(b) $F(x) = \chi_K(x) = \begin{cases} +\infty &, x \notin K \\ 0 &, x \in K \end{cases}$ for $K \subseteq \mathbb{R}^n$ convex.

$$\boxed{F^*(y) = \sup_{x \in K} \langle x, y \rangle}$$

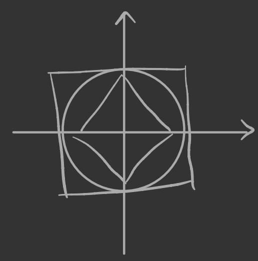
"Fantastic"

Proof: $F^*(y) = \sup_{x \in \mathbb{R}^n} \{\langle x, y\rangle - F(x)\} = \sup_{x \in \mathbb{R}^n} \begin{Bmatrix} -\infty &, x \notin K \\ \langle x, y \rangle - 0 &, x \in K \end{Bmatrix}$

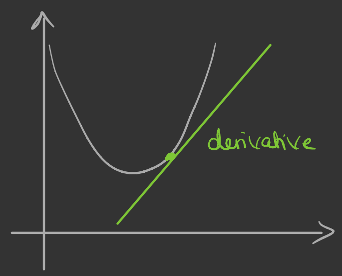
Exercise: Let $B_p = \{x \in \mathbb{R}^n \mid \|x\|_p \leq 1\}$, $1 \leq p \leq +\infty$

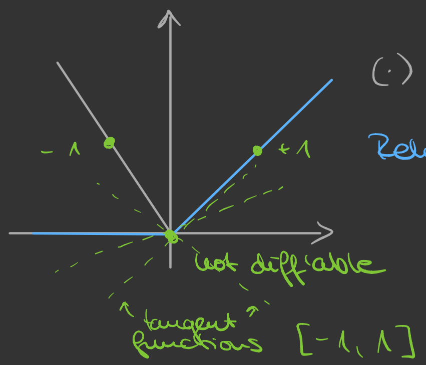
Now let $K = B_p$, $F(x) = \chi_K(x)$.

Then $F^*(y) = ?$



5.4. The Subdifferential





Def. 5.19: The subdifferential of a convex function $F: \mathbb{R}^N \to \mathbb{R}$ at a point $x \in \mathbb{R}^N$ is defined by

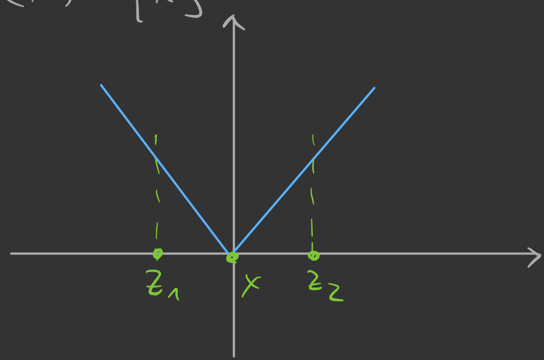
$$\partial F(x) = \{ v \in \mathbb{R}^N : F(z) - F(x) \geq \langle v, z - x \rangle \ \forall z \in \mathbb{R}^N \} \subseteq \mathbb{R}^N$$

Every element of $\partial F(x)$ is called subgradient.

Exercise: For $n = 1$, $F(x) = |x|$.

Prove that $\partial F(-1) = \{-1\}$ and $\partial F(1) = \{1\}$

but $\partial F(0) = [-1, 1]$.



Exercise: If F is convex, $\partial F(x) \neq \emptyset$.

Exercise: If F is diff'able, then $\partial F(x) = \{ \nabla F(x) \}$.

$\Rightarrow$ subdifferential as generalization.

Thm 5.20: A vector $x \in \mathbb{R}^n$ is a minimum

$$\Longleftrightarrow \quad 0 \in \partial F(x).$$

Proof: "$\Leftarrow$": $0 \in \partial F(x)$

$$\Rightarrow F(z) - F(x) \geq \langle \underset{\overset{\shortparallel}{0}}{v}, z - x \rangle = 0$$

$$\Rightarrow F(z) \geq F(x) \quad \forall z \in \mathbb{R}^n$$

$\Rightarrow$ x is a point of minimum for F.

"$\Rightarrow$" $F(z) \geq F(x) \quad \forall z \in \mathbb{R}^n$

$$\Rightarrow F(z) - F(x) \geq \langle 0, z - x \rangle = 0.$$

$$\Rightarrow v \in \partial F(x).$$

Thm 5.21: Let $F : \mathbb{R}^n \to (-\infty, \infty]$ convex, $x, y \in \mathbb{R}^N$. Then the following two conditions are equivalent:

(i) $y \in \partial F(x)$

(ii) $F(x) + F^*(y) = \langle x, y \rangle$ ($\geq$ given by Fenchel's)

If additionally F is lower semicontinuous, then (i) and (ii) are equivalent to

(iii) $x \in \partial F^*(y)$.

Remark: If F is convex and lower semicontinuous

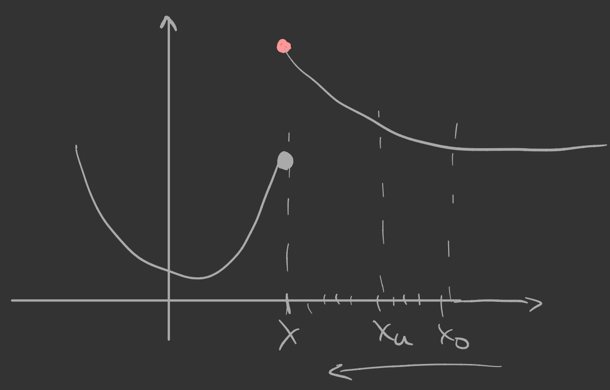
" ∂F is the inverse of ∂F^* "

in the sense that

$$x \in \partial F^*(y) \quad \Leftrightarrow \quad y \in \partial F(x).$$

Remark: Using the subgradient, we can define the <u>subgradient descent</u> as $x_{u+1} = x_u + \lambda \cdot \xi_{n}$ $\xi_n \in \partial F(x^u)$

Remark:



- lower semicontinuous
  $\liminf_n F(x_u) \geq F(x)$

- upper semicontinuous
  $\limsup_n F(x_u) \leq F(x)$

Vorlesung 22

08.07.2020

Proof: (Thm. 5.21)

(i) $\Rightarrow$ (ii): By assumption $y \in \partial F(x)$. This means by def.

$$F(z) - F(x) \geq \langle y, z-x \rangle \quad \forall z, x \in \mathbb{R}^N.$$

$$\Leftrightarrow \quad F(z) - F(x) \geq \langle y, z \rangle - \langle y, x \rangle$$

$$\Leftrightarrow \quad \langle x, y \rangle - F(x) \geq \langle z, y \rangle - F(z).$$

In other words, x is a maximum for the function

$$z \longmapsto \langle z, y \rangle - F(z).$$

We can write

$$\langle x, y \rangle - F(x) \geq \sup_z \langle z, y \rangle - F(z)$$

$$= F^*(y) \quad \text{by definition.}$$

$$\Leftrightarrow \quad F^*(y) + F(x) \leq \langle x, y \rangle$$

Fenchel gives "$\geq$"

$$\Rightarrow \quad F^*(y) + F(x) = \langle x, y \rangle.$$

(ii) $\Rightarrow$ (i) works analogously.

(i), (ii) $\Leftrightarrow$ (iii) F is l.s.c. and convex $\Rightarrow F^{**} = F$.

Using (ii) we have

$$\underbrace{F^{**}}_{F}(x) + F^*(y) = \langle x, y \rangle$$

By applying (i) $\Leftrightarrow$ (ii) for F^* (which is indeed convex) we can obtain that

$$F^*(y) + (F^*)^*(x) = \langle y, x \rangle$$

$$\Leftrightarrow \quad y \in \partial F^*(x)$$

$\blacksquare$

Def.: For a function $F: \mathbb{R}^N \to (-\infty, +\infty]$, we define

$$P_F(z) \in \underset{x \in \mathbb{R}^N}{\operatorname{argmin}}\ F(x) + \frac{1}{2}\|x - z\|_2^2 .$$

$\rightsquigarrow$ might be empty

the so called <u>proximal mapping</u> associated to F.

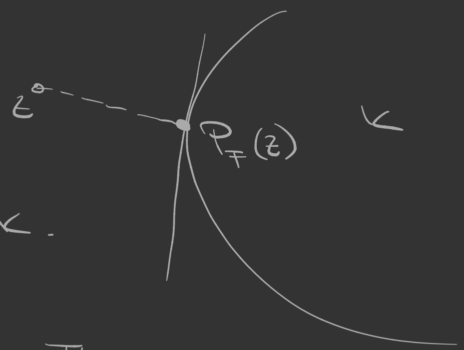
Property: For F convex, the map $x \longmapsto F(x) + \frac{1}{2}\|x - z\|_2^2$ is always strictly convex (because $x \longmapsto \|x\|_2^2$ is strictly convex).

$$\Rightarrow \underset{\substack{\uparrow \\ \text{unique}}}{P_F(z)} = \underset{x \in \mathbb{R}^N}{\operatorname{argmin}}\ F(x) + \frac{1}{2}\|x - z\|_2^2 .$$

convex fct
+ strictly conv. fct
= strictly conv. funct.

Example: $F = \chi_K$, K convex, then

$$P_F(z) = P_K(z) = \underset{x \in K}{\operatorname{argmin}} \|z - x\|_2^2$$



$\Rightarrow$ this is the orthogonal projection onto K.

Prop. 5.22: Let $F: \mathbb{R}^N \to (-\infty, \infty]$ be convex. Then

$$\underset{\substack{\uparrow \\ \text{equality}}}{x =} P_F(z) \iff \underset{\substack{\uparrow \\ \text{inclusion}}}{z \in}\ \underbrace{x + \partial F(x)}_{} = \{x + v \mid v \in \partial F(x)\}.$$

Proof: $x = P_F(z) = \underset{\xi}{\operatorname{argmin}}\ F(\xi) + \frac{1}{2}\|\xi - z\|_2^2$.

$$\overset{5.20}{\iff} 0 \in \partial\left(F(\cdot) + \frac{1}{2}\|\cdot - z\|_2^2\right)(x)$$

But

$$\partial\left(F(\cdot) + \frac{1}{2}\|\cdot - z\|_2^2\right)(x)$$

$$\overset{\text{diff'able}}{=} \partial F(x) + \frac{\partial}{\partial x} \frac{1}{2}\|x - z\|_2^2$$

$$= \partial F(x) + (x - z)$$

Since $0 \in \partial F(x) + x - z$

$$\iff z \in x + \partial F(x).$$

$\square$

Remark: We can characterize the proximal mapping as

$$P_F = (\mathrm{Id} + \partial F)^{-1}$$

$(\mathrm{Id} + \partial F)(x) = x + \partial F(x)$

Thm 5.23: Moreau's Identity

If F convex + lsc, then $\forall z \in \mathbb{R}^n$:

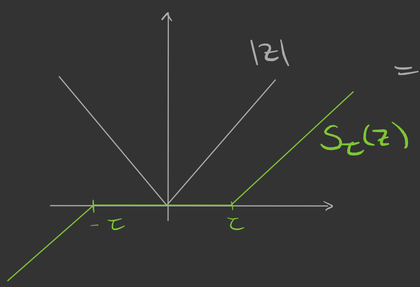
$$z = P_F(z) + P_{F^*}(z).$$

If we have one of these functions, we immediately have the other.

Example: $F(x) = |x|$, $x \in \mathbb{R}$, $N = 1$. $\tau F(x) = \tau |x|$ for $\tau > 0$.

Then $P_{\tau F}(z) = \underset{x \in \mathbb{R}}{\operatorname{argmin}} \left\{ \frac{1}{2} |x - z|^2 + \tau |x| \right\}$

$0 \in \partial\left(\frac{1}{2}|\cdot - z|^2 - \tau|\cdot|\right) P_{\tau F}(z)$



$$= \begin{cases} z - \tau, & z \geq \tau \\ 0, & |z| \leq \tau \\ z + \tau, & z \leq -\tau \end{cases}$$

$=: S_\tau(z)$ "soft-thresholding".

Example: $z \longmapsto \underset{x \geq 0}{\operatorname{argmin}} \frac{1}{2} |x - z|^2 + \tau |x| = \mathrm{ReLU}(z - \tau)$.

$\rightsquigarrow$ Hence, we can interpret each layer of a n.n. as an optimization problem.

Proof (Thm 5.23)

Let $x = P_F(z)$ and $y = z - x$, $z = x + y$, $x = z - y$

$$\overset{(5.22)}{\Rightarrow} z \in x + \partial F(x)$$

$$\Rightarrow z - x = y \in \partial F(x)$$

Since F is l.s.c and convex,

$$y \in \partial F(x) \overset{5.22}{\Longleftrightarrow} x \in \partial F^*(y)$$

$$\| $$

$$z - y \in \partial F^*(y) \Longleftrightarrow z \in y + F^*(y)$$

$\Leftrightarrow \quad y = P_{F^*}(z).$

$\Rightarrow \quad z = x + y = P_F(z) + P_{F^*}(z).$

$\boxtimes$

<u>Non-expansive maps</u>:

A function $F: \mathbb{R}^N \to \mathbb{R}^N$ is <u>non-expansive</u>, if

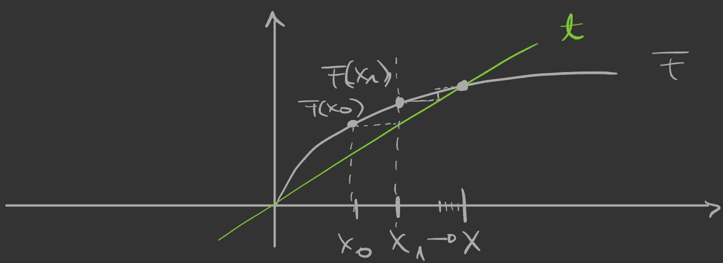
$$\| F(x) - F(y) \|_2 \leq \| x - y \|_2 .$$

If $\exists \; 0 < L < 1$ s.t.

$$\| F(x) - F(y) \|_2 \leq L \| x - y \|_2$$

then the map is called a <u>contraction mapping</u>.

<u>Remark</u>: What do we love about contractions?



$x_{n+1} = F(x_n)$

with $x = F(x)$ being a <u>fixed-point</u>.

We can prove that we will always converge to a unique fixed-point. (<u>Banach fixed-point theorem</u>)

If we weaken the requirements, and let $L = 1$, we often can prove the following:

<u>Thm</u>: (Opial's Fixed point Theorem)

Let $F: \mathbb{R}^N \to \mathbb{R}^N$ be a map. If

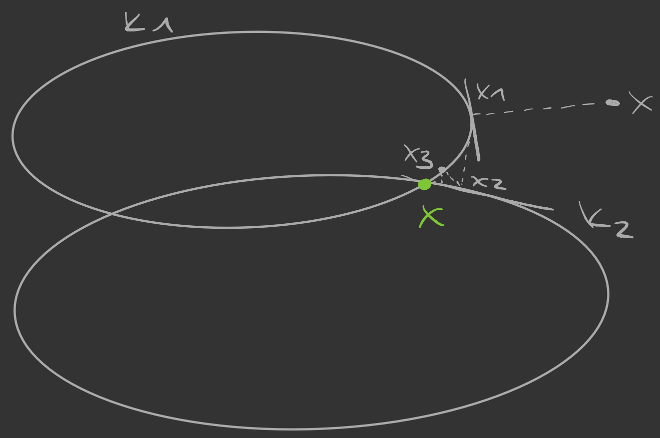
1) F is non-expansive.
2) $\| F(x^n) - x^n \|_2 \to 0$
3) $\operatorname{Fix} F = \{ x \in \mathbb{R}^N \mid x = F(x) \} \neq \emptyset$.

Then

$$x^{n+1} = F(x^n) \to \bar{x} = F(\bar{x}).$$

Property: Composition of non-expansive maps is non-expansive.

Example: K_1, K_2 convex, then P_{K_1}, P_{K_2} is non-expansive, and so is $P_{K_1} \circ P_{K_2}$. We have the POCS-alg.:



$$x_{n+1} = P_{K_2} \circ P_{K_1}(x_n)$$

with $x = (P_{K_2} \circ P_{K_1})(x) \in K_1 \cap K_2$.

Thm 5.24: For a convex function $F: \mathbb{R}^N \to (-\infty, \infty]$, the proximal map is a contraction, i.e.

$$\| P_F(z) - P_F(z') \|_2 \le \| z - z' \|_2 \qquad \forall z, z' \in \mathbb{R}^N.$$

Proof: (skipped)

Remark: We have seen that $P_{\chi_K}(z) = P_K(z)$ and with this we know that orthogonal projections are non-expansive.

5.5 Convex Optimization Problems

An optimization problem is of form

$$\min_{x \in \mathbb{R}^N} F_0(x) \text{ subject to } Ax = y, \quad \rightsquigarrow \text{linear constraint} \Rightarrow \text{exact match}$$

$$F_j(x) \le b_j, \; j \in [M] \quad \rightsquigarrow \text{inequality constraints} \Rightarrow \text{non-exact match}$$

choose the best of the feasible points

where the function $F_0: \mathbb{R}^N \to (-\infty, \infty]$ is called objective function

A point x satisfying linear and inequality is called feasible, the constraints are called feasible, if such an $x \in \mathbb{R}^N$ exists.

we define the set of feasible points

$$K = \{x \in \mathbb{R}^N \mid Ax = y,\ F_j(x) \le b_j,\ j \in [M]\}.$$

The problem is called convex if F_0 and the constraints are convex. Then the set K is convex and we can formulate the equivalent, unconstraint optimization problem

$$\min_{x \in \mathbb{R}^N} \underbrace{F_0(x) + \chi_K(x)}_{=: F(x) \text{ convex}}.$$

If $F_0, F_1, \dots, F_m$ are linear, we call the problem a linear optimization problem.

Def.: For a convex optimization problem, we define the Lagrange function for $x \in \mathbb{R}^N$, $\xi \in \mathbb{R}^m$, $\nu \in \mathbb{R}^M$, $\nu_\ell \ge 0$, $\ell \in [M]$, by

$$L(x, \xi, \nu) = F_0(x) + \xi^T (Ax - y) + \sum_{\ell=1}^{M} \nu_\ell (F_\ell(x) - b_\ell).$$

linear

or $\quad L(x, \xi, \nu) = F_0(x) + \langle \xi, Ax - y \rangle + \langle \nu, F(x) - b \rangle$

The variables ξ, ν are called Lagrange multipliers, and $F(x) := (F_1(x), \dots, F_M(x))^T$ and $b = (b_1, \dots, b_M)^T$.

Def. We define the Lagrange dual function by

$$H(\xi, \nu) = \inf_{x \in \mathbb{R}^N} L(x, \xi, \nu) \quad , \xi \in \mathbb{R}^m, \nu \in \mathbb{R}^M, \nu \ge 0.$$

Property: The dual function is always concave because it is pointwise infimum of a family of affine functions, even if the original problem is not convex.

<u>Property</u>: The dual function provides a bound on the optimal value of $F_0(x^*)$ of the minimization problem, i.e. $\forall \xi \in \mathbb{R}^m$, $\nu \geq 0$, it holds that

$$\boxed{\underbrace{H(\xi,\nu)}_{=\inf_{x\in K} L(x,\xi,\nu)} \leq \underbrace{F(x^*)}_{=\inf_{x\in K} F_0(x)}.}$$

<u>Proof</u>: Indeed, if x is a feasible point, then $Ax=b$, and $F_\ell(x) \leq b_\ell$, so that $\forall \xi \in \mathbb{R}^m$ and $\nu \geq 0$

$$\xi^T \underbrace{(Ax-y)}_{=0 \text{ for } x \text{ feasible}} + \sum_{\ell=1}^{M} \nu_\ell \underbrace{(F_\ell(x)-b_\ell)}_{\leq 0 \text{ by def.}} \leq 0.$$

Vorlesung 23

optimal value of <u>primal problem</u> (convex problem)

$$F_0(x^*) = \min_{x \in K} F_0(x) \quad \text{s.t.} \quad Ax = y,\ F(x) \le b.$$

$$\Big\updownarrow \quad H(\xi, \nu) \le F_0(x^*) \qquad (*)$$

optimal value of <u>dual problem</u> (concave problem)

$$H(\xi^*, \nu^*) = \max_{(\xi,\nu),\, \nu \ge 0} H(\xi, \nu)$$

"max-min-problem" $\rightsquigarrow$

$$= \max_{(\xi,\nu),\, \nu\ge 0} \ \min_{x \in \mathbb{R}^n} L(x, \xi, \nu).$$

$$\text{for } L(x,\xi,\nu) = F_0(x) + \langle \xi, Ax - y\rangle + \langle \nu, F(x) - b\rangle.$$

We call the triple (x^*, ξ^*, ν^*) the <u>primal-dual optimal solution</u>.

By $(*)$ we have in particular that

$$H(\xi^*, \nu^*) \le F_0(x^*)$$

and in addition $\forall x \in K$ ← feasible set

$$\langle \xi, \underbrace{Ax - y}_{=0,\text{ for } x \in K} \rangle + \langle \nu, \underbrace{F(x) - b}_{\le 0,\text{ for } x \in K,\text{ since } F(x) \le b.} \rangle \le 0$$

Hence, we have <u>weak duality</u>, i.e.

$$\max_{\xi,\nu} H(\xi, \nu) \le \min_{x \in K} F_0(x).$$

For some problems, we have strong duality i.e.

$$\max_{\xi, \nu} H(\xi, \nu) = \min_{x \in K} F_0(x).$$

↳ actually happens quite often

Theorem 5.25: Slater's constraint qualification.

Let $F_0, F_1, \dots, F_M$ be convex and $\mathrm{dom}(F_0) = \mathbb{R}^N$. If there exists a point $x \in \mathbb{R}^N$ s.t. $Ax = y$ and $F_j(x) < b_j$, $j \in [M]$, then strong duality holds.
In particular, all linear programs such that $Ax = y$ has solutions, fulfill the strong duality.

What do we love about strong duality?

- good for proofs.
- we can already gain insights about the optimal solutions within K, e.g. extreme points / points in the middle.

$$\xi^* = Ax^* - y \quad \longleftrightarrow \quad A^T(\xi^* + y) = x^*$$

- we can use numerical methods and make "the problems meet" iteratively, i.e. "sandwich" the solution between the concave function H, and the convex function F.

Example: linear program (no inequ. constraints)

(x^*, ξ^*) primal-dual solution

$$L(x, \xi) = F_0(x) + \langle \xi, Ax - y \rangle$$

$$\sup_{\xi \in \mathbb{R}^m} L(x, \xi) = \sup_{\xi \in \mathbb{R}^m} \left(F_0(x) + \langle \xi, \underbrace{Ax - y}_{\text{if this is } \neq 0, \text{ i.e. } x \notin K} \rangle \right)$$

choose ξ, s.t. $\langle \cdot , \cdot \rangle$ gets larger ...
$\xi = t(Ax - y)$ for $t \to \infty$.

$$\Rightarrow \quad \sup_{\xi \in \mathbb{R}^m} L(x,\xi) = \begin{cases} F_0(x) & , \text{ indep. of } \xi. \\ +\infty & , \; Ax \neq y \end{cases}$$

Now, adding the infimum we see, that we arrive again at the primal problem:

$$\inf_{x \in \mathbb{R}^N} \sup_{\xi \in \mathbb{R}^m} L(x,\xi) = \begin{cases} \inf_{x \in \mathbb{R}^N} F_0(x) & , \; x \in K \\ +\infty & , \; x \in K \end{cases}$$

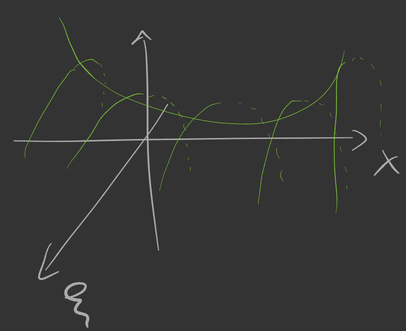
$\Rightarrow$ For strong duality problems, we have

$$\underbrace{\max_{\xi} \min_{x} L(x,\xi)}_{H(\xi^*)} = \underbrace{\min_{x} \max_{\xi} L(x,\xi)}_{F_0(x^*)}$$

called a <u>saddle point problem</u>. The name here refers to the fact that

$$L(x,\xi)$$

convex wrt x — concave wrt ξ

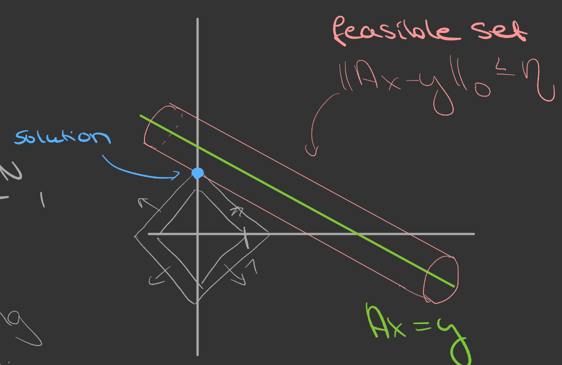


<u>One (theoretical) application of strong duality</u>

<u>Theorem</u>: Let $\|\cdot\|_0, \|\cdot\|_1$ be two norms on $\mathbb{R}^N$. Let $A \in \mathbb{R}^{m \times N}$, $y \in \mathbb{R}^m$ and $\eta > 0$. Consider the following optimization problem:

$$(*) \quad \min_{x \in \mathbb{R}^N} \|x\|_1 \quad \text{s.t.} \quad \|Ax-y\|_0 \leq \eta.$$

Assume that there exists a point $x \in \mathbb{R}^N$, s.t. $\|Ax-y\|_0 < \eta$. $\rightarrow$ gives strong duality.



Consider an optimal solution x^*, then $\exists \lambda \geq 0$, s.t. x^* is also the minimizer of the unconstraint optimization problem

$$(**) \quad \min_{x \in \mathbb{R}^N} \|x\|_1 + \lambda \|Ax-y\|_0 .$$ ⤳ now, unconstraint

Conversely, for $\lambda > 0$, the minimizes of $(**)$ $x^\#$ is solution to $(*)$ for a suitable choice of η.

In other words: we can make these problems equivalent by choice of corresponding pairs (η, λ).

Proof: Let x^* be the solution of $(*)$.

$F_0(x) = \|x\|_1$, $F_1(x) = \|Ax-y\|_0$, $b_1 = \eta$.

$$L(x,\xi) \overset{\text{Def.}}{=} F_0(x) + \nu(F_1(x) - b_1)$$
$$= \|x\|_1 + \nu(\|Ax-y\|_0 - \eta).$$

By assumption we have an x s.t. $\|Ax-y\|_0 < \eta$ (5.25)

$\Rightarrow$ Strong duality

$\Rightarrow \exists x^* > 0$, s.t. $L(x^*, \nu^*) \leq L(x, \nu^*) \quad \forall x \in \mathbb{R}^N$.

Indeed: $$\underbrace{\min_x \underbrace{\max_\nu L(x,\nu)}_{= L(x,\nu^*)}}_{= L(x^*,\nu^*)}$$

$\Rightarrow x \longmapsto L(x,\nu^*)$ has x^* as minimizer.

We have $L(x,\nu^*) = \|x\|_1 + \nu^*(\|Ax-y\|_0 - \eta)$, and ↖ indep. of x

hence minimizes of $x \mapsto L(x,\nu^*)$ must be the minimizes of

$$x \longmapsto \|x\|_1 + \nu^* \|Ax-y\|_0 .$$

Choosing $\lambda = \gamma^*$, we get that x^* is also a solution of (**) with such λ. ∎

<u>Remark</u>: For convex constraint sets we can use the alternating projection. The more complex the projections get (e.g. on L_1-ball), calculation of the orthogonal projection is difficult.
Then, using an unconstraint optimization problem is numerically much easier, than on a constraint set. Here we can e.g. use the "usual" gradient descent.

<u>More general</u>: Consider optimization problems of the form

$$\min_{x \in \mathbb{R}^N} F(Ax) + G(x)$$

with A some matrix, F and G convex functions.

[This is a generalization to the problems we considered so far, e.g. $F(Ax) = \chi_{\{y\}}(Ax)$ and $G(x) = \|x\|_1$.]

Introduce $z = Ax$. Then we have

unconstraint problem to augmented constraint problem

$$\min_{x \in \mathbb{R}^N} F(Ax) + G(x)$$

$$\Longleftrightarrow \min_{z, x \in \mathbb{R}^N} F(z) + G(x) \quad \text{s.t.} \quad (Ax - z) \equiv 0.$$

We use the dual function

$$
\begin{aligned}
H(\xi) &= \inf_{x,z} F(z) + G(x) + \langle \xi, Ax - z \rangle \\
&= \inf_{x,z} \left\{ F(z) + G(x) + \langle A^T \xi, x \rangle - \langle \xi, z \rangle \right\} \\
&= \underbrace{- \sup_z \left\{ \langle \xi, z \rangle - F(z) \right\}}_{\text{Fenchel duality for } F} \underbrace{- \sup_x \left\{ - \langle A^T \xi, x \rangle - G(x) \right\}}_{\text{Fenchel duality for } G}
\end{aligned}
$$

$$= -F^*(\xi) - G^*(-A^T\xi).$$

Theorem 5.29:

Given strong duality we have that

$$\min_x \; F(Ax) + G(x) \qquad \text{primal problem}$$

$=$ (through duality and Fenchel duality)

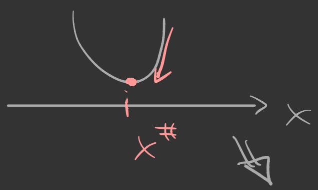
$$\max_\xi \; -F^*(\xi) - G(-A^T\xi) \qquad \text{dual problem}$$

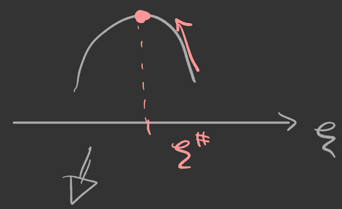
Vorlesung 24

15.07.2020

Recall: If strong duality holds, we have

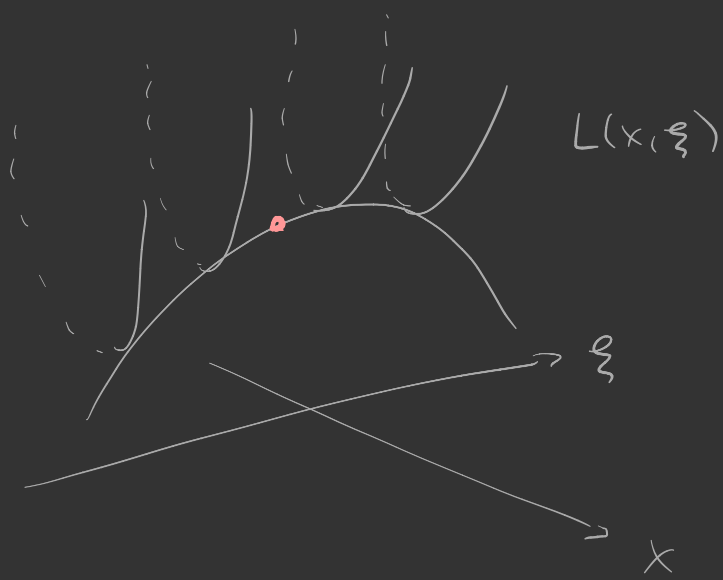
$$\underbrace{\min_x \; F(Ax) + G(x)}_{\text{convex primal problem}} = \underbrace{\max_\xi \; -F^*(\xi) - G^*(-A^T\xi)}_{\text{concave dual problem}}$$





$(x^\#, \xi^\#)$ primal-dual solution

Lagrangian function: Saddle point problem

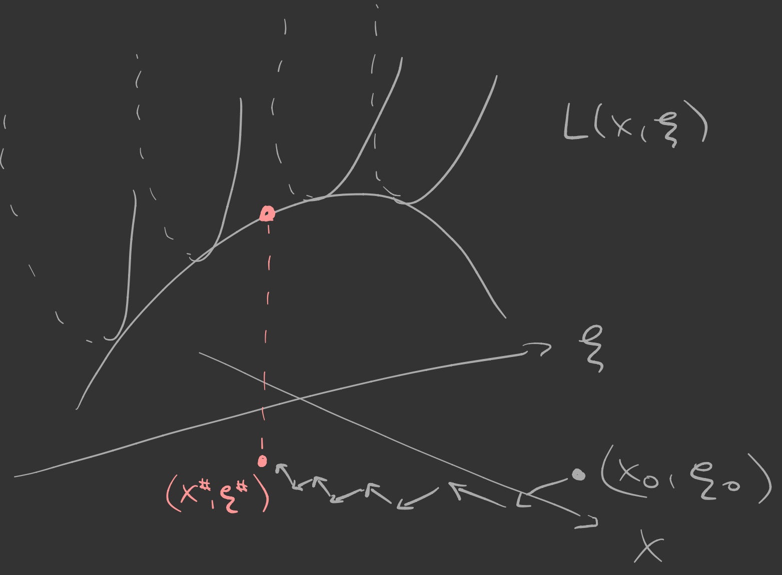


Let's look at an algorithm to solve such a problem:

6.3. A Primal Dual Algorithm

Approach of an "alternating gradient descent" using proximal maps

- ▷ want to maximize in ξ-direction
- ▷ want to minimize in x-direction



Define:

① $P_G(\tau; z) = \underset{x}{\operatorname{argmin}} \left(\tau \cdot G(x) + \frac{1}{2}\|x - z\|_2^2 \right)$

② $P_{F^*}(\sigma, \xi) = \underset{y}{\operatorname{argmin}} \left(\sigma F^*(y) + \frac{1}{2}\|y - \xi\|_2^2 \right)$

Whenever P_G and P_{F^*} are easy to compute (directly or using Moreau's identity).

We perform the following algorithm: <u>Chambolle-Pock Alg.</u>

$$\begin{cases} (x^0, \xi^0) \text{ arbitrary}, \quad \tau, \sigma > 0, \quad \theta \in [0,1] \\ \xi^{n+1} = P_{F^*}(\sigma; \xi^n + \sigma A \bar{x}^n) \\ x^{n+1} = P_G(\tau; x^n - \tau A^T \xi^{n+1}) \\ \bar{x}^{n+1} = x^{n+1} + \theta(x^{n+1} - x^n) \end{cases}$$

⇝ move upwards into ξ-direction

⇝ move downwards into x-direction

⇝ correction

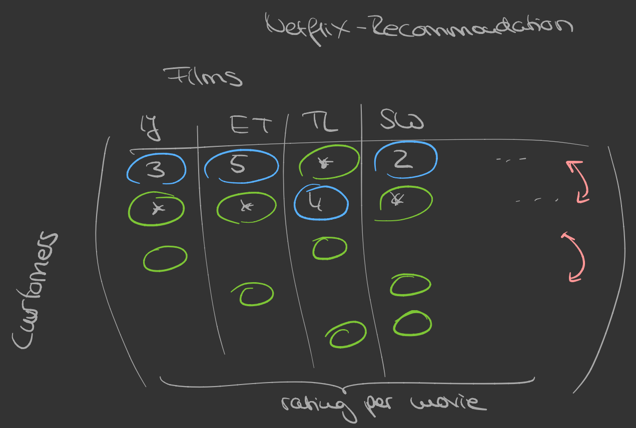
When does this actually converge?

<u>Thm 6.7</u>: If F, G are convex, l.s.c., then for $\theta = 1$, $\tau, \sigma > 0$ and $\tau \cdot \sigma \|A\|^2 < 1$, then

$$(x^n, \xi^n) \longrightarrow (x^\#, \xi^\#) \quad \text{for } n \to \infty.$$

Remark: Depending on F and G we can apply it on lots of problems. The art lies in computing the proximal mappings.

Example: Recall the Netflix problem.



* for unseen movies

want to predict the rating in a collaborative fashion for unseen movies

$\Rightarrow$ "Matrix completion problem" (low-rank matrix completion)

$\Omega \underset{\neq}{\subseteq} I \times J$ (small), we know A_{ij} for $(i,j) \in \Omega$.

How can we find A_{lk}, $(l,k) \in \Omega^c$?

$$X = \underset{\{Z_{ij},\,(i,j)\in\Omega\} = y}{\arg\min} \ \operatorname{rank}(Z)$$

convex relaxation

$$\boxed{X = \underset{\{Z_{ij} \mid (i,j)\in\Omega\} = y}{\arg\min} \ \|Z\|_*}$$

$\Rightarrow$ Falls into the realm of the Chambolle-Pock algorithm.

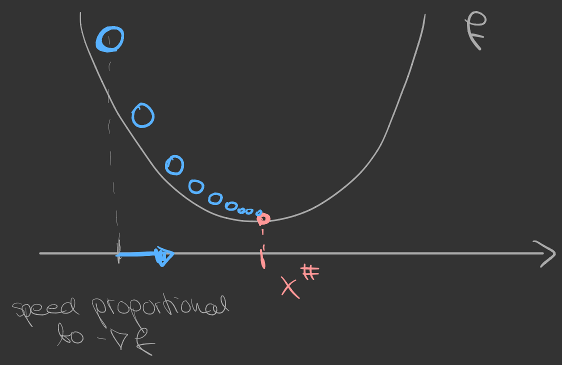
6.1. Gradient descent

We deal with simple unconstraint problems of the form

$$\underset{x \in \mathbb{R}^N}{\arg\min} \ f(x)$$

with f smooth, $f \in \mathcal{C}^2(\mathbb{R}^N)$, strongly convex.

Remark: These are very strong assumptions



$\Rightarrow$ solving a differential equation

$$\begin{cases} \frac{d}{dt} x(t) = -\nabla f(x(t)) \\ x(0) = x^{(0)} \end{cases}$$

Natural discretication: <u>explicit Euler Scheme</u>

$$x^{(n+1)} = x^{(n)} - \alpha^{(n)} \nabla f(x^{(n)})$$

which uses the fact that

$$\frac{d}{dt} x(t^{(n)}) \approx \frac{x(t^{(n+1)}) - x(t^{(n)})}{\alpha^{(n)}}, \quad t^{(n+1)} = t^{(n)} + \alpha^{(n)}.$$

Now, assume that $\exists$ $t \mapsto x(t)$ solving the ODE. We want to prove that

$$\lim_{t\to\infty} x(t) = x^{\#} = \underset{x}{\operatorname{argmin}}\, f(x).$$

twice continuously diff'able

<u>Lemma 6.1</u>: For a function $f \in \mathcal{C}^2(\mathbb{R}^N)$ and f has a Lipschitz-continuous gradient, i.e. $\exists L > 0$

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L \|x-y\| \qquad \forall x,y \in \mathbb{R}^N,$$

then it holds that

$$|f(y) - f(x) - \langle \nabla f(x), y-x \rangle| \le \frac{L}{2} \|x-y\|_2^2.$$

<u>Proof</u>: $\forall x,y \in \mathbb{R}^N$, we can use Taylor expansion with integral-remainder, i.e.

$$f(y) = f(x) + \int_0^1 \langle \nabla f(x+\tau(y-x)), (y-x) \rangle \, d\tau$$

$$= f(x) + \langle \nabla f(x), y-x \rangle - \langle \nabla f(x), y-x \rangle$$

$$+ \int_0^1 \dots \, d\tau$$

$$= f(x) + \langle \nabla f(x), y-x \rangle$$

$$+ \int_0^1 \langle \nabla f(x+\tau(y-x)) - \nabla f(x), y-x \rangle \, d\tau$$

$\varphi(\tau) = x + \tau(y-x)$, $\varphi'(\tau) = (y-x)$

$$= \int_0^1 \langle \nabla f(\varphi(\tau)), \varphi'(\tau) \rangle \, d\tau$$

$$= \int_{\varphi(0)}^{\varphi(1)} \nabla f(t) \, dt = \int_x^y \nabla f(t) \, dt$$

$$= \int_x^y \frac{f^{(0+1)}(t)}{0!} (y-t)^0 \, dt \quad \leftarrow \text{typical "integral form" of the remainder}$$

$$\iff \left| f(y) - \left(f(x) + \langle \nabla f(x), y-x \rangle \right) \right|$$

$$= \left| \int_0^1 \langle \nabla f(x+\tau(y-x)) - \nabla f(x), y-x \rangle \, d\tau \right|$$

$$\leq \int_0^1 \left| \langle \nabla f(x+\tau(y-x)) - \nabla f(x), y-x \rangle \right| d\tau$$

$$\overset{\text{C.S.}}{\leq} \int_0^1 \|\nabla f(x+\tau(y-x)) - \nabla f(x)\|_2 \|y-x\|_2 \, d\tau$$

L-Lipschitz gradient

$$\leq \int_0^1 L \|\cancel{x} + \tau(y-x) - \cancel{x}\| \, \|y-x\|_2 \, d\tau$$

$$\leq L \cdot \|x-y\|^2 \int_0^1 \tau \, d\tau$$

$$= \frac{L}{2} \|x-y\|^2.$$

$\blacksquare$

Lemma 6.2 : Co-Coercivity condition

For a convex function $f \in \mathcal{C}^2(\mathbb{R}^N)$ with a Lipschitz-continuous gradient, we have

$$\|\nabla f(x) - \nabla f(y)\|_2^2 \leq L \langle x-y, \nabla f(x) - \nabla f(y) \rangle.$$

Remark: We have

$$\|\nabla f(x) - \nabla f(y)\|_2^2 \leq L \langle x-y, \nabla f(x) - \nabla f(y) \rangle.$$

$$\overset{\text{C.S.}}{\Longrightarrow} \quad \|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x-y\|.$$

But the lemma gives us the opposite direction under specific conditions.

Proof: skipped.

Theorem 6.3: Assume $f \in \mathcal{C}^2(\mathbb{R}^N)$, strongly convex, i.e.

$$\nabla^2 f(x) \geq \frac{\gamma}{2} I, \quad \rightsquigarrow \quad v^T \nabla^2 f(x) v \geq \gamma \|v\|^2$$

and $\nabla^2 f(x) \leq L I.$ $\rightsquigarrow$ implies L-Lipschitz continuity

Then the solution to the ODE always converges to the minimizes of f, i.e.

$$\lim_{t \to \infty} x(t) = x^* = \underset{x \in \mathbb{R}^N}{\operatorname{argmin}} f(x)$$

Proof: $\nabla f(x^*) = 0$ because x^* is minimizes and f strongly convex. Let's denote

$$g(x) = \frac{1}{2} \|\nabla f(x)\|_2^2 .$$

Then

chain rule

$$\frac{\partial}{\partial t} g(x(t)) = \langle \nabla g(x(t)), \frac{d}{dt} x(t) \rangle$$

$$= \nabla f(x(t))^T \nabla^2 f(x(t)) \frac{d}{dt} x(t)$$

Def. of the ODE

$$\frac{d}{dt} x(t) = - \nabla f(x(t)) \qquad = - \underbrace{\nabla f(x(t))^T}_{v^T} \nabla^2 f(x(t)) \underbrace{\nabla f(x(t))}_{v}$$

Lipschitz ass. for $\nabla^2 f(x)$

$$\leq - \gamma/2 \|\nabla f(x(t))\|_2^2$$

$$= - \gamma g(x(t)).$$

$$\Rightarrow \frac{d}{dt} g(x(t)) \leq - \gamma g(x(t)).$$

$$\Rightarrow \int \frac{\frac{d}{dt} g(x(t))}{g(x(t))} dt \leq - \int \gamma \, dt \quad \dots$$

$$\Rightarrow \frac{1}{2} \|\nabla f(x)\|^2 \overset{\text{Def.}}{=} g(x(t))$$

$$\leq g(x(0)) e^{-\gamma t}$$

$$\overset{\text{Def.}}{=} \frac{1}{2} \|\nabla f(x(0))\|_2^2 e^{-\gamma t} \xrightarrow{t \to \infty} 0, \quad \gamma > 0.$$

$$\Rightarrow \quad t \longmapsto \frac{1}{2}\|\nabla f(x(t))\|_2^2 \xrightarrow{t\to\infty} 0$$

$$\Rightarrow \|\nabla f(x(t)) - \underbrace{\nabla f(x^*)}_{=0}\|_2^2 \xrightarrow{t\to\infty} 0 . \qquad (*)$$

Now, we consider

$$\frac{d}{dt}\frac{1}{2}\|x(t)-x^*\|^2 = \langle x(t)-x^*, \frac{d}{dt}x(t)\rangle$$

$$\overset{\text{ODE}}{=} -\langle x(t)-x^*, \nabla f(x(t)) - \underbrace{\nabla f(x^*)}_{=0}\rangle$$

$$\overset{\text{Lemma 6.2}}{\leq} -\frac{1}{L}\|\nabla f(x(t)) - \nabla f(x^*)\|_2^2 \leq 0$$

can only decrease wrt time

$$\Rightarrow \frac{1}{2}\|x(t)-x^*\|_2^2 \leq \frac{1}{2}\|x(0)-x^*\|_2^2 \in \mathbb{R} \text{ constant.}$$

value at time 0

$\Rightarrow$ $t \longmapsto x(t)$ is always bounded, i.e. defining

$(t_n)_n$, $t_n \geq t_{n+1} \geq t_{n+2}$ ..... and $(x(t_m))_m$.

Then this sequence is bounded, and hence there exists a subsequence $(x(t_{n_k}))_k$ that converges, i.e. $\exists \bar{x}$ s.t.

Bolzano-Weierstraß

$$\lim_{k\to\infty} x(t_{n_k}) = \bar{x}.$$

$$\Rightarrow \lim_{k\to\infty} \|\nabla f(x(t_{n_k})) - \nabla f(x^*)\|_2$$

$$= \|\nabla f(\bar{x}) - \nabla f(x^*)\|_2$$

$$\Rightarrow \nabla f(\bar{x}) = 0$$

strong convexity

$$\Rightarrow \bar{x} = x^*.$$

$\blacksquare$

Vorlesung 25

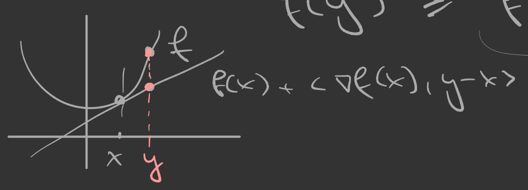
21.07.2020

We analyze now the algorithm, called <u>gradient descent</u>

$$\begin{cases} x^{(n+1)} = x^{(n)} - \alpha^{(n)} \nabla f(x^{(n)}) \\ x^{(0)} \text{ is given.} \end{cases}$$

We have proven that if f is convex and f is diffable. Then

$$f(y) \geq \underbrace{f(x) + \langle \nabla f(x), y-x \rangle}_{\text{affine lower bound for } f} \quad \forall x,y \in \mathbb{R}^d$$

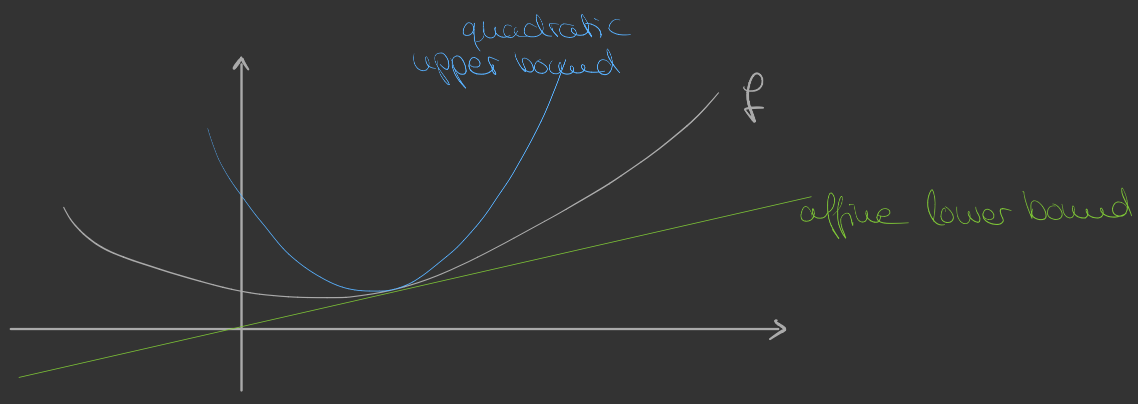


If f has additionally an L-Lipschitz-continuous gradient, $L>0$, i.e.

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

then by lemma 6.1, we have the upper bound

$$f(y) \leq f(x) + \langle \nabla f(x), y-x \rangle + \frac{L}{2} \| y-x \|_2^2 .$$



<u>Theorem 6.4</u>: Assume that f is convex and finite at every $x \in \mathbb{R}^N$, a finite solution x^* exists, and $\nabla f(x)$ is globaly L-Lipschitz continuous, $L>0$. Then for $\alpha = \alpha^n < \frac{1}{L}$ the gradient descent

$$\begin{cases} x^{(n+1)} = x^{(n)} - \alpha^{(n)} \nabla f(x^{(n)}) \\ x^{(0)} \text{ is given.} \end{cases}$$

converges to x^*.

Proof:

① Step: We wish to bound from above $|f(x^{(n+1)}) - f(x^*)|$ in terms of $\|x^{(n)} - x^*\|_2^2$ and $\|x^{(n+1)} - x^*\|_2^2$. To simplify notation, write $x = x^{(n)}$, $x^+ = x^{(n+1)}$, $\alpha = \alpha^{(n)}$.

By Lem. 6.1., we have

$$
\begin{aligned}
f(x^+) &\leq f(x) + \langle \nabla f(x), x^+ - x \rangle + \frac{L}{2}\| x^+ - x\|_2^2 \\
&\overset{\text{Def } x^+}{=} f(x) + \langle \nabla f(x), -\alpha \nabla f(x) \rangle + \frac{L}{2}\| -\alpha \nabla f(x)\|_2^2 \\
&= f(x) - \alpha \|\nabla f(x)\|_2^2 + \frac{\alpha^2 L}{2}\|\nabla f(x)\|_2^2 \\
&= f(x) - \alpha\left(1 - \frac{\alpha L}{2}\right)\|\nabla f(x)\|_2^2 \\
&\overset{\alpha \leq \frac{1}{L},\ \frac{\alpha L}{2} \leq \frac{1}{2}}{\leq} f(x) - \frac{\alpha}{2}\|\nabla f(x)\|_2^2 \\
&\overset{\text{convexity} \Rightarrow \text{affine lower bound}}{\leq} f(x^*) + \langle \nabla f(x), x - x^* \rangle - \frac{\alpha}{2}\|\nabla f(x)\|_2^2 \\
&\overset{\star}{=} f(x^*) + \frac{1}{2\alpha}\left(\|x - x^*\|_2^2 - \|\underbrace{x - \alpha\nabla f(x)}_{= x^+} - x^*\|_2^2 \right) \\
&= f(x^*) + \frac{1}{2\alpha}\left(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2 \right).
\end{aligned}
$$

In summary, we have obtained

$$f(x^{(n+1)}) - f(x^*) \leq \frac{1}{2\alpha}\left(\|x^{(n)} - x^*\|_2^2 - \|x^{(n+1)} - x^*\|_2^2\right).$$

② Step: We show that $f(x^{(n)}) - f(x^*) \xrightarrow{n \to \infty} 0$

$$
\begin{aligned}
\sum_{n=0}^{\infty} f(x^{(n+1)}) - f(x^*) &\leq \frac{1}{2\alpha} \sum_{n=0}^{\infty} \|x^{(n)} - x^*\|_2^2 - \|x^{(n+1)} - x^*\|_2^2 \\
&\qquad \text{Telescopic sum} \\
&= \frac{1}{2\alpha}\|x^{(0)} - x^*\|_2^2 < \infty.
\end{aligned}
$$

$\Rightarrow f(x^{(n+1)}) - f(x^*))$ has to be a null sequence.

$\sum_{i=0}^{n} a_n < \infty$, $a_n \geq 0$

$\Rightarrow \lim_n a_n = 0$.

Using the co-coercivity lemma analogously to prior proof, we can show that if x^* is the unique minimizer, then

$$x^{(n)} \to x^*.$$

$\boxtimes$

<u>Remark</u>: If f is non-convex, we might only end up at a local minimum.
The behavior can be improved using e.g. "inertial terms" but they are not reliable.

6.2. Stochastic gradient descent

already used probability for stable SVD, overcoming curse of dimensionality and for Johnson-Lindenstrauss.

$$f(x) = \mathbb{E}\, f(x,\cdot) = \int_\Omega f(x,\omega)\, d\mathbb{P}(\omega)$$

If $f(x,\cdot)$ is having a density ϱ_f not depending on x, we can write

identically distributed wrt x.

$$f(x) = \int_{\mathbb{R}} f(x,y)\, \varrho_f(y)\, dy.$$

If $\Omega = \mathbb{N}$, then

$$f(x) = \sum_{k=0}^{\infty} f(x, \omega^{(k)}) \cdot p^{(k)} \qquad \text{with} \quad \sum_{k=0}^{\infty} p^{(k)} = 1.$$

Additionally, we have that

$$\nabla f(x) = \int_\Omega \nabla_x f(x,\omega)\, d\mathbb{P}(\omega).$$

<u>Problem</u>: Lots of gradients would need to be calculated, to get $\nabla f(x)$.

<u>Example</u>: Assume we have a feed forward neural network NN.

$$x \longmapsto NN(x, (W,\theta)) = y.$$

with weights W, biases θ as parameters. We can understand a L-hidden layer NN as the iterative algorithm

$$
\begin{cases}
x^{(n+1)} = \sigma\left(W^{(n)} x^{(n)} + \theta^{(n)}\right) & \\
y = x^{(L)} & \text{output} \\
x^{(0)} = x & \text{input}
\end{cases}
$$

Given $x_1, \dots, x_N$ images of animals and $y_1, \dots, y_N$ the description of the animals, we want to find (W^*, θ^*) s.t.

$$NN(x_j, (W^*, \theta^*)) = y_j \quad \forall\, j \in [N].$$

We can write

$$
\begin{cases}
z^{(\ell)} = (x^{(\ell)}, y^{(\ell)}) \quad , \quad \ell \in [N] \\
f((W,\theta), z^{(\ell)}) = \| NN(x^{\ell}, (W,\theta)) - y^{\ell} \|_2^2 \\
f(W,\theta) = \frac{1}{N} \sum_{\ell=1}^{N} f((W,\theta), z^{(\ell)}) \\
\qquad = \sum_{\ell=1}^{N} f((W,\theta), z^{(\ell)}) \cdot p^{(\ell)} \quad , \quad p^{(\ell)} = \frac{1}{N}.
\end{cases}
$$

Then training the neural network boils down to reducing the discrepancy. Our pairs of training data are stochastic values, since they are aquired in a random fashion.

If we want to use gradient descent,

$$\nabla_{(W,\theta)} f(W,\theta) = \frac{1}{N} \sum_{i=1}^{N} \nabla_{(W,\theta)} f((W,\theta), z^{(\ell)})$$

one for each data pair

would involve an huge amount of gradients including the calculation of the gradient of the NN including chain rule, also called <u>backpropagation</u>.

Then our iteration would read

$$\left(w^{(n+1)}, \theta^{(n+1)}\right) = \left(w^{(n)}, \theta^{(n)}\right) + \alpha^{(n)} \nabla_{(w,\theta)} f\left(\left(w^{(n)}, \theta^{(n)}\right)\right)$$

This computational effort is out of discussion.

What can we do instead?

<u>Definition</u>: For $f(x) = \int_\Omega f(x,\omega)\, d\mathbb{P}(\omega)$ we consider the <u>stochastic gradient descent</u> (SGD)

single gradient on one data point

$$\tilde{x}^{(n+1)} = \tilde{x}^{(n)} - \alpha^{(n)} \nabla_x f(\tilde{x}^{(n)}, \omega^{(n)})$$

where $\omega^{(n)}$ is drawn at random according to $\mathbb{P}$.

$\hookrightarrow$ can work due to LLN

<u>Example</u>: (continued)

$$\left(\tilde{w}^{(n+1)}, \tilde{\theta}^{(n+1)}\right) = \left(\tilde{w}^{(n)}, \tilde{\theta}^{(n)}\right)$$

$$- \alpha^{(n)} \nabla_{(w,\theta)} \left(\| NN(\tilde{x}^{(\ell)}, (\tilde{w}^{(n)}, \tilde{\theta}^{(n)})) - y^{(\ell)} \|_2^2 \right),$$

where $\omega^{(\ell)} = (x^{(\ell)}, y^{(\ell)})$ are picked at random.

<u>Theorem 6.5</u>:

We consider the stochastic gradient descent (SGD) defined by

$$\begin{cases} x^{(n+1)} = x^{(n)} + \alpha \nabla_x f(x^{(n)}, \omega^{(n)}) \\ x_0 \text{ is given.} \end{cases}$$

Let each of the $f(\cdot, \omega)$ be convex with $x \mapsto \nabla f(x,\omega)$ $L(\omega)$-Lipschitz continuous, $L(\omega) > 0$ and $L = \sup_\omega L(\omega) < +\infty$ and let

$$f(x) = \mathbb{E}_\omega [f(x,\omega)]$$

be γ-strongly convex. Set $\sigma^2 = \underbrace{\mathbb{E}_\omega\left[\| \nabla_x f(x^*, \omega)) \|_2^2 \right]}_{\text{variance around the optimizer}}$

where $x^* = \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} f(x)$. Suppose $0 < \alpha < \frac{1}{L}$. $\longleftrightarrow \alpha L < 1$

Then the SGD satisfies

$$\mathbb{E}\left[\| x^{(n)} - x^* \|_2^2 \right] \leq \underbrace{\left(1 - 2\alpha\gamma(1 - \alpha \cdot L)\right)^n}_{n \to \infty \;\longrightarrow\; 0} \cdot \| x^{(0)} - x^* \|_2^2 + \underbrace{\frac{\alpha \cdot \sigma^2}{\gamma(1 - \alpha \cdot L)}}_{\text{small by choice of } \alpha}$$

where the expectation is over all the sequence of sampling $(\omega^{(\ell)})_{\ell \in \mathbb{N}}$.

<u>Corollary 6.6</u>: For any desired ε, using stepsize of

$$\alpha = \frac{\gamma \cdot \varepsilon}{2\varepsilon\gamma L + 2\sigma^2}$$

we have that after

$$n \geq 2 \log(2\varepsilon_0 / \varepsilon) \left(\frac{L}{\gamma} + \frac{\sigma^2}{\gamma^2 \varepsilon} \right)$$

iterations, the SGD fulfills that

$$\mathbb{E}_\omega\left[\| x^{(n)} - x^* \|_2^2 \right] \leq \varepsilon \longrightarrow \text{in expectation "only"}$$

not converging, but getting close in a neighbourhood

where $\varepsilon_0 = \| x^0 - x^* \|_2^2$.

<u>In short</u>: SGD doesn't converge for fixed α, but we can choose α and n s.t. we get arbitrarily close to the true global minimizer x^*.

Vorlesung 26

22.07.2020

Proof: for Thm 6.5

$\omega^{(n)}$ are randomly sampled from Ω at each iteration $n \in \mathbb{N}$.

For simplicity, we denote

$$f_n(x) = f(x, \omega^{(n)})$$

Recall that $f(x) = \mathbb{E}_\omega[f(x,\omega)]$. Then

$$x^{(n+1)} = x^{(n)} - \alpha \nabla_x f_n(x^{(n)}).$$

Consider

$$\|x^{(n+1)} - x^*\|_2^2 = \| x^{(n)} - \alpha \nabla_x f_n(x^{(n)}) - x^*\|_2^2$$

$$= \| x^{(n)} - x^*\|_2^2 - 2\alpha < x^{(n)} - x^*, \nabla_x f_n(x^{(n)}) >$$

$(a+b)^2 \leq 2(a^2+b^2)$, $a,b \in \mathbb{R}$

$\Leftrightarrow a^2 + b^2 + 2ab \leq 2(a^2+b^2)$

$\Leftrightarrow 2ab \leq a^2 + b^2$

$\Leftrightarrow 0 \leq a^2 - 2ab + b^2 = (a-b)^2$ ✓

$$+ \alpha^2 \| \nabla_x f_n(x^{(n)}) - \nabla_x f_n(x^*) + \nabla_x f_n(x^*)\|_2^2$$

$$\leq \| x^{(n)} - x^*\|_2^2 - 2\alpha < x^{(n)} - x^*, \nabla_x f_n(x^{(n)}) >$$

$$+ 2\alpha^2 \|\nabla_x f_n(x^{(n)}) - \nabla_x f_n(x^*)\|_2^2$$

$$+ 2\alpha^2 \| \nabla_x f_n(x^*)\|_2^2$$

Lipschitz-prop. + Co-coercivity

$$\leq \| x^{(n)} - x^*\|_2^2 - 2\alpha < x^{(n)} - x^*, \nabla_x f_n(x^{(n)}) >$$

$$+ 2\alpha^2 L(\omega^{(n)}) < x^{(n)} - x^*, \nabla_x f_n(x^{(n)}) - \nabla_x f_n(x^*) >$$

$$+ 2\alpha^2 \| \nabla_x f_n(x^*)\|_2^2$$

Taking the expectation on both sides:

$$\mathbb{E}_\omega\left[\| x^{(n+1)} - x^*\|_2^2\right] \leq \| x^{(n)} - x^*\|_2^2 - 2\alpha < x^{(n)} - x^*, \mathbb{E}\left[\nabla_x f_n(x^{(n)})\right] >$$

$$+ 2\alpha^2 \mathbb{E}\left[L(\omega^{(n)}) < x^{(n)} - x^*, \nabla_x f_n(x^{(n)}) - \nabla_x f_n(x^*) >\right]$$

$$+ 2\alpha^2 \mathbb{E}\left[\| \nabla_x f_n(x^*)\|_2^2\right]$$

indep. of ω

deped. on ω

$\mathbb{E}_\omega[\|\nabla_x f(x^*,\omega)\|_2^2] = \sigma^2$

$\mathbb{E}_\omega[\nabla_x f(x,\omega)] = \nabla_x f(x)$

$$\leq \|x^{(n)} - x^*\|_2^2 - 2\alpha \langle x^{(n)} - x^*, \nabla_x f(x^{(n)}) \rangle$$
$$+ 2\alpha^2 \underbrace{\mathbb{E}\Big[L(\omega^{(n)}) \underbrace{\langle x^{(n)} - x^*, \nabla_x f_n(x^{(n)}) - \nabla_x f_n(x^*) \rangle}_{\text{co-coercivity } =: \Gamma(\omega) \geq \frac{1}{L}\|\nabla_x f_n(x^{(n)}) - \nabla_x f(x^*)\|_2^2 \geq 0}\Big]}_{\text{non-negative } \leq \sup_{\omega\in\Omega} L(\omega)\, \mathbb{E}[\Gamma(\omega)]}$$
$$+ 2\alpha^2\sigma^2$$

$$\leq \|x^{(n)} - x^*\|_2^2 - 2\alpha \langle x^{(n)} - x^*, \nabla_x f(x^{(n)}) \rangle$$
$$+ 2\alpha^2 \cdot L \langle x^{(n)} - x^*, \nabla_x f(x^{(n)}) - \nabla_x f(x^*) \rangle$$
$$+ 2\alpha^2\sigma^2$$

$F(x) \geq F(y) + \langle \nabla F(y), y - x \rangle + \frac{\gamma}{2}\|x-y\|_2^2$

$\nabla_x f(x^*) = 0, \ f(x^{(n)}) - f(x^*) \geq 0$

$\star$

$$\leq \|x^{(n)} - x^*\|_2^2 - 2\alpha \cdot \underset{\text{strong convexity}}{\underset{\uparrow}{\gamma}} (1 - \alpha L) \|x^{(n)} - x^*\|_2^2$$
$$+ 2\alpha^2\sigma^2.$$

Summary:

$$\mathbb{E}_\omega\left[\|x^{(n+1)} - x^*\|_2^2\right] \leq \|x^{(n)} - x^*\|_2^2 - 2\alpha \cdot \gamma (1 - \alpha L) \|x^{(n)} - x^*\|_2^2$$
$$+ 2\alpha^2\sigma^2.$$

$$= (1 - 2\alpha\gamma(1 - \alpha L)) \|x^{(n)} - x^*\|_2^2 + 2\alpha^2\sigma^2.$$

Iterating this over n, we get that

$$\mathbb{E}_{\omega^{(n-1)}} \mathbb{E}_{\omega^{(n)}} \left[\|x^{(n+1)} - x^*\|_2^2\right]$$
$$\leq (1 - 2\alpha\gamma(1 - \alpha \cdot L))\, \mathbb{E}_{\omega^{(n-1)}}\left[\|x^{(n)} - x^*\|_2^2\right] + 2\alpha^2\sigma^2$$
$$\leq (1 - 2\alpha\gamma(1 - \alpha L))^n \|x^{(0)} - x^*\|_2^2$$
$$+ 2 \underbrace{\sum_{k=0}^{n-1} (\underbrace{1 - 2\alpha\gamma(1 - \alpha L)}_{<1})^k}_{\leq \frac{1}{1 - (1 - 2\alpha\gamma(1 - \alpha L))}} \alpha^2\sigma^2$$

$\sum_{i=0}^{\infty} \alpha^k \leq \frac{1}{1-\alpha}$ for $\alpha < 1$.

$$\leq (1 - 2\alpha\gamma(1 - \alpha L))^n \|x^{(0)} - x^*\|_2^2 + \frac{\alpha\sigma^2}{\gamma(1 - \alpha L)}.$$

6.4. Nonconvex optimization

Other approach: Convexify the problem: "Graduating Non-Convexity"

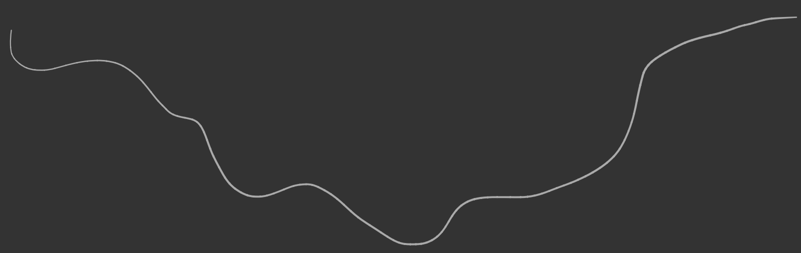
It can happen that $f: \mathbb{R}^d \to \mathbb{R}$ is nonconvex, but $\exists\, \omega > 0$, s.t. $\forall y \in \mathbb{R}^d$:

$$f_{y,\omega}(x) = f(x) + \underbrace{\omega \|x-y\|_2^2}_{\text{adding a strictly convex parabola}}$$

is (strictly) convex. Such functions f are called ω-convex.

(Also see "Graduating Non-Convexity" in Lecture 20).

Examples: ▷ $f \in \mathcal{C}^1$, i.e. continuously diff'able with piece-wise continuous and bounded second derivative.
$\Leftrightarrow$ not "too" sharp edges



Hence, we can use the following algorithm:

$$x^{(n+1)} = \underset{x}{\operatorname{argmin}}\ f(x) + \omega \|x - x^{(n)}\|_2^2$$

$$= P_{\omega^{-1} f}(x^{(n)}).$$ $\leftarrow$ proximal mapping, well defined since f is ω-convex

If $x^{(n)} \to \bar{x}$, then $\|x - x^{(n)}\| \to 0$, i.e. the collection becomes less and less relevant. Then this approach boils down to gradient descent, i.e.

$$\nabla f(x^{(n+1)}) + 2\omega(x^{(n+1)} - x^{(n)}) = 0$$

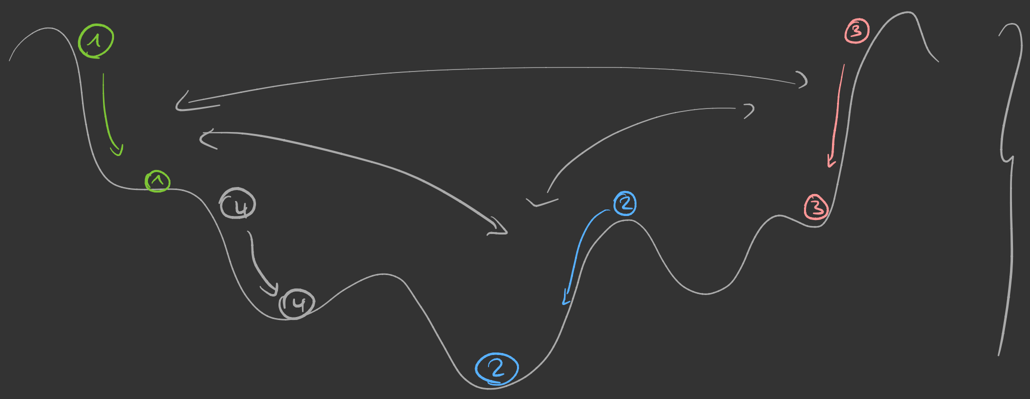
$$\Leftrightarrow \quad x^{(n+1)} - x^{(n)} = -\frac{1}{2\omega} \nabla f(x^{(n+1)})$$

$$\Leftrightarrow \quad x^{(n+1)} = x^{(n)} - \frac{1}{2\omega} \nabla f(\underset{\text{implicit}}{x^{(n+1)}}).$$

"Implicit Gradient Descent"

One can prove that $x^{(k)} \to \bar{x}$ where $\bar{x}$ is a steady state, i.e. $\nabla f(\bar{x}) = 0$, but no convergence to global minimizer.

<u>Other methods</u>: <u>Consensus based optimization</u>



running multiple GD-algorithms

"Best Case": One ball has a better value than any other

Adding shared information between all algorithms x^i, $i \in [N]$

global thinking

stochastic + deterministic

local thinking

$$X^i_{t+\Delta t} = X^i_t - \underbrace{\lambda \cdot \Delta t \cdot (X^i_t - m_{\alpha,t})}_{\text{deterministic, drift towards } m_{\alpha,t}}$$

$$+ \underbrace{\sigma \sqrt{\Delta t}\, |X^i_t - m_{\alpha,t}| \mathcal{U}_t}_{\text{stochastic, } \mathcal{U}_t \sim N(0,1)\text{, exploration unless } x^i_t \neq m_{\alpha,t} \text{ i.e. not agreed yet}}$$

with $$m_{\alpha,t} = \frac{1}{\sum_{i=1}^{n} e^{-\alpha f(x^i_t)}} \sum_{i=1}^{n} x^i_t e^{-\alpha f(x^i_t)}$$

weighted center of mass, where small values have a larger value

This algorithm actually works quite well.

One more advantage: we don't need any gradient.