A "crash course" in first-order continuous optimization

Zev Woodstock
woodstzc(at)jmu.edu*

1 Introduction

These notes are supplementary material to a "crash course" I am teaching in Spring of 2026. The topic is *continuous optimization*, with an emphasis on constrained, first-order optimization algorithms. These notes are not meant to be used as a standalone resource. Please cite peer-reviewed material. The reference book for this class is *Convex Analysis and Monotone Operator Theory*, 2nd edition, by Heinz H. Bauschke and Patrick L. Combettes, published by Springer.

If unspecified, $\mathcal{H}$ is a real finite-dimensional vector space in Section 1 and a real finite-dimensional Hilbert space from Sections 2 onward (e.g., $\mathbb{R}^n$ with the Euclidean inner product is fine). While this class sticks to finite dimensions, virtually all of these results also apply to real (infinite-dimensional) Hilbert spaces, modulo minor adjustments detailed in the class book.

1.1 Optimization terminology and the extended real line

Notation 1.1 We will work with the **extended real line**, i.e., $[-\infty, +\infty] := \mathbb{R} \cup \{-\infty, +\infty\}$. Algebra on this field follows most "natural" rules one could expect (e.g., for $x \in \mathbb{R}$, $x + \infty = \infty$). However, the following quantities are **undefined**:

- Any subtraction of infinities: "$+\infty - (+\infty)$"
- Zero times infinity: "$0 \cdot (\pm\infty)$"
- Any quotient of infinities: "$\pm\infty / \pm\infty$, $\pm\infty / \mp\infty$, ..."

As a result, if we work with extended-real-valued functions, we must be sure to avoid anything which is undefined (e.g., the objective function $f(x) + g(x)$ could be undefined if there exists z such that $g(z) = -\infty$ and $f(z) = \infty$.)

*Please report typos/errors found in these notes. Homework solutions should be handed in during our lab meetings, or to the dropbox outside of my office (Roop 310).

Definition 1.2 Given a real vector space $\mathcal{H}$, a function $f\colon \mathcal{H} \to [-\infty, +\infty]$, and a set $C \subset \mathcal{H}$, consider the following optimization problem.

$$\underset{x \in C}{\text{minimize}} \;\; f(x) \tag{1}$$

We call f the **objective function**. We call C a **constraint**. For any $x \in C$, we say x is **feasible**. Otherwise, for $x \in \mathbb{R}^n \setminus C$, x is infeasible. If a point $x^* \in C$ satisfies

$$(\forall x \in C) \quad f(x^*) \leq f(x), \tag{2}$$

we call x^* a **solution** to the optimization problem (1).

For this class, we consider minimization; to maximize f, just use the objective function $-f$.

Definition 1.3 For $I \subset [-\infty, +\infty]$, $a \in [-\infty, +\infty]$ is a **lower bound** (**upper bound**) if, for every $\xi \in I$, $a \leq \xi$ ($a \geq \xi$). The **greatest lower bound**, or **infimum**, of the set I is denoted $\inf I$. Analogously, the **least upper bound**, or **supremum**, of the set I is denoted $\sup I$. In general, $\inf I, \sup I \in [-\infty, +\infty]$. If, additionally, $\inf I \in I$ ($\sup I \in I$), we call it the **minimum** (**maximum**), and denote it $\min I$ ($\max I$). In these cases, we say the infimum (supremum) is *attained*.

A few things to mention:

(i) For $I \neq \varnothing$, we have $\inf I \leq \sup I$. For the empty set, $\inf \varnothing = +\infty$ and $\sup \varnothing = -\infty$.

(ii) While the $\inf$ and $\sup$ are always defined, $\max$ and $\min$ may not exist (e.g., consider $I = (0, 1)$ has $\inf I = 0$ and $\sup I = 1$. However, since $0, 1 \notin I$, neither $\max I$ nor $\min I$ exist.)

(iii) Let $f\colon \mathbb{R}^n \to [-\infty, +\infty]$. We adopt the notation that $\inf_{x \in C} f(x) = \inf\{f(x) \,|\, x \in C\}$.

(iv) It is common in optimization literature to abuse notation, and use

$$\min_{x \in C} f(x) \tag{3}$$

to describe the optimization problem (1). Technically, $\min_{x \in C} f(x)$ is not an optimization problem – it is the optimal value of the objective function at a solution, which may or may not exist.

The following theorem is often used as a tool to ensure that a solution to an optimization problem exists. Regretfully, this class does not have enough time to detail the topics of compact/closed/lsc. However, since the following theorem is referenced a few times in the class, I will provide its statement here.[1]

Theorem 1.4 (Weierstraß) *Let $f\colon \mathcal{H} \to [-\infty, +\infty]$ be lower semicontinuous and let C be a compact subset of $\mathcal{H}$. Suppose that $C \cap \operatorname{dom} f \neq \varnothing$. Then f achieves its infimum over C.*

[1] Write me if you are interested in learning more about existence of solutions to optimization problems! For unbounded problems, analytic notions of "coercivity" and "recession cones" can also yield existence results.

Definition 1.5 Let $f\colon \mathcal{H} \to [-\infty, +\infty]$. We will use the following terms.

(i) The **domain** of f is

$$\operatorname{dom} f = \big\{x \in \mathcal{H} \;\big|\; f(x) < +\infty\big\} \tag{4}$$

(ii) The **epigraph** of f is

$$\operatorname{epi} f = \big\{(x, \xi) \in \mathcal{H} \times \mathbb{R} \;\big|\; f(x) \leq \xi\big\} \tag{5}$$

(iii) The function f is **proper** if $\operatorname{dom} f \neq \varnothing$ and it never outputs the value $-\infty$ (i.e., $-\infty \notin f(\mathcal{H})$).

(iv) The function f is **lower semicontinuous** (sometimes abbreviated "lsc") at $x \in \mathcal{H}$ if, for every sequence $(x_n)_{n\in\mathbb{N}}$ satisfying $x_n \to x$, we have $f(x) \leq \liminf f(x_n)$

(v) The **argmin** of f over C is

$$\operatorname*{Argmin}_{x\in C} f(x) = \left\{x \in C \;\middle|\; f(x) = \min_{z\in C} f(z)\right\}; \tag{6}$$

If the set under an argmin is not mentioned, then usually we assume $C = \mathcal{H}$.

For this class, we will predominantly consider proper and lsc functions. A few things to note about the lsc assumption: (1) every continuous function is lsc, and (2) lower semicontinuity basically allows for a jump-discontinuity to occur at $x \in \mathcal{H}$, but requires that f takes the lowest possible limiting value at x (cf. the figures drawn in class, or here[2]).

1.2 Inner product and norms

Definition 1.6 Let $\mathcal{H}$ be a real finite-dimensional vector space. A **scalar product** (sometimes called **inner product**) is a function $\langle \cdot \mid \cdot \rangle\colon \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ which satisfies the following properties.

(i) $(\forall x \in \mathcal{H} \setminus \{0\}) \quad \langle x \mid x\rangle > 0$

(ii) $(\forall x, y \in \mathcal{H}) \quad \langle x \mid y\rangle = \langle y \mid x\rangle$

(iii) $(\forall x, y, z \in \mathcal{H})(\forall \alpha \in \mathbb{R}) \quad \langle \alpha x + y \mid z\rangle = \alpha\langle x \mid z\rangle + \langle y \mid z\rangle$

Exercise 1.7 Let $\mathbf{0} \in \mathcal{H}$ be the zero element of $\mathcal{H}$. Prove that, for every $x \in \mathcal{H}$, $\langle \mathbf{0} \mid x\rangle = 0$.

Exercise 1.8 Consider $\mathcal{H} = \mathbb{R}^n$. For two vectors $x, y \in \mathbb{R}^n$, the *dot product* is given by $\langle x \mid y\rangle = x^\top y$. Prove that the dot product on $\mathbb{R}^n$ is a scalar product.

[2] https://en.wikipedia.org/wiki/Semi-continuity#/media/File:Lower_semi.svg

Exercise 1.9 Consider the vector space of matrices $\mathbb{R}^{n\times n}$. For two matrices $A = (a_{i,j})_{1\leq i,j\leq n}$ and $B = (b_{i,j})_{1\leq i,j\leq n}$, the *Frobenius inner product* is given by

$$\langle A \mid B\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}b_{i,j} \tag{7}$$

Show (7) is an inner product.

Proposition 1.10 (Cauchy-Schwarz) *For every $x, y \in \mathcal{H}$,*

$$\langle x \mid y\rangle^2 \leq \langle x \mid x\rangle\langle y \mid y\rangle. \tag{8}$$

Proof. If $y = 0$, (8) holds. Now suppose that $y \neq 0$. By Definition 1.6, $\langle y \mid y\rangle > 0$. Set $\alpha = \langle x \mid y\rangle/\langle y \mid y\rangle$. First, we find

$$\begin{aligned} 0 &\leq \langle x - \alpha y \mid x - \alpha y\rangle && (9)\\ &= \langle x \mid x\rangle - 2\alpha\langle x \mid y\rangle + \alpha^2\langle y \mid y\rangle && (10)\\ &= \langle x \mid x\rangle - 2\alpha\langle x \mid y\rangle + \alpha\langle x \mid y\rangle && (11)\\ &= \langle x \mid x\rangle - \alpha\langle x \mid y\rangle. && (12)\end{aligned}$$

Rearranging the inequality, we find that

$$\begin{aligned} &\frac{\langle x \mid y\rangle^2}{\langle y \mid y\rangle} = \alpha\langle x \mid y\rangle \leq \langle x \mid x\rangle && (13)\\ \Leftrightarrow\quad &\langle x \mid y\rangle^2 \leq \langle y \mid y\rangle\langle x \mid x\rangle. && (14)\end{aligned}$$

□

Definition 1.11 Let $\mathcal{H}$ be a real finite-dimensional vector space. A function $\|\cdot\|\colon \mathcal{H} \to \mathbb{R}$ is a **norm** if the following hold.

(i) $(\forall x \in \mathcal{H}) \quad \|x\| = 0 \Rightarrow x = 0$

(ii) $(\forall x, y \in \mathcal{H}) \quad \|x + y\| \leq \|x\| + \|y\|$

(iii) $(\forall x \in \mathcal{H})(\forall \alpha \in \mathbb{R}) \quad \|\alpha x\| = |\alpha|\|x\|$

A norm is a way to measure magnitude of vectors, or the distance from one vector to another $\|x - y\|$.

Exercise 1.12 Let $\mathcal{H}$ be a real finite-dimensional vector space, and let $\langle\cdot \mid \cdot\rangle$ be a scalar product on $\mathcal{H}$. Show that the norm defined by

$$\|\cdot\|\colon \mathcal{H} \to \mathbb{R}\colon x \mapsto \sqrt{\langle x \mid x\rangle} \tag{15}$$

satisfies the properties in Definition 1.11.

The **Euclidean norm** on $\mathbb{R}^n$, given by $(\xi_1, \ldots, \xi_n) \mapsto \sqrt{\xi_1^2 + \cdots + \xi_n^2}$, arises from the dot product. Exercise 1.12 yields the following formulation of the Cauchy-Schwarz inequality

$$(\forall x, y \in \mathcal{H}) \quad \langle x \mid y \rangle \leq \|x\|\|y\|. \tag{C-S}$$

While the actual definition can get quite technical, for our class, when we say "**Hilbert space**", we are referring to the finite-dimensional vector space $\mathcal{H}$, equipped with a scalar product $\langle \cdot \mid \cdot \rangle$ and a norm who arises from the scalar product via $\|\cdot\| = \sqrt{\langle \cdot \mid \cdot \rangle}$. Some examples are $\mathbb{R}^n$ under the Euclidean inner product, or the space of real $n \times m$ matrices under the Frobenius inner product.

Exercise 1.13 Let $(x_1, x_2, x_3) \in \mathbb{R}^3$. Prove that

$$x_1^2 - 2x_2 + 8x_3 \leq 5\sqrt{x_1^4 + x_2^2 + 4x_3^2}. \tag{16}$$

Can the coefficient 5 in (16) be reduced?

Fact 1.14 For all $x, y \in \mathcal{H}$,

$$\|x + y\|^2 = \|x\|^2 + 2\langle x \mid y \rangle + \|y\|^2 \tag{17}$$

and

$$\|x - y\|^2 = \|x\|^2 - 2\langle x \mid y \rangle + \|y\|^2 \tag{18}$$

Proof. Since $\|x + y\| = \sqrt{\langle x + y \mid x + y \rangle}$, using Definition 1.6 provides

$$\begin{aligned} \|x + y\|^2 &= \langle x + y \mid x + y \rangle && (19)\\ &= \langle x \mid x + y \rangle + \langle y \mid x + y \rangle && (20)\\ &= \langle x \mid x \rangle + \langle x \mid y \rangle + \langle y \mid x \rangle + \langle y \mid y \rangle && (21)\\ &= \|x\|^2 + 2\langle x \mid y \rangle + \|y\|^2, && (22) \end{aligned}$$

which shows (17). To show (18), we can apply the same argument to x and $-y$, making use of the fact that $2\langle x \mid -y \rangle = -2\langle x \mid y \rangle$. □

Exercise 1.15 Let $x, y \in \mathcal{H}$. Prove the *parallelogram identity*:

$$\|x + y\|^2 + \|x - y\|^2 = 2\|x\| + 2\|y\|^2 \tag{23}$$

and the *polarization identity*

$$4\langle x \mid y \rangle = \|x + y\|^2 - \|x - y\|^2. \tag{24}$$

Remark 1.16 Regardless of the specific space $\mathcal{H}$, the inner product can be used to gain intuition about the *angle* between two vectors $x, y \in \mathcal{H}$ in the following sense.

- If $\langle x \mid y \rangle = 0$, then x and y form a *right* angle. In other words, x is *orthogonal* to y.
- If $\langle x \mid y \rangle > 0$, then x and y form an *acute* angle.
- If $\langle x \mid y \rangle < 0$, then x and y form an *obtuse* angle.

2 Convexity

Definition 2.1 A set $C \subset \mathcal{H}$ is **convex** if, for every $x, y \in C$

$$(\forall \alpha \in]0,1[) \quad \alpha x + (1-\alpha)y \in C. \tag{25}$$

A function f is **convex** if epi f is convex.

Proposition 2.2 *$f\colon \mathcal{H} \to [-\infty, +\infty]$ is convex if and only if*

$$(\forall x, y \in \operatorname{dom} f) \quad (\forall \alpha \in]0,1[) \qquad f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y). \tag{26}$$

Proof. First, we note that if f is identically $+\infty$, then dom $f = \varnothing$ if and only if epi $f = \varnothing$, so (26) is vacuously true. Now assume that dom $f \neq \varnothing$. Let (x, ξ) and (y, η) be in epi f and let $\alpha \in]0,1[$.
($\Rightarrow$) Assume that epi f is convex. Then

$$\alpha(x,\xi) + (1-\alpha)(y,\eta) = (\alpha x + (1-\alpha)y, \alpha\xi + (1-\alpha)\eta) \in \operatorname{epi} f. \tag{27}$$

Therefore, $f(\alpha x + (1-\alpha)y) \leq \alpha\xi + (1-\alpha)\eta$. Taking the limit as $\xi \searrow f(x)$ and $\eta \searrow f(y)$ yields (26).
($\Leftarrow$) Assume that (26) holds. By definition, $f(x) \leq \xi$ and $f(y) \leq \eta$. So, by (26),

$$\begin{aligned} f(\alpha x + (1-\alpha)y) &\leq \alpha f(x) + (1-\alpha) f(y) && (28)\\ &\leq \alpha\xi + (1-\alpha)\eta. && (29) \end{aligned}$$

Therefore, $(\alpha x + (1-\alpha)y, \alpha\xi + (1-\alpha)\eta) \in \operatorname{epi} f$ which completes the proof. □

Definition 2.3 Let $\rho > 0$ and let $x \in \mathcal{H}$. A **closed ball** of radius ρ is $B(x;\rho) = \big\{ z \in \mathcal{H} \;\big|\; \|x - z\| \leq \rho \big\}$.

Definition 2.4 Let $f\colon \mathcal{H} \to [-\infty, +\infty]$ and let $x \in \mathcal{H}$. x is a **local minimizer** of f if there exists $\rho > 0$ such that

$$(\forall z \in \mathcal{H} \cap B(x;\rho)) \quad f(x) \leq f(z). \tag{30}$$

x is a **global minimizer** of f if $x \in \operatorname*{Argmin}_{x \in \mathcal{H}} f(z)$, i.e.,

$$(\forall z \in \mathcal{H}) \quad f(x) \leq f(z). \tag{31}$$

If no prefix is used with the term 'minimizer', this usually refers to a global minimizer.

Exercise 2.5 Let f be a convex and proper function. Prove that every local minimizer is a global minimizer.

The proof of Fact 2.5 is left as an exercise. *hint:* Try a proof by contradiction, and use convexity.

Definition 2.6 Let $C \subset \mathcal{H}$ be nonempty.

(i) The **indicator function** of C is

$$\iota_C \colon \mathcal{H} \to [-\infty, +\infty] : x \mapsto \begin{cases} 0 & \text{if } \ x \in C \\ +\infty & \text{if } \ x \notin C. \end{cases} \tag{32}$$

(ii) The **distance function** of C is

$$\operatorname{dist}_C(x) = \inf_{c \in C} \|x - c\| \tag{33}$$

(iii) Suppose that C is also closed. A **projection** of $x \in \mathcal{H}$ onto C is a solution to the minimization problem

$$\operatorname*{minimize}_{z \in C} \ \|x - z\|. \tag{34}$$

In other words, it minimizes the distance from x to C (33), i.e., it is a closest point to x that resides within C.

A solution to (34) is a "closest" point to x that resides in C.

Fact 2.7 Let $C \subset \mathcal{H}$ be nonempty and let $x \in \mathcal{H}$.

(i) Without loss of generality, constrained optimization can be rephrased as unconstrained optimization via changing the objective function:

$$\inf_{x \in C} f(x) = \inf_{x \in \mathcal{H}} f(x) + \iota_C(x). \tag{35}$$

The objective function $f + \iota_C$ on the righthand side, although a bit fancier, allows us to rephrase the constraint on the lefthand side.

(ii) C is convex if and only if its indicator function ι_C is convex.

(iii) C is closed if and only if its indicator function ι_C is lsc.

(iv) Suppose that C is closed. Then a solution to (34) exists.

(v) Suppose that C is convex. If a solution to (34) exists, it is guaranteed to be unique.

The proofs of (ii) and (iii) follow from the fact that $\operatorname{epi} \iota_C = C \times [0, +\infty[$. Loosely speaking, the proof of (iv) follows from the Weierstraß theorem (compactness is achieved by intersecting C with $\big\{ y \in \mathcal{H} \;\big|\; \|x - y\| \leq \eta \big\}$ for $\eta > 0$) and (v) follows from the fact that the norm is *strictly convex* – (a notion we have not yet defined, but we will see later in Definition 4.1).

Definition 2.8 Let $C \subset \mathcal{H}$ be nonempty, closed, and convex. In view of Fact 2.7(iv)–(v), for every $x \in \mathcal{H}$ there is a unique point, $\operatorname{Proj}_C(x) \in C$, that solves (34). This implicitly defines the **projection operator** of C.

$$\operatorname{Proj}_C \colon \mathcal{H} \to \mathcal{H} \colon x \mapsto \operatorname{Proj}_C(x) \qquad \text{(solution to (34))} \tag{36}$$

Note: if $x \in C$, then $\operatorname{Proj}_C x = x$.

Fact 2.9 (Inner-product characterization of the projection) Let $C \subset \mathcal{H}$ be nonempty, closed, and convex. Let $x, p \in \mathcal{H}$. Then:

$$\operatorname{Proj}_C x = p \quad \Leftrightarrow \quad \Big[p \in C \quad \text{and} \quad (\forall c \in C) \quad \langle x - p \mid c - p \rangle \leq 0 \Big]. \tag{37}$$

For "convex optimization," it is standard to look at the following class of functions:

$$\Gamma_0(\mathcal{H}) = \big\{ f \colon \mathcal{H} \to]-\infty, +\infty] \;\big|\; f \text{ is proper, lower semicontinuous, and convex} \big\}. \tag{38}$$

The following functions live in $\Gamma_0(\mathcal{H})$:

(i) ι_C, where C is nonempty, closed, and convex

(ii) Exponentials: e^x

(iii) Log-barriers $f(x) = \begin{cases} -\ln(x) & \text{if } x > 0 \\ +\infty & \text{otherwise.} \end{cases}$

(iv) Any norm: $\|\cdot\|$ (e.g., $\|\cdot\|_1$ which promotes sparsity, $\|\cdot\|_{\text{nuclear}}$ which promotes low-rank)

(v) Hinge-Loss, ReLU, KL-Divergence, ...

(vi) Given a collection of functions $(f_i)_{i=1}^m$ in $\Gamma_0(\mathcal{H})$, if their domains intersect (i.e., $\bigcap_{i=1}^m \operatorname{dom} f_i \neq \varnothing$), we can remain in $\Gamma_0(\mathcal{H})$ via the following operations.

  (a) $\max\{f_1, \ldots, f_m\}$

  (b) Positive linear combinations: $\lambda_1 f_1 + \cdots + \lambda_m f_m$, where $\{\lambda_i\}_{i=1}^m$ are positive.

  (c) Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two finite-dimensional real vector spaces. Let $b \in \mathcal{H}_2$ and let $A \colon \mathcal{H}_1 \to \mathcal{H}_2$ be a linear operator (e.g., a matrix from $\mathbb{R}^n$ to $\mathbb{R}^m$). If $f_1 \in \Gamma_0(\mathcal{H}_2)$, and there exists at least one $z \in \mathcal{H}_1$ such that $Az + b \in \operatorname{dom} f_1$, then $g(\cdot) = f_1(A \cdot + b) \in \Gamma_0(\mathcal{H}_1)$.

Exercise 2.10 The **Minkowski sum** of two subsets A, B of $\mathcal{H}$ is given by

$$A + B = \big\{ a + b \;\big|\; a \in A \;\text{ and }\; b \in B \big\}. \tag{39}$$

Assume that A and B are convex. Prove that $A + B$ is convex.

Exercise 2.11 Prove that every norm $\|\cdot\|$ is convex using Definition 1.11.

Exercise 2.12 Suppose that $C \subset \mathcal{H}$ is convex and nonempty. Prove that ι_C is a convex function.

3 What is Differentiability?

There are a lot of ML engineers who brush off the mathematical details of what it means for a function to be differentiable. Algorithmic differentiation (sometimes misleadingly-called "automatic" differentiation) is only guaranteed to work when certain theoretical conditions about the *existence* of a gradient hold. This part of the class is dedicated to explaining that differentiability is not a freebie.[3]

To start our discussion on differentiability, we will begin with a few preliminaries from analysis.

Definition 3.1 Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces, and let $A\colon \mathcal{H}_1 \to \mathcal{H}_2$. Then A is **linear** if, for every $\alpha \in \mathbb{R}$ and every $x, y \in \mathcal{H}_1$,

$$A(\alpha x) = \alpha A(x) \quad \text{and} \quad A(x+y) = A(x) + A(y). \tag{40}$$

The **adjoint** of a linear operator is the unique operator, denoted A^*, such that for all $(x, z) \in \mathcal{H}_1 \times \mathcal{H}_2$,

$$\langle Ax \mid z \rangle_{\mathcal{H}_2} = \langle x \mid A^* z \rangle_{\mathcal{H}_1} \tag{41}$$

where $\langle \cdot \mid \cdot \rangle_{\mathcal{H}_i}$ denotes the scalar product in $\mathcal{H}_i$.

Theorem 3.2 (Riesz-Fréchet representation) *Let $A\colon \mathcal{H} \to \mathbb{R}$ be linear. Then there exists a unique vector $u \in \mathcal{H}$ such that, for every $x \in \mathcal{H}$, $A(x) = \langle u \mid x \rangle$.* [4]

Although at first-glance it looks unrelated, Theorem 3.2 is a central notion for defining the gradient. A necessary (albeit insufficient) condition for the existence of a gradient is the existence of a directional derivative, defined below.

Definition 3.3 Let $f\colon \mathcal{H} \to]-\infty, +\infty]$ be proper. The **directional derivative** of f at $x \in \operatorname{dom} f$ in the direction $y \in \mathcal{H}$ is

$$f'(x; y) = \lim_{\alpha \searrow 0} \frac{f(x + \alpha y) - f(x)}{\alpha}. \tag{42}$$

From Definition 3.3, we point out a few things.

(i) The limit in (42) might not exist.

(ii) If f is convex, then $f'(x; y) \in [-\infty, +\infty]$.

(iii) Even if a directional derivative exists, it might not exist in $\mathbb{R}$ (since it could be $+\infty$ or $-\infty$).

[3] What happens if we use algorithmic differentiation (AKA, 'backpropagation') anyways? $\nabla f(x) = 0$ becomes meaningless, and artificial solutions can occur. Learn more from Jérôme Bolte's talk (especially slides 10-15)

[4] Theorem 3.2 requires additional assumptions if $\mathcal{H}$ is infinite dimensional.

Definition 3.4 Let $x \in \operatorname{dom} f$. If $f'(x;\cdot)$ is linear, we say f is **differentiable at** x. In this case, the unique vector provided by Theorem 3.2 is called the **gradient** of f at x and denoted $\nabla f(x)$.

$$f'(x;\cdot) = \lim_{\alpha \searrow 0} \frac{f(x+\alpha\,\cdot) - f(x)}{\alpha} = \langle \nabla f(x) \mid \cdot \rangle \tag{43}$$

If f is differentiable at every $x \in \operatorname{dom} f$, we say that f is **differentiable**.

Exercise 3.5 Prove that $\nabla(\frac{1}{2}\|\cdot\|^2)(x) = x$.

All of the properties we know and love about differentiability (chain rule, product rule, etc.) have to be proven. Here is an example below.

Proposition 3.6 *Let $A\colon \mathcal{H}_1 \to \mathcal{H}_2$ be a linear operator (with adjoint denoted A^*), let $b \in \mathcal{H}_2$, and let $f\colon \mathcal{H} \to \mathbb{R}$ be proper and differentiable. Set $g = f(Ax+b)$. Then g is differentiable and*

$$\nabla g = A^*(\nabla f(A\cdot + b)). \tag{44}$$

Proof. Since $\operatorname{dom} f = \mathcal{H}_2$, $\operatorname{dom} g \neq \varnothing$ so we let $x \in \operatorname{dom} g$. By definition,

$$\begin{aligned}
g'(x;y) &= \lim_{\alpha \searrow 0} \frac{g(x+\alpha y) - g(x)}{\alpha} && (45)\\
&= \lim_{\alpha \searrow 0} \frac{f(A(x+\alpha y)+b) - f(Ax+b)}{\alpha} && (46)\\
&= \lim_{\alpha \searrow 0} \frac{f(Ax+b+\alpha Ay) - f(Ax+b)}{\alpha} && (47)\\
&= f'(Ax+b;Ay). && (48)
\end{aligned}$$

So the directional derivative of g exists. Now, since f is differentiable,

$$g'(x;y) = f'(Ax+b;Ay) = \langle \nabla f(Ax+b) \mid Ay \rangle = \langle A^*(\nabla f(Ax+b)) \mid y \rangle. \tag{49}$$

Hence the directional derivative of g is linear and g is differentiable. The specific form of the gradient is constructed in (49) □

Algorithmic differentiation tools use results like Proposition 3.6 to approximate a gradient of a function by reading its machine code. However, these subroutines do not check the theoretical conditions required for their theorems (e.g., *f must be differentiable*) – this must be done (and is oftentimes unjustly ignored) by the user.

Definition 3.7 Let f be proper and differentiable. f is **smooth** ("L-smooth") if there exists $L > 0$ such that

$$(\forall x, y \in \mathcal{H}) \quad \|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|. \tag{50}$$

Another name for this property is to say that "∇f is **Lipschitz continuous** with constant L."

Exercise 3.8 Construct a function that is differentiable and nonsmooth.

Fact 3.9 If f is smooth, then the following inequality holds

$$(\forall (x,y) \in \mathcal{H}^2) \quad f(y) - f(x) \leq \langle \nabla f(x) \mid y - x \rangle + \frac{L}{2}\|y-x\|^2. \tag{51}$$

Sometimes, if f satisfies (51), then f is also called "smooth," depending on the article/book. If f is convex, then (51) is equivalent to (50). To learn more about smoothness, including the aforementioned proof of equivalence, see "On the Fenchel Duality between Strong Convexity and Lipschitz Continuous Gradient," X. Zhou, arXiv:1803.06573 (preprint, 2018).

Proposition 3.10 *Let $f \colon \mathcal{H} \to]-\infty, +\infty]$ be proper and convex. Then,*

$$(\forall x \in \operatorname{dom} f)(\forall y \in \mathcal{H}) \qquad f'(x; y-x) + f(x) \leq f(y). \tag{52}$$

Proof. By Proposition 2.2, for every $\alpha \in]0,1[$,

$$\begin{aligned} f(x + \alpha(y-x)) - f(x) &= f\left((1-\alpha)x + \alpha y\right) - f(x) && (53)\\ &\leq (1-\alpha)f(x) + \alpha f(y) - f(x) && (54)\\ &= \alpha(f(y) - f(x)). && (55) \end{aligned}$$

Therefore,

$$\frac{f(x + \alpha(y-x)) - f(x)}{\alpha} \leq f(y) - f(x). \tag{56}$$

Taking the limit as $\alpha \searrow 0$ implies $f'(x; y-x) \leq f(y) - f(x)$, which in turn yields (52). □

Corollary 3.11 *Let $f \colon \mathcal{H} \to]-\infty, +\infty]$ be proper and convex. If f is differentiable at an interior point x of its domain, then*

$$(\forall y \in \mathcal{H}) \quad \langle y - x \mid \nabla f(x) \rangle + f(x) \leq f(y). \tag{57}$$

When the lefthand side of (57) is viewed as a function of y, we see it is the first-order Taylor series approximation of f. Therefore, it follows from (52) that a convex differentiable function always remains above its first-order Taylor approximation! This is the motivating idea in defining a (convex) subgradient

Definition 3.12 Let $f \colon \mathcal{H} \to]-\infty, +\infty]$. A vector g is a **subgradient** of f at $x \in \mathcal{H}$ if

$$(\forall y \in \mathcal{H}) \quad \langle y - x \mid g \rangle + f(x) \leq f(y). \tag{58}$$

The **subdifferential** of f at x is the set of all subgradients, denoted $\partial f(x)$.

Remark 3.13 The notion of subdifferential in Definition 3.12 is the most common one for convex functions, i.e., $f \in \Gamma_0(\mathcal{H})$. However, if f is nonconvex, oftentimes the subdifferential of Definition 3.12 is the empty set! To address this issue, the optimization community has proposed more general notions of a *subgradient* (e.g., Clarke or Mordukhovich subdifferentials). These generalized subgradients have two useful properties:

- They are more likely to be nonempty, even for nonconvex functions. This allows us to look for stationary points (potential minimizers), even for such nonconvex functions.
- If $f \in \Gamma_0(\mathcal{H})$, then most of the time, these notions of subdifferential yield the same set as the subdifferential in Definition 3.12.

Example 3.14 As shown in class,

$$\partial(|\cdot|)(x) = \begin{cases} -1 & \text{if } x < 0 \\ [-1,1] & \text{if } x = 0 \\ 1 & \text{if } x > 0. \end{cases} \tag{59}$$

This leads to the following fundamental theorem for optimization.

Theorem 3.15 (Fermat's Rule) *Let $f\colon \mathcal{H} \to]-\infty,+\infty]$ be proper. Then x is a minimizer of f if and only if $0 \in \partial f(x)$.*

Proof. By definition,

$$0 \in \partial f(x) \Leftrightarrow (\forall y \in \mathcal{H}) \quad \langle 0 \mid y - x\rangle + f(x) \leq f(y) \tag{60}$$

$$\Leftrightarrow (\forall y \in \mathcal{H}) \quad f(x) \leq f(y). \tag{61}$$

□

Definition 3.16 A point $x \in \mathcal{H}$ is a **stationary point** if $0 \in \partial f(x)$.[5]

Unlike differentiable functions, there are technical conditions we must check in order to get the "standard" rules one would hope for. The following theorem demonstrates some conditions required to simplify computing the subdifferential of a sum of functions.

Theorem 3.17 (Sum rule) *Let $f, g \in \Gamma_0(\mathcal{H})$ and suppose that one of the following holds:*

(i) *The interior of* dom g *intersects with* dom f

(ii) dom $g = \mathcal{H}$

(iii) *The relative interiors*[6] *of* dom f *and* dom g *intersect.*

Then $\partial(f+g) = \partial f + \partial g$.

Remark 3.18 If f is convex and differentiable at $x \in \mathcal{H}$, then $\partial f(x) = \{\nabla f(x)\}$.

[5] If f is nonconvex, the same definition is used, just with a different notion of "subdifferential" – usually with a prefix referring to the type of subdifferential, e.g., "Clarke-stationary point".

[6] The notion of "relative interior" is beyond this crash course, but its definition can be found, e.g., in our textbook

Remark 3.19 (How do we measure the success of an algorithm?) In the analysis of algorithms, we can consider a variety of quantities to measure "how optimal" an approximate solution $\tilde{x}$ is from a true solution $x^\star$. Here are a few options:

(i) Function-value optimality $f(\tilde{x}) - f(x^\star)$ in tandem with feasibility $\operatorname{dist}_C(x) = \inf_{c \in C} \|x - c\|$.

(ii) The distance to a solution $\inf \|\tilde{x} - x^\star\|$.

(iii) Define an operator $F \colon \mathcal{H} \to \mathcal{H}$ such that $Fx = x$ if and only if x is a solution; then quantify $\|x - Fx\|$.

These are great candidates for analysis if the objective function is convex; however, barring some very restrictive settings, if your objective function is nonconvex, it becomes nearly impossible to construct an algorithm that guarantees that the optimality measures above converge to zero. To quote R. T. Rockafellar, one of the pioneers of modern optimization,

"...in fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity."

In these settings of nonconvexity (and potentially nonsmoothness), instead of comparing to a global optimum, we often try to analyze a measure of having a *stationary point* (cf. Definition 3.16), potentially with a generalized subdifferential if the objective function is not differentiable.

Remark 3.20 You may be familiar with the gradient descent algorithm

$$x_{t+1} = x_t - \gamma_t \nabla f(x_t), \tag{62}$$

which in general will converge for small $\varepsilon > 0$, stepsize sequences satisfying $\gamma_t \in [\varepsilon, (2/L) - \varepsilon]$ on L-smooth functions f.[7] A *subgradient descent algorithm* is an optimization algorithm that uses subgradients instead of gradients; so, for instance, a subgradient verison of (62) is $x_{t+1} = x_t - \lambda_t g_t$, where $g_t \in \partial f(x_t)$. In general, this class of algorithm may still converge, but there are *much* more limiting restrictions on the stepsizes λ_t you may choose. For more on this topic (e.g., options for stepsizes when f is convex), see Section 29.6 on "Subgradient projectors" in the class book.

4 Proximity Operators

We will refer to $\mathcal{H}$ as a "finite dimensional real **Hilbert space**." When we say "Hilbert space," we are referring to both a vector space and an inner product $(\mathcal{H}, \langle \cdot \mid \cdot \rangle)$, which also gives rise to a norm via $\|\cdot\| = \sqrt{\langle \cdot \mid \cdot \rangle}$ (see Exercise 1.12).

[7] See this reference for a proof, as well as various convergence rates for convex and nonconvex settings: https://www.mit.edu/ gfarina/2024/67220s24_L07_gradient_descent/L07.pdf

Definition 4.1 Let $f\colon \mathcal{H} \to]-\infty,+\infty]$. Then f is **strictly convex** if, for *distinct* points $x, y \in \mathcal{H}$ $(x \neq y)$, we have

$$(\forall \alpha \in]0,1[) \quad f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha) f(y). \tag{63}$$

Exercise 4.2 Let $f\colon \mathcal{H} \to]-\infty,+\infty]$ be strictly convex and suppose that at least one minimizer exists. Prove that the minimizer is unique.

Proposition 4.3 *Let $f \in \Gamma_0(\mathcal{H})$ and let $x \in \mathcal{H}$. The following optimization problem has a unique solution*

$$\operatorname*{minimize}_{z\in\mathcal{H}} \; f(z) + \frac{1}{2}\|x-z\|^2 \tag{$*$}$$

The proof idea of Proposition 4.3 comes from the fact that the norm is strictly convex, which implies that the objective function in $(*)$ is convex.

Definition 4.4 Let $f \in \Gamma_0(\mathcal{H})$ and let $x \in \mathcal{H}$. The solution to $(*)$ is the **proximal point** of f at x, denoted $\operatorname{Prox}_f(x) \in \mathcal{H}$. This defines an operator $\operatorname{Prox}_f\colon \mathcal{H} \to \mathcal{H}$ where x maps to $\operatorname{Prox}_f(x)$.

Example 4.5 Let C be a nonempty closed convex subset of $\mathcal{H}$. Then, for $x \in \mathcal{H}$, we have

$$\min_{z\in\mathcal{H}} \iota_C(z) + \frac{1}{2}\|x-z\|^2 = \min_{z\in C} \frac{1}{2}\|x-z\|^2. \tag{64}$$

However, the minimizer of (64) will be the same as in (34). Therefore, the proximity operator of ι_C is actually just the projection operator!

$$\operatorname{Prox}_{\iota_C} = \operatorname{Proj}_C. \tag{65}$$

Exercise 4.6 Let $f \in \Gamma_0(\mathcal{H})$ and let $x, p \in \mathcal{H}$. Show that

$$p = \operatorname{Prox}_f x \qquad \Leftrightarrow \qquad x - p \in \partial f(p). \tag{66}$$

Definition 4.7 Let $\mathcal{H}_1, \ldots, \mathcal{H}_m$ be real Hilbert spaces with inner products denoted $\langle \cdot \mid \cdot \rangle_{\mathcal{H}_1}, \ldots, \langle \cdot \mid \cdot \rangle_{\mathcal{H}_m}$. We can construct another Hilbert space, called the **direct sum** (sometimes also called a *product space*) as follows. Our vector space is denoted $\bigoplus_{i=1}^m \mathcal{H}_i = \mathcal{H}_1 \times \cdots \times \mathcal{H}_m$, and the inner product on the direct sum is given by

$$\langle \cdot \mid \cdot \rangle_{\bigoplus_{i=1}^m \mathcal{H}_i}\colon \bigoplus_{i=1}^m \mathcal{H}_i \times \bigoplus_{i=1}^m \mathcal{H}_i \to \mathbb{R}\colon ((x_i)_{i=1}^m, (y_i)_{i=1}^m) \mapsto \sum_{i=1}^m \langle x_i \mid y_i \rangle_{\mathcal{H}_i}. \tag{67}$$

If, for every $i \in I$, $f_i\colon \mathcal{H}_i \to]-\infty,+\infty]$, then the function

$$f\colon \bigoplus_{i=1}^m \mathcal{H}_i \to]-\infty,+\infty]\colon (x_i)_{i=1}^m \mapsto \sum_{i=1}^m f_i(x_i) \tag{68}$$

is called a **separable sum**.

Proposition 4.8 (Proximity Operator of a Separable Sum) *Let $\mathcal{H}_1, \dots, \mathcal{H}_m$ be real Hilbert spaces, and for every $i \in \{1, \dots, m\}$, let $f_i \in \Gamma_0(\mathcal{H}_i)$. Let f be the separable sum of $(f_i)_{i=1}^m$ in the form (68). Then*

$$\left(\forall (x_i)_{i=1}^m \in \bigoplus_{i=1}^m \mathcal{H}_i\right) \quad \operatorname{Prox}_f((x_i)_{i=1}^m) = (\operatorname{Prox}_{f_1}(x_1), \operatorname{Prox}_{f_2}(x_2), \dots, \operatorname{Prox}_{f_m}(x_m)). \tag{69}$$

Proof. For notational convenience, set $\mathcal{H} = \bigoplus_{i=1}^m \mathcal{H}_i$. Let $x = (x_i)_{i=1}^m \in \mathcal{H}$. Then

$$\min_{z \in \mathcal{H}} f(z) + \frac{1}{2}\|x - z\|_{\mathcal{H}}^2 = \min_{\substack{z_1 \in \mathcal{H}_1 \\ \vdots \\ z_m \in \mathcal{H}_m}} \sum_{i=1}^m f_i(z_i) + \frac{1}{2}\|x_i - z_i\|_{\mathcal{H}_i}^2 \tag{70}$$

$$= \sum_{i=1}^m \min_{z_i \in \mathcal{H}_i} f_i(z_i) + \frac{1}{2}\|x_i - z_i\|_{\mathcal{H}_i}^2. \tag{71}$$

Since all of the summands in (70) are only concerned with a single optimization variable z_i (as opposed to the entire vector $z \in \mathcal{H}$), we actually have a sum of independent optimization problems! This observation permits us to commute the sum and the minimization between (70) and (71). Finally, we see that the solution in each of the subproblems in (71) is precisely the proximity operator of f_i at x_i. Since the solution in each component is $\operatorname{Prox}_{f_i}(x_i)$, we have are done since we have just shown (69). □

Example 4.9 Let $\gamma > 0$. Let's compute the proximity operator of my favorite nonsmooth function $\gamma|\cdot|$. Let $x, p \in \mathbb{R}$. Combiniing Example 3.14 and the characterization in Exercise 4.6, we find

$$p = \operatorname{Prox}_{\gamma|\cdot|} x \Leftrightarrow x - p \in \partial(\gamma|\cdot|)(p) \tag{72}$$

$$\Leftrightarrow x - p \in \partial(|\cdot|)(p) = \begin{cases} -\gamma & \text{if } p < 0 \\ [-\gamma, \gamma] & \text{if } p = 0 \\ \gamma & \text{if } p > 0. \end{cases} \tag{73}$$

Now, (73) looks a bit funny, since we don't know what p is to begin with. However, if we cobble together the case analysis, we will be able to formally invert this inclusion. Let's consider the first case: If $p < 0$, then

$$x - p \in \{-\gamma\} \quad \Leftrightarrow \quad p = x + \gamma, \tag{74}$$

which means $p < 0$ if and only if $x < -\gamma$. Note we have translated a condition on p to a condition on x. Similarly, for $p > 0$

$$x - p \in \{\gamma\} \quad \Leftrightarrow \quad p = x - \gamma, \tag{75}$$

and $p > 0$ if and only if $x > \gamma$. Finally, if $p = 0$

$$x - p \in [-\gamma, \gamma] \quad \Leftrightarrow \quad x \in [-\gamma, \gamma]. \tag{76}$$

Note that for each of the three cases, we were able to translate the condition on p in (73) to a condition on x! Combining these three observations, we find that

$$p = \operatorname{Prox}_{\gamma|\cdot|}(x) \begin{cases} x+\gamma & \text{if } x < -\gamma \\ 0 & \text{if } -\gamma \leq x \leq \gamma \\ x-\gamma & \text{if } \gamma < x. \end{cases} \tag{77}$$

This operation in (77) is known as the **soft thresholder**.

Exercise 4.10 Compute the proximity operator of the **one norm**

$$\|\cdot\|_1 \colon \mathbb{R}^n \to \mathbb{R} \colon (x_i)_{i=1}^n \mapsto \sum_{i=1}^n |x_i| \tag{78}$$

hint: Proposition 4.8

Remark 4.11 Collectively, humans know how to compute lot of proximity operators. Two libraries I have used are below.

(i) `proximity-operator.net`

(ii) ProximalOperators.jl (on Github)

Definition 4.12 Let $T \colon \mathcal{H} \to \mathcal{H}$ be an operator and let $x \in \mathcal{H}$. If $Tx = x$, x is a **fixed point** of T.

Proposition 4.13 *Let $f \in \Gamma_0(\mathcal{H})$ and let $x \in \mathcal{H}$. Then*

$$\operatorname{Prox}_f(x) = x \quad \Leftrightarrow \quad x \text{ minimizes } f. \tag{79}$$

Proof. First, note that $\operatorname{dom}(\frac{1}{2}\|\cdot - x\|^2) = \mathcal{H}$, so by the Sum Rule (Theorem 3.17(ii)), we know that

$$\partial\left(f + \frac{1}{2}\|\cdot - x\|^2\right) = \partial f + \partial\left(\frac{1}{2}\|\cdot - x\|^2\right).$$

In a matter which proceeds similarly to Exercise 3.5, we observe from Remark 3.18 that, for every $z \in \mathcal{H}$, $\partial\left(\frac{1}{2}\|\cdot - x\|^2\right)(z) = \{z - x\}$. Combining these facts, along with using Fermat's rule (Theorem 3.15) twice, we find

$$\begin{aligned} \operatorname{Prox}_f(x) = x &\Leftrightarrow x \text{ minimizes } f(\cdot) + \frac{1}{2}\|\cdot - x\|^2 && (80)\\ &\Leftrightarrow 0 \in \partial\left(f(\cdot) + \frac{1}{2}\|\cdot - x\|^2\right)(x) && (81)\\ &\Leftrightarrow 0 \in \partial f(x) + \partial\left(\frac{1}{2}\|\cdot - x\|^2\right) && (82)\\ &\Leftrightarrow 0 \in \partial f(x) + \{x - x\} && (83)\\ &\Leftrightarrow x \text{ minimizes } f. && (84) \end{aligned}$$

□

Proposition 4.13 tells us that the fixed-points of a proximity operator are precisely the minimizers of our function. Spoiler alert: this motivates a fixed-point algorithm! It turns out that, if we just repeatedly apply the proximity operator, we will converge to a solution of our problem. This will be detailed in Theorem 4.19.

There are a **lot** of awesome properties of the proximity operator. However, in the interest of time (and practicability), we now shift focus towards algorithms.

4.1 Firmly nonexpansive operators: useful tools for convergence

In optimization, there are a variety of methods to quantify "convergence" of an algorithm. One method – *primal convergence* shows that the function value of the objective (1) approaches the optimal value, i.e., $f(x_n) - \inf_{x\in C} f(x) \to 0$. However, this could be misleading.

Exercise 4.14 *(extra credit)* Construct a convex function f with at least one minimizer and a sequence $(x_n)_{n\in\mathbb{N}}$ such that $f(x_n) - \inf_{x\in\mathcal{H}} f(x) \to 0$, and for every minimizer z, $\|x_n - z\| \not\to 0$.

There are solutions to Exercise 4.14 (e.g., in Bauschke/Combettes' book mentioned in the introduction) which demonstrate that, no matter how low the objective value can be, our iterates could still be arbitrarily far away from the actual minimizers. As a result, some like to instead show that, for a minimizer $z \in C$, the sequence of iterates actually approaches the minimizer, i.e., $\|x_n - z\| \to 0$. As we will see later, for the proximal point algorithm (PPA), we have both. We'll begin with the notion of a *firmly nonexpansive operator*, which can be used to show convergence for both PPA and a simple variant of gradient descent.

Definition 4.15 An operator $T\colon \mathcal{H} \to \mathcal{H}$ is *firmly nonexpansive* if

$$(\forall x, y \in \mathcal{H}) \quad \|Tx - Ty\|^2 \leq \|x - y\|^2 - \|(x - Tx) - (y - Ty)\|^2. \tag{85}$$

Fact 4.16 If $f \in \Gamma_0(\mathcal{H})$, then:

- For every $\gamma > 0$, $\operatorname{Prox}_{\gamma f}$ is firmly nonexpansive.
- If f is L-smooth, then for every $\alpha \in \left]0, L^{-1}\right]$, $\alpha\nabla f$ and $\operatorname{Id} - \alpha\nabla f(x)$ is firmly nonexpansive.

In consequence, Proposition 4.17 establishes convergence of the proximal point algorithm and the gradient descent algorithm (although wider stepsize ranges are also available for gradient descent; e.g., see https://www.mit.edu/ gfarina/2024/67220s24_L07_gradient_descent/L07.pdf

Proposition 4.17 *Let $T\colon \mathcal{H} \to \mathcal{H}$ be firmly nonexpansive, let $x_0 \in \mathcal{H}$, and let*

$$(\forall t \in \mathbb{N}) \quad x_{n+1} = Tx_n. \tag{86}$$

Suppose that the fixed point set of T *is nonempty, i.e.,* $\operatorname{Fix} T := \left\{z \in \mathcal{H} \;\middle|\; Tz = z\right\} \neq \varnothing$. *Then $(x_n)_{n\in\mathbb{N}}$ converges to a fixed point of T.*

While the full proof of this relies on a tool called *Fejér monotonicity*, we can at least easily show that (x_t) must converge:

First, let $z \in \operatorname{Fix} T$, so $Tz = z$, and let $t \in \mathbb{N}$. By (86) and firm nonexpansiveness,

$$\|x_{t+1} - z\|^2 = \|Tx_t - Tz\|^2 \tag{87}$$
$$\leq \|x_t - z\|^2 - \|x_t - Tx_t - z - Tz\|^2 \tag{88}$$
$$= \|x_t - z\|^2 - \|x_t - Tx_t\|^2 \tag{89}$$
$$\leq \|x_t - z\|^2. \tag{90}$$

The argument (90) above establishes that the sequence of real numbers $(\|x_t - z\|^2)_{t\in\mathbb{N}}$ is decreasing; also, since the norm is always nonnegative, this sequence is also bounded below. By the monotonic convergence theorem (From calc 1), we know $(\|x_t - z\|^2)_{t\in\mathbb{N}}$ must converge to a real number β. Using positivity of the norm for the lower bound, then rearranging the inequality (89) for the upper bound, we find that

$$0 \leq \|x_t - x_{t+1}\|^2 = \|x_t - Tx_t\|^2 \leq \|x_{t+1} - z\|^2 - \|x_t - z\|^2. \tag{91}$$

By the Squeeze theorem, since $\lim_{t\to\infty} 0 = 0$ and $\lim_{t\to\infty} \|x_{t+1} - z\|^2 - \|x_t - z\|^2 = \beta - \beta = 0$, we can conclude that $\lim_{t\to\infty} \|x_t - x_{t+1}\|^2 = 0$. Therefore, $(x_t)_{t\in\mathbb{N}}$ converges.

Again, the full proof of the below theorem uses slightly more complicated machinery, including the notion of *Fejér monotonicity*.

Remark 4.18 (Comment on convergence proofs for prox-based algorithms) The proximal point algorithm (and many of its relatives) can be proven to converge using the following template[8].

(i) Show that the algorithm is **Fejér monotone**, i.e., for a solution to your problem $z \in \mathcal{H}$, we have

$$\|x_{n+1} - z\| \leq \|x_n - z\| \tag{92}$$

Oftentimes, the properties of $(T_n)_{n\in\mathbb{N}}$ actually reveal a strictly negative term being added to $\|x_n - z\|$ on the upper bound in (92). This can sometimes be used to obtain specific rates of convergence.

(ii) Assume that a cluster point of the algorithm exists. Show that it is a solution of our problem.

Using existing theory about Fejér monotonicity, we can conclude $(x_n)_{n\in\mathbb{N}}$ converges to a solution of our problem. Note that we never actually had to prove the algorithm converges.

Theorem 4.19 (Proximal Point Algorithm (Martinet, 1970)) *Let $f \in \Gamma_0(\mathcal{H})$ have at least one minimizer. Let $(\gamma_n)_{n\in\mathbb{N}}$ be a sequence in $]0, +\infty[$ such that $\sum_{n\in\mathbb{N}} \gamma_n = +\infty$, and let $x_0 \in \mathcal{H}$. Set*

$$(\forall n \in \mathbb{N}) \quad x_{n+1} = \operatorname{Prox}_{\gamma_n f}(x_n). \tag{93}$$

Then the following hold.

[8] See Chapter 5 of the course textbook for further details

(i) $f(x_n)_{n\in\mathbb{N}}$ *converges monotonically to* $\min f(\mathcal{H})$.

(ii) $(x_n)_{n\in\mathbb{N}}$ *converges to a minimizer of* f.

Remark 4.20 (A retrospective on Theorem 4.19) The Proximal Point Algorithm tells us that, as long as we can compute the proximity operator of our objective function, we can declare victory. That's it, right? Well, it turns out that the story is not quite so simple. In practice, one can typically compute the proximity operator of each summand in an optimization problem. However, computing the prox of their sum is much trickier. Tune in next time, for an introduction to *splitting algorithms*!

5 Splitting Algorithms

Let's start with a motivating example (from "Learning with optimal interpolation norms" by Combettes, McDonald, Miccheli, & Pontil, in *Numerical Algorithms*, 2019).

Example 5.1 (linear SVM training) Given two datasets $\mathcal{D}_1, \mathcal{D}_2 \subset \mathcal{H}$, let's consider the problem of training a sparse linear separator

$$\underset{x\in\mathcal{H}}{\text{minimize}} \quad \left(\sum_{d\in\mathcal{D}_1} \max\{0, 1 - \langle d \mid x\rangle\}\right) + \left(\sum_{d\in\mathcal{D}_2} \max\{0, 1 + \langle d \mid x\rangle\}\right) + \lambda\|x\|_1. \tag{94}$$

A solution, x^*, to (94) is used to *classify* unobserved data u by looking at the sign of a scalar product: if $\langle u \mid x^*\rangle > 0$, we predict $u \in \mathcal{D}_1$; for $\langle u \mid x^*\rangle < 0$, we predict $u \in \mathcal{D}_2$. The first two sums are composed of **hinge loss** functions based on our two datasets. The final term $\|\cdot\|_1$ is used to promote **sparsity** of x. Loosely speaking, x is "sparse" if it has a small number of nonzero components. The larger value of λ, the more sparse x^* becomes. Once a sparse solution x^* is found, this makes classification predictions very efficient (particularly in high-dimensional settings), because we only have to look at the nonzero components of x^* in order to compute $\langle u \mid x^*\rangle$.

If we look at the libraries in Remark 4.11, the objective function in Example 5.1 does not appear anywhere! We can find the prox of an individual hinge loss and the prox of $\|\cdot\|_1$. However it looks like we can't compute the prox of the sum! This exemplifies a more general problem, an instance of which is summarized below.

Question 5.2

Given $f, g \in \Gamma_0(\mathcal{H})$ *and access to compute* Prox_f *and* Prox_g*, can we efficiently compute* $\operatorname{Prox}_{f+g}$*?*

If a generic answer to Question 5.2 were available, we could just apply the Proximal Point Algorithm (Theorem 4.19) to declare victory. For some settings, we have a positive answer to this question:

Example 5.3 Let U and V be orthogonal vector subspaces of $\mathcal{H}$. Then (as demonstrated in class)

$$\text{Prox}_{\iota_U+\iota_V} = \text{Prox}_{\iota_{U\cap V}} = \text{Proj}_{U\cap V} = \text{Proj}_U \circ \text{Proj}_V = \text{Prox}_{\iota_U} \circ \text{Prox}_{\iota_V}. \tag{95}$$

Hence, for this setting, the prox of the sum is expressed as the composition of prox operators.

However, outside of special instances like Example 5.3, a satisfactory answer to Question 5.2 has not been found. Therefore, the research community has circumvented this issue by producing algorithms which converge to minimizers of $(f+g)$ without requiring an answer to Question 5.2. Loosely speaking, this class of algorithms are known as *splitting algorithms*, and their hallmark characteristic is that they *only rely on operators associated with the individual summands of the objective function*. Below, we present the Forward-Backward algorithm, which minimizes a sum of a smooth function $g \in \Gamma_0(\mathcal{H})$ and another (potentially nonsmooth) function $f \in \Gamma_0(\mathcal{H})$. Instead of requiring Prox_{f+g}, we only rely on Prox_f and evaluating ∇g. The specific form below comes from Proposition 28.13 in Bauschke/Combettes' book.

Theorem 5.4 (Forward-Backward Algorithm) *Let $f \in \Gamma_0(\mathcal{H})$, let $L > 0$, and let $g \in \Gamma_0(\mathcal{H})$ be L-smooth. Let $\gamma \in \,]0, 2/L[$ and set $\delta = 2 - \gamma L/2$. Let $(\lambda_n)_{n\in\mathbb{N}}$ be a sequence in $[0,\delta]$ such that $\sum_{n\in\mathbb{N}} \lambda_n(\delta - \lambda_n) = +\infty$ and let $x_0 \in \mathcal{H}$. Suppose $(f+g)$ has a minimizer. Iterate*

$$\begin{array}{l} \textit{for } \; n = 0, 1, \ldots \\ \quad \left\lfloor \begin{array}{ll} y_n = x_n - \gamma\nabla g(x_n) & \textit{\# Gradient (forward) step} \\ x_{n+1} = x_n + \lambda_n(\text{Prox}_{\gamma f} y_n - x_n). & \textit{\# Prox (backward) step} \end{array} \right. \end{array} \tag{96}$$

Then the following hold:

(i) $(x_n)_{n\in\mathbb{N}}$ *converges to a minimizer of* $f+g$.

(ii) $(f(x_n) + g(x_n)) - \inf_{x\in\mathcal{H}} f(x) + g(x) \searrow 0$.

The proof idea follows from the Fejer machinery discussed in Remark 4.18. We can provide an intuition which demonstrates that fixed-points of the Forward-backward operators in (96) are minimizers of $f+g$. Since g is differentiable, we can use the sum rule. By Fermat's rule, we have

$$\begin{aligned} x \;\text{ minimizes }\; f+g &\Leftrightarrow 0 \in \partial(f+g)(x) && (97)\\ &\Leftrightarrow 0 \in \partial f(x) + \partial g(x) && (98)\\ &\Leftrightarrow 0 \in \partial f(x) + \{\nabla g(x)\} && (99)\\ &\Leftrightarrow -\nabla g(x) \in \partial f(x). && (100) \end{aligned}$$

Multiplying by γ and adding x, then using our characterization from Exercise 4.6 we find, for every $n \in \mathbb{N}$,

$$\begin{aligned} x \;\text{ minimizes }\; f+g &\Leftrightarrow x - \gamma\nabla g(x) \in x + \partial\gamma f(x). && (101)\\ &\Leftrightarrow x = \text{Prox}_{\gamma f}(x - \gamma\nabla g(x)) && (102)\\ &\Leftrightarrow \lambda_n\Big(\text{Prox}_{\gamma f}\big(x - \gamma\nabla g(x)\big) - x\Big) = 0 && (103)\\ &\Leftrightarrow x + \lambda_n\Big(\big(\text{Prox}_{\gamma f}(x - \gamma\nabla g(x)\big) - x\Big) = x, && (104) \end{aligned}$$

i.e., x is a fixed-point of the algorithm (96).

Example 5.5 Let C be a closed convex subset of $\mathcal{H}$, let $f = \iota_C$, and let $\lambda_n \equiv 1/L$. Then, it follows from Example 4.5 that the popular **projected gradient descent algorithm**

$$(\forall n \in \mathbb{N}) \quad x_{n+1} = \operatorname{Proj}_C(x_n - \gamma \nabla g(x_n)) \tag{105}$$

is a special case of the Forward-Backward algorithm (Theorem 5.4).

Example 5.6 (LASSO (Tibshirani)) The Forward-backward algorithm can be applied to solve the LASSO problem,

$$\operatorname*{minimize}_{z \in \mathcal{H}} \ \|Ax - b\|^2 + \|x\|_1. \tag{106}$$

Now, suppose that we can only use proximity operators – for instance, either I must (because all of my functions are nonsmooth, as in Example 5.1), or I have read some literature telling me that I may get better convergence behavior[9]. Even for the LASSO problem, we could compute this prox:

Exercise 5.7 Let $A\colon \mathbb{R}^n \to \mathbb{R}^m$ be a matrix and let $b \in \mathbb{R}^m$. Compute the proximity operator of

$$\frac{1}{2}\|A(\cdot) + b\|^2.$$

hint: Proposition 3.6 and Exercises 4.6 and 3.5.[10]

One of the standard "prox-only" splitting algorithms for solving

$$\operatorname*{minimize}_{x \in \mathcal{H}} \ f(x) + g(x) \tag{107}$$

is the following (whose form is simplified from Corollary 28.3 in the book).

Theorem 5.8 (Douglas-Rachford algorithm) *Let $f, g \in \Gamma_0(\mathcal{H})$ such that a minimizer of $(f+g)$ exists, let $(\lambda_n)_{n\in\mathbb{N}}$ be a sequence in $[0,2]$ such that $\sum_{n\in\mathbb{N}} \lambda_n(2-\lambda_n) = +\infty$, and let $\gamma > 0$. Let $x_0 \in \mathcal{H}$ and set*

$$\begin{aligned} &\textit{for } \ n = 0, 1, \ldots \\ &\quad \left\lfloor \begin{aligned} y_n &= \operatorname{Prox}_{\gamma g} x_n \\ z_n &= \operatorname{Prox}_{\gamma f}(2y_n - x_n) \\ x_{n+1} &= x_n + \lambda_n(z_n - y_n). \end{aligned} \right. \end{aligned} \tag{108}$$

Then $(x_n)_{n\in\mathbb{N}}$ converges to a minimizer of $f+g$.

Below is an indication as to why this algorithm works.

[9] e.g., Combettes & Glaudin, "Proximal activation of smooth functions in splitting algorithms for convex image recovery", *SIAM J. Imaging Sci., 2019*, or Briceño-Arias et al., "A random block-coordinate Douglas-Rachford splitting method with low computational complexity for binary logistic regression", *Comput. Optim. Appl.*, 2019.

[10] (or, if you must, use this link for the solution: http://tiny.cc/y3oy001)

Exercise 5.9 (optional) Let $R_1 = 2\text{Prox}_{\gamma f} - \text{Id}$ and let $R_2 = 2\text{Prox}_{\gamma g} - \text{Id}$. Show that the Douglas-Rachford algorithm is of the form

$$x_{n+1} = x_n + \frac{\lambda_n}{2}\left(R_1(R_2 x_n) - x_n\right). \tag{109}$$

From Exercise 5.9, we can show an indication as to why fixed points of the Douglas-Rachford algorithm yield solutions of (107). Let $x \in \mathcal{H}$ and $n \in \mathbb{N}$. Then, using Exercise 5.9, x is a fixed-point of the Douglas Rachford algorithm if and only if

$$\begin{aligned}
x = x + \frac{\lambda_n}{2}\left(R_1(R_2 x) - x\right) &\Leftrightarrow 0 = \frac{\lambda_n}{2}\left(R_1(R_2 x) - x\right) && (110)\\
&\Leftrightarrow x = R_1(R_2 x) && (111)\\
&\Leftrightarrow x = R_1(2\text{Prox}_{\gamma g} x - x) && (112)\\
&\Leftrightarrow x = 2\text{Prox}_{\gamma f}(2\text{Prox}_{\gamma g} x - x) + x - 2\text{Prox}_{\gamma g} x && (113)\\
&\Leftrightarrow \text{Prox}_{\gamma g} x = \text{Prox}_{\gamma f}(2\text{Prox}_{\gamma g} x - x) && (114)\\
&\Leftrightarrow \begin{cases} y = \text{Prox}_{\gamma g} x \\ y = \text{Prox}_{\gamma f}(2y - x). \end{cases} && (115)
\end{aligned}$$

Therefore, in view of Exercise 4.6

$$\begin{aligned}
x = x + \frac{\lambda_n}{2}\left(R_1(R_2 x) - x\right) &\Leftrightarrow \begin{cases} x - y \in \partial\gamma g(y) \\ (2y - x) - y =\in \partial\gamma f(y). \end{cases} && (116)\\
&\Leftrightarrow \begin{cases} \dfrac{x - y}{\gamma} \in \partial g(y) \\ \dfrac{y - x}{\gamma} \in \partial f(y). \end{cases} && (117)
\end{aligned}$$

Adding implies that $0 \in \partial f(y) + \partial g(y)$. One statement I did not mention is that $\partial(f) + \partial(g) \subset \partial(f+g)$ basically always holds (while achieving equality for the reverse inclusion requires extra hypotheses, as discussed in the Sum Rule Theorem 3.17). Hence, by Fermat's rule (Theorem 3.15), $y = \text{Prox}_{\gamma g} x$ is a minimizer of $f + g$!

5.1 What about a sum of more than two functions?

Thanks to the Douglas-Rachford algorithm (Theorem 5.8), we can minimize a sum of two functions using their proximity operators. But this still does not give us a route to solve Example 5.1. What about an arbitrary number of functions? For $f_1, \ldots, f_m \in \Gamma_0(\mathcal{H})$, I want to minimize

$$\sum_{i=1}^{m} f_i(x). \tag{118}$$

We will start with the following construction.

Exercise 5.10 Let $\mathcal{H}$ be a Hilbert space, and consider the m-fold direct sum $\boldsymbol{\mathcal{H}} = \oplus_{i=1}^m \mathcal{H}$. Let us define the **diagonal subspace**:

$$D = \{(x_i)_{i=1}^m \in \boldsymbol{\mathcal{H}} \mid (\forall i, j \in \{1, \dots, m\}) \quad x_i = x_j\}. \tag{119}$$

Show that the projection onto D is obtained by averaging all of the components:

$$(\forall (x_i)_{i=1}^m \in \boldsymbol{\mathcal{H}}) \quad \text{Proj}_D((x_i)_{i=1}^m) = \left(\frac{1}{m}\sum_{i=1}^m x_i\right)_{i=1}^m \tag{120}$$

Example 5.11 (Product space technique) Let $(f_i)_{i=1}^m$ be functions on $\Gamma_0(\mathcal{H})$, and suppose I want to

$$\operatorname*{minimize}_{x\in\mathcal{H}} \quad \sum_{i=1}^n f_i(x). \tag{121}$$

We are going to solve (121) by reformulating it on the product space $\boldsymbol{\mathcal{H}} = \bigoplus_{i=1}^m \mathcal{H}$. Set

$$f\colon \boldsymbol{\mathcal{H}} \to \left]-\infty, +\infty\right] : (x_i)_{i=1}^m \mapsto \sum_{i=1}^m f_i(x_i) \quad \text{and} \tag{122}$$

$$g = \iota_D, \quad \text{where } D \text{ is defined in (119).} \tag{123}$$

Then

$$\operatorname*{minimize}_{x\in\boldsymbol{\mathcal{H}}} \; f(x) + g(x) = \operatorname*{minimize}_{\substack{x_1\in\mathcal{H}\\ \vdots \\ x_m\in\mathcal{H}\\ (\forall i,j)\;\; x_i = x_j}} \quad \sum_{i=1}^m f_i(x_i) = \operatorname*{minimize}_{x\in\mathcal{H}} \quad \sum_{i=1}^m f_i(x). \tag{124}$$

So, using this product space $\boldsymbol{\mathcal{H}}$, we were able to write our original problem (121) in the form (107) which is tractable with the Douglas Rachford algorithm (Theorem 5.8). Now, we only need the proximity operators of f and g. However, from Proposition 4.8 and Exercise 5.10 we know that

$$\text{Prox}_f(x_i)_{i=1}^m = (\text{Prox}_{f_1}(x_1), \dots, \text{Prox}_{f_m}(x_m)) \tag{125}$$

$$\text{Prox}_g(x_i)_{i=1}^m = \text{Proj}_D((x_i)_{i=1}^m) = \left(\frac{1}{m}\sum_{i=1}^m x_i\right)_{i=1}^m, \tag{126}$$

so we can use DR to minimize (121). This yields the following algorithm.

Algorithm 5.12 (Douglas-Rachford in a product space) Let $(f_i)_{i=1}^m$ be functions in $\Gamma_0(\mathcal{H})$ such that a minimizer of (118) exists, let $(\lambda_n)_{n\in\mathbb{N}}$ be a sequence in $[0, 2]$ such that $\sum_{n\in\mathbb{N}} \lambda_n(2 - \lambda_n) = +\infty$, and let $\gamma > 0$. Let $\{x_{i,0}\}_{i=1}^m \subset \mathcal{H}$ and set

$$\begin{array}{l}
\text{for } \; n = 0, 1, \dots \\
\quad y_n = \frac{1}{m}\sum_{i=1}^m x_{i,n} \qquad\qquad\qquad\qquad \text{\# Prox}_{\gamma g} \text{ step} \\
\quad \text{for } \; i = 1, \dots, m \\
\quad\quad z_{i,n} = \text{Prox}_{\gamma f_i}(2y_n - x_{i,n}) \qquad \text{\# Prox}_{\gamma f} \text{ step} \\
\quad\quad x_{i,n+1} = x_{i,n} + \lambda_n(z_{i,n} - y_n).
\end{array} \tag{127}$$

One drawback is that storage scales linearly with the number of summands in our objective. Unfortunately, I am not aware of many prox-based methods which can avoid this issue.

Exercise 5.13 If $\mathcal{H}$ has dimension d, show that the storage required by the product-space Douglas-Rachford algorithm (127) scales like $\mathcal{O}(md)$.

5.2 Practitioners' notes and algorithmic advances

Remark 5.14 (Computing a prox) If you find yourself needing to compute the prox of a function, you can oftentimes avoid computing it directly. The easier route is to use existing theory to cobble together the prox of your specific function. The prox operators of many "base" nonlinear functions $f \in \Gamma_0(\mathcal{H})$ appearing in optimization are available in the libraries discussed in Remark 4.11. From there, you can use some of the following rules to compute the prox of your "fancier" function g in terms the prox of your "base" function f. These results (and many, many more) appear in Chapter 24 of the class book. Let $z, u \in \mathcal{H}$.

(i) Let $g = f(\cdot - z)$. Then $\operatorname{Prox}_{\gamma g}(x) = z + \operatorname{Prox}_{\gamma f}(x - z)$.

(ii) Let $g = f + \frac{\alpha}{2}\|\cdot - z\|^2 + \langle \cdot \mid u\rangle$. Then $\operatorname{Prox}_{\gamma g} x = \operatorname{Prox}_{\gamma(\gamma\alpha+1)^{-1} f}((\gamma\alpha + 1)^{-1}(x + \gamma(\alpha z - u)))$.

(iii) Let $g(x) = f(-x)$. Then $\operatorname{Prox}_g(x) = -\operatorname{Prox}_f(-x)$.

(iv) Let $\mathcal{H} = \mathbb{R}^{n\times m}$ be the real Hilbert space of $n \times m$ matrices under the Frobenius norm. Let $r = \min\{m, n\}$ and, for $x \in \mathcal{H}$, we denote its reduced SVD with $x = U \operatorname{diag}(\sigma_1(x), \ldots, \sigma_r(x))V^\top$. Let $f \in \Gamma_0(\mathbb{R})$ be an even function[11]. If, for every $x \in \mathcal{H}$, $g(x) = \sum_{i=1}^r f(\sigma_i(x))_{i=1}^r$, then

$$\operatorname{Prox}_{\gamma g} = U \operatorname{diag}(\operatorname{Prox}_f \sigma_1(x), \ldots, \sigma_k(x), 0, \ldots, 0)V^\top, \tag{128}$$

where $k = \operatorname{rank}(x)$.[12]

(v) Let $g = f \circ L$ where L is an invertible linear operator such that $L^{-1} = L^*$ (e.g., Fourier transform, DCT, or an orthogonal wavelet transform). Then $\operatorname{Prox}_{\gamma g} x = L^*(\operatorname{Prox}_{\gamma f}(Lx))$.

The book has extensive results on how to handle more arcane linear operators than in (v). In the worst case, one may need to "split" a function f from its linear operator $L\colon \mathcal{H} \to \mathcal{G}$, using the graph of the linear operator via the following technique. By setting $G = \{(x, y) \in \mathcal{H} \times \mathcal{G} \mid Lx = y\}$, we can reformulate a problem on $\mathcal{H}$ with a generic linear operator as a problem on the product space $\mathcal{H} \bigoplus \mathcal{G}$ as follows

$$\inf_{x\in\mathcal{H}} f(Lx) = \inf_{\substack{(x,y)\in\mathcal{H}\bigoplus\mathcal{G} \\ Lx=y}} f(y) = \inf_{(x,y)\in\mathcal{H}\bigoplus\mathcal{G}} f(y) + \iota_G(x, y). \tag{129}$$

[11] An **even** function satisfies $f(-x) = f(x)$

[12] This result is used to derive the prox of the nuclear norm $\|\cdot\|_{\text{nuc}} = \sum_{i=1}^r |\sigma_i(\cdot)|$ which is used in low-rank recovery problems in data science and image processing.

With the reformulation (129) (where, in the final infimum, f is technically viewed as a function which maps (x, y) to $f(y)$), one can use any basic splitting algorithm. Formulae for computing $\text{Prox}_{\iota_G} = \text{Proj}_G$ are presented in Example 29.19 in the book (note they all require inverting a linear operator).

5.3 A few keywords to look up

In the last several decades, there have been a **lot** of algorithmic "bells and whistles" folks have added to prox-based algorithms for solving problems of the form (118) (and more generic models as well). I will hand-wavily discuss a few of them with terminology one could use to search around in the literature.

(i) **Parallelism**: For most prox-based algorithms, the lion's share of computational time is usually devoted to computing the prox operators (since the other operations are often simple linear algebra operations). Structures like the product-space DR algorithm in (127) allow the prox operators in the inner loop to be activated in parallel.

(ii) **Block-activation** While parallelized algorithms are helpful, processing every Prox_{f_i} for $i \in \{1, \dots, m\}$ may be intractable or prohibitively slow. For some data science applications (e.g., in Example 5.1), this corresponds to processing a full epoch over a dataset in every iteration. This issue has given rise to the advent of **block-iterative** (or **block-activated**) algorithms where, instead of activating every operator $(\text{Prox}_{f_i})_{i=1}^m$ at every iteration $n \in \mathbb{N}$, we only activate a subset $I_n \subset \{1, \dots, m\}$ of them. By re-using old iterates for the "non-activated" terms in $\{1, \dots, m\} \setminus I_n$, we save a lot of computational effort. There are both *deterministic* and *stochastic* variants of these algorithms. For deterministic algorithms, one can select I_n such that we expect the computations for $(\text{Prox}_{f_i})_{i \in I_n}$ to finish at roughly the same time.

(iii) **Asynchrony** In most block-activated methods, we must wait until all of the calculations of $(\text{Prox}_{f_i})_{i \in I_n}$ are completed before one can perform a *synchronization* step to complete one iteration. However, asynchronous algorithms circumvent this waiting issue.

(iv) **Extrapolation** Many iterative schemes use updates of the form

$$x_{n+1} = x_n + \lambda_n d_n,$$

where d_n is the "direction" we travel from the current iterate, and λ_n can be viewed as a step-size. For most of our algorithms, λ_n has a fixed upper bound. However (particularly in early parts of an optimization algorithm), large steps can yield favorable convergence. So, there are algorithms out there which allow one to use larger values of λ_n which are computed on-the-fly during your iteration (rather than pre-scheduled stepsizes, as in the vanilla DR/FB algorithms).

(v) **Acceleration** is a method for improving the convergence rate of an algorithm.

These terms are often turned into adjectives in the literature, for instance many **Projective Splitting** algorithms[13] are *parallel, block-activated, and asynchronous*.

6 Frank-Wolfe algorithms

Suppose we want to solve

$$\underset{x\in C}{\text{minimize}}\ g(x), \tag{130}$$

where g is L-smooth, and C is nonempty, convex and compact. We could try applying the forward-backward algorithm (i.e., projected gradient descent of Example 5.5); however, sometimes Proj_C is very computationally expensive, or it is not possible to compute. Frank-Wolfe algorithms can help with that.

Instead of relying on a projection operator, Frank-Wolfe algorithms rely on the so-called "linear minimization operator", or LMO.

Definition 6.1 Let C be convex and compact. Any operator that, for a given input $x \in \mathcal{H}$, yields a vector in $\text{Argmin}_{c\in C}\langle x \mid c\rangle$ (cf. Definition 1.5) is called a *linear minimization oracle*. In particular

$$\text{LMO}_C(x) \in \underset{c\in C}{\text{Argmin}}\langle x \mid c\rangle. \tag{131}$$

On many applications, evaluating LMO_C is computationally cheaper than evaluating Proj_C. In fact, it is an open question as to whether or not there are *any* compact convex sets such that Proj_C is actually faster than LMO_C. This improvement in speed has led to increased interest in Frank-Wolfe (also called, "Conditional gradient") algorithms in the past decade or so. Here's a basic version:

Pick any initial iterate $x_0 \in C$ and iterate

$$(\forall t \in \mathbb{N}) \qquad \begin{cases} v_t \leftarrow \text{LMO}_C(\nabla g(x_t)) \\ x_{t+1} = x_t + \gamma_t(v_t - x_t). \end{cases} \tag{FW}$$

Notes:

- In order for an operation (131) to be well-defined, compactness is often assumed (see the Weierstraß Theorem 1.4).
- By the convexity of C, as long as $\gamma_t \in [0, 1]$, we know $x_{t+1} = (1 - \gamma_t)x_t + \gamma_t v_t$, so the iterates $\{x_t\}_{t\in\mathbb{N}} \subset C$, are always feasible.

[13] e.g., "Asynchronous block-iterative primal-dual decomposition methods for monotone inclusions" by Combettes & Eckstein, in *Math. Program.*, 2018, or "Projective splitting with forward steps" by Jonstone and Eckstein in *Math. Program.*, 2022.

- There are several common choices for stepsizes γ_t:
  - A pre-defined stepsize is often called "open loop." One of the most common is $(\forall t \in \mathbb{N}) \quad \gamma_t = \frac{2}{t+2}$.
  - The "short-step" stepsize is adaptively chosen based on terms that can be easily compued on the fly as the algorithm runs: $\gamma_t = \min\left\{1, \frac{\langle \nabla g(x_t) | x_t - v_t \rangle}{L\|x_t - v_t\|^2}\right\}$. For scenarios when L is not known, there are shortsep variants called "backtracking" that can adaptively estimate L on-the-fly (and they're still proven to converge under mild assumptions).
  - The "line search" stepsize is typically more expensive, but iteratively evaluates $(\forall t \in \mathbb{N}) \quad \gamma_t = \operatorname{Argmin}_{\gamma \in [0,1]} g(x_t + \gamma(v_t - x_t))$.
- The "v_t" step can be considered as finding an extreme point of C that best aligns with the direction of steepest local descent, i.e., $-\nabla f(x_t)$.

Theorem 6.2 *Let $g\colon \mathcal{H} \to \mathbb{R}$ be convex and L-smooth, let $C \subset \mathcal{H}$ be convex and compact, and let $x_0 \in C$. Let $D = \sup_{z_1, z_2 \in C} \|z_1 - z_2\|$ be the diameter of C.*[14] *Consider the step size $(\forall t \in \mathbb{N})$ $\gamma_t = \frac{2}{t+3}$. Then the iterates of* (FW) *satisfy*

$$g(x_t) - \min_{z \in C} g(z) \leq \frac{2LD^2}{t+3}. \tag{132}$$

See this talk by S. Pokutta's for a good introduction; proof of the above theorem appears on slide 7.

Exercise 6.3 Consider the smoothness inequality (51), where $y \leftarrow x_{t+1}$ and $x \leftarrow x_t$, under the Frank-Wolfe framework (FW). If we want to find a good bound on our progress measure $f(x_{t+1}) - f(x_t)$, it would be good if we can find an ideal stepsize $\gamma_t \in [0, 1]$ that makes our upper bound from (51) the *smallest*. So:

(i) Minimize the upper bound over all $\gamma_t \in [0, 1]$ within above framework. This is a 1-D constrained optimization problem, and you can use tools from Calculus 1 to solve it in closed-form.

(ii) Discuss this solution's relationship with the short-step FW stepsize.

[14] D is finite because C is bounded.