Convex Risk Minimization and Conditional Probability Estimation

Matus Telgarsky, Miro Dudík, Robert Schapire

Question.

Setting. Predictors $(f_i)_{i\geq 1}$ with:

- f_i *linear*: $f_i = \sum_{h\in\mathcal{H}} w_i(h)h$, where $\mathcal{H} \ni h : \mathcal{X} \to [-1,+1]$.
- $(f_i)_{i\geq 1}$ *minimize risk*: $\mathcal{R}(f_i) \to \inf_{f \text{ linear}} \mathcal{R}(f)$, where

$$\mathcal{R}(f) \;:=\; \int \ell(-yf(x))d\mu(x,y) \;\;,$$

and ℓ has $\ell'' > 0$, $\lim_{r\to-\infty} \ell(r) = 0$, and some technical conditions. (Examples: logistic, exponential).

Motivation.
Convex risk $\mathcal{R}$ only a surrogate; want to say more about $(f_i)_{i\geq 1}$.

Goal.
A unique limit object which describes classification properties of $(f_i)_{i\geq 1}$.

Obstruction. Unboundedness and/or infinite dimension, for example:

- Unconstrained/unregularized risk minimization (boosting, empirical folklore, etc).
- Weakening constraints/regularization.
- Even: logistic regression with $d = n = 1$.

Resolution.

Conditional probability models. Given $f : \mathcal{X} \to \mathbb{R}$, define η_f as

$$\eta_f(x,y) \;:=\; \frac{1}{1+\exp(-yf(x))} \qquad \text{logistic } \ell;$$

$$\eta_f(x,y) \;:=\; \frac{1}{1+\frac{\ell'(-yf(x))}{\ell'(yf(x))}} \qquad \text{generic } \ell.$$

We'll exhibit a unique $\bar\eta$ with $\eta_{f_i} \to \bar\eta$ in $L_1(\mu)$.

Risk minimization. Given $(\ell, \mathcal{H}, \mu)$, there exists a unique $\bar\eta$ so that every $(f_i)_{i\geq 1}$ with f_i linear and $\mathcal{R}(f_i) \to \inf_{f \text{ linear}} \mathcal{R}(f)$ has

$$\int |\eta_{f_i} - \bar\eta| d\mu \;\to\; 0 \;\;.$$

- Note: potentially $|\mathcal{H}| = \infty$ and $\{\eta_f : f \text{ linear}\}$ not compact; i.e., $\bar\eta$ nontrivial.

Empirical risk minimization. Additionally, when $|\mathcal{H}| < \infty$, with probability at least $1-\delta$ over a draw of size $n \geq \Omega(\ln(1/\delta))$, each linear $\hat f$ has

$$\int |\eta_{\hat f} - \bar\eta| d\mu \;\leq\; \mathcal{O}\left(g_1\left(\widehat{\mathcal{R}}_n(\hat f)\right) \sqrt{\widehat{\mathcal{R}}_n(\hat f) - \inf_{f \text{ linear}} \widehat{\mathcal{R}}_n(f) \;+\; \frac{\ln(n/\delta)}{n}} \right)$$

with (nondecreasing) g_1, $\mathcal{O}$, and Ω independent of $\hat f$ and the sample.

- Note: direct proof via Rademacher introduces norms.

Application: Classification.

Observation.
$f(x)$ and $\eta_f(x,1)$ have the same sign.
Let $\mathcal{R}_{\mathrm{z}}$ denote zero-one loss; does $\eta_{f_i} \to \bar\eta$ imply $\mathcal{R}_{\mathrm{z}}(f_i)$ converges?

Guarantee. If $\eta_{f_i} \to \bar\eta$, then

$$\limsup_{i\to\infty} \left|\mathcal{R}_{\mathrm{z}}(f_i) - \mathcal{R}_{\mathrm{z}}(\bar\eta - 1/2)\right| \;\leq\; \limsup_{i\to\infty} \left| \int_{\bar\eta=1/2} g_2(f_i) d\mu \right| \;\;.$$

- RHS can be zero.
- RHS can be positive and tight; but LHS's inner term can diverge.
- To prove consistency, need RHS zero and $\bar\eta$ agrees with $\Pr[Y = 1|X]$.

Proof sketches I: $\bar\eta$ via duality.

Duality.

$$\begin{aligned} & \inf\{\mathcal{R}(f) \;:\; f \text{ linear}\} \\ = \;& \max\left\{ -\int \ell^*(q)d\mu \;:\; q \in L_{\beta^*}(\mu) \;\;, \right. \\ & \qquad \left. \int yf(x)q(x,y)d\mu(x,y) = 0 \;\; \text{for every linear } f \right\} \;\;, \end{aligned}$$

where

- ℓ^* is conjugate to ℓ, meaning $\ell^*(s) := \sup_{r\in\mathbb{R}}(rs - \ell(r))$;
- $L_{\beta^*}(\mu)$ is the natural Orlicz space for ℓ^*;
- the constraint means: a feasible q *decorrelates* linear functions and μ.

Constructing $\bar\eta$. Dual optimum $\bar q$ always exists! Define

$$\bar\eta(x,y) \;:=\; \frac{1}{1+\frac{\bar q(x,y)}{\bar q(x,-y)}} \;\;.$$

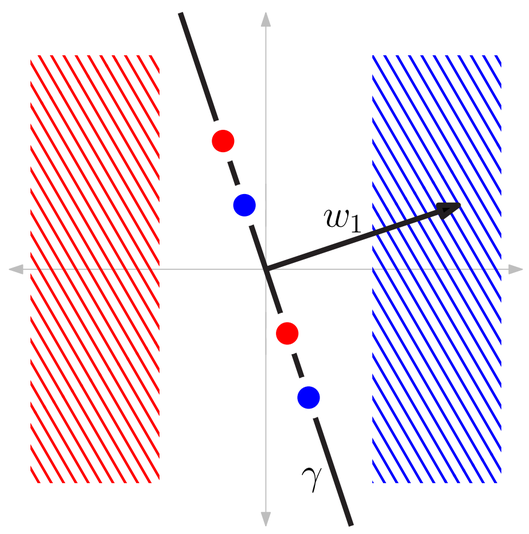
Optimality conditions. Generally no optimal linear predictor. Instead,

$$\bar f(x) \;:=\; \begin{cases} -(\ell')^{-1}(\bar q(x,1)) & \text{when } \bar\eta(x,1) \in (0,1), \\ +\infty & \text{when } \bar\eta(x,1) \geq 1, \\ -\infty & \text{when } \bar\eta(x,1) \leq 0 \end{cases}$$

satisfies roughly $\bar q(x,y) = \ell'(-y\bar f(x))$ and $\bar\eta = \eta_{\bar f}$.

Proof sketches II: Easy and difficult sets.

Stylized example.



Easy points:
w_1 perfectly separates the two big clouds.

Difficult points:
all predictors err on γ.

General case.
Define $\mathcal{D} := \{(x,y) : \bar\eta(x,y) \in (0,1)\}$.

- $\mathcal{D}$ is difficult; $\bar\eta$ (and all linear predictors) make mistakes.
- $\mathcal{D}^c$ is easy; $\bar\eta$ is perfect.

Handling easy set $\mathcal{D}^c$.

- Small $\mathcal{R}(f)$
  $\implies$ large (unnormalized) margins on most of $\mathcal{D}^c$
  $\implies$ $\eta_f \approx 1 \approx \bar\eta$ on most of $\mathcal{D}^c$.
- When $|\mathcal{H}| < \infty$, can control these margins via VC theorem.

Handling difficult set $\mathcal{D}$.

- Risk along $\mathcal{D}$ looks like a *bowl* around $\bar f$:
  for any linear f nontrivial along $\mathcal{D}$,
  $\mathcal{R}(\bar f + f) > \mathcal{R}(\bar f)$ along $\mathcal{D}$.
- Via Taylor's theorem, $\eta_{f_i} \to \bar\eta$.
- When $|\mathcal{H}| < \infty$,
  can roughly show $\mathcal{R}(\bar f + f) \geq \mathcal{R}(\bar f) + \Omega(\|f - \bar f\|_1^2)$ along $\mathcal{D}$;
  rate follows by local Rademacher tools.