Convex Risk Minimization
and
Conditional Probability Estimation

Matus Telgarsky
Miro Dudík
Robert Schapire
**Setting.** Convex risk minimizing sequences of linear predictors.

**Goal.** Convergence properties relevant to classification.

**Obstruction.** Unboundedness and infinite dimension.
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**: \(\text{Risk}(f_i) \to \inf_{f \text{ linear}} \text{Risk}(f)\), where

\[
\text{Risk}(f) := \int \ell(-yf(x))d\mu(x, y)
\]

for certain losses \(\ell\) with \(\ell'' > 0\) and \(\lim_{r \to -\infty} \ell(r) = 0\).
Predictors \((f_i)_{i \geq 1}\) with:

- \(f_i\) linear: 
  \[ f_i = \sum_{h \in \mathcal{H}} w_i(h) h, \text{ where } \mathcal{H} \ni h : \mathcal{X} \to [-1, +1]. \]

- \((f_i)_{i \geq 1}\) minimize risk: 
  \[ \text{Risk}(f_i) \to \inf_{f \text{ linear}} \text{Risk}(f), \text{ where} \]
  \[ \text{Risk}(f) := \int \ell(-yf(x))d\mu(x, y) \]
  for certain losses \(\ell\) with \(\ell'' > 0\) and \(\lim_{r \to -\infty} \ell(r) = 0\).

Conditional probability models:

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

\[ \eta_f(x, y) := \frac{1}{1 + \frac{\ell'(-yf(x))}{\ell'(yf(x))}} \quad \text{generic } \ell. \]
There exists a unique $\bar{\eta}$ so that every $(f_i)_{i \geq 1}$ with $f_i$ linear and

$$\text{Risk}(f_i) \rightarrow \inf_{f \text{ linear}} \text{Risk}(f)$$

satisfies

$$\eta f_i \rightarrow \bar{\eta}.$$
There exists a unique $\bar{\eta}$ so that every $(f_i)_{i \geq 1}$ with $f_i$ linear and

$$\text{Risk}(f_i) \longrightarrow \inf_{f \text{ linear}} \text{Risk}(f)$$

satisfies

$$\eta f_i \longrightarrow \bar{\eta}.$$

**Motivation:**

- **Risk** is only a surrogate; need to say more about $(f_i)_{i \geq 1}$.
- $(\eta_i)_{i \geq 1}$ and $\bar{\eta}$ capture what is needed for classification.
There exists a unique $\bar{\eta}$ so that every $(f_i)_{i \geq 1}$ with $f_i$ linear and

$$\text{Risk}(f_i) \longrightarrow \inf_{f \text{ linear}} \text{Risk}(f)$$

satisfies

$$\eta f_i \longrightarrow \bar{\eta}.$$  

**Generality:**

- Infinite dimensional and unbounded.
- Applies to no/weakening regularization.
Additionally, when $|\mathcal{H}| < \infty$, with $\Pr \geq 1 - \delta$ over a draw of $n \geq \Omega(\ln(1/\delta))$ examples, each linear $f$ has

$$\int |\eta_f - \bar{\eta}| d\mu \leq O \left( g(\hat{\text{Risk}}(f)) \sqrt{\text{Excess Risk}(f) + \frac{\ln(n/\delta)}{n}} \right),$$

with (nondecreasing) $g$ and $O$ independent of $f$ and the sample.
Additionally, when \(|\mathcal{H}| < \infty\), with \(\Pr \geq 1 - \delta\) over a draw of \(n \geq \Omega(\ln(1/\delta))\) examples, each linear \(f\) has

\[
\int |\eta_f - \bar{\eta}| d\mu \leq O \left( g(\hat{\text{Risk}}(f)) \sqrt{\text{Excess Risk}(f) + \frac{\ln(n/\delta)}{n}} \right),
\]

with (nondecreasing) \(g\) and \(O\) independent of \(f\) and the sample.

**Generality:**
- Finite dimensional but still unbounded.
Step 1. Craft $\tilde{\eta}$ via the dual optimum.
Step 1. Craft $\bar{\eta}$ via the dual optimum.

Step 2.
Step 1. Craft $\tilde{\eta}$ via the dual optimum.

Step 2.
Step 1. Craft $\bar{\eta}$ via the dual optimum.

Step 2.