Chapter 4

Empirical risk minimization

Chapter summary

- Convexification of the risk: for binary classification, optimal predictions can be achieved with convex surrogates.
- Risk decomposition: the risk can be decomposed into the sum of the approximation and estimation errors.
- Rademacher complexity: To study estimation errors and compute expected uniform deviations of real-valued outputs, Rademacher complexities are a very flexible and powerful tool that allows obtaining dimension-independent concentration inequalities.
- Relationship with asymptotic statistics: classical asymptotic results provide a finer picture of the behavior of empirical risk minimization as they provide asymptotic limits of performance as a well-defined constant times $1/n$, but they may not, in general, characterize small-sample effects.

As outlined in Chapter 2, given a joint distribution p on $\mathcal{X} \times \mathcal{Y}$, and n independent and identically distributed observations from p, our goal is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ with minimum risk $\mathcal{R}(f) = \mathbb{E}[\ell(y, f(x))]$, or equivalently minimum expected excess risk:

$$\mathcal{R}(f) - \mathcal{R}^* = \mathcal{R}(f) - \inf_{g \text{ measurable}} \mathcal{R}(g).$$

In this chapter, we will consider methods based on empirical risk minimization. Before looking at the necessary probabilistic tools, we will first show how problems where the output space is not a vector space, such as binary classification with $\mathcal{Y} = \{-1, 1\}$, can be reformulated as real-valued outputs, with so-called "convex surrogates" of loss functions.

4.1 Convexification of the risk

In this section, for simplicity, we focus on binary classification where $\mathcal{Y} = \{-1, 1\}$ with the 0-1 loss, but many of the concepts extend to the more general structured prediction set-up (see Chapter 13).

As our goal is to estimate a binary-valued function, the first idea that comes into mind is to minimize the empirical risk over a hypothesis space of binary-valued functions f (or equivalently, the subsets of $\mathcal{X}$ by considering the set $\{x \in \mathcal{X}, f(x) = 1\}$). However, this approach leads to a combinatorial problem that can be computationally intractable. Moreover, how to control the capacity (i.e., how to regularize) for these types of hypothesis spaces needs to be clarified. Learning a real-valued function instead through the framework of convex surrogates simplifies and overcomes this problem as it convexifies the problem. Classical penalty-based regularization techniques can then be used for theoretical analysis (this chapter) and algorithms (Chapter 5).

This choice of treating classification problems through real-valued prediction functions allows us to avoid introducing Vapnik-Chervonenkis dimensions (see Vapnik and Chervonenkis, 2015) to obtain general convergence results for empirical risk minimization in this chapter, where we will use the generic tool of Rademacher complexities in Section 4.5.

Instead of learning $f : \mathcal{X} \to \{-1, 1\}$, we will thus learn a function $g : \mathcal{X} \to \mathbb{R}$ and define $f(x) = \text{sign}(g(x))$ where

$$\text{sign}(a) = \begin{cases} 1 & \text{if } a > 0 \\ 0 & \text{if } a = 0 \\ -1 & \text{if } a < 0. \end{cases}$$

The convention $\text{sign}(0) = 0$ implies that the prediction can never be correct when $g(x) = 0$. Within our context, for maximally ambiguous observations, this corresponds to choosing none of the two labels (other conventions consider taking $+1$ or -1 uniformly at random).

The risk of the function $f = \text{sign} \circ g$, still denoted $\mathcal{R}(g)$ (⚠ note the slight overloading of notations $\mathcal{R}(g) = \mathcal{R}(\text{sign} \circ g)$), is then equal to:

$$\mathcal{R}(g) = \mathbb{P}(\text{sign}(g(x)) \neq y) = \mathbb{E}[1_{\text{sign}(g(x))\neq y}] = \mathbb{E}[1_{yg(x)\leqslant 0}] = \mathbb{E}\big[\Phi_{0-1}(yg(x))\big],$$

where $\Phi_{0-1} : \mathbb{R} \to \mathbb{R}$, with $\Phi_{0-1}(u) = 1_{u \leqslant 0}$ is called the "margin-based" 0-1 loss function or simply the 0-1 loss function.

⚠ Note the slightly overloaded notation above where the 0-1 loss function is defined on $\mathbb{R}$, compared to the 0-1 loss function from Chapter 2, which is defined on $\{-1, 1\} \times \{-1, 1\}$.

In practice, for empirical risk minimization, we then minimize with respect to the function $g : \mathcal{X} \to \mathbb{R}$ the corresponding empirical risk $\frac{1}{n}\sum_{i=1}^{n} \Phi_{0-1}(y_i g(x_i))$. The function Φ_{0-1} is not continuous (and thus also non-convex) and leads to difficult optimization problems.

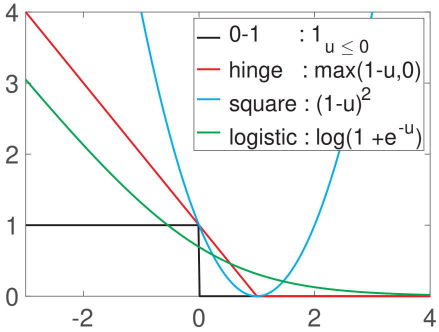


Figure 4.1: Classical convex surrogates for binary classification with the 0-1 loss.

4.1.1 Convex surrogates

A key concept in machine learning is the use of *convex surrogates*, where we replace Φ_{0-1} by another function Φ with better numerical properties (all will be convex). See classic examples below, plotted in Figure 4.1.

Instead of minimizing the classical risk $\mathcal{R}(g)$ or its empirical version, one then minimizes the "Φ-risk" (and its empirical version) defined as

$$\mathcal{R}_\Phi(g) = \mathbb{E}[\Phi(yg(x))].$$

In this context, the function g is sometimes called the *score function*.

The critical question we tackle in this section is: does it make sense to convexify the problem? In other words, does it lead to good predictions for the 0-1 loss?

Classical examples. We first review the primary examples used in practice:

- **Quadratic / square loss:** $\Phi(u) = (u-1)^2$, leading to, since we have $y^2 = 1$, $\Phi(yg(x)) = (y - g(x))^2 = (g(x) - y)^2$. We get back least-squares, ignore that the labels have to belong to $\{-1, 1\}$, and take the sign of $g(x)$ for the prediction. Note the overpenalization for a large positive value of $yg(x)$ that will not be present for the other losses below (which are non-increasing).
- **Logistic loss:** $\Phi(u) = \log(1 + e^{-u})$, leading to

  $$\Phi(yg(x)) = \log(1 + e^{-yg(x)}) = -\log\Big(\frac{1}{1 + e^{-yg(x)}}\Big) = -\log(\sigma(yg(x))),$$

  where: $\sigma(v) = \frac{1}{1+e^{-v}}$ is the sigmoid function. Note the link with maximum likelihood estimation, where we define the model through

  $$\mathbb{P}(y = 1|x) = \sigma(g(x)) \text{ and } \mathbb{P}(y = -1|x) = \sigma(-g(x)) = 1 - \sigma(g(x)).$$

  The risk is then the negative conditional log-likelihood $\mathbb{E}[-\log p(y|x)]$. It is also often called the cross-entropy loss.[1] See more details about probabilistic methods in Chapter 14.

[1] See https://en.wikipedia.org/wiki/Logistic_regression for details.

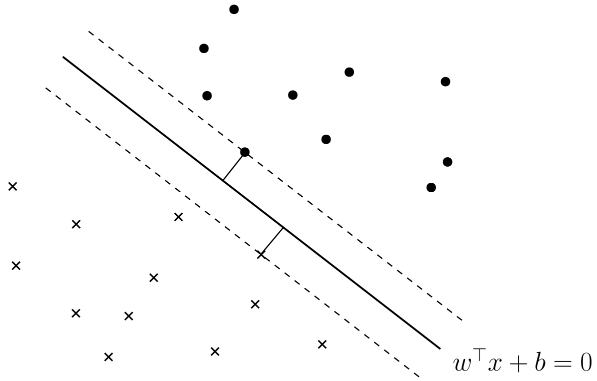
- **Hinge loss:** $\Phi(u) = \max(1-u, 0)$. With linear predictors, this leads to the support vector machine, and $yg(x)$ is often called the "margin" in this context. This loss has a geometric interpretation (see Section 4.1.2 below).[2]
- **Squared hinge loss:** $\Phi(u) = \max(1-u, 0)^2$. This is a smooth counterpart to the regular hinge loss.
- **Exponential loss:** $\Phi(u) = \exp(-u)$. This loss is often used within the boosting framework presented in Section 10.3, in particular through the Adaboost algorithm (Section 10.3.4).

This section analyzes precisely how replacing the 0-1 loss with convex surrogates still leads to optimal predictions. This allows us to only focus on *real-valued prediction functions* in the rest of this book. We will consider loss function $\ell(y, f(x))$ that will be the square loss $(y - f(x))^2$ for regression, and any of the ones above for binary classification, that is, $\Phi(yf(x))$. We will consider alternatives and extensions in Chapter 13.

4.1.2 Geometric interpretation of the support vector machine (♦)

Given its historical importance, this section provides a geometrical perspective on the hinge loss to highlight why it leads to a learning architecture called the "support vector machine" (SVM). We consider n observations $(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}$, for $i = 1, \dots, n$.

Separable data (Vapnik and Chervonenkis, 1964). We first assume that the data are separable by an affine hyperplane, that is, there exist $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that for all $i \in \{1, \dots, n\}$, $y_i(w^\top x_i + b) > 0$. Among the infinitely many separating hyperplanes, we aim to select the one for which the closest points from the dataset are the farthest.



The distance from x_i to the hyperplane $\{x \in \mathbb{R}^d,\ w^\top x + b = 0\}$ is equal to $\frac{|w^\top x_i + b|}{\|w\|_2}$, and thus, this minimal distance is

$$\min_{i \in \{1,\dots,n\}} \frac{y_i(w^\top x_i + b)}{\|w\|_2},$$

[2] See also https://en.wikipedia.org/wiki/Support_vector_machine for details.

and we thus aim at maximizing this quantity. Because of the invariance by rescaling (that is, we can multiply w and b by the same scalar constant without modifying the affine separator), this problem is equivalent to minimizing $\|w\|_2$ with the constraint that $\min_{i\in\{1,\dots,n\}} y_i(w^\top x_i + b) \geqslant 1$, and thus to the following problem:

$$\min_{w\in\mathbb{R}^d,\ b\in\mathbb{R}} \frac{1}{2}\|w\|_2^2 \ \text{ such that } \ \forall i \in \{1,\dots,n\},\ y_i(w^\top x_i + b) \geqslant 1. \tag{4.1}$$

General data (Cortes and Vapnik, 1995). When a hyperplane may not separate data, then we can introduce so-called "slack variables" $\xi_i \geqslant 0$, $i = 1,\dots,n$, allowing the constraint $y_i(w^\top x_i + b) \geqslant 1$ to be violated, by introducing the modified constraint $y_i(w^\top x_i + b) \geqslant 1 - \xi_i$ instead. The overall amount of slack is then minimized, leading to the following problem (with $C > 0$):

$$\min_{w\in\mathbb{R}^d,\ b\in\mathbb{R},\ \xi\in\mathbb{R}^n} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^n \xi_i \ \text{ such that } \ \forall i \in \{1,\dots,n\},\ y_i(w^\top x_i + b) \geqslant 1 - \xi_i,\ \xi_i \geqslant 0. \tag{4.2}$$

With $\lambda = \frac{1}{nC}$, the problem above is equivalent to

$$\min_{w\in\mathbb{R}^d,\ b\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n (1 - y_i(w^\top x_i + b))_+ + \frac{\lambda}{2}\|w\|_2^2,$$

which is exactly an ℓ_2-regularized empirical risk minimization with the hinge loss for the prediction function $f(x) = w^\top x + b$.

Lagrange dual and "support vectors" (♦). The problem in Eq. (4.2) is a linearly constrained convex optimization problem and can be analyzed using Lagrangian duality (see, e.g., Boyd and Vandenberghe, 2004). We consider non-negative Lagrange multipliers α_i and β_i, $i \in \{1,\dots,n\}$, and the following Lagrangian:

$$\mathcal{L}(w,b,\xi,\alpha,\beta) \ = \ \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i\big(y_i(w^\top x_i + b) - 1 + \xi_i\big) - \sum_{i=1}^n \beta_i\xi_i.$$

Minimizing with respect to $\xi \in \mathbb{R}^n$ leads to the equality constraints that for all $i \in \{1,\dots,n\}$, $\alpha_i + \beta_i = C$, while minimizing with respect to b leads to the constraint $\sum_{i=1}^n y_i\alpha_i = 0$. Finally, minimizing with respect to w can be done in closed form as $w = \sum_{i=1}^n \alpha_i y_i x_i$. Overall, this leads to the dual optimization problem:

$$\max_{\alpha\in\mathbb{R}^n} \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y_i y_j x_i^\top x_j \ \text{ such that } \ \sum_{i=1}^n y_i\alpha_i = 0 \text{ and } \forall i \in \{1,\dots,n\},\ \alpha_i \in [0,C]. \tag{4.3}$$

As we will show in Chapter 7 for all ℓ_2-regularized learning problems with linear predictors, the optimization problem only depends on the dot-products $x_i^\top x_j$, $i,j = 1,\dots,n$. The optimal predictor can be written as a linear combination of input data points x_i,

$i = 1, \dots, n$. Moreover, for optimal primal and dual variables, the "complementary slackness" conditions for linear inequality constraints lead to $\alpha_i\big(y_i(w^\top x_i + b) - 1 + \xi_i\big) = 0$ and $(C - \alpha_i)\xi_i = 0$. This implies that $\alpha_i = 0$ as soon as $y_i(w^\top x_i + b) < 1$, and thus many of the α_i's are equal to zero, and the optimal predictor is a linear combination of only some of the data points x_i's which are then called "support vectors". The sparsity of the α_i's can be leveraged computationally (Platt, 1998), but statistically, given that the number of support vectors is proportional to the number n of observations (Steinwart, 2003), this sparsity alone cannot directly justify the potential superiority of the hinge loss over other convex surrogates.

4.1.3 Conditional Φ-risk and classification calibration (♦)

From margin bounds to convergence to optimal predictions. All of the convex surrogates presented in Section 4.1.1 are upper-bounds on the 0-1 loss or can be made so with rescaling. This simple fact allows us to get a variety of "margin bounds" where the 0-1 risk is upper-bounded by the Φ-risk. When the Φ-risk is equal to zero, which can only occur for problems with deterministic labels, this leads to a guarantee that the resulting classifier is the optimal one. In non-deterministic settings, however, the Φ-risk will be strictly positive, and while the margin bound shows that the error is controlled, it does not lead to guarantees to be close to the optimal predictions. In this section, we study the tools dedicated to obtaining such guarantees.

If we denote $\eta(x) = \mathbb{P}(y = 1|x) \in [0, 1]$, then we have, $\mathbb{E}[y|x] = 2\eta(x) - 1$, and, as seen in Chapter 2:

$$\mathcal{R}(g) = \mathbb{E}[\Phi_{0-1}(yg(x))] = \mathbb{E}[\mathbb{E}[1_{\mathrm{sign}(g(x)) \neq y}|x]] \geqslant \mathbb{E}[\min(\eta(x), 1 - \eta(x))] = \mathcal{R}^*,$$

and one best classifier is $f^*(x) = \mathrm{sign}(2\eta(x) - 1) = \mathrm{sign}(\mathbb{E}[y|x])$. Note that there are **many** potential other functions $g(x)$ than $2\eta(x) - 1$ so that $f^*(x) = \mathrm{sign}(g(x))$ is optimal. The first (minor) reason is the arbitrary choice of prediction for $\eta(x) = 1/2$. The other reason is that $g(x)$ has to have the same sign as $2\eta(x) - 1$, which leads to many possibilities beyond $2\eta(x) - 1$.

This section aims to ensure that the minimizers of the expected Φ-risk lead to optimal predictions.

Square loss. Before moving on to general functions Φ, the square loss leads to simple arguments. Indeed, as seen in Chapter 2, the function minimizing the expected Φ-risk is then $g(x) = \mathbb{E}(y|x) = 2\eta(x) - 1$, and taking its sign leads to the optimal prediction. Thus, using the square loss for binary classification leads to the optimal prediction in the population case.

General losses. To study the impact of using the Φ-risk, we first look at the conditional risk for a given x (as for the 0-1 loss, the function g that will minimize the Φ-risk can be determined by looking at each x separately).

Definition 4.1 (conditional Φ-risk) *Let $g : \mathcal{X} \to \mathbb{R}$, we define the conditional Φ-risk as*

$$\mathbb{E}[\Phi(yg(x))|x] = \eta(x)\Phi(g(x)) + (1-\eta(x))\Phi(-g(x)) \text{ which we denote } C_{\eta(x)}(g(x)),$$

with $C_\eta(\alpha) = \eta\Phi(\alpha) + (1-\eta)\Phi(-\alpha)$.

The least we can expect from a convex surrogate is that in the population case, where all x's decouple, the optimal $g(x)$ obtained by minimizing the conditional Φ-risk exactly leads to the same prediction as the Bayes predictor (at least when this prediction is unique). In other words, since the prediction is $\text{sign}(g(x))$, we want that for any $\eta \in [0,1]$ (below $\mathbb{R}_+^*$ is the set of strictly positive numbers, with a similar notation for $\mathbb{R}_-^*$):

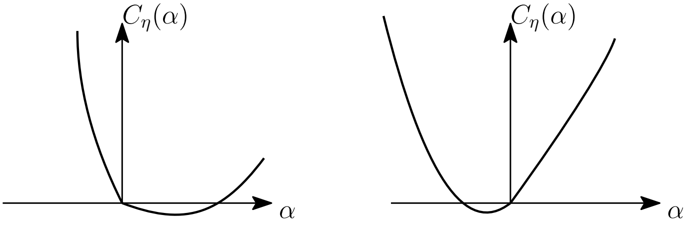
$$\text{(positive optimal prediction)} \qquad \eta > 1/2 \;\Leftrightarrow\; \arg\min_{\alpha\in\mathbb{R}} C_\eta(\alpha) \subset \mathbb{R}_+^* \tag{4.4}$$

$$\text{(negative optimal prediction)} \qquad \eta < 1/2 \;\Leftrightarrow\; \arg\min_{\alpha\in\mathbb{R}} C_\eta(\alpha) \subset \mathbb{R}_-^*. \tag{4.5}$$

A function Φ that satisfies these two statements is said *classification-calibrated*, or simply *calibrated*. It turns out that when Φ is convex, a simple sufficient and necessary condition is available:

Proposition 4.1 (Bartlett et al., 2006) *let $\Phi : \mathbb{R} \to \mathbb{R}$ convex. The surrogate function Φ is classification-calibrated if and only if Φ is differentiable at 0 and $\Phi'(0) < 0$.*

Proof Since Φ is convex, so is C_η for any $\eta \in [0,1]$, and thus we simply consider left and right derivatives at zero to obtain conditions about the location of minimizers, with the two possibilities below (minimizer in $\mathbb{R}_+^*$ if and only if the right derivative at zero is strictly negative, and minimizer in $\mathbb{R}_-^*$ if and only if the left derivative at zero is strictly positive):



$$\arg\min_{\alpha\in\mathbb{R}} C_\eta(\alpha) \subset \mathbb{R}_+^* \;\Leftrightarrow\; (C_\eta)_+(0)' = \eta\Phi'_+(0) - (1-\eta)\Phi'_-(0) < 0 \tag{4.6}$$

$$\arg\min_{\alpha\in\mathbb{R}} C_\eta(\alpha) \subset \mathbb{R}_-^* \;\Leftrightarrow\; (C_\eta)_-(0)' = \eta\Phi'_-(0) - (1-\eta)\Phi'_+(0) > 0. \tag{4.7}$$

(a) Assume Φ is calibrated. By letting η tend to $\frac{1}{2}+$ in Eq. (4.6), this leads to $(C_{1/2})_+(0)' = \frac{1}{2}\big[\Phi'_+(0) - \Phi'_-(0)\big] \leqslant 0$. Since Φ is convex, we always have the inequality $\Phi'_+(0) - \Phi'_-(0) \geqslant 0$. Thus, the left and right derivatives are equal, which implies that Φ is differentiable at 0. Then $C'_\eta(0) = (2\eta-1)\Phi'(0)$, and from Eq. (4.4) and Eq. (4.6), we need to have $\Phi'(0) < 0$.

(b) Assume Φ is differentiable at 0 and $\Phi'(0) < 0$, then $C'_\eta(0) = (2\eta-1)\Phi'(0)$; Eq. (4.4) and Eq. (4.5) are then direct consequences of Eq. (4.6) and Eq. (4.7).

■

Note that the proposition above excludes the convex surrogate $u \mapsto (-u)+ = \max\{-u, 0\}$, which is not differentiable at zero, but that all examples from Section 4.1.1 are calibrated.

We now assume that Φ is classification-calibrated and convex, that is, Φ is convex, Φ differentiable at 0, and $\Phi'(0) < 0$.

4.1.4 Relationship between risk and Φ-risk (♦♦)

Now that we know that for any $x \in \mathcal{X}$, minimizing $C_{\eta(x)}(g(x))$ with respect to $g(x)$ leads to the optimal prediction through $\text{sign}(g(x))$, we would like to make sure that an explicit control of the excess Φ-risk (which we aim to do with empirical risk minimization using tools from later sections) leads to an explicit control of the original excess risk. In other words, we are looking for an increasing function $H : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\mathcal{R}(g) - \mathcal{R}^* \leq H\big[\mathcal{R}_\Phi(g) - \mathcal{R}_\Phi^*\big]$, where $\mathcal{R}_\Phi^*$ is the minimum possible Φ-risk. The function H is often called the *calibration function.* This section shows that this calibration is the identity for the hinge loss (support vector machine), while it can be the square root for smooth convex surrogates such as the square and logistic losses.

As opposed to the least-squares regression case, where the loss function used for testing is directly the one used within empirical risk minimization, there are two notions here: the testing *error* $\mathcal{R}(g)$, which is obtained after thresholding at zero the function g, and the quantity $\mathcal{R}_\Phi(g)$, which is sometimes called the testing *loss*.

We first start with a simple lemma expressing the excess risk, as well as an upper bound (adapted from Theorem 2.2 from Devroye et al., 1996), that we will need for comparison inequalities below:

Lemma 4.1 *For any function $g : \mathcal{X} \to \mathbb{R}$, and for a Bayes predictor $g^* : \mathcal{X} \to \mathbb{R}$, we have:*

$$\mathcal{R}(g) - \mathcal{R}(g^*) = \mathbb{E}\big[1_{g(x)g^*(x)<0} \cdot |2\eta(x) - 1|\big].$$

Moreover, we have $\mathcal{R}(g) - \mathcal{R}(g^*) \leqslant \mathbb{E}\big[|2\eta(x) - 1 - g(x)|\big]$.

Proof We express the excess risk as:

$$\mathcal{R}(g) - \mathcal{R}(g^*) \quad = \mathbb{E}\big[\mathbb{E}\big[1_{\text{sign}(g(x))\neq y} - 1_{\text{sign}(g^*)(x)\neq y}\big|x\big]\big] \text{ by definition of the 0-1 loss.}$$

For any given $x \in \mathcal{X}$, we can look at the two possible cases for the signs of $\eta(x) - 1/2$ and $g(x)$ that lead to different predictions for g and g^*, namely (a) $\eta(x) > 1/2$ and $g(x) < 0$, and (b) $\eta(x) < 1/2$ and $g(x) > 0$ (equality cases are irrelevant). For the first case the expectation with respect to y is $\eta(x) - (1 - \eta(x)) = 2\eta(x) - 1$, while for the second case, we get $1 - 2\eta(x)$. By combining these two cases into the condition $g(x)g^*(x) < 0$ and the conditional expectation $|2\eta(x) - 1|$, we get the first result.

For the second result, we use the fact that if $g(x)g^*(x) < 0$, then, by splitting the cases in two (the first one being $\eta(x) > 1/2$ and $g(x) < 0$, the second one being $\eta(x) < 1/2$

and $g(x) > 0$), we get $|2\eta(x) - 1| \leqslant |2\eta(x) - 1 - g(x)|$, and thus the second result. ■

Note that for any function $b : \mathbb{R} \to \mathbb{R}$ that preserves the sign (that is $b(\mathbb{R}_+^*) \subset \mathbb{R}_+^*$ and $b(\mathbb{R}_-^*) \subset \mathbb{R}_-^*$), we have $\mathcal{R}(g) - \mathcal{R}(g^*) \leqslant \mathbb{E}[|2\eta(x) - 1 - b(g(x))|]$.

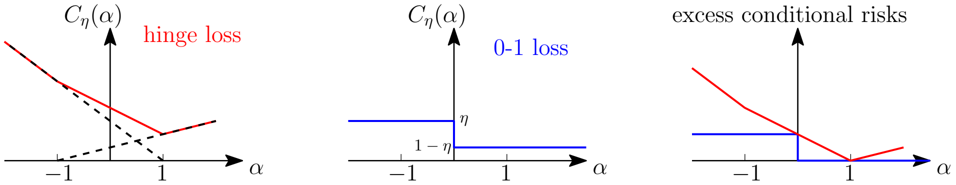
We see that the excess risk is the expectation of a quantity $|2\eta(x) - 1)| \cdot 1_{g(x)g^*(x)<0}$, which is equal to 0 if the classification through $g(x)$ is the same as the Bayes predictor and equal to $|2\eta(x) - 1|$ otherwise. The excess conditional Φ-risk is the quantity

$$\eta(x)\Phi(g(x)) + (1 - \eta(x))\Phi(-g(x)) - \inf_{\alpha} \{\eta(x)\Phi(\alpha) + (1 - \eta(x))\Phi(-\alpha)\},$$

which, as a function of $g(x)$, is the deviation between a convex function (of $g(x)$) and its minimum value. We simply need to relate it to the quantity $|2\eta(x) - 1)| \cdot 1_{g(x)g^*(x)<0}$ above for any $x \in \mathcal{X}$ and take expectations.

Zhang (2004a) and Bartlett et al. (2006) propose a general framework. We will only consider the hinge loss and smooth losses for simplicity (they already cover all cases from Section 4.1.1).

- For the hinge loss $\Phi(\alpha) = (1 - \alpha)_+ = \max\{1 - \alpha, 0\}$, we can easily compute the minimizer of the conditional Φ-risk (which leads to the minimizer of the Φ-risk). Indeed, we need to minimize $\eta(x)(1 - \alpha)_+ + (1 - \eta(x))(1 + \alpha)_+$, which is a piecewise affine function with kinks at -1 and 1, with a minimizer attained at $\alpha = 1$ for $\eta(x) > 1/2$ (see below), and symmetrically at $\alpha = -1$ for $\eta(x) < 1/2$, with a minimum conditional Φ-risk equal to $2\min\{1 - \eta(x), \eta(x)\}$. The two excess risks are plotted below for the hinge loss and the 0-1 loss, for $\eta(x) > 1/2$, showing pictorially that the conditional excess Φ-risk is greater than the excess risk.



  This leads to the calibration function $H(\sigma) = \sigma$ for the hinge loss.

  Note that when the Bayes risk is zero (but not in other cases), that is, $\eta(x) \in \{0, 1\}$ almost surely, then using the fact that the hinge loss is an upper-bound on the 0-1 loss is enough to show that the excess risk is less than the excess Φ-risk (indeed, the two optimal risks $\mathcal{R}^*$ and $\mathcal{R}_\Phi^*$ are equal to zero).

- For the square loss $\Phi(v) = (v - 1)^2$, we can use directly Lemma 4.1, to get, using Jensen's inequality

$$\mathcal{R}(g) - \mathcal{R}(g^*) \leqslant \mathbb{E}\big[|2\eta(x) - 1 - g(x)|\big] \leqslant \big(\mathbb{E}\big[|2\eta(x) - 1 - g(x)|^2\big]\big)^{1/2} = \big(\mathcal{R}_\Phi(g) - \mathcal{R}_\Phi^*\big)^{1/2},$$

  which is exactly a calibration result which we extend below to smooth losses.

- We consider smooth losses of the form (up to additive and multiplicative constants) $\Phi(v) = a(v) - v$, where $a(v) = \frac{1}{2}v^2$ for the quadratic loss, $a(v) = 2\log(e^{v/2} + e^{-v/2})$

for the logistic loss. We assume that a is even, $a(0) = 0$, a is β-smooth (that is, as defined in Chapter 5, $a''(v) \leqslant \beta$ for all $v \in \mathbb{R}$). This implies[3] that for all $v \in \mathbb{R}$, $a(v) - \alpha v - \inf_{w \in \mathbb{R}} \{a(w) - \alpha w\} \geqslant \frac{1}{2\beta}|\alpha - a'(v)|^2$, leading to:

$$\begin{aligned}
\mathcal{R}_\Phi(g) - \mathcal{R}_\Phi^* &= \mathbb{E}\big[a(g(x)) - (2\eta(x) - 1)g(x) - \inf_{w \in \mathbb{R}} \{a(w) - (2\eta(x) - 1)w\}\big] \\
&\geqslant \frac{1}{2\beta}\mathbb{E}\big[\big|2\eta(x) - 1 - a'(g(x))\big|^2\big] \text{ by the property above,} \\
&\geqslant \frac{1}{2\beta}\big(\mathbb{E}\big[\big|2\eta(x) - 1 - a'(g(x))\big|\big]\big)^2 \text{ by Jensen's inequality,} \\
&\geqslant \frac{1}{2\beta}\big(\mathcal{R}(g) - \mathcal{R}^*\big)^2,
\end{aligned}$$

using Lemma 4.1, and the fact that a' is sign-preserving (since $a'(0) = 0$). This leads to the calibration function $H(\sigma) = \sqrt{2\sigma}$ for the square loss and $H(\sigma) = 2\sqrt{\sigma}$ for the logistic loss.

Exercise 4.1 *(♦) Show that if a^* is the Fenchel conjugate of a, then for any function $g : \mathcal{X} \to \mathbb{R}$, we have $a^*\big(\mathcal{R}(g) - \mathcal{R}^*\big) \leqslant \mathcal{R}_\Phi(g) - \mathcal{R}_\Phi^*$.*

Exercise 4.2 *(♦♦) We consider a convex function $\Phi : \mathbb{R} \to \mathbb{R}$ which is differentiable at zero with $\Phi'(0) < 0$. Define $G(z) = \Phi(0) - \inf_{\alpha \in \mathbb{R}} \big\{\frac{1+z}{2}\Phi(\alpha) + \frac{1-z}{2}\Phi(-\alpha)\big\}$. Show that G is convex, $G(0) = 0$, and $G\big[\mathcal{R}(g) - \mathcal{R}^*\big] \leqslant \mathcal{R}_\Phi(g) - \mathcal{R}_\Phi^*$ for any function $g : \mathcal{X} \to \mathbb{R}$. Compute G for the exponential loss.*

We can make the following observations:

- For the (non-smooth) hinge loss, the calibration function is identity, so if the excess Φ-risk goes to zero at a specific rate, the excess risk goes to zero at the same rate. In contrast, for smooth losses, the upper bound only ensures a (worse) rate with a square root. Therefore, when going from the excess Φ-risk to the excess risk, that is, after thresholding the function g at zero, the observed rates may be worse. However, as shown in Chapter 5, smooth losses can be easier to optimize. There is, thus, a trade-off between these two types of losses.
- Note that the noiseless case where $\eta(x) \in \{0, 1\}$ (zero Bayes risk) leads to a stronger calibration function, as well as a series of intermediate "low-noise" conditions (see Bartlett et al., 2006, for details, as well as the exercise below).

Exercise 4.3 *(♦) Assume that $|2\eta(x) - 1| > \varepsilon$ almost surely, for some $\varepsilon \in (0, 1]$. Show that for any smooth convex classification calibrated function $\Phi : \mathbb{R} \to \mathbb{R}$ of the form $\Phi(v) = a(v) - v$ above, then we have $\mathcal{R}(g) - \mathcal{R}(g^*) \leqslant \frac{\varepsilon}{a^*(\varepsilon)}\big[\mathcal{R}_\Phi(g) - \mathcal{R}_\Phi^*\big]$ for any function $g : \mathcal{X} \to \mathbb{R}$.*

[3] Using the Fenchel conjugate $a^* : \mathbb{R} \to \mathbb{R}$ which is $(1/\beta)$-strongly convex (see Chapter 5), we have: $a(v) - \alpha v - \inf_{w \in \mathbb{R}} \{a(w) - \alpha w\} = a(v) - \alpha v + a^*(\alpha) = a^*(\alpha) - a^*(a'(v)) - (\alpha - a'(v))(a^*)'(a'(v)) \geqslant \frac{1}{2\beta}|\alpha - a'(v)|^2$, where a^* is the Fenchel conjugate of a (Boyd and Vandenberghe, 2004).

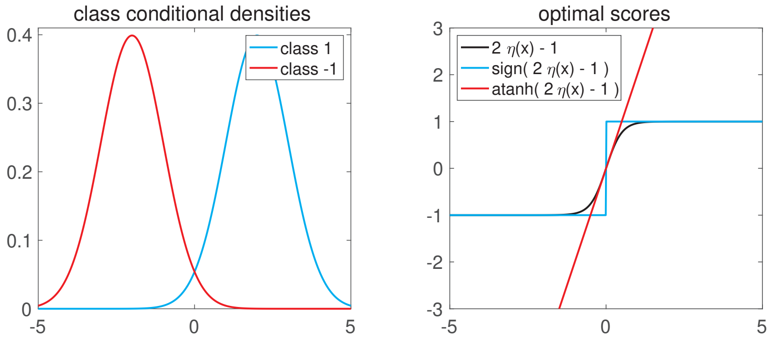


Figure 4.2: Optimal score functions for Gaussian class-conditional densities in one dimension. Left: conditional densities, right: optimal score functions for the square loss ($g^*(x) = 2\eta(x) - 1$), the hinge loss ($g^*(x) = \text{sign}(2\eta(x) - 1)$) and the logistic loss ($g^*(x) = \text{atanh}(2\eta(x) - 1)$).

Impact on approximation errors (♦). For the same classification problem, several convex surrogates can be used. While the Bayes classifier is always the same, that is, $f^*(x) = \text{sign}(2\eta(x)-1)$, the minimizer of the testing Φ-risk will be different. For example, for the hinge loss, the minimizer $g(x)$ is exactly $\text{sign}(2\eta(x) - 1)$, while for losses of the form like above $\Phi(v) = a(v)-v$, we have $a'(g(x)) = 2\eta(x)-1$, and thus for the square loss $g(x) = 2\eta(x) - 1$, while for the logistic loss, one can check that $g(x) = \text{atanh}(2\eta(x) - 1)$ (hyperbolic arc tangent). See examples in Figure 4.2, with $\mathcal{X} = \mathbb{R}$ and Gaussian class conditional densities.

The choice of surrogates will have an impact since to attain the minimal Φ-risk, different assumptions are needed on the class of functions used for empirical risk minimization, that is, $\text{sign}(2\eta(x) - 1)$ has to be in the class of functions we use (for the hinge loss), or $2\eta(x) - 1$ for the square loss, or $\text{atanh}(2\eta(x) - 1)$ for the logistic loss.

Exercise 4.4 *For the logistic loss, show that for data generated with class-conditional densities of $x|y = 1$ and $x|y = -1$, which are Gaussians with the same covariance matrix, the function $g(x)$ minimizing the expected logistic loss is affine in x (this model is often referred to as linear discriminant analysis). Provides an extension to the multi-class setting.*

Beyond calibration and loss consistency. The main property proved in this section is $\mathcal{R}(g) - \mathcal{R}^* \leq H\big[\mathcal{R}_\Phi(g) - \mathcal{R}_\Phi^*\big]$ for any prediction function $g : \mathcal{X} \to \mathbb{R}$, for a function H which tends to zero at zero. When the space of functions chosen for g is flexible enough to reach the minimizer of $\mathcal{R}_\Phi$, e.g., for kernel methods (Chapter 7) or neural networks with sufficiently many neurons (Chapter 9), then g will reach the minimum risk $\mathcal{R}(g)$. Such properties will also be available for structured prediction in Chapter 13.

However, it is common in practice, in particular in high dimensions, to use a restricted class of models, in particular linear models, where reaching the minimum Φ-risk is not possible anymore. In such setups, a more refined notion of consistency can be defined and studied, see, e.g., Long and Servedio (2013).

4.2 Risk minimization decomposition

We consider a family $\mathcal{F}$ of prediction functions $f : \mathcal{X} \to \mathbb{R}$. Empirical risk minimization aims to compute

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \widehat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)),$$

with algorithms presented in Chapter 5. We consider loss functions that are defined for real-valued outputs even for binary classification problems through the use of surrogates presented in Section 4.1.1.

We can decompose the risk as follows into two terms:

$$\begin{aligned} \mathcal{R}(\hat{f}) - \mathcal{R}^* &= \Big\{ \mathcal{R}(\hat{f}) - \inf_{f' \in \mathcal{F}} \mathcal{R}(f') \Big\} + \Big\{ \inf_{f' \in \mathcal{F}} \mathcal{R}(f') - \mathcal{R}^* \Big\} \\ &= \quad \text{estimation error} \qquad + \ \text{approximation error}. \end{aligned}$$

A classic example is the situation where a subset of $\mathbb{R}^d$ parameterizes the family of functions, that is, $\mathcal{F} = \{f_\theta, \ \theta \in \Theta\}$, for $\Theta \subset \mathbb{R}^d$. This includes neural networks (Chapter 9) and the simplest case of linear models of the form $f_\theta(x) = \theta^\top \varphi(x)$, for a particular feature vector $\varphi(x)$ (such as in Chapter 3). We will use linear models with Lipschitz-continuous loss functions as a motivating example, most often with constraints or penalties on the ℓ_2-norm $\|\theta\|_2$, but other norms can be considered as well (such as the ℓ_1-norm in Chapter 8).

We now turn separately to the approximation and estimation errors.

4.3 Approximation error

The approximation error $\inf_{f \in \mathcal{F}} \mathcal{R}(f) - \mathcal{R}^*$ is deterministic and depends on the underlying distribution, and the class $\mathcal{F}$ of functions: the larger the class, the smaller the approximation error.

Bounding the approximation error requires assumptions on the Bayes predictor (sometimes also called the "target function") f^*, and hence on the testing distribution.

In this section, we will focus on $\mathcal{F} = \{f_\theta, \ \theta \in \Theta\}$, for $\Theta \subset \mathbb{R}^d$ (we will consider infinite-dimensions in Chapter 7), and convex Lipschitz-continuous losses, assuming that θ_* is the minimizer of $\mathcal{R}(f_\theta)$ over $\theta \in \mathbb{R}^d$, which is assumed to exist (typically, θ_* does not belong to Θ). This implies that the approximation error decomposes into

$$\inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \mathcal{R}^* = \Big(\inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \inf_{\theta \in \mathbb{R}^d} \mathcal{R}(f_\theta) \Big) + \Big(\inf_{\theta \in \mathbb{R}^d} \mathcal{R}(f_\theta) - \mathcal{R}^* \Big).$$

- The second term $\inf_{\theta \in \mathbb{R}^d} \mathcal{R}(f_\theta) - \mathcal{R}^*$ is the incompressible approximation error coming from the chosen set of models f_θ. For flexible models such as kernel methods

(Chapter 7) or neural networks (Chapter 9), this incompressible error can be made as small as desired.

- The function $\theta \mapsto \mathcal{R}(f_\theta) - \inf_{\theta \in \mathbb{R}^d} \mathcal{R}(f_\theta)$ is a positive function on $\mathbb{R}^d$, which can be typically upper bounded by a specific norm (or its square) $\Omega(\theta - \theta_*)$, and we can see the first term above $\inf\limits_{\theta \in \Theta} \mathcal{R}(f_\theta) - \inf\limits_{\theta \in \mathbb{R}^d} \mathcal{R}(f_\theta)$ as a "distance" between θ_* and Θ.

  For example, if the loss which is considered is G-Lipschitz-continuous with respect to the second variable (which is possible for regression or when using a convex surrogate for binary classification as presented in Section 4.1), we have,

  $$\mathcal{R}(f_\theta) - \mathcal{R}(f_{\theta'}) \quad = \quad \mathbb{E}\big[\ell(y, f_\theta(x)) - \ell(y, f_{\theta'}(x))\big] \leqslant G\mathbb{E}\big[|f_\theta(x) - f_{\theta'}(x)|\big],$$

  and thus, this second part of the approximation error is upper bounded by G times the distance between f_{θ_*} and $\mathcal{F} = \{f_\theta, \ \theta \in \Theta\}$, for a particular pseudo-distance $d(\theta, \theta') = \mathbb{E}\big[|f_\theta(x) - f_{\theta'}(x)|\big]$.

  A classical example will be $f_\theta(x) = \theta^\top \varphi(x)$, and $\Theta = \{\theta \in \mathbb{R}^d, \ \|\theta\|_2 \leqslant D\}$, leading to the upper bound[4]

  $$\inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \inf_{\theta \in \mathbb{R}^d} \mathcal{R}(f_\theta) \leqslant G \inf_{\|\theta\|_2 \leqslant D} \mathbb{E}\big[\|\varphi(x)\|_2\big] \cdot \|\theta - \theta_*\|_2 \leqslant G\,\mathbb{E}\big[\|\varphi(x)\|_2\big](\|\theta_*\|_2 - D)_+,$$

  which is equal to zero if $\|\theta_*\|_2 \leqslant D$ (well-specified model).

  Exercise 4.5 *Show that for $\Theta = \{\theta \in \mathbb{R}^d, \ \|\theta\|_1 \leqslant D\}$ (ℓ_1-norm instead of the ℓ_2-norm), we have*

  $$\inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \inf_{\theta \in \mathbb{R}^d} \mathcal{R}(f_\theta) \leqslant G\,\mathbb{E}\big[\|\varphi(x)\|_\infty\big](\|\theta_*\|_1 - D)_+.$$

  Generalize to all norms.

4.4 Estimation error

We will consider general techniques and apply them as illustrations to linear models with bounded ℓ_2-norm by D and G-Lipschitz-losses. See further applications in Chapter 7 and Chapter 9.

The estimation error is often decomposed using $g_{\mathcal{F}} \in \arg\min_{g \in \mathcal{F}} \mathcal{R}(g)$ the minimizer of the expected risk for our class of models and $\hat{f} \in \arg\min_{f \in \mathcal{F}} \widehat{\mathcal{R}}(f)$ the minimizer of

[4] The identity $\inf_{\|\theta\|_2 \leqslant D} \|\theta - \theta_*\|_2 = (\|\theta_*\|_2 - D)_+$ can be shown by looking for the optimal θ proportional to θ_* and optimizing with respect to the proportionality constant.

the empirical risk:

$$\begin{aligned}\mathcal{R}(\hat{f}) - \inf_{f \in \mathcal{F}} \mathcal{R}(f) &= \mathcal{R}(\hat{f}) - \mathcal{R}(g_{\mathcal{F}}) \\ &= \big\{\mathcal{R}(\hat{f}) - \widehat{\mathcal{R}}(\hat{f})\big\} + \big\{\widehat{\mathcal{R}}(\hat{f}) - \widehat{\mathcal{R}}(g_{\mathcal{F}})\big\} + \big\{\widehat{\mathcal{R}}(g_{\mathcal{F}}) - \mathcal{R}(g_{\mathcal{F}})\big\} \\ &\leqslant \sup_{f \in \mathcal{F}} \big\{\mathcal{R}(f) - \widehat{\mathcal{R}}(f)\big\} + \big\{\widehat{\mathcal{R}}(\hat{f}) - \widehat{\mathcal{R}}(g_{\mathcal{F}})\big\} + \sup_{f \in \mathcal{F}} \big\{\widehat{\mathcal{R}}(f) - \mathcal{R}(f)\big\} \\ &\leqslant \sup_{f \in \mathcal{F}} \big\{\mathcal{R}(f) - \widehat{\mathcal{R}}(f)\big\} + 0 + \sup_{f \in \mathcal{F}} \big\{\widehat{\mathcal{R}}(f) - \mathcal{R}(f)\big\} \text{ by definition of } \hat{f}. \quad (4.8)\end{aligned}$$

This is often further upper-bounded by $2 \sup_{f \in \mathcal{F}} \big|\widehat{\mathcal{R}}(f) - \mathcal{R}(f)\big|$. We can make the following observations:

- The key tool to remove the statistical dependence between $\widehat{\mathcal{R}}$ and $\hat{f}$ is to take a *uniform* bound.
- When $\hat{f}$ is not the global minimizer of $\widehat{\mathcal{R}}$ but satisfies $\widehat{\mathcal{R}}(\hat{f}) \leqslant \inf_{f \in \mathcal{F}} \widehat{\mathcal{R}}(f) + \varepsilon$, then the *optimization error* ε has to be added to the bound above (see more details in Chapter 5).
- The uniform deviation grows with the "size" of $\mathcal{F}$, is a random quantity (because of its dependence on data), and usually decays with n. See the examples below.
- A key issue is that we need a *uniform control* for all $f \in \mathcal{F}$: with a single f, we could apply any concentration inequality to the random variable $\ell(y, f(x))$ to obtain a bound in $O(1/\sqrt{n})$; however, when controlling the maximal deviations over many functions f, there is always a small chance that one of these deviations get large. We thus need explicit control of this phenomenon, which we now tackle by first showing that we can focus on the expectation alone.

Since the estimation error is a random quantity, we need to bound it using probabilistic tools. This can be done either in high probability or in expectation. In the next section, we show how concentration inequalities allow us to focus on control in expectation.

4.4.1 Application of McDiarmid's inequality

Let $H(z_1, \dots, z_n) = \sup_{f \in \mathcal{F}} \big\{\mathcal{R}(f) - \widehat{\mathcal{R}}(f)\big\}$, where the random variables $z_i = (x_i, y_i)$ are independent and identically distributed, and $\widehat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i))$. We let ℓ_∞ denote the maximal absolute value of the loss functions for all (x, y) in the support of the data generating distribution and $f \in \mathcal{F}$ (for most loss functions, this is a consequence of having bounded prediction functions).

For a single function $f \in \mathcal{F}$, we can control the deviation between $\widehat{\mathcal{R}}(f)$, which is an empirical average of bounded independent random variables, and its expectation $\mathcal{R}(f)$ through Hoeffding's inequality, presented in detail and proved in Section 1.2.1: for any $\delta \in (0, 1)$, with probability greater than $1 - \delta$,

$$\mathcal{R}(f) - \widehat{\mathcal{R}}(f) \leqslant \frac{\ell_\infty \sqrt{2}}{\sqrt{n}} \sqrt{\log \frac{1}{\delta}}.$$

Such a control can be extended beyond a single function f. When changing a single $z_i \in \mathcal{X} \times \mathcal{Y}$ into $z_i' \in \mathcal{X} \times \mathcal{Y}$, the deviation in H is almost surely at most $\frac{2}{n}\ell_\infty$.[5] Thus, applying McDiarmid's inequality (see Section 1.2.2 in Chapter 1), with probability greater than $1-\delta$, we have:

$$H(z_1,\dots,z_n) - \mathbb{E}[H(z_1,\dots,z_n)] \leqslant \frac{\ell_\infty\sqrt{2}}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}}.$$

We thus only need to bound the expectation of $\sup_{f\in\mathcal{F}}\big\{\mathcal{R}(f) - \widehat{\mathcal{R}}(f)\big\}$ and of the similar quantity $\sup_{f\in\mathcal{F}}\big\{\widehat{\mathcal{R}}(f) - \mathcal{R}(f)\big\}$ (which will typically have the same bound), and add on top of it $\frac{\ell_\infty\sqrt{2}}{\sqrt{n}}\sqrt{\log\frac{2}{\delta}}$, to ensure a high-probability bound.[6]

We now provide a series of bounds to bound these expectations, from simple to more refined, culminating in Rademacher complexities in Section 4.5.

4.4.2 Easy case I: quadratic functions

We will show what happens with a quadratic loss function and an ℓ_2-ball constraint. We remember that in this case $\ell(y,\theta^\top\varphi(x)) = (y-\theta^\top\varphi(x))^2$. From that, we get

$$\begin{aligned}\widehat{\mathcal{R}}(f) - \mathcal{R}(f) &= \theta^\top\Big(\frac{1}{n}\sum_{i=1}^n \varphi(x_i)\varphi(x_i)^\top - \mathbb{E}\big[\varphi(x)\varphi(x)^\top\big]\Big)\theta \\ &\qquad -2\theta^\top\Big(\frac{1}{n}\sum_{i=1}^n y_i\varphi(x_i) - \mathbb{E}\big[y\varphi(x)\big]\Big) + \Big(\frac{1}{n}\sum_{i=1}^n y_i^2 - \mathbb{E}\big[y^2\big]\Big).\end{aligned}$$

Hence, the supremum can be upper-bounded in closed form as

$$\begin{aligned}\sup_{\|\theta\|_2\leqslant D}|\mathcal{R}(f) - \widehat{\mathcal{R}}(f)| &\leqslant D^2\Big\|\frac{1}{n}\sum_{i=1}^n \varphi(x_i)\varphi(x_i)^\top - \mathbb{E}\big[\varphi(x)\varphi(x)^\top\big]\Big\|_{\rm op} \\ &\qquad +2D\Big\|\frac{1}{n}\sum_{i=1}^n y_i\varphi(x_i) - \mathbb{E}\big[y\varphi(x)\big]\Big\|_2 + \Big|\frac{1}{n}\sum_{i=1}^n y_i^2 - \mathbb{E}\big[y^2\big]\Big|,\end{aligned}$$

where $\|M\|_{\rm op}$ is the operator norm of the matrix M defined as $\|M\|_{\rm op} = \sup_{\|u\|_2=1}\|Mu\|_2$ (for which we have $|u^\top Mu| \leqslant \|M\|_{\rm op}\|u\|_2^2$ for any vector u).

Thus, to get a uniform bound, we simply need to upper-bound the three *non-uniform* expectations of deviations, and therefore of order $O(1/\sqrt{n})$, and we get an overall uniform deviation bound. This case gives the impression that it should be possible to get such a rate in $O(1/\sqrt{n})$ for other types of losses than the quadratic loss. However, closed-form calculations are impossible, so we must introduce new tools.

[5] For a fixed function $f \in \mathcal{F}$, only one term in the average is changed, with absolute value less than ℓ_∞, thus a deviation of at most $\frac{2}{n}\ell_\infty$. This can be extended to the supremum by a simple computation left as an exercise.

[6] When combining two bounds in probability, the union bound leads to the term $2/\delta$ instead of $1/\delta$, see Section 1.2.1.

Exercise 4.6 *(♦) Provide an explicit bound on* $\sup_{\|\theta\|_2 \leqslant D} |\mathcal{R}(f) - \widehat{\mathcal{R}}(f)|$ *above, and compare it to using Rademacher complexities in Section 4.5. The concentration of averages of matrices from Section 1.2.6 can be used.*

⚠ Note that from now on, in the sections below, unless otherwise stated, we do not require the loss to be convex.

4.4.3 Easy case II: Finite number of models

We assume in this section that the loss functions are bounded between 0 and ℓ_∞, using the upper-bound $2\sup_{f\in\mathcal{F}} \big|\widehat{\mathcal{R}}(f) - \mathcal{R}(f)\big|$ on the estimation error, and the union bound:

$$\mathbb{P}\Big(\mathcal{R}(\hat{f}) - \inf_{f\in\mathcal{F}} \mathcal{R}(f) \geqslant t\Big) \leqslant \mathbb{P}\Big(2\sup_{f\in\mathcal{F}} \big|\widehat{\mathcal{R}}(f) - \mathcal{R}(f)\big| \geqslant t\Big) \leqslant \sum_{f\in\mathcal{F}} \mathbb{P}\Big(2\big|\widehat{\mathcal{R}}(f) - \mathcal{R}(f)\big| \geqslant t\Big).$$

We have, for $f \in \mathcal{F}$ fixed, $\widehat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i))$, and we can apply Hoeffding's inequality from Section 1.2.1 to bound each $\mathbb{P}\big(2\big|\widehat{\mathcal{R}}(f) - \mathcal{R}(f)\big| \geqslant t\big)$, leading to

$$\mathbb{P}\Big(\mathcal{R}(\hat{f}) - \inf_{f\in\mathcal{F}} \mathcal{R}(f) \geqslant t\Big) \leqslant \sum_{f\in\mathcal{F}} 2\exp(-2n(t/2)^2/\ell_\infty^2) = 2|\mathcal{F}|\exp(-nt^2/(2\ell_\infty^2)).$$

Thus, by setting $\delta = 2|\mathcal{F}|\exp(-nt^2/2\ell_\infty^2)$, and finding the corresponding t, with probability greater than $1-\delta$, we get (using $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$):

$$\begin{aligned}\mathcal{R}(\hat{f}) - \inf_{f\in\mathcal{F}} \mathcal{R}(f) &\leqslant t = \frac{\sqrt{2}\ell_\infty}{\sqrt{n}}\sqrt{\log\frac{2|\mathcal{F}|}{\delta}} = \frac{\sqrt{2}\ell_\infty}{\sqrt{n}}\sqrt{\log(2|\mathcal{F}|) + \log\frac{1}{\delta}} \\ &\leqslant \sqrt{2}\ell_\infty\sqrt{\frac{\log(2|\mathcal{F}|)}{n}} + \frac{\sqrt{2}\ell_\infty}{\sqrt{n}}\sqrt{\log\frac{1}{\delta}}.\end{aligned}$$

Exercise 4.7 *(♦) In terms of expectation, show that (using the proof of the max of random variables from Section 1.2.4 in Chapter 2, which applies because bounded random variables are sub-Gaussian):*

$$\mathbb{E}\big[\mathcal{R}(\hat{f}) - \inf_{f\in\mathcal{F}} \mathcal{R}(f)\big] \leqslant 2\mathbb{E}\Big[\sup_{f\in\mathcal{F}} \big|\widehat{\mathcal{R}}(f) - \mathcal{R}(f)\big|\Big] \leqslant \ell_\infty\sqrt{\frac{2\log(2|\mathcal{F}|)}{n}}.$$

Thus, according to the bound, learning is possible when the logarithm $\log(|\mathcal{F}|)$ of the number of models is small compared to n. This is the first generic control of uniform deviations.

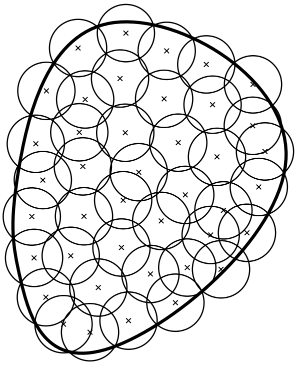
⚠ Note that this is only an upper bound, and learning is possible with infinitely many models (the most classical scenario). See below.

4.4.4 Beyond finitely many models through covering numbers (♦)

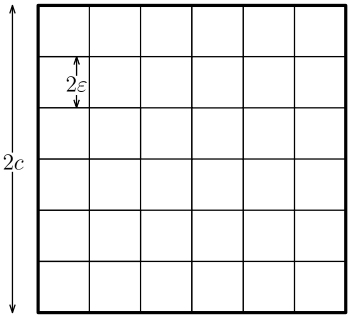
The simple idea behind covering numbers is to deal with function spaces with infinitely many elements by approximating them through a finite number of elements. This is often referred to as an "ε-net argument."

For simplicity, we assume that the loss functions are regular, for example, that they are G-Lipschitz-continuous with respect to their second argument.

Covering numbers. We assume there exists $m = m(\varepsilon)$ elements $f_1, \dots, f_m$ such that for any $f \in \mathcal{F}$, $\exists i \in \{1, \dots, m\}$ such that $d(f, f_i) \leqslant \varepsilon$. The minimal possible number $m(\varepsilon)$ is the *covering number* of $\mathcal{F}$ at precision ε. See an example below in two dimensions of a covering with Euclidean balls.



The covering number $m(\varepsilon)$ is a non-increasing function of ε. Typically, $m(\varepsilon)$ grows with ε as a power ε^{-d} when $\varepsilon \to 0$, where d is the underlying dimension. Indeed, for the ℓ_∞-metric, if (in a certain parameterization) $\mathcal{F}$ is included in a ball of radius c in the ℓ_∞-ball of dimension d, it can be easily covered by $(c/\varepsilon)^d$ cubes of length 2ε. See below.



Given that all norms are equivalent in dimension d, we get the same dependence in ε^{-d} of $m(\varepsilon)$ for all bounded subsets of a finite-dimensional vector space, and thus $\log m(\varepsilon)$ grows as $d \log \frac{1}{\varepsilon}$ when ε tends to zero.

For some sets (e.g., all Lipschitz-continuous functions in d dimensions) $\log m(\varepsilon)$ grows faster, for example as ε^{-d}. See, e.g., Wainwright (2019).

ε-net argument. Given a cover of $\mathcal{F}$, for all $f \in \mathcal{F}$, and with $(f_i)_{i\in\{1,\dots,m(\varepsilon)\}}$ the associated cover elements,

$$\begin{aligned}\big|\widehat{\mathcal{R}}(f) - \mathcal{R}(f)\big| &\leqslant \big|\widehat{\mathcal{R}}(f) - \widehat{\mathcal{R}}(f_i)\big| + \big|\widehat{\mathcal{R}}(f_i) - \mathcal{R}(f_i)\big| + \big|\mathcal{R}(f_i) - \mathcal{R}(f)\big| \\ &\leqslant 2G\varepsilon + \sup_{i\in\{1,\dots,m(\varepsilon)\}} \big|\widehat{\mathcal{R}}(f_i) - \mathcal{R}(f_i)\big|.\end{aligned}$$

This implies that using bounds on the expectation of the maximum (Section 1.2.4), which apply because bounded random variables are sub-Gaussian (with the sub-Gaussianity parameter proportional to the almost sure bound):

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}} \big|\widehat{\mathcal{R}}(f)-\mathcal{R}(f)\big|\Big] \leqslant 2G\varepsilon + \mathbb{E}\Big[\sup_{i\in\{1,\dots,m(\varepsilon)\}} \big|\widehat{\mathcal{R}}(f_i)-\mathcal{R}(f_i)\big|\Big] \leqslant 2G\varepsilon + 2\ell_\infty\sqrt{\frac{2\log(2m(\varepsilon)))}{n}}.$$

Therefore, if $m(\varepsilon) \sim \varepsilon^{-d}$, ignoring constants, we need to balance $\varepsilon + \sqrt{d\log(1/\varepsilon)/n}$, which leads to, with a choice of ε proportional to $1/\sqrt{n}$, to a rate proportional to $\sqrt{(d/n)\log(n)}$, which shows that the dependence in n is also close to $1/\sqrt{n}$. Unfortunately, unless refined computations of covering numbers or more advanced tools (such as "chaining") are used, this often leads to a non-optimal dependence on dimension and/or number of observations (see, e.g., Wainwright, 2019, for examples of these refinements).

One powerful tool that allows sharp bounds at a reasonable cost is Rademacher complexities (Boucheron et al., 2005) or Gaussian complexities (Bartlett and Mendelson, 2002). In this chapter, we will focus on Rademacher complexity.

4.5 Rademacher complexity

We consider n independent and identically distributed random variables $z_1, \dots, z_n \in \mathcal{Z}$, and a class $\mathcal{H}$ of functions from $\mathcal{Z}$ to $\mathbb{R}$. In our context, the space of functions is related to the learning problem as: $z = (x, y)$, and $\mathcal{H} = \{(x, y) \mapsto \ell(y, f(x)),\ f \in \mathcal{F}\}$.

Our goal in this section is to provide an upper-bound on $\sup_{f\in\mathcal{F}} \big\{\mathcal{R}(f) - \widehat{\mathcal{R}}(f)\big\}$, which happens to be equal to

$$\sup_{h\in\mathcal{H}} \Big\{\mathbb{E}[h(z)] - \frac{1}{n}\sum_{i=1}^n h(z_i)\Big\},$$

where $\mathbb{E}[h(z)]$ denotes the expectation with respect to a variable having the same distribution as all z_i's.

We denote by $\mathcal{D} = \{z_1, \dots, z_n\}$ the data. We define the *Rademacher complexity* of the class of functions $\mathcal{H}$ from $\mathcal{Z}$ to $\mathbb{R}$:

$$\mathrm{R}_n(\mathcal{H}) = \mathbb{E}_{\varepsilon,\mathcal{D}}\Big(\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i h(z_i)\Big), \tag{4.9}$$

where $\varepsilon \in \mathbb{R}^n$ is a vector of independent Rademacher random variable (that is, taking values -1 or 1 with equal probabilities), which is also independent of $\mathcal{D}$. It is a deterministic quantity that only depends on n and $\mathcal{H}$.

In words, the Rademacher complexity is equal to the expectation of the maximal dot-product between values of a function h at the observations z_i and random labels. It measures the "capacity" of the set of functions $\mathcal{H}$. We will see later that it can be computed (or simply upper-bounded) in many interesting cases, leading to powerful bounds. The term "Rademacher average" is also commonly used.

⚠ Be careful with the two notations $\mathrm{R}_n(\mathcal{H})$ (Rademacher complexity) and $\mathcal{R}(f)$ (risk of the prediction function f), not to be confused with the feature norm R often used with linear models.

Exercise 4.8 *Show the following properties of Rademacher complexities (see Bartlett and Mendelson, 2002, for more details):*

(a) *If* $\mathcal{H} \subset \mathcal{H}'$, *then* $\mathrm{R}_n(\mathcal{H}) \leqslant \mathrm{R}_n(\mathcal{H}')$.

(b) $\mathrm{R}_n(\mathcal{H} + \mathcal{H}') = \mathrm{R}_n(\mathcal{H}) + \mathrm{R}_n(\mathcal{H}')$.

(c) *If* $\alpha \in \mathbb{R}$, $\mathrm{R}_n(\alpha\mathcal{H}) = |\alpha|\mathrm{R}_n(\mathcal{H})$.

(d) *If* $h_0 : \mathcal{Z} \to \mathbb{R}$, $\mathrm{R}_n(\mathcal{H} + \{h_0\}) = \mathrm{R}_n(\mathcal{H})$.

(e) $\mathrm{R}_n(\mathcal{H}) = \mathrm{R}_n(\text{convex hull}(\mathcal{H}))$.

Exercise 4.9 (Massart's lemma) *If* $\mathcal{H} = \{h_1, \dots, h_m\}$, *and almost surely we have the bound* $\frac{1}{n}\sum_{i=1}^n h_j(x_i)^2 \leqslant R^2$ *for all* $j \in \{1, \dots, m\}$, *then the Rademacher complexity of the class of functions* $\mathcal{H}$ *satisfies* $\mathrm{R}_n(\mathcal{H}) \leqslant \sqrt{\frac{2\log m}{n}}R$.

4.5.1 Symmetrization

First, we relate the Rademacher complexity to the uniform deviation through a general "symmetrization" property, which shows that the Rademacher complexity directly controls the expected uniform deviation.

Proposition 4.2 (symmetrization) *Given the Rademacher complexity of* $\mathcal{H}$ *defined in Eq. (4.9), we have:*

$$\mathbb{E}\Big[\sup_{h\in\mathcal{H}}\Big(\frac{1}{n}\sum_{i=1}^n h(z_i)-\mathbb{E}[h(z)]\Big)\Big] \leqslant 2\mathrm{R}_n(\mathcal{H}) \ , \ \mathbb{E}\Big[\sup_{h\in\mathcal{H}}\Big(\mathbb{E}[h(z)]-\frac{1}{n}\sum_{i=1}^n h(z_i)\Big)\Big] \leqslant 2\mathrm{R}_n(\mathcal{H}).$$

Proof (♦) Let $\mathcal{D}' = \{z_1', \dots, z_n'\}$ be an independent copy of the data $\mathcal{D} = \{z_1, \dots, z_n\}$. Let $(\varepsilon_i)_{i\in\{1,\dots,n\}}$ be i.i.d. Rademacher random variables, which are also independent of $\mathcal{D}$ and $\mathcal{D}'$. Using that for all i in $\{1, \dots, n\}$, $\mathbb{E}[h(z_i')|\mathcal{D}] = \mathbb{E}[h(z)]$, we have:

$$\begin{aligned}
\mathbb{E}\Big[\sup_{h\in\mathcal{H}}\Big(\mathbb{E}[h(z)] - \frac{1}{n}\sum_{i=1}^n h(z_i)\Big)\Big] &= \mathbb{E}\Big[\sup_{h\in\mathcal{H}}\Big(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[h(z_i')|\mathcal{D}] - \frac{1}{n}\sum_{i=1}^n h(z_i)\Big)\Big] \\
&= \mathbb{E}\Big[\sup_{h\in\mathcal{H}}\Big(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[h(z_i') - h(z_i)\big|\mathcal{D}\big]\Big)\Big],
\end{aligned}$$

by definition of the independent copy $\mathcal{D}'$. Then

$$\mathbb{E}\bigg[\sup_{h\in\mathcal{H}}\bigg(\mathbb{E}[h(z)]-\frac{1}{n}\sum_{i=1}^n h(z_i)\bigg)\bigg] \leqslant \mathbb{E}\bigg[\mathbb{E}\bigg(\sup_{h\in\mathcal{H}}\bigg(\frac{1}{n}\sum_{i=1}^n\big[h(z_i')-h(z_i)\big]\bigg)\bigg|\mathcal{D}\bigg)\bigg],$$

using that the supremum of the expectation is less than the expectation of the supremum. Thus, by the towering law of expectation, we get

$$\mathbb{E}\bigg[\sup_{h\in\mathcal{H}}\bigg(\mathbb{E}[h(z)]-\frac{1}{n}\sum_{i=1}^n h(z_i)\bigg)\bigg] \leqslant \mathbb{E}\bigg[\sup_{h\in\mathcal{H}}\bigg(\frac{1}{n}\sum_{i=1}^n\big[h(z_i')-h(z_i)\big]\bigg)\bigg].$$

We can now use the symmetry of the laws of ε_i and $h(z_i')-h(z_i)$, to get:

$$\begin{aligned}
& \mathbb{E}\bigg[\sup_{h\in\mathcal{H}}\bigg(\mathbb{E}[h(z)]-\frac{1}{n}\sum_{i=1}^n h(z_i)\bigg)\bigg] \\
\leqslant\ & \mathbb{E}\bigg[\sup_{h\in\mathcal{H}}\bigg(\frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(h(z_i')-h(z_i)\big)\bigg)\bigg] \\
\leqslant\ & \mathbb{E}\bigg[\sup_{h\in\mathcal{H}}\bigg(\frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(h(z_i)\big)\bigg)\bigg] + \mathbb{E}\bigg[\sup_{h\in\mathcal{H}}\bigg(\frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(-h(z_i)\big)\bigg)\bigg] \\
=\ & 2\mathbb{E}\bigg[\sup_{h\in\mathcal{H}}\bigg(\frac{1}{n}\sum_{i=1}^n \varepsilon_i h(z_i)\bigg)\bigg] = 2\mathrm{R}_n(\mathcal{H}).
\end{aligned}$$

The reasoning is essentially identical for $\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n h(z_i)-\mathbb{E}[h(z)]\right)\right] \leqslant 2\mathrm{R}_n(\mathcal{H})$. ■

The lemma above only bounds the expectation of the deviation between the empirical average and the expectation by the Rademacher average. Together with concentration inequalities from Section 1.2, we can obtain high-probability bounds, as done in Section 4.4.1 with McDiarmid's inequality.

Exercise 4.10 *(♦) The Gaussian complexity of a class of functions $\mathcal{H}$ from $\mathcal{Z}$ to $\mathbb{R}$ is defined as $G_n(\mathcal{H}) = \mathbb{E}_{\varepsilon,\mathcal{D}}\big(\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \varepsilon_i h(z_i)\big)$, where $\varepsilon\in\mathbb{R}^n$ is a vector of independent Gaussian variables with mean zero and variance one. Show that (a) $\mathrm{R}_n(\mathcal{H}) \leqslant \sqrt{\frac{\pi}{2}}\cdot G_n(\mathcal{H})$ and (b) $G_n(\mathcal{H}) \leqslant \sqrt{2\log(2n)}\cdot \mathrm{R}_n(\mathcal{H})$.*

Empirical Rademacher complexities (♦). The Rademacher complexity $\mathrm{R}_n(\mathcal{H})$ defined in Eq. (4.9) is a *deterministic* quantity that depends on the distribution of inputs. When using bound in high-probability through MacDiarmid's inequality in Section 4.4.1, we obtained that if $|h(z)| \leqslant \ell_\infty$ for all $h\in\mathcal{H}$, then with probability greater than $1-\delta$, for all $h\in\mathcal{H}$,

$$\mathbb{E}[h(z)] \leqslant \frac{1}{n}\sum_{i=1}^n h(z_i) + 2\mathrm{R}_n(\mathcal{H}) + \frac{2\ell_\infty}{\sqrt{n}}\sqrt{\log(1/\delta)}.$$

While we provide below estimates based on simple information on the input distribution, an empirical version can be defined that does not take the expectation with respect to the data, that is,

$$\hat{\mathrm{R}}_n(\mathcal{H}) = \mathbb{E}_\varepsilon\Big(\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i h(z_i)\Big), \tag{4.10}$$

which is now a random quantity that is computable from the training data and the class of functions. We can also use MacDiarmid's inequality to bound the difference between $\mathrm{R}_n(\mathcal{H})$ and $\hat{\mathrm{R}}_n(\mathcal{H})$, obtain a similar high-probability bound as above, that is,

$$\mathbb{E}[h(z)] \leqslant \frac{1}{n}\sum_{i=1}^n h(z_i) + 2\hat{\mathrm{R}}_n(\mathcal{H}) + 4\frac{2\ell_\infty}{\sqrt{n}}\sqrt{\log(2/\delta)},$$

which is now computable.

4.5.2 Lipschitz-continuous losses

A particularly appealing property in our context is the following property, sometimes called the "contraction principle," using a simple proof from Meir and Zhang (2003, Lemma 5); see also Ledoux and Talagrand (1991, Section 4.5). See Prop. 4.4 below for a similar result for the Rademacher complexity defined with absolute values (and then with an extra factor of 2).

Proposition 4.3 (Contraction principle - Lipschitz-continuous functions) *Given any functions b, $a_i : \Theta \to \mathbb{R}$ (no assumption) and $\varphi_i : \mathbb{R} \to \mathbb{R}$ any 1-Lipschitz-functions, for $i = 1, \dots, n$, we have, for $\varepsilon \in \mathbb{R}^n$ a vector of independent Rademacher random variables:*

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

Proof (♦) We consider a proof by induction on n. The case $n = 0$ is trivial, and we show how to go from $n \geqslant 0$ to $n+1$. We thus consider $\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_{n+1}}\Big[\sup_{\theta\in\Theta}\Big\{b(\theta)+\sum_{i=1}^{n+1}\varepsilon_i\varphi_i(a_i(\theta))\Big\}\Big]$ and compute the expectation with respect to ε_{n+1} explicitly, by considering the two potential values with probability $1/2$:

$$\begin{aligned}
&\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_{n+1}}\Big[\sup_{\theta\in\Theta}\Big\{b(\theta) + \sum_{i=1}^{n+1}\varepsilon_i\varphi_i(a_i(\theta))\Big\}\Big] \\
&= \frac{1}{2}\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_n}\Big[\sup_{\theta\in\Theta}\Big\{b(\theta)+\sum_{i=1}^n\varepsilon_i\varphi_i(a_i(\theta)) + \varphi_{n+1}(a_{n+1}(\theta))\Big\}\Big] \\
&\qquad + \frac{1}{2}\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_n}\Big[\sup_{\theta\in\Theta}\Big\{b(\theta)+\sum_{i=1}^n\varepsilon_i\varphi_i(a_i(\theta)) - \varphi_{n+1}(a_{n+1}(\theta))\Big\}\Big],
\end{aligned}$$

which is equal to

$$\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_n}\bigg[\sup_{\theta,\theta'\in\Theta}\bigg\{\frac{b(\theta)+b(\theta')}{2} + \sum_{i=1}^n \varepsilon_i \frac{\varphi_i(a_i(\theta))+\varphi_i(a_i(\theta'))}{2} + \frac{\varphi_{n+1}(a_{n+1}(\theta))-\varphi_{n+1}(a_{n+1}(\theta'))}{2}\bigg\}\bigg],$$

by assembling the terms. By taking the supremum over (θ,θ') and (θ',θ), we get

$$\begin{aligned}&\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_n}\bigg[\sup_{\theta,\theta'\in\Theta}\bigg\{\frac{b(\theta)+b(\theta')}{2} + \sum_{i=1}^n \varepsilon_i \frac{\varphi_i(a_i(\theta))+\varphi_i(a_i(\theta'))}{2} + \frac{|\varphi_{n+1}(a_{n+1}(\theta))-\varphi_{n+1}(a_{n+1}(\theta'))|}{2}\bigg\}\bigg]\\ \leqslant\ &\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_n}\bigg[\sup_{\theta,\theta'\in\Theta}\bigg\{\frac{b(\theta)+b(\theta')}{2} + \sum_{i=1}^n \varepsilon_i \frac{\varphi_i(a_i(\theta))+\varphi_i(a_i(\theta'))}{2} + \frac{|a_{n+1}(\theta)-a_{n+1}(\theta')|}{2}\bigg\}\bigg],\end{aligned}$$

using Lipschitz-continuity. We can redo the same sequence of *equalities* with φ_{n+1} being the identity to obtain that the last expression above is equal to

$$\begin{aligned}&\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_n}\mathbb{E}_{\varepsilon_{n+1}}\bigg[\sup_{\theta\in\Theta}\bigg\{b(\theta)+\varepsilon_{n+1}a_{n+1}(\theta)+\sum_{i=1}^n\varepsilon_i\varphi_i(a_i(\theta))\bigg\}\bigg]\\ \leqslant\ &\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_n,\varepsilon_{n+1}}\bigg[\sup_{\theta\in\Theta}\bigg\{b(\theta)+\varepsilon_{n+1}a_{n+1}(\theta)+\sum_{i=1}^n\varepsilon_i a_i(\theta)\bigg\}\bigg] \text{ by the induction hypothesis,}\end{aligned}$$

which leads to the desired result. ■

We can apply the contraction principle above to supervised learning situations where $u_i \mapsto \ell(y_i, u_i)$ is G-Lipschitz-continuous for all i almost surely (which is possible for regression or when using a convex surrogate for binary classification as presented in Section 4.1), leading to, by the contraction principle (applied conditioned on the data $\mathcal{D}$ to $b=0$, $\Theta = \{(f(x_1),\dots,f(x_n)),\ f\in\mathcal{F}\}\subset\mathbb{R}^n$ and $a_i(\theta)=\theta_i$, $\varphi_i(u_i)=\ell(y_i,u_i)$):

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

which leads to

$$\mathrm{R}_n(\mathcal{H}) \leqslant G\cdot\mathrm{R}_n(\mathcal{F}). \tag{4.11}$$

Thus, the Rademacher complexity of the class of prediction functions controls the uniform deviations of the empirical risk. We consider simple examples below but give before without proof a contraction result that we will need in Section 9.2.3 (See proof by Ledoux and Talagrand, 1991, Theorem 4.12).

Proposition 4.4 (Contraction principle - absolute values) *Given any functions $a_i : \Theta \to \mathbb{R}$ (no assumption) and $\varphi_i : \mathbb{R} \to \mathbb{R}$ any 1-Lipschitz-functions such that $\varphi_i(0) = 0$, for $i = 1, \dots, n$, we have, for $\varepsilon \in \mathbb{R}^n$ a vector of independent Rademacher random variables:*

$$\mathbb{E}_\varepsilon \bigg[\sup_{\theta \in \Theta} \Big| \sum_{i=1}^n \varepsilon_i \varphi_i(a_i(\theta)) \Big| \bigg] \leqslant 2\, \mathbb{E}_\varepsilon \bigg[\sup_{\theta \in \Theta} \Big| \sum_{i=1}^n \varepsilon_i a_i(\theta) \Big| \bigg].$$

4.5.3 Ball-constrained linear predictions

We now assume that $\mathcal{F} = \{ f_\theta(x) = \theta^\top \varphi(x),\ \Omega(\theta) \leqslant D \}$ where Ω is a norm on $\mathbb{R}^d$. We denote the design matrix by $\Phi \in \mathbb{R}^{n \times d}$. We have (with expectations both with respect to ε and the data):

$$\begin{aligned} \mathrm{R}_n(\mathcal{F}) &= \mathbb{E}\bigg[\sup_{\Omega(\theta) \leqslant D} \Big\{ \frac{1}{n} \sum_{i=1}^n \varepsilon_i \theta^\top \varphi(x_i) \Big\} \bigg] = \mathbb{E}\bigg[\sup_{\Omega(\theta) \leqslant D} \frac{1}{n} \varepsilon^\top \Phi \theta \bigg] \\ &= \frac{D}{n} \mathbb{E}\Big[\Omega^*(\Phi^\top \varepsilon) \Big], \end{aligned}$$

where $\Omega^*(u) = \sup_{\Omega(\theta) \leqslant 1} u^\top \theta$ is the *dual norm* of Ω. For example, when Ω is the ℓ_p-norm, with $p \in [1, \infty]$, then Ω^* is the ℓ_q-norm, where q is such that $\frac{1}{p} + \frac{1}{q} = 1$, e.g., $\| \cdot \|_2^* = \| \cdot \|_2$, $\| \cdot \|_1^* = \| \cdot \|_\infty$, and $\| \cdot \|_\infty^* = \| \cdot \|_1$. For more details, see Boyd and Vandenberghe (2004).

Thus, computing Rademacher complexities is equivalent to computing expectations of norms. When $\Omega = \| \cdot \|_2$, we get:

$$\begin{aligned} \mathrm{R}_n(\mathcal{F}) &= \frac{D}{n} \mathbb{E}\left[\| \Phi^\top \varepsilon \|_2 \right] \leqslant \frac{D}{n} \sqrt{\mathbb{E}\left[\| \Phi^\top \varepsilon \|_2^2 \right]} \text{ by Jensen's inequality,} \\ &= \frac{D}{n} \sqrt{\mathbb{E}\left[\mathrm{tr}[\Phi^\top \varepsilon \varepsilon^\top \Phi] \right]} = \frac{D}{n} \sqrt{\mathbb{E}\left[\mathrm{tr}[\Phi^\top \Phi] \right]} \text{ using that } \mathbb{E}[\varepsilon \varepsilon^\top] = I, \\ &= \frac{D}{n} \sqrt{\sum_{i=1}^n \mathbb{E}(\Phi^\top \Phi)_i} = \frac{D}{n} \sqrt{\sum_{i=1}^n \mathbb{E} \| \varphi(x_i) \|_2^2} = \frac{D}{\sqrt{n}} \sqrt{\mathbb{E} \| \varphi(x) \|_2^2}. \end{aligned} \tag{4.12}$$

We thus obtain a *dimension-independent* Rademacher complexity that we can use in the summary in Section 4.5.4 below.

Exercise 4.11 *(ℓ_1-norm) Assume that almost surely $\| \varphi(x) \|_\infty \leqslant R$. Show that the Rademacher complexity $\mathrm{R}_n(\mathcal{F})$ for $\mathcal{F} = \{ f_\theta(x) = \theta^\top \varphi(x),\ \Omega(\theta) \leqslant D \}$ with $\Omega = \| \cdot \|_1$ is upper-bounded by $RD\sqrt{\frac{2 \log(2d)}{n}}$.*

Exercise 4.12 *(♦) Let $p > 1$, and q such that $1/p + 1/q = 1$. Assume that almost surely $\| \varphi(x) \|_q \leqslant R$. Show that the Rademacher complexity $\mathrm{R}_n(\mathcal{F})$ for $\mathcal{F} = \{ f_\theta(x) = \theta^\top \varphi(x),\ \Omega(\theta) \leqslant D \}$, with $\Omega = \| \cdot \|_p$, is upper-bounded by $\frac{RD}{\sqrt{n}} \frac{1}{\sqrt{p-1}}$ (hint: use Exercise 1.19). Recover the result of Exercise 4.11 by taking $p = 1 + \frac{1}{\log(2d)}$.*

4.5.4 Putting things together (linear predictions)

We now consider a linear model based on some feature map $\varphi : \mathcal{X} \to \mathbb{R}^d$ and apply the Rademacher results from the previous section to obtain a bound on the estimation error. We then look at the approximation error.

Estimation error. With all the elements above, we can now propose the following general result (where no convexity of the loss function is assumed) for the estimation error. Note that there is no explicit dependence on the underlying dimension d, which will be important in Chapter 7 where we consider infinite-dimensional feature spaces.

Proposition 4.5 (Estimation error) *Assume a G-Lipschitz-continuous loss function, linear prediction functions with $\mathcal{F} = \{f_\theta(x) = \theta^\top \varphi(x),\ \|\theta\|_2 \leqslant D\}$, where $\mathbb{E}\|\varphi(x)\|_2^2 \leqslant R^2$. Let $\hat{f} = f_{\hat{\theta}} \in \mathcal{F}$ be the minimizer of the empirical risk, then:*

$$\mathbb{E}\big[\mathcal{R}(f_{\hat{\theta}})\big] \leqslant \inf_{\|\theta\|_2 \leqslant D} \mathcal{R}(f_\theta) + \frac{4GRD}{\sqrt{n}}.$$

Proof Using Prop. 4.2 to relate the uniform deviation to the Rademacher average, Eq. (4.11) to take care of the Lipschitz-continuous loss, and Eq. (4.12) to account for the ℓ_2-norm constraint, we get the desired result. Note that the factor of 4 comes from symmetrization (Prop. 4.2, which leads to a factor of 2), and Eq. (4.8) in Section 4.4 (which leads to another one). ■

Approximation error. If we assume that there exists a minimizer θ_* of $\mathcal{R}(f_\theta)$ over $\mathbb{R}^d$, the approximation error (of using a ball of θ rather than the whole $\mathbb{R}^d$) is upper-bounded by, following derivations from Section 4.3 (using Cauchy-Schwarz and Jensen's inequalities):

$$\begin{aligned}
\inf_{\|\theta\|_2 \leqslant D} \mathcal{R}(f_\theta) - \mathcal{R}(f_{\theta_*}) &\leqslant G \inf_{\|\theta\|_2 \leqslant D} \mathbb{E}\big[|f_\theta(x) - f_{\theta_*}(x)|\big] \\
&= G \inf_{\|\theta\|_2 \leqslant D} \mathbb{E}\big[|\varphi(x)^\top(\theta - \theta_*)|\big] \\
&\leqslant G \inf_{\|\theta\|_2 \leqslant D} \|\theta - \theta_*\|_2 \mathbb{E}\big[\|\varphi(x)\|_2\big] \leqslant GR \inf_{\|\theta\|_2 \leqslant D} \|\theta - \theta_*\|_2.
\end{aligned}$$

This leads to

$$\mathbb{E}\big[\mathcal{R}(f_{\hat{\theta}})\big] - \mathcal{R}(f_{\theta_*}) \leqslant GR \inf_{\|\theta\|_2 \leqslant D} \|\theta - \theta_*\|_2 + \frac{4GRD}{\sqrt{n}} = GR(\|\theta_*\|_2 - D)_+ + \frac{4GRD}{\sqrt{n}}.$$

We see that for $D = \|\theta_*\|_2$, we obtain the bound $\frac{4GR\|\theta_*\|_2}{\sqrt{n}}$, but this setting requires to know $\|\theta_*\|_2$ which is not possible in practice. If D is too large, the estimation error increases (overfitting). At the same time, if D is too small, the approximation error can quickly kick in (with a value that does not go to zero when n tends to infinity), leading to underfitting. Note that on top of this approximation error, we need to add the incompressible one due to the choice of a linear model.

Exercise 4.13 *We consider a learning problem with 1-Lipschitz-continuous loss (with respect to the second variable), with a function class $f_\theta(x) = \theta^\top \varphi(x)$, with $\|\theta\|_1 \leqslant D$, and $\varphi : \mathcal{X} \to \mathbb{R}^d$ with $\|\varphi(x)\|_\infty$ almost surely less than R. Given the expected risk $\mathcal{R}(f_\theta)$ and the empirical risk $\widehat{\mathcal{R}}(f_\theta)$. Show that $\mathbb{E}\big[\mathcal{R}(f_{\hat{\theta}})\big] \leqslant \inf_{\|\theta\|_1 \leqslant D} \mathcal{R}(f_\theta) + 4RD\sqrt{2\log(2d)/n}$.*

4.5.5 From constrained to regularized estimation (♦)

In practice, it is preferable to penalize by the norm $\Omega(\theta)$ instead of constraining. While the respective sets of solutions when letting the respective constraint and regularization parameters vary are the same, the main reason is that the hyperparameter is easier to find, and the optimization is typically easier. For simplicity, we only consider the ℓ_2-norm in this section.

We now denote $\hat{\theta}_\lambda$ the minimizer of

$$\widehat{\mathcal{R}}(f_\theta) + \frac{\lambda}{2}\|\theta\|_2^2. \tag{4.13}$$

If the loss is always positive, then $\frac{\lambda}{2}\|\hat{\theta}_\lambda\|_2^2 \leqslant \widehat{\mathcal{R}}(f_{\hat{\theta}_\lambda}) + \frac{\lambda}{2}\|\hat{\theta}_\lambda\|_2^2 \leqslant \widehat{\mathcal{R}}(f_0)$, leading to a bound $\|\hat{\theta}_\lambda\|_2 = O(1/\sqrt{\lambda})$. Thus, with $D = O(1/\sqrt{\lambda})$ in the bound above, this leads to a deviation of $O(1/\sqrt{\lambda n})$, which is not optimal.

We now give an interesting stronger result using the strong convexity of the squared ℓ_2-norm (with now a convex loss), adapted from Sridharan et al. (2009); Bartlett et al. (2005).

Proposition 4.6 (Fast rates for regularized objectives) *Assume the loss function is G-Lipschitz-continuous and **convex**, with linear prediction functions defined by $f_\theta(x) = \theta^\top \varphi(x)$, for $\theta \in \mathbb{R}^d$, where $\|\varphi(x)\|_2 \leqslant R$ almost surely. Let $\hat{\theta}_\lambda \in \mathbb{R}^d$ be the minimizer of the regularized empirical risk in Eq. (4.13), then:*

$$\mathbb{E}\big[\mathcal{R}(f_{\hat{\theta}_\lambda})\big] \leqslant \inf_{\theta \in \mathbb{R}^d} \big\{\mathcal{R}(f_\theta) + \frac{\lambda}{2}\|\theta\|_2^2\big\} + \frac{32 G^2 R^2}{\lambda n}.$$

Proof (♦) Let $\mathcal{R}_\lambda(f_\theta) = \mathcal{R}(f_\theta) + \frac{\lambda}{2}\|\theta\|_2^2$, with minimum value $\mathcal{R}_\lambda^*$ attained at θ_λ^*. We consider the convex set $\mathcal{C}_\varepsilon = \{\theta \in \mathbb{R}^d,\ \mathcal{R}_\lambda(f_\theta) - \mathcal{R}_\lambda^* \leqslant \varepsilon\}$ for an $\varepsilon > 0$ to be chosen later. If $\hat{\theta}_\lambda \notin \mathcal{C}_\varepsilon$, then, by convexity, there has to be an η in the segment $[\theta_\lambda^*, \hat{\theta}_\lambda]$ such that $\mathcal{R}_\lambda(f_\eta) - \mathcal{R}_\lambda^* = \mathcal{R}_\lambda(f_\eta) - \mathcal{R}_\lambda(f_{\theta_\lambda^*}) = \varepsilon$, and $\widehat{\mathcal{R}}_\lambda(f_\eta) \leqslant \widehat{\mathcal{R}}_\lambda(f_{\theta_\lambda^*})$.[7] This implies that

$$\mathcal{R}_\lambda(f_\eta) - \widehat{\mathcal{R}}_\lambda(f_\eta) + \widehat{\mathcal{R}}_\lambda(f_{\theta_\lambda^*}) - \mathcal{R}_\lambda(f_{\theta_\lambda^*}) = \mathcal{R}_\lambda(f_\eta) - \mathcal{R}_\lambda(f_{\theta_\lambda^*}) + \widehat{\mathcal{R}}_\lambda(f_{\theta_\lambda^*}) - \widehat{\mathcal{R}}_\lambda(f_\eta) \geqslant \varepsilon. \tag{4.14}$$

[7] This can be shown by taking η at the intersection of the segment $[\theta_\lambda^*, \hat{\theta}_\lambda]$ and the set $\partial\mathcal{C}_\varepsilon$ (the boundary of $\mathcal{C}_\varepsilon$).

By strong convexity, we have, $\mathcal{R}_\lambda(f_\theta) - \mathcal{R}_\lambda^* \geqslant \frac{\lambda}{2}\|\theta - \theta_\lambda^*\|_2^2$ for all θ, and thus $\mathcal{C}_\varepsilon$ is included in the ℓ_2-ball of center θ_λ^* and radius $\sqrt{2\varepsilon/\lambda}$. Thus, from Eq. (4.14), we get $\sup_{\|\eta-\theta_\lambda^*\|_2 \leqslant \sqrt{2\varepsilon/\lambda}} \Big\{ \mathcal{R}_\lambda(f_\eta) - \mathcal{R}_\lambda(f_{\theta_\lambda^*}) - \big[\widehat{\mathcal{R}}_\lambda(f_\eta) - \widehat{\mathcal{R}}_\lambda(f_{\theta_\lambda^*})\big] \Big\} \geqslant \varepsilon$. Using Section 4.5.3, we have

$$\begin{aligned} & \mathbb{E}\Big[\sup_{\|\eta-\theta_\lambda^*\|_2 \leqslant \sqrt{2\varepsilon/\lambda}} \Big\{ \mathcal{R}_\lambda(f_\eta) - \mathcal{R}_\lambda(f_{\theta_\lambda^*}) - \big[\widehat{\mathcal{R}}_\lambda(f_\eta) - \widehat{\mathcal{R}}_\lambda(f_{\theta_\lambda^*})\big] \Big\} \Big] \\ \leqslant \ & 2\mathbb{E}\Big[\sup_{\|\eta-\theta_\lambda^*\|_2 \leqslant \sqrt{2\varepsilon/\lambda}} \Big\{ \frac{1}{n}\sum_{i=1}^n \varepsilon_i[\ell(y_i, \varphi(x_i)^\top \eta) - \ell(y_i, \varphi(x_i)^\top \theta_\lambda^*)] \Big\} \Big] \leqslant 2GR\sqrt{2\varepsilon/\lambda}. \end{aligned}$$

Moreover, by McDiarmid's inequality,

$$\mathbb{P}\Big(\mathcal{R}_\lambda(f_\eta) - \widehat{\mathcal{R}}_\lambda(f_\eta) + \widehat{\mathcal{R}}_\lambda(f_{\theta_\lambda^*}) - \mathcal{R}_\lambda(f_{\theta_\lambda^*}) \geqslant 2GR\sqrt{2\varepsilon/\lambda} + t\frac{2\frac{GR}{\sqrt{n}}\sqrt{2\varepsilon/\lambda}}{\sqrt{2n}} \Big) \leqslant e^{-t^2}.$$

Thus, if $\varepsilon \geqslant 2\frac{GR}{\sqrt{n}}\sqrt{2\varepsilon/\lambda}(1 + \frac{t}{\sqrt{2}})$, that is, if $\varepsilon \geqslant 8\frac{G^2R^2}{\lambda n}(2+t^2)$, we have the high probability bound $\mathbb{P}\big(\mathcal{R}_\lambda(f_{\hat{\theta}_\lambda}) - \mathcal{R}_\lambda^* > \varepsilon\big) \leqslant e^{-t^2}$. This leads to, by integration, the final bound $\mathbb{E}\big[\mathcal{R}_\lambda(f_{\hat{\theta}_\lambda}) - \mathcal{R}_\lambda^*\big] \leqslant \frac{32G^2R^2}{\lambda n}$. ■

Note that we obtain a "fast rate" in $O(R^2/(\lambda n))$, which has a better dependence in n but depends on λ, which can be very small in practice. One classical choice of λ that we have seen in Chapter 3 also applies here, as $\lambda \propto \frac{GR}{\sqrt{n}\|\theta_*\|}$, leading to the slow rate

$$\mathbb{E}\big[\mathcal{R}(f_{\hat{\theta}_\lambda})\big] \leqslant \mathcal{R}(f_{\theta_*}) + O\Big(\frac{GR}{\sqrt{n}}\|\theta_*\|_2\Big).$$

This result is similar to the one obtained in Chapter 3 for ridge (least-squares) regression, but now for all Lipschitz-continuous losses. Note that the amount of regularization to get the result above still depends on the unknown quantity $\|\theta_*\|_2$. Below, we consider the general case of penalization by a norm, where we will obtain similar results but with a hyperparameter that does not depend on the unknown norm of $\|\theta_*\|_2$.

Exercise 4.14 *(♦♦) Extend the result in Prop. 4.6 to features that are almost surely bounded in ℓ_p-norm by R, and a regularizer ψ which is strongly-convex with respect to the ℓ_p-norm, that is, such that for all $\theta, \eta \in \mathbb{R}^d$, $\psi(\theta) \geqslant \psi(\eta) + \psi'(\eta)^\top(\theta - \eta) + \frac{\mu}{2}\|\theta - \eta\|_p^2$, where $\psi'(\eta)$ is a subgradient of ψ at η.*

Norm-penalized estimation. (♦♦) While Proposition 4.6 considered squared ℓ_2-norm penalization and relied on specific properties of the ℓ_2-norm, we now consider penalization by *any* (non-squared) norm. That is, we now focus on the following objective function:

$$\widehat{\mathcal{R}}_\lambda(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, \theta^\top \varphi(x_i)) + \lambda\Omega(\theta), \tag{4.15}$$

for any norm Ω on $\mathbb{R}^d$, with Ω^* denoting the dual norm. The following proposition provides an estimation rate in $O(1/\sqrt{n})$.

Proposition 4.7 (Norm-penalized estimation) *Assume that the unregularized risk $\mathcal{R}_0(\theta) = \mathbb{E}_{p(x,y)}\big[\ell(y, \theta^\top \varphi(x))\big]$ is minimized at some $\theta_* \in \mathbb{R}^d$, and that the function $\theta \mapsto \ell(y, \theta^\top \varphi(x))$ is GR-Lipschitz continuous in θ for $\Omega(\theta) \leqslant 2\Omega(\theta_*)$, and that $\Omega^*(\varphi(x)) \leqslant R$ almost surely. Denote $\rho_\Omega = \sup_{\Omega^*(z_1),\dots,\Omega^*(z_n) \leqslant 1} \mathbb{E}_\varepsilon\big[\Omega^*\big(\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i z_i\big)\big]$ where $\varepsilon \in \{-1,1\}^n$ is a vector of independent Rademacher random variables. For any $\delta \in (0,1)$, and for $\lambda = \frac{GR}{\sqrt{n}}(\rho_\Omega + 4\sqrt{2\log\frac{1}{\delta}})$, with probability at least $1-\delta$, any minimizer $\hat\theta_\lambda$ of Eq. (4.15) satisfies:*

$$\mathcal{R}(\hat\theta_\lambda) \leqslant \mathcal{R}(\theta_*) + \Omega(\theta_*)\frac{3GR}{\sqrt{n}}\Big(\rho_\Omega + 4\sqrt{2\log\frac{1}{\delta}}\Big).$$

Proof We consider θ_λ^* a minimizer of the population regularized risk $\mathcal{R}_\lambda(\theta) = \mathcal{R}(\theta) + \lambda\Omega(\theta)$. It satisfies $\Omega(\theta_\lambda^*) \leqslant \Omega(\theta_*)$. Moreover, ρ_Ω is such that the Rademacher complexity of the set of linear predictors such that $\Omega(\theta) \leqslant D$ for $D \leqslant 2\Omega(\theta_*)$, is less than $\frac{\rho_\Omega GRD}{\sqrt{n}}$ (see Section 4.5.3). For example, for the ℓ_2-norm, we have $\rho_\Omega = 1$, while for the ℓ_1-norm, we have $\rho_\Omega = \sqrt{2\log(2d)}$. In terms of losses, for the logistic loss, we have $G = 1$, while for the square loss (with a factor of 1/2) with a model $y = \varphi(x)^\top\theta^* + \varepsilon$ with $|\varepsilon| \leqslant \sigma$ almost surely, we get $G = \sigma + 3R\Omega(\theta^*)$.

Using McDiarmid's inequality like in Section 4.4.1, by fixing any θ_0 such that $\Omega(\theta_0) \leqslant D$, with probability greater than $1-e^{-t^2}$, for all θ such that $\Omega(\theta) \leqslant 2\Omega(\theta_*)$, $\mathcal{R}(\theta) - \mathcal{R}(\theta_0) \leqslant \widehat{\mathcal{R}}(\theta) - \widehat{\mathcal{R}}(\theta_0) + \frac{\rho_\Omega GRD}{\sqrt{n}} + t\frac{2GRD\sqrt{2}}{\sqrt{n}}$.

We consider the set $\mathcal{C}_{\nu,\varepsilon} = \big\{\theta \in \mathbb{R}^d,\ \Omega(\theta) \leqslant 2\Omega(\theta_\lambda^*),\ \mathcal{R}_\lambda(\theta) - \mathcal{R}_\lambda(\theta_\lambda^*) \leqslant \varepsilon\big\}$. This is a convex set, with boundary $\partial\mathcal{C}_{\nu,\varepsilon} = \big\{\theta \in \mathbb{R}^d,\ \Omega(\theta) \leqslant 2\Omega(\theta_\lambda^*),\ \mathcal{R}_\lambda(\theta) - \mathcal{R}_\lambda(\theta_\lambda^*) = \varepsilon\big\}$ for a well-chosen ε (that is, the saturated constraint has to be one on the expected risk). Indeed, if $\Omega(\theta) = 2\Omega(\theta_\lambda^*)$, then, using that the optimality conditions for θ_λ^* implies that $\Omega^*(\mathcal{R}'(\theta_\lambda^*)) \leqslant \lambda$:

$$\begin{aligned}
\mathcal{R}_\lambda(\theta) - \mathcal{R}_\lambda(\theta_\lambda^*) &= \mathcal{R}(\theta) - \mathcal{R}(\theta_\lambda^*) + \lambda\Omega(\theta) - \lambda\Omega(\theta_\lambda^*) \text{ by definition,}\\
&\geqslant \mathcal{R}'(\theta_\lambda^*)^\top(\theta - \theta_\lambda^*) + \lambda\Omega(\theta) - \lambda\Omega(\theta_\lambda^*) \text{ by convexity,}\\
&\geqslant -\Omega^*(\mathcal{R}'(\theta_\lambda^*)) \cdot \Omega(\theta - \theta_\lambda^*) + \lambda\Omega(\theta) - \lambda\Omega(\theta_\lambda^*)\\
&\qquad\qquad\qquad\qquad \text{by definition of the dual norm,}\\
&\geqslant -\lambda\Omega(\theta - \theta_\lambda^*) + \lambda\Omega(\theta) - \lambda\Omega(\theta_\lambda^*) \text{ by optimality of } \theta_\lambda^*,\\
&\geqslant 2\lambda\Omega(\theta) - 2\lambda\Omega(\theta_\lambda^*) \text{ by the triangular inequality,}\\
&= 2\lambda\Omega(\theta_\lambda^*) \text{ since we have assumed } \Omega(\theta) = 2\Omega(\theta_\lambda^*).
\end{aligned}$$

Thus, we must impose that $\varepsilon \leqslant 2\Omega(\theta_\lambda^*)$.

We now show that with high probability, we must have $\hat\theta_\lambda \in \mathcal{C}_{\nu,\varepsilon}$. If $\hat\theta_\lambda \notin \mathcal{C}_{\nu,\varepsilon}$, since $\theta_\lambda^* \in \mathcal{C}_{\nu,\varepsilon}$, there has to be an element θ in the segment $[\theta_\lambda^*, \hat\theta_\lambda]$ which is in $\partial\mathcal{C}_{\nu,\varepsilon}$. Since our risks are convex, we have $\widehat{\mathcal{R}}_\lambda(\theta) \leqslant \max\{\widehat{\mathcal{R}}_\lambda(\theta^*_\lambda), \widehat{\mathcal{R}}_\lambda(\hat{\theta}_\lambda)\} = \widehat{\mathcal{R}}_\lambda(\theta^*_\lambda)$. Thus

$$\widehat{\mathcal{R}}(\theta^*_\lambda) - \widehat{\mathcal{R}}(\theta) - \mathcal{R}(\theta^*_\lambda) + \mathcal{R}(\theta) = \widehat{\mathcal{R}}_\lambda(\theta^*_\lambda) - \widehat{\mathcal{R}}_\lambda(\theta) - \mathcal{R}_\lambda(\theta^*_\lambda) + \mathcal{R}_\lambda(\theta) \geqslant -\mathcal{R}_\lambda(\theta^*_\lambda) + \mathcal{R}_\lambda(\theta) = \varepsilon.$$

Thus if we take, $\varepsilon \geqslant \frac{\rho_\Omega GRD}{\sqrt{n}} + t\frac{2GRD\sqrt{2}}{\sqrt{n}}$, with $D = 2\Omega(\theta^*_\lambda)$, this can only happen with probability less than $\exp(-t^2)$. This leads to the constraint $\varepsilon \geqslant \frac{2GR\Omega(\theta^*_\lambda)}{\sqrt{n}}(\rho_\Omega + 4t\sqrt{2})$. Thus, we can take $\lambda = \frac{GR}{\sqrt{n}}(\rho_\Omega + 4t\sqrt{2})$, and with probability greater than $1 - e^{-t^2}$ we have

$$\mathcal{R}_\lambda(\hat{\theta}_\lambda) - \mathcal{R}_\lambda(\theta^*_\lambda) \leqslant 2\lambda\Omega(\theta^*_\lambda) \leqslant 2\lambda\Omega(\theta^*).$$

Overall, denoting $\delta = e^{-t^2}$, we get that with probability greater than $1 - \delta$

$$\mathcal{R}(\hat{\theta}_\lambda) \leqslant \mathcal{R}(\theta_*) + \Omega(\theta_*)\frac{3GR}{\sqrt{n}}\Big(\rho_\Omega + 4\sqrt{2\log\frac{1}{\delta}}\Big).$$

for $\lambda = \frac{GR}{\sqrt{n}}(\rho_\Omega + 4\sqrt{2\log\frac{1}{\delta}})$. Note that we could get a result in expectation (left as an exercise). The key here is that the value of λ does not depend on $\Omega(\theta^*)$. ■

4.5.6 Extensions and improvements

In this chapter, we have focused on the simplest situations for empirical risk minimization technique: regression or binary classification with i.i.d. data. Statistical learning theory investigates many more complex cases along several lines:

- **Slower rates than $1/\sqrt{n}$:** In this chapter, we primarily studied the estimation error that decays as $1/\sqrt{n}$. When balancing it with approximation error (by adapting norm constraints or regularization parameters), we will obtain slower rates, but with weaker assumptions, in Chapter 7 (kernel methods) and Chapter 9 (neural networks).

- **Faster rates with discrete outputs:** Further analysis can be carried through when dealing with binary classification, or more generally discrete outputs, with potentially different convergence rates for the convex surrogate and the original loss function (i.e., after thresholding, where sometimes exponential rates can be obtained). This is often done under so-called "low noise" conditions (see, e.g., Koltchinskii and Beznosova, 2005; Audibert and Tsybakov, 2007), as briefly exposed in Section 4.1.4.

- **Other generic learning theory frameworks:** In this chapter, we have focused primarily on the tools of Rademacher averages to obtain generic learning bounds. Other frameworks lead to similar bounds but from different mathematical perspectives. For example, PAC-Bayesian analysis (Catoni, 2007; Zhang, 2006) is described in Section 14.4, while stability-based arguments (Bousquet and Elisseeff, 2002) lead to similar results (see exercise below).

Exercise 4.15 *(♦) We consider a learning algorithm and a distribution p on (x, y) such that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and two outputs $f, f' : \mathcal{X} \to \mathcal{Y}$ of the learning algorithm on datasets of n observations which differ by a single observation, $|\ell(y, f(x)) - \ell(y, f'(x))| \leqslant \beta_n$, an assumption referred to as "uniform stability". Show that the expected deviation between the expected risk and the empirical risk of the algorithm's output is bounded by β_n. With the same assumptions as in Prop. 4.6, show that we have $\beta_n = \frac{2G^2R^2}{\lambda n}$ (see Bousquet and Elisseeff, 2002, for more details).*

- **Beyond independent observations:** Much of statistical learning theory deals with the simplifying assumptions that observations are i.i.d. from the same distribution as the one used during the testing phase. This leads to the reasonably simple results presented in this chapter. Several lines of work deal with situations when data are not independent: among them, online learning presented in Chapter 11 shows that many classical algorithms are indeed robust to such dependence. Another avenue coming from statistics is to make some assumptions on the dependence between observations, the most classical one being that the sequence of observations $(x_i, y_i)_{i \geqslant 1}$ form a Markov chain, and thus satisfies "mixing conditions" (see, e.g., Mohri and Rostamizadeh, 2010).
- **Mismatch between training and testing distributions:** In many application scenarios, the testing distribution may deviate from the training distribution: the input distribution of x may be different while the conditional distribution of y given x remains the same, a situation commonly referred to as "covariate shift", or the entire distribution of (x, y) may deviate (often referred to as the need for "domain adaptation"). If no assumption is made on the proximity of these two distributions, no guarantee can be obtained. Several ideas have been explored to derive algorithms and/or guarantees, such as importance reweighting (Sugiyama et al., 2007) or finding projections of the data with similar test and train distributions (Ganin et al., 2016).
- **Semi-supervised learning:** In many applications, many unlabelled observations are available (that is, only with the input x being available). To leverage the abundance of unlabelled data, some assumptions are typically made to show an improvement of learning algorithms, such as the "cluster assumption" (points in the same class tend to cluster together) or "low-density separation" (for classification, decision boundaries tend to be in regions with few input observations). Many algorithms exist, such as Laplacian regularization (see Cabannes et al., 2021, and references therein) or discriminative clustering (Xu et al., 2004; Bach and Harchaoui, 2007).

4.6 Model selection (♦)

Throughout this chapter, we have considered a family $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathcal{Y}$ and have obtained generalization bounds for the minimizer $\hat{f} \in \mathcal{F}$ of the empirical risk $\widehat{\mathcal{R}}$. Assuming that the loss function $\ell(y, f(x))$ is almost surely in $[-\ell_\infty, \ell_\infty]$, we have obtained in Section 4.4.1, together with Rademacher complexities in Section 4.5, a bound of the form

$$\sup_{f \in \mathcal{F}} \big| \mathcal{R}(f) - \widehat{\mathcal{R}}(f) \big| \leqslant 2\mathrm{R}_n(\mathcal{H}) + \frac{\ell_\infty}{\sqrt{n}} \sqrt{2 \log \frac{2}{\delta}}$$

with probability greater than $1-\delta$, with the Rademacher complexity $\mathrm{R}_n(\mathcal{H})$ of the class of functions $\mathcal{H} = \{(x,y) \mapsto \ell(y, f(x)),\ f \in \mathcal{F}\}$. From Eq. (4.8), this leads to, with probability greater than $1-\delta$:

$$\mathcal{R}(\hat{f}) \leqslant \widehat{\mathcal{R}}(\hat{f}) + 2\mathrm{R}_n(\mathcal{H}) + \frac{\ell_\infty}{\sqrt{n}} \sqrt{2 \log \frac{2}{\delta}}, \tag{4.16}$$

which is a data-dependent generalization bound in high probability. Moreover, at the end of Section 4.5.1, we have seen that we could use the empirical Rademacher complexity, which can be more easily computed (with fewer assumptions).

We now consider a finite (but potentially large) number m of models $\mathcal{F}_1, \dots, \mathcal{F}_m$, together with their associated loss function spaces $\mathcal{H}_1, \dots, \mathcal{H}_m$ and their generalization bounds for the empirical risk minimizer based on Rademacher complexities. In this section, we consider how to choose the best corresponding empirical risk minimizer among $\hat{f}_1, \dots, \hat{f}_m$. We consider two approaches, either based on minimizing a penalized data-generalization bound (also referred to as structural risk minimization) or simply using a validation set.

In both cases, we consider a set of positive weights $\pi_1, \dots, \pi_m$ that sum to one. We can typically choose $\pi_i = 1/m$ for all $i \in \{1, \dots, m\}$, we can also consider other choices, in particular when m gets large and we are willing to put more prior weight on certain models.

4.6.1 Structural risk minimization

We minimize the data-dependent generalization bounds plus an additional parameter to take into account the prior on models, that is,

$$\hat{\imath} = \arg \min_{i \in \{1, \dots, m\}} \Big\{ \widehat{\mathcal{R}}(\hat{f}_i) + 2\mathrm{R}_n(\mathcal{H}_i) + \frac{2\ell_\infty}{\sqrt{n}} \sqrt{2 \log(1/\pi_i)} \Big\}.$$

We can then use Eq. (4.16) for each of the m models, and with $\pi_i \delta$ instead of δ, use the union bound to get, with probability greater than $1-\delta$:

$$\mathcal{R}(\hat{f}_{\hat{\imath}}) \quad \leqslant \quad \min_{i \in \{1, \dots, m\}} \Big\{ \inf_{f_i \in \mathcal{F}_i} \mathcal{R}(f_i) + 2\mathrm{R}_n(\mathcal{H}_i) + \frac{2\ell_\infty}{\sqrt{n}} \sqrt{2 \log(1/\pi_i)} \Big\} + \frac{2\ell_\infty}{\sqrt{n}} \sqrt{2 \log(2/\delta)}.$$

For example, when $\pi_1 = \cdots = \pi_m = 1/m$, we thus obtain that the model selection procedure pays an extra price of $\sqrt{\log(m)/n}$ on top of the individual generalization bounds.

4.6.2 Selection based on validation set

We assume here that we have kept a proportion $\rho \in (0,1)$ of the training data as a validation set (assuming for simplicity that ρn is an integer). We then have an empirical risk based on $(1-\rho)n$ observations, we which now denote $\widehat{\mathcal{R}}^{(\text{training})}_{(1-\rho)n}$, and a validation empirical risk denoted $\widehat{\mathcal{R}}^{(\text{validation})}_{\rho n}$. Given the m minimizers $\hat{f}_i$ of the training empirical risks $\widehat{\mathcal{R}}^{(\text{training})}_{(1-\rho)n}(f_i)$ over $f_i \in \mathcal{F}_i$, for $i \in \{1,\dots,m\}$, we choose $\hat{\imath}$ that minimizes the following criterion

$$\widehat{\mathcal{R}}^{(\text{validation})}_{\rho n}(\hat{f}_i) + \frac{2\ell_\infty}{\sqrt{\rho n}}\sqrt{2\log(1/\pi_i)}$$

(for uniform weights $\pi_1 = \cdots = \pi_m = 1/m$, this is simply the minimizer of the validation risk). We can then simply use Hoeffding's inequality and the union bound to get the generalization bound:

$$\mathcal{R}(\hat{f}_{\hat{\imath}}) \leqslant \min_{i\in\{1,\dots,m\}} \mathcal{R}(\hat{f}_i) + \frac{2\ell_\infty}{\sqrt{\rho n}}\sqrt{2\log(1/\pi_i)} + \frac{2\ell_\infty}{\sqrt{\rho n}}\sqrt{2\log(2/\delta)},$$

which shows an extra price proportional to $1/\sqrt{\rho n}$, highlighting the fact that the validation set proportion $\rho \in (0,1)$ should not be too small. We can also obtain a result similar to the one in the section above by using the same generalization bounds based on Rademacher averages, that is,

$$\begin{aligned}\mathcal{R}(\hat{f}_{\hat{\imath}}) \;\; &\leqslant \;\; \min_{i\in\{1,\dots,m\}} \Big\{ \inf_{f_i\in\mathcal{F}_i} \mathcal{R}(f_i) + 2\mathrm{R}_{(1-\rho)n}(\mathcal{H}_i) + \frac{2\ell_\infty}{\sqrt{\rho n}}\sqrt{2\log(1/\pi_i)}\Big\} \\ &\qquad + \frac{2\ell_\infty}{\sqrt{\rho n}}\sqrt{2\log(2/\delta)} + \frac{2\ell_\infty}{\sqrt{(1-\rho)n}}\sqrt{2\log(2/\delta)}.\end{aligned}$$

Suppose ρ is bounded away from 0 and 1. In that case, we obtain a slightly worse bound than when using data-dependent bounds, with the difference that the performance of validation methods is, in practice, much better than the bound guarantees (while data-dependent bound optimization may not exhibit such adaptivity).

4.7 Relationship with asymptotic statistics (♦)

In this last section, we will relate the non-asymptotic analysis presented in this chapter to results from asymptotic statistics (see the comprehensive book by Van der Vaart (2000), which presents this large literature).

To make this concrete, we assume that we have a set of models $\mathcal{F} = \{f_\theta : \mathcal{X} \to \mathbb{R},\ \theta \in \mathbb{R}^d\}$ parameterized by a vector $\theta \in \mathbb{R}^d$. We consider the empirical risk and expected risks (with a slight overloading of notations):

$$\mathcal{R}(\theta) = \mathcal{R}(f_\theta) = \mathbb{E}\big[\ell(y, f_\theta(x))\big] \;\text{ and }\; \widehat{\mathcal{R}}(\theta) = \widehat{\mathcal{R}}(f_\theta) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, f_\theta(x_i)).$$

We assume that we have a loss function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ (such as for regression or any of the convex surrogates for classification), which is sufficiently differentiable with respect to the second variable, so that results from Van der Vaart (2000) apply (e.g., Theorems 5.21 or 5.41 on "M-estimation", which cover empirical risk minimization). In this section, we will only report their final result and provide an intuitive justification.

We assume that $\theta_* \in \mathbb{R}^d$ is a minimizer of $\mathcal{R}(\theta)$ and that the Hessian $\mathcal{R}''(\theta_*)$ is positive-definite (it has to be positive semi-definite as θ_* is a minimizer, we assume invertibility on top of it).

We let $\hat{\theta}_n$ denote a minimizer of $\widehat{\mathcal{R}}$. Since $\mathcal{R}'(\theta_*) = 0$, and $\widehat{\mathcal{R}}'(\theta_*) = \frac{1}{n}\sum_{i=1}^n \frac{\partial \ell(y, f_\theta(x))}{\partial \theta}$, by the law of large numbers, $\widehat{\mathcal{R}}'(\theta_*)$ tends to $\mathcal{R}'(\theta_*) = 0$ (e.g., almost surely), and we should thus expect that $\hat{\theta}_n$ (which is defined through $\widehat{\mathcal{R}}'(\hat{\theta}_n) = 0$) tends to θ_* (all these statements can be made rigorous, see Van der Vaart (2000)).

Then, a Taylor expansion of $\widehat{\mathcal{R}}'$ around θ_* leads to

$$0 = \widehat{\mathcal{R}}'(\hat{\theta}_n) \approx \widehat{\mathcal{R}}'(\theta_*) + \widehat{\mathcal{R}}''(\theta_*)(\hat{\theta}_n - \theta_*).$$

By the law or large numbers, $\widehat{\mathcal{R}}''(\theta_*)$ tends to $H(\theta_*) = \mathcal{R}''(\theta_*)$ when n tends to infinity, and thus we obtain:

$$\hat{\theta}_n - \theta_* \approx \mathcal{R}''(\theta_*)^{-1}\widehat{\mathcal{R}}'(\theta_*) = H(\theta_*)^{-1}\widehat{\mathcal{R}}'(\theta_*).$$

Moreover, $\widehat{\mathcal{R}}'(\theta_*)$ is the average of n i.i.d. random vectors, and by the central limit theorem, it is asymptotically Gaussian with mean zero and covariance matrix equal to $\frac{1}{n}G(\theta_*) = \frac{1}{n}\mathbb{E}\big[\big(\frac{\partial \ell(y, f_\theta(x))}{\partial \theta}\big)\big(\frac{\partial \ell(y, f_\theta(x))}{\partial \theta}\big)^\top\big|_{\theta=\theta_*}\big]$. Therefore, we (intuitively) obtain that $\hat{\theta}_n$ is asymptotically Gaussian with mean θ_* and covariance matrix $\frac{1}{n}H(\theta_*)^{-1}G(\theta_*)H(\theta_*)^{-1}$.

This asymptotic result has the nice consequence that:

$$\begin{aligned}
\mathbb{E}\big[\|\hat{\theta}_n - \theta_*\|_2^2\big] &\sim \frac{1}{n}\operatorname{tr}\big[H(\theta_*)^{-1}G(\theta_*)H(\theta_*)^{-1}\big] \\
\mathbb{E}\big[\mathcal{R}(\hat{\theta}_n) - \mathcal{R}(\theta_*)\big] &\sim \frac{1}{n}\operatorname{tr}\big[H(\theta_*)^{-1}G(\theta_*)\big].
\end{aligned}$$

For example, for well-specified linear regression (like analyzed in Chapter 3), it turns out that we have $G(\theta_*) = \sigma^2 H(\theta_*)$ (proof left as an exercise), and thus we recover the rate $\sigma^2 d/n$.

Benefits of the asymptotic analysis. As shown above, the asymptotic analysis gives a precise picture of the asymptotic behavior of empirical risk minimization. Much more than simply providing an upper-bound on $\mathbb{E}\big[\mathcal{R}(\hat{\theta}_n) - \mathcal{R}(\theta_*)\big]$, it gives also a limit Gaussian distribution for $\hat{\theta}_n$, and a *fast rate* as $O(1/n)$. Moreover, because we have limits, we can compare limits between various learning algorithms and claim (asymptotic) superiority or inferiority of one method over another, which comparing upper bounds cannot achieve.

Thus, an asymptotic analysis does not suffer from the traditional looseness of non-asymptotic bounds that rely on crude approximations, and, while they are valid even for small n and often exhibit the desired behavior of n, are overly pessimistic.

Pitfalls of the asymptotic analysis. The main drawback of this analysis is that it is... asymptotic. That is, n tends to infinity, and it is impossible to tell without further analysis when the asymptotic behavior will kick in. Sometimes, this is for reasonably small n, sometimes for large n. Further asymptotic expansions can be carried out, but small sample effects are hard to characterize, particularly when the underlying dimension d gets large.

Bridging the gap. Studying the validity of the asymptotic expansion described above can be done in several ways. See, e.g., Ostrovskii and Bach (2021a) (and references therein) for finite-dimensional models, and Chapter 7 for results similar to $\sigma^2 d/n$ when the dimension of the feature space gets infinite.

4.8 Summary

In this chapter, we have first introduced convex surrogates for binary classification problems to avoid performing optimization on functions with values in $\{-1, 1\}$. This comes with generalization guarantees that will be extended in Chapter 13 to multiple categories and, more generally, to structured output spaces.

The chapter's core was dedicated to introducing Rademacher complexities, which are flexible tools to study estimation errors in many settings. This led to simple bounds for linear models and ball constraints, which will be extended to infinite-dimensional settings in Chapter 7 and neural networks in Chapter 9. Other frameworks exist to obtain similar bounds, such as the PAC-Bayes framework presented in Section 14.4, often leading to tighter bounds.

While the present chapter was dedicated to the statistical analysis of empirical risk minimizers, the next chapter is dedicated to optimization algorithms traditionally dedicated to finding approximate such minimizers, with notably stochastic gradient descent, which also naturally exhibits good generalization performance.