Convex Analysis and Optimization for Econ-CS

August 9, 2021

Denizalp Goktas

Contents

1 Preliminaries

1.1 Metric Spaces

A **metric space** is an **ordered tuple** (E, d) that consists of a set E and , a function, i.e., a **metric**, $d : E \times E \to \mathbb{R}_+$ such that for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in E$ the following hold:

1. (Non-Degeneracy) $d(\boldsymbol{x}, \boldsymbol{y}) = 0 \iff \boldsymbol{x} = \boldsymbol{y}$
2. (Triangle Inequality) $d(\boldsymbol{x}, \boldsymbol{z}) \geq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$
3. (Symmetry) $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$

A sequence $\{\boldsymbol{x}_n\}_n \subset E$ is said to be a **Cauchy sequence** if for all $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $n, m > N$, we have $d(\boldsymbol{x}_n, \boldsymbol{x}_m) < \epsilon$.

A metric space (E, d) is said to be **complete** if any Cauchy sequence $\{\boldsymbol{x}_n\}_n \subset E$, $\boldsymbol{x}_n \to \boldsymbol{x}$ for $\boldsymbol{x} \in E$. That is, a metric space is complete if any Cauchy sequence in the set has a limit within the set. An example of a complete metric space is $(\mathbb{R}^n, d)$ with $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$.

Let (E, d_E) and (F, d_F) be Banach spaces. A function $f : E \to F$ is said to be continuous if for all $\varepsilon > 0$, there exists $\delta > 0$ such that if $d_F(f(\boldsymbol{x}), f(\boldsymbol{y})) < \epsilon$, then $d_E(\boldsymbol{x}, \boldsymbol{y}) < \delta$ for all $\boldsymbol{x}, \boldsymbol{y} \in E$.

A function $T : E \to F$ is said to be **linear** if it satisfies:

1. (Linearity) $\forall \boldsymbol{x}, \boldsymbol{y} \in E,\ \ T(\boldsymbol{x} + \boldsymbol{y}) = T(\boldsymbol{x}) + T(\boldsymbol{y})$
2. (Homogeneity) $\forall \boldsymbol{x} \in E, c \in \mathbb{R},\ \ T(c\boldsymbol{x}) = cT(\boldsymbol{x})$

1.2 Normed Spaces

A **Normed Space** is an ordered tuple $(E, ||.||)$ that consists of a set E and a function, i.e., a **norm**, $||.||: E \to \mathbb{R}_+$ such that for all $\boldsymbol{x}, \boldsymbol{y} \in E$ the following hold:

1. (Normalized) $||\boldsymbol{x}|| = 0 \iff \boldsymbol{x} = 0$
2. (Homogeneity) $\forall c \in \mathbb{R},\ \ ||c\boldsymbol{x}|| = |c|||\boldsymbol{x}||$
3. (Triangle Inequality) $||\boldsymbol{x} + \boldsymbol{y}|| \geq ||\boldsymbol{x}|| + ||\boldsymbol{y}||$

Note that any norm $||.||$ induces a metric $d(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||$ such that (E, d) is a valid metric space. That is, any normed space is also a metric space.

If E is complete with respect to the metric d, then $(E, ||.||)$ is said to be a **Banach space**, or a complete normed space. More often than not, we assume that E is a **vector space**, e.g., $\mathbb{R}^n$, in which case $(E, ||.||)$ is called a **normed vector space (NVS)**.

Let $(E, ||.||_E)$ and $(F, ||.||_F)$ be Banach spaces. An example of a complete Banach space is $(\mathbb{R}^n, ||.||_2)$.

A function $f : E \to F$ is said to be continuous if for all $\varepsilon > 0$, there exists $\delta > 0$ such that if $||f(\boldsymbol{x}) - f(\boldsymbol{y})|| < \epsilon$, then $||\boldsymbol{x} - \boldsymbol{y}|| < \delta$ for all $\boldsymbol{x}, \boldsymbol{y} \in E$. A function $T : E \to F$ is said to be **bounded** if there exists $M < \infty$ such that for all $\boldsymbol{x} \in E$:

$$||T(\boldsymbol{x})||_F < M||\boldsymbol{x}||_E \tag{1}$$

1.3 Topological Spaces

A **topological space** is an ordered tuple $(E, \mathcal{T})$ that consists of a set E and and a collection $\mathcal{T}$ of (open) subsets of E, i.e., **a topology**, such that:

1. E is closed under arbitrary unions, i.e., if $\{U_\alpha\}_{\alpha \in A}$, then $\bigcup_{\alpha \in A} U_\alpha \in \mathcal{T}$

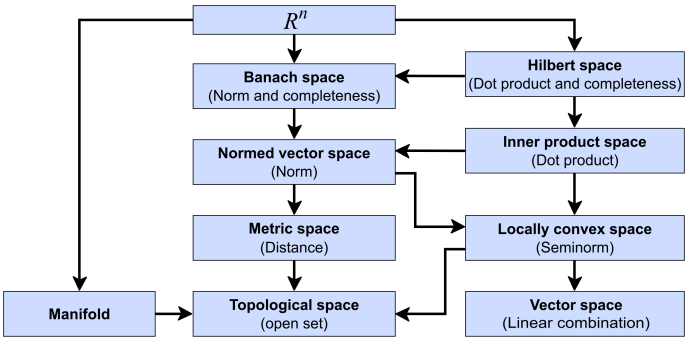

Figure 1: The relationship between different mathematical spaces. An arrow indicates that a space is a kind of another space, e.g., a metric space is a kind of topological space.

2. E is closed under finite intersections, i.e., if $\{U_i\}_{i=1}^n$, then $\bigcap_{i=1}^n U_i \in \mathcal{T}$

If $\boldsymbol{x} \in E$ (or $X \subset E$), a **neighborhood** of $\boldsymbol{x}$ (or E) is a set $N \subset E$ such that $\boldsymbol{x} \in N^\circ$ (or $X \subset N^\circ$) Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be two topological spaces. A function $f : E \to F$ is said to be **continuous** if for all $U \subset F$, $f^{-1}(U) \in E$ and $f^{-1}(U)$ is open. That is, a function f is continuous if for all $\boldsymbol{x} \in E$, $f^{-1}(U)$ is a neighborhood of $\boldsymbol{x}$, for every neighborhood U of $f(\boldsymbol{x})$.

Remark 1.1. *Any normed space is a metric space, and any metric space is a topological space. The relationship between different mathematical spaces can be observed in fig. 1.*

1.4 Derivatives

An operator $f : E \to F$ is said to be **(Fréchet) differentiable at $\boldsymbol{a}$** if there exists a bounded linear operator, named the **Fréchet derivative at $\boldsymbol{a}$**, $Df(\boldsymbol{a}) : E \to F$, such that:

$$\lim_{\|\boldsymbol{h}\| \to 0} \frac{\|f(\boldsymbol{a}+\boldsymbol{h}) - f(\mathbf{a}) - (Df(\mathbf{a}))\,\boldsymbol{h}\|_F}{\|\boldsymbol{h}\|_E} = 0 \tag{2}$$

If the Frćhet derivative exists for all $\boldsymbol{a} \in \mathrm{dom}(f) = E$, then f is said to be (Frćhet) differentiable. A differentiable function f is said to be C^1 if $Df : E \to \mathcal{L}(E, F)$, where $\mathcal{L}(E, F)$ is the space of continuous (=bounded) linear operators from E to F, is continuous. If $E = \mathbb{R}^n$ and $F = \mathbb{R}$, then we denote the Frćhet derivative Df by ∇f and call it the **gradient**.

An operator $f : E \to F$ is said to be **(Gâteau) differentiable** at $\boldsymbol{a}$, if for all $\boldsymbol{x} \in E$, there exists a bounded linear operator, named the **Gâteau derivative at $\boldsymbol{a}$ in the direction of $\boldsymbol{x}$**, $D_{\boldsymbol{x}} f(\boldsymbol{a}) : \mathbb{R}^n \to \mathbb{R}^n$, such that:

$$D_{\boldsymbol{x}} f(\boldsymbol{a}) = \frac{f(\boldsymbol{a}+t\boldsymbol{x}) - f(\mathbf{a})}{t} \tag{3}$$

If the f is Gâteau differentiable for all $\boldsymbol{a} \in \mathrm{dom}(f) = E$, then f is said to be (Gâteau) differentiable. If $E = \mathbb{R}^n$ and $F = \mathbb{R}$, then we denote the Gâteau derivative in the direction of $\boldsymbol{x}$ $D_{\boldsymbol{x}} f$ by $\nabla_{\boldsymbol{x}} f$ and call it the **directional derivative of f in the direction of $\boldsymbol{x}$**.

For the purposes of this document, we will consider the Banach space $(\mathbb{R}^n, ||.||_2)$ and will often deal with **functionals** which are functions of the form $f : E \to \mathbb{R}$, i.e., functions that map from a Banach space to the set of reals. As a result will only deal with gradients and directional gradients. Unfortunately, often the functions that we deal with might not be differentiable, in this case regular calculus will not get us far and we will have to use subdifferential calculus, which can be considered generalization of calculus to non-differentiable functions.

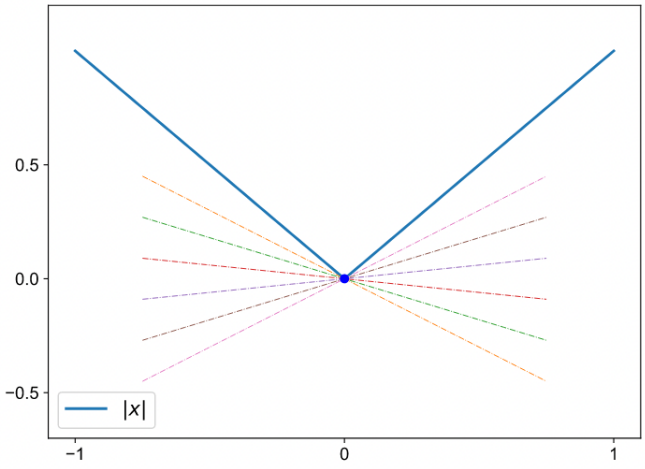

Figure 2: Different subgradients of the absolute value function $f(x) = |x|$ at $x = 0$. The set of subifferentials at $x = 0$ is given by $\partial_x f(0) = \{y \mid |y| < 1\}$

1.5 Subdifferential Calculus

Recall that a functional $f : \mathbb{R}^n \to \mathbb{R}$ is **convex** if:

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \lambda \in (0,1) \qquad f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \geq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) \tag{4}$$

We say that a vector $\boldsymbol{h} \in \mathbb{R}^n$ is a **subgradient** of a continuous, convex function $f : \mathbb{R}^n \to \mathbb{R}$ at $\boldsymbol{a} \in \mathrm{dom}(f)$ if for all $\boldsymbol{x} \in \mathrm{dom}(f)$:

$$f(\boldsymbol{x}) \geq f(\boldsymbol{a}) + \boldsymbol{h}^T(\boldsymbol{x} - \boldsymbol{a}) \tag{5}$$

A function can have multiple subgradients at a point $\boldsymbol{a}$, e.g., fig. 2. The set of all subgradients $\boldsymbol{h}$ at a point $\boldsymbol{a}$ satisfying the above condition for a function f of x is called the **subdiffential of** f **at** $\boldsymbol{a}$ and is denoted $\partial_{\boldsymbol{x}} f(\boldsymbol{a}) = \{\boldsymbol{h} | f(\boldsymbol{x}) \geq f(\boldsymbol{a}) + \boldsymbol{h}^T(\boldsymbol{x} - \boldsymbol{a})\}$. If f is convex and differentiable, then its subdifferential at any point is equal to its gradient, i.e., $\forall \in \mathbb{R}^n, \;\; \partial_{\boldsymbol{x}} f(\boldsymbol{a}) = \{\nabla_{\boldsymbol{x}} f(\boldsymbol{a})\}$.

Subgradients satisfy the additivity property:

$$\partial_{\boldsymbol{x}}(f + g)(\boldsymbol{a}) = \partial_{\boldsymbol{x}} f(\boldsymbol{a}) + \partial_{\boldsymbol{x}} g(\boldsymbol{a}) \tag{6}$$

where the right-hand sum is the minkowski sum, i.e., $\partial_{\boldsymbol{x}} f(\boldsymbol{a}) + \partial_{\boldsymbol{x}} g(\boldsymbol{a}) = \{\boldsymbol{b} + \boldsymbol{c} \mid \boldsymbol{b} \in \partial_{\boldsymbol{x}} f(\boldsymbol{a}), \boldsymbol{c} \in \partial_{\boldsymbol{x}} g(\boldsymbol{a})\}$. Subgradients satisfy also the composition property:

$$\partial_{\boldsymbol{x}} g \circ f(\boldsymbol{a}) = g' \circ \partial_{\boldsymbol{x}} f(\boldsymbol{a}) \tag{7}$$

where g is a differentiable function with derivative g' and $g' \circ \partial_{\boldsymbol{x}} f(\boldsymbol{a}) = \{g'(\boldsymbol{b}) \mid \boldsymbol{b} \in \partial_{\boldsymbol{x}} f(\boldsymbol{a})\}$.

2 Convex Sets, Convex Functions and Dual Representations

2.1 Convex Sets and Dual Representation

We define the **dual space** E^* of E as the space of all continuous linear functionals on E. Note that for $E = \mathbb{R}^n$, $E^* = \mathbb{R}^n$ since any linear functional on $E = \mathbb{R}^n$ is simply a a dot product with another vector in $\mathbb{R}^n$. That is, for $E = \mathbb{R}^n$, any continuous linear functional $f \in E^*$ on E can be simply expressed as $f(\boldsymbol{x}) = \langle \boldsymbol{c}, \boldsymbol{x} \rangle$ for some $c \in E$. Given $f \in E^*$ and $x \in E$, by convention, we write $f(x) = \langle f, x \rangle$ and make no distinction between the "functional form" of the function f its "vector form".

An **(affine) hyperplane** is a subset $H \subset E$, of the form $H = \{\boldsymbol{x} \in E; f(\boldsymbol{x}) = \alpha\}$ where $f \in E^*$. A **halfspace** is defined as the space on either sides of a hyperplane. The **closed upper halfspace** (resp. **lower halfspace**) defined by a hyperplane $H = \{\boldsymbol{x} \in E \mid f(\boldsymbol{x}) = \alpha\}$ is given by $\{\boldsymbol{x} \in E \mid f(\boldsymbol{x}) \geq \alpha\}$ (resp. $\{\boldsymbol{x} \in E \mid f(\boldsymbol{x}) \leq \alpha\}$). The halfspaces are closed (instead of open) if the inequality holds strictly.

A set $S \subset E$ is said to be **convex** if for any collection of points in S, $\{x_i\}_1^n \subset S$, and non-negative numbers $\{\lambda_i\}_1^n \subset \mathbb{R}_+$ s.t. $\sum_{i=1}^n \lambda_i = 1$, we have $\sum_{i=1}^n \lambda_i x_i \in S$ We now introduce **the separating hyperplane theorem** (or **Hahn Banach's geometric form**), one of the most important results in functional analysis.

Theorem 2.1 (Separating Hyperplane Theorem)**.** *Let $C \subset E$ and $D \subset E$ be two nonempty convex subsets such that $C \cap D = \emptyset$. Then, there exists a hyperplane H that separates A and B.*

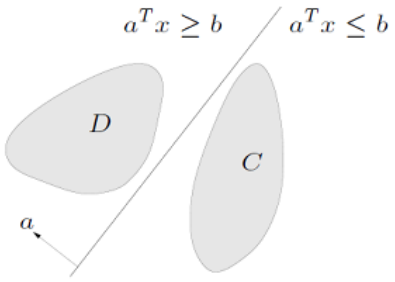

Figure 3: The Seperating Hyperplane Theorem

There are two equivalent ways to represent a convex set $S \subset E$:

1. The union of points in the set (standard/primal representation)
2. The intersection of halfspaces containing the set (the dual representation)

At the core of convex programming duality and many duality relationships in mathematics lies this representation duality that exists for convex sets.

Note: Duality is very overused word, in general, however, it is used to note a way to group elements of a set in pairs, e.g., Linear Programming duality allows us to connect two distinct problems together.

The standard/primal representation of sets is the one that we are used to and is simply the collection of all points contained in the set. However, it turns out that by the separating hyperplane theorem, any convex set can also be represented as an intersection of halfspaces that contain it. The hyperplanes that define the halfspaces in which the set is contained are called **supporting hyperplanes**.

Theorem 2.2 (Convex Set Duality)**.** *Every closed set can be expressed as the intersection of all closed hyperspaces containing it.*

Proof. Let $S \subset E$ be a closed convex set and let $\mathcal{H}$ be the collection of halfspaces that contain S, i.e., $\forall H \in \mathcal{H}, S \subseteq H$. Clearly, we have $S \subseteq \bigcap_{H \in \mathcal{H}} H$ since every H contains S by definition. We then just have to prove no point outside of S is part of the intersection of halfspaces to prove equality.

Let $x \notin S$. Since $\{x\}$ and S are closed and convex sets, by the seperating hyperplane theorem, there exists a hyperplane and a as a result a halfspace H s.t. $H \in \mathcal{H}$ and $x \notin H$. Hence, we have that $x \notin \bigcap_{H \in \mathcal{H}} H$ □

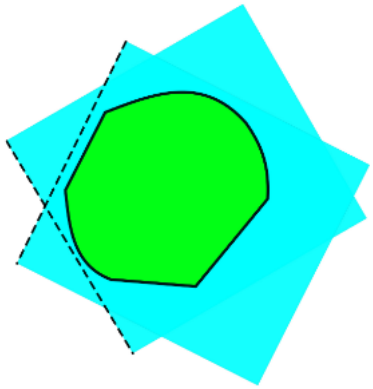

Figure 4: Representing a convex set as intersection of halfspaces containing it.

3 Convex Functions and Dual Representations

Often, in convex analysis we study functions via a set representation of their graph. A set representation of a function is equivalent to the standard definition of a function, however, it is often more convenient to consider the set representation as it provides geometric intuition. A convex functional f can be represented as a set in two ways. Set representation of functions come in handy in convex analysis as they allow us to understand the geometric properties of functions better.

The standard/primal set representation of a functional $f : E \to \mathbb{R}$ is through its **epigraph**, $\text{epi} f$, which is the set of points lying above the function's graph. That is:

$$\text{epi} f = \{(\boldsymbol{x}, y) \in E \times \mathbb{R} \mid f(\boldsymbol{x}) \leq y\} \tag{8}$$

The epigraph allows us to represent a function as the set of points above its graph. Assuming that f is continuous (actually lower semi continuity is enough), its epigraph $\text{epi} f$ is also closed. Now, since we can represent

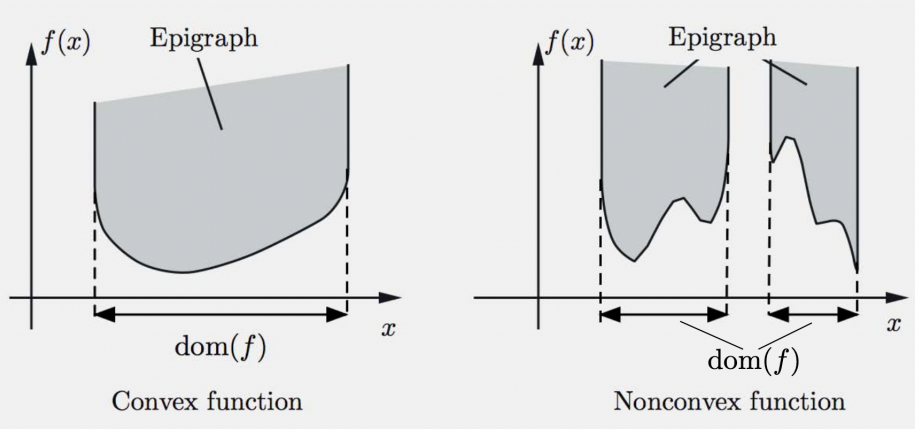

Figure 5: Epigraph of a function

a function as a union of points above it graph, i.e., the epigraph, assuming that f is convex, we should also be able to represent it as intersection of halfspaces that contain it! In fact, since f is convex, then its epigraph is also convex, hence there must exist a collection of halfspaces $\mathcal{H}$ whose intersection is equal to the epigraph of f. We now introduce the Fenchel conjugate which is very closely related to the dual representation of the epigraph of f. The **Fenchel conjugate** of a functional $f : E \to \mathbb{R}$ is function on the dual space defined as $f^* : E^* \to \mathbb{R}$ and is given by:

$$f^*(\boldsymbol{y}) = \sup_{\boldsymbol{x}\in\mathbb{R}^n} \{\langle \boldsymbol{y}, \boldsymbol{x}\rangle - f(\boldsymbol{x})\} = -\inf_{\boldsymbol{x}\in\mathbb{R}^n} \{f(\boldsymbol{x}) - \langle \boldsymbol{y}, \boldsymbol{x}\rangle\} \tag{9}$$

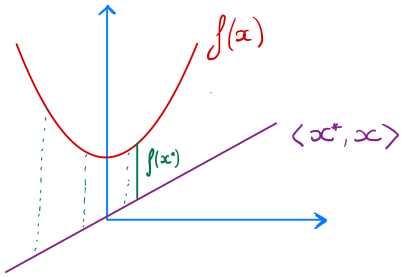

Figure 6: The Fenchel conjugate of a function f gives the largest difference value between the function and a line through the origin.

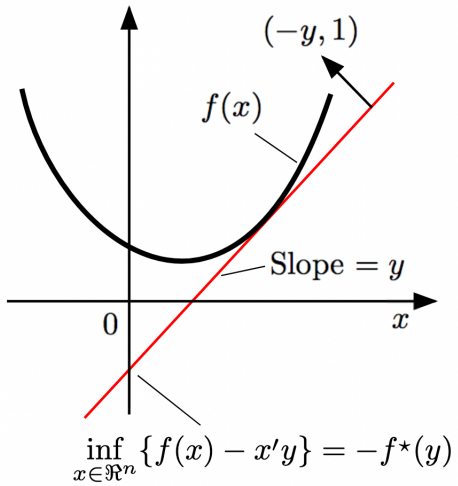

Figure 7: The Fenchel conjugate, $f^*(\boldsymbol{x}^*)$ at $\boldsymbol{x}^*$, encodes the supporting hyperplane of the function f, with $f^*(\boldsymbol{x}^*)$ being the y-offset of the hyperplane and $\boldsymbol{x}^*$ being the slope.

Example: The convex conjugate of an affine function $f(x) = \langle a, x\rangle - b$ is

$$f^*(x^*) = \begin{cases} b, & x^* = a \\ +\infty, & x^* \neq a. \end{cases} \tag{10}$$

The convex conjugate of the absolute value function $f(x) = |x|$ is

$$f^*(x^*) = \begin{cases} 0, & |x^*| \leq 1 \\ \infty, & |x^*| > 1. \end{cases} \tag{11}$$

The most important fact about the Fenchel conjugate f^* is that it encodes the supporting hyperplanes of the epigraph of the function f, $\mathrm{epi}f$. That is, for any $\boldsymbol{x}^* \in E^* = \mathbb{R}^n$, denote the maximizer of $f^*(\boldsymbol{x}^*)$ by $\boldsymbol{x}$, i.e., $\boldsymbol{x} = \arg\sup_{\boldsymbol{x}\in\mathbb{R}^n} \{\langle \boldsymbol{x}^*, \boldsymbol{x}\rangle - f(\boldsymbol{x})\}$, the following relationship holds:

$$\mathrm{epi}f = \bigcap_{\{x^* | f^*(x*)<\infty\}} \{(x, y) \mid \langle \boldsymbol{x}, \boldsymbol{x}^*\rangle - f^*(\boldsymbol{x}^*) \leq \boldsymbol{y}\} \tag{12}$$

Finally, one can relate the arguments of a function and it Fenchel conjugate via subgradients, as shown by the following theorem.

Theorem 3.1 (Conjugate Subgradient Theorem)**.** *Suppose f and f^* are convex conjugates; then we have*

$$f(\boldsymbol{x}) = \sup_{\boldsymbol{x}^*} \langle \boldsymbol{x}, \boldsymbol{x}^*\rangle - f^*(\boldsymbol{x}^*) \tag{13}$$

$$f^*(\boldsymbol{x}^*) = \sup_{\boldsymbol{x}} \langle \boldsymbol{x}, \boldsymbol{x}^*\rangle - f(\boldsymbol{x}) \tag{14}$$

and furthermore, for the $\boldsymbol{x}$ and $\boldsymbol{x}^*$ solving them respectively,

$$\boldsymbol{x}^* \in \partial_{\boldsymbol{x}} f(\boldsymbol{x}) \tag{15}$$

$$\boldsymbol{x} \in \partial_{\boldsymbol{x}} f^*(\boldsymbol{x}^*) \tag{16}$$

where $\partial_{\boldsymbol{x}} f(\boldsymbol{x})$ is the subdifferential of f.

4 Lagrangian Duality Theory

4.1 Definitions

Consider the optimization problem P, given by the ordered tuple $P = (m, p, \{f_i\}_{i=0}^{p})$, where $m, p \in \mathbb{N}, p \geq m$ and $\forall i \in [p], \quad f_i : \mathbb{R}^n \to \mathbb{R}$, called the **primal** problem:

$$\min_{\boldsymbol{x}} \quad f_0(\boldsymbol{x}) \tag{17}$$

$$\text{Constrained by} \quad f_i(\boldsymbol{x}) \leq 0 \qquad \forall i \in \{1, \ldots, m\} \tag{18}$$

$$\text{And} \quad f_i(\boldsymbol{x}) = 0 \qquad \forall i \in \{m+1, \ldots, p\} \tag{19}$$

when m and p are clear from context, we simply denote $P = (\{f_i\}_{i=0}^{p})$.

A vector $\boldsymbol{x} \in \mathbb{R}^n$ is said to be **feasible** if it satisfies eqs. (18) to (19). We assume that there exists a feasible $\boldsymbol{x}$, otherwise the problem is not interesting since it does not have solution![1]

4.2 The Lagrangian

We define the Lagrangian function, $L : \mathbb{R}^n \times \mathbb{R}_+^m \times \mathbb{R}^{p-m}$, corresponding to the above optimization problem P as follows:

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f_0(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}) + \sum_{i=m+1}^{p} \mu_i f_i(\boldsymbol{x}) \tag{20}$$

where $\boldsymbol{\lambda} \in \mathbb{R}_+^m$ and $\boldsymbol{\mu} \in \mathbb{R}^{p-m}$ are called **slack variables**. These variables are called slack variables because by setting them wisely we obtain a function whose minimum corresponds exactly to that of the problem P. We now show how we should set these slack variables such as to obtain a function whose minima corresponds to the minima of the optimzation problem P. Observe that for every feasible $\boldsymbol{x} \in \mathbb{R}^n$, and for all $\boldsymbol{\lambda} \in \mathbb{R}_+^m$, $\boldsymbol{\mu} \in \mathbb{R}^{p-m}$, $f_0(\boldsymbol{x})$ is bounded below by the Lagrangian, that is:

$$\forall \lambda \in \mathbb{R}_+^m, \mu \in \mathbb{R}^{p-m} \qquad f_0(\boldsymbol{x}) \geq L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \tag{21}$$

Taking the supremum over all $\boldsymbol{\lambda} \in \mathbb{R}_+^m$, $\boldsymbol{\mu} \in \mathbb{R}^{p-m}$, we get:

$$f_0(\boldsymbol{x}) \geq \sup_{\boldsymbol{\lambda} \geq \boldsymbol{0}, \boldsymbol{\mu}} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \tag{22}$$

Before we move further, let's look further at the right hand side quantity in the above:

$$\sup_{\boldsymbol{\lambda} \geq \boldsymbol{0}, \boldsymbol{\mu}} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \sup_{\boldsymbol{\lambda} \geq \boldsymbol{0}, \boldsymbol{\mu}} \left[f_0(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}) + \sum_{i=m+1}^{p} \mu_i f_i(\boldsymbol{x}) \right] \tag{23}$$

$$= f_0(\boldsymbol{x}) + \sup_{\boldsymbol{\lambda} \geq \boldsymbol{0}} \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}) + \sup_{\boldsymbol{\mu}} \sum_{i=m+1}^{p} \mu_i f_i(\boldsymbol{x}) \tag{24}$$

$$= f_0(\boldsymbol{x}) + \sum_{i=1}^{m} \sup_{\lambda_i \geq 0} \lambda_i f_i(\boldsymbol{x}) + \sum_{i=m+1}^{p} \sup_{\mu_i} \mu_i f_i(\boldsymbol{x}) \tag{25}$$

[1] This also explain why we use a minimum instead of a supremum.

Observe the following:

$$\sup_{\lambda_i \geq 0} \lambda_i f_i(\boldsymbol{x}) = \begin{cases} 0 & \text{if } f_i(\boldsymbol{x}) \leq 0 \\ \infty & \text{Otherwise} \end{cases} \tag{26}$$

$$\sup_{\mu_i} \mu_i f_i(\boldsymbol{x}) = \begin{cases} 0 & \text{if } f_i(\boldsymbol{x}) = 0 \\ \infty & \text{Otherwise} \end{cases} \tag{27}$$

By considering the extended real-line, $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$, we can replace the supremum by a maximum, since the supremum exists and can re-express the supremum over the Lagrangian as a maximum as follows:

$$\max_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\mu}} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \begin{cases} f_0(\boldsymbol{x}) & \text{if } \forall i \in [0, m], \;\; f_i(\boldsymbol{x}) \leq 0 \text{ and } \forall i \in [m, p-m], \;\; f_i(\boldsymbol{x}) = 0 \\ \infty & \text{Otherwise} \end{cases} \tag{28}$$

$$= \begin{cases} f_0(\boldsymbol{x}) & \text{if } \boldsymbol{x} \text{ is feasible} \\ \infty & \text{Otherwise} \end{cases} \tag{29}$$

That is, by taking the maximum over the slack variables $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ we essentially obtain a function where all feasible values $\boldsymbol{x}$ of the program P corresponds to the values of $f_0(\boldsymbol{x})$, and to ∞ for all infeasible values. As a result, we have:

$$\min_{\substack{\forall i \in \{1,\ldots,m\} f_i(\boldsymbol{x}) \leq 0 \\ \forall i \in \{m+1,\ldots,p\} f_i(\boldsymbol{x}) = 0}} f_0(\boldsymbol{x}) = \min_{\boldsymbol{x}} \max_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\mu}} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \tag{30}$$

4.3 The Lagrangian Dual

[DENI: *Add some more intuition on the dual.*]The above relation suggests that if we could switch the order of the min and max on the right hand-side, we could obtain a dual maximization problem that is somehow related to our original minimization problem P. This is the intuition behind deriving another program D called **the dual** from the Lagrangian. The dual variables, i.e., the Lagrangian Slack variables, often have meaning and can be used to solve for an additional problem related to the original optimzation problem P. **The Lagrange dual function** of a program P with Lagrangian function L is defined as:

$$g(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \min_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \tag{31}$$

Note that more generally, the minimum is replaced with an infimum in the case if a minimum of the Lagrangian does not exist but since here we consider the extended set of reals, $\bar{\mathbb{R}}$, instead of the reals, $\mathbb{R}$, we can use a min. The **dual program**, D of the primal program is given as:

$$\max_{\boldsymbol{\lambda}, \boldsymbol{\mu}} \qquad g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \tag{32}$$

$$\text{Constrained by} \qquad \boldsymbol{\lambda} \geq \mathbf{0} \tag{33}$$

Note that the dual function g is concave, even when the initial problem is not convex, because it is a point-wise minimum of affine functions. Note that under no assumptions on the Lagrangian we have that weak programming duality holds which relates the optimal values of the primal and dual programs:

Theorem 4.1. *Weak Programming Duality*
Let $\boldsymbol{x}^*$ be any feasible solution to the primal program P, and $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ be any feasible solution to the dual program of D. Then $f_0(\boldsymbol{x}^*) \geq g(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$

Proof.

$$\max_{\boldsymbol{\lambda}\geq 0,\boldsymbol{\mu}} L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) \geq L(\boldsymbol{x},\boldsymbol{\lambda}',\boldsymbol{\mu}') \qquad \forall \boldsymbol{x},\boldsymbol{\lambda}' \geq \mathbf{0},\boldsymbol{\mu}' \tag{34}$$

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{\lambda}\geq 0,\boldsymbol{\mu}} L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) \geq \min_{\boldsymbol{x}} L(\boldsymbol{x},\boldsymbol{\lambda}',\boldsymbol{\mu}') \qquad \forall \boldsymbol{\lambda}' \geq \mathbf{0},\boldsymbol{\mu}' \tag{35}$$

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{\lambda}\geq 0,\boldsymbol{\mu}} L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) \geq \max_{\boldsymbol{\lambda}\geq 0,\boldsymbol{\mu}} \min_{\boldsymbol{x}} L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) \tag{36}$$

$$f_0(\boldsymbol{x}^*) \geq g(\boldsymbol{\lambda}^*,\boldsymbol{\mu}^*) \tag{37}$$

□

In order for the solutions of the primal and dual problems to be related, one would expect that the optimal value of the primal and dual problems should be equal to each other, i.e., the min-max of the Lagrangian should be equal to its max-min. Surprisingly, the conditions for this to be true can formalized in a game theoretic manner! In fact, the Lagrangian's min-max optimization can be seen as a game in between two opponents trying to minimize (resp. maximize) a payoff function. This intuition can be formalized with a generalized version of the min-max theorem which applies beyond Nash equilibria for 2 person zero-sum games but to arbitrary convex-concave payoff functions:

Theorem 4.2. *Minimax Theorem Let $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ be compact convex sets. If $f : X \times Y \to \mathbb{R}$ is a continuous function that is concave-convex, i.e. $f(\cdot, y) : X \to \mathbb{R}$ is concave for fixed y, and $f(x,\cdot) : Y \to \mathbb{R}$ is convex for fixed x Then we have that*

$$\max_{x\in X} \min_{y\in Y} f(x,y) = \min_{y\in Y} \max_{x\in X} f(x,y) \tag{38}$$

Going back to our Lagrangian, one can construct sets

$$X = \left\{ (\boldsymbol{\lambda},\boldsymbol{\mu}) \mid \min_{\boldsymbol{x}} L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) > -\infty \right\} \tag{39}$$

$$Y = \left\{ \boldsymbol{x} \mid \max_{\boldsymbol{\lambda}\geq 0,\boldsymbol{\mu}} L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) < \infty \right\} \tag{40}$$

Hence, if X and Y are convex and compact, by the min-max theorem, we have:

$$\min_{\boldsymbol{x}\in Y} \max_{(\boldsymbol{\lambda},\boldsymbol{\mu})\in X} L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) = \max_{(\boldsymbol{\lambda},\boldsymbol{\mu})\in X} \min_{\boldsymbol{x}\in Y} L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) \tag{41}$$

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{\lambda}\geq 0,\boldsymbol{\mu}} L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) = \max_{\boldsymbol{\lambda}\geq 0,\boldsymbol{\mu}} \min_{\boldsymbol{x}} L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) \tag{42}$$

where the last line holds since the optimization problems' optimizers do not change when the domain changes. This means that the Lagrangian can be seen as the payoff function of a two-player zero-sum game, where both players are trying to maximize their worst-case payoff or minimize their opponents' best-case payoff. When min-max value of the Lagrangian is equal to the max-min value, we say that strong duality holds for the program P. We will provide conditions on the program P, in sections 5 to 6.

5 Linear Programming

Linear Programming is a mathematical method that allows to find the variables that maximize or minimize a linear function that is constrained by a set of linear constraints. That is, linear programming refers to the set of methods that allow us to solve problems defined by the inputs $(\boldsymbol{A}, \boldsymbol{b}.\boldsymbol{c})$ and that can be expressed in the following canonical form :

Primal:

$$\min_{\boldsymbol{x}} \quad \boldsymbol{b}^T\boldsymbol{x} \tag{43}$$

$$\text{Constrained by} \quad \boldsymbol{A}^T\boldsymbol{x} \geq \boldsymbol{c} \tag{44}$$

$$\text{And} \quad \boldsymbol{x} \geq \boldsymbol{0} \tag{45}$$

We will now derive the **dual** program of the problem above using the machinery introduced in the previous section. Since the above expression is the canonical form for any linear program, the dual that we derive will also give us a formula to easily find the dual of any linear program. Let's first calculate the Lagrangian of the above program:

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{b}^T\boldsymbol{x} + \boldsymbol{\lambda}^T(\boldsymbol{c} - \boldsymbol{A}^T\boldsymbol{x}) - \boldsymbol{\mu}^T\boldsymbol{x} \tag{46}$$

$$= \boldsymbol{x}^T\boldsymbol{b} + \boldsymbol{c}^T\boldsymbol{\lambda} - \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{\lambda} - \boldsymbol{x}^T\boldsymbol{\mu} \tag{47}$$

$$= \boldsymbol{c}^T\boldsymbol{\lambda} + \boldsymbol{x}^T(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{\lambda}) - \boldsymbol{x}^T\boldsymbol{\mu} \tag{48}$$

We can then calculate the dual objective function g:

$$g(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{\boldsymbol{x}}\{\boldsymbol{c}^T\boldsymbol{\lambda} - \boldsymbol{x}^T(\boldsymbol{A}\boldsymbol{\lambda} - \boldsymbol{b}) - \boldsymbol{\mu}^T\boldsymbol{x}\} \tag{49}$$

Notice that this function is exactly Lagrangian form of the following maximization problem:

$$\max_{\boldsymbol{\lambda}} \quad \boldsymbol{c}^T\boldsymbol{\lambda} \tag{50}$$

$$\text{Constrained by} \quad \boldsymbol{A}\boldsymbol{\lambda} \leq \boldsymbol{b} \tag{51}$$

$$\text{and} \quad \boldsymbol{\mu} \geq \boldsymbol{0} \tag{52}$$

Adding to this program the dual program feasibility constraint, we obtain **the dual of the standard form** of Linear programming where we renamed the slack variable $\boldsymbol{\lambda}$ as $\boldsymbol{y}$. The dual variables $\boldsymbol{y}$ are called **the dual variables**.

Dual:

$$\max_{\boldsymbol{y}} \quad \boldsymbol{c}^T\boldsymbol{y} \tag{53}$$

$$\text{Constrained by} \quad \boldsymbol{A}\boldsymbol{y} \leq \boldsymbol{b} \tag{54}$$

$$\text{And} \quad \boldsymbol{y} \geq \boldsymbol{0} \tag{55}$$

Note that different authors might refer to the minimization problem as the dual and the maximization problem as the primal. The expressions $\boldsymbol{c}^T\boldsymbol{x}$ and $\boldsymbol{b}^T\boldsymbol{y}$ are respectively called the **objectives** of the primal and dual problems. Solutions $\boldsymbol{x}^*$ for the primal that satisfy the primal constraints $\boldsymbol{A}\boldsymbol{x}^* \leq \boldsymbol{b}, \quad \boldsymbol{x}^* \geq \boldsymbol{0}$ are called **feasible solutions**. Similarly, solutions $\boldsymbol{y}^*$ for the dual that satisfy the dual constraints $\boldsymbol{A}^T\boldsymbol{y}^* \leq \boldsymbol{c}, \quad \boldsymbol{y}^* \geq \boldsymbol{0}$ are called **feasible solutions**. $\boldsymbol{y}^*$. A linear program is said to be a **feasible program** iff there exists variables for the program that are feasible, otherwise the program is said to be an **infeasible program**. A feasible variable $\boldsymbol{x}^*$ for the primal is called optimal iff $\forall \boldsymbol{x} \in \{\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \quad \boldsymbol{x} \geq \boldsymbol{0}\}, \quad \boldsymbol{c}^T\boldsymbol{x}^* \geq \boldsymbol{c}^T\boldsymbol{x}$. A linear program in the primal form is said to be bounded iff $\forall \boldsymbol{x} \in \{\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \quad \boldsymbol{x} \geq \boldsymbol{0}\}, \quad \boldsymbol{c}^T\boldsymbol{x} < \infty$. A feasible variable $\boldsymbol{y}^*$ for the dual is called optimal iff $\forall \boldsymbol{y} \in \{\boldsymbol{y} \mid \boldsymbol{A}^T\boldsymbol{y} \leq \boldsymbol{c}, \quad \boldsymbol{y} \geq \boldsymbol{0}\}, \quad \boldsymbol{b}^T\boldsymbol{x}^* \geq \boldsymbol{b}^T\boldsymbol{x}$. A linear program in the primal form is said to be bounded iff $\forall \boldsymbol{y} \in \{\boldsymbol{y} \mid \boldsymbol{A}^T\boldsymbol{y} \leq \boldsymbol{c}, \quad \boldsymbol{y} \geq \boldsymbol{0}\}, \quad \boldsymbol{b}^T\boldsymbol{y} < \infty$

There are many polynomial time algorithms to solve linear programs such as the simplex or big M algorithms. These algorithms generally take as input $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c})$ and return a tuple $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ which are respectively the optimal variables for the primal and dual problems.

Below are the two most important results from Linear Programming Duality that confirm our initial motivation of the dual using the min-max theorem:

Theorem 5.1. *Strong Programming Duality for Linear Programming*
Let $\boldsymbol{x}^*$ be any feasible solution to the primal of a linear program P, and $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ be any feasible solution to the dual program of P. Let $f_0(\boldsymbol{x}^*)$ be the objective of the primal and let $g(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ be the objective of the dual, then $f_0(\boldsymbol{x}^*) = g(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$.

6 Convex Programming

Convex Programming is a mathematical method that allows to find the variables that maximize or minimize a function that is constrained by a set of convex inequality constraints and affine equality constraints. That is, convex programming refers to the set of methods that allow us to solve problems defined by the inputs $(f_0, f_1, \ldots, f_m)$ and that can be expressed in the following canonical primal form:

Primal:

$$\min_{\boldsymbol{x}} \quad f_0(\boldsymbol{x}) \tag{56}$$

$$\text{Constrained by} \quad f_i(\boldsymbol{x}) \leq 0 \qquad \forall i \in \{1, \ldots, m\} \tag{57}$$

$$\text{And} \quad h_i(\boldsymbol{x}) = 0 \qquad \forall i \in \{m+1, \ldots, p\} \tag{58}$$

Note that different authors might refer to the minimization problem as the dual and the maximization problem as the primal.

It is harder to derive the dual in closed form like we did for linear programming, however this can be done by going through the lagrangian or using shortcuts with the help of Frenchel conjugates. More information about finding the dual of a convex primal program can be found in section 3 of [1].

Definition 6.1. ***Slater's condition*** *We say that the problem satisfies Slater's condition if it is strictly feasible, that is:*

$$\exists x_0 \in \mathcal{D} : f_i(x_0) < 0, \quad i = 1, \ldots, m, \quad h_i(x_0) = 0, \quad i = 1, \ldots, p$$

We can replace the above by a weak form of Slater's condition, where strict feasibility is not required whenever the function f_i is affine.

For our purposes, we present the following duality theorem to confirm our intuition from the minimax theorem that a dual exists with the same objective value.

Theorem 6.2. ***Strong duality via Slater condition***
If the primal problem (8.1) is convex, and satisfies the weak Slater's condition, then strong duality holds. That is, let $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ be the optimal variables for the primal and dual respectively, then $f_0(\boldsymbol{x}^*) = g(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$.

7 Properties, Characteristics and Comparative Statics of Optimization Problems

In this section, we first discuss the properties of optimization problem in an abstract setting and then provide optimality conditions for convex optimization problems.

7.1 Continuity, Existence, and Uniqueness

We now consider a more generally stated optimization problem in terms of a constraint set defined by some correspondence $C : \Theta \rightrightarrows X$. Consider the class of parametric constrained optimization problems defined over the convex set of parameters Θ.

$$\max_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{\theta})} f(\boldsymbol{x}, \boldsymbol{\theta}) \tag{59}$$

where $\mathcal{X} : \Theta \rightrightarrows X$ is a non-empty and compact-valued correspondence.

The **value function** of this optimization problem is defined as $V(\boldsymbol{\theta}) = \max_{\boldsymbol{x} \in C(\boldsymbol{\theta})} f(\boldsymbol{x}, \boldsymbol{\theta})$, while it's **solution correspondence** is defined as $\mathcal{X}^*(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{x} \in C(\boldsymbol{\theta})} f(\boldsymbol{x}, \boldsymbol{\theta})$. We denote any of the solutions outputted by the solution correspondence $\mathcal{X}^*(\boldsymbol{\theta})$ at some $\boldsymbol{\theta} \in \Theta$ by $\boldsymbol{x}^*(\boldsymbol{\theta})$, i.e., $\boldsymbol{x}^*(\boldsymbol{\theta}) \in \mathcal{X}^*(\boldsymbol{\theta})$. Additionally, if f is strictly concave, then the solution is unique for all $\boldsymbol{\theta} \in \Theta$, in which case we call the solution correspondence, the **solution function** and simply denote $\boldsymbol{x}^*(\boldsymbol{\theta})$, since $\mathcal{X}^*(\boldsymbol{\theta}) = \{\boldsymbol{x}^*(\boldsymbol{\theta})\}$.

The maximum theorem provides us with a characterization of the continuity and uniqueness properties of the value and solution mappings.

Theorem 7.1 (Maximum Theorem)**.** *Consider the class of parametric constrained optimization problems*

$$\max_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{\theta})} f(\boldsymbol{x}, \boldsymbol{\theta}) \tag{60}$$

defined over the set of parameters Θ. Suppose that (1) $\mathcal{X} : \Theta \rightrightarrows X$ is continuous (i.e. lsc and usc) and compact-valued, and (2) $f : X \times \Theta \to \mathbb{R}$ is a continuous function. Let $\mathcal{X}^*(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{\theta})} f(\boldsymbol{x}, \boldsymbol{\theta})$ and, $V(\boldsymbol{\theta}) = \max_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{\theta})} f(\boldsymbol{x}, \boldsymbol{\theta})$, then:

1. *$\mathcal{X}^*(\boldsymbol{\theta})$ is non-empty for every $\boldsymbol{\theta} \in \Theta$*
2. *$\mathcal{X}^*$ is upper semi-continuous (and thus continuous if $\mathcal{X}^*$ is singleton-valued)*
3. *V is continuous*

Often, we are interested in knowing if the value function is convex or if the solutions are unique. The following result allows us to get a better understanding of the properties of optimization problems.

Theorem 7.2. *Consider the class of parametric constrained optimization problems defined over the convex set of parameters Θ.*

$$\max_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{\theta})} f(\boldsymbol{x}, \boldsymbol{\theta}) \tag{61}$$

Suppose that (1) $\mathcal{X} : \Theta \rightrightarrows X$ is continuous and compact-valued, and (2) $f : X \times \Theta \to \mathbb{R}$ is a continuous function. Let $\boldsymbol{x}^*(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{\theta})} f(\boldsymbol{x}, \boldsymbol{\theta})$ and, $V(\boldsymbol{\theta}) = \max_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{\theta})} f(\boldsymbol{x}, \boldsymbol{\theta})$, then:

1. *If $f(\cdot, \boldsymbol{\theta})$ is a quasi-concave function in $\boldsymbol{x}$ for each $\boldsymbol{\theta}$, and $\mathcal{X}$ is convex valued, then $\boldsymbol{x}^*$ is convex-valued.*

2. *If $f(\cdot, \boldsymbol{\theta})$ is a strictly quasi-concave function in $\boldsymbol{x}$ for each $\boldsymbol{\theta}$, and D is convex-valued, then $\boldsymbol{x}^*(\boldsymbol{\theta})$ is single-valued.*

3. *If f is a concave function in $(\boldsymbol{x}, \boldsymbol{\theta})$ and $\mathcal{X}$ is convex-valued, then V is a concave function and $\boldsymbol{x}^*$ is convex-valued.*

4. *If f is a strictly concave function in $(\boldsymbol{x}, \boldsymbol{\theta})$ and $\mathcal{X}$ is convex-valued, then V is strictly concave and $\boldsymbol{x}^*$ is a function.*

7.2 Optimality Conditions and Characterization of Solutions

Often, it might not be clear what properties the optimal solutions to a convex program might satisfy. As a result, to prove that a solution to a particular convex program satisfies certain properties, one uses the optimality conditions for the convex program given via the Lagrangian. We now introduce one of the most important results in convex analysis, the Karush–Kuhn–Tucker theorem. The Karush–Kuhn–Tucker theorem provides optimality conditions for a convex program via its associated Lagrangian.

Theorem 7.3 (Karush–Kuhn–Tucker theorem)**.** *Let $L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ be the Lagrangian function corresponding to the optimization problem $P = (m, p, \{f_i\}_{i=0}^p)$ given by:*

$$\min_{\boldsymbol{x}} \quad f_0(\boldsymbol{x}) \tag{62}$$

$$\textit{Constrained by} \quad f_i(\boldsymbol{x}) \leq 0 \qquad \forall i \in \{1, \ldots, m\} \tag{63}$$

$$\textit{And} \quad f_i(\boldsymbol{x}) = 0 \qquad \forall i \in \{m+1, \ldots, p\} \tag{64}$$

Suppose that for all $i = 1, \ldots, p$, $f_i(\boldsymbol{x})$'s are all convex and there exists a $\boldsymbol{x}$ such that $\forall i = 1, \ldots, m,\ \ f_i(\boldsymbol{x}) < 0$, i.e., there exists an interior point, then the optimal $\boldsymbol{x}^\star$ has an associated $(\boldsymbol{\mu}^\star, \boldsymbol{\lambda}^\star)$ such that $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$ is a saddle point of $L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ and satisfies the following conditions:

1. $0 \in \partial_{\boldsymbol{x}} f_0(\boldsymbol{x}) + \sum_{i=1}^m \lambda_i^* \partial_{\boldsymbol{x}} f_i(\boldsymbol{x}) + \sum_{i=m+1}^p \mu_i^* \partial_{\boldsymbol{x}} f_i(\boldsymbol{x})$ *(Stationarity)*

2. $\forall i \in \{1, \ldots, m\}, \quad \lambda_i^* f_i(\boldsymbol{x}^*) = 0$ *(Complementary Slackness)*

3. $\forall i \in \{1, \ldots, m\}, \quad f_i(\boldsymbol{x}^*) \leq 0$ *and* $\forall i \in \{m+1, \ldots, p\} \quad f_i(\boldsymbol{x}^*) = 0$ *(Primal Feasibility)*

4. $\forall i \in \{1, \ldots, m\}, \quad \lambda_i^* \geq 0$ *(Dual Feasibility)*

Proof. Since we have a strictly feasible point, it must be that Slater's condition holds and hence strong duality, which gives:

$$f_0(\boldsymbol{x}^*) = g(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \tag{65}$$

$$f_0(\boldsymbol{x}^*) = \min_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \tag{66}$$

$$f_0(\boldsymbol{x}^*) = \min_{\boldsymbol{x}} \left\{ f_0(\boldsymbol{x}) + \sum_{i=1}^m \lambda_i^* f_i(\boldsymbol{x}) + \sum_{i=m+1}^p \mu_i^* f_i(\boldsymbol{x}) \right\} \tag{67}$$

$$f_0(\boldsymbol{x}^*) \leq f_0(\boldsymbol{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\boldsymbol{x}^*) + \sum_{i=m+1}^p \mu_i^* f_i(\boldsymbol{x}^*) \tag{68}$$

$$f_0(\boldsymbol{x}^*) \leq f_0(\boldsymbol{x}^*) \tag{69}$$

This means that all inequalities are actually equalities which implies that $f_0(\boldsymbol{x}^*) = g(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = L(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$. As a result, we have $\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$. Hence, the stationarity conditions for $\boldsymbol{x}^*$ are given by:

$$0 \in \partial_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \tag{70}$$

$$0 \in \partial_{\boldsymbol{x}} f_0(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i \partial_{\boldsymbol{x}} f_i(\boldsymbol{x}) + \sum_{i=m+1}^{p} \mu_i \partial_{\boldsymbol{x}} f_i(\boldsymbol{x}) \tag{71}$$

The rest of the conditions can be obtained from definitions.

□

7.3 Comparative Statics

More than often, we are interested in understanding how the value function or the solution function changes as the parameters $\boldsymbol{\theta}$ of the problem changes. Two important tools allow us to obtain answers to these questions, 1) the inverse function theorem, and 2) the envelope theorem.

7.3.1 Inverse Function Theorem

TODO The inverse function theorem (IFT) allows us to compute the derivative of the solution function.

7.3.2 Envelope Theorems

While the IFT allows us to compute the derivative of the solution function, it does not provide any information on the derivative of the value function. Danskin's theorem [2] offers insights into optimization problems when the constraint set is fixed, i.e., the optimization problem is of the form $\max_{\boldsymbol{x} \in X} f(\boldsymbol{x}, \boldsymbol{\theta})$, where $X \subset \mathbb{R}^m$ is compact and non-empty. Among other things, Danskin's theorem allows us to compute the gradient of the objective function of this optimization problem with respect to $\boldsymbol{\theta}$.

Theorem 7.4 (Danskin's Theorem)**.** *Consider an optimization problem of the form:* $\max_{\boldsymbol{x} \in X} f(\boldsymbol{x}, \boldsymbol{\theta})$*, where* $X \subset \mathbb{R}^n$ *is compact and non-empty. Suppose that* X *is convex and that* f *is concave in* $\boldsymbol{x}$*. Let* $V(\boldsymbol{\theta}) = \max_{\boldsymbol{x} \in X} f(\boldsymbol{x}, \boldsymbol{\theta})$ *and* $\mathcal{X}^*(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{x} \in X} f(\boldsymbol{\theta}, \boldsymbol{x})$*. Then* V *is differentiable at* $\widehat{\boldsymbol{\theta}}$*, if the solution correspondence* $\mathcal{X}^*(\widehat{\boldsymbol{\theta}})$ *is a singleton: i.e.,* $\mathcal{X}^*(\widehat{\boldsymbol{\theta}}) = \{\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}})\}$*. Additionally, the gradient at* $\widehat{\boldsymbol{\theta}}$ *is given by* $V'(\widehat{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{\theta}} f(\widehat{\boldsymbol{\theta}}, \boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}))$*.*

Example 7.5 (Shepherd's Lemma)**.** *Consider the expenditure minimization problem:*

$$\min_{\boldsymbol{x}_i \in \mathbb{R}^n : u_i(\boldsymbol{x}_i) \geq \nu_i} \boldsymbol{p} \cdot \boldsymbol{x}_i \tag{72}$$

The value function associated with the expenditure minimization problem is called the expenditure function and is denoted as follows:

$$e_i(\boldsymbol{p}, \nu_i) = \min_{\boldsymbol{x}_i \in \mathbb{R}^n : u_i(\boldsymbol{x}_i) \geq \nu_i} \boldsymbol{p} \cdot \boldsymbol{x}_i \tag{73}$$

The solution function associated with the expenditure minimization problem is called the Hicksian demand and is denoted as follows:

$$h_i(\boldsymbol{p}, \nu_i) = \arg\min_{\boldsymbol{x}_i \in \mathbb{R}^n : u_i(\boldsymbol{x}_i) \geq \nu_i} \boldsymbol{p} \cdot \boldsymbol{x}_i \tag{74}$$

By the envelope theorem, the derivative of the expenditure function in $\boldsymbol{p}$ *is given by:*

$$\nabla_{\boldsymbol{p}} e_i(\boldsymbol{p}, \nu_i) = \boldsymbol{x}_i^\star(\boldsymbol{p}) + \lambda(\boldsymbol{x}_i(\boldsymbol{p}), \boldsymbol{p})(0) \tag{75}$$

$$= h_i(\boldsymbol{p}, \nu_i) \tag{76}$$

Unfortunately, Danskin's theorem does not hold when the constraint set X is replaced by a correspondence,i.e., when the inner problem is $\max_{\boldsymbol{x}\in\mathcal{X}(\boldsymbol{\theta})} f(\boldsymbol{\theta},\boldsymbol{x})$.

Example 7.6 (Danskin's theorem does not apply to min-max games with dependent strategy sets)**.** *Consider the optimization problem:*

$$\max_{x\in\mathbb{R}:x+\theta\geq 0} -x^2+x+2\theta+2 \ . \tag{77}$$

The solution to this problem is unique, given any $\theta\in\Theta$, meaning the solution correspondence $X^*(\theta)$ is singleton-valued. We denote this unique solution by $x^*(\theta)$. After solving, we find that

$$x^*(\theta)=\begin{cases} 1/2 & if\,\theta\geq -1/2 \\ -\theta & if\,\theta<-1/2 \end{cases} \tag{78}$$

The value function $V(\theta)=\max_{x\in\mathbb{R}:x+\theta\geq 0} -x^2+x+2\theta+2$ is then given by:

$$\begin{aligned} V(\theta) &= f(\theta,x^*(\theta)) && (79)\\ &= -x^*(\theta)^2+x^*(\theta)+2\theta+2 && (80)\\ &= \begin{cases} -1/4+1/2+2\theta+2 & if\,\theta\geq -1/2 \\ -\theta^2-\theta+2\theta+2 & if\,\theta<-1/2 \end{cases} && (81)\\ &= \begin{cases} 9/4+2\theta & if\,\theta\geq -1/2 \\ -\theta^2+\theta+2 & if\,\theta<-1/2 \end{cases} && (82) \end{aligned}$$

The derivative of this value function is:

$$\frac{dV}{d\theta}=\begin{cases} 2 & if\,\theta\geq -1/2 \\ 1-2\theta & if\,\theta<-1/2 \end{cases} \tag{83}$$

However, the derivative predicted by Danskin's theorem is 2.

N.B. *For simplicity, we do not assume the constraint set is compact in this example. Compactness of the constraint set can be used to guarantee existence of a solution, but as a solution to this particular problem always exists, we can do away with this assumption.*

The following theorem called **the envelope theorem**, due to Milgrom and Segal [3], generalizes Danskin's theorem to handle parameterized constraints:

Theorem 7.7 (Envelope Theorem [3])**.** *Consider the maximization problem*

$$V(\boldsymbol{\theta})=\max_{\boldsymbol{x}\in X} f(\boldsymbol{x},\boldsymbol{\theta}),\ \textit{subject to}\ g_k(\boldsymbol{x},\boldsymbol{\theta})\geq 0,\ \textit{for all}\ k=1,\dots,K \ , \tag{84}$$

where $X\subseteq\mathbb{R}^m$.

Define the solution correspondence $X^*(\boldsymbol{\theta})=\arg\max_{\boldsymbol{x}\in\Theta:\boldsymbol{g}(\boldsymbol{x},\boldsymbol{\theta})\geq\mathbf{0}} f(\boldsymbol{x},\boldsymbol{\theta})$. Suppose that 1. $f,g_1,\dots,g_K$ are continuous and concave in $\boldsymbol{y}$; 2. $\nabla_{\boldsymbol{\theta}}f,\nabla_{\boldsymbol{\theta}}g_1,\dots,\nabla_{\boldsymbol{\theta}}g_K$ are continuous in $(\boldsymbol{x},\boldsymbol{\theta})$; and 3. $\forall\boldsymbol{\theta}\in\Theta,\exists\boldsymbol{x}\in X$ s.t. $g_k(\boldsymbol{x},\boldsymbol{\theta})>0$, for all $k=1,\dots,K$., then the value function V is absolutely continuous, and at any point $\widehat{\boldsymbol{\theta}}\in\Theta$ where V is differentiable:

$$\nabla_{\boldsymbol{\theta}}V(\widehat{\boldsymbol{\theta}})=\nabla_{\boldsymbol{\theta}}L(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}),\boldsymbol{\lambda}^*(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}),\widehat{\boldsymbol{\theta}}),\widehat{\boldsymbol{\theta}})=\nabla_{\boldsymbol{\theta}}f\left(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}),\widehat{\boldsymbol{\theta}}\right)+\sum_{k=1}^{K}\lambda_k^*(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}),\widehat{\boldsymbol{\theta}})\nabla_{\boldsymbol{\theta}}g_k\left(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}),\widehat{\boldsymbol{\theta}}\right) \ , \tag{85}$$

where $\boldsymbol{\lambda}^*(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}),\widehat{\boldsymbol{\theta}})=\left(\lambda_1^*(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\theta}})),\dots,\lambda_K^*(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}),\widehat{\boldsymbol{\theta}})\right)^T\in\Lambda^*(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}),\widehat{\boldsymbol{\theta}})$ are the Lagrange multipliers associated associated with $\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}})\in X^*(\widehat{\boldsymbol{\theta}})$.

Proof. First, we have $f, g_1, \ldots, g_K$ are continuous and concave in $\boldsymbol{y}$ and $\forall \boldsymbol{\theta} \in \Theta, \exists \boldsymbol{x} \in X$ s.t. $g_k(\boldsymbol{x}, \boldsymbol{\theta}) > 0$, for all $k = 1, \ldots, K$. We can restate it as $\tilde{f}, \tilde{g_1}, \ldots, \tilde{g_K}$ are continuous and convex in $\boldsymbol{y}$ and $\forall \boldsymbol{\theta} \in \Theta, \exists \boldsymbol{x} \in X$ s.t. $\tilde{g_k}(\boldsymbol{x}, \boldsymbol{\theta}) < 0$, for all $k = 1, \ldots, K$, where $\tilde{f} = -f$ and $\tilde{g_k} = -g_k \; \forall k = 1, \ldots, K$. Thus, we can apply the K.K.T theorem to the convex program $(\tilde{f}, \tilde{\boldsymbol{g}})$ and equivalently to the concave program $(f, \boldsymbol{g})$: Let $\boldsymbol{x}^* \in \max_{\boldsymbol{x} \in X : \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\theta}) \geq 0} f(\boldsymbol{x}, \boldsymbol{\theta})$ be the optimal solution to the value function problem, we have $\forall \widehat{\boldsymbol{\theta}} \in \Theta, \exists \boldsymbol{\lambda}^*(\boldsymbol{x}(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}}) \geq 0$ s.t.

$$V(\widehat{\boldsymbol{\theta}}) = L(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}), \boldsymbol{\lambda}^*(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}})$$

where L is the Lagrangian associated with the concave program $(f, \boldsymbol{g})$.

Then,

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} V(\widehat{\boldsymbol{\theta}}) &= \nabla_{\boldsymbol{\theta}} L(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}), \boldsymbol{\lambda}^*(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}}) + \\
&\quad \nabla_{\boldsymbol{\theta}} \boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}) \nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}), \boldsymbol{\lambda}^*(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}}), \boldsymbol{x}^*(\widehat{\boldsymbol{\theta}})) + \\
&\quad \nabla_{\boldsymbol{\theta}} \boldsymbol{\lambda}^*(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}}) \nabla_{\boldsymbol{\lambda}} L(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}), \boldsymbol{\lambda}^*(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}}), \boldsymbol{\lambda}^*(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}})) \\
&= \nabla_{\boldsymbol{\theta}} L(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}), \boldsymbol{\lambda}^*(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}}) \\
&= \nabla_{\boldsymbol{\theta}} \left(f\left(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}}\right) + \sum_{k=1}^{K} \lambda_k^*(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}}) g_k\left(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}}\right) \right) \\
&= \nabla_{\boldsymbol{\theta}} f\left(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}}\right) + \sum_{k=1}^{K} \lambda_k^*(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}}) \nabla_{\boldsymbol{\theta}} g_k\left(\boldsymbol{x}^*(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}}\right)
\end{aligned}$$

□

To grasp the intuition for the ET, think about a simple one-dimensional optimization problem with no constraints:[2]

$$V(\boldsymbol{\theta}) = \max_{\boldsymbol{x} \in \mathbb{R}} f(\boldsymbol{x}, \theta)$$

where $\boldsymbol{\theta} \in [0, 1]$ and f is differentiable. If the solution $\boldsymbol{x}^*(\boldsymbol{\theta})$ is differentiable, then $V(\theta) = f\left(\boldsymbol{x}^*(\theta), \boldsymbol{\theta}\right)$ is differentiable. Applying the chain rule, we get:

$$V'(\theta) = \underbrace{\frac{\partial f\left(\boldsymbol{x}^*(\theta), \boldsymbol{\theta}\right)}{\partial \boldsymbol{x}}}_{=0 \text{ (at an optimum)}} \times \frac{\partial \boldsymbol{x}^*(\boldsymbol{\theta})}{\partial \theta} + \frac{\partial f\left(\boldsymbol{x}^*(\boldsymbol{\theta}), \theta\right)}{\partial \boldsymbol{\theta}} = \frac{\partial f\left(\boldsymbol{x}^*(\theta), \theta\right)}{\partial \theta}.$$

A change in θ has two effects on the value function: (i) a direct effect $f_\theta\left(\boldsymbol{x}^*(\theta), \theta\right)$, and (ii) an indirect effect $f_{\boldsymbol{x}}\left(\boldsymbol{x}^*(\theta), \theta\right) \frac{\partial \boldsymbol{x}^*(\theta)}{\partial \theta}$. The ET tells us that under certain conditions, we can ignore the indirect effect and focus on the direct effect. In problems with constraints, there is also a third effect the change in the constraint set. If constraints are binding (some λ 's are positive), this effect is accounted for by the ET above.

Example 7.8. *Consider the optimization problem:*

$$\max_{x \in \mathbb{R} : 2x + 2\theta \geq 0} -x^2 + x + 2\theta + 2 \tag{86}$$

For clarity, let $f(x; \theta) = -x^2 + x + 2\theta + 2$ and $g(x; \theta) = 2x + 2\theta$. Clearly, the constraints define a convex set, i.e., g is concave, and f is concave. f and g are also continuously differentiable. Additionally, for all $\theta \in \Theta$, there exists $x > -\theta$ such that there exists an interior solution. The Lagrangian for the above problem is:

$$L(x, \lambda; \theta) = x^2 - x - 2\theta - 2 - \lambda(2x + 2\theta) \tag{87}$$

[2]This simple proof below can be generalized to the constrained case by going through the Lagrangian

From the KKT stationarity conditions, we obtain:

$$\frac{\partial L}{\partial x} = 2x^\star - 1 - 2\lambda^\star := 0 \tag{88}$$

Solving for the optimal value of the Lagrange multiplier $\lambda^\star$:

$$\lambda^\star = x^* - \frac{1}{2} \tag{89}$$

We now solve for the optimal variable x^*. Note that without constraints, the objective function achieves a maximum at $x^* = 1/2$:

$$\frac{df}{dx} = -2x + 1 := 0 \tag{90}$$

$$x^* = \frac{1}{2} \tag{91}$$

Since the constraint g requires that $x \geq -\theta$ and f is decreasing for $x \geq \frac{1}{2}$, the solution function x^* is given as:

$$x^*(\theta) = \begin{cases} \frac{1}{2} & if\, \theta \geq -\frac{1}{2} \\ -\theta & if\, \theta < -1/2 \end{cases} \tag{92}$$

and the Lagrange multiplier solution function is given by:

$$\lambda(x^*(\theta), \theta) = x^*(\theta) - \frac{1}{2} \tag{93}$$

The value function is given by:

$$V(\theta) = f(x^*(\theta), \theta) \tag{94}$$

$$= -x^*(\theta)^2 + x^*(\theta) + 2\theta + 2 \tag{95}$$

$$= \begin{cases} -\frac{1}{4} + \frac{1}{2} + 2\theta + 2 & if\, \theta \geq -\frac{1}{2} \\ -\theta^2 - \theta + 2\theta + 2 & if\, \theta < -1/2 \end{cases} \tag{96}$$

$$= \begin{cases} \frac{9}{4} + 2\theta & if\, \theta \geq -\frac{1}{2} \\ -\theta^2 + \theta + 2 & if\, \theta < -1/2 \end{cases} \tag{97}$$

Hence, by the Envelope theorem, the derivative of the value function for any $\theta \neq -\frac{1}{2}$ is given by:

$$\frac{\partial V(\theta)}{\partial \theta} = \frac{\partial f}{\partial \theta} + \lambda(x^*(\theta), \theta)\left(\frac{\partial g}{\partial \theta}\right) \tag{98}$$

$$= 2 + 2\left(x^*(\theta) - \frac{1}{2}\right) \tag{99}$$

$$= 2 + 2x^\star(\theta) - 1 \tag{100}$$

$$= 1 + 2x^\star(\theta) \tag{101}$$

$$= \begin{cases} 2 & if\, \theta \geq -\frac{1}{2} \\ 1 - 2\theta & if\, \theta < -\frac{1}{2} \end{cases} \tag{102}$$

One can easily verify that derivative given by the envelope theorem is exactly equal to the derivative we could calculate by differentiating eq. (97)*. A 2D and 3D geometric animation of these functions can be found here and here.*

[Deni: *Add integral envelope theorem*]

8 Gradient Descent

In this section, we assume that we are dealing with a minimization problem whose objective function is differentiable and convex.

8.1 Unconstrained Optimization

We first study solving unconstrained minimization problems of the following type using gradient descent:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \tag{103}$$

where we assume that $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable and convex function.
Gradient descent is an iterative method defined as follows:

$$\boldsymbol{x}(t) = \boldsymbol{x}(t-1) - \gamma_t \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t-1)) \qquad \forall t = 1, \dots \tag{104}$$

$$\forall \boldsymbol{x}(0) \in \mathbb{R}^n \tag{105}$$

8.1.1 Smooth Objective

We will prove the following result regarding the convergence properties of gradient descent:

Theorem 8.1. *Suppose the function $f : \mathbb{R}^n \to \mathbb{R}$ in (103) is convex and differentiable, and $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ is L-Lipschitz (i.e., $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n,\ ||\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{y})|| \leq L||\boldsymbol{x} - \boldsymbol{y}||$). Let $\boldsymbol{x}^* \in \mathbb{R}^n$ be the minimizer of f. Then, if we run the process in (104) up to time t, with $\forall t = 1, \dots, \quad \gamma_t = \frac{1}{L}$, the process will yield a solution $f(\boldsymbol{x}(t))$ which satisfies:*

$$f(\boldsymbol{x}(t)) - f(\boldsymbol{x}^*) \leq L\frac{||\boldsymbol{x}(0) - \boldsymbol{x}^*||}{2t} \tag{106}$$

We define the **rate of convergence** of an iterative process as a function $\tau(t)$ that satisfies $f(\boldsymbol{x}(t)) - f(\boldsymbol{x}^*) = O(\tau(t))$. Intuitively, an implication of the above result is that gradient descent is guaranteed to converge and that it converges with rate $O\left(\frac{1}{t}\right)$

Proof. The assumption that $\nabla_{\boldsymbol{x}} f$ is L-Lipschitz continuous implies that $\nabla^2_{\boldsymbol{x}} f \leq L\boldsymbol{I}$, where $\nabla^2_{\boldsymbol{x}} f$ is the Hessian of f, and $\boldsymbol{I}$ is the identity matrix. We can then use a taylor expansion of f around $\boldsymbol{x}$ to approximate any $f(\boldsymbol{y})$:

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^T \nabla^2_{\boldsymbol{x}} f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) \tag{107}$$

$$\leq f(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}L(\boldsymbol{y} - \boldsymbol{x})^T \boldsymbol{I}(\boldsymbol{y} - \boldsymbol{x}) \tag{108}$$

$$= f(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}L||\boldsymbol{y} - \boldsymbol{x}||^2 \tag{109}$$

Plugging $\boldsymbol{y} = \boldsymbol{x}(t+1)$ and $\boldsymbol{x} = \boldsymbol{x}(t)$, we get the following:

$$f(\boldsymbol{x}(t+1)) \leq f(\boldsymbol{x}(t)) + \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t))^T(\boldsymbol{x}(t+1) - \boldsymbol{x}(t)) + \frac{1}{2}L||\boldsymbol{x}(t+1) - \boldsymbol{x}(t)||^2 \tag{110}$$

$$= f(\boldsymbol{x}(t)) + \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t))^T(\boldsymbol{x}(t) - \gamma_t \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t)) - \boldsymbol{x}(t)) + \frac{1}{2}L||\boldsymbol{x}(t) - \gamma_t \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t)) - \boldsymbol{x}(t)||^2 \tag{111}$$

$$= f(\boldsymbol{x}(t)) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t))^T \gamma_t \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t)) + \frac{1}{2}L||\gamma_t \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t))||^2 \tag{112}$$

$$= f(\boldsymbol{x}(t)) - \gamma_t||\nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t))||^2 + \frac{1}{2}\nabla^2_{\boldsymbol{x}} L\gamma_t^2||\nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t))||^2 \tag{113}$$

$$= f(\boldsymbol{x}(t)) - (1 - \frac{1}{2}\gamma_t L)\gamma_t||\nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t))||^2 \tag{114}$$

Since $\gamma_t = \frac{1}{L}$, we know that $-(1 - \frac{1}{2}\gamma_t L) = -\frac{1}{2}$, which then gives us:

$$f(\boldsymbol{x}(t+1)) \leq f(\boldsymbol{x}(t)) - \frac{1}{2}\gamma_t ||\nabla_{\boldsymbol{x}}^T f(\boldsymbol{x}(t))||^2 \tag{115}$$

Since $\gamma_t ||\nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t))||^2$ is always strcitly positive unless $\boldsymbol{x}(t) = \boldsymbol{x}^*$ this inequality implies that gradient descent converges to the optimal value of the minimand since at each iteration it strictly decreases.

Next, we bound the value of $f(\boldsymbol{x}(t+1))$. First note by convexity, we have:

$$f(\boldsymbol{x}^*) \geq f(\boldsymbol{x}(t)) + \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t))^T (\boldsymbol{x}^* - \boldsymbol{x}(t)) \tag{116}$$

$$f(\boldsymbol{x}(t)) \leq f(\boldsymbol{x}^*) + \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t))^T (\boldsymbol{x}(t) - \boldsymbol{x}^*) \tag{117}$$

Plugging this into 115, we obtain:

$$f(\boldsymbol{x}(t+1)) \leq f(\boldsymbol{x}^*) + \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t))^T (\boldsymbol{x}(t) - \boldsymbol{x}^*) - \frac{1}{2}\gamma_t ||\nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t))||^2 \tag{118}$$

$$f(\boldsymbol{x}(t+1)) - f(\boldsymbol{x}^*) \leq \frac{1}{2\gamma_t} \left(2\gamma_t \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t))^T (\boldsymbol{x}(t) - \boldsymbol{x}^*) - \gamma_t^2 ||\nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t))||^2 \right) \tag{119}$$

$$f(\boldsymbol{x}(t+1)) - f(\boldsymbol{x}^*) \leq \frac{1}{2\gamma_t} \left(2\gamma_t \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t))^T (\boldsymbol{x}(t) - \boldsymbol{x}^*) - \gamma_t^2 ||\nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t))||^2 - ||\boldsymbol{x}(t) - \boldsymbol{x}^*||^2 + ||\boldsymbol{x}(t) - \boldsymbol{x}^*||^2 \right) \tag{120}$$

Note that we have the following expansion:

$$-||\boldsymbol{x} - \gamma_t \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) - \boldsymbol{x}^*|| = -\left(||\boldsymbol{x}(t) - \boldsymbol{x}^*||^2 - 2\gamma_t \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t))^T (\boldsymbol{x}(t) - \boldsymbol{x}^*) + \gamma_t^2 ||\nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t))||^2 \right) \tag{121}$$

$$= \gamma_t^2 ||\nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t))||^2 + 2\gamma_t \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t))^T (\boldsymbol{x}(t) - ||\boldsymbol{x}(t) - \boldsymbol{x}^*||^2 \tag{122}$$

Going back to equation 120, we get:

$$f(\boldsymbol{x}(t+1)) - f(\boldsymbol{x}^*) \leq \frac{1}{2\gamma_t} \left(||\boldsymbol{x}(t) - \boldsymbol{x}^*||^2 - ||\boldsymbol{x}(t) - \gamma_t \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) - \boldsymbol{x}^*|| \right) \tag{123}$$

$$= \frac{1}{2\gamma_t} \left(||\boldsymbol{x}(t) - \boldsymbol{x}^*||^2 - ||\boldsymbol{x}(t+1) - \boldsymbol{x}^*|| \right) \qquad \text{(By definition of } \boldsymbol{x}(t+1)) \tag{124}$$

Since this inequality holds between any consecutive time periods, we can sum the change in the value of f the first T iterations:

$$\sum_{t=0}^{T-1} f(\boldsymbol{x}(t+1)) - f(\boldsymbol{x}^*) \leq \sum_{t=0}^{T-1} \frac{1}{2\gamma_t} \left(||\boldsymbol{x}(t) - \boldsymbol{x}^*||^2 - ||\boldsymbol{x}(t+1) - \boldsymbol{x}^*|| \right) \tag{125}$$

$$= \frac{1}{2\gamma_t} \left(||\boldsymbol{x}(0) - \boldsymbol{x}^*||^2 - ||\boldsymbol{x}(T) - \boldsymbol{x}^*|| \right) \tag{126}$$

$$= \frac{1}{2\gamma_t} \left(||\boldsymbol{x}(0) - \boldsymbol{x}^*||^2 \right) \tag{127}$$

Since f is strictly decreasing on each iteration, we have:

$$\sum_{t=0}^{T-1} f(\boldsymbol{x}(t+1)) - f(\boldsymbol{x}^*) \geq T \left(f(\boldsymbol{x}(T)) - f(\boldsymbol{x}^*) \right) \tag{128}$$

We can then conclude:

$$f(\boldsymbol{x}(T)) - f(\boldsymbol{x}^*) \leq \frac{||\boldsymbol{x}(0) - \boldsymbol{x}^*||^2}{2\gamma_t T} = L \frac{||\boldsymbol{x}(0) - \boldsymbol{x}^*||^2}{2T} \tag{129}$$

□

An implication of the above theorem is that we achieve an $\boldsymbol{x}(t)$ that is at an ϵ distance from the minimum in $O\left(\frac{1}{\epsilon}\right)$. Such a bound is called a **sub-linear convergence bound**.

8.1.2 Strongly Convex and Smooth Objective

It turns out we can achieve an even better bound if our objective function is both Lipschitz smooth and strongly convex:

Theorem 8.2. *Let f be L-Lipschitz smooth and m-strongly convex. Let $\boldsymbol{x}^* \in \mathbb{R}^n$ be the minimizer of f. Then, if we run the process in (104) up to time t, with $\forall t = 1, 2, \ldots, \ \gamma_t = \frac{1}{L}$, then for any iterate $\boldsymbol{x}(t)$, the following holds:*

$$||\boldsymbol{x}(t) - \boldsymbol{x}^*|| \leq \left(1 - \frac{m}{L}\right)^t ||\boldsymbol{x}(0) - \boldsymbol{x}^*|| \tag{130}$$

Proof.

$$\begin{aligned} ||\boldsymbol{x}(t) - \boldsymbol{x}^*|| &= ||\boldsymbol{x}(t-1) - \gamma_t \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t-1)) - \boldsymbol{x}^*|| && (131)\\ &= ||\boldsymbol{x}(t-1) - \boldsymbol{x}^*||^2 - 2\gamma_t \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t-1))^T (\boldsymbol{x}(t-1) - \boldsymbol{x}^*) + \gamma_t^2 ||\nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t-1))||^2 && (132) \end{aligned}$$

Note that from strong convexity we have:

$$\begin{aligned} f(\boldsymbol{y}) &\geq f(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} f(\boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y}) + \frac{m}{2}||\boldsymbol{y} - \boldsymbol{x}|| && (133)\\ &\geq f(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{y})(\boldsymbol{y} - \boldsymbol{x}) + \frac{m}{2}||\boldsymbol{y} - \boldsymbol{x}|| && (134) \end{aligned}$$

Plugging $\boldsymbol{y} = \boldsymbol{x}(t-1)$ and $\boldsymbol{x} = \boldsymbol{x}(t)$, we get:

$$f(\boldsymbol{x}(t-1)) \geq f(\boldsymbol{x}(t)) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t-1))(\boldsymbol{x}(t-1) - \boldsymbol{x}(t)) + \frac{m}{2}||\boldsymbol{x}(t-1) - \boldsymbol{x}(t)|| \tag{135}$$

Re-organizing this expression we get:

$$f(\boldsymbol{x}(t-1)) - f(\boldsymbol{x}(t)) - \frac{m}{2}||\boldsymbol{x}(t-1) - \boldsymbol{x}(t)|| \geq -\nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t-1))(\boldsymbol{x}(t-1) - \boldsymbol{x}(t)) \tag{136}$$

Multiplying both sides by $2\gamma_t$, we get:

$$2\gamma_t f(\boldsymbol{x}(t-1)) - 2\gamma_t f(\boldsymbol{x}(t)) - m\gamma_t ||\boldsymbol{x}(t-1) - \boldsymbol{x}(t)|| \geq -2\gamma_t \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t-1))(\boldsymbol{x}(t-1) - \boldsymbol{x}(t)) \tag{137}$$

Going back to 132 and using the above, we get:

$$\begin{aligned} ||\boldsymbol{x}(t) - \boldsymbol{x}^*|| &= ||\boldsymbol{x}(t-1) - \boldsymbol{x}^*||^2 - 2\gamma_t \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t-1))^T (\boldsymbol{x}(t-1) - \boldsymbol{x}^*) + \gamma_t^2 ||\nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t-1))||^2 && (138)\\ &\leq (1 - \gamma_t m)||\boldsymbol{x}(t-1) - \boldsymbol{x}^*||^2 - 2\gamma_t \left(f(\boldsymbol{x}(t-1)) - f(\boldsymbol{x}(t))\right) + \gamma_t^2 ||\nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t-1))||^2 && (139) \end{aligned}$$

Furthermore, we know that for any L-smooth function f, we have $\forall \boldsymbol{x}, \ \ f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \leq \frac{1}{2L}||\nabla_{\boldsymbol{x}} f(\boldsymbol{x})||^2$ which then gives us:

$$\begin{aligned} ||\boldsymbol{x}(t) - \boldsymbol{x}^*|| &\leq (1 - \gamma_t m)\, ||\boldsymbol{x}(t-1) - \boldsymbol{x}^*||^2 - 2\gamma_t \left(f(\boldsymbol{x}(t-1)) - f(\boldsymbol{x}(t))\right) + 2\gamma_t^2 L \left(f(\boldsymbol{x}) - f(\boldsymbol{x}^*)\right) && (140)\\ &= (1 - \gamma_t m)\, ||\boldsymbol{x}(t-1) - \boldsymbol{x}^*||^2 - 2\gamma_t (1 - \gamma_t L) \left(f(\boldsymbol{x}) - f(\boldsymbol{x}^*)\right) && (141)\\ &= \left(1 - \frac{m}{L}\right) ||\boldsymbol{x}(t-1) - \boldsymbol{x}^*||^2 - 2\frac{1}{L}\left(1 - \frac{L}{L}\right) \left(f(\boldsymbol{x}) - f(\boldsymbol{x}^*)\right) && (142)\\ &= \left(1 - \frac{m}{L}\right) ||\boldsymbol{x}(t-1) - \boldsymbol{x}^*||^2 && (143) \end{aligned}$$

Unrolling the left hand side as above up to time $t = 0$, we obtain the desired result. □

An implication of the above theorem is that we achieve an $\boldsymbol{x}(t)$ that is at an ϵ distance from the minimum in $O\left(\log\left(\frac{1}{\epsilon}\right)\right)$. Such a bound is called a **linear convergence bound**.

8.2 Constrained Optimization

In this section, we study the optimization problems of the following type:

$$\min_{\boldsymbol{x} \in X} f(\boldsymbol{x}) \tag{144}$$

where we assume that $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable and convex function and X is a convex set.
The projection of a point $\boldsymbol{y}$ onto a set X, is defined as:

$$\Pi_X(\boldsymbol{y}) = \arg\min_{\boldsymbol{x} \in X} ||\boldsymbol{x} - \boldsymbol{y}||^2 \tag{145}$$

We now introduce **projected gradient descent** which allows us to solve optimization problems with constraints as in 144:

$$\boldsymbol{x}(t) = \boldsymbol{x}(t-1) - \Pi_X \left(\boldsymbol{x}(t-1) - \gamma_t \nabla_{\boldsymbol{x}} f(\boldsymbol{x}(t-1)) \right) \qquad \forall t = 1, \dots \tag{146}$$

$$\forall \boldsymbol{x}(0) \in X \tag{147}$$

We now introduce convergence results regarding projected gradient descent. We skip the proofs as they are similar to those for gradient descent modulo the projection operator.

Theorem 8.3. *Suppose the function $f : \mathbb{R}^n \to \mathbb{R}$ in (144) is convex and differentiable, and $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ is L-Lipschitz. Let $\boldsymbol{x}^* \in \mathbb{R}^n$ be the minimizer of f. Then, if we run the process in (146) up to time t, with $\forall t = 1, \dots, \quad \gamma_t = \frac{1}{L}$, the process will yield a solution $f(\boldsymbol{x}(t))$ which satisfies:*

$$f(\boldsymbol{x}(t)) - f(\boldsymbol{x}^*) \leq 2L \frac{||\boldsymbol{x}(0) - \boldsymbol{x}^*||}{t} \tag{148}$$

Theorem 8.4. *Let f be L-Lipschitz smooth and m-strongly convex. Let $\boldsymbol{x}^* \in \mathbb{R}^n$ be the minimizer of f. Then, if we run the process in (146) up to time t, with $\forall t = 1, 2, \dots, \ \gamma_t = \frac{1}{L}$, then for any iterate $\boldsymbol{x}(t)$, the following holds:*

$$||\boldsymbol{x}(t) - \boldsymbol{x}^*|| \leq (1 - \frac{m}{L})^t ||\boldsymbol{x}(0) - \boldsymbol{x}^*|| \tag{149}$$

9 Subgradient Methods

So far we have considered only optimization problems with differentiable objectives. In this section, we introduce the tools to solve optimization problems with objective functions that are not necessarily differentiable.

9.1 Subgradient Descent

Consider the optimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \tag{150}$$

where f is a convex function that is not necessarily differentiable.

The conventional method to solve the above problem is the **subgradient method**:

$$\boldsymbol{x}(t) = \boldsymbol{x}(t-1) - \eta_t \boldsymbol{h}(t-1) \qquad t = 0, 1, \dots \tag{151}$$

$$\boldsymbol{x}(0) \in V \tag{152}$$

where $\boldsymbol{h}(t) \in \partial_{\boldsymbol{x}}$ and η_t is the learning rate at time t.

The subgradient method is not a descent method and as a result the last iterate might not be the best estimate of the minimizer of f. As a result, the subgradient method requires us to keep track of the best iterate $\boldsymbol{x}_{\text{best}}^{(T)}$ up to time T. That is:

$$\boldsymbol{x}_{\text{best}}^{(T)} = \arg\min_{t\in[T]} f(\boldsymbol{x}(t)) \tag{153}$$

Theorem 9.1. *Consider the iterative process 169. Let f be a convex function $f : \mathbb{R}^n \to \mathbb{R}$ that is L-Lipschitz and $\boldsymbol{h}(t) \in \partial_{\boldsymbol{x}} f(\boldsymbol{x}(t))$. Let $\boldsymbol{x}^*$ be the minimizer of f. Assume that the step sizes γ_t satisfy the following:*

$$\sum_{k=1}^{t} \gamma_k^2 \leq \infty \qquad \sum_{k=1}^{t} \gamma_k = \infty \tag{154}$$

Then, we have:

$$\lim_{k\to\infty} f(\boldsymbol{x}_{\text{best}}^{(k-1)}) = f(\boldsymbol{x}^*) \tag{155}$$

furthermore, the following convergence bounds hold:

$$f(\boldsymbol{x}_{\text{best}}^{(k-1)}) - f(\boldsymbol{x}^*) \leq \frac{||\boldsymbol{x}(0) - \boldsymbol{x}^*||^2 + L^2 \sum_{k=1}^{t} \gamma_k^2}{2\left(\sum_{k=1}^{t} \gamma_k\right)} \tag{156}$$

$$\leq \frac{L||\boldsymbol{x}(0) - \boldsymbol{x}^*||}{\sqrt{t}} \tag{157}$$

Proof.

$$||\boldsymbol{x}(t) - \boldsymbol{x}^*|| = ||\boldsymbol{x}(t-1) - \gamma_t \boldsymbol{h}(t-1) - -\boldsymbol{x}^*||^2 \tag{158}$$

$$= ||\boldsymbol{x}(t-1) - \boldsymbol{x}^*||^2 - 2\gamma_t \boldsymbol{h}(t)^T(\boldsymbol{x}(t-1) - \boldsymbol{x}^*) + \gamma_t^2 ||\boldsymbol{h}(t-1)||^2 \tag{159}$$

$$\leq ||\boldsymbol{x}(t-1) - \boldsymbol{x}^*||^2 - 2\gamma_t (f(\boldsymbol{x}(t-1)) - f(\boldsymbol{x}^*)) + \gamma_t^2 ||\boldsymbol{h}(t-1)||^2 \tag{160}$$

where the last inequality was derived from the definition of the subgradient. Applying the inequality above recursively, we obtain:

$$||\boldsymbol{x}(t) - \boldsymbol{x}^*|| \leq ||\boldsymbol{x}(0) - \boldsymbol{x}^*||^2 - \sum_{k=1}^{t} 2\gamma_k (f(\boldsymbol{x}(k-1)) - f(\boldsymbol{x}^*)) + \sum_{k=1}^{t} \gamma_k^2 ||\boldsymbol{h}(k-1)||^2 \tag{161}$$

Since $||\boldsymbol{x}(t) - \boldsymbol{x}^*|| \geq 0$ and let $||\boldsymbol{x}(0) - \boldsymbol{x}^*|| = R$, for clarity, we have:

$$\sum_{k=1}^{t} 2\gamma_k (f(\boldsymbol{x}(k-1)) - f(\boldsymbol{x}^*)) \leq R^2 + \sum_{k=1}^{t} \gamma_k^2 ||\boldsymbol{h}(k-1)||^2 \tag{162}$$

Note that we have:

$$\sum_{k=1}^{t} \gamma_k (f(\boldsymbol{x}(k-1)) - f(\boldsymbol{x}^*)) \geq \left(\sum_{k=1}^{t} \gamma_k \right) \min_k (f(\boldsymbol{x}(k-1)) - f(\boldsymbol{x}^*)) \tag{163}$$

$$= \left(\sum_{k=1}^{t} \gamma_k \right) (f(\boldsymbol{x}_{\text{best}}^{(k-1)}) - f(\boldsymbol{x}^*)) \tag{164}$$

Hence, we get the following bound:

$$f(\boldsymbol{x}_{\text{best}}^{(k-1)}) - f(\boldsymbol{x}^*) \leq \frac{R^2 + \sum_{k=1}^{t} \gamma_k^2 ||\boldsymbol{h}(k-1)||^2}{2 \left(\sum_{k=1}^{t} \gamma_k \right)} \tag{165}$$

Since f is L-Lipschitz, we know that $||\boldsymbol{h}(k-1)|| \leq L$, which gives u:

$$f(\boldsymbol{x}_{\text{best}}^{(k-1)}) - f(\boldsymbol{x}^*) \leq \frac{R^2 + L^2 \sum_{k=1}^{t} \gamma_k^2}{2 \left(\sum_{k=1}^{t} \gamma_k \right)} \tag{166}$$

Since we assumed that:

$$\sum_{k=1}^{t} \gamma_k^2 \leq \infty \qquad\qquad \sum_{k=1}^{t} \gamma_k = \infty \tag{167}$$

as $t \to \infty$, $\lim_{k \to \infty} f(\boldsymbol{x}_{\text{best}}^{(k-1)}) = f(\boldsymbol{x}^*)$ □

9.2 Projected Subgradient Methods

The subgradient method presented in the previous section can be extended to constrained optimization problem with non-differentiable objective functions.

Consider the optimization problem:

$$\min_{\boldsymbol{x} \in X} f(\boldsymbol{x}) \tag{168}$$

where f is a convex function that is not necessarily differentiable and X is a convex set.

The conventional method to solve the above problem is the **subgradient method**:

$$\boldsymbol{x}(t) = \boldsymbol{x}(t-1) - \Pi_X \left(\boldsymbol{x}(t-1) - \eta_t \boldsymbol{h}(t-1)\right) \qquad t = 1, 2, \ldots \tag{169}$$

$$\boldsymbol{x}(0) \in V \tag{170}$$

where $\boldsymbol{h}(t) \in \partial_{\boldsymbol{x}} f(\boldsymbol{x}(t))$, Π_X is the projection operator onto the set X and η_t is the learning rate at time t. It turns out that the convergence bound for projected subgradient descent is the same as the convergence bound for subgradient descent.

10 The Proximal Gradient Method

10.0.1 The Proximal Operator

Given a function $f : E \to (-\infty, \infty]$, the **proximal mapping** is the operator prox given by:

$$\text{prox}_f(\boldsymbol{x}) = \arg\min_{\boldsymbol{u} \in E} \left\{ f(\boldsymbol{u}) + \frac{1}{2} \|\boldsymbol{u} - \boldsymbol{x}\|_2^2 \right\} \tag{171}$$

The proximal operator essential returns the minimizer of a regularized version of the function f.

Example 10.1. *Consider $f : \mathbb{R} \to (-\infty, \infty]$ s.t. $f(x) = 0$, then*

$$\text{prox}_f(x) = \arg\min_{u \in \mathbb{R}} \left\{ f(u) + \frac{1}{2} \|u - x\|_2^2 \right\} = \arg\min_{u \in \mathbb{R}} \left\{ \frac{1}{2} \|u - x\|_2^2 \right\} = \{x\}$$

Example 10.2. *Consider $f : \mathbb{R} \to (-\infty, \infty]$ s.t. $f(x) = \mathrm{I}_C$, where I_C is the indicator function of some set $C \subset \mathbb{R}$, then*

$$\text{prox}_f(x) = \arg\min_{u \in \mathbb{R}} \left\{ \mathrm{I}_C(u) + \frac{1}{2} \|u - x\|_2^2 \right\} = \arg\min_{u \in C} \left\{ \frac{1}{2} \|u - x\|_2^2 \right\} = \Pi_C(x)$$

Example 10.3. *Consider the functions:*

$$f_1 = \begin{cases} 0 & x \neq 0 \\ -1/2 & x = 0 \end{cases}$$

$$f_2 = \begin{cases} 0 & x \neq 0 \\ 1/2 & x = 0 \end{cases}$$

To compute the proximal of f_1, note that $\text{prox}_{f_1}(x) = \arg\min_{u \in \mathrm{R}} \tilde{f}_1(u, x)$, where

$$\tilde{f}_1(u, x) \doteq f_1(u) + \frac{1}{2}(u - x)^2 = \begin{cases} \frac{1}{2}(x^2 - 1), & u = 0 \\ \frac{1}{2}(u - x)^2, & u \neq 0 \end{cases}$$

which allows us to compute the proximal of f_1 as:

$$\text{prox}_{f_1}(x) = \begin{cases} \{0\} & |x| < 1 \\ \{x\} & |x| > 1 \\ \{0, x\} & |x| = 1 \end{cases}$$

To compute the proximal of f_2, note that $\text{prox}_{f_2}(x) = \arg\min_{u \in \mathrm{R}} \tilde{f}_2(u, x)$, where

$$\tilde{f}_2(u, x) \doteq f_2(u) + \frac{1}{2}(u - x)^2 = \begin{cases} \frac{1}{2}(x^2 + 1), & u = 0 \\ \frac{1}{2}(u - x)^2, & u \neq 0 \end{cases}$$

A similar argument shows that:

$$\text{prox}_{f_2}(x) = \begin{cases} \{x\} & x \neq 0 \\ \emptyset \text{[Amy : ill-defined]} & x = 0 \end{cases}$$

Consider a function $f : E \to [-\infty, \infty]$. f is said to be a **proper function** if:

$$f(x) > -\infty \qquad \forall x \in E \tag{172}$$

and

$$f(x) < \infty \qquad \exists x \in E \tag{173}$$

That is, a convex function is proper if never takes a negative infinite value and its effective domain [Amy: *please define!*] is non-empty, i.e., $\text{dom}(f) \neq \emptyset$.

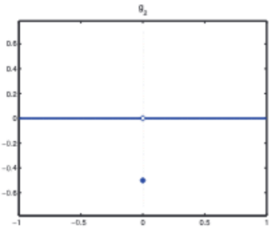
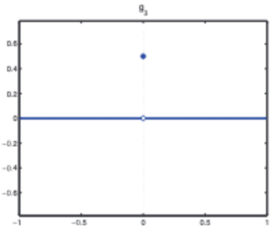

Figure 6.1. *The left and right images are the plots of the functions g_2 and g_3, respectively, with $\lambda = 0.5$ from Example 6.2.*

Theorem 10.4 (First Prox Theorem). *Let $f : E \to (-\infty, \infty]$ be a proper closed and convex function. Then, $\operatorname{prox}_f(\boldsymbol{x})$ is a singleton for any $\boldsymbol{x} \in E$.*

Proof. Let $\tilde{f}(\boldsymbol{x}, \boldsymbol{u}) = f(\boldsymbol{u}) + \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{x}\|_2^2$, then $\operatorname{prox}_f(\boldsymbol{x}) = \arg\min_{\boldsymbol{u} \in E} \tilde{f}(\boldsymbol{x}, \boldsymbol{u})$. Note that $\tilde{f}(\boldsymbol{x}, \boldsymbol{u})$ is strongly convex in $\boldsymbol{u}$ since it is the sum of a closed function and strongly convex function. This means that $\tilde{f}(\boldsymbol{x}, \boldsymbol{u})$ is strictly convex in $\boldsymbol{u}$, which implies that $\operatorname{prox}_f(\boldsymbol{x}) = \arg\min_{\boldsymbol{u} \in E} \tilde{f}(\boldsymbol{x}, \boldsymbol{u})$ is singleton valued for all $\boldsymbol{x} \in E$. □

As we will consider mostly closed and convex functions, from now on we will assume that prox_f is a vector-valued function, rather than a set-valued mapping, i.e., we will write $\operatorname{prox}_f(\boldsymbol{x}) = \boldsymbol{y}$ rather than $\operatorname{prox}_f(\boldsymbol{x}) = \{\boldsymbol{y}\}$.

A function $f : E \to (-\infty, \infty]$ is said to be **coercive** if $f(\boldsymbol{x}) \to \infty$ as $\|\boldsymbol{x}\| \to \infty$.

Theorem 10.5. *Let $f : E \to (-\infty, \infty]$ be a proper closed function such that $\tilde{f}(\boldsymbol{x}, \boldsymbol{u}) = f(\boldsymbol{u}) + \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{x}\|_2^2$ is coercive in $\boldsymbol{u}$ for all $\boldsymbol{x} \in E$. Then $\operatorname{prox}_f(\boldsymbol{x})$ is non-empty for all $\boldsymbol{x} \in E$.*

Proof. $\tilde{f}(\boldsymbol{x}, \boldsymbol{u})$ is closed since it is the sum of two closed functions. Since $\tilde{f}(\boldsymbol{x}, \boldsymbol{u})$ is coercive and closed in $\boldsymbol{u}$, its minimum always exists. □

An illustration of the above theorem can be found in example 10.3.

Corollary 10.6. *Let $f : E \to (-\infty, \infty]$ be a continuous function, then $\operatorname{prox}_f(\boldsymbol{x})$ is non-empty for all $\boldsymbol{x} \in E$.*

We summarize in the table below closed form expression for the proximal operators of important classes of functions:

f	$\operatorname{prox}_f(\boldsymbol{x})$
c	$\boldsymbol{x}$
$\langle \boldsymbol{a}, \boldsymbol{x} \rangle + b$	$\boldsymbol{x} - \boldsymbol{a}$
$1/2\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T \boldsymbol{x} + c$	$(\boldsymbol{A} + \boldsymbol{I})^{-1}(\boldsymbol{x} - \boldsymbol{b})$

We also note the following proximal calculus rules, which can allow one to derive closed form expressions for complicated functions. A more complete review of these rules and more closed form expressions for proximal operator can be found here.

$f(\mathbf{x})$	$\text{prox}_f(\mathbf{x})$	Assumptions
$\sum_{i=1}^m f_i(\mathbf{x}_i)$	$\text{prox}_{f_1}(\mathbf{x}_1) \times \cdots \times \text{prox}_{f_m}(\mathbf{x}_m)$	
$g(\lambda\mathbf{x} + \mathbf{a})$	$\frac{1}{\lambda}\left[\text{prox}_{\lambda^2 g}(\lambda\mathbf{x} + \mathbf{a}) - \mathbf{a}\right]$	$\lambda \neq 0, \mathbf{a} \in \mathbb{E}$, g proper
$\lambda g(\mathbf{x}/\lambda)$	$\lambda\text{prox}_{g/\lambda}(\mathbf{x}/\lambda)$	$\lambda \neq 0$, g proper
$g(\mathbf{x}) + \frac{c}{2}\|\mathbf{x}\|^2 + \langle \mathbf{a}, \mathbf{x}\rangle + \gamma$	$\text{prox}_{\frac{1}{c+1}g}(\frac{\mathbf{x}-\mathbf{a}}{c+1})$	$\mathbf{a} \in \mathbb{E}$, $c > 0$, $\gamma \in \mathbb{R}$, g proper
$g(\mathcal{A}(\mathbf{x}) + \mathbf{b})$	$\mathbf{x} + \frac{1}{\alpha}\mathcal{A}^T(\text{prox}_{\alpha g}(\mathcal{A}(\mathbf{x}) + \mathbf{b}) - \mathcal{A}(\mathbf{x}) - \mathbf{b})$	$\mathbf{b} \in \mathbb{R}^m$, $\mathcal{A} : \mathbb{V} \to \mathbb{R}^m$, g proper closed convex, $\mathcal{A} \circ \mathcal{A}^T = \alpha I$, $\alpha > 0$
$g(\|\mathbf{x}\|)$	$\text{prox}_g(\|\mathbf{x}\|)\frac{\mathbf{x}}{\|\mathbf{x}\|}$, $\mathbf{x} \neq \mathbf{0}$ $\{\mathbf{u} : \|\mathbf{u}\| = \text{prox}_g(0)\}$, $\mathbf{x} = \mathbf{0}$	g proper closed convex, $\text{dom}(g) \subseteq [0, \infty)$

Figure 8: Summary of proximal calculus rules

We now introduce an important result regarding the proximal operator.

Theorem 10.7 (Second Prox Theorem)**.** *Let $f : E \to (-\infty, \infty]$ be a proper closed and convex function. Then for any $\boldsymbol{x}, \boldsymbol{v} \in E$, the following three claims are equivalent:*

1. $\boldsymbol{v} = \text{prox}_f(\boldsymbol{x})$
2. $\boldsymbol{x} - \boldsymbol{v} \in \partial f(\boldsymbol{v})$
3. $\langle \boldsymbol{x} - \boldsymbol{v}, \boldsymbol{y} - \boldsymbol{v}\rangle \leq f(\boldsymbol{y}) - f(\boldsymbol{v})$ *for any* $\boldsymbol{y} \in E$

Proof. By definition, $\boldsymbol{v} = \text{prox}_f(\boldsymbol{x})$ if and only if

$$\boldsymbol{v} = \arg\min_{\boldsymbol{u}} \left\{ f(\boldsymbol{u}) + \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{x}\|^2 \right\},$$

which, by the KKT optimality conditions theorem 7.3 and the sum rule of subdifferential calculus, is equivalent to the relation

$$\mathbf{0} \in \partial f(\boldsymbol{v}) + \boldsymbol{v} - \boldsymbol{x}$$

We have thus shown the equivalence between claims (i) and (ii). Finally, by the definition of the subgradient, the membership relation of claim (ii) is equivalent to (iii). □

An important consequence of the above theorem is that minimizing prox_f is equivalent to minimizing f. The above theorem allows us to show that the proximal operator is non-expansive:

Theorem 10.8. *Let f be a proper closed and convex function. Then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{E}$,*
(a) (firm nonexpansivity)

$$\left\langle \boldsymbol{x} - \boldsymbol{y}, \text{prox}_f(\boldsymbol{x}) - \text{prox}_f(\boldsymbol{y})\right\rangle \geq \left\|\text{prox}_f(\boldsymbol{x}) - \text{prox}_f(\boldsymbol{y})\right\|^2$$

(b) (nonexpansivity)

$$\left\|\operatorname{prox}_f(\boldsymbol{x}) - \operatorname{prox}_f(\boldsymbol{y})\right\| \leq \|\boldsymbol{x} - \boldsymbol{y}\|.$$

Proof. (a) Denoting $\boldsymbol{u} = \operatorname{prox}_f(\boldsymbol{x}), \boldsymbol{v} = \operatorname{prox}_f(\boldsymbol{y})$, by the equivalence of (i) and (ii) in the second prox theorem (Theorem 6.39), it follows that

$$\boldsymbol{x} - \boldsymbol{u} \in \partial f(\boldsymbol{u}), \boldsymbol{y} - \boldsymbol{v} \in \partial f(\boldsymbol{v})$$

Thus, by the subgradient inequality,

$$f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \langle \boldsymbol{x} - \boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u} \rangle,$$
$$f(\boldsymbol{u}) \geq f(\boldsymbol{v}) + \langle \boldsymbol{y} - \boldsymbol{v}, \boldsymbol{u} - \boldsymbol{v} \rangle.$$

Summing the above two inequalities, we obtain

$$0 \geq \langle \boldsymbol{y} - \boldsymbol{x} + \boldsymbol{u} - \boldsymbol{v}, \boldsymbol{u} - \boldsymbol{v} \rangle$$

which is the same as

$$\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{u} - \boldsymbol{v} \rangle \geq \|\boldsymbol{u} - \boldsymbol{v}\|^2$$

that is,

$$\left\langle \boldsymbol{x} - \boldsymbol{y}, \operatorname{prox}_f(\boldsymbol{x}) - \operatorname{prox}_f(\boldsymbol{y}) \right\rangle \geq \left\|\operatorname{prox}_f(\boldsymbol{x}) - \operatorname{prox}_f(\boldsymbol{y})\right\|^2$$

(b) If $\operatorname{prox}_f(\boldsymbol{x}) = \operatorname{prox}_f(\boldsymbol{y})$, then the inequality is obvious. Assume that $\operatorname{prox}_f(\boldsymbol{x}) \neq \operatorname{prox}_f(\boldsymbol{y})$. Using (a) and the Cauchy-Schwarz inequality, it follows that

$$\begin{aligned}\left\|\operatorname{prox}_f(\boldsymbol{x}) - \operatorname{prox}_f(\boldsymbol{y})\right\|^2 &\leq \langle \operatorname{prox}_h(\boldsymbol{x}) - \operatorname{prox}_h(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \\ &\leq \|\operatorname{prox}_h(\boldsymbol{x}) - \operatorname{prox}_h(\boldsymbol{y})\| \cdot \|\boldsymbol{x} - \boldsymbol{y}\|.\end{aligned}$$

Dividing by $\|\operatorname{prox}_h(\boldsymbol{x}) - \operatorname{prox}_h(\boldsymbol{y})\|$, the desired result is established. □

The following identity given by the Moreau Decomposition theorem holds for the proximal of a function f and its conjugate $f^\star$:

Theorem 10.9 (Moreau Decomposition)**.** *Let $f : E \to (-\infty, \infty]$ be proper closed and convex. Then for any $\boldsymbol{x} \in E$,*

$$\operatorname{prox}_f(\boldsymbol{x}) + \operatorname{prox}_{f^\star}(\boldsymbol{x}) = \boldsymbol{x}$$

Proof. Let $\boldsymbol{x} \in E$ and denote $\boldsymbol{u} = \operatorname{prox}_f(\boldsymbol{x})$. Then by the equivalence between claims (i) and (ii) in the second prox theorem theorem 10.7, it follows that $\boldsymbol{x} - \boldsymbol{u} \in \partial f(\boldsymbol{u})$, which by the conjugate subgradient theorem (theorem 3.1) is equivalent to $\boldsymbol{u} \in \partial f^*(\boldsymbol{x} - \boldsymbol{u})$. Using the second prox theorem again, we conclude that $\boldsymbol{x} - \boldsymbol{u} = \operatorname{prox}_{f^\star}(\boldsymbol{x})$. Therefore

$$\operatorname{prox}_f(\boldsymbol{x}) + \operatorname{prox}_{f^\star}(\boldsymbol{x}) = \boldsymbol{u} + (\boldsymbol{x} - \boldsymbol{u}) = \boldsymbol{x}$$

□

Facts about the proximal operator can be used to guarantee the existence of a well-defined smoothed version of any function f called the **Moreau envelope** which has the same minimizers as f. Given a proper closed convex function $f : \mathbb{E} \to (-\infty, \infty]$ and $\mu > 0$, the Moreau envelope of f is the function

$$M_f^\mu(\boldsymbol{x}) = \min_{\boldsymbol{u} \in \mathbb{E}} \left\{ f(\boldsymbol{u}) + \frac{1}{2\mu}\|\boldsymbol{x} - \boldsymbol{u}\|^2 \right\}$$

The parameter μ is called the smoothing parameter. By the first prox theorem (theorem 10.4), the minimization problem of the Moreau envelope has a unique solution, given by $\operatorname{prox}_{\mu f}(\boldsymbol{x})$.

10.0.2 Proximal Gradient Descent

Consider the following unconstrained problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) = g(\boldsymbol{x}) + h(\boldsymbol{x}) \tag{174}$$

where $g : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable and $h : \mathbb{R}^n \to \mathbb{R}$ is closed, convex, and possibly non-differentiable.

The proximal gradient algorithm is given as follows:

$$\boldsymbol{x}(t) = \text{prox}_{\eta_t h}\left(\boldsymbol{x}(t-1) - \eta_t \nabla g(\boldsymbol{x}(t-1))\right) \qquad \forall t \in \mathbb{N}_+ \tag{175}$$

$$\boldsymbol{x}(0) \in \mathbb{R}^n \tag{176}$$

where $\forall t \in \mathbb{N}_+$, $\eta_t > 0$ is a variable learning rate.

Note that we can re-write the proximal gradient update step as:

$$\boldsymbol{x}(t) = \arg\min_{\boldsymbol{u}\in\mathbb{R}^n} \left\{ h(\boldsymbol{u}) + \frac{1}{2\eta_t} \left\|\boldsymbol{u} - \boldsymbol{x}(t-1) + \eta_t \nabla_{\boldsymbol{x}} g(\boldsymbol{x}(t-1))\right\|_2^2 \right\} \tag{177}$$

$$= \arg\min_{\boldsymbol{u}\in\mathbb{R}^n} \left\{ h(\boldsymbol{u}) + g(\boldsymbol{x}(t-1)) + \nabla_{\boldsymbol{x}} g(\boldsymbol{x}(t-1))\left(\boldsymbol{u} - \boldsymbol{x}(t-1)\right) + \frac{1}{2\eta_t}\left\|\boldsymbol{u} - \boldsymbol{x}(t-1)\right\|_2^2 \right\} \tag{178}$$

That is, $\boldsymbol{x}(t)$ minimizes $h(\boldsymbol{u})$ plus a simple quadratic model of $g(\boldsymbol{u})$ around $\boldsymbol{x}$.

Note that when $h = 0$ we recover gradient descentm and when $h = \mathrm{I}_C$, i.e., the indicator function of some set $C \subset \mathbb{R}^n$, we recover projected gradient descent.

The following convergence result holds for the proximal gradient method:

Theorem 10.10. *Consider the unconstrained optimization problem in eq.* (174)*. Suppose that $\nabla_{\boldsymbol{x}} g(\boldsymbol{x})$ is Lipschitz continuous with constant $L > 0$, for all $t \in \mathbb{N}_+$, $\eta_t = \frac{1}{L}$, and that the minimum value of $\min_{\boldsymbol{x}\in\mathbb{R}} f(\boldsymbol{x})$ is attained at $\boldsymbol{x}^\star$ (not necessarily unique), then the following convergence bound holds:*

$$f(\boldsymbol{x}(t)) - f(\boldsymbol{x}^\star) \leq O\left(\frac{1}{t}\right) \tag{179}$$

References

[1] Richard Cole et al. *Convex Program Duality, Fisher Markets, and Nash Social Welfare*. 2016. arXiv: 1609.06654 [cs.GT].

[2] John M. Danskin. "The Theory of Max-Min, with Applications". In: *SIAM Journal on Applied Mathematics* 14.4 (1966), pp. 641–664. ISSN: 00361399. URL: http://www.jstor.org/stable/2946123.

[3] Paul Milgrom and Ilya Segal. "Envelope theorems for arbitrary choice sets". In: *Econometrica* 70.2 (2002), pp. 583–601.