

Adaptivity in Machine Learning

Francis Bach

September 25, 2024

Based on book “Learning Theory from First Principles”, available at https://www.di.ens.fr/~fbach/lftp_book.pdf

Outline of the class:

- Lecture 1: How to get generalization bounds, the SGD way
- Lecture 2: Adaptivity of kernel methods to smoothness
- Lecture 3: Adaptivity of neural networks to linear latent variables

Remain as simple as possible. Can look at special topics chapter for deeper analysis.

1 Lecture 1: Simple generalization bounds with SGD (linear models)

1.1 Classical machine learning set up

- Observed data: $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$ i.i.d. from a given distribution
- Infinite amount of testing data from the same distribution
- Goal: estimate a prediction function $f : \mathcal{X} \rightarrow \mathcal{Y}$
- Loss function $\ell(y, z)$ (running example of least-squares)
- Expected risk: $\mathcal{R}(f) = \mathbb{E}[\ell(y, f(x))]$.  Randomness
- Empirical risk: $\widehat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i))$
- Bayes predictor and Bayes risk: minimizer

$$f_*(x) \in \arg \min_{z \in \mathcal{Y}} \mathbb{E}[\ell(y, z)|x]$$

and minimal value $\mathcal{R}_*$ of $\mathcal{R}$ over all functions from $\mathcal{X}$ to $\mathcal{Y}$. Goal of machine learning, achieve the Bayes risk

- Regression: $\mathcal{Y} = \mathbb{R}$, and the usual loss is $\ell(y, z) = (y - z)^2$, with $f_*(x) = \mathbb{E}[y|x]$. Absolute loss can also be considered.
- Classification: $\mathcal{Y} = \{-1, 1\}$, with $\ell(y, z) = 1_{y \neq z}$. Use of convex surrogates (with plot): square, logistic, hinge, each with its own interpretation, and optimal $f_*(x)$.
For logistic regression, $\ell(y, f(x)) = \log(1 + \exp(-yf(x)))$, with $f_*(x) = 2 \operatorname{atanh}(\mathbb{E}[y|x])$.
For hinge loss, $\ell(y, f(x)) = (1 - yf(x))_+$, with $f_*(x) = \operatorname{sign}(\mathbb{E}[y|x])$.
Calibration functions exist. Focus only on real-valued predictions. Many other examples (Chapter 13 on structured prediction)
- Two classical frameworks for learning methods: (1) local averaging (which simply replaces $p(y|x)$ by a local approximation based on data), and (2) empirical risk minimization.

1.2 Empirical risk minimization

- Consider a set $\mathcal{F}$ of functions / models from $\mathcal{X}$ to $\mathbb{R}$, typically $\mathcal{F} = \{f_\theta, \theta \in \Theta\}$
- Classical risk decomposition (estimation and approximation errors), for $f \in \mathcal{F}$:

$$\mathcal{R}(f) - \mathcal{R}_* = \left\{ \mathcal{R}(f) - \inf_{f' \in \mathcal{F}} \mathcal{R}(f') \right\} + \left\{ \inf_{f' \in \mathcal{F}} \mathcal{R}(f') - \mathcal{R}_* \right\}$$

 Randomness, dependence on number of observations, and “size” of $\mathcal{F}$

- Exact empirical risk minimizer $\hat{f} \in \arg \min_{f \in \mathcal{F}} \widehat{\mathcal{R}}(f)$

- Approximate empirical risk minimizer $\widehat{\mathcal{R}}(\hat{f}) \leq \min_{f \in \mathcal{F}} \widehat{\mathcal{R}}(f) + \varepsilon$ optimization error
⚠ optimization error may not always go to zero! Has to be part of the analysis
- Approximation error dealt with in next lecture
- Estimation error, with $f_{\mathcal{F}}^* \in \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}(f)$:

$$\begin{aligned} \mathcal{R}(\hat{f}) - \mathcal{R}(f_{\mathcal{F}}^*) &= \left\{ \mathcal{R}(\hat{f}) - \widehat{\mathcal{R}}(\hat{f}) \right\} + \left\{ \widehat{\mathcal{R}}(\hat{f}) - \widehat{\mathcal{R}}(f_{\mathcal{F}}^*) \right\} + \left\{ \widehat{\mathcal{R}}(f_{\mathcal{F}}^*) - \mathcal{R}(f_{\mathcal{F}}^*) \right\} \\ &\leq 2 \sup_{f \in \mathcal{F}} |\mathcal{R}(f) - \widehat{\mathcal{R}}(f)| + \varepsilon \end{aligned}$$

- Classical analysis: bound uniform deviations (statistics) and optimization errors (optimization) separately

1.3 Classical statistical analysis for estimation error

- Focus on G -Lipschitz-continuous loss functions (logistic, hinge, or quadratic once reduced to a compact set)
- Focus on “linear” predictors: $f_{\theta}(x) = \varphi(x)^{\top} \theta$, with $\|\varphi(x)\|_2 \leq R$ almost surely. Consider the upper-bound $\Theta = \{\theta, \|\theta\|_2 \leq D\}$. ⚠ Can be made more general, can be infinite-dimensional (see next lecture)
- Focus on bounds in expectation $\mathbb{E} \left[\sup_{f \in \mathcal{F}} |\mathcal{R}(f) - \widehat{\mathcal{R}}(f)| \right]$.
- Classical symmetrization result leading to Rademacher complexity:

$$\mathbb{E}_{\mathcal{D}} \left[\sup_{f \in \mathcal{F}} |\mathcal{R}(f) - \widehat{\mathcal{R}}(f)| \right] \leq 2 \cdot \mathbb{E}_{\mathcal{D}, \varepsilon} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \ell(y_i, f(x_i)) \right| \right]$$

- Contraction principle:

$$\mathbb{E}_{\mathcal{D}, \varepsilon} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \ell(y_i, f(x_i)) \right| \right] \leq 2G \cdot \mathbb{E}_{\mathcal{D}, \varepsilon} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right]$$

- Uniform deviations, with closed-form maximization:

$$\begin{aligned} \mathbb{E}_{\mathcal{D}} \left[\sup_{f \in \mathcal{F}} |\mathcal{R}(f) - \widehat{\mathcal{R}}(f)| \right] &\leq 4G \cdot \mathbb{E}_{\mathcal{D}, \varepsilon} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right] \\ &\leq 4G \cdot \mathbb{E}_{\mathcal{D}, \varepsilon} \left[\sup_{\|\theta\|_2 \leq D} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \varphi(x_i)^{\top} \theta \right| \right] \\ &\leq \frac{4GDR}{\sqrt{n}} \end{aligned}$$

⚠ No explicit dependence on dimension!

1.4 Subgradient method

- Given $F : \mathbb{R}^d \rightarrow \mathbb{R}$ convex, differentiable, B -Lipschitz-continuous (gradients bounded by B in ℓ_2 -norm),

$$\theta_k = \Pi_{\|\cdot\|_2 \leq D}(\theta_{k-1} - \gamma F'(\theta_{k-1})).$$

Constant step-size for simplicity.

- Lemma about convexity: $F(\theta') - F(\theta) \leq F'(\theta')^\top (\theta' - \theta)$
- For any θ such that $\|\theta\|_2 \leq D$, we have:

$$\begin{aligned} \|\theta_k - \theta\|_2^2 &\leq \|\theta_{k-1} - \gamma F'(\theta_{k-1}) - \theta\|_2^2 \\ &\leq \|\theta_{k-1} - \theta\|_2^2 - 2\gamma F'(\theta_{k-1})^\top (\theta_{k-1} - \theta) + \gamma^2 \|F'(\theta_{k-1})\|_2^2 \\ &\leq \|\theta_{k-1} - \theta\|_2^2 - 2\gamma [F(\theta_{k-1}) - F(\theta)] + \gamma^2 B^2 \end{aligned}$$

leading to

$$\begin{aligned} [F(\theta_{k-1}) - F(\theta)] &\leq \frac{1}{2\gamma} \|\theta_{k-1} - \theta\|_2^2 - \frac{1}{2\gamma} \|\theta_k - \theta\|_2^2 + \frac{1}{2\gamma} B^2 \\ F\left(\frac{1}{k} \sum_{i=0}^{k-1} \theta_i\right) - F(\theta) &\leq \frac{1}{2\gamma k} \|\theta_0 - \theta\|_2^2 + \frac{1}{2\gamma} B^2 \\ &\leq \frac{1}{2\gamma k} 4D^2 + \frac{1}{2\gamma} B^2 \\ &\leq \frac{2BD}{\sqrt{k}} \text{ with } \gamma = 2D/(B\sqrt{k}) \end{aligned}$$

- Application to machine learning, with $F(\theta) = \widehat{\mathcal{R}}(f_\theta)$, and $B = GR$, $k = n$ iterations: expected estimation error less than

$$\frac{4GDR}{\sqrt{n}} + \frac{2GDR}{\sqrt{n}} = \frac{6GDR}{\sqrt{n}}$$

but $O(n^2)$ calls to gradient of individual loss functions.

NB: can be done as well without the orthogonal projection.

Note the dependence in D of the estimation error.

1.5 Stochastic gradient descent

- Two classical set ups: single pass or multiple passes. Focus on single pass (can obtain the other as special case) where $F(\theta) = \mathcal{R}(f_\theta)$ is the *expected* risk.
- Assumptions: at time k , $\mathbb{E}[g_k | \mathcal{F}_{k-1}] = F'(\theta_{k-1})$, and $\|g_k\|_2^2 \leq B^2$ almost surely.
- Iteration: $\theta_k = \theta_{k-1} - \gamma g_k$

- Exact “same” proof with additional expectations leads to

$$\mathbb{E} \left[F \left(\frac{1}{n} \sum_{i=0}^{n-1} \theta_i \right) \right] - F(\theta) \leq \frac{6GDR}{\sqrt{n}}$$

with $O(n)$ accesses to local gradients.



Bound on expected risk!

- Classical extensions: strongly-convex, smoothness, variance reduction, mirror descent
- Other benefits: extend to multivariate outputs

2 Lecture 2: Adaptivity of kernel methods to smoothness

- Recall on loss functions, empirical risk, and expected risks. Model $f_\theta : \mathcal{X} \rightarrow \mathbb{R}$, $\theta \in \Theta$
- Decomposition between estimation and approximation errors:

$$\begin{aligned}\mathcal{R}(f_\theta) - \mathcal{R}_* &= \left\{ \mathcal{R}(\theta) - \inf_{\theta' \in \Theta} \mathcal{R}(f_{\theta'}) \right\} + \left\{ \inf_{\theta' \in \Theta} \mathcal{R}(f_{\theta'}) - \mathcal{R}_* \right\} \\ &= \text{estimation error} + \text{approximation error}\end{aligned}$$

- Summary of last lecture: For linear models $f_\theta(x) = \theta^\top \varphi(x)$, the estimation error after ERM or SGD on the ball of radius D is proportional to $\frac{GDR}{\sqrt{n}}$, when all features are bounded in ℓ_2 -norm by R , and a G -Lipschitz-continuous function.

⚠ No explicit dependence on dimension!

⚠ Linear in $D/\sqrt{n}$

- Goals of this lecture:
 - Show that infinite-dimensional Hilbert spaces are computationally feasible.
 - Deal with approximation error (requires assumption on f_* based on the existence and boundedness of s -th order derivatives).
 - Show (partial) adaptivity of kernel methods.

2.1 Kernel trick

- Now assume that $\varphi(x) \in \mathcal{H}$ Hilbert space, and consider f parameterized by $\theta \in \mathcal{H}$, as

$$f(x) = \langle \theta, \varphi(x) \rangle.$$

Define a space of function for which the function evaluations at a given x are bounded linear operators (this excludes spaces which are too big).

- Constrained ERM: $\min_{\|\theta\|_{\mathcal{H}} \leq D} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \theta, \varphi(x_i) \rangle).$

Representer theorem (proof by Pythagore argument): $\theta = \sum_{i=1}^n \alpha_i \varphi(x_i)$, and everything depends on the kernel function $k(x, x') = \langle \varphi(x), \varphi(x') \rangle$, since $f(x) = \langle \theta, \varphi(x) \rangle = \sum_{i=1}^n \alpha_i k(x, x_i)$, and $\|\theta\|_{\mathcal{H}}^2 = \sum_{i,j=1}^n \alpha_i \alpha_j K_{ij}$, where $K_{ij} = k(x_i, x_j)$.

Kernel trick: only need to know the kernel function and not the feature vector.

- SGD starting from $\theta_0 = 0$:

$$\theta_i = \theta_{i-1} - \gamma \ell'(y_i, \langle \theta_{i-1}, \varphi(x_i) \rangle) \varphi(x_i)$$

can be written as $\theta_i = \sum_{j=1}^i \alpha_j \varphi(x_j)$, with a new iteration

$$\alpha_i = -\gamma \ell' \left(y_i, \sum_{j=1}^{i-1} \alpha_j k(x_j, x_i) \right)$$

Complexity is $O(n^2)$ after n iterations but several methods exist to lower the cost (random features, column sampling).

2.2 Approximation / estimation trade-off for kernel methods

- Goal: optimize D (radius of ball). What is meant by adaptivity? With a single hyperparameter, can benefit from faster rates when available. Still needs some form of validation to find that hyperparameter.
- Estimation error proportional to $\frac{GRD}{\sqrt{n}}$ (as seen in last lecture for ERM or SGD)
- Approximation error, for Θ ball of radius D and center 0:

$$\begin{aligned} \inf_{\theta' \in \Theta} \mathcal{R}(f_{\theta'}) - \mathcal{R}_* &= \inf_{\theta' \in \Theta} \mathcal{R}(f_{\theta'}) - \mathcal{R}(f_*) \\ &= \inf_{\theta' \in \Theta} \mathbb{E} \left[\ell(y, f_{\theta'}(x)) - \ell(y, f_*(x)) \right] \leq G \inf_{\theta' \in \Theta} \mathbb{E} [|f_{\theta'}(x) - f_*(x)|] \\ &\leq G \inf_{\theta' \in \Theta} \left(\mathbb{E} [|f_{\theta'}(x) - f_*(x)|^2] \right)^{1/2} \\ &\leq \inf_{\|\theta'\|_{\mathcal{H}} \leq D} \|f_{\theta'} - f_*\|_{L_2(p)} \end{aligned}$$

- The excess risk can then be upper-bounded as (up to universal constants), with

$$\hat{f}_D \in \operatorname{argmin}_{\|\theta'\|_{\mathcal{H}} \leq D} \hat{\mathcal{R}}(f_{\theta'})$$

or by single pass SGD on the ball Θ :

$$\begin{aligned} \mathcal{R}(\hat{f}_D) - \mathcal{R}_* &\leq \frac{GRD}{\sqrt{n}} + \inf_{\|\theta\|_{\mathcal{H}} \leq D} \|f_{\theta} - f_*\|_{L_2(p)} \\ \inf_{D \geq 0} \mathcal{R}(\hat{f}_D) - \mathcal{R}_* &\leq \inf_{\theta \in \mathcal{H}} \|f_{\theta} - f_*\|_{L_2(p)} + \frac{GR}{\sqrt{n}} \|\theta\|_{\mathcal{H}} \\ &\leq \left(\inf_{\theta \in \mathcal{H}} \left\{ \|f_{\theta} - f_*\|_{L_2(p)}^2 + \frac{G^2 R^2}{n} \|\theta\|_{\mathcal{H}}^2 \right\} \right)^{1/2} \end{aligned}$$

- Goal: how to approximate

$$A(\lambda) = \inf_{\theta \in \mathcal{H}} \|f_{\theta} - f_*\|_{L_2(p)}^2 + \lambda \|\theta\|_{\mathcal{H}}^2$$

where $f_{\theta}(x) = \langle \theta, \varphi(x) \rangle$.

Given some (natural) assumptions on f_* , optimal excess risk proportional to $A(G^2 R^2 / n)^{1/2}$.

2.3 Kernels for non-parametric estimation in one dimension

- Simple possible set-up: $\mathcal{X} = [0, 1]$, and p uniform on $[0, 1]$.
- Using Fourier series expansions $f(x) = \sum_{m \in \mathbb{Z}} \hat{f}_m e^{2im\pi x}$, define the norm of the Hilbert space $\mathcal{H}$ as

$$\|f\|_{\mathcal{H}}^2 = \sum_{m \in \mathbb{Z}} \frac{1}{c_m} |(\hat{f})_m|^2,$$

with dot-product $\langle f, g \rangle = \sum_{m \in \mathbb{Z}} \frac{1}{c_m} (\hat{f})_m^* (\hat{g})_m$, for $c_m > 0$.

If $\frac{1}{c_m} \sim (1 + m^{2s})$, this is the Sobolev space of functions with square-integrable s -th derivative, with the constraint $s > 1/2$ (so that $\sum_{m \in \mathbb{Z}} c_m$ is finite)

- Explicit feature map and kernel: $\varphi_m(x) = c_m e^{2im\pi x}$, for $m \in \mathbb{Z}$, so that

$$\langle \varphi(x), \varphi(x') \rangle = \sum_{m \in \mathbb{Z}} c_m e^{2im\pi(x-x')} = k(x, x')$$

$$f(x) = \sum_{m \in \mathbb{Z}} \hat{f}_m e^{2im\pi x} = \sum_{m \in \mathbb{Z}} \frac{\hat{f}_m}{c_m} c_m e^{2im\pi x} = \langle f, \varphi(x) \rangle.$$

Note that kernel can be obtained in closed form by Fourier series summations for simple sequences (c_m) .

- Decomposition of optimal predictor: f_* can be expanded in Fourier series

$$f_*(x) = \sum_{m \in \mathbb{Z}} (\hat{f}_*)_m e^{2im\pi x}.$$

- This leads to

$$\begin{aligned} A(\lambda) &= \inf_{\theta \in \mathcal{H}} \|f_\theta - f_*\|_{L_2(p)}^2 + \lambda \|\theta\|_{\mathcal{H}}^2 \\ &= \inf_{\hat{\theta} \in \mathbb{C}^{\mathbb{Z}}} \sum_{m \in \mathbb{Z}} |\hat{\theta}_m - (\hat{f}_*)_m|^2 + \lambda \sum_{m \in \mathbb{Z}} \frac{1}{c_m} |\hat{\theta}_m|^2 \\ &= \inf_{\hat{\theta} \in \mathbb{C}^{\mathbb{Z}}} \sum_{m \in \mathbb{Z}} \left\{ |(\hat{f}_*)_m|^2 - 2\hat{\theta}_m^* (\hat{f}_*)_m + (1 + \lambda c_m^{-1}) |\hat{\theta}_m|^2 \right\} \end{aligned}$$

Minimizer characterized by $\theta_m(1 + \lambda c_m^{-1}) = (\hat{f}_*)_m$, leading to optimal value

$$\begin{aligned} A(\lambda) &\leq \sum_{m \in \mathbb{Z}} \left\{ |(\hat{f}_*)_m|^2 - \frac{|(\hat{f}_*)_m|^2}{1 + \lambda c_m^{-1}} \right\} \\ &= \sum_{m \in \mathbb{Z}} \frac{\lambda c_m^{-1} |(\hat{f}_*)_m|^2}{1 + \lambda c_m^{-1}}. \end{aligned}$$

- Assumption: $\sum_{m \in \mathbb{Z}} (1 + m^{2t}) |(\hat{f}_*)_m|^2$ finite for $t \geq 0$, that is, t -th derivative of f_* is square integrable. We get:

$$\begin{aligned} A(\lambda) &\leq \sum_{m \in \mathbb{Z}} \frac{\lambda c_m^{-1} |(\hat{f}_*)_m|^2}{1 + \lambda c_m^{-1}} = \sum_{m \in \mathbb{Z}} \frac{\lambda c_m^{-1} m^{-2t}}{1 + \lambda c_m^{-1}} m^{2t} |(\hat{f}_*)_m|^2 \\ &\leq \sup_{m \in \mathbb{Z}} \frac{\lambda(1 + m^{2t})^{-1}}{\lambda + c_m} \sum_{m \in \mathbb{Z}} (1 + m^{2t}) |(\hat{f}_*)_m|^2. \end{aligned}$$

Two cases:

- If $t \geq s$, then f_* is part of the function space we use for modelling (we have a well-specified model), and thus $A(\lambda) \leq \lambda \|f_*\|_{\mathcal{H}}^2$.
- If $t < s$,

$$\begin{aligned} A(\lambda) &\leq \sup_{m \in \mathbb{Z}} \frac{\lambda(1 + m^{2t})^{-1}}{\lambda + c_m} \sum_{m \in \mathbb{Z}} (1 + m^{2t}) |(\hat{f}_*)_m|^2 \\ &\leq \sup_{m \in \mathbb{Z}} \frac{\lambda(1 + m^{2t})^{-1}}{\lambda^{1-t/s} c_m^{t/s}} \sum_{m \in \mathbb{Z}} (1 + m^{2t}) |(\hat{f}_*)_m|^2 \\ &\leq O(\lambda^{t/s}) \sum_{m \in \mathbb{Z}} m^{2t} |(\hat{f}_*)_m|^2. \end{aligned}$$

Using lemma: $a + b \geq \frac{t}{s}a + (1 - \frac{t}{s})b \geq a^{t/s} b^{1-t/s}$.

- Thus, the excess risk is less than a constant times $n^{-1/2}$ if $t > s$ and $n^{-t/2s}$, for $t \in (1, s)$. That is, faster rates with more derivatives (i.e., t bigger).
- More precise results for least-squares (see book and references therein), in particular with the possibility to take s large and have a rate that does not degrade with s , and for which we get optimal behavior with respect to the model class.

2.4 Extensions beyond dimension one

- Translation invariant kernel on $\mathbb{R}^d$, $k(x, y) = q(x - y)$, with q having non-negative Fourier transform
- Convergence rates depend on decay of Fourier transform $\hat{q}(\omega)$.
- Abel kernel: $q(x) = \exp(-\|x\|_2)$, $\hat{q}(\omega) \propto \frac{1}{1 + \|\omega\|_2^2}$, corresponds to all s -th order derivatives being bounded with $s = d/2 + 1/2 > d/2$.
- Similar developments as for one dimension with rate $n^{-t/2s}$, but with now constraint that $s > d/2$. Similar adaptivity.

3 Lecture 3: Adaptivity of neural networks to latent variables