CPSC 532D — 17. IMPLICIT REGULARIZATION

Danica J. Sutherland

University of British Columbia, Vancouver

Fall 2023

We've seen some examples so far of settings where there's more than one empirical risk minimizer; this often happens with *interpolation*, when you can achieve $L_S(h) = 0$ in more than one way, some of which are awful, but $\mathcal{A}$ often picks decent ones. In particular, we saw some explicit examples with polynomial regression.

One way to choose between ERMs (or near-ERMs) is regularized loss minimization, where we prefer solutions with e.g. a small norm. But often we don't do that, and we just run gradient descent to minimize $L_S(h)$. Doing this doesn't just get us *any* arbitrary ERM; it gets us a particular one, decided on by our choice of algorithm. The idea that our optimization algorithm or other such "implementation details" can actually choose for us which of the "equally valid solutions" we end up with is called the *implicit regularization* of the algorithm: we don't explicitly write down a regularizer, but the choice of algorithm has a similar effect.

It's also sometimes called the implicit bias *of the algorithm, in the sense that the algorithm has a certain inductive bias towards certain kinds of solutions. That can sometimes cause confusion with the concept of the same name from social science, though, and just generally kind of imply that it's "bad" when actually often the presence of this implicit regularization is "good."*

In our discussion of neural tangent kernels, we mentioned that we could solve the ODE for gradient flow to say *which* ERM we end up at. We didn't prove this, though, and it only applied to "kernel gradient flow" which is not really the algorithm we usually use. What happens for actual problems, with finite learning rates?

1 GRADIENT DESCENT FOR LINEAR REGRESSION

We're going to optimize the function

$$f(w) = L_S^{\mathrm{sq}}(x \mapsto w \cdot x) = \frac{1}{m}\left\|Xw - y\right\|^2,$$

where $X \in \mathbb{R}^{m \times d}$ is the matrix stacking up S_x and $y \in \mathbb{R}^m$ is the vector form of S_y.

It's possible to use this form to handle kernels, too. If there's a finite-dimensional embedding ϕ, we could just collect $\phi(x_i)$ in rows of X and find w. If we instead write $f_\alpha(x) = \sum_i \alpha_i k(x_i, x)$ and do gradient descent on α, notice the training set loss becomes $L_S(f_\alpha) = \frac{1}{m}\left\|K\alpha - y\right\|^2)$ and so the rest of the analysis will apply with $X = K$ – which will potentially give a *different* solution than the kernel gradient descent version. Implicit regularization is highly algorithm-specific.

This agrees with "kernel gradient descent" as in our NTK discussion for finite-dimensional kernels.

In any case, we have

$$\nabla f(w) = \frac{2}{m} X^{\mathsf{T}}(Xw - y),$$

which notice is $\frac{2}{m}\left\|X^{\mathsf{T}}X\right\|$-smooth, so f is convex and β-smooth, thus small-learning-rate gradient descent finds a global optimum. In the traditional $m > d$ case when X is full-rank, there's a unique solution to this problem, typically with $Xw \neq y$ but always having $X^{\mathsf{T}}(Xw - y) = 0$. In high-dimensional settings $d > m$, though, it's

For more, visit `https://cs.ubc.ca/~dsuth/532D/23w1/`.

possible to achieve $Xw = y$ (interpolation) in infinitely many ways. Which one does gradient descent find?

1.1 *Characterizing the solution*

We'll run constant-learning-rate gradient descent (not SGD, not projected) starting from w_1. To avoid writing $2/m$ everywhere, let's absorb it into the learning rate:

$$w_{t+1} = w_t - \eta_0 \nabla f(w_t) = w_t - \frac{2\eta_0}{m} X^\mathsf{T}(Xw_t - y) = w_t - \eta X^\mathsf{T}(Xw_t - y),$$

where our "real" learning rate is η_0 and $\eta = 2\eta_0/m$. Then, unrolling this iteration, the $(t+1)$st iterate is

$$\begin{aligned} w_{t+1} &= (I - \eta X^\mathsf{T} X) w_t + \eta X^\mathsf{T} y \\ &= (I - \eta X^\mathsf{T} X)^2 w_{t-1} + (I - \eta X^\mathsf{T} X)\eta X^\mathsf{T} y + \eta X^\mathsf{T} y \\ &= (I - \eta X^\mathsf{T} X)^3 w_{t-2} + (I - \eta X^\mathsf{T} X)^2 \eta X^\mathsf{T} y + (I - \eta X^\mathsf{T} X)\eta X^\mathsf{T} y + \eta X^\mathsf{T} y \\ &= (I - \eta X^\mathsf{T} X)^t w_1 + \eta \sum_{k=0}^{t-1} (I - \eta X^\mathsf{T} X)^k X^\mathsf{T} y. \end{aligned} \tag{1}$$

The SVD is the single most useful tool you probably didn't really internalize from undergrad linear algebra (at least, I didn't). It's worth getting used to; it comes up all the time.

To analyze this, we'll use the singular value decomposition (SVD), specifically what Wikipedia calls the "compact SVD." We're going to decompose $X = U\Sigma V^\mathsf{T}$: if X is $m \times d$ of rank $r \le \min(m, d)$, then Σ is an $r \times r$ diagonal matrix with positive entries on the diagonal, U is $m \times r$, and V is $d \times r$. This can also be written $X = \sum_{i=1}^{r} \sigma_i U_{:,i} V_{:,i}^\mathsf{T}$, where $\sigma_i = \Sigma_{ii}$ is the ith singular value (typically sorted in descending order); there we've decomposed X into a sum of rank-one matrices $U_{:,i} V_{:,i}^\mathsf{T}$. Importantly, we have that $U^\mathsf{T} U = I_r = V^\mathsf{T} V$, i.e. the columns of U are r orthonormal vectors in $\mathbb{R}^m$, and the columns of V are r orthonormal vectors in $\mathbb{R}^d$. Thus, for instance, $XX^\mathsf{T} = U\Sigma V^\mathsf{T} V \Sigma U^\mathsf{T} = U\Sigma^2 U^\mathsf{T}$, and similarly $X^\mathsf{T} X = V\Sigma^2 V^\mathsf{T}$. We also have that the operator norm of X is $\|X\| = \sigma_1$, the largest singular value.

These are valid (truncated) eigendecompositions for XX^T and $X^\mathsf{T}X$, showing that the nonzero singular values are the square roots of the nonzero eigenvalues of $X^\mathsf{T}X$ or XX^T (which are the same), and that the corresponding left/right singular vectors are the corresponding eigenvectors of XX^T / $X^\mathsf{T}X$.

In this compact SVD, VV^T is a $d \times d$ matrix of rank r, and in fact it's the matrix of an *orthogonal projection* since $(VV^\mathsf{T})(VV^\mathsf{T}) = V(V^\mathsf{T} V)V^\mathsf{T} = VV^\mathsf{T}$ and $(VV^\mathsf{T})^\mathsf{T} = VV^\mathsf{T}$. It projects onto the row space of X. Similarly, UU^T is the orthogonal projection onto the column space of X. Also, $I - VV^\mathsf{T}$ is the orthogonal projection onto the null space, and $I - UU^\mathsf{T}$ that onto the left null space: note that

$$V^\mathsf{T}(I - VV^\mathsf{T}) = V^\mathsf{T} - \underbrace{V^\mathsf{T} V}_{I} V^\mathsf{T} = 0,$$

so $X(I - VV^\mathsf{T}) = U\Sigma V^\mathsf{T}(I - VV^\mathsf{T}) = 0$.

Since $X^\mathsf{T} X = V\Sigma^2 V^\mathsf{T}$, we have that

$$\begin{aligned} (I - \eta X^\mathsf{T} X)^k &= \left(I - VV^\mathsf{T} + VV^\mathsf{T} - \eta V\Sigma^2 V^\mathsf{T}\right)^k \\ &= \left(\left(I - VV^\mathsf{T}\right) + V\left(I - \eta\Sigma^2\right)V^\mathsf{T}\right)^k. \end{aligned}$$

Because $(\mathrm{I}-\mathrm{V}\mathrm{V}^\top)\mathrm{V} = 0 = \mathrm{V}(\mathrm{I}-\mathrm{V}\mathrm{V}^\top)$, we have that

$$\begin{aligned}
&\left((\mathrm{I}-\mathrm{V}\mathrm{V}^\top)+\mathrm{V}\mathrm{A}\mathrm{V}^\top\right)^2 \\
&\qquad = \underbrace{(\mathrm{I}-\mathrm{V}\mathrm{V}^\top)^2}_{\mathrm{I}-\mathrm{V}\mathrm{V}^\top} + \underbrace{(\mathrm{I}-\mathrm{V}\mathrm{V}^\top)\mathrm{V}}_{0}\mathrm{A}\mathrm{V}^\top + \mathrm{V}\mathrm{A}\underbrace{\mathrm{V}^\top(\mathrm{I}-\mathrm{V}\mathrm{V}^\top)}_{0} + \mathrm{V}\mathrm{A}\underbrace{\mathrm{V}^\top\mathrm{V}}_{\mathrm{I}}\mathrm{A}\mathrm{V}^\top \\
&\qquad = (\mathrm{I}-\mathrm{V}\mathrm{V}^\top) + \mathrm{V}\mathrm{A}^2\mathrm{V},
\end{aligned}$$

and iterating the product k times does the same thing:

$$(\mathrm{I}-\eta\mathrm{X}^\top\mathrm{X})^k = \left(\mathrm{I}-\mathrm{V}\mathrm{V}^\top\right) + \mathrm{V}\left(\mathrm{I}-\eta\Sigma^2\right)^k\mathrm{V}^\top. \tag{2}$$

Plugging (2) into the second term of (1),

$$\begin{aligned}
\eta\sum_{k=0}^{t-1}(\mathrm{I}-\eta\mathrm{X}^\top\mathrm{X})^k\mathrm{X}^\top y &= \eta\sum_{k=0}^{t-1}\left((\mathrm{I}-\mathrm{V}\mathrm{V}^\top)+\mathrm{V}(\mathrm{I}-\eta\Sigma^2)^k\mathrm{V}^\top\right)\mathrm{V}\Sigma\mathrm{U}^\top y \\
&= \eta\sum_{k=0}^{t-1} 0 + \mathrm{V}(\mathrm{I}-\eta\Sigma^2)^k\Sigma\mathrm{U}^\top y \\
&= \eta\mathrm{V}\left[\sum_{k=0}^{t-1}(\mathrm{I}-\eta\Sigma^2)^k\right]\Sigma\mathrm{U}^\top y.
\end{aligned} \tag{3}$$

This sum of powers looks analogous to what you might remember as a geometric series: for $|q| < 1$, $\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$.

LEMMA 1 (Neumann series). *Let* A *be a symmetric matrix with* $-\mathrm{I} \prec \mathrm{A} \prec \mathrm{I}$, *i.e. its eigenvalues are all in* $(-1,1)$. *Then* $\sum_{k=0}^{\infty}\mathrm{A}^k = (\mathrm{I}-\mathrm{A})^{-1}$.

Proof. Note that $\mathrm{I}-\mathrm{A}$ has eigenvalues in $(0,2)$ and is hence invertible. We have

$$(\mathrm{I}-\mathrm{A})\sum_{k=0}^{\mathrm{N}}\mathrm{A}^k = (\mathrm{I}-\mathrm{A})\sum_{k=0}^{\mathrm{N}}\mathrm{A}^k = \sum_{k=0}^{\mathrm{N}}\mathrm{A}^k - \sum_{k=1}^{\mathrm{N}+1}\mathrm{A}^k = \mathrm{I}-\mathrm{A}^{\mathrm{N}+1}.$$

Letting the eigenvalues of A be λ_i and corresponding eigenvectors v_i, we can write $\mathrm{A}^{\mathrm{N}+1} = \sum_i \lambda_i^{\mathrm{N}+1}v_iv_i^\top$; since $|\lambda_i| < 1$, $\lambda_i^{\mathrm{N}+1} \to 0$. Thus $\mathrm{A}^{\mathrm{N}+1} \to 0$, and

$$(\mathrm{I}-\mathrm{A})\sum_{k=0}^{\infty}\mathrm{A}^k = \lim_{\mathrm{N}\to\infty}(\mathrm{I}-\mathrm{A})\sum_{k=0}^{\mathrm{N}} = \mathrm{I} - \lim_{\mathrm{N}\to\infty}\mathrm{A}^{\mathrm{N}+1} = \mathrm{I}.$$

Left-multiply the equation above by $(\mathrm{I}-\mathrm{A})^{-1}$. □

Returning to (3), we can apply Lemma 1 as long as the eigenvalues of $(\mathrm{I}-\eta\Sigma^2)$ are in $(-1,1)$. We always have that

$$\lambda_{\max}(\mathrm{I}-\eta\Sigma^2) = 1-\eta\lambda_{\min}(\Sigma^2) < 1,$$

since $\eta > 0$, and $\lambda_{\min}(\Sigma^2) > 0$ since we're using the compact SVD. We thus only need

$$\lambda_{\min}(I - \eta\Sigma^2) = 1 - \eta\lambda_{\max}(\Sigma^2) = 1 - \eta\sigma_{\max}(X)^2 > -1,$$

which holds when

$$\eta < \frac{2}{\sigma_{\max}(X)^2} \quad \text{or} \quad \eta_0 < \frac{m}{\sigma_{\max}(X)^2}.$$

Thus for small η, we have from (3) that

$$\begin{aligned}\eta\sum_{k=0}^{\infty}(I - \eta X^\mathsf{T} X)^k X^\mathsf{T} y &= \eta V\Big(I - (I - \eta\Sigma^2)\Big)^{-1}\Sigma U^\mathsf{T} y \\ &= \eta V(\eta\Sigma^2)^{-1}\Sigma U^\mathsf{T} y = V\Sigma^{-1}U^\mathsf{T} y = X^\dagger y,\end{aligned}$$

where $X^\dagger = V\Sigma^{-1}U^\mathsf{T}$ is the Moore-Penrose pseudoinverse of X.

There's one other term in (1). Applying (2), we get

$$(I - \eta X^\mathsf{T} X)^t w_1 = (I - VV^\mathsf{T})w_1 + V \underbrace{(I - \eta\Sigma^2)^t}_{\to 0 \text{ for small } \eta} V^\mathsf{T} w_1.$$

We've thus at last proved the following:

THEOREM 2. *Let $X \in \mathbb{R}^{m\times d}$ have pseudoinverse $X^\dagger$, orthogonal projection onto its null space $(I - VV^\mathsf{T})$, and largest singular value $\sigma_{\max}(X)$. Let $y \in \mathbb{R}^m$. If $\eta < m/\sigma_{\max}(X)^2$, the gradient descent process*

$$w_{t+1} = w_t - \frac{2\eta}{m}X^\mathsf{T}(Xw - y)$$

converges to

$$w_\infty = (I - VV^\mathsf{T})w_1 + X^\dagger y.$$

1.2 *Discussion*

In the traditional setting where $\operatorname{rank}(X) = d$, $VV^\mathsf{T} = I$ and so we (unsurprisingly) obtain the unique minimizer $X^\dagger y$. In this case, we can write this as

$$X^\dagger y = V\Sigma^{-1}U^\mathsf{T} y = V\Sigma^{-2}V^\mathsf{T} V\Sigma U^\mathsf{T} y = (X^\mathsf{T} X)^{-1}X^\mathsf{T} y,$$

since $X^\mathsf{T} X = V\Sigma^2 V^\mathsf{T}$ is $d \times d$ of rank d and thus invertible.

Otherwise, though, $VV^\mathsf{T} \neq I$, and if we use a nonzero initialization the component in the null space of X persists through optimization (as it must: each step of gradient descent adds something in the row space of X).

In the usual high-dimensional setting with features in general position, $\operatorname{rank}(X) = m$, in which case $X^\dagger y = X^\mathsf{T}(XX^\mathsf{T})^{-1}y$ is *not* the only solution. We can characterize the solution that gradient descent finds like this:

PROPOSITION 3. *Let $X \in \mathbb{R}^{m\times d}$ have compact SVD $X = U\Sigma V^\mathsf{T}$, and let $y \in \mathbb{R}^m$. Then*

$$\operatorname*{arg\,min}_{w:\, Xw=y} \|w - w_1\| = \{X^\dagger y + (I - VV^\mathsf{T})w_1\}.$$

In particular, $X^\dagger y$ is the minimum-norm interpolator.

Proof. First, we characterize the set of possible interpolators:

$$y = \mathrm{X}w \quad \text{implies} \quad \mathrm{X}^\dagger y = \mathrm{X}^\dagger \mathrm{X} w = \mathrm{V}\mathrm{V}^\mathsf{T} w,$$
$$\text{so} \quad \{w \in \mathbb{R}^d : \mathrm{X}w = y\} = \{\mathrm{X}^\dagger y + q : \mathrm{V}\mathrm{V}^\mathsf{T} q = 0\}.$$

We also have, since $\mathrm{V}\mathrm{V}^\mathsf{T}\mathrm{X}^\dagger = \mathrm{X}^\dagger$ and $\mathrm{V}\mathrm{V}^\mathsf{T} q = 0$, that

$$\begin{aligned}\left\|\mathrm{X}^\dagger y + q - w_1\right\|^2 &= \left\|\mathrm{V}\mathrm{V}^\mathsf{T}(\mathrm{X}^\dagger y + q - w_1)\right\|^2 + \left\|(\mathrm{I} - \mathrm{V}\mathrm{V}^\mathsf{T})(\mathrm{X}^\dagger y + q - w_1)\right\|^2 \\ &= \left\|\mathrm{X}^\dagger y - \mathrm{V}\mathrm{V}^\mathsf{T} w_1\right\|^2 + \left\|q - (\mathrm{I} - \mathrm{V}\mathrm{V}^\mathsf{T})w_1\right\|^2.\end{aligned}$$

Our choice of q does not affect the first term; the second is uniquely minimized by picking $q = (\mathrm{I} - \mathrm{V}\mathrm{V}^\mathsf{T})w_1$. □

1.3 *What about SGD?*

Suppose that rather than stepping along $\nabla f(w_t)$, we step along $\hat{g}_t$ such that $\mathbb{E}[\hat{g}_t \mid w_t] = \nabla f(w_t)$ and the $\hat{g}_t \mid w_t$ are independent of one another:

$$w_{t+1} = w_t - \eta \hat{g}_t.$$

But then, taking the expectation of both sides,

$$\begin{aligned}\mathbb{E}[w_{t+1}] = \mathbb{E}[w_t - \eta\hat{g}_t] = \underset{w_1, \hat{g}_{1:t-1}}{\mathbb{E}}[w_t - \eta \underset{\hat{g}_t}{\mathbb{E}}[\hat{g}_t \mid w_t]] &= \underset{w_1, \hat{g}_{1:t-1}}{\mathbb{E}}[w_t - \eta \nabla f(w_t)]] \\ = \mathbb{E}\, w_t - \eta \mathbb{E}\left[\frac{2}{m}\mathrm{X}^\mathsf{T}(\mathrm{X}w_t - y)\right] &= \mathbb{E}\, w_t - \frac{2\eta}{m}\mathrm{X}^\mathsf{T}(\mathrm{X}\,\mathbb{E}\, w_t - y).\end{aligned}$$

Thus $\mathbb{E}\, w_t$ follows exactly the same update formula as w_t did for gradient descent, and so we immediately know that if $\eta \le m/\sigma_{\max}(\mathrm{X})^2$,

This property depended on $\nabla f(w)$ being affine in w! It's not true for all f.

$$\mathbb{E}\, w_t \to (\mathrm{I} - \mathrm{V}\mathrm{V}^\mathsf{T})\,\mathbb{E}\, w_1 + \mathrm{X}^\dagger y.$$

This expectation thing isn't the whole story. If you imagine using $\hat{g}_t = \nabla f(w_t) + \xi_t$ for some zero-mean noise ξ_t, any component of ξ_t that lies in the null space of X will necessarily "stick around" unless a later ξ_t cancels it out. (So, if we only add noise on the third timestep and never again, it'll definitely stick.) But, because we assumed $\mathbb{E}[\hat{g}_t \mid w_t] = \nabla f(w_t)$, ξ_t has to be zero mean, and so the *mean* of the final iterate agrees with gradient descent.

Proving high-probability (or stronger) results on the distribution of w_∞ would require much stronger assumptions about the gradient samplers and/or the function being optimized. The final iterate of constant learning rate SGD doesn't even necessarily converge: it can bounce around indefinitely.

This kind of motivates the average-iterate bound we did before.

There are now some asides poking further at this result; main content resumes on page 7.

1.4 *Aside: convergence speed*

Assume $w_1 = 0$ for simplicity. Then we can use the formula for partial sums derived in the proof of Lemma 1 inside (3) to see that

$$\begin{aligned} w_t &= \eta \mathrm{V} \left[\sum_{k=0}^{t-2} (\mathrm{I} - \eta \Sigma^2)^k \right] \Sigma \mathrm{U}^\mathsf{T} y \\ &= \eta \mathrm{V} (\eta \Sigma^2)^{-1} \left[\mathrm{I} - (\mathrm{I} - \eta \Sigma^2)^{t-1} \right] \Sigma \mathrm{U}^\mathsf{T} y \\ &= \mathrm{X}^\dagger y - \mathrm{V} \Sigma^{-2} (\mathrm{I} - \eta \Sigma^2)^{t-1} \Sigma \mathrm{U}^\mathsf{T} y, \end{aligned}$$

and so

Diagonal matrices commute.

$$\begin{aligned} \mathrm{X}^\dagger y - w_t &= \mathrm{V} (\mathrm{I} - \eta \Sigma^2)^{t-1} \Sigma^{-1} \mathrm{U}^\mathsf{T} y \\ &= \mathrm{V} (\mathrm{I} - \eta \Sigma^2)^{t-1} \mathrm{V}^\mathsf{T} \mathrm{V} \Sigma^{-1} \mathrm{U}^\mathsf{T} y = \mathrm{V} (\mathrm{I} - \eta \Sigma^2)^{t-1} \mathrm{V}^\mathsf{T} \mathrm{X}^\dagger y. \end{aligned}$$

Thus

$$\left\| \mathrm{X}^\dagger y - w_1 \right\| \le \left\| \mathrm{V} (\mathrm{I} - \eta \Sigma^2)^{t-1} \mathrm{V} \right\| \left\| \mathrm{X}^\dagger y \right\|.$$

Letting $\operatorname{diag}(\Sigma) = (\sigma_1, \dots, \sigma_r)$ with $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > 0$, we have

$$\left\| \mathrm{V} (\mathrm{I} - \eta \Sigma^2)^{t-1} \mathrm{V} \right\| = \max\left(\left| 1 - \eta \sigma_1^2 \right|, \left| 1 - \eta \sigma_r^2 \right| \right)^{t-1} = \left(\frac{1}{\mu} \right)^{t-1},$$

defining $\mu = \frac{1}{\max(1 - \eta \sigma_1^2, 1 - \eta \sigma_r^2)}$, which has $\mu > 1$ when $\eta < 2/\sigma_1^2$. Thus

$$\left\| \mathrm{X}^\dagger y - w_1 \right\| \le \frac{1}{\mu}^{t-1} \left\| \mathrm{X}^\dagger y \right\|,$$

and we can achieve $\left\| \mathrm{X}^\dagger y - w_t \right\| \le \varepsilon$ if the number of gradient steps we take is $t - 1 \ge \frac{1}{\log \mu} \log \frac{\| \mathrm{X}^\dagger y \|}{\varepsilon}$. Similarly,

$$\begin{aligned} y - \mathrm{X} w_t &= (\mathrm{I} - \mathrm{U}\mathrm{U}^\mathsf{T}) y + \mathrm{U}\mathrm{U}^\mathsf{T} y - \mathrm{X} w_t \\ &= (\mathrm{I} - \mathrm{U}\mathrm{U}^\mathsf{T}) y + \mathrm{X} (\mathrm{X}^\dagger y - w_t), \end{aligned}$$

and since $\mathrm{U}\mathrm{U}^\mathsf{T} \mathrm{X} = \mathrm{X}$ we have

$$\left\| y - \mathrm{X} w_t \right\|^2 = \left\| (\mathrm{I} - \mathrm{U}\mathrm{U}^\mathsf{T}) y \right\|^2 + \left\| \mathrm{X} (\mathrm{X}^\dagger y - w_t) \right\|^2 \le \left\| (\mathrm{I} - \mathrm{U}\mathrm{U}^\mathsf{T}) y \right\|^2 + \left(\frac{1}{\mu^{t-1}} \sigma_1 \left\| \mathrm{X}^\dagger y \right\| \right)^2,$$

or, multiplying by $2/m$,

$$f(w_t) \le f(\mathrm{X}^\dagger y) + \frac{2 \sigma_1^2 \left\| \mathrm{X}^\dagger y \right\|^2}{m} \mu^{-2(t-1)}.$$

So, we can guarantee $f(w_t) \le f(\mathrm{X}^\dagger y) + \varepsilon$ in $\frac{1}{2 \log \mu} \log \frac{2 \sigma_1^2 \| \mathrm{X}^\dagger y \|^2}{m \varepsilon}$ gradient steps, a so-called "linear rate". Using $\left\| \mathrm{X}^\dagger y \right\| \le \left\| \mathrm{X}^\dagger \right\| \left\| y \right\| = \frac{1}{\sigma_r} \left\| y \right\|$ lets us see that this rate depends on the condition number $\frac{\sigma_1}{\sigma_r}$.

In the low-dimensional regime $\operatorname{rank}(\mathrm{X}) = d$, we already knew this linear rate would be achieved: the Hessian of our objective is $\nabla^2 f(w) = \frac{2}{m} \mathrm{X}^\mathsf{T} \mathrm{X} = \frac{2}{m} \mathrm{V} \Sigma^2 \mathrm{V}^\mathsf{T}$, so our objective is always $\frac{2 \sigma_1^2}{m}$-smooth, and if $\operatorname{rank}(\mathrm{X}) = d$ then f is also $\frac{2 \sigma_d^2}{m}$-strongly convex.

We briefly mentioned, but didn't fully prove, before that gradient descent gets a linear rate on smooth, strongly convex objectives. But in the high-dimensional regime where $\operatorname{rank}(X) < d$, though, $X^\mathsf{T} X$ is singular, so f is *not* strongly convex; even so, we get the fast linear rate.

Bach [Bach23, near (11.2)] shows that the PL inequality *is satisfied here, implying a linear rate [KNS16], in the interpolating case.*

Notice that the case $\operatorname{rank}(X) = d$ implies that $f(X^\dagger y) = 0$, i.e. the interpolating setting. In fact, it's often the case that optimization algorithms do better in this interpolating case than you might otherwise expect [e.g. VBS19].

1.5 *Aside: learning rate bound*

One last question: how stringent is the requirement that $\eta < m/\sigma_{\max}(X)^2$? That is, how does $\sigma_{\max}(X)$ behave? That's going to depend on the data distribution and how d changes with m. (If we think about a fixed d, eventually we have $m \geq d$ and the result becomes uninteresting.)

If we assume $x_i \sim \mathcal{N}(\mu, \Sigma)$, we can use $X = Z + \mathbf{1}_m \mu^\mathsf{T}$ for $Z \sim \mathcal{N}(0, \Sigma)$, and so

$$\|X\| \leq \|Z\| + \|\mathbf{1}_m\| \left\|\mu\right\| = \|Z\| + \sqrt{m} \left\|\mu\right\|.$$

Thus

$$\frac{m}{\sigma_{\max}(X)^2} = \frac{m}{\|X\|^2} \geq \frac{1}{\left(\|Z\|/\sqrt{m} + \left\|\mu\right\|\right)^2}.$$

Thus, we know that the threshold for the learning rate in Theorem 2 is at least constant if $\left\|\mu\right\| = \mathcal{O}(1)$ and $\frac{1}{m}\|Z\|^2 = \mathcal{O}_p(1)$. For $\|Z\|$, it turns out [KL17] that

$$\frac{1}{m}\|Z\|^2 = \left\|\frac{1}{m} Z Z^\mathsf{T}\right\| \leq \|\Sigma\| + \left\|\frac{1}{m} Z Z^\mathsf{T} - \Sigma\right\| = \|\Sigma\| + \mathcal{O}_p\left(\sqrt{\frac{\operatorname{Tr}(\Sigma)\,\|\Sigma\|}{m}} + \frac{\operatorname{Tr}(\Sigma)}{m}\right),$$

and so we can use at least a constant learning rate if

$$\left\|\mu\right\| = \mathcal{O}(1), \qquad \|\Sigma\| = \mathcal{O}(1), \qquad \text{and} \qquad \operatorname{Tr}(\Sigma) = \mathcal{O}(m).$$

Notice that if we use $d = \Theta(m)$ (called *proportional asymptotics*), then $\Sigma = I_d$ satisfies the Σ conditions, but the amount of variance in any *single* direction can't grow with m. We also can't pick something like $\mu = \mathbf{1}$.

2 SEPARABLE LOGISTIC REGRESSION

Now let's consider logistic regression: for $y_i \in \{-1, 1\}$,

$$f(w) = \frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp(-y_i x_i^\mathsf{T} w)).$$

We're also going to assume that the data is linearly separable: there is some w^* such that $y_i x_i^\mathsf{T} w^* > 0$ for all i. Then, it's possible to drive $f(w)$ arbitrarily close to zero, but never to actually reach it: we only get $\log(1 + \exp(-t)) \to 0$ for $t \to \infty$, so we need $\|w\| \to \infty$. A solution of the form cw^* for $c \to \infty$ would work, but potentially so would many other solutions, since there are probably many possible perfect linear separators on this dataset. Which one does gradient descent find?

We're going to approach this informally, for time and simplicity. Soudry et al. [Sou+18] and Gunasekar et al. [GLSS18] handle it in full, and Ji and Telgarsky [JT19] approach the non-separable case; Bach [Bach23, Section 11.1.2] gives an overview including a few things we aren't covering here.

Notice that

$$\nabla f(w) = -\frac{1}{m}\sum_{i=1}^{m} \frac{\exp(-y_i x_i^\mathsf{T} w)}{1+\exp(-y_i x_i^\mathsf{T} w)} y_i x_i.$$

We know that we'll get $\|w_t\| \to \infty$ from the argument above; it's reasonable to expect, then, that we'll have $\frac{w_t}{\|w_t\|} \to v$ for some $\|v\| = 1$, and $y_i x_i^\mathsf{T} v > 0$ for all i since otherwise we wouldn't approach a minimizer. This gives us, roughly speaking,

$$\nabla f(\|w_t\| \, v) \sim -\frac{1}{m}\sum_{i=1}^{m} \frac{\exp(-y_i \|w_t\| \, x_i^\mathsf{T} v)}{1+\exp(-y_i \|w_t\| \, x_i^\mathsf{T} v)} y_i x_i \sim -\frac{1}{m}\sum_{i=1}^{m} \exp(-y_i \|w_t\| \, x_i^\mathsf{T} v) y_i x_i,$$

since $\frac{t}{1+t} = t + \mathcal{O}(t^2)$ and we'll eventually have $\exp(-y_i \|w_t\| \, x_i^\mathsf{T} v) \ll 1$.

The asymptotic ratio between the size of the gradient contributions from x_i and x_j is

$$\frac{\exp(-y_i \|w_t\| \, x_i^\mathsf{T} v) \, \|x_i\|}{\exp(-y_j \|w_t\| \, x_j^\mathsf{T} v) \, \|x_j\|} = \frac{\|x_i\|}{\|x_j\|} \exp\left(-\|w_t\| \, (y_i x_i^\mathsf{T} v - y_j x_j^\mathsf{T} v)\right).$$

As $\|w_t\| \to \infty$, this ratio goes to 0 if $y_i x_i^\mathsf{T} v > y_j x_j^\mathsf{T} v$, or ∞ if the order is reversed; it is $\|x_i\| / \|x_j\| \in (0, \infty)$ if and only if $y_i x_i^\mathsf{T} v = y_j x_j^\mathsf{T} v$. So, for whatever v we have, let $\mathcal{I}_v$ be the set of indices such that $y_i x_i^\mathsf{T} v$ is minimized. Only these terms really matter:

$$\nabla f(\|w_t\| \, v) \sim -\frac{1}{m}\sum_{i \in \mathcal{I}_v} \exp(-y_i \|w_t\| \, x_i^\mathsf{T} v) y_i x_i.$$

So, if gradient descent diverges in a direction v, the dominant direction in which w_t moves is a (positive) linear combination of the points $\{x_i : i \in \mathcal{I}_v\}$. Let's scale that direction to have unit margin, $\hat{v} = v / (\min_{i \in [m]} y_i x_i^\mathsf{T} v)$; this will still be a linear combination of those same points. Thus, we know that

$$\hat{v} = \sum_{i=1}^{m} \alpha_i y_i x_i \quad \text{with } \forall i,\ (\alpha_i \ge 0 \text{ and } y_i x_i^\mathsf{T} \hat{v} = 1) \text{ or } (\alpha_i = 0 \text{ and } y_i x_i^\mathsf{T} \hat{v} > 1). \tag{4}$$

It turns out that these are exactly the optimality conditions for the hard SVM / margin maximization problem, as we'll show below. (If you want to refresh your memory, our notes on SVMs were here.) Thus, $\hat{v}$ is exactly the hard SVM, i.e. gradient descent with logistic regression on separable data eventually maximizes the margin.

If you recall, the hard SVM is *also* a minimum-norm hinge loss interpolator. That's kind of neat, that we get a minimum-norm interpolator in both cases, although here it's the minimum-norm interpolator for a *different* loss than the one we're explicitly minimizing.

Soudry et al. [Sou+18] give a real proof, and also show that the convergence of gradient descent to that solution is quite slow: typically $\left\| \frac{w_t}{\|w_t\|} - \frac{\hat{v}}{\|\hat{v}\|} \right\| = \mathcal{O}\left(\frac{1}{\log t}\right)$.

We'll now derive the KKT conditions, which is slightly outside the scope of this class; main content resumes on page 11.

2.1 *Aside: KKT conditions*

Consider a general optimization problem

$$f^* = \min_{x \in \mathbb{R}^d} f(x) \quad \text{s.t. } \forall i,\ h_i(x) \le 0 \quad \text{and} \quad \forall j,\ \ell_j(x) = 0.$$

The Karush-Kuhn-Tucker conditions are a set of conditions on *both* the primal variable x and dual variables α, β corresponding to each constraint. There are a few variants; one is

- Stationarity: $x \in \arg\min_{x'} f(x') + \sum_i \alpha_i h_i(x') + \sum_j \beta_j \ell_j(x')$. For instance, this holds if $0 \in \partial f(x) + \sum_i \alpha_i \partial h_i(x) + \sum_i \beta_j \partial \ell_j(x)$.
- Primal feasibility: all of the $h_i(x) \le 0$ and $\ell_j(x) = 0$.
- Dual feasibility: all of the $\alpha_i \ge 0$.
- Complementary slackness: for all i, $\alpha_i h_i(x) = 0$.

For stationarity: recall (from our (S)GD notes) that if a function is convex and differentiable at x, $\partial f(x) = \{\nabla f(x)\}$. This is *not* necessarily true for nonconvex functions, where subgradients might not exist even where the gradient does: for example, at a suboptimal local minimum, $\nabla f(x) = 0$ but there is no tangent globally lower bounding f, and hence no subgradients. If everything is convex, then (sub)gradients are a great way to check stationarity.

The Wikipedia article is messy on this point; they state it with subgradients, but then their discussion about sufficiency only applies to a (more common) version where you use gradients *in the stationarity condition. In that version, if strong duality holds, you still get necessity, but not sufficiency without more assumptions.*

For this version of the conditions, any x, α, β satisfying the conditions are optimal, and if strong duality holds then any optimal solution must satisfy the conditions.

Applying them to the max-margin problem

$$\min_w \frac{1}{2} \|w\|^2 \quad \text{s.t. } \forall i,\ y_i x_i^\mathsf{T} w \ge 1,$$

stationarity gives $w + \sum_i \alpha_i(-y_i x_i) = 0$, primal feasibility is $1 - y_i x_i^\mathsf{T} w \le 0$, dual feasibility is $\alpha_i \ge 0$, and complementary slackness is $\alpha_i(1 - y_i x_i^\mathsf{T} w) = 0$. These four conditions, after rearranging a bit, exactly agree with (4).

Proofs

To see why the KKT conditions work, recall that the Lagrangian is given by

$$\mathcal{L}(x, \alpha, \beta) = f(x) + \sum_i \alpha_i h_i(x) + \sum_j \beta_j \ell_j(x),$$

and the Lagrange dual is

$$g(\alpha, \beta) = \min_x \mathcal{L}(x, \alpha, \beta), \qquad g^* = \max_{\alpha, v} g(\alpha, \beta) \ \text{ s.t. } \forall i,\ \alpha_i \ge 0.$$

Notice that stationarity is exactly saying that $x \in \arg\min_{x'} \mathcal{L}(x', \alpha, \beta)$.

It always holds that $f^* \ge g^*$. One interesting consequence is that we have for any x, α, β that

$$f(x) - f^* \le f(x) - g^* \le f(x) - g(\alpha, \beta).$$

So the *duality gap* $f(x) - g(\alpha, \beta)$ upper-bounds the suboptimality of $f(x)$. The same is true for $g(\alpha, \beta)$, since

$$g^* - g(\alpha, \beta) \le f^* - g(\alpha, \beta) \le f(x) - g(\alpha, \beta).$$

If we ever have $f(x) = g(\alpha, \beta)$, we know both x and α, β are optimal.

To show sufficiency of the KKT conditions: suppose that (x^*, α^*, β^*) satisfy the KKT conditions. Then

$$g(\alpha^*, \beta^*) = \min_x \mathcal{L}(x, \alpha^*, \beta^*) = \mathcal{L}(x^*, \alpha^*, \beta^*),$$

where the second equality is by stationarity. Because complementary slackness implies that each $\alpha_i^* h_i(x^*) = 0$, and primal feasibility requires each $\ell_j(x^*) = 0$, we have that

$$\mathcal{L}(x^*, \alpha^*, \beta^*) = f(x^*) + \sum_i \alpha_i^* h_i(x^*) + \sum_j \beta_j \ell_j(x^*) = f(x^*).$$

But now we've shown $g(\alpha^*, \beta^*) = f(x^*)$, and everything is feasible, so x^* is a primal solution, and (α^*, β^*) a dual solution.

To show necessity of the KKT condtions, we'll need *strong duality*, where $f^* = g^*$. Two important cases where strong duality holds are

- Linearity constraint qualification: if all the constraints are affine.
- Slater's condition: if f and the h_i are convex, the ℓ_j are affine, and there exists at least one x with all $\ell_j(x) = 0$, any affine h_i having $h_i(x) \le 0$, and any non-affine h_i having $h_i(x) < 0$.

If we have strong duality, x^* is primal optimal, and (α^*, β^*) is dual optimal, then

$$f(x^*) = f^* = g^* = g(\alpha^*, \beta^*) = \min_x f(x) + \sum_i \alpha_i^* h_i(x) + \sum_j \beta_j^* \ell_j(x).$$

We can upper-bound this minimum by plugging in any specific value of x, say x^*; this gives

$$f(x^*) \le f(x^*) + \sum_i \alpha_i^* h_i(x^*) + \sum_j \beta_j^* \ell_j(x^*).$$

If x^* is optimal it must be primal-feasible, so $\ell_j(x^*) = 0$ and $h_i(x^*) \le 0$; likewise α_i^* is dual-feasible, so $\alpha_i^* \ge 0$. But that means the first sum is nonpositive, and the second sum is zero, giving

$$f(x^*) \le f(x^*) + \sum_i \alpha_i^* h_i(x^*) + \sum_j \beta_j^* \ell_j(x^*) \le f(x^*).$$

Since obviously $f(x^*) = f(x^*)$, we must therefore have $\sum_i \alpha_i^* h_i(x^*) = 0$. But we just said each $\alpha_i h_i(x^*) \le 0$, so if they sum to zero they must each be zero: that's exactly complementary slackness. We also showed that

$$\min_x f(x) + \sum_i \alpha_i^* h_i(x) + \sum_j \beta_j^* \ell_j(x) = f(x^*) + \sum_i \alpha_i^* h_i(x^*) + \sum_j \beta_j^* \ell_j(x^*),$$

i.e. that $x^* \in \arg\min_x \mathcal{L}(x, \alpha^*, \beta^*)$; this is stationarity. Since everything must be feasible, we've shown all the KKT conditions hold.

Most importantly, Lyu and Li [LL20] and Ji and Telgarsky [JT20] study small-learning-rate gradient descent on L-*homogeneous* networks, those satisfying $h(x; \alpha w) = \alpha^L h(x; w)$ for $\alpha > 0$; this is true e.g. for (leaky)-ReLU networks. (We'll describe the [LL20] results.) Their analysis is in terms of the *normalized margin*

$$\bar{\gamma}(w) = \frac{\min_{i \in [m]} y_i h(x_i; w)}{\|w\|_2^L}.$$

This normalization is exactly the one that makes $\bar{\gamma}(\alpha w) = \bar{\gamma}(w)$. They show, using an approach like that of Section 2, that gradient flow or small-learning-rate gradient descent (under some additional regularity conditions) monotonically increase the log-sum-exp version of normalized margin, which means they approximately monotonically increase the normalized margin, which roughly means that it finds a local maximum (ish) of the normalized margin.

They talk about convergence to a "KKT point"; this is using the version of the KKT conditions where stationarity is defined by gradients, not subgradients, and hence isn't sufficient for optimality in nonconvex problems.

This is a kind of margin maximization, but in general it's *not* margin maximization in an RKHS. Compare this to training a very wide network with square loss, in which case the implicit regularization prefers solutions with minimal NTK norm distance from the initialization. Knowing these results, you can ask questions like what this margin maximization actually does on particular models [e.g. Fre+23].

There's been a bunch of recent work trying to figure out the implicit regularization of Adam, rather than SGD, on homogeneous networks; some recent papers are [WMCL21; Wan+22; CKS23].

There's also a *ton* more work in this area; Vardi [Var22] gives a survey.

REFERENCES

[Bach23] Francis Bach. *Learning Theory from First Principles*. April 2023 draft.

[CKS23] Matias D. Cattaneo, Jason M. Klusowski, and Boris Shigida. *On the Implicit Bias of Adam*. 2023. arXiv: 2309.00079.

[Fre+23] Spencer Frei, Gal Vardi, Peter L. Bartlett, Nathan Srebro, and Wei Hu. "Implicit Bias in Leaky ReLU Networks Trained on High-Dimensional Data." *ICLR*. 2023. arXiv: 2210.07082.

[GLSS18] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. "Characterizing Implicit Bias in Terms of Optimization Geometry." *ICML*. 2018. arXiv: 1802.08246.

[JT19] Ziwei Ji and Matus Telgarsky. "The implicit bias of gradient descent on nonseparable data." *COLT*. 2019. arXiv: 1803.07300.

[JT20] Ziwei Ji and Matus Telgarsky. "Directional convergence and alignment in deep learning." *NeurIPS*. 2020. arXiv: 2006.06657.

[KL17] Vladimir Koltchinskii and Karim Lounici. "Concentration Inequalities and Moment Bounds for Sample Covariance Operators." *Bernoulli* 23.1 (2017), pages 110–133. arXiv: 1405.2468.

[KNS16] Hamed Karimi, Julie Nutini, and Mark Schmidt. "Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition." *ECML PKDD*. 2016. arXiv: 1608.04636.

[LL20] Kaifeng Lyu and Jian Li. "Gradient Descent Maximizes the Margin of Homogeneous Neural Networks." *ICLR*. 2020. arXiv: 1906.05890.

[Sou+18] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. "The Implicit Bias of Gradient Descent on Separable Data." *JMLR* (2018). arXiv: 1710.10345.

[Var22] Gal Vardi. *On the Implicit Bias in Deep-Learning Algorithms*. 2022. arXiv: 2208.12591.

[VBS19] Sharan Vaswani, Francis Bach, and Mark Schmidt. "Fast and Faster Convergence of SGD for Over-Parameterized Models and an Accelerated Perceptron." *AISTATS*. 2019. arXiv: 1810.07288.

[Wan+22] Bohan Wang, Qi Meng, Huishuai Zhang, Ruoyu Sun, Wei Chen, Zhi-Ming Ma, and Tie-Yan Liu. "Does Momentum Change the Implicit Regularization on Separable Data?" *NeurIPS*. 2022. arXiv: 2110.03891 [cs.LG].

[WMCL21] Bohan Wang, Qi Meng, Wei Chen, and Tie-Yan Liu. "The Implicit Bias for Adaptive Optimization Algorithms on Homogeneous Neural Networks." *ICML*. 2021. arXiv: 2012.06244.