Appl. Comput. Math., V.15, N.1, 2016, pp.3-43

# A PRIMER ON MONOTONE OPERATOR METHODS
***SURVEY***

ERNEST K. RYU[1], STEPHEN BOYD[2]

Abstract. This tutorial paper presents the basic notation and results of monotone operators and operator splitting methods, with a focus on convex optimization. A very wide variety of algorithms, ranging from classical to recently developed, can be derived in a uniform way. The approach is to pose the original problem to be solved as one of finding a zero of an appropriate monotone operator; this problem in turn is then posed as one of finding a fixed point of a related operator, which is done using the fixed point iteration. A few basic convergence results then tell us conditions under which the method converges, and, in some cases, how fast. This approach can be traced back to the 1960s and 1970s, and is still an active area of research. This primer is a self-contained gentle introduction to the topic.

Keywords: Monotone Operators, Convex Optimization, Splitting Methods, Fixed Point Iteration.

AMS Subject Classification: 47H05, 47H09, 47H10, 47N10, 65K05, 90C25.

## 1. Introduction

In the field of convex optimization, there are a myriad of seemingly disparate algorithms each with its specific setting and convergence properties. It is possible to understand, derive, and analyze many of these methods in a unified manner, using the abstraction of monotone operators and a single approach. First, the problem at hand is expressed as finding a zero of a monotone operator. This problem is in turn transformed into finding a fixed point of a related function. The fixed point is then found by the fixed point iteration, yielding an algorithm for the original problem. This single approach yields many different algorithms, with different convergence conditions, depending on how the first and second steps are done (*i.e.*, the selection of the monotone operator and fixed point function). It recovers many classical and modern algorithms along with conditions under which they converge.

The idea of this basic approach is not new, and several surveys based on it have already been written, e.g., by [7, 33, 34, 36, 44]. Several surveys that rigorously develop the theory behind monotone operators also have been written, e.g., by [3, 7, 17, 103].

In fact, the ideas can be further traced back to the 1960s and 1970s. In the 1960s, the notion of monotone operators was first formulated and studied [71,87,88]. Much of the initial work was done in the context of functional analysis and partial differential equations [24–26], but it was soon noticed that the theory is relevant to convex functions and convex optimization [71,90,110]. In the 1970s, iterative algorithms constructed from monotone operators and fixed point functions were itroduced [81,85,86,117]. Since then, this field has grown considerably and is still an active area of research.

This paper is not meant to be a survey of the huge literature in these and related areas. Rather, this paper is a gentle and self-contained introduction and tutorial, for the reader whose main interest is in understanding convex optimization algorithms, or even developing her own. We do not focus on the mathematical details. When a result is simple to show, we do so;

[1]Institute for Computational and Mathematical Engineering, Stanford University, e-mail: eryu@stanford.edu
[2]Institute for Computational and Mathematical Engineering, Stanford University, e-mail: boyd@stanford.edu
*Manuscript received 29 September 2015.*

but in other cases we do not. We will often merely state certain regularity conditions without fully explaining how they exclude pathologies; the reader can consult the references listed for the mathematical details. These same references can also be consulted for details of individual contributions to the field; in the sequel, we cite only a few earliest contributions.

We first review the basic tools that lay the foundation for later sections. In §2, we introduce relations and functions, and in §3, we introduce the ideas of Lipschitz constant, and nonexpansive and contractive operators. In §4, we introduce the concept of a monotone operator, which generalizes the idea of a monotone increasing function in several ways, and give some important examples, such as the subdifferential mapping. We then introduce the fixed point iteration and discuss its convergence properties in §5. In fact, all algorithms presented in this paper are instances of the fixed point iteration.

After this background, we present several approaches to transform the problem of finding a zero of a monotone operator into a fixed point equation. In §6, we introduce the resolvent and Cayley relations and describe the proximal point method. In §7, we describe operator splitting methods, and using them we derive a wide variety of algorithms, including the gradient method, the method of multipliers, and the alternating diections method of multipliers, using the basic approach.

We assume that the reader has had some exposure to convex optimization and convex analysis, such as the subdifferential of a function, and optimality conditions. Standard references on convex optimization and convex analysis (which contain far more than the reader needs) include [4, 13–18, 20, 94, 111, 120]. Other than this basic background, this tutorial is more or less self-contained.

## 2. Relations

We define the notation of relations and functions here. A relation, point-to-set mapping, set-valued mapping, multi-valued function, correspondence, or operator $R$ on $\mathbf{R}^n$ is a subset of $\mathbf{R}^n \times \mathbf{R}^n$. We will overload function and matrix notation and write $R(x)$ and $Rx$ to mean the set $\{y \mid (x, y) \in R\}$.

If $R(x)$ is a singleton or empty for any $x$, then $R$ is a function or single-valued with a domain $\{x \mid R(x) \neq \emptyset\}$. In this case, we may mix functions and relations and write (with some abuse of notation) $R(x) = y$ (function notation) although $R(x) = \{y\}$ (relation or multi-valued function notation) would be strictly correct.

Also, we extend the familiar set-image notation for functions to relations: for $S \subseteq \mathbf{R}^n$, we define $R(S) = \cup_{s\in S} R(s)$.

**Simple examples.** The empty relation is $R = \emptyset$; the full relation is $R = \mathbf{R}^n \times \mathbf{R}^n$. More useful examples include the zero relation (function) $0 = \{(x, 0) \mid x \in \mathbf{R}^n\}$ and the identity relation (function) $I = \{(x, x) \mid x \in \mathbf{R}^n\}$.

**Subdifferential.** A more interesting relation is the subdifferential relation $\partial f$ of a function $f : \mathbf{R} \to \mathbf{R} \cup \{\infty\}$, defined by

$$\partial f = \{(x, g) \mid x \in C,\ \forall z \in \mathbf{R}^n\ f(z) \geq f(x) + g^T(z - x)\}.$$

The set $\partial f(x)$ is the subdifferential of $f$ at $x$. Any $g \in \partial f(x)$ is called a subgradient of $f$ at $x$. The subdifferential $\partial f(x)$ is always a well-defined closed convex set for any function $f$ at any point $x \in \mathbf{R}^n$, but it can be empty. When $f$ is convex, $\partial f(x) \neq \emptyset$ for any $x \in \mathbf{relint}\, C$, where **relint** denotes the relative interior.

**Operations on relations.** We extend many notions for functions to relations. For example, the domain of a relation $R$ is defined as

$$\mathbf{dom}\, R = \{x \mid R(x) \neq \emptyset\}.$$

If $R$ and $S$ are relations, we define the composition $RS$ as

$$RS = \{(x,z) \mid \exists y \ (x,y) \in S,\ (y,z) \in R\},$$

and their sum as

$$R + S = \{(x, y+z) \mid (x,y) \in R,\ (x,z) \in S\}.$$

We overload addition and scalar multiplication and inequality operations to handle sets (as well as mixtures of sets and points) in the standard way.

**Inverse relation.** The inverse relation of $R$ is defined as

$$R^{-1} = \{(x,y) \mid (y,x) \in R\}.$$

This always exists, even when $R$ is a function that is not one-to-one. As a note of caution, the inverse relation is not quite an inverse in the usual sense, as we can have $R^{-1}R \neq I$. The zero relation is such an example.

However, we do have $R^{-1}Rx = x$ when $R^{-1}$ is a function and $x \in \mathbf{dom}\, R$. To see this, first note that $R^{-1}y = \{x\}$ if and only if $y \in R\tilde{x}$ holds only for $\tilde{x} = x$. Clearly if $x \in \mathbf{dom}\, R$, then $x \in R^{-1}Rx$. Now assuming $R^{-1}$ is a function, we get

$$\begin{aligned} R^{-1}Rx \ni \tilde{x} \quad &\Leftrightarrow \quad R^{-1}y = \tilde{x},\ y \in Rx \text{ for some } y \\ &\Leftrightarrow \quad y \in R\tilde{x},\ y \in Rx \text{ for some } y \\ &\Rightarrow \quad \tilde{x} = x. \end{aligned}$$

**Zeros of a relation.** When $R(x) \ni 0$, we say that $x$ is a zero of $R$. The zero set of a relation $R$ is $\{x \mid (x,0) \in R\} = R^{-1}(\{0\})$, which we also write (slightly confusingly) as $R^{-1}(0)$. We will see that many interesting problems can be posed as finding zeros of a relation.

**Inverse of subdifferential.** As another example, consider $(\partial f)^{-1}$, the inverse of the subdifferential. We have

$$\begin{aligned} (u,v) \in (\partial f)^{-1} \quad &\Leftrightarrow \quad (v,u) \in \partial f \\ &\Leftrightarrow \quad u \in \partial f(v) \\ &\Leftrightarrow \quad 0 \in \partial f(v) - u \\ &\Leftrightarrow \quad v \in \underset{x}{\operatorname{argmin}}(f(x) - u^T x). \end{aligned}$$

So we can write $(\partial f)^{-1}(u) = \operatorname{argmin}_x (f(x) - u^T x)$. (The righthand side is the set of minimizers.)

We can relate this to $f^*$, the conjugate function of $f$, defined as

$$f^*(y) = \sup_x (y^T x - f(x)).$$

The last line in the equivalences given above for $(u,v) \in (\partial f)^{-1}$ says that $v$ maximizes $(u^T x - f(x))$ over $x$, *i.e.*, $f^*(u) = u^T v - f(v)$. Thus we have

$$(u,v) \in (\partial f)^{-1} \quad \Leftrightarrow \quad f(v) + f^*(u) = v^T u.$$

A function $f : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ is said to be closed if its epigraph

$$\mathbf{epi}\, f = \{(x,t) \in \mathbf{R}^{n+1} \mid x \in \mathbf{dom}\, f,\ f(x) \leq t\}$$

is closed; it is called proper if its domain is nonempty. When $f$ is convex closed proper (CCP), $f^*$ is CCP and $f^{**} = f$; *i.e.*, the conjugate is CCP and the conjugate of the conjugate function is the original function [111, Theorem 12.2]. So when $f$ is CCP we have

$$v^T u = f(v) + f^*(u) = f^{**}(v) + f^*(u) \quad \Leftrightarrow \quad (v,u) \in (\partial f^*)^{-1} \quad \Leftrightarrow \quad (u,v) \in \partial f^*,$$

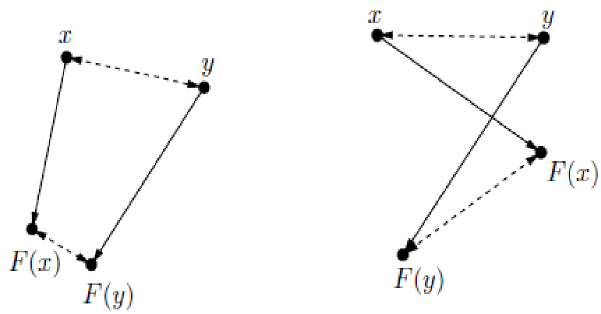
and we can write the simple formula $(\partial f)^{-1} = \partial f^*$.

## 3. Nonexpansive mappings and contractions

A relation $F$ on $\mathbf{R}^n$ has Lipschitz constant $L$ if for all $u \in F(x)$ and $v \in F(y)$ we have

$$\|u - v\|_2 \leq L\|x - y\|_2.$$

This implies that $F$ is a function, since if $x = y$ we must have $u = v$. When $L < 1$, $F$ is called a contraction; when $L = 1$, $F$ is called nonexpansive. Mapping a pair of points by a contraction reduces the distance between them; mapping them by a nonexpansive operator does not increase the distance between them. See Fig. 1.




(A) $F$ is contractive. (B) $F$ is nonexpansive.

Figure 1. Examples of a nonexpansive and a contractive mapping. The dotted lines denote the distances between $x$ and $y$ and $F(x)$ and $F(y)$.

**Basic properties.** If $F$ has Lipschitz constant $L$ and $\tilde{F}$ has Lipschitz constant $\tilde{L}$, then composition $F\tilde{F}$ has Lipschitz constant $L\tilde{L}$. Thus, the composition of nonexpansive operators is nonexpansive; the composition of a contraction and a nonexpansive operator is a contraction.

If $F$ has Lipschitz constant $L$ and $\tilde{F}$ has Lipschitz constant $\tilde{L}$, and $\alpha,\ \tilde{\alpha} \in \mathbf{R}$, then $\alpha F + \tilde{\alpha}\tilde{F}$ has Lipschitz constant $|\alpha|L + |\tilde{\alpha}|\tilde{L}$. Thus a weighted average of nonexpansive operators $F$ and $\tilde{F}$, *i.e.*, $\theta F + (1-\theta)\tilde{F}$ with $\theta \in [0,1]$, is also nonexpansive. If in addition one of them is a contraction, and $\theta \in (0,1)$, the weighted average is a contraction.

**Fixed points.** We say $x$ is a fixed point of $F$ if $x = F(x)$. If $F$ is nonexpansive and $\mathbf{dom}\, F = \mathbf{R}^n$, then its set of fixed points

$$\{x \in \mathbf{dom}\, F \mid x = F(x)\} = (I - F)^{-1}(0),$$

is closed and convex. Certainly, the fixed point set $X = \{x \mid F(x) = x\}$ can be empty (for example, $F(x) = x + 1$ on $\mathbf{R}$) or contain many points (for example, $F(x) = x$). If $F$ is a contraction and $\mathbf{dom}\, F = \mathbf{R}^n$, its has exactly one fixed point.

Let us show this. Suppose $F : \mathbf{R}^n \to \mathbf{R}^n$ is nonexpansive. That $X$ is closed follows from the fact that $F - I$ is a continuous function. Now suppose that $x,\ y \in X$, *i.e.*, $F(x) = x$, $F(y) = y$, and $\theta \in [0,1]$. We'll show that $z = \theta x + (1-\theta)y \in X$. Since $F$ is nonexpansive we have

$$\|Fz - x\|_2 \leq \|z - x\|_2 = (1-\theta)\|y - x\|_2,$$

and similarly, we have

$$\|Fz - y\|_2 \leq \theta\|y - x\|_2.$$

So the triangle inequality

$$\|x - y\|_2 \leq \|Fz - x\|_2 + \|Fz - y\|_2$$

holds with equality, which means the inequalities above hold with equality and $Fz$ is on the line segment between $x$ and $y$. From $\|Fz-y\|_2 = \theta\|y-x\|_2$, we conclude that $Fz = \theta x+(1-\theta)y = z$. Thus $z \in X$.

Next suppose $F : \mathbf{R}^n \to \mathbf{R}^n$ is a contraction with contraction factor $L$. Let $x$ and $\tilde{x}$ be fixed points. Then

$$\|x - \tilde{x}\|_2 = \|Fx - F\tilde{x}\|_2 \leq L\|x - \tilde{x}\|_2,$$

a contradiction unless $x = \tilde{x}$. We show the existence of a fixed point later.

**Averaged operators.** We say an operator $F$ is averaged if $F = (1-\theta)I+\theta G$ for some $\theta \in (0, 1)$, where the implicitly defined $G$ is nonexpansive. In other words, taking a weighted average of $I$ and a nonexpansive operator $G$ gives an averaged operator $F$. Clearly, $F$ is nonexpansive and has the same fixed points as $G$.

When operators $F$ and $\tilde{F}$ are averaged operators, so is $F\tilde{F}$. Interested readers can find a proof in [33, 37].

## 3.1. Examples

**Affine functions.** An affine function $F(x) = Ax + b$ has (smallest) Lipschitz constant $L = \|A\|_2$, the spectral norm or maximum singular value of $A$.

**Differentiable functions.** A differentiable function $F : \mathbf{R}^n \to \mathbf{R}^n$ is Lipschitz with parameter $L$ if and only if $\|DF(x)\|_2 \leq L$ for all $x$.

To see this, first assume $\|Df(x)\|_2 \leq L$ and define

$$g(t) = (Fx - Fy)^T F(tx + (1 - t)y).$$

By the mean value theorem and Cauchy-Schwartz inequality we have

$$\begin{aligned}
\|Fx - Fy\|_2^2 &= g(1) - g(0) = g'(\xi) = \\
&= (Fx - Fy)^T DF(\xi x + (1 - \xi)y)(x - y) \leq \\
&\leq \|Fx - Fy\|_2 \|DF(\xi x + (1 - \xi)y)(x - y)\|_2 \leq \\
&\leq \|Fx - Fy\|_2 \|DF(\xi x + (1 - \xi)y)\|_2 \|x - y\|_2 \leq \\
&\leq L\|Fx - Fy\|_2 \|x - y\|_2
\end{aligned}$$

for some $\xi \in [0, 1]$, and we conclude $F$ has Lipschitz parameter $L$.

On the other hand, if we assume $F$ has Lipschitz parameter $L$. Then

$$\|DF(x)v\|_2 = \lim_{h\to 0} \frac{1}{h} \|F(x + hv) - F(x)\|_2 \leq L\|v\|_2$$

and we conclude that $\|DF(x)\|_2 \leq L$ for all $x$.

**Projections.** The projection of $x$ onto a nonempty closed convex set $C$ is defined as

$$\Pi_C(x) = \operatorname*{argmin}_{z\in C} \|z - x\|_2,$$

which always exists and is unique. We can interpret $\Pi_C$ as the point in $C$ closest to $x$. Let us show that $\Pi_C$ is nonexpansive. (We will later derive the same result using the idea of a resolvent.)

Reorganizing the optimality condition for the projection [20, p. 139], we get that for any $u \in \mathbf{R}^n$ and $v \in C$ we have

$$(v - \Pi_C u)^T (\Pi_C u - u) \geq 0. \tag{1}$$

Now for any $x$, $y \in \mathbf{R}^n$, we get

$$(\Pi_C y - \Pi_C x)^T (\Pi_C x - x) \geq 0$$
$$(\Pi_C x - \Pi_C y)^T (\Pi_C y - y) \geq 0$$

using (1), and by adding these two we get

$$(\Pi_C x - \Pi_C y)^T (x - y) \geq \|\Pi_C x - \Pi_C y\|_2^2. \tag{2}$$

Finally, we apply the Cauchy-Schwartz inequality to conclude

$$\|\Pi_C x - \Pi_C y\|_2 \leq \|x - y\|_2.$$

Overprojection with respect to a nonempty closed convex $C$ is defined as

$$Q_C = 2\Pi_C - I.$$

Let us show that $Q_C$ is also nonexpansive:

$$\begin{aligned}\|Q_C x - Q_C y\|_2^2 &= \|2(\Pi_C x - \Pi_C y) - (x - y)\|_2^2 = \\ &= 4\|\Pi_C x - \Pi_C y\|_2^2 - 4(\Pi_C x - \Pi_C y)^T (x - y) + \|x - y\|_2^2 \leq \\ &\leq \|x - y\|_2^2,\end{aligned}$$

where the last line follows from (2). This implies that $\Pi_C = 1/2I + 1/2Q_C$ is an averaged operator. See Fig. 2. (We will later derive this result as well, using the idea of the Cayley operator.)

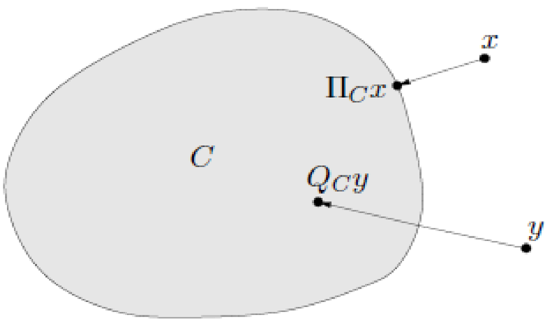



Figure 2. Illustration of projection and overprojection of points $x$ and $y$ onto $C$, respectively.

## 4. Monotone operators

A relation $F$ on $\mathbf{R}^n$ is called monotone if it satisfies

$$(u - v)^T (x - y) \geq 0$$

for all $(x, u),\ (y, v) \in F$. In circuit theory, such a relation is called incrementally passive [87,132]. In multi-valued function notation, monotonicity can be expressed as

$$(Fx - Fy)^T (x - y) \geq 0$$

for all $x,\ y \in \mathbf{dom}\, F$. (The left-hand side is a subset of $\mathbf{R}$, so the inequality means that this subset lies in $\mathbf{R}_+$.)

**Maximality.** The relation $F$ is maximal monotone if there is no monotone operator that properly contains it (as a relation, *i.e.*, subset of $\mathbf{R}^n \times \mathbf{R}^n$). In other words, if the monotone operator $F$ is not maximal, then there is $(x, u) \notin F$ such that $F \cup \{(x, u)\}$ is still monotone.

Maximality looks like a technical detail, but it turns out to be quite critical for the things we will do. In careful treatments of the topics of this paper, much effort goes into showing that various relations of interest are not just monotone but also maximal, under appropriate conditions.

**Strong monotonicity.** A relation $F$ is said to be strongly monotone or coercive with parameter $m > 0$ if

$$(Fx - Fy)^T(x - y) \geq m\|x - y\|_2^2$$

for all $x, y \in \mathbf{dom}\, F$.

When $F$ is strongly monotone with parameter $m$ and also Lipschitz with constant $L$, we have the lower and upper bounds

$$m\|x - y\|_2^2 \leq (Fx - Fy)^T(x - y) \leq L\|x - y\|_2^2.$$

(The right hand inequality follows immediately from the Cauchy-Schwarz inequality.) We will refer to $\kappa = L/m \geq 1$ as a condition number of $F$.

4.1. **Basic properties**

**Sum and scalar multiple.** If $F$ and $G$ are monotone, so is $F + G$. If $F$ and $G$ are maximal monotone and if $\mathbf{dom}\, F \cap \mathbf{int\, dom}\, G \neq \emptyset$, then $F + G$ is maximal monotone [114]. If in addition $F$ is strongly monotone with parameter $m$ and $G$ with parameter $\tilde{m}$ (which we can take to be zero if $G$ is merely monotone), then $F + G$ is strongly monotone with parameter $m + \tilde{m}$. For $\alpha > 0$, $\alpha F$ is strongly monotone with parameter $\alpha m$.

**Inverse.** If $F$ is (maximal) monotone, then $F^{-1}$ is (maximal) monotone.

If $F$ is strongly monotone with parameter $m$, then $F^{-1}$ is a function with Lipschitz constant $L = 1/m$. To see this, suppose $u \in F(x)$ and $v \in F(y)$. Then we have

$$\|u - v\|_2\|x - y\|_2 \geq (u - v)^T(x - y) \geq m\|x - y\|_2^2,$$

where the left hand inequality is the Cauchy-Schwarz inequality, and the right hand inequality is the definition of strong monotonicity. We see immediately that if $u = v$ we must have $x = y$, which means that $F^{-1}$ is a function and we can write $x = F^{-1}u$, $y = F^{-1}v$. Dividing the inequality above by $\|x - y\|_2$ (when $x \neq y$) we get

$$\|u - v\|_2 \geq m\|x - y\|_2 = m\|F^{-1}u - F^{-1}v\|_2,$$

which shows that $F^{-1}$ has Lipschitz constant $1/m$.

In general, however, that $F$ is Lipschitz with parameter $L$ does not necessarily imply that $F^{-1}$ is strongly monotone with parameter $1/L$.

**Congruence.** If the relation $F$ on $\mathbf{R}^s$ is monotone and $M \in \mathbf{R}^{s\times t}$, then so is the relation $G$ on $\mathbf{R}^t$ given by

$$G(x) = M^T F(Mx).$$

If $F$ is strongly monotone with parameter $m$, and $M$ has rank $t$ (so $s \geq t$), then $G$ is strongly monotone with parameter $m\sigma_{\min}^2$, where $\sigma_{\min}$ is the smallest singular value of $M$. If $F$ has Lipschitz constant $L$, then $G$ has Lipschitz constant $L\sigma_{\max}^2$, where $\sigma_{\max} = \|M\|_2$ is the largest singular value of $M$.

**Concatenation.** Let $A$ and $B$ be operators on $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively. Then the operator on $\mathbf{R}^{n+m}$

$$C(x, y) = \{(u, v) \mid u \in Ax,\ v \in By\}$$

is monotone if $A$ and $B$ are and maximal monotone if $A$ and $B$ are. We call $C$ a concatenated operator and use the notation

$$C(x, y) = \begin{bmatrix} Ax \\ By \end{bmatrix}$$

and

$$C = \begin{bmatrix} A \\ B \end{bmatrix}$$

to denote this concatenation.

### 4.2. **Subdifferential operator**

Suppose $f : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$. Then $\partial f(x)$ is a monotone operator. If $f$ is CCP then $\partial f$ is maximal monotone. See Fig. 3 for an example.

To prove monotonicity, add the inequalities

$$f(y) \geq f(x) + \partial f(x)^T (y - x), \qquad f(x) \geq f(y) + \partial f(y)^T (x - y),$$

which hold by definition of subdifferentials, to get

$$(\partial f(x) - \partial f(y))^T (x - y) \geq 0.$$

This holds even when $f$ is not convex.

To establish that $\partial f$ is maximal, we show that for any $(\tilde{x}, \tilde{g}) \notin \partial f$ there is $(x, g) \in \partial f$ such that

$$(g - \tilde{g})^T (x - \tilde{x}) < 0,$$

*i.e.*, $\partial f \cup \{(\tilde{x}, \tilde{g})\}$ is not monotone. Let

$$x = \underset{z}{\operatorname{argmin}} \left( f(z) + (1/2)\|z - (\tilde{x} + \tilde{g})\|_2^2 \right).$$

Then

$$0 \in \partial f(x) + x - \tilde{x} - \tilde{g}$$
$$-(x - \tilde{x}) = g - \tilde{g}, \;\; g \in \partial f(x).$$

Since we assumed $(\tilde{x}, \tilde{g}) \notin \partial f$, either $x \neq \tilde{x}$ or $g \neq \tilde{g}$. So we have

$$(g - \tilde{g})^T (x - \tilde{x}) = -\|x - \tilde{x}\|_2^2 = -\|g - \tilde{g}\|_2^2 < 0.$$

This result was first presented in [113] and [111, p. 340].

As we will see a few times throughout this paper, subdifferential operators of CCP functions, a subclass, enjoy certain properties general maximal monotone operators do not.

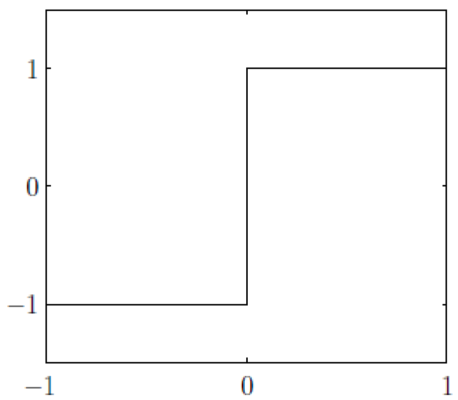



Figure 3. Plot of $\partial f$ for $f(x) = |x|$.

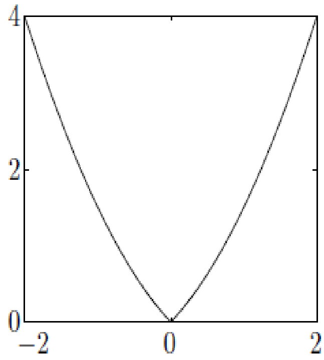
**Differentiability.** A convex function $f$ is differentiable at $x$ if and only if $\partial f(x)$ is a singleton [111, Theorem 25.1]. When $f$ is known or assumed to be differentiable, we write $\nabla f$ instead of $\partial f$.

**Strong convexity and strong smoothness.** We say a CCP $f$ is strongly convex with parameter $m$ if $f(x) - m\|x\|_2^2$ is convex, or equivalently if $\partial f$ is strongly monotone with parameter $m$. When $f$ is twice continuously differentiable, strong convexity is equivalent to $\nabla^2 f(x) \succeq mI$ for all $x$.

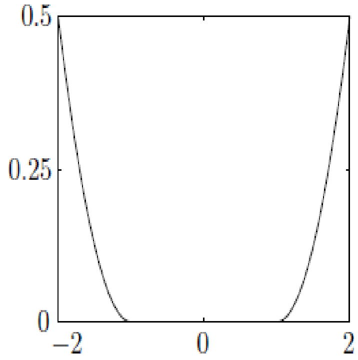
On the other hand, We say a CCP $f$ is strongly smooth with parameter $L$ if $f(x) - L\|x\|_2^2$ is concave or equivalently if $f$ is differentiable and $\nabla f$ is Lipschitz with parameter $L$. When $f$ is twice continuously differentiable, strong smoothness is equivalent to $\nabla^2 f(x) \preceq LI$ for all $x$.

For a CCP functions, strong convexity and strong smoothness are dual properties; a CCP $f$ is strongly convex with parameter $m$ if and only if $f^\star$ is strongly smooth with parameter $L = 1/m$, and vice versa. We discuss these claims in the appendix.

For example, $f(x) = x^2/2 + |x|$, where $x \in \mathbf{R}$, is strongly convex with parameter 1 but not strongly smooth. Its conjugate is $f^*(x) = ((|x| - 1)_+)^2/2$, where $(\cdot)_+$ denotes the positive part, and is strongly smooth with parameter 1 but not strongly convex. See Fig. 4.




(A) $f(x) = x^2/2 + |x|$ is strongly convex but not strongly smooth.

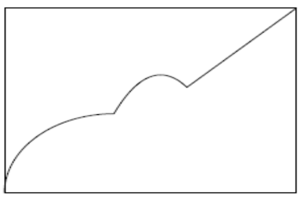



(B) $f^*(x) = ((|x| - 1)_+)^2/2$ is strongly smooth but not strongly convex.

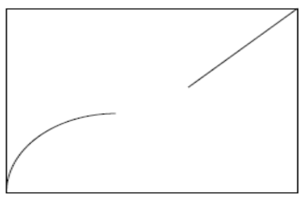
Figure 4. Example of $f$ and its conjugate $f^\star$.

## 4.3. Examples

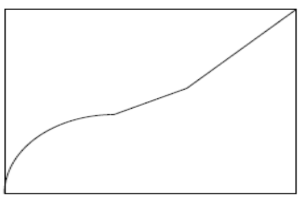
**Relations on R.** We describe this informally. A relation on $\mathbf{R}$ is monotone if it is a curve in $\mathbf{R}^2$ that is always nondecreasing; it can have horizontal (flat) portions and also vertical (infinite slope) portions. If it is a continuous curve with no end points, then it is maximal monotone. It is strongly monotone with parameter $m$ if it maintains a minimum slope $m$ everywhere; it has Lipschitz constant $L$ if its slope is never more than $L$. See Fig. 5.



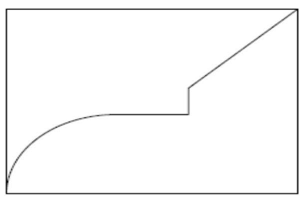
(A) Not monotone.



(B) Monotone but not maximal.



(C) Maximal monotone function.



(D) Maximal monotone but not a function.

Figure 5. Examples of operators on $\mathbf{R}$.

**Continuous functions.** A continuous monotone function $F : \mathbf{R}^n \to \mathbf{R}^n$ (with $\mathbf{dom}\, F = \mathbf{R}^n$) is maximal.

Let us show this. Assume for contradiction that there is a pair $(\tilde{x}, \tilde{u}) \notin F$, such that

$$(\tilde{u} - F(x))^T(\tilde{x} - x) \geq 0$$

for all $x \in \mathbf{R}^n$, *i.e.*, $F \cup \{(\tilde{x}, \tilde{u})\}$ is monotone. Into $x$ we plug in $\tilde{x} - t(z - \tilde{x})$, where $z \in \mathbf{R}^n$ is not yet specified, to get

$$(\tilde{u} - F(\tilde{x} - t(z - \tilde{x})))^T (z - \tilde{x}) \geq 0$$

for all $t > 0$. We take the limit $t \to 0^+$ and use the continuity of $F$ to get

$$(\tilde{u} - F(\tilde{x}))^T (z - \tilde{x}) \geq 0$$

for all $z$. Since $z$ is arbitrary, it must be that $\tilde{u} = F(\tilde{x})$. This is a contradiction, and we conclude $F$ is maximal.

**Affine functions.** An affine function $F(x) = Ax + b$ is maximal monotone if and only if $A + A^T \succeq 0$. It is a subdifferential operator of a CCP function if and only if $A = A^T$ and $A \succeq 0$. It is strongly monotone with parameter $m = \lambda_{\min}(A + A^T)/2$.

**Differentiable function.** A differentiable function $F : \mathbf{R}^n \to \mathbf{R}^n$ is monotone if and only if $DF(x) + DF(x)^T \succeq 0$ for all $x$. It is strongly monotone with parameter $m$ when $DF(x) + DF(x)^T \succeq 2mI$ for all $x$.

To see this, first assume $DF(x) + DF(x)^T \succeq 0$ for all $x$ and define $g(t) = (x - y)^T F(tx + (1 - t)y)$. Then by the mean value theorem

$$\begin{aligned}(x - y)^T (Fx - Fy) &= g(1) - g(0) = g'(\xi) = \\ &= (x - y)^T DF(\xi x + (1 - \xi)y)(x - y) = \\ &= \frac{1}{2}(x - y)^T (DF(\xi x + (1 - \xi)y) + DF(\xi x + (1 - \xi)y)^T)(x - y) \geq 0\end{aligned}$$

for some $\xi \in [0, 1]$, and we conclude monotonicity. The claim regarding strong monotonicity follows from the same argument.

On the other hand, assume $F$ is monotone. Then

$$\begin{aligned}\frac{1}{2}v^T (DF(x) + DF(x)^T)v &= v^T DF(x)v = \\ &= \lim_{h \to 0} \frac{1}{h^2}(x + hv - x)^T (F(x + hv) - F(x)) \geq 0,\end{aligned}$$

and we conclude that $DF(x) + DF(x)^T \succeq 0$ for all $x$.

A continuously differentiable monotone function $F : \mathbf{R}^n \to \mathbf{R}^n$ is a subdifferential operator of a CCP function if and only if $DF(x)$ is symmetric for all $x \in \mathbf{R}^n$. When $n = 3$, this condition is equivalent to the so-called curl-less condition discussed in the context of electromagnetic potentials [1, §10.21 and §12.14].

**Projections.** Monotonicity of projections follow immediately from (2):

$$(\Pi_C x - \Pi_C y)^T (x - y) \geq \|\Pi_C x - \Pi_C y\|_2^2 \geq 0.$$

**Normal cone operator.** Let $C \subseteq \mathbf{R}^n$ be a closed convex set. Its normal cone operator $N_C$ is defined as

$$N_C(x) = \begin{cases} \emptyset & x \notin C \\ \{y \mid y^T (z - x) \leq 0 \ \forall z \in C\} & x \in C. \end{cases}$$

For $x \in \mathbf{int}\, C$, $N_C(x) = \{0\}$; $N_C(x)$ is nontrivial only when $x$ is on the boundary of $C$. As it turns out, $N_C(x)$ is the subdifferential mapping of the (convex) indicator function of $C$, defined by $I_C(x) = 0$, $\mathbf{dom}\, I_C = C$. In other words, $N_C = \partial I_C$.

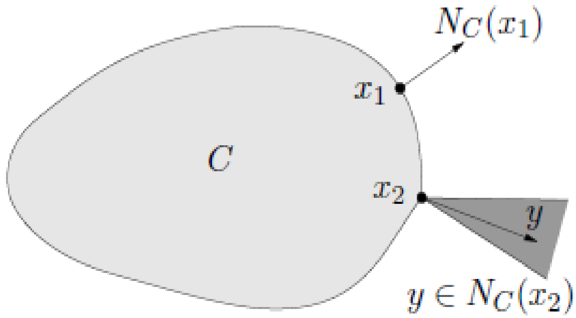



Figure 6. Roughly speaking, $N_C(x)$ is the collection of outward pointing directions with respect to $C$ at point $x$. At $x_1$, there is only one such direction, which can be scaled by any nonnegative value. At $x_2$, there are many such directions.

**Saddle subdifferential.** Suppose that $f : \mathbf{R}^m \times \mathbf{R}^n \to \mathbf{R} \cup \{\pm\infty\}$. The saddle subdifferential relation is defined as

$$F(x,y) = \begin{bmatrix} \partial_x f(x,y) \\ \partial_y(-f(x,y)) \end{bmatrix},$$

nonempty for $(x,y)$ for which $\partial_x f(x,y)$ and $\partial_y(-f(x,y))$ are nonempty. The zero set of $F$ is the set of saddle points of $f$, *i.e.*,

$$(x,y) \in F^{-1}(0) \quad \Leftrightarrow \quad f(x,\tilde{y}) \le f(x,y) \le f(\tilde{x},y) \text{ for all } (x,\tilde{y}),(\tilde{x},y) \in \mathbf{R}^m \times \mathbf{R}^n.$$

When $f$ is convex in $x$ for each $y$, concave in $y$ for each $x$, and satisfies certain regularity conditions, $F$ is maximal [112].

**KKT operator.** Consider the problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \le 0, \quad i = 1,\dots,m \\ & h_i(x) = 0, \quad i = 1,\dots,p, \end{array}$$

where $x$ is the optimization variable, $f_i$ is CCP for $i = 0,\dots,m$, and $h_i$ is affine for $i = 1,\dots,p$. The associated Lagrangian

$$L(x,\lambda,\nu) = \begin{cases} f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x) & \text{for } \lambda \succeq 0 \\ -\infty & \text{otherwise} \end{cases}$$

is a saddle function, and we define the KKT operator as

$$T(x,\lambda,\nu) = \begin{bmatrix} \partial_x L(x,\lambda) \\ -F(x) + N_{\{\lambda \ge 0\}} \\ -H(x) \end{bmatrix},$$

where

$$F(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix}, \qquad H(x) = \begin{bmatrix} h_1(x) \\ \vdots \\ h_p(x) \end{bmatrix}.$$

The operator $T$, a special case of the saddle subdifferential, is monotone. Furthermore, $0 \in T(x^\star,\lambda^\star,\nu^\star)$ if and only if the primal dual pair solves the optimization problem (*i.e.*, the zero set is the set of optimal primal dual pairs).

**Subdifferential of the dual function.** Consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b, \end{array}$$

where $f$ is a strictly convex function, $x \in \mathbf{R}^n$ is the optimization variable, $A \in \mathbf{R}^{m\times n}$, and $b \in \mathbf{R}^m$.

Its dual problem is

$$\text{maximize} \quad g(y) = -(f^*(-A^T y) - y^T b),$$

where $y \in \mathbf{R}^m$ is the optimization variable.

The subdifferential of the dual function, $\partial(-g)$, can be interpreted as the multiplier to residual mapping. Let

$$F(y) = b - Ax, \qquad x = \operatorname*{argmin}_z L(z, y),$$

where $L(x, y) = f(x) + y^T(Ax - b)$ is the Lagrangian. In other words, $F$ maps the dual or multiplier vector $y$ into the associated primal residual $b - Ax$, where $x$ is found by minimizing the Lagrangian. Since $x$ minimizes $L(x, y)$, we have

$$0 \in \partial f(x) + A^T y \quad \Leftrightarrow \quad x = (\partial f)^{-1}(-A^T y),$$

and we plug this back into $F$ and get

$$F(y) = b - A(\partial f)^{-1}(-A^T y) = \partial_y(b^T y + f^\star(-A^T y)) = \partial(-g).$$

## 5. Fixed point iteration

In this section we discuss the main (meta) algorithm of this paper. Recall that $x$ is a fixed point of $F : \mathbf{R}^n \to \mathbf{R}^n$ if $x = Fx$. The algorithm fixed point iteration is

$$x^{k+1} = Fx^k,$$

where $x^0 \in \mathbf{R}^n$ is some starting point, and is used to find a fixed point of $F$. This algorithm, also called the Picard iteration, dates back to [6, 80, 104].

Using the fixed point iteration involves two steps. The first is to find a suitable $F$ whose fixed points are solutions to the problem at hand. We will see examples of this in §6 and §7. The second is to show that the iteration actually converges to a fixed point. (Clearly, the algorithm stays at a fixed point if it starts at a fixed point.) In this section, we will show two simple conditions that guarantee convergence, although these two are not the only possible approaches.

### 5.1. Contractive operators

Suppose that $F : \mathbf{R}^n \to \mathbf{R}^n$ is a contraction with Lipschitz constant $L$ (with $L < 1$) for some norm $\|\cdot\|$ (which need not be the Euclidean norm). In this setting, the fixed point iteration, also called the contraction mapping algorithm in this context, converges to the unique fixed point of $F$.

Let us show this. The sequence $x^k$ is Cauchy. To see this, we note that

$$\begin{aligned} \|x^{k+l} - x^k\| &= \|(x^{k+l} - x^{k+l-1}) + \cdots + (x^{k+1} - x^k)\| \leq \\ &\leq \|x^{k+l} - x^{k+l-1}\|_2 + \cdots + \|x^{k+1} - x^k\| \leq \\ &\leq (L^{l-1} + \cdots + 1)\|x^{k+1} - x^k\| \leq \\ &\leq \frac{1}{1-L}\|x^{k+1} - x^k\|, \end{aligned}$$

for $l \geq 1$. In the third line we use

$$\|x^{k+1} - x^k\| = \|Fx^k - Fx^{k-1}\| \leq L\|x^k - x^{k-1}\|.$$

Therefore $x^k$ converges to a point $x^\star$. It follows that $x^\star$ is the fixed point of $F$ (which we already know is unique) since

$$\begin{aligned}\|Fx^\star - x^\star\| &\le \|x^{k+1} - Fx^\star\| + \|x^{k+1} - x^\star\| \le \\ &\le L\|x^k - x^\star\| + \|x^{k+1} - x^\star\| \to 0.\end{aligned}$$

So a fixed point exists. Note that we can also conclude

$$\|x^k - x^\star\| \le \frac{1}{1-L}\|x^{k+1} - x^k\|,$$

so we have a nonheuristic stopping criterion for the contraction mapping algorithm.

Furthermore, the distance to the fixed point $x^\star$ decreases at each step. (An algorithm with this property is called Fejér monotone.) To see this, we simply note that

$$\|x^{k+1} - x^\star\| = \|Fx^k - Fx^\star\| \le L\|x^k - x^\star\|.$$

This also shows that $\|x^k - x^\star\| \le L^k\|x^0 - x^\star\|$, *i.e.*, convergence is at least geometric, with factor $L$.

**Gradient method.** Consider the problem

$$\text{minimize} \quad f(x),$$

where $x \in \mathbf{R}^n$ is the optimization variable, and $f$ is a CCP function on $\mathbf{R}^n$.

Assume $f$ is differentiable. Then $x$ is a solution if and only if

$$0 = \nabla f(x) \quad \Leftrightarrow \quad x = (I - \alpha\nabla f)(x)$$

for any nonzero $\alpha \in \mathbf{R}$. In other words, $x$ is a solution if and only if it is a fixed point of the mapping $I - \alpha\nabla f$.

The fixed point iteration for this setup is

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

This algorithm, which dates back to [27], is called the gradient method or gradient descent, and $\alpha$ is called the step size in this context.

Now assume $f$ is strongly convex and strongly smooth with parameters $m$ and $L$, respectively. Then $I - \alpha\nabla f$ is Lipschitz with parameter $L_{\mathrm{GM}} = \max\{|1-\alpha m|, |1-\alpha L|\}$. Let us prove this assuming $f$ is twice continuously differentiable (although it is still true without this assumption). Then $D(I - \alpha\nabla f) = I - \alpha\nabla^2 f$, and therefore

$$(1-\alpha L)I \preceq D(I - \alpha F) \preceq (1-\alpha m)I,$$

where (somewhat confusingly) $I$ also denotes the identity matrix here. So $\|D(I-\alpha\nabla f)(x)\|_2 \le \max\{|1-\alpha m|, |1-\alpha L|\}$ for all $x$, and we conclude $I - \alpha\nabla f$ is Lipschitz with parameter $L_{\mathrm{GM}}$.

So under these assumptions, $I - \alpha\nabla f$ is a contractive operator for $\alpha \in (0, 2/L)$. Consequently, the solution $x^\star$ exists, and gradient method converges geometrically with rate

$$\|x^k - x^\star\| \le L_{\mathrm{GM}}^k\|x^0 - x^\star\|.$$

The value of $\alpha$ that minimizes $L_{\mathrm{GM}}$ is $2/(L+m)$, and the corresponding optimal contraction factor is $(\kappa-1)/(\kappa+1) = 1 - 2/\kappa + O(1/\kappa^2)$.

**Forward step method.** Consider the problem of finding an $x \in \mathbf{R}^n$ that satisfies

$$0 = F(x),$$

where $F : \mathbf{R}^n \to \mathbf{R}^n$.

By the same argument, $x$ is a solution if and only if it is a fixed point of $I - \alpha F$ for any nonzero $\alpha \in \mathbf{R}$. The fixed point iteration for this setup is

$$x^{k+1} = x^k - \alpha Fx^k,$$

which we call the forward step method.

Now assume $F$ is strongly monotone and Lipschitz with parameters $m$ and $L$, respectively. Also assume $\alpha > 0$. Then

$$\begin{aligned}\|(I-\alpha F)x-(I-\alpha F)y\|_2^2 &= \|x-y+\alpha Fx-\alpha Fy\|_2^2 = \\ &= \|x-y\|_2^2 - 2\alpha(Fx-Fy)^T(x-y)+\alpha^2\|Fx-Fy\|_2^2 \le \\ &\le (1-2\alpha m+\alpha^2L^2)\|x-y\|_2^2.\end{aligned}$$

So for $\alpha \in (0, 2m/L^2)$ the iteration is a contraction. Consequently, the solution exist and the method converges geometrically to the solution.

This result is, however, weaker than what we had for the gradient method: the values of $\alpha$ for which the iteration converges is more restrictive, the contraction factor is worse for all values of $\alpha > 0$, and the optimal contraction factor $1 - m^2/L^2 = 1 - 1/\kappa^2$, given by $\alpha = m/L^2$, is worse.

Furthermore, the forward step method in general will not converge without the strong monotonicity assumption. (We will soon see that the gradient method converges without strong convexity, albeit slowly.) For example,

$$F(x,y) = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix},$$

which is the KKT operator of the problem of minimizing $x$ subject to $x = 0$, is monotone but not strongly monotone and has Lipschitz constant 1. By computing the singular values, we can verify that $I - \alpha F$ is an expansion for any $\alpha \neq 0$, *i.e.*, all singular values are greater than 1. Therefore the iterates of the forward step method on $F$ diverge away from the solution (unless we start at the solution).

### 5.2. **Averaged operators**

Suppose that $F : \mathbf{R}^n \to \mathbf{R}^n$ is averaged. Then the fixed point iteration, also called the damped, averaged, or Mann-Krasnosel'skii iteration in this context [77,83], converges to a solution if one exists.

Let us be specific. Assume the set of fixed points $X$ is nonempty. Then we can conclude that $x^k \to x^\star$ for some $x^\star \in X$. Moreover, the algorithm is Fejér monotone, *i.e.*, $\mathbf{dist}(x^k, X) = \inf_{z\in X}\|x-z\|_2 \to 0$ monotonically. We also have

$$\|Fx^k - x^k\|_2 \to 0$$

with rate

$$\min_{j=0,\dots,k} \|Fx^j - x^j\|_2^2 = O(1/k). \tag{3}$$

In other words, the algorithm produces points for which the fixed point condition $x = F(x)$ holds arbitrarily closely with rate $O(1/k)$. (The quantity in (3) is what we care about since our stopping criterion will be $\|Fx^k - x^k\|_2 \le \epsilon$.)

When $F : \mathbf{R}^n \to \mathbf{R}^n$ is nonexpansive but not averaged, the fixed point iteration need not converge to $X$, even when $X$ is nonempty. Simple examples: $F$ is a rotation about some line, or a reflection through a plane. In this case, then we can use the averaged operator $G = (1-\theta)I + \theta F$ with $\theta \in (0, 1)$ in the fixed point iteration to find a fixed point of $F$.

**Gradient method.** Again consider the problem

$$\text{minimize} \quad f(x),$$

where $x \in \mathbf{R}^n$ is the optimization variable, and $f$ is a CCP function on $\mathbf{R}^n$.

Assume $f$ is differentiable. Then as discussed before, the gradient method

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

is a fixed point iteration for this problem for a nonzero $\alpha \in \mathbf{R}$.

Now assume $f$ is strongly smooth with parameter $L$. By the same argument as before, $I - \alpha\nabla f$ is Lipschitz with parameter $L_{\mathrm{GM}} = \max\{1, |1-\alpha L|\}$ and therefore is nonexpansive for $\alpha \in (0, 2/L]$. So it is averaged for $\alpha \in (0, 2/L)$ since

$$(I - \alpha\nabla f) = (1-\theta)I + \theta(I - 2/L\nabla f),$$

where $\theta = \alpha L/2 < 1$. Consequently, $x^k \to x^\star$ for some solution $x^\star$, if one exists, with rate

$$\min_{j=0,\ldots,k} \|\nabla f(x^j)\|_2^2 = O(1/k),$$

for any $\alpha \in (0, 2/L)$. Of course, this rate is worse than what we had when we also assumed strong convexity.

The (sub)gradient method with constant step size, in general, does not converge if $\partial f$ is not Lipschitz. For example, if the (sub)gradient method is applied to the function $\|x\|_1$ and some starting point $x^0 \neq 0$, then the method fails to converge for almost all values of $\alpha$.

**Convergence proof.** We will use the identity

$$\|(1-\theta)a + \theta b\|_2^2 = (1-\theta)\|a\|_2^2 + \theta\|b\|_2^2 - \theta(1-\theta)\|a-b\|_2^2, \tag{4}$$

which holds for any $\theta \in \mathbf{R}$, $a, b \in \mathbf{R}^n$. (It can be verified by expanding both sides as a quadratic function of $\theta$.) For $\theta \in (0,1)$, the first two terms on the righthand side correspond to Jensen's inequality, applied to the convex function $\|\cdot\|_2^2$. The third term on the righthand side improves the basic Jensen inequality.

Now let $F = (1-\theta)I + \theta G$ be an averaged operator, where $\theta \in (0,1)$ and $G$ is nonexpansive. Consider the fixed point iteration

$$x^{k+1} = F(x^k) = (1-\theta)x^k + \theta G(x^k),$$

and let $x^\star \in X$. Using our identity (4), we have

$$\begin{aligned}\|x^{k+1} - x^\star\|_2^2 &= (1-\theta)\|x^k - x^\star\|_2^2 + \theta\|G(x^k) - x^\star\|_2^2 - \theta(1-\theta)\|G(x^k) - x^k\|_2^2 \leq \\ &\leq (1-\theta)\|x^k - x^\star\|_2^2 + \theta\|x^k - x^\star\|_2^2 - \theta(1-\theta)\|G(x^k) - x^k\|_2^2 = \\ &= \|x^k - x^\star\|_2^2 - \theta(1-\theta)\|G(x^k) - x^k\|_2^2, \end{aligned} \tag{5}$$

where we use $\|G(x^k) - x^\star\|_2 \leq \|x^k - x^\star\|_2$ in the second line. This shows that the fixed point ieration with an averaged operator is Fejér monotone, *i.e.*, $\mathbf{dist}(x^k, X)$ decreases at each step.

Iterating the inequality above, we have

$$\|x^{k+1} - x^\star\|_2^2 \leq \|x^0 - x^\star\|_2^2 - \theta(1-\theta)\sum_{j=0}^{k} \|G(x^j) - x^j\|_2^2,$$

so

$$\sum_{j=0}^{k} \|G(x^j) - x^j\|_2^2 \leq \frac{\|x^0 - x^\star\|_2^2}{\theta(1-\theta)},$$

and thus

$$\|G(x^k) - x^k\|_2 \to 0.$$

This also implies

$$\min_{j=0,\ldots,k} \|G(x^j) - x^j\|_2^2 \leq \frac{\|x^0 - x^\star\|_2^2}{(k+1)\theta(1-\theta)}. \tag{6}$$

As convergence rates go, this is pretty bad; it corresponds to the subgradient method.

Let's show that $x^k \to x^\star$ for some $x^\star \in X$. By picking a point in $x \in X$ and applying (5) we see that the iterates $x^k$ lie within the compact set $\{z \mid \|z - x\|_2 \leq \|x^0 - x\|_2\}$ and therefore must have a limit point. This limit point $x^\star$ must be in $X$, *i.e.*, must satisfy $F(x^\star) - x^\star = 0$, as $F(x^k) - x^k \to 0$ and $F - I$ is continuous. Finally, applying (5) to this limit point $x^\star \in X$, we conclude that $\|x^k - x^\star\|$ monotonically decreases to 0, *i.e.*, the entire sequence converges to $x^\star$. So $\mathbf{dist}(x^k, X) \le \|x^k - x^\star\|_2 \to 0$.

The choice $\theta = 1/2$ maximizes $\theta(1-\theta)$, and therefore minimizes the righthand side of (6). To the extent that the proof predicts actual convergence rates (which it does not in general), this would be the optimal choice of $\theta$. So when we construct a averaged operator from a nonexpansive one, we can expect the choice $\theta = 1/2$, which corresponds to the simple iteration

$$x^{k+1} = (1/2)x^k + (1/2)G(x^k)$$

to come up.

### 5.3. Examples

**Dual ascent.** Consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b, \end{array}$$

where $x \in \mathbf{R}^n$ is the optimization variable, $f$ is a CCP function on $\mathbf{R}^n$, $A \in \mathbf{R}^{m\times n}$, and $b \in \mathbf{R}^m$. Its dual is

$$\text{maximize} \quad g(y),$$

where $g(y) = -(f^*(-A^T y) - y^T b)$ and $y \in \mathbf{R}^m$ is the optimization variable.

Assume that $f$ is strongly convex with parameter $m$ and that $\sigma_{\max}$, the maximum singular value of $A$, is positive. Then $\partial f^* = (\partial f)^{-1}$ is Lipschitz with parameter $1/m$ and $\partial(-g)$ is Lipschitz with parameter $\sigma_{\max}^2/m$.

The gradient method applied to $-g$ (which is the fixed point iteration on $I + \alpha\nabla g$) becomes

$$\begin{aligned} x^{k+1} &= \operatorname*{argmin}_x L(x, y^k) \\ y^{k+1} &= y^k + \alpha(Ax^{k+1} - b), \end{aligned}$$

where $L(x,y)$ denotes the Lagrangian. This method is called the Uzawa method [2] or dual ascent [125, 127]. Assume strong duality holds and optimal primal and dual solutions exist. Then dual ascent converges for $\alpha \in (0, 2\sigma_{\max}^2/m)$.

Furthermore, if we assume $f$ is strongly smooth with parameter $L$ and $\sigma_{\min}$, the smallest singular value of $A$, is positive then $-g(y)$ is strongly convex with parameter $\sigma_{\min}^2/L$, and we get geometric convergence.

**Projections onto convex sets.** Consider the problem of finding an $x \in C \cap D$, where $C$ and $D$ are nonempty closed convex sets. This is also called the convex feasibility problem.

Recall that $\mathbf{dist}(x, X)$ is the distance of $x$ to the set $X$, *i.e.*,

$$\mathbf{dist}(x, X) = \inf_{z\in X} \|x - z\|_2.$$

If $X$ is a nonempty closed convex set, then $f(x) = 1/2\,\mathbf{dist}^2(x, X)$ is CCP and strongly smooth with parameter 1, and we have $\nabla f(x) = (I - \Pi_X)(x)$. See [7, §12.4] for a proof.

Let $\theta \in (0,1)$. Then $x \in C \cap D$ if and only if $x$ is the solution to the optimization problem

$$\text{minimize} \quad (\theta/2)\,\mathbf{dist}^2(x, C) + ((1-\theta)/2)\,\mathbf{dist}^2(x, D)$$

with optimal value 0.

The objective of the optimization problem is CCP and strongly smooth with parameter 1. So we can use the gradient method with step size 1 to get (parallel) projections onto convex sets:

$$\begin{aligned} x_C^{k+1} &= \Pi_C x^k \\ x_D^{k+1} &= \Pi_D x^k \\ x^{k+1} &= \theta x_C^{k+1} + (1-\theta)x_D^{k+1}. \end{aligned}$$

If $C \cap D \neq \emptyset$, then $x^k \to x^\star$ for some $x^\star \in C \cap D$. This method dates back to [32]. See [9,10,47] for an overview of projection methods for the convex feasibility problem.

## 6. Resolvent and Cayley operator

The resolvent of a relation $A$ on $\mathbf{R}^n$ is defined as

$$R = (I + \alpha A)^{-1},$$

where $\alpha \in \mathbf{R}$. The Cayley operator, reflection operator, or reflected resolvent of $A$ is defined as

$$C = 2R - I.$$

When we are considering the resolvents or Cayley operators of multiple relations, we denote them with a subscript, as in $R_A$ or $C_A$.

When $\alpha > 0$, we have the following.

- If $A$ is monotone, then $R$ and $C$ are nonexpansive functions.
- If $A$ is maximal monotone, then $\mathbf{dom}\, R = \mathbf{dom}\, C = \mathbf{R}^n$.
- $0 \in A(x)$ if and only if $x = R_A(x) = C_A(x)$.

Let's first show that $R$ is nonexpansive. Suppose $(x, u) \in R$ and $(y, v) \in R$. By definition of $R$, we have

$$u + \alpha A(u) \ni x, \qquad v + \alpha A(v) \ni y.$$

Subtract these to get

$$u - v + \alpha(Au - Av) \ni x - y. \tag{7}$$

Multiply by $(u - v)^T$, and use monotonicity of $A$ to get

$$\|u - v\|_2^2 \leq (x - y)^T (u - v). \tag{8}$$

Now we apply Cauchy-Schwarz and divide by $\|u - v\|_2$ to get

$$\|u - v\|_2 \leq \|x - y\|_2,$$

*i.e.*, $R$ is nonexpansive.

Next, let's show that $C = 2R - I$ is nonexpansive. Using the inequality (8), we get

$$\begin{aligned} \|Cx - Cy\|_2^2 &= \|2(u - v) - (x - y)\|_2^2 = \\ &= 4\|u - v\|_2^2 - 4(x - y)^T (u - v) + \|x - y\|_2^2 \leq \\ &\leq \|x - y\|_2^2, \end{aligned}$$

*i.e.*, $C$ is nonexpansive. Since $R$ and $C$ are nonexpansive, they are single-valued.

That $\mathbf{dom}\, R = \mathbf{dom}\, C = \mathbf{R}^n$, called the Minty surjectivity theorem [89], is harder to show, and we skip the proof. Interested readers can refer to [4, §6.2], [7, §21], or [3, §4.1].

Finally, we prove the third claim:

$$\begin{aligned} 0 \in A(x) \quad &\Leftrightarrow \quad x \in (I + A)(x) \\ &\Leftrightarrow \quad (I + A)^{-1}(x) \ni x \\ &\Leftrightarrow \quad x = R_A(x), \end{aligned}$$

where the last line uses the fact that $R_A$ is a function. The statement about $C_A$ follows simply by definition.

**When the resolvent is a contraction.** If $A$ is strongly monotone with parameter $m$, then $R$ is Lipschitz with parameter $L = 1/(1 + \alpha m)$. To show this, we observe that $I + \alpha A$ is strongly monotone with parameter $1 + \alpha m$. Therefore its inverse, $R$, has Lipschitz constant $1/(1 + \alpha m)$:

$$\|Rx - Ry\|_2 \leq \frac{1}{1 + \alpha m} \|x - y\|_2.$$

**When the Cayley operator is a contraction.** When $A$ is merely strongly monotone, $C$ need not be a contraction. But if $A$ is strongly monotone with parameter $m$ and also has Lipschitz constant $L$, then $C$ is a contraction with Lipschitz constant

$$L_C = \left(1 - \frac{4\alpha m}{(1+\alpha L)^2}\right)^{1/2}.$$

To prove this, let $(x,u) \in R$ and $(y,v) \in R$. Then

$$u - v + \alpha(Au - Av) = x - y \tag{9}$$

and multiply by $(u-v)^T$ to get

$$\|u-v\|_2^2 + \alpha(u-v)^T(Au - Av) = (u-v)^T(x-y).$$

Using strong monotonicity, we get

$$(1+\alpha m)\|u-v\|_2^2 \le (u-v)^T(x-y), \tag{10}$$

which is a strengthened form of (8). Expanding $\|Cx - Cy\|_2^2$ as before, but using the sharper inequality (10), we get

$$\|Cx - Cy\|_2^2 \le \|x-y\|_2^2 - 4\alpha m\|u-v\|_2^2.$$

Now take the norm of (9) to get

$$\|u-v\|_2 + \alpha\|Au - Av\|_2 \ge \|x-y\|_2.$$

Using the Lipschitz inequality we get

$$\|u-v\|_2 \ge \frac{1}{1+\alpha L}\|x-y\|_2.$$

Combined with our inequality above we get

$$\|Cx - Cy\|_2^2 \le \|x-y\|_2^2 - \frac{4\alpha m}{(1+\alpha L)^2}\|x-y\|_2^2 = \left(1 - \frac{4\alpha m}{(1+\alpha L)^2}\right)\|x-y\|_2^2,$$

which is a strict contraction for all positive values of $m$, $L$, and $\alpha$. The choice $\alpha = 1/L$ is optimal and yields a contraction factor of

$$\sqrt{1 - 1/\kappa} = 1 - 1/2\kappa + O(1/\kappa^2).$$

On the other hand, when $A$ is a subdifferential operator of a CCP function that is strongly convex and strongly smooth with parameters $m$ and $L$, respectively, the contraction factor can be further improved to

$$L_C = \max\left\{\left|\frac{1-\alpha L}{1+\alpha L}\right|, \left|\frac{1-\alpha m}{1+\alpha m}\right|\right\},$$

which is strictly smaller than the previous one [58]. The optimal choice $\alpha = 1/\sqrt{mL}$ yields a contraction factor of

$$(\sqrt{\kappa}-1)/(\sqrt{\kappa}+1) = 1 - 2/\sqrt{\kappa} + O(1/\kappa),$$

which, again, is better than the previous one.

The difference between the contraction factors, the first for general monotone operators and the second specifically for subdifferential operators, is not an artifact of the proof. See [58] for further details.

**Zero set of a maximal monotone operator.** If $A$ is maximal monotone, then $R_A$ is nonexpansive with $\mathbf{dom}\, R_A = \mathbf{R}^n$ for any $\alpha > 0$. Since the set of fixed points of $R_A$ is closed and convex, the zero set of $A$ is closed and convex.

If $A$ is maximal and strongly monotone, then there is exactly one fixed point of $A$, as $R_A$ is a contraction.

**Cayley operator identities.** When a monotone operator $A$ is maximal and single-valued and $\alpha \geq 0$, we have

$$C_A = (I - \alpha A)(I + \alpha A)^{-1}.$$

This follows simply from

$$\begin{aligned} C_A &= 2(I + \alpha A)^{-1} - I = \\ &= 2(I + \alpha A)^{-1} - (I + \alpha A)(I + \alpha A)^{-1} = \\ &= (I - \alpha A)(I + \alpha A)^{-1}. \end{aligned}$$

When a monotone operator $A$ is maximal (but not necessarily single-valued) and $\alpha > 0$, we have

$$C_A(I + \alpha A) = I - \alpha A. \tag{11}$$

To see this, first note the assumptions make $(I + \alpha A)^{-1}$ a function. So for any $x \in \mathbf{dom}\, A$, we have

$$\begin{aligned} C_A(I + \alpha A)(x) &= 2(I + \alpha A)^{-1}(I + \alpha A)(x) - (I + \alpha A)(x) = \\ &= 2I(x) - (I + \alpha A)(x) = \\ &= (I - \alpha A)(x). \end{aligned}$$

For any $x \notin \mathbf{dom}\, A$, both sides are empty sets.

## 6.1. Examples

**Matrices.** It is easier to see why $R$ and $C$ are nonexpansive when the operator is linear. If $F$ is a symmetric matrix and $F \geq 0$, its eivenvalues are positive. If $\alpha \geq 0$, then (the inverse defining) $R = (I + \alpha F)^{-1}$ exists, and has eigenvalues in $(0, 1]$. It follows that the matrix

$$C = 2R - I = (I - \alpha F)(I + \alpha F)^{-1} = (I + \alpha F)^{-1}(I - \alpha F),$$

which is called the Cayley transform of $F$, has eigenvalues in $(-1, 1]$.

**Subdifferential mapping.** Suppose $f$ is convex and $\alpha > 0$. Let us work out what $(I + \alpha\partial f)^{-1}$ is:

$$\begin{aligned} z = (I + \alpha\partial f)^{-1}(x) \quad &\Leftrightarrow \quad z + \alpha\partial f(z) \ni x \\ &\Leftrightarrow \quad 0 \in \partial_z \left(\alpha f(z) + (1/2)\|z - x\|_2^2\right) \\ &\Leftrightarrow \quad z = \underset{u}{\operatorname{argmin}} \left(f(u) + (1/2\alpha)\|u - x\|_2^2\right). \end{aligned}$$

(Another way to see that the argmin is unique is to note that the function on the righthand side is strictly convex.) The resolvent $R = (I + \alpha\partial f)^{-1}$ is called the proximal operator or proximity operator associated with $f$, with parameter $\alpha$. When $f$ is CCP, $\mathbf{dom}\, R = \mathbf{R}^n$, even if $f$ takes on infinite values, *i.e.*, $R(x)$ is defined for all $x \in \mathbf{R}^n$ even when $\mathbf{dom}\, f \neq \mathbf{R}^n$.

**Normal cone operator.** Suppose $C$ is closed, convex, and with the normal cone operator $N_C$. Then by noting that $N_C(x) = \partial I_C(x)$ we conclude

$$(I + \alpha N_C)^{-1} = \Pi_C$$

for any $\alpha > 0$. Of course, the overprojection operator, $Q_C = 2\Pi_C - I$, is the Cayley operator of $N_C$.

**Subdifferential of the dual function.** Consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b, \end{array}$$

where $f$ is a CCP function on $\mathbf{R}^n$, $x \in \mathbf{R}^n$ is the optimization variable, $A \in \mathbf{R}^{m\times n}$, and $b \in \mathbf{R}^m$. Its dual is

$$\text{maximize} \quad g(y),$$

where $g(y) = -(f^*(-A^T y) - y^T b)$ and $y \in \mathbf{R}^m$ is the optimization variable.

Assume $f$ is strictly convex, and let $\alpha > 0$. Let us examine $R_{\partial(-g)}(y) = v$. Remember that

$$\partial(-g)(v) = b - Ax, \qquad \partial f(x) + A^T v \ni 0.$$

So we have

$$\begin{aligned} v = R_{\partial(-g)}(y) \quad &\Leftrightarrow \quad v + \alpha\partial(-g)(v) = y \\ &\Leftrightarrow \quad v + \alpha(b - Ax) = y, \quad \partial f(x) + A^T v \ni 0. \end{aligned}$$

Reorganizing, we see that $x$ satisfies

$$\partial f(x) + A^T(y + \alpha(Ax - b)) \ni 0,$$

and we get

$$\begin{aligned} x &= \operatorname*{argmin}_z L_\alpha(z, y) \\ v &= y + \alpha(Ax - b), \end{aligned}$$

where

$$L_\alpha(x, y) = f(x) + y^T(Ax - b) + (\alpha/2)\|Ax - b\|_2^2$$

is the augmented Lagrangian. The first step above is minimizing the augmented Lagrangian; the second is a multiplier update.

**KKT operator for linearly constrained problems.** Consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b, \end{array}$$

where $f$ is a CCP function on $\mathbf{R}^n$, $x \in \mathbf{R}^n$ is the optimization variable, $A \in \mathbf{R}^{m\times n}$, and $b \in \mathbf{R}^m$. Consider the KKT operator

$$T(x, y) = \begin{bmatrix} \partial f(x) + A^T y \\ b - Ax \end{bmatrix},$$

which corresponds to the Lagrangian

$$L(x, y) = f(x) + y^T(Ax - b).$$

Let $\alpha > 0$. Then

$$R_T(x, y) = (u, v) \quad \Leftrightarrow \quad \begin{bmatrix} x \\ y \end{bmatrix} \in \begin{bmatrix} u \\ v \end{bmatrix} + \alpha \begin{bmatrix} \partial f(u) + A^T v \\ b - Au \end{bmatrix}.$$

The second line gives us

$$v = y + \alpha(Au - b),$$

and with substitution the first line gives us

$$0 \in \partial f(u) + A^T y + \alpha A^T(Au - b) + \frac{1}{\alpha}(u - x).$$

Reorganizing, we get

$$\begin{aligned} u &= \operatorname*{argmin}_z \left(L_\alpha(z, y) + (1/2\alpha)\|z - x\|_2^2\right) \\ v &= y + \alpha(Au - b). \end{aligned}$$

This is quite similar to the resolvent of the subdifferential of the dual function. The first step above is minimizing the augmented Lagrangian with an additional regularization term; the second is a multiplier update.

### 6.2. **Proximal point method**

Consider the problem of finding an $x$ that satisfies

$$0 \in A(x),$$

where $A$ is maximal monotone.

Let $\alpha > 0$. As discussed before, $0 \in A(x)$ if and only if $x = C(x)$. The fixed point iteration for this setup is

$$x^{k+1} = C(x^k),$$

which we call the Cayley method. This, however, may not converge when $C$ is merely nonexpansive. A simple example is when $A = N_{\{0\}}$ and $C = Q_{\{0\}}$.

Likewise, $0 \in A(x)$ if and only if $x = R(x)$, when $\alpha > 0$. The associated fixed point iteration is

$$x^{k+1} = R(x^k),$$

which is called the proximal point method or proximal minimization and was first presented in [22, 85, 86, 117]. The proximal point method, on the other hand, always converges to a solution if one exists, as $R$ is an averaged operator.

**Role of maximality.** A fixed point iteration $x^{k+1} = F(x^k)$ becomes undefined if its iterates ever escapes the domain of $F$. (This is why we assumed $\mathbf{dom}\, F = \mathbf{R}^n$ in §5.) When $A$ is maximal and $\alpha > 0$, the Cayley iteration and the proximal point method does not have this problem.

So we assume maximality out of theoretical necessity, but in practice the non-maximal monotone operators, such as the subdifferential operator of a nonconvex function, are usually ones we cannot efficiently compute the resolvent for anyways. In other words, there is little need to consider resolvents or Cayley operators of non-maximal monotone operators, theoretically or practically.

**Method of multipliers.** Consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b, \end{array}$$

where $f$ is a CCP function on $\mathbf{R}^n$, $x \in \mathbf{R}^n$ is the optimization variable, $A \in \mathbf{R}^{m\times n}$, and $b \in \mathbf{R}^n$. Write $g$ for the dual function of the dual optimization problem.

Assume $f$ is strictly convex. Then a dual variable $y$ is optimal if and only if $0 \in -\nabla g$. Writing out the proximal point method $y^{k+1} = R_{-\nabla g}(y^k)$, we get

$$\begin{aligned} x^{k+1} &= \operatorname*{argmin}_x L_\alpha(x, y^k) \\ y^{k+1} &= y^k + \alpha(Ax^{k+1} - b). \end{aligned}$$

This method is called the method of multipliers and was first presented in [67, 106, 115].

Assume strong duality holds and a primal and dual solution exists. Then $y^k \to y^\star$ for some optimal dual variable $y^\star$, as the method is an instance of the proximal point method. Using standard arguments from convex analysis, one can also show that $x^k \to x^\star$.

When $f$ is not strictly convex, the minimizer of the augmented Lagrangian may not be unique. If so, choose any minimizer, *i.e.*, let

$$\begin{aligned} x^{k+1} &\in \operatorname*{argmin}_x L_\alpha(x, y^k) \\ y^{k+1} &= y^k + \alpha(Ax^{k+1} - b), \end{aligned}$$

and we retain all the desirable convergence properties. (However, making this statement precise goes beyond the scope of this paper.)

**Proximal method of multipliers.** Consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b, \end{array}$$

where $f$ is a CCP function on $\mathbf{R}^n$, $x \in \mathbf{R}^n$ is the optimization variable, $A \in \mathbf{R}^{m\times n}$, and $b \in \mathbf{R}^n$. Write $T(x,y)$ for the associated KKT operator, where $y \in \mathbf{R}^m$.

A primal-dual pair $(x,y)$ is optimal if and only if $0 \in T(x,y)$. Writing out the proximal point method $(x^{k+1}, y^{k+1}) = R_T(x^k, y^k)$, we get

$$\begin{aligned} x^{k+1} &= \underset{x}{\operatorname{argmin}} \left( L_\alpha(x, y^k) + (1/2\alpha)\|x - x^k\|_2^2 \right) \\ y^{k+1} &= y^k + \alpha(Ax^{k+1} - b). \end{aligned}$$

We then scale the equality constraint $Ax = b$ and re-parameterize to get the algorithm

$$\begin{aligned} x^{k+1} &= \underset{x}{\operatorname{argmin}} \left( L_{\alpha_1}(x, y^k) + (\alpha_2/2)\|x - x^k\|_2^2 \right) \\ y^{k+1} &= y^k + \alpha_1(Ax^{k+1} - b), \end{aligned}$$

where $\alpha_1 > 0$ and $\alpha_2 > 0$. Assume strong duality holds and primal and dual solutions exist. Then $x^k \to x^\star$ and $y^k \to y^\star$ for some optimal $x^\star$ and $y^\star$. This method, called the proximal method of multipliers, was first presented in [116, 118].

The proximal method of multipliers has two advantages over the method of multipliers. The first is that the primal iterates $x^k$ are uniquely defined and converge to a solution, regardless of whether $f$ is strictly convex or the solution is unique. The second is that the subproblem of minimizing the Lagrangian is better conditioned due to the additional regularizing term.

**Iterative refinement.** Consider the problem of finding an $x \in \mathbf{R}^n$ that solves the linear system $Ax = b$, where $A \in \mathbf{R}^{n\times n}$ and $b \in \mathbf{R}^n$. Of course, this is equivalent to finding a zero of the operator

$$F(x) = Ax - b.$$

Assume $A^T + A \succeq 0$. Then $F$ is maximal monotone, and we can apply the proximal point method with $R_F = (I + 1/\varepsilon F)^{-1}$ to get

$$\begin{aligned} r^k &= Ax^k - b \\ x^{k+1} &= x^k - (A + \varepsilon I)^{-1} r^k, \end{aligned}$$

an instance of iterative refinement [64,84,130]. We can interpret each iteration to be refining the iterate $x^k$ by correcting for its residual $r^k$. If a solution exists, then $x^k \to x^\star$ for some solution $x^\star$ for any $\varepsilon > 0$,

Iterative refinement is useful when $A$ is either singular or has a large condition number. In such a case, approximately solving $Ax = b$ using $(A + \varepsilon I)^{-1}$ may be easier than directly using $A^{-1}$. In particular, $A + \varepsilon I$ will always have well-defined LU factorization as all its leading principal minors are nonsingular [63, Theorem 3.2.1]. The factorization can be computed once and reused every iteration.

**Generalized proximal point method.** Let $F$ be a maximal monotone operator on $\mathbf{R}^n$, and $L \in \mathbf{R}^{n\times n}$ be invertible. Then $L^{-T}FL^{-1}$ is monotone and maximal [7, Theorem 24.5]. The proximal point method

$$y^{k+1} = (I + \alpha L^{-T}FL^{-1})^{-1} y^k$$

with $\alpha > 0$ converges to a zero of $L^{-T}FL^{-1}$ if one exists, and for any zero $y^\star$ of $L^{-T}FL^{-1}$

$$\|y^k - y^\star\|_2$$

decreases monotonically.

Let us re-parameterize the iteration with

$$Lx^k = y^k.$$

Then we have

$$\begin{aligned} y^{k+1} + \alpha L^{-T} F L^{-1} y^{k+1} &\ni y^k \\ (A + \alpha F) x^{k+1} &\ni A x^k, \end{aligned}$$

where $A = L^T L$. Of course, $A \succ 0$. Conversely, we can start with a positive definite $A \in \mathbf{R}^{n \times n}$, and let $L = A^{1/2}$. We call the iteration

$$x^{k+1} = (A + \alpha F)^{-1} A x^k,$$

the generalized proximal point method, and its iterates $x^k$ inherit the convergence properties of $y^k$. In particular, $x^k$ converges to a zero of $F$ if one exists, and for any zero $x^\star$ of $F$

$$\|x^k - x^\star\|_A = \sqrt{(x^k - x^\star)^T A (x^k - x^\star)} = \|y^k - Lx^\star\|_2$$

decreases monotonically, where $\|\cdot\|_A$ is the $A$-norm (cf. [69, Theorem 5.3.2] or [63, p. 626]).

## 7. Operator splitting

In this section, we consider the problem of finding a zero of a monotone operator that admits a splitting into two or three maximal monotone operators. In other words, we wish to find an $x$ that satisfies $0 \in (A+B)(x)$ or $0 \in (A+B+C)(x)$, where $A$, $B$, and $C$ are maximal monotone.

The idea is to transform this problem into a fixed-point equation with operators constructed from $A$, $B$, $C$, their resolvents, and Cayley operators. We then apply the fixed point iteration and discuss its convergence. In practice, these methods will be useful when the operators used are efficient to compute.

### 7.1. **Forward-backward splitting**

Consider the problem of finding an $x$ that satisfies

$$0 \in (A + B)(x),$$

where $A$ and $B$ are maximal monotone.

Assume $A$ is single-valued and $\alpha > 0$. Then we have

$$\begin{aligned} 0 \in (A+B)(x) \quad &\Leftrightarrow \quad 0 \in (I + \alpha B)(x) - (I - \alpha A)(x) \\ &\Leftrightarrow \quad (I + \alpha B)(x) \ni (I - \alpha A)(x) \\ &\Leftrightarrow \quad x = (I + \alpha B)^{-1}(I - \alpha A)(x). \end{aligned}$$

So $x$ is a solution if and only if it is a fixed point of $R_B(I - \alpha A)$.

The fixed point iteration for this setup is

$$x^{k+1} = R_B(x^k - \alpha A x^k).$$

This method, first presented in [100], is called forward-backward splitting.

Assume that $A$ is a subdifferential operator with Lipschitz parameter $L$ and that $\alpha \in (0, 2/L)$. Or assume that $A$ is respectively strongly monotone and Lipschitz with parameters $m$ and $L$ and that $\alpha \in (0, 2m/L^2)$. Then the forward step $I - \alpha A$ is averaged, as discussed in §5. The backward step $(I+\alpha B)^{-1}$ is averaged for any $\alpha > 0$. So when $I - \alpha A$ is averaged, the composition $(I+\alpha B)^{-1}(I - \alpha A)$ is an averaged operator, and the iterates converge to a solution if one exists.

The steps $I - \alpha A$ and $(I + \alpha B)^{-1}$ are respectively called forward and backward steps in analogy to the forward and backward Euler methods used to solve differential equations. See §3.2.2 of [44] or §4.1.1 of [99] for a discussion on this interpretation.

**Proximal gradient method.** Consider the problem

$$\text{minimize} \quad f(x) + g(x),$$

where $x \in \mathbf{R}^n$ is the optimization variable and $f$ and $g$ are CCP functions on $\mathbf{R}^n$. Of course, $x$ is a solution if and only if $x$ satisfies

$$0 \in (\partial f + \partial g)(x),$$

assuming $\textbf{relint dom}\, f \cap \textbf{relint dom}\, g \neq \emptyset$.

Assume $f$ is differentiable. Then forward-backward splitting applied to $\nabla f + \partial g$ is

$$x^{k+1} = \underset{x}{\operatorname{argmin}} \left( f(x^k) + \nabla f(x^k)^T (x - x^k) + g(x) + \frac{1}{2\alpha} \|x - x^k\|_2^2 \right),$$

which is called called the proximal gradient method [12, 36, 42]. Unlike the proximal point method, we use the first-order approximation of $f$ about $x^k$ in the minimization. When a solution exists, $f$ is strongly smooth with parameter $L$, and $\alpha \in (0, 2/L)$, the proximal gradient method converges.

### 7.2. Forward-backward-forward splitting

Again, consider the problem of finding an $x$ that satisfies

$$0 \in (A + B)(x),$$

where $A$ and $B$ are maximal monotone.

Assume $A$ is Lipschitz with parameter $L$ (and therefore single-valued) and let $\alpha \in (0, 1/L)$. Then the function $I - \alpha A$ is a one-to-one mapping. To see this, consider any distinct $x$ and $y$ and we have

$$\begin{aligned} \|(I - \alpha A)x - (I - \alpha A)y\|_2 &= \|x - y - \alpha(Ax - Ay)\|_2 \geq \\ &\geq \|x - y\|_2 - \alpha \|(Ax - Ay)\|_2 \geq \\ &\geq (1 - \alpha L)\|x - y\|_2 > 0. \end{aligned}$$

We used the reverse triangle inequality and Lipschitz continuity of $A$ on the second and third lines, respectively. So $(I - \alpha A)x \neq (I - \alpha A)y$, and we conclude $I - \alpha A$ is one-to-one.

As discussed in §7.1, $x$ is a solution if and only if it is a fixed point of $R_B(I - \alpha A)$. Since $I - \alpha A$ is one-to-one, we have

$$\begin{aligned} 0 \in (A + B)(x) \quad &\Leftrightarrow \quad x = R_B(I - \alpha A)(x) \\ &\Leftrightarrow \quad (I - \alpha A)x = (I - \alpha A)R_B(I - \alpha A)(x) \\ &\Leftrightarrow \quad x = ((I - \alpha A)R_B(I - \alpha A) + \alpha A)(x). \end{aligned}$$

So $x$ is a solution if and only if it is a fixed point of $(I - \alpha A)R_B(I - \alpha A) + \alpha A$.

The fixed point iteration for this setup is

$$\begin{aligned} x^{k+1/2} &= R_B(x^k - \alpha A x^k) \\ x^{k+1} &= x^{k+1/2} - \alpha(A x^{k+1/2} - A x^k). \end{aligned}$$

This method, developed by Tseng [126], is called forward-backward-forward splitting.

The mapping $(I - \alpha A)R_B(I - \alpha A) + \alpha A$ is not nonexpansive, so the convergence results of §5 do not apply. (It is, however, nonexpansive towards any solution.) Nevertheless, forward-backward-forward splitting converges when $A$ is Lipschitz with parameter $L$ and $\alpha \in (0, 1/L)$. (This is a weaker assumption than what was necessary for forward-backward splitting to converge.)

Let us be specific. Assume the set of fixed points $X$ is nonempty. Then $x^k \to x^\star$ for some $x^\star \in X$. Moreover, the iteration is Fejér monotone, *i.e.*, $\textbf{dist}(x^k, X) \to 0$ monotonically. Furthermore,

$$\|x^{k+1/2} - x^k\|_2 \to 0$$

with rate

$$\min_{j=0,\ldots,k} \|x^{k+1/2} - x^k\|_2^2 = O(1/k).$$

**Convergence proof.** Assume a solution $x^\star$ exists. Since $A + B$ is monotone and $0 \in A(x^\star) + B(x^\star)$, we have

$$(x^{k+1/2} - x^\star)^T(Ax^{k+1/2} + Bx^{k+1/2}) \in (x^{k+1/2} - x^\star)^T(Ax^{k+1/2} + Bx^{k+1/2} - Ax^\star - Bx^\star) \geq 0.$$

Using this inequality, we get

$$\begin{aligned}\|x^{k+1} - x^\star\|_2^2 = \|x^k - x^\star\|_2^2 + \alpha^2\|Ax^{k+1/2} - Ax^k\|_2^2 - \|x^{k+1/2} - x^k\|_2^2 \\ - 2\alpha(x^{k+1/2} - x^\star)^T(Ax^{k+1/2} + Bx^{k+1/2}) \leq \\ \leq \|x^k - x^\star\|_2^2 - (1 - \alpha^2 L^2)\|x^{k+1/2} - x^k\|_2^2,\end{aligned}$$

where we use Lipschitz continuity of $A$ and the preliminary inequality we proved. Using the same argument as in §5.2, we get the desired results.

**Extragradient method.** Consider the problem of finding an $x$ that satisfies

$$0 = A(x),$$

where $A$ is maximal monotone and single-valued.

The extragradient method,

$$\begin{aligned} x^{k+1/2} &= x^k - \alpha A x^k \\ x^{k+1} &= x^k - \alpha A x^{k+1/2}, \end{aligned}$$

a special case of forward-backward-forward splitting with $B = 0$, converges to a solution if $A$ is Lipschitz with parameter $L$ and $\alpha \in (0, 1/L)$. This method was first presented in [76].

**Combettes-Pesquet.** Consider the problem of solving

$$0 \in (A + B + C)(x),$$

where $A$, $B$, and $C$ are maximal monotone.

Then $x$ is a solution if and only if

$$\begin{aligned} &0 \in Ax + u + Cx \\ &u \in Bx \end{aligned}$$

for some $u$. In turn, $x$ is a solution if and only if

$$0 \in \begin{bmatrix} Ax \\ B^{-1}u \end{bmatrix} + \begin{bmatrix} u + Cx \\ -x \end{bmatrix}$$

for some $u$.

Forward-backward-forward splitting on this setup is

$$\begin{aligned} x^{k+1/2} &= R_A(x^k - \alpha u^k - \alpha C x^k) \\ u^{k+1/2} &= R_{B^{-1}}(u^k + \alpha x^k) \\ x^{k+1} &= x^{k+1/2} - \alpha(u^{k+1/2} - u^k + Cx^{k+1/2} - Cx^k) \\ u^{k+1} &= u^{k+1/2} + \alpha(x^{k+1/2} - x^k), \end{aligned}$$

which we call the Combettes-Pesquet method [35]. If $C$ is Lipschitz with parameter $L$ and if a solution exists, this method converges for $0 < \alpha < 1/(1 + L)$.

### 7.3. **Peaceman-Rachford and Douglas-Rachford splitting**

Again, consider the problem of finding an $x$ that satisfies

$$0 \in (A + B)(x),$$

where $A$ and $B$ are maximal monotone.

**Peaceman-Rachford splitting.** The key fixed point result is

$$0 \in (A + B)(x) \quad \Leftrightarrow \quad C_A C_B(z) = z,\ x = R_B(z)$$

for any $\alpha > 0$. Let us show this:

$$\begin{aligned} 0 \in Ax + Bx \quad &\Leftrightarrow \quad 0 \in (I + \alpha A)x - (I - \alpha B)x \\ &\Leftrightarrow \quad 0 \in (I + \alpha A)x - C_B(I + \alpha B)x \\ &\Leftrightarrow \quad 0 \in (I + \alpha A)x - C_B z,\ z \in (I + \alpha B)x \\ &\Leftrightarrow \quad C_B z \in (I + \alpha A)R_B z,\ x = R_B z \\ &\Leftrightarrow \quad R_A C_B z = R_B z,\ x = R_B z \\ &\Leftrightarrow \quad C_A C_B z = z,\ x = R_B z, \end{aligned}$$

where we have used (11).

The fixed point iteration for this setup is

$$\begin{aligned} x^{k+1/2} &= R_B(z^k) \\ z^{k+1/2} &= 2x^{k+1/2} - z^k \\ x^{k+1} &= R_A(z^{k+1/2}) \\ z^{k+1} &= 2x^{k+1} - z^{k+1/2}. \end{aligned}$$

This method, first presented in [74, 81, 101], is called Peaceman-Rachford splitting.

Without further assumptions, the mapping $C_A C_B$ is only guaranteed to be nonexpansive for $\alpha \geq 0$. So the iteration need not converge.

**Douglas-Rachford splitting.** To ensure convergence, we average the nonexpansive mapping. Clearly, for any $\alpha > 0$ we have

$$x \in (A + B)(x) \quad \Leftrightarrow \quad (1/2I + 1/2C_A C_B)(z) = z,\ x = R_B(z).$$

The fixed point iteration for this setup is

$$\begin{aligned} x^{k+1/2} &= R_B(z^k) \\ z^{k+1/2} &= 2x^{k+1/2} - z^k \\ x^{k+1} &= R_A(z^{k+1/2}) \\ z^{k+1} &= z^k + x^{k+1} - x^{k+1/2}. \end{aligned}$$

This method, first presented in [41, 81], is called Douglas-Rachford splitting. For $\alpha > 0$, the mapping is averaged and the iterates converge to a solution, if one exists.

We can think of $x^{k+1/2}$ and $x^{k+1}$ as estimates of a solution, with slightly different properties. For example, if $R_B$ is a projection onto a constraint set, then the estimate $x^{k+1/2}$ satisfies these constraints exactly.

**Geometric convergence.** As discussed before, $C_A$ and $C_B$ are always nonexpansive. So Peaceman-Rachford and Douglas-Rachford converge geometrically when either $C_A$ or $C_B$ is contractive. This, for example, happens if $A$ is strongly monotone and Lipschitz.

With a more refined analysis, one can find other conditions that ensure geometric convergence. For example, if $f$ is a strongly convex CCP function, $g$ is a strongly smooth CCP function, $A = \partial f$, and $B = \partial g$, then $1/2I + 1/2C_AC_B$ is a contraction, and Douglas-Rachford converges geometrically [39, 57].

**Consensus optimization.** Consider the problem

$$\text{minimize} \quad \textstyle\sum_{i=1}^m f_i(x),$$

where $x \in \mathbf{R}^n$ is the optimization variable and $f_1, f_2, \dots, f_m$ are CCP functions on $\mathbf{R}^n$. This problem is equivalent to

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m f_i(x_i) \\ \text{subject to} & x_1 = x_2 = \cdots = x_m, \end{array}$$

where $x_1, x_2, \dots, x_m \in \mathbf{R}^n$ are the optimization variables. In turn, this problem is equivalent to finding $x_1, x_2, \dots, x_m \in \mathbf{R}^n$ that satisfies

$$0 \in \begin{bmatrix} \partial f_1(x_1) \\ \partial f_2(x_2) \\ \vdots \\ \partial f_m(x_m) \end{bmatrix} + N_{\{x_1 = x_2 = \cdots = x_m\}}(x_1, x_2, \dots, x_m),$$

assuming $\bigcap_{i=1}^m \textbf{relint}\,\textbf{dom}\, f_i \neq \emptyset$.

The constraint $x_1 = x_2 = \cdots = x_m$ is called the consensus constraint, and the projection onto it is simple averaging:

$$\Pi(x_1, x_2, \dots, x_m) = (\overline{x}, \overline{x}, \dots, \overline{x}), \qquad \overline{x} = \frac{1}{m}\sum_{i=1}^m x_i.$$

So when we apply Douglas-Rachford splitting to this setup,

$$\begin{aligned} x_i^{k+1} &= \underset{x}{\operatorname{argmin}} \left( f_i(x) + (1/2\alpha)\|x - z_i^k\|_2^2 \right), \quad i = 1, 2, \dots, m \\ z_i^{k+1} &= z_i^k + 2\overline{x}^{k+1} - \overline{z}^k - x_i^{k+1}, \quad i = 1, 2, \dots, m, \end{aligned}$$

where $\overline{x}^k$ is the average of $x_1^k, x_2^k, \dots, x_m^k$, and $\overline{z}^k$ is defined similarly. If a solution exists, this method converges for $\alpha > 0$. An advantage of this method is that it is conducive to parallelization since minimization step splits. See [19, 98, 99] for further discussions.

More generally, one can solve

$$0 \in \sum_{i=1}^n A_i x,$$

where $A_1, \dots, A_n$ are maximal monotone, with this approach.

### 7.4. Davis-Yin three-operator splitting

So far we have looked at splitting schemes with two operators. Finding a splitting scheme with three or more operators that cannot be reduced to a splitting of two has been a major open problem for a while. Here, we briefly present Davis and Yin's recent breakthrough [40].

Consider the problem of solving

$$0 \in (A + B + C)(x),$$

where $A$, $B$, and $C$ are maximal monotone.

Assume $C$ is single-valued and let $\alpha > 0$. Then we have

$$0 \in (A + B + C)(x) \quad \Leftrightarrow \quad Tz = z,\ x = R_B(z),$$

where

$$T = C_A(C_B - \alpha C R_B) - \alpha C R_B.$$

Let us show this:

$$\begin{aligned}
0 \in Ax + Bx + Cx \quad &\Leftrightarrow \quad 0 \in (I + \alpha A)x - (I - \alpha B)x + \alpha Cx \\
&\Leftrightarrow \quad 0 \in (I + \alpha A)x - C_B(I + \alpha B)x + \alpha Cx \\
&\Leftrightarrow \quad 0 \in (I + \alpha A)x - C_B z + \alpha Cx, \; z \in (I + \alpha B)x \\
&\Leftrightarrow \quad (C_B - \alpha C R_B)z \in (I + \alpha A)R_B z, \; x = R_B z \\
&\Leftrightarrow \quad R_A(C_B - \alpha C R_B)z = R_B z, \; x = R_B z \\
&\Leftrightarrow \quad (C_A(C_B - \alpha C R_B) - \alpha C R_B)z = z, \; x = R_B z,
\end{aligned}$$

where we have used (11).

Of course, this also means $x$ is a solution if and only if $z$ is a fixed point of $1/2I + 1/2T$ with $x = R_B z$. The fixed point iteration for this setup is

$$\begin{aligned}
x^{k+1/2} &= R_B(z^k) \\
z^{k+1/2} &= 2x^{k+1/2} - z^k \\
x^{k+1} &= R_A(z^{k+1/2} - \alpha C x^{k+1/2}) \\
z^{k+1} &= z^k + x^{k+1} - x^{k+1/2}.
\end{aligned}$$

This splitting scheme reduces to Douglas-Rachford when $C = 0$ and to forward-backward when $B = 0$.

Assume that $C$ is a subdifferential operator with Lipschitz parameter $L$ and that $\alpha \in (0, 2/L)$. Or assume that $C$ is respectively strongly monotone and Lipschitz with parameters $m$ and $L$ and that $\alpha \in (0, 2m/L^2)$. Davis and Yin showed that under these assumptions $1/2I + 1/2T$ is averaged, which implies convergence. (However, $T$ itself may not be nonexpansive without further assumptions.)

### 7.5. **Examples**

**Iterative shrinkage-thresholding algorithm.** Consider the problem

$$\text{minimize} \quad f(x) + \lambda \|x\|_1,$$

where $x \in \mathbf{R}^n$ is the optimization variable, $f$ is a CCP function on $\mathbf{R}^n$, and $\lambda > 0$.

Assume $f$ is differentiable. Then the proximal gradient method applied to this problem is

$$x^{k+1} = S_{\alpha\lambda}(x^k - \alpha \nabla f(x^k)),$$

which is called the Iterative Shrinkage-Thresholding Algorithm (ISTA) [12,36]. Here, $S_\kappa$ is called the soft thresholding operator or shrinkage operator and defined element-wise as

$$S_\kappa(x)_i = \mathbf{sign}(x_i)(|x_i| - \kappa)_+$$

for $i = 1, 2, \ldots, n$. See Fig. 6. If a solution exists, $f$ is strongly smooth with parameter $L$, and $\alpha \in (0, 2/L)$, then ISTA converges.

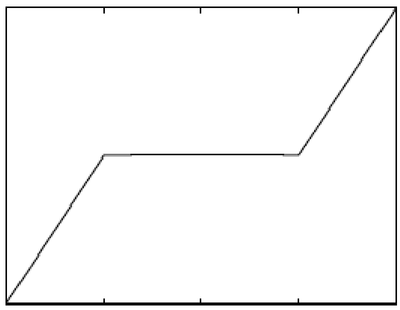



Figure 7. The soft thresholding operator.

**Projected gradient method.** Consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in C, \end{array}$$

where $x \in \mathbf{R}^n$ is the optimization variable, $C \subseteq \mathbf{R}^n$ is a nonempty closed convex set, and $f$ is a CCP function on $\mathbf{R}^n$.

Assume $f$ is differentiable. Then the proximal gradient method applied to this setup is

$$x^{k+1} = \Pi_C(x^k - \alpha \nabla f(x^k)).$$

This algorithm, first presented in [62, 79], is also called the projected gradient method. If a solution exists, $f$ is strongly smooth with parameter $L$, and $\alpha \in (0, 2/L)$, then this method converges.

**Projections onto convex sets.** Consider the problem of finding an $x \in C \cap D$, where $C$ and $D$ are nonempty closed convex sets, *i.e.*, the convex feasibility problem. Note that $x \in C \cap D$ if and only if $x$ is the solution to the optimization problem

$$\begin{array}{ll} \text{minimize} & (1/2)\,\mathbf{dist}^2(x, D) \\ \text{subject to} & x \in C \end{array}$$

with optimal value 0.

Since the objective of the optimization problem is CCP and strongly smooth with parameter 1, we can use the proximal gradient method with step size 1 to get serial or alternating projections onto convex sets:

$$x^{k+1} = \Pi_C \Pi_D x^k.$$

If $C \cap D \neq \emptyset$, then $x^k \to x^\star$ for some $x^\star \in C \cap D$. This method dates back to [129, Theorem 13.7] and [30, 66].

**Another alternating projections method.** Again, consider the convex feasibility problem of finding an $x \in C \cap D$, where $C$ and $D$ are nonempty closed convex sets. This problem is equivalent to finding an $x$ satisfying

$$0 \in (N_C + N_D)(x).$$

Applying Douglas-Rachford splitting to this setup gives us

$$\begin{aligned} x^{k+1/2} &= \Pi_D(z^k) \\ x^{k+1} &= \Pi_C(2x^{k+1/2} - z^k) \\ z^{k+1} &= z^k + x^{k+1} - x^{k+1/2}, \end{aligned}$$

which resembles but is not the same as Dykstra's method of alternating projections [8, 11, 21]. This method converges if a solution exists and is generally faster than classical alternating projections (although the analysis doesn't tell us why).

**Chambolle-Pock.** Consider the optimization problem

$$\text{minimize} \quad f(x) + g(Mx),$$

where $x \in \mathbf{R}^n$ is the optimization variable, $M \in \mathbf{R}^{m\times n}$, and $f$ and $g$ are CCP functions on $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively. This setup has been studied in several contexts such as image processing [23, 48, 133].

The optimality condition for this problem is

$$0 \in \partial f(x) + M^T \partial g(Mx),$$

assuming $\mathbf{relint\,dom}\, f \cap \mathbf{relint\,dom}\, g(M\cdot) \neq \emptyset$. This holds if and only if

$$\begin{aligned} 0 &\in \partial f(x) + M^T u \\ Mx &\in (\partial g)^{-1}(u) \end{aligned}$$

for some $u \in \mathbf{R}^m$. So $x$ is a solution if and only if

$$0 \in \begin{bmatrix} \partial f(x) \\ \partial g^*(u) \end{bmatrix} + \begin{bmatrix} 0 & M^T \\ -M & 0 \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix}$$

for some $u \in \mathbf{R}^m$.

We now apply the generalized proximal point method with

$$A = \begin{bmatrix} I & -\alpha M^T \\ -\alpha M & I \end{bmatrix},$$

where $\alpha$ satisfies $0 < \alpha < 1/\|M\|_2$. (One can show $A$ is positive definite using its Schur complement [69, Theorem 7.7.6].) Writing out the iteration we get

$$\alpha \begin{bmatrix} \partial f(x^{k+1}) \\ \partial g^*(u^{k+1}) \end{bmatrix} + \begin{bmatrix} x^{k+1} \\ u^{k+1} - 2\alpha M x^{k+1} \end{bmatrix} \ni \begin{bmatrix} x^k - \alpha M^T u^k \\ u^k - \alpha M x^k \end{bmatrix}.$$

This simplifies to

$$\begin{aligned} x^{k+1} &= R_{\partial f}(x^k - \alpha M^T u^k) \\ u^{k+1} &= R_{\partial g^*}(u^k + \alpha M(2x^{k+1} - x^k)), \end{aligned}$$

which we call the Chambolle-Pock method [28]. This method converges for $0 < \alpha < 1/\|M\|_2$, if a solution exists and $\mathbf{relint\,dom}\, f \cap \mathbf{relint\,dom}\, g(M\cdot) \neq \emptyset$.

Sometimes, it may be computationally easier to evaluate the resolvent of $\partial g^*$ than to evaluate the resolvent of $M^T \partial g M$. For a simple example, think of $g(x) = \|x\|_1$. This algorithm can be useful in such a setting.

**A first-order method for LPs.** Consider the problem

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b \\ & x \succeq 0, \end{array}$$

where $x \in \mathbf{R}^n$ is the optimization variable, $A \in \mathbf{R}^{m\times n}$, $b \in \mathbf{R}^m$, and $c \in \mathbf{R}^n$.

The KKT operator associated with this problem is $T = T_1 + T_2$, where

$$T_1(x,\nu,\lambda) = \begin{bmatrix} c + A^T\nu - \lambda \\ -(Ax - b) \\ x \end{bmatrix} \qquad T_2(x,\nu,\lambda) = \begin{bmatrix} 0 \\ 0 \\ N_{\{\lambda \succeq 0\}} \end{bmatrix},$$

and $(x,\nu,\lambda)$ is primal-dual optimal if and only if $0 \in (T_1 + T_2)(x,\nu,\lambda)$.

When we apply forward-backward-forward splitting to this setup we get

$$\begin{aligned}
x^{k+1/2} &= x^k - \alpha(c + A^T\nu^k - \lambda^k)\\
\nu^{k+1/2} &= \nu^k + \alpha(Ax^k - b)\\
\lambda^{k+1/2} &= (\lambda^k - \alpha x^k)_+\\
x^{k+1} &= x^k - \alpha(c + A^T\nu^{k+1/2} - \lambda^{k+1/2})\\
\nu^{k+1} &= \nu^k + \alpha(Ax^{k+1/2} - b)\\
\lambda^{k+1} &= (\lambda^k - \alpha x^{k+1/2})_+ + \alpha^2(c + A^T\nu^k - \lambda^k),
\end{aligned}$$

where $(\cdot)_+$ takes the positive part, element-wise.

If a primal solution exists, a dual solution exists due to standard LP duality [14, 82]. Then this method converges for $\alpha \in \left(0, 1/\sqrt{\sigma_{\max}^2 + 1}\right)$, where $\sigma_{\max}$ is the largest singular value of $A$, and has rate of convergence

$$\min_{j=0,\dots,k} \left\{\|Ax^j - b\|_2^2 + \|c + A^T\nu^j - \lambda^j\|_2^2 + \|\min\{\lambda^j/\alpha, x^j\}\|_2^2\right\} = O(1/k).$$

The first term enforces primal feasibility, the second term dual feasibility, and the third term complementary slackness and $x \succeq 0$.

**Convex-concave games.** A (zero-sum, two-player) game on $\mathbf{R}^m \times \mathbf{R}^n$ is defined by its payoff function $f : \mathbf{R}^m \times \mathbf{R}^n \to \mathbf{R} \cup \{\pm\infty\}$. The meaning is that player 1 chooses a value (or move) $x \in \mathbf{R}^m$, and player 2 chooses a value (or move) $v \in \mathbf{R}^n$; based on these choices, player 1 makes a payment to player 2, in the amount $f(x, y)$. The goal of player 1 is to minimize this payment, while the goal of player 2 is to maximize it. See [20, 55] for further discussions.

We say that $(x, y)$ is a solution of the game if it is a saddle point of $f$. Let

$$F(x, y) = \begin{bmatrix} \partial_x f(x, y) \\ \partial_y(-f(x, y)) \end{bmatrix}$$

be the saddle subdifferential of $f$. Then $(x, y)$ is a solution of the game if and only if $0 \in F(x, y)$.

Assume that $f(x, y)$ is CCP and strongly smooth with parameter $L_1$ for each $y$ as a function of $x$ and that $-f(x, y)$ is CCP and strongly smooth with parameter $L_2$ for each $x$ as a function of $y$. Then $F$ is Liptschitz with parameter $L_1 + L_2$, and we can find a solution with the extragradient method (or equivalently with forward-backward-forward splitting):

$$\begin{aligned}
x^{k+1/2} &= x^k - \alpha\nabla_x f(x^k, y^k)\\
y^{k+1/2} &= y^k + \alpha\nabla_y f(x^k, y^k)\\
x^{k+1} &= x^k - \alpha\nabla_x f(x^{k+1/2}, y^{k+1/2})\\
y^{k+1} &= y^k + \alpha\nabla_y f(x^{k+1/2}, y^{k+1/2}),
\end{aligned}$$

which converges for $\alpha \in (0, 1/(L_1 + L_2))$, if a solution exists.

**Complementarity problem.** Consider the problem of finding an $x \in \mathbf{R}^n$ that satisfies

$$\begin{aligned}
&x \in K\\
&F(x) \in K^\star\\
&x^T F(x) = 0,
\end{aligned}$$

where $K$ is a nonempty closed convex cone, and $F$ is an operator that is single-valued on $K$. This problem is called the complementarity problem, and is sometimes written more concisely as finding an $x \in \mathbf{R}^n$ that satisfies

$$x \in K \perp F(x) \in K^\star.$$

It is not too hard to show that this is equivalent to finding an $x$ that satisfies

$$0 \in (F + N_K)(x),$$

where $N_K$ is the normal cone operator. Many problems in mechanics, economics, and game theory are naturally posed as complementarity problems. See [38, 49, 50] for more details.

Assume $F$ is maximal monotone and Lipschitz with parameter $L$. Then we can use forward-backward-forward splitting to solve the complementarity problem:

$$\begin{aligned} x^{k+1/2} &= \Pi_K(x^k - \alpha F x^k) \\ x^{k+1} &= x^{k+1/2} - \alpha(F x^{k+1/2} - F x^k), \end{aligned}$$

which converges for $\alpha \in (0, 1/L)$, if a solution exists. (When $F$ is not maximal monotone, complementarity problems are hard; there are no known polynomial time algorithms to solve them [31].)

**Quasidefinite systems.** Consider the problem of finding an $x \in \mathbf{R}^n$ that solves the linear system

$$Kx = b,$$

where $b \in \mathbf{R}^n$. The matrix $K \in \mathbf{R}^{n\times n}$ is (symmetric) quasidefinite, *i.e.*,

$$K = \begin{bmatrix} -A & C \\ C^T & B \end{bmatrix},$$

where $A \in \mathbf{R}^{m\times m}$ and $B \in \mathbf{R}^{(n-m)\times(n-m)}$ are positive definite and $C \in \mathbf{R}^{m\times(n-m)}$. For further discussions on quasidefinite matrices, see [56, 128].

Define

$$J = \begin{bmatrix} -I_m & 0 \\ 0 & I_{(n-m)} \end{bmatrix},$$

where $I_m$ and $I_{(n-m)}$ are the $m \times m$ and $(n-m)\times(n-m)$ identity matrices, respectively. Now consider the operator

$$F(x) = J(Kx - b),$$

which is monotone since $JK + (JK)^T \succ 0$. Since $J$ is invertible, $x$ is a solution if and only if $F(x) = 0$.

Write

$$K_1 = \begin{bmatrix} -A & 0 \\ 0 & B \end{bmatrix}, \qquad K_2 = \begin{bmatrix} 0 & C \\ C^T & 0 \end{bmatrix}$$

(so that $K = K_1 + K_2$). Now we have the splitting $F = F_1 + F_2$ with $F_1(x) = JK_1x$ and $F_2(x) = JK_2x - Jb$, and we can apply Peaceman-Rachford splitting to get

$$x^{k+1} = \tilde{b} + (J + \alpha K_2)^{-1}(J - \alpha K_1)(J + \alpha K_1)^{-1}(J - \alpha K_2)x^k,$$

where

$$\tilde{b} = \alpha(J + \alpha K_2)^{-1}(I + (J - \alpha K_1)(J + \alpha K_1)^{-1})b.$$

This method converges to the solution for all $\alpha > 0$.

**ADMM.** Consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) + g(z) \\ \text{subject to} & Ax + Bz = c, \end{array}$$

where $x \in \mathbf{R}^m$ and $z \in \mathbf{R}^n$ are the optimization variables, $A \in \mathbf{R}^{l\times m}$, $B \in \mathbf{R}^{l\times n}$, $c \in \mathbf{R}^l$, and $f$ and $g$ are respectively CCP functions on $\mathbf{R}^m$ and $\mathbf{R}^n$. Its dual problem is

$$\text{maximize} \quad -f^*(-A^T\nu) - g^*(-B^T\nu) + c^T\nu,$$

where $\nu \in \mathbf{R}^l$ is the optimization variable.

The subdifferential operator of the dual function

$$F(\nu) = -A\partial f^*(-A^T\nu) - B\partial g^*(B^T\nu) - c$$

admits the splitting $F = F_1 + F_2$, where

$$F_1(\nu) = -A\partial f^*(-A^T\nu) - c \qquad F_2(\nu) = -B\partial g^*(B^T\nu).$$

Assume that $f$ and $g$ are strictly convex, so that the argmin are well-defined. Applying Douglas-Rachford splitting to $F = F_1 + F_2$, we get

$$\begin{aligned}
\zeta^{k+1} &= R_{F_2}(y^k)\\
\xi^{k+1} &= R_{F_1}(2\zeta^{k+1} - y^k)\\
y^{k+1} &= y^k + \xi^{k+1} - \zeta^{k+1}.
\end{aligned}$$

Evaluating the resolvents involves a minimization step, as discussed in §6.1. Making these explicit we get

$$\begin{aligned}
\tilde{z}^{k+1} &= \underset{z}{\operatorname{argmin}} \left( g(z) + (y^k)^T Bz + \frac{\alpha}{2}\|Bz\|_2^2 \right)\\
\zeta^{k+1} &= y^k + \alpha B\tilde{z}^{k+1}\\
\tilde{x}^{k+1} &= \underset{x}{\operatorname{argmin}} \left( f(x) + (y^k + 2\alpha B\tilde{z}^{k+1})^T(Ax - c) + \frac{\alpha}{2}\|Ax - c\|_2^2 \right)\\
\xi^{k+1} &= y^k + \alpha(A\tilde{x}^{k+1} - c) + 2\alpha B\tilde{z}^{k+1}\\
y^{k+1} &= y^k + \alpha(A\tilde{x}^{k+1} + B\tilde{z}^{k+1} - c).
\end{aligned}$$

Since $\zeta^k$ and $\xi^k$ iterates no longer have any explicit dependence, they can be removed. Next substitute $y^k = \alpha u^k + \alpha(A\tilde{x}^k - c)$

$$\begin{aligned}
\tilde{z}^{k+1} &= \underset{z}{\operatorname{argmin}} \left( g(z) + \frac{\alpha}{2}\|A\tilde{x}^k + Bz - c + u^k\|_2^2 \right)\\
\tilde{x}^{k+1} &= \underset{x}{\operatorname{argmin}} \left( f(x) + \frac{\alpha}{2}\|Ax + B\tilde{z}^{k+1} - c + u^{k+1}\|_2^2 \right)\\
u^{k+1} &= u^k + A\tilde{x}^k + B\tilde{z}^{k+1} - c.
\end{aligned}$$

Finally, we swap the order of the $u^{k+1}$ and $\tilde{x}^{k+1}$ update to get the correct dependency and substitute $\tilde{x}^k = x^{k+1}$ and $\tilde{z}^k = z^k$ to get the alternating direction method of multipliers (ADMM):

$$\begin{aligned}
x^{k+1} &= \underset{x}{\operatorname{argmin}} \left( f(x) + \frac{\alpha}{2}\|Ax + Bz^k - c + u^k\|_2^2 \right)\\
z^{k+1} &= \underset{z}{\operatorname{argmin}} \left( g(z) + \frac{\alpha}{2}\|Ax^{k+1} + Bz - c + u^k\|_2^2 \right)\\
u^{k+1} &= u^k + Ax^{k+1} + Bz^{k+1} - c,
\end{aligned}$$

which was first presented in [54,61]. Assume strong duality holds and primal and dual solutions exist. Then for any $\alpha > 0$ we have $x^k \to x^\star$, $z^k \to z^\star$, and $u^k \to u^\star$, where $(x^\star, z^\star)$ is primal optimal and $\alpha u^\star$ is dual optimal.

Recently, ADMM has gained a wide popularity. For an overview, interested readers can refer to [19, 45, 46, 99]. It is worth noting that there are other ways to analyze ADMM. One approach avoids discussing monotone operators and relies on first principles [19,54,61]. Another views ADMM as the proximal point method applied to the so called splitting operator [43]. Another obtains ADMM by applying Douglas-Rachford splitting to the primal optimization problem [131]. Here, we followed the approach of [53], to derive ADMM by applying Douglas-Rachford splitting to the dual problem.

## 8. Further topics

In this primer we have covered the basic ideas of monotone operators, convergence of fixed point iteration, and applications to a solving variety of convex optimization problems. The same basic components that we have described can be assembled in other ways to derive still more algorithms.

We conclude by listing here some further topics that are closely related to, or an extension of, the material we have covered.

**Inexact solves.** When evaluating the operators (especially the resolvents) it may be useful to do so approximately. Say we perform the (approximate) proximal point method

$$x^{k+1} = R(x^k) + \varepsilon^k,$$

where $\varepsilon^k$ denotes the error. Under certain assumptions on $\|\varepsilon^k\|$ one can prove convergence results. Analysis of algorithms using monotone operators is more amenable to this type of error analysis than analysis using first principles [46, 117].

**Varying averaging factors and step sizes.** To find a fixed point of a nonexpansive mapping $F$, one can use the iteration

$$x^{k+1} = (1-\theta^k)x^k + \theta^k F(x^k),$$

where $\theta^k$ varies each iteration. Likewise, to find a zero of a maximal monotone operator $A$, one can use the iteration

$$x^{k+1} = (I + \alpha^k A)^{-1}(x^k),$$

where $\alpha^k$ varies each iteration. One could analyze these approaches by making a few modification to our proofs. Alternatively, one can view these as applications of different (but related) operators with common fixed points [117].

**Preconditioning.** The convergence speed of all of the algorithms developed in this paper can be improved by transforming the variables with an appropriate linear operator, called a preconditioner. For example, the optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b, \end{array}$$

where $f$ is a function on $\mathbf{R}^n$, $x \in \mathbf{R}^n$ is the optimization variable, $A \in \mathbf{R}^{m\times n}$, and $b \in \mathbf{R}^m$, is equivalent to the problem

$$\begin{array}{ll} \text{minimize} & f(Ex) \\ \text{subject to} & AEx = b, \end{array}$$

where $E \in \mathbf{R}^{n\times n}$ is nonsingular. Choosing $E$ well and applying the algorithm to the transformed problem can improve the performance [51, 58, 59, 96, 97, 105].

**Cutting plane methods.** If $F$ is a monotone operator $0 \in F(x^\star)$, then for any $y \in F(x)$, we have

$$y^T x^\star \geq y^T x.$$

In other words, every evaluation of $F$ gives a cutting plane for $x^\star$, which eliminates a half-space from our search for $x^\star$. By judiciously choosing points to evaluate and accumulating the cutting planes, one can localize a small set in which $x^\star$ must lie in [29, 65, 68, 73, 78, 91, 92, 95, 107, 108].

**Hilbert spaces.** Often the theory of monotone operators is developed in the broader setting of Hilbert spaces, and a new set of challenges arise in these infinite dimensional settings. For example, all fixed points iterations we discussed only converge to a solution weakly unless we make additional assumptions. When an iteration is a strict contraction, we get strong convergence [3, 7, 33, 34, 36, 44, 103].

**Variational inequalities.** Given a set $C \subseteq \mathbf{R}^n$ and a function $F : C \to \mathbf{R}^n$, a variational inequality problem is to find a solution $x^\star$ that satisfies

$$(y - x^\star)^T F(x^\star) \geq 0$$

for all $y \in C$. When $C$ is convex and $F$ is monotone, this problem reduces to solving $0 \in F(x) + N_C(x)$. However, some interesting problems can be posed with variational inequalities but not with monotone operators [49, 60, 70, 75, 122].

**Partial inverse.** Let an operator $T$ on $\mathbf{R}^n$ be monotone, $A$ a subspace of $\mathbf{R}^n$, and $A^\perp$ the orthogonal complement of $A$. Then we call

$$T_A = \{(\Pi_A x + \Pi_{A^\perp} y, \Pi_A y + \Pi_{A^\perp} x) \,|\, (x, y) \in T\},$$

also a monotone operator, the partial inverse of $T$ with respect to $A$. If $A = \{0\}$ then $T_A = T^{-1}$, and if $A = \mathbf{R}^n$ then $T_A = T$; hence the name. The partial inverse is not only central in the study of dualities of monotone operators [102, 109] (another topic we missed) but is also useful in generating many optimization algorithms [123, 124].

**Existence of solutions.** In §5.1, we did prove a contraction mapping has a fixed point, but for the most part we assumed a solution exists and focused on finding it. It is, however, possible to prove the existence of solutions in more general settings, and this is one of the main uses of monotone operator theory in differential equations [52, 121].

## 9. Appendix

In this section, we discuss the equivalent definitions of strong convexity and strong smoothness. Interested readers can refer to [7, 15, 72, 93, 94, 119] for further details.

**Strong convexity.** A CCP function $f$ on $\mathbf{R}^n$ is strongly convex with parameter $m$ if any of the following equivalent conditions are satisfied:

(1) $f(x) - m/2\|x\|_2^2$ is convex.
(2) $f(x) - m/2\|x - x_0\|_2^2$ is convex for all $x_0 \in \mathbf{R}^n$.
(3) $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y) - \theta(1-\theta)m/2\|x - y\|_2^2$ for all $x, y \in \mathbf{R}^n$ and $\theta \in [0, 1]$.
(4) $f(y) \geq f(x) + \partial f(x)^T (y - x) + m/2\|y - x\|_2^2$ for all $x, y \in \mathbf{R}^n$.
(5) $\partial f$ is strongly monotone with parameter $m$, *i.e.*, $(\partial f(x) - \partial f(y))^T (x - y) \geq m\|x - y\|_2^2$ for all $x, y \in \mathbf{R}^n$.
(6) $\nabla^2 f(x) \succeq mI$ for all $x \in \mathbf{R}^n$, if $f$ is twice continuously differentiable.

Let's prove this. Conditions (1) and (2) are equivalent as the two functions only differ by an affine function. Equivalence of conditions (1) and (3) follow simply from algebra. Equivalence of conditions (1) and (6) follow from the fact that twice continuously differentiable functions are convex if and only if its Hessian is positive semidefinite everywhere.

Assume condition (1) and (3). For any $\varepsilon > 0$ and $x, y \in \mathbf{R}^n$, the definition of subdifferentials gives us

$$\begin{aligned} f(y) + \varepsilon \partial f(y)^T (x - y) = f(y) + \partial f(y)^T (\varepsilon x + (1 - \varepsilon y) - y) \leq \\ \leq f(\varepsilon x + (1 - \varepsilon) y). \end{aligned}$$

Now we divide by $\varepsilon$ and apply condition (3) to get

$$\begin{aligned} \partial f(y)^T (x - y) &\leq \frac{1}{\varepsilon} f(\varepsilon x + (1-\varepsilon) y) - \frac{1}{\varepsilon} f(y) \leq \\ &\leq f(x) - f(y) - (1 - \varepsilon)\frac{m}{2}\|x - y\|_2^2. \end{aligned}$$

Finally we take the limit $\varepsilon \to 0^+$ to get condition (4).

Assuming condition (4), we have

$$f(x) \geq f(y) + \partial f(y)^T(x-y) + \frac{m}{2}\|x-y\|_2^2$$
$$f(y) \geq f(x) + \partial f(x)^T(y-x) + \frac{m}{2}\|y-x\|_2^2.$$

Adding these two lines we get

$$(\partial f(x) - \partial f(y))^T(x-y) \geq m\|x-y\|_2^2,$$

and we conclude condition (5).

Assume condition (5), and assume $\mathbf{dom}\, f$ is not a singleton as otherwise condition (1) holds trivially. Consider two points $x, y \in \mathbf{relint}\,\mathbf{dom}\, f$ and the function $g(\theta) = f(\theta x + (1-\theta)y)$ for $\theta \in [0,1]$.

Then we have

$$(\theta_2 - \theta_1)m\|x-y\|_2^2 \leq (\partial f(\theta_2 x + (1-\theta_2)y) - \partial f(\theta_1 x + (1-\theta_1)y))^T(x-y) =$$
$$= (\partial g(\theta_2) - \partial g(\theta_1))$$

for all $\theta_1, \theta_2 \in [0,1]$ and $\theta_2 \geq \theta_1$ [111, Theorem 23.9]. Note that $g'_+(\theta) \subseteq \partial g(\theta)$, where $g'_+$ is the right derivative of $g$ [111, p. 299]. So

$$g'_+(\theta_2) - g'_+(\theta_1) \geq (\theta_2 - \theta_1)m\|x-y\|_2^2. \tag{12}$$

We integrate (12) with respect to $\theta_2$ on $[\theta, 1]$ and let $\theta_1 = \theta$ to get

$$g(1) \geq g(\theta) + g'_+(\theta)(1-\theta) + (1-\theta)^2\frac{m}{2}\|x-y\|_2^2. \tag{13}$$

(This is justified by Corollary 24.2.1 of [111].) Likewise, we can integrate (12) with respect to $\theta_1$ on $[0, \theta]$ and set $\theta_2 = \theta$ to get

$$g(0) \geq g(\theta) - g'_+(\theta)\theta + \theta^2\frac{m}{2}\|x-y\|_2^2. \tag{14}$$

We multiply the (13) by $\theta$ and (14) by $1-\theta$ and add the results to get

$$\theta g(1) + (1-\theta)g(0) \geq g(\theta) + \theta(1-\theta)\frac{m}{2}\|x-y\|_2^2.$$

Finally, plugging $f$ back in gives condition (3). So condition (3) holds for any $x, y \in \mathbf{relint}\,\mathbf{dom}\, f$.

Finally, we extend this result to all of $\mathbf{dom}\, f$. Consider any two points $x, y \in \mathbf{dom}\, f$. Pick an arbitrary point $z \in \mathbf{relint}\,\mathbf{dom}\, f$ (which exists by [111, Theorem 6.2]) and define

$$x_\varepsilon = (1-\varepsilon)x + \varepsilon z$$
$$y_\varepsilon = (1-\varepsilon)y + \varepsilon z.$$

Clearly, we have $x_\varepsilon \to x$ as $\varepsilon \to 0$ and $x_\varepsilon \in \mathbf{relint}\,\mathbf{dom}\, f$ for $\varepsilon \in (0,1]$ [111, Theorem 6.1] and the same can be said for $y_\varepsilon$. So the condition (3) holds for $x_\varepsilon$ and $y_\varepsilon$ for $\varepsilon \in (0,1]$. We take the limit $\varepsilon \to 0^+$ and conclude condition (3) for $x$ and $y$ (where we use continuity of CCP functions restricted to a line [111, Corollary 7.5.1]).

**Strong smoothness.** A CCP function $f$ on $\mathbf{R}^n$ is strongly smooth with parameter $L$ if any of the following equivalent conditions are satisfied:

(1) $f(x) - L/2\|x\|_2^2$ is concave.
(2) $f(x) - L/2\|x - x_0\|_2^2$ is concave for all $x_0 \in \mathbf{R}^n$.
(3) $f(\theta x + (1-\theta)y) \geq \theta f(x) + (1-\theta)f(y) - \theta(1-\theta)L/2\|x-y\|_2^2$ for all $x, y \in \mathbf{R}^n$ and $\theta \in [0,1]$.
(4) $f$ is differentiable and $f(y) \leq f(x) + \nabla f(x)^T(y-x) + L/2\|x-y\|_2^2$ for all $x, y \in \mathbf{R}^n$.
(5) $f$ is differentiable and $(\nabla f(x) - \nabla f(y))^T(x-y) \leq L\|x-y\|_2^2$ for all $x, y \in \mathbf{R}^n$.
(6) $\partial f$ is Lipschitz with parameter $L$.
(7) $f$ is differentiable and $\nabla f$ is Lipschitz with parameter $L$.

(8) $f$ is differentiable and $1/L\|\nabla f(x)-\nabla f(y)\|_2^2 \leq (\nabla f(x)-\nabla f(y))^T(x-y)$ for all $x, y \in \mathbf{R}^n$.
(9) $\nabla^2 f \preceq LI$ for all $x \in \mathbf{R}^n$, if $f$ is twice continuously differentiable.

Parts of this equivalence can be found in [119, Proposition 12.60], [94, Theorem 2.1.5], and [15, p. 433]. The full set of equivalence follows from [7, Theorem 18.15].

**Duality of strong monotonicity and strong smoothness.** Condition (8) of strong smoothness is called cocoercivity or inverse strong monotonicity. For general monotone operators, cocoercivity is by definition the dual property of strong monotonicity; a monotone operator $F$ is cocoercive if and only if $F^{-1}$ is strongly monotone. In general, cocoercivity is a stronger assumption than Lipschitz continuity, *i.e.*, strong monotonicity and cocoercivity are not dual properties in general.

For subdifferential operators of CCP functions, however, cocoercivity and Lipschitz continuity are equivalent, *i.e.*, conditions (6) and (8) are equivalent. This result is referred to as the Baillon-Haddad theorem [5].

Let $f$ is a CCP function. Then using the identity $(\partial f)^{-1} = \partial f^*$, we see that $\partial f$ is strongly monotone with parameter $m$ (*i.e.*, satisfies condition (5) of strong convexity) if and only if $\partial f^\star$ satisfies condition (8) of strong smoothness with $L = 1/m$. So $f$ is strongly convex with parameter $m$ if and only if $f^\star$ is strongly smooth with parameter $L = 1/m$.

## Acknowledgements

We thank Neal Parikh for his work on an earlier version of this paper, used as class notes for EE364b at Stanford. We also thank Shuvomoy Das Gupta, Pontus Giselsson, and Heinz Bauschke for very helpful feedback. We are especially grateful to Patrick Combettes as his feedback improved this paper significantly. Stephen Boyd was partially supported by the DARPA X-DATA program. Ernest Ryu was supported by the Math+X fellowship from the Simons Foundation.

## References

[1] Apostol, T.M. *Calculus*, V.2, 2nd edition, 1969.
[2] Arrow, K.J., Hurwicz, L., Uzawa, H. *Studies in Linear and Non-Linear Programming*, 1972.
[3] Artacho, F.J.A., Borwein, J.M., Martín-Márquez, V., Yao, L. Applications of convex analysis within mathematics, *Math. Program., Ser. B*, V.148, N.1–2, 2014, pp.49-88.
[4] Auslender, A., Teboulle, M. *Asymptotic Cones and Functions in Optimization and Variational Inequalities*, 2003.
[5] Baillon, J.B., Haddad, G. Quelques propriétés des opérateurs angle-bornés et $n$-cycliquement monotones, *Isr. J. Math.*, V.26, N.2, 1977, pp.137-150.
[6] Banach, S. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales, *Fund. Math.*, V.3, N.1, 1922, pp.133-181.
[7] Bauschke, H., Combettes, P. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2011.
[8] Bauschke, H.H., Borwein, J.M. Dykstra's alternating projection algorithm for two sets, *J. Approx. Theory*, V.79, N.3, 1994, pp.418-443.
[9] Bauschke, H.H., Borwein, J.M. On projection algorithms for solving convex feasibility problems, *SIAM Rev.*, V.38, N.3, 1996, pp.367-426.
[10] Bauschke, H.H., Borwein, J.M., Lewis, A.S. On the method of cyclic projections for convex sets in Hilbert space, *Contemp. Math.*, V.204, 1997.
[11] Bauschke, H.H., Koch, V.R. Projection methods: Swiss Army knives for solving feasibility and best approximation problems with halfspaces. In: *Infinite Products and Their Applications*, *Contemporary Mathematics*, V.636, pp.1-40.
[12] Beck, A., Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM J. on Imaging Sci.*, V.2, N.1, 2009.
[13] Ben-Tal, A., Nemirovski, A. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, 2001.
[14] Bertsekas, D.P. *Convex Optimization Theory*, 2009.
[15] Bertsekas, D.P. *Convex Optimization Algorithms*, 2015.
[16] Bertsekas, D.P., Nedić, A., Ozdaglar, A.E. *Convex analysis and optimization*, 2003.

[17] Borwein, J., Lewis, A. *Convex Analysis and Nonlinear Optimization: Theory and Examples*, 2006.
[18] Borwein, J., Vanderwerff, J. *Convex Functions: Constructions, Characterizations and Counterexamples*, 2010.
[19] Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers, *Found. Trends Mach. Learn.*, V.3, N.1, 2011, pp.1-122.
[20] Boyd, S., Vandenberghe, L. *Convex Optimization*, 2004.
[21] Boyle, J.P., Dykstra, R. A method for finding projections onto the intersection of convex sets in Hilbert spaces. In: *Advances in Order Restricted Statistical Inference*, *Lecture Notes in Statistics*, V.37, pp.28-47, 1986.
[22] Brezis, H., Lions, P.L. Produits infinis de resolvantes, *Isr. J. Math.*, V.29, N.4, 1978, pp.329-345.
[23] Briceño-Arias, L.M., Combettes, P.L. A monotone+skew splitting model for composite monotone inclusions in duality, *SIAM J. Optim.*, V.21, N.4, 2011, pp.1230-1250.
[24] Browder, F.E. Nonlinear elliptic boundary value problems, *Bull. Amer. Math. Soc.*, V.69, N.6, 1963, pp.862-874.
[25] Browder, F.E. The solvability of non-linear functional equations, *Duke Math. J.*, V.30, N.4, 1963, pp.557-566.
[26] Browder, F.E. Variational boundary value problems for quasi-linear elliptic equations of arbitrary order, V.50, N.1, 1963, pp.31-37.
[27] Cauchy, M.A. Méthode générale pour la résolution des systémes d'équations simultanées, *C. R. Hebd. Séances d'Acad. Sci.*, V.25, 1847, pp.536-538.
[28] Chambolle, A., Pock, T. A first-order primal-dual algorithm for convex problems with applications to imaging, *J. Math. Imaging Vis.*, V.40, N.1, 2011, pp.120-145.
[29] Cheney, E.W., Goldstein, A.A. Newton's method for convex programming and Tchebycheff approximation, *Numer. Math.*, V.1, N.1, 1959, pp.253-268.
[30] Cheney, E.W., Goldstein, A.A. Proximity maps for convex sets, *Proc. Amer. Math. Soc.*, V.10, N.3, 1959, pp.448-450.
[31] Chung, S.J. NP-completeness of the linear complementarity problem, *J. Optim. Theory Appl.*, V.60, N.3, 1989, pp.393-399.
[32] Cimmino, G. Calcolo approssimato per le soluzioni dei sistemi di equazioni lineari, *La Ric. Sci.*, V.9, 1938, pp.326-333.
[33] Combettes, P.L. Solving monotone inclusions via compositions of nonexpansive averaged operators, *Optimization*, V.53, N.5–6, 2004, pp.475-504.
[34] Combettes, P.L., Pesquet, J.C. Proximal splitting methods in signal processing. In: *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp.185-212, 2011.
[35] Combettes, P.L., Pesquet, J.C. Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators, *Set-Valued Var. Anal.*, V.20, N.2, 2012, pp.307-330.
[36] Combettes, P.L., Wajs, V.R. Signal recovery by proximal forward-backward splitting, *Multiscale Model. Simul.*, V.4, N.4, 2005, pp.1168-1200.
[37] Combettes, P.L., Yamada, I. Compositions and convex combinations of averaged nonexpansive operators, *J. Math. Anal. Appl.*, V.425, N.1, 2015, pp.55-70.
[38] Cottle, R., Pang, J., Stone, R. *The Linear Complementarity Problem*, 2009.
[39] Davis, D., Yin, W. Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions, CAM Report 14–58, UCLA, 2014.
[40] Davis, D., Yin, W. A three-operator splitting scheme and its optimization applications, 2015.
[41] Douglas, J., Rachford, H.H. On the numerical solution of heat conduction problems in two and three space variables, *Trans. Amer. Math. Soc.*, V.82, 1956, pp.421-439.
[42] Duchi, J., Singer, Y. Efficient online and batch learning using forward backward splitting, *J. Mach. Learn. Res.*, V.10, 2009, pp.2899-2934.
[43] Eckstein, J. The Lions-Mercier splitting algorithm and the alternating direction method are instances of the proximal point algorithm, Technical Report LIDS-P-1769, MIT, 1988.
[44] Eckstein, J. *Splitting methods for monotone operators with applications to parallel optimization*, Ph.D. thesis, MIT, 1989.
[45] Eckstein, J. Augmented Lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results, RUTCOR Research Report RRR, Rutgers University, 2012.
[46] Eckstein, J., Bertsekas, D.P. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators, *Math. Program.*, V.55, N.1–3, 1992.
[47] Escalante, R., Raydan, M. *Alternating Projection Methods*, 2011.
[48] Esser, E., Zhang, X., Chan, T.F. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science, *SIAM J. on Imaging Sci.*, V.3, N.4, 2010, pp.1015-1046.

[49] Facchinei, F., Pang, J. *Finite-Dimensional Variational Inequalities and Complementarity Problems*, 2003.

[50] Ferris, M.C., Pang, J.S. Engineering and economic applications of complementarity problems, *SIAM Rev.*, V.39, N.4, 1997, pp.669-713.

[51] Fougner, C., Boyd, S. Parameter selection and pre-conditioning for a graph form solver, 2015.

[52] Francaů, J. Monotone operators: A survey directed to applications to differential equations, *Apl. Mat.*, V.35, N.4, 1990, pp.257-301.

[53] Gabay, D. Applications of the method of multipliers to variational inequalities. In: *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, 1983.

[54] Gabay, D., Mercier, B. A dual algorithm for the solution of nonlinear variational problems via finite element approximation, *Comput. Math. Appl.*, V.2, N.1, 1976, pp.17-40.

[55] Ghosh, A., Boyd, S. Minimax and convex-concave games, 2003. Stanford University EE392o Lecture Notes.

[56] Gill, P.E., Saunders, M.A., Shinnerl, J.R. On the stability of Cholesky factorization for symmetric quasidefinite systems, *SIAM J. Matrix Anal. Appl.*, V.17, N.1, 1996, pp.35-46.

[57] Giselsson, P. Tight global linear convergence rate bounds for Douglas-Rachford splitting, 2015.

[58] Giselsson, P., Boyd, S. Metric selection in Douglas-Rachford splitting and ADMM, 2014.

[59] Giselsson, P., Boyd, S. Metric selection in fast dual forward backward splitting, *Automatica*, 2015.

[60] Glowinski, R., Lions, J.L., Trémolières, R. *Numerical Analysis of Variational Inequalities*, 1981.

[61] Glowinski, R., Marroco, A. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problmes de Dirichlet non linéaires, *Rev. Fr. d'Autom. Inform. Rech. Oper. Anal. Numer.*, V.9, N.2, 1975, pp.41-76.

[62] Goldstein, A.A. Convex programming in Hilbert space, *Bull. Amer. Math. Soc.*, V.70, N.5, 1964, pp.709-710.

[63] Golub, G.H., Van Loan, C.F. *Matrix Computations*, 4th edition, 2012.

[64] Golub, G.H., Wilkinson, J.H. Note on the iterative refinement of least squares solution, *Numer. Math.*, V.9, N.2, 1966, pp.139-148.

[65] Gomory, R.E. Outline of an algorithm for integer solutions to linear programs, *Bull. Amer. Math. Soc.*, V.64, N.5, 1958, pp.275-278.

[66] Gubin, L.G., Polyak, B.T., Raik, E.V. The method of projections for finding the common point of convex sets, *Zh. Vychisl. Mat. Mat. Fiz.*, V.7, N.6, 1967, pp.1-24.

[67] Hestenes, M.R. Multiplier and gradient methods, *J. Optim. Theory Appl.*, V.4, N.5, 1969, pp.303-320.

[68] Hiriart-Urruty, J.B., Lemaréchal, C. *Convex Analysis and Minimization Algorithms*, V.2, 1993.

[69] Horn, R.A., Johnson, C.R. *Matrix Analysis*, 1985.

[70] Huang, Z.Y., Noor, M.A. Explicit parallel resolvent methods for system of general variational inclusions, *TWMS J. Pure Apple. Math.*, V.4, N.2, 2013, pp.159-168.

[71] Kachurovskii, R.I. Monotone operators and convex functionals, *Usp. Mat. Nauk*, V.15, N.4, 1960, pp.213-215.

[72] Kakade, S.M., Shalev-Shwartz, S., Tewari, A. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization, Technical report, Toyota Technological Institute, 2009.

[73] Kelley, J.E. The cutting-plane method for solving convex programs, *J. Soc. Ind. Appl. Math.*, V.8, N.4, 1960, pp.703-712.

[74] Kellogg, R.B. A nonlinear alternating direction method, *Math. Comput.*, V.23, N.105, 1969, pp.23-27.

[75] Kinderlehrer, D., Stampacchia, G. *An Introduction to Variational Inequalities and Their Applications*, 2000.

[76] Korpelevich, G.M. The extragradient method for finding saddle points and other problems, V.12, 1976, pp.747-756.

[77] Krasnosel'skii, M.A. Two remarks on the method of successive approximations, *Usp. Mat. Nauk*, V.10, N.1, 1955, pp.123-127.

[78] Levin, A.Y. On an algorithm for the minimization of convex function, *Dokl. Akad. Nauk SSSR*, V.160, N.6, 1965, pp.1244-1247.

[79] Levitin, E.S., Polyak, B.T. Constrained minimization methods, *Zh. Vychisl. Mat. Mat. Fiz.*, V.6, N.5, 1966, pp.787-823.

[80] Lindelöf, E. Sur l'applications de la méthode des approximations successives aux équations différentielles ordinaires du premier ordre, *C. R. Hebd. Séances d'Acad. Sci.*, V.118, 1894, pp.454-456.

[81] Lions, P.L., Mercier, B. Splitting algorithms for the sum of two nonlinear operators, *SIAM J. Numer. Anal.*, V.16, N.6, 1979, pp.964-979.

[82] Luenberger, D.G., Ye, Y. *Linear and Nonlinear Programming*, 2008.

[83] Mann, W.R. Mean value methods in iteration, *Proc. Amer. Math. Soc.*, V.4, N.3, 1953, pp.506-510.

[84] Martin, R.S., Peters, G., Wilkinson, J.H. Iterative refinement of the solution of a positive definite system of equations, *Numer. Math.*, V.8, N.3, 1966, pp.203-216.

[85] Martinet, B. Régularisation d'inéquations variationnelles par approximations successives, *Rev. Fr. d'Inform. Rech. Oper. Sér. Rouge*, V.4, N.3, 1970, pp.154-158.

[86] Martinet, B. Determination approchée d'un point fixe d'une application pseudo-contractante, *C. R. l'Acad. Sci. Sér. A*, V.274, 1972, pp.163-165.
[87] Minty, G.J. Monotone networks, *Proc. R. Soc. Lond. A: Math. Phys. Eng. Sci*, V.257, N.1289, 1960, pp.194-212.
[88] Minty, G.J. On the maximal domain of a "monotone" function., *Mich. Math. J.*, V.8, N.2, 1961, pp.135-137.
[89] Minty, G.J. Monotone (nonlinear) operators in Hilbert space, *Duke Math. J.*, V.29, N.3, 1962, pp.341-346.
[90] Minty, G.J. On the monotonicity of the gradient of a convex function., *Pac. J. Math.*, V.14, N.1, 1964, pp.243-247.
[91] Mitchell, J.E. Cutting plane methods and subgradient methods. In: *Tutorials in Operations Research: Decision Technologies and Applications*, pp.34-61, 2009.
[92] Nemirovski, A. Efficient methods in convex programming, 1995. Lecture notes.
[93] Nemirovski, A.S., Yudin, D.B. *Problem Complexity and Method Efficiency in Optimization*, 1983.
[94] Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*, 2004.
[95] Newman, D.J. Location of the maximum on unimodal surfaces, *J. Assoc. Comput. Mach.*, V.12, N.3, 1965, pp.395-398.
[96] O'Donoghue, B., Chu, E., Parikh, N., Boyd, S. Operator splitting for conic optimization via homogeneous self-dual embedding, 2013.
[97] Osborne, E.E. On pre-conditioning of matrices, *J. Assoc. Comput. Mach.*, V.7, N.4, 1960, pp.338-345.
[98] Parikh, N., Boyd, S. Block splitting for distributed optimization, *Math. Program. Comput.*, V.6, N.1, 2014, pp.77-102.
[99] Parikh, N., Boyd, S. Proximal algorithms, *Found. Trends Optim.*, V.1, N.3, 2014, pp.127-239.
[100] Passty, G.B. Ergodic convergence to a zero of the sum of monotone operators in Hilbert space, *J. Math. Anal. Appl.*, V.72, N.2, 1979, pp.383-390.
[101] Peaceman, D.W., Rachford, H.H. The numerical solution of parabolic and elliptic differential equations, *J. Soc. Ind. Appl. Math.*, V.3, N.1, 1955.
[102] Pennanen, T. Dualization of generalized equations of maximal monotone type, *SIAM J. Optim.*, V.10, N.3, 2000, pp.809-835.
[103] Phelps, R.R. *Convex Functions, Monotone Operators and Differentiability*, 2nd edition, 1993.
[104] Picard, E. Mémoire sur la théorie des équations aux dérivées partielles et la méthode des approximations successives, *J. Math. Pures Appl. 4éme Sér.*, V.6, 1890, pp.145-210.
[105] Pock, T., Chambolle, A. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In: *Proceedings of the 2011 International Conference on Computer Vision*, pp.1762-1769, 2011.
[106] Powell, M.J.D. A method for nonlinear constraints in minimization problems. In: *Optimization: Symposium of the Institute of Mathematics and Its Applications, University of Keele, England, 1968*, pp.283-298, 1969.
[107] Remez, E.Y. Sur le calcul effectif des polynomes d'approximation de Tchebychef, *C. R. l'Acad. Sci.*, V.199, 1934, pp.337-340.
[108] Remez, E.Y. Sur un procédé convergent d'approximations successives pour déterminer les polynomes d'approximation, *C. R. l'Acad. Sci.*, V.198, 1934, pp.2063-2065.
[109] Robinson, S.M. Composition duality and maximal monotonicity, *Math. Program.*, V.85, N.1, 1999, pp.1-13.
[110] Rockafellar, R.T. Characterization of the subdifferentials of convex functions., *Pac. J. Math.*, V.17, N.3, 1966, pp.497-510.
[111] Rockafellar, R.T. *Convex Analysis*, 1970.
[112] Rockafellar, R.T. Monotone operators associated with saddle-functions and minimax problems. In: *Nonlinear Functional Analysis, Part 1*, *Proceedings of Symposia in Pure Mathematics*, V.18, pp.241-250, 1970.
[113] Rockafellar, R.T. On the maximal monotonicity of subdifferential mappings., *Pac. J. Math.*, V.33, N.1, 1970, pp.209-216.
[114] Rockafellar, R.T. On the maximality of sums of nonlinear monotone operators, *Trans. Am. Math. Soc.*, V.140, 1970, pp.75-88.
[115] Rockafellar, R.T. The multiplier method of Hestenes and Powell applied to convex programming, *J. Optim. Theory Appl.*, V.12, N.6, 1973, pp.555-562.
[116] Rockafellar, R.T. Augmented Lagrangians and applications of the proximal point algorithm in convex programming, *Math. Oper. Res.*, V.1, N.2, 1976, pp.97-116.
[117] Rockafellar, R.T. Monotone operators and the proximal point algorithm, *SIAM J. Control Optim.*, V.14, N.5, 1976, pp.877-898.
[118] Rockafellar, R.T. Monotone operators and augmented Lagrangian methods in nonlinear programming. In: *Nonlinear Programming 3*, pp.1-25, 1978.
[119] Rockafellar, R.T., Wets, R.J.B. *Variational Analysis*, 1998.
[120] Ruszczyński, A. *Nonlinear Optimization*, 2006.
[121] Showalter, R.E. *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations*, 1997.

[122] Sofonea, M., Matei, A. *Variational Inequalities with Applications: A Study of Antiplane Frictional Contact Problems*, 2009.
[123] Spingarn, J.E. Partial inverse of a monotone operator, *Appl. Math. Optim.*, V.10, N.1, 1983, pp.247-265.
[124] Spingarn, J.E. Applications of the method of partial inverses to convex programming: Decomposition, *Math. Program.*, V.32, N.2, 1985, pp.199-223.
[125] Tseng, P. Dual ascent methods for problems with strictly convex costs and linear constraints: A unified approach, *SIAM J. Control Optim.*, V.28, N.1, 1990, pp.214-242.
[126] Tseng, P. A modified forward-backward splitting method for maximal monotone mappings, *SIAM J. Control Optim.*, V.38, N.2, 2000, pp.431-446.
[127] Tseng, P., Bertsekas, D.P. Relaxation methods for problems with strictly convex separable costs and linear constraints, *Math. Program.*, V.38, N.3, 1987, pp.303-321.
[128] Vanderbei, R.J. Symmetric quasidefinite matrices, *SIAM J. Optim.*, V.5, N.1, 1995, pp.100-113.
[129] von Neumann, J. *Functional Operators. Volume II. The Geometry of Orthogonal Spaces*, 1950.
[130] Wilkinson, J.H. *Rounding Errors in Algebraic Processes*, 1963.
[131] Yan, M., Yin, W. Self equivalence of the alternating direction method of multipliers, 2014.
[132] Zames, G. On the input-output stability of time-varying nonlinear feedback systems part one: Conditions derived using concepts of loop gain, conicity, and positivity, *IEEE Trans. Autom. Control*, V.11, N.2, 1966, pp.228-238.
[133] Zhu, M., Chan, T.F. An efficient primal-dual hybrid gradient algorithm for total variation image restoration, CAM Report 08–34, UCLA, 2008.

**Ernest K. Ryu** - Ph.D. candidate at the Institute for Computational and Mathematical Engineering at Stanford University. His current research focus is on convex optimization, stochastic optimization, scientific computing, and monotone operators.

**Stephen Boyd** - Samsung Professor of Engineering, and Professor of Electrical Engineering in the Information Systems Laboratory at Stanford University. He has courtesy appointments in the Department of Management Science and Engineering and the Department of Computer Science, and is a member of the Institute for Computational and Mathematical Engineering. His current research focus is on convex optimization applications in control, signal processing, and circuit design.