

Lecture 1 — February 18th

*Lecturer: Francis Bach**Scribe: Benjamin Goehry, Antoine Havet-Morel*

1.1 Motivation

Nowadays, data are everywhere and it has become crucial (as well industrially as scientifically) to learn from this always growing amount of data. As the way to obtain data has improved, the major problem faced in this context is that we collect many instances of a single phenomenon which can be quite complex in terms of data : that is we get n observations of an object in a d -dimensional space with n and d “quite big”.

For instance when someone is searching information on a web search engine; in order to recommend the most likely interesting web pages the search engine has a database composed of n users of the engine and d variables assigned to existing web pages indicating if they were visited or not. The same situation appears in marketing where the goal is to make the “best” personalized recommendation to one potential buyer knowing some information about him/her and other users.

Such a problem appears also in scientific contexts, for example when we aim at visual object recognition, d is the “size” of the picture we want to give information about and n is the number of pictures we already have information about or in bioinformatics when it comes about dealing with the millions of proteins playing important roles and represented by very complex structured data.

In this course we will thus consider a large-scale machine learning context, that is :

- the number of observations n will be large.
- the dimension d of each observation will be large as well.

Our main objective will be to show how to deal with the problem of machine learning in this large-scale context with statistical and optimizational tools allowing a tractable running-time complexity; the ideal complexity being the time needed to read the data : $O(nd)$, or more generally the number of non-zeros when dealing with sparse inputs.

1.2 Supervised Learning

Context : We’re looking at n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1, \dots, n$ i.i.d and knowing x we would like to give a prediction of an unseen y thanks to a function $\theta^\top \Phi(x)$ linear in the “features” $\Phi(x) \in \mathbb{R}^d$.

To perform this task, the most common approach is regularized empirical risk minimization which main goal is to find $\hat{\theta}$ which minimizes the regularized empirical risk, that is, to

solve the optimization problem :

$$\min_{\theta \in \mathbb{R}^d} \left(\frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i)) + \mu \Omega(\theta) \right),$$

where $\ell : \mathcal{Y} \times \mathbb{R} \rightarrow \mathbb{R}^+$ is a loss function and $\Omega(\theta)$ is called a regularizer. The solution to this optimization problem is then called “*empirical risk minimizer*” and abbreviated ERM.

Note 1.2.1 *In order to choose Φ well-adapted to the problem, deep learning techniques can be used but that is out of the scope of this course and Φ is supposed to be known in the following classes. See, e.g., [1].*

There are two fundamental questions : how to compute $\hat{\theta}$ and how to analyze its statistical properties. These questions can be tackled separately but dealing with it simultaneously actually allows much better results in term of running-time complexity. To perform this task, the two following quantities will be considered :

- the empirical risk $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$ (computable with the only data).
- the expected risk $f(\theta) = \mathbb{E}_{(x,y)} [\ell(y_i, \theta^\top \Phi(x_i))]$ (which is practically unknown because we do not know the distribution the data are sampled from).

Our ERM as defined previously is often used in practice. There is an equivalent definition using Lagrange duality :

$$\min_{\theta \in \mathbb{R}^d} \left(\frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i)) \right) \text{ such that } \Omega(\theta) \leq D.$$

This definition is easier to analyze in theory. It can be used in practice as well but one needs to be really careful about the choice of D .

1.2.1 Usual losses

The most common loss in regression is the quadratic loss function. Let $y \in \mathbb{R}$ and $\hat{y} = \theta^\top \Phi(x)$ our prediction. The quadratic loss function is then defined as

$$\ell(y, \hat{y}) = \frac{1}{2} (y - \theta^\top \Phi(x))^2.$$

In classification we use the binary loss defined by $\ell(y, \hat{y}) = \mathbb{1}_{y\hat{y} < 0}$ which is the most *natural* loss (leading to the usual error rate).

Nevertheless, this loss function is neither differentiable nor continuous and hence there is no way to effectively optimize the empirical risk using it. The trick is to use instead convex loss functions known as a *convex surrogates* (see Figure 1.1). The three convex losses we will consider in this case are :

- the hinge loss : $\ell(y, \theta^\top \Phi(x)) = \max(\{1 - y\theta^\top \Phi(x), 0\})$.
- the logistic loss: $\ell(y, \theta^\top \Phi(x)) = \log(1 + \exp(-y\theta^\top \Phi(x)))$.
- the quadratic loss function (already mentioned).

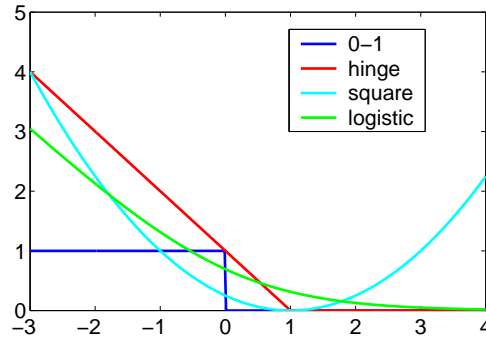


Figure 1.1. Convex surrogates for binary classification.

1.2.2 Usual regularizers

The goal of a regularizer is to avoid overfitting by penalizing values of the parameter θ which would account well for the learning sample but would show poor generalization performances. The most commonly used regularizers are :

- the euclidean norm $\|\theta\|_2^2 = \sum_{j=1}^d |\theta_j|^2$ which is convex and hence numerically feasible and is convenient because of the *representer theorem* related to it existing in RKHS-theory [3].
- the lasso ℓ_1 -norm $\|\theta\|_1 = \sum_{j=1}^d |\theta_j|$ which induces sparsity allowing not only variable selection but also model selection (see, e.g., [4] and references therein for structured situations).

When looking for theoretical properties of an optimization problem, it is often easier to use the constrained version of it but when concentrating on the computation of solutions to the problem, it is easier to consider the regularized version.

1.3 General assumptions

We then make the following additional assumption on the features :

$$\exists R > 0 / \forall x \in \mathbb{R}^d, \|\Phi(x)\|_2 \leq R.$$

We define the loss for a single observation:

$$\forall i \in \{1, \dots, n\}, f_i(\theta) = \ell(y_i, \theta^\top \Phi(x_i))$$

so that $\forall i \in \{1, \dots, n\}, f(\theta) = \mathbb{E}[f_i(\theta)]$.

We then assume that f_i, f and \hat{f} have the following properties :

- they are convex on \mathbb{R}^d .
- they are Lipschitz-continuous, smooth and/or strongly convex.

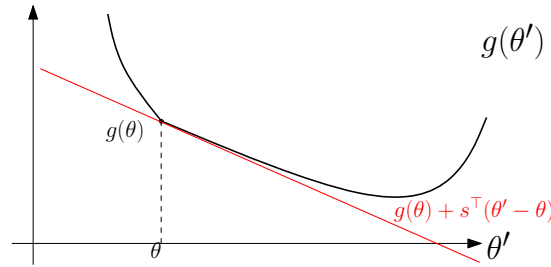


Figure 1.2. Subgradient.

1.3.1 Convexity

Definition 1.1 (Convexity) A function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex, if and only if :

- Without further assumption :

$$\forall (\theta_1, \theta_2) \in (\mathbb{R}^d)^2, \alpha \in [0, 1], \quad g(\alpha\theta_1 + (1 - \alpha)\theta_2) \leq \alpha g(\theta_1) + (1 - \alpha)g(\theta_2).$$

- Assuming differentiability :

$$\forall (\theta_1, \theta_2) \in (\mathbb{R}^d)^2, \quad g(\theta_1) \geq g(\theta_2) + g'(\theta_2)^\top (\theta_1 - \theta_2).$$

- If twice differentiable :

$$\forall \theta \in \mathbb{R}^d, \quad g''(\theta) \succcurlyeq 0 \text{ positive semi-definite Hessian.}$$

The main reasons why we use convex functions is that a local minimum is in fact a global minimum and that it allows to use convex duality to solve optimization problems.

1.3.2 Subgradients and subdifferentials

Definition 1.2 (Subgradients and subdifferentials) Let $g : \mathbb{R}^d \rightarrow \mathbb{R}$ be a convex function. $s \in \mathbb{R}^d$ is said to be a subgradient of g at point $\theta \in \mathbb{R}^d$ if and only if:

$$\forall \theta' \in \mathbb{R}^d, \quad g(\theta') \geq g(\theta) + s^\top (\theta' - \theta)$$

The set of all the subgradients of g at point $\theta \in \mathbb{R}^d$ written $\partial g(\theta)$ is called the subdifferential of at point θ . See Figure 1.2.

Example 1.3.1 The function $g : \mathbb{R} \rightarrow \mathbb{R}$ defined by $g(x) = |x| = \max(-x, x)$ is such that :

$$\forall \theta \in \mathbb{R}, \quad \partial g(\theta) = \begin{cases} \{-1\} & \text{if } \theta < 0. \\ [-1, +1] & \text{if } \theta = 0. \\ \{+1\} & \text{if } \theta > 0. \end{cases}$$

Example 1.3.2 With the hinge loss $h(u) = \max\{1 - u, 0\}$, we have $\partial h(1) = [-1, 0]$.

Theorem 1.3 If $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is a convex differentiable function at point $\theta \in \mathbb{R}^d$, then, $\partial g(\theta) = \{g'(\theta)\}$.

Theorem 1.4 For any convex function defined on \mathbb{R}^d , the differential is non-empty at all points θ .

1.3.3 Lipschitz continuity

Definition 1.5 (Lipschitz continuity) *If $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is a convex differentiable function, then, the two following properties are equivalent :*

$$\forall \theta \in \mathbb{R}^d, \|\theta\|_2 \leq D \Rightarrow \|g'(\theta)\|_2 \leq B$$

\Leftrightarrow

$$\forall (\theta, \theta') \in (\mathbb{R}^d)^2, \|\theta\|_2, \|\theta'\|_2 \leq D \Rightarrow |g(\theta) - g(\theta')| \leq B\|\theta - \theta'\|_2.$$

In this case, we say that g has gradients uniformly bounded by B on the ball of center 0 and radius D or equivalently that g is B -Lipschitz continuous on the ball of center 0 and radius D (with respect to $\|\cdot\|_2$).

Example 1.3.3 *Coming back to our ERM, let g be defined by $g(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$. If we assume that the loss is differentiable and G -Lipschitz and that the data are R -bounded, we easily compute the following upper bound $\forall (\theta, \theta') \in (\mathbb{R}^d)^2$:*

$$\begin{aligned} |f(\theta) - f(\theta')| &\leq \frac{1}{n} \sum_{i=1}^n |\ell(y_i, \theta^\top \Phi(x_i)) - \ell(y_i, \theta'^\top \Phi(x_i))| \\ &\leq \frac{1}{n} \sum_{i=1}^n G |\theta^\top \Phi(x_i) - \theta'^\top \Phi(x_i)| \\ &\leq \frac{1}{n} \sum_{i=1}^n RG \|\theta - \theta'\|_2 = RG \|\theta - \theta'\|_2. \end{aligned}$$

Thus we get that g is B -Lipschitz continuous with $B = GR$.

1.3.4 Smoothness

Definition 1.6 *A function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is L -smooth if and only if it is differentiable and its gradient is L -Lipschitz-continuous with respect to $\|\cdot\|_2$:*

$$\forall (\theta_1, \theta_2) \in (\mathbb{R}^d)^2, \|g'(\theta_1) - g'(\theta_2)\|_2 \leq L\|\theta_1 - \theta_2\|_2.$$

Theorem 1.7 *If $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is twice differentiable then g is L -smooth if and only if $\forall \theta \in \mathbb{R}^d, g''(\theta) \preceq L \cdot \text{Id}$.*

Example 1.3.4 *Coming back to our ERM with g defined by $g(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$ and assuming that the loss ℓ is twice differentiable L_{loss} -smooth and that the data are R -bounded, we can use the previous theorem.*

We compute the first and second derivative of g with respect to θ .

For the first derivative we get :

$$g'(\theta) = \frac{1}{n} \sum_{i=1}^n \ell'(y_i, \theta^\top \Phi(x_i)) \Phi(x_i).$$

Hence for the second derivative, where ℓ'' is the second-derivative with respect to the second variable:

$$g''(\theta) = \frac{1}{n} \sum_{i=1}^n \ell''(y_i, \theta^\top \Phi(x_i)) \Phi(x_i) \Phi(x_i)^\top.$$

Since the loss ℓ is L_{loss} -smooth so that $\ell'' \preceq L_{\text{loss}}$ and writing $\Sigma = \frac{1}{n} \sum_{i=1}^n \Phi(x_i) \Phi(x_i)^\top$ the uncentered empirical covariance matrix of the features, it comes :

$$g''(\theta) \preceq L_{\text{loss}} \frac{1}{n} \sum_{i=1}^n \Phi(x_i) \Phi(x_i)^\top = L_{\text{loss}} \Sigma.$$

Furthermore, writing $\lambda_{\max}(S)$ the largest eigenvalue of any symmetric matrix S and using that the data are R -bounded we finally get that :

$$\begin{aligned} \lambda_{\max}(L_{\text{loss}} \Sigma) &\leq \text{Tr} \left(L_{\text{loss}} \frac{1}{n} \sum_{i=1}^n \Phi(x_i) \Phi(x_i)^\top \right) \\ &= L_{\text{loss}} \frac{1}{n} \sum_{i=1}^n \text{Tr} (\Phi(x_i)^\top \Phi(x_i)) \\ &\leq L_{\text{loss}} \frac{1}{n} \sum_{i=1}^n R^2 = L_{\text{loss}} R^2. \end{aligned}$$

Thus, g is L -smooth with $L = L_{\text{loss}} R^2$.

1.3.5 Strong convexity

Definition 1.8 A function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is μ -strongly convex if and only if

$$\forall (\theta_1, \theta_2) \in (\mathbb{R}^d)^2, g(\theta_1) \geq g(\theta_2) + g'(\theta_2)^\top (\theta_1 - \theta_2) + \frac{\mu}{2} \|\theta_1 - \theta_2\|_2^2.$$

Example 1.3.5 Coming back to our ERM with g defined by $g(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \theta^\top \Phi(x_i))$ and assuming that the loss ℓ is twice differentiable μ_{loss} -strongly convex, we can use the previous theorem.


Reusing the calculations of the previous example and since the loss ℓ is μ_{loss} -strongly convex so that $\ell'' \succeq \mu_{\text{loss}} \cdot \text{Id}$ and writing again $\Sigma = \frac{1}{n} \sum_{i=1}^n \Phi(x_i) \Phi(x_i)^\top$ the uncentered empirical covariance matrix of the features, it comes :

$$g''(\theta) \succeq \mu \frac{1}{n} \sum_{i=1}^n \Phi(x_i) \Phi(x_i)^\top = \mu_{\text{loss}} \Sigma.$$

Furthermore, writing $\lambda_{\min}(S)$ the smallest eigenvalue of any symmetric matrix S and assuming that the uncentered empirical covariance matrix of the features is invertible ($\Leftrightarrow \lambda_{\min}(\Sigma) > 0$) we finally get that :

$$\lambda_{\min}(\mu_{\text{loss}}\Sigma) \geq \mu_{\text{loss}}\lambda_{\min}(\Sigma) > 0.$$

Thus, g is μ -strictly convex with $\mu = \mu_{\text{loss}}\lambda_{\min}(\Sigma)$.

 Note that strong-convexity is a strong assumption. In particular, in a machine learning context, this imposes that $n > d$. Moreover, in practice, even when n is much larger than d , μ is tiny.

1.4 Analysis of ERM

Approximation and estimation errors: $\Theta = \{\theta \in \mathbb{R}^d, \Omega(\theta) \leq D\}$

$$f(\hat{\theta}) - \min_{\theta \in \mathbb{R}^d} (f(\theta)) = \underbrace{\left[f(\hat{\theta}) - \min_{\theta \in \Theta} (f(\theta)) \right]}_{\text{estimation error}} + \underbrace{\left[\min_{\theta \in \Theta} (f(\theta)) - \min_{\theta \in \mathbb{R}^d} (f(\theta)) \right]}_{\text{approximation error}}$$

The estimation error is due to the fact that $\hat{\theta}$ has been computed using \hat{f} instead of f and the approximation error is due the fact that the minimization is made only on Θ and not on the whole parameter space \mathbb{R}^d .

Note 1.4.1 *Instead of comparing the performance of the estimator $\hat{\theta}$ with the best linear prediction $\min_{\theta \in \mathbb{R}^d} (f(\theta))$, it is also possible to compare it to the best (non-linear) prediction.*

1.4.1 Uniform deviation bounds

Let $\hat{\theta} \in \arg \min_{\theta \in \Theta} (\hat{f}(\theta))$ be the empirical (constrained) risk minimizer, and $\theta_{\Theta}^* \in \arg \min_{\theta \in \Theta} (f(\theta))$.

From the definitions of $\hat{\theta}$ and θ_{Θ}^* , we directly deduce that :

$$\hat{f}(\hat{\theta}) - \hat{f}(\theta_{\Theta}^*) \leq 0.$$

Using this inequality and rewriting the estimation error as :

$$f(\hat{\theta}) - \min_{\theta \in \Theta} (f(\theta)) = \left[f(\hat{\theta}) - \hat{f}(\hat{\theta}) \right] + \left[\hat{f}(\hat{\theta}) - \hat{f}(\theta_{\Theta}^*) \right] + \left[\hat{f}(\theta_{\Theta}^*) - f(\theta_{\Theta}^*) \right],$$

we get :

$$\begin{aligned} f(\hat{\theta}) - \min_{\theta \in \Theta} (f(\theta)) &\leq \sup_{\theta \in \Theta} \left[f(\theta) - \hat{f}(\theta) \right] + \sup_{\theta \in \Theta} \left[\hat{f}(\theta) - f(\theta) \right] \\ &\leq 2 \sup_{\theta \in \Theta} \left(|f(\theta) - \hat{f}(\theta)| \right). \end{aligned}$$

1.5 Slow rate for supervised learning

We will show what happens with a quadratic loss function. We remember that in this case $\ell(y, \theta^\top \Phi(x)) = \frac{1}{2}(y - \theta^\top \Phi(x))^2$. From that we get

$$\begin{aligned} \hat{f}(\theta) - f(\theta) &= \frac{1}{2}\theta^\top \left(\frac{1}{n} \sum_{i=1}^n \Phi(x_i)\Phi(x_i)^\top - \mathbb{E}[\Phi(X)\Phi(X)^\top] \right) \theta \\ &\quad - \theta^\top \left(\frac{1}{n} \sum_{i=1}^n y_i \Phi(x_i) - \mathbb{E}[Y\Phi(X)] \right) + \frac{1}{2} \left(\frac{1}{n} \sum_{i=1}^n y_i^2 - \mathbb{E}[Y^2] \right). \end{aligned}$$

Hence

$$\begin{aligned} \sup_{\|\theta\|_2 \leq D} (|f(\theta) - \hat{f}(\theta)|) &\leq \frac{D^2}{2} \left\| \frac{1}{n} \sum_{i=1}^n \Phi(x_i)\Phi(x_i)^\top - \mathbb{E}[\Phi(X)\Phi(X)^\top] \right\|_{\text{op}} \\ &\quad + D \left\| \frac{1}{n} \sum_{i=1}^n y_i \Phi(x_i) - \mathbb{E}[Y\Phi(X)] \right\|_2 + \frac{1}{2} \left| \frac{1}{n} \sum_{i=1}^n y_i^2 - \mathbb{E}[Y^2] \right|. \end{aligned}$$

$\sup_{\|\theta\|_2 \leq D} (|f(\theta) - \hat{f}(\theta)|) \leq O(1/\sqrt{n})$ with high probability from 3 concentration inequalities.

This particular case gives the impression that it should be possible to get such a rate in $O(1/\sqrt{n})$ for other type of losses than the quadratic loss... See [5] for more details.

Note that in this section, we do not require the loss to be convex.

Definition 1.9 (Rademacher complexity) *The Rademacher complexity of the class of functions $(X, Y) \mapsto \ell(Y, \theta^\top \Phi(X))$ is defined as :*

$$R_n = \mathbb{E} \left[\sup_{\theta \in \Theta} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i f_i(\theta) \right) \right].$$

The “Empirical” Rademacher average \hat{R}_n of the class of functions $(X, Y) \mapsto \ell(Y, \theta^\top \Phi(X))$ is defined as :

$$\hat{R}_n = \mathbb{E} \left[\sup_{\theta \in \Theta} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i f_i(\theta) \right) \middle| \mathcal{D} \right].$$

Theorem 1.10

$$\mathbb{E} \left[\sup_{\theta \in \Theta} (f(\theta) - \hat{f}(\theta)) \right] = \mathbb{E} \left[\sup_{\theta \in \Theta} (\hat{f}(\theta) - f(\theta)) \right] \leq 2R_n.$$

Proof Let $\mathcal{D}' = \{x'_1, y'_1, \dots, x'_n, y'_n\}$ an independent copy of the data $\mathcal{D} = \{x_1, y_1, \dots, x_n, y_n\}$, with corresponding loss functions $f'_i(\theta)$. Let $(\varepsilon_i)_{i \in \{1, \dots, n\}}$ be i.i.d random variables uniformly

distributed in $\{-1,1\}$ and independent of \mathcal{D} and \mathcal{D}' . It then comes :

$$\begin{aligned}
\mathbb{E} \left[\sup_{\theta \in \Theta} f(\theta) - \hat{f}(\theta) \right] &= \mathbb{E} \left[\sup_{\theta \in \Theta} \left(f(\theta) - \frac{1}{n} \sum_{i=1}^n f_i(\theta) \right) \right] \\
&= \mathbb{E} \left[\sup_{\theta \in \Theta} \left(\frac{1}{n} \sum_{i=1}^n \mathbb{E} [f'_i(\theta) - f_i(\theta) | \mathcal{D}] \right) \right] \\
&\leq \mathbb{E} \left[\mathbb{E} \left[\sup_{\theta \in \Theta} \left(\frac{1}{n} \sum_{i=1}^n (f'_i(\theta) - f_i(\theta)) \right) \middle| \mathcal{D} \right] \right] \\
&= \mathbb{E} \left[\sup_{\theta \in \Theta} \left(\frac{1}{n} \sum_{i=1}^n (f'_i(\theta) - f_i(\theta)) \right) \right] \\
&= \mathbb{E} \left[\sup_{\theta \in \Theta} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i (f'_i(\theta) - f_i(\theta)) \right) \right] \text{ by symmetry of the } \varepsilon_i \text{ law} \\
&\leq 2 \mathbb{E} \left[\sup_{\theta \in \Theta} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i f_i(\theta) \right) \right] = 2R_n.
\end{aligned}$$

■

And so it seems that assuming Lipschitz-continuity of the loss function only affords to get a slow rate for the bound on the estimation error.

Lemma 1.11 (Ledoux-Talagrand refined by Meir and Zhang [2])

$$\hat{R}_n \leq G \mathbb{E}_\varepsilon \left[\sup_{\|\theta\|_2 \leq D} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i \theta^\top \Phi(x_i) \right) \right].$$

Proof Given any $b, a_i : \Theta \rightarrow \mathbb{R}$ (no assumption) and $\varphi_i : \mathbb{R} \rightarrow \mathbb{R}$ any 1-Lipschitz-functions, $i = 1, \dots, n$

$$\mathbb{E}_\varepsilon \left[\sup_{\theta \in \Theta} \left(b(\theta) + \sum_{i=1}^n \varepsilon_i \varphi_i(a_i(\theta)) \right) \right] \leq \mathbb{E}_\varepsilon \left[\sup_{\theta \in \Theta} \left(b(\theta) + \sum_{i=1}^n \varepsilon_i a_i(\theta) \right) \right].$$

We proceed by induction on $n \in \mathbb{N}$.

The base case $n = 0$ is immediatly satisfied.

Assuming the result holds for $n \in \mathbb{N}$, it comes for $n + 1$:

$$\begin{aligned}
& \mathbb{E}_{\varepsilon_1, \dots, \varepsilon_{n+1}} \left[\sup_{\theta \in \Theta} \left(b(\theta) + \sum_{i=1}^{n+1} \varepsilon_i \varphi_i(a_i(\theta)) \right) \right] \\
&= \mathbb{E}_{\varepsilon_1, \dots, \varepsilon_n} \left[\sup_{\theta, \theta' \in \Theta} \left(\frac{b(\theta) + b(\theta')}{2} + \sum_{i=1}^n \varepsilon_i \frac{\varphi_i(a_i(\theta)) + \varphi_i(a_i(\theta'))}{2} + \frac{\varphi_{n+1}(a_{n+1}(\theta)) - \varphi_{n+1}(a_{n+1}(\theta'))}{2} \right) \right] \\
&= \mathbb{E}_{\varepsilon_1, \dots, \varepsilon_n} \left[\sup_{\theta, \theta' \in \Theta} \left(\frac{b(\theta) + b(\theta')}{2} + \sum_{i=1}^n \varepsilon_i \frac{\varphi_i(a_i(\theta)) + \varphi_i(a_i(\theta'))}{2} + \frac{|\varphi_{n+1}(a_{n+1}(\theta)) - \varphi_{n+1}(a_{n+1}(\theta'))|}{2} \right) \right] \\
&\leq \mathbb{E}_{\varepsilon_1, \dots, \varepsilon_n} \left[\sup_{\theta, \theta' \in \Theta} \left(\frac{b(\theta) + b(\theta')}{2} + \sum_{i=1}^n \varepsilon_i \frac{\varphi_i(a_i(\theta)) + \varphi_i(a_i(\theta'))}{2} + \frac{|a_{n+1}(\theta) - a_{n+1}(\theta')|}{2} \right) \right] \\
&= \mathbb{E}_{\varepsilon_1, \dots, \varepsilon_n} \mathbb{E}_{\varepsilon_{n+1}} \left[\sup_{\theta \in \Theta} \left(b(\theta) + \varepsilon_{n+1} a_{n+1}(\theta) + \sum_{i=1}^n \varepsilon_i \varphi_i(a_i(\theta)) \right) \right] \\
&\leq \mathbb{E}_{\varepsilon_1, \dots, \varepsilon_n, \varepsilon_{n+1}} \left[\sup_{\theta \in \Theta} \left(b(\theta) + \varepsilon_{n+1} a_{n+1}(\theta) + \sum_{i=1}^n \varepsilon_i a_i(\theta) \right) \right] \text{ by our induction hypothesis.}
\end{aligned}$$

By mathematical induction, the desired result is then true for all $n \in \mathbb{N}$. ■

Theorem 1.12 *If the regularizer is the Euclidean norm $\|\cdot\|_2$ and that the loss is G -Lipschitz so that f and \hat{f} are GR -Lipschitz on the set $\Theta = \{\|\theta\|_2 \leq D\}$ then for $0 < \delta < 1$, with probability greater than $1 - \delta$, we get :*

$$\sup_{\theta \in \Theta} (|\hat{f}(\theta) - f(\theta)|) \leq \frac{\ell_0 + GRD}{\sqrt{n}} \left(2 + \sqrt{2 \log \frac{1}{\delta}} \right).$$

Furthermore, it holds the following upper bound :

$$\mathbb{E} \left[\sup_{\theta \in \Theta} (|\hat{f}(\theta) - f(\theta)|) \right] \leq \frac{2GRD}{\sqrt{n}} \leq \frac{2(\ell_0 + GRD)}{\sqrt{n}}.$$

Proof Let's define $Z = \sup_{\theta \in \Theta} (|f(\theta) - \hat{f}(\theta)|)$.

By changing the pair (x_i, y_i) , relying on the G -Lipschitz continuity of the loss on the set $\Theta = \{\|\theta\|_2 \leq D\}$ and the R -boundedness of the features, Z may only change by :

$$\begin{aligned}
\frac{2}{n} \sup (|\ell(Y, \theta^\top \Phi(X))|) &= \frac{2}{n} \sup (|\ell(Y, 0)| + |\ell(Y, \theta^\top \Phi(X)) - \ell(Y, 0)|) \\
&\leq \frac{2}{n} (\sup (|\ell(Y, 0)| + G |\theta^\top \Phi(X)\theta|)) && \text{by } G\text{-Lipschitz continuity of } \ell \\
&\leq \frac{2}{n} (\sup (|\ell(Y, 0)| + GRD)) && \text{by Cauchy-Schwarz inequality} \\
&\leq \frac{2}{n} (\ell_0 + GRD) = c.
\end{aligned}$$

with $\sup |\ell(Y, 0)| = \ell_0$ so that Z is a Lipschitz-continuous function in the (x_i, y_i) .

Hence, the MacDiarmid inequality yields that with probability greater than $1 - \delta$, it holds :

$$Z \leq \mathbb{E}Z + c\sqrt{\frac{n}{2} \log\left(\frac{1}{\delta}\right)} \leq 2R_n + (\ell_0 + GRD) \sqrt{\frac{2}{n} \log\left(\frac{1}{\delta}\right)}.$$

Noticing that $R_n = \mathbb{E} \left[\hat{R}_n \right]$ and remembering Ledoux-Talagrand result given by the previous lemma, it then comes :

$$\begin{aligned} R_n &\leq G\mathbb{E} \left[\sup_{\|\theta\|_2 \leq D} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i \theta^\top \Phi(x_i) \right) \right] \\ &= G\mathbb{E} \left[\left\| D \frac{1}{n} \sum_{i=1}^n \varepsilon_i \Phi(x_i) \right\|_2 \right] \text{ by Cauchy-Schwarz inequality} \\ &\leq GD \sqrt{\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \Phi(x_i) \right\|_2^2 \right]} \text{ by Jensen's inequality} \\ &\leq \frac{GRD}{\sqrt{n}} \text{ by using } \|\Phi(x)\|_2 \leq R \text{ and independence of the } \varepsilon_i. \end{aligned}$$

Using that $\mathbb{E} \left[\sup_{\theta \in \Theta} (f(\theta) - \hat{f}(\theta)) \right] = \mathbb{E} \left[\sup_{\theta \in \Theta} (\hat{f}(\theta) - f(\theta)) \right] \leq 2R_n$, we get the expected estimation error :

$$\mathbb{E} \left[\sup_{\theta \in \Theta} (|\hat{f}(\theta) - f(\theta)|) \right] \leq \frac{2GRD}{\sqrt{n}} \leq \frac{2(\ell_0 + GRD)}{\sqrt{n}}.$$

Replacing the later in the MacDiarmid inequality, we have with probability greater than $1 - \delta$:

$$\begin{aligned} \sup_{\theta \in \Theta} (|f(\theta) - \hat{f}(\theta)|) &\leq \frac{1}{\sqrt{n}} \left(2GRD + (GRD + \ell_0) \sqrt{2 \log\left(\frac{1}{\delta}\right)} \right) \\ &\leq \frac{1}{\sqrt{n}} (\ell_0 + GRD) \left(2 + \sqrt{2 \log\left(\frac{1}{\delta}\right)} \right). \end{aligned}$$

■

Finally getting back to our uniform deviation bounds for the ERM, we get :

Theorem 1.13

$$\begin{aligned} f(\hat{\theta}) - \min_{\theta \in \Theta} (f(\theta)) &\leq 2 \sup_{\theta \in \Theta} (|\hat{f}(\theta) - f(\theta)|) \\ &\leq \frac{2}{\sqrt{n}} (\ell_0 + GRD) \left(2 + \sqrt{2 \log\left(\frac{1}{\delta}\right)} \right). \end{aligned}$$

So far, we have considered an exact minimizer $\hat{\theta} \in \arg \min_{\theta \in \Theta} (\hat{f}(\hat{\theta}))$ of \hat{f} but we can't always afford exact minimization and instead compute an inexact minimizer $\eta \in \Theta$ of \hat{f} . The decomposition of the excess risk yields then :

$$f(\eta) - \min_{\theta \in \Theta} (f(\theta)) \leq 2 \sup_{\theta \in \Theta} (|\hat{f}(\theta) - f(\theta)|) + [\hat{f}(\eta) - \hat{f}(\hat{\theta})].$$

And the upper bound on $\sup_{\theta \in \Theta} (|\hat{f}(\theta) - f(\theta)|)$ with high probability ensures that we only need to optimize with precision $\frac{2}{\sqrt{n}} (\ell_0 + GRD)$: this shows again that it's worth considering simultaneously statistical analysis and optimization of our estimator.

1.6 Fast rate for supervised learning

Motivation from mean estimation. We sometimes have better than slow rate convergence if we compute things differently.

Let's have a mean estimator $\hat{\theta}$ such that :

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n z_i = \arg \min_{\theta \in \mathbb{R}} \left(\frac{1}{2n} \sum_{i=1}^n (\theta - z_i)^2 \right) = \hat{f}(\theta).$$

If we use previous results we get something like :

- $f(\theta) = \frac{1}{2} \mathbb{E} [(\theta - z)^2] = \frac{1}{2} (\theta - \mathbb{E}[z])^2 + \frac{1}{2} \text{Var}(z) = \hat{f}(\theta) + O(1/\sqrt{n})$.
- $f(\hat{\theta}) = \frac{1}{2} (\hat{\theta} - \mathbb{E}[z])^2 + \frac{1}{2} \text{Var}(z) = f(\mathbb{E}[z]) + O(1/\sqrt{n})$.

But we can get way better bound using directly :

$$\begin{aligned} f(\hat{\theta}) - f(\mathbb{E}[z]) &= \frac{1}{2} (\hat{\theta} - \mathbb{E}[z])^2 \\ \mathbb{E} [f(\hat{\theta}) - f(\mathbb{E}[z])] &= \frac{1}{2} \mathbb{E} \left[\left(\frac{1}{n} \sum_{i=1}^n z_i - \mathbb{E}[z] \right)^2 \right] = \frac{1}{2n} \text{Var}(z). \end{aligned}$$

We get a bound only on $\hat{\theta}$ instead of a uniform bound.

General situation. Let f be the expected risk, \hat{f} the empirical risk. We supposed furthermore that :

- Features are bounded and the function loss is Lipschitz.
- f is convex.
- Regularized risks: $f^\mu(\theta) = f(\theta) + \frac{\mu}{2} \|\theta\|_2^2$ and $\hat{f}^\mu(\theta) = \hat{f}(\theta) + \frac{\mu}{2} \|\theta\|_2^2$.

As defined, $f^\mu(\theta)$ is μ -strongly convex.

Under these assumptions we get the following result :

Theorem 1.14 (Sridharan, Srebro, Shalev-Shwartz (2008)) For any $a > 0$, with probability greater than $1 - \delta$, for all $\theta \in \mathbb{R}^d$:

$$f^\mu(\hat{\theta}) - \min_{\eta \in \mathbb{R}^d} (f^\mu(\eta)) \leq \frac{8(1 + \frac{1}{a})G^2 R^2(32 + \log(\frac{1}{\delta}))}{\mu n}.$$

In contrast to Lipschitz-continuous functions, we observe that strongly convex functions give a fast rate convergence for supervised learning.

Though the μ term adds a bias to our error so it is important that μ decreases with n to reduce the approximation error.

Bibliography

- [1] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.
- [2] R. Meir and T. Zhang. Generalization error bounds for Bayesian mixture algorithms. *The Journal of Machine Learning Research*, 4, 839-860, 2003.
- [3] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
- [4] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity- inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012b.
- [5] S. Boucheron, O. Bousquet, G. Lugosi, et al. Theory of classification: A survey of some recent advances. *ESAIM Probability and statistics*, 9:323–375, 2005.