Chapter 15

Lower bounds on performance

Chapter summary

- Statistical lower bounds: for least-squares regression, the optimal performance of supervised learning with target functions that are linear in some feature vector or in Sobolev spaces on $\mathbb{R}^d$ happens to be achieved by several algorithms presented earlier in the book. The lower bounds can be obtained through information theory or Bayesian analysis.
- Optimization lower bounds: for the classical problem classes from Chapter 5, hard functions can be designed so that gradient-descent-based algorithms that linearly combine gradients are shown to be optimal.
- Lower bounds for stochastic gradient descent: The rates proportional to $O(1/\sqrt{n})$ for convex functions and $O(1/n\mu)$ for μ-strongly convex problems are optimal.

In this textbook, we have shown various convergence rates for statistical procedures when the number of observations n goes to infinity, and optimization methods, as the number of iterations t goes to infinity. Most were non-asymptotic upper bounds on the error measures, with a precise dependence on the problem parameters (e.g., smoothness of the target function or the objective function).

In this chapter, we are looking at lower bounds on performance, that is, we aim to show that for a particular problem class and a specific class of algorithms, the error measures cannot go to zero too quickly. Lower bounds are useful, in particular when they match upper bounds up to constants (we can then claim that we have an "optimal" method). They sometimes provide hard problems (like for optimization), sometimes not (when they are based on information theory, such as for prediction performance).

> Lower bounds will be obtained in a "minimax" setting where we look at the worst-case performance over the entire problem class. As for upper bounds, looking at worst-case performance is, in essence, pessimistic, and algorithms often behave better than their bounds. The key is to identify classes of problems that are not too large (or the bounds will be very bad) but still contain interesting problems.

The chapter is naturally divided into three sections: Section 15.1 considers statistical lower bounds, Section 15.2 considers optimization lower bounds, while Section 15.3 considers lower bounds for stochastic gradient methods. All of these provide bounds related to the setups encountered in earlier chapters.

15.1 Statistical lower bounds

In this section, our goal is to obtain lower bounds for regression problems in $\mathbb{R}^d$ with the square loss when assuming the target function $f^* : \mathbb{R}^d \to \mathbb{R}$ (here the conditional expectation of y given x) is in a particular set, such as:

- linear function of some d-dimensional features, that is, $f_*(x) = \langle \theta_*, \varphi(x) \rangle$, for $\theta_* \in \mathbb{R}^d$, potentially in a ℓ_2-ball, and/or with less than k non-zero elements,
- functions with all partial derivatives up to order s bounded in L_2-norm (e.g., Sobolev spaces).

Since we are looking for lower bounds, we are free to make extra assumptions (that can only make the problem simpler) and lower the lower bounds. For example, we will focus on Gaussian noise with constant variance σ^2 that is independent of x.

We can either consider fixed design assumptions or random designs with the simplest input distributions (that can only make the problem simpler).

Classification. Lower bounds for classification problems are more delicate and out of scope (see, e.g., Yang, 1999). However, we can get lower bounds for the convex surrogates that are typically used (but note that this does not translate to lower bounds for the 0-1 loss), see for example Section 15.3 for Lipschitz-continuous loss functions.

15.1.1 Minimax lower bounds

We consider a set of probability distributions indexed by some set Θ (that can characterize input distributions and the smoothness of the target function). We consider some data $\mathcal{D}$, generated from this distribution, and we denote $\mathbb{E}_\theta$ expectations with respect to data coming from the distribution indexed by θ.

We consider an estimator $\mathcal{A}(\mathcal{D})$ of $\theta \in \Theta$, with some squared distance d^2 between two elements of Θ, so that $d(\theta, \theta')^2$ measures the performance of θ' when the true estimator is θ. The performance of $\mathcal{A}$ when the data come from θ_* is

$$\mathbb{E}_{\theta_*}\big[d(\theta_*, \mathcal{A}(\mathcal{D}))^2\big].$$

The goal is to find an algorithm so that $\sup_{\theta_* \in \Theta} \mathbb{E}_{\theta_*}\big[d(\theta_*, \mathcal{A}(\mathcal{D}))^2\big]$ is as small as possible, and the lower bound of performance is thus:

$$\inf_{\mathcal{A}} \sup_{\theta_* \in \Theta} \mathbb{E}_{\theta_*}\big[d(\theta_*, \mathcal{A}(\mathcal{D}))^2\big]. \tag{15.1}$$

This is often referred to as "minimax" lower bounds.

Since by Markov's inequality, $\mathbb{E}_{\theta_*}\big[d(\theta_*, \mathcal{A}(\mathcal{D}))^2\big] \geqslant A\,\mathbb{P}_{\theta_*}\big(d(\theta_*, \mathcal{A}(\mathcal{D}))^2 > A\big)$, it is sufficient to lower bound

$$\inf_{\mathcal{A}} \sup_{\theta_* \in \Theta} \mathbb{P}_{\theta_*}\big(d(\theta_*, \mathcal{A}(\mathcal{D}))^2 > A\big), \tag{15.2}$$

for some arbitrary $A > 0$. This will be useful for techniques based on information theory.

We will see two principles for obtaining statistical minimax lower bounds:

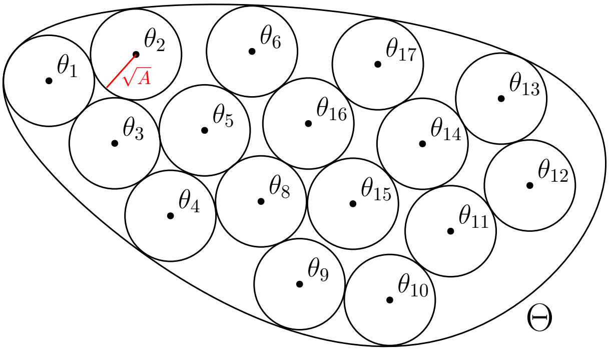
- **Reduction to a hypothesis test:** by selecting a finite subset $\{\theta_1, \dots, \theta_M\}$ of distributions Θ which is maximally spread, a good estimator leads to a good hypothesis test that can identify which θ_j (among the M possibilities) was used to generate the data. We can then use information theory to lower-bound the probability of error of such a test. This versatile technique can handle most situations, from fixed to random design.
- **Bayesian analysis:** We can lower bound the supremum for all Θ by any expectation over a distribution supported on Θ. Once we have an expectation, we can use the same decision-theoretic argument as the ones we used to compute the Bayes risk is Chapter 4, e.g., for Hilbertian or Euclidean performance measures, the optimal estimator is the conditional expectation $\mathbb{E}[\theta_* | \mathcal{D}]$. The key is choosing distributions so they can be computed in closed form. This approach is less flexible but the simplest in situations where it can be applied (fixed design regression on balls, with potentially sparse assumptions).

15.1.2 Reduction to a hypothesis test

The principle is simple: pack the set Θ with "balls" of some radius $4A$, that is, find $\theta_1, \dots, \theta_M \in \Theta$ such that

$$\forall i \neq j, \ d(\theta_i, \theta_j)^2 \geqslant 4A, \tag{15.3}$$

and transform the estimation problem into a hypothesis test, that is, an algorithm going from the data $\mathcal{D}$ to one out of M potential outcomes (see illustration below).



Then, because we take the supremum over a smaller set:

$$\sup_{\theta_* \in \Theta} \mathbb{P}_{\theta_*}\big(d(\theta_*, \mathcal{A}(\mathcal{D}))^2 > A\big) \geqslant \max_{j \in \{1,\dots,M\}} \mathbb{P}_{\theta_j}\big(d(\theta_j, \mathcal{A}(\mathcal{D}))^2 > A\big). \tag{15.4}$$

Any algorithm $A(\mathcal{D}) \in \Theta$ gives a "test", that is, a function $g \circ \mathcal{A} : \mathcal{D} \to \{1, \dots, m\}$ defined as

$$g(\mathcal{A}(\mathcal{D})) = \arg \min_{j \in \{1,\dots,m\}} d(\theta_j, \mathcal{A}(\mathcal{D})) \in \{1, \dots, m\},$$

where ties are broken arbitrarily (e.g., by selecting the minimal index). Because of the packing condition in Eq. (15.3), the performance of $\mathcal{A}$ can be lower-bounded by the classification performance of $g \circ \mathcal{A}$ (with the 0-1 loss).

Indeed, if, for some $j \in \{1, \dots, M\}$, $g(\mathcal{A}(\mathcal{D})) \neq j$, there exists $k \neq j$, such that $d(\theta_k, \mathcal{A}(\mathcal{D})) < d(\theta_j, \mathcal{A}(\mathcal{D}))$. Moreover, using the triangle inequality for d, we get:

$$d(\theta_j, \theta_k)^2 \leqslant 2\big[d(\theta_j, \mathcal{A}(\mathcal{D}))^2 + d(\theta_k, \mathcal{A}(\mathcal{D}))^2\big],$$

then,

$$\begin{aligned} d(\theta_j, \mathcal{A}(\mathcal{D}))^2 &\geqslant \frac{1}{2} d(\theta_j, \theta_k)^2 - d(\theta_k, \mathcal{A}(\mathcal{D}))^2 \\ &> \frac{1}{2} d(\theta_j, \theta_k)^2 - d(\theta_j, \mathcal{A}(\mathcal{D}))^2 \text{ by the choice of } k, \end{aligned}$$

which implies $d(\theta_j, \mathcal{A}(\mathcal{D}))^2 > \frac{1}{4} d(\theta_j, \theta_k)^2 \geqslant A$. Thus, we have the following inequality for the probabilities of these two events:

$$\mathbb{P}_{\theta_j}\big(d(\theta_j, \mathcal{A}(\mathcal{D}))^2 > A\big) \geqslant \mathbb{P}_{\theta_j}\big(g(\mathcal{A}(\mathcal{D})) \neq j\big),$$

leading to, using Eq. (15.2) and Eq. (15.4),

$$\begin{aligned} \inf_{\mathcal{A}} \sup_{\theta_* \in \Theta} \mathbb{E}_{\theta_*}\big[d(\theta_*, \mathcal{A}(\mathcal{D}))^2\big] &\geqslant A \cdot \inf_h \max_{j \in \{1,\dots,M\}} \mathbb{P}_{\theta_j}\big(g(\mathcal{D}) \neq j\big) \\ &\geqslant A \cdot \inf_h \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_{\theta_j}\big(g(\mathcal{D}) \neq j\big), \end{aligned} \tag{15.5}$$

where h is any function from $\mathcal{D}$ to $\{1,\dots,M\}$. We have thus lower-bounded the minimax statistical performance by the minimax performance of a hypothesis test $h : \mathcal{D} \to \{1,\dots,M\}$. Information theory can then be used to lower-bound this minimax error. We first provide a quick review of information theory (see Cover and Thomas, 1999, for more details).

15.1.3 Review of information theory

Entropy. Given a random variable y taking finitely many values in $\mathcal{Y}$, its entropy is equal to

$$H(y) = -\sum_{y' \in \mathcal{Y}} \mathbb{P}(y = y') \log \mathbb{P}(y = y').$$

Since $\mathbb{P}(y = y') \in [0,1]$, the entropy is always non-negative. Moreover, using Jensen's inequality for the logarithm, we have $H(y) = \sum_{y' \in \mathcal{Y}} \mathbb{P}(y = y') \log \frac{1}{\mathbb{P}(y=y')} \leqslant \log \big(\sum_{y' \in \mathcal{Y}} \mathbb{P}(y = y') \frac{1}{\mathbb{P}(y=y')} \big) = \log |\mathcal{Y}|$.

The entropy $H(y)$ represents the uncertainty associated with the random variable y, going from $H(y) = 0$ if y is deterministic (that is $\mathbb{P}(y = y') = 1$ for some $y' \in \mathcal{Y}$), to $\log |\mathcal{Y}|$ when y has a uniform distribution.

Joint and conditional entropies. Given two random variables x, y with finitely many values in $\mathcal{X}$ and $\mathcal{Y}$, we can define the joint entropy

$$H(x,y) = -\sum_{x' \in \mathcal{X}} \sum_{y' \in \mathcal{Y}} \mathbb{P}(x = x', y = y') \log \mathbb{P}(x = x', y = y').$$

It can be decomposed as

$$\begin{aligned} H(x,y) &= -\sum_{x' \in \mathcal{X}} \sum_{y' \in \mathcal{Y}} \mathbb{P}(y = y', x = x') \log \big[\mathbb{P}(y = y' | x = x') \mathbb{P}(x = x') \big] \\ &= -\sum_{x' \in \mathcal{X}} \sum_{y' \in \mathcal{Y}} \mathbb{P}(y = y', x = x') \log \mathbb{P}(y = y' | x = x') \\ &\qquad\qquad - \sum_{x' \in \mathcal{X}} \sum_{y' \in \mathcal{Y}} \mathbb{P}(y = y', x = x') \log \mathbb{P}(x = x') \\ &= \sum_{x' \in \mathcal{X}} \mathbb{P}(x = x') \log H(y | x = x') + H(x), \end{aligned}$$

where $H(y|x = x')$ is the entropy of the conditional distribution of y given $x = x'$. By defining the conditional entropy $H(y|x)$ as $H(y|x) = \sum_{x' \in \mathcal{X}} \mathbb{P}(x = x') H(y|x = x')$, we exactly have:

$$H(x,y) = H(y|x) + H(x).$$

This leads to a first version of Fano's inequality, which lower bounds the probability that $y \neq \hat{y}$ from the conditional entropy $H(y|\hat{y})$; the main idea is that if y remains very uncertain given $\hat{y}$, then the probability that they are equal cannot be too large.

Proposition 15.1 (Fano's inequality) *If the random variables y and $\hat{y}$ have values in the same finite set $\mathcal{Y}$, then*

$$\mathbb{P}(\hat{y} \neq y) \geqslant \frac{H(y|\hat{y}) - \log 2}{\log |\mathcal{Y}|}.$$

Proof Let $e = 1_{y \neq \hat{y}} \in \{0, 1\}$ be the indicator function of errors; by decomposing the joint entropy through conditional and marginal entropies in the two different ways, we get:

$$H(e|\hat{y}) + H(y|e, \hat{y}) = H(e, y|\hat{y}) = H(y|\hat{y}) + H(e|y, \hat{y}).$$

We then have $H(e|y, \hat{y}) = 0$ (since e is deterministic given y and $\hat{y}$), $H(e|\hat{y}) \leqslant H(e) \leqslant \log 2$ (because $e \in \{0, 1\}$), and $H(y|e, \hat{y}) = \mathbb{P}(e = 1)H(y|\hat{y}, e = 1) + \mathbb{P}(e = 0)H(y|\hat{y}, e = 0) = \mathbb{P}(e = 1)H(y|\hat{y}, e = 1) + 0 \leqslant \mathbb{P}(\hat{y} \neq y) \log |\mathcal{Y}|$. Expressing $\mathbb{P}(\hat{y} \neq y)$ in function of the other quantities leads to the desired result. ■

Data processing inequality. A fundamental result in information theory allows to lower-bound conditional entropies where conditional independencies are present. That is, if we have three random variables x, y, z, such that z and x are conditionally independent given y, then $H(x|z) \geqslant H(x|y)$: in words, the uncertainty of x given z has to be larger than the uncertainty of $x|y$, which is "normal" because the statistical dependence between x and z is occurring through y. In other words, the sequence $x \to y \to x$ forms a Markov chain.

The data processing inequality is a simple application of the concavity of the entropy as a function of the probability mass function; indeed, using that by conditional independence $\mathbb{P}(x = x'|z = z') = \sum_{y' \in \mathcal{Y}} \mathbb{P}(x = x', y = y'|z = z') = \sum_{y' \in \mathcal{Y}} \mathbb{P}(x = x'|y = y')\mathbb{P}(y = y'|z = z')$, and Jensen's inequality for the function $t \mapsto -t \log t$, we have:

$$\begin{aligned} H(x|z) &= \sum_{z' \in \mathcal{Z}} \mathbb{P}(z = z') H(x|z = z') \\ &\geqslant \sum_{z' \in \mathcal{Z}} \mathbb{P}(z = z') \sum_{y' \in \mathcal{Y}} \mathbb{P}(y = y'|z = z') H(x|y = y') \\ &= \sum_{y' \in \mathcal{Y}} \mathbb{P}(y = y') H(x|y = y') = H(x|y). \end{aligned}$$

This leads immediately to the following full version of Fano's inequality:

Proposition 15.2 (Fano's inequality) *If the random variable y and $\hat{y}$ have values in the same finite set $\mathcal{Y}$, and if we have a Markov chain $y \to z \to \hat{y}$, then*

$$\mathbb{P}(\hat{y} \neq y) \geqslant \frac{H(y|\hat{y}) - \log 2}{\log |\mathcal{Y}|} \geqslant \frac{H(y|z) - \log 2}{\log |\mathcal{Y}|}.$$

We need a last concept from information theory, namely mutual information and Kullback-Leibler divergence, both for discrete and continuous-valued random variables.

Mutual information. Given two random variables x and y, then we can define their mutual information as

$$I(x,y) = H(x) - H(x|y) = H(x) + H(y) - H(x,y) = H(y) - H(y|x).$$

This can be seen as the uncertainty reduction in x when observing y. It is symmetric, always less than $\log|\mathcal{X}|$ and $\log|\mathcal{Y}|$. Moreover, it can be written as:

$$\begin{aligned} I(x,y) &= H(x) + H(y) - H(x,y) \\ &= \sum_{x'\in\mathcal{X}}\sum_{y'\in\mathcal{Y}} \mathbb{P}(x=x',y=y')\log\frac{\mathbb{P}(x=x',y=y')}{\mathbb{P}(x=x')\mathbb{P}(y=y')}, \end{aligned}$$

which can be seen as the Kullback-Leibler (KL) divergence between the distribution of (x,y) and the product of marginals of x and y. Indeed, given two distributions on $\mathcal{Z}$, p and q (which are non-negative functions on $\mathcal{Z}$ that sum to one), then the KL divergence is defined as

$$D_{\mathrm{KL}}(p||q) = \sum_{z\in\mathcal{Z}} p(z)\log\frac{p(z)}{q(z)}.$$

The KL divergence is always non-negative by convexity of the function $t \mapsto t\log t$, and equal to zero, if and only if $p = q$. Moreover, the KL divergence is jointly convex in (p,q). Thus, one can see the mutual information between the KL divergences between the joint distribution of (x,y) and the corresponding product of marginals (which is thus non-negative).

From discrete to continuous distributions. Many of the information theory concepts can be extended to continuous random variables on $\mathbb{R}^d$ by replacing the probability mass function with the probability density with respect to some base measures. Then, many properties (which were obtained through convex arguments) extend. In particular, the data processing inequality and Fano's inequality when z is continuous-valued (see more details by Cover and Thomas, 1999).

Moreover, the KL divergence between two distributions can be defined as

$$D_{\mathrm{KL}}(p||q) = \mathbb{E}_p\Big[\log\frac{dp}{dq}(x)\Big].$$

A short calculation shows that for two normal distributions of means μ_1,μ_2 and equal covariance matrices Σ, the KL divergence is equal to $\frac{1}{2}(\mu_1-\mu_2)^\top\Sigma^{-1}(\mu_1-\mu_2)$.

15.1.4 Lower-bound on hypothesis testing based on information theory

We consider a joint random variable $(y,\mathcal{D})$ distributed as y uniform in $\{1,\dots,M\}$, and, given $y=j$, $\mathcal{D}$ distributed as the distribution associated with θ_j. We consider $\hat{y} = h(\mathcal{D})$. This defines a Markov chain: $y \to \mathcal{D} \to h(\mathcal{D})$, that is, even for a randomized test h (with extra randomization), $h(\mathcal{D})$ is independent of y given $\mathcal{D}$. By construction, the last term in Eq. (15.5) is exactly the probability that $\hat{y} \neq y$. This is exactly what Fano's inequality from information theory gives us, leading to the following corollary.

Corollary 15.1 (Fano's inequality for multiple hypothesis testing) *Given M probability distributions p_j on $\mathcal{D}$, then*

$$\inf_h \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_{p_j}\big(h(\mathcal{D}) \neq j\big) \geqslant 1 - \frac{1}{M^2 \log M} \sum_{j,j'=1}^{M} D_{\mathrm{KL}}(p_j || p_{j'}) - \frac{\log 2}{\log M}. \tag{15.6}$$

Proof We consider a joint random variable $(y, \mathcal{D})$ distributed as y uniform in $\{1, \dots, M\}$, and, given $y = j$, $\mathcal{D}$ distributed as the distribution p_j. We have:

$$\begin{aligned} H(y|z) &= H(y) - I(y,z) = \log M - \frac{1}{M} \sum_{j=1}^{M} D_{\mathrm{KL}}\big(p_j \big\| \frac{1}{M} \sum_{j'=1}^{M} p_{j'}\big) \\ &\geqslant \log M - \frac{1}{M^2} \sum_{j,j'=1}^{M} D_{\mathrm{KL}}(p_j \| p_{j'}), \end{aligned}$$

by the convexity of the Kullback-Leibler divergence. We can then apply Prop. 15.2 and conclude. ■

Using Gaussian noise to compute KL divergences. For regression with Gaussian errors such as $y_i = f_\theta(x_i) + \varepsilon_i$, with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, then, for fixed designs (all x_i's deterministic), we get exactly

$$D_{\mathrm{KL}}(p_{\theta_j} || p_{\theta_{j'}}) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} \big[f_{\theta_j}(x_i) - f_{\theta_{j'}}(x_i)\big]^2 = \frac{n}{2\sigma^2} d(\theta_j, \theta_{j'})^2,$$

where $d(\theta, \theta')^2 = \frac{1}{n} \sum_{i=1}^{n} \big[f_\theta(x_i) - f_{\theta'}(x_i)\big]^2$ is the empirical mean squared difference between two models.

For random designs, we consider distributions on $(x_i, y_i)_{i=1,\dots,n}$. If we consider a common distribution p for x, then

$$D_{\mathrm{KL}}(p_{\theta_j} || p_{\theta_{j'}}) = \frac{1}{2\sigma^2} \int_{\mathcal{X}} \big[f_{\theta_j}(x) - f_{\theta_{j'}}(x)\big]^2 dp(x) = \frac{1}{2\sigma^2} \|f_{\theta_j} - f_{\theta_{j'}}\|^2_{L_2(p)},$$

which we define to be $\frac{1}{2\sigma^2} d(\theta_j, \theta_{j'})^2$.

Overall, to obtain a lower bound with Gaussian noise, we need to find $\theta_1, \dots, \theta_M$ in Θ such that:

- $\displaystyle \frac{1}{M^2} \sum_{j,j'=1}^{M} \frac{n}{2\sigma^2} d(\theta_j, \theta_{j'})^2 \leqslant \log(M)/4.$ and $\log 2 / \log M \leqslant 1/4$ (that is $M \geqslant 16$), so that Eq. (15.6) leads to a lower bound of $A/2$.

- $\min_{j \neq k} d(\theta_j, \theta_k)^2 \geqslant 4A$, so that we can apply Eq. (15.5).

Then, the minimax lower bound is $A/2$. Thus, the lower bound is essentially the largest possible A for a given M such that we can find M points in Θ, which are all $2\sqrt{A}$ apart. There are two main tools to find such packings: (1) a direct volume argument and (2) using Varshamov-Gilbert's lemma. We present them before going over examples.

Volume argument. The following lemma provides the simplest argument.

Lemma 15.1 (Packing ℓ_2-balls) *Let M be the maximal number of elements of the Euclidean ball of radius 1, which are at least 2ε-apart in ℓ_2-norm. Then $(2\varepsilon)^{-d} \leqslant M \leqslant (1+\varepsilon^{-1})^d$.*

Proof Let $\theta_1, \dots, \theta_M$ be the corresponding M points.

(a) All balls of center θ_j and radius ε are disjoint and included in the ball of radius $1+\varepsilon$. Thus, the sum of the volumes of the small balls is smaller than the volume of the large balls, that is, $M\varepsilon^d \leqslant (1+\varepsilon)^d$.

(b) Since M is maximal, for any θ such that $\|\theta\|_2 \leqslant 1$, there exists a $j \in \{1, \dots, M\}$ such that $\|\theta_j - \theta\|_2 \leqslant 2\varepsilon$ (otherwise, we can add a new point to $\{\theta_1, \dots, \theta_M\}$ and M is not maximal). Thus, the ball of radius 1 is covered by the M balls of radius θ_j and radius 2ε. Thus, by using volumes, we get $1 \leqslant M(2\varepsilon)^d$. ■

Packing with Varshamov-Gilbert lemma. The maximal number of points in the hypercube $\{0,1\}^d$ that are at least $d/4$-apart in Hamming loss (i.e., ℓ_1-distance) is greater than than $\exp(d/8)$, with a nice probabilistic argument.

Lemma 15.2 (Varshamov-Gilbert's lemma) *For any $\alpha \in (0,1)$, there exists a subset $\mathcal{B}$ of the hypercube $\{0,1\}^d$ such that*

(a) for all $x, x' \in \mathcal{B}$ such that $x \neq x'$, $\|x - x'\|_1 \geqslant (1-\alpha)\frac{d}{2}$,

(b) $|\mathcal{B}| \geqslant \exp(d\alpha^2/2)$.

Proof We consider the largest family satisfying (a). By maximality, the union of ℓ_1-balls of radius $(1-\alpha)\frac{d}{2}$ includes all of $\{0,1\}^d$. Therefore, by comparing cardinalities,

$$2^d \leqslant \sum_{x \in \mathcal{B}} \big|\big\{y \in \{0,1\}^d, \|y - x\|_1 \leqslant (1-\alpha)\frac{d}{2}\big\}\big|.$$

Consider a random variable z, which is binomial with parameters d and $1/2$ (that is, the sum of d independent uniform Bernoulli random variables). Then,

$$2^{-d}\big|\big\{y \in \{0,1\}^d, \|y-x\|_2^2 = \|y-x\|_1 \leqslant (1-\alpha)\frac{d}{2}\big\}\big| = \mathbb{P}\big(z \leqslant (1-\alpha)\frac{d}{2}\big) = \mathbb{P}\big(z \geqslant (1+\alpha)\frac{d}{2}\big).$$

Using Hoeffding's inequality (Prop. 1.1), we get $\mathbb{P}(z \geqslant (1+\alpha)\frac{d}{2}) = \mathbb{P}(\frac{z}{d} - \frac{\mathbb{E}[z]}{d} \geqslant \alpha\frac{d}{2}) \leqslant \exp(-2d(\alpha/2)^2) = \exp(-d\alpha^2/2)$. This leads to the result. ■

15.1.5 Examples

Fixed design linear regression. We consider linear regression with $\Phi \in \mathbb{R}^{n\times d}$ a design matrix with $\frac{1}{n}\Phi^\top\Phi = I$ (which imposes $n \geqslant d$). We consider the ball $\Theta = \{\theta \in \mathbb{R}^d, \|\theta\|_2 \leqslant D\}$. By rotational invariance of the Gaussian distribution of the noise variable ε, we can assume that the first d rows of Φ are equal to $\sqrt{n}I$ and the rest of the rows are equal to zero. Thus we can assume the model $y = \theta_* + \frac{1}{\sqrt{n}}\varepsilon$, where $\varepsilon \in \mathbb{R}^d$ with normal distribution with mean zero and covariance $\sigma^2 I$, and $y \in \mathbb{R}^d$. We are thus in the situation where $d(\theta,\theta')^2 = \|\theta-\theta'\|_2^2$.

In order to find M points in $\Theta = \{\theta \in \mathbb{R}^d,\ \|\theta\|_2 \leqslant D\}$, we consider the $M \geqslant \exp(d/8)$ elements $x_1,\dots,x_M$ of $\{0,1\}^d$ from Lemma 15.2 with $\alpha = 1/2$, and define $\theta_i = \beta(2x_i - 1_d) \in \{-\beta,\beta\}$. Thus $\|\theta_i\|_2^2 = \beta^2 d$, and, for $i \neq j$,

$$\|\theta_i - \theta_j\|_2^2 \leqslant 4\beta^2 d \leqslant 32\beta^2\log(M) \text{ and } \|\theta_i-\theta_j\|_2^2 \geqslant \beta^2 d.$$

We thus need, $\beta^2 d \leqslant D^2$, and $32\beta^2\log(M)\frac{n}{2\sigma^2} \leqslant \frac{\log M}{4}$, that is, $64\beta^2\frac{n}{\sigma^2} \leqslant 1$. Thus, the optimal rate is greater than

$$\frac{1}{8}\beta^2 d \geqslant \frac{1}{8}\min\Big\{D^2, \frac{\sigma^2 d}{64n}\Big\}.$$

Therefore, when $D^2 \geqslant \frac{\sigma^2 d}{64n}$, we get a lower bound of $\dfrac{\sigma^2 d}{512n}$, which is the upper-bound obtained in Chapter 3 (note that in Section 3.7 we provided a sharper lower-bound using similar tools as in Section 15.1.6).

The sparse regression setting could also be considered with the same tool, but the proof is simpler with the Bayesian arguments from Section 15.1.6. We now turn to the random design setting.

Exercise 15.1 *Use Lemma 15.1 instead of Lemma 15.2 to obtain the same result.*

Random design linear regression. We consider the same model as above, but with (x_i,y_i) sampled i.i.d. from a given distribution such that $\mathbb{E}[\varphi(x)\varphi(x)^\top] = I$, so that $d(\theta,\theta')^2 = \|\theta-\theta'\|_2^2$. Thus, the result above for fixed design regression also applies to the random design setting.

Non-parametric estimation with Hilbert spaces (♦). We consider random design regression with a fixed distribution for the inputs, with Gaussian independent noise and target functions in a certain ellipsoid of $L_2(p)$. That is, we assume that there exists a compact self-adjoint operator T on $L_2(p)$ such that $\langle\theta, T^{-1}\theta\rangle_{L_2(p)} \leqslant D^2$. We denote by $(\lambda_m)_{m\geqslant 1}$ the non-increasing sequence of eigenvalues of T, with the associated eigenvectors ψ_m in $L_2(p)$.

We consider a certain integer K, and $M \geqslant \exp(K/8)$ elements $x_1,\dots,x_M$ of $\{0,1\}^K$. We define $\theta_i = \beta\sum_{m=1}^K (2(x_i)_m - 1)\psi_m$. Then $\langle\theta, T^{-1}\theta\rangle_{L_2(p)} = \beta^2\sum_{m=1}^K \lambda_m^{-1} \leqslant K\beta^2\lambda_K^{-1}$, and, for $i \neq j$,

$$\|\theta_i - \theta_j\|_{L_2(p)}^2 \leqslant 4\beta^2 K \leqslant 32\beta^2 \log(M) \text{ and } \|\theta_i - \theta_j\|_{L_2(p)}^2 \geqslant \beta^2 K.$$

We thus need, $\beta^2 K \leqslant D^2 \lambda_K$, and $32\beta^2 \log(M) \frac{n}{2\sigma^2} \leqslant \frac{\log M}{4}$, that is, $64\beta^2 \frac{n}{\sigma^2} \leqslant 1$. Thus, the minimax lower bound is greater than

$$\frac{1}{8}\beta^2 K \geqslant \frac{1}{8} \min\Big\{D^2 \lambda_K, \frac{\sigma^2 K}{64n}\Big\}.$$

We can now specialize to Sobolev spaces where it can be shown that for compact supports with piecewise smooth boundaries. The sum of all L_2-norms of partial derivatives corresponds to an operator for which $\lambda_K \geqslant C \cdot K^{-\alpha}$, with $\alpha = 2s/d$ when all s-th order derivatives are taken, for a constant C (Adams and Fournier, 2003). The lower bound becomes

$$\max_{K \geqslant 1} \frac{1}{8} \min\Big\{D^2 C K^{-\alpha}, \frac{\sigma^2 K}{64n}\Big\},$$

which can be balanced to obtain $K \propto \big(\frac{nD^2}{\sigma^2}\big)^{1/(1+\alpha)}$, leading to lower bound proportional to

$$D^{2/(1+\alpha)} \Big(\frac{\sigma^2}{n}\Big)^{\alpha/(1+\alpha)}.$$

For $\alpha = 2s/d$, we get $\alpha/(1+\alpha) = \frac{2s}{2s+d}$, and the lower matches the upper-bound obtained with kernel ridge regression in Chapter 7. It turns out that the lower bound on the minimax rate for Lipschitz-continuous functions is the same as for $s = 1$ (Tsybakov, 2008, Section 2.6).

15.1.6 Minimax lower bounds through Bayesian analysis

We can use a Bayesian analysis as outlined for least-squares in Section 3.7. We consider a particular probability distribution $p(\theta_*)$ whose support is included in Θ. Then we have:

$$\inf_{\mathcal{A}} \sup_{\theta_* \in \Theta} \mathbb{E}_{\theta_*}\big[d(\theta_*, \mathcal{A}(\mathcal{D}))^2\big] \geqslant \inf_{\mathcal{A}} \mathbb{E}_{p(\theta_*)} \mathbb{E}_{\theta_*}\big[d(\theta_*, \mathcal{A}(\mathcal{D}))^2\big].$$

This reasoning is particularly simple when the optimal algorithm $\mathcal{A}$ is simple to estimate, which is the case in particular where d is a Euclidean norm so that $\mathcal{A}^*(\mathcal{D}) = \mathbb{E}\big[\theta_* | \mathcal{D}\big]$. If the prior $p(\theta_*)$ and the likelihood $p(\mathcal{D}|\theta_*)$ are simple enough, then the conditional expectation can be computed in closed form. In Section 3.7, these were all Gaussians, which was possible for the prior distribution on Θ because Θ was unbounded. When dealing with bounded balls, we need to use different distributions, as used originally by Donoho and Johnstone (1994).

Least-squares on a Euclidean ball. We consider linear regression with a fixed design like in the previous section (with a bound $\|\theta_*\|_2 \leqslant D$), which corresponds to the model $y = \theta_* + \frac{1}{\sqrt{n}}\varepsilon$, where $\varepsilon \in \mathbb{R}^d$ with normal distribution with mean zero and covariance $\sigma^2 I$, and $y, \theta_* \in \mathbb{R}^d$.

We then consider a prior distribution on θ_* as $\theta_* = \beta x$, where $x \in \{-1,1\}^d$ are independent Rademacher random variables. We need $\beta^2 d \leqslant D^2$ to be in the correct set. We then need to compute $\mathbb{E}[\theta_*|y]$. The posterior probability of θ_* is supported on $\beta\{-1,1\}^n$. Moreover, given the independence by component, we can treat each separately. Then, by keeping only terms that depend on the posterior value, we get:

$$\mathbb{P}((\theta_*)_i = \pm\beta|y_i) \propto \exp(-\frac{n}{2\sigma^2}(y_i - \pm\beta)^2) \propto \exp(\pm\frac{n}{\sigma^2}y_i\beta).$$

Thus,

$$\mathbb{E}\big[(\theta_*)_i|y_i\big] = \beta\frac{\exp(\frac{n}{\sigma^2}y_i\beta) - \exp(\frac{-n}{\sigma^2}y_i\beta)}{\exp(\frac{n}{\sigma^2}y_i\beta) + \exp(\frac{-n}{\sigma^2}y_i\beta)} = \beta\frac{1 - \exp(\frac{-2n}{\sigma^2}y_i\beta)}{1 + \exp(\frac{-2n}{\sigma^2}y_i\beta)} = \beta\big[2\text{sigmoid}\big(\frac{2n}{\sigma^2}y_i\beta\big) - 1\big],$$

where $\text{sigmoid}(\alpha) = 1/(1+\exp(-\alpha))$.

The posterior variance for the i-th component is equal to

$$\begin{aligned}
\mathbb{E}\big[\big((\theta_*)_i - \mathbb{E}\big[(\theta_*)_i|y_i\big]\big)^2\big] &= \frac{1}{2}\mathbb{E}_{\varepsilon_i}\big(\beta - \beta\big[2\text{sigmoid}\big(2\frac{n}{\sigma^2}\beta(\beta + \varepsilon_i/\sqrt{n})\big) - 1\big]\big)^2 \\
&\quad + \frac{1}{2}\mathbb{E}_{\varepsilon_i}\big(-\beta - \beta\big[2\text{sigmoid}\big(2\frac{n}{\sigma^2}\beta(-\beta + \varepsilon_i/\sqrt{n})\big) - 1\big]\big)^2 \\
&= 4\beta^2\mathbb{E}_{\varepsilon_i\sim\mathcal{N}(0,\sigma^2)}\Big[\big(\text{sigmoid}\big(-2\frac{n}{\sigma^2}\beta^2 + 2\frac{\sqrt{n}}{\sigma^2}\beta\varepsilon_i\big)\big)^2\Big] \\
&= 4\beta^2\mathbb{E}_{\tilde{\varepsilon}_i\sim\mathcal{N}(0,1)}\Big[\big(\text{sigmoid}(-2\frac{n}{\sigma^2}\beta^2 + 2\frac{\beta\sqrt{n}}{\sigma}\tilde{\varepsilon}_i)\big)^2\Big].
\end{aligned}$$

We consider the function $\psi : \alpha \mapsto \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,1)}\big[\big(\text{sigmoid}(-2\alpha^2 + 2\alpha\varepsilon)\big)^2\big]$. We have $\psi(0) = 1/4$, and $\psi(\alpha) \to 0$ when $\alpha \to +\infty$, and $\psi(\alpha) \geqslant \frac{1}{4}\mathbb{P}_{\varepsilon\sim\mathcal{N}(0,1)}(\varepsilon > \alpha) \geqslant \frac{1}{8}\exp(-\alpha^2)$, by using simple Gaussian tail bounds (and since the sigmoid function is greater than $1/2$ for positive numbers).

Thus, the total posterior variance $\mathbb{E}\big[\big\|\theta_* - \mathbb{E}\big[\theta_*|y\big]\big\|_2^2\big]$ is greater than

$$\frac{\beta^2 d}{2}\exp(-n\beta^2/\sigma^2) = \frac{\sigma^2 d}{n} \times \frac{\beta^2 n}{2\sigma^2}\exp(-n\beta^2/\sigma^2),$$

which is maximized for $\beta^2 \propto \sigma^2/n$, and thus if $\sigma^2 d/n$ is smaller than D^2, we obtain the usual $\sigma^2 d/n$, while if it is greater then D^2, we take $\beta_2^2 = D^2/d$, to obtain the lower bound

$$D^2\exp(-4nD^2/(\sigma^2 d)) \geqslant D^2\exp(-4),$$

which leads to the same bound as the previous section but with a more direct argument.

Sparse case (♦). To deal with the sparse case, we could consider a prior on θ_* that only selects k non-zero elements out of d and perform an analysis based on the posterior probability of θ_*. Following Donoho and Johnstone (1994), it is easier to divide the set of d variables into k blocks of size d/k (for simplicity, we assume that d/k is an integer).

We then consider a prior probability defined independently on each of the k blocks by selecting one of the d/k variables uniformly at random and setting its value to β. In contrast, all others are set to zero.

To compute the posterior probability of θ_*, we can treat each block independently and sum the posterior variances; we thus consider the first block, composed of d/k variables, and compute the probability that the selected variable is the j-th one, which is proportional to

$$\exp(-n/(2\sigma^2)(y_j-\beta)^2)\prod_{i\neq j}\exp(-n/(2\sigma^2)(y_i)^2) \propto \exp(n\beta y_j/\sigma^2).$$

The conditional expectation of θ_* then satisfies

$$\mathbb{E}[(\theta_*)_i|y] = \beta\frac{\exp(n\beta y_i/\sigma^2)}{\sum_{j=1}^{d/k}\exp(n\beta y_j/\sigma^2)}.$$

To compute the posterior variance, we need to sample from the prior θ_*. By symmetry, we may consider that $\theta_1=\beta$. If $y_1 \leqslant \max_{j\neq 1} y_j$, then

$$\mathbb{E}[(\theta_*)_1|y] = \beta\frac{\exp(n\beta y_1/\sigma^2)}{\sum_{j=1}^{d/k}\exp(n\beta y_j/\sigma^2)} \leqslant \beta t\frac{\exp(n\beta y_1/\sigma^2)}{\exp(n\beta y_1/\sigma^2)+\exp(n\beta\max_{j\neq 1} y_j/\sigma^2)} \leqslant \beta/2,$$

and then the risk is at least $(\beta-\mathbb{E}[(\theta_*)_1|y])^2 \geqslant \beta^2/4$.

In order to lower-bound the probability that $y_1 \leqslant \max_{j\neq 1} y_j$, we can consider the events $\{y_1 \leqslant \beta\}$ and $\{\beta \leqslant \max_{j\neq 1} y_j\}$. The probability that $y_1=\beta+\varepsilon_1$ is less than β is greater than 1/2. Moreover, by independence of all y_j, $j\neq 1$,

$$\mathbb{P}\big(\{\beta \leqslant \max_{j\neq 1} y_j\}\big) \geqslant 1-\big(1-\mathbb{P}_{t\sim\mathcal{N}(0,1)}(t\geqslant \beta\sqrt{n}/\sigma)\big)^{d/k-1}.$$

Thus, the lower bound is greater than

$$k\frac{\beta^2}{8}\Big[1-\big(1-\mathbb{P}_{t\sim\mathcal{N}(0,1)}(t\geqslant\beta\sqrt{n}/\sigma)\big)^{d/k-1}\Big] \geqslant k\frac{\beta^2}{8}\Big[1-\Big(1-\frac{1}{2}\exp(-\beta^2 n/\sigma^2)\Big)^{d/k-1}\Big],$$

using the Gaussian tail bound $\mathbb{P}_{t\sim\mathcal{N}(0,1)}(t\geqslant z)\geqslant\frac{1}{2}\exp(-z^2)$. We can then consider $\beta^2=\frac{\sigma^2}{n}\sqrt{2\log(d/k)}$, leading to a lower bound

$$\frac{\sigma^2 k}{4n}\log(d/k)\big[(1-(1-\tfrac{1}{2}(k/d))^{d/k-1}\big],$$

which is greater than $\frac{\sigma^2 k}{8n}\log(d/k)$ if $k\leqslant 2d$. We obtain the same lower-bound as the upper-bound for ℓ_0-penalty-based methods in Chapter 8.

15.2 Optimization lower bounds

In this section, we consider ways of obtaining lower bounds of performance for optimization algorithms corresponding to upper-bounds derived in Chapter 5 for gradient-based algorithms. While the statistical lower bounds from the previous section were not obtained by explicitly building hard problems, the algorithmic lower bounds of this section will explicitly build such hard problems.

15.2.1 Convex optimization

To obtain computational lower bounds for convex optimization, which is notoriously hard in general in computer science, we will rely on a simple model of computation; that is, we will restrict ourselves to methods that access gradients of the objective function and combine them linearly to select a new query point.

We follow the results from Nesterov (2018, Section 2.1.2) and Bubeck (2015, Section 3.5), and assume that we want to minimize a convex function F defined on $\mathbb{R}^d$. The algorithm starts from $\theta_0 = 0$ and can only query points in the span of the observed gradients or some sub-gradients of F at the previously observed points.

The key is finding functions with the proper regularity properties, for which we know a few iterations provably lead to suboptimal performance. These functions will only reveal one new variable at each iteration and, after k iterations, can only achieve the minimum on the first k variables.

Non-smooth functions. We consider the following function, which will be dedicated to a given number of iterations k:

$$F(\theta) = \eta \max_{i \in \{1,\dots,k+1\}} \theta_i + \frac{\mu}{2}\|\theta\|_2^2,$$

for $k < d$, and η, μ positive parameters that will be set later.

The subdifferential of $F(\theta)$ is equal to

$$\mu\theta + \eta \cdot \mathrm{hull}\big(\big\{e_i,\ \theta_i = \max_{i' \in \{1,\dots,k+1\}} \theta_{i'}\big\}\big),$$

which is bounded in ℓ_2-norm on the ball of radius R, by $\mu R + \eta$ (here e_i denotes the i-th basis vector). We consider the oracle where the output gradient is $\mu\theta + \eta e_i$, where i is the smallest index within maximizers of $\theta_{i'}$.

Starting from $\theta_0 = 0$, θ_1 is supported on the first variable, and by recursion, after $k \leqslant d$ steps of subgradient descent, θ_k is supported on the first k variables. Since $k < d$, then $(\theta_k)_{k+1} = 0$, so $F(\theta_k) \geqslant 0$. Minimizing over the span of the first $k+1$ variables leads to, by symmetry, $\theta_* = \kappa \sum_{i=1}^{k+1} e_i$, for a certain κ which minimizes $\eta\kappa + \frac{(k+1)\mu}{2}\kappa^2$, so that $\kappa = -\frac{\eta}{\mu(k+1)}$, and thus $\theta_* = -\frac{\eta}{\mu(k+1)} \sum_{i=1}^{k+1} e_i$, with value $F(\theta_*) = -\frac{\eta^2}{2\mu(k+1)}$. Thus

$$F(\theta_k) - F(\theta_*) \geqslant 0 - F(\theta_*) = \frac{\eta^2}{2\mu(k+1)},$$

with $\|\theta_*\|_2^2 = \frac{\eta^2}{\mu^2(k+1)}$.

In order to build a B-Lipschitz-continuous function on a ball of center 0 and radius D, we can take $\eta = B/2$, and $D = B/(2\mu)$, and we get a lower bound of $\frac{B^2}{8\mu k}$.

With $\mu = \frac{B}{D}\frac{1}{1+\sqrt{k+1}}$ and $\eta = B\frac{\sqrt{k+1}}{1+\sqrt{k+1}}$, we also get a B-Lipschitz continuous function, and we get the lower bound $\frac{DB}{2(1+\sqrt{k+1})}$, which is valid as long as $k < d$.

⚠ The lower bounds are only valid for $k < d$ because there exist algorithms that are linearly convergent in this setting with a constant that depends on d, such as the ellipsoid method or the center of mass method (see Bubeck, 2015, for details).

Smooth functions (♦). We consider a sequence of quadratic functions on $\mathbb{R}^d$. We need that the gradient for iterates supported on the first i components is supported on the first $i+1$ components, so that the k-th iterate starting from 0 only has its first k coordinates that can be non-zero. We consider the example from Nesterov (2018, Section 2.1.2), and highlight the main arguments without proof:

$$F_k(\theta) = \frac{L}{4}\Big\{\frac{1}{2}\Big[\theta_1^2 + \theta_k^2 + \sum_{i=1}^{k-1}(\theta_i - \theta_{i+1})^2\Big] - \theta_1\Big\}.$$

The function F_k is convex and smooth, with a smoothness constant less than L. Moreover, its global minimizer is attained at $\theta_*^{(k)}$ such that $(\theta_*^{(k)})_i = 1 - \frac{i}{k+1}$ for $i \in \{1, \dots, k\}$ and 0 otherwise, with an optimal value of $F_k(\theta_*^{(k)}) = \frac{L}{8}\frac{-k}{k+1}$, and with

$$\|\theta_*^{(k)}\|_2^2 = \sum_{i=1}^{k}\big(1 - \frac{i}{k+1}\big)^2 \leqslant \frac{k+1}{3}.$$

By construction, if θ is supported in the first i components for $i < k$, then $F_k'(\theta)$ is supported on the first $i+1$ components. Thus, the i-th iterate is supported on the first i components, and therefore the lowest attainable value is $F_i(\theta_*^{(i)})$.

Given this set of functions, for a given k such that $k \leqslant \frac{d-1}{2}$, we consider F_{2k+1}, for which $\theta_*^{(2k+1)}$ is the global minimizer with value $\frac{L}{8}\frac{-2k-1}{2k+2}$, while after k iterations, we can only achieve $F_k(\theta_*^{(k)}) = \frac{L}{8}\frac{-k}{k+1}$. Thus, we have:

$$\frac{F_{2k+1}(\theta_k) - F_{2k+1}^*}{\|\theta_0 - \theta_*\|_2^2} \geqslant \frac{L}{8}\frac{\frac{1}{k+1} - \frac{1}{2k+2}}{\frac{2k+2}{3}} \geqslant \frac{3L}{32}\frac{1}{(k+1)^2}.$$

We thus obtain the lower bounds corresponding to the upper bounds obtained from Nesterov acceleration.

⚠ The number of iterations has to be less than half the dimension for the lower bound to hold.

Smooth strongly-convex functions (♦). Following Nesterov (2018), we consider a function defined on the space ℓ_2 of square-summable sequences as

$$F(\theta) = \frac{L-\mu}{4}\Big\{\frac{1}{2}\Big[\theta_1^2 + \sum_{i=1}^{\infty}(\theta_i - \theta_{i+1})^2\Big] - \theta_1\Big\} + \frac{\mu}{2}\|\theta\|_2^2.$$

This function is L-smooth and μ-strongly convex. Its global minimizer is θ_* such that

$$(\theta_*)_k = \Big(\frac{1-\sqrt{\mu/L}}{1+\sqrt{\mu/L}}\Big)^k = q^k,$$

with $\|\theta_*\|_2^2 = \sum_{k=1}^{\infty} q^{2k} = \frac{q^2}{1-q^2}$. Moreover, it can be shown that $\|\theta_k - \theta_*\|_2^2 \geqslant \sum_{i=k+1}^{\infty} q^{2i} = q^{2k}\|\theta_*\|_2^2$. This leads to $F(\theta_k) - F_* \geqslant \frac{\mu}{2}\|\theta_k - \theta_*\|_2^2 \geqslant q^{2k}\|\theta_0 - \theta_*\|_2^2$.

15.2.2 Non-convex optimization (♦)

While upper and lower bounds can have good behavior with respect to dimension in the convex case, this is not the case when removing the convexity assumption. In this section, we show that when optimizing a Lipschitz-continuous function on a compact subset of $\mathbb{R}^d$, we cannot hope to have guarantees which are not exponential in dimension.

⚠ This does not mean that all problem instances will require exponential time, but that in the worst-case sense, for any algorithm, there will always be a bad function.

We consider minimizing a function F on a bounded subset Θ of $\mathbb{R}^d$, based only on function evaluations, a problem often referred to as zero-th order optimization or derivative-free optimization (see algorithms for convex functions in Section 11.2). No convexity is assumed in this section, so we should not expect fast rates and, again, no efficient algorithms that can provably find a global minimizer. Clearly, such algorithms are not made to be used to find millions of parameters for logistic regression or neural networks. Still, they are often used for hyperparameter tuning (regularization parameters, size of neural network layers, etc.). See, e.g., Snoek et al. (2012) for applications.

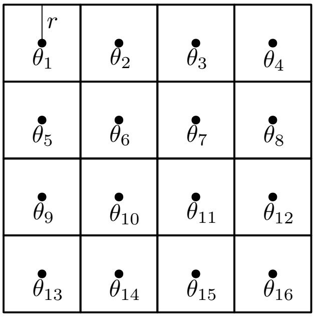
We will assume some regularity for the functions we want to minimize, typically bounded derivatives. We will thus assume that $f \in \mathcal{F}$, for a space $\mathcal{F}$ of functions from Θ to $\mathbb{R}$. We will take a worst-case approach, where we characterize convergence over all members of $\mathcal{F}$. That is, we want our guarantees to hold for *all* functions in $\mathcal{F}$. Note that this worst-case analysis may not predict well what is happening for a particular function; in particular, it is (by design) pessimistic.

An algorithm $\mathcal{A}$ will be characterized by (a) the choice of points $\theta_1, \dots, \theta_n \in \Theta$ to query the function, and (b) the algorithm to output a candidate $\hat{\theta} \in \Theta$ such that $F(\hat{\theta}) - \inf_{\theta \in \Theta} F(\theta)$ is small. The estimate $\hat{\theta}$ can only depend on $(\theta_i, F(\theta_i))$, for $i \in \{1, \dots, n\}$. In this section, the choice of points $\theta_1, \dots, \theta_n$ is made once (without seeing any function values).[1]

Given a selection of points and the algorithm $\mathcal{A}$, the rate of convergence is the supremum over all functions $F \in \mathcal{F}$ of the error $F(\hat{\theta}) - \inf_{\theta \in \Theta} F(\theta)$. This is a function $\varepsilon_n(\mathcal{A})$ of the number n of sampled points (and of the class of functions $\mathcal{F}$). The optimal algorithm (minimizing $\varepsilon_n(\mathcal{A})$) will lead to a rate we denote $\varepsilon_n^{\text{opt}}$, and which we aim to characterize.

Direct lower/upper bounds for Lipschitz-continuous functions. The argument is particularly simple for a bounded metric space Θ with distance δ, and $\mathcal{F}$ the class of L-Lipschitz-continuous functions, that is, such that for all $\theta, \theta' \in \Theta$, $|F(\theta) - F(\theta')| \leqslant L\delta(\theta, \theta')$. This is a very large set of functions, so we expect weak convergence rates.

Like in Section 4.4.4, we will need to cover the set Θ with balls of a given radius. The minimal radius r of a cover of Θ by n balls of radius r is denoted $r_n(\Theta, \delta)$. This corresponds to n ball centers $\theta_1, \dots, \theta_n$. See example below for the unit cube $\Theta = [0,1]^2$ and the metric obtained from the ℓ_∞-norm, with $n = 16$, and $r_n([0,1]^2, \ell_\infty) = 1/8$.



More generally, for the unit cube $\Theta = [0,1]^d$, we have $r_n([0,1]^d, \ell_\infty) \approx \frac{1}{2} n^{-1/d}$ (which is not an approximation when n is the d-th power of an integer). For other normed metrics (since all norms are equivalent), the scaling as $r_n \sim \text{diam}(\Theta) n^{-1/d}$ is the same on any bounded set in $\mathbb{R}^d$ (with an extra constant that depends on d).

Naive algorithm. Given the ball centers $\theta_1, \dots, \theta_n$, outputting the minimum of function values $F(\theta_i)$ for $i = 1, \dots, n$, leads to an error which is less than $Lr_n(\Theta, \delta)$, as the optimal $\theta_* \in \Theta$ is at most at distance $r_n(\Theta, \delta)$ from one of the cluster centers, let's say θ_k, and thus $F(\theta_k) - F(\theta_*) \leqslant L\delta(\theta_k, \theta_*) \leqslant Lr_n(\Theta, \delta)$. This provides an upper bound on $\varepsilon_n^{\text{opt}}$. The algorithm we just described seems naive, but it turns out to be optimal for this class of problems.

Lower bound. Consider any optimization algorithm, with its first n point queries and its estimate $\hat{\theta}$. By considering the functions that are zero in these $n+1$ points, the algorithm can only output an arbitrary fixed real number for the optimal value (let's say

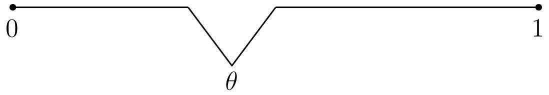
[1] It turns out that going *adaptive*, where the point θ_{i+1} is selected after seeing $(\theta_j, F(\theta_j))$ for all $j \leqslant i$, does not bring much (at least in the worst-case sense) (Novak, 2006).

zero). We now simply need to construct a function $F \in \mathcal{F}$ such that F is zero at these points but maximally smaller than zero at a different point.

Given the $n+1$ points above, there is at least a point $\eta \in \Theta$ which is at a distance at most $r_{n+1}(\Theta, \delta)$ from all of them (otherwise, we obtain a cover of Θ with $n+1$ points). We can then construct the function

$$F(\theta) = \ -L\big(r_{n+1}(\Theta,\delta) - \delta(\theta,\eta)\big)_+ = -L \max\big\{r_{n+1}(\Theta,\delta) - d(\theta,\eta), 0\big\},$$

which is L-Lipschitz-continuous, equal to zero on all points of the algorithm and the output point $\hat{\theta}$, and with minimum value $-Lr_{n+1}(\Theta,\delta)$ attained at η. Thus, we must have $\varepsilon_n^{\rm opt} \geqslant 0 - (-Lr_{n+1}(\Theta,\delta)) = Lr_{n+1}(\Theta,\delta)$. This difficult function is plotted below in one dimension.



Thus, the performance of any algorithm from n function values has to be larger than $Lr_{n+1}(\Theta,\delta)$. Thus, so far, we have shown that

$$Lr_{n+1}(\Theta,\delta) \leqslant \varepsilon_n^{\rm opt} \leqslant Lr_n(\Theta,\delta).$$

For $\Theta \subset \mathbb{R}^d$, $r_n(\Theta,\delta)$ is typically of order $\operatorname{diam}(\Theta) n^{-1/d}$, and thus the difference between n and $n+1$ above is negligible. Note that the rate in $n^{-1/d}$ is *very* slow and symptomatic of the classical curse of dimensionality. The appearance of a covering number is not totally random here and comes from the equivalence in terms of worst-case guarantees between optimization and uniform approximation (Novak, 2006).

Random search. We can have a similar bound up to logarithmic terms for random search, that is, after selecting independently n points $\theta_1, \dots, \theta_n$, uniformly at random in Θ, and selecting the points with smallest function value $F(\theta_i)$. The performance can be shown to be proportional to $L\operatorname{diam}(\Theta)(\log n)^{1/d} n^{-1/d}$ in high probability, leading to an additional logarithmic term (the proof can be obtained with a simple covering argument, see exercise below). Therefore, random search is optimal up to logarithmic terms for optimizing this very large class of functions.

To go beyond Lipschitz-continuous functions, we can leverage smoothness like in supervised learning and hopefully avoid the dependence in $n^{-1/d}$. This can be done by a somewhat surprising equivalence between worst-case guarantees from optimization and worst-case guarantees for uniform approximation.[2]

Exercise 15.2 *(♦) Consider sampling independently and uniformly in Θ n points $\theta_1, \dots, \theta_n$.*

(a) For a given L-Lipschitz-continuous function F, show that the worst-case performance of outputting the lower function value is less than $L \max_{\theta \in \Theta} \min_{i \in \{1,\dots,n\}} \delta(\theta, \theta_i)$.

[2] See https://francisbach.com/optimization-is-as-hard-as-approximation/ for more details, as well as Novak (2006).

(b) Considering an optimal cover with m points and radius $r = r_m(\Theta, d)$, show that

$$\mathbb{P}\Big(\max_{\theta \in \Theta} \min_{i \in \{1,\dots,n\}} \delta(\theta, \theta_i) \geqslant 2r \Big) \leqslant m(1 - 1/m)^n.$$

(c) By the appropriate choice of m, show that when $r \sim m^{-1/d}\mathrm{diam}(\mathcal{X})$, we get an overall performance proportional to $L\big(\frac{\log n}{n}\big)^{1/d}$ with probability greater than $1 - \frac{\log n}{n}$.

15.3 Lower bounds for stochastic gradient descent (♦)

In this section, our goal is to show that the convergence rates for stochastic gradient descent (SGD) shown in Section 5.4 are "optimal", in a sense to be made precise. We consider a class $\mathcal{F}$ of functions, here the convex B-Lipschitz-continuous functions on the ball of center zero and radius D (for the Euclidean norm). We consider the class $\mathcal{A}$ of algorithms that can sequentially access independent random, unbiased estimates of the gradients of a function F in $\mathcal{F}$, with squared norm bounded by B^2. We denote $A_t(F) \in \mathbb{R}^d$ the output of algorithm A. Our goal is to find upper and lower bounds on

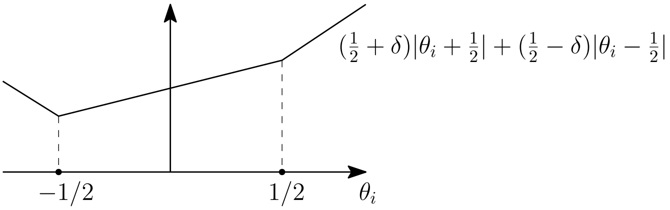
$$\varepsilon_t(\mathcal{A}, \mathcal{F}) = \inf_{A \in \mathcal{A}} \sup_{f \in \mathcal{F}} \mathbb{E}\big[F(A_t(F)) - \inf_{\|\theta\|_2 \leqslant D} F(\theta)\big].$$

SGD is an algorithm in $\mathcal{A}$ achieving a bound proportional to $BD/\sqrt{t}$, thus, up to a constant, $\varepsilon_t(\mathcal{A}, \mathcal{F}) \leqslant BD/\sqrt{t}$. We now prove a matching lower-bound by exhibiting a set of functions that will make any algorithm have this desired performance. Note that, as opposed to Section 15.2.1 on deterministic convex optimization, we make no assumption on the running-time complexity of algorithms in $\mathcal{A}$.

We follow the exposition from Agarwal et al. (2012) and consider a function

$$F_\alpha(\theta) = \frac{B}{2d} \sum_{i=1}^d \Big\{ \big(\frac{1}{2} + \alpha_i \delta\big) \cdot \big|\theta_i + \frac{1}{2}\big| + \big(\frac{1}{2} - \alpha_i \delta\big) \cdot \big|\theta_i - \frac{1}{2}\big| \Big\}, \tag{15.7}$$

with $\alpha \in \{-1, 1\}^d$ a well chosen vector and $\delta \in (0, 1/4]$, and $B > 0$. One element of the sum is plotted below.



The function F_α is convex and Lipschitz-continuous with gradients bounded in L_2-norm by $B/(2\sqrt{d})$. Moreover, the global minimizer of F_α is $\theta = -\frac{\alpha}{2}$, with an optimal value equal to $F_\alpha^* = \frac{B}{4}(1 - 2\delta)$. That is, minimizing F_α on $[-1/2, 1/2]^d$ exactly corresponds to finding an element of the hypercube α. Moreover, it turns out that minimizing it approximately also leads to identifying α among a set of α's, which are sufficiently different.

Lemma 15.3 *If $\alpha, \beta \in \{-1,1\}^d$, and $F_\alpha(\theta) - F_\alpha^* \leqslant \varepsilon$, then $F_\beta(\theta) - F_\beta^* \geqslant \frac{B\delta}{2d}\|\alpha-\beta\|_1 - \varepsilon$.*

Proof (♦) We have: $F_\beta(\theta) - F_\beta^* = F_\beta(\theta) + F_\alpha(\theta) - F_\beta^* - F_\alpha^* + [F_\alpha^* - F_\alpha(\theta)]$. We then notice that for all $\theta \in \mathbb{R}^d$,

$$F_\beta(\theta) + F_\alpha(\theta) - F_\beta^* - F_\alpha^* \geqslant \frac{B}{2d} \sum_{i,\ \alpha_i \neq \beta_i} \big\{ \big|\theta_i + \frac{1}{2}\big| + \big|\theta_i - \frac{1}{2}\big| + 2\delta - 1 \big\} \geqslant \frac{B\delta}{2d}\|\alpha - \beta\|_1.$$

■

Thus, if we consider M points $\alpha^{(1)}, \dots, \alpha^{(M)} \in \{-1,1\}^d$ such that $\|\alpha^{(i)} - \alpha^{(j)}\|_1 \geqslant \frac{d}{2}$ (with potentially $M \geqslant \exp(d/8)$ such points from Lemma 15.2), then, if $\varepsilon < \frac{B\delta}{8}$, because of Lemma 15.3, minimizing up to ε exactly identifies which of the functions $F_{\alpha^{(i)}}$ was being minimized.

Moreover, if $\hat\theta$ is random then, denoting $\mathcal{A} = \{\alpha^{(1)}, \dots, \alpha^{(M)}\}$, following the same reasoning as in Section 15.1.2:

$$\sup_{\alpha \in \mathcal{A}} \mathbb{E}_\alpha\big[F_\alpha(\hat\theta) - F_\alpha^*\big] \geqslant \varepsilon \cdot \sup_{\alpha \in \mathcal{A}} \mathbb{P}_\alpha\big(F_\alpha(\hat\theta) - F_\alpha^* > \varepsilon\big) \geqslant \varepsilon \cdot \frac{1}{|\mathcal{A}|} \sum_{\alpha \in \mathcal{A}} \mathbb{P}_\alpha\big(F_\alpha(\hat\theta) - F_\alpha^* > \varepsilon\big).$$

From an estimate $\hat\theta$, we can build a test $g(\hat\theta) \in \mathcal{A}$ by selecting the (unique if $\varepsilon < \frac{B\delta}{8}$) $\alpha \in \mathcal{A}$ such that $F_\alpha(\hat\theta) - F_\alpha^* \leqslant \varepsilon$ if it exists, and uniformly at random in $\mathcal{A}$ otherwise. Therefore, the minimax performance is greater than ε times the probability of a mistake in the best possible test.

We consider the following stochastic oracle:

(1) pick some coordinate $i \in \{1, \dots, d\}$ uniformly at random,

(2) draw a Bernoulli random variable $b_i \in \{0,1\}$ with parameter $\frac{1}{2} + \alpha_i\delta$,

(3) consider $\hat F(\theta) = b_i\big|\theta_i + \frac{1}{2}\big| + (1-b_i)\big|\theta_i - \frac{1}{2}\big|$, with gradient with components

$$\hat{F}'_\alpha(\theta)_i = \frac{B}{2}\big[b_i \operatorname{sign}(\theta_i + 1/2) + (1-b_i)\operatorname{sign}(\theta_i - 1/2)\big].$$

The stochastic gradients have an ℓ_2-norm bounded by B and are unbiased. Moreover, observation of the gradient for a $\theta \in [-1/2, 1/2]^d$ reveals the outcome of the Bernoulli random variable b_i.

Therefore, after t steps, we can apply Fano's inequality to the following set-up: the random variable $\alpha \in \mathcal{A}$ is uniform, and given α, we sample independently t times, one variable i in $\{1, \dots, d\}$ and observe (a potentially noisy version of) a Bernoulli random variable b, with parameter α_i.

We then need to upper bound the mutual information between α and (i,b) and multiply the result t times because each of the t gradients is sampled independently.

The mutual information can be decomposed as

$$I(\alpha, (i,b)) \quad = \quad I(\alpha, i) + I(\alpha, b|i) = 0 + \mathbb{E}_i\mathbb{E}_\alpha\big[D_{\mathrm{KL}}(p(b|i,\alpha)||p(b|i))\big],$$

where $p(b|i,\alpha)$ and $p(b|i)$ denotes the probability distribution of b. Thus, by convexity of the KL divergence,

$$\begin{aligned} I(\alpha,(i,b)) &= \mathbb{E}_i\mathbb{E}_\alpha\Big[D_{\mathrm{KL}}\Big(p(b|i,\alpha)\Big\|\frac{1}{|\mathcal{A}|}\sum_{\alpha'\in\mathcal{A}}p(b|i,\alpha')\Big)\Big] \\ &\leqslant \frac{1}{|\mathcal{A}|}\sum_{\alpha'\in\mathcal{A}}\mathbb{E}_i\mathbb{E}_\alpha\big[D_{\mathrm{KL}}(p(b|i,\alpha)||p(b|i,\alpha'))\big]. \end{aligned}$$

Since $p(b|i,\alpha)$ is Bernoulli random variable with parameter $\frac{1}{2}+\delta$ or $\frac{1}{2}-\delta$, the KL divergences above are bounded by the KL between two Bernoulli random variables with the two different parameters, that is,

$$\begin{aligned} I(\alpha,(i,b)) &\leqslant \big(\frac{1}{2}+\delta\big)\log\frac{\frac{1}{2}+\delta}{\frac{1}{2}-\delta}+\big(\frac{1}{2}-\delta\big)\log\frac{\frac{1}{2}-\delta}{\frac{1}{2}+\delta}=2\delta\log\frac{1+2\delta}{1-2\delta} \\ &= 2\delta\log\Big(1+\frac{4\delta}{1-2\delta}\Big)\leqslant\frac{8\delta^2}{1-2\delta}\leqslant 16\delta^2 \text{ if } \delta\in[0,1/4]. \end{aligned}$$

Therefore, applying Corollary 15.1, the minimax lower bound is greater than

$$\varepsilon\Big(1-\frac{16t\delta^2-\log 2}{\log M}\Big)\geqslant\varepsilon\Big(1-\frac{16t\delta^2-\log 2}{d/8}\Big).$$

Thus, we need $256t\delta^2\geqslant d$, and then $B\delta/4$ is the lower bound on the rate so that the lower bound is

$$\frac{1}{16}\sqrt{\frac{d}{t}},$$

which is the desired lower-bound (up to a constant) $\varepsilon_t(\mathcal{A},\mathcal{F})\geqslant DB/\sqrt{t}$ where D is the diameter of the set of θ. The lower-bound is thus the same as the upper-bound achieved by stochastic gradient descent in Chapter 5. The result above can be extended to strongly-convex problems (Agarwal et al., 2012).

15.4 Conclusion

This chapter was entirely dedicated to lower bounds of performance associated with the upper bounds presented in the rest of the book. Statistical lower bounds are obtained by reducing the learning problem to a hypothesis test where information theory is brought to bear. In comparison, optimization lower bounds are obtained by designing functions that are explicitly hard to optimize for the proposed computational model of combining gradients linearly.