Empirical Minimization

Jerome Le Ny

May 18, 2007

1 Introduction

This report is a summary of the paper [BM06] of Peter Bartlett and Shahar Mendelson on Empirical Minimization.

Typical algorithms in machine learning minimize an empirical loss. This is a particular case of minimum contrast estimation in statistics. To recall the set-up, a learning algorithm is presented with a set of i.i.d. input-output pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, drawn from a probability distribution P on the space $\mathcal{X} \times \mathcal{Y}$, with P unknown. We would like to find a function $g : \mathcal{X} \to \mathcal{Y}$ which, given a new input X, accurately predicts the output Y, where (X, Y) is again distributed according to P. Accurately means that we have a loss function $L : \mathcal{Y}^2 \to [0, 1]$, and we would like to find g which minimizes the risk $PL(g(X), Y)$ [1]. Since P in unknown and only the sequence of samples is given, in empirical risk minimization we choose g to minimize the sample average of $L(g(x), y)$ instead, hoping that this function will "generalize well" for new inputs.

The general problem that we are led to study is as follows. Given i.i.d. samples $X_1, \ldots, X_n$ (which are the pairs of inputs/outputs above) with values in a set $\mathcal{X}$, we define the empirical measure $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$. We compute an empirical minimizer $\hat{f}$ over a class $\mathcal{F}$ of real-valued functions on $\mathcal{X}$

$$P_n \hat{f} = \inf_{f \in \mathcal{F}} P_n f,$$

or a ρ-approximate minimizer verifying

$$P_n \hat{f} \leq \inf_{f \in \mathcal{F}} P_n f + \rho.$$

We then want to estimate the expectation of this empirical minimizer

$$\mathbb{E}\left[\hat{f}(X) | X_1, \ldots, X_n\right]$$

which we just denote $P\hat{f}$ in the following. The authors argue that by considering relative loss functions instead of loss functions in the scenario described above, it is natural to assume that for every $f \in \mathcal{F}$, $Pf \geq 0$, although the functions in $\mathcal{F}$ can take negative values.

The paper considers several approaches to estimating the expectation of the empirical minimizer, which we review in the following. Roughly, the idea is that there are relevant mild assumptions that one can make

[1] For P a probability measure, in the following I will write $Pf := \int f dP$

on the class $\mathcal{F}$ which help improve significantly the traditonal upper bounds on $P\hat{f}$. It is also possible to give upper bounds directly on the empirical minimizer instead of bounds valid for the whole class $\mathcal{F}$, and an example is given where this approach gives much better results.

Throughout this paper we assume that $\mathcal{F}$ is a uniformly bounded class of functions and we denote $b_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \|f\|_\infty$. The proofs rely heavily on the functional version of Bernstein's inequality due to Talagrand [Tal94], which requires this uniform bound.

2 Structural Approach to Comparing the Empirical and Actual Measures

In the classical approach, we compare the deviation between Pf and P_nf over the entire class $\mathcal{F}$, assuming that the algorithm might return any function of the class for $\hat{f}$. Then if $P_n\hat{f}$ is small, and the worst case deviation is small, $P\hat{f}$ will be small obviously.

2.1 The Uniform Law of Large Numbers

We say that a class $\mathcal{F}$ satisfies the uniform law of large numbers with respect to a probability measure P if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr(\|P - P_n\|_{\mathcal{F}} \geq \epsilon) = 0,$$

where $\|P - P_n\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |Pf - P_nf|$. This leads to the following notion of similarity, uniform on the entire class, between the empirical and actual measures.

Definition 2.1. Given an integer n and a probability measure P, we say that the empirical and actual structures on $\mathcal{F}$ are (λ, δ)-close if

$$\Pr(\|P - P_n\|_{\mathcal{F}} \geq \lambda) \leq \delta.$$

In particular, if $0 \in \mathcal{F}$ and the empirical and actual structures are (λ, δ)-close, then a ρ-empirical minimizer verifies $P\hat{f} \leq \inf_{f\in\mathcal{F}} P_nf + \lambda + \rho \leq \lambda + \rho$ with probability at least $1 - \delta$.

In the following, let $\sigma_{\mathcal{F}}^2 = \sup_{f\in\mathcal{F}} \mathrm{var}[f]$ and recall $b_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \|f\|_\infty$.

The paper provides a set of results to evaluate the benefits and limitations of using this notion of similarity.

Theorem 2.1. *There exists absolute constants C, C', c, c' for which the following holds.*

1. *For any class of functions $\mathcal{F}$, and every $0 < \delta < 1$, the empirical and actual structures are (λ_n, δ)-close provided that*
$$\lambda_n \geq C \max\left\{ \mathbb{E}\|P - P_n\|_{\mathcal{F}}, \sigma_{\mathcal{F}} \sqrt{\frac{\log(1/\delta)}{n}}, \frac{b_{\mathcal{F}} \log(1/\delta)}{n} \right\}.$$
2. *Let $\mathcal{F}$ be a class of functions such that $b_{\mathcal{F}} \leq 1$. Then*
$$\mathbb{E}\|P - P_n\|_{\mathcal{F}} \geq c\frac{\sigma_{\mathcal{F}}}{\sqrt{n}}.$$

Furthermore, for every integer $n \geq 1/\sigma_{\mathcal{F}}^2$, with probability at least c',

$$\|P - P_n\|_{\mathcal{F}} \geq C' \, \mathbb{E}\|P - P_n\|_{\mathcal{F}}.$$

A conclusion we can draw from this theorem is the following. The only assumption made here is that the class $\mathcal{F}$ is uniformly bounded with uniformly bounded variances. The notion of similarity in definition 2.1 involves bounding $\mathbb{E}\|P - P_n\|_{\mathcal{F}}$, and we obtain empirical and actual structures which are close uniformly on $\mathcal{F}$. On the other hand, λ_n cannot decrease faster than $1/\sqrt{n}$, and it is impossible to use this notion of similarity to obtain an asymptotic result stronger than $P\hat{f} \leq 1/\sqrt{n}$.

In practice this approach is too conservative and we would like to find better bounds. This is done in the following with some additional assumptions on the function class $\mathcal{F}$. The approach in section 2.2 is still based on considering the distance between the measures on the whole class, whereas in section 3 we describe the results that the authors obtained by also taking into account the properties of the empirical minimizer.

2.2 Isomorphic Coordinate Projections

Let us denote the sample by $\tau = (X_1, \ldots, X_n)$ and define on the class of functions $\mathcal{F}$ the (random) coordinate projection $\Pi_\tau : f \to (f(X_1), \ldots, f(X_n))$. Here we will consider the following notion of similarity between the empirical and actual measures.

Definition 2.2. We say that the coordinate projection Π_τ is an ϵ-isomorphism if for every $f \in \mathcal{F}$,

$$(1 - \epsilon)Pf \leq P_n f \leq (1 + \epsilon)Pf.$$

We will also need the following definition, which says that the variance of the elements of $\mathcal{F}$ decreases as their expectation decreases.

Definition 2.3. We say that $\mathcal{F}$ is a (β, B)-Bernstein class with respect to the probability measure P (where $0 < \beta \leq 1$ and $B \geq 1$), if every f in $\mathcal{F}$ satisfies

$$P(f^2) \leq B(Pf)^\beta.$$

Note that if f belong to a Bernstein class, then necessarily $Pf \geq 0$. The paper gives references showing that certain relative loss functions are Bernstein.

We will need an additional assumption on the class $\mathcal{F}$. We say that $\mathcal{F}$ is *star-shaped around* 0 if for every $0 \leq a \leq 1$ and any $f \in \mathcal{F}, af \in \mathcal{F}$. Let $\mathcal{F}_\lambda = \{f \in \mathcal{F} : Pf = \lambda\}$. First we have:

Lemma 2.2. *Let $\mathcal{F}$ be star-shaped around 0 and let $\tau \in \mathcal{X}^n$. For any $\lambda > 0$ and $0 < \epsilon < 1$, the projection Π_τ is an ϵ-isomorphism of $\mathcal{F}_\lambda$ if and only if it is an ϵ-isomorphism of $\{f \in \mathcal{F} : Pf \geq \lambda\}$.*

With these definitions, the authors show the following result.

Theorem 2.3. *There is an absolute constant c for which the following holds. Let $\mathcal{F}$ be a class of functions, such that for every $f \in \mathcal{F}$, $\|f\|_\infty \leq b_{\mathcal{F}}$. Assume that $\mathcal{F}$ is a (β, B)-Bernstein class and suppose that*

$\lambda, x, n, 0 < \epsilon < 1$ *and* $0 < \theta < 1$ *satisfy*

$$\lambda \geq c \max\left\{ \frac{b_{\mathcal{F}}\, x}{n\theta^2\epsilon}, \left(\frac{Bx}{n\theta^2\epsilon^2}\right)^{1/(2-\beta)} \right\}.$$

Then

1. *If* $E\|P - P_n\|_{\mathcal{F}_\lambda} \geq (1+\theta)\lambda\epsilon$, *then*

$$Pr\left(\Pi_\tau \text{ is not an } \epsilon\text{-isomorphism of } \mathcal{F}_\lambda\right) \geq 1 - e^{-x}.$$

2. *If* $E\|P - P_n\|_{\mathcal{F}_\lambda} \leq (1-\theta)\lambda\epsilon$, *then*

$$Pr\left(\Pi_\tau \text{ is an } \epsilon\text{-isomorphism of } \mathcal{F}_\lambda\right) \geq 1 - e^{-x},$$

and if moreover $\mathcal{F}$ *is star-shaped around* 0 *and* $0 < \lambda < 1$, *every* $f \in \mathcal{F}$ *satisfies*

$$Pr\left(Pf \leq \max\left\{\frac{P_n f}{1-\epsilon}, \lambda\right\}\right) \geq 1 - e^{-x}.$$

Note that the second assertion of 2 follows immediately from the first and lemma 2.2, considering the sets $\{f \in \mathcal{F} : Pf < \lambda\}$ and $\{f \in \mathcal{F} : Pf \geq \lambda\}$ separately.

Example 2.1. If $\mathcal{F}$ consists of nonnegative functions uniformly bounded by $b_{\mathcal{F}}$ (properties frequently assumed for loss functions), then clearly $\mathcal{F}$ is a $(1, b_{\mathcal{F}})$-Bernstein class. If in addition $b_{\mathcal{F}} = 1$, the condition on λ in the theorem becomes

$$\lambda \geq c\frac{x}{n\theta^2\epsilon^2},$$

which improves, with these additional assumptions, on the $1/\sqrt{n}$ bound of the previous paragraph.

To close this paragraph, I'm adding another theorem proved in this paper, which is similar to theorem 2.3, with an additional final part which will be used in section 3.1 to compare the different approaches. First, let

$$\xi_n(r) = E\sup\{Pf - P_n f : f \in \mathcal{F}, Pf = r\}, \quad \text{and}$$
$$\xi_n(r_1, r_2) = E\sup\{Pf - P_n f : f \in \mathcal{F}, r_1 \leq Pf \leq r_2\}.$$

Here is the theorem. Again, note that the first part is similar to the last part of theorem 2.3, and the second part is obtained by fixing the values of ϵ, θ and taking $\lambda = r'$ in the first part.

Theorem 2.4. *There is an absolute constant* c *for which the following holds. Let* $\mathcal{F}$ *be a* (β, B)*-Bernstein class of functions bounded by* $b_{\mathcal{F}}$ *which is star-shaped around* 0. *Then for any* $0 < \theta, \epsilon, \lambda < 1$ *satisfying*

$$\lambda \geq \max\left\{ \frac{\xi_n(\lambda)}{(1-\theta)\epsilon}, c\frac{b_{\mathcal{F}}\, x}{n\theta^2\epsilon}, c\left(\frac{Bx}{n\theta^2\epsilon^2}\right)^{1/(2-\beta)} \right\}.$$

every $f \in \mathcal{F}$ *satisfies*

$$Pr\left(Pf \leq \max\left\{\frac{P_n f}{1-\epsilon}, \lambda\right\}\right) \geq 1 - e^{-x}.$$

In particular, there is an absolute constant c such that if

$$r' = \max\left\{\inf\left\{r > 0 : \xi_n(r) \le r/4\right\}, \frac{c\, b_{\mathcal{F}}\, x}{n}, c\left(\frac{Bx}{n}\right)^{1/(2-\beta)}\right\}$$

then a ρ-approximate empirical minimizer $\hat{f} \in \mathcal{F}$ satisfies

$$Pr\left(P\hat{f} \le \max\{2\rho, r'\}\right) \ge 1 - e^{-x}.$$

3 Direct Approach

This part deals with a direct analysis of the empirical mimizer, and an example in section 3.1 shows that this approach can yield in certain cases much sharper estimates than by means of a structural result which holds for every function in the class. The goal is to show that the dominant term in the upper bound on $P\hat{f}$ is, roughly,

$$\mathrm{argmax}_{r>0}(\xi_n(r) - r).$$

The theorem below will give upper and lower bounds on $P\hat{f}$, not using this maximizer directly, but values of r that almost maximize this quantity. For $\epsilon > 0$, define

$$r_{\epsilon,+} = \sup\left\{0 \le r \le b_{\mathcal{F}} : \xi_n(r) - r \ge \sup_{s>0}(\xi_n(s) - s) - \epsilon\right\},$$
$$r_{\epsilon,-} = \inf\left\{0 \le r \le b_{\mathcal{F}} : \xi_n(r) - r \ge \sup_{s>0}(\xi_n(s) - s) - \epsilon\right\}.$$

The theorem shows that, with a suitable choice of ϵ, not too small, $P\hat{f}$ is approximately between $r_{\epsilon,-}$ and $r_{\epsilon,+}$.

Theorem 3.1. *For any $c_1 > 0$, there is a constant c, depending only on c_1, such that the following holds. Let $\mathcal{F}$ be a (β, B)-Bernstein class that is star-shaped around 0, and such that for every $f \in \mathcal{F}$, $\|f\|_\infty \le b_{\mathcal{F}}$. For x given, let*

$$r' = \max\left\{\inf\left\{r > 0 : \xi_n(r) \le r/4\right\}, \frac{c\, b_{\mathcal{F}}(x + \log n)}{n}, c\left(\frac{B(x + \log n)}{n}\right)^{1/(2-\beta)}\right\}$$

For $0 \le \rho \le r'/2$, let $\hat{f}$ denote a ρ-approximate empirical risk minimizer. If

$$\epsilon \ge c\left(\max\left\{\sup_{s>0}(\xi_n(s) - s), r'^{\beta}\right\}\frac{(B+b)(x+\log n)}{n}\right)^{1/2} + \rho,$$

1. *then*
$$Pr\left(P\hat{f} \le \max\left\{\frac{1}{n}, r_{\epsilon,+}\right\}\right) \ge 1 - e^{-x}.$$
2. *If moreover*
$$\xi_n(0, c_1/n) < \sup_{s>0}(\xi_n(s) - s) - \epsilon,$$
then
$$Pr\left(P\hat{f} \ge r_{\epsilon,-}\right) \ge 1 - e^{-x}.$$

Example 3.1. The authors mention that it is easy to see, for the case of nonnegative functions as in example 2.1, that this theorem gives the same upper bound. Somehow I don't find it that easy.

Remark. Note the important difference in this theorem with the previous results: it gives bounds valid for $P\hat{f}$ only, not for the whole class $\mathcal{F}$.

3.1 Comparison of the Approaches

Recall the assumptions made: $\mathcal{F}$ is a star-shaped class of uniformly bounded functions which satisfies a Bernstein condition (which implies $Pf \geq 0$, for every $f \in \mathcal{F}$). In this part we mention the existence of such a class $\mathcal{F}$ for which the direct approach yields a much better upper bound on $P\hat{f}$ than the structural approach. The theorem is as follows.

Theorem 3.2. *There is an absolute constant c for which the following holds. If $0 < \delta < 1$ and n is sufficiently large, there is a probability measure P and a $(1,2)$-Bernstein, star-shaped class $\mathcal{F}$, with $b_{\mathcal{F}} = 1$, such that*

1. *For every $X_1, \ldots, X_n$, there is a function $f \in \mathcal{F}$ with $Pf = 1/4$ and $P_n f = 0$.*
2. *For the class $\mathcal{F}$, the function ξ_n satisfies:*

$$\xi_n(r) = \begin{cases} (n+1)r & \text{if } 0 < r \leq 1/n \\ r & \text{if } 1/n < r \leq 1/4 \\ 0 & \text{if } r > 1/4. \end{cases}$$

   Thus, $\inf\{r > 0 : \xi_n(r) \leq r/4\} = 1/4$.

3. *If $\hat{f}$ is a ρ-approximate minimizer, where $0 \leq \rho < 1/8$, then with probability larger than $1 - \delta$,*

$$\frac{1}{n}\left(1 - c\sqrt{\frac{\log n}{n} - \rho}\right) \leq P\hat{f} \leq \frac{1}{n}.$$

Part 1 says that no coordinate projection can be an ϵ-isomorphism (for any $0 < \epsilon < 1$). In particular, part 2 says that theorem 2.4 can only give the upper bound $1/4$. More generally, any kind of similar structural approach, proving inequalities valid for the whole class $\mathcal{F}$, can only provide trivial bounds in this case. On the other hand, part 3 is proved in the paper as a direct application of theorem 3.1, which is thus seen to provide a much better upper bound in this case.

References

[BM06] P.L. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, July 2006.

[Tal94] M. Talagrand. Sharper bounds for gaussian and empirical processes. *Annals of Probability*, 22:28–76, 1994.