Lecture 11: High-dimensional non-convex dynamics

SGD on NNs: if we are free to set architecture and initialization, then we can emulate any polynomial time algorithm.

↳ this is very different than what we do in practice

↳ there is some "hyperparameter tuning" but the space of search is relatively constrained: we only use some standard architectures and "generic" initializations (fully connected / CNNs / transformers) (iid initializat° e.g. Gaussian where we can tune the variance)

To understand what SGD does on regular NNs, one need to study the dynamics

↳ high-dimensional (many parameters)

\+ non-convex

↳ this is hard!

While it is generically intractable to study these dynamics, there has been a lot of interesting work in the past few years.

Today: we will focus on a very simple setting that presents nonetheless a very interesting phenomenology + is highly instructive.

"Online SGD on non-convex losses from high-dim inference"
Ben Arous, Gheissari, Jagannath, 2021.

Setting: $x \sim N(0, I_d)$

$y = f(\langle w_*, x \rangle)$ for some $\|w_*\|_2 = 1$

We learn this data using a "single neuron"

$\hat{f}(x) = f(\langle w, x \rangle)$ using SGD from a random initialization

$w^0 \sim \mathrm{Unif}(\mathbb{S}^{d-1})$

Test error:

$$L(w) = \frac{1}{2}\mathbb{E}\left[\left(f(\langle w_*, x \rangle) - f(\langle w, x \rangle)\right)^2\right]$$

Hermite polynomials & the information exponent

$\{He_k\}_{k=0}^{\infty}$ the orthogonal basis of Hermite polynomials (basis of $L^2(\mathbb{R}, N(0,1))$)

He_k is a degree-k polynomial such that

$$\mathbb{E}_{G\sim N(0,1)}\left[He_k(G)\,He_j(G)\right] = k!\,\mathbb{1}[j=k]$$

$He_0(x)=1 \quad He_1(x)=x \quad He_2(x)=x^2-1 \quad He_3(x)=x^3-3x$

Properties: (i) $\frac{d}{dx}He_k(x) = k\,He_{k-1}(x)$

↓ integration by part

(ii) $\mathbb{E}[He_k(G)\,g(G)] = \mathbb{E}[g^{(k)}(G)]$

(iii) $He_{k+1}(x) = x\,He_k(x) - k\,He_{k-1}(x)$

Any function $f:\mathbb{R}\to\mathbb{R}$ s.t. $\mathbb{E}[f(G)^2]<\infty$ can be decomposed as

$$f(x) = \sum_{k=0}^{\infty} \frac{\mu_k(f)}{k!}\,He_k(x) \quad \text{s.t.} \quad \mu_k(f) = \mathbb{E}[f(G)\,He_k(G)]$$

$$\left(\text{if } k \text{ times diff} = \mathbb{E}[f^{(k)}(G)]\right)$$

Def: [Information exponent] We say that $f \in L^2(\mathbb{R}, N(0,1))$ has information exponent k_* if

$$k_* = \min \{ k \in \mathbb{N} : \mathbb{E}[f(G) He_k(G)] \neq 0 \}$$

Lemma: Let k_* be the info exponent of f. Then

$$L(w) = \frac{1}{2} \mathbb{E}\left[\left(f(\langle w_*, x \rangle) - f(\langle w, x \rangle) \right)^2 \right] = 1 - O(\langle w, w_* \rangle^{k_*})$$

Proof:

$$\frac{1}{2} \| f(\langle w_*, \cdot \rangle) - f(\langle w, \cdot \rangle) \|_{L^2}^2 = 1 - \mathbb{E}[f(\langle w_*, x \rangle) f(\langle w, x \rangle)]$$

$$= 1 - \sum_{k=k_*}^{\infty} \frac{\mu_k^2}{(k!)^2} \mathbb{E}[He_k(\langle w_*, x \rangle) He_k(\langle w, x \rangle)]$$

$$\mathbb{E}[He_k(\langle w_*, x \rangle) He_k(\langle w, x \rangle)] \qquad G_1, G_2 \sim N(0,1)$$

$$= \mathbb{E}\left[He_k(G_1) He_k\left(\langle w, w_* \rangle G_1 + \sqrt{1 - \langle w, w_* \rangle^2}\, G_2 \right) \right]$$

Int by part on G_1 (property (ii)) ↓

$$= \mathbb{E}\left[\frac{d^k}{dG_1^k} He_k\left(\langle w, w_* \rangle G_1 + \sqrt{1 - \langle w, w_* \rangle^2}\, G_2 \right) \right]$$

$$= \langle w, w_* \rangle^k \underbrace{\mathbb{E}[He_k^{(k)}(G)]}_{= k!} \quad \text{(property (i))}$$

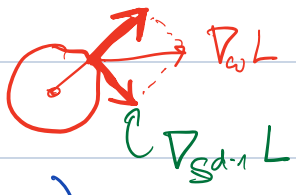
Hence $L(w) = 1 - \sum_{k=k_*}^{\infty} \frac{\mu_k^2(f)}{k!} \langle w_*, w \rangle^k$ □

For simplicity, we will set $f(x) = \frac{He_k(x)}{\sqrt{k!}}$ from now on.

[results will hold for more general f of I.E. k]

Test error: $L(w) = 1 - \langle w, w_* \rangle^k = 1 - m^k \qquad m = \langle w, w_* \rangle$

We minimize over $w \in \mathbb{S}^{d-1}$. We will consider "spherical" gradient descent (gradient $\in$ tangent space of sphere manifold)



$$\nabla_{\mathbb{S}^{d-1}} L(w) = (Id - ww^T) \nabla_w L(w)$$
$$= (Id - ww^T)\left(-k \langle w, w_* \rangle^{k-1} w_*\right)$$
$$= -k\, m^{k-1} (w_* - m\, w)$$

The gradient along the signal direction is

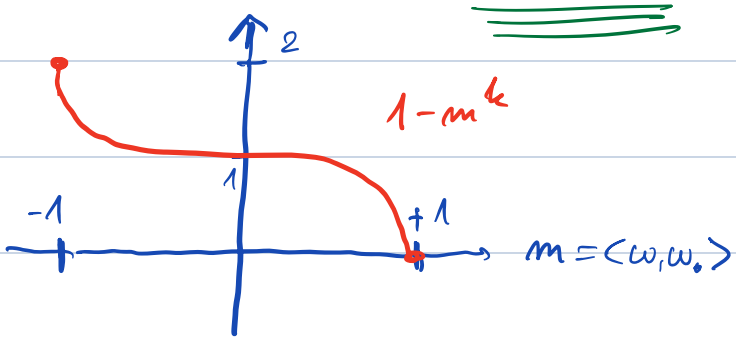
$$\langle w_*, \nabla_{\mathbb{S}^{d-1}} L(w) \rangle = -k\, m^{k-1} (1 - m^2)$$

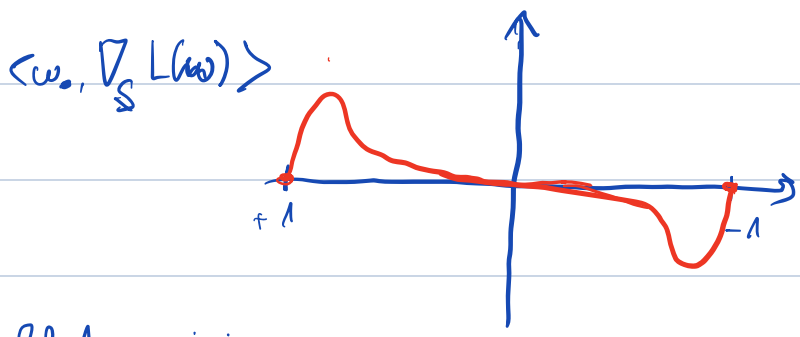
Population landscape

non-convex

If k is odd

$L(m)$





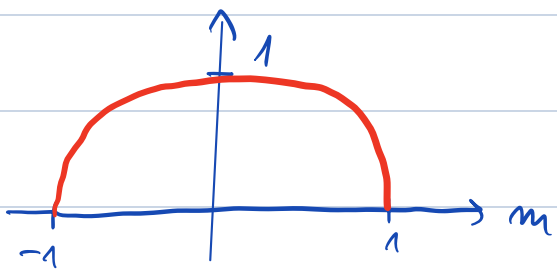
$\omega = \omega_*$ global minimum

$\langle \omega, \omega_* \rangle = 0$ stationary submanifold

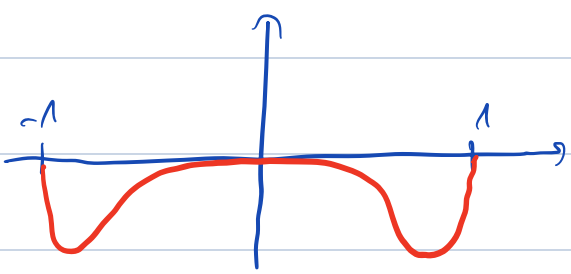
$\omega = -\omega_*$ unstable stationary point (global maxima)

If k is even

$L(m)$



$\langle \omega_*, \nabla_S L(\omega) \rangle$

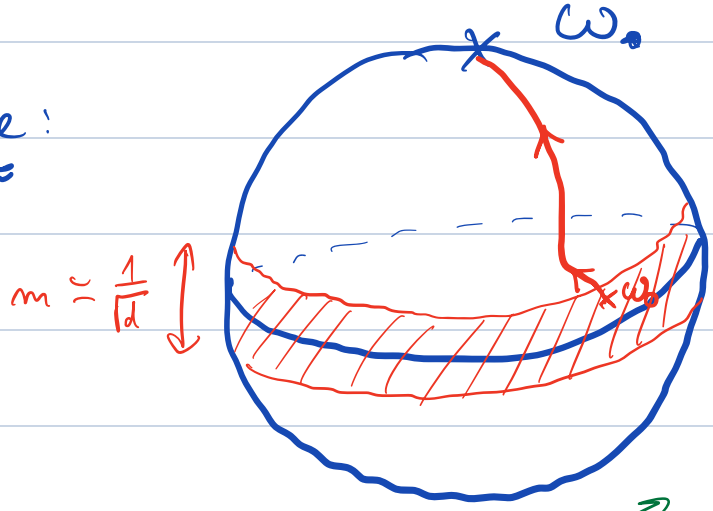


$\omega = \pm \omega_*$ 2 global minima

$\langle \omega, \omega_* \rangle = 0$ stationary submanifold

we will assume $\langle \omega^0, \omega_* \rangle > 0$ (prob $1/2$) and $\omega^t \to +\omega_*$

Lanscape:



w.p. $\geq \frac{1}{2} + \eta$

will assume $\langle w^0, w_* \rangle \gtrsim \frac{c}{\sqrt{d}}$

with $w_* \sim \mathrm{Unif}(S^{d-1})$ w.h.p. $\langle w^0, w_* \rangle \leq \frac{C}{\sqrt{d}}$

At initialization, $\| \nabla_S L(w^0) \|_2 \simeq \frac{1}{d^{\frac{k-1}{2}}}$

$$\langle w_*, \nabla_S L(w^0) \rangle \simeq \frac{1}{d^{\frac{k-1}{2}}}$$

very small gradient: dynamic initialized near a saddle-pt.
SGD will take a long time to escape this generic initialization

We want to study online SGD on this problem and in particular establish # of SGD steps to reach the global minimum.

Gradient flow limit

A standard approach is to approximate SGD by its continuous

limit (step size $\to 0$): $\dot{\omega}^t = -\nabla_S L(\omega^t)$

We only need to track $m_t = \langle \omega^t, \omega_* \rangle$

$$m_0 = \frac{c}{\sqrt{d}} \qquad \dot{m}_t = k (m_t)^{k-1} (1 - m_t^2)$$

As long as m_t small $\quad \dot{m}_t \simeq k (m_t)^{k-1}$

$$\frac{dm_t}{m_t^{k-1}} = k\, dt$$

If $k = 1$: $m_T \simeq m_0 + kT \to$ converge in $\boxed{T = O_d(1)}$

If $k = 2$: $m_T \simeq m_0 e^{kT} \to$ converge in $\boxed{T = O_d(\log d)}$

If $k > 2$: $\frac{1}{k-2}\left[\frac{1}{m_0^{k-2}} - \frac{1}{m_T^{k-2}}\right] = kT \to m_T \simeq \frac{1}{(d^{\frac{k}{2}-1} - kT)^{\frac{1}{k-2}}}$

$\to$ converge in $\boxed{T = d^{\frac{k}{2}-1}}$

Comparing k steps of SGD with step size η, we can show the propagation of error: with high proba ($\sqrt{\frac{T}{\eta}} \times \eta$, T = steps)

$$|L(\underbrace{\omega^k}_{\text{SGD}}) - L(\underbrace{\omega^{k\eta}}_{\text{GF}})| \leq e^{ck\eta}\sqrt{d\eta}$$

For $k = 1$, this shows that SGD with step size $\eta = \Theta_d(\frac{1}{d})$ achieves small test error in $\Theta_d(d)$ steps / samples

[note that $\Omega(d)$ samples is information-theoretic optimal]

For $k > 1$, because of propagation of error, we cannot approximate discrete SGD by GF throughout the entire dynamic.

We need a different analysis!

Projected online SGD

We consider the following algorithm: initialize at ω^0

At each step t:

$$\begin{cases} \tilde{\omega}^{t+1} = \omega^t - \eta \nabla_S \ell(\omega^t, x^t) \\ \omega^{t+1} = \dfrac{\tilde{\omega}^{t+1}}{\|\tilde{\omega}^{t+1}\|_2} \end{cases}$$

) project back on the sphere

$$\ell(\omega^t, x^t) = \frac{1}{2}\left(f(\langle \omega_*, x_t \rangle) - f(\langle \omega, x_t \rangle)\right)^2$$

$(x_t)_{iid} \sim N(0, I_d)$

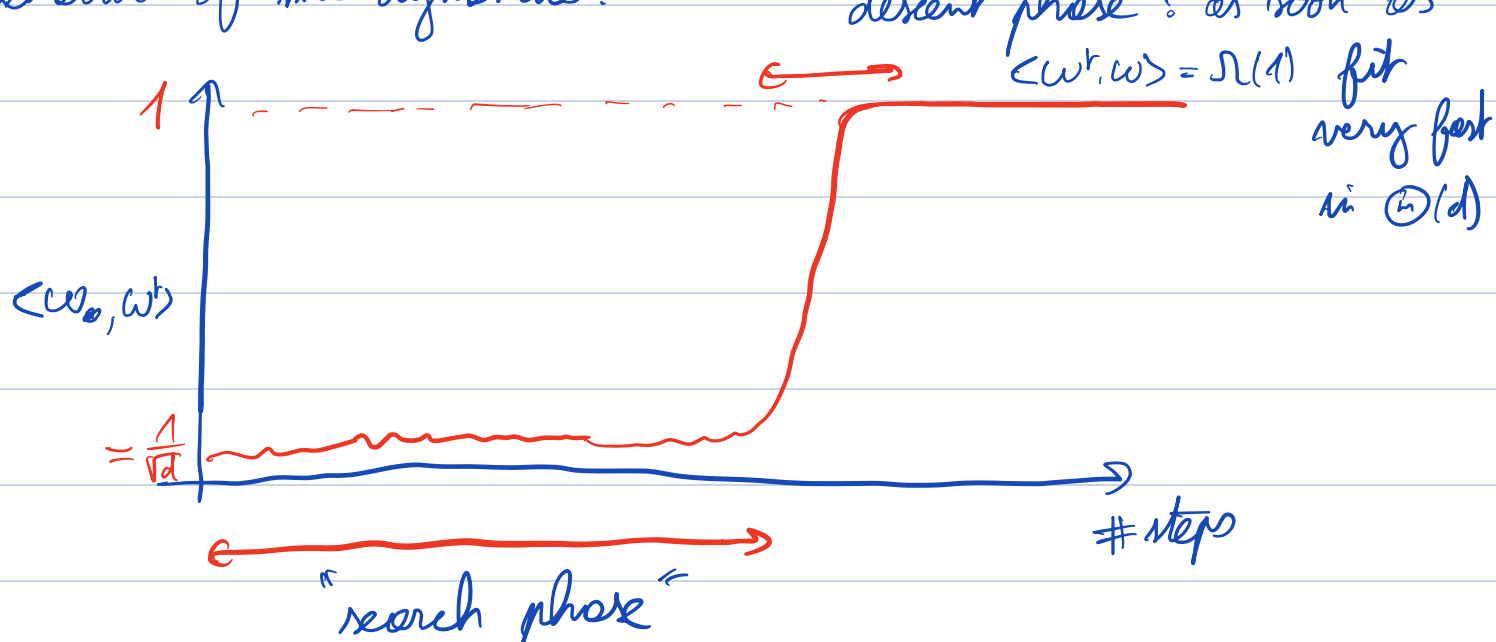
Thm: [Ben Arous, Gheissari, Jagannath, '21]

If we choose

$k=1$	$\eta = \Theta(\frac{1}{d})$	$T = \Theta(d)$
$k=2$	$\eta = \tilde{\Theta}(\frac{1}{d})$	$T = \Theta(d \log d)$
$k>2$	$\eta = \tilde{\Theta}(d^{-\frac{k}{2}})$	$T = \tilde{\Theta}(d^{k-1})$

Then online SGD with step size η and T steps will have $\langle w_*, w^T \rangle \approx 1$ w.h.p.

Behavior of the dynamics:



$\tilde{\Theta}(d^{k-1})$ samples spent to get correlated to the signal.

Main difficulty in this problem is leaving the high entropy region of the initialization. As soon as one has left this region, the "fitting" is fast.

Rmk: Remember the case of learning k-parities. This is the same phenomenon!

Proof of the Theorem

Note that $\eta T \simeq d^{\frac{I}{2}-1}$ which matches the continuous time guarantee from GF

However GF will not be a good approximation of SGD for $\eta T \gg 1$

Instead, we will (approximately) decompose the dynamics into a deterministic drift induced by the signal that drives dynamics to $\longrightarrow w_*$

+ a mean zero noise term.

Dynamics: $$w^{t+1} = \frac{w^t - \eta \nabla_S \ell(w^t, x^t)}{\| w^t - \eta \nabla_S \ell(w^t, x^t) \|_2}$$

Below we will focus on search phase, i.e., getting $\langle w^t, w_* \rangle = \Omega_d(1)$

(descent phase can be studied by GF)

For η small enough, $w^{t+1} \approx w^t - \eta \nabla_S \ell(w^t, x^t)$ (*)

In these notes, we will ignore the projection step and directly consider (*).

$$w^{t+1} = w^t - \eta \nabla_S \ell(w^t, x^t) = w_0 - \eta \sum_{s=0}^{t} \nabla_S \ell(w^s, x^s)$$

It is enough to track $m^t = \langle w^t, w_* \rangle$

$$m_{t+1} = m_0 + \eta \sum_{s=0}^{t} g^s$$

$$g^s := -\langle w_*, \nabla_S \ell(w^s, x^s) \rangle = \left(f(\langle w_*, x^s \rangle) - f(\langle w^s, x^s \rangle)\right) f'(\langle w^s, x^s \rangle) \left(x^s - w^s \langle w^s, x^s \rangle\right)$$

Denote $\bar{g}^s = \mathbb{E}_{x^s}[g^s] = -\langle w_*, \nabla_S L(w^s) \rangle$

$$= k\, m_s^{k-1} (1 - m_s^2)$$

Then

$$m_{t+1} = \underbrace{m_0}_{= \frac{c}{\sqrt{d}}} + \underbrace{\eta \sum_{s=0}^{t} \bar{g}^s}_{D^t:\ \text{"deterministic drift"}} + \underbrace{\eta \sum_{s=0}^{t} g^s - \bar{g}^s}_{M^t:\ \text{martingale}}$$

M^t is a martingale $\left(M^t = \sum_{s=0}^{t} X_s \qquad \mathbb{E}[X_t \mid X_{\le t-1}] = 0\right)$

① Controlling the martingale:

We can use Doob's maximal inequality: for every $p \ge 1$,

$$\mathbb{P}\left(\max_{0 \le t \le T} |M_t| \ge \lambda\right) \le \frac{p\, \mathbb{E}[|M_T|^p]}{(p-1)\,\lambda^p}$$

One can show [using hypercontractivity of low-degree polynomials on Gaussian data $\|f\|_{L^q}^2 \le (q-1)^k \|f\|_{L^2}^2$ for any degree k polynomial f]

$$\mathbb{E}[|M_T|^p] \lesssim C_p\, \eta^p\, T^{\frac{p}{2}}$$

Hence, with prob $1-\delta$,

$$\frac{C \eta^2 T}{\lambda^2} = \delta$$

$$\max_{0 \le t \le T} |M_t| \le C \eta \sqrt{T}\, \delta^{-\frac{1}{2}}$$

dependency can be improved to $\log\left(\frac{1}{\delta}\right)$

In particular if $C\eta\sqrt{T} \leq \frac{m_0}{2}$

that is $\boxed{\eta\sqrt{T} = O\left(\frac{1}{\sqrt{d}}\right)}$ (A1)

Then: $\frac{m_0}{2} + D^t \leq m_t \leq \frac{3}{2}m_0 + D^t$ for all $t \leq T$.

and we can neglect the Martingale / noise term

(2) Controlling the drift term

In the search phase (m^t small) $\bar{g}^s \simeq k\, m_s^{k-1}$

Hence for all $t \leq T$ (under (A1))

$$m_t \geq \frac{m_0}{2} + \eta k \sum_{s=0}^{t-1} m_s^{k-1}$$

We can study this sequence and lower bound their growth. The Lemma below can be seen as a discrete lower bound version of Bihari-LaSalle inequality.

Lemma: For a sequence $(u_t)_{t \in \mathbb{N}}$ that satisfies

$$u_t \geq a_0 + a_1 \sum_{s=0}^{t-1} u_s^{k-1}$$

Then u_t is lower bounded by

$k=2$: $u_t \geq a_0 (1 + a_1)^t$

$k>2$: $u_t \geq \min\left(1, \frac{1}{\left(a_0^{-(k-2)} - \frac{k-2}{(1+a_1)} a_1 t\right)^{\frac{1}{k-2}}}\right)$

For $k \geq 2$: $m_0^{-(k-2)} \simeq d^{\frac{k}{2}-1}$

Hence $m_T = \Omega(1)$ as long as

$$\eta T = \Omega\left(d^{\frac{k}{2}-1}\right)$$ (A2)

In fact, if $t = o\left(d^{\frac{k}{2}-1}\right)$ then $m_t \simeq \frac{1}{\sqrt{d}}$

↳ most of the search phase $m_t \simeq \frac{1}{\sqrt{d}}$

Combining (A1) and (A2)

$$\eta\sqrt{T} = \widetilde{\Theta}\left(\frac{1}{\sqrt{d}}\right) \qquad \eta T = \widetilde{\Theta}\left(d^{\frac{k}{2}-1}\right)$$

$$\Rightarrow \boxed{T = \widetilde{\Theta}\left(d^{k-1}\right) \qquad \eta = \widetilde{\Theta}\left(d^{-\frac{k}{2}}\right)}$$

for $k > 2$

For $k = 2$, then $u_T \gtrsim m_0 e^{\eta T}$

$$\rightarrow \eta T = \Theta(\log d)$$

$$\Rightarrow \boxed{T = \widetilde{\Theta}(\log d) \qquad \eta = \widetilde{\Theta}\left(d^{-1}\right)}$$

Landscape concentration

The above guarantees are for online SGD

that is # samples = # steps

What about if we reuse samples?

Batch 1 SGD choose one random sample from the empirical distribution $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$

We can do the same drift + martingale decomposition but now

$$\bar{g}^s \longrightarrow \hat{g}_n^s = \langle w_*, \nabla_s \hat{\mathbb{E}}_n[\ell(w^s, x)] \rangle$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left(y_i - f(\langle w^s, x_i \rangle)\right) f'(\langle w^s, x_i \rangle)\left(x_i - w^s \langle w^s, x_i \rangle\right)$$

If we can show that $\hat{g}_n(w) \geq c\, \bar{g}(w)$

for all $w \in \mathbb{S}^{d-1}$ $\langle w, w_* \rangle \geq \frac{c}{\sqrt{d}}$

then the same analysis as for one-pass SGD goes through

For a given $w \in \mathbb{S}^{d-1}$, $\hat{g}_n(w)$ is an empirical average with mean $\bar{g}(w)$

For simplicity take f with $\|f\|_\infty, \|f'\|_\infty \le C$.

$$\mathbb{P}\left(\left| \hat{g}_n(w) - \bar{g}(w) \right| \ge \sqrt{\mathrm{Var}(\hat{g}_n(w))}\, t \right) \le C e^{-ct^2}$$

Hence $$\mathbb{P}\left(\min_{\substack{w \in S^{d-1} \\ \langle w, w_* \rangle \ge d^{-\gamma}}} \hat{g}_n(w) - \bar{g}(w) \ge \sqrt{c \frac{d \log d}{n}} \right) \le \frac{C}{d^C}$$

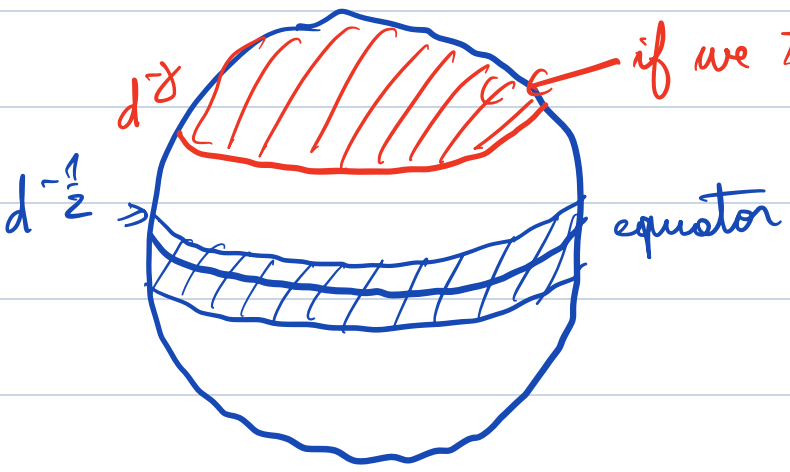
[using a standard uniform convergence bound]

Hence for $\frac{1}{2} \le \gamma \le 0$

then if we take $\sqrt{\frac{d}{n}} \le \frac{1}{d^{\gamma(k-1)}}$

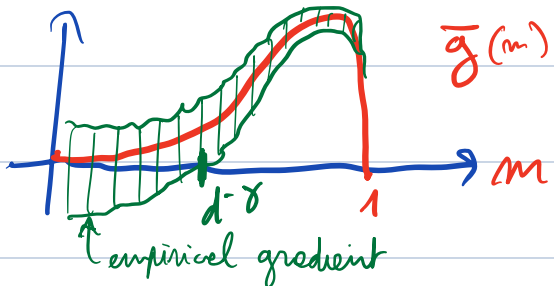
$$\Rightarrow \quad n \ge d^{2\gamma(k-1)+1}$$

then the landscape is "benign" for any $\langle w, w_* \rangle \ge d^{-\gamma}$



if we initialize on this cap, then

$w^t \to w_*$



If $\gamma = \frac{1}{2}$, $\boxed{n \gtrsim d^k}$ random initialization falls in this cap w.h.p.

and $w_t \longrightarrow w_*$ in d^{k-1}

↳ so in fact see each sample only once during the dynamics w.h.p.
and we didn't gain anything compared to online SGD

Open question: can we show that batch-1 multi-pass SGD still succeeds with high-probability for $n \ll d^{k-1}$ despite the landscape being not benign?

How tight are these guarantees?

Online SGD requires $n = T \gtrsim d^{k-1}$ samples

and runtime $Td \gtrsim d^k$ (each step compute a d-dim gradient)

* for parity fcts: SQ lower bound is $\Omega(d^k)$
  → match best SQ runtime

* for Hermite-k: CSQ lower bound is $\Omega(d^{k/2})$

  ↳ modifying the dynamics, online SGD succeeds with $n = T \simeq d^{k/2}$ (runtime $d^{\frac{k}{2}+1}$)

  [Damian, Nichani, Ge, Lee, '23]

* for Hermite-k: SQ lower bound is much smaller $\Omega(d)$

  ↳ taking L^1 loss $\ell(y, \hat{y}) = |y - \hat{y}|$

  online SGD succeeds in $n \simeq T \simeq d$

  ↳ reusing samples with L^2 loss

full-batch GD with $n = \Theta(d)$ and $T = \Theta_d(1)$ succeeds

* There exists single index functions that seem to have a fundamental statistical/computational trade-off

$n = \Theta(d)$ is info-theoretically optimal

$n = \Omega(d^{k/2})$ samples is necessary (and sufficient) for a polynomial-time algo to exists.

(under low-degree poly conjecture)

[See Kostas and Leda presentation !!!]