Project in mathematics
Proximal operators and Von Neumann inequality

Linn Öström, linn.ostrom@math.lth.se

July 15, 2020

In this project in mathematics I'll consider two distinct topics, proximal operators and the proof of the von Neumann inequality. For the proximal operators I'll present a summary of readings and solve some general convex optimization problems, while for the proof I'll replicate the proof shown in [1]. The sections will thereby divided into these two topics.

Proximal operators

Proximal operators or more specifically proximal gradient methods are a central algorithms in solving non-differetiable convex optimization problem. In [2] the method is referred to as the classical choice of algorithm for solving non-smooth optimization problems, just as Newtons method is for solving unconstrained smooth optimization problems in lower dimensions.

In this short summary I'll summarize important concepts and results for proximal operators. This work has been heavily influenced by *Proximal Operators* [2] and *First-order methods in Optimization* [3]. All definitions and theorems are presented in a similar manner as in [2] and [3]. All spaces considered in this summary are Hilbert spaces if nothing else is stated, and when referring to a general norm what is meant is the Euclidean norm $||\cdot||_2$.

The basics

Before heading into the main topic proximal operators, we need to define some preliminary concepts frequently used in optimization. Starting off with the definition of a convex function, the epigraph $\mathbf{epi}\, f$ and the domain $\mathbf{dom}\, f$ of a function.

Definition: The *effective domain* $\mathbf{dom}\, f$ of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the set

$$\mathbf{dom}\, f = \{\boldsymbol{x} \in \mathbb{R}^n |\, f(\boldsymbol{x}) < \infty\}$$

i.e. thus consists of the set of points that maps f onto finite values.

Definition: A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *convex* if $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $\lambda \in [0, 1]$ we have that

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}). \tag{1}$$

If the inequality is strict we say that the function is *strictly convex*. Further if the inequality is mirrored we get the definition of a concave function, and strictly concave function. Since proximal operators mostly are useful for convex functions, we will only consider these when deriving the proximal operators in this summary. A very similar definition of convexity holds for convex sets.

Definition: A set $\mathcal{C}$ is said to be convex if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$ and $\lambda \in [0, 1]$ it holds that

$$\lambda \boldsymbol{x} + (1 - \lambda)\boldsymbol{y} \in \mathcal{C}.$$

Some examples of convex sets (left and middle) and non-convex set (right) are shown in the figure below.

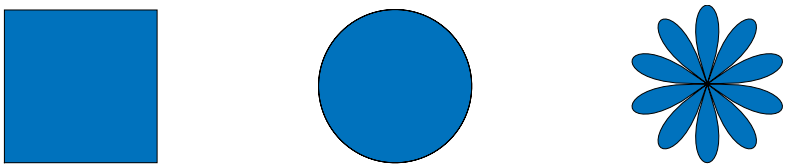

Figure 1: Example of convex and non-convex sets.

Definition: A *epigraph* of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the defined as the set

$$\mathbf{epi}\, f = \{(\boldsymbol{x}, t) \in \mathbb{R}^n \times \mathbb{R} | f(x) \leq t\}. \tag{2}$$

Definition: A $\alpha-$*sublevel* set for a function f is defined as the set

$$\mathcal{S}_\alpha = \{\boldsymbol{x} \in \mathrm{dom}\, f | f(\boldsymbol{x}) \leq \alpha\}$$

Definition: A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *closed* if for each $\alpha \in \mathbb{R}$, the sublevel set $\mathcal{S}_\alpha = \{\boldsymbol{x} \in \mathrm{dom}\, f | f(\boldsymbol{x}) \leq \alpha\}$ is a closed set.

Definition: A convex function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *proper* its effective domain is nonempty and $f(x) > -\infty$, $\forall x \in \mathbf{dom}\, f$.

The function f is a closed proper convex function if the epigraph set (in Eq. (2)) is a closed convex set. Some examples of convex functions are shown in the left and middle plots in Figure 2. Note that the right figure is not convex as the chord between the point the points $x = -1$ and $x = 1$ will lie below the graph, thus its epi-graph is a non-convex set.

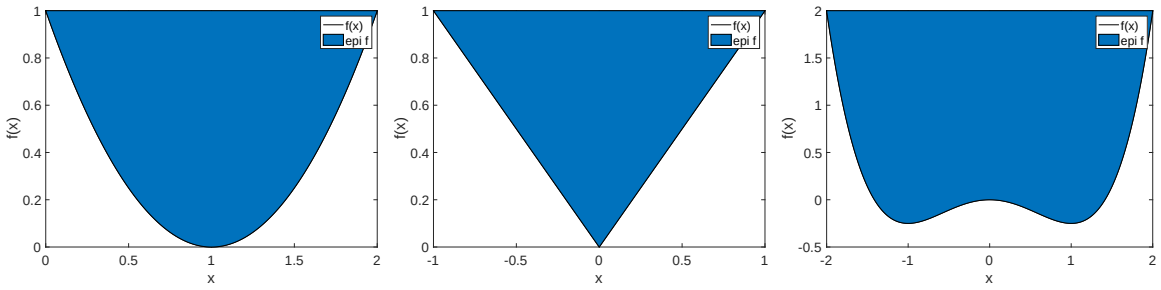

Figure 2: Example of convex and non-convex functions, at the left $f_1(x) = (x-1)^2$, in the middle $f_2(x) = |x|$ and to the right $f_3(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2$. Convex functions are shown in left and middle, while the right function is not.

The functions shown in the figures are; at the left $f_1(x) = (x-1)^2$, in the middle $f_2(x) = |x|$ and to the right $f_3(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2$. Further, the corresponding epigraphs of the functions are filled in blue. Again note that the epigraphs in the left and middle figures are convex sets while this does not hold for the right function. Next, we will define the subdifferential of a function f, which is particularly useful for non-differential functions.

Definition: The *Subdifferential* of a convex function f at $\boldsymbol{x}$ is defined by

$$\partial f(\boldsymbol{x}) = \{\boldsymbol{y} | f(\boldsymbol{z}) \geq f(\boldsymbol{x}) + \boldsymbol{y}^T(\boldsymbol{z} - \boldsymbol{x}), \text{ for all } \boldsymbol{z} \in \mathbf{dom}\, f\}$$

thus note that this operator maps points to sets. Further, note that if f is differentaible we analogously get $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$ for all $\boldsymbol{x}$. Let us consider som examples of this subdifferential for the non-differetiable and differetiable functions $f_1(x) = |x|, f_2(x) = x^2$ in $\mathbb{R}$. The subdifferential the two functions thus becomes

$$\partial f_1(x) = \begin{cases} \{-1\} & \text{if } x < 0 \\ \{1\} & \text{if } x > 0 \\ \{y \in \mathbb{R} | -1 \leq y \leq 1\} & \text{if } x = 0 \end{cases} \qquad \partial f_2(x) = \{2x\}, \quad \forall x \in \mathbb{R}.$$

For $\partial f_1(x)$ the subdifferential consists of a single element except at the non-differential point $x = 0$ where the subdifferential instead consist of an uncoutable set $\{y \in \mathbb{R} | -1 \leq y \leq 1\}$, a illustration of the subdifferential for some functions are shown in Figure 3. Convex envelop-connection to subdieffertial?. Now we are ready to dive into the main topic in this summary i.e. the proximal operator.

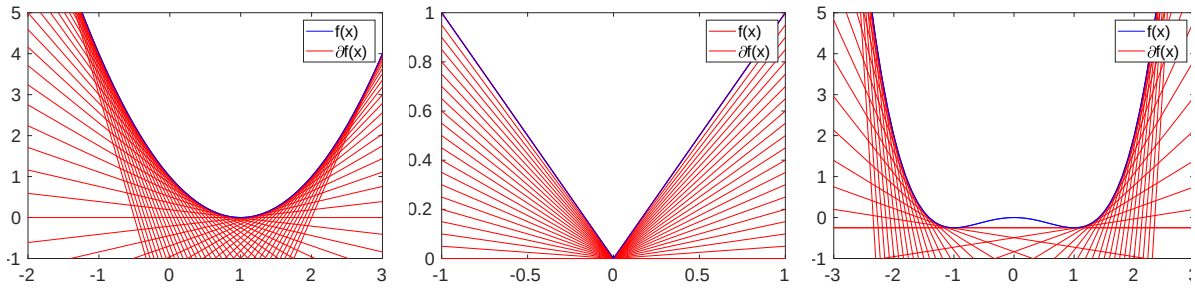

Figure 3: Example subdifferentials of convex and non-convex functions, at the left $f_1(x) = (x-1)^2$, in the middle $f_2(x) = |x|$ and to the right $f_3(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2$. Here the sampled subdifferential are shown as tangents.

The proximal operator

Definition: A proximal operator, or proximal mapping, of a function f is given by

$$\text{prox}_f(\boldsymbol{x}) = \underset{\boldsymbol{u} \in \mathbb{R}^n}{\text{argmin}} \left\{ f(\boldsymbol{u}) + \frac{1}{2}||\boldsymbol{u} - \boldsymbol{x}||_2^2 \right\}, \qquad \text{for any } \boldsymbol{x} \in \mathbb{R}^n. \tag{3}$$

The proximal operator maps a vector $\boldsymbol{x} \in \mathbb{R}^n$ onto subset of the space. This operator has been shown to been useful in convex optimization, when one targets to solve a either non-smooth, non-differentaible, constrained or large-scale problem [2]. Locally, the proximal operator essentially solves a smaller convex minimization problem instead of the might harder one given by f. From the construction of the operator one can thus instead use simpler and less computational effort to solve the optimization-problem by solving smaller sub-problems. In general one often instead consider the proximal operator of the scaled function λf with a parameter $\lambda > 0$, and thus one simply ends up with the similar expression

$$\text{prox}_{\lambda f}(\boldsymbol{x}) = \underset{\boldsymbol{u} \in \mathbb{R}}{\text{argmin}} \left\{ f(\boldsymbol{u}) + \frac{1}{2\lambda}||\boldsymbol{u} - \boldsymbol{x}||_2^2 \right\}, \qquad \text{for any } \boldsymbol{x} \in \mathbb{R}^n. \tag{4}$$

Where inverse λ has the geometrical interpretation of being a step-size, i.e. a large λ yields a smaller penalization for moving far from $\boldsymbol{x}$ and a small λ penalizes large movements from $\boldsymbol{x}$. We will consider this effect in more detail later on.

Properties of the operator and the First Proximal Theorem

The proximal operator has the following useful properties:

Postcomposition: If f has the form $f(x) = \alpha\phi(x) + b$ with $\alpha > 0$ then

$$\mathbf{prox}_{\lambda f}(u) = \mathbf{prox}_{\lambda\alpha\phi}(u)$$

Precomposition: If $f = \phi(\alpha x + b)$, $\alpha \neq 0$ then

$$\mathbf{prox}_{\lambda f}(u) = \frac{1}{\alpha^2}\left(\mathbf{prox}_{\lambda\alpha^2\phi}(\phi(\alpha u + b) - b\right)$$

and if Q is an orthogonal matrix then $f(x) = \phi(Qx)$ then $\mathbf{prox}_{\lambda f}(u) = Q^T\mathbf{prox}_{\lambda\phi}(Qu)$.

Affine addition: If $f(x) = \phi(x) + a^T x + b$ then

$$\mathbf{prox}_{\lambda f}(u) = \mathbf{prox}_{\lambda\phi}(u - \alpha a)$$

Regularization: If $f(x) = f(x) = \phi(x) + (\rho/2)||x - a||_2^2$

$$\mathbf{prox}_{\lambda f}(u) = \mathbf{prox}_{\lambda/(1+\lambda\rho)f}\left(\frac{1}{1+\lambda\rho}u + \rho\frac{\lambda}{1+\lambda\rho}a\right)$$

Further, when considering separable functions i.e. $f : \mathbb{R}^n \to \mathbb{R}$ where $f(x_1, \ldots, x_n) = \sum_{i=1}^m f_i(x_i)$ for any $\boldsymbol{x} \in \mathbb{R}^n$, we will show that evaluating the proximal operator of a separable function reduces to evaluating the proximal operators for each of the separable parts. The proximal operator for a separable function is given by

$$\text{prox}_f(x_1, \ldots, x_n) = \text{argmin}_{u_1,\ldots,u_n}\left[\sum_{i=1}^m f_i(x_i) + \frac{1}{2}||u_i - x_i||_2^2\right] \tag{5}$$

$$= \prod_{i=1}^n \text{argmin}_{u_i}\left[\sum_{i=1}^m f_i(x_i) + \frac{1}{2}||u_i - x_i||_2^2\right] = \prod_{i=1}^n \text{prox}_{f_i}(x_i) \tag{6}$$

since the squared norm is also separable. Further, if f, f_i $i = 1, \ldots, n$ are proper closed convex and separable, and each f_i also univariate then (6) becomes

$$\text{prox}_f(x_1, \ldots, x_n) = (\text{prox}_{f_i}(x_i))_{i=1}^n. \tag{7}$$

Include connection between proximal operator and the subdifferential (i.e. resolvent) here?

Now we will state and prove a very important theorem of the operator, for specific functions. Reflecting back to the definition of the operator, the size/cardinality of the subspace, in Eq. 3, indeed depends on the function f. We will now show that if f is a proper, closed and convex function, then the proximal set consist of a singleton (i.e. one point). We will state this theorem and subsequently prove it. But, before this we will state the existence and uniqueness of a minimizer to closed strongly convex functions.

Theorem: (*Existence and uniqueness of a minimizer to closed strongly convex functions*). Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a proper closed $\sigma-$strongly convex function, $\sigma > 0$. Then

(a) f has a unique minimizer

(b) $f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \geq \frac{\sigma}{2}||\boldsymbol{x} - \boldsymbol{x}^*||$ for all $\boldsymbol{x} \in \text{dom } f$, where $\boldsymbol{x}^*$ is the unique minimizer of f.

Proof. In this proof we refer to theorems and lemmas stated in [3].
Since f is a proper and strongly convex, the domain $\mathbf{dom}\, f$ is a convex and non-empty set. Thus, from Theorem 3.17 the relative domain of $\mathbf{dom}\, f$ is nonempty. Take $\boldsymbol{x}_0 \in \text{ri(dom)}$. Similarly, using Theorem 3.18 we have that the subdifferential of the point $\boldsymbol{x}_0$ is non-empty. Let $\boldsymbol{y} \in \partial f(\boldsymbol{x}_0)$, now we can use the equivalences in Theorem 5.24 (first-order characteristics of strong convexity) which states the equivalence of $\alpha-$strongly convex functions and property (ii), i.e. for any $\boldsymbol{x} \in \text{dom}, \boldsymbol{y} \in \text{dom}\, \partial f$ and $y \in \partial f(\boldsymbol{x})$ it holds that $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{y}^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{\sigma}{2}||\boldsymbol{y} - \boldsymbol{x}||^2$. Thus we have that

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0) + \boldsymbol{y}^T(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{\sigma}{2}||\boldsymbol{x} - \boldsymbol{x}_0||^2 \geq f(\boldsymbol{x}_0) + \boldsymbol{y}^T(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{C\sigma}{2}||\boldsymbol{x} - \boldsymbol{x}_0||_2^2$$

where we have used that all norms are equivalent in finite dimensional spaces, and thus there exists such a $C > 0$. Not finished yet. This proof was quite long.. □

Theorem: (*First Proximal Theorem*). Let $f : \mathbb{R}^n \to \mathbb{R}$ be a proper closed and convex function. Then the proximal set $\text{prox}_f(\boldsymbol{x})$ is a singleton-set for any $\boldsymbol{x} \in \mathbb{R}^n$.

Proof. Recall the definition of the proximal operator $\text{prox}_f(\boldsymbol{x}) = \underset{\boldsymbol{u}\in\mathbb{R}^n}{\text{argmin}} \tilde{f}(\boldsymbol{u}, \boldsymbol{x})$ where $\tilde{f}(\boldsymbol{u}, \boldsymbol{x}) = f(\boldsymbol{u}) + \frac{1}{2}||\boldsymbol{u} - \boldsymbol{x}||_2^2$. The norm is a convex function since for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $\lambda \in [0, 1]$ it holds that

$$||\lambda\boldsymbol{x} + (1 - \lambda)\boldsymbol{y}|| \leq \lambda||\boldsymbol{x}|| + (1 - \lambda)||\boldsymbol{y}||$$

using the absolute scalability and the triangle inequality of the norm. Thus, the function $\tilde{f}(\boldsymbol{u}, \boldsymbol{x})$ is convex. Further, the norm $||\cdot||_2$ is a proper function and we are given that f is proper, thus $\tilde{f}$ is proper. Last, the α-sublevel set (for any $\alpha \in \mathbb{R}$)

$$\mathcal{S}_\alpha = \{\boldsymbol{x} \in \mathbf{dom}\tilde{f}\,|\tilde{f}(\boldsymbol{x}) \leq \alpha\}$$

is closed. Thus using the Theorem of existence and uniqueness of a minimizer to closed strongly convex functions, we have that $\tilde{f}(\boldsymbol{u}, \boldsymbol{x}) = f(\boldsymbol{u}) + \frac{1}{2}||\boldsymbol{u} - \boldsymbol{x}||_2^2$ has a unique minimizer. Detta kanske är lite väl kort förklarat? □

Now we have established they nice property that the proximal operator has only one point for *nice* functions. Hereafter, I will assume that all presented functions f are such nice function, given that nothing else is stated. Next we will derive the proximal operator of two different scaled functions λf.

The proximal operator of $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}$

$$\mathbf{prox}_{\lambda f}(\boldsymbol{x}) = \underset{\boldsymbol{u}\in\mathbb{R}^n}{\text{argmin}}\{f(\boldsymbol{u}) + \frac{1}{2\lambda}||\boldsymbol{u} - \boldsymbol{x}||_2^2\} \quad \text{for any } \boldsymbol{x} \in \mathbb{R}^n$$

Let us first consider the argument inside the brackets more closely, i.e. $f(\boldsymbol{u}) + \frac{1}{2\lambda}||\boldsymbol{u} - \boldsymbol{x}||_2^2$, this expands to

$$\begin{aligned}\tilde{f}(\boldsymbol{u}, \boldsymbol{x}) &= f(\boldsymbol{u}) + \frac{1}{2\lambda}||\boldsymbol{u} - \boldsymbol{x}||_2^2 = \frac{1}{2}||\boldsymbol{u}||_2^2 + \frac{1}{2\lambda}||\boldsymbol{u} - \boldsymbol{x}||_2^2 \\ &= \frac{1}{2}\boldsymbol{u}^T\boldsymbol{u} + \frac{1}{2\lambda}(\boldsymbol{u} - \boldsymbol{x})^T(\boldsymbol{u} - \boldsymbol{x}) = \frac{1}{2}\boldsymbol{u}^T\boldsymbol{u} + \frac{1}{2\lambda}(\boldsymbol{u}^T\boldsymbol{u} + \boldsymbol{x}^T\boldsymbol{x} - 2\boldsymbol{u}^T\boldsymbol{x}).\end{aligned}$$

Since we want to find the minimizer of this quantity, and the function is differentiable we compute the derivative w.r.t. $\boldsymbol{u}$ (i.e. using first order conditions).

$$\frac{\partial f(\boldsymbol{u}, \boldsymbol{x})}{\partial \boldsymbol{u}} = \boldsymbol{u} + \frac{1}{2\lambda}(2\boldsymbol{u} - 2\boldsymbol{x}) = \boldsymbol{u} + \frac{1}{\lambda}(\boldsymbol{u} - \boldsymbol{x}) \quad \Rightarrow \boldsymbol{u} = \frac{\boldsymbol{x}}{1 + \lambda}$$

Thus the proximal operator of $\lambda f_1(x)$ is given by $\mathbf{prox}_{\lambda f}(\boldsymbol{x}) = \frac{\boldsymbol{x}}{1+\lambda}$

The proximal operator of $f(\boldsymbol{x}) = ||\boldsymbol{x}||_1$

For this function we can use that the function is separable, and using that the proximal operator is separable. Thus we can find $\mathbf{prox}_{\lambda f}(\boldsymbol{x})$ by considering each $\mathbf{prox}_{\lambda f}(x_i)$ separately, which is given by

$$\mathbf{prox}_{\lambda f}(x_i) = \underset{u_i \in \mathbb{R}}{\operatorname{argmin}} \ \lambda |u_i| + \frac{1}{2}(u_i - x_i)^2.$$

Now using the first order condition and the sub-differential of the absolute-value function, we have the three cases, $u_i > 0$, $u_i = 0$ and $u_i < 0$. These gives the differentials

$$\begin{aligned} u_i > 0: &\quad 0 = \lambda + u_i - x_i \iff u_i = x_i - \lambda \quad \text{if } x_i > \lambda \\ u_i < 0: &\quad 0 = -\lambda + u_i - x_i \iff u_i = x_i + \lambda \quad \text{if } x_i < \lambda \end{aligned}$$

for the last case $u_i = 0$, i.e. $|x_i| \leq \lambda$, the sub-differential is independent of u_i and will be zero for all elements in this interval thus $u_i = 0$ is a minimizer in this interval. All these cases can be combined into the expressed of $\mathbf{prox}_{\lambda f}(x_i)$

$$\mathbf{prox}_{\lambda f}(x_i) = \operatorname{sign}(x_i)(|x_i| - \lambda). \tag{8}$$

Next we will consider how these operators are useful in solving convex optimization problems.

Proximal algorithms

In this section we will link these proximal operators with fixed points algorithms, starting off with the central theorem:

Theorem: The point $\boldsymbol{x}^*$ minimizes f iff $\boldsymbol{x}^* = \mathbf{prox}_{\lambda f}(\boldsymbol{x}^*)$. The point $\boldsymbol{x}^*$ is called a *fixed point.*

Proof. ($\Rightarrow$): Assume $\boldsymbol{x}^*$ minimizes $f : \mathbb{R}^n \to \mathbb{R}$, then $f(\boldsymbol{x}) \geq f(\boldsymbol{x})^*$ for all $\boldsymbol{x} \in \mathbb{R}^n$. Thus

$$f(\boldsymbol{x}) + \frac{1}{2\lambda}||\boldsymbol{x} - \boldsymbol{x}^*||_2^2 \geq f(\boldsymbol{x}^*) + \frac{1}{2\lambda}||\boldsymbol{x}^* - \boldsymbol{x}||_2^2 \tag{9}$$

for any $\boldsymbol{x} \in \mathbb{R}^n$. Thus $\boldsymbol{x}^*$ minimizes (9) and hence $\boldsymbol{x}^* = \mathbf{prox}_{\lambda f}(\boldsymbol{x}^*)$.
($\Leftarrow$): Assume $\tilde{\boldsymbol{x}} = \mathbf{prox}_{\lambda f}(\boldsymbol{u})$ then $\mathbf{0} \in \partial f(\tilde{\boldsymbol{x}}) + (\tilde{\boldsymbol{x}} - \boldsymbol{u})$. Now simply taking $\tilde{\boldsymbol{x}} = \boldsymbol{v} = \boldsymbol{x}^*$ we get that $\mathbf{0} \in \partial f(\boldsymbol{x}^*)$. Since f is a convex function we have that $\boldsymbol{x}^*$ minimizes f. □

Now we have established that we can find the minimizer of f by finding a fixed point of its proximal operator. Thus if we apply $\mathbf{prox}_{\lambda f}$ repeatedly we would recover its fixed point.
Below not done—>

$$||\mathbf{prox}_f(\boldsymbol{x}) - \mathbf{prox}_f(\boldsymbol{y})||_2^2 \leq (\boldsymbol{x} - \boldsymbol{y})^T(\mathbf{prox}_f(\boldsymbol{x}) - \mathbf{prox}_f(\boldsymbol{y}))$$

$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

$$\boldsymbol{x}^{k+1} = (1 - \alpha)\boldsymbol{x}^k + \alpha N(\boldsymbol{x}^k)$$

$\boldsymbol{x}^{k+1} := T(\boldsymbol{x}^k)$, $||T(\boldsymbol{x}^k) - \boldsymbol{x}^k|| \to 0$ as $k \to \infty$ thus $\boldsymbol{x}^k$ converges to a fixed point of T.
<—Above not done

Let us consider some examples of how the updates can be performed using the proximal operator of the scaled function f, here considering the functions $f_1(x) = \frac{1}{2}x^2$ and $f_2(x) = |x|$, this done by using proximal algorithms. The *simplest* proximal method is given by

$$\boldsymbol{x}^{k+1} = \mathbf{prox}_{\lambda f}(\boldsymbol{x}^k), \tag{10}$$

in which the proximal operator is evaluated in the current point/solution $\boldsymbol{x}^k$ and finds the next solution $\boldsymbol{x}^{k+1}$ by solving $\mathbf{prox}_{\lambda f}(\boldsymbol{x}^k)$. Thus the proximal operator suggest a new point closer to the minimum. Some examples of points $\boldsymbol{x}^k$ and updated points $\boldsymbol{x}^{k+1}$ are shown in Figure 8. Here the origin of the red arrow indicates the position $\boldsymbol{x}^k$ and the end of the arrow indicates the position of new point $\boldsymbol{x}^{k+1}$. These plots where generated by $\lambda = 5$, and their corresponding proximal operators. These proximal operators are derived in the previous section.

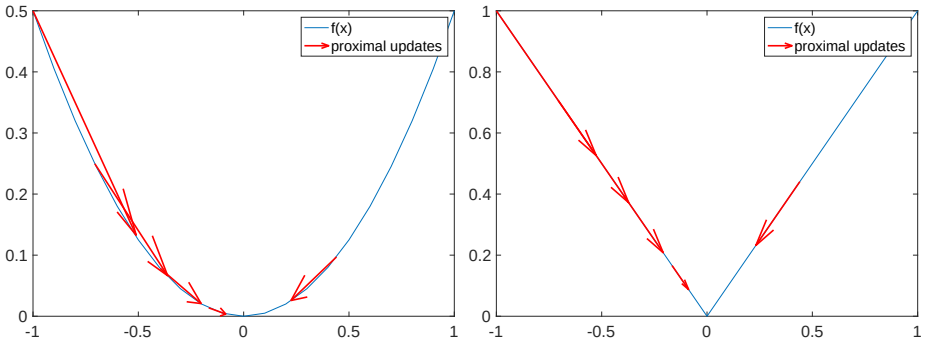

Figure 4: Functions $f : \mathbb{R} \to \mathbb{R}$ ($f_1(x) = \frac{1}{2}x^2$ right and $f_2(x) = |x|$ at left), red arrows indicates the updated solutions using the proximal algorithm in Eq. (10)

We could now also consider the aforementioned functions in a higher dimension, as they are fulfill the requirements for separable functions and (7). Thus, each update is performed in each variable separately and a visualization of the updates are shown in the figure below, note that the simplest proximal algorithm is considered.

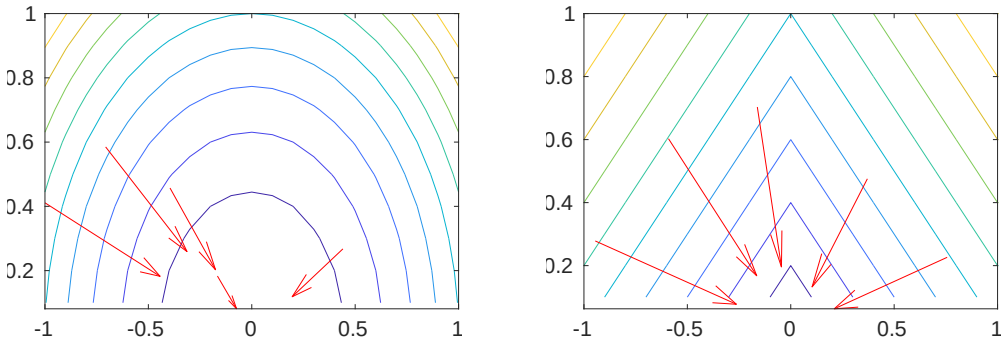

Figure 5: Functions $f : \mathbb{R}^2 \to \mathbb{R}$ (left $f_1(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2)$ and to the right $f_2(x_1, x_2) = |x_1| + |x_2|$), updates from the proximal algorithm is indicated by the arrows in red.

Again, expanding to $\mathbb{R}^3$ we get the similar updates shown in Figure 8 (left middle). In the right figure we consider the function $f(x_1, x_2, x_3) = f_1(x_1, x_2, x_3) + f_2(x_1, x_2, x_3)$ and the proximal update

$$\mathbf{prox}_{\lambda f_1 + \lambda f_2}(\boldsymbol{x}) = \mathbf{prox}_{\lambda f_1}(\boldsymbol{x}) + \mathbf{prox}_{\lambda f_2}(\boldsymbol{x})$$

.

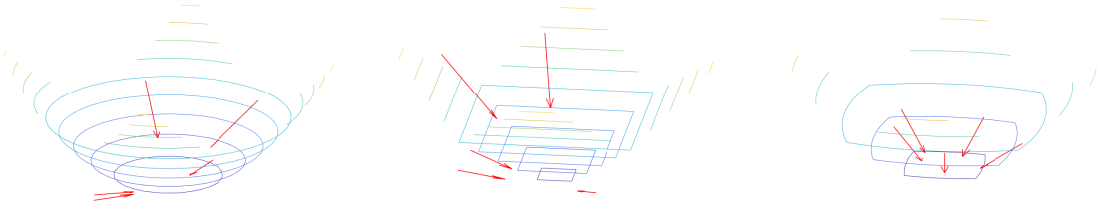

Figure 6: Functions $f : \mathbb{R}^2 \to \mathbb{R}$ (left $f_1(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2)$ and to the right $f_2(x_1, x_2) = |x_1| + |x_2|$), updates from the proximal algorithm is indicated by the arrows in red.

Hyperparameter choice

This part is not done yet! In this small section we will consider the effect of λ has on the proximal algorithm. Now we will consider a common structure of convex optimization problem.

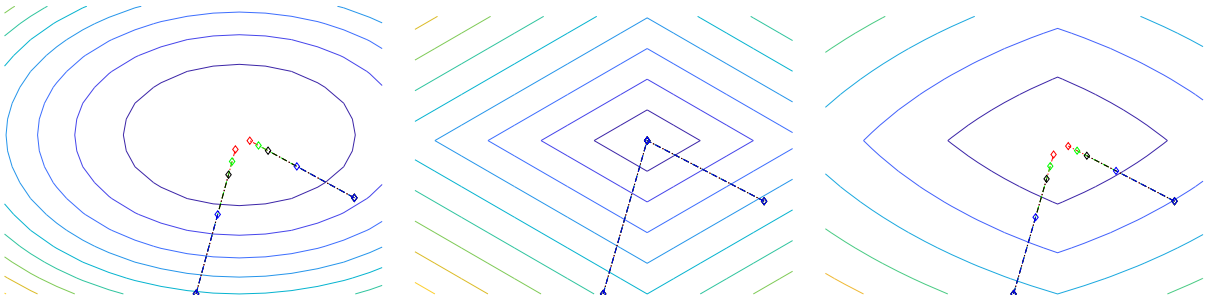

Figure 7: Functions $f : \mathbb{R}^2 \to \mathbb{R}$ (left $f_1(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2)$ and to the right $f_2(x_1, x_2) = |x_1| + |x_2|$), updates from the proximal algorithm is indicated by the arrows in red.

Often, the objective function $f(x)$ can be decomposed into a convex differentiable $g(x)$ and a convex non-differentaible function $h(x)$. This convex optimization problem can be solved using the proximal gradient algorithm given by

$$x^{k+1} = \mathbf{prox}_{t_k h}(x^k - t_k \nabla g(x^k)))$$

where t_k is the step size and $\nabla g(x^k)$ is the gradient of the differntiable function g.

MATLAB examples of some general problems

Let us consider a similar example as shown in equation (XX), where we have a mix of a convex differentiable function given by the norm and the convex and non-differentialble ℓ_1-norm. Thus we aim to find the optimal x^* to the objective

$$\min_x \frac{1}{2}||Ax - b||_2^2 + ||x||_1$$

where the matrix $A \in \mathbb{R}^{200\times 100}$ is a randomly generated matrix having $A_{ij} \sim \mathcal{U}(0, 1)$ for experiment 1 and $A_{ij} \sim \mathcal{N}(0, 1)$ in experiment 2. The proximal operator and proximal algorithm is given by

$$\begin{aligned}
x^{k+1} &= \text{prox}_{t_k h}(x^k - t_k \nabla g(x^k)) \\
x^{k+1} &= \text{prox}_{t_k h}(x^k - t_k A^T(Ax^k - b)) \\
x^{k+1} &= \big(\max(x_i^k - t_k A^T(Ax_i^k - b), t_k)\text{sign}(x_i^k - t_k A^T(Ax_i^k - b)\big)_i^n
\end{aligned}$$

from equation XX and YY. The proximal algorithm is implemented in MATLAB, and run through the code shown below. The convergence of the algorithm for the two different matrix generations are shown in the figures below.

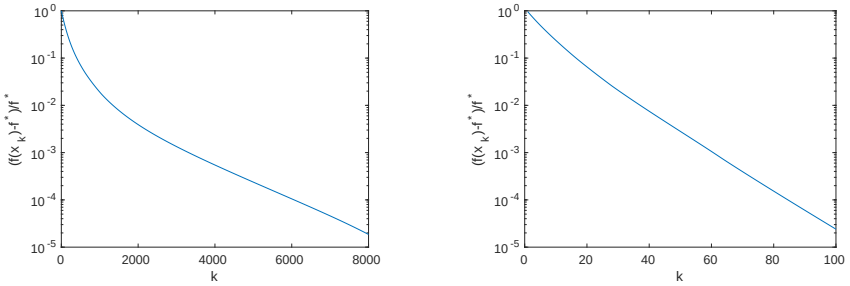

Figure 8: Convergence of the proximal algorithms for $A_{ij} \sim \mathcal{U}(0,1)$ in the left plot and $A_{ij} \sim \mathcal{N}(0,1)$ in the right.

Von Neumann's inequality

In this part we will prove the inequality

$$\operatorname{tr}(AB) \leq \sum_{i=1} \sigma_i^A \sigma_i^B$$

where σ_i^A, σ_i^B are the sorted (in decreasing order) singular values of matrix A and B respectively.

Proof. In this proof I'll consider square matrices A, B, however for non-square matrices the proof is very similar. First let us formulate the matrices singular value decompositions; $A = U_A S_A V_A$ and $B = U_B S_B V_B$, where U_A, V_A, U_B, V_B are unitary matrices and S_A, S_B are diagonal matrices with the matrices singular values in sorted descending order on the diagonal. Further, we will use that the trace operator is invariant to cyclic permutations, i.e. $\operatorname{tr}(ABCD) = \operatorname{tr}(BCDA)$. Thus we have that

$$\operatorname{tr}(AB) = \operatorname{tr}(U_A S_A V_A U_B S_B V_B) = \operatorname{tr}(S_A V_A U_B S_B V_B U_A)$$

and we now define the unitary matrices $V = V_B U_A = (v_{rs})$ and $U = (U_A V_B)^T = (u_{rs})$. These matrices are indeed unitary since products of unitary matrices are unitary. I.e. if U, V are unitary matrices ($UU^* = VV^* = I$) then $UV = VU$ are unitary matrices,

$$(UV)(UV)^* = UVV^*U^* = UI_{n\times n}U^* = I$$

thus V, U are unitary. The trace can now be written as:

$$\operatorname{tr}(AB) = \sum_{i=1}^{n}\sum_{j=1}^{n} u_{ij}v_{ij}\sigma_i^A\sigma_j^B \leq \sum_{i=1}^{n}\sum_{j=1}^{n} |u_{ij}v_{ij}|\sigma_i^A\sigma_j^B.$$

Next, using Young's inequality $|xy| \leq \frac{1}{2}|x|^2 + \frac{1}{2}|y|^2$ we have that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} |u_{ij}v_{ij}|\sigma_i^A\sigma_j^B \leq \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{2}|u_{ij}|^2\sigma_i^A\sigma_j^B + \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{2}|v_{ij}|^2\sigma_i^A\sigma_j^B.$$

Now by the Lemma in [1] for doubly-stochastic matrices, we have that

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{2}|u_{ij}|^2\sigma_i^A\sigma_j^B + \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{2}|v_{ij}|^2\sigma_i^A\sigma_j^B \\ &\leq \frac{1}{2}\Big(\sum_{i=1}^{n} \frac{1}{2}\sigma_i^A\sigma_i^B + \sum_{i=1}^{n} \frac{1}{2}\sigma_i^A\sigma_i^B\Big) = \sum_{i=1}^{n} \sigma_i^A\sigma_i^B.\end{aligned}$$

□

References

[1] L. Mirsky, "A trace inequality of john von neumann," vol. 79, pp. 303–306, Springer-Verlag, 1975.

[2] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, p. 127–239, Jan. 2014.

[3] A. Beck, *First-Order Methods in Optimization*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2017.