mitchell.krock@colorado.edu

# APPM 5720

## Convex Optimization

Dr. Stephen Becker • Fall 2018 • University of Colorado Boulder

Last Revision: February 10, 2019

## Table of Contents

**Abstract**

These notes are intended as a resource for APPM 5720, Convex Optimization.

# 1 Introduction

We will be studying the following problem: $\min\limits_{x \in \mathbb{R}^n} f(x)$ subject to $x \in C \subseteq \mathbb{R}^n$ with $f : \mathbb{R}^n \to \mathbb{R}$.

**Remark 1.1.** Note that $\min -f(x) = -\max f(x)$, so this formulation covers maximization as well.

**Definition 1.1** (Lipschitz Continuity)**.** A function is **Lipschitz continuous** if there exists $L > 0$ such that for all $x, y$, we have $|f(x) - f(y)| \leq L\|x - y\|$. The choice of norm doesn't matter since all norms are equivalent on a finite dimensional vector space. (Slightly stronger than uniform continuity).

**Definition 1.2** ($\epsilon$-solution)**.** An $\epsilon$-solution is a point $x$ such that $f(x) \leq \left(\min\limits_x f(x)\right) + \epsilon$.

**Example 1.1.** Suppose $f$ is Lipschitz continuous with respect to $\|\cdot\|_\infty$, and $C = \{x \in \mathbb{R}^n \mid \|x\|_\infty \leq 1\}$. Choose $2p + 1$ points in each direction:

$$x_1 \in \left\{-1, \frac{-p+1}{p}, \ldots, \frac{p-1}{p}, 1\right\}$$

The solution $x^*$ is within $\frac{1}{2} \cdot \frac{1}{p}$ of a gridpoint in $x_1$. My error $|f(x_{best}) - f(x^*)| \leq \frac{1}{2p}L$, so we take $\frac{L}{2p}$ for our $\epsilon$. Choose $p = \frac{L}{2} \cdot \frac{1}{\epsilon}$; the number of evaluations is $(2p + 1)^n$. Thus to find an $\epsilon$-solution, we need $(\frac{L}{\epsilon} + 1)^n$ evaluations.

## 1.1 Solutions to Optimization Problems

- If $x^*$ is a **solution**, then it's a **global minimizer**. In particular, $x^*$ is **feasible**, meaning it satisfies the constraints $x^* \in C$, and $f(x^*) \leq f(x)$ for all $x \in C$.
- If $x^*$ is a **strict global minimizer**, then it's a unique solution.
- If $x^*$ is a **local minimizer**, then $x^*$ is feasible and $\exists\ \epsilon > 0$ so that $f(x^*) \leq f(x)$ for all $x \in C \cap B_\epsilon(x^*)$.
- If $x^*$ is a **strict local minimizer**, then $x^*$ is feasible and $\exists\ \epsilon > 0$ so that $f(x^*) < f(x)$ for all $x \in C \cap B_\epsilon(x^*)$.
- If $x^*$ is a **isolated local minmizer**, then $x^*$ is a local minimizer and there exists $\delta > 0$ so that no other local minimizers exists within $B_\delta(x^*)$.

**Example 1.2.** Let's take a look at some examples.

- $\min\limits_x \frac{1}{2}\|x\|^2$. Then $x^* = 0$ is a unique global minimizer.
- $\min\limits_{x \in \mathbb{R}} x^2(x-1)^2$. Then $x^* = 1$ is global minimizer and also an isolated local minimum.
- $\min\limits_x$ has no minimizer
- $\min\limits_{x \in (0,1)} x$ has no minimizer. (Need an infimum).

**Remark 1.2.** When considering $\min\limits_{x \in C} f(x)$, it's nice to assume the set $C$ is closed and nonempty.

**Definition 1.3** (Compact)**.** A set $C$ is **compact** if it's closed and bounded. (This is an easy definition for finite dimensional vector space).

**Theorem 1.1** (Weierstrass Theorem)**.** Consider $\min_{x\in C} f(x)$ where $f$ is continuous and $C$ is compact. A solution necessarily exists.

*Proof.* We claim that $f(C)$ is bounded. For a contradiction, assume there exists $(x_k)$ such that $|f(x_k)| \to \infty$. From compactness, there exists $x_{k_\ell} \to x \in C$. By continuity of $f$ we get $\lim_{\ell\to\infty} f(x_{k_\ell}) = f(x)$, a contradiction.
Let $f^* = \inf_{x\in C} f(x)$. We create $(x_k)$ so that $f(x_k) \to f^*$. and the result follows by sequential continuity. □

**Example 1.3.** Isolated local minimizer $\implies$ strict local minimizer. The converse is not true: consider

$$f(x) = \begin{cases} x^4\cos(1/x) + 2x^4, & x \neq 0 \\ 0, & x = 0 \end{cases}$$

Here $x^* = 0$ is a strict local minimizer but not an isolated local minimizer because of oscillation near $x = 0$.

**Example 1.4.** Consider $\max_{(x,y)\in\mathbb{R}^2} 21x + 11y$ such that $\vec{x} \succeq 0$ componentwise and $7x + 4y \leq 13$ and $x, y \in \mathbb{Z}$. The solution to the linear program without $x, y \in \mathbb{Z}$ is $x = 13/7$ and $y = 0$. If we consider the integer constraint, the solution is $(x, y) = (0, 3)$. (The integer constraints are difficult!)

**Definition 1.4** (Convex Set)**.** $C$ is a **convex set** if for all $x, y \in C$ and all $t \in [0, 1]$, we have $tx + (1-t)y \in C$. Geometrically, the line segment between points $x$ and $y$ is contained in $C$.

**Definition 1.5** (Convex Function)**.** $f : \mathbb{R}^n \to \mathbb{R}$ is a **convex function** if for all $x, y \in \mathbb{R}^n$ and all $t \in [0, 1]$, we have $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$. Geometrically, the value $f(tx + (1-t)y)$ lies below the line segment connecting $f(x)$ and $f(y)$ for all $t \in [0, 1]$.

**Definition 1.6** (Graph)**.** We define the **graph** of a function by $\text{graph}(f) = \{(x, t) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) = t\}$.

**Definition 1.7** (Epigraph)**.** We define the **epigraph** of a function by $\text{epi}(f) = \{(x, t) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \leq t\}$.

**Theorem 1.2.** A function $f$ is convex $\iff$ $\text{epi}(f)$ is a convex set.

**Theorem 1.3.** Let $f$ be a convex function and $C$ a convex set Then all local solutions to $\min_{x\in C} f(x)$ are global.

*Proof.* Suppose there exists $\epsilon > 0$ and a local solution $x$ such that for all $y \in B_\epsilon(x)$, $f(x) \leq f(y)$. Suppose $x$ is not a global solution so there exists a feasible point $z$ such that $f(z) < f(x)$. Consider the point $\left(1 - \frac{\epsilon}{2}\right)x + \frac{\epsilon}{2}(z - x)$ which lies between $x$ and $z$. By definition of convexity,

$$f\left(\left(1 - \frac{\epsilon}{2}\right)x + \frac{\epsilon}{2}(z - x)\right) \leq \left(1 - \frac{\epsilon}{2}\right)f(x) + \frac{\epsilon}{2}f(z) < \left(1 - \frac{\epsilon}{2}\right)f(x) + \frac{\epsilon}{2}f(x) = f(x),$$

which contradicts the assumption that $x$ is a local solution. □

## 1.2 Overview

We consider in optimization a multi-objective $f : \mathbb{R}^n \to \mathbb{R}^m$ (not in this class!) or a single objective $f : \mathbb{R}^n \to \mathbb{R}$. The single objective functions are characterized into convex and nonconvex problems:

1. (Convex)
   - $f$ smooth or not smooth
   - Constrained or non constrained

- Conic programs (e.g. LP, QP, SDP)

2. (Nonconvex)

- Find a local solution (called *non-linear programming*)
- Final a global solution (small dimensions). There are provable results (typically in small dimensions) and heuristic results (e.g. genetic algorithms) more useful in larger dimensions. Finally, there are integer/combinatorial problems for which the dual problem (CPLEX/Gurobi)

In this course we will study the first four bullet points.

## 2 B&V Chapter 2: Convex Sets

**Definition 2.1.** A **convex combination** is $\sum_{i=1}^{m} t_i x_i,\ t_i \geq 0, \sum_{i=1}^{n} t_i = 1$.

**Definition 2.2.** For all $t \in \mathbb{R}$, $tx + (1-t)y$ is a linear combination. If $x, y \in C$ and $tx + (1-t)y \in C$ for all $t<$ then $C$ is an **affine space**. If $C$ is affine, then for any $x_0 \in C$, $C - x_0$ is linear.

**Definition 2.3.** For all $t \in \mathbb{R}_+$, $tx$ is a **conic combination**. Similarly, $tx + sy$ is a **convex conic combination**. If $C$ contains all its conic combinations, then $C$ is a **cone**. If $C$ also contains all of its convex combinations, it's a **convex cone**.

**Example 2.1.** The simplest example of a cone is $\mathbb{R}^n_+ = \{x \in \mathbb{R}^n \mid x_i \geq 0\}$. The Lorentz cone $\{(x,t) \in \mathbb{R}^{n+1} \mid \|x\| \leq t\}$ is a convex, pointed cone. (Also called the second-order norm when $\|\cdot\|_2$.) $S^n_+ = \{X = X^T \in \mathbb{R}^{n \times n} \mid X \succeq 0\}$ is the positive semidefinite cone.

**Remark 2.1.** For matrices we define $\langle X, Y \rangle = \sum_i \sum_j X_{ij} Y_{ij} = \text{vec}(X)^T \text{vec}(Y) = \text{trace}(X^T Y)$.

**Remark 2.2.** If $X^T X = X X^T$ then $X$ is a normal matrix and hence diagonalizable. Symmetric matrices are therefore diagonalizable. $X$ is psd $\iff$ it has an orthonormal basis of eigenvectors and positive eigenvalues.

**Definition 2.4** (Standard Form of a Linear Program)**.** The standard form for a linear program is $\min_x c^T x$; $Ax = b, x \geq 0$.

**Remark 2.3.** We handle the constraint $Ax \geq b$ by introducing a slack variable $\begin{pmatrix} x \\ s \end{pmatrix}$ and then write our problem as $[A, I] \begin{pmatrix} x \\ s \end{pmatrix} = b$ with $Ax + s = b, s \geq 0$.

**Example 2.2.** Consider $\|x\|_1$ such that $Ax = b$. We formulate this as

$$[A, -A] \begin{pmatrix} x_+ \\ x_- \end{pmatrix} = b, \quad x_+ \geq 0,$$

In particular, the problem is $\min_{x_+, x_-} \mathbf{1}^T \begin{pmatrix} x_+ \\ x_- \end{pmatrix}$ such that $[A, -A] \begin{pmatrix} x_+ \\ x_- \end{pmatrix} = b$.

**Definition 2.5.** The **interior** of a set $C$ is the set of all points $x \in C$ such that $\exists \epsilon > 0$ so that $B_\epsilon(x) \subseteq C$.

**Definition 2.6.** The **relative interior** of $C$ is $\{x \in C \mid \exists \epsilon > 0 \text{ such that } B_\epsilon(x) \cap \text{AffineHull}(C) \subseteq C\}$

**Remark 2.4.** We refer exclusively to polyhedrons, and a polytope is a bounded polyhedron.

**Definition 2.7.** A **simplex** is a polyhedra defined by the convex hull of up to $n$ points.

**Example 2.3.** The unit simplex is $\{0, e_1, e_2, \ldots, e_n\}$, while the probability simplex is $\{e_1, e_2, \ldots, e_n\}$.

**Definition 2.8.** The **convex hull** of $C$ is the smallest convex set containing $C$. In other words, for all $\{x_1, \ldots, x_m\} \subseteq C$, all conic combinations of $\{x_1, \ldots, x_m\}$ are in the convex hull.

## 2.1 Operations that Preserve Convexity

Assume $C_i$ are convex. The following operations preserve convexity:

- $\bigcap\limits_{i \in A} C_i$ is convex
- $C_1 \cup C_2$ is not necessarily convex.
- Image of an affine function $x \mapsto Ax + b$.
- The Cartesian product $C_1 \times C_2 = \{(x, y) \mid x \in C_1, y \in C_2\}$
- The Minkowski sum of sets $C_1 + C_2 = \{x \mid x = y + z \text{ where } y \in C_1, z \in C_2\}$
- The perspective function

Thus scaling, translation, and projections preserve convexity.

## 2.2 Generalized Inequalities

Assume $K$ is

- a cone
- convex
- closed
- solid (interior is nonempty)
- pointed (no lines in the cone)

$X \succeq_K 0$ means $x \in K$ (e.g. $K = S_+^n$). $X \succeq_K Y$ means $X - Y \succeq_K 0$.

**Example 2.4.** The following are examples of proper cones:

- The nonnegative orthant $\mathbb{R}_+^n$
- The second-order cone $\{(x, t) \mid \|x\| \leq t\}$
- The semi-definite cone $S_+^n$.

### 2.3 Separating Hyperplanes

When can I separate a point $x_0$ and a set $C$? We're searching for the closest point $x \in C$ to $x_0$; i.e. $\underset{x\in C}{\arg\min} \frac{1}{2}\|x-x_0\|^2$.

**Definition 2.9** (Chebyshev Set)**.** A set $C$ is called a **chebyshev set** if there exists a unique solution $x = \underset{x\in C}{\arg\min} \frac{1}{2}\|x-x_0\|^2$ for all $x_0 \in \mathbb{R}^n$.

**Remark 2.5.** A closed, convex set is Chebyshev.

**Theorem 2.1** (Matzin)**.** In $\mathbb{R}^2$, all Chebyshev sets are convex. (under Euclidean norm).

**Theorem 2.2.** In a finite dimensional Hilbert space, all Chebyshev sets are convex.

**Theorem 2.3** (Separating Hyperplane Theorem)**.** If $C$ and $D$ are convex and $C \cap D \neq 0$, then there exists $a \neq 0 \in \mathbb{R}$ and $b \in \mathbb{R}$ such that

$$\begin{aligned} a^T x \leq b, &\quad \forall x \in C \\ a^T x \geq b, &\quad \forall x \in D \end{aligned}$$

**Definition 2.10** (Alternative (Strong))**.** Consider sets $C, D$. An **alternative** means that either $C$ is empty and $D$ is non-empty or either $C$ is nonempty and $D$ is empty.

**Theorem 2.4** (Theorem of Alternatives)**.** The sets $\{x : Ax < b\}$ and $\{\lambda \mid \lambda \geq 0, A^T\lambda = 0, \lambda^T b \leq 0\}$ are alternatives.

**Theorem 2.5** (Fredholm Alternative)**.** The sets $\{x \mid Ax = b\}$ and $\{\lambda \mid A^T\lambda = 0, \lambda^T b \neq 0\}$

*Proof.* Is $C$ empty; i.e. is $b \in \operatorname{ran}(A)$? Remember that $\mathbb{R}^m = \operatorname{ran}(A) \oplus \ker(A^T)$, which essentially gives the proof. □

**Theorem 2.6** (Farkas Lemma)**.** The sets $\{x \geq 0 \mid Ax = b\}$ and $\{\lambda \mid A^T\lambda \geq 0, \lambda^T b < 0\}$ are alternatives.

## 3 B&V Chapter 3: Convex Functions

Also supplemented with textbooks from Bauschke+Combettes and Beck.

**Definition 3.1** (Convex)**.** A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if $\operatorname{dom}(f)$ is **convex** and for all $x, y \in \operatorname{dom}(f)$ and all $t \in [0,1]$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

**Definition 3.2** (Convex)**.** $f : \mathbb{R}^n$ is **convex** $\iff$ $\operatorname{epi}(f) = \{(x,t) \in \mathbb{R}^{n+1} \mid f(x) \leq t\}$ is convex.

**Definition 3.3** (Strictly Convex)**.** A function $f : \mathbb{R}^n \to \mathbb{R}$ is **strictly convex** if $\operatorname{dom}(f)$ is convex and for all $x, y \in \operatorname{dom}(f)$ and all $t \in (0,1)$,

$$f(tx + (1-t)y) < tf(x) + (1-t)f(y).$$

**Definition 3.4** ($\mu$-strongly Convex)**.** A function $f : \mathbb{R}^n \to \mathbb{R}$ is **$\mu$-strongly convex** if $\operatorname{dom}(f)$ is convex and if there exists $\mu > 0$ such that $x \mapsto f(x) - \frac{1}{2}\mu\|x\|^2$.

**Remark 3.1.** $f$ is convex $\iff$ for all directions $d$ and for all $x_0 \in \mathbb{R}^n$ and for all $t \in \mathbb{R}$, $\phi(t) = f(x_0 + td)$ is convex.

**Theorem 3.1** (Alexanderov's Theorem)**.** If $f$ is convex, then $f''$ exists almost everywhere.

**Remark 3.2.** Consider **extended value function** $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$. The problems $\min\limits_{x \in \text{dom}(f)} f(x)$ and

$$\min_{x \in \mathbb{R}^n} \tilde{f}(x) = \begin{cases} f(x) & x \in \text{dom}(f) \\ \infty & \text{else} \end{cases}$$

are equivalent. So if $f : \mathbb{R}^n \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, we define $\text{dom}(f) = \{x \mid f(x) < \infty\}$.

**Definition 3.5** (Indicator function)**.** We use the indicator function $\delta_C(x) = \begin{cases} 0, & x \in C \\ \infty, & x \notin C \end{cases}$

**Example 3.1.** Consider $f : \mathbb{R}^2 \to \mathbb{R}$. Suppose that for all $x$, $\phi(y) = f(x, y)$ and for all $y$, $\psi(x) = f(x, y)$ is continuous. Then $f$ need not be continuous; e.g. $f(x, y) = \begin{cases} 0, & (x, y) = 0 \\ \frac{2x}{x^2+y^2}, & \text{else} \end{cases}$ In a similar fashion, if $f(x, y)$ is marginally convex in $x$ and marginally convex in $y$, then $f(x, y)$ is not necessarily jointly convex. (e.g. $f(x, y) = xy$ is not convex.)

**Definition 3.6** (Lower Semicontinuous)**.** $f : \mathbb{R}^n \to \overline{R}$ is **lower semicontinuous** at $x$ if for all $(x_n)$ such that $x_n \to x$, then $f(x) \leq \liminf f(x_n)$. Equivalently, $f$ is lower semicontinuous $\iff$ $\text{epi}(f)$ is closed. Equivalently, $f$ is lower semicontinuous if for all $\alpha \in \mathbb{R}$, then $\text{lev}_\alpha = \{x \mid f(x) \leq \alpha\}$ is closed.

**Definition 3.7** (Proper)**.** $f : \mathbb{R}^n \to \overline{R}$ is **proper** if

- $f : \mathbb{R}^n \to \mathbb{R} \cup \infty$
- $f$ is not $\equiv \infty$ i.e. $\text{dom}(f) \neq \varnothing$

**Definition 3.8.** $\Gamma(\mathbb{R}^n) = \{f : \mathbb{R}^n \to \overline{R} \mid f \text{ is convex and lower semicontinuous}\}$.

**Definition 3.9.** $\Gamma_0(\mathbb{R}^n) = \Gamma(\mathbb{R}^n) \cap \{f \text{ is proper}\}$.

**Example 3.2.** Consider $f(x) = \delta_C(x)$. Then $f$ is convex $\iff$ $C$ is convex, and $f$ is lower semicontinuous $\iff$ $C$ is closed, and finally $f$ is proper $\iff$ $C$ is nonempty.

**Theorem 3.2** (Cor 8.30 in B+C first edition)**.** *If $f$ is proper and convex, then $f$ is continuous on the interior of its domain.*

## 3.1 First Order Conditions

For now assume $f$ is differentiable; i.e. that $\nabla f$ exists. Assume $\text{dom}(f)$ is convex and open.

**Theorem 3.3.** *$f$ is convex $\iff$ $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$ for all $x, y \in \text{dom}(f)$. This means that $f$ lies above its tangent lines.*

**Theorem 3.4.** *$f$ is convex $\iff$ $\langle y - x, \nabla f(y) - \nabla f(x) \rangle \geq 0$ for all $x, y \in \text{dom}(f)$. This means that $\nabla f$ is monotone.*

**Theorem 3.5.** *Also, and additionally assuming $\nabla^2 f$ exists, $f$ is convex $\iff$ $\nabla^2 f \succeq 0$.*

**Remark 3.3.** If you want strict convexity change $\geq 0$ to $>$ above except with the Hessian. Note that $f(x) = x^4$ is strictly convex but $f''(x) = 12x^2 = 0$ at $x = 0$.

**Remark 3.4.** For strong convexity, $f$ is $\mu$-strongly convex $\iff$ for all $x \in \text{dom}(f)$, we have $\nabla^2 f(x) \succeq \mu I$.

*Proof.* Observe that if $g(x) = f(x) - \frac{\mu}{2}\|x\|^2$, then $\nabla^2 g(x) = \nabla^2 f(x) - \mu I$. □

**Definition 3.10** (Subdifferential)**.** Let $f : \mathbb{R}^n \cup \infty$ be proper. The **subdifferential** is

$$\partial f(x) = \{d \mid f(y) \geq f(x) + \langle d, y - x\rangle \forall y \in \mathbb{R}^n\}$$

**Theorem 3.6.** If $f$ is proper and convex, then $\partial f(x) \neq \varnothing$ for all $x \in \text{relint}(\text{dom}(f))$.

**Theorem 3.7.** Assume $f$ is proper and convex and $x \in \text{dom}(f)$. If $f$ is continuous at $x$, then $f$ is differentiable at $x \iff \partial f(x)$ is a singleton containing the gradient.

**Example 3.3.** Consider $f(x) = |x|$. Then $\partial f(x) = \begin{cases} \{-1\} & x < 0 \\ \{1\} & x > 0 \\ [-1, 1] & x = 0 \end{cases}$.

**Theorem 3.8** (Fermat's Rule)**.** $x$ minimizes $f(x) \iff f(y) \geq f(x)$ for all $y \iff 0 \in \partial f(x)$.

**Theorem 3.9.** If $f$ is proper, then $\partial f$ is monotone so that for all $x, y$ and for all $\partial_x \in \partial f(x)$ and $\partial_y \in \partial f(y)$, we have $\langle x - y, \partial_x - \partial_y\rangle \geq 0$.

**Remark 3.5.** Note that $\partial(f + g)(x)$ does not necessarily equal $\partial f(x) + \partial g(x)$.

**Theorem 3.10.** Suppose $f, g \in \Gamma_0(\mathbb{R}^n)$. If $\text{relint}(\text{dom}(g) \cap \text{relint}(\text{dom}(f))) \neq \varnothing$, then $\partial(f + g)(x) = \partial f(x) + \partial g(x)$ for all $x$. We call this the (French form) of a Constraint Qualification; it is otherwise known as Slater's condition.

**Example 3.4.** If $g(x) = \delta_C(x)$ and $f(x) = \delta_D(x)$ where $C \cap D = \varnothing$, then $f + g$ is not proper.

$$\partial g(x) = \{d \mid \forall y, g(y) \geq g(x) + \langle d, y - x\rangle\} = \begin{cases} \varnothing & x \notin C \\ \{d \mid \forall y \in C, \langle d, y - x\rangle \leq 0\} & x \in C \end{cases}$$

We call this the normal cone.

**Example 3.5.** Two sets $C = B_1(1)$ and $D = B_1(-1)$ are closed balls of radius 1 in the Cartesian plane centered at 1 and $-1$ respectively. (The circles touch at $x = 0$). If $g = \delta_C$, what is $\partial g(0)$? It's the normal vector pointing left.

$$\partial g(0) = \mathbb{R}_- \times \{0\}, \quad \partial f(0) = \mathbb{R}_+ \times \{0\}$$

We calculate $\partial g(0) + \partial f(0) = \mathbb{R} \times \{0\}$, whereas $\partial(f + g)(0) = \partial(\delta_{\{0\}}) = \mathbb{R}^2$

## 3.2 Further Properties of Convex functions

$F : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous if $\exists L > 0$ such that for all $x, y \in \mathbb{R}^n$, $\|F(x) - F(y)\| \leq L\|x - y\|$. A very standard assumption is $F = \nabla f$ is $L$-Lipschitz where $\nabla f(x) : \mathbb{R}^n \to \mathbb{R}^n$ and $\nabla^2 f(x) : \mathbb{R}^n \to \mathbb{R}^{n \times n}$. ($\langle \nabla f(x), d\rangle$ versus $\langle d, \nabla^2 f(x) d\rangle$.

**Theorem 3.11.** Assume $\nabla^2 f$ exists. Then $f$ is a L-lipschitz continuous $\iff \nabla^2 f(x) \preceq LI$.

**Theorem 3.12** (Descent Lemma)**.** If $\nabla f$ is L-Lipschitz, then $f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|x - y\|^2$

*Proof.* Taylor expanding $f$ about $x$ gives $f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\langle y - x, \nabla^2 f(\xi)(y - x)\rangle$, but by assumption $\nabla^2 f(\xi) - L \preceq 0$. □

**Theorem 3.13.** If $f$ is $\mu$-strongly convex, then $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|x - y\|^2$.

**Remark 3.6.** If $x \in \mathbb{R}^n$, then $\log(x) = \sum_{i=1}^n \log(x_i)$.

**Example 3.6.** The following are convex functions:

- $f(x) = \|x\|$
- $f(x) = \max\limits_i \{x_1, \ldots, x_n\}$
- $f(x, y) = x^2/y,\ y > 0$. This is the 1D version of the (convex) matrix fractional function $f(x, Y) = x^T Y^{-1} x$ for $Y \succeq 0$.
- The linear fractional function $g(x) = \frac{Ax+b}{c^T x+d}$ where $c^T x + d > 0$.
- The log-sum-exp function $f(x) = \frac{1}{\alpha} \log\left(\sum_{i=1}^p e^{\alpha x_i}\right)$, $\alpha \geq 0$. Note that for a fixed $x$, $f(x) \xrightarrow{\alpha \to \infty} \max\limits_i \{x_1, \ldots, x_n\}$.
- The geometric mean $f(x) = \left(\prod_{i=1}^n x_i\right)^{1/n}$ is concave.
- $\log(\det(X))$ is concave on $X \in S_{++}^n$.

**Remark 3.7.** Convexity implies that all sub-level sets $\{x : f(x) \leq \alpha\}$ are convex, but the converse is not necessarily true (look at a simple concave function!)

**Definition 3.11** (Quasi-convex)**.** A function is **quasi-convex** if all sub-level sets are convex.

**Example 3.7.** The number of nonzeros $\|x\|_0$ (in 1D) is quasi-convex but not convex. (Not in higher dimensions!)

**Example 3.8.** In sparse regression, want to solve the NP-hard problem

$$\min_x \frac{1}{2}\|Ax - b\|^2 \text{ s.t. } \|x\|_0 \leq k$$

**Remark 3.8.** B&V claim that quasi-convex optimization is as easy as convex optimization.

$$\min_x f(x) \iff \min_{(x,t)} t \text{ s.t. } f(x) \leq t$$

Here $f(x)$ is quasi-convex and $f(x) \leq t$ is a convex set.

**Remark 3.9.** Unlike convex functions, quasi-convex functions are not additive.

## 3.3 Preserving Convexity

- If $f, g$ are convex, so is $\alpha f + \beta g$ where $\alpha, \beta \geq 0$.
- Consider the composition $f = h \circ g$. Then, $f$ is convex if either $g$ is convex and $h$ is convex and non-decreasing or if $g$ is concave and $h$ is concave and non-increasing.
- Consider $g : \mathbb{R}^n \to \mathbb{R}^m$ and $h : \mathbb{R}^m \to \mathbb{R}$. If $h$ is convex and $g$ is affine, then $f = h \circ g$ is convex.

**Remark 3.10.** The maximum $f(x) = \max\limits_{t \in I} f_t(x)$ where $f_t$ are convex is again convex. The minimum does not behave as nicely.

**Theorem 3.14.** If $f(x, y)$ is jointly convex in $(x, y)$ and $C \neq \varnothing$ is a convex set, then $g(x) = \inf\limits_{y \in C} f(x, y)$ is convex (assuming it's never $-\infty$).

**Example 3.9.** $x \mapsto \operatorname{dist}(x, C)$ for a convex set $C$ is convex. Note that $f(x,y) = [I, -I][x,y]^T = \|x-y\|$ is convex from an earlier theorem.

**Definition 3.12** (Fenchel-Legendre Transform)**.** $f^*(y) = \sup_x \langle y, x\rangle - f(x)$. This is trivially convex in $y$ and we're taking a supremum; $f^*$ is always convex even if $f$ isn't.

**Example 3.10.** Let $f$ be differentiable and strictly convex. (Thus $f'$ is increasing). Our problem is $-\left(\inf_x f(x) - y \cdot x\right)$. Taking the derivative gives $f'(x) - y \implies y = f'(x) \implies x = (f')^{-1}(y)$. Substituting this $x$ into the objective gives the Legendre transform.

**Example 3.11.** For $f(x) = |x|$, we calculate $f^*(y) = \delta_{\{y||y|\leq 1\}}$.

**Example 3.12.** Recall Holder's inequality: $|\langle x, y\rangle| \leq \|x\|_p \|y\|_q$ for $\frac{1}{p} + \frac{1}{q} = 1$. For $f(x) = \|x\|_p$ in $\mathbb{R}^n$, we have

$$f^*(y) = \sup_x \langle y, x\rangle - f(x) \leq \|x\|_p - \|y\|_q - f(x) = \|x\|_p\|y\|_q - \|x\|_p$$

and if $\|y\|_q \leq 1$ this is $\leq 0$, and since $f^*(y) \geq 0$, we get equality. If $\|y\|_q > 1$, then $f^*(y) = \infty$. Therefore $f^*(y) = \delta_{\{y|\|y\|_q \leq 1\}}$.

**Theorem 3.15.** If $f$ is proper, then $f$ is lower semicontinuous and convex $\iff f = f^{**}$.

**Remark 3.11.** There are special rules for finding the conjugate of $f \circ A$ where $A$ is linear. If $f(u,y) = f_1(u) + f_2(v)$ is separable, then $f^*(x,y) = f_1^*(x) + f_2^*(y)$. Not true for general sums. For $f(x) = \frac{1}{2}\|x\|_2^2$, we have $f^*(x) = f(x)$.

**Theorem 3.16** (Fenchel's Inequality)**.** For all $x, y$ $f(x) + f^*(y) \geq \langle x, y\rangle$. Also called Young's inequality if the functions are differentiable.

*Proof.* Immediate: $f^*(y) = \sup_x \langle y, x\rangle - f(x) \geq \langle x, y\rangle - f(x)$. □

**Corollary 3.1.** Suppose $f \in \Gamma_0(\mathbb{R}^n)$. Then $f(x) + f^*(y) = \langle y, x\rangle \iff y \in \partial f(x) \iff x \in (\partial f)^{-1}(y)$.

**Corollary 3.2.** Suppose $f \in \Gamma_0(\mathbb{R}^n)$. Then $f(x) + f^*(y) = \langle x, y\rangle \iff x \in \partial f^*(y)$.

**Corollary 3.3.** Suppose $f \in \Gamma_0(\mathbb{R}^n)$. Then $\partial f^* = (\partial f)^{-1}$.

**Remark 3.12.** When do minimizers exist and when are they unique? Our set of minimizers $S = \arg\min_x f(x)$ is convex (if $f$ is convex).

**Corollary 3.4** (11.8 B+C)**.** Suppose $f : \mathbb{R}^n \to (-\infty, \infty]$ is proper. If $f$ is strictly convex, then there exists at most 1 minimizer.

**Theorem 3.17.** If $f$ is convex and $g$ is strictly convex, then $f + g$ is strictly convex. If $f$ is convex and $g$ is strongly convex, then $f + g$ is strongly convex.

**Definition 3.13** (Coercive)**.** $f : \mathbb{R}^n \to [-\infty, \infty]$ is **coercive** if $\lim_{\|x\|\to\infty} f(x) = \infty$.

**Theorem 3.18.** $f$ is coercive $\iff$ all sub-level sets are bounded.

**Theorem 3.19.** If $f \in \Gamma_0(\mathbb{R}^n)$, then $f$ is coercive $\iff$ at least one sub-level set is bounded and nonempty.

**Theorem 3.20.** Suppose $f \in \Gamma_0(\mathbb{R}^n)$ and $C$ is any closed, convex set such that $C \cap \operatorname{dom}(f) \neq \varnothing$. If either

- $f$ is coercive

- $C$ is bounded

are true, then there exists a minimizer of $f$ over $C$.

**Corollary 3.5.** Suppose $f \in \Gamma_0(\mathbb{R}^n)$. If $f$ is strongly convex, then $f$ is strictly convex and $f$ is coercive; i.e. there exists a unique minimizer.

**Remark 3.13.** If $f$ is coercive and convex and $g$ is convex, if $f+g$ coercive? Recall that $f(x) = \|x\| \iff f^*(y) = \delta_{\{y \mid \|y\|_* \leq 1\}}(y)$. This motivates the counterexample $\|x\| - \langle x, y\rangle$ which is not coercive.

**Definition 3.14** (Supercoercive)**.** $f : \mathbb{R}^n \to [-\infty, \infty]$ is **supercoercive** if $\lim\limits_{\|x\| \to \infty} \frac{f(x)}{\|x\|} = \infty$.

**Theorem 3.21.** If $f$ is supercoercive and convex and $g$ is convex, then $f+g$ is supercoercive.

**Theorem 3.22.** Strong convexity $\implies$ supercoercive.

**Example 3.13.** Find $\operatorname*{arg\,min}\limits_x \frac{1}{2}\|x-y\|_2^2 + \|x\|_1$. Since the objective is strongly convex, there is a unique minimizer $p$. Since the Frobenius norm is squared, the objective is $\frac{1}{2}\sum |x_i - y_i|^2 + \sum |x_i|$ and hence separable. So now we do scalar minimization. We want 0 inside the subdifferential of $\operatorname*{arg\,min}\limits_{x \in \mathbb{R}} \frac{1}{2}(x-y)^2 + |x|$ at $p$:

$$0 \in p - y + \begin{cases} \{-1\} & p < 0 \\ \{1\} & p > 0 \\ [-1, 1] & p = 0 \end{cases}$$

So $p = y - \partial|p|$. If $p > 0$, then $\partial|p| = 1$ so $p = y - 1$ and thus need $y > 1$. Similarly if $p < -$, then $p = y + 1$ and need $y < -1$. If $p = 0$, then $0 = y - \partial|p|$ so need $y \in [-1, 1]$.

$$p(y) = \begin{cases} y - 1 & y > 1 \\ y + 1 & y < 1 \\ 0 & y \in [-1, 1] \end{cases} = \operatorname{sign}(y)\lfloor |y| - 1 \rfloor_+$$

The final operation is known as soft-thresholding or shrinkage.

**Definition 3.15** (Proximity Operator)**.** The **prox** of $f$ is $y \mapsto \operatorname*{arg\,min}\limits_x \frac{1}{2}\|x - y\|^2 + f(x)$.

**Theorem 3.23.** If $f \in \Gamma_0(\mathbb{R}^n)$, then prox exists and is single-valued.

**Example 3.14.** If $f = \delta_C(x)$, then $\operatorname*{prox}\limits_f(y) = \operatorname*{arg\,min}\limits_{x \in C} \frac{1}{2}\|x - y\|^2$ is a Euclidean projection.

**Remark 3.14.** Observe that

$$\begin{aligned} p = \operatorname*{arg\,min}_x \frac{1}{2}\|x - y\|^2 + f(x) &\iff 0 \in \partial f(p) + p - y \\ &\iff y \in (I + \partial f)(p) \\ &\iff p = (I + \partial f)^{-1}(y) \end{aligned}$$

**Theorem 3.24** (Moreau's Decomposition)**.** Suppose $f \in \Gamma_0(\mathbb{R}^n)$. Then $y = \operatorname*{prox}\limits_f(y) + \operatorname*{prox}\limits_{f^*}(y)$. Moreover,

$$f\left(\operatorname*{prox}_f(y)\right) + f^*\left(\operatorname*{prox}_f(y)\right) = \langle \operatorname*{prox}_f(y), \operatorname*{prox}_f(y) \rangle$$

**Remark 3.15.** $\underset{f+g}{\text{prox}} \neq \underset{f}{\text{prox}} + \underset{g}{\text{prox}}$

# 4 Gradient Descent

Assume $f \in \Gamma_0(\mathbb{R}^n)$ and $f$ is real-valued where $\nabla f$ is $L$-Lipschitz. To minimize $f$, our first attempt is the taylor expansion

$$x_{k+1} = \arg\min_x f(x_k) + \langle f(x_k), x - x_k \rangle = \pm\infty$$

A second attempt is the (convex) quadratic taylor expansion

$$x_{k+1} = \arg\min_x f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}\langle x - x_k, \nabla^2 f(x_k)(x - x_k) \rangle$$

Guided by Fermat's principle, we check when 0 is in the subdiffferential:

$$0 = \nabla f(x_k) + \nabla^2 f(x_k)(x - x_k)$$

Solve this linear equation $\nabla^2 f(x_k)x = \nabla^2 f(x_k)x_k - \nabla f(x_k) \implies x = x_k - \nabla^2 f(x_k)\nabla f(x_k)$. Newton's method! This is the generalization of rootfinding a real-valued function $F(x) = 0$ via $x_{k+1} = x_k - \frac{F(x_t)}{F'(x_t)}$.
But we're still not at gradient descent. Our third and final attempt comes from bounding our function above by a quadratic depending on the Lipschitz constant $L$. Fact: $f(x) \leq Q_k(x)$, where

$$x_{k+1} = Q_k(x) = \arg\min f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}\langle x - x_k, LI(x - x_k) \rangle$$

The proof of this fact comes from the fact that $\nabla f$ L-Lipschitz $\implies \nabla^2 f(x_k) \preceq LI$. This leads to the Gradient Descent updates

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$

**Remark 4.1.** In optimization, Newton's method is a second order method but for root-finding it is first order.

**Remark 4.2.** Frank-Wolfe, or Conditional Gradient Descent, for $\min_{x \in X} f(x)$ where $C$ is compact or $\min_x f(x) + g(x)$, where $g$ is supercoercive.

$$S = \arg\min_x f(x_k) + \langle f(x_k), x - x_k \rangle$$

Since $C$ is compact, the function values don't blow up to $\pm\infty$.

**Remark 4.3.** Gradient Descent is a specific type of Majorization-Minimization (MM) algorithm where

$$\begin{cases} f(x) \leq Q_k(x) \\ f(x_k) = Q_k(x_k) \\ x_{k+1} \in \arg\min_x Q_k(x) \end{cases}$$

Note that EM algorithm and difference-of-convex (DC) programs fall under this MM framework.

## 4.1 Convergence Analysis

(Following Nesterov and Vanderberghe's notes EE236C.) Consider $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$. Let $x^*$ be the optimal solution.

$$\begin{aligned}
f(x_{k+1}) &\le f(x_k) + \langle f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \quad \text{(Descent Lemma)}\\
&= f(x_k) + \langle \nabla f(x_k), -\frac{1}{L}\nabla f(x_k)\rangle + \frac{1}{2L}\|\nabla f(x_k)^2\|^2\\
&= f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 \quad \text{(Descent Property!)}\\
&\le f(x^*) + \langle \nabla f(x_k), x_k - x^*\rangle - \frac{1}{2L}\|\nabla f(x_k)\|^2 \quad \text{(Convexity)}\\
&= f(x^*) + \frac{L}{2}\left(\|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L}\nabla f(x_k)\|^2\right)\\
&= f(x^*) + \frac{L}{2}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right)
\end{aligned}$$

Using a telescoping series,

$$\begin{aligned}
\frac{1}{k}\sum_{i=1}^{k} f(x_i) - f(x^*) &\le \frac{L}{2k}\sum_{i=1}^{k}\left(\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2\right)\\
&= \frac{L}{2k}\left(\|x_0 - x^*\|^2 - \|x_k - x^*\|^2\right) \le \frac{L}{2k}\|x_0 - x^*\|^2
\end{aligned}$$

We can also bound the LHS of the above equation using the Descent Property:

$$f(x_k) - f^* = \frac{1}{k}\sum_{i=1}^{k} f(x_k) - f^* \le \frac{1}{k}\sum_{i=1}^{k} f(x_i) - f(x^*)$$

and putting the two bounds together gives

$$e_k = f(x_k) - f^* \le \frac{L}{2k}\|x_0 - x^*\|^2$$

**Definition 4.1** (Linear Convergence)**.** We have **linear convergence** of $(e_k)$ if there exists $\rho < 1$ such that $e_{k+1} \le \rho e_k$.

**Definition 4.2** (Sublinear Convergence)**.** We have **sublinear convergence** of $(e_k)$ if there does not exist $\rho < 1$ such that $e_{k+1} \le \rho e_k$ but there exists $\rho_k \to 1$ such that $e_{k+1} \le \rho_k e_k$.

**Definition 4.3** (Superlinear Convergence)**.** We have **superlinear convergence** of $(e_k)$ if there does not exist $\rho < 1$ such that $e_{k+1} \le \rho e_k$ but there exists $\rho_k \to 0$ such that $e_{k+1} \le \rho_k e_k$.

**Example 4.1.** Here's an example of linear convergence: $e_k = (0.9)^k$. Here's an example of sublinear convergence. Take $e_k = \frac{1}{k}$ or $e_k = \frac{1}{\sqrt{k}}$ or $e_k = \frac{1}{k^2}$ with the function $\sum_{k=1}^{\infty} e_k$. Here's an example of superlinear convergence: $e_k = (0.9)^{2^k}$.

**Remark 4.4.** How long to reach an $\epsilon$-solution? Gradient Descent converges like $O(\frac{1}{k})$ iterations, or $f(x_k) - f(x^*) \le \frac{c}{k} < \epsilon$. Need $k > c/\epsilon$. That is $O(\frac{1}{\epsilon})$ or $O(\frac{1}{k})$. Note that $O(\frac{1}{k^2}) \iff O(\frac{1}{\sqrt{\epsilon}})$.

Convergence rates for subgradient descent and stochastic gradient descent are $O(\frac{1}{\epsilon^2})$. Under this rate, If $x$ is an $\epsilon = 1$ solution and we want an $\epsilon = 10^{-2}$. This requires $10,000T$ iterations, where $T$ is the time it took you to get to the $\epsilon = 1$ solution. If we have $O(1/\epsilon)$ like gradient descent, we need $100T$ more iterations. If we have $O(1/\sqrt{\epsilon})$ like in accelerated gradient descent, requires $10T$ more iterations. Under a linear convergence (e.g. gradient descent with extra assumption), this requires $2 \cdot C$ more. Under quadratic method (e.g. Newton's method), we require only 1 more iteration to reach an $\epsilon = 10^{-2}$ solution.

**Remark 4.5.** First order methods have access to $\nabla f(x)$ and $f(x)$. Can improve gradient descent $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$ by using an exact line search. For example, take $f(x) = \frac{1}{2}\|Ax - b\|^2$

**Theorem 4.1.** For any fixed 1st order method, there exists $f \in \Gamma_0(\mathbb{R}^n)$ with $\nabla f$ $L$-Lipschitz such that

$$f(x_k) - f(x^*) \geq \frac{3}{32} L \frac{\|x_0 - x^*\|^2}{k^2}$$

and $\|x_k - x^*\| \geq \frac{1}{8}\|x_0 - x^*\|^2$ for $k \leq \frac{1}{2}(n-1)$.

*Proof.* Assume $x_k \in \text{span}\{x_0, \nabla f(x_0), \nabla f(x_1), \ldots, \nabla f(x_{k-1})\}$ and WLOG take $x_0 = 0$. Let $A = \text{tridiag}(-1, 2, -1)$ and define $f(x) = \frac{L}{4}\left(\langle x, Ax\rangle - \langle e_1, x\rangle\right)$. Observe that $\nabla f(x) = \frac{L}{4}(Ax - e_1)$ and $\nabla^2 f(x) = A \succeq 0$ by Gershgorin Theorem. Now,

$$\text{span}\{x_0, \nabla f(x_0), \nabla f(x_1), \ldots, \nabla f(x_{k-1})\} = \{e_1, e_2, \ldots\}$$

Rest omitted. □

**Theorem 4.2** (Nesterov Acceleration)**.** Pick $x_0$ and then set $y_0 = x_0$. Nesterov's optimal method (1983) is

$$\begin{cases} x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k) \\ y_{k+1} = x_{k+1} + \frac{k+1}{k+4}(x_{k+1} - x_k) \end{cases}$$

This converges in $O(\frac{1}{k^2})$.

**Remark 4.6.** A momentum method looks like

$$x_{k+1} = x_k - t_k \nabla f(x_k) + s_k(x_k - x_{k-1})$$

Might correct the gradient descent zig-zag behavior of e.g. $\frac{1}{2}(x_1^2 + 10x_2^2)$ compared to $\frac{1}{2}\|x\|^2$.

**Example 4.2.** Consider $f(x, y) = \begin{cases} \sqrt{x^2 + \gamma y^2} & |y| \leq x \\ \frac{x + \gamma|y|}{\sqrt{1+\gamma}} & |y| \geq x \end{cases}$
Note that the function is not bounded below as $x \to -\infty$ in $R_2$. Let $x_0 = (\gamma, 1)$. Then the gradient descent with the exact line search converges to $(0, 0)$ rather than this global minimum.

## 4.2 Inequalities

- The Descent Lemma $f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|x - y\|^2$ if $\nabla f$ is L-Lipschitz. (No convexity needed)
- $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|x - y\|^2$ if $f$ is $\mu$-strongly convex.

These two imply that

$$\begin{cases} \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \\ \mu\|x - y\|^2 \\ \frac{\mu L}{\mu + L}\|x - y\|^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|^2 \end{cases} \leq \langle \nabla f(x) - \nabla f(y), x - y\rangle \leq \begin{cases} L\|x - y\|^2 \\ \frac{1}{\mu}\|\nabla f(x) - \nabla f(y)\|^2 \end{cases}$$

**Definition 4.4** (Polyak-Lojasiewicz)**.** Let $x^* \in \arg\min_x f(x)$. We say that $f$ satisfies the Polyak-Lojasiewicz inequality if there exists $\mu > 0$ such that for all $x$, we have

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f(x^*))$$

**Theorem 4.3.** If $f$ is $\mu$-strongly convex, then $f$ satisfies the PL inequality with constant $\mu$.

**Theorem 4.4.** Suppose $f$ is $\mu$-PL and $\nabla f$ is $L$-Lipschitz. Then if $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$,

$$f(x_k) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right)(f(x_{k-1}) - f(x^*)) \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f(x^*))$$

This is a linear rate of convergence depending upon the condition number $\left(1 - \frac{\mu}{L}\right)^{-k}$.

*Proof.* In the previous proof of gradient descent we showed that the Descent property is satisfied; i.e. $f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L}\|\nabla f(x_k)\|^2$. From our inequality, we have

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L}\|\nabla f(x_k)\|^2 \leq -\frac{\mu}{L}(f(x_k) - f(x^*)) \qquad \square$$

**Remark 4.7.** Under strong convexity, we get that the sequence of objective $f(x_k)$ converges, but also that the sequence $(x_k)$ converges.

## 5 Convex Optimization Problems

This follows Chapter 4 in B&V, primarily 4.1 and 4.2. Define the Non-linear programming (NLP) problem

$$p^* = \begin{cases} \min f_0(x) \\ f_i(x) \leq 0, \quad i \in \{1, 2, \ldots, m\} \\ h_i(x) = 0 \quad i \in \{1, 2, \ldots, p\} \end{cases}$$

where $p^* = \infty$ if infeasible and $p^* = -\infty$ if unbounded.
For a convex function $f$, the corresponding problem is

$$p^* = \begin{cases} \min f_0(x) \\ f_i(x) \leq 0, \quad i \in \{1, 2, \ldots, m\} \\ a_i^T x = b_i \quad i \in \{1, 2, \ldots, p\} \end{cases}$$

where $f_i$ are convex. Note that a constrained optimization problem must be over an affine set, and recall that an affine transformation preserves convexity. The epigraph trick refers to:

$$p^* = \begin{cases} \min_x f_0(x) \\ f_i(x) \leq 0, \quad i \in \{1, 2, \ldots, m\} \\ h_i(x) = 0 \quad i \in \{1, 2, \ldots, p\} \end{cases} = \begin{cases} \min_t t \\ f_0(x) \leq t \\ f_i(x) \leq 0, \quad i \in \{1, 2, \ldots, m\} \\ h_i(x) = 0 \quad i \in \{1, 2, \ldots, p\} \end{cases}$$

Here $h_i$ are affine.

## 5.1 Tricks

There are a few tricks we can commonly use.

- Above, for the epigraph trick, we introduced slack variables.
- For $\|x\|_1$ we can write $\|x\|_1 = \mathbf{1}^T(x^+ - x^-)$, where $x = x^+ - x^-,\ x^+ \geq 0, x^- \geq 0$
- $\min\limits_x f(x) + g(x)$ is equivalent to $\min\limits_{x,z} f(x) + g(z)$ such that $x = z$.
- $\min\limits_{x,y} f(x,y) = \min\limits_x \left(\min\limits_y f(x,y)\right)$. This is always true, unlike e.g. weak duality $\sup\limits_y \inf\limits_x f(x,y) \leq \inf\limits_x \sup\limits_y f(x,y)$.

## 5.2 Optimality Conditions

**Theorem 5.1.** Consider $\min\limits_{x \in C} f(x)$, where $f$ and $C$ are convex. If $\nabla f$ exists, then $x^*$ is optimal $\iff$ $x^* \in C$ and the Euler property is satisfied: for all $y \in C$, $\langle \nabla f(x^*), y - x^* \rangle \geq 0$.

*Proof.* ($\Leftarrow$) By convexity of $f$,

$$f(y) \geq f(x^*) + \langle \nabla f(x^*), y - x^* \rangle \geq f(x^*)$$

($\Rightarrow$) For a contradiction, suppose there is $y \in C$ such that $\langle \nabla f(x^*), y - x^* \rangle < 0$. Define

$$\phi(t) = f(x^* + t(y - x^*))$$

so that

$$\phi'(0) = \langle \nabla f(x^*), y - x^* \rangle < 0$$

by assumption. But

$$\phi'(0) = \lim_{h \to 0^+} \frac{\phi(h) - \phi(0)}{h} < 0$$

Therefore for some $h$, we have $\phi(h) < \phi(0) = f(x^*)$, a contradiction. □

**Remark 5.1.** If $f$ is not convex, such a point isn't a global minimum necessarily. We call it a *stationary point.*

**Remark 5.2.** The Euler inequality is a special case of an *variational inequality.* Variational Inequality (VI): find $x \in C$ such that $\langle F(x), y - x \rangle \geq 0$ for all $y \in C$ where $F : \mathbb{R}^n \to \mathbb{R}^n$. If $F = \nabla f$, this is the Euler inequality. A special VI is the linear complementary problem (LCP): find $x \geq 0$ such that $x \perp F(x) = Ax + b$ with $Ax + b \geq 0$. (This is like KKT for linear programming. Either $x_i = 0$ or $(Ax + b)_i = 0$ or both.)

**Remark 5.3.** Let's consider some special cases.

- Unconstrained $C = \mathbb{R}^n$. Pick $y = x^* - \epsilon \nabla f(x^*)$ for $\epsilon > 0$. The Euler inequality becomes

  $$-\epsilon \|\nabla f(x^*)\|^2 \geq 0 \implies \nabla f(x^*) = 0$$

  This is a special case of Fermat's rule.

- Equality constraints $C = \{y \mid Ay = b\}$. The Euler inequality: $Ax^* = b$ for all $y$ such that $Ay = b$,

  $$\langle f(x^*), y - x^* \rangle \geq 0$$

Choose $y = x^* + v$ where $v \in \text{Null}(A)$. Thus

$$\langle f(x^*), y - x^* \rangle = \langle \nabla f(x^*), v \rangle \geq 0, \quad \forall v \in \text{Null}(A)$$

and since the gradient a subspace $-v$ is in Null(A); i.e.

$$\langle f(x^*), y - x^* \rangle = \langle \nabla f(x^*), v \rangle = 0, \quad \forall v \in \text{Null}(A)$$

This is equivalent to

$$\nabla f(x^*) \perp \text{Null}(A) \iff \nabla f(x^*) \in \text{Range}(A^T) \iff \nabla f(x^*) \perp \text{Null}(A) \iff \nabla f(x^*) + A^T v = 0, \, \forall v \in \text{Null}(A)$$

and we recognize the final equation as a Lagrange multiplier.

## 5.3 Special Types of Optimization Problems

1. Linear Programming (LP).

$$\begin{cases} \min \langle c, x \rangle \\ Gx \leq h \\ Ax = b \end{cases} \quad \text{or} \quad \begin{cases} \min \langle c, x \rangle \\ x \geq 0 \\ Ax = b \end{cases} \quad \text{or} \quad \begin{cases} \min \langle c, x \rangle \\ Ax = b \end{cases}$$

2. Quadratic Problem (QP).

$$\begin{cases} \min \frac{1}{2} \langle x, Px \rangle + \langle q, x \rangle + r \\ Gx \leq h \\ Ax = b \end{cases}$$

Convex $\iff$ $P \succeq 0$. Notice that we have a quadratic objective but linear constraints.

*S-Lemma*: A special case where we can solve nonconvex quadratic optimization problems.

The closely related QCQP (Quad Constrained QP) looks like:

$$\begin{cases} \min \frac{1}{2} \langle x, Px \rangle + \langle q, x \rangle + r \\ \frac{1}{2} \langle x, P_i x \rangle + \cdots \leq 0 \\ Ax = b \end{cases}$$

The second-order cone program (SOCP) is always convex:

$$\begin{cases} \min \langle c_0, x \rangle \\ \|A_i x + b\|_2 \leq \langle c_i, x \rangle + d_i, \quad i = 1, 2, \ldots, m \\ Fx = b \end{cases}$$

3. Conic Program (CP).

$$\begin{cases} \min\limits_x f_0(x) \\ f_i(x) \preceq_{K_i} 0 \quad i = 1, 2, \ldots \\ Ax = b \end{cases}$$

$\min_x f_0(x)$ where $f_0$ is convex. The conic form of the above equations is $f_i(x) \preceq_{K_i} 0$; i.e. $f_i(x) \in K_i$. Here $f_i(x)$ are $K_i$-convex, meaning that

$$f(tx + (1-t)y) \preceq_K tf(x) + (1-t)f(y)$$

Note that we require a proper cone (closed, convex, solid, pointed). An alternate form where all $f_i$ for $i = 0, 1, 2, \ldots$ are linear is

$$\begin{cases} \min\langle c, x\rangle \\ Fx + g \preceq_K 0 \\ Ax = b \end{cases} \quad \text{or} \quad \begin{cases} \min\langle c, x\rangle \\ x \succeq_K 0 \\ Ax = b \end{cases}$$

Note that if $K_1$ and $K_2$ are proper cones, then $K_1 \times K_2$ is a proper cone, hence the single cone $K$ in the two generalized inequalities above.

**Remark 5.4.** In nonlinear nonnegative least squares, $\min_{x\geq 0} \frac{1}{2}\|Ax - b\|^2$. Here, the quadratic term looks like $\frac{1}{2}\langle x, A^T Ax\rangle$ which is positive. (Since the Frobenius norm is squared, we always have PD).

**Remark 5.5.** To encode $X \succeq 0$ and $Y \succeq 0$, can use $\begin{pmatrix} X & 0 \\ 0 & Y \end{pmatrix} \succeq 0$. Alternatively, $Z = \begin{pmatrix} X & W \\ W^T & Y \end{pmatrix}$ where $Z \succeq 0$. Can force $w = 0$ with a linear constraint. (Note: not useful in software!)

**Remark 5.6.** What is a linear function on matrices?

- Vector case: Consider a linear operator $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^m$; defined by $\mathcal{A}(x) = Ax$.
- Matrix case: Consider a linear operator $\mathcal{A} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$; defined by $\mathcal{A}X = A\text{vec}(X)$, where $A$ is $m \times (n_1n_2)$. Call the $i$th row of $A$ by $a_i^T \in \mathbb{R}^{n_1n_2\times 1}$. Define $A_i = \text{mat}(a_i)$.

$$(A\text{vec}(x))_i = a_i^T \text{vec}(X) = \langle A_i, X\rangle = \text{trace}(A_i^T X)$$

**Example 5.1.** An example of a conic program is the semi-definite program (SDP).

$$\begin{cases} \min_{X=X^T\succeq 0} \langle C, X\rangle \\ \langle A_i, X\rangle = b_i, \quad i = 1, 2, \ldots, m \end{cases}$$

This is the dual problem of the Linear Matrix Inequality (LMI):

$$\begin{cases} \min_{x\in\mathbb{R}^n} \langle c, x\rangle \\ Ax = b \\ G + \sum_{i=1}^m x_i F_i \preceq 0 \end{cases}$$

**Definition 5.1** (Adjoint)**.** The **adjoint** of $\mathcal{A} : \mathbb{R}^{n_1 n_2} \to \mathbb{R}^m$ is defined by $\mathcal{A}^* : \mathbb{R}^m \to \mathbb{R}^{n_1n_2} : y \mapsto \sum_{i=1}^m y_i A_i$.

**Remark 5.7.** Here's some common software for these types of problems.

- SDP: SeDuMi, SDPT3, Mosek
- QCQP: Mosek, Gurobi
- LP: Gurobi, CPLEX

### 5.4 Schur Complement

Given a matrix $\begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$, its Schur complement is $S = C - B^T A^{-1} B$. Note that solving the sadlle point system

$$\begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} c \\ d \end{pmatrix}$$

requires solving a $(2n) \times (2n)$ system for about $8 \cdot O(n^3)$. Instead, solve $Ax + By = c \implies x = A^{-1}(c - By)$. Then substituting into $B^T x = d \implies BA^{-1}(c - By) = d \implies B^T A^{-1} By = \tilde{d}$ which is an $n \times n$ system.

**Theorem 5.2.** Suppose $A \succ 0$. Then $\begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \succeq 0 \iff C - B^T A^{-1} B \succeq 0$.

**Theorem 5.3** (Matrix Inversion Lemma)**.** Here, $A$ and $C$ are square matrices of different dimension, and $U, V$ are rectangular of appropriate dimension. Also known as Sherman Morriwon Woodbury identity.

$$(A + UCV^T)^{-1} = A^{-1} - A^{-1}U(C^{-1} + V^T A^{-1} U)^{-1} V^T A^{-1}$$

## 6 Duality

This follows Chapter 5 in B&V. Consider the problem

$$(P) = \begin{cases} \min\limits_{x \in D} f_0(x) \\ f_i(x) \leq 0, \quad i \in \{1, 2, \ldots, m\} \\ h_i(x) = 0 \quad i \in \{1, 2, \ldots, p\} \end{cases}$$

where $D$ is the domain of $f$ and $f$ is not necessarily convex.

**Definition 6.1** (Lagrangian)**.** The **Lagrangian** of (P) is the function

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{\infty} \lambda_i f_i(x) + \sum_{i=1}^{\infty} \nu_i h_i(x)$$

We refer to $\lambda \in \mathbb{R}^m$ and $\nu \in \mathbb{R}^p$ as dual variables or Lagrange multipliers.

**Remark 6.1.** There can be many Lagrangians depending upon how you formulate the optimization problem!

**Remark 6.2.** To create the Lagrangian for an unconstrained problem $\min\limits_x f(x) + g(x)$, consider rewriting the problem as $\min\limits_{x,z} f(x) + g(z)$ subject to $x = z$.

**Definition 6.2** (Dual function)**.** The **dual function** is defined as $g(\lambda, \nu) = \inf\limits_{x \in D} L(x, \lambda, \nu)$. Note that $g$ is concave even if (P) isn't convex.

**Remark 6.3.** Recall that $g(y) = \inf\limits_{x \in C} f(y, x)$ where $C$ is convex and $f$ is jointly convex implies $g$ is convex. On the other hand, $g(y) = \sup\limits_{x \in A} f(y, x)$ where $A$ is convex and $f$ is convex in $y$, then $g$ is convex.

**Definition 6.3** (Dual Problem)**.** We define the **dual problem** as $d^* = \max\limits_{\lambda \geq 0, \nu \in \mathbb{R}^p} g(\lambda, \nu)$; a convex problem.

**Theorem 6.1** (Weak Duality)**.** $d^* \leq p^*$ where $d^*$ is defined above and $p^* = \max\limits_{x \text{ feasible}} f_0(x)$

*Proof.* Consider $\tilde{x}$ feasible.

$$g(\lambda,\nu) = \inf_{x \in D} L(x,\lambda,\nu) \leq L(\tilde{x},\lambda,\nu) = f_0(\tilde{x}) + \underbrace{\sum_{i=1}^{m} \lambda_i f_i(\tilde{x})}_{\leq 0} + \underbrace{\sum_{i=1}^{p} \nu_i h_i(\tilde{x})}_{=0} \leq f_0(\tilde{x})$$

So $g(\lambda,\nu) \leq p^* = \inf_{x \text{ feasible}} f_0(x)$. Taking a supremum over $\lambda \geq 0$ and $\nu$ gives the result. □

**Remark 6.4.** If we also have strong duality ($d^* = p^*$), then usually it's equivalent to solve the dual. This is useful for structural reasons (Fourier transform, smoothness, shifting around linear operators). Even if (P) isn't convex, we can use this to help find lower bounds.

**Example 6.1.** Consider (P) $= \min_{x=[x_1,x_2]^T \succeq 0} \begin{cases} 3x_1 + 2x_2 \\ x_1 + 2x_2 \geq 5 \\ x_2 \leq 2 \end{cases}$ and a point $\tilde{x} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$. Here $f_0(\tilde{x}) = 7$ so $p^* \leq 7$ by definition. Rewrite the objective as $f_0(x) = 2x_1 + (x_1 + 2x_2) \geq 0 + 5 = 5$.
A further rewrite gives $3(x_1 + 2x_2) + 4(-x_2) \geq 3 \cdot 4 + 4 \cdot 2 = 15 - 8 = 7$ and we have attained the optimal solution. Here 3 and 4 are the dual variables.

**Example 6.2.** Here's duality for LP's. Consider $\min_x \langle c, x \rangle$ such that $Ax = b$ and $x \geq 0$. Form the Lagrangian $L(x,\lambda,\nu) = \langle c, x \rangle - \langle \lambda, x \rangle + \langle \nu, Ax - b \rangle$.

$$g(\lambda,\nu) = \min_x L(x,\lambda,\nu) = \min_x \langle c - \lambda + A^T\nu, x \rangle - \langle \nu, b \rangle = -\langle \nu, b \rangle + \delta_{\{c-\lambda+A^T\nu=0\}}(\lambda,\nu)$$

Hence the dual problem of an LP is again an LP:

$$\max_{\lambda \geq 0,\nu} g(\lambda,\nu) = \begin{cases} \max -\langle \nu, b \rangle \\ c + A^T\nu = \lambda \\ \lambda \geq 0 \end{cases} = \begin{cases} \min \langle \nu, b \rangle \\ c + A^T\nu \geq 0 \end{cases}$$

Alternatively, with a different standard form, the primal and dual problems are

$$\begin{cases} \max \langle c, x \rangle \\ x \geq 0 \\ Ax \leq b \end{cases} = \begin{cases} \max \langle b, y \rangle \\ y \geq 0 \\ A^T y \leq c \end{cases}$$

**Remark 6.5.** Some ideas when going from the primal to dual:

- Minima go to maxima and vice versa
- Variables go to constraints and vice versa
- objective/RHS are flipped

(SOB) For the primal variables, we call $x_i \geq 0$ *sensible*, $x_i$ unconstrained *odd*, and $x_i \leq 0$ *bizarre.*
When maximizing, a *sensible constraint* is $a_i^T x \leq b_i$, a *odd constraint* is $a_i^T x = b_i$, and a *bizarre constraint* is $a_i^T x \geq b_i$.

- A dual constraint is S/O/B if the primal variable is S/O/B.

**Example 6.3.** Consider $\min_{x\in\mathbb{R}^3}\begin{cases}3x_1+5x_2+x_3\\ x_1+x_2x_3=2:y_1\\ 2x_1-3x_2\le 0:y_2\\ x_1\ge 0, x_2\ge 0, x_3\in\mathbb{R}.\end{cases}$ Note $y_1$ is odd and $y_2$ is bizarre. The dual is $\min_{y\in\mathbb{R}^2}\begin{cases}2y_1\\ y_1+2y_3\le 3\\ y_1\le 5\\ y_1-3y_2=1\\ y_1\in\mathbb{R}, y_2\le 0.\end{cases}$

## 6.1 Slater's Condition

Weak duality always holds ($d^*\le p^*$). What about $d^*=p^*$ (**strong duality**)? Slater's condition, a constraint qualification (CQ) for (P) $=\begin{cases}\min f_0(x)\\ f_i(x)\le 0\\ Ax=b\end{cases}$ can guarantee strong duality if there exists $x\in\operatorname{relint}(\operatorname{dom}(f_0))$, $f_i(x)<0$ and if $f_i$ is affine $f_i(x)\le 0$, and $Ax=b$.

**Theorem 6.2.** If (P) is convex and Slater's condition holds, then $d^*=p^*$ (**strong duality**) and if $p^*<\infty$, then there exists a dual optimal point.

**Theorem 6.3.** LPs are nice. Either $p^*=d^*$ and there exists optimal primal and dual points, or $d^*=-\infty$ and $p^*=\infty$.

*Proof.* Use Slater's condition twice. □

**Remark 6.6.** SDPs aren't nice. Consider

$$\min_{X=X^T\succeq 0}\left\langle\begin{pmatrix}1&0\\0&0\end{pmatrix},X\right\rangle \text{ such that } \left\langle\begin{pmatrix}0&1\\1&0\end{pmatrix},X\right\rangle=2$$

Write in the form $X=\begin{pmatrix}a&1\\1&b\end{pmatrix}$ to satisfy the equality constraint. The problem is $\min a$ such that $\begin{pmatrix}a&1\\1&b\end{pmatrix}\succeq 0$. Looking at $\begin{pmatrix}a&1\\1&b\end{pmatrix}$, we have $a+b\ge 0$ and $ab\ge 1$ looking at trace and determinant, respectively. We conclude that $a,b\ge 0$ and $a\ge 1/b$.

Optimal solution: $\lim_{\epsilon\to 0}\begin{pmatrix}\epsilon&1\\1&1/\epsilon\end{pmatrix}$ gives $p^*=0$. Note that $d^*=0$ also and there exists a dual optimal point. The dual is degenerate.

**Remark 6.7.** Consider $\min_{x\in D} f_0(x)$ with $f_1(x)\le 0$ and $h_1(x)=0$. Let $G=\{(f_1(x),h_1(x),f_0(x))\mid x\in D\}$.

$$p^*=\inf_{(u,v,t)\in G,u\le 0,v=0} t$$

Forming the Lagrangian, we see

$$g(\lambda,\nu)=\min_{x\in D} f_0(x)+\lambda f_1(x)+\nu h_1(x)=\inf_{(u,v,t)\in G}\langle(\lambda,\nu,1),(u,v,t)\rangle$$

If $g(\lambda,\nu)>-\infty$ (i.e. $\lambda,\nu$ are feasible dual points), then we have a supporting hyperplane for $G$:

$$g(\lambda,\nu)\le\langle(\lambda,\nu,1),(u,v,t)\rangle,\quad\forall(u,v,t)\in G$$

See Ch. 5.3 in B&V for illustrations. There is also a minimum common points / maximum crossing interpretation explained by Bertsekas.

**Theorem 6.4** (Slater's Theorem)**.** If (P) is convex and Slater's condition holds ($\exists$ strictly feasible point), then $d^* = p^*$ and if $d^* = p^* > -\infty$ the dual optimal solution is achieved.

*Proof.* Consider $\mathcal{A} = G + (\mathbb{R}_+ \times \{0\} \times \mathbb{R}_+) = \{(u, v, t) \mid \exists x \in D, f_i(x) \leq u, h_i(x) = v, f_0(x) \leq t\}$. $\mathcal{A}$ is convex since $f_0$ and $f_i$ and $h_i$ are convex. Let $B = \{(0, 0, s) \in \mathbb{R}^3 \mid s < p^*\}$, where

$$p^* = \inf_{(u,v,t)\in G, u\leq 0, v=0} t = \inf_{(u,v,t)\in \mathcal{A}, u\leq 0, v=0} t$$

Since $\mathcal{A}, B$ are convex sets that don't intersect, we can separate them with a hyperplane that has normal vector $(\lambda^*, \nu^*, \mu^*)$ and we scale the third component: $(\lambda^*, \nu^*, 1)$. (See remark below). This means that

$$\langle(\lambda^*, \nu^*, 1), (u, v, t)\rangle \begin{cases} \geq C & (u, v, t) \in \mathcal{A} \\ \leq C & (u, v, t) \in B \end{cases}$$

Fix $\lambda^* \geq 0$. Looking at the $B$ inequality, we get $p^* \leq C$. If $x \in D$, then $(f_1(x), h_1(x), f_0(x)) \in G \subseteq \mathcal{A}$. So $\lambda^* f_1(x) + \nu^* h_1(x) + f_0(x) \geq C \geq p^*$. On the other hand,

$$g(\lambda, \nu) = \inf_{x\in D} L(x, \lambda^*, \nu^*) = \inf_{x\in D} \lambda^* f_1(x) + \nu^* h_1(x) + f_0(x) \geq p^*$$

The result follows from weak duality. □

**Remark 6.8.** If $\mu = 0$, there is a vertical supporting hyperplane and no strictly feasible point.

### 6.2 Saddle Point

Here's the saddle point interpretation of strong duality. Consider again (P) $p^* = \begin{cases} \min f_0(x) \\ f_i(x) \leq 0 \\ Ax = b \end{cases}$. The Lagrangian for this problem is

$$L(x, \lambda, \nu) = f_0(x) + \sum_i \lambda_i f_i(x) + \nu^T(Ax - b)$$

Note that we can write

$$p^* = \min_x \sup_{\lambda\geq 0,\nu} L(x, \lambda, \nu)$$

On the other hand,

$$d^* = \max_{\lambda\geq 0,\nu} g(\lambda, \nu) = \max_{\lambda\geq 0,\nu} \left(\min_x L(x, \lambda, \nu)\right)$$

and we obtain the weak duality condition by the *strong saddle point property*:

$$d^* = \sup_{\lambda\geq 0,\nu} \inf_x L(x, \lambda, \nu) \leq \inf_x \sup_{\lambda\geq 0,\nu} L(x, \lambda, \nu)$$

We call this inequality above the weak saddle point property.

**Example 6.4.** Consider $\min_x \|x\|_1$ such that $\|Ax - b\|^2 \leq \epsilon^2$. The Lagrangian is

$$L(x, \lambda) = \|x\|_1 + \lambda(\|Ax - b\|^2 - \epsilon^2)$$

If we know the optimal dual value, then $\min_x L(x, \lambda^*) = \min_x \overline{\lambda}\|x\|_1 + \frac{1}{2}\|Ax - b\|^2$ where $\overline{\lambda} = \frac{2}{\lambda}$. The latter problem is easier to solve because of no constraints and it's easier to tune $\lambda$ than $\epsilon$.

## 6.3 Game Theory

Consider a finite 2-person, 0-sum game. If I win \$5 $\iff$ you lose \$5. Let $P_{ij} =$ \$ paid from player 1 to player 2 assuming player 1 picks $i$ and player 2 picks $j$. (Rock Paper Scissors!)

$$\begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}$$

Since the matrix $P = -P^T$ is skew symmetric, we know it's a fair game. If player 1 choses a vector in $\{e_1, e_2, e_3\}$ and $v$ does the same, the outcome is $u^T P v$. We pick according to a probability density; i.e. $u \in \Delta = \{(u_1, u_2.u_3)^T \mid u_i \geq 0, u_1 + u_2 + u_3 = 1\}$. Then the expected outcome is $u^T P v$.

Suppose player 2 knows the strategy $u$ of player 1. The best choice for player 1, assuming player 2 knows their strategy, is:

$$p_1^* = \min_{u \in \Delta} \left( \max_{v \in \Delta} u^T P v \right)$$

Similarly,

$$p_2^* = \max_{v \in \Delta} \left( \min_{u \in \Delta} u^T P v \right)$$

Then $p_2^* \leq p_1^*$, so $p_1^*$ is better than $p_2^*$ for player 2.

- Recall that saddle points $(u^*, v^*) \implies$ strong duality $p_1^* = p_2^*$.
- Slater's condition (P) $\implies$ strong duality and $v^*$ exists
- Slater's condition (D) $\implies$ strong duality and $u^*$ exists
- Slater's condition (P & D) $\implies$ saddle points $(u^*, v^*)$

**Remark 6.9.** Here's an example of a convex program without strong duality. (It's has an empty interior).

$$\min_{y \geq 0, x} e^{-x}, \text{ s.t. } x^2/y \leq 0$$

Here's a nonconvex (eigenvalue) problem that has strong duality

$$\min_{\|x\|^2 \leq 1} x^T A x + 2\langle b, x \rangle, \quad A \in S^n \setminus S^n_+$$

## 6.4 Fenchel-Rockefeller Duality

See Bauschke+Combettes 2011 for this section. Consider (P) $= \min\limits_x f(x) + g(L(x))$, where $L$ is a linear operator and we allow $f, g$ to take $\pm\infty$. Here, the dual problem is

$$(D) = \min_v f^*(L^* v) + g^*(-v) = -d^*$$

**Remark 6.10.** Can quickly find dual function using this trick. That is, $g(x) = \|x\|$ and $g^*(v) = \delta_{\|x\|_\infty \leq 1}$.

**Remark 6.11.** Recall Moreau's Theorem, which says $y = \operatorname*{prox}\limits_f(y) + \operatorname*{prox}\limits_{f^*}(y)$. This helps find the proximal operator of a conjugate.

**Theorem 6.5** (18.15)**.** Suppose $f \in \Gamma_0(\mathbb{R}^n)$. Then, $f$ is differentiable and $\nabla f$ is L-Lipschitz $\iff$ $f^*$ is $\frac{1}{L}$-Lipschitz and strongly convex.

**Theorem 6.6** (15.23)**.** If $0 \in \text{relint}(\text{dom}(g) - L\text{dom}(f))$, then $p^* = \inf f + g + L = -\min f^* \circ L^* + \overline{g}^* = d^*$, where $\overline{g}(v) = g(-v)$. In other words, we have strong duality where the dual is achieved.

**Theorem 6.7** (15.24, 6.19)**.** In finite dimensions, it's sufficient to show $\text{relint}(\text{dom}(g)) \cap L(\text{relint}(\text{dom}(f))) \neq \varnothing$

**Theorem 6.8** (15.25)**.** In finite dimensions, if $f$ is polyhedral (i.e. $\text{epi}(f)$ is polyhedron), then it's sufficient to show $\text{dom}(g) \cap L(\text{dom}(f)) \neq \varnothing$

## 6.5 Connection with Lagrangian Duality

Given a dual solution, can we retrieve the primal solution?

1. Consider (P) $= \min f(x) + g(z)$ such that $Lx = z$. Then,

$$L(x, z, \nu) = f(x) + g(z) + \langle z - Lx, \nu \rangle$$

and

$$h_{\text{dual}}(\nu) = \inf_{x,z} L(x, z, \nu) = \Big(\inf_x f(x) + \langle -Lx, \nu \rangle\Big) + \Big(\inf_z g(z) + \langle z, \nu \rangle\Big) = -f^*(L^*\nu) - g^*(-\nu)$$

2. Here's the saddle point interpretation. Using the fact that $f^** = f$ for $f$ lower semi continuous,

$$\min f(x) + g(Lx) = \min_x f(x) + \sup_v \langle v, Lx \rangle - g^*(v)$$

so that

$$p^* = \min_x \sup_v f(x) + \langle v, Lx \rangle - g^*(v).$$

Note that

$$d^* = \sup_v \min_x f(x) + \langle v, Lx \rangle - g^*(v) = \sup_v \min_x f(x) + \langle L^*v, x \rangle - g^*(v) = \sup_v -f^*(-L^*v) - g^*(v)$$

**Definition 6.4** (Saddle Point)**.** $(x^*, v^*)$ **saddle point** of the convex-concave function $F(x, v)$ in the sense that

$$x^* = \min_x F(x, \nu^*), \quad v^* = \max_v F(x^*, v)$$

**Theorem 6.9** (19.1)**.** TFAE:

1. There is no duality gap and $x, v$ are primal, dual optimal points.
2. $L^*v \in df(x)$ and $-v \in dg(Lx)$
3. $x \in df^*(L^*v)$ and $Lv \in dg^*(-x)$.

**Remark 6.12.** For $f \in \Gamma_0(\mathbb{R}^n)$, recall that $df^* = (df)^{-1}$.

**Example 6.5.** Consider $\arg\min_x \frac{1}{2}\|x - y\|^2$ subject to $\|x\|_2 \leq \tau$. The solution is $x = \frac{y}{\max\{1, \|y\|/\tau\}}$. Changing the constraint so that $\frac{1}{2}\|x\|^2 \leq \frac{1}{2}\tau^2$, we get

$$L(x, \lambda) = \frac{1}{2}\|x - y\|^2 + \frac{\lambda}{2}(\|x\|^2 - \tau^2)$$

Then the dual function is

$$g(\lambda) = \min_x L(x, \lambda)$$

and $0 = \nabla L \implies x = (1+\lambda)^{-1} y \implies g(\lambda) = x = (1+\lambda)^{-1} y$. Then solving the dual is a 1D optimization problem.

**Example 6.6.** Consider $\text{prox}_{\lambda\|x\|_1}$; i.e. the problem $\arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|x-y\|^2 + \lambda\|x\|_1$. Since this is separable in $x$, we can examine the 1D operator

$$\arg\min_{x \in \mathbb{R}^n} \frac{1}{2}(x-y)^2 + \lambda|x| = \text{sign}(y)\lfloor |y| - \lambda \rfloor_+$$

### 6.6 KKT conditions

Consider the (P) $= \begin{cases} \min_x f_0(x) \\ f_i(x) \leq 0, & i = 1, 2, \ldots, m \\ h_i(x) = 0, & i = 1, 2, \ldots, p. \end{cases}$ Then $(x^*, \lambda^*, \nu^*)$ satisfies the **KKT conditions** if:

1. Stationarity: $\nabla_x L(x^*, \lambda^*, \nu^*) = 0$ or $0 \in \partial_x L(x^*, \lambda^*, \nu^*)$
2. Primal feasibility: $f_i(x^*) \leq 0$ and $h_i(x^*) = 0$
3. Dual feasiliity: $\lambda^* \geq 0$
4. Complementary-slackness: for all $i \in \{1, 2, \ldots, m\}$, we have $f_i(x^*)\lambda_i^* = 0$.

**Theorem 6.10.** Consider a nonconvex problem (P). If we have strong duality and $x^*$ and $(\lambda^*, \nu^*)$ are primal-dual optimal, then $(x^*, \lambda^*, \nu^*)$ necessarily satisfies the KKT conditions.

**Theorem 6.11.** If (P) is convex and $x^*$ and $(x^*, \lambda^*, \nu^*)$ satisfy the KKT conditions, then $x^*$ is primal optimal and $(\lambda^*, \nu^*)$ are dual optimal and $p^* = d^*$. (In the convex case we also have sufficiency for the theorem directly above this one).

*Proof.* Observe that

$$d^* \geq g(\lambda^*, \nu^*) = \inf_x L(x, \lambda^*, \nu^*) = L(x^*, \lambda^*, \nu^*) = f_0(x^*) + \sum \lambda_i^* f_i(x^*) + \sum \nu_i^* h_i(x^*) = f(x^*) \geq p^*$$

and thus $d^* = p^*$ and we have a primal optimal solution. □

**Theorem 6.12.** If (P) is convex and Slater's condition holds, then the KKT conditions are necessary for all primal optimal solutions.

**Example 6.7.** From Moreau's identity, we know $\underset{\|\cdot\|_\infty}{\text{prox}}$ and $\underset{\|\cdot\|_1 \leq 1}{\text{prox}}$ have the same solution. The first problem $\arg\min_{\|x\|_1 \leq 1} \frac{1}{2}\|x-y\|^2$ is now not separable due to the norm. Let's form the Lagrangian,

$$\arg\min_x \frac{1}{2}\|x-y\|^2 + \lambda(\|x\|_1 - 1)$$

which has (componentwise) solution $x_\lambda := \text{sign}(y)\lfloor |y| - \lambda \rfloor_+ = \begin{cases} y - \lambda & y > \lambda \\ y + \lambda & y < \lambda \\ 0 & y \in [-\lambda, \lambda] \end{cases}$

Now, instead of solving $\max_\lambda g(\lambda)$ we will search for the KKT conditions on $g(\lambda)$ for $\lambda$.

1. $x = x_\lambda$
2. $\|x_\lambda\| = 1$
3. $\lambda \geq 0$
4. $\lambda = 0$ or $\|x_\lambda\|_1 = 1$

We rule out $\lambda = 0$ since this means $x = y$ and $y$ might not be feasible. Of course if $y$ is feasible we pick $\lambda = 0$. To solve $\|x_\lambda\| = 1$ subject to $\lambda \geq 0$, we consider the scalar equation

$$\sum_{i=1}^{n} \lfloor |y_i| - \lambda \rfloor_+ = 1$$

This is a monotonic decreasing function of $\lambda$ with breakpoints $|y_i|$. Can solve with bisection method, and then it's strictly linear (i.e. no floor function) on that subinterval. The fast version of this algorithm finds the median in $O(n)$ time.

**Remark 6.13.** Suppose we have optimal saddle points; i.e. primal/dual optimal pairs $x^*, (\lambda^*, \nu^*)$ where strong duality holds.

$$f_0(x^*) = p^* = d^* = \sup_{\lambda \geq 0} g(\lambda, \nu) = g(\lambda^*, \nu^*) = \inf_x L(x, \lambda, \nu) \leq L(x^*, \lambda^*, \nu^*) = f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{n} \nu_i^* h_i(x^*) \leq f_0(x^*)$$

This means all the inequalities above are equalities. In particular, we must have $\sum_{i=1}^n \lambda_i^* f_i(x^*) = 0$ which is complementary slackness (CS). Note that if $\lambda_i \geq 0$ and $f_i(x) \leq 0$, from CS we have

$$\sum_{i=1}^{m} \lambda_i f_i(x) = 0$$

or equivalently, $\lambda_i \geq 0$ and $f_i(x) \leq 0$ and $\lambda_i f_i(x) = 0$ for all $i$.

**Example 6.8.** For the LP $\min\limits_X \langle \lambda, X \rangle = 0$ subject to $\lambda, X \geq 0$, we have $\lambda_i X_i = 0$ so $\lambda_i = 0$ or $X_i = 0$.

**Example 6.9.** Consider $\min \frac{1}{2}\langle x, Px \rangle + \langle q, x \rangle + r$. This has a closed form solution. If we also add the constraint $Ax = b$, there is also a closed form solution.

$$L(x, \nu) = \frac{1}{2}\langle x^*, Px^* \rangle + \langle q, x^* \rangle + \langle \nu, Ax^* - b \rangle$$

Now we check for a stationary/critical point:

$$0 = Px^* + q + A^*\nu$$

and ensure that $Ax^* = b$. This is equivalent to $\begin{pmatrix} P & A^* \\ A & 0 \end{pmatrix} \begin{pmatrix} x^* \\ \nu^* \end{pmatrix} = \begin{pmatrix} -q \\ b \end{pmatrix}$ and so we need to solve a linear system.

## 6.7 Perturbation and Sensitivity Analysis

Consider (P) with $p^*(0,0) = \begin{cases} \min\limits_x f_0(x) \\ f_i(x) \leq 0, \quad i = 1, 2, \ldots, m \\ h_i(x) = 0, \quad i = 1, 2, \ldots, p. \end{cases}$ and the perturbed version P$(u, v)$ with $\begin{cases} p^*(u, v) = \min\limits_x f_0(x) \\ f_i(x) \leq u_i, \quad i = 1, 2, . \\ h_i(x) = v_i, \quad i = 1, 2, . \end{cases}$

**Remark 6.14.** Can we drop constraints if they're not active? No, consider $\min\limits_x x$ such that $x \geq 0$ and $x^2 \geq 1$ versus $\min\limits_x x$ such that $x^2 \geq 1$.

**Theorem 6.13.** If (P) is convex and we have Slater's condition, then if a constraint isn't tight, it's OK to drop it.

**Remark 6.15.** Assume Slater's condition and convexity. Then,

$$p^*(0,0) = g(\lambda^*, \nu^*) = \inf_x L(x, \lambda^*, \nu^*) \leq L(x, \lambda^*, \nu^*)$$

Assuming $x$ is feasible for $P_{u,v}$, we have

$$L(x, \lambda^*, \nu^*) = h_0(x) + \sum_i \lambda_i g_i(x) + \sum_i \nu_i^* h_i(x) \leq f_0(x) + \langle \lambda, u \rangle + \langle \nu, v \rangle$$

In other words, for all feasible $x$ for $P_{u,v}$, we have

$$p^*(u,v) = \inf_x f_0(x) \geq p^*(0,0) - \langle \lambda^*, u \rangle - \langle \nu^*, v \rangle$$

**Remark 6.16.** We state some interpretations of the perturbed problem:

- If $\lambda_i^* \gg 0$ and we tighten $f_i(x) \leq 0$ to $u_i < 0$, then $p^*(u,0) \gg p^*(0,0)$
- If $\lambda^*$ is almost zero and we loosen the $i$th constraint so $u_i > 0$ and $\lambda_i^* \approx 0 > 0$. We get

  $$p^*(u,v) \leq p^*(0,0)$$

  If $\lambda_i^* = 0$, then $p^*(u,0) \geq p^*(0,0) \implies p^*(u,0) = p^*(0,0)$.
- WIth equality $\nu_i^* \gg 0$ and $v_i < 0$ or $\nu_i^* \ll 0$ and $v_i > 0$, we get $p^*(0,v) \gg p^*(0,0)$
- With equality, $|v_i^*| \approx 0$ and $v_i^* ? 0$ and $v_i > 0$ or $v_i^* < 0$ and $v_i < 0 \implies p^*(0,v) \approx p^*(0,0)$

## 6.8 Local Analysis

$p^*(u,v)$ is a convex function in $u$ and $v$ assuming that (P) is convex. If we further assume differentiability, then

$$\frac{\partial p^*}{\partial u_i}(0,0) = -\lambda^*, \quad \frac{\partial p^*}{\partial v_i}(0,0) = -\nu^*$$

# 7 Non-smooth Optimization

Let $f \in \Gamma(\mathbb{R}^n)$, but have no Lipschitz continuous gradient or maybe no gradient at all. We do, however, have a sugradient. This motivates a subgradient method

$$x_{k+1} = x_k - t_k d_k$$

where $d_k \in df(x_k)$. In the smooth case, if $\nabla f$ is Lipschitz, then we expect $f(x_k) - f(x^*) = O(1/k)$ with gradient descent or $O(1/k^2)$ with Nesterov accelerated gradient descent. If we also have strong convexity, we expect linear convergence i.e. $O(\beta^k)$ for some $\beta < 1$.

**Remark 7.1.** This subgradient method is not a descent method; i.e. the subgradients are not necessarily descent directions. Subgradient method suffers from poor convergence results.

**Example 7.1.** Consider $f(x,y) = |x| + 2|y|$. We compute $df(1,0) = \{(1,s) \mid |s| \leq 2\}$. So $(1,2) \in df(1,0)$, can draw a picture with level curves.

**Theorem 7.1** (8.13 in Beck)**.** Let $f$ be Lipschitz and we consider the polyak stepsize $t_k = \dfrac{f(x_k) - f(x^*)}{\|d_k\|^2}$. Then,

$$f(x_{\text{best after } k}) - f(x^*) = O(1/\sqrt{k})$$

**Theorem 7.2** (8.25 in Beck)**.** Let $f$ be Lipschitz and we consider the stepsize that diminishes to zero but not too fast, i.e. $\dfrac{\sum_{i=1}^k t_i^2}{\sum_{i=1}^k t_i} \xrightarrow{k\to\infty} 0$, then $f(x_{\text{best after } k}) \xrightarrow{k\to\infty} f(x^*)$.

**Theorem 7.3** (Bubeck)**.** Choose $t_k \propto \frac{1}{\sqrt{k}}$. Then we have the ergodic result $f\left(\frac{1}{k}\sum_{i=1}^k x_i\right) - f(x^*) \leq \frac{c}{\sqrt{k}}$. Assumes boundedness.

**Theorem 7.4** (8.31 in Beck)**.** If $f$ is $\mu$-strongly convex and $t_k = \frac{2}{\mu(k+1)}$. Then convergence is $O(1/k)$.

**Remark 7.2.** Recall that $d_k \in df(x_k) \iff f(y) \geq f(x_k) + \langle d_k, y - x_k\rangle$ for all $y$. The subgradient method

$$x_{k+1} = x_k - t_k d_k$$

is the same as

$$x_{k+1} \in \arg\min \frac{1}{2}\|x - (x_k - t_k d_k)\|^2$$

Suppose we add the (true) constraint $f(x) \geq f(x_\ell) + \langle d_\ell, x - x_\ell\rangle$ for $\ell = 0, 1, \ldots, k-1$. Look up cutting plane/bundle methods.

## 7.1 Proximal Gradient

Consider $\min\limits_x \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1 = f(x) + g(x)$, where $\nabla f$ is Lipschitz and $g$ is non-smooth but nice.
Recall the gradient descent method for a smooth function $f$:

$$x_{k+1} = \arg\min_x f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{L}{2}\|x - x_k\|^2 = x_k - \frac{1}{L}\nabla f(x_k)$$

Note that the objective of the arg min is $\geq f(x)$ for all $x$ by construction. What about the same for $g$, i.e.

$$x_{k+1} = \arg\min_x f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{L}{2}\|x - x_k\|^2 + g(x)$$

which is $\geq f(x) + g(x)$ for all $x$. After rewriting, this is the same as

$$\arg\min_x \frac{L}{2}\|x - (x_k - \frac{1}{L}\nabla f(x_k))\|^2 + g(x) = \arg\min_x \frac{1}{2}\|x - y\|^2 + \frac{1}{L}g(x) = \operatorname{prox}_{g/L}(y)$$

where $y = x_k - \frac{1}{L}\nabla f(x_k)$.

**Remark 7.3.** How to solve the proximal minimization? We need

$$\arg\min_x \frac{1}{2}\|x - y\|^2 + \frac{1}{L}g(x)$$

and so

$$0 \in x - y + \frac{1}{L} dg(x)$$

This means $y \in (I + dg)(x)$ or equivalently $x = (I + dg)^{-1} y$. Connections to forward/backward Euler:

$$x_{k+1} = y - \frac{1}{L} dg(x_{k+1}) = x_k - \frac{1}{L} \left( \underbrace{\nabla f(x_k)}_{\text{explicit}} + \underbrace{dg(x_{k+1})}_{\text{implicit}} \right)$$

**Remark 7.4.** Some options for picking the line search: find $t$ where

- A curvilinear line search would be
$$x(t) = \operatorname*{prox}_{tg}(x_k - t\nabla f(x_k))$$
- $\tilde{x} = \operatorname*{prox}_{t_0 g}(x_k - t_0 \nabla f(x_k))$. Then $x(t) = x_k + t(\tilde{x} - x_k)$.

**Remark 7.5.** With ADMM/Douglas Rachford, our objective is find $0 \in d(f+g)(x)$, the latter of which equals $df(x) + dg(x) = \nabla f(x) + dg(x)$.

$$x \in \frac{1}{L}\nabla f(x) + x + dg(x) \iff \left(I - \frac{1}{L}\nabla f\right) x \in \left(I + \frac{1}{L} dg\right) x \iff \left(I + \frac{1}{L} dg\right)^{-1} \left(I - \frac{1}{L}\nabla f\right) x = x$$

This $T(x) = x$ is a fixed point problem, so the algorithm is $x_{k+1} = T(x_k)$. Notice these updates in $T$ correspond exactly to the proximal gradient updates. May be known as a "forward backward" operator.
For a unique solution to exist for such a fixed point problem, need $\|T(x) - T(y)\| \leq \rho \|x - y\|$ where $\rho < 1$.

# 8 Unconstrained Optimization Algorithms

Consider $\min_x f(x)$ where $f$ is smooth and convex. We have already see (Nesterov's accelerated) gradient descent $x_{k+1} = x_k - t_k \nabla f(x_k)$ and Newton's method $x_{k+1} = x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k)$. These both come from the same framework: $x_{k+1} = \arg\min$ second order Taylor Expansion of $f$ at $x_k$. Here, gradient descent says $\nabla^2 f(x_k) \approx t_k^{-1} I$.

## 8.1 Conjugate Gradient

Our next new algorithm is conjugate gradient. Conjugate gradient descent originated in the 50's (linear case) and a nonlinear version began to be studied in the 60s.

1. Linear case: Consider $\min_x \frac{1}{2}\|\tilde{A}x - \tilde{b}\|_2^2$. Setting the gradient to zero gives $\tilde{A}^T(\tilde{A}x - \tilde{b}) = 0 \implies \tilde{A}^T\tilde{A}x = \tilde{A}^T\tilde{b}$. This is the linear system CG solves. Need $A \succ 0$ so that our objective is strongly convex. There are options if $A \succeq 0$, however.

   Define the residual $r(x) = Ax - b$ and $r_k = r(x_k)$. We will find a set of "conjugate directions" $\{p_i\}$ that are A-orthogonal; i.e. $p_i^T A p_j = \langle p_i \mid A \mid p_j \rangle = 0$ for all $i, j$. The algorithm is:

$$x_{k+1} = x_k + \alpha_k p_k$$

   where $\alpha_k$ is chosen via exact linesearch. This gives $\alpha_k = \frac{-\langle r_k, p_k \rangle}{\langle p_k | A | p_k \rangle}$

**Theorem 8.1.** Let $x_n = x^*$ be the true minimizer.

*Proof.* Since $A \succ 0$, $\{p_i\}_{i=0}^{n-1}$ is a basis. In particular,

$$x^* - x_0 = \sigma_0 p_0 + \cdots + \sigma_{n-1} p_{n-1}$$

but we can easiliy find the coefficients by multiplyting on the left by $p_k^T A$:

$$p_k^T A(x^* - x_0) = \sigma_k p_k^T A p_k \implies \sigma_k = \frac{p_k A^T (x^* - x_0)}{\langle p_k \mid A \mid p_k \rangle}$$

Via our method,

$$x_k - x_0 = \alpha_0 p_0 + \cdots + \alpha_{k-1} p_{k-1} \implies p_k^T A(x_k - x_0) = 0$$

and therefore we get

$$\sigma_k = \frac{p_k A^T (x^* - x_0)}{\langle p_k \mid A \mid p_k \rangle} = \frac{p_k A^T (x^* - x_k)}{\langle p_k \mid A \mid p_k \rangle}$$

□

**Remark 8.1.** CG is ultimately equivalent to the Lanczos procedure for finding eigenvectors.

**Remark 8.2.** What we proposed earlier is a conjugate direction method where we have assumed $\{p_k\}$ are A-orthogonal. How to find these directions? $p_k = r_k + \beta_k p_{k-1}$. Choose $\beta_k$ so that $\beta_k = \frac{\langle r_k | A | p_{k-1} \rangle}{\langle p_{k-1} | A | p_{k-1} \rangle}$ and thus $\langle p_k \mid A \mid p_{k-1} \rangle = 0$,

**Remark 8.3.** We have the following result about convergence of CG: $\|x_k - x^*\|_A \leq 2 \left( \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right)^k \|x_0 - x^*\|_A$ where $\kappa$ is the condition number of the matrix $A$ and $\| \cdot \|_A = \sqrt{\langle \cdot \mid A \mid \cdot \rangle}$

**Remark 8.4.** Observe that $x_0 \cup \text{span}\{p_0, \ldots, p_{k-1}\} = \text{span}\{x_0, Ax_0, \ldots, A^{k-1} x_0\}$ and the RHS is a Krylov subspace. This is true for $k \leq n$ since the Krylov subspace eventually spans the whole space.

**Remark 8.5.** In practice, we use **preconditioning** to solve $(AP)(P^{-1}x) = b$ instead of $Ax = b$. Of course, it's now important to invert $P$ easily; common choices are diagonal matrices.

## 8.2 Nonlinear CG

Here we consider $\min_x f(x)$ where $f$ is not quadratic.

1. replace $r(x)$ with $\nabla f(x)$
2. Choose $\alpha_k$ via linesearch (sensitive!)
3. Choose $\beta_k$. Fletcher-Reeves says $\beta_{k+1}^{FB} = \frac{\|\nabla f(x_{k+1})\|^2}{\|\nabla f(x_k)\|^2}$. See Nocedal + Wright for other choices.

**Remark 8.6.** Hoger + Zhang study modern non-linear CG but their method has lots of parameters. Nemirovsky and Yuden showed that it's inferior to gradient descent on some problems. Not a lot of theory overall. Doesn't work well with constraints.

### 8.3 Quasi-Newton Methods

Again see Nocedal + Wright. We build a quadratic approximation

$$m_k(p) = f(x_k) + \langle \nabla f(x_k), p\rangle + \frac{1}{2}\langle p \mid B_k \mid p\rangle$$

where $p = \tilde{x}_{k+1} - x_k$ is a stepsize and $B_k \approx \nabla^2 f(x_k)$. (If $\beta_k = LI$ we have gradient descient, and if $\beta_k = \nabla^2 f(x_k)$, we have Newton's method). Choose $p_k = \arg\min_p m_k(p) = -\beta_k^{-1}\nabla f(x_k)$.

Define $x_{k+1} = x_k + \alpha p_k$ and $s_k = x_{k+1} - x_k$ and $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$, where $\alpha$ is a stepsize. Note that $m_k(0) = f(x_k)$ and $\nabla m_k(0) = \nabla f(x_k)$, so our quadratic approximation matches slope correctly of the true function at $x_k$. Let's impose that $\nabla m$ and $\nabla f$ agree at the previous point. At $x_{k+1}$, we want

$$\nabla m_{k+1}(-s_k) = \nabla f(x_k)$$

The LHS equals

$$\nabla f(x_{k+1}) - \beta_{k+1} s$$

and so in our notation, we want $B_{k+1}s_k = y_k$, which is the secant equation. If the secant equation is satisfied, we have a quasi-Newton method. Note that if $\langle s_k, y_k\rangle > 0$ (curvature condition), then $B_{k+1} \succ 0$.

**Remark 8.7.** $B_{k+1}$ is symmetric so there are $\frac{n(n+1)}{2}$ degrees of freedom. There are $n$ constraints. If $n > 1$, we have several choices for solving the secant equation $B_{k+1}s_k = y_k$. If we have $B_k$, we choose $B_{k+1}$ so that $B_{k+1} \succeq 0$, $B_{k+1}s_k = y_k$, and $B_k$ and $B_{k+1}$ are close. Defining "close" gives rise to the different quasi-Newton methods.

1. BFGS: make $H_{k+1} = B_{k+1}^{-1}$ close to $H_k = B_k^{-1}$. With $\rho_k = \frac{1}{\langle y_k, s_k\rangle}$, the update looks like:

   $$H_{k+1} = (I - \rho_i s_k y_k^T) H_k (I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T$$

   That is, $H_{k+1} = H_k +$ rank 2 perturbation. Under this perturbed form, we can invert via the matrix inversion lemma and hence work with the inverse $B_k$ as well. Usually initialize $H_0 = \frac{\langle y_0, s_0\rangle}{\langle y_0, y_0\rangle} I$,

   $$\tilde{x}_{k+1} = x_k - H_k \nabla f(x_k) \quad x_{k+1} = x_k + t(\tilde{x}_{k_1} - x_k)$$

2. Others include SR1, ...

If $n = 1$, we have the secant method in 1D.

**Theorem 8.2.** If $0 \prec \mu I \preceq \nabla^2 f(x) \preceq LI$, then $x_k \to x^*$ and hence $f(x_k) \to f(x^*)$ by continuity.

**Remark 8.8.** Convergence without strong convexity is unknown. Convergence to a stationary point without convexity is unknown. Quasi-newton methods can achieve superlinear convergence with some extra assumptions.

**Remark 8.9.** Recall gradient descent with no constraints

$$x_{k+1} = x_k - L^{-1}\nabla f(x_k)$$

and with constraints

$$x_{k+1} = \operatorname*{Proj}_C \left(x_k - L^{-1}\nabla f(x_k)\right)$$

Can we similarly get a quasi-Newton method with constraints, e.g.

$$x_{k+1} = \operatorname*{Proj}_C \left(x_k - B_k^{-1}\nabla f(x_k)\right)$$

Recall our quasi-Newton update was

$$x_{k+1} = \arg\min_{x \in C} f(x_k) + \langle f(x_k), x - x_k \rangle + \frac{1}{2}\|x - x_k\|^2_{B_k}$$

This last term messes up our projection; we would no longer have the standard Euclidean projection and must project with respect to $\| \cdot \|^2_B = \langle \cdot \mid B \mid \cdot \rangle$

**Example 8.1.** Consider $\min_x \frac{1}{2}\|Ax - b\|^2 + g(x)$ where $g$ is differentiable and small. Here $A$ is $m \times n$ and $x \in \mathbb{R}^n$. First, we consider Newton's method. The cost of $\nabla$ is $O(mn)$. For $\nabla^2$, forming it is $O(mn^2)$, storage is $O(n^2)$, and inversion is $O(n^3)$.
For a quasi-Newton method, the update $H_k$ is $O(n^2)$, we compute the inverse for free, and storage is $O(n^2)$.
Still more costly that gradient descent.

## 8.4 Limited Memory Quasi Newton

These are state of the art tools! Let $v_k = I - \rho_k y_k s_k^T$. Then the BFGS update was

$$H_{k+1} = v_k^T H_k v_k + \rho_k s_k s_k^T$$

We need a matrix and vector product:

$$H_{k+1} v = v_k^T H_k v_k \cdot z + \rho_k S_k s_k^T z$$

Because of the form for $v_k$ above, we can apply $v_k \cdot z$ in $O(n)$ operations. The sum on the RHS is also $O(n)$, and overall the cost (and storage) for $H_{k+1} z$ is $O(kn)$. This is ineffective for large $k$. The L-BFGS algorithm, we recurse with this formula for $k+1, \ldots, k-M$ with maybe $M = 5, 10, 20$. With costs of $O(20 \cdot n)$, this starts to look more like gradient descent.

**Remark 8.10.** If you apply the memoryless case $M = 1$ to a quadratic, we get back the conjugate gradient method.

## 8.5 Newton-CG

Also known as Inexact Newton or Matrix-Free Newton.

$$x_{k+1} = x_k - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$$

Note that $\left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$ is the solution of the system $\left(\nabla^2 f(x_k)\right) z = \nabla f(x_k)$. Use CG to (approximately) solve this linear system.

## 8.6 Nonlinear Least Squares

This is nonconvex and we are interested in finding a local minimum.
For example, $\min_x \frac{1}{2}\|A(x) - y\|^2$ where $A(\cdot)$ is nonlinear rather than linear. Let $f(x) = \frac{1}{2}\sum_{j=1}^m r_j^2(x) = \frac{1}{2}\|\mathbf{r}(x)\|_2^2$ where $r_j(\cdot)$ is nonlinear and $\mathbf{r} : \mathbb{R}^n \to \mathbb{R}^m$. Let $J(x) = \begin{pmatrix} \nabla r_1(x)^T \\ \vdots \\ \nabla r_m(x)^T \end{pmatrix}$ be the $m \times n$ Jacobian matrix. We claim that

$\nabla f(x) = J(x)^T \mathbf{r}(x)$. Indeed,

$$\nabla f(x) = \frac{1}{2} \sum_{j=1}^{m} \nabla r_j^2(x) = \sum_{j=1}^{m} r_j(x) \nabla r_j(x) = J(x)^T \mathbf{r}(x)$$

We calculate $\nabla^2 f(x) = J(x)^T J(x) + \sum_{j=1}^{m} r_j(x) \nabla^2 r_j(x)$.

**Example 8.2.** Consider the PDE constrained optimization problem: $\min\limits_{u,\theta} \frac{1}{2}\|Au - y\|^2$ such that $u_{tt} = F(u_t, u, u_x, u_{xx}, \theta)$ where $\theta$ is a vector of parameters and $F$ describes the PDE e.g. $u_{tt} = c^2(\theta)u_{xx}$. The matrix $A$ describes your observations and $y$ are your actual observations.

**Remark 8.11.** Gauss-Newton drops the second sum in $\nabla^2 f(x)$ for nonlinear least squares. To be precise, let $r_k = r(x_k)$ and $J_k = J(x_k)$.

$$\tilde{x}_{k+1} = x_k - (J_k^T J_k)^{-1} \nabla f(x_k) = x_k - (J_k^T J_k)^{-1}(J_k^T r_k)$$

and then maybe followed by a line search:

$$x_{k+1} = x_k + \gamma(\tilde{x}_{k+1} - x_k)$$

Note that $\nabla^2 f(x) = J(x)^T J(x)$ is positive definite in this case, and the method is hence a descent method. Another derivation: $f(x) = \frac{1}{2}\|\mathbf{r}(x)\|^2$ where we use a first order taylor series:

$$\mathbf{r}(x_k + p) \approx r_k + J_k \cdot p$$

The model is $\min\limits_{p} \frac{1}{2}\|r_k + J_k p\|^2$, which we can solve in 1 step using Newton. The popular variant is Levenberg-Marguardt, a “trust region” version of Gauss Newtonm where we solve

$$\min \frac{1}{2}\|r_k + J_k p\|^2$$

such that $\frac{1}{2}\|p^2\| \leq \delta^2$, which is naturally a ridge regression when written in Lagrangian form.

$$\tilde{x}_{k+1} = x_k - (J_k^T J_k + \lambda I)^{-1}(J_k^T r_k)$$

where $\lambda$ is chosen using some trust region methods.

# 9 Constrained Optimization Algorithms

## 9.1 L-BFGS-B

We have seen limited memory BFGS i.e. (L-BFGS). The appended B refers to “bounded” and allows simple bounds of the form $\ell_i \leq x_i \leq u_i$.

## 9.2 Penalty methods

$\min\limits_{x} f_0(x)$ such that $\frac{1}{2}\|Ax - b\|^2 \leq \epsilon^2$. Instead, solve with a quadratic penalty $\min\limits_{x} f_0(x) + \frac{\mu}{2}\|Ax - b\|^2$.

$$L(x, \lambda) = f(x) + \lambda \left( \frac{1}{2}\|Ax - b\|^2 \right)$$

If I know a dual optimal point $\lambda^*$, then $x^* \in \arg\min L(x,\lambda^*)$ from stationarity. This motivates the quadratic penalty method

$$\min_x f(x) + \frac{\mu}{2}\|Ax-b\|^2$$

for $\min\limits_x f(x)$ such that $Ax = b$. In reality, we don't know $\lambda^*$ and so we take $\mu \to \infty$ i.e.. $(\mu_k)$ such that $\mu_k \to \infty$ and

$$x_k \in \arg\min_x f(x) + \frac{\mu_k}{2}\|Ax-b\|^2$$

**Theorem 9.1.** If $\mu_k \to \infty$ and $x_k \in \arg\min f(x) + \frac{\mu_k}{2}\left(\sum i = 1^m h_i^2(x)\right)$, then if $\{x_k\}$ has a limit point $x^*$, then $x^*$ solves the original problem.

**Remark 9.1.** Exact Penalty Methods: Don't square the norm in the penalty. These have the property that if $\mu$ is large enough, then $x_m$ solves the original problem. (In general, a penalty is exact $\implies$ it's not smooth).

**Remark 9.2.** For inequality constraints $f_i(x)$, we use a penalty of the form $\frac{\mu}{2}\left(\lfloor f_i(x)\rfloor_+\right)^2$

**Example 9.1.** Consider $\min \frac{1}{2}x^TPx$ where $P \succeq 0$ such that $Ax = b$ where $A$ is $m \times n$ and $m < n$. To solve

$$\min_x \frac{1}{2}x^TPx + \frac{\mu}{2}\|Ax-b\|^2$$

we find a sequence $x_k$ that satisfies $0 = Px + \mu A^T(Ax-b)$ i.e. $(P+\mu A^TA)x = \mu A^Tb$. Note that as $\mu \to \infty$ presents an ill-conditioned problem.

**Remark 9.3.** Penalty methods, augmented lagrangian, and SQP are from the 80's. Interior Point Methods were from the 80's and 90's. Proximal gradient descent, Doughlas-Rachford/ADMM, and primal dual versions are much more recent from 00's.

## 9.3 Augmented Lagrangian

The key observation is that the following problems are equivalent: $\min\limits_{Ax=b} f(x) \iff \min\limits_{Ax=b} f(x) + \frac{\mu}{2}\|Ax-b\|^2$ Forming the **augmented lagrangian** of the RHS:

$$L_\mu(x,\nu) = f(x) + \frac{\mu}{2}\|Ax-b\|^2 + \langle \nu, Ax-b\rangle; \quad g(\nu) = \min_x L_\mu(x,\nu)$$

**Theorem 9.2.** If $x^*$ is the unique minmizer of $L_\mu(x,\nu)$, then $\nabla g(\nu) = \nabla_\nu L_\mu(x^*,\nu)$. Furthermore if $L_\mu$ is $\mu$-strongly convex in $x$, then $\nabla g$ is $\frac{1}{\mu}$-Lipschitz continuous.

The previous theorem motivates a **gradient ascent** on $g(\nu)$:

$$\nu_{k+1} = \nu_k + \mu\left(Ax_k - b\right); \quad x_k = \arg\min_x L_\mu(x,\nu_k)$$

**Remark 9.4.** This is a compromise with the $\mu \to \infty$ in the quadratic penalty method (ill-conditioned) and solving the dual problem.

**Remark 9.5.** Seen in LANCELOT and MINOS software for NLP (nonconvex and find a local minimizer). With inequality constraints, one idea is to convert all inequality constraints to bound constraints using slack variables.

$$f_i(x) \le 0 \iff f_i(x) + s_i = 0,\ s_i \ge 0$$

Then L-BFGS-B is available.

### 9.4 SQP

SQP stands for Sequential Quadratic Programming. It is a variant of Newton's method. For SQP with equality constraints, consider $\min\limits_x f_0(x)$ such that $h_i(x) = 0$ for $i = 1, \dots, m$ and let $h(x) = \begin{pmatrix} h_1(x) \\ \vdots \\ h_m(x) \end{pmatrix}$. Let's for the Lagrangian $L(x, \nu) = f_0(x) + \sum \nu_i h_i(x)$ and look at the KKT conditions:

- Stationarity: $0 = \nabla f(x) + \sum \nu_i \nabla h_i(x)$
- Feasibility: $h(x) = 0$
- dual feasibility (needs inequality constraints)
- complementary slackness (needs inequality constraints)

Let $J(x) = \begin{pmatrix} \nabla h_1(x)^T \\ \dots \\ \nabla h_m(x)^T \end{pmatrix}$ and define $F(x, \nu) = \begin{pmatrix} \nabla f_0(x) + J(x)\nu \\ h(x) \end{pmatrix}$. From KKT, we have $F(x, \nu) = 0$. Moreover, we compute

$$F'(x, \nu) = \begin{pmatrix} \nabla^2_{xx} L(x, \nu) & J(x) \\ J^T(x) & 0 \end{pmatrix}$$

and use Newton's method on the big system:

$$\begin{pmatrix} x_{k+1} \\ \nu_{k+1} \end{pmatrix} = \begin{pmatrix} x_k \\ \nu_k \end{pmatrix} - (F'(x, \nu))^{-1} F(x_k, \nu_k)$$

**Remark 9.6.** Some SQP programs include SSNOPT (state of the art), KNITRO, TRON.

## 10 Derivatives

### 10.1 Gradients

We will talk about: avoiding gradients, computing them analytically, approximating with finite differences, and automatic differentiation (e.g. back propagation or adjoint state method)

1. Derivative Free Optimization (DFO). Assume $f(x)$ is computable for $x \in \mathbb{R}^n$ with $n \in (10, 100)$. The conventional method is model-based : we have $\{x_i, f(x_i)\}_{i=1}^{\ell}$ interpolated with a polynomial (or something nicer) and use this fit as a proxy. (See Nocedal + Wright or Scheinberg, Conn, and Vicente DFO textbook) This would be applicable if you observe a noisy version of $f$ e.g. $f(x) = \int h(x, s)\, ds$ from Monte Carlo estimates, or also $\min\limits_\theta \frac{1}{2}\|A \cdot u(\theta) - y\|^2 = \min\limits_\theta f(\theta)$ where $u(\theta)$ solves the wave equation.

   Also includes coordinate descent or pattern based searches. The Nelder-Mead simplex reflection also falls in this category. Implicit filtering is like finite difference but with $h$ not taken to be small.

2. Analytic Computation. Let's summarize some basis calculus properties:
   - Product rule $(fg)' = f'g + g'f$
   - Leibniz rule

$$\frac{d}{dx}\left(\int_{a(x)}^{b(x)} f(x, t)\, dt\right) = f\big(x, b(x)\big) \cdot \frac{d}{dx} b(x) - f\big(x, a(x)\big) \cdot \frac{d}{dx} a(x) + \int_{a(x)}^{b(x)} \frac{\partial}{\partial x} f(x, t)\, dt.$$

- Chain rule $\frac{d}{dx}(g \circ f)(x) = g'(f(x)) \cdot f'(x)$
- Total Derivative. If $z = f(x, y)$ and $x = g(t)$ and $y = h(t)$, then $\frac{dz}{dt} = \frac{df}{dx}\frac{dx}{dt} + \frac{df}{dy}\frac{dy}{dt}$.
- Implicit Differentiation. For $F(x, y) = 0 = F(x, f(x))$, we know $\frac{dF}{dx} = 0$ since $F = 0$. From the total derivative,
$$\frac{dF}{dx} = \frac{dF}{dx}\frac{dx}{dx} + \frac{dF}{dy}\frac{dy}{dx}$$
and so $\frac{dy}{dx} = -\left(\frac{dF}{dy}\right)^{-1}\frac{dF}{dx}$

Our main new result is the multidimensional chain rule. Let $h = g \circ f$ where $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^p$. Define the Jacobian as $(J(h(x)))_{ij} = \frac{dh_i(x)}{dx_j}$ i.e. $J(h(x)) = \begin{pmatrix} \nabla h_1(x)^T \\ \vdots \\ \nabla h_p(x)^T \end{pmatrix}$ Then, the rule is

$$(Jh)(x) = (Jg)(f(x)) \cdot (Jf)(x)$$

**Remark 10.1.** Let $p = 1$ and then $\nabla h(x) = (Jh(x))^T = (Jf(x))^T \cdot (Jg(f(x)))^T = (Jf(x))^T \nabla g(f(x))$.

**Example 10.1.** Let $f(x) = Ax - b$ where $A \in \mathbb{R}^{m \times n}$ and consider $h(x) = g(f(x))$. We see $(Jf)(x) = A$ and immediately get $\nabla h(x) = A^T \nabla g(Ax - b)$. So if $h(x) = \frac{1}{2}\|Ax - b\|^2$ then $\nabla h(x) = A^T(Ax - b)$

**Example 10.2.** Consider $f(U, V) = \frac{1}{2}\|UV^T - B\|_F^2$ where $B \in \mathbb{R}^{n_1 \times n_2}$ and $U \in \mathbb{R}^{n_1 \times r}$ and $V \in \mathbb{R}^{n_2 \times r}$. Let's view it just as a single function of $U$ and take the gradient. We will identify the gradient under the Taylor series expansion
$$f(u + \Delta) = f(u) + \langle \nabla f(u), \Delta \rangle + O(\|\Delta\|^2)$$
where $\langle \cdot, \cdot \rangle$ indicates matrix inner product product $\langle X, Y \rangle = \mathrm{tr}(X^T Y) = \mathrm{vec}(X)^T \mathrm{vec}(Y)$.

$$\begin{aligned} f(U + \Delta) &= \frac{1}{2}\|(U + \Delta)V^T - B\|_F^2 \\ &= \frac{1}{2}\|UV^T - B\|_F^2 + \frac{1}{2}\|\Delta V^T\|_F^2 + \langle UV^T - B, \Delta V^T \rangle \\ &= f(U) + O(\|\Delta\|^2) + \mathrm{tr}((UV^T - B)^T \Delta V^T) \\ &= f(U) + O(\|\Delta\|^2) + \mathrm{tr}(V^T(UV^T - B)^T \Delta) \\ &= f(U) + O(\|\Delta\|^2) + \mathrm{tr}(((UV^T - B)V)^T \Delta) \end{aligned}$$

**Remark 10.2.** For more about calculating gradients, see the *Matrix Cookbook* by Peterson and Peterson. Also see *A Matrix Handbook for Statisticians* by Seber.

**Example 10.3.** Consider the Speelpenning function $g(y) = \prod_{i=1}^n g_i(y)$ where $g_i(y) = y - t_i$ for $y \in \mathbb{R}^1$.

$$g'(y) = \sum_{i=1}^n \prod_{j \neq i}^n y - t_j$$

This naive implementation is $O(n^2)$. An alternative is $O(n)$ via

$$g'(y) = \sum_{i=1}^n \frac{g(y)}{y - t_i}$$

but this is numerically unstable near $t_i$. A better way is to store $f^{(k)} = \prod_{i=1}^{k}(y - t_i)$ and $r^{(k)} = \prod_{i=k}^{n}(y - t_i)$ and we get an (more stable) $O(n)$ computational method:

$$g'(y) = \sum_{i=1}^{n} f^{(i-1)} r^{(i-1)}$$

The tradeoff is a requirement for more memory. This falls under Algorithmic Differentiation or Automatic Differentiation (AD) or Backpropogation (AD in reverse mode).

**Example 10.4.** Let $f(x_1, x_2) = x_1x_2 + \sin(x_1) = w_1w_2 + \sin(w_2) = w_3 + w_4 = w_5$.

Loop over dimension of input (here $k = 1, 2$).

- Let $\widehat{w_i} = \dfrac{dw_i}{dx_1}$. Write $f(x_1, x_2) = w_5(w_3(w_1(x), w_2(x)), w_4(w_1(x)))$. Then we calculate

$$\frac{df}{dx_i} = \frac{dw_5}{dw_3}\left(\frac{dw_3}{dw_1}\frac{dw_1}{dx_1} + \frac{dw_3}{dw_2}\frac{dw_2}{dx_1}\right) + \frac{dw_5}{dw_4}\left(\frac{dw_4}{dw_1}\frac{dw_1}{dx_1}\right)$$

  We find $\widehat{w_1} = \frac{dw_1}{dx_1} = \frac{dx_1}{dx_1} = 1$ and $\widehat{w_2} = \frac{dw_2}{dx_1} = 0$ and $w_3 = w_1w_2$ and $\widehat{w_3} = \widehat{w_1}w_2 + w_1\widehat{w_2}$ and $w_4 = \sin(w_1)$ and $\widehat{w_4} = \cos(w_1)\widehat{w_1}$ and $w_5 = w_3 + w_4$ and $\widehat{w_5} = \widehat{w_3} + \widehat{w_4}$. We pass along the $w_i$ and their derivatives $\widehat{w_i}$ in our code.

**Remark 10.3.** Suppose $f : \mathbb{R}^n \to \mathbb{R}$, let $C_f$ be the cost of evaluating $f(x)$. Then the cost of $f'$ via forward mode ADI is $n \cdot C_f$. This is impractical for large problems (hence backpropogation in deep learning).

**Remark 10.4.** Reverse mode AD, consider the adjoint $\overline{w_j} = \frac{dy_i}{dw_j}$ where $y \in \mathbb{R}^m$. Here we loop over output variables, and the cost is $m \cdot C_f$, so if $f : \mathbb{R}^n \to \mathbb{R}$ then $\nabla f$ costs about the same as $f(x)$.

**Example 10.5.** We return to the example $f(x_1, x_2) = x_1x_2 + \sin(x_1) = w_1w_2 + \sin(w_2) = w_3 + w_4 = w_5$. The steps are as follows:

- Make a forward pass and store values $\{w_i\}$
- Reverse the tree and compute $\overline{w_i}$
- Reverse the tree and compute $\overline{w_5} = \frac{dy}{dw_5} = 1$. Then,

$$\overline{w_3} = \frac{dy}{dw_3} = \frac{dy}{dw_5}\frac{dw_5}{dw_3} = \overline{w_5}$$

  and similarly

$$\overline{w_4} = \overline{w_5}$$

  For $w_4 = \sin(w_1 a)$, we get

$$\overline{w_1^a} = \frac{dy}{dw_1^a} = \frac{dy}{dw_4} \cdot \frac{dw_4}{dw_1^a} = \overline{w_4} \cdot \cos(w_1^a)$$

  and we compute the last term with our values. We also have $w_3 = w_1^b w_2$.

$$\overline{w_2} = \frac{dy}{dw_2} = \frac{dy}{dw_3}\frac{\partial w_3}{\partial w_2} = \overline{w_3}\frac{\partial w_3}{\partial w_2} = \overline{w_3}w_1^b$$

and $\overline{w_1^b} = \overline{w_3} w_2$. Finally, let $w_1 = w_1^a$.

$$\overline{w_1} = \frac{dy}{dw_1} = \frac{dy}{dw_1^a}\frac{dw_1^a}{dw_1} + \frac{dy}{dw_1^b}\frac{dw_1^b}{dw_1} = \overline{w_1^a}\cdot 1 + \overline{w_1^b}\cdot 1$$

**Remark 10.5.** Code for ADI in ADIFOR (Fortran), Adigator (Matlab), and PyTorch (Python).

**Example 10.6.** Adjoint-State Method for PDE-constrained optimization. Consider $\min_{p.u} g(u,p)$ such that $f(u,p) = 0$ where $u \in L^2([a,b])$, $f(u,p) = 0$ is a PDE, and $p \in \mathbb{R}^n$ are parameters. We need $\nabla_p g = \frac{dg}{dp} = g_p \in \mathbb{R}^n$.

For an explicit example, consider $\min g(u,p)$ such that $Au = b$ where $u \in \mathbb{R}^M$ and $A \in \mathbb{R}^{M\times M}$. Let $A$ and/or $b$ depend on $p$.

$$\frac{dg}{d_p} = g_p + g_u\frac{du}{dp}$$

Computing $g_p$ and $g_u$ is straightforward from the cost function. Computing $u_p$ is more difficult since it's implicit.

$$Au = b \implies A\cdot\frac{du}{dp_i} + \frac{dA}{dp_i}u = \frac{db}{dp_i} \implies \frac{du}{dp_i} = A^{-1}\left(-\frac{dA}{dp_i}u + \frac{db}{dp_i}\right)$$

This was for a single component $p_i$ and so we get $u_p = A^{-1}(b_p - A_p u)$. Therefore,

$$\nabla_g = \underbrace{g_p}_{1\times n} + \underbrace{g_u}_{1\times m}\underbrace{A^{-1}}_{m\times m}\underbrace{(b_p - A_p u)}_{m\times n}$$

Denote $\lambda^T = g_u A^{-1} \implies A^T\lambda = g_u^T$. This is called the adjoint equation, which is the same dimension as the system $Ax = b$ but requires only solving a single system of linear equations.

# 11 Chapter 9, 10, 11 in B+V

## 11.1 Ch. 9: Newton's Method

We are solving the unconstrained problem $\min f(x)$ where $f \in C^2$. The method is $x_{k+1} = x_k + \Delta x_{nt}$ where

$$\Delta x_{nt} = \arg\min_{\Delta x} f(x_k) + \langle \nabla f(x_k), \Delta x\rangle + \frac{1}{2}\langle \Delta x \mid \nabla^2 f(x_k) \mid \Delta x\rangle = -\nabla^2 f(x_k)^{-1}\nabla f(x_k)$$

**Example 11.1.** This works for nonconvex optimization if we are careful. Consider the quadratic form $f(x) = \frac{1}{2}x\begin{pmatrix} 1 & 0 \\ -1 & 0\end{pmatrix}x$. Note that $x = 0$ is a saddle point, and in general it's hard to avoid saddle points. Newton's method would take you to the saddle point in one step, whereas a less powerful method like gradient descent can avoid it.

**Remark 11.1.** If $f$ is strongly convex, then $\Delta x_{nt}$ is a descent direction, i.e., there exists $t > 0$ such that $f(x_k + t\Delta x_{nt}) < f(x_k)$. It's sufficient to check that $\langle \nabla f(x_k), \Delta x_{nt}\rangle < 0$. Since $f$ is strongly convex, it's inverse Hessian is positive definite and we get $\langle \nabla f(x_k), -\nabla^2 f(x_k)^{-1}\nabla f(x_k)\rangle < 0$.

**Remark 11.2.** Newton's method is **affine invariant** and unaffected by preconditioning matrices. Define a new norm for $H = H^T \succeq 0$:

$$\|x\|_H^2 := \langle x \mid H \mid x\rangle$$

Let $H_k = \nabla^2 f(x_k)$. The **Newton decrement** is $\lambda(x) = \|\Delta x_{nt}\|_{\nabla^2 f(x)}$ helps measure convergence.

The **damped/guarded** Newton method is as follows:

- Find $\Delta x_{nt}$
- Stop if $\frac{\lambda^2}{2} < \epsilon$
- If not, use a backtracking linesearch $x_{k+1} = x_k + t\Delta x_{nt}$ initialized at $t = 1$.

In the 90's, Nesterov and Nemirovskii developed **self-concordance**. We say $f : \mathbb{R} \to \mathbb{R}$ is self-concordant if it's convex and $|f'''(x)| \le 2\left(f''(x)\right)^{3/2}$. This is affine invariant in the sense that $g(x) = f(ax+b)$ is self-concordant $\iff$ $f$ is self-concordant.

$$|g'''(x)| \le 2(g''(x))^{3/2} \iff |f'''(x)| = |a^3 g'''(x)| \le 2a^3(g''(x))^{3/2} = 2(a^2 g''(x))^{3/2} = 2(f''(x))^{3/2}$$

The factor of 2 out front is "standard self-concordant".

**Example 11.2.** Quadratic and linear functions are trivially self-concordant since their third derivative vanishes. For the function $f(x) = -\log(x)$, we get the equality $|f'''(x)| = 2(f''(x))^{3/2}$.

**Remark 11.3.** If $f$ is self-concordant then $af$ is self-concordant if $a \ge 1$. If $f, g$ are self-concordant then so is $f + g$. The function $f : \mathbb{R}^n \to \mathbb{R}$ is self-concordant if $p(t) = f(x + tv)$ is self-concordant for all $x, v$.

**Example 11.3.** If $X$ is a matrix, then $f(x) = -\log\det X$ is self-concordant.

**Remark 11.4.** Suppose $f$ is strictly convex and self-concordant. This assumption is equivalent to $\nabla^2 f(x) \succ 0$. Then if $\lambda(x) \le 0.68$ we have $f(x) - \min_x f(x) \le \lambda^2(x)$.

**Theorem 11.1.** Assume $f$ is strictly convex and self-concordant. Let $x_0 \in \operatorname{dom}(f)$ be known and the sublevel set $S = \{x \mid f(x) \le f(x_0)\}$ is closed and $f$ is bounded below. (All together, these assumptions imply that there exists a unique minimizer). Then $\exists 0 < \eta < 1/4$ and $\gamma > 0$ such that

- Damped Newton Phase I: $\lambda(x_k) > \eta$ and $f(x_{k+1}) - f(x_k) \le -\gamma$
- Quadratic Convergence Phase II: The $t = 1$ stepsize is selected and $\lambda(x_k) < \epsilon$.

$$2\lambda(x_{k+1}) \le (2\lambda(x_k))^2$$

**Example 11.4.** For $\epsilon = 10^{-10}$ it would take 5.05 iterations in phase II quadratic convergence. For tolerance $\epsilon = 10^{-20}$ need 6.05 iterations. For tolerance $\epsilon = 10^{-40}$ need 7.05 iterations. (Phase I is the time-consuming part!)

**Theorem 11.2.** We can state a convergence result assuming self-concordancy. Start at $x_0$ and let $p^* = \min_x f(x)$. The number of iterations needed to reach an $\epsilon$ solution is

$$\frac{1}{\gamma}\left(f(x_0) - p^*\right) + \log\log(1/\epsilon)$$

for some constant $\gamma = 375$.

## 11.2 Ch. 10: Equality Constrained Minimization

Consider $\min_{Ax=b} f(x)$, somewhat akin to a SQP.

- Motivation 1: $x_{k+1} = x_k + \Delta x$ where

$$\Delta x = \arg\min_{\Delta x} f(x_k) + \langle \nabla f(x_k), \Delta x\rangle + \frac{1}{2}\langle \Delta x \mid \nabla^2 f(x_k) \mid \Delta x\rangle$$

  such that $A(x_k + \Delta) = b$. Assuming $Ax_k = b$ (so we want $A\Delta x = 0$, we use KKT:

$$\begin{pmatrix} \nabla^2 f(x_k) & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ w \end{pmatrix} = \begin{pmatrix} -\nabla f(x_k) \\ 0 \end{pmatrix}$$

- Motivation 2: Apply KKT directly to our minimization problem.

$$\begin{cases} 0 = \nabla f(x) + A^T \nu \\ Ax = b \end{cases}$$

  Linearize KKT by

$$\nabla f(x + \Delta x) \approx \nabla f(x_k) + \nabla^2 f(x_k)\Delta x$$

- Motivation 3: Regarding the linear system $Ax = b$, let $F$ be a parity check $AF = 0$; i.e. $F$ is a basis for null($A$). (Note: null($A$) is not trivial or else this problem is pointless.) Let $x = Fz + x_p$, where $Ax_p = b$. Our minimization problem becomes

$$\min_z f(Fz + x_p)$$

### 11.3 Ch. 11: Inequality Constrained Minimization

Consider the problem $\begin{cases} \min_x f_0(x) \\ f_i(x) \leq 0, \quad i = 1, 2, \ldots, m \\ Ax = b,\ f_i \text{ convex and } C^2 \end{cases}$ We assume there exists $x^* \in \arg\min(\text{problem})$ and a strictly feasible point $f_i(x) < 0$ for $i = 1, \ldots, m$. The KKT conditions read as follows:

- Primal feasibility: $Ax = b$
- Dual feasibility: $\lambda \geq 0$
- Stationarity: $\nabla f_0(x) + \sum_i \lambda_i \nabla f_i(x) + A^T \nu = 0$
- Complementary slackness: $\lambda_i f_i(x) = 0$ for $i = 1, \ldots, m$

Recall the quadratic penalty method, where we penalize $\lfloor f_1(x) \rfloor_+^2$ and want $f_1(x) \leq 0$. We will use a log-barrier $-\log(-f_1(x))$. In particular we consider $-\frac{1}{t}\log(-f_1(x))$ and plot this in the $[f_1(x), -\log(-f_1(x))]$ plane for various values of $t$. Note that as $t \to \infty$ this starts to look like the indicator function.
Formally, we define the barrier $\frac{1}{t}\phi(x) = \sum_{i=1}^m -\frac{1}{t}\log(-f_i(x))$. Note if $f_i$ are convex and $f_i(x) < 0$, then $\phi$ is convex.

$$x^*(t) := \arg\min_{Ax=b}\ t f_0(x) + \phi(x)$$

Call $\{x^*(t)\}_{t>0}$ the central path. The KKT conditions for $x = x^*(t)$ read:

- Feasibity: $Ax = b$ and $f_1(x) < 0$

- Stationarity: $0 = t\nabla f_0(x) + \sum_{i=1}^m \left(-\frac{1}{f_i(x)}\nabla f_i(x)\right) + A^T\hat{\nu}$. Rewrite as $\nu = \hat{\nu}/t$ and let $\lambda_i = -\frac{1}{t}\frac{1}{f_i(x)} > 0$

$$0 = \nabla f_0(x) + \sum_{i=1}^m \left(-\frac{1}{tf_i(x)}\nabla f_i(x)\right) + A^T\nu = \nabla f_0(x) + \sum_{i=1}^m \lambda_i \nabla f_i(x) + A^T\nu$$

Therefore $(\lambda, \nu)$ are feasible dual points for our problem.

$$g(\lambda,\nu) \le d^* \le p^* \le f_0(x) \implies f_0(x) - p^* \le f_0(x) - g(\lambda,\nu) \le \frac{m}{t}$$

So $f_0(x) - p^* \le \frac{m}{t}$ gives us a bound on suboptimality.

This early method was determined by Fiacco and McCormic in the 60s but not seriously analyzed until 80s and 90s.

**Remark 11.5.** Another viewpoint is changing $\lambda_i f_i(x) = 0$ into $\lambda_i f_i(x) = -\frac{1}{t}$. This is important since $f_i(x)$ is nonzero under the log barrier and hence we can solve for $\lambda_i$ explicitly. In this case, we solve $x^*(t) = \arg\min_{Ax=b} tf_0(x) + \phi(x)$, update $t \leftarrow \mu t$ for $\mu > 1$ and repeat the process. The linear solve is with Newton's method, starting at the previous point. To select a strictly feasible initial point $t_0$, we find $x_0$ in Phase I and in Phase II use the regular barrier method. Phase I: pick any $\overline{x}$ and solve $\min_{x,s} s$ with $Ax = b$ and $f_i(x) \le s$ for all $i = 1, \dots, m$. For Phase II: choose $\overline{x}$ and let $s = \max_{i=1,\dots,m} f_i(x)) + c$ where $c > 0$.

**Remark 11.6.** Picking the barrier function. For the LP $\min_{x\ge 0,\ Ax=b} c^T x$, use $\phi(x) = -\frac{1}{t}\sum_{i=1}^m \log(-f_i(x))$. For the SDP $\min_{X\succeq 0,\ A(X)=b} \langle C, X\rangle$ we use $\phi(x) = -\frac{1}{t}\log\det(X)$

**Remark 11.7.** Completxity Analysis: assuming $f_i$ are nice i.e. $\min_{Ax=b} tf_0(x) + \phi(x)$ has a unique solution and $tf_0(x) + \phi(x)$ is self-concordant. For $f_0(x^*(t)) - p^* \le \frac{m}{t}$, we find $x^*(t)$ in $f_0(x_0) - p^* + 6$ iterations. The 6 comes from a $\log\log(\epsilon)$ term. The number of total Newton Steps is given by 11.2.7 in the textbook:

$$\frac{\log\left(\frac{m}{t_0\epsilon}\right)}{\log\mu}\left(\frac{m(\mu-1)-\log\mu}{\gamma} + 6\right)$$

For e.g. the gradient method, could get $e_k = (1-10^{-6})^k$. For the logbarrier method, $e_k \le \frac{1}{2^k}$.

**Remark 11.8.** A practical IPM will be a primal dual method. In particular, use a predictor corrector method, the Mehotra PC method. See Steve Wright's textbook *IPM for linear programs.*

## 11.4 Primal Dual Methods

This is 11.7 in B&V. In the previous log-barrier method, we have an outer loop updating $t \leftarrow \mu t$ and in inner loop performing the Newton steps. In the primal dual method, however, there is a single loop. In general, primal dual is faster but can lose feasibility and hence there is less theory. Sometimes the primal dual will work without strict feasibility.

Goal: set the residual $r_t(x, \lambda, \nu)$ where $r_t = \begin{pmatrix} \nabla f_0(x) + Df(x)^T\lambda + A^T\nu \\ -\text{diag}(\lambda)\cdot f(x) - \frac{1}{L}\mathbf{1} \\ Ax - b \end{pmatrix}$

We compute the Lagrangian:

$$L = f_0(x) + \sum_i \lambda_i f_i(x) + \nu^T(Ax - b)$$

From stationarity, we have $0 = \nabla f_0(x) + \sum \lambda_i \nabla f_i(x) + A^T\nu$. Also need primal and dual feasibility and stationarity. Changing the stationary condition to $f(x) \cdot \lambda = -\frac{1}{t}$ rather than $f(x) \cdot \lambda = 0$. Under this selection, an initial feasible point will result in a sequence of feasible points. We can hence drop the inequality constraints dealiing with feasibility and now only have equality constraints.
How to solve $r_t = 0$? Use Newton's method $0 = T(y + \Delta y) \approx T(y) + (DT)\Delta y \implies \Delta y = -(DT)^{-1}T(y)$. (Think $\Delta y = y_{k+1} - y_k$.) In our case, $T = r_t$ and $y = (x, \lambda, \nu)^T$, and our Newton step is

$$\begin{pmatrix} \nabla^2 f_0(x) + \sum \lambda_i \nabla^2 f_i(x) & Df(x)^T & A^T \\ -\mathrm{diag}(\lambda) Df(x) & -\mathrm{diag}(f(x)) & 0 \\ A & 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta \lambda \\ \Delta \nu \end{pmatrix} = - \begin{pmatrix} r_{\text{dual}} \\ r_{\text{central}} \\ r_{\text{primal}} \end{pmatrix}$$

Note that the lower right hand block on the leftmost matrix is simply diagonal and zeros.

# 12 First Order Methods

We have already studied Gradient descent and Nesterov's accelerated version. In a similar vein, we saw Proximal Gradient descent and the related FISTA algorithm.
Here we take a look at the proximal point method. The problem is $\min\limits_x f(x)$ where $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. Note that

$$\min_y \left( \min_x f(x) + \frac{1}{2}\|x - y\|^2 \right)$$

is an equivalent problem. The proximal point algorithm is as follows:

- Solve $x_{k+1} = \arg\min f(x) + \frac{\mu}{2}\|x - y_k\|^2 = \operatorname*{prox}\limits_{\mu^{-1}df}(y_k)$
- Set $y_{k+1} = y_k - \left(\frac{1}{\mu} \cdot \mu\right)(y_k - y_{k-1}) = x_{k+1}$

In other words, $y_{k+1} = (I + \mu^{-1}df)^{-1}(y_k)$. This is proximal gradient descent with zero gradient, or alternatively, an implicit Euler scheme. We call $\phi(y) := \min\limits_x f(x) + \frac{\mu}{2}\|x - y\|$ the **infimal convolution** or the **Moreau envelope** of $f$. Note that

$$\nabla \phi = \nabla \min_x \psi(x, y) = \frac{d\psi}{dy}(x^*, y) = \mu(y - x^*)$$

**Remark 12.1.** In the adapt reg by Allen-Zhu and Hazan, choose $x_{k+1} \approx \min\left(f(x) + \frac{\mu}{2}\|x\|^2\right)$ and then $\mu_{k+1} = \mu_{k/2}$. Also look at the Universal Catalyst problem, which takes any method converging linearly and essentially creates a Nesterov accelerated version.

In the Conditional Gradient Method (Frank-Wolfe), we consider the constrained problem $\min\limits_{x \in C} f(x)$ where $C$ is compact. More generally, $\min_x f(x) + g(x)$ where $g$ is coercive. The algorithm looks like

$$\begin{cases} s_k \in \arg\min\limits_x g(x) + f(x_k) + \langle df(x_k), x - x_k \rangle \\ x_{k+1} = (1 - \gamma_k)x_k + \gamma_k \cdot s_k \end{cases}$$

Stepsize $\gamma_k = \frac{2}{k+2}$.

## 12.1 ADMM

Alternating Direction Method of Multipliers or Douglas-Rachford. We'll follow `https://web.stanford.edu/~boyd/papers/admm_distr_stats.html`. The problem is $\min\limits_{Ax=b} f(x)$. To solve via the dual, first set up

$$L(x,y) = f(x) + \langle y, Ax-b\rangle \implies g(y) = \inf_x L(x,y) = -f^*(-A^Ty) - b^Ty$$

The gradient ascent steps would be as follows:

$$\begin{cases} x_{k+1} = \arg\min\limits_x L(x,y_k) \\ y_{k+1} = y_k + z(Ax_{k+1}-b) \end{cases}$$

Why do this? If $f(x) = \sum_{i=1}^n f_i(x_i)$ is separable, then

$$x_{k+1} = \arg\min_x \sum_{i=1}^n f_i(x_i) + (A^Ty)^Tx$$

is a separable problem. Also might be called the dual decomposition and is related to specialized methods for LPs; e.g. Dantzig-Wolfe decomposition.

Another way of motivating the setup is through the Augmented Lagrangian with $\min\limits_{Ax=b} f(x) + \frac{\rho}{2}\|Ax-b\|^2$ with $\rho > 0$. In particular, we apply the dual decomposition:

$$\begin{cases} x_{k+1} = \arg\min L_\rho(x,y_k) \\ y_{k+1} = y_k + \rho(Ax_{k+1}-b) \end{cases}$$

But this is harder to solve since it's now coupled with constraints.

The ideas is to write $x_k = \begin{pmatrix} x_k^{(1)} \\ x_k^{(2)} \end{pmatrix}$ and then

$$\begin{cases} x_{k+1}^{(1)} = \arg\min\limits_{x^{(1)}} L_\rho(x^{(1)}, x_k^{(2)}, y_k) \\ x_{k+2}^{(2)} = \arg\min\limits_{x^{(2)}} L_\rho(x_{k+1}^{(1)}, x^{(2)}, y_k) \end{cases}$$

In the general case for $\min\limits_{Ax+Bz=c} f(x)+g(z)$, we have the augmented lagrangian

$$L_\rho(x,z,y) = f(x) + g(z) + \frac{\rho}{2}\|Ax+Bz-c\|^2 + \langle y, Ax+Bz-c\rangle$$

and want to deteremine $(x,z) \leftarrow \arg\min L_\rho(x,z,y)$. The ADMM solution is

$$\begin{cases} x_{k+1} \in \arg\min\limits_x L_\rho(x,z_k,y_k) \\ z_{k+1} \in \arg\min\limits_z L_\rho(x_{k+1},z,y_k) \\ y_{k+1} = y_k + \rho(Ax_{k+1}+Bz_{k+1}-c) \end{cases}$$

Note that the $(x,z)$ updates follow a Gauss-Siedel form. The convergence rate is akin to that of a first order method.

**Remark 12.2.** Consider the more general problem $\min\limits_x \sum_{i=1}^3 f_i(x)$. The trick is to life to a bigger space via

$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ with $F(\mathbf{x}) = \sum_{i=1}^{3} f_i(x_i)$. Now, we have

$$\min_x F(\mathbf{x}) + G(\mathbf{x})$$

where $G(\cdot)$ is the consensus indicator function of $\{\mathbf{x} : x_1 = x_2 = x_3\}$. The updates are now ADMM in the form of proximal operators. The proximal operator of $F(\cdot)$ is separable by construction and can often can be parallelized in practice. The proximal operator of $G(\cdot)$ is a projection onto the set $\{\mathbf{x} : x_1 = x_2 = x_3\}$ which amounts to averaging each component.

**Remark 12.3.** Note that no assumptions about the smoothness of $f$, $g$ were made in ADMM. For gradient descent we have an idea on how to select the stepsize, but in ADMM the convergence depends on the mysterious parameter $\rho$.

## 12.2 Douglas-Rachford

This follows B+C 2nd edition 2017, 20.3. Analysis of Douglas-Rachford from Lions, Mercie 1979. Let $f, g \in \Gamma_0(\mathbb{R}^n)$, and assume that $d(f+g) = df + dg$ and there exist primal and dual saddle points. The primal problem is

$$(P) = \min_x f(x) + g(x)$$

and the dual problem is

$$(D) = \min_u -f^*(-u) + g^*(u)$$

The algorithm is

$$\begin{cases} x_k = \underset{\rho g}{\operatorname{prox}}(y_k) \\ z_k = \underset{\rho f}{\operatorname{prox}}(2x_k - y_k) \\ y_{k+1} = y_k + \lambda(z_k - x_k) \end{cases}$$

Here $\lambda \in (0,2)$ is a relaxation parameter and $\rho > 0$. ($\rho$ might be the inverse of $\rho$ in a previous section.)

**Theorem 12.1.** If $y_k \to y$ and $x = \underset{\rho g}{\operatorname{prox}}(y)$, then $x$ is a primal optimal point.

**Remark 12.4.** Let's motivate the updates:

$$\begin{aligned} 0 &\in df(x) + dg(x) \\ x - dg(x) &\in x + df(x) \\ 2x - (I + dg)(x) &\in (I + df)(x) \\ x &= (I + df)^{-1}(2x - (I + dg)x) \end{aligned}$$

Identify $y = (I + dg)(x)$ and $x = (I + dg)^{-1}(y)$ and $z = (I + df)^{-1}(2x - (I + dg)x)$, giving the fixed point equation

$$\begin{cases} 0 = z - x \\ y \leftarrow y + z - x \end{cases}$$

**Remark 12.5.** ADMM is a special case of Douglas Rachford.

### 12.3 Primal-Dual Methods

See Becker, Schmidt 2014. The problem is $\min_x g(x) + h(Ax)$. Note that $\text{prox}_{hA}(\cdot)$ is not easy even if $\text{prox}_h(\cdot)$ is easy. In 2011 Chamballe and Pack suggest a primal-dual hybrid gradient method and a preconditioned ADMM. Another is from Laurent + Condat 2011, where we consider

$$\min_x f(x) + g(x) + h(Ax)$$

where $f, g \in \Gamma_0(\mathbb{R}^n)$ and $h \in \Gamma_0(\mathbb{R}^m)$ and $A$ is $m \times n$ and $\nabla f$ is L-Lipscihtz and $g, h$ have easy-to-compute proximal operators. Again, the issue is that $h \circ A$ might not have a nice proximal operator.
Note that $h(z) = (h^*)^* = \sup_y \langle y, z\rangle - h^*(y)$. We can therefore write

$$p = \min_x \max_y (f+g)(x) + \langle Ax, y\rangle - h^*(y)$$

Assuming constraint qualifications so that the subdifferential of sum is sum of subdifferentials, we have the monotone inclusion problem

$$0 \in df(x) + dg(x) + A^* dh(Ax)$$

With $y = dh(Ax)$, we rewrite this as

$$0 \in \nabla f(x) + dg(x) + A^* y$$

and observe

$$y \in dh(Ax) \iff Ax \in dh^{-1}(y) \iff Ax \in dh^*(y)$$

Slightly abusing notation, we rewrite the two above equations as

$$-\begin{pmatrix} \nabla f & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \in \begin{pmatrix} dg & A^* \\ -A & dh^* \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

or in shorthand $-T_2\mathbf{x} = T_1\mathbf{x}$. Note that the leftmost matrix is not symmetric (e.g. not a Hessian matrix) and hence this monotone inclusion problem is a bit more general than a typical minimization problem.
For $0 \in \nabla f(x) + dg(x)$, we have proximal gradient descent:

$$(I - \nabla f)(x) = (I + dg)(x) \implies x = (I + dg)^{-1}(I - \nabla f)(x)$$

In our problem, the update is

$$\mathbf{x} = (I + T_1)^{-1}(I - T_2)\mathbf{x}$$

and hence we get an update of the form $\mathbf{x}_{k+1} = (I + T_1)^{-1}(I - T_2)\mathbf{x}_k$. Note that $(I + T_1)^{-1}$ involves

$$\begin{pmatrix} I + dg & A^* \\ -A & I + dh^* \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix}$$

and this is difficult to solve. We could instead consider $0 \in V^{-1}(T_1\mathbf{x} + T_2\mathbf{x})$ for $V \succ 0$ so that our updates take the form

$$\mathbf{x}_{k+1} = (V + T_1)^{-1}(V - T_2)\mathbf{x}_k$$

We choose $V = \begin{pmatrix} I/\tau & -A^* \\ -A & I/\sigma \end{pmatrix}$ for tuning parameters $\tau, \sigma$. If $\sigma\tau > \|A\|^2$, then $V \succ 0$. Now when computing the inverse $(V+T_1)^{-1}$ we have a system that can be solved efficiently with a back substitution:

$$\begin{pmatrix} I/\tau + dg & 0 \\ -2A & I/\sigma + dh^* \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix}$$

### 12.4 Other

Alternating minmization and coordinate descent are two other common minimization techniques. We consider $\min\limits_{(x,y)} f(x,y)$ where $(x,y) \in \Omega_x \times \Omega_y$. The updates are in the Gauss-Siedel style where we always use the latest update:

$$\begin{cases} x_{k+1} = \arg\min\limits_{x \in \Omega_x} f(x, y_k) \\ y_{k+1} = \arg\min\limits_{y \in \Omega_y} f(x_{k+1}, y) \end{cases}$$

The Jacobi style doesn't use the most recent update and hence is natural to parallelize.
A nonconvex version is the PALM by Botte, Sebach, and Teboulle. Here,

$$\begin{cases} x_{k+1} = \arg\min\limits_{x \in \Omega_x} f(x, y_k) + \frac{\mu}{2}\|x - x_k\|^2 \\ y_{k+1} = \arg\min\limits_{y \in \Omega_y} f(x_{k+1}, y) + \frac{\mu}{2}\|y - y_k\|^2 \end{cases}$$

**Example 12.1.** Consider projection onto the set $C_1 \cap C_2$ given that projection onto the sets $C_1, C_2$ are straightforward. One solution is POCS (Projection Onto Convex Sets):

$$x_{k+1} = (P_{C_1} \circ P_{C_2}) x_k$$

A simple parallel method is

$$x_{k+1} = \frac{1}{2}(P_{C_1} + P_{C_2}) x_k$$

and this is straightforward to generalize to more than 2 projections.

## 13 Linear Programming

### 13.1 Simplex Method

Developed by Dantzig in late 1940s. One of the best ways to solve LPs. Software: CPLEX.
We say $v \in C$ is a vertex if and only if $\begin{cases} v + h \in C \\ v - h \in C \end{cases} \implies h = 0$
The idea is we always have a vertex solution. See Nocedal+Wright textbook for an explanation of the algorithm.

### 13.2 Integer Linear Programming

This is Chapter 23 in Robert Vanderbei's LP book. We consider an Integer Linear Program (ILP) $\begin{cases} \max c^T x \\ Ax \leq 0 \\ x \geq 0 \\ x \in \mathbb{Z} \end{cases}$
Variants are the mixed ILP and binary ILP. Slow for large problems but can find the solution despite nonconvexity.

**Example 13.1.** Consider $p^* = \max_{x,y} 17x + 12y$ such that $10x + 7y \leq 40$ and $x + y \leq 5$ and $x, y$ are nonnegative integers. The Branch-and-Bound method is as follows:

- Solve the relaxation (LP) without $x, y \in \mathbb{Z}$ so that $(x, y) = (\frac{5}{3}, \frac{10}{3})$
- Now, branch $\min_{x \in C_1 \cup C_2} f(x)$ into $\min_{x \in C_1} f(x)$ and $\min_{x \in C_2} f(x)$. Note that in the context of integer programming, $\{x \leq 1\} \cup \{x > 1\} = \{x \leq 1\} \cup \{x \geq 2\}$ so we can exclude large regions. Branching our guess $(\frac{5}{3}, \frac{10}{3})$ into $\{x \leq 1\} \cup \{x \geq 2\}$ gives $p_1 = (1, 4)$ with $p = 65$, and $p_2 = (2, 2.86)$ with $p = 68.29$.
- The latter is not an integer solution, branch into $\{y \leq 2\} \cup \{y \geq 3\}$. Solving this $y \leq 2$ branch gives $p_3 = (2.6, 2)$ with $p = 68.2$, and the $y \geq 3$ branch is infeasible. Split $(2.6, 2)$ into $\{x \leq 2\} \cup \{x \geq 3\}$. The $\{x \leq 2\}$ gives the integer solution $p_4 = (2, 2)$ with $p = 58$, and we can prune this branch since we already have a better solution. The $\{x \geq 3\}$ gives $(3.1, 1.43)$ with $p = 68.14$. The $y \geq 2$ branch is infeasible, and the $y \leq 1$ branch is $(3, 3.1)$ with $p = 68.1$. Split $(3, 3.1)$ into $\{x \leq 3\} \cup \{x \geq 4\}$. Solve $x \leq 3$ with $(3, 1)$ and $p = 63$, this branch is pruned. For $x \geq 4$, we get $(4, 0)$ with $p = 58$, and this is the solution.

**Remark 13.1.** Good software for ILP include CPLEX, Gurobi, Xpress. Also try GLP which might be slower.

## 14 Stochastic Methods

Consider $\min_x f(x)$. Note that $f(x) = \mathbb{E}_\xi F(x; \xi)$ could be a random variable; this is the Stochastic Approximation (SA). Earliest example of SA was Robbins, Monroe, and Polyak. Alternatively, we formulate the problem in the context of empirical risk minimization, where $f(x) = \frac{1}{N}\sum_{i=1}^N f_i(x)$ with $f_i = F(x; \xi_i)$. This is the Sample Average Approximation (SAA).
Define the error $e_k = f(x_k) - f^*$. We may want to show $L^1$ convergence; i.e. $\mathbb{E}[|e_k|] \xrightarrow{k\to\infty} 0$. Convergence in $L^2$ would be $\mathbb{E}[|e_k|^2] \xrightarrow{k\to\infty} 0$. For convergence in probability, $e_k \xrightarrow{P} 0$, we show $\lim_{k\to\infty} \mathbb{P}(|e_k| > \epsilon) = 0$, for all $\epsilon > 0$. Finally, for almost sure convergence denoted $e_k \xrightarrow{a.s.} 0$, we show $\mathbb{P}(\lim_{k\to\infty} e_k = 0) = 1$.

**Example 14.1.** Consider $e_k = \begin{cases} 1 & \text{w.p. } 1/k \\ 0 & \text{w.p. } 1 - 1/k \end{cases}$ Note that $\mathbb{E}[e_k] = \frac{1}{k} \to 0$. It is a fact that $e_k \xrightarrow{L^p} 0 \implies e_k \xrightarrow{P} 0$. Don't think it converges a.s.
Alternatively, $e_k = \begin{cases} \sqrt{k} & \text{w.p. } 1/k \\ 0 & \text{w.p. } 1 - 1/k \end{cases}$ Here, we have $\mathbb{E}[e_k] = \frac{1}{\sqrt{k}} \to 0$, but this doesn't converge in $L^2$.

Let's introduce Stochastic Gradient Descent (SGD). Here,

$$x_{k+1} = x_k - t_k d_k$$

where $t_k$ is the stepsize and $\mathbb{E}[d_k] = \nabla f(x_k)$. For the SAA setup, we would select $d_k = \nabla f_i(x_k)$ where $d_k \sim \text{Unif}[1, \ldots, N]$. In practice, we draw multiple indices, called a minibatch.

**Theorem 14.1** (Bottou, Curtic, Nocedal $\sim$ 2017)**.** Let $\nabla f$ be L-Lipschitz. Assume $f$ is $\mu$ strongly convex (or $\mu$ PL inequality), and $f \geq 0$. We also make an assumption on the variance of $d_k$ in addition to the first moment condition $\mathbb{E}[d_k] = f_k$, namely that $\mathbb{E}[\|d_k\|^2] \leq M + \|\nabla f(x_k)\|^2$. Part 1: if $t_k = t = \frac{1}{L}$,

$$\mathbb{E}[f(x_k) - f^*] \leq \frac{tLM}{2\mu} + (1 - t\mu)^{k-1}\left(f(x_k) - f^* - \frac{tLM}{4\mu}\right)$$

Note that the term $\frac{tLM}{2\mu}$ is constant and this is not convergence to zero.
Part 2: Choose $t_k = \frac{\beta}{\gamma+k}$ where $\gamma > 0$ and $\beta > \frac{1}{\mu}$. Then

$$\mathbb{E}[f(x_k - x^*)] \leq \frac{\nu}{\gamma + k}$$

where $\nu$ is some constant depending on $\gamma$ and $\beta$. So we do converge if we use a decreasing step size.

**Remark 14.1.** Useful in low-precision settings with lots of similar training data.

## 14.1 Variance Reduced Methods for SAA

Assume $f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x) < \infty$. For example, in the context of general linearized models, $f_i(x) = L(a_i^T x - b_i)$. In SAGA, we first pick $x^{(1)}$ and store $\{\nabla f_i(x^{(i)})\}_{i=1}^{N}$ in a $n \times N$ table. For $k = 1, 2, \ldots$, we draw $j \sim \text{Unif}[1, \ldots, N]$ and define $\overline{z} = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x^{(i)})$ via the table.

$$\begin{cases} x_{k+1} = x_k - t_k(\nabla f_j(x_k) - \nabla f_j(x^{(j)}) + \overline{z}) \\ x^{(j)} = x_k \end{cases}$$

Note the addition of the control variates in the $x_{k+1}$ step. In addition to SAG and SAGA, there is SVRG.

**Theorem 14.2.** For appropriate $t$, this converges linearly.

**Remark 14.2.** Define $\overline{x_k} = \frac{1}{k}\sum_{j=1}^{k} x_j$. Sometimes we can get better convergence rate using the average iterates.