Adaptivity in Machine Learning
CUSO winter school

Francis Bach

February 4, 2026

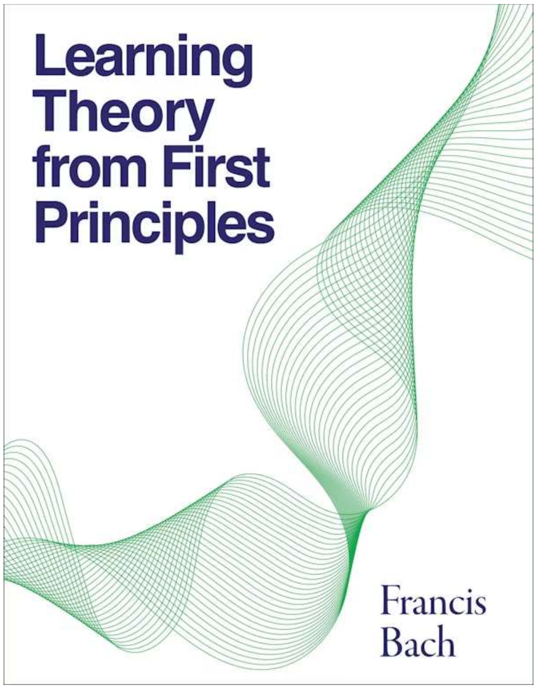
Notes updated everyday!

Based on book "Learning Theory from First Principles", available at `https://www.di.ens.fr/~fbach/ltfp_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)

Chapters 2, 4, and 5

1.1 Classical machine learning set up (Chapter 2)

- 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 $f : \mathcal{X} \to \mathcal{Y}$
- 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

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

  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 $\ell(y, z) = 1_{y=z}$. 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 $f_*(x) = 2\text{atanh}(\mathbb{E}[y|x])$.

  For hinge loss, $\ell(y, f(x)) = (1 - yf(x))_+$, with $f_*(x) = \text{sign}(\mathbb{E}[y|x])$.

  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 $f \in \mathcal{F}$:

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

  ⚠ 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 $\widehat{\mathcal{R}}(\hat{f}) \leqslant \min_{f\in\mathcal{F}} \widehat{\mathcal{R}}(f) + \varepsilon$ 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 $f^*_{\mathcal{F}} \in \operatorname{argmin}_{f\in\mathcal{F}} \mathcal{R}(f)$:

$$\begin{aligned}\mathcal{R}(\hat{f}) - \mathcal{R}(f^*_{\mathcal{F}}) &= \Big\{\mathcal{R}(\hat{f}) - \widehat{\mathcal{R}}(\hat{f})\Big\} + \Big\{\widehat{\mathcal{R}}(\hat{f}) - \widehat{\mathcal{R}}(f^*_{\mathcal{F}})\Big\} + \Big\{\widehat{\mathcal{R}}(f^*_{\mathcal{F}}) - \mathcal{R}(f^*_{\mathcal{F}})\Big\} \\ &\leqslant 2\sup_{f\in\mathcal{F}} \big|\mathcal{R}(f) - \widehat{\mathcal{R}}(f)\big| + \varepsilon\end{aligned}$$

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

1.3 Classical statistical analysis for estimation error (Chapter 4)

- 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 $\|\varphi(x)\|_2 \leqslant R$ almost surely. Consider the upper-bound $\Theta = \{\theta, \|\theta\|_2 \leqslant D\}$. ⚠ Can be made more general, can be infinite-dimensional (see next lecture)
- Focus on bounds in expectation $\mathbb{E}\Big[\sup_{f\in\mathcal{F}} \big|\mathcal{R}(f) - \widehat{\mathcal{R}}(f)\big|\Big]$.
- Classical symmetrization result leading to Rademacher complexity:

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

- Contraction principle:

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

- Uniform deviations, with closed-form maximization:

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

  ⚠ No explicit dependence on dimension!

1.4 Subgradient method (Chapter 5)

- Given $F : \mathbb{R}^d \to \mathbb{R}$ convex, differentiable, B-Lipschitz-continuous (gradients bounded by B in ℓ_2-norm),
$$\theta_k = \Pi_{\|\cdot\|_2 \leqslant D}\big(\theta_{k-1} - \gamma F'(\theta_{k-1})\big).$$
Constant step-size for simplicity.

- Lemma about convexity: $F(\theta') - F(\theta) \leqslant F'(\theta')^\top(\theta' - \theta)$

- For any θ such that $\|\theta\|_2 \leqslant D$, we have:
$$\begin{aligned}
\|\theta_k - \theta\|_2^2 &\leqslant \|\theta_{k-1} - \gamma F'(\theta_{k-1}) - \theta\|_2^2 \\
&\leqslant \|\theta_{k-1} - \theta\|_2^2 - 2\gamma F'(\theta_{k-1})^\top(\theta_{k-1} - \theta) + \gamma^2 \|F'(\theta_{k-1})\|_2^2 \\
&\leqslant \|\theta_{k-1} - \theta\|_2^2 - 2\gamma\big[F(\theta_{k-1}) - F(\theta)\big] + \gamma^2 B^2
\end{aligned}$$
leading to
$$\begin{aligned}
\big[F(\theta_{k-1}) - F(\theta)\big] &\leqslant \frac{1}{2\gamma}\|\theta_{k-1} - \theta\|_2^2 - \frac{1}{2\gamma}\|\theta_k - \theta\|_2^2 + \frac{1}{2\gamma}B^2 \\
F\Big(\frac{1}{k}\sum_{i=0}^{k-1}\theta_i\Big) - F(\theta) &\leqslant \frac{1}{2\gamma k}\|\theta_0 - \theta\|_2^2 + \frac{\gamma}{2}B^2 \\
&\leqslant \frac{1}{2\gamma k}4D^2 + \frac{\gamma}{2}B^2 \\
&\leqslant \frac{2BD}{\sqrt{k}} \text{ with } \gamma = 2D/(B\sqrt{k})
\end{aligned}$$

- 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 (Chapter 5)

- 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 $\|g_k\|_2^2 \leqslant B^2$ almost surely.

- Iteration: $\theta_k = \theta_{k-1} - \gamma g_k$

- Exact "same" proof with additional expectations leads to

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

  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 $f_\theta : \mathcal{X} \to \mathbb{R}$, $\theta \in \Theta$
- Decomposition between estimation and approximation errors:

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

- Summary of last lecture:
  (1) 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.
  (2) Single-pass constrained (on a ball of radius D) SGD, $\mathbb{E}[\mathcal{R}(f_{\hat{\theta}_n})] - \inf_{\|\theta\| \leqslant D} \mathcal{R}(f_\theta) \leqslant \frac{GRD}{\sqrt{n}}$.
  (3) Unconstrained single-pass SGD: for all $\theta \in \mathbb{R}^d$, $\mathbb{E}[\mathcal{R}(f_{\bar{\theta}_n})] \leqslant \mathcal{R}(f_\theta) + \frac{1}{2\gamma n}\|\theta - \theta_0\|^2 + \frac{\gamma G^2 R^2}{2}$ for $\bar{\theta}_n$ averaged iterate.

  ⚠ No explicit dependence on dimension!

  ⚠ Linear in $D/\sqrt{n}$

  ⚠ Many improvements (e.g., fast rate with strong convexity)

- 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

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

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

- Penalized ERM (or constrained ERM): $\displaystyle\min_{\theta \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^n \ell(y_i, \langle \theta, \varphi(x_i) \rangle) + \frac{\lambda}{2}\|\theta\|_{\mathcal{H}}^2$.

  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.

- Alternative "algorithmic" version. 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

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

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 for constrained optimization), λ (regularization parameter for penalized optimization), or γ (step-size of SGD).

  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.

Classical analysis for constrained optimization. (not done in lecture)

- 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:

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

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

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

or by single pass SGD on the ball Θ:

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

- Goal: how to approximate
$$A(\lambda) = \inf_{\theta \in \mathcal{H}} \quad \|f_\theta - f_*\|^2_{L_2(p)} + \lambda \|\theta\|^2_{\mathcal{H}}$$
where $f_\theta(x) = \langle \theta, \varphi(x) \rangle$.
Given some (natural) assumptions on f_*, optimal excess risk proportional to $A(G^2R^2/n)^{1/2}$.

SGD way (simpler). We have, after n iterations of SGD started at $\theta_0 = 0$, for any $\theta \in \mathcal{H}$,
$$\mathbb{E}[\mathcal{R}(f_{\bar{\theta}_n})] \leqslant \mathcal{R}(f_\theta) + \frac{1}{2\gamma n}\|\theta\|^2 + \frac{\gamma G^2 R^2}{2},$$
with, if ℓ is G-Lipschitz-continuous,
$$\begin{aligned}\mathcal{R}(f_\theta) - \mathcal{R}(f_*) &= \mathbb{E}[\ell(y, \langle \varphi(x), \theta \rangle) - \ell(y, f_*(x))] \\ &\leqslant G\mathbb{E}[|\langle \varphi(x), \theta \rangle) - f_*(x)|] \leqslant G\sqrt{\mathbb{E}[|f_\theta - f_*|^2]} = G\|f_\theta - f_*\|_{L_2(p)} \\ &\leqslant \frac{K}{2}\|f_\theta - f_*\|^2_{L_2(p)} + \frac{G^2}{2K}\end{aligned}$$
for any K (which will be minimized later). We then get, with $K = \frac{1}{\gamma R^2}$,
$$\begin{aligned}\mathbb{E}[\mathcal{R}(f_{\bar{\theta}_n})] &\leqslant \mathcal{R}(f_*) + \frac{K}{2}\|f_\theta - f_*\|^2_{L_2(p)} + \frac{1}{2\gamma n}\|\theta\|^2 + \frac{G^2}{2K} + \frac{\gamma G^2 R^2}{2} \\ &= \mathcal{R}(f_*) + \frac{1}{2\gamma R^2}\|f_\theta - f_*\|^2_{L_2(p)} + \frac{1}{2\gamma n}\|\theta\|^2 + \gamma G^2 R^2.\end{aligned}$$
If
$$A(\lambda) = \inf_{\theta \in \mathcal{H}} \left\{ \|f_\theta - f_*\|^2_{L_2(p)} + \lambda\|\theta\|^2 \right\} \leqslant \rho \lambda^\kappa,$$
we get the two bounds
$$\begin{aligned}\mathbb{E}[\mathcal{R}(f_{\bar{\theta}_n})] - \mathcal{R}(f_*) &\leqslant \frac{1}{2\gamma R^2} A(R^2/n) + \gamma G^2 R^2 \leqslant \frac{1}{2\gamma R^2}\rho(R^2/n)^\kappa + \gamma G^2 R^2 \\ &\leqslant \sqrt{2\rho} G (R^2/n)^{\kappa/2}.\end{aligned}$$
NB: same bound can be obtained with constrained empirical risk optimization and optimizing over D.

⚠ f_* is not in general of the form $f_*(x) = \langle \varphi(x), \theta_* \rangle$.

If well specified, then $A(\lambda) \leqslant \lambda \|\theta_*\|^2$, and the excess risk is proportional to $1/\sqrt{n}$.

Alternative not done in lecture. Alternatively, we can obtain a better bound if ℓ is H-smooth, that is, second derivatives are bounded, then, by optimality $\mathbb{E}[\ell'(y, f_*(x))h(x)] = 0$ for any function h, and
$$\begin{aligned}\mathcal{R}(f_\theta) - \mathcal{R}(f_*) &= \mathbb{E}[\ell(y, \langle \varphi(x), \theta \rangle) - \ell(y, f_*(x))] \\ &\leqslant \mathbb{E}[\ell'(y, f_*(x))(\langle \varphi(x), \theta \rangle) - f_*(x)) + \frac{H}{2}|\langle \varphi(x), \theta \rangle) - f_*(x)|^2] \\ &= \frac{H}{2}\mathbb{E}[|\langle \varphi(x), \theta \rangle) - f_*(x)|^2] = \frac{H}{2}\|f_\theta - f_*\|^2_{L_2(p)}.\end{aligned}$$
Thus, we get:
$$\mathbb{E}[\mathcal{R}(f_{\bar{\theta}_n})] \leqslant \mathcal{R}(f_*) + \inf_{\theta \in \mathcal{H}} \left\{ \frac{H}{2}\|f_\theta - f_*\|^2_{L_2(p)} + \frac{1}{2\gamma n}\|\theta\|^2 \right\} + \frac{\gamma G^2 R^2}{2}.$$

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). For example, for Sobolev cases, the kernel is a Bernoulli polynomial.

- 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

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

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

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

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

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

  Two cases:

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

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

  Using lemma (based on Jensen's inequality): $a + b \geqslant \frac{t}{s}a + (1-\frac{t}{s})b \geqslant a^{t/s} b^{1-t/s}$.

- Thus, the excess risk is less than a constant times $n^{-1/2}$ if $t > s$ and $n^{-t/2s}$, for $t \in (1, s)$. Two facts: (1) misspecified leads to slower rates, (2) 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

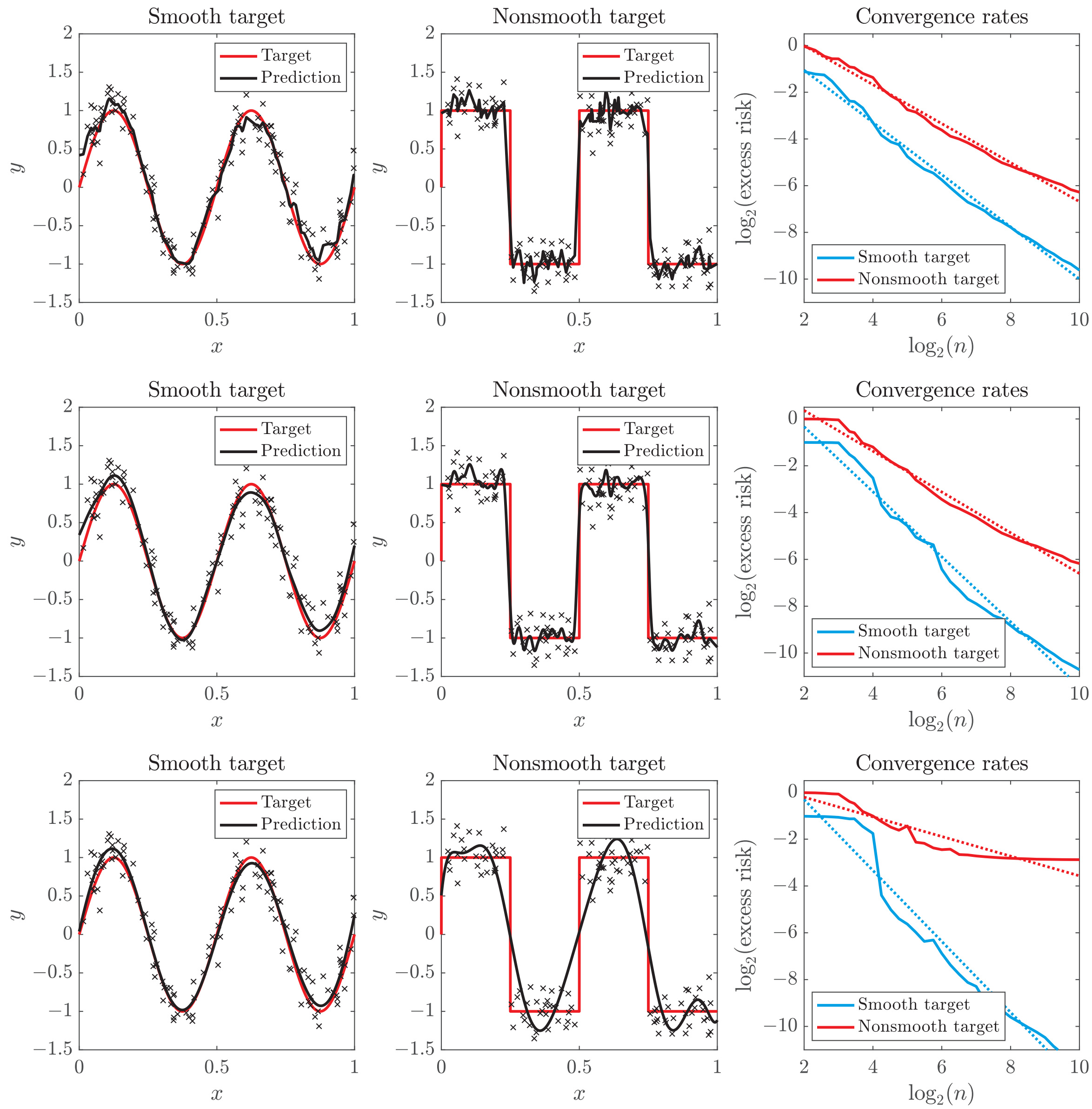
- Convergence rates depend on decay of Fourier transform $\hat{q}(\omega)$.

- 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.

- ⚠ If $t = 1$ (only first-order partial derivatives), then the rate of estimation for the Abel kernel is $n^{-1/(d+1)}$, which is very slow, but unavoidable. See next lecture for adapativity for linear substructures.

Comparison of three kernels: Sobolev space of order 1 (top), Matern kernel corresponding to the Sobolev space of order 3 (middle), and Gaussian kernel (bottom). We consider two different target functions and



3 Lecture 3: Adaptivity of neural networks to latent variables

- This lecture: focus on one-hidden layer neural networks. Simplest model, on which some results may be obtained.
- Prediction function:
$$f(x) = \sum_{j=1}^{m} \eta_j \sigma(w_j^\top x + b_j),$$
with the activation function being often the rectified linear unit, that is, $\sigma(u) = \max\{u, 0\} = (u)_+$ (key property = homogeneity).
- Model $f_\theta(x)$ which is *nonlinear* in $\theta = \{(w_j), (b_j), (\eta_j)\} \in \mathbb{R}^{m(d+2)}$.
- Three types of error: Estimation error (uniform deviations + optimization) + Approximation error

3.1 Optimization

- Gradient descent or SGD only reaches a stationary point. No global minima
- Chizat & Bach (2018): for overparameterized networks (that is, m goes to infinity), then the stationary points of the gradient flow have to be global minimizers. Proof based on a mean field limit since neurons decouple
$$f(x) = \frac{1}{m} \sum_{j=1}^{m} \eta_j \sigma(w_j^\top x + b_j) = \frac{1}{m} \sum_{j=1}^{m} \Phi(\theta_j) = \int \Phi(\theta) d\mu(\theta).$$
We have
$$\mathcal{R}(f) = \mathcal{R}\Big(\int \Phi(\theta) d\mu(\theta) \Big)$$
with $\mathcal{R}$ convex. Gradient flow on $\Theta = \{w_j\}_{j=1,\dots,m}$ with $\theta_j = (w_j, b_j, \eta_j)$ can be interpreted as a Wasserstein gradient flow, with bad stationary points, but global convergence when initialized with a measure with full mass.

3.2 Uniform deviations

- Assume $\|x\|_2 \leqslant R = 1$ almost surely. Constrained the norm $\|w_j\|^2 + b_j^2$ to be less than one (by renormalization, since ReLU is homogeneous), and add a constraint $\|\eta\|_1 \leqslant D$, which corresponds to an ℓ_2-constraint on $\theta = \{(w_j), (b_j), (\eta_j)\} \in \mathbb{R}^{m(d+2)}$ after optimizing over scale, since $\alpha_j^{-1} \eta_j (\alpha_j w_j^\top x + \alpha_j b_j)_+ = \eta_j (w_j^\top x + b_j)_+$, and
$$\inf_{\alpha_j > 0} \alpha_j^{-2} \eta_j^2 + \alpha_j^2 \|w_j\|_2^2 + \alpha_j^2 b_j^2 = 2|\eta_j| \sqrt{\|w_j\|^2 + b_j^2}.$$
- Uniform deviations (no proof on board):
$$\sup_{\theta \in \Theta} |\mathcal{R}(f_\theta) - \widehat{\mathcal{R}}(f_\theta)| \leqslant \text{ universal constant } \times \frac{GRD}{\sqrt{n}}$$

- Proof using Rademacher complexity (see book) and contraction principles
- Independent of the number of neurons m
- ⚠ The number of parameters is not what counts! The norm $\|\eta\|_1$ matters.

3.3 Approximation error - infinitely many neurons

- Assume all neurons $(w_j, b_j) \in K$, with $K = \{(w, b),\ \max\{\|w\|_2, b\} \leqslant 1\}$.
- Variation norm:

$$\|f\|^2 = \inf \int_K |d\eta(w,b)| \text{ such that } \|x\|_2 \leqslant R \Rightarrow f(x) = \int_K (w^\top x + b)_+ d\eta(w,b)$$

  Mean field limit of $\|\eta\|_1$ when considering $d\eta(w,b) = \sum_{j=1}^{m} \eta_j \delta_{(w_j,b_j)}$.

  ⚠ Key property of the total variation: if the function depends on a projection, we can do the expansion in low dimension, that is, if $f(x) = g(a^\top x)$, with $g : \mathbb{R} \to \mathbb{R}$ and $\|a\| = 1$, then $\|f\|^2 \leqslant \|g\|^2$.

- Approximation of one-dimensional linear functions: proof by image
- Approximation of one-dimensional function using Taylor series with integral remainder, assuming for simplicity that f is twice differentiable and $R = 1$:

$$f(x) = f(-1) + (x+1)f'(-1) + \int_{-1}^{x} (x-b) f''(b) dt = f(-1) + (x+1)f'(-1) + \int_{-1}^{1} (x-b)_+ f''(b) db$$

  If constants and linear functions can be approximated by two ReLUs, we get a bound on $\|f\|$ based on the L_1-norm of f'' and values of $f(1)$ and $f'(-1)$, and because of Poincaré inequality, $\|f\| \leqslant \square \int_{-1}^{1} |f(t)|dt + \square \int_{-1}^{1} |f''(t)|dt$, and thus it is less than the Sobolev-2 norm defined as $\sqrt{\square \int_{-1}^{1} |f(t)|^2 dt + \square \int_{-1}^{1} |f''(t)|^2 dt}$.

  Thus, all approximation theorems for Sobolev spaces extend to one-dimension neural networks.

- Extensions to more than one dimension using Fourier transform. Give only formula.

$$f(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{f}(\omega) e^{i\omega^\top x} d\omega,$$

  and decompose $x \mapsto e^{i\omega^\top x}$ with norm proportional to $1 + \|\omega\|^2$, leading to an upper bound

$$\|f\| \leqslant \square \int_{\mathbb{R}^d} \|\hat{f}(\omega)\|(1 + \|\omega\|_2^2) d\omega,$$

  often called Barron norms. They can also be upper bounded by Sobolev norms of order $\frac{d}{2} + \frac{3}{2}$. This implies that approximation theorems for Sobolev spaces extend, ⚠ for better or for worse (adaptivity to smoothness but still pay the curse of dimensionality).

- **Adaptivity to linear structures**. Simply does it when $f_*(x) = g(a^\top x)$ and escape the dependence in d in the rate, not necessarily in the constants.