Basic definitions

For convenience, bake the training set into the predictor:

\[ f(w) := \begin{bmatrix} f(x_1; w) \\ \vdots \\ f(x_n; w) \end{bmatrix} \in \mathbb{R}^{n \times p} \]

We'll be considering squared loss regression:

\[ R(\alpha f(w)) := \frac{1}{2} \| \alpha f(w) - y \|^2, \quad R(t) := R(\alpha f(t)) = R(\alpha f(w(t))), \]

where \( \alpha > 0 \) is a scale factor we'll optimize for later.

We'll consider gradient flow:

\[ J_w := \begin{bmatrix} \nabla f(x_1; w) \\ \vdots \\ \nabla f(x_n; w) \end{bmatrix} \]

\[ J_t := J_{w_t} \]

\[ \dot{u}(t) := -\nabla u R(\alpha f_0(u(t))) = -\alpha J_0^T \nabla R(\alpha f_0(u(t))). \]

Tangent model

As usual in NTK, the proof tracks a tangent model / linearization at initialization:

\[ f_0(u) := f(w(0)) + J_0(u - w(0)). \]

\[ \dot{u}(t) := -\nabla u R(\alpha f_0(u(t))) = -\alpha J_0^T \nabla R(\alpha f_0(u(t))). \]

Both gradient flows have the same initial condition:

\[ u(0) = w(0), \quad f_0(u(0)) = f_0(w(0)) = f(w(0)). \]
Main theorem

- The goal, as usual in this literature, is to prove we minimize the empirical risk.
- As a byproduct, the proof shows we stay close to initialization and to the tangent model.
- We need some further notation:

\[
\begin{align*}
\text{rank}(J_0) &= n, \\
\sigma_{\text{min}} := \sigma_{\text{min}}(J_0) &= \sqrt{\lambda_{\text{min}}(J_0 J_0^T)} = \sqrt{\lambda_n(J_0 J_0^T)} > 0, \\
\sigma_{\text{max}} := \sigma_{\text{max}}(J_0) &> 0, \\
\|J_w - J_v\| &\leq \beta \|w - v\|.
\end{align*}
\]

We'll discuss these assumptions and estimate them in the shallow case shortly.
- The proof and theorem are not tied to the shallow case, but we'll only make it concrete in that case.

Theorem. Suppose \(\alpha \geq \beta \sqrt{288\sigma_{\text{max}}^2 R(0) / \sigma_{\text{min}}^2}\). Then

\[
\begin{align*}
\max \{R(\alpha f(w(t))), R(\alpha f_0(u(t)))\} &\leq R(0) \exp(-t\alpha^2 \sigma_{\text{min}}^2 / 2), \\
\max \{|w(t) - w(0)|, |u(t) - w(0)|\} &\leq \frac{3\sqrt{8\sigma_{\text{max}}^2 R(0)}}{\alpha \sigma_{\text{min}}^2}.
\end{align*}
\]

Remark. For a reasonable tradeoff, let’s set \(\kappa := \sigma_{\text{max}} / \sigma_{\text{min}}\) (“condition number”? and \(\alpha_0 := \kappa / \sigma_{\text{min}} = \sigma_{\text{max}} / \sigma_{\text{min}}\), whereby

\[
\begin{align*}
\max \{R(\alpha f(w(t))), R(\alpha f_0(u(t)))\} &\leq R(0) \exp(-t\kappa^2 / 2), \\
\max \{|w(t) - w(0)|, |u(t) - w(0)|\} &\leq \frac{3\sqrt{8R(0)}}{\alpha_0^2},
\end{align*}
\]

and the precondition on \(\alpha\) is met once \(\sigma_{\text{min}} = \Omega(\beta \sqrt{R(0)})\). We’ll revisit this in the shallow case.

Singular value assumptions

- We’ll prove in the homework that the assumptions imply \(\alpha f(w(t)) \approx y\) and \(\alpha f(u(t)) \approx y\) for sufficiently large \(t\); that is to say, we’ve baked in a strong “approximation/representation” assumption, that we can exactly fit any training set!
- Since \(J \in \mathbb{R}^{n \times p}\) and \(\text{rank}(J_0) = n\), let’s reason about the eigenvalues via \(\sigma_{\text{min}} = \sqrt{\lambda_{\text{min}}(J_0 J_0^T)}\).
- \((J_0 J_0^T)_{ij} = \nabla f(x_i; w(0))^T \nabla f(x_j; w(0)).\) Therefore, as before, this is something like \(\approx m \mathbb{E} \nabla f(x_i; w)^T \nabla f(x_j; w)\).
- (Shallow case.) Again treating the singular values as \(\sqrt{\lambda_{j}(J_0 J_0^T)}\), we hope that with some concentration inequalities, we can have \(\sigma_{\text{min}}\) and \(\sigma_{\text{max}}\) to be \(\tilde{O}(\sqrt{m})\). In particular, \(J_0 J_0^T\) is positive definite (not just positive semi-definite). This will be deferred to homework, but for now let’s have a \(\sqrt{m}\) ballpark in mind.

Smoothness assumptions

- For the smoothness calculation in the shallow case

\[
\sum_j s_j \sigma(w_j^T x)
\]

with \(\beta_0\)-smooth activations,

\[
\begin{align*}
\|J_w - J_v\|_2^2 &\leq \sum_{i,j} \|x_i\|^2 (\sigma'(w_j^T x_i) - \sigma'(v_j^T x_i))^2 \\
&\leq \sum_{i,j} \|x_i\|^4 \beta_0^2 \|w_j - v_j\|^2 \\
&= \beta_0^2 \|X\|_F^4 \|w - v\|^2.
\end{align*}
\]

Thus \(\beta = \beta_0 \|X\|_F^2\) suffices.
- Consequently, \(\beta\)-smoothness of \(J_w\) captures the “activations hardly changing” setting we want, and we expect our bounds to have \(\|J_t - J_0\|\) small.
Remark (tradeoff revisited).

For the tradeoff, we set $\alpha := \frac{\sigma_{\text{max}}}{\sigma^2_{\text{min}}}$, and got: if $\sigma_{\text{min}} \geq \beta \sqrt{288} R(0)$,

$$\max \{ R(\alpha f(w(t))), R(\alpha f_0(u(t))) \} \leq R(0) \exp(-t(\sigma_{\text{max}}/\sigma_{\text{min}})^2/2),$$
$$\max \{ \|w(t) - w(0)\|, \|u(t) - w(0)\| \} \leq 3\sqrt{8 R(0)},$$

- In the shallow case, $\sigma_{\text{min}} = \Theta(\sqrt{m})$ and $\beta = \Theta(n)$ and $R(0) = O(n)$, so $\sqrt{m} = \Omega(n^{1.5})$ suffices. For regression, people have gotten the dependence down to $m = \Omega(n)$, and for classification $m = \Omega(\ln n)$.
- Even as we increase $m$, the total weight motion is a constant; there is redundancy in the weights. In many setting it is possible to prove individual weights move no more than $1/\sqrt{m}$.

General proof scheme.

- First we will show that good stuff happens near initialization. Indeed, define some ball of nice iterates

$$\|w - w(0)\| \leq B := \frac{\sigma_{\text{min}}}{2\beta}.$$

and the earliest exit time from this regime:

$$T := \inf \{ t \geq 0 : \|w(t) - w(0)\| > B \}$$

- The proof will hold for $t \in [0, T]$. But we will also prove $T = \infty$!

The evolution in prediction space is

$$\frac{d}{dt} \alpha f(w(t)) = \alpha J_t \dot{w}(t) = -\alpha^2 J_t J_t^T \nabla R(\alpha f(w(t))),$$
$$= -\alpha^2 J_t J_t^T (\alpha f(w(t)) - y),$$
$$\frac{d}{dt} \alpha f_0(u(t)) = \frac{d}{dt} \alpha f(w_0) + J_0(u(t) - w_0))$$
$$= \alpha J_0 \dot{u}(t) = -\alpha^2 J_0 J_0^T \nabla R(\alpha f_0(u(t))).$$

The first one is complicated because we don't know how $J_t$ evolves.

But the second one can be written

$$\frac{d}{dt} [\alpha f_0(u(t))] = -\alpha^2 (J_0 J_0^T) [\alpha f_0(u(t))] + \alpha^2 (J_0 J_0^T) y,$$

which is a concave quadratic in the predictions $\alpha f_0(u(t))$.

The original NTK paper, (Jacot, Gabriel, and Hongler 2018), had as its story that GF follows a gradient in kernel space. Seeing the evolution of $\alpha f_0(u(t))$ makes this clear!

Let's fantasize a little and suppose $(J_w J_w)^T$ is also positive semi-definite. Do we still have a nice convergence theory?

Lemma. Suppose $z(t) = -Q(t) \nabla R(z(t))$ and $\lambda := \inf_{t \in [0, \tau]} \lambda_{\text{min}} Q(t) > 0$. Then for any $t \in [0, \tau]$,

$$R(z(t)) \leq R(z(0)) \exp(-2t \lambda).$$

Remark. A useful consequence is

$$\|z(t) - y\| \leq \|z(0) - y\| \exp(-t \lambda).$$

Proof. Mostly just repeating our old strong convexity steps,

$$\frac{d}{dt} \frac{1}{2} \|z(t) - y\|^2 = \langle -Q(t)(z(t) - y), z(t) - y \rangle$$
$$\leq -\lambda_{\text{min}} (Q(t)) \|z(t) - y, z(t) - y \|$$
$$\leq -2 \lambda \|z(t) - y\|^2 / 2,$$

and Grönwall's inequality completes the proof.
We can also prove this setting implies we stay close to initialization.

**Lemma.** Suppose \( \dot{v}(t) = -S(t)^T \nabla R(g(v(t))) \), where \( S_t S_t^T = Q_t \), and \( \lambda_i(Q_t) \in [\lambda, 1] \) for \([0, \tau]\). Then for \( t \in [0, \tau] \),

\[
\|v(t) - v(0)\| \leq \frac{\sqrt{\lambda_1}}{\lambda} \|g(v(0)) - y\| 
\preceq \sqrt{\lambda_1} \int_0^t \exp(-s\lambda) \, ds 
\leq \frac{\sqrt{\lambda_1}}{\lambda} \|g(v(0)) - y\| \leq \frac{\sqrt{2\lambda_1 R(g(v(0)))}}{\lambda}.
\]

**Proof.**

\[
\|v(t) - v(0)\| = \left\| \int_0^t \dot{v}(s) \, ds \right\| \leq \int_0^t \|\dot{v}(s)\| \, ds 
\leq \frac{\sqrt{\lambda_1}}{\lambda} \|g(v(s)) - y\| \, ds 
\leq \frac{\sqrt{\lambda_1}}{\lambda} \|g(v(0)) - y\| \leq \frac{\sqrt{2\lambda_1 R(g(v(0)))}}{\lambda}.
\]

**Lemma.** Suppose \( \|w - w(0)\| \leq B = \frac{\sigma_{\min}}{2\beta} \).

\( \sigma_{\min}(J_w) \geq \sigma_{\min} - \beta \|w - w(0)\| \geq \sigma_{\min} \)

\( \sigma_{\max}(J_w) \leq \frac{3\sigma_{\max}}{2} \).

**Proof.** For the upper bound,

\[
\|J_w\| \leq \|J_0\| + \|J_w - J_0\| \leq \|J_0\| + \beta \|w - w(0)\| \leq \sigma_{\max} + \frac{\sigma_{\min}}{2}.
\]

For the lower bound, given vector \( v \) define \( A_v := J_0^T v \) and \( B_v := (J_w - J_0)^T v \), and thus

\[
\sigma_{\min}(J_w)^2 = \min_{\|v\|=1} v^T J_w J_w^T v 
= \min_{\|v\|=1} ( (J_0 + J_w - J_0)^T J_0 + (J_w - J_0)^T ) v 
= \min_{\|v\|=1} \|A_v\|^2 + 2A_v^T B_v + \|B_v\|^2 
\geq \min_{\|v\|=1} \|A_v\|^2 - 2\|A_v\| \cdot \|B_v\| + \|B_v\|^2 
= \min_{\|v\|=1} (\|A_v\| - \|B_v\|)^2 \geq \min_{\|v\|=1} (\sigma_{\min} - \beta B)^2 = \left(\frac{\sigma_{\min}}{2}\right)^2.
\]

**Where does this leave us?**

We can apply the previous two lemmas to the tangent model \( u(t) \), since for any \( t \geq 0 \),

\[
\dot{u}(t) = -\alpha J_0^T \nabla (\alpha f(u(t))), \quad \frac{d}{dt} \alpha f(u(t)) = -\alpha^2 (J_w J_w^T) \nabla (\alpha f(u(t))).
\]

Thus since \( Q_0 := \alpha^2 J_0 J_0^T \) satisfies \( \lambda_i(Q_0) \in 2[\sigma_{\min}^2, \sigma_{\max}^2] \),

\[
R(\alpha f(u(t))) \leq R(\alpha f(w(0))) \exp(-2t\alpha^2 \sigma_{\min}^2), 
\|
\|u(t) - u(0)\| \leq \frac{\sqrt{2\sigma_{\max}^2 R(\alpha f(w(0)))}}{\alpha \sigma_{\min}^2}. 
\]

How about \( w(t) \)?

Let’s relate \((J_w J_w^T)\) to \((J_0 J_0^T)\).

Using this, for \( t \in [0, T] \),

\[
\dot{w}(t) = -\alpha J_w^T \nabla (\alpha f(w(t))), \quad \frac{d}{dt} \alpha f(w(t)) = -\alpha^2 (J_w J_w^T) \nabla (\alpha f(w(t))).
\]

Thus since \( Q_t := \alpha^2 J_t J_t^T \) satisfies \( \lambda_i(Q_t) \in \alpha^2[\sigma_{\min}^2/4, 9\sigma_{\max}^2/4] \),

\[
R(\alpha f(w(t))) \leq R(\alpha f(w(0))) \exp(-t\alpha^2 \sigma_{\min}^2/2), 
\|
\|w(t) - w(0)\| \leq \frac{\sqrt{3} \alpha \sigma_{\max}^2 R(\alpha f(w(0)))}{\alpha \sigma_{\min}^2} =: \star.
\]

Moreover, note that if

\[
\alpha \geq \frac{\beta \sqrt{288\sigma_{\max}^2 R(0)}}{\sigma_{\min}^3},
\]

then \( \star \leq B = \frac{\sigma_{\min}}{2\beta} \), and \( T = \infty \).

This completes the proof.
On the downside, this means the proof is insensitive to any benefits of the network $f$ over the tangent model $f_0$!

Assumptions can fail in many cases we care about, but proof only needs them inside the $B$-ball.

This also implies closeness of $w(t)$ and $u(t)$, since

$$\|w(t) - u(t)\| \leq \|w(t) - w(0)\| + \|u(t) - w(0)\|,$$

$$\|g(w(t)) - g_0(u(t))\| \leq \|g(w(t)) - y\| + \|g_0(w(0)) - y\|,$$

and $u(0) = w(0)$.

---

A polynomial $p(x)$ is $L$-homogeneous when all monomials have the same degree; by the earlier calculation,

$$p(\alpha x) = \sum_{j=1}^{r} m_j(\alpha x) = \alpha^L \sum_{j=1}^{r} m_j(x).$$

Norms are 1-homogeneous, meaning $\|\alpha x\| = \alpha \|x\|$ for $\alpha > 0$. But they moreover satisfy a stronger property $\|\alpha x\| = |\alpha| \cdot \|x\|$ when $\alpha < 0$. Also, $\ell_p$ norms are obtained by taking the root of a homogeneous polynomial, which in general changes the degree of a homogeneous function.

---

Homogeneity in math is often tied to polynomials and generalizations thereof.

---

**Remarks (high-level on the general result).**

- On the downside, this means the proof is insensitive to any benefits of the network $f$ over the tangent model $f_0$!
- Assumptions can fail in many cases we care about, but proof only needs them inside the $B$-ball.
- This also implies closeness of $w(t)$ and $u(t)$, since

$$\|w(t) - u(t)\| \leq \|w(t) - w(0)\| + \|u(t) - w(0)\|,$$

$$\|g(w(t)) - g_0(u(t))\| \leq \|g(w(t)) - y\| + \|g_0(w(0)) - y\|,$$

and $u(0) = w(0)$.

---

**Examples.**

- A polynomial $p(x)$ is $L$-homogeneous when all monomials have the same degree; by the earlier calculation,

$$p(\alpha x) = \sum_{j=1}^{r} m_j(\alpha x) = \alpha^L \sum_{j=1}^{r} m_j(x).$$

- Norms are 1-homogeneous, meaning $\|\alpha x\| = \alpha \|x\|$ for $\alpha > 0$. But they moreover satisfy a stronger property $\|\alpha x\| = |\alpha| \cdot \|x\|$ when $\alpha < 0$. Also, $\ell_p$ norms are obtained by taking the root of a homogeneous polynomial, which in general changes the degree of a homogeneous function.

---

**Example (continued).**

- Layers of a ReLU network are 1-homogeneous:

$$f(x; (W_1, \ldots, \alpha W_i, \ldots, W_L))$$

$$= W_L \sigma(W_{L-1}\sigma(\ldots \alpha W_i \sigma(\ldots W_1 x \ldots) \ldots))$$

$$= \alpha W_L \sigma(W_{L-1}\sigma(\ldots W_i \sigma(\ldots W_1 x \ldots) \ldots))$$

$$= \alpha f(x; w).$$

The entire network is $L$-homogeneous:

$$f(x; \alpha w) = f(x; (\alpha W_1, \ldots, \alpha W_L))$$

$$= \alpha W_L \sigma(\alpha W_{L-1}\sigma(\ldots \sigma(\alpha W_1 x) \ldots))$$

$$= \alpha^L W_L \sigma(W_{L-1}\sigma(\ldots \sigma(W_1 x) \ldots))$$

$$= \alpha^L f(x; w).$$

---

Homework will cover some nonsmooth architectures that are not positive homogeneous!
Positive homogeneity and gradients

Let $A_i$ be a diagonal matrix with activations on the diagonal:

$$A_i = \text{diag}(\sigma'(W_i \sigma(\ldots \sigma(W_1 x) \ldots))) ,$$

(note we’ve baked in $x_i$ and so $\sigma(r) = r \sigma'(r)$ implies layer $i$ outputs

$$x \mapsto A_i W_i \sigma(\ldots \sigma(W_1 x) \ldots)) = A_i W_i A_{i-1} W_{i-1} \cdots A_1 W_1 x,$$

and the network outputs

$$f(x; w) = W_L A_{L-1} W_{L-1} A_{L-2} \cdots A_1 W_1 x.$$ 

and the gradient is

$$\frac{d}{dW_i} f(x; w) = (W_L A_{L-1} \cdots W_{i+1} A_i)^{T}(A_{i-1} W_{i-1} \cdots W_1 x)^{T} .$$

Additionally

$$\left\langle W_i, \frac{d}{dW_i} f(x; w) \right\rangle = f(x; w) = \left\langle W_{i+1}, \frac{d}{dW_{i+1}} f(x; w) \right\rangle .$$

We can make this calculation much more general...

**Lemma** (Lyu and Li 2019). If $f$ is $L$-homogeneous and locally Lipschitz, then $\langle x, \partial^a f(x) \rangle = \{L f(x)\}$.

**Proof.** Suppose $f$ differentiable at $w$; then

$$\left. \frac{d}{d\alpha} f(\alpha w) \right|_{\alpha=1} = \langle \nabla f(w), w \rangle ,$$

$$\left. \frac{d}{d\alpha} \alpha L f(w) \right|_{\alpha=1} = L \alpha^{L-1} f(w) \bigg|_{\alpha=1} = L f(w) ,$$

and these are equal so $\langle \nabla f(w), w \rangle = L f(w)$.

For the nondifferentiable case, if $w_i \to w$ and $\nabla f(w_i) \to s$,

$$L f(w) = \lim_{i} L f(w_i) = \lim_{i} \langle \nabla f(w_i), w_i \rangle = \langle s, w \rangle .$$

Similarly, $s = \sum_{i=1}^{k} \beta_k s_i$ with $\sum_{i} \beta_i = 1$ and $s_i$ as preceding gives

$$\langle s, w \rangle = \sum_{i} \beta_i \langle s_i, w \rangle = \sum_{i} \beta_i L f(w) = L f(w) .$$

8c. Norm preservation

If predictions are positive homogeneous in each layer, it gives an interesting property.

**Lemma** (Simon S. Du, Hu, and Lee 2018). Suppose for $\alpha > 0$, $f(x; (W_L, \ldots, \alpha W_i, \ldots, W_1)) = \alpha f(x; w)$ (each layer is 1-homogeneous). Then for every pair of layers $(i, j)$, the gradient flow maintains

$$\frac{1}{2} \|W_j(t)\|^2 - \frac{1}{2} \|W_j(0)\|^2 = \frac{1}{2} \|W_i(t)\|^2 - \frac{1}{2} \|W_i(0)\|^2 .$$

**Remark.** We’ll assume a risk of the form $E_k \ell(y_k f(x_k; w))$, but it holds more generally.
8d. Smoothness inequality adapted to ReLU

Let’s consider: single hidden ReLU layer, only bottom trainable:

\[ f(x; w) := \frac{1}{\sqrt{m}} \sum_j a_j \sigma(\langle x, w_j \rangle), \quad a_j \in \{\pm 1\}. \]

Let \( W_s \in \mathbb{R}^{m \times d} \) denote parameters at time \( s \), suppose \( \|x\| \leq 1 \).

\[ \frac{df(x; W)}{dW} = \begin{bmatrix} a_1 \sigma'(w_1^T x)/\sqrt{m} \\ \vdots \\ a_m \sigma'(w_m^T x)/\sqrt{m} \end{bmatrix}, \]

\[ \left\| \frac{df(x; W)}{dW} \right\|_F^2 = \sum_j \left\| a_j \sigma'(w_j^T x)/\sqrt{m} \right\|_2^2 \leq \frac{1}{m} \sum_j \|x\|_2^2 \leq 1. \]

We’ll use the logistic loss, whereby

\[ \ell(z) = \ln(1 + \exp(-z)), \]
\[ \ell'(z) = \frac{-\exp(-z)}{1 + \exp(-z)} \in (-1, 0), \]
\[ R(W) := \frac{1}{n} \sum_k \ell(y_k f(x_k; W)). \]

A key fact (can be verified with derivatives) is

\[ |\ell'(z)| = -\ell'(z) \leq \ell(z), \]

whereby

\[ \frac{dR}{dW} = \frac{1}{n} \sum_k \ell'(y_k f(x_k; W)) y_k \nabla f(x_k W), \]
\[ \left\| \frac{dR}{dW} \right\|_F \leq \frac{1}{n} \sum_k |\ell'(y_k f(x_k; W))| \cdot \|y_k \nabla f(x_k W)\|_F \leq \frac{1}{n} \sum_k |\ell'(y_k f(x_k; W))| \leq \min\left\{1, R(W)\right\}. \]
Lemma (Lemma 2.6, Ji and Telgarsky 2019) If \( \eta \leq 1 \), for any \( Z \),
\[
\| W_t - Z \|^2_F + \eta \sum_{i<t} R(i)(W_i) \leq \| W_0 - Z \|^2_F + 2\eta \sum_{i=t} R(i)(Z),
\]
where \( R(i)(W) = \frac{1}{n} \sum_k \ell(y_k \langle W, \nabla f(x_k; W) \rangle) \).

Remarks.

- \( R(i)(W_i) = R(W_i) \).
- \( R(i)(Z) \approx R(Z) \) if \( W_i \) and \( Z \) have similar activations.
- (Ji and Telgarsky 2019) uses this in a proof scheme like (Chizat and Bach 2019): consider those iterations where the activations are similar, and then prove it actually happens a lot. (Ji and Telgarsky 2019), with additional work, can use this to prove low test error.

Proof. Using the squared distance potential as usual,
\[
\| W_{i+1} - Z \|^2_F = \| W_i - Z \|^2_F - 2\eta \langle \nabla R(W_i), W_i - Z \rangle + \eta^2 \| \nabla R(W_i) \|^2_F,
\]
where \( \| \nabla R(W_i) \|^2_F \leq \| \nabla R(W_i) \|^2_F \leq R(W_i) = R(i)(W_i) \), and
\[
\sum_k \ell(y_k \langle W_i, \nabla f(x_k; W_i) \rangle) \leq \sum_k \ell(y_k \langle W_i, \nabla f(x_k; W_i) \rangle - y_k f(x_k; W_i))
\]
\[
= n \left( R(i)(Z) - R(i)(W_i) \right).
\]
Together,
\[
\| W_{i+1} - Z \|^2_F = \| W_i - Z \|^2_F + 2\eta \left( R(i)(Z) - R(i)(W_i) \right) + \eta R_i(W_i);
\]
applying \( \sum_{i<t} \) to both sides gives the bound.

8e. Omitted topics

- Mean-field perspective (Chizat and Bach 2018; Mei, Montanari, and Nguyen 2018): as \( m \to \infty \), gradient descent mimics a Wasserstein flow on a distribution over nodes (random features).
- Landscape analysis. (E.g., all local optima are global.)
  - Matrix completion: solve (under RIP)
    \[
    \min_{X \in \mathbb{R}^{d \times r}} \sum_{(i,j) \in S} (M_{i,j} - XX^T)^2.
    \]
    Recently it was shown that all local optima are global, and so gradient descent from random initialization suffices (Ge, Lee, and Ma 2016).
  - For linear networks optimized with the squared loss, local optima are global, but there are bad saddle points (Kawaguchi 2016).
  - Width \( n \) suffices with general losses and networks (Nguyen and Hein 2017).

- There is also work on residual networks but I haven’t looked.

(Farthom omitted work.)

- Margin perspective: gradient descent on standard networks seems to maximize margins (Bartlett, Foster, and Telgarsky 2017), and in some cases we can prove this [Telgarsky (2013); Soudry, Hoffer, and Srebro (2017); Ji and Telgarsky (2018); etc].
- Multiple layers. Some work analyzes multiple layers, but AFAIK none of this work (in the nonlinear case) shows any benefit to depth, indeed the theorems degrade with depth (happy to be wrong here!) (Allen-Zhu, Li, and Song 2018; Du et al. 2018).
**Acceleration.** Consider gradient descent with momentum: $w_0$ arbitrary, and thereafter

$$v_{i+1} := w_i - \eta_i \nabla R(w_i), \quad w_{i+1} := v_{i+1} + \gamma_i(v_{i+1} - v_i)$$

This seems to help in deep learning (even in stochastic case), but no one knows why.

If set $\eta_i = 1/\beta$ and $\gamma_i = i/(i + 3)$ (constants matter) and $R$ convex, $R(w_i) - \inf_w R(w) \leq O(1/t^2)$ ("Nesterov's accelerated method"). This rate is tight amongst algorithms outputting iterates in the span of gradients, under some assumptions people treat as standard.

**Beyond NTK.** A very limited amount of work studies nonlinear cases beyond what is possible with the NTK (Allen-Zhu and Li 2019).

---

**References I**


---

**References II**


---

**References III**

Ge, Rong, Jason D. Lee, and Tengyu Ma. 2016. “Matrix Completion Has No Spurious Local Minimum.” In NIPS.


———. 2019. “Polylogarithmic Width Suffices for Gradient Descent to Achieve Arbitrarily Small Test Error with Shallow Relu Networks.”

Kawaguchi, Kenji. 2016. “Deep Learning Without Poor Local Minima.” In NIPS.

Lyu, Kaifeng, and Jian Li. 2019. “Gradient Descent Maximizes the Margin of Homogeneous Neural Networks.”
References IV


Telgarsky, Matus. 2013. “Margins, Shrinkage, and Boosting.” In ICML.