Lecture 10

Last time
- Strict saddles
- Measure zero sets
- Escaping saddles

Today
- Missing lemmas
- SGD
- Expected descent

Missing lemmas

Let's finish the proofs from last time.

Lemma (Null): The set S has zero measure.

Proof: The key idea is to show that two points separated by a negative-curvature direction cannot both stay in K. Recall that at $w_\# = 0$,

$$\nabla T(w_\#) = \begin{pmatrix} I_m - \eta D_- & \\ & I_{d-m} - \eta D_+ \end{pmatrix}.$$

Consider four points

$w \in (u, v) \in K$, $\tilde{w} \in (\tilde{u}, \tilde{v}) \in K$, $T(w) = (u^+, v^+)$, & $T(\tilde{w}) = (\tilde{u}^+, \tilde{v}^+)$

($u \in \mathbb{R}^m$)

Plus, the four distances

$U = \|u - \tilde{u}\|$, $V = \|v - \tilde{v}\|$, $U^+ = \|u^+ - \tilde{u}^+\|$ & $V^+ = \|v^+ - \tilde{v}^+\|$.

$U^+ \geq (1+\frac{\eta\gamma}{2})U$,

Claim: Suppose $U \geq V$, then $U^+ \geq V^+$.

Proof of the claim: By the fundamental theorem of calculus

$$T(w) - T(\tilde{w})$$
$$= \nabla T(0)(w-\tilde{w}) + \int_0^1 [\nabla T(\tilde{w}+t(w-\tilde{w})) - \nabla T(0)](w-\tilde{w})\,dt.$$

This segment is in K.

$e = (e_u, e_v)$

By the definition of K,

$$\|e\| \leq \int_0^1 \|\nabla T(\tilde{w}+t(w-\tilde{w})) - \nabla T(0)\| \, \|w-\tilde{w}\| \, dt$$
$$\leq \frac{\eta\gamma}{4} \|w-\tilde{w}\|$$
$$\leq \frac{\eta\gamma}{4} (U+V)$$
$$\leq \frac{\eta\gamma}{2} U.$$

Then,

$$
\begin{aligned}
U^+ &= \|u^+ - \tilde{u}^+\| \\
&= \|(I - \eta D_-)(u - \tilde{u}) + e_u\| \\
&\geq (1 + \eta\gamma)\|u - \tilde{u}\| - \|e_u\| \qquad \lambda_{\min}(I - \eta D_-) \geq (1+\eta\gamma) \text{ \& reverse triangle} \\
&\geq (1+\eta\gamma)\, U - \frac{\eta\gamma}{2} U \\
&\geq \left(1 + \frac{\eta\gamma}{2}\right) U \\
&\geq V + \|e\| \qquad \|I - \eta D_+\|_{op} \leq 1 \text{ \& triangle} \\
&\geq \|(I - \eta D_+)(v - \tilde{v}) + e_v\| \\
&= V^+.
\end{aligned}
$$

□

Fix $q \in \mathbb{R}^m \setminus \{0\}$, we will show that for any $x \in \mathbb{R}^d$, the line $x + \text{span}\{(q, 0)\}$ can only intersect S in at most one point. Suppose seeking contradiction suppose there are two points $w_0, \tilde{w}_0 \in S$

$(t_q, 0)$

such that $U_0 = \|w_0 - \tilde{w}_0\| > 0 = V_0$. Then, inductively applying the claim, we get that $\|w_k - \tilde{w}_k\| \geq U_k \geq (1 + \frac{\eta\gamma}{2})^k U_0 \to \infty$. Thus one of the trajectories has to leave K, which contradicts $w_0, \tilde{w} \in S$. The conclusion follows by Fubini's. $\square$

Lemma (Image of a null): Suppose $F: \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz, and $\mathcal{A}$ has zero measure. Then $F(\mathcal{A})$ has zero measure. $\dashv$

Proof: Let $\varepsilon > 0$, and let $B_{r_1}(c_1), B_{r_2}(c_2), \ldots$ a covering of $\mathcal{A}$ with volume at most $\frac{\varepsilon}{(2L)^d}$. Discard any ball that doesn't intersect $\mathcal{A}$. For all j, pick $w_j \in \mathcal{A} \cap B_{r_j}(c_j)$. Then,

$\|F(w) - F(w_j)\| \leq L\|w - w_j\|$

$w \in \mathcal{A} \cap B_j(c_j)$

$$\leq L\left(\|w - c_j\| + \|c_j - w_j\|\right)$$

$$\leq 2L r_i.$$

Then, $\{B(F(w_j), 2Lr_i)\}$ covers $F(\mathcal{A})$, and

$$\sum_{j=1}^{\infty} \mathrm{vol}\left(B(F(w_j), 2Lr_i)\right) = (2L)^d \sum \mathrm{vol}\left(B(c_j, r_j)\right)$$

$$\leq \varepsilon. \qquad \square$$

Stochastic Gradient Descent

Today, we shift gears a bit and start talking about random problems. Suppose there is a distribution $\mathcal{D}$ over some space $\mathcal{Z}$ (We will avoid worrying to much about measure theoretic details), and instead of having an exact gradient oracle $w \longmapsto \nabla f(w)$ we have access to a noisy oracle

$$w \longmapsto g(w, z), \quad z \sim \mathcal{D}.$$

We will make the following assumption

(Unbiased) $\nabla f(w) = \mathbb{E}\, g(w, z)$.

This suggests a natural iterative method

(SGD) Draw a fresh sample z_k,
update $w_{k+1} \leftarrow w_k - \eta_k\, g(w_k, z_k)$.

There are a number of concrete examples.

Example (Coordinate descent): Imagine d is too large, so instead of computing all coordinates of $\nabla f(\cdot)$, we use

$$i \sim \mathrm{Unif}[d]$$

$$g(w, i) = d\, \partial_i f(w)\, e_i.$$

In that case,

$$\mathbb{E}\, g(w, i) = \frac{1}{d} \sum_{i=1}^{d} d\, \partial_i f(w)\, e_i = \nabla f(w).$$

Example (Finite sums): Suppose we can

write our loss as

$$f(w) = \frac{1}{n} \sum_{j=1}^{n} f_j(w).$$

← All of ML training.

Then, we can write

$$j \sim \text{Unif}([n]),$$

$$g(w, j) = \nabla f_j(w).$$

← What is the advantage over taking a full gradient?

Once more

$$\mathbb{E}\, g(w, j) = \frac{1}{n} \sum_{j=1}^{n} \nabla f_j(w) = \nabla f(w).$$

→

Example (stochastic Programming):

Suppose we can write our loss as

$$f(w) = \mathbb{E}_{z \sim D} f(w, z).$$

Define

$$g(w, z) = \nabla f(w, z).$$

Under mild assumptions (Leibniz rule)

$$\mathbb{E}\, g(w, z) = \nabla f(w).$$

→

Conditional expectation

Within the analysis, we will use the notion of conditional expectation of the samples we have seen

$$\mathcal{F}_k = \sigma(w_0, z_0, \dots, z_{k-1}).$$

That is $\mathbb{E}[\cdot \mid \mathcal{F}_k]$. We will use three properties. Given r.v. $X_1, \dots, X_n$ and $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ we have

(Linearity) $\mathbb{E}[\sum \alpha_i X_i \mid \mathcal{F}_k] = \sum \alpha \mathbb{E}[X_i \mid \mathcal{F}_k]$.

(Tower) $\mathbb{E}[\mathbb{E}[X_1 \mid \mathcal{F}_k]] = \mathbb{E}[X_1]$.

Expectation over r.v. in $\mathcal{F}_k$.

Suppose a r.v. Y is determined by $\mathcal{F}_k$,

(Pull-out) $\mathbb{E}[YX_1 \mid \mathcal{F}_k] = Y\mathbb{E}[X_1 \mid \mathcal{F}_k]$.

Expected descent

As a first step towards proving a guarantee, we establish descent under the following assumption.

Assumption (SGD): The function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and $f^* = \inf f > -\infty$. Further, $g : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}^d$ satisfies (unbiased) and

$$\mathbb{E}_z\left[\|g(w,z) - \nabla f(w)\|^2\right] \le \sigma^2 \quad \forall w \in \mathbb{R}^d.$$

Lemma (ED). Suppose Assumption (SGD) holds. Then, at each k we have

$$\mathbb{E}\left[f(w_{k+1}) \mid \mathcal{F}_k\right] \le f(w_k) - \eta_k\left(1 - \frac{L\eta_k}{2}\right)\|\nabla f(w_k)\|^2 + \frac{L\eta_k^2\sigma^2}{2}.$$

Proof: Using Taylors Theorem

$$f(w_{k+1}) \le f(w_k) - \eta_k \langle \nabla f(w_k), \underbrace{g_k}_{g(w_k, z_k)} \rangle + \frac{L\eta_k^2}{2}\|g_k\|^2. \quad (*)$$

Let's bound the norm $\|g_k\|^2$ in expectation

$$\mathbb{E}\left[\|g_k\|^2 \mid \mathcal{F}_k\right] \quad (**)$$

$$= \mathbb{E}\left[\|\underbrace{g_k - \nabla f(w_k)}_{\xi_k} + \nabla f(w_k)\|^2 \mid \mathcal{F}_k\right]$$

$$\overset{\text{(Linearity)}}{=} \|\nabla f(w_k)\|^2 + 2\langle \nabla f(w_k), \mathbb{E}\left[\xi_k \mid \mathcal{F}_k\right] \rangle + \mathbb{E}\left[\|\xi_k\|^2 \mid \mathcal{F}_k\right]$$

(pull-out)

$$\overset{\text{(unbiased)}}{=} \|\nabla f(w_k)\|^2 + \mathbb{E}\left[\|\xi_k\|^2 \mid \mathcal{F}_k\right] \le \|\nabla f(w_k)\|^2 + \sigma^2.$$

Combining (ü) and (♡) gives the result. □

Next class we will use this lemma to prove a rate for $\|\nabla f(w_k)\|$.