Learning without Concentration

Shahar Mendelson

Main message I

- Learning is possible even in situations in which concentration is simply FALSE!!!

- Unnecessary assumptions give short-term dividends but may cause long term damage.

- Make sure that you focus on the problem you want to solve rather than on previously used methods – the two are not the same....

Main message II

In the context of Learning Theory, unnecessary assumptions/methods include:

- Uniformly bounded classes (or with a well behaved envelope function)
- Lipschitz losses
- Concentration methods
- Contraction methods
- Combinatorial dimension
- ?????

The goal (restricted...)

Given a class of functions F a target Y and a loss functional ℓ, let $\hat{f}$ be the empirical minimizer.

Show that with probability $1 - \delta$

$$\|\hat{f} - f^*\|^2_{L_2(\mu)} \leq \text{error}(F, N, \delta).$$

ERM

For every $f \in F$ set $\mathcal{L}_f = \ell(f(X) - Y) - \ell(f^*(X) - Y)$.

ERM produces $\hat{f}$ that lives in the (random) region of F in which

$$P_N \mathcal{L}_f \leq 0, \quad \text{while for every } f \in F, \ \ \mathbb{E}\mathcal{L}_f \geq 0.$$

The point is to exploit the gap between the empirical mean and the actual one to pin-point $\hat{f}$.

The standard method of analysis I

An exclusion argument:

with sufficiently high probability, if $\|f - f^*\|_{L_2}^2 \geq \rho_N$ then

$$\frac{1}{2}\mathbb{E}\mathcal{L}_f \leq P_N\mathcal{L}_f.$$

This implies $\|\hat{f} - f^*\|_{L_2}^2 \leq \rho_N$.

However, what is usually studied is the TWO SIDED version – showing that with high probability,

$$\sup_{\|f-f^*\|_{L_2}^2 \geq \rho_N} \left| \frac{P_N\mathcal{L}_f}{\mathbb{E}\mathcal{L}_f} - 1 \right| \leq 1/2.$$

The standard method of analysis II

Set $\xi = f^*(X) - Y$ and note that $(f(X) - Y)^2 - (f^*(X) - Y)^2$

$$= ((f - f^*)(X) + \xi)^2 - \xi^2 = (f - f^*)^2(X) + 2\xi(f - f^*)(X).$$

For every sample $(X_i, Y_i)_{i=1}^N$ let $\phi_i(t) = \ell(t + \xi_i) - \ell(\xi_i)$. Note that

$$P_N \mathcal{L}_f = \frac{1}{N} \sum_{i=1}^N \phi_i \left((f - f^*)(X_i) \right).$$

A 'simple' solution: use symmetrization, followed by contraction to 'kill the loss' and concentration to study the resulting linearized process.

This approach forces one to impose (possibly needless) assumptions on the class/loss.

An outcome

The following result is from [BBM]-2005:

Let $\mathcal{B}_{f^*}$ be the $L_2(\mu)$ ball of radius 1, centred in f^*. For any $r > 0$, set

$$k_N(r) = \mathbb{E} \sup_{f \in F \cap 2r\mathcal{B}_{f^*}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i (f - f^*)(X_i) \right|$$

('localized Rademacher complexity'), and put

$$k_N^*(\gamma) = \inf \left\{ r > 0 : k_N(r) \leq \gamma r^2 \sqrt{N} \right\}$$

If F is a convex class of functions bounded by 1 and Y is bounded by 1, then for every $t > 0$, with probability at least $1 - c_0 \exp(-t)$,

$$\|\hat{f} - f^*\|_{L_2}^2 \leq c_1 \max \left\{ \left(k_N^*(c_2)\right)^2, \frac{t}{N} \right\}.$$

Weaknesses

- The target and class have to be bounded in L_∞ and the loss has to be Lipshcitz.
- For example, the result is useless when dealing with linear regression with gaussian noise...
- Even in seemingly bounded cases, the estimate does not scale properly with the L_∞ bound (scales like R^2 – which is wrong) nor with the 'noise level' $\sigma = \mathrm{var}(\xi)$.

Example: The rate for ERM in RB_1^n (when $N \leq cn^2$) is $R^2\sqrt{\frac{\log(cn/\sqrt{N})}{N}}$, while main term in the minimax rate is $\sigma R\sqrt{\frac{\log(cn\sigma/\sqrt{N}R)}{N}}$ (actually there are four ranges depending on N, n, R and σ).

Heavy tailed classes

Just to clarify: If the class exhibits a very nice behaviour (e.g. subgaussian....) an accurate two-sided ratio estimate is possible but requires technical acrobatics.

Otherwise, two-sided concentration or ratio estimates are simply FALSE, even for a single function!!!!

Is learning still possible outside the subgaussian case, or even in a heavy-tailed problem?

(the meaning 'heavy tailed random variables' $Pr(|X| > t) \leq c/(1+t^{\beta})$ for some $\beta > 0$.....)

The difference between upper and lower estimates

In a concentration result, the upper and lower tails appear equivalent, but this is misleading! For example, the statement $\left|\frac{1}{N}\sum_{i=1}^N f^2(X_i) - \mathbb{E}f^2\right| \leq \frac{1}{2}\mathbb{E}f^2$ consists of two different inequalities.

While

$$\frac{1}{N}\sum_{i=1}^N f^2(X_i) \leq \frac{3}{2}\mathbb{E}f^2$$

only holds with constant probability (Chebyshev is sharp),

$$\frac{1}{N}\sum_{i=1}^N f^2(X_i) \geq \frac{1}{2}\mathbb{E}f^2$$

holds with probability at least $1 - 2\exp(-cN)$.

Thankfully, we only need the lower tail...

Splitting the excess loss - the two regimes

- A multiplier process $f \to \frac{1}{N}\sum_{i=1}^{N} \xi_i (f - f^*)(X_i)$.
- A quadratic process $f \to \frac{1}{N}\sum_{i=1}^{N} (f - f^*)^2(X_i)$.

- The interaction of the class with the noise is only via the 'multiplier' term.
- 'low noise' problems depend only on the quadratic component. (e.g., in a noise-free problem mistakes can only occur by selecting functions with a perfect fit on the data).
- For higher 'noise levels', in which the the multiplier component is dominant mistakes may occur because of the interaction of the noise with class member.

Controlling the multiplier component

Let $\xi = f^*(X) - Y$. Roughly put, $\alpha_N^*(\kappa, \delta)$ is the 'localization level' in which, with probability $1 - \delta$,

$$\sup_{f \in F \cap s\mathcal{B}_{f^*}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i \xi_i (f - f^*)(X_i) \right| \sim s^2 \sqrt{N}.$$

(should be compared with:

$$\max\{\|\xi\|_{L_\infty}, \sup_{f \in F} \|f\|_{L_\infty}\} \sup_{f \in F \cap s\mathcal{B}_{f^*}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i (f - f^*)(X_i) \right| \sim s^2 \sqrt{N},$$

in [BBM]/contraction-concentration based argument).

The idea: show that when $\|f - f^*\|_{L_2} \geq \alpha_N^*(\kappa, \delta)$,

$$\left| \frac{2}{N} \sum_{i=1}^N \xi_i (f - f^*)(X_i) \right| \leq \kappa \|f - f^*\|_{L_2}^2.$$

Controlling the quadratic component

Roughly put, $\beta_N^*(\kappa)$ is the 'localization level' in which

$$\mathbb{E} \sup_{f \in F \cap r\mathcal{B}_{f^*}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i (f - f^*)(X_i) \right| \sim \kappa \sqrt{N} r.$$

(should be compared with

$$\sup_{f \in F} \|f\|_{L_\infty} \sup_{f \in F \cap r\mathcal{B}_{f^*}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i (f - f^*)(X_i) \right| \sim \sqrt{N} r^2,$$

in [BBM]/contraction-concentration based argument).

The idea: show that when $\|f - f^*\|_{L_2} \geq \beta_N^*(\kappa)$,

$$\frac{1}{N} \sum_{i=1}^N (f - f^*)^2(X_i) \geq \kappa \|f - f^*\|_{L_2}^2.$$

The small-ball assumption

Let H a class of functions. Set

$$Q_H(u) = \inf_{h \in H} Pr(|h| \geq u \|h\|_{L_2}).$$

Let $F - F = \{f - h : f, h \in F\}$. Assume that there is some $u > 0$ for which

$$Q_{F-F}(u) > 0.$$

Learning without concentration

- Let $F \subset L_2$ be a closed, convex class of functions and set $Y \in L_2$ to be the unknown target.
- Fix $\tau > 0$ for which $Q_{F-F}(2\tau) > 0$.
- Set $\kappa < \tau^2 Q_{F-F}(2\tau)/16$.

For every $\delta > 0$, with probability at least $1 - \delta - \exp(-NQ^2_{F-F}(2\tau)/2)$,

$$\|\hat{f} - f^*\|_{L_2} \leq 2 \max\left\{ \alpha^*_N(\kappa, \delta/4), \beta^*_N\left(\frac{\tau Q_{F-F}(2\tau)}{16}\right)\right\}.$$

This estimate actually gives the minimax rates (e.g for independent gaussian noise), under rather minimal assumptions on the class.