Lecture 9. (Sketch.)

Today we'll begin segment 2 of the course: optimization and online learning.

We'll start with the Perceptron algorithm, which is in the online setting, and easy to jump into.

1. Basics of optimization and online learning.

- Batch optimization: $((x_i, y_i))_{i=1}^m$ given, approximately solve $\inf_{f \in \mathcal{F}} \widehat{\mathcal{R}}_\ell(f(x_i), y_i)$.
  - Standard approach: parameterize $\mathcal{F}_p$ by $w \in \mathbb{R}^p$, approximately solve the infimum via continuous optimization over $\mathbb{R}^p$. Common methods include gradient descent (GD) and stochastic gradient descent. ML lately does not make widespread use of second-order methods.

    Remark. The above comment reflects the classical view of GD as nothing more than an optimizer, but recently there's an exciting *implicit regularization* perspective.
- Online learning: process data in a (possibly adversarial) stream.
  1. Initialize prediction model.
  2. For $t = 1, \ldots$:
     2.1 Receive x_i; predict $\hat{y}_i$.
     2.2 Suffer loss $\ell(\hat{y}_i, y_i)$ (nature chooses $\hat{y}_i$ given y_i!); update model.

2. Linearly separable data.

- For today, we assume a linear model: $f(x) = \langle w, x \rangle$.
  - Given a univariate convex loss ℓ, then $\ell(f(x)y) = \ell(\langle w, x \rangle\, y)$ is convex in w.
- Often in optimization, we aim to prove that the iterates w_i converge to some approximate optimum $\bar{w}$, or that we approximately minimize the convex risk $\widehat{\mathcal{R}}$ upon which we run gradient descent. Today we'll aim for something different, namely a guarantee on a nonconvex, nondifferentiable objective which we are not directly minimizing, and we'll also use a different assumption.
  - This assumption is **linear separability**: $\exists \bar{u}, \|\bar{u}\| = 1, \exists \gamma > 0$ s.t. $\langle \bar{u}, xy \rangle \geq \gamma \; \forall (x, y)$.
    [In class, pictures were drawn, and the two cases $y \in \pm 1$ were discussed.]

We can rewrite linear separability as an optimization problem:

$$\gamma := \max_{\|u\| \leq 1} \min_i \langle u, xy \rangle \, ; \qquad (1)$$

separablility means $\gamma > 0$. Scaling both sides by $1/\gamma$,

$$1 = \max_{\|u\| \leq 1/\gamma} \min_i \langle u, xy \rangle$$

This suggests an equivalent constrained form:

$$\min_w \frac{1}{2}\|w\|^2 \qquad \text{s.t. } 1 \leq \langle w, x_i y_i \rangle \;\; \forall i, \qquad (2)$$

and its Lagrangian

$$\min_w \sup_{C > 0} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^n \max\{0, 1 - \langle w, x_i y_i \rangle\}. \qquad (3)$$

This last form is the "hard margin SVM".

Exercise: Prove 1, 2, 3 are equivalent.

We can relax the final Lagrangian into the familiar soft-margin SVM:

$$\inf_w \frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \ell_h(-\langle w, x_i y_i\rangle),$$

where $C \geq 0$ and $\ell_h(z) := \max\{0, 1+z\}$ is the *hinge loss.*

This form is also called the "soft margin SVM".

[In class, we discussed this a little geometrically, and running gradient descent on it and its dual.]

3. Perceptron.

Consider a simpler approach: let's run SGD on the objective after stripping away $\|w\|^2$ and the "$1+$" in the hinge loss. That means SGD on the ReLU loss:

$$w' := w - \partial_w(w \mapsto \sigma_r(\langle w, -xy\rangle)) = w + xy\mathbb{1}[\langle w, -xy\rangle \geq 0],$$

where $\sigma_r(z) := \max\{0, z\}$, and usually the initial point is $w_0 = 0$.
Convention: always take the subgradient 1 at 0 (this matters a lot)...

Geometric interpretation:

- When $\mathbb{1}[\langle w, xy\rangle \leq 0]$ (mistake), we rotate towards the example.
- Otherwise, we do nothing.

Note. We predict with $\hat{y} := 2 \cdot \mathbb{1}[\langle w, x\rangle \geq 0] - 1$, so $\mathbb{1}[\hat{y} \neq y] \neq \mathbb{1}[\langle w, xy\rangle \leq 0]$ when $w = 0$ and $y = +1$;

Remark (optimization)**.** We defined the iteration as SGD on the ReLU loss:

$$w_i := w_{i-1} - \partial_w(w \mapsto \sigma_r(\langle w, -x_i y_i\rangle))(w_{i-1}).$$

The optimization view, then, would be that SGD will be approximately minimizing the objective

$$\inf_w \sum_{i=1}^n \sigma_r(\langle w, -x_i y_i\rangle.$$

- 0 is the optimal objective value (i.e., because $\sigma_r \geq 0$ and since $w = 0$ attains this lower bound).
- Therefore $w_0 = 0$, the standard initialization, is optimal!
- By choosing 1 as the subgradient at 0, we have deliberately moved away from the global optimum!
  - The explanation is that the ReLU loss is only used as a surrogate potential function in this problem; it is not a quantity we actually care about.

Theorem. Suppose linear separability (i.e., $\langle \bar{u}, xy\rangle \geq \gamma > 0$), $\|xy\| \leq 1$, and all (w_i, y_i) are given by Perceptron. Then

$$\sum_{i\geq 1} \mathbb{1}[\hat{y}_i \neq y_i] \leq \frac{1}{\gamma^2}.$$

Remark. In the first lecture, we said that learning requires "coherence" between past and future. In this case, $(\bar{u}, \gamma)$ provide that coherence: they guarantee that (some) good choices in the past will be good in the future.

Proof. Define the set $M_t := \{i \leq t : \mathbb{1}[\langle w_{i-1}, x_i y_i \rangle \leq 0\}$, a superset of the iterations making mistakes up through time t. Momentarily we'll show $|M_t| \leq 1/\gamma^2$ for arbitrary t, which proves the result since $|M_t|$ increases monotonically and $\sum_{i=1}^t \mathbb{1}[\hat{y}_i \neq y_i] \leq |M_t|$.

Continuing, let's go back to our intuition: mistakes rotate us towards $x_i y_i$, which we can take more generally to mean "rotation towards correctness", or in other words $\bar{u}$. This suggests a potential function

$$\frac{1}{2}\left\| \frac{w_i}{\|w_i\|} - \bar{u} \right\|^2 = 1 - \left\langle \frac{w_i}{\|w_i\|}, \bar{u} \right\rangle.$$

Note. We won't show this quantity converges to anything, it's just a proof technique and intuition.

- Indeed, we will not in general converge to $\bar{u}$; consider $x_i y_i = (1,0)$ for $i \geq 1$, which means $w_i = (1,0)$ for $i \geq 1$, but we can choose any $\bar{u}$ with positive coordinates to satisfy the conditions of the theorem.

To lower bound $\langle w_t, \bar{u} \rangle$, note by induction $w_t := \sum_{i \in M_t} x_i y_i$, thus

$$\langle w_t, \bar{u} \rangle = \sum_{i \in M_t} \langle x_i y_i, \bar{u} \rangle \geq \gamma |M_t|.$$

To upper bound $\langle w_t, \bar{u} \rangle \leq \|w_t\|$, since $\langle w_{i-1}, x_i y_i \rangle \leq 0$ when $i \in M_i$,

$$\begin{aligned}
\|w_i\|^2 &= \|w_{i-1} + x_i y_i \mathbb{1}[i \in M_i]\|^2 \\
&= \|w_{i-1}\|^2 + 2\left\langle x_i y_i \mathbb{1}[i \in M_i], w_{i-1} \right\rangle + \|x_i y_i \mathbb{1}[i \in M_i\|^2 \\
&= \|w_{i-1}\|^2 + 2\left\langle w_{i-1}, x_i y_i \right\rangle \mathbb{1}[i \in M_i] + \mathbb{1}[i \in M_i]\|x_i y_i\|^2 \\
&\leq \|w_{i-1}\|^2 + 0 + \mathbb{1}[i \in M_i],
\end{aligned}$$

which by induction and the monotonicity property $M_i \subseteq M_t$ gives

$$\|w_t\|^2 \leq \|w_0\|^2 + \sum_{i<t} \mathbb{1}[i \in M_t] = |M_t|.$$

Combining the upper and lower bounds,

$$\gamma |M_t| \leq \langle w_t, \bar{u} \rangle \leq \sqrt{|M_t|} \qquad \Longrightarrow \qquad |M_t| \leq \frac{1}{\gamma^2}.$$

Remark.

- As with SVM, Perceptron can be kernelized. In particular, given a kernel function $k(\cdot,\cdot)$,

  $$w_t := \sum_{i \in M_t} x_i y_i \qquad \text{becomes} \qquad w_t := \sum_{i \in M_t} y_i k(x_i, \cdot),$$

  and $\langle w_t, x \rangle = \sum_{i \in M_t} y_i k(x_i, x)$.
- Many parts of the Perceptron proof go through in the nonconvex case. Perhaps we'll see more of it in time to come...