpha^2}\theta\right) \tag{9.137}$$

Plugging in θ_α, we get

$$\nabla Q(\theta_\alpha) = \text{arcsinh}\left(\frac{1}{2\alpha^2}\theta_\alpha\right) = -4X^\top v' \tag{9.138}$$

Thus, θ_α satisfies both KKT conditions. Even further, since our optimization problem (9.98) is convex (we do not formally argue this), we conclude that θ_α is a global minimum. □

9.4 Implicit regularization towards max-margin solutions in classification

We now switch our focus to classification problems. We consider linear models (though these results also apply to nonlinear models with a weaker version of the conclusion). We assume that our data is separable and will prove that gradient descent converges to the max-margin solution. This result holds for any initialization and does not require any additional regularization; we only require the use of gradient descent and the standard logistic loss function. The results in this subsection are originally given by Soudry et al. [2018], and our exposition heavily depends on those in [Ji and Telgarsky, 2018, Telgarsky, 2021].

Assume we have data $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$, where $x^{(i)} \in \mathbb{R}^d$ and $y^{(i)} \in \{\pm 1\}$. We consider the linear model $h_w(x) = w^\top x$ and the cross entropy loss function $\widehat{L}(w) = \sum_{i=1}^n \ell\left(y^{(i)}, h_w\left(x^{(i)}\right)\right)$, where $\ell(t) = \log(1 + \exp(-t))$ is the logistic loss.

As we have separable data, there can be multiple global minima, as you can trivially take an infinite number of separators. (More formally, there are an infinite number of unit vectors $\bar{w}$ such that $\bar{w}^\top x^{(i)} y^{(i)} > 0$ for all i) as one can perturb any strict separator while still maintaining a separation of classes.) Then, we can scale the separator to make the loss arbitrarily small—we have that $\widehat{L}(\alpha \bar{w}) \to 0$ as $\alpha \to \infty$. Thus, informally, for any unit vector $\bar{w}$ that separate the data, $\infty \cdot \bar{w}$ is a global minimum.

We would like to understand which global minimum gradient descent converges to. We will now show that it finds the max-margin solution. Before we can do so, we recall/introduce the following definitions.

Definition 9.22 (Margin)**.** Let $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ be given data. Assuming $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ is linearly separable, the *margin* is defined as

$$\min_{i \in [n]} y^{(i)} w^\top x^{(i)} \tag{9.139}$$

Definition 9.23 (Normalized Margin)**.** Let $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ be given data. Assuming $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ is linearly separable, the *normalized margin* is defined as

$$\gamma(w) = \frac{\min_{i \in [n]} y^{(i)} w^\top x^{(i)}}{\|w\|_2} \tag{9.140}$$

Definition 9.24 (Max-Margin Solution)**.** Using the normalized margin γ defined in Definition 9.23, we define a *max-margin solution* as

$$\bar{\gamma} = \max_w \gamma(w) \tag{9.141}$$

and let w^* be the unit-norm maximizer. [4]

Using these definitions, we claim the following result.

Theorem 9.25. *Gradient flow converges to the direction of max-margin solution in the sense that*

$$\gamma(w(t)) \to \bar{\gamma} \text{ as } t \to \infty \tag{9.142}$$

where $w(t)$ is the iterate at time t.

The following observations provide some intuition for Theorem 9.25.

1. $\widehat{L}(w(t)) \to 0$ by a standard optimization argument. Namely, if the objective is monotone decreasing at each iteration, $\widehat{L}(w(t)) \approx 0$ for large enough t.
2. Using a Taylor expansion, we can show that $\ell(z) = \log(1 + \exp(-z)) \approx \exp(-z)$ for large z. Thus, logistic loss is close to exponential loss when z is very large.

[4] The normalized margin $\bar{\gamma}$ is scale-invariant. For $c \neq 0$, $\gamma(cw) = \min_{i \in [n]} \frac{y^{(i)} c w^\top x^{(i)}}{\|cw\|_2} = \min_{i \in [n]} \frac{y^{(i)} w^\top x^{(i)}}{\|w\|_2} = \gamma(w)$.

3. Using observation 1, we see that $\|w(t)\|_2 \to \infty$ because if $\|w(t)\|_2$ were instead bounded, then the loss $\widehat{L}(w(t))$ will be bounded below by a constant that is strictly greater than zero, contradicting observation 1. Formally, if $\|w(t)\|_2 \le B$, then

$$|y^{(i)} w^\top x^{(i)}| \le B\|x^{(i)}\|, \tag{9.143}$$

and therefore we get

$$\widehat{L}(w(t)) \ge \sum_{i=1}^{n} \exp\left(-B\|x^{(i)}\|_2\right) > 0. \tag{9.144}$$

4. Suppose we have w such that $\|w\|_2 = q$ is very big. Then, using observation 2, we see that

$$\widehat{L}(w) = \sum_{i=1}^{n} \ell(y^{(i)} w^\top x^{(i)}) \tag{9.145}$$

$$\approx \sum_{i=1}^{n} \exp\left(-y^{(i)} w^\top x^{(i)}\right) \tag{9.146}$$

$$\log \widehat{L}(w) \approx \log \sum_{i=1}^{n} \exp\left(-y^{(i)} w^\top x^{(i)}\right) \tag{9.147}$$

$$= \log \sum_{i=1}^{n} \exp\left(-q y^{(i)} \bar{w}^\top x^{(i)}\right) \tag{9.148}$$

$$\approx \max_{i\in[n]} -q y^{(i)} \bar{w}^\top x^{(i)} \tag{9.149}$$

where $\bar{w} = \frac{w}{\|w\|_2}$ and the last step holds because the log of a sum of exponentials (*log-sum-exp*) is a smooth approximation to the maximum function. To motivate this claim, observe that:

$$\log \sum_{i=1}^{n} \exp(q u_i) \ge q \max_i u_i \tag{9.150}$$

$$\log \sum_{i=1}^{n} \exp(q u_i) \le \log\left(n \exp(q \max_i u_i)\right) \tag{9.151}$$

$$= \log n + q \max_i u_i \tag{9.152}$$

$$\approx q \max_{i\in[n]} u_i + o(q) \text{ as } q \to \infty \tag{9.153}$$

Thus, minimizing the loss is the same as

$$\min_w \left(\max_{i\in[n]} -q y^{(i)} \bar{w}^\top x^{(i)}\right) \tag{9.154}$$

which can be reformulated as

$$\max_w \left(\min_{i\in[n]} q y^{(i)} \bar{w}^\top x^{(i)}\right) \tag{9.155}$$

The above observations heuristically demonstrate that minimizing the logistic loss (with gradient descent) is equivalent (in the limit) to maximizing the margin. Below, we prove Theorem 9.25 rigorously for the exponential loss function $\ell(t) = \exp(-t)$, which is nearly the same as the logistic loss.

Proof of Theorem 9.25. We begin by defining the smooth margin as

$$\tilde{\gamma}(w) \triangleq \frac{-\log \hat{L}(w)}{\|w\|_2} \tag{9.156}$$

$$= \frac{-\log\left(\sum_{i=1}^n \exp(-y^{(i)} w^\top x^{(i)})\right)}{\|w\|_2}. \tag{9.157}$$

Note that $\tilde{\gamma}(w)$ approximates $\gamma(w)$ by the log-sum-exp approximation. To make this precise, recall that $\gamma(w) \geq \tilde{\gamma}(w)$ because $y^{(i)} w^\top x^{(i)} \geq \gamma(w)\|w\|_2$ for all i.

Then, since $\gamma(w) \leq \bar{\gamma}$ by definition, it suffices to show that

$$\lim_{t\to\infty} \tilde{\gamma}(w(t)) = \bar{\gamma}. \tag{9.158}$$

Let $\dot{w}(t) = -\nabla \hat{L}(w(t))$. Then,

$$\frac{\partial}{\partial t}\left(-\log \hat{L}(w(t))\right) = \left\langle \nabla\left(-\log \hat{L}(w(t))\right), \dot{w}(t)\right\rangle \tag{9.159}$$

$$= \left\langle -\frac{\nabla \hat{L}(w(t))}{\hat{L}(w(t))}, \dot{w}(t)\right\rangle \tag{9.160}$$

$$= \frac{\|\nabla \hat{L}(w(t))\|_2^2}{\hat{L}(w(t))} \tag{9.161}$$

$$= \frac{\|\dot{w}(t)\|_2^2}{\hat{L}(w(t))} \geq 0 \tag{9.162}$$

This result tells us that the log loss is decreasing at each infinitesimal step of the gradient flow. By integrating (9.162), we can also evaluate the log loss at time T:

$$-\log \hat{L}(w(T)) = -\log \hat{L}(w(0)) + \int_0^T \frac{\partial}{\partial t}\left(-\log \hat{L}(w(t))\right) dt \tag{9.163}$$

$$= -\log \hat{L}(w(0)) + \int_0^T \frac{\|\dot{w}(t)\|_2^2}{\hat{L}(w(t))} dt. \tag{9.164}$$

While the derivation above tells us how the numerator of (9.156) is changing, we have yet to relate this to the denominator, i.e. the norm of w. Recall that w^* is the direction of the max-margin solution. Then, we have

$$\|\dot{w}(t)\|_2 \geq \langle \dot{w}(t), w^*\rangle \qquad \text{(Cauchy-Schwarz)} \tag{9.165}$$

$$= \left\langle -\nabla \hat{L}(w(t)), w^*\right\rangle \tag{9.166}$$

$$= \left\langle \sum_{i=1}^n y^{(i)} \exp(-y^{(i)} w^\top x^{(i)}) \cdot x^{(i)}, w^*\right\rangle \tag{9.167}$$

$$= \sum_{i=1}^n y^{(i)} \exp(-y^{(i)} w^\top x^{(i)}) \cdot \left\langle w^*, x^{(i)}\right\rangle \tag{9.168}$$

$$\geq \bar{\gamma} \sum_{i=1}^n \exp(-y^{(i)} w^\top x^{(i)}) \tag{9.169}$$

$$= \bar{\gamma} \cdot \hat{L}(w(t)). \tag{9.170}$$

This shows that $\dot{w}(t)$ is correlated to w^*, and that this correlation depends on $\bar{\gamma}$ and the loss. In addition, $\dot{w}(t)$ is not too small compared to the loss.

Next, we substitute (9.165) into the second term of the right-hand-side of (9.163):

$$\int_0^T \frac{\|\dot{w}(t)\|_2^2}{\hat{L}(w(t))} dt \geq \bar{\gamma} \cdot \int_0^T \|\dot{w}(t)\|_2 dt \tag{9.171}$$

$$\geq \bar{\gamma} \cdot \left\| \int_0^T \dot{w}(t) dt \right\|_2 \tag{9.172}$$

$$= \bar{\gamma} \|w(T)\|_2. \tag{9.173}$$

Applying this bound to the RHS of (9.163), we obtain

$$-\log \hat{L}(w(T)) \geq -\log \hat{L}(w(0)) + \bar{\gamma}\|w(T)\|_2. \tag{9.174}$$

Dividing both sides by $\|w(T)\|_2$,

$$-\frac{\log \hat{L}(w(T))}{\|w(T)\|_2} \geq -\frac{\log \hat{L}(w(0))}{\|w(T)\|_2} + \bar{\gamma}. \tag{9.175}$$

Since $\lim_{T\to\infty} \|w(T)\|_2 = \infty$, we know that the first term on the RHS of (9.175) goes to 0 in the limit. Thus,

$$\lim_{T\to\infty} \left(-\frac{\log \hat{L}(w(T))}{\|w(T)\|_2} \right) \geq \bar{\gamma}. \tag{9.176}$$

Recognizing the LHS as the definition of the smooth margin, i.e. (9.156), we conclude that

$$\lim_{T\to\infty} \tilde{\gamma}(w(T)) \geq \bar{\gamma}. \tag{9.177}$$

Meanwhile, since we also know that

$$\bar{\gamma} \geq \gamma(w(T)) \geq \tilde{\gamma}(w(T)), \tag{9.178}$$

we conclude by the squeeze theorem that

$$\lim_{T\to\infty} \gamma(w(T)) = \lim_{T\to\infty} \tilde{\gamma}(w(T)) = \bar{\gamma}. \tag{9.179}$$

□

9.5 Implicit regularization effect of noise in SGD

In the previous section, we discussed implicit regularization via initialization and the implicit regularization of gradient descent for logistic loss-minimizing classifiers. In the sequel, we will move forward to the implicit regularization effect of SGD noise. Starting from the quadratic case, we analyze how the SGD noise will affect the optimization solution, and present (heuristically) a result for non-quadratic loss functions. In particular, the main (heuristic) results are:

1. On the one dimensional quadratic function, the iterate can be disentangled into a contraction part and a stochastic part, the latter of which is characterized by the Ornstein–Uhlenbeck (OU) process. The noise makes the iterate bounce around the global minimum.

2. On the multi-dimensional quadratic function, the iterate can be disentangled into multiple separate 1-D OU processes. The noise makes the iterate bounce around the global minimum, while the fluctuation is closely related to the shape of the noise.

3. On non-quadratic functions, SGD with *label noise* on empirical loss $\hat{L}(\theta)$ converges to a stationary point of the regularized loss $\hat{L}(\theta) + \lambda \mathrm{tr}(\nabla^2 \hat{L}(\theta))$, which is mainly due to the accumulation of a third order effect.