Lecture 14. (Sketch.)

- No class next Wednesday.
- When should we have final project presentations?

1. Convex "margin losses".

- In general, our losses have the form $\ell(y, \hat{y})$.
- In the binary classification case (our lazy focus), we can simplify with univariate "margin" losses: $(y, \hat{y}) \mapsto \ell(-y\hat{y})$.
  - We want $\ell(-y\hat{y})$ large when $\hat{y} \neq y$; thus ℓ nondecreasing.
  - Another convention is to drop "$-$" and have nonincreasing losses.
- Some examples:
  - Least squares: $y = \pm 1$, so $(y + \hat{y})^2 = (y(1 + \hat{y}))^2 = (1 + \hat{y})^2$; so $\ell_{\text{ls}}(z) := (1 + z)^2/2$.
  - Logistic loss $\ell_{\log}(z) = \ln(1 + \exp(z))$. Most common in practice (multiclass case is "cross entropy" used in neural networks). When $z \geq 0$, it's essentially affine; when $z < 0$ it is similar to exp.
  - Ramp loss $\ell_\gamma(z) := \max\{0, \min\{1, 1 + z/\gamma\}\}$, which is nonconvex. We'll use this in generalization analysis.

Remark (smoothness and convexity of risks/losses).

- If ℓ is convex, $\mathcal{R}_\ell$ might still be non-convex (as a function of the model parameters).
- If ℓ is β-smooth and the predictor f is differentiable in the parameters w,

$$\begin{aligned}&\left\|\ell'(-yf(w;x))(-y\nabla_w f(w;x)) - \ell'(-yf(w';x))(-y\nabla_{w'} f(w';x))\right\| \\ &= \left\|\ell'(-yf(w;x))\nabla_w f(w;x)) - \ell'(-yf(w';x))\nabla_{w'} f(w';x))\right\|.\end{aligned}$$

  In the linear case $f(w;x) = \langle w, x\rangle$, this becomes

$$\begin{aligned}&\left\|\ell'(-yf(w;x))x - \ell'(-yf(w';x))x\right\| \\ &\leq \|x\|\left\|\ell'(-y\langle w, x\rangle) - \ell'(y\langle w, x\rangle)\right\| \leq \|x\|^2\beta\left\|w - w'\right\|,\end{aligned}$$

  meaning smoothness of ℓ is inherited by the risk $\widehat{\mathcal{R}}$ as a function of the model parameters.

Remark (more on smoothness, given its role in optimization).

Here are two standard ways to get a smooth optimization problem. Note that these are mainly "of theoretical interest"; e.g., I don't know of anyone applying smoothing when training (non-smooth) ReLU networks.

- **Mollification.** replace f with $w \mapsto \mathbb{E}_\xi f(w + \xi)$, where $\xi \sim \mathcal{N}(0, \sigma^2 I)$ for some some standard deviation σ.
- **Moreau-Yosida regularization.** Recall the Fenchel/convex conjugate $g^*(s) = \sup_x \langle x, s\rangle - g(x)$. If g is λ-sc, then g^* is λ^{-1}-smooth, Thus replacing objective f with $(f^* + \|\cdot\|^2/(2\beta))^*$ gives a nearby β-smooth objective. A related topic are "proximal point methods".

2. Classification & convex losses: separable case.

The rest of the lecture considers the relationship of convex risk minimization and zero-one/classification minimization. Namely, consider the relationship

$$\mathcal{R}_z(f) - \inf_g \mathcal{R}_z(g) \qquad \text{vs.} \qquad \mathcal{R}_\ell(f) - \inf_g \mathcal{R}_\ell(g)$$

We minimize the right side because it seems more tractable, however we care about the left side. ($\mathcal{R}_z(f) = \Pr[f(x) \neq y]$.)

Note. The infimum is over *all* (measurable) functions.

Remark. Much of this literature considers the RHS tractable (e.g., convex opt over $\mathbb{R}^d$ with d small), supposes the right hand side is not 0, and considers the distribution not the training set. It seems deep learning has shaken up the dominant paradigm here: now the RHS is 0 over the training set, and it's not convex but it seems "tractable in practice".

Theorem. Suppose $\ell \geq 0$ and ℓ nondecreasing. Then

$$\Pr[f(X) \neq Y] = \mathcal{R}_z(f) \leq \frac{\mathcal{R}_\ell(f)}{\ell(0)}.$$

If $\exists \bar{f} \in \mathcal{F}$ with $\bar{f}(x)y > 0$ almost surely and $\mathcal{F}$ is closed under multiplication by constants and $\lim_{z\to-\infty} \ell(z) = 0$, then

$$\mathcal{R}_z(f) - \inf_{g \text{ meas.}} \mathcal{R}_z(g) \leq \frac{1}{\ell(0)} \left(\mathcal{R}_\ell(f) - \inf_{h\in\mathcal{F}} \mathcal{R}_\ell(h) \right).$$

Remark.

- The last inequality is pretty good, and the technical-looking assumptions aren't so bad (e.g., often satisfied by neural networks on finite training sets).

Proof. For the first part, by Markov's inequality,

$$\begin{aligned}\Pr[f(X) \neq Y] &\leq \Pr[-f(X)Y \geq 0] \leq \Pr[\ell(-f(X)Y) \geq \ell(0)] \\ &\leq \frac{\mathbb{E}\ell(-f(X)Y)}{\ell(0)}.\end{aligned}$$

The second part follows if we can show the two infima are 0. Since $x \mapsto \operatorname{sgn}(f(x))$ is measurable and agrees with y almost surely,

$$0 \leq \inf_{g \text{ meas.}} \Pr[g(X) \neq Y] \leq \Pr[\operatorname{sgn}(\bar{f}(X)) \neq Y] = 0,$$

On the other hand, let $\epsilon > 0$ be arbitrary, choose τ so that $r \leq \tau$ implies $\ell(r) \leq \epsilon/2$, and choose c large enough so that $\Pr[|c\bar{f}(X)| \leq \tau] \leq \epsilon/(2\ell(0))$. Then, since $c\bar{f} \in \mathcal{F}$,

$$\begin{aligned}0 &\leq \inf_{f\in\mathcal{F}} \mathcal{R}_\ell(f) \leq \mathcal{R}_\ell(c\bar{f}) \\ &= \mathbb{E}\left(\ell(-c\bar{f}(X)Y)\mathbb{1}[|c\bar{f}(X)| \leq \tau]\right) + \mathbb{E}\left(\ell(-c\bar{f}(X)Y)\mathbb{1}[|c\bar{f}(X)| > \tau]\right) \\ &= \ell(0)\Pr[|c\bar{f}(X)| \leq \tau] + \mathbb{E}\left(\ell(r)\right) \leq \epsilon.\end{aligned}$$

2. The general case.

Define *regression function* $\bar{p}(x) := \Pr[Y = 1|X = x]$, and *bayes predictor* $\bar{h}(x) := \operatorname{sgn}(2\bar{p} - 1)$.

- $\bar{h}$ is the best classifer: for any other $h : X \to \{-1, +1\}$,

$$\begin{aligned}\mathcal{R}_z(h) - \mathcal{R}_z(\bar{h}) &= \mathbb{E}\left(\Pr[h(X) \neq Y|X] + \Pr[\bar{h}(X) \neq Y|X]\right) \\ &= \mathbb{E}\left(\mathbb{1}[h(X) \neq \bar{h}(X)]\left(\max\{\bar{p}(X), 1-\bar{p}(X)\} - \min\{\bar{p}(X), 1-\bar{p}(X)\}\right)\right) \\ &= \mathbb{E}\left(\mathbb{1}[h(X) \neq \bar{h}(X)]|2\bar{p}(X) - 1|\right) \geq 0.\end{aligned}$$

- $\inf_{h \text{ meas.}} \mathcal{R}_z(h) = \mathcal{R}_z(\bar{h}) = 0$ iff $\bar{p} \in \{0, 1\}$ a.e..

In the theorem earlier today, we had $\mathcal{R}_z(\bar{h}) = 0$.

If $\mathcal{R}_z(\bar{h}) > 0$ things are much nastier. We'll discuss negative and positive results.

2a. General case: negative results.

First, a remark on computation.

Remark. It is possible that there exists a linear predictor f with $\widehat{\mathcal{R}}_z(\mathrm{sgn}(f)) = 0.01$, but it is NP-hard to find a linear predictor g with $\widehat{\mathcal{R}}(\mathrm{sgn}(g)) \leq 0.49$ (Guruswami and Raghavendra 2006).

Remark. Some statements are on $\mathcal{R}$, some on $\widehat{\mathcal{R}}$. Can instantiate the $\widehat{\mathcal{R}}$ statements with an empirical measure, but it can be brittle; e.g., demanding the same instance to appear multiple times with different labels.

Minimizing a convex risk does not circumvent this hardness barrier: convex risk minimization can be fooled arbitrarily badly.

Theorem (see (Ben-David et al. 2012) for a similar construction). Let $\epsilon \in (0,1)$ and $r \in [0,1]$ be given. There exists a discrete probability distribution on $n = \mathcal{O}(1/\epsilon)$ examples $((x_i, y_i))_{i=1}^n$ satisfying:

- $x_i \in \mathbb{R}$, $y_i = +1$ (univariate, positive labels).
- There exists a linear predictor $\bar{w} \in \mathbb{R}$ with $\mathcal{R}_z(\bar{w}) \leq \epsilon$.
- For any convex loss $\ell : \mathbb{R} \to \mathbb{R}_{\geq 0}$ with $\min \partial\ell(0) \geq r \cdot \max \partial\ell(0) > 0$, the convex risk $\mathcal{R}_\ell$ has minimizers, and every minimizer $\hat{w}$ has $\mathcal{R}_z(\hat{w}) \geq 1 - \epsilon$.

Remark (intuition). Suppose (as above) $\min \partial\ell(0) > 0$.

- For $m > 0$, note $\ell(m) \geq \ell(0) + m \cdot \min \partial\ell(0)$; linear in m.
- On the other hand, $\ell(-m) \geq 0$; increasing m doesn't help much. A single wrong prediction can matter more than many correct ones.

Proof.

- Pick n with $\epsilon/2 \leq 1/n \leq \epsilon$ (thus $n = \mathcal{O}(1/\epsilon)$).
- Place $(x_1, \ldots, x_{n-1})$ at $+1$ with $y_i = +1$.
- Place $x_n := -c$ with $c := n/r$, and $y_n := +1$. *[Picture drawn in class.]*
- Note that any $\bar{w} > 0$ is globally minimizes $\mathcal{R}_z$, attaining $\mathcal{R}(z) = 1/n \leq \epsilon$.
- Let any convex $\ell : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be given as specified in the theorem; it remains to show that $\mathcal{R}_\ell$ behaves badly.
- Note that $\mathcal{R}_\ell$ possess minimizers: e.g., it possesses bounded sublevel sets by applying the subgradient rule to $\mathcal{R}_\ell(\pm m)$ for sufficiently large m, and comparing this to $\mathcal{R}_\ell(0)$.
- It remains to show that all minimizers are bad; namely, any minimizer $\hat{w}$ has $\hat{w} < 0$, and thus $\mathcal{R}_\ell(\hat{w}) \geq (n-1)/n \geq 1 - \epsilon$.

Proof (continued).

It will now be shown that any $w \geq 0$ can not be a minimizer, thus completing the proof (since minimizers exist, thus must be negative).

- Note that $0 \notin \partial\mathcal{R}_\ell(0)$: setting $\alpha := \max \partial\ell(0)$,

$$\begin{aligned}
\min \mathcal{R}_\ell(0) &= \min\left(\frac{n-1}{n}\partial(w \mapsto \ell(-w))(0) + \frac{1}{n}\partial(w \mapsto \ell(cw))(0)\right) \\
&= \min\left(-\frac{n-1}{n}\partial\ell(0) + \frac{c}{n}\partial\ell(0)\right) \\
&\geq -\frac{n-1}{n}\alpha + \frac{c}{n}(r\alpha) = \frac{\alpha}{n} > 0.
\end{aligned}$$

- On the other hand, Since $s := \min \partial\mathcal{R}_\ell(w) > 0$ as above, every $w > 0$ has

$$\mathcal{R}_\ell(w) \geq \mathcal{R}_\ell(0) + s(w - 0) > \mathcal{R}_\ell(0).$$

2b. General case: positive results.

A key thing went wrong in this example:

- **Weak representation power.** Predictions on $+1$ constrained predictions on $-c$; this is similar to the "separation condition" in Stone-Weierstrass. **Note:** for linear predictors, these constraints are easy to understand, but for, say, deep networks, no one understands.

Remark. We will only be able to say something strong when the predictor class is essentially everything (e.g., all continuous functions). It's not clear how to prove anything more sensitive.

So, the predictors need to be expressive. What about the loss function?

Consider some x and $\bar{p}$ with $p := \bar{p}(x) = \Pr[Y = 1|X = x] \neq 1/2$.

- The error of a predictor f conditioned on x is
$$\ell(-f(x))p + \ell(f(x))(1-p).$$
- Suppose the function class is so powerful that $f(x)$ can be set to any $\alpha \in \mathbb{R}$. Then we want ℓ to agree with $\bar{h}(x) = \operatorname{sgn}(2p-1)$:
$$\inf_{\alpha\in\mathbb{R}} \ell(-\alpha)p + \ell(\alpha)(1-p) < \inf_{\substack{\alpha\in\mathbb{R} \\ (2p-1)\leq 0}} \ell(-\alpha)p + \ell(\alpha)(1-p).$$

Let's call this condition the "majority vote condition": ℓ should prefer answers that agree with majority vote on the conditional distribution $\Pr[Y|X = x]$.

In the literature, this is called **classification calibration** (Zhang 2004; Bartlett, Jordan, and McAuliffe 2006).

If $\mathcal{F}$ is expressive and ℓ agrees with majority vote, minimizing $\mathcal{R}_\ell$ over $\mathcal{F}$ will pick $f \in \mathcal{F}$ that agrees with $\bar{h}$ everywhere, and $\mathcal{R}_z(f)$ is also small.

Theorem (See (Zhang 2004) for similar version and proof). Suppose $\ell : \mathbb{R} \to \mathbb{R}_{\geq 0}$ convex, $\min \partial\ell(0) > 0$. Define the function

$$G_\ell(p) := \ell(0) - \inf_{\alpha\in\mathbb{R}} \left(p\ell(-\alpha) + (1-p)\ell(\alpha)\right).$$

Suppose there exist $c \geq 0, s \geq 1$ so that $\forall p \in [0,1]$, $|2p-1| \leq cG_\ell(p)^{1/s}$. Then

$$\mathcal{R}_z(f) - \inf_{g \text{ meas}}(\operatorname{sgn}(g)) \leq c\left(\mathcal{R}_\ell(f) - \inf_{g \text{ meas}} \mathcal{R}_\ell(g)\right)^{1/s}.$$

Remark.

Here are some standard losses and their constants; see (Zhang 2004) for more discussion. In particular, these numbers done mean one is really better than another...

- Squared loss $z \mapsto (1+z)^2/2$: $c = 1, s = 2$.
- Hinge loss $z \mapsto \max\{0, 1+z\}$: $c = s = 1$.
- Logistic loss $z \mapsto \ln(1+\exp(z))$: $c = \sqrt{2}$, $s = 2$.
- Exponential loss $z \mapsto \exp(z)$: $c = \sqrt{2}, s = 2$.
- Impagliazzo-Zhang loss
$$z \mapsto \begin{cases} 0 & z < -1, \\ (1+z)^2 & z \in [-1, +1], \\ 4z & z > 1, \end{cases}$$
has $c = s = 1$. (Impagliazzo 1995)

Proof. Starting from an earlier calculation, setting $h = \text{sgn}(f)$ for convenience,

$$
\begin{aligned}
&\mathcal{R}_z(h) - \mathcal{R}(\bar{h}) \\
&= \mathbb{E}\left(\mathbb{1}[\bar{h}(X) \neq h(X)]|2\bar{p}(X) - 1|\right) \\
&\leq \mathbb{E}\left(\mathbb{1}[f(X)(2\bar{p}(X) - 1) \leq 0]|2\bar{p}(X) - 1|\right) \\
&\leq c\mathbb{E}\Big(\mathbb{1}[f(X)(2\bar{p}(X) - 1) \leq 0]^{1/s} \\
&\qquad \cdot \left(\ell(0) - \inf_\alpha(\bar{p}(X)\ell(-\alpha) + (1 - \bar{p}(X))\ell(\alpha))\right)^{1/s}\Big) \\
&\leq c\left(\mathbb{E}\mathbb{1}[f(X)(2\bar{p}(X) - 1) \leq 0]\left(\ell(0) - \inf_\alpha(\bar{p}(X)\ell(-\alpha) + (1 - \bar{p}(X))\ell(\alpha))\right)\right)^{1/s}
\end{aligned}
$$

The proof is done if we can show $f(x)(2\bar{p}(x) - 1) \leq 0$ implies $\ell(0) \leq \bar{p}(x)\ell(-f(x)) + (1 - \bar{p}(x))\ell(f(x))$.

Proof (continued).

Choose any $s \in \partial\ell(0)$ with $s > 0$. Taking the convex combination (with weights $p := \bar{p}(X)$ and $1 - p$) of the two subgradient inequalities

$$
\begin{aligned}
\ell(-\alpha) &\geq \ell(0) + s(-\alpha - 0), \\
\ell(\alpha) &\geq \ell(0) + s(\alpha - 0),
\end{aligned}
$$

gives

$$
\begin{aligned}
p\ell(-\alpha) + (1 - p)\ell(\alpha) &\geq \ell(0) + ps\alpha - (1 - p)s\alpha \\
&= \ell(0) + s\alpha(2p - 1) \\
&\geq \ell(0).
\end{aligned}
$$

Remark.

Can generalize these concepts to multiclass and others.

The key principle is that the predictor defines a conditional probability model of Y given X which agrees with the true $\Pr[Y|X]$.

[I can elaborate, sorry. Didn't type up all of my notes...]

References

Bartlett, Peter L., Michael I. Jordan, and Jon D. McAuliffe. 2006. "Convexity, Classification, and Risk Bounds." *Journal of the American Statistical Association* 101 (473): 138–56.

Ben-David, Shai, David Loker, Nathan Srebro, and Karthik Sridharan. 2012. "Minimizing the Misclassification Error Rate Using a Surrogate Convex Loss." In *ICML*.

Guruswami, Venkatesan, and Prasad Raghavendra. 2006. "Hardness of Learning Halfspaces with Noise." In *FOCS*.

Impagliazzo, Russell. 1995. "Hard-Core Distributions for Somewhat Hard Problems." In *FOCS*, 538–45.

Zhang, Tong. 2004. "Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization." *The Annals of Statistics* 32: 56–85.