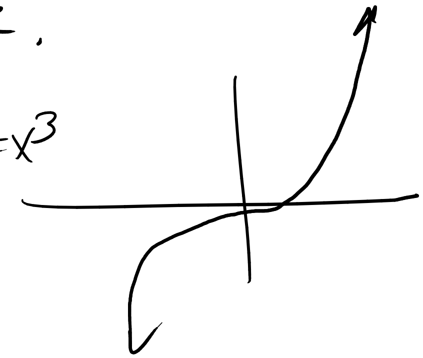
Non-convex optimization.

When objective is non-convex, generally the best we can hope for is to find a stationary point: $\|\nabla f\| = 0$, or an ε-appx stationary point, $\|\nabla f\| \leq \varepsilon$.

Stationary points need not be local min: $y = x^3$



Recall Lemma 7 from last class:

Lemma 7: Let f be β-smooth (not nec. cvx). If $\alpha_t = \alpha < \frac{1}{\beta}$, then $\forall T \geq 1$ G.D. satisfies

$$\min_{t<T} \|\nabla f(w_t)\|^2 \leq \frac{2}{\alpha T}\left(f(w_0) - f(w_T)\right)$$

So, GD on smooth objectives efficiently finds stationary points.

There are many problems where all stationary points are global optima.

Let's consider a few: Let $\mathcal{X}^* := \{x^* : f(x^*) = \min_x f(x)\}$.

Let $\Pi(x) :=$ projection of x onto $\mathcal{X}^*$.

① (Strong) convexity: for some $\mu \geq 0$ ($\mu > 0$: strong),

$\forall x, y,$ $$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|x - y\|^2$$

② weak strong convexity: for some $\mu > 0$, if $f^* = \min_x f(x)$,

$\forall x,$ $$f^* \geq f(x) + \langle \nabla f(x), \pi(x) - x \rangle + \frac{\mu}{2} \|x - \pi(x)\|^2.$$

③ Restricted secant inequality:

$\forall x,$ $$\langle \nabla f(x), x - \pi(x) \rangle \geq \mu \|x - \pi(x)\|^2.$$

④ Error bound inequality:

$\forall x,$ $$\|\nabla f(x)\| \geq \mu \|x - \pi(x)\|.$$

⑤ Polyak-Lojasiewicz: (PL)

$\forall x,$ $$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu \cdot (f(x) - f^*)$$

Clear that PL $\Rightarrow$ all local min are global. However:

<u>Thm</u> (Karimi et al '16): If f is β-smooth, then:

$$① \Rightarrow ② \Rightarrow ③ \Rightarrow (④ \Leftrightarrow ⑤).$$

So, PL inequality is quite crucial.

<u>Thm</u>. Suppose f is β-smooth, and $\exists x^*$ st $f(x^*) = \min_x f(x)$.

Then GD w/ $\alpha \leq \frac{1}{\beta} \wedge \frac{1}{\mu}$ $f(x_k) - f^* \leq (1-\alpha\mu)^k \cdot (f(x_0) - f^*)$.

<u>Pf</u>. By defn, $x_{k+1} = x_k - \alpha \nabla f(x_k)$.

By β-smoothness,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\beta}{2}\|x-y\|^2 \quad \forall x, y.$$

$$\Rightarrow f(x_{k+1}) - f(x_k) \leq \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{\beta}{2}\|x_{k+1} - x_k\|^2$$

$$= -\alpha\|\nabla f(x_k)\|^2 + \frac{\beta\alpha^2}{2}\|\nabla f(x_k)\|^2$$

$$\leq -\frac{\alpha}{2}\|\nabla f(x_k)\|^2 \quad \text{as } \alpha \leq \frac{1}{\beta}.$$

$$\leq -\alpha\mu(f(x_k) - f^*) \quad \text{by PL.}$$

$$\Rightarrow f(x_{k+1}) - f^* \leq (1-\alpha\mu) f(x_k) + \alpha\mu f^* - f^*$$

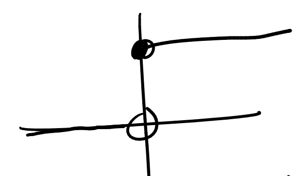
$$= (1-\alpha\mu) f(x_k) - (1-\alpha\mu) f^* = (1-\alpha\mu)(f(x_k) - f^*).$$

$\Rightarrow$ unrolling, we see $f(x_k) - f^* \leq (1-\alpha\mu)^k (f(x_0) - f^*)$. $\blacksquare$

So, with PL on smooth objectives, can get "linear convergence"
to get $f(w_t) - f^* \le \varepsilon$, only need $O(\log \frac{1}{\varepsilon})$ iterations.

Exercise Show that there exists a β-smooth convex function on $\mathbb{R}$ for which PL is not satisfied.

So establishing PL is clearly very powerful. Unsurprisingly, it often does not hold, and if it does, can be very hard to show.
Moreover, for many problems of interest, we do not have β-smoothness. eg, neural networks with ReLU activations,
since $\varphi(t) = \max(0,t)$ has $\varphi'(t) = \mathbb{1}(t \ge 0)$ (subderivative).
Clearly φ' is not Lipschitz, so φ is not β-smooth.



We'll next describe some alternative non-convex approaches which
1. allow for non-smooth objectives, &/or
2. allow for situations where "all local min are good" (but not nec. global opt.)

Surrogate objectives.

In classification, we are most interested in the 0-1 loss:

$$\mathbb{1}(y \neq f(x)) = \begin{cases} 1: & y \neq f(x), \\ 0: & y = f(x) \end{cases} = \mathbb{1}(y \cdot f(x) < 0)$$

for $y, f(x) \in \{-1, 1\}$

[In general: $\mathbb{1}(A) = \begin{cases} 1 & \text{if } A \text{ occurs}, \\ 0 & \text{otherwise.} \end{cases}$]

Unfortunately, $\mathbb{1}(\cdot)$ is discontinuous, piecewise constant, so its derivative is zero a.e $\rightarrow$ gradient descent is useless.

So instead, we minimize convex surrogates for the 0-1 loss:

Call $\ell(t)$ a convex surrogate for $g(t)$ if:

1. $\cdot$ ℓ is convex
2. $\cdot$ $g(t) \leq \ell(t)$ $\forall t$.

Thus, minimizing ℓ implies minimizing g.

We will typically also require ℓ is differentiable/smooth so we can apply G.D. to ℓ.

Common convex surrogates to 0-1 loss $\mathbb{1}(t < 0) = \begin{cases} 1: & t < 0, \\ 0: & t \geq 0 \end{cases}$:

- $\exp(-t)$ "exponential loss"
- $\max(0, 1-t)$ "hinge loss"

- $\log(1 + \exp(-t))$ "logistic / cross entropy"

(note ② not satisfied, but $\mathbb{1}(t < 0) \leq 2 \cdot \log(1 + \exp(-t))$)

More generally, there are a large family of surrogates to 0-1 loss.

Let $\ell(t)$ be convex, decreasing, and twice differentiable. Then:

Assume $\ell'(0) \neq 0$.

① $\ell''(t) \geq 0$ (so $-\ell'$ is decreasing)

② $-\ell'(t) \geq 0$.

Then: $\{t < 0\} = \{-\ell'(t) \geq -\ell'(0)\}$ by ①

If we can show $-\ell'(t) \to 0$, then by

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y_i f(x_i)) = \frac{1}{n}\sum_{1}^{n} \mathbb{1}(y_i \cdot f(x) < 0)$$

$$= \hat{\mathbb{P}}(y \cdot f(x) < 0)$$

$$= \hat{\mathbb{P}}(-\ell'(y \cdot f(x)) \geq -\ell'(0))$$

$$\leq \frac{\hat{\mathbb{E}}(-\ell'(y \cdot f(x)))}{-\ell'(0)} \quad \text{by Markov \& } \ell'(0) \neq 0.$$

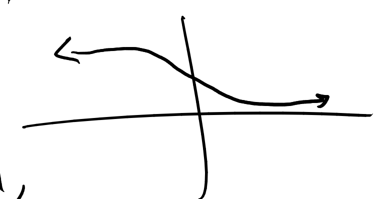
$$= \frac{1}{-\ell'(0)} \cdot \frac{1}{n}\sum_{1}^{n} -\ell'(y_i f(x_i))$$

Thus, for <u>any</u> convex, decreasing, twice differentiable ℓ st $\ell'(0) \neq 0$,
$-\ell'(t)$ is surrogate for 0-1 loss.

① $\ell(t) = \exp(-t) \Rightarrow -\ell'(t) = \ell(t)$.

② $\ell(t) = \log(1+\exp(-t)) \Rightarrow -\ell'(t) = \dfrac{\exp(-t)}{1+\exp(-t)} = \dfrac{1}{1+\exp(t)}$.

Here $-\ell'$ is <u>not</u> convex.



However, if we do G.D. on ℓ (which is convex),
i.e. G.D. on $\hat{L}(W) := \frac{1}{n}\sum_i^n \ell(y_i f(x_i; W))$,
and we somehow found that the loss
$\hat{G}(W) := \frac{1}{n}\sum_i^n -\ell'(y_i f(x_i; W))$ were small, <u>then</u> we
would know that the 0-1 loss is small as well.
(In the homework, we'll see why this is useful.)

Def. $f: \mathbb{R}^p$ is called (g,h,v)-proxy convex if, $\exists\, g,h: \mathbb{R}^p \to \mathbb{R}$ st $\exists v$ s.t. $\forall w$,

$$h(v) \geq g(w) + \langle \nabla f(w), v - w \rangle.$$

Thm: If f is (g,h,v)-proxy-convex and $\|\nabla f(w)\| \leq L_1$ for all w, Then for any $w(0)$, $\eta > 0$, $T \geq 1$,

$$\min_{t<T} g(w_t) \leq h(v) + \frac{\eta L_1^2}{2} + \frac{\|w(0) - v\|^2}{2\eta T}.$$

Remark. For small step sizes η and large T, G.D. on f finds $g(w_t)$ which is close to $h(v)$.

useful if we can show there are families of related objective functions $\{g, h\}$ for which are "meaningful". i.e., $v = \arg\min f$,

$h(w) = \sqrt{f(w)}$ and $g(w) = \frac{1}{2} f(w)$ would imply

$$\min_{t<T} \frac{1}{2} f(w_t) \leq \min_{w} \sqrt{f(w)} + \varepsilon.$$

Not a global min, but a fcn of it. If $\min_w f(w)$ is very small, then good.

In remainder, we will prove that GD converges to a global min for 2-layer leaky ReLU nets on linearly sep. data.

Two proof techniques:

(1) Proxy convexity] lecture

(2) Perceptron argument.] homework

2-layer leaky nets: $w_j \in \mathbb{R}^d$, $W \in \mathbb{R}^{m \times d}$ (rows w_j), $a_j \in \{\pm 1/\sqrt{m}\}$, m = # neurons. $\varphi(t) = \max(t, \alpha t)$, $\alpha \in (0,1]$

$$f(x; W) = \sum_{j=1}^{m} a_j \varphi(\langle w_j, x \rangle).$$

Only training first layer.

$\ell(t) = \log(1+\exp(-t))$; $\|x_i\| \le 1$

$\exists v^* \in \mathbb{R}^d$ st. $\|v^*\| = 1$, $\forall i$, $\langle y_i x_i, v^* \rangle \geq \gamma$, $\gamma \in (0,1]$.

G.D. on empirical risk $\hat{L}(W) = \frac{1}{n} \sum_{i}^{n} \ell(y_i f(x_i; W))$.

Remark We will consider parameterizing the objective fcn $\hat{L}(W)$ by a matrix W, while we typically think of vectors in $\mathbb{R}^d$ for some d.
You can cast the problem to that form by considering $vec(W) \in \mathbb{R}^{d \cdot m}$: stack each column of W one after another.
Everything works out — can check via matrix calculus.

- $\langle W, V \rangle := tr(W^T V) \quad \left(= tr(V^T W) = tr(W V^T)\right)$, $W, V \in \mathbb{R}^{m \times d}$

$$= \langle vec(W), vec(V) \rangle, \quad vec(W), vec(V) \in \mathbb{R}^{dm}$$

- $\|W\|_F^2 = \|vec(W)\|_2^2$

Lemma. If $g:\mathbb{R}^d \to \mathbb{R}$ is L-positively homogeneous, ($g(\alpha w) = \alpha^L g(w)$ for $\alpha \geq 0$)
then $\langle \nabla g(w), w \rangle = L g(w)$

pf: Exercise.

Remark: $\varphi(t) = \max(t, \alpha t)$ is 1-positively homogeneous.

Our goal will be to show $\hat{L}(w)$ is proxy convex for appropriate proxies g, h, i.e. want to show, for some g, h,

$$\langle \nabla \hat{L}(W), w - V \rangle \geq g(W) - h(V).$$

$$\nabla \hat{L}(W) = \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(y_i f(x_i; W))$$

$$= \frac{1}{n} \sum_1^n \ell'(y_i f(x_i; W))\, y_i \nabla f(x_i; W).$$

Now: $f(x_i; W) = \sum_j a_j \varphi(\langle w_j, x \rangle)$ is 1-homog. since φ is. So,

$$\langle \nabla f(x_i; W), W \rangle = f(x_i, W).$$

Thus: $\boxed{\begin{aligned}\langle \nabla \hat{L}(W), W-V\rangle &= \frac{1}{n}\sum_1^n \ell'(y_i f(x_i;W))\, y_i \langle \nabla f(x_i;W), W-V\rangle \\ &= \frac{1}{n}\sum_1^n \ell'(y_i f(x_i;W)) \cdot \big(y_i f(x_i;W) - y_i\langle \nabla f(x_i;W), V\rangle\big)\end{aligned}}$

So now, which V do we want to choose?

Calculus $\Rightarrow$ $\nabla f(x;W) = D_x^W a x^T$, $D_x^W = \mathrm{diag}(\varphi'(\langle w_j, x\rangle))$.

$\in \mathbb{R}^{m \times d}$

$$\begin{aligned} y_i \langle \nabla f(x_i;W), V\rangle &= y_i \cdot \mathrm{tr}\big((D_{x_i}^W a x_i^T)^T V\big) \\ &= y_i\, \mathrm{tr}(x_i a^T D_{x_i}^W V) \\ &= y_i\, \mathrm{tr}(a^T D_{x_i}^W V x_i) \\ &= y_i\, a^T D_{x_i}^W V x_i \\ &= \sum_j a_j \varphi'(\langle w_j, x_i\rangle)\langle v_j, y_i x_i\rangle. \end{aligned}$$

know: $\langle v^*, y_i x_i\rangle \geq \gamma > 0$. So let $v_j := \rho \cdot a_j \cdot v^*$, $\rho > 0$.

Since $\varphi'(t) \geq \alpha > 0$ $\forall t$,

$$y_i \langle \nabla f(x_i;W), V\rangle = \sum_{j=1}^m a_j \varphi'(\langle w_j, x_i\rangle) \cdot \langle v_j, y_i x_i\rangle = \sum_{j=1}^m a_j^2 \cdot \rho \cdot \varphi'(\langle w_j, x_i\rangle)\langle v^*, y_i x_i\rangle \geq \frac{\rho}{m}\sum_{j=1}^m \alpha\gamma = \rho\alpha\gamma.$$

Continuing from $\boxed{}$: Since ℓ is decreasing, $-\ell'(z) < 0$. So,

$$\boxed{-\ell'(y_i f(x_i;W)) \cdot y_i \langle \nabla f(x_i;W), V \rangle \geq -\ell'(y_i f(x_i;W)) \cdot \rho \alpha \gamma.}$$

Thus,

$$\langle \nabla \hat{L}(W), W - V \rangle \geq \frac{1}{n} \sum_1^n \ell'(y_i f(x_i;W)) \cdot (y_i f(x_i;W) - \rho \alpha \gamma).$$

since ℓ is convex, $\quad \ell'(z_1) \cdot (z_1 - z_2) \geq \ell(z_1) - \ell(z_2)$.

$$\geq \frac{1}{n} \sum_1^n \left[\ell(y_i f(x_i;W)) - \ell(\rho \alpha \gamma) \right].$$

$$= \hat{L}(W) - \ell(\rho \alpha \gamma).$$

Therefore, $\hat{L}(W)$ is $(\hat{L}, \ell(\rho \alpha \gamma), V)$-proxy convex with $V = V(\rho)$ having rows $v_j = \rho a_j v^*$.

Since $\|\nabla f(x_i;W)\|_F = \|\mathrm{Diag}(\varphi'(\langle w_j, x_i \rangle)) a x_i^T\|_F$

$$\leq 1 \quad (\underline{\text{Exercise}}!)$$

Thus the theorem guarantees that for any $T \geq 1$, $w(0)$, $\eta > 0$, $\varepsilon > 0$,

$$\min_{t<T} \hat{L}(W_t) \leq \ell(\rho \alpha \gamma) + \frac{\eta}{2} + \frac{\|W(0) - V\|_F^2}{2\eta T}.$$

Since $\ell(t) \leq 2\exp(-t)$, taking $\rho = \alpha^{-1}\gamma^{-1}\log\left(\frac{6}{\varepsilon}\right)$ guarantees $\ell(\rho\alpha\gamma) \leq \frac{\varepsilon}{3}$.

For $\eta \leq \frac{\varepsilon}{2}$, this means

$$\min_{t<T} \hat{L}(W_t) \leq \frac{2\varepsilon}{3} + \frac{\|W(0) - V\|_F^2}{2\eta T}.$$

For $T \geq 2\eta^{-1}\|W(0) - V\|_F^2 \varepsilon^{-1}$, $\min_{t<T} \hat{L}(W_t) \leq \varepsilon$. $\square$

All together:

<u>Theorem</u>. Suppose $\|x_i\| \leq 1$, $y_i \langle v^*, x_i \rangle \geq \gamma$ for all i for some $\|v^*\| = 1$. Then GD on 2-layer leaky ReLU net satisfies the following. For any $\varepsilon > 0$, if $\eta \leq \varepsilon/2$ then for $V \in \mathbb{R}^{m \times d}$ with rows $v_j = a_j \alpha^{-1}\gamma^{-1}\log\left(\frac{6}{\varepsilon}\right)$, if $T \geq 2\eta^{-1}\|W(0) - V\|_F^2 \varepsilon^{-1}$, then $\min_{t<T} \hat{L}(W_t) \leq \varepsilon$.

Remark I'm not aware of proofs with any of:

① - φ is ReLU rather than leaky ReLU

② - bias terms included

③ - training both first and second layer weights

④ - training > 2 layers net.

difficulty: 4 > 1 > 3 > 2.

If you can come up with a proof in any of these settings, this would probably be sufficient for a NeurIPS paper! (except maybe ②) Requires:

1. Random initialization
2. Only assume linear separability of training data.
3. No "NTK". Ideally allow for constant number of neurons.

You can work on this instead of choosing paper for final project if you'd like.