Lecture 13. (Sketch.)

- Homework was due today. *[Some homework solutions discussed in class.]*

1. Smoothness, sparsification, and the Maurey Lemma.

Fix any $z \in \mathbb{R}^d$ with $\|z\| \le R$; GD with $w_0 := 0$ gives us $w_t := -\frac{1}{\beta}\sum_{i<t} \nabla f(w_i)$ with

$$f(w_t) \le f(z) + \frac{\beta R^2}{2t}.$$

Speaking vaguely (but making things precise momentarily), if $\nabla f(w_i)$ is "simple", then so is w_t (by induction), and we've given the existence of "simple" approximate optima to f.

Remark.

Never underestimate the power of simply writing down a gradient on paper.

- In the least squares case $f(w) := \|Xw - y\|^2/2$ with $w_0 = 0$,

$$w_t := -\frac{1}{\beta}\sum_{i<t} \nabla f(w_i) = X^\top \left(-\frac{1}{\beta}\sum_{i<t}(Xw_i - y) \right);$$

  i.e., $w_t \in \text{im}(X^\top) = \ker(X)^\perp$; amongst other things, this implies w_t converges to the minimum norm solution.

- In the case of neural networks, a few recent results crucially rely upon simply writing down the gradient and staring at it a certain way.

Lemma (Maurey). Let β-smooth f and matrix $V \in \mathbb{R}^{d\times n}$ be given. For any $\alpha \in \Delta_n$, and any integer k, $\exists \hat\alpha \in \Delta_n \cap \mathbb{Z}^n/k$ with

$$f(V\hat\alpha) \le f(V\alpha) + \frac{\beta}{2k}\max_i \|V_{:,i}\|^2.$$

In particular, $\forall \alpha \in \Delta_n$, integer k, $\exists \hat\alpha \in \Delta_n \cap \mathbb{Z}^n/k$ satisfies

$$\|V\alpha - V\hat\alpha\|^2 \le \frac{1}{k}\max_i \|V_{:,i}\|^2.$$

Remarks.

- Note that $\hat\alpha \in \Delta_n \cap \mathbb{Z}^n/k$ is k-sparse:

$$1 = \sum_{i=1}^n \hat\alpha_i \ge \sum_{i=1}^n \frac{1}{k}\mathbb{1}\left[\hat\alpha_i \ge \frac{1}{k}\right] = \sum_{i=1}^n \frac{1}{k}\mathbb{1}\left[\hat\alpha_i > 0\right] = \frac{1}{k}\left|\{i : \hat\alpha_i > 0\}\right|.$$

- This lemma highlights another power of smoothness: it allows us to sparsify convex hulls! We'll use this property in the statistical/generalization part of the class.

- It's also used in a neural network uniform approximation proof Barron (1993).

Remarks (continued).

- Another interpretation of the result: for any $x \in V\Delta_n$, there exists $\hat{x} \in V\Delta_n$ which is k-sparse and satisfies
$$\|x - \hat{x}\|^2 \le \frac{1}{k} \max_i \|V_{:,i}\|^2.$$
This has no dependence on the dimensions of V, but only the norms of its columns!
- To highlight this lack of dimension dependence, consider a set U with $\sup_{v \in U} \|v\| \le R < \infty$ (but potentially $|U| = \infty$). For any $x \in \text{cl}(\text{conv}(U))$ and $\epsilon > 0$, by definition, there exists a subset $(u_1, \ldots, u_n)$ and $\alpha \in \Delta_n$ with
$$\|x - x_n\| \le \epsilon \qquad \text{where } x_n := \sum_{i=1}^n \alpha_i u_i.$$
Now we can apply the Maurey Lemma to x_n, and obtain x_k which is k-sparse with
$$\|x_n - x_k\|^2 \le \frac{R^2}{k}, \qquad \|x - x_k\|^2 \le 2\epsilon^2 + \frac{2R^2}{k}.$$

Proof. Let β-smooth f, $\alpha \in \Delta_n$, integer k be given. Define r.v. X with $\Pr[X = \mathbf{e}_i] = \alpha_i$, whereby
$$\mathbb{E}X = \sum_{i=1}^n \alpha_i \mathbf{e}_i = \alpha, \qquad \mathbb{E}VX = V\mathbb{E}X = V\alpha.$$
Let $(X_1, \ldots, X_n)$ be k iid copies of X, define $Y := \sum_i X_i/k$ (thus again $\mathbb{E}Y = \alpha$ and $\mathbb{E}VY = V\alpha$), and
$$\begin{aligned}
\mathbb{E}\|V\alpha - VY\|^2 &= \frac{1}{k^2}\mathbb{E}\left\langle \sum_{i=1}^k (V\alpha - VX_i), \sum_{j=1}^k (V\alpha - VX_j) \right\rangle \\
&= \frac{1}{k^2}\mathbb{E}\left(\sum_{i=1}^k \|V\alpha - VX_i\|^2 + \sum_{i \ne j} \langle V\alpha - VX_i, V\alpha - VX_j \rangle \right) \\
&= \frac{1}{k}\mathbb{E}\|V\alpha - VX_1\|^2 = \frac{1}{k}\mathbb{E}\left(\|V\alpha\|^2 - 2\langle VX_1, V\alpha\rangle + \|VX_1\|^2 \right) \\
&= \frac{1}{k}\left(\left(\sum_{i=1}^n \alpha_i \|V\mathbf{e}_i\|^2 \right) - \|V\alpha\|^2 \right) \le \frac{1}{k}\|\alpha\|_1 \max_i \|V\mathbf{e}_i\|^2.
\end{aligned}$$

Proof (continued).

Thus
$$\mathbb{E}\|V\alpha - VY\|^2 \le \frac{1}{k}\|\alpha\|_1 \max_i \|V_{:,i}\|^2.$$
By the probabilistic method (min is at most the expectation), there exists a $y_0 \in \Delta_n \cap \mathbb{Z}^n/k$ satisfying this bound, which gives the second part of the lemma. For the first part, since f is β-smooth,
$$\begin{aligned}
\mathbb{E}f(VY) &\le \mathbb{E}\left(f(V\alpha) + \langle \nabla f(V\alpha), VY - V\alpha\rangle + \frac{\beta}{2}\|V\alpha - VY\|^2 \right) \\
&\le f(V\alpha) + 0 + \frac{\beta}{2k}\max_i \|V_{:,i}\|^2,
\end{aligned}$$
which now (by the probabilistic method) gives a y_1 satisfying the first part of the theorem.

Note. We did not use a single y_0 for both parts.

2. Constructing sparse covers.

Natural greedy approach:

1. $w_0 := V\hat{\alpha}_0$ for some (sparse) $\hat{\alpha}_0 \in \Delta_n$.
2. For $i \in \{1, \ldots, t\}$:
   - 2.1 $u_i := \arg\min_{v \in \{V\mathbf{e}_1, \ldots, V\mathbf{e}_n\}} \langle w_{i-1} - V\alpha, v\rangle$.
   - 2.2 $w_i := (1 - \eta_i)w_{i-1} + \eta_i u_i \in V\Delta_n$.

To generalize this, note
$w_{i-1} - V\alpha = \nabla_w(w \mapsto \|w - V\alpha\|^2/2)(w_{i-1})$.

Frank-Wolfe / **conditional gradient** method.

1. $w_0 \in S$.
2. For $i \in \{1, \ldots, t\}$:
   - 2.1 $u_i := \arg\min_{v \in S} \langle \nabla f(w_{i-1}), v\rangle$. (Assume minimum exists.)
   - 2.2 $w_i := (1 - \eta_i)w_{i-1} + \eta_i u_i$.

Remark (constrained optimization).

Frank-Wolfe is performing optimization constrained to a set S.

- To do so, it must compute $\arg\min_{v\in S}\langle\nabla f(w_{i-1}), v\rangle$, which is a linear objective subject to a convex constraint.
- A standard competing approach, which we will discuss soon, is projected gradient descent, which must compute $\arg\min_{v\in S}\|v-w\|^2$, a quadratic objective subject to a convex constraint.

Frank-Wolfe literature often focuses on this distinction, naming many examples where the former is more tractable than the latter. Personally, I have found Frank-Wolfe very easy and convenient to implement a number of times. It is nice that it never leaves the constraint set.

Theorem.

Suppose f is β-smooth and convex, S is closed and bounded with $D := \sup_{v,v'\in S}\|v-v'\| < \infty$. Let $(w_i)_{i\leq t}$ be given by Frank-Wolfe with $\eta_i := 2/(i+1)$. Then

$$f(w_t) \leq f(z) + \frac{2\beta D^2}{t+1}.$$

Remark (Comparison to Maurey).

- To make the comparison, set $S := \{V\mathbf{e}_1, \ldots, V\mathbf{e}_n\}$.
- Maurey was non-constructive, did not need f to be convex, and slightly tighten the constants in the bound (but that may be analytic coincidence).
- Maurey gave a discrete solution in $V\hat{\alpha}$ with $\hat{\alpha}\in\Delta_n\cap\mathbb{Z}^n/k$. Frank-wolfe had $V\alpha'$ where α' has support size at most k but is real-valued. The exact properties of α' will be discussed momentarily.

Remarks (continued).

- (Step size.) Using $\eta_i := 1/i$ incurs a factor $\ln(t)$ in the bound. The weighting here can be shown inductively to satisfy

  $$w_t := \frac{\sum_{i=1}^t i v_i}{\sum_{i=1}^t i} = \frac{2}{t(t+1)}\sum_{i=1}^t i v_i,$$

  which puts more wait on later choices, and is sometimes called "polynomial weighting".
- To simplify, let's consider $S := \{w : \|w\|_2 \leq 1\}$. Then

  $$v := -\frac{\nabla f(w)}{\|\nabla f(w)\|}, \qquad w' := w - \eta\left(\frac{\nabla f(w)}{\|\nabla f(w)\|} + w\right),$$

  where the final "$+w$" is not present in gradient descent.

Remarks (continued).

- Recall the definition of dual norm:

  $$\|s\|_* := \arg\max\{\langle w, s\rangle : \|w\| \leq 1\}.$$

  Consequently, if $S := \{w : \|w\| \leq 1\}$ (now an arbitrary norm), then

  $$v := \arg\min_{v\ inS}\langle\nabla f(w), v\rangle$$

  satisfies $\langle\nabla f(w), u\rangle = -\|\nabla f(w)\|_*$. (We might elaborate upon this sort of analysis in the homework.)
- (Stopping conditions.) We mentioned that duality gap is the best way to construct stopping conditions, but that it's generally computationally infeasible. In the case of Frank-Wolfe, it ends up being tractable and clean (we might have a homework problem on this).

Proof. Let $w \in S$ and $\eta \in [0,1]$ be arbitrary, and set $u := \arg\min\{\langle \nabla f(w), v\rangle$ and $w' := (1-\eta)w + \eta u$. For any z,

$$\begin{aligned} f(w') - f(z) &\le f(w) - f(z) + \langle \nabla f(w), \eta(u-w)\rangle + \frac{\beta\eta^2\|u-w\|^2}{2} \\ &\le f(w) - f(z) + \eta \min_{v\in S} \langle \nabla f(w), v - w\rangle + \frac{\beta\eta^2 D^2}{2} \\ &\le f(w) - f(z) + \eta \langle \nabla f(w), z - w\rangle + \frac{\beta\eta^2 D^2}{2} \\ &\le f(w) - f(z) + \eta(f(z) - f(w)) + \frac{\beta\eta^2 D^2}{2} \\ &= (1-\eta)\left(f(w) - f(z)\right) + \frac{\beta\eta^2 D^2}{2}. \end{aligned}$$

Now consider $i = 1$; then $\eta_i = 2/(i+1) = 1$, so the above gives

$$f(w_1) - f(z) \le 0 + \frac{\beta\eta^2 D^2}{2} \le \frac{2\beta D^2}{i+1}.$$

Proof (continued).

When $i > 1$, the inductive hypothesis and preceding inequality together give

$$\begin{aligned} f(w') - f(z) &\le (1 - 2/(i+1))\left(f(w) - f(z)\right) + \frac{2\beta\eta^2 D^2}{(i+1)^2} \\ &\le (1 - 2/(i+1))\left(\frac{2\beta D^2}{i}\right) + \frac{2\beta\eta^2 D^2}{(i+1)^2} \\ &\le \frac{2\beta D^2}{i+1}\left(\frac{i-1}{i} + \frac{1}{i+1}\right). \end{aligned}$$

References

Barron, Andrew R. 1993. "Universal Approximation Bounds for Superpositions of a Sigmoidal Function." *IEEE Transactions on Information Theory* 39 (3): 930–45.