Consistency of convex ERM for classification problems, part 1.

Administrative/Meta:

- Key reference for today, one of my favorite papers of all time: Zhang (2004b).
- Dmitry Yarotsky has used my construction to prove that multiplication can be approximated to accuracy $\epsilon > 0$ using a ReLU network of size $\mathcal{O}(\ln(1/\epsilon))$: "Error bounds for approximations with deep ReLU networks".

Convex ERM

Let's take a step back for a moment and remember why we are minimizing convex functions. Our running example is empirical risk minimization, namely the minimization over some function class $\mathcal{F}$ of

$$\widehat{\mathcal{R}}(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(-y_i f(x_i)),$$

where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1,+1\}$. This procedure is selecting a predictor $\mathcal{F} \ni f : \mathbb{R}^d \to \mathbb{R}$; for instance, we've extensively discussed the case of linear predictors $\mathcal{F} := \{x \mapsto \langle w, x\rangle : w \in \mathbb{R}^d\}$, and we'll allow the notation

$$\widehat{\mathcal{R}}(w) := \widehat{\mathcal{R}}(x \mapsto \langle w, x\rangle) = \frac{1}{n}\sum_{i=1}^{n} \ell(-y_i \langle w, x_i\rangle).$$

(Recall that "linear" can be quite flexible by working with x_i in some richer space, for instance in boosting and SVMs.)

To convert this into a linear *classifier*, one needs a map to $\{-1,+1\}$, for instance $x \mapsto \operatorname{sgn}(\langle w, x\rangle)$.

Key question: how well does this convex procedure solve the original problem, namely the minimization of the classification error

$$\widehat{\mathcal{R}_z}(f) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[\operatorname{sgn}(f(x_i)) \neq y_i] \ ?$$

This and the subsequent lecture will give one negative and one positive result to this end.

- **Negative scenario** (informally stated)**.** For any $\epsilon \in (0,1)$ and any loss ℓ with $\min \partial\ell(0) > 0$ (which holds for every loss we've tried), there exists a set of points so that convex risk minimization (cf. ()) over linear classifiers gives a linear classifier $\bar{w}$ with classification error $\widehat{\mathcal{R}_z}(\bar{w}) \geq 1 - \epsilon$, despite the existence of a linear classifier $\hat{w}$ with classification error $\widehat{\mathcal{R}_z}(\hat{w}) \leq \epsilon$.
- **Positive scenario** (informally stated)**.** Rather than considering a fixed set of predictors (which are defeated by the preceding result), if we instead consider a sequence of predictors which in the limit are dense in continuous functions, the solution to () achieves classification error arbitrarily close to the optimal error in ().

Example losses.

- **Squared loss** $\ell(z) = (1+z)^2/2$. Typically we consider the squared loss to have form $(y_i - \langle w, x_i \rangle)^2/2$; this is equal to $\ell(-y_i \langle w, x_i \rangle)$ when $y_i \in \{-1, +1\}$, which holds throughout today.
- **Hinge loss** $\ell(z) = \max\{0, 1+z\}$.
- **Logistic loss** $\ell(z) = \ln(1 + \exp(z))$. This loss mimics the hinge loss, but is smooth. The logistic loss may seem to have some benefits due to having nonzero gradients everywhere and being smooth. However, there is no complete theory on the relative benefits of losses; one of the only discussions of this type is by Zhang (2004b).
- **Exponential loss** $\ell(z) = \exp(z)$. This loss is only used in AdaBoost, and practical implementations of AdaBoost these days typically use logistic loss or squared loss.
- **Impagliazzo/Zhang loss**

  $$\ell(z) := \begin{cases} 0 & \text{when } z < -1, \\ (1+z)^2 & \text{when } z \in [-1, +1], \\ 4z & \text{when } z > 1. \end{cases}$$

  This is another smooth analog to the hinge loss. It can be reverse-engineered from a proof from the hardness amplification literature Impagliazzo (1995), and also appears concluding a discussion of losses as a choice which simultaneously attains best possible scenarios Zhang (2004b Section 3.6).

Remark.

- Notice that all preceding convex loss functions satisfy $\partial\ell(0) > 0$.
- The ideas for the "positive scenario" do carry over to multi-class classification (Zhang 2004a), and we'll discuss this in the next lecture.
- In a sense, we can't hope for convex ERM to do much, since it is NP-hard to find a linear classifier with error below 0.49 when the best possible error is 0.01 (Guruswami and Raghavendra 2006). Thus, barring a proof that P=NP, any method for this setting must either achieve poor error, or it must take long to converge.

A negative scenario

Theorem (See Ben-David et al. (2012) for a similar result)**.** Let $\epsilon \in (0,1)$ and scalar $r > 0$ be given. There exists a set of $n = \mathcal{O}(1/\epsilon)$ labeled points $((x_i, y_i))_{i=1}^n$ satisfying the following conditions.

- $x_i \in \mathbb{R}$ and $y_i = +1$. (That's right, no negative examples!)
- There exists a linear predictor $\hat{w}$ with $\widehat{\mathcal{R}}_z(\hat{w}) \leq \epsilon$.
- Choose any convex loss $\ell : \mathbb{R}_+ \to \mathbb{R}$ with $\min(\partial\ell)(0) \geq r \cdot \max(\partial\ell)(0) > 0$. Then minimizers to $\widehat{\mathcal{R}}$ exist, and every minimizer $\bar{w}$ must satisfy $\widehat{\mathcal{R}}_z(\bar{w}) \geq 1 - \epsilon$.

Proof. *[picture showing the location and probability mass of the two distinct data point.]*

Choose any integer n with $\epsilon/2 \leq 1/n \leq \epsilon$, meaning $n = \mathcal{O}(1/\epsilon)$. Place $n-1$ points at $+1$ with label $+1$, and 1 point at $-c$ where $c := n/r$, again with label $+1$; any predictor $\hat{w} > 0$ achieves error $1/n \leq \epsilon$.

The convex empirical risk of $w \in \mathbb{R}^d$ is

$$\widehat{\mathcal{R}}(w) := \frac{1}{n}\left(\ell(cw) + \sum_{i=1}^{n-1} \ell(-w)\right) = \frac{1}{n}\ell(cw) + \frac{n-1}{n}\ell(-w).$$

[The proof of existence of minimizers is omitted; for instance, the conditions on ℓ imply bounded level sets.]
First note that $\min \partial\widehat{\mathcal{R}}(0) > 0$:

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

This in turn implies that every minimizer $\bar{w}$ has $\bar{w} < 0$.

- First note that $\bar{w} \leq 0$: choosing any subgradient $s \in \partial\widehat{\mathcal{R}}(0) > 0$,

$$\mathcal{R}(\bar{w}) \geq \mathcal{R}(0) + s(\bar{w} - 0) \qquad \Longrightarrow \qquad \bar{w} \leq (\mathcal{R}(\bar{w}) - \mathcal{R}(0))/s \leq 0.$$

- Since $0 \notin \partial\widehat{\mathcal{R}}(0)$, the first-order necessary and sufficient optimality conditions imply 0 is not optimal.

Together, these two points imply $\bar{w} < 0$. This means any optimal choice $\bar{w}$ will only correctly classify the lone point at $-c$, giving $\widehat{\mathcal{R}_z}(\bar{w}) = 1 - 1/n \geq 1 - \epsilon$. $\square$

A positive scenario

[In class, about 20 minutes were spent explaining how the "negative scenario" can be defeated, which naturally leads to the positive scenario. This material is deferred to the notes for the next lecture.]

[Note to future matus: the class liked this topic a lot and there was a lot of time spent on various questions, and based on a vote it was split into two lectures. Not sure what to do next year, though.]

References

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. 2004a. "Statistical Analysis of Some Multi-Category Large Margin Classification Methods." *JMLR* 5: 1225–51.

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