Project in mathematics
Proximal operators and Von Neumann inequality

Linn Öström, linn.ostrom@math.lth.se

July 7, 2020

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

Proximal operators

Short intro-text about proximal operators and why they are useful

Introduction

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 directly states as in them. All spaces considered in this summary are Hilbert spaces of nothing else is stated and when referring to a general norm, what is mean 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 and the epigraph $\mathbf{epi}\, f$, and the domain $\mathbf{dom}\, f$ of a function.

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

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

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

Definition: A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is said to be *convex* if $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $\lambda \in [0, 1]$ is holds 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. Since proximal operators only are helpful for convex functions, we will only consider these when deriving the proximal operators. A very similar definition is for convex sets.

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

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

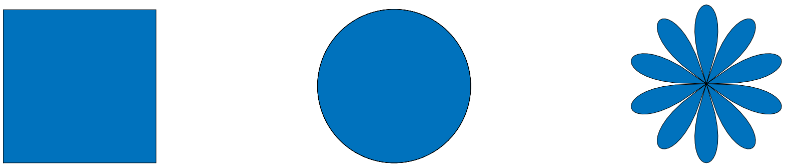

Figure 1: Example of convex and non-convex sets.

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

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

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

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

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

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

Definition: A convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is said to be *proper* its effective domain is nonempty and $\{-\infty\} \notin \textbf{dom} f$.

The epigraph set (in Eq. (2)) is a nonempty closed convex set if f is a closed proper convex function. Some examples of convex functions are shown in the left and middle figures 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.

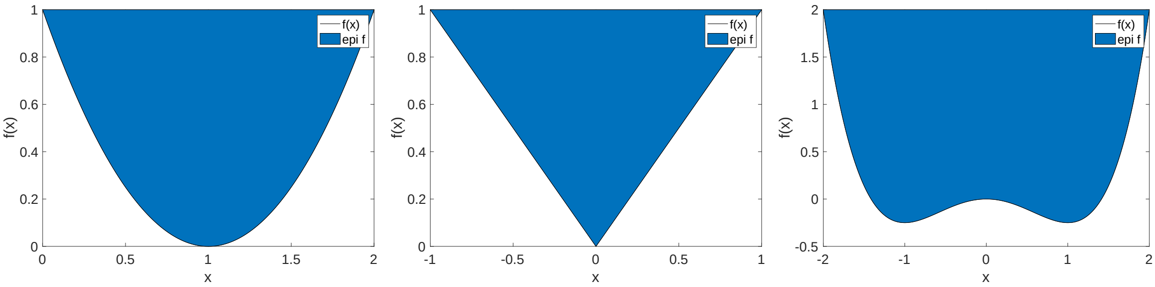

Figure 2: Example of convex and non-convex functions, at the left $f_1(x) = (x-1)^2$, in the middle $f_2(x) = |x|$ and to the right $f_3(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2$. Convex functions are shown in left and middle, while the right function is not.

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 shown in filled 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 non-differential function f

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

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

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 $\boldsymbol{x}$.

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

The proximal operator

Next we will define the proximal operator. Expand with into-text/recap intro but not in more detail

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

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

thus the proximal operator maps a vector $\boldsymbol{x}\in\mathbb{R}^n$ and maps it onto subset of this 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 constrcuction of the operator one can thus instead use simpler and less computational effort to solve the optimization-problem by solving smaller sub-problems. Further one can also consider the proximal operator of the scaled function λf with a parameter $\lambda > 0$, and thus one simply ends up with the similar expression

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

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 the current point and a small λ penalizes large movements from $\boldsymbol{x}$. 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

$$\boldsymbol{x}^{k+1} = \mathbf{prox}_{\lambda f}(\boldsymbol{x}^k), \tag{5}$$

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 5. 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 subsequent section *Examples: Deriving the proximal operator for a set of functions*. Next we will consider which for the proximal operator takes when we consider separable functions i,e. $f : \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ where

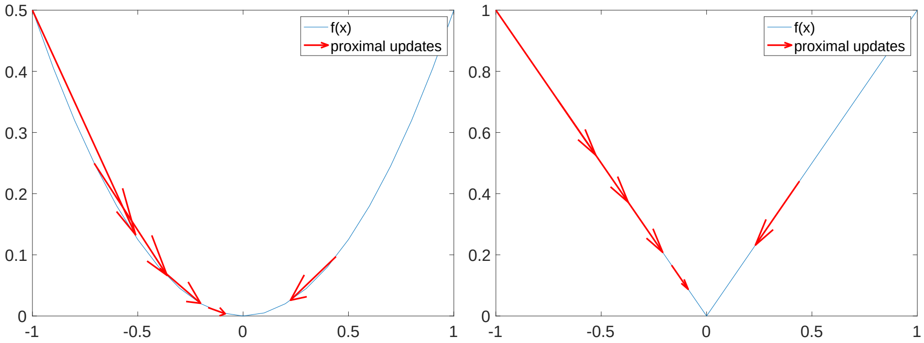

Figure 3: Functions $f : \mathbb{R}\to\mathbb{R}$ ($f_1(x) = \frac{1}{2}x^2$ right and $f_2(x) = |x|$ at left), red arrows indicates the updated solutions using the proximal algorithm in Eq. (5)

$$f(x_1,\ldots,x_n) = \sum_{i=1}^{m} f_i(x_i)$$

for any $\boldsymbol{x} \in \mathbb{R}^n$. Then the proximal operator takes the form

$$\begin{aligned}
\operatorname{prox}_f(x_1, \dots, x_n) &= \operatorname{argmin}_{u_1,\dots,u_n} \Big[\sum_{i=1}^{m} f_i(x_i) + \frac{1}{2}||u_i - x_i||_2^2 \Big] && (6)\\
&= \prod_{i=1}^{n} \operatorname{argmin}_{u_1} \Big[\sum_{i=1}^{m} f_i(x_i) + \frac{1}{2}||u_i - x_i||_2^2 \Big] && (7)\\
&= \prod_{i=1}^{n} \operatorname{prox}_{f_i}(x_i) && (8)
\end{aligned}$$

since the squared norm is also separable. Further, if f, f_i $i = 1, \dots, n$ are proper closed convex and separable, and each f_i also univariate then (8) becomes

$$\operatorname{prox}_f(x_1, \dots, x_n) = (\operatorname{prox}_{f_i}(x_i))_{i=1}^{n}. \qquad (9)$$

Thus we could now consider the aforementioned functions in a higher dimension, as they are fulfill the requirements for (9). 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.

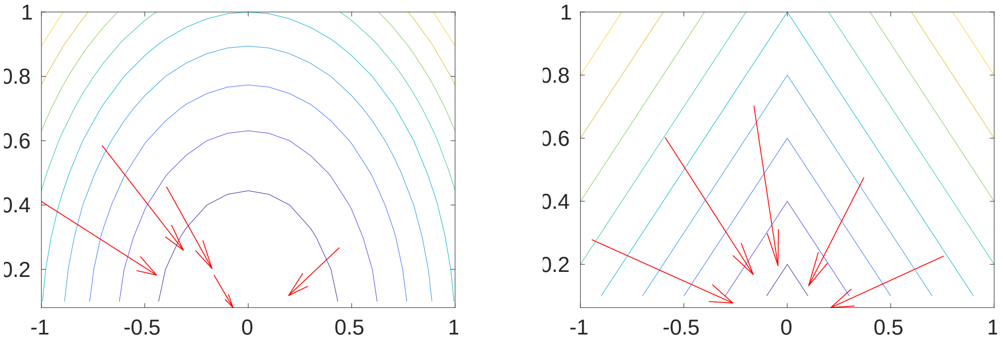

Figure 4: Functions $f : \mathbb{R}^2 \to \mathbb{R}$ (left $f_1(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2)$ and to the right $f_2(x_1, x_2) = |x_1| + |x_2|$), updates from the proximal algorithm is indicated by the arrows in red.

This size 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 set consist of a singleton (i.e. one point). We will state this as a Theorem and subsequently prove it. 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 $\boldsymbol{x} \in \operatorname{dom} f$, where $\boldsymbol{x^*}$ is the unique minimizer of f.

OBS: Pelle, tycker du att det vore bra om jag gör beviset för denna också?

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

Proof. Recall the definition of the proximal operator $\operatorname{prox}_f(\boldsymbol{x}) = \underset{\boldsymbol{u} \in \mathbb{R}^n}{\operatorname{argmin}} \tilde{f}(\boldsymbol{u}, \boldsymbol{x})$ 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

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

where we have used the triangle inequality. 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 $\tilde{f}$ 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 f is such nice function, given that nothing else is stated.

Examples: Deriving the proximal operator for a set of functions

In these example we will consider the proximal operator of the 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. $\boldsymbol{u}$.

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

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 a little trick, using that the proximal is separable for both variables $\boldsymbol{x}$, and $\boldsymbol{y}$. Thus we can find $\mathbf{prox}_{\lambda f}(\boldsymbol{x})$ by considering each $\mathbf{prox}_{\lambda f}(x_i)$ separately, which is given by

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

Now we can use the first order condition and the sub-differential of the absolute-value function. We thus 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 $\mathbf{prox}_{\lambda f}(x_i)$

$$\mathbf{prox}_{\lambda f}(x_i) = \text{sign}(x_i)(|x_i| - \lambda) \tag{10}$$

Examples: MATLAB implementations of some examples

Again considering the aforementioned functions, f_1 and f_2 we will now consider them in higher dimension and ...

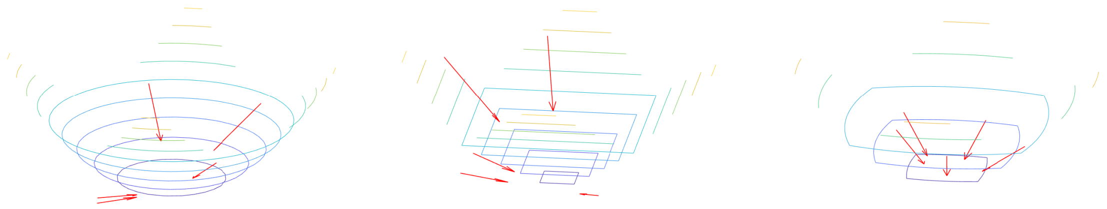

Figure 5: Functions $f : \mathbb{R}^2 \to \mathbb{R}$ (left $f_1(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2)$ and to the right $f_2(x_1, x_2) = |x_1| + |x_2|$), updates from the proximal algorithm is indicated by the arrows in red.

Von Neuman norm

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

$A = U_A S_A V_A$ and $B = U_B S_B V_B$,
$\operatorname{tr}(ABCD) = \operatorname{tr}(BCDA)$

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

Here introducing paper things...

we define $V = V_B U_A = (v_{rs})$ and $U = (U_A V_B)^T = (u_{rs})$

If U, V are unitary matrices $(UU^* = VV^*) = I)$ then $UV = VU$ are unitary matrices,

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

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

Now 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 Lemma .... we have that

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

Appendix

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.