Some exercise solutions

Michel Baes, Patrick Cheridito

December 20, 2018

Open sets and algebraically open sets Let E be a real vector space. A set $S \subseteq E$ is said to be *algebraically open* if the intersection of S and any straight line of E is open according to the induced topology on the line (recall that the empty set is itself open).

1. Assume that E is a normed space. Prove that if S is open, then it is algebraically open. Prove that the converse holds when E is finite dimensional and S is convex.
2. Find an example of a (obviously non-convex) set in $\mathbb{R}^2$ that is algebraically open but not open with the usual Euclidean topology of $\mathbb{R}^2$.

`Lecture_1/AlgebraicallyOpenSets`

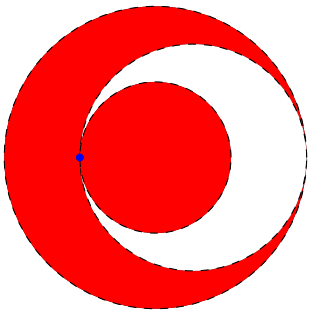
Open sets and algebraically open sets 1. The intersection of an open set and a straight line is open in that line. Indeed, let $S \subseteq E$ be an open set and $L = \{x_0 + tv : t \in \mathbb{R}\} \subseteq E$ be a line). Here x_0 is a point of L and v is a vector parallel to L of norm 1. If $S \cap L = \emptyset$, there is nothing to prove. Otherwise, let $a \in S \cap L$, and let $\epsilon > 0$ be such that the open ball B centered at a of radius ϵ is in S. Since $a \in L$, we can write $L = \{a + tv : t \in \mathbb{R}\}$. Now, $B \cap L = \{a + tv : ||a + tv - a|| = t||v|| = t < \epsilon\} = a+]-\epsilon, \epsilon[v$, proving that $S \cap L$ contains a open segment containing a.

In other words, every open set is algebraically open (even if it is non convex).

Assume now that S is convex and algebraically open, and that $E \equiv \mathbb{R}^n$. Let $a \in S$ and let $e_1, \ldots, e_n$ be n linearly independent vectors. As S is algebraically open, there exists $t_i^-, t_i^+ > 0$ such that S contains the open segment $I_i := a+]-t_i^- e_i, t_i^+ e_i[$. By convexity, the set S contains the convex hull of $\bigcup_{i=1}^n I_i$, which is an open set. Hence S is open.

2. We denote by $B(a, r)$ the open Euclidean ball with center a and radius r and by $B[a, r]$ its closure. Define

$$S = (B((0,0),2) \backslash B[(1/2,0),3/2]) \cup B((0,0),1) \cup \{(-1,0)\},$$



The set S is not open, because it is not possible to find any open ball centered at $p := (-1, 0)$ contained in S. Any line that intersects S which does not contain p is open, as proved in Item 1, because $S \backslash \{p\}$ is open. If a line L contains p, this line might be vertical, in which case $(L \cap S)$ is an open interval. Or this line is not vertical. Then its intersection with $B(0, 2) \backslash B[-1/2, 3/2]$ and $B(0, 1)$ consists of two intervals with, with p as end point. The union of these two half-open interval is an open interval.

Strong Separation Theorem Let us consider two *closed*, *convex* sets $A, B \subseteq \mathbb{R}^n$ with A *compact*.

(a) Show that $A - B$ is convex and closed.
Give a counterexample (with A, B disjoint, closed and convex), that compactness is needed in order to ensure that $A - B$ is closed.

(b) Suppose that A and B are disjoint. Show that those two sets can be *strongly* separated. (*Hints:* analogously to what we have done for the "open-set case" in the lecture, the proof consists again in two steps: first show that we can strongly separate $A - B$ from 0, and then use this strong separation to obtain the desired strong separation for A and B. For the strong separation of $A - B$ and 0, note that there exists an open ball centered in 0 and disjoint from $A - B$.)

`Lecture_2/ExtensionSeparationTheorem`

Strong Separation Theorem Let $A, B \subseteq \mathbb{R}^n$ be two disjoint, closed, convex sets with A compact (recall that in $\mathbb{R}^n$ compact is equivalent to closed and bounded, hence we just assume A to be bounded in addition).

(a) $A - B$ is clearly convex, since it is a Minkowski sum. To show that the set is also closed, we need to recall the topological definitions of closedness and compactness of a set $A \subseteq \mathbb{R}^n$:

- ◇ A set A is *closed* $\Leftrightarrow$ the limit of every converging sequence of elements of A also belongs to A.
- ◇ A set A is *compact* $\Leftrightarrow$ every sequence of elements of A has a convergent subsequence (whose limit is in A).

To show that $A - B$ is closed we therefore consider a convergent sequence $(x_i) := a_i - b_i$ in $A - B$ and show that its limit $\overline{x}$ also belongs to $A - B$.

Since A is compact, the sequence (a_i) has a convergent subsequence (a_{i_j}) with limit $\overline{a} \in A$. Let us consider the corresponding subsequence (b_{i_j}), i.e. $b_{i_j} = a_{i_j} - x_{i_j}$. Since by assumption we have the convergences $x_i \overset{i \to +\infty}{\longrightarrow} \overline{x}$ (hence $x_{i_j} \overset{i_j \to +\infty}{\longrightarrow} \overline{x}$) and $a_{i_j} \overset{i_j \to +\infty}{\longrightarrow} \overline{a}$, the sequence (b_{i_j}) also converges:

$$b_{i_j} = a_{i_j} - x_{i_j} \overset{i_j \to +\infty}{\longrightarrow} \overline{a} - \overline{x} =: \overline{b} \in B \text{ (this limit is in } B, \text{ since } B \text{ is closed).}$$

But then we are done, since from $\overline{a} \in A, \overline{b} \in B$ it follows that $\overline{x} = \overline{a} - \overline{b} \in A - B$.

As a counterexample showing that the compactness of A (i.e. closed and bounded in $\mathbb{R}^n$) is needed consider $A := \{x \in \mathbb{R}^2_{++} \mid x_2 \geq \frac{1}{x_1}\}$ and $B := \{x \in \mathbb{R}^2 \mid x_2 = 0\}$. The set $A - B$ is then the upper open halfspace $\{x \in \mathbb{R}^2 \mid x_2 > 0\}$, which is not closed.

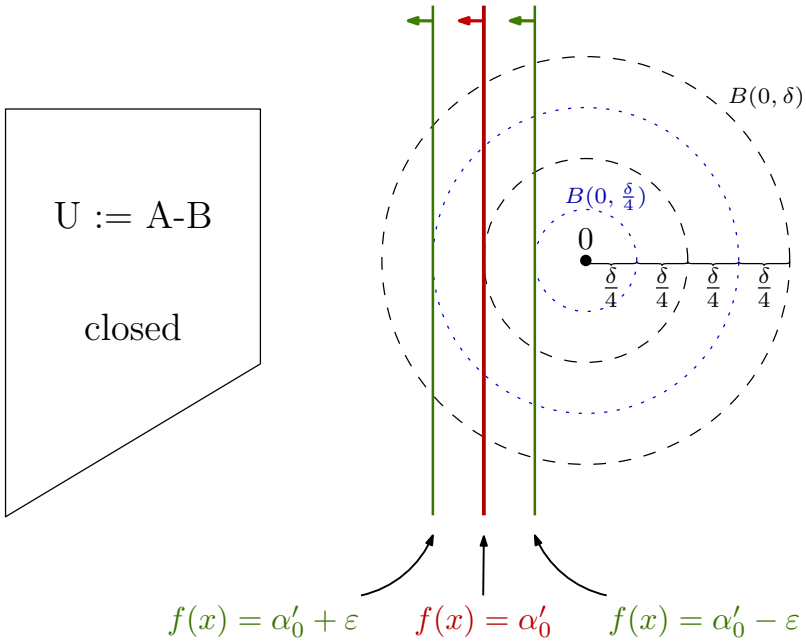
(b) We want here to find a strong separation of A and B, i.e. a linear function f and numbers $\alpha_0, \varepsilon \neq 0$ s.t. for each $\alpha \in]\alpha_0 - \varepsilon, \alpha_0 + \varepsilon[$ we have $f(b \in B) \leq \alpha \leq f(a \in A)$. Analogously to what we have done for the "open-set case" in the lecture, the proof consists again in two steps: first we show that we can strongly separate $A - B$ from 0 (since $A \cap B = \emptyset$ we have that $0 \notin A - B$), and then we use this separation to obtain the desired separation for A and B.

We show that $A - B$ can be strongly separated from 0: for this purpose, we could replicate the proof of "*Separating one point from an open convex set*" seen in the lecture using the Minkowski function μ_{A-B} of $A - B$, exploiting the fact that now $\mu_{A-B}(y \notin A - B) > 1$, since $A - B$ is closed.

However, let us consider the following alternative proof, which does not use Minkowski functions and, despite of the mathematical definitions, is somehow graphically easier to understand.

Since $A - B$ is closed (i.e. its complement is open) and $0 \notin A - B$, there exists a $\delta > 0$ s.t. the whole open ball $B(0, \delta)$ centered at 0 does not intersect $A - B$. Applying the result of "*Separating a convex set from an open convex set*" (where $A - B$ is the convex set and $B(0, \delta)$ the open set), we have a linear function f s.t. $f(x \in B(0, \delta)) < f(y \in A - B)$. Having a look at the picture below, let us now define:

$$\alpha_0' := \sup_{x \in B(0, \frac{\delta}{2})} f(x) \quad , \quad \alpha_0' - \varepsilon := \sup_{x \in B(0, \frac{\delta}{4})} f(x) \quad , \quad \alpha_0' + \varepsilon = \sup_{x \in B(0, \frac{3}{4}\delta)} f(x).$$



We hence get:

$$f\left(x \in B\left(0, \frac{\delta}{4}\right)\right) \leq \alpha \leq f(y \in A - B), \text{ for all } \alpha \in]\alpha_0' - \varepsilon, \alpha_0' + \varepsilon[, \qquad (*)$$

i.e. a strong separation of $A - B$ and 0 as desired.

We now turn this into a strong separation for A and B: for any pair $a \in A, b \in B$ build the point $y := a - b \in A - B$ and fix the point $x := 0 \in B(0, \frac{\delta}{4})$. By this choice and linearity of f, $(*)$ now reads:

$\forall a \in A, \forall b \in B$:

$$\begin{aligned}
& 0 = f(0) \leq \alpha \leq f(a-b) = f(a) - f(b), && \forall \alpha \in]\alpha_0' - \varepsilon, \alpha_0' + \varepsilon[\\
\implies\ & f(b) \leq \alpha + f(b) \leq f(a), && \forall \alpha \in]\alpha_0' - \varepsilon, \alpha_0' + \varepsilon[\\
\implies\ & f(b) \leq \alpha + \sup_{\tilde{b} \in B} f(\tilde{b}) \leq f(a), && \forall \alpha \in]\alpha_0' - \varepsilon, \alpha_0' + \varepsilon[\\
\implies\ & f(b) \leq \alpha \leq f(a), && \forall \alpha \in] \underbrace{\alpha_0' + \sup_{\tilde{b} \in B} f(\tilde{b}) - \varepsilon}_{=:\alpha_0}, \underbrace{\alpha_0' + \sup_{\tilde{b} \in B} f(\tilde{b}) + \varepsilon}_{\alpha_0}[
\end{aligned}$$

which is the desired strong separation of A and B.

Automatic backward differentiation [1] Assume that a function $f : \mathbb{R}^n \to \mathbb{R}$ is given as in Algorithm [1], where every step of the loop consists of a *simple* operation g_i. In general,

[1] This method is also known as the *Backpropagation Algorithm* in the context of Artificial Neural Networks

the operation g_i uses the data $(v_{1-n}, \cdots, v_0) \equiv (x_1, \ldots, x_n)$ and what is already computed $(v_1, ..., v_{i-1})$ to compute v_i, which can then be used in subsequent operations. The last operation computes v_m, which is our output $f(x)$. Usually, g_i takes only very few inputs and its differential can be computed in closed form.

Algorithm 1 Function evaluation

Input: $x \in \mathbb{R}^n$
Output: $f(x) \in \mathbb{R}$
Begin
  for $i := 1, 2, \ldots, n$,
    $v_{i-n} := x_i$
  end for
  for $i := 1, 2, \ldots, m$,
    $v_i := g_i(v_{1-n}, v_{2-n}, \ldots, v_{i-1})$
  end for
  $f(x) := v_m$
end

For differentiating this function, we can apply the backward mode of automatic differentiation described below in Algorithm [2]. The key to efficiency in this algorithm is, of course, that the inner loop (in the variable j) runs in reality only through the indices j of those variables v_j that actually influence v_i.

1. Prove that this algorithm indeed gives the gradient of f in x.
   Hint: you can represent the value of f at x in function of v in $m+1$ different ways. For instance, $f(x) = \Phi_{m+1}(v_{1-n}, \ldots, v_m) := v_m$, or $f(x) = \Phi_m(v_{1-n}, \ldots, v_{m-1}) := g_m(v_1, \ldots, v_{m-1})$, or $f(x) = \Phi_{m-1}(v_{1-n}, \ldots, v_{m-2}) := g_m(v_{1-n}, \ldots, v_{m-2}, g_{m-1}(v_{1-n}, \ldots, v_{m-2}))$, etc... Now, what does $\bar{v}_j$ represents at iteration i?
2. Use the Automatic Differentiation procedure to compute the gradient of the function implemented in `Nesterov_Chebysev.m`.
   Write and test a Matlab code and compare its output and computation time with the more standard procedure where you approximate each component of the gradient by $(f(x + te_i) - f(x))/t$, where $t = \sqrt{\epsilon_{\text{mach}}}$ and ϵ_{mach} is the machine precision of the floating-point system you are using. Alternatively, you can use Python or R, but you would need to rewrite `Nesterov_Chebysev.m` in the language you wish to use; note that vector components are numbered from 1 in Matlab and in R, but from 0 in Python.

`Lecture_3/AutomaticDifferentiation`

Automatic backward differentiation 1. Let us prove that at the end of loop i for $1 \leq i \leq m+1$, the variable $\bar{v}_j$ contains the value of

$$\frac{\partial \Phi_i(v_{1-n}, \ldots, v_{i-1})}{\partial v_j}$$

Algorithm 2 Function and gradient evaluation

Input: $x \in \mathbb{R}^n$
Output: $f(x) \in \mathbb{R}, \nabla f(x) \in \mathbb{R}^n$
Begin
//Function evaluation:
for $i := 1, 2, \ldots, n,$
$v_{i-n} := x_i$
end for
for $i := 1, 2, \ldots, m,$
$v_i := g_i(v_{1-n}, v_{2-n}, \ldots, v_{i-1})$
end for
$f(x) := v_m$
end //Adjoint derivative computation:
Begin
for $i := 1, 2, \ldots, n+m,$
$\bar{v}_{i-n} := 0$
end for
$\bar{v}_m := 1$
//Backwards loop:
for $i := m, m-1, \ldots, 1,$
for $j := -n+1, -n+2, \ldots, i-1,$
$\bar{v}_j := \bar{v}_j + \bar{v}_i \frac{\partial g_i}{\partial v_j}(v_{1-n}, v_{2-n}, \ldots, v_{i-1})$
end for
end for
for $i := 1, 2, \ldots, n,$
$\nabla f_i(x) := \bar{v}_{i-n}$
end for
end

for every $1-n \leq j \leq i-1$. When $i := m+1$, this assertion is trivial, as $\Phi_{m+1}(v_{1-n}, \ldots, v_m) = v_m$. Assuming that this assertion is true for $i+1$, let us show it for i. Since

$$\Phi_i(v_{1-n}, \ldots, v_{i-1}) = \Phi_{i+1}(v_{1-n}, \ldots, v_{i-1}, f_i(v_{1-n}, \ldots, v_{i-1})),$$

we have for all $1-n \leq j \leq i-1$:

$$\begin{aligned}\frac{\partial \Phi_i(v_{1-n}, \ldots, v_{i-1})}{\partial v_j} &= \frac{\partial \Phi_{i+1}(v_{1-n}, \ldots, v_{i-1}, f_i(v_{1-n}, \ldots, v_{i-1}))}{\partial v_j} \\ &\quad + \frac{\partial \Phi_{i+1}(v_{1-n}, \ldots, v_{i-1}, f_i(v_{1-n}, \ldots, v_{i-1}))}{\partial v_i} \cdot \frac{\partial f_i(v_{1-n}, \ldots, v_{i-1})}{\partial v_j} \\ &= \bar{v}_j + \bar{v}_i \cdot \frac{\partial f_i(v_{1-n}, \ldots, v_{i-1})}{\partial v_j},\end{aligned}$$

proving that $\bar{v}_j$ is correctly updated. Now, when $i = 1$, the function Φ_1 only depends on $(v_{1-n}, \ldots, v_0) = (x_1, \ldots, x_n)$ and is therefore equal to the function f. Thus, the components $(\bar{v}_{1-n}, \ldots, \bar{v}_0)$ are the components of the gradient of f.

2. Here is a possible Matlab implementation:

```
function [f,grad] = Nesterov_Chebysev_grad(x)
%
% INPUT
%%%%%%%%%
% x: n * 1 vector in [-1,1]^n
%%%%%%%%%
% OUTPUT
%%%%%%%%%
% f = ( x(1) - 1 )^2 * (( x(2) - T_3(x(1)) )^2+1)
%              * ... * (( x(n) - T_3(x(n-1)) )^2+1),
%     where T_3(x) = 4x^3-x is the third Chebysev polynomial of the first
%     kind. Note that this function is minimized in x = ones(n,1)
% grad = f'(x)

% Remark that each line of computation has its own name.

n = size(x,1); Chebysev = zeros(n-1,1); factors = zeros(n,1);
factors(1,1) = ( x(1,1) - 1 )^2; for i = 2:n
    Chebysev(i-1,1) = 4 * x(i-1,1)^3 - 3 * x(i-1,1);
    factors(i,1) = factors(i-1,1) * ((x(i,1) - Chebysev(i-1,1))^2+1);
end f = factors(n,1);


bar_Chebysev = zeros(n-1,1); bar_factors = zeros(n,1); bar_x =
zeros(n,1); grad = zeros(n,1);

bar_factors(n,1) = 1;

for i = n:-1:2,
%     factors(i,1) = factors(i-1,1) * ((x(i,1) - Chebysev(i-1,1))^2+1);
```

```
    bar_factors(i-1,1) = bar_factors(i-1,1) + ...
       bar_factors(i,1) * ((x(i,1) - Chebysev(i-1,1))^2+1);
    bar_x(i,1) = bar_x(i,1) + ...
       bar_factors(i,1) * factors(i-1,1) * 2 * (x(i,1) - Chebysev(i-1,1));
    bar_Chebysev(i-1,1) = bar_Chebysev(i-1,1) + ...
       bar_factors(i,1) * factors(i-1,1) * 2 * (Chebysev(i-1,1) - x(i,1));

%     Chebysev(i-1,1) = 4 * x(i-1,1)^3 - 3 * x(i-1,1);

    bar_x(i-1,1) = bar_x(i-1,1) + bar_Chebysev(i-1,1) * (12 * x(i-1,1)^2 - 3);
end

% factors(1,1) = ( x(1,1) - 1 )^2;
bar_x(1,1) = bar_x(1,1) + bar_factors(1,1) * 2 * (x(1,1) - 1);

grad = bar_x;
```

A nonconvex problem with strong duality Let $A \in \mathbb{S}^n$ and $b \in \mathbb{R}^n$. Consider the problem

$$\begin{array}{ll} \inf & x^T A x + b^T x \\ \text{s.t.} & x^T x - 1 \leq 0, \\ & x \in \mathbb{R}^n. \end{array} \tag{1}$$

Observe that this problem is convex for $A \in \mathbb{S}^n_+$, whereas for $A \notin \mathbb{S}^n_+$, it is nonconvex. Prove that for (1) the strong duality holds. We suggest you to follow these steps:

i) Prove the result for $A \in \mathbb{S}^n_+$.

ii) If A is not positive definite, prove that the primal optimal value does not change if we replace the inequality in the constraint by an equality.

iii) By observing, for any real α, that the equality

$$\min\{x^T A x + b^T x \ : \ x^T x = 1\} = -\alpha + \min\{x^T (A + \alpha I) x + b^T x \ : \ x^T x = 1\}$$

holds, prove the desired result (for A not necessarily positive semidefinite).

`Lecture_4/NonconvexProblemWithStrongDuality`

A nonconvex problem with strong duality

i) If A is positive semidefinite, then $x^T A x + b^T x$ is convex and $-x^T x$ is concave. More over the feasible set is compact. Hence the perturbation function is LSC and convex, i.e. strong duality holds.

ii) Let x be a optimal solution. Assume that $x^T x < 1$, then x must be a local minimum of $f(x) := x^T A x + b^T x$, with $\nabla_x f(x) = 2Ax + b = 0$ and $\nabla^2_x f(x) = 2A \succeq 0$, i.e. A is positive semidefinite.

iii) First consider the problem

$$\begin{array}{ll} \inf & x^T B x + b^T x \\ \text{s.t.} & x \in \mathbb{R}^n. \end{array}$$

with $B \in \mathbb{S}^n_{++}$. Then from the optimality conditions it follows that the optimal solution is attained at $x^\star$ if $\nabla_x((x^\star)^T B x^\star + b^T x^\star) = 2Bx^\star + b = 0$. Since B is strictly positive definite and therefore invertible, we can write $x^\star$ in terms of B and b:

$$x^\star = \frac{B^{-1}b}{2}.$$

Now assume $A \notin \mathbb{S}^n_+$. Let $\lambda^\star < 0$ be the smallest eigenvalue of A and consider the problem

$$\begin{aligned} -\alpha + \inf\ & x^T(A + \alpha I)x + b^T x \\ \text{s.t. } & x^T x = 1 \\ & x \in \mathbb{R}^n. \end{aligned} \tag{2}$$

with $\alpha > -\lambda^\star$. Note that then $A + \alpha I$ is strictly positive. If we wouldn't have any constraint the optimal solution $x^\star$ of (2) would be $x^\star = \frac{(A+\alpha I)^{-1}b}{2}$, or given a spectral decomposition $A = \sum_i \lambda_i v_i v_i^T$ of A,

$$x^\star = \frac{1}{2}\sum_i \frac{v_i^T b}{\lambda_i + \alpha} v_i.$$

Suppose that there exists an $\alpha^\star > -\lambda^\star$ for which $\|x^\star\| = 1$. Then (2) can be rewritten as:

$$\begin{aligned} -\alpha + \inf\ & x^T(A + \alpha I)x + b^T x \\ \text{s.t. } & x^T x \leq 1 \\ & x \in \mathbb{R}^n. \end{aligned}$$

for which strong duality holds because of (i) (since $A + \alpha^\star I \succ 0$). It remains to check that such an $\alpha^\star$ exists.

Consider the function

$$\Phi(\alpha) = \|(A + \alpha I)^{-1}b\|_2^2.$$

Let $\sum_i \lambda_i v_i v_i^T$ be the spectral decomposition of A. Then

$$\Phi(\alpha) = \sum_i \frac{(v_i^T b)^2}{(\lambda_i + \alpha)^2}.$$

Note that $\Phi(\alpha)$ is a continuous function. And furthermore

$$\lim_{\alpha \downarrow \lambda^\star} \Phi(\alpha) = \infty \text{ and } \lim_{\alpha \to \infty} \Phi(\alpha) = 0.$$

It follows that an $\alpha^\star$ must exist such that $\Phi(\alpha^\star) > 1$.

Strong convexity and Lipschitz continuity of the gradient Let $\omega : \mathbb{R} \to \mathbb{R}_+$ be a convex function that equals zero only in 0. A proper function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is *ω-strongly convex* if for every $x \in \text{relint}(\text{dom} f), y \in \text{dom} f$, we have:

$$f(y) - f(x) - \langle f'(x), y - x\rangle \geq \omega(||y - x||),$$

where $f'(x) \in \partial f(x)$. Show that the conjugate f_* satisfies:

$$f_*(v) - f_*(u) - \langle f'_*(u), v - u\rangle \leq \omega_*(||v-u||_*) \quad \forall u, v \in \mathrm{dom} f_*,$$

where ω_* is the conjugate of ω.

Note: this exercise proved to have important consequences. It is at the basis of *Smoothing Techniques*, which are now largely used to solve efficiently some large-scale nondifferentiable convex problems.

A frequently used particular case of the above exercise is given when we take $\omega(x) = \frac{\sigma}{2}||x||_2^2$ for some $\sigma > 0$: the conjugate of a σ-strongly convex function is a $\frac{1}{\sigma}$-smooth function (because $\omega^*(x) = \frac{1}{2\sigma}||x||_2^2$). The case $\omega(x) = \sum_i x_i \log(x_i)$ for $x_i \geq 0$ and $\sum_i x_i = 1$, $\omega(x) = +\infty$ otherwise plays an important role in Smoothing Techniques.

`Lecture_5/StrongConvexity`

Strong convexity and Lipschitz continuity of the gradient Let us check first that the problem $\max_x \langle s, x\rangle - f(x)$ has only one solution for every $s \in \mathbb{R}^n$. We show that $\langle s, x\rangle - f(x)$ is strictly concave, i.e. that $f(x)$ is strictly convex. Suppose on the contrary that there exists $x \neq y \in \mathrm{dom} f$ such that

$$f(x + \lambda(y-x)) = (1-\lambda)f(x) + \lambda f(y) = f(x) + \lambda\langle f'(x), y - x\rangle$$

for all $\lambda \in [0, 1]$. Then

$$0 = f(y) - f(x) - \langle f'(x), y - x\rangle \geq \omega(||y - x||),$$

and $||y - x|| = 0$, that is $x = y$, a contradiction.

As $\max_x \langle s, x\rangle - f(x)$ has only one solution x_s, according to the optimality conditions, $s = f'(x_s)$. In view of the duality correspondence between a function and its conjugate, $x_s = f'_*(s)$. and $f(f'_*(s)) + f_*(s) = \langle s, f'_*(s)\rangle$.

Let $u, v \in \mathrm{dom} f$ and $h := v - u$. We have:

$$\begin{aligned}
f_*(v) - f_*(u) - \langle f'_*(u), v - u\rangle &= \max_x \langle v, x\rangle - f(x) - f_*(u) - \langle f'_*(u), v - u\rangle \\
&= \max_x \langle u + h, x\rangle - f(x) - \langle u, f'_*(u)\rangle + f(f'_*(u)) - \langle f'_*(u), h\rangle \\
&= \max_x \langle u, x - f'_*(u)\rangle - f(x) + f(f'_*(u)) + \langle h, x - f'_*(u)\rangle \\
&\leq \max_x -\omega(||x - f'_*(u)||) + \langle h, x - f'_*(u)\rangle \\
&\leq \max_{t \geq 0} -\omega(t) + ||h||_* t \\
&= \omega_*(||h||_*) = \omega_*(||v - u||_*).
\end{aligned}$$

The first inequality is justified by $f'(f'_*(u)) = u$:

$$f(f'_*(u)) + \langle u, x - f'_*(u)\rangle - f(x) \leq -\omega_*(||x - f'_*(u)||).$$

Design of a cylindrical can The goal is to design a cylindrical can with height h and radius r such that the volume is at least V and the total surface area is minimal.

a) The problem can be posed as follows:

$$\begin{array}{ll} \min & 2\pi(r^2+rh) \\ \text{subject to} & \pi r^2 h \geq V \\ & r > 0 \\ & h > 0 \end{array}$$

Why is this problem not convex? How can this problem be transformed into a convex optimization problem? (Many different possibilities for treating the problem exist. One of them starts by taking *logarithms* of the variables.)

b) Solve the transformed convex problem, in order to give the optimal h and r in terms of V.

`Lecture_6/CylindricalCan`

Design of cylindrical can The proposed problem has a nonconvex objective function. Also, the constraint $\pi r^2 h \geq V$ is not convex. Defining $a = \ln(r)$ and $b = \ln(h)$, we can rewrite our problem as:

$$\begin{array}{l} \min\ \ln(\exp(2a) + \exp(a+b)) \\ \text{s. t. } \ln(\pi) + 2a + b \geq \ln(V). \end{array}$$

We have simply taken the logarithm of the objective function, which is now convex as verified below, and of the nonconvex constraint, which is now linear. The positivity constraints are now useless.

The convexity of the objective function comes from the convexity of the function $f : \mathbb{R}^n \to \mathbb{R}$ $x \mapsto f(x) = \ln\left(\sum_{j=1}^n \exp(x_j)\right)$, which can be deduced, for instance, by verifying that its Hessian is positive semidefinite.

We compute for any $x \in \mathbb{R}^n$:

$$\frac{\partial f(x)}{\partial x_i} = \frac{\exp(x_i)}{\sum_{j=1}^n \exp(x_j)},$$

then

$$\frac{\partial^2 f(x)}{\partial x_i^2} = \frac{\exp(x_i)\sum_{j=1}^n \exp(x_j) - \exp(x_i)\cdot\exp(x_i)}{\left(\sum_{j=1}^n \exp(x_j)\right)^2},$$

and

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{-\exp(x_i)\cdot\exp(x_j)}{\left(\sum_{j=1}^n \exp(x_j)\right)^2},$$

Denoting $v := f'(x)$, we observe that $\sum_{i=1}^n v_i = 1$, $v_i > 0$ for all i, and that $f''(x) = \mathrm{diag}(v) - vv^T$. To prove that $f''(x)$ is positive semidefinite, we take an arbitrary $u \in \mathbb{R}^n$ and compute:

$$u^T f''(x) u = \sum_{i=1}^n u_i^2 v_i - \left(\sum_{i=1}^n u_i v_i\right)^2.$$

The fact that this number is always nonnegative follows from Cauchy-Schwarz's inequality $\sum_{i=1}^n a_i^2 \cdot \sum_{i=1}^n b_i^2 \geq \left(\sum_{i=1}^n a_i b_i\right)^2$, applied to $a_i := u_i\sqrt{v_i}$ and $b_i := \sqrt{v_i}$ (remember that $\sum_{i=1}^n v_i = 1$).

We can now use the KKT conditions to solve this cylindrical can problem. We have:

$$1 + \frac{\exp(a^*)}{\exp(a^*) + \exp(b^*)} - 2\lambda = 0,$$

$$\frac{\exp(b^*)}{\exp(a^*) + \exp(b^*)} - \lambda = 0,$$

$$2a^* + b^* = \ln\left(\frac{V}{\pi}\right),$$

where the Lagrange multiplier λ must be nonnegative. Adding the two first identities shows that $\lambda = \frac{2}{3}$, so that $\exp(b^*) = 2\exp(a^*)$, or $b^* = a^* + \ln(2)$. The last identity implies $a^* = \ln\sqrt[3]{\frac{V}{2\pi}}$, and

$$r^* = \sqrt[3]{\frac{V}{2\pi}}, \quad h^* = 2\sqrt[3]{\frac{V}{2\pi}}:$$

an optimal can must be as tall as it is wide.

Sketch of an alternative solution Arguing that the inequality constraint $\pi r^2 h \geq V$ must be tight at the optimum, we deduce $h = \frac{V}{\pi r^2}$ at the optimum. Substituting h by this value in the optimum, we need to minimize $2\pi(r^2 + r^{-1})$, which is convex. It remains to determine for which r the gradient of this function vanishes.

Expected Shortfall as a linear optimization problem Let $x \in \mathbb{R}^d$ be a decision vector and Y a random vector taking values in $\mathbb{R}^m$, with probability density p. The loss incurred by the decision x under the circumstance y (a realization of the random vector Y) is given by $f(x, y)$. The set of all events that make us face a loss larger than L when we take the decision x is therefore $E(x, L) := \{y : f(x, y) \geq L\}$, and its probability is:

$$\psi(x, L) := \int_{E(x,L)} p(y)dy.$$

Observe that $\psi(x, L)$ decreases to 0 as L increases.

We assume that ψ is continuous in L for every x.

The *Value-at-Risk* of x at probability level $\alpha \in]0, 1[$ is defined in this notation as:

$$\mathrm{VaR}_\alpha(x) = \min\{L : \psi(x, L) \leq \alpha\}.$$

(Typically, α is a small number, such as 0.01 or 0.005). The expected shortfall at level $\alpha \in]0, 1[$ is defined as:

$$\mathrm{ES}_\alpha(x) := \frac{1}{\alpha}\int_{E(x,\mathrm{VaR}_\alpha(x))} f(x, y)p(y)dy.$$

This formula can be interpreted as follows: we *average* all losses that exceed $L = \mathrm{VaR}_\alpha(x)$. Note that then the correct probability distribution is $p(y)/\alpha$ and not $p(y)$, because we need its integral on its domain $[L, \infty[$ to equal 1.

Let

$$F_\alpha(x, L) := L + \frac{1}{\alpha} \int [f(x, y) - L]^+ p(y) dy.$$

Here, $[t]^+ := \max\{t, 0\}$.

1. Show that $L \mapsto F_\alpha(x, L)$ is convex on $\mathbb{R}$ for every x.
2. Define $G(L) := \int [f(x, y) - L]^+ p(y) dy$. Prove that the subgradient of G is $\partial G(L) := \{-\psi(x, L)\}$, and therefore that G is differentiable (and even continuously differentiable because $\psi(x, L)$ is continuous in L). Hint: the function $z \mapsto \varphi(z) := (z - \bar{L})^+ - (z - L)^+$ should appear naturally in your computations. Consider it for $\bar{L} > L$ first, then for $\bar{L} < L$; observe that an integral of the form $\int \varphi(z) p(z) dz$ can be split into three simple terms.
3. Prove now that $\mathrm{ES}_\alpha(x) = \min_{L \in \mathbb{R}} F_\alpha(x, L)$ and that $\mathrm{VaR}_\alpha(x)$ is a minimizer.
4. Verify that $\mathrm{ES}_\alpha(x) = F_\alpha(x, \mathrm{VaR}_\alpha(x))$.
   (It is trivial from the above, but it is nice to see all the ingredients united in a single simple formula.)
5. The integral in $F_\alpha(x, L)$ might be difficult to evaluate. Suggest a way to discretize the integral that makes the resulting optimization problem linear.
6. Suppose that $f(x, y)$ is also convex in x for every y. Prove that $F_\alpha(x, L)$ is convex in (x, L) and that $\mathrm{ES}_\alpha(x)$ is convex in x.

(Note: Rockafellar and Uryasev defined the function $F_\alpha(x, L)$ and studied its properties as above in their paper "Conditional value-at-risk for general loss distributions". Their results, rederived in this exercise, have had and still have a tremendous impact in Quantitative Risk Management.)

`Lecture_7/CVaRLinear`

Expected Shortfall as a linear optimization problem 1. To prove the convexity of F in L, let us first observe that the function $L \mapsto [f(x, y) - L]^+$ is convex for any $x \in \mathbb{R}^d, y \in \mathbb{R}^m$. Indeed, $[f(x, y) - L]^+ = \max\{f(x, y) - L, 0\}$ is the maximum of two linear functions. Let $L_0, L_1 \in \mathbb{R}$ and $0 \le \lambda \le 1$. Since $p(y) \ge 0$

$$((1 - \lambda)[f(x, y) - L_0]^+ + \lambda [f(x, y) - L_1]^+ - [f(x, y) - (1 - \lambda)L_0 + \lambda L_1]^+) p(y) \ge 0.$$

Integrating in y, dividing by $\alpha > 0$, we get that $L \mapsto F(x, L) - L$ is convex.

2. Let us fix $x \in \mathbb{R}^d$ and $L, \bar{L} \in \mathbb{R}$ with $\bar{L} > L$. Then

$$\begin{aligned}
G(\bar{L}) - G(L) &= \int_{E(x,\bar{L})} \left([f(x,y) - \bar{L}]^+ - [f(x,y) - L]^+\right) p(y)dy \\
&\quad + \int_{E(x,L)\setminus E(x,\bar{L})} \left([f(x,y) - \bar{L}]^+ - [f(x,y) - L]^+\right) p(y)dy \\
&\quad + \int_{\mathbb{R}^m \setminus E(x,L)} \left([f(x,y) - \bar{L}]^+ - [f(x,y) - L]^+\right) p(y)dy \\
&= \int_{E(x,\bar{L})} (L - \bar{L}) p(y)dy + \int_{E(x,L)\setminus E(x,\bar{L})} (f(x,y) - \bar{L}) p(y)dy \\
&= \psi(x,\bar{L})(L - \bar{L}) + \int_{E(x,L)\setminus E(x,\bar{L})} (f(x,y) - \bar{L}) p(y)dy \\
&\geq \psi(x,\bar{L})(L - \bar{L}) + \int_{E(x,L)\setminus E(x,\bar{L})} (L - \bar{L}) p(y)dy = \psi(x,L)(L - \bar{L}).
\end{aligned}$$

If $\bar{L} < L$,

$$\begin{aligned}
G(\bar{L}) - G(L) &= \int_{E(x,L)} \left([f(x,y) - \bar{L}]^+ - [f(x,y) - L]^+\right) p(y)dy \\
&\quad + \int_{E(x,\bar{L})\setminus E(x,L)} \left([f(x,y) - \bar{L}]^+ - [f(x,y) - L]^+\right) p(y)dy \\
&\quad + \int_{\mathbb{R}^m \setminus E(x,\bar{L})} \left([f(x,y) - \bar{L}]^+ - [f(x,y) - L]^+\right) p(y)dy \\
&= \int_{E(x,L)} (L - \bar{L}) p(y)dy + \int_{E(x,\bar{L})\setminus E(x,L)} (f(x,y) - L) p(y)dy \\
&= \psi(x,L)(L - \bar{L}) + \int_{E(x,\bar{L})\setminus E(x,L)} (f(x,y) - L) p(y)dy \\
&\geq \psi(x,L)(L - \bar{L}).
\end{aligned}$$

Hence $-\psi(x,L) \in \partial G(L)$. It remains to prove that no other scalar belongs to $\partial G(L)$. Let $v \in \partial G(L)$. For every $\bar{L} > L$

$$G(\bar{L}) - G(L) = \psi(x,\bar{L})(L - \bar{L}) + \int_{E(x,L)\setminus E(x,\bar{L})} (f(x,y) - \bar{L}) p(y)dy \geq v(\bar{L} - L);$$

then $-\psi(x,\bar{L}) \geq v$. Similarly, with $L > \bar{L}$, we obtain $-\psi(x,\bar{L}) \leq v$. Thus $v = -\psi(x,\bar{L})$, so $\partial G(L) = \{-\psi(x,L)\}$, and, as it is a singleton, G is differentiable.

3. We fix $x \in \mathbb{R}^d$. The optimality condition to the problem $\inf_{L\in\mathbb{R}} F_\alpha(x,L)$ reads $\frac{\partial F(x,L^*)}{\partial L} = 0$, or $1 - \frac{\psi(x,L^*)}{\alpha} = 0$, or $\psi(x,L^*) = \alpha$, which is satisfied for $L^* = \mathrm{VaR}_\alpha(x)$.
On the one hand, by definition:

$$\mathrm{ES}_\alpha(x) := \frac{1}{\alpha} \int_{E(x,\mathrm{VaR}_\alpha(x))} f(x,y)p(y)dy.$$

On the other hand:

$$\begin{aligned} F_\alpha(x, \mathrm{VaR}_\alpha(x)) &= \mathrm{VaR}_\alpha(x) + \frac{1}{\alpha}\int [f(x,y) - \mathrm{VaR}_\alpha(x)]^+ p(y)dy \\ &= \mathrm{VaR}_\alpha(x) + \frac{1}{\alpha}\int_{E(x,\mathrm{VaR}_\alpha(x))} (f(x,y) - \mathrm{VaR}_\alpha(x))\, p(y)dy \\ &= \mathrm{VaR}_\alpha(x) + \frac{1}{\alpha}\int_{E(x,\mathrm{VaR}_\alpha(x))} f(x,y)p(y)dy - \mathrm{VaR}_\alpha(x)\frac{\psi(x,\mathrm{VaR}_\alpha(x))}{\alpha} \\ &= \frac{1}{\alpha}\int_{E(x,\mathrm{VaR}_\alpha(x))} f(x,y)p(y)dy. \end{aligned}$$

Therefore, $\mathrm{ES}_\alpha(x) = \inf_{L\in\mathbb{R}} F_\alpha(x, L)$.

4. We have proved above that $\mathrm{ES}_\alpha(x) = F_\alpha(x, \mathrm{VaR}_\alpha(x))$.
5. Many possibilities exist to discretize the integral $\int [f(x,y) - L]^+ p(y)dy$. The simplest consists in sampling N points $\{y_1, \ldots, y_N\}$ according to the probability density p and to approximate the integral by

$$\int [f(x,y) - L]^+ p(y)dy \approx \frac{1}{N}\sum_{i=1}^{N} [f(x,y_i) - L]^+.$$

As proved above, computing $\mathrm{ES}_\alpha(x)$ is equivalent to minimize $L \mapsto F_\alpha(x, L)$ over $\mathbb{R}$. With our approximation of the integral, this reduces to:

$$\mathrm{ES}_\alpha(x) \approx \min_{L\in\mathbb{R}} L + \frac{1}{N\alpha}\sum_{i=1}^{N} [f(x,y_i) - L]^+$$

It remains to express the piecewise linear functions $L \mapsto [f(x,y_i) - L]^+$ in the form of a linear problem. Observe that $[f(x,y_i) - L]^+ \le t_i$ if and only if $f(x,y_i) - L \le t_i$ and $0 \le t_i$. Thus

$$\begin{aligned} \mathrm{ES}_\alpha(x) \approx \min\ & L + \tfrac{1}{N\alpha}\textstyle\sum_{i=1}^{N} t_i \\ \text{s.t. } & t_i + L \ge f(x,y_i), \qquad \text{for } 1 \le i \le N \\ & t_i \ge 0, \qquad \text{for } 1 \le i \le N. \end{aligned}$$

The objective is linear in the variables $L, t_1, \ldots, t_N$, as well as the constraints. So we have a linear optimization problem.

6. The joint convexity of $F_\alpha(x, L)$ is immediate form the convexity of $(x, L) \mapsto [f(x,y) - L]^+$ for every fixed y. It remains to use the nonnegativity of $p(y)$ and the linearity of the integral as in Item 1 to get the convexity of $(x, L) \mapsto F_\alpha(x, L)$.
We can deduce the convexity of $\mathrm{ES}_\alpha(x)$ using directly the definition of convexity. Let $x_0, x_1 \in \mathbb{R}^d$ and $\lambda \in [0, 1]$. We denote $x_\lambda := (1-\lambda)x_0 + \lambda x_1$. We have:

$$\begin{aligned} (1-\lambda)\mathrm{ES}_\alpha(x_0) + \lambda\mathrm{ES}_\alpha(x_1) &= (1-\lambda)F_\alpha(x_0, \mathrm{VaR}_\alpha(x_0)) + \lambda F_\alpha(x_1, \mathrm{VaR}_\alpha(x_1)) \\ &\ge F_\alpha(x_\lambda, (1-\lambda)\mathrm{VaR}_\alpha(x_0) + \lambda\mathrm{VaR}_\alpha(x_1)) \\ &\ge \inf_L F_\alpha(x_\lambda, L) = \mathrm{ES}_\alpha(x_\lambda). \end{aligned}$$

Thus $\mathrm{ES}_\alpha(x)$ is a convex function of x.

Bounding portfolio risk with incomplete covariance information We have to manage a portfolio of n assets. The quantity x_i denotes the amount of asset i we currently own. This quantity can be negative in case of short-selling. We would like to have more information on the risk of our positions. In the classical Markowitz setting, the risk of our portfolio is described by the quantity $x^T \Sigma x$, where Σ is the $n \times n$ correlation matrix between the different assets.

Suppose that we know the matrix Σ only (very) partially. We know its diagonal elements Σ_{ii}, which are the squares of the price volatilities of the assets. The off-diagonal entries of Σ are only known by their sign (or, in some cases, this sign is not even known). Intuitively, if $\Sigma_{ij} > 0$, then the price of assets i and j will tend to rise and fall together, like the price of oil and the price of an Exxon share.

Write a piece of Matlab code that computes the worst-case variance of our portfolio return and the corresponding matrix Σ. Compare this risk to the one where the matrix Σ is diagonal.

Test your code with the following instance:

$$x = \begin{pmatrix} 0.1 \\ 0.2 \\ -0.05 \\ 0.1 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 0.2 & + & + & \\ + & 0.1 & - & - \\ + & - & 0.3 & + \\ & - & + & 0.1 \end{pmatrix}.$$

Empty spots in Σ mean that the sign of the corresponding coefficient is unspecified.

`Lecture_8/SedumiIncompleteCovariance`

Bounding portfolio risk with incomplete covariance information

Modeling Let us call S the unknown covariance matrix. Note that S is symmetric and positive semidefinite.

We need to maximize $x^T S x = \langle x x^T, S \rangle_F$, a linear function of the variable S, subject to the diagonal constraints $S_{ii} = \Sigma_{ii}$, and the sign constraints $S_{ij} \gtrless 0$.

For fixing notation, let us denote by U an $n \times n$ matrix containing $-1, 0$, and 1; if $U_{ij} = -1$, then S_{ij} is required to be non-positive, if $U_{ij} = 1$, we want that $S_{ij} \geq 0$, and $U_{ij} = 0$ if the sign of S_{ij} is unspecified. By convention, we will also assume that the diagonal elements of U are null.

Let $I_\pm := \{\{i, j\} : U_{ij} = \pm 1\}$. We can model our optimization problem as:

$$\begin{array}{ll} \max & \langle x x^T, S \rangle_F \\ \text{s.t.} & S_{ii} = \Sigma_{ii} \quad 1 \leq i \leq n \\ & S_{ij} \geq 0 \quad \{i, j\} \in I_+ \\ & S_{ij} \leq 0 \quad \{i, j\} \in I_- \\ & S \in \mathbb{S}^n_+. \end{array}$$

It remains to introduce a few variables to separate the semidefinite constraints from the

sign constraints:

$$\begin{aligned} \max\ & \langle xx^T, S\rangle_F \\ \text{s.t. } & S_{ii} = \Sigma_{ii} \quad 1 \le i \le n \\ & S_{ij} - u_{ij} = 0 \quad \{i,j\} \in I_+ \\ & S_{ij} + v_{ij} = 0 \quad \{i,j\} \in I_- \\ & u_{ij} \ge 0 \quad \{i,j\} \in I_+ \\ & v_{ij} \le 0 \quad \{i,j\} \in I_- \\ & u, v \ge 0, \quad S \in \mathbb{S}^n_+. \end{aligned}$$

Code To avoid "for" loops, which slow down Matlab, we construct the matrix A_{sedumi} of our problem using the Matlab function `sparse` with appropriate indices vectors. DO NOT FORGET that Sedumi minimizes functions.

```
function S = Incomplete_Covariance(x,Sigma,U)
% function S = Incomplete_Covariance(x,Sigma,U)
%%%%%%%%%%
% INPUT
%%%%%%%%%%
% x: n * 1 vector: positions on each asset in the portfolio
% Sigma: n * 1 vector: squared volatilities
%                      of each asset in the portfolio
% U: n * n matrix: desired sign pattern of S
% OUTPUT
%%%%%%%%%%
% S: worse-case covariance matrix for the portfolio,
%    satifying diagonal and sign constraints

n = size(x,1);
[i_plus,j_plus] = find(U==1);
[i_minus,j_minus] = find(U==-1);
n_plus = size(i_plus,1);
n_minus = size(i_minus,1);

col_u_plus = 0;
col_u_minus = col_u_plus + n_plus;
col_S = col_u_minus + n_minus;
n_cols = col_S + n^2;
n_rows = n + n_plus + n_minus;

AA = spalloc(n_rows,n_cols,n + 2*(n_plus + n_minus));
bb = zeros(n_rows,1);
cc = zeros(n_cols,1);
KK.l = n_plus + n_minus;
KK.s = n;

cc(col_S+1:col_S+n^2,1) = reshape(-x*x',n^2,1);
```

```
row = 0;

% Constraint S(i,i) = Sigma(i,1)
all_int = [1:n]';
diag_int = (all_int-1)*n + all_int;
AA(row+1:row+n, col_S+1:col_S+n^2) = sparse(all_int,diag_int,ones(n,1));
bb(row+1:row+n,1) = Sigma;
row = row + n;

% Constraints S(i,j) - u_plus(i,j) = 0
plus_int = n * (i_plus-1) + j_plus;
AA(row+1:row+n_plus, col_u_plus+1:col_u_plus+n_plus) = -speye(n_plus);
AA(row+1:row+n_plus, col_S+1:col_S+n^2) =...
sparse([1:n_plus]',plus_int,ones(n_plus,1),n_plus,n^2);
row = row + n_plus;

% Constraints S(i,j) + u_minus(i,j) = 0
minus_int = n * (i_minus-1) + j_minus;
AA(row+1:row+n_minus, col_u_minus+1:col_u_minus+n_minus) = ...
speye(n_minus);
AA(row+1:row+n_minus, col_S+1:col_S+n^2) = ...
sparse([1:n_minus]',minus_int,ones(n_minus,1),n_minus,n^2);
row = row + n_minus;

xx = sedumi(AA,bb,cc,KK);
S = mat(xx(col_S+1:col_S+n^2,1));
```

Efficiency of the Maximum Volume Inscribed Ellipsoid In this exercise we want to prove the following geometrical result. Suppose that P is a polyhedron in $\mathbb{R}^n$, bounded, symmetric around the origin and described as

$$P = \left\{x \mid -1 \leq a_i^T x \leq 1, i = 1, \ldots, p\right\}.$$

Let

$$\mathcal{E} = \left\{x \mid x^T Q^{-1} x \leq 1\right\},$$

where $Q \in \mathbb{S}^n_{++}$, be the maximum volume ellipsoid centered in the origin, inscribed in P.

Theorem: The ellipsoid $\sqrt{n}\mathcal{E} = \left\{x \mid x^T Q^{-1} x \leq n\right\}$ (i.e. the ellipsoid $\mathcal{E}$ scaled by a factor $\sqrt{n}$ around the origin) then *contains* P.

To prove this, proceed as follows:

(a) Show the equivalent characterization: $\mathcal{E} \subseteq P \iff a_i^T Q a_i \leq 1, \forall i = 1, \ldots, p$.

(b) It is well-known that the volume of $\mathcal{E}$ is proportional to $\sqrt{\det Q}$, so we can find the maximum volume ellipsoid $\mathcal{E}$ inside P by solving the following convex optimization problem:

$$\begin{array}{ll} \min & \log\left(\det\left(Q^{-1}\right)\right) \\ \text{s.t.} & a_i^T Q a_i \leq 1, \forall i = 1, \ldots, p \end{array}$$

The variable is the matrix $Q \in \mathbb{S}^n$ and the domain of the objective function is $\mathbb{S}^n_{++}$. Derive the Lagrange dual problem $\max_{u \succcurlyeq 0} \min_{Q \succ 0} L(Q, u)$.

(*Hint:* the gradient of the convex objective function $Q \mapsto \log\left(\det\left(Q^{-1}\right)\right)$ is given by $-Q^{-1}$).

(c) Note that Slater's condition holds for the primal problem above (e.g. $a_i^T Q a_i < 1$ for $Q = \varepsilon I, \varepsilon > 0$ small enough), so we have strong duality, and the KKT conditions are necessary and sufficient for optimality. What are the KKT conditions for this primal problem?
Suppose Q^* is optimal (i.e. describing our maximal volume ellipsoid $\mathcal{E} \subseteq P$). Use the KKT conditions to show that $x \in P \Longrightarrow x^T (Q^*)^{-1} x \leq n$, which is the desired result.

`Lecture_9/MaximumVolumeInscribedEllipsoid`

Efficiency of the Maximum Volume Inscribed Ellipsoid (a) $x \in \mathcal{E}$ iff $x = Q^{1/2} y$ for some y with $\|y\| \leq 1$, so we have $\mathcal{E} \subseteq P$ iff

$$\|y\| \leq 1 \Rightarrow -1 \leq a_i^T Q^{1/2} y \leq 1, i = 1, \ldots, p.$$

In other words, for $i = 1, \ldots, p$:

$$\sup_{\|y\| \leq 1} a_i^T Q^{1/2} y = \left\| Q^{1/2} a_i \right\| \leq 1, \quad \inf_{\|y\| \leq 1} a_i^T Q^{1/2} y = -\left\| Q^{1/2} a_i \right\| \geq -1,$$

i.e. $a_i^T Q a_i = \left\| Q^{1/2} a_i \right\|^2 \leq 1$.

(b) We first notice that $\log(\det(Q^{-1})) = \log(\det(Q)^{-1}) = -\log(\det(Q))$ is a convex function, and by minimizing it we are indeed maximizing $\log(\det(Q))$, which is an increasing function in $\det(Q)$. Hence, the given optimization problem is really looking for a Q *maximizing* the volume of $\mathcal{E}$, subject to $\mathcal{E} \subseteq P$.
The dual function is

$$h(u) := \min_{Q \succ 0} L(Q, u) = \min_{Q \succ 0} \left\{ \log(\det(Q^{-1})) + \sum_{i=1}^{p} u_i (a_i^T Q a_i - 1) \right\}.$$

Since $L(Q, u)$ is convex in Q, we minimize it by setting $\nabla_Q L(Q^*, u) \stackrel{!}{=} 0$ (the matrix-gradient ∇_Q is just like any usual vector-gradient: we compute its entries by $\partial / \partial q_{i,j}$). Using the gradient given in the hint and noting that $a_i^T Q a_i = \mathrm{Tr}(Q a_i a_i^T)$ has gradient $a_i a_i^T$, we obtain the necessary and sufficient condition

$$-(Q^*)^{-1} + \sum_{i=1}^{p} u_i a_i a_i^T = 0.$$

Since every positive definite matrix is invertible and its inverse is also positive definite, we have:

$$h(u) = L(Q^*, u) = \begin{cases} \log\left(\det\left(\sum_{i=1}^p u_i a_i a_i^T\right)\right) + n - \sum_{i=1}^p u_i, & \text{if } \sum_{i=1}^p u_i a_i a_i^T \succ 0 \\ -\infty, & \text{otherwise.} \end{cases}$$

The resulting dual problem is:

$$\begin{array}{ll} \max & \log\left(\det\left(\sum_{i=1}^p u_i a_i a_i^T\right)\right) + n - \sum_{i=1}^p u_i \\ \text{s.t.} & u \succeq 0. \end{array}$$

(c) The KKT conditions for the primal problem are:

(i) *(Feasibility)* $Q^* \succ 0, \quad a_i^T Q^* a_i \leq 1,\ i = 1, \ldots, p$

(ii) *(KKT Conditions)* $(Q^*)^{-1} = \sum_{i=1}^p u_i^* a_i a_i^T, \quad u^* \succeq 0$

(iii) *(Complementarity)* $u_i^*(1 - a_i^T Q^* a_i) = 0,\ i = 1, \ldots, p$

We now multiply equation (ii) with Q on the left and then take the trace:

$$n = \mathrm{Tr}(Q^*(Q^*)^{-1}) = \sum_{i=1}^p u_i^* \mathrm{Tr}(Q^* a_i a_i^T) = \sum_{i=1}^p u_i^* a_i^T Q^* a_i = \sum_{i=1}^p u_i^*$$

(the last equality comes from (iii): $a_i^T Q^* a_i = 1$ when $u_i^* \neq 0$). Finally we note that (ii) also implies that

$$\underbrace{\left|a_i^T x\right| \leq 1, i = 1, \ldots, p}_{x \in P} \Longrightarrow \underbrace{x^T (Q^*)^{-1} x = \sum_{i=1}^p u_i^* x^T a_i a_i^T x = \sum_{i=1}^p u_i^* (a_i^T x)^2 \leq \sum_{i=1}^p u_i^* = n\,.}_{x \in \sqrt{n}\mathcal{E}}$$

The Golden Search We want to minimize the convex function $f : [a, b] \to \mathbb{R}$. We have at our disposal an oracle of order 0: given an input point in $[a, b]$, this oracle only returns the value of the f at that point, but none of its subgradients. We assume that f has only one minimizer x^* on $[a, b]$. We compare in this exercise two minimization algorithms: the trisection method, which intuitively seems the most natural, and the golden section method, which performs considerably better.

1. Consider Algorithm 1.

Algorithm 3 TrisectionMethod$(f, [a, b])$

$a_0 := a$, $b_0 := b$, $\Delta_0 := [a_0, b_0]$.
for $k \geq 0$
- Compute $f_k^- := f\left(\frac{2a_k}{3} + \frac{b_k}{3}\right)$ and $f_k^+ := f\left(\frac{a_k}{3} + \frac{2b_k}{3}\right)$.
- **if** $f_k^- \leq f_k^+$, set $a_{k+1} := a_k$ and $b_{k+1} := \frac{a_k}{3} + \frac{2b_k}{3}$.
  else set $a_{k+1} := \frac{2a_k}{3} + \frac{b_k}{3}$ and $b_{k+1} := b_k$.
  end if
- Set $\Delta_{k+1} := [a_{k+1}, b_{k+1}]$.

end for

Show that this algorithm converges linearly with convergence ratio $\sqrt{2/3}$: prove that after N function evaluations, we have $|x_N - x^*| \leq \left(\frac{2}{3}\right)^{\lfloor N/2 \rfloor} \cdot |b - a|$, where x_N is an arbitrary point in $\Delta_{\lfloor N/2 \rfloor}$. Why is the above method obviously not the optimal one?

Algorithm 4 GoldenSectionMethod$(f, [a, b])$

$a_0 := a$, $b_0 := b$, $\Delta_0 := [a_0, b_0]$, $\lambda := \frac{\sqrt{5}-1}{2}$.
for $k \geq 0$
- Compute $f_k^- := f(\lambda a_k + (1-\lambda)b_k)$ and $f_k^+ := f((1-\lambda)a_k + \lambda b_k)$.
- **if** $f_k^- \leq f_k^+$, set $a_{k+1} := a_k$ and $b_{k+1} := (1-\lambda)a_k + \lambda b_k$.
- **else** set $a_{k+1} := \lambda a_k + (1-\lambda)b_k$ and $b_{k+1} := b_k$.
- **end if**
- Set $\Delta_{k+1} := [a_{k+1}, b_{k+1}]$.

end for

2. Let $\lambda := \frac{\sqrt{5}-1}{2}$ be the *golden number*, for which we have $\lambda + 1 = \frac{1}{\lambda}$, and consider Algorithm 2.

   Show that, except at iteration $k = 0$, one of the two values $f_k^- := f(\lambda a_k + (1-\lambda)b_k)$ and $f_k^+ := f((1-\lambda)a_k + \lambda b_k)$ has already been computed (i.e. coincides with some f_j^+ or f_j^- for a $j < k$).

   Show that this algorithm converges linearly with convergence ratio λ: prove that after $N \geq 2$ function evaluations, we have $|x_N - x^*| \leq \lambda^{N-1}|b-a|$, where x_N is an arbitrary point in Δ_{N-1}.

`Lecture_10/GoldenSection`

The Golden Search 1. First, for every iteration k, we have $x^* \in \Delta_k$. To show this, we set $x_k^- := 2a_k/3 + b_k/3$ and $x_k^+ := a_k/3 + 2b_k/3$. Suppose that $f_k^- = f(x_k^-) \leq f_k^+ = f(x_k^+)$, and that there exists $x \geq x_k^+$ for which $f(x) < f(x_k^-)$. Let $\lambda := \frac{x_k^+ - x_k^-}{x - x_k^-} > 0$, so that $x_k^+ = \lambda x + (1-\lambda)x_k^-$. By convexity of f, we have:

$$f(x_k^+) \leq \lambda f(x) + (1-\lambda) f(x_k^-) < \lambda f(x_k^+) + (1-\lambda) f(x_k^-) \leq f(x_k^+),$$

a contradiction. If $f_k^- \geq f_k^+$, we proceed symmetrically.

At every iteration, we perform two evaluations of f. After iteration k, we have computed $2k$ values of f. Also, at every iteration, the length of the interval is scaled by $2/3$. Thus, the length of the interval Δ_k is $|b-a|(2/3)^k$, and $|x_{2k} - x^*| \leq |b-a|(2/3)^k$, which proves the linear convergence of the algorithm with convergence ratio $\sqrt{2/3} \approx 0.8165$.

A major drawback of this method is that we compute many evaluations of f that are not used anymore afterwards.

2. Notice that Algorithm 2 is exactly Algorithm 1 except we replaced $2/3$ with the golden number λ. Therefore, we can prove exactly as above that $x^* \in \Delta_k$ for every iteration k and that the length of Δ_k equals $\lambda^k |b-a|$.

   It remains to show that either f_k^- or f_k^+ has already been computed at a previous stage. After iteration $k-1$, we have at our disposal the value of f at points $x_{k-1}^- := \lambda a_{k-1} + (1-\lambda)b_{k-1}$ and $x_{k-1}^+ := (1-\lambda)a_{k-1} + \lambda b_{k-1}$. Suppose that $a_k = a_{k-1}$ and

therefore $b_k = x_{k-1}^+$. Then:

$$\begin{aligned} x_k^+ &:= (1-\lambda)a_k + \lambda b_k = (1-\lambda)a_{k-1} + \lambda\Big((1-\lambda)a_{k-1} + \lambda b_{k-1}\Big) \\ &= (1-\lambda^2)a_{k-1} + \lambda^2 b_{k-1} = \lambda a_{k-1} + (1-\lambda)b_{k-1}, \end{aligned}$$

this last equality comes from the fact that $\lambda^2 = 1 - \lambda$. Therefore, only one function evaluation is needed for each iteration $k \geq 1$. After iteration $k \geq 1$, the function f has been evaluated in exactly $k+2$ points, whereas the desired result. Note that the convergence ratio is $\lambda \approx 0.6180 < \sqrt{2/3}$.

An optimal gradient method for smooth convex problems *(bonus exercise)* This method is due to Arkadi Nemirovski (private communication). It might have been published in Russian in the eighties.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex and differentiable function with a Lipschitz continuous gradient, i.e. there exists a constant $L > 0$ for which every point $x, y \in \mathbb{R}^n$ satisfies

$$||f'(x) - f'(y)||_2 \leq L||x - y||_2.$$

We would like to minimize the function f over $\mathbb{R}^n$. We assume that $f^* := \min\{f(x) : x \in \mathbb{R}^n\}$ is finite and that $R := \min\{||x||_2 : f(x) = f^*\}$ is finite as well, or in other words that there exists a minimizer in the ball $B[0, R]$.

A minimizing sequence full of properties Let $x_0 = 0, x_1, x_2, \ldots$ be a sequence of points of $\mathbb{R}^n$ that satisfy the following three properties. Here, we abbreviate $f'(x_t)$ by g_t.

1. $\langle g_t, x_t \rangle = 0$ for every $t \geq 0$.
2. There exists a sequence $\lambda_0, \lambda_1, \lambda_2, \ldots$ of positive numbers (still to be determined) for which $\langle \sum_{k=0}^{t-1} \lambda_k g_k, g_t \rangle = 0$ for every $t \geq 0$.
3. $f(x_{t+1}) \leq f(x_t) - \frac{||g_t||_2^2}{2L}$ for every $t \geq 0$.

Let $\epsilon_t := f(x_t) - f^*$.

1. Show that $\{\epsilon_t : t \geq 0\}$ is a decreasing sequence.
2. Show that

$$\sum_{k=0}^{t} \lambda_k \epsilon_k \leq R\sqrt{\sum_{k=0}^{t} \lambda_k^2 ||g_k||_2^2} \leq R\sqrt{2L \sum_{k=0}^{t} \lambda_k^2 (\epsilon_k - \epsilon_{k+1})}.$$

Hint: we suggest you to use successively the convexity of f, and the three properties of the sequence $\{x_t : t \geq 0\}$ in the given order.

Defining a sequence λ_t that works

Observe that the sum on the right-hand side is:

$$S_t := \lambda_0^2 \epsilon_0 + (\lambda_1^2 - \lambda_0^2)\epsilon_1 + (\lambda_2^2 - \lambda_1^2)\epsilon_2 + \cdots + (\lambda_t^2 - \lambda_{t-1}^2)\epsilon_t - \lambda_t^2 \epsilon_{t+1},$$

and that this number is nonnegative (why?). We denote the left-hand side by $V_t := \sum_{k=0}^{t} \lambda_k \epsilon_k$. There is a natural choice for the constants $\lambda_0, \lambda_1, \ldots$ so that $V_t - \lambda_t^2 \epsilon_{t+1} = S_t$:

We can take $\lambda_0 = 1$, $\lambda_{k+1} = \frac{1}{2}(1 + \sqrt{1 + 4\lambda_k^2})$ (why?). Show that $\lambda_k \geq \frac{k+1}{2}$ for every $k \geq 0$.

Prove that with this choice of λ_k, we have $\sqrt{V_t} \leq R\sqrt{2L}$ and thus:

$$\epsilon_{t+1} = f(x_{t+1}) - f^* \leq \frac{V_t}{\lambda_t^2} \leq \frac{2LR^2}{\lambda_t^2} \leq \frac{8LR^2}{(t+1)^2}.$$

Observe that this result reaches the optimal complexity for L-smooth convex problems with a first-order oracle.

We can construct such a minimizing sequence

Show that if $\hat{x}_t := x_t - \frac{g_t}{L}$, we have $f(\hat{x}_t) \leq f(x_t) - \frac{||g_t||_2^2}{2L}$.

Define now $x_{t+1} := \arg\min \left\{ f(x) : x \in \text{span}\{\hat{x}_t, \sum_{k=0}^t \lambda_k g_k\} \right\}$.

Prove that:

1. $\langle g_{t+1}, x_{t+1} \rangle = 0$.
2. $\langle \sum_{k=0}^t \lambda_k g_k, g_{t+1} \rangle = 0$.
3. $f(x_{t+1}) \leq f(x_t) - \frac{||g_t||_2^2}{2L}$.

Note: the drawback of this method is that one needs to solve at every step a 2D convex optimization problem exactly. Nesterov's accelerated method, which is based on entirely different principles, does not have this unpleasant feature.

`Lecture_10/OptimalGradientMethodForSmoothConvexProblems`

An optimal gradient method for smooth convex problems A minimizing sequence full of properties

Let us recall the three properties that the sequence of points $\{x_t : t \geq 0\}$ must satisfy. Remember that we have abbreviated $f'(x_t)$ by g_t.

1. $\langle g_t, x_t \rangle = 0$ for every $t \geq 0$.
2. For a sequence $\lambda_0, \lambda_1, \lambda_2, \ldots$ of positive numbers still to determine, $\langle \sum_{k=0}^{t-1} \lambda_k g_k, g_t \rangle = 0$ for every $t \geq 0$.
3. $f(x_{t+1}) \leq f(x_t) - \frac{||g_t||_2^2}{2L}$ for every $t \geq 0$.

Let $\epsilon_t := f(x_t) - f^*$.

1. Let $t \geq 0$. Since $f(x_{t+1}) \leq f(x_t) - \frac{||g_t||_2^2}{2L} \leq f(x_t)$, we have, by subtracting f^* on each side, that: $\epsilon_{t+1} = f(x_{t+1}) - f^* \leq f(x_t) - f^* = \epsilon_t$. Therefore, the sequence $\{\epsilon_t : t \geq 0\}$ is decreasing.
2. Let x^* be a minimizer of f for which $||x^*||_2 = R$ (that is, let x^* be the smallest solution to our problem) and let $t \geq 0$. By convexity of the function f, we have

$$\epsilon_t = f(x_t) - f^* \leq \langle g_t, x_t - x^* \rangle = -\langle g_t, x^* \rangle.$$

This last equality comes from the first property of the sequence $\{x_t : t \geq 0\}$. Since $\lambda_k \geq 0$, we deduce:

$$\sum_{k=1}^t \lambda_k \epsilon_k \leq -\left\langle \sum_{k=1}^t \lambda_k g_k, x^* \right\rangle \leq \left\| \sum_{k=1}^t \lambda_k g_k \right\|_2 \cdot ||x^*||_2 = \left\| \sum_{k=1}^t \lambda_k g_k \right\|_2 R.$$

The second inequality follows from Cauchy-Schwarz. Here is where the second property of $\{x_t : t \geq 0\}$ is used.
Note that

$$\left\|\sum_{k=1}^{t} \lambda_k g_k\right\|_2^2 = \left\|\sum_{k=1}^{t-1} \lambda_k g_k + \lambda_t g_t\right\|_2^2$$
$$= \left\|\sum_{k=1}^{t-1} \lambda_k g_k\right\|_2^2 + 2\left\langle \sum_{k=1}^{t-1} \lambda_k g_k, \lambda_t g_t \right\rangle + \lambda_t^2 ||g_t||_2^2 = \left\|\sum_{k=1}^{t-1} \lambda_k g_k\right\|_2^2 + \lambda_t^2 ||g_t||_2^2.$$

We can apply the same reasoning successively to $\left\|\sum_{k=1}^{t-1} \lambda_k g_k\right\|_2^2$, $\left\|\sum_{k=1}^{t-2} \lambda_k g_k\right\|_2^2$, etc..., to get:

$$\left\|\sum_{k=1}^{t} \lambda_k g_k\right\|_2^2 = \sum_{k=1}^{t} \lambda_k^2 ||g_k||_2^2.$$

Therefore,

$$\sum_{k=1}^{t} \lambda_k \epsilon_k \leq -\left\langle \sum_{k=1}^{t} \lambda_k g_k, x^* \right\rangle \leq R\sqrt{\sum_{k=1}^{t} \lambda_k^2 ||g_k||_2^2}.$$

It remains to use the third property of $\{x_t : t \geq 0\}$: we have $\frac{||g_k||_2^2}{2L} \leq f(x_k) - f(x_{k+1}) = (f(x_k) - f^*) - (f(x_{k+1}) - f^*) = \epsilon_k - \epsilon_{k+1}$. Thus:

$$\sum_{k=1}^{t} \lambda_k \epsilon_k \leq R\sqrt{2L \sum_{k=1}^{t} \lambda_k^2 (\epsilon_k - \epsilon_{k+1})}.$$

Fixing a sequence λ_t that works

We denote by S_t the sum of the right-hand side, that is:

$$S_t = \sum_{k=1}^{t} \lambda_k^2 (\epsilon_k - \epsilon_{k+1})$$

We have shown that $\epsilon_k - \epsilon_{k+1} \geq 0$ for every $k \geq 0$. Thus, $S_t \geq 0$. We can rewrite S_t as follows:

$$S_t := \lambda_0^2 \epsilon_0 + (\lambda_1^2 - \lambda_0^2)\epsilon_1 + (\lambda_2^2 - \lambda_1^2)\epsilon_2 + \cdots + (\lambda_t^2 - \lambda_{t-1}^2)\epsilon_t - \lambda_t^2 \epsilon_{t+1}.$$

Let $V_t := \sum_{k=0}^{t} \lambda_k \epsilon_k$, that is, the left-hand side of our inequality. If we set

$$\lambda_0 = 1, \quad \lambda_1 = \lambda_1^2 - \lambda_0^2, \quad , \ldots, \lambda_t = \lambda_t^2 - \lambda_{t-1}^2,$$

we get that $S_t = V_t - \lambda_t^2 \epsilon_{t+1}$, and our inequality can be rewritten as $V_t \leq R\sqrt{2LS_t}$.
For any $k \geq 0$, the number λ_{k+1} satisfies a second-order equation, whose only nonnegative solution is readily computed: $\lambda_{k+1} = \frac{1}{2}(1 + \sqrt{1 + 4\lambda_k^2})$.

Note that $\lambda_0 = 1 \geq \frac{0+1}{2} = \frac{1}{2}$. Assume that $\lambda_k \geq \frac{k+1}{2}$. Then:

$$\lambda_{k+1} = \frac{1}{2}\left(1 + \sqrt{1 + 4\lambda_k^2}\right) \geq \frac{1}{2}\left(1 + \sqrt{1 + (k+1)^2}\right) \geq \frac{k+2}{2}.$$

This last inequality comes from:

$$1 + \sqrt{1 + (k+1)^2} \geq k + 2 \iff 1 + (k+1)^2 \geq (k+1)^2.$$

So, $\lambda_{k+1} \geq \frac{k+2}{2}$, and by induction, $\lambda_t \geq \frac{t+1}{2}$ for every $t \geq 0$.
Let us get back to our inequality $V_t \leq R\sqrt{2LS_t}$. Since $S_t \leq V_t$, we can relax it it by $V_t \leq R\sqrt{2LV_t}$ or $\sqrt{V_t} \leq R\sqrt{2L}$, or $V_t \leq 2LR^2$. Now, since $S_t = V_t - \lambda_t^2\epsilon_{t+1} \geq 0$, we have:

$$\frac{(t+1)^2}{4}\epsilon_{t+1} \leq \lambda_t^2\epsilon_{t+1} \leq V_t \leq 2LR^2,$$

That is:

$$f(x_{t+1}) - f^* \leq \frac{8LR^2}{(t+1)^2}.$$

Such a method would be **optimal**: it is not possible to conceive a significantly faster method for the unconstrained minimization of a smooth convex function.

We can construct such a minimizing sequence It remains to construct an actual sequence of points that satisfies the three desired properties.

Suppose that $\hat{x}_t := x_t - \frac{g_t}{L}$. Since the function f has a Lipschitz continuous gradient, we can write:

$$\begin{aligned} f(\hat{x}_t) &\leq f(x_t) + \langle f'(x_t), \hat{x}_t - x_t\rangle + \frac{L}{2}||\hat{x}_t - x_t||_2^2 \\ &= f(x_t) - \frac{1}{L}||g_t||_2^2 + \frac{L}{2L^2}||g_t||_2^2 = f(x_t) - \frac{1}{2L}||g_t||^2. \end{aligned}$$

Define now $x_{t+1} := \arg\min\left\{f(x) : x \in \text{span}\{\hat{x}_t, \sum_{k=0}^t \lambda_k g_k\}\right\}$. To simplify the notation, we define $\Pi := \text{span}\{\hat{x}_t, \sum_{k=0}^t \lambda_k g_k\}$. The optimality conditions show that for every $y \in \Pi$, we have $\langle f'(x_{t+1}), y - x_{t+1}\rangle \geq 0$. As $\Pi = \Pi + \{x_{t+1}\}$, we can reformulate this condition as $\langle f'(x_{t+1}), y\rangle \geq 0$ for every $y \in \Pi$. Finally, since $\Pi = -\Pi$, we get $\langle f'(x_{t+1}), y\rangle = \langle g_{t+1}, y\rangle = 0$ for all $y \in \Pi$.

1. We have $\langle g_{t+1}, x_{t+1}\rangle = 0$, because $x_{t+1} \in \Pi$.
2. We have $\langle \sum_{k=0}^t \lambda_k g_k, g_{t+1}\rangle = 0$, because $\sum_{k=0}^t \lambda_k g_k \in \Pi$.
3. Finally, $f(x_{t+1}) = \min\{f(x) : x \in \Pi\} \leq f(\hat{x}_t) \leq f(x_t) - \frac{1}{2L}||g_t||_2^2$.

Therefore, this sequence satisfies everything we need: it defines an optimal first-order method.

Newton Decrement (Boyd) Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a twice differentiable μ-strongly convex function with respect to the Euclidean norm, with $\mu > 0$. For every $x \in \text{dom} f$, we define the *local norm* of $d \in \mathbb{R}^n$ as:

$$||d||_x := \langle f''(x)d, d\rangle^{1/2},$$

where $\langle \cdot, \cdot \rangle$ represents the standard dot product.

1. Show that the dual norm of $||\cdot||_x$ satisfies for all $d \in \mathbb{R}^n$:

$$||d||_{x*} := \langle f''(x)^{-1}d, d\rangle^{1/2}.$$

2. The *Newton decrement* of a point x is defined as

$$\lambda(x) := ||x_+ - x||_x,$$

where $x_+ := x - f''(x)^{-1}f'(x)$ is the result of a Newton step made from x. Show that these alternative definitions of the Newton decrement are equivalent.

$$\lambda(x) = ||f'(x)||_{x*} = \sup_{\langle f''(x)v, v\rangle = 1} -\langle f'(x), v\rangle = \sup_{v \neq 0} \frac{-\langle f'(x), v\rangle}{\langle f''(x)v, v\rangle^{1/2}}.$$

Give an interpretation of the last equality.

`Lecture_12/NewtonDecrement`

Newton Decrement Observe that the Hessian of f is invertible everywhere on its domain, due to the strong convexity of f.

1. By definition,

$$||d||_{x*} = \sup_{||u||_x^2 = 1} \langle d, u\rangle.$$

The KKT conditions for the above problem read as follows:

$$d - 2\lambda^* f''(x)u^* = 0, \quad \langle f''(x)u^*, u^*\rangle = 1$$

Taking the scalar product of the first equation with u^* and simplifying with the second one, we obtain $\lambda^* = \langle d, u^*\rangle/2$. As $f''(x)$ is invertible, we get:

$$f''(x)^{-1}d - \langle d, u^*\rangle u^* = 0.$$

We take the scalar product of this equation with d to get

$$||d||_{x*} = \langle d, u^*\rangle = \langle d, f''(x)^{-1}d\rangle^{1/2}.$$

2. We have:

$$\begin{aligned}\lambda(x)^2 = ||x_+ - x||_x^2 &= \langle f''(x)(f''(x)^{-1}f'(x)), f''(x)^{-1}f'(x)\rangle \\ &= \langle f'(x), f''(x)^{-1}f'(x)\rangle = ||f'(x)||_{x*}^2.\end{aligned}$$

The second equality is a rewriting of the fact proved in the previous item. The last equality is trivial:

$$\sup_{||v||_x = 1} -\langle f'(x), v\rangle = \sup_{v \neq 0} -\langle f'(x), \frac{v}{||v||_x}\rangle.$$

This last equality corresponds to the best improvement of the Newton decrement under the metric derived from the Hessian matrix. The maximizer v^*, which coincides with the Newton step direction, corresponds to the direction in which this best improvement is done.

Universal approximation in Hilbert spaces Let H be a Hilbert space and $G \subseteq H$ be a bounded set in H, say $G \subseteq B[0, R]$. Let x be a point in the closure of $\text{conv}(G)$ and let $\delta > R^2 - ||x||^2$. Prove that for every positive integer n, there exists n point $p_1, \ldots, p_n \in G$ and nonnegative numbers $\lambda_1, \ldots, \lambda_n$ summing up to 1 for which:

$$\left\| x - \sum_{i=1}^{n} \lambda_i p_i \right\|^2 \leq \frac{\delta}{n}.$$

We suggest the following plan for your proof. Let $\epsilon > 0$. Take $x_\epsilon \in \text{conv}(G)$ so that $||x - x_\epsilon|| < \epsilon$ and define $x_1, \ldots, x_M \in G$ so that $x_\epsilon = \sum_{i=1}^{M} \gamma_i x_i$, where the γ_i's are positive and sum up to 1.

1. Suppose that we pick randomly a point from $S_M := \{x_1, \ldots, x_M\}$, so that x_i is taken with probability γ_i. What is the expectation of this random variable?
2. Suppose that we pick randomly n points $p_1, \ldots, p_n$ from S_M, independently (we can therefore pick the same point more than once), with the same distribution as above. What is the expectation of their average $\bar{p} := \frac{p_1 + \cdots + p_n}{n}$?
3. Verify that for every $i \neq j$ we have $\mathbb{E}[\langle x_\epsilon - p_i, x_\epsilon - p_j \rangle] = 0$.
4. Prove that $\mathbb{E}[||x_\epsilon - \bar{p}||^2] \leq (R^2 - ||x_\epsilon||^2)/n$.
5. Since $\delta > R^2 - ||x||^2$, there is some $0 < \gamma < 1$ for which $\gamma\delta \geq R^2 - ||x||^2$. Prove that if $\epsilon \leq \min\left\{\frac{2R}{n}, \frac{\delta(1-\gamma)}{4R}\right\}$, we have $\mathbb{E}[||x - \bar{p}||^2] \leq \frac{\delta}{n}$.
6. Deduce that there exists $y \in H$ that is a convex combination of at most n points of G, and for which $||x - y||^2 \leq \frac{\delta}{n}$.

Note: this result is used in a critical way to relate, in a shallow neural network with sigmoid activation functions, the number of neurons and the optimal approximation accuracy of this network.

The formal statement is as follows.

> Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function (the function to be approximated by a neural network) that has a Fourier transform $\hat{f}$ defined on $\mathbb{R}^n$. Assume that the functions $\omega \mapsto \exp(i\omega^T x)\hat{f}'(\omega)$ are integrable for every $x \in \mathbb{R}^n$ and that these integrals are uniformly bounded by a constant C. Let $\phi : \mathbb{R} \to [0, 1]$ be a measurable function for which $\lim_{t\to\infty} \phi(t) = 1$ and $\lim_{t\to-\infty} \phi(t) = 0$ (this ϕ is the activation function of the neurons). Let $f_n(x; A, b, c) := \sum_{i=1}^{n} c_i \phi(\langle a_i, x \rangle + b_i) + c_0$ for a real matrix $A = [a_1, a_2, \ldots, a_n]$ and vectors b, and c of appropriate size (this $f_n(\cdot; A, b, c)$ is the function implemented by a shallow neural network with n nodes and parameters A, b, c). Finally, let μ_r be an arbitrary probability measure on the ball $B(0, r)$ for some fixed radius $r > 0$.
>
> Then for every $n \geq 1$ there exist A, b, c such that
>
> $$\int_{B(0,r)} ||f_n(x; A, b, c) - f(x)||^2 d\mu_r(x) \leq \frac{4r^2C^2}{n}.$$

Here is quite a rough proof sketch.

We can assume that $f(0) = 0$ as it suffices to adjust c_0 appropriately.

Let G_F be the set of functions of the form $cF(\langle a, x\rangle + b)$ for $|c| \leq 2C$. By the result proved in the above exercise, it suffices to verify that f is in the closure of $\text{conv}(G_\phi)$ (and to prove that we can take δ as small as $4r^2C^2$, but this is actually pretty elementary).

That statement is first proved for $G_{\cos}$ by using the inverse Fourier transform of $\hat{f}$. Denoting by "step" the function $\mathbf{1}_{\{t\geq 0\}}$, we can show that $G_{\cos}$ is in the closure of $\text{conv}(G_{\text{step}})$ by representing the function "cos" as a limit (in the $L^2(\mu_r)$ sense) of functions in $\text{conv}(G_{\text{step}})$. Finally, we can show that the function "step" can be written as the limit (also in the $L^2(\mu_r)$ sense) of a sequence of functions in $\text{conv}(G_\phi)$.

`Lecture_10/HilbertConvexApproximation`

Universal approximation in Hilbert spaces Let H be a Hilbert space and $G \subseteq H$ be a bounded set in H, say $G \subseteq B[0, R]$. Let x be a point in the closure of $\text{conv}(G)$ and let $\delta > R^2 - ||x||^2$. Prove that for every positive integer n, there exists n point $p_1, \ldots, p_n \in G$ and nonnegative numbers $\lambda_1, \ldots, \lambda_n$ summing up to 1 for which:

$$\left\| x - \sum_{i=1}^{n} \lambda_i p_i \right\|^2 \leq \frac{\delta}{n}.$$

We suggest the following plan for your proof. Let $\epsilon > 0$. Take $x_\epsilon \in \text{conv}(G)$ so that $||x - x_\epsilon|| < \epsilon$ and define $x_1, \ldots, x_M \in G$ so that $x_\epsilon = \sum_{i=1}^{M} \gamma_i x_i$, where the γ_i's are positive and sum up to 1.

1. Let X be a random vector that equals x_i with probability γ_i. Then $\mathbb{E}[X] = \sum_{i=1}^{M} \gamma_i x_i$, which is precisely x_ϵ.
2. Since $p_i \sim X$, where X is the random vector defined above, $\mathbb{E}[\bar{p}] := \frac{1}{n}\mathbb{E}[p_1 + \cdots + p_n] = \frac{n}{n}\mathbb{E}[X] = x_\epsilon$.
3. The orthogonality property shown here will prove very useful later to simplify calculations considerably.

   Let $i \neq j$ and p_i, p_j be two independent random vectors distributed as X. Then:

$$\begin{aligned}
\mathbb{E}[\langle x_\epsilon - p_i, x_\epsilon - p_j\rangle] &= \mathbb{E}[||x_\epsilon||^2] - \mathbb{E}[\langle p_i, x_\epsilon\rangle] - \mathbb{E}[\langle x_\epsilon, p_j\rangle] + \mathbb{E}[\langle p_i, p_j\rangle] \\
&= ||x_\epsilon||^2 - \langle \mathbb{E}[X], x_\epsilon\rangle - \langle x_\epsilon, \mathbb{E}[X]\rangle + \mathbb{E}[\langle p_i, p_j\rangle] \\
&= -||x_\epsilon||^2 + \mathbb{E}[\langle p_i, p_j\rangle] = -||x_\epsilon||^2 + \sum_{i,j=1}^{M} \gamma_i \gamma_j \langle x_i, x_j\rangle \\
&= -||x_\epsilon||^2 + \left\langle \sum_{i=1}^{M} \gamma_i x_i, \sum_{j=1}^{M} \gamma_j x_j \right\rangle = -||x_\epsilon||^2 + \langle x_\epsilon, x_\epsilon\rangle = 0
\end{aligned}$$

4. In order to use the orthogonality shown above, we split the squared norm $||x_\epsilon - \bar{p}||^2$ into:

$$\left\langle x_\epsilon - \sum_{i=1}^{n} \frac{p_i}{n}, x_\epsilon - \sum_{j=1}^{n} \frac{p_j}{n} \right\rangle = \left\langle \sum_{i=1}^{n} \frac{x_\epsilon - p_i}{n}, \sum_{j=1}^{n} \frac{x_\epsilon - p_j}{n} \right\rangle.$$

Taking expectations, the terms corresponding to $i \neq j$ are null. It remains:]

$$\mathbb{E}[||x_\epsilon - \bar{p}||^2] = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}[\langle x_\epsilon - p_i, x_\epsilon - p_i \rangle] = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}[||x_\epsilon - p_i||^2] = \frac{1}{n} \mathbb{E}[||x_\epsilon - X||^2].$$

By definition,

$$\begin{aligned} \mathbb{E}[||x_\epsilon - X||^2] &= \sum_{i=1}^{M} \gamma_i ||x_\epsilon - x_i||^2 = \sum_{i=1}^{M} \gamma_i ||x_\epsilon||^2 - 2 \sum_{i=1}^{M} \gamma_i \langle x_\epsilon, x_i \rangle + \sum_{i=1}^{M} \gamma_i ||x_i||^2 \\ &= ||x_\epsilon||^2 - 2\langle x_\epsilon, x_\epsilon \rangle + \sum_{i=1}^{M} \gamma_i ||x_i||^2 \leq -||x_\epsilon||^2 + R^2. \end{aligned}$$

5. So far, we know that $\mathbb{E}[||x_\epsilon - \bar{p}||^2] \leq \frac{R^2 - ||x_\epsilon||^2}{n}$. This inequality needs two modifications to prove the bound we want: first, we are interested by a bound on $\mathbb{E}[||x - \bar{p}||^2]$ and not on $\mathbb{E}[||x_\epsilon - \bar{p}||^2]$; second, we want to exploit that $R^2 - ||x||^2 \leq \gamma\delta$, so we need to modify the numerator $R^2 - ||x_\epsilon||^2$ appropriately. Here is how to work out the first modification:

$$\begin{aligned} \mathbb{E}[||x - \bar{p}||^2] &= \mathbb{E}[||x - x_\epsilon||^2] + 2\mathbb{E}[\langle x - x_\epsilon, x_\epsilon - \bar{p} \rangle] + \mathbb{E}[||x_\epsilon - \bar{p}||^2] \\ &= \mathbb{E}[||x - x_\epsilon||^2] + 2\langle x - x_\epsilon, x_\epsilon - \mathbb{E}[\bar{p}] \rangle + \mathbb{E}[||x_\epsilon - \bar{p}||^2] \\ &= \mathbb{E}[||x - x_\epsilon||^2] + \mathbb{E}[||x_\epsilon - \bar{p}||^2] \\ &\leq \epsilon^2 + \frac{R^2 - ||x_\epsilon||^2}{n}. \end{aligned}$$

The second modification results from a lower bound for $||x_\epsilon||^2$:

$$||x_\epsilon||^2 \geq (||x|| - ||x - x_\epsilon||)^2 \geq ||x||^2 - 2||x|| \cdot ||x - x_\epsilon|| \geq ||x||^2 - 2||x||\epsilon \geq ||x||^2 - 2R\epsilon.$$

Hence:

$$\mathbb{E}[||x - \bar{p}||^2] \leq \epsilon^2 + \frac{R^2 - ||x||^2 + 2R\epsilon}{n} \leq \epsilon^2 + \frac{\delta\gamma + 2R\epsilon}{n}.$$

Now, we use the bounds on ϵ:

$$\epsilon^2 + \frac{\delta\gamma + 2R\epsilon}{n} \leq \epsilon \cdot \frac{2R}{n} + \frac{\delta\gamma + 2R\epsilon}{n} = \frac{\delta\gamma + 4R\epsilon}{n} \leq \frac{\delta}{n}.$$

The first inequality comes from $\epsilon \leq \frac{2R}{n}$ and the second one from $\epsilon \leq \frac{\delta(1-\gamma)}{4R}$.

6. Since the average $\bar{p}$ is at an expected distance from x that is smaller than $\sqrt{\delta/n}$, there must be a convex combination y of n points of G at most at a distance $||x - y||$ smaller than $\sqrt{\delta/n}$.