A Geometric Approach to Statistical Learning Theory

Shahar Mendelson

Centre for Mathematics and its Applications

The Australian National University

Canberra, Australia

THE AUSTRALIAN
NATIONAL UNIVERSITY

What is a learning problem

- A class of functions F on a probability space (Ω, μ)
- A random variable Y one wishes to estimate
- A loss functional ℓ
- The information we have: a sample $(X_i, Y_i)_{i=1}^{n}$

Our goal: with high probability, find a good approximation to Y in F with respect to ℓ, that is

- Find $f \in F$ such that $\mathbb{E}\ell(f(X), Y))$ is "almost optimal".
- f is selected according to the sample $(X_i, Y_i)_{i=1}^{n}$.

Example

Consider

- The random variable Y is a fixed function $T : \Omega \to [0,1]$ (that is $Y_i = T(X_i)$).
- The loss functional is $\ell(u,v) = (u-v)^2$.

Hence, the goal is to find some $f \in F$ for which

$$\mathbb{E}\ell(f(X), T(X)) = \mathbb{E}(f(X) - T(X))^2 = \int_\Omega (f(t) - T(t))^2 d\mu(t)$$

is as small as possible.

To select f we use the sample $X_1, ..., X_n$ and the values $T(X_1), ..., T(X_n)$.

A variation

Consider the following *excess loss*:

$$\bar{\ell}_f(x) = (f(x) - T(x))^2 - (f^*(x) - T(x))^2 = \ell_f(x) - \ell_{f^*}(x),$$

where f^* minimizes $\mathbb{E}\ell_f(x) = \mathbb{E}(f(X) - T(X))^2$ in the class.

The difference between the two cases:

Our Goal

- Given a fixed sample size, how close to the optimal can one get using empirical data?
- How does the specific choice of the loss influence the estimate?
- What parameters of the class F are important?
- Although one has access to a random sample, the measure which generates the data is not known.

The algorithm

Given a sample $(X_1, ..., X_n)$, select $\hat{f} \in F$ which satisfies

$$\text{argmin}_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \ell_f(X_i),$$

that is, $\hat{f}$ is the "best function" in the class on the data.

The hope is that with high probability $\mathbb{E}(\ell_{\hat{f}} | X_1, ..., X_n) = \int \ell_{\hat{f}}(t) d\mu(t)$ is close to the optimal.

In other words, hopefully, with high probability, the empirical minimizer of the loss is "almost" the best function in the class with respect to the loss.

Back to the squared loss

In the case of the squared excess loss -

$$\bar{\ell}_f(x) = (f(x) - T)^2 - (f^*(x) - T(x))^2,$$

since the second term if the same for every $f \in F$, the empirical minimization selects

$$\text{argmin}_{f \in F} \sum_{i=1}^{n} (f(X_i) - T(X_i))^2$$

and the question is

how to relate this *empirical distance to the “real” distance we are interested in.*

Analyzing the algorithm

For a second, let's forget the loss, and from here on, to simplify notation, denote by G the loss class.

We shall attempt to connect $\frac{1}{n}\sum_{i=1}^{n} g(X_i)$ (i.e. the random, empirical structure on G) to $\mathbb{E}g$.

We shall examine various notions of similarity of the structures.

Note: in the case of an excess loss, $0 \in G$ and our aim is to be as close to 0 as possible. Otherwise, our aim is to approach $g^* \neq 0$.

A road map

- Asking the “correct” question - beware of loose methods of attack.
- Properties of the loss and their significance.
- Estimating the complexity of a class.

A little history

Originally, the study of $\{0,1\}$-valued classes (e.g. Perceptrons) used the *uniform law of large numbers*:

$$Pr\left(\exists g \in G \left|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}g\right| \geq \varepsilon\right),$$

which is a *uniform measure of similarity*.

If the probability of this is small, then for every $g \in G$, the empirical structure is "close" to the real one. In particular, this is true for the empirical minimizer, and thus, on the good event,

$$\mathbb{E}\hat{g} \leq \frac{1}{n}\sum_{i=1}^{n} \hat{g}(X_i) + \varepsilon.$$

In a minute: this approach is suboptimal!!!!

Why is this bad?

Consider the excess loss case.

- We hope that the algorithm will get us close to 0...
- So, it seems likely that we would only need to control the part of G which is not too far from 0.
- No need to control functions which are far away, while in the ULLN, we control *every* function in G.

Why would this lead to a better bound?

Well, first, the set is smaller...

More important:

- functions close to 0 in expectation are likely to have a small variance (under mild assumptions)...

On the other hand,

- Because of the CLT, for every fixed function $g \in G$ and n large enough, with probability $1/2$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} g(X_i) - \mathbb{E}g \right| \sim \sqrt{\frac{\mathrm{var}(g)}{n}},$$

So?

- Control over the entire class $\Longrightarrow$ control over functions with nonzero variance $\Longrightarrow$ rate of convergence can't be better than $c/\sqrt{n}$.

- If $g^* \neq 0$, we can't hope to get a faster rate than $c/\sqrt{n}$ using this method.

- This shows the statistical limitation of the loss.

What does this tell us about the loss?

- To get faster convergence rates one has to consider the excess loss.
- We also need a condition that would imply that if the expectation is small, the variance is small (e.g $\mathbb{E}\ell_f^2 \leq B\mathbb{E}\ell_f$ - A Bernstein condition).

  It turns out that this condition is connected to convexity properties of ℓ at 0.
- One has to connect the richness of G to that of F (which follows from a Lipshitz condition on ℓ).

Localization - Excess loss

There are several ways to localize.

- It is enough to bound

$$Pr\left(\exists g \in G \;\; \frac{1}{n}\sum_{i=1}^{n} g(X_i) \leq \varepsilon \;\; \mathbb{E}g \geq 2\varepsilon\right)$$

- This event upper bounds the probability that the algorithm fails. If this probability is small , and since $n^{-1}\sum_{i=1}^{n}\hat{g}(X_i) \leq \varepsilon$, then $\mathbb{E}\hat{g} \leq 2\varepsilon$.

- Another (similar) option: relative bounds:

$$Pr\left(\exists g \in G \;\; \left|\frac{n^{-1/2}\sum_{i=1}^{n}(g(X_i) - \mathbb{E}g)}{\sqrt{\mathrm{var}(g)}}\right| \geq \varepsilon\right)$$

Comparing structures

Suppose that one could find r_n, for which, with high probability, for every $g \in G$ with $\mathbb{E}g \geq r_n$,

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

(here, $1/2$ and $3/2$ can be replaced by $1-\varepsilon$ and $1+\varepsilon$).

Then if $\hat{g}$ was produced by the algorithm it can either

- have a "large expectation" - $\mathbb{E}\hat{g} \geq r_n$, $\Longrightarrow$

- The structures are similar and thus $\mathbb{E}\hat{g} \leq \frac{2}{n}\sum_{i=1}^{n} \hat{g}(X_i)$,

Or

- have a "small expectation" $\Longrightarrow \mathbb{E}\hat{g} \leq r_n$,

Comparing structures II

Thus, with high probability

$$\mathbb{E}\hat{g} \le \max\left\{ r_n,\ \frac{2}{n}\sum_{i=1}^{n} \hat{g}(X_i) \right\}.$$

This result is based on a *ratio limit theorem*, because we would like to show that

$$\sup_{g \in G, \mathbb{E}g \ge r_n} \left| \frac{n^{-1}\sum_{i=1}^{n} g(X_i)}{\mathbb{E}g} - 1 \right| \le \varepsilon.$$

This normalization is possible if $\mathbb{E}g^2$ can be bounded using $\mathbb{E}g$ (which is a property of the loss). Otherwise, one needs a slightly different localization.

Star shape

If G is star-shaped, its "relative richness" increases as r becomes smaller.

Why is this better?

Thanks to a star-shape assumption, our aim is to find the smallest r_n such that with high probability,

$$\sup_{g \in G,\ \mathbb{E}g = r} \left| \frac{1}{n} \sum_{i=1}^{n} g(X_i) - \mathbb{E}g \right| \leq \frac{r}{2}.$$

This would imply that the error of the algorithm is at most r.

For the non-localized result, to obtain the same error, one needs to show

$$\sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^{n} g(X_i) - \mathbb{E}g \right| \leq r,$$

where the supremum is on a much larger set, and includes functions with a “large” variance.

BTW, even this is not optimal....

It turns out that a structural approach (uniform or localized) does not give the best result that one could get on the error of the EM algorithm.

A sharp bound follows from a direct analysis of the algorithm (under mild assumptions) and depends on the behavior of the (random) function

$$\hat{\phi}_n(r) = \sup_{g \in G,\ \mathbb{E}g = r} \left| \frac{1}{n} \sum_{i=1}^{n} g(X_i) - \mathbb{E}g \right| .$$

Application of concentration

Suppose that we can show that with high probability, for every r,

$$\sup_{g \in G,\ \mathbb{E}g=r} \left| \frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}g \right| \sim \mathbb{E} \sup_{g \in G,\ \mathbb{E}g=r} \left| \frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}g \right| \equiv \phi_n(r)$$

i.e. that the expectation is a good estimation of the random variable. Then,

- The uniform estimate: the error is close to $\phi_n(1)$.
- The localized estimate: the error is close to the fixed point: $\phi_n(r^*) = r^*/2$.
- The direct analysis ......

A short summary

- bounding the right quantity - one needs to understand
$$Pr\left(\sup_{g\in A}\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}g\right| \geq t\right).$$
- For an estimate on EM, $A \subset G$. The smaller we can take A - the better the bound!

- loss class vs. excess loss: being close to 0 (hopefully) implies small variance (property of the loss).

- One has to connect the "complexity" of $A \subset G$ to the complexity of the subset of the base class F that generated it.
- If one considers excess loss (better statistical error), there is a question of the approximation error.

Estimating the complexity

- Concentration:

$$\sup_{g\in A}\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}g\right| \sim \mathbb{E}\sup_{g\in A}\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}g\right| .$$

- symmetrization

$$\mathbb{E}\sup_{g\in A}\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}g\right| \sim \mathbb{E}_X\mathbb{E}_\varepsilon\sup_{g\in A}\left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i g(X_i)\right| = R_n(A).$$

- ε_i are independent, $Pr(\varepsilon = 1) = Pr(\varepsilon = -1) = 1/2$.

Estimating the Complexity II

- For $\sigma = (X_1, ..., X_n)$ consider $P_\sigma A = \{(g(X_1), ..., g(X_n))\}$.

- Then

$$R_n(A) = \mathbb{E}_X \left(\mathbb{E}_\varepsilon \sup_{v \in P_\sigma A} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i v_i \right| \right).$$

- For every (random) coordinate projection $V = P_\sigma A$, the complexity parameter $\mathbb{E}_\varepsilon \sup_{v \in P_\sigma A} \left| n^{-1} \sum_{i=1}^{n} \varepsilon_i v_i \right|$ measures the correlation of V with a "random noise".

- The noise model: a random point in $\{-1, 1\}^n$ (the n-dimensional combinatorial cube), i.e., $(\varepsilon_1, ..., \varepsilon_n)$. $R_n(A)$ measures how V is correlated with this noise.

Example - A large set

- If the class of functions is bounded by 1 then for any sample $(X_1, ..., X_n)$, $P_\sigma A \subset [-1,1]^n$.
- For $V = [-1,1]^n$ (or $V = \{-1,1\}^n$),

  $$\frac{1}{n}\mathbb{E}_\varepsilon \sup_{v \in P_\sigma A} \left| \sum_{i=1}^n \varepsilon_i v_i \right| = 1,$$

  (If there are many "large" coordinate projections, R_n does not converge to 0 as $n \to \infty$!).
- Question: what subsets of $[-1,1]^n$ are big in the context of this complexity parameter?

Combinatorial dimensions

Consider a class of $\{-1,1\}$-valued functions. Define the Vapnik-Chervonenkis dimension of A by

$$vc(A) = \sup \left\{ |\sigma| \mid P_\sigma A = \{-1,1\}^{|\sigma|} \right\}.$$

In other words, $vc(A)$ is the largest dimension of a coordinate projection of A which is the entire (combinatorial) cube.

There is a real-valued analog of the Vapnik-Chervonenkis dimension, which is called the *combinatorial dimension.*

Combinatorial dimensions II

The combinatorial dimension:

For every ε, it measures the largest dimension $|\sigma|$ of a “cube” of side length ε that can be found in a coordinate projection $P_\sigma A$.

Some connection between the parameters

- If $vc(A) \leq d$ then $R_n(A) \leq c\sqrt{d/n}$.

- If $vc(A,\varepsilon)$ is the combinatorial dimension of A at scale ε, then
$$R_n(A) \leq \frac{C}{\sqrt{n}} \int_0^\infty \sqrt{vc(A,\varepsilon)} d\varepsilon.$$

Note: These bounds on R_n are (again) not optimal and can be improved in various ways. For example:

- The bounds take into account the worst case projection - not the average projection.

- The bound does not take into account the diameter of A.

Important

The vc dimension (combinatorial dimension) and other related complexity parameters (e.g covering numbers) are only ways of upper bounding R_n.

Sometimes such a bound is good, but at times it is not.

Although the connections between the various complexity parameters are very interesting and nontrivial, for SLT it is always best to try and bound R_n directly.

Again, the difficulty of the learning problem is captured by the "richness" of a random coordinate projection of the loss class.

Example: Error rates for VC classes

Let

- F be a class of $\{0,1\}$-valued functions with $vc(F) \leq d$ and $T \in F$. (Proper learning)
- ℓ is the squared loss and G is the loss class. Note, $0 \in G$!
- $H = \mathrm{star}(G,0) = \{\lambda g \mid g \in G\}$ is the star-shaped hull of G.

Example: Error rates for VC classes II

Then:

- Since $\ell_f(x) \geq 0$ and functions in F are $\{0,1\}$-valued, then $\mathbb{E}h^2 \leq \mathbb{E}h$.

- The error rate is upper bounded by the fixed point of

$$\mathbb{E} \sup_{h \in H,\ \mathbb{E}h = r} \left| n^{-1} \sum_{i=1}^{n} \varepsilon_i h(X_i) \right| = R_n(H_r),$$

  i.e. when

$$R_n(H_r) = \frac{r}{2}.$$

- The next step is to relate the complexity of H_r to the complexity of F.

Bounding $R_n(H_r)$

- F is small in the appropriate sense.
- ℓ is a Lipschitz function, and thus $G = \ell(F)$ is not much larger than F.
- The star-shaped hull of G is not much larger than G.

In particular, for every $n \geq d$, with probability larger than $1 - \left(\frac{ed}{n}\right)^{c'd}$, if $\mathbb{E}_n \hat{g} \leq \inf_{g \in G} \mathbb{E}_n g + \rho$, then

$$\mathbb{E}\hat{g} \leq c \max\left\{\frac{d}{n}\log\left(\frac{n}{ed}\right), \rho\right\}.$$