Spring 6-1-2017

# Generalized Differential Calculus and Applications to Optimization

R. Blake Rector
*Portland State University*

# Generalized Differential Calculus and Applications to Optimization

by

Robert Blake Hayden Rector

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
in
Mathematical Sciences

Dissertation Committee:
Mau Nam Nguyen, Chair
Robert Bass
Gerardo Lafferriere
Bin Jiang
Tugrul Daim

Portland State University
2017

**Abstract**

This thesis contains contributions in three areas: the theory of generalized calculus, numerical algorithms for operations research, and applications of optimization to problems in modern electric power systems. A geometric approach is used to advance the theory and tools used for studying generalized notions of derivatives for nonsmooth functions. These advances specifically pertain to methods for calculating subdifferentials and to expanding our understanding of a certain notion of derivative of set-valued maps, called the coderivative, in infinite dimensions. A strong understanding of the subdifferential is essential for numerical optimization algorithms, which are developed and applied to nonsmooth problems in operations research, including nonconvex problems. Finally, an optimization framework is applied to solve a problem in electric power systems involving a smart solar inverter and battery storage system providing energy and ancillary services to the grid.

*This thesis is dedicated to my grandfather, Dr. Robert W. Rector, who inspired me to—among other things—study mathematics.*

## Table of Contents

## List of Tables

## List of Figures

# 1

# Introduction

*"Waste not, want not" goes the old saying. If you use a resource wisely, you will never be in need. In this phrase lie the seeds of optimization...*

This thesis concerns generalized differential calculus and applications of optimization to location problems and electric power systems. Let us begin with a brief discussion of the key terms in this sentence. *Generalized differential calculus* is a generalization of classical calculus. In particular, this includes the study of generalized notions of derivatives for nonsmooth functions and set-valued mappings. These topics generally fall under the subject of variational or nonsmooth analysis. The term convex analysis is used when such a study is restricted to functions that are convex. Optimization problems ask for the best solution from a given set of feasible solutions; such problems are ubiquitous in applied science, business, engineering, economics, social sciences and everyday life. *Optimization* is the process by which solutions to optimization problems are found.

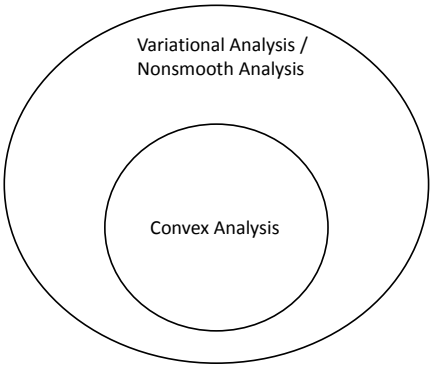
Recently, nonsmooth analysis and optimization have become increasingly important for applications to many new fields such as computational statistics, machine learning, and sparse optimization. The work in this dissertation provides a more complete picture of generalized differentiation and develops nonsmooth optimization methods to solve facility location problems. The optimization techniques and methods developed significantly contribute to the field of variational analysis and nonsmooth optimization as well as their applications. In addition, the applications of optimization to

electric power systems developed here will play an important role in the evolution of modern energy markets and responsive electric grids. Collectively, this work develops theory, numerical methods, and a practical solution, all related to optimization.

**1.0.1. Convex Analysis, Nonsmooth Analysis, and Variational Analysis.** Inspired by Fenchel [**18**], in the 1960's Rockafellar [**66**] and Moreau [**45**] independently laid the groundwork for the field of convex analysis with their development of the idea of the *subdifferential*. The subdifferential generalizes the idea of the derivative in classical calculus from differentiable functions to functions that are convex but not necessarily differentiable. The subdifferential of a convex function is a *set*, rather than a single value as in the classical case. Related to the subdifferential, the geometric idea of the *normal cone* to a convex set goes back to Minkowski [**37**].

Before Rockafellar and Moreau, generalized differentiation ideas had been discussed in mathematics and applied sciences, for example by Saks [**70**] and Sobelev [**75**]. However, these generalized derivatives "ignore sets of measure zero" and hence are not useful in optimization theory, where the behavior of functions at individual points is of critical importance. Convex analysis contains the theoretical framework for the numerical methods of convex optimization. The presence of convexity makes it possible not only to comprehensively investigate qualitative properties of optimal solutions and derive necessary and sufficient conditions for optimality but also to develop effective numerical algorithms for solving convex optimization problems, even with nondifferentiable objective functions. Convex analysis and convex optimization have an increasing impact on many areas of mathematics and applications including automatic control systems, estimation and signal processing, communications and networks, electronic circuit design, data analysis and modeling, statistics, machine learning, economics and finance.

The beauty and applications of convex analysis motivated the search for a new theory to deal with broader classes of functions and sets where convexity is not assumed. This search has been the inspiration for the developments of variational analysis, also known as nonsmooth analysis, initiated in the early 1970's. Variational analysis has now become a well-developed field with many applications, especially to optimization theory; see the cornerstone references [**9, 38, 64, 65**] for more history and recent developments of the field.




**1.0.2. Optimization.** Optimization is the process by which solutions to optimization problems are found. In mathematics, an *optimization problem* includes three things: an objective function, a constraint set, and a desired outcome. For example, if $f\colon \mathbb{R}^n \to \mathbb{R}$, then an optimization problem can be stated as:

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

$$\text{subject to } x \in C \subset \mathbb{R}^n.$$

This means: find the function input $\bar{x} \in C$ such that $f(\bar{x}) \leq f(x)$ for all $x \in C$. Such an $\bar{x}$ would be called a *solution* to the optimization problem. If $f$ is a convex function and $C$ is a convex set, then the optimization problem is a *convex optimization problem*. Optimization problems may be further delineated by the nature of the objective function as *differentiable* (*smooth*) or *non-differentiable* (*nonsmooth*). The latter types of objective functions are of particular interest in convex optimization.

In practice, optimization problems are often so complicated that they cannot be solved exactly with analytic methods. In these cases, a computer program can be written

to employ an optimization algorithm to find an approximate solution. *Optimization algorithms* are typically iterative procedures that start with an initial guess $x_0$ for the solution and update that guess based on the logic provided by the algorithm. In this way, a sequence of function inputs $x_0, x_1, x_2, ...$ is derived that converges to the solution $\bar{x}$. The optimization algorithms developed in this thesis utilize the theory from convex, nonsmooth or variational analysis.

The field of *operations research* applies optimization to solve practical problems from industry or logistics. *Location problems* are among the oldest class of problems studied in operations research, and are discussed in detail in this dissertation.

**1.0.3. Electric Power Systems.** The generation, transmission, distribution and consumption of electric power has shown itself to be an essential part of modern life. Despite some incremental improvements, the electric power system in the year 2000 was essentially the same as the electric power system originally built in the early 1900s. The operation of this historical legacy system was straight forward: build large enough generation, transmission and distribution facilities so that consumer peak demand can be met by increasing or decreasing generation as needed. Consumer demand for electricity varies throughout the day; it is not uncommon for peak demand to be twice that of minimum demand over a 24 hour period.

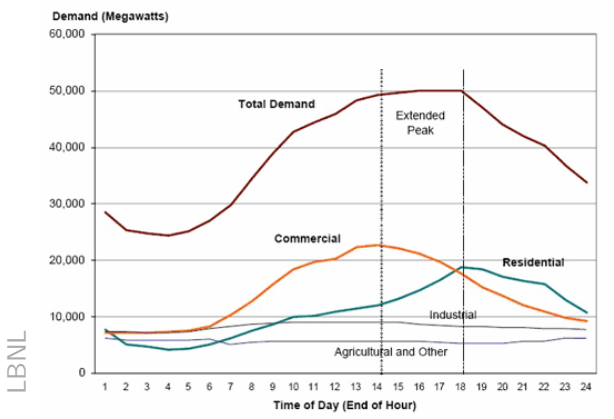



FIGURE 1.1. Electricity demand in California on a hot day.

In recent decades, two major advances have caused system planners and operators to rethink this legacy electric power system. These advances are (1) the significant (and rapidly increasing) presence of "variable generation" such as wind and solar, and (2) the advent of distributed energy resources, including rooftop solar, storage, and "controllable loads", that is, devices that have some flexibility in how much and when they consume power, such as smart appliances, thermostats, heating and cooling systems, other types of demand response, and electric vehicles. This new paradigm provides a natural opportunity for optimization, including the minimization of system cost, the maximization of revenue for individual participants, or the minimization of emissions.

**1.0.4. Overview of Research.** The research presented in this thesis is generally organized into Chapters 2, 3, and 4 as follows:

- Chapter 2: Use the *optimal value function* and *coderivatives* to establish subdifferential formulas in locally convex topological vector spaces. Develop formulas for coderivatives in infinite dimensions.
- Chapter 3: Apply *nonsmooth optimization* techniques to location problems:
  - Apply *Nesterov's smoothing technique* and *accelerated gradient method* to facility location problems.
  - Apply *DC programming* techniques to problems of multifacility location and clustering.
- Chapter 4: Apply optimization techniques to a problem in electric power systems: develop an *optimal control scheme* for a smart solar inverter and battery-storage system operating in a transactive control setting.

The remainder of the present chapter is devoted to providing an overview of some basic tools of convex analysis and optimization.

## 1.1. Basic Tools of Convex Analysis and Optimization

This section contains definitions and preliminary facts that will be useful later in the document. Applications of convex analysis typically take place in the finite dimensional space $\mathbb{R}^n$. For greater generality (and, we feel, greater clarity), we provide the basic definitions of convex analysis in the more abstract setting of locally convex topological vector spaces. We begin with definitions and facts from convex analysis and proceed to some basic algorithms for convex and non-convex optimization.

**1.1.1. Definitions.** In these definitions, we consider $X$ and $Y$ to be Hausdorff locally convex topological vector spaces over $\mathbb{R}$. A *locally convex topological vector space* is a topological vector space where the topology can be generated by a basis consisting of translations of balanced, absorbent, convex sets. We use $X^*$ and $Y^*$ to denote the topological dual of $X$ and $Y$ respectively: the set of all continuous linear maps into $\mathbb{R}$. We equip $Y$ with a partial ordering as follows. Let $Y_+$ be a cone in $Y$. This means that $\alpha y \in Y_+$ for all $y \in Y_+$ and all $\alpha \geq 0$. Then, for $y, z \in Y$, we say

$$y \geq z \iff z \leq y \iff y - z \in Y_+.$$

Whenever $Y = \mathbb{R}$ we assume the typical ordering $Y_+ = [0, +\infty)$. When $Y = \overline{\mathbb{R}} := (-\infty, \infty]$, we assume $Y_+ = [0, \infty]$. (The equal sign proceeded by the colon := means that we are defining the object on the left as the object on the right.) We say a map $\phi \colon Y \to \mathbb{R}$ is *non-decreasing* if $y \leq z$ implies $f(y) \leq f(z)$.

A map $f \colon X \to Y$ is *convex* if $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$ for all $x_1, x_2 \in X$ and all $\lambda \in (0, 1)$. When $Y = \overline{\mathbb{R}}$, we define the *domain* of $f$ to be the set $\operatorname{dom} f := \{x \in X \mid f(x) < \infty\}$. The *epigraph* of $f$ is the set $\operatorname{epi} f := \{(x, y) \in X \times Y \mid y \geq f(x)\}$.

A set $\Omega \in X$ is *convex* if for all $x_1, x_2 \in \Omega$ and all $\lambda \in (0,1)$ we have $\lambda x_1 + (1-\lambda)x_2 \in \Omega$. An important geometric property of convex functions is that $f$ is convex if and only if its epigraph is convex.

We use the double arrows $\rightrightarrows$ to denote set-valued maps. This means, for example, the map $G\colon X \rightrightarrows Y$ has $G(x) \subseteq Y$ for each $x \in X$. We allow the possibility that $G(x)$ is empty. The *domain* of a set-valued map is the set of all inputs that produce nonempty output, which we denote $\operatorname{dom} G := \{x \in X \mid G(x) \neq \emptyset\}$. The *graph* of a set-valued map $G\colon X \rightrightarrows Y$ is $\operatorname{gph} G := \{(x,y) \in X \times Y \mid y \in G(x)\}$, which is the same as $\{(x,y) \in X \times Y \mid x \in \operatorname{dom} G, y \in G(x)\}$. If the set $\operatorname{gph} G$ is convex in $X \times Y$ then we say that $G$ is a *convex set-valued map*. The *subdifferential* of a convex function $f\colon X \to \overline{\mathbb{R}}$ at $\bar{x} \in \operatorname{dom} f$ is the set

$$\partial f(\bar{x}) := \{x^* \in X^* \mid f(x) \geq f(\bar{x}) + \langle x^*, x - \bar{x}\rangle \text{ for all } x \in X\}.$$

Elements of this set are called the *subgradients* of $f$ at $\bar{x}$. In this way, the operator $\partial f$ is a set-valued map $\partial f\colon X \rightrightarrows X^*$. (The overline $\bar{x}$ is often used to denote a point of interest, not necessarily being the solution to an optimization problem.)

Another important set-valued map into $X^*$ is that of the normal cone. Let $\Omega \subset X$ be convex and $\bar{x} \in \Omega$. Then the *normal cone to* $\Omega$ *at* $\bar{x}$ is defined by

$$N(\bar{x}; \Omega) := \{x^* \in X^* \mid \langle x^*, x - \bar{x}\rangle \leq 0 \text{ for all } x \in \Omega\}.$$

The following proposition provides a useful representation of the subdifferential via the normal cone and epigraph is easily proven.

**Proposition 1.1.1.** *Let $f\colon X \to \overline{\mathbb{R}}$ be convex and let $\bar{x} \in \operatorname{dom} f$. Then we have*

$$\partial f(\bar{x}) = \{x^* \in X^* \mid (x^*, -1) \in N\left((\bar{x}, f(\bar{x})); \operatorname{epi} f\right)\}.$$

**Proof.** For notation, set $W = \{x^* \in X^* \mid (x^*, -1) \in N\left((\bar{x}, f(\bar{x})); \text{epi} f\right)\}$. Let $x^* \in \partial f(\bar{x})$ and pick any $(x, \lambda) \in \text{epi} f$. Then we have

$$\begin{aligned}\langle (x^*, -1), (x, \lambda) - (\bar{x}, f(\bar{x}))\rangle &= \langle x^*, x - \bar{x}\rangle - (\lambda - f(\bar{x})) \\ &\leq \langle x^*, x - \bar{x}\rangle - (f(x) - f(\bar{x})) \leq 0,\end{aligned}$$

where the last inequality holds because $x^* \in \partial f(\bar{x})$, so $x^* \in W$.

For the reverse containment, let $u^* \in W$. Since $\langle (u^*, -1), (x, \lambda) - (\bar{x}, f(\bar{x}))\rangle \leq 0$ for all $(x, \lambda) \in \text{epi} f$, we have

$$\langle x^*, x - \bar{x}\rangle - (f(x) - f(\bar{x})) \leq 0$$

for all $x \in \text{dom}\, f$, so $u^* \in \partial f(\bar{x})$. This completes the proof. □

Two more important definitions for our geometric approach are those of the indicator function and the support function. Let $\Omega \subset X$. The *indicator function* $\delta_\Omega \colon X \to \overline{\mathbb{R}}$ is defined by setting $\delta_\Omega(x)$ equal to 0 if $x \in \Omega$ and to $\infty$ if $x \notin \Omega$. It follows directly from the definition that $\partial \delta_\Omega(\bar{x}) = N(\bar{x}; \Omega)$ whenever $\Omega$ is convex. The *support function* $\sigma_\Omega \colon X^* \to \overline{\mathbb{R}}$ is defined as $\sigma_\Omega(x^*) := \sup\{\langle x^*, x\rangle \mid x \in \Omega\}$. If $\bar{x} \in \Omega$, it follows from the definition that $x^* \in N(\bar{x}; \Omega) \iff \sigma_\Omega(x^*) = \langle x^*, \bar{x}\rangle$.

**1.1.2. Optimal Value Function.** The following function plays an important role in obtaining the subdifferential formulas in our generalized calculus setting (see Chapter 2).

**Definition 1.1.2.** *Given* $G \colon X \rightrightarrows Y$ *and* $\phi \colon X \times Y \to \mathbb{R}$*, we denote the* optimal value function $\mu \colon X \to \overline{\mathbb{R}}$ *by*

$$\mu(x) := \inf\{\phi(x, y) \mid y \in G(x)\}.$$

It can be helpful to think of $\phi(x, \cdot)$ as the "objective function" and $G(x)$ as the "constraint set". We have as a *standing assumption* in this document that $\mu(x) > -\infty$ for all $x \in X$.

**Definition 1.1.3.** *The* solution set *of* $\mu$ *at* $\bar{x}$ *is* $M(\bar{x}) := \{y \in Y \mid \mu(\bar{x}) = \phi(\bar{x}, y)\}$.

**1.1.3. Optimization Algorithms.** Here we introduce some basic convex optimization algorithms. These are provided for reference.

1.1.3.1. *Subgradient Method.* The subgradient method is a standard method for solving nonsmooth optimization problems. It is outlined as follows.

Input $x_0 \in \mathbb{R}^n$. Then update

$$x_{k+1} := x_k - t_k w_k \qquad \text{where } \; w_k \in \partial f(x_k),$$

and where $t_k > 0$ is a pre-determined step size. When $f$ is convex, the subgradient method converges for all initial values $x_0$, as long as the sequence of step sizes $(t_k)$ is chosen so that $\sum_{k=1}^{\infty} t_k = \infty$ and $\sum_{k=1}^{\infty} t_k^2 < \infty$. In general, the subgradient method is known to have a convergence rate of order $\left(\frac{1}{\sqrt{k}}\right)$.

1.1.3.2. *Stochastic Subgradient Method.* The stochastic subgradient method is a variation on the subgradient method. It is particularly well-suited for problems where the objective function is a sum of convex functions.

**Definition 1.1.4.** *Let* $f \colon \mathbb{R}^n \to \mathbb{R}$ *be a convex function. A vector valued random variable* $\tilde{V} \in \mathbb{R}^n$ *is called a* NOISY UNBIASED SUBGRADIENT *of* $f$ *at* $\bar{x}$ *if the expected value* $E(\tilde{V}) \in \partial f(\bar{x})$. *That is*

$$\left\langle E(\tilde{V}), x - \bar{x} \right\rangle \leq f(x) - f(\bar{x}) \;\; \textit{for all} \;\; x \in \mathbb{R}^n.$$

The stochastic subgradient method is outlined as follows.

Input $x_0 \in \mathbb{R}^n$. Then update

$$x_{k+1} := x_k - t_k \tilde{v_k} \qquad \text{where} \quad E(\tilde{v_k}) \in \partial f(x_k)$$

and $t_k > 0$ is a pre-determined step size.

Convex optimization is a growing field with many other available algorithms such as smoothing methods, proximal point methods, bundle methods, majorization minimization methods, and more. It has a wide range of applications including machine learning and computational statistics, optimal control, neural network training, data mining, engineering, and economics.

1.1.3.3. *Optimization Beyond Convexity.* We may often be presented with an optimization problem where the objective function is not convex. No complete theory exists for finding solutions to these types of optimization problems, but certain results may be obtained using the tools of convex analysis. A main focus of this thesis is to develop optimization algorithms for optimization problems in which the objective function is not necessarily convex. We present below one such class of non-convex optimization problems, and an algorithm for find their solutions.

**DC Programming:**

DC programming stands for "Difference of Convex" programming. It offers a method to solve the following types of optimization problems.

Let $f(x) = g(x) - h(x)$, where $g\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ and $h\colon \mathbb{R}^n \to \mathbb{R}$ are convex functions. Then $f$ is called a *DC function* since it is the difference of convex functions.

The problem

$$\text{minimize } f(x) = g(x) - h(x),\ x \in \mathbb{R}^n, \tag{1.1.1}$$

is a *DC optimization problem.*

The framework for DC programming was constructed by Tao and An in their papers [**79, 80**] in the late 1990's; its essential elements are presented below.

**The DC Programming Algorithm (DCA):**

One of the key components in the DCA is the *Fenchel conjugate* $\varphi^*$ of a convex function $\varphi\colon \mathbb{R}^n \to (-\infty, +\infty]$, defined by

$$\varphi^*(v) := \sup\{\langle v, x\rangle - \varphi(x) \mid x \in \mathbb{R}^n\}.$$

If $\varphi$ is proper, i.e., $\operatorname{dom}\varphi \neq \emptyset$, then $\varphi^*\colon \mathbb{R}^n \to (-\infty, +\infty]$ is also a convex function. Some other important properties of the Fenchel conjugate of a convex function are given in the following proposition.

**Proposition 1.1.5.** *Let $\varphi\colon \mathbb{R}^n \to (-\infty, +\infty]$ be a convex function.*

(i) *Given any $x \in \operatorname{dom}\varphi$, one has that $y \in \partial\varphi(x)$ if and only if*

$$\varphi(x) + \varphi^*(y) = \langle x, y\rangle .$$

(ii) *If $\varphi$ is proper and lower semicontinuous, then for any $x \in \operatorname{dom}\varphi$ one has that $y \in \partial\varphi(x)$ if and only if $x \in \partial\varphi^*(y)$.*

(iii) *If $\varphi$ is proper and lower semicontinuous, then $(\varphi^*)^* = \varphi$.*

The DC optimization problem (1.1.1) possesses useful optimality conditions, one of which is given in the next proposition.

**Proposition 1.1.6.** *If $\bar{x} \in \operatorname{dom} f$ is a local minimizer of* (1.1.1)*, then*

$$\partial h(\bar{x}) \subset \partial g(\bar{x}). \tag{1.1.2}$$

Any point $\bar{x} \in \operatorname{dom} f$ satisfying condition (1.1.2) is called a *stationary point* of (1.1.1). One says that $\bar{x}$ a *critical point* of (1.1.1) if $\partial g(\bar{x}) \cap \partial h(\bar{x}) \neq \emptyset$. It is obvious that every stationary point $\bar{x}$ is a critical point, but the converse is not true in general.

The *Toland dual* of (1.1.1) is the problem

$$\text{minimize } h^*(y) - g^*(y),\ y \in \mathbb{R}^n. \tag{1.1.3}$$

Using the convention $(+\infty) - (+\infty) = +\infty$, we have the following relationship between a DC optimization problem and its Toland dual.

**Proposition 1.1.7.** *Considering the function $f = g - h$ given in* (1.1.1)*, one has*

$$\inf\{g(x) - h(x) \mid x \in \mathbb{R}^n\} = \inf\{h^*(y) - g^*(y) \mid y \in \mathbb{R}^n\}.$$

The DCA is based on Toland's duality theorem and Proposition 1.1.5. The idea of the DCA is to construct two sequences $\{x_k\}$ and $\{y_k\}$ such that the real sequences $g(x_k) - h(x_k)$ and $h^*(y_k) - g^*(y_k)$ are both monotone decreasing and every cluster point $\bar{x}$ of $\{x_k\}$ is a critical point of problem (1.1.1). Similarly, every cluster point $\bar{y}$ of $\{y_k\}$ is a critical point of (1.1.3), i.e., $\partial g^*(\bar{y}) \cap \partial h^*(\bar{y}) \neq \emptyset$.

The DCA is summarized as follows:

**Step 1.** Choose $x_0 \in \operatorname{dom} g$.

**Step 2.** For $k \geq 0$, use $x_k$ to find $y_k \in \partial h(x_k)$.

**Step 3.** Use $y_k$ to find $x_{k+1} \in \partial g^*(y_k)$.

**Step 4.** Increase $k$ by 1 and go back to **Step 2**.

In the case where we cannot find $y_k$ or $x_{k+1}$ exactly, we can find them approximately by solving a convex optimization problem. This idea is explored in Chapter 3.

# 2

## Generalized Differential Calculus

The term *generalized differential calculus* refers to calculus rules and generalized derivatives developed for nonsmooth functions and set-valued mappings. These functions and mappings arise naturally in many applications. The study of generalized differential calculus provides the mathematical foundation for nonsmooth optimization.

In this chapter we present results from a geometric approach to convex analysis and generalized differential calculus. The term *geometric approach* was coined by B.S. Mordukhovich. It builds on the concepts of normal cone, optimal value function and coderivative to provide an easy way to prove new and existing generalized calculus results. These new proofs become so easy, in fact, that they may now be taught to beginning graduate or even undergraduate students, a challenge previously typically avoided even for advanced graduate courses. As in the previous chapter, we assume $X$ and $Y$ to be Hausdorff locally convex topological vector spaces over $\mathbb{R}$ unless otherwise stated.

This chapter contains two sections. In section 2.1, we present a fundamental result relating the subdifferential of the optimal value function to the coderivative. While this result itself is not new, we apply it in new ways to derive formulas for the subdifferentials of various convex functions. Having thus seen the importance of coderivatives, in section 2.2 we derive new formulas for coderivatives of various set-valued mappings.

## 2.1. A Geometric Approach to Subdifferential Calculus

In this section we prove subdifferential calculus rules via the optimal value function and coderivatives. The subdifferential of the optimal value function $\mu$ is related to coderivatives by way of a fundamental theorem (Theorem 2.1.8), which we sometimes refer to as "the fundamental theorem". The proof of the fundamental theorem uses the subdifferential sum rule, which is proven using the normal cone intersection rule. The normal cone intersection rule is proven in a new way, using support functions and the convex extremal principal. We first state a definition and then proceed to the theorems.

**Definition 2.1.1.** *We say that two nonempty sets $\Omega_1, \Omega_2 \in X$ form an* extremal system *if for any neighborhood $V$ of the origin there exists a vector $a \in V$ such that*

$$(\Omega_1 + a) \cap \Omega_2 = \emptyset. \tag{2.1.4}$$

The next theorem is part of what is known as the *convex extremal principal*, a consequence of the classical separation principle. For its proof and further discussion, see the paper by Mordukhovich, Nam, Rector and Tran [**39**].

**Theorem 2.1.2.** *Let $\Omega_1, \Omega_2 \subset X$ be nonempty and convex. If $\Omega_1$ and $\Omega_2$ form an extremal system and $\operatorname{int}(\Omega_1 - \Omega_2) \neq \emptyset$, then $\Omega_1$ and $\Omega_2$ can be separated, i.e., there exists some nonzero $x^* \in X^*$ such that*

$$\sup_{x \in \Omega_1} \langle x^*, x \rangle \leq \inf_{x \in \Omega_2} \langle x^*, x \rangle . \tag{2.1.5}$$

We apply this result in the proof of the following theorem regarding support functions to intersections of sets. As mentioned, we use this result to prove the normal cone intersection rule, which in turn is used to prove the subdifferential sum rule.

**Theorem 2.1.3.** *Let $\Omega_1, \Omega_2 \subset X$ be nonempty and convex. Suppose that either $(\operatorname{int}\Omega_2) \cap \Omega_1 \neq \emptyset$ or $(\operatorname{int}\Omega_1) \cap \Omega_2 \neq \emptyset$. Then for any $x^* \in \operatorname{dom}(\sigma_{\Omega_1 \cap \Omega_2})$ there are $x_1^*, x_2^* \in X^*$ such that $x^* = x_1^* + x_2^*$ and*

$$\sigma_{\Omega_1 \cap \Omega_2}(x^*) = \sigma_{\Omega_1}(x_1^*) + \sigma_{\Omega_2}(x_2^*). \tag{2.1.6}$$

**Proof.** First we note that for any $x_1^*, x_2^* \in X^*$ with $x_1^* + x_2^* = x^*$, we have

$$\langle x_1^*, x\rangle + \langle x_2^*, x\rangle \leq \sigma_{\Omega_1}(x_1^*) + \sigma_{\Omega_2}(x_2^*)$$

for any $x \in \Omega_1 \cap \Omega_2$. So the "$\leq$" inequality is established in (2.1.6).

To prove the other direction, we will apply the convex extremal principal Theorem 2.1.2 to obtain the elements $x_1^*, x_2^* \in X^*$ required to prove the "$\geq$" inequality. First, we set up the application of this theorem. Let $x^* \in \operatorname{dom}(\sigma_{\Omega_1 \cap \Omega_2})$ and set $\alpha = \sigma_{\Omega_1 \cap \Omega_2}(x^*)$, so

$$\langle x^*, x\rangle - \alpha \leq 0$$

for all $x \in \Omega_1 \cap \Omega_2$. Next, we define the two sets to which we will ultimately apply Theorem 2.1.2. Let

$$\Theta_1 = \Omega_1 \times [0, \infty),$$

$$\Theta_2 = \{(x, \lambda) \in X \times \mathbb{R} \mid x \in \Omega_2,\ \lambda \leq \langle x^*, x\rangle - \alpha\}.$$

We can see from the construction of $\Theta_1$ and $\Theta_2$ that

$$(\Theta_1 + (0, \gamma)) \cap \Theta_2 = \emptyset \text{ for any } \gamma > 0,$$

so, $\Theta_1$ and $\Theta_2$ form an extremal system. To apply Theorem 2.1.2, we need to show that $\operatorname{int}(\Theta_1 - \Theta_2) \neq \emptyset$. We will use the assumption that $(\operatorname{int}\Omega_2) \cap \Omega_1 \neq \emptyset$.

The interior of $\Theta_2$ is expressed as

$$\operatorname{int}(\Theta_2) = \{(x, \lambda) \in X \times \mathbb{R} \mid x \in \operatorname{int}(\Omega_2),\ \lambda < \langle x^*, x\rangle - \alpha\}.$$

Let $x' \in (\operatorname{int}\Omega_2) \cap \Omega_1$. We know $\langle x^*, x'\rangle - \alpha \leq 0$. Pick any $\lambda'$ such that $\lambda' < \langle x^*, x'\rangle - \alpha$. Then $(x', \lambda') \in \operatorname{int}(\Theta_2)$. Since $(x', 0) \in \Theta_1$, we have $(0, -\lambda') \in \Theta_1 - \operatorname{int}(\Theta_2)$. Since $\Theta_1 - \operatorname{int}(\Theta_2) \subset \operatorname{int}(\Theta_1 - \Theta_2)$, we can conclude that $\operatorname{int}(\Theta_1 - \Theta_2) \neq \emptyset$. Therefore we can apply Theorem 2.1.2 to get nonzero $(z^*, \beta) \in X^* \times \mathbb{R}$ that separates $\Theta_1$ and $\Theta_2$:

$$\langle z^*, x_1\rangle + \beta\lambda_1 \leq \langle z^*, x_2\rangle + \beta\lambda_2 \text{ for all } (x_1, \lambda_1) \in \Theta_1,\ (x_2, \lambda_2) \in \Theta_2. \tag{2.1.7}$$

It follows from the structure of $\Theta_1$ that $\beta \leq 0$. If $\beta = 0$, then we would have

$$\langle z^*, x_1\rangle \leq \langle z^*, x_2\rangle \text{ for all } x_1 \in \Omega_1,\ x_2 \in \Omega_2,$$

which yields $z^* = 0$ because $0 \in \operatorname{int}(\Omega_1 - \Omega_2)$, a contradiction. Thus $\beta < 0$.

Proceeding, take $(x, 0) \in \Theta_1$ and $(y, \langle x^*, y\rangle - \alpha) \in \Theta_2$. Then, as per (2.1.7) we get

$$\langle z^*, x\rangle \leq \langle z^*, y\rangle + \beta(\langle x^*, y\rangle - \alpha).$$

It follows that

$$\alpha \geq \left\langle \frac{z^*}{\beta} + x^*, y \right\rangle + \left\langle \frac{-z^*}{\beta}, x \right\rangle \text{ for all } x \in \Omega_1,\ y \in \Omega_2.$$

Setting $x_1^* = \frac{z^*}{\beta} + x^*$ and $x_2^* = \frac{-z^*}{\beta}$, we get the desired conclusion

$$\sigma_{\Omega_1 \cap \Omega_2}(x^*) \geq \sigma_{\Omega_1}(x_1^*) + \sigma_{\Omega_2}(x_2^*).$$

This completes the proof. □

Now we are ready to prove the normal cone intersection rule. The theorem after it is known as the subdifferential sum rule.

**Theorem 2.1.4.** *Let $\Omega_1, \Omega_2 \subseteq X$ be convex with $\operatorname{int}(\Omega_1) \cap \Omega_2 \neq \emptyset$ or $\Omega_1 \cap \operatorname{int}(\Omega_2) \neq \emptyset$. Then for any $\bar{x} \in \Omega_1 \cap \Omega_2$ we have*

$$N(\bar{x}; \Omega_1 \cap \Omega_2) = N(\bar{x}; \Omega_1) + N(\bar{x}; \Omega_2).$$

**Proof.** First we prove the set inclusion "$\subset$". Let $x^* \in N(\bar{x}; \Omega_1 \cap \Omega_2)$, so $\sigma_{\Omega_1 \cap \Omega_2}(x^*) = \langle x^*, \bar{x} \rangle$. By Theorem 2.1.3, there exists $x_1^*, x_2^* \in X^*$ with $x^* = x_1^* + x_2^*$ such that

$$\sigma_{\Omega_1 \cap \Omega_2}(x^*) = \sigma_{\Omega_1}(x_1^*) + \sigma_{\Omega_2}(x_2^*).$$

Since $\langle x^*, \bar{x} \rangle = \langle x_1^*, \bar{x} \rangle + \langle x_2^*, \bar{x} \rangle$, and since $\sigma_{\Omega_1}(x_1^*) \geq \langle x_1^*, \bar{x} \rangle$ and $\sigma_{\Omega_2}(x_2^*) \geq \langle x_2^*, \bar{x} \rangle$, this implies that $\sigma_{\Omega_1}(x_1^*) = \langle x_1^*, \bar{x} \rangle$ and $\sigma_{\Omega_2}(x_2^*) = \langle x_2^*, \bar{x} \rangle$. Thus we have $x_1^* \in N(\bar{x}; \Omega_1)$ and $x_2^* \in N(\bar{x}; \Omega_2)$, so

$$x^* \in N(\bar{x}; \Omega_1) + N(\bar{x}; \Omega_2),$$

and we have the desired set inclusion.

The opposite set inclusion is straightforward from the definition: let $u_1^* \in N(\bar{x}; \Omega_1)$ and $u_2^* \in N(\bar{x}; \Omega_2)$. Then $\langle u_1^*, u - \bar{x} \rangle + \langle u_2^*, u - \bar{x} \rangle \leq 0$ for all $u \in \Omega_1 \cap \Omega_2$. So $u_1^* + u_2^* \in N(\bar{x}; \Omega_1 \cap \Omega_2)$ and the proof is complete. □

**Theorem 2.1.5.** *Let $f, g \colon x \to (-\infty, \infty]$ be convex. Assume that $f$ or $g$ are continuous at some point in $\operatorname{dom} f \cap \operatorname{dom} g$. Then we have*

$$\partial(f + g)(\bar{x}) = \partial f(\bar{x}) + \partial g(\bar{x}) \tag{2.1.8}$$

*for all $\bar{x} \in \operatorname{dom}(f) \cap \operatorname{dom}(g)$.*

**Proof.** Let $\bar{x} \in \operatorname{dom} f \cap \operatorname{dom} g$ be fixed for the entire proof. Since the inclusion "$\subset$" in (2.1.8) can be easily checked by the definition, we now concentrate on proving the opposite inclusion. Pick any $x^* \in \partial(f+g)(\bar{x})$. We will show how the geometric results of Theorem 2.1.4 can be used in verifying $x^* \in \partial f(\bar{x}) + \partial g(\bar{x})$. Having

$$\langle x^*, x - \bar{x} \rangle \leq (f+g)(x) - (f+g)(\bar{x}),$$

define the following convex subsets of $X \times \mathbb{R} \times \mathbb{R}$:

$$\Omega_1 := \big\{(x, \lambda_1, \lambda_2) \mid \lambda_1 \geq f(x)\big\},$$
$$\Omega_2 := \big\{(x, \lambda_1, \lambda_2) \mid \lambda_2 \geq g(x)\big\}.$$

It follows from the definition that $(x^*, -1, -1) \in N((\bar{x}, f(\bar{x}), g(\bar{x})); \Omega_1 \cap \Omega_2)$. The fact that $f$ or $g$ is continuous at a point in $\operatorname{dom} f \cap \operatorname{dom} g$ means that $\operatorname{int}(\Omega_1) \cap \Omega_2 \neq \emptyset$ or $\Omega_1 \cap \tau(\Omega_2) \neq \emptyset$, so we can apply Theorem 2.1.4 to get

$$(x^*, -1, -1) \in N\big((\bar{x}, f(\bar{x}), g(\bar{x})); \Omega_1\big) + N\big((\bar{x}, f(\bar{x}), g(\bar{x})); \Omega_2\big), \tag{2.1.9}$$

which tells us therefore that

$$(x^*, -1, -1) = (x_1^*, -\lambda_1, -\lambda_2) + (x_2^*, -\gamma_1, -\gamma_2)$$

with $(x_1^*, -\lambda_1, -\lambda_2) \in N((\bar{x}, f(\bar{x}), g(\bar{x})); \Omega_1)$ and $(x_2^*, -\gamma_1, -\gamma_2) \in N((\bar{x}, f(\bar{x}), g(\bar{x})); \Omega_2)$. By the construction of $\Omega_1$ and $\Omega_2$ we have $\lambda_2 = \gamma_1 = 0$ and hence find dual elements $(x_1^*, -1) \in N((\bar{x}, f(\bar{x})); \operatorname{epi} f)$ and $(x_2^*, -1) \in N((\bar{x}, g(\bar{x})); \operatorname{epi} g)$ satisfying the relationships

$$x_1^* \in \partial f(\bar{x}), \quad x_2^* \in \partial g(\bar{x}), \text{ and } x^* = x_1^* + x_2^*.$$

This shows that $x^* \in \partial f(\bar{x}) + \partial g(\bar{x})$. Thus (2.1.8) holds and the proof is complete. $\square$

The following proposition sets up our use of the subdifferential sum rule in the proof of the fundamental theorem that follows. Recall that the optimal value function $\mu$ was defined in Definition 1.1.2.

**Proposition 2.1.6.** *Let $\phi\colon X \times Y \to \overline{\mathbb{R}}$ be convex and let $G\colon X \rightrightarrows Y$ have convex graph. Then for any $\bar{x} \in X$ and $\bar{y} \in M(\bar{x})$, we have the subdifferential representation*

$$\partial\mu(\bar{x}) = \left\{ x^* \in X^* \;\middle|\; (x^*, 0) \in \partial(\phi + \delta_{\operatorname{gph} G})(\bar{x}, \bar{y}) \right\}. \tag{2.1.10}$$

**Proof.** Let $\bar{y} \in M(\bar{x})$ be fixed for the entire proof.

For the first set inclusion, let $x^* \in \partial\mu(\bar{x})$. It follows from the definitions that

$$\langle x^*, x - \bar{x} \rangle \leq \mu(x) - \phi(\bar{x}, \bar{y}) \leq \phi(x, y) - \phi(\bar{x}, \bar{y}) \text{ for all } y \in G(x).$$

Since $(\bar{x}, \bar{y}) \in \operatorname{gph} G$, we can add the indicator function on the right hand side to get

$$\langle (x^*, 0), (x, y) - (\bar{x}, \bar{y}) \rangle \leq (\phi + \delta_{\operatorname{gph} G})(x, y) - (\phi + \delta_{\operatorname{gph} G})(\bar{x}, \bar{y})$$

for all $(x, y) \in X \times Y$, which proves $(x^*, 0) \in \partial(\phi + \delta_{\operatorname{gph} G})(\bar{x}, \bar{y})$.

For the reverse containment, let $(u^*, 0) \in \partial(\phi + \delta_{\operatorname{gph} G})(\bar{x}, \bar{y})$. Then we have $\langle u^*, x - \bar{x} \rangle \leq (\phi + \delta_{\operatorname{gph} G})(x, y) - (\phi + \delta_{\operatorname{gph} G})(\bar{x}, \bar{y})$ for all $(x, y) \in X \times Y$. This implies

$$\langle u^*, x - \bar{x} \rangle \leq \phi(x, y) - \mu(\bar{x}) \text{ for all } y \in G(x).$$

Since this holds for all $x \in X$, and since $\phi$ is continuous on its domain, taking the inf on the right hand side over all $y \in G(x)$ yields

$$\langle u^*, x - \bar{x} \rangle \leq \mu(x) - \mu(\bar{x}) \text{ for all } x \in X,$$

which means $u^* \in \partial\mu(\bar{x})$. This completes the proof. □

Next we define the coderivative of a set-valued map.

**Definition 2.1.7.** *Given a set-valued map $G\colon X \rightrightarrows Y$ and a point $(\bar{x}, \bar{y}) \in \operatorname{gph} G$, the* coderivative *is a set-valued map $D^*G(\bar{x}, \bar{y})\colon Y^* \rightrightarrows X^*$ defined by*

$$D^*G(\bar{x}, \bar{y})(y^*) := \{x^* \in X^* \mid (x^*, -y^*) \in N((\bar{x}, \bar{y}); \operatorname{gph} G)\}.$$

Next we are ready to prove the fundamental theorem which gives us a precise representation of the subdifferential of the optimal value function using coderivatives. This theorem has been stated and proven in a variety of settings, including: Asplund Spaces by Nam, Hoang and Rector in [**49**]; Hausdorff locally convex topological vector spaces by An and Yen in [**1**]; and in $\mathbb{R}^n$ by Mordukhovich and Nam [**41**]. Our presentation here follows the paper [**39**] by Mordukhovich, Nam, Rector and Tran.

**Theorem 2.1.8.** *Let $G\colon X \rightrightarrows Y$ be a convex set-valued map and $\phi\colon X \times Y \to \overline{\mathbb{R}}$ be a proper convex function. If at least one of the following conditions is satisfied:*

(i) $\operatorname{int}(\operatorname{gph} G) \cap \operatorname{dom} \phi \neq \emptyset$,

(ii) *$\phi$ is continuous at a point $(\bar{u}, \bar{v}) \in \operatorname{gph} G$,*

*then for any $\bar{x} \in X$ and $\bar{y} \in M(\bar{x})$ we have*

$$\partial\mu(\bar{x}) = \bigcup_{(x^*, y^*) \in \partial\phi(\bar{x}, \bar{y})} \{x^* + D^*G(\bar{x}, \bar{y})(y^*)\},$$

*where $\mu\colon X \to \overline{\mathbb{R}}$ is the optimal value function defined using $G$ and $\phi$.*

**Proof.** First we note that condition (i) implies that $\delta_{\operatorname{gph} G}$ is continuous at a point in $\operatorname{dom} \phi \cap \operatorname{dom}(\delta_{\operatorname{gph} G})$, and condition (ii) implies that $\phi$ is continuous at a point in the same intersection. Either way, we will be able to apply the subdifferential sum rule Theorem 2.1.5 to the function $(\phi + \delta_{\operatorname{gph} G})$.

By (2.1.10), we have that $u^* \in \partial\mu(\bar{x})$ is equivalent to $(u^*, 0) \in \partial(\phi + \delta_{\operatorname{gph} G})(\bar{x}, \bar{y})$. Under the set equality in Theorem 2.1.5, this is the same as saying that there exists $(x^*, y^*) \in \partial\phi(\bar{x}, \bar{y})$ such that $(u^*, 0) \in (x^*, y^*) + \partial\delta_{\operatorname{gph} G}(\bar{x}, \bar{y})$. Since $\partial\delta_{\operatorname{gph} G}(\bar{x}, \bar{y}) = N((\bar{x}, \bar{y}); \operatorname{gph} G)$, which follows from the definition, this is equivalent to $(u^* - x^*, -y^*) \in N((\bar{x}, \bar{y}); \operatorname{gph} G)$, which means $u^* - x^* \in D^*G(\bar{x}, \bar{y})(y^*)$ by the definition of the coderivative. Thus we see that $u^* \in \partial\mu(\bar{x})$ is equivalent to there being some $(x^*, y^*) \in \partial\phi(\bar{x}, \bar{y})$ such that $u^* \in x^* + D^*G(\bar{x}, \bar{y})(y^*)$, and this completes the proof. □

The following corollary treats objective functions that do not change their definitions with the choice of constraint set $G(x)$. This corollary is used repeatedly in the proofs of the results in the remainder of this section.

**Corollary 2.1.9.** *Let $G\colon X \rightrightarrows Y$ be a convex set-valued map and $\psi\colon Y \to \overline{\mathbb{R}}$ be a proper convex function. Define $\mu(x)$ as above but using $\psi$ instead of $\phi$, so $\mu(x) = \inf\{\psi(y) \mid y \in G(x)\}$. If at least one of the following conditions is satisfied:*

(i) *there exists $\bar{u} \in X$ and $\bar{v} \in \operatorname{dom}\psi$ such that $(\bar{u}, \bar{v}) \in \operatorname{int}(\operatorname{gph} G)$,*

(ii) *there exists $(\bar{u}, \bar{v}) \in \operatorname{gph} G$ such that $\psi$ is continuous at $\bar{v}$,*

*then for any $\bar{x} \in X$ and $\bar{y} \in M(\bar{x})$ we have*

$$\partial\mu(\bar{x}) = \bigcup_{y^* \in \partial\psi(\bar{y})} D^*G(\bar{x}, \bar{y})(y^*).$$

**Proof.** The result follows from an application of Theorem 2.1.8. The equivalence of the conditions (i) and (ii) from the two theorems is clear. Define the function $\phi\colon X \times Y \to \overline{\mathbb{R}}$ as $\phi(x, y) = \psi(y)$. Then we have

$$\partial\mu(\bar{x}) = \bigcup_{(x^*, y^*) \in \partial\phi(\bar{x}, \bar{y})} \{x^* + D^*G(\bar{x}, \bar{y})(y^*)\}. \tag{2.1.11}$$

The subdifferential $\partial\phi(\bar{x},\bar{y})$ is equal to the set of all $(x^*,y^*) \in X^* \times Y^*$ such that $\langle (x^*,y^*),(x,y)-(\bar{x},\bar{y})\rangle \leq \phi(x,y)-\phi(\bar{x},\bar{y})$ for all $(x,y) \in X \times Y$, which equals

$$\{(x^*,y^*) \in X^* \times Y^* \mid \langle x^*, x-\bar{x}\rangle + \langle y^*, y-\bar{y}\rangle \leq \psi(y)-\psi(\bar{y}) \text{ for all } (x,y) \in X \times Y\}.$$

Note that for any fixed $r \in \mathbb{R}$ we only have $\langle x^*, x-\bar{x}\rangle \leq r$ for all $x \in X$ if $x^* \equiv 0$. Thus if $(x^*,y^*) \in \partial\phi(\bar{x},\bar{y})$ we must have $x^* \equiv 0$. So we can re-write (2.1.11) as

$$\begin{aligned}\partial\mu(\bar{x}) &= \bigcup_{(x^*,y^*)\in\partial\phi(\bar{x},\bar{y})} \{D^*G(\bar{x},\bar{y})(y^*)\} \\ &= \bigcup_{y^*\in\partial\psi(\bar{y})} \{D^*G(\bar{x},\bar{y})(y^*)\},\end{aligned}$$

and the proof is complete. □

To help understand the optimal value function, the coderivative, and the fundamental theorem, we provide the following example.

**Example 2.1.10.** In this example we have $X = \mathbb{R}$ and $Y = \mathbb{R}$. We maintain the use of $X$ and $Y$ (rather than just writing $\mathbb{R}$) to help illustrate the previous presentation.

Let $G\colon X \rightrightarrows Y$ be the convex set-valued map defined by

$$G(x) := \left[\frac{1}{2}|x-2| + \frac{1}{2}(x-2),\, \infty\right).$$

Let $\psi\colon Y \to \overline{\mathbb{R}}$ be defined by

$$\psi(x) = |x+1| + |x-1| - 2.$$

Then it follows that the optimal value function $\mu\colon X \to \overline{\mathbb{R}}$ has the closed form expression

$$\mu(x) = |x-3| + (x-3) + 1.$$

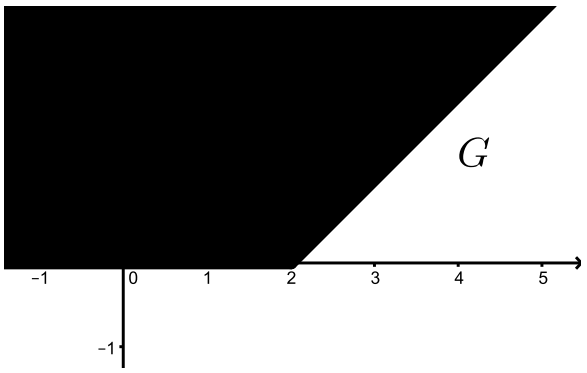



FIGURE 2.1. The set-valued map G.

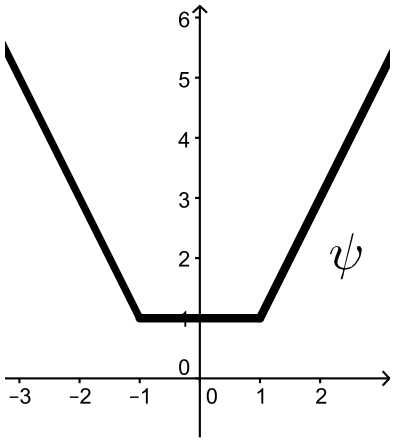



FIGURE 2.2. The objective function $\psi$.

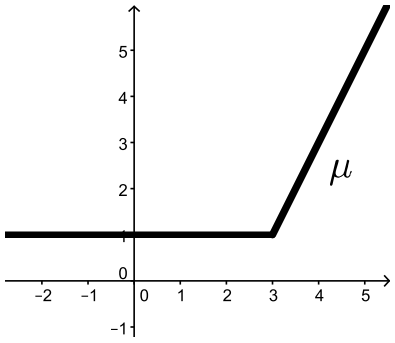



FIGURE 2.3. The resulting optimal value function $\mu$.

We can then use the 2.1.11 version of the fundamental theorem to calculate the subdifferential of $\mu$ at $\bar{x}$ as follows.

Let $\bar{x} = 1$. Then $M(\bar{x}) = [0, 1]$. Choose $\bar{y} = 0 \in M(\bar{x})$. Then $\partial\psi(\bar{y}) = \{0\}$ and $D^*G(\bar{x}, \bar{y})(0) = \{0\}$. So $\partial\mu(\bar{x}) = \{0\}$, as expected. If we were to use a different $\bar{y} \in M(\bar{x})$, say $\bar{y} = 1$, then we would have $\partial\psi(\bar{y}) = [0, 2]$. We can check that $D^*G(\bar{x}, \bar{y})(0) = \{0\}$ for all $y^* \in [0, 2]$, confirming that $\partial\mu(\bar{x}) = \{0\}$, as expected.

Let us calculate $\partial\mu(\bar{x})$ again but this time for $\bar{x} = 3$. In this case we have $M(\bar{x}) = \{1\}$, so use $\bar{y} = 1$. Then we can see $\partial\psi(\bar{y}) = [0, 2]$. A quick calculation shows us that for each $y^* \in [0, 2]$, we have $D^*G(\bar{x}, \bar{y})(y^*) = [0, y^*]$. Thus, taking the union over all $y^* \in [0, 2]$ shows us that $\partial\mu(\bar{x}) = [0, 2]$, as expected.

This concludes the example.

Next, we state a lemma that will be used in a subdifferential chain rule for affine transformations. This lemma also serves to give better insight into the coderivative map. Let $B\colon X \to Y$ be given by $B(x) = A(x)+b$ where $b \in Y$ is fixed and $A\colon X \to Y$ is linear. Recall the *adjoint* of $A$ is defined as $A^*\colon Y^* \to X^*$ where $A^*(y^*) = y^* \circ A$.

**Lemma 2.1.11.** *Let $B\colon X \rightrightarrows Y$ be an affine set-valued map given by $B(x) = \{A(x)+ b\}$ where $A\colon X \to Y$ is linear and $b \in Y$ is fixed. Then the coderivative of $B$ at $(\bar{x}, \bar{y}) \in \operatorname{gph} B$ is given by the formula*

$$D^*B(\bar{x}, \bar{y})(y^*) = \{A^*(y^*)\}$$

**Proof.** Let $\bar{x} \in X$ and $\bar{y} = A(\bar{x}) + b$. Recall the coderivative of $B$ at $(\bar{x}, \bar{y}) \in \operatorname{gph} B$ is a set-valued map $D^*B(\bar{x}, \bar{y})\colon Y^* \rightrightarrows X^*$ given by

$$D^*B(\bar{x}, \bar{y})(y^*) = \{x^* \in X^* \mid (x^*, -y^*) \in N((\bar{x}, \bar{y}); \operatorname{gph} B)\}.$$

Thus we need to show that $(x^*, -y^*) \in N((\bar{x}, \bar{y}); \operatorname{gph} B)$ is equivalent to $x^* = A^*(y^*)$.

Let $(x^*, y^*) \in N((\bar{x}, \bar{y}); \operatorname{gph} B)$. Since $\operatorname{gph} B$ is convex, this means that

$$\langle x^*, x - \bar{x} \rangle + \langle y^*, y - \bar{y} \rangle \leq 0 \qquad \text{for all } x, y \text{ such that } y = A(x) + b.$$

Since $\bar{y} = A(\bar{x})$, this means

$$\langle x^*, x - \bar{x} \rangle + \langle y^*, A(x - \bar{x}) \rangle \leq 0 \qquad \text{for all } x \in X,$$

which is the same as saying

$$\langle x^*, x - \bar{x} \rangle + \langle A^*(y^*), x - \bar{x} \rangle \leq 0 \qquad \text{for all } x \in X.$$

So $\langle x^* + A^*(y^*), x - \bar{x} \rangle \leq 0$ for all $x \in X$, which is equivalent to $x^* + A^*(y^*) \equiv 0$, so $x^* = -A^*(y^*)$. Thus we have deduced the logical equivalence $(x^*, y^*) \in N((\bar{x}, \bar{y}); \operatorname{gph} B) \iff x^* = -A^*(y^*)$, and hence can conclude $D^*B(\bar{x}, \bar{y})(y^*) = \{A^*(y^*)\}$, as desired. This completes the proof. □

This allows us to prove the following chain rule for affine transformations.

**Corollary 2.1.12.** *Let $X$ and $Y$ be locally convex topological vector spaces. Let $B\colon X \rightrightarrows Y$ be an affine set-valued map given by $B(x) = \{A(x) + b\}$ where $A\colon X \to Y$ is linear and $b \in Y$ is fixed. Let $\psi\colon Y \to \overline{\mathbb{R}}$ be a proper convex function. [Note that the optimal value function $\mu$ in this case becomes simply $\mu(x) = \inf\{\psi(y) \mid y = A(x) + b\} = \psi(A(x) + b)$.] If there exists some point $\bar{v} \in A(X) + b$ such that $\psi$ is continuous at $\bar{v}$, then for any $\bar{x} \in X$ with $\bar{y} = A(\bar{x}) + b$, we have*

$$\partial\mu(\bar{x}) = A^*(\partial\psi(\bar{y})).$$

**Proof.** By Corollary (2.1.9) we have

$$\partial\mu(\bar{x}) = \bigcup_{y^* \in \partial\psi(\bar{y})} D^*B(\bar{x}, \bar{y})(y^*).$$

By Lemma (2.1.11) we have $D^*B(\bar{x}, \bar{y})(y^*) = \{A^*(y^*)\}$. Thus,

$$\partial\mu(\bar{x}) = A^*(\partial\psi(\bar{y})),$$

which completes the proof. □

With the next corollary we obtain the results of [**33**] but prove using the optimal value function, thus showing the usefulness of our geometric approach. First we state a lemma.

**Lemma 2.1.13.** *Let $\phi\colon Y \to \overline{\mathbb{R}}$ be non-decreasing, $\bar{y} \in \operatorname{dom}\phi$, and $y^* \in \partial\phi(\bar{y})$. Then we have $\langle y^*, z\rangle \geq 0$ for all $z \in Y_+$.*

**Proof.** Let $z \in Y_+$. Note that $\bar{y} \geq \bar{y} - z$, and so

$$\phi(\bar{y}) \geq \phi(\bar{y} - z) \geq \phi(\bar{y}) - \langle y^*, z\rangle,$$

which is to say, $\langle y^*, z\rangle \geq 0$. □

**Corollary 2.1.14.** *Let $X$ and $Y$ be locally convex topological vector spaces. Let $f\colon X \to Y$ be convex and $\phi\colon Y \to \overline{\mathbb{R}}$ be convex and non-decreasing. If there exists $x \in X$ such that $\phi$ is continuous at some $y \in Y$ with $y \geq f(x)$, then for all $\bar{x} \in X$ we have*

$$\partial(\phi \circ f)(\bar{x}) = \bigcup_{y^* \in \partial\phi(f(\bar{x}))} \partial(y^* \circ f)(\bar{x}).$$

**Proof.** Define $G\colon X \rightrightarrows Y$ by $G(x) = \{y \in Y \mid y \geq f(x)\}$. Then $\mu = (\phi \circ f)$ and so by Corollary 2.1.9 we have

$$\partial(\phi \circ f)(\bar{x}) = \bigcup_{y^* \in \partial\phi(f(\bar{x}))} D^*G(\bar{x}, f(\bar{x}))(y^*).$$

Let $y^* \in \partial\phi(f(\bar{x}))$. We will prove the desired result by showing $D^*G(\bar{x}, f(\bar{x}))(y^*) = \partial(y^* \circ f)(\bar{x})$.

Let $x^* \in D^*G(\bar{x}, f(\bar{x}))(y^*)$. This implies that $\langle y^*, y\rangle \geq \langle y^*, f(\bar{x})\rangle + \langle x^*, x - \bar{x}\rangle$ for all $x \in X$ and $y \geq f(x)$. Let $h \in X$. Set $x = \bar{x} + h$ and $y = f(x)$. Since $y \geq f(x)$, this means

$$\langle y^*, f(\bar{x} + h)\rangle \geq \langle y^*, f(\bar{x})\rangle + \langle x^*, h\rangle\,,$$

so $x^* \in \partial(y^* \circ f)(\bar{x})$.

For the reverse containment, let $x^* \in \partial(y^* \circ f)(\bar{x})$. This implies that $\langle y^*, f(\bar{x} + h) - f(\bar{x})\rangle \geq \langle x^*, h\rangle$ for all $h \in X$. Let $x \in X$ and $y \geq f(x)$. Set $h = x - \bar{x}$, so $f(\bar{x} + h) = f(x)$. Note that by Lemma 2.1.13 we have $\langle y^*, y\rangle \geq \langle y^*, f(x)\rangle$ and so

$$\langle y^*, y - f(\bar{x})\rangle \geq \langle y^*, f(x) - f(\bar{x})\rangle \geq \langle x^*, x - \bar{x}\rangle\,,$$

which means $(x^*, -y^*) \in N((\bar{x}, f(\bar{x})); \operatorname{gph} G)$. Thus $x^* \in D^*G(\bar{x}, f(\bar{x}))(y^*)$, which completes the proof. □

The next corollary provides a useful formula for the subdifferential of a max of convex functions. This generalizes Theorem IV4.3.2 in Hiriart-Urruty [**21**], and is proven using a much simpler technique. It is essentially an application of Corollary 2.1.14.

**Corollary 2.1.15.** *Let $f_i\colon X \to \mathbb{R}$ be continuous convex functions for $i = 1, ..., m$. Define $f\colon X \to \mathbb{R}$ by $f_{max} = \max\{f_1, ..., f_m\}$. Let $I(x) := \{i \mid f_i(x) = f_{max}(x)\}$ be*

*the* active index set. *Then for* $\bar{x} \in X$ *we have*

$$\partial f_{max}(\bar{x}) = \text{co}\{\partial f_i(\bar{x}) \mid i \in I(\bar{x})\}.$$

**Proof.** Define $f\colon X \to \mathbb{R}^m$ by $f(x) = (f_1(x), ..., f_m(x))$ and $g\colon \mathbb{R}^m \to \mathbb{R}$ by $g(u_1, ..., u_m) = \max\{u_1, ..., u_m\}$. Then we have $f_{max} = g \circ f$. It is well known [**41**] that the subdifferential of $g$ at $(u_1, ..., u_m) \in \mathbb{R}^m$ can be written as $\partial g(u_1, ..., u_m) = \text{co}\{e_i \mid u_i = g(u_1, ..., u_m)\}$, where $e_1, ..., e_m$ is the standard basis on $\mathbb{R}^m$. Thus, for $x \in X$, the subdifferential of $g$ at $f(x) = (f_1(x), ..., f_m(x))$ is

$$\begin{aligned} \partial g(f(x)) &= \partial g(f_1(x), ..., f_m(x)) \\ &= \text{co}\{e_i \mid i \in I(x)\} \\ &= \left\{ \sum_{i \in I(x)} \lambda_i e_i \mid \lambda_i \geq 0, \sum_{i \in I(x)} \lambda_i = 1 \right\}. \end{aligned} \tag{2.1.12}$$

Since $g$ is non-decreasing (using $Y_+ = [0, \infty)^m$), continuous at all $u \in \mathbb{R}^m$, and there is a point $\bar{x} \in X$ such that each $f_i$ is continuous (and hence finite) at $\bar{x}$, then we can apply Corollary 2.1.14 to get

$$\partial(f_{max})(\bar{x}) = \partial(g \circ f)(\bar{x}) = \bigcup_{y^* \in \partial g(f(\bar{x}))} \partial(y^* \circ f)(\bar{x}).$$

We will complete the proof by showing that this union is equal to $\text{co}\{\partial f_i(\bar{x}) \mid i \in I(\bar{x})\}$. First, let us consider $y^* \in \partial g(f(\bar{x}))$ and the resulting $y^* \circ f\colon X \to \mathbb{R}$.

Let $(\lambda_i)_{i\in I(x)}$ be the associated scalars from the representation (2.1.12). Then we have

$$\begin{aligned}(y^* \circ f)(x) &= y^*(f_1(x), ..., f_m(x)) \\ &= \sum_{i\in I(x)} \lambda_i \langle e_i, (f_1(x), ..., f_m(x))\rangle \\ &= \sum_{i\in I(x)} \lambda_i f_i(x) = \left(\sum_{i\in I(x)} \lambda_i f_i\right)(x).\end{aligned}$$

Thus, we can write

$$\partial(y^* \circ f)(x) = \partial\left(\sum_{i\in I(x)} \lambda_i f_i\right)(x) = \sum_{i\in I(x)} \lambda_i \partial f_i(x).$$

Therefore

$$\bigcup_{y^*\in\partial g(f(\bar{x}))} \partial(y^* \circ f)(\bar{x}) \subseteq \text{co}\{\partial f_i(\bar{x}) \mid i \in I(\bar{x})\}.$$

Furthermore, since every $(\lambda_i)_{i\in I(x)}$ with $\lambda_i \geq 0, \sum_{i\in I(x)} \lambda_i = 1$ defines an element $y^* \in \partial g(f(\bar{x}))$, we also have the opposite inclusion

$$\bigcup_{y^*\in\partial g(f(\bar{x}))} \partial(y^* \circ f)(\bar{x}) \supseteq \text{co}\{\partial f_i(\bar{x}) \mid i \in I(\bar{x})\},$$

and so can conclude the desired set equality

$$\begin{aligned}\partial(f_{max})(\bar{x}) &= \bigcup_{y^*\in\partial g(f(\bar{x}))} \partial(y^* \circ f)(\bar{x}) \\ &= \text{co}\{\partial f_i(\bar{x}) \mid i \in I(\bar{x})\}.\end{aligned}$$

This completes the proof. □

The next corollary also generalizes and provides a simpler proof for Theorem IV4.5.1 in [**21**].

**Corollary 2.1.16.** *Let $A\colon Y \to X$ be a surjective linear map. Use $A$ to define a set-valued map $G\colon X \rightrightarrows Y$ by $G(x) := A^{-1}(x) = \{y \in Y \mid A(x) = y\}$. Let $\psi\colon Y \to \overline{\mathbb{R}}$ be a proper convex function and $\mu$ be the optimal value function associated with $\psi$ and $G$. If there exists $\bar{v} \in Y$ such that $\psi$ is continuous at $\bar{v}$, then for any $\bar{x} \in X$ and $\bar{y} \in M(\bar{x})$ we have*

$$\partial\mu(\bar{x}) = (A^*)^{-1}(\partial\psi(\bar{y})).$$

**Proof.** Since $A$ is surjective, the condition that $\psi$ is continuous at some $\bar{v} \in Y$ implies the regularity condition (ii) from Corollary 2.1.9. Therefore we have the formula

$$\partial\mu(\bar{x}) = \bigcup_{y^* \in \partial\psi(\bar{y})} D^*G(\bar{x}, \bar{y})(y^*).$$

All that needs to be shown is that $D^*G(\bar{x}, \bar{y}) = (A^*)^{-1}$. Note that $(A^*)^{-1}\colon Y^* \rightrightarrows X^*$ and is defined by $(A^*)^{-1}(y^*) = \{x^* \in X^* \mid A^*(x^*) = y^*\}$. To show that $D^*G(\bar{x}, \bar{y}) = (A^*)^{-1}$, we will show that $(x^*, y^*) \in N((\bar{x}, \bar{y}); \operatorname{gph} G)$ is equivalent to $A^*(x^*) = -y^*$. This is done easily with the following string of logical equivalences:

$$\begin{aligned}
&(x^*, y^*) \in N((\bar{x}, \bar{y}); \operatorname{gph} G) \\
&\iff \langle (x^*, y^*), (x, y) - (\bar{x}, \bar{y}) \rangle \leq 0 && \forall (x, y) \in \operatorname{gph} G \\
&\iff \langle x^*, x - \bar{x} \rangle + \langle y^*, y - \bar{y} \rangle \leq 0 && \text{for all } x \in X \text{ and } y \in Y \text{ with } x = A(y) \\
&\iff \langle x^*, A(y) - A(\bar{y}) \rangle + \langle y^*, y - \bar{y} \rangle \leq 0 && \text{for all } y \in Y \text{ (since } A \text{ is surjective)} \\
&\iff \langle A^*(x^*), y - \bar{y} \rangle + \langle y^*, y - \bar{y} \rangle \leq 0 && \text{for all } y \in Y \\
&\iff A^*(x^*) \equiv -y^*.
\end{aligned}$$

Thus $D^*G(\bar{x},\bar{y}) = (A^*)^{-1}$, so

$$\partial\mu(\bar{x}) = \bigcup_{y^*\in\partial\psi(\bar{y})} D^*G(\bar{x},\bar{y})(y^*) = (A^*)^{-1}(\partial\psi(\bar{y})),$$

which completes the proof. □

## 2.2. Coderivative Rules

As in the previous section, we continue to assume that our spaces $X, Y$ and $Z$ are Hausdorff locally convex topological vector spaces. Our goal in this section is to derive various formulas for coderivatives where the set-valued map involves, for example, the *sum*, *intersection*, or *composition* of other set-valued maps. Such formulas may be useful for future researchers using coderivatives in variational analysis. The formulas presented here represent improvements to the formulas previously found in the literature, and the first time these formulas have been presented in infinite dimensions.

First, we note that the normal cone intersection rule Theorem 2.1.4 generalizes as follows.

**Proposition 2.2.1.** *Let $A_1, ..., A_n$ be convex subsets of $X$ with $A_1 \cap \operatorname{int}(A_2) \cap \cdots \cap \operatorname{int}(A_n) \neq \emptyset$. Define $A = A_1 \cap \cdots \cap A_n$. Then for any $x \in A$ we have*

$$N(x;A) = N(x;A_1) + \cdots + N(x;A_n).$$

Next we state the first of our coderivative formulas.

Let $F_1, F_2 \colon X \rightrightarrows Y$. Define $(F_1 + F_2) \colon X \rightrightarrows Y$ by $(F_1 + F_2)(x) = F_1(x) + F_2(x)$. It is straightforward to check that $F_1 + F_2$ is convex when $F_1$ and $F_2$ are. Also, $\operatorname{dom}(F_1 + F_2) = \operatorname{dom}(F_1) \cap \operatorname{dom}(F_2)$. For $(\bar{x},\bar{y}) \in \operatorname{gph}(F_1 + F_2)$, define

$$S(\bar{x},\bar{y}) = \{(\bar{y}_1,\bar{y}_2) \in Y \times Y \mid \bar{y} = \bar{y}_1 + \bar{y}_2 \text{ and } \bar{y}_1 \in F_1(\bar{x}),\ \bar{y}_2 \in F_2(\bar{x})\}.$$

**Theorem 2.2.2.** *Let $F_1, F_2 \colon X \rightrightarrows Y$ be convex set-valued graphs. Assume one of the following holds:*

**(i)** $\operatorname{int}(\operatorname{dom} F_1) \cap \operatorname{dom} F_2 \neq \emptyset$ *and* $\operatorname{int}(\operatorname{gph} F_1) \neq \emptyset$ *or*

**(ii)** $\operatorname{dom} F_1 \cap \operatorname{int}(\operatorname{dom} F_2) \neq \emptyset$ *and* $\operatorname{int}(\operatorname{gph} F_2) \neq \emptyset$.

*Then for all $(\bar{x}, \bar{y}) \in \operatorname{gph}(F_1 + F_2)$ and $v \in Y^*$ we have*

$$D^*(F_1 + F_2)(\bar{x}, \bar{y})(v) = D^*F_1(\bar{x}, \bar{y}_1)(v) + D^*F_2(\bar{x}, \bar{y}_2)(v)$$

*for each $(\bar{y}_1, \bar{y}_2) \in S(\bar{x}, \bar{y})$.*

**Proof.** Let $v \in Y^*$ and $(\bar{x}, \bar{y}) \in \operatorname{gph}(F_1 + F_2)$ and $(\bar{y}_1, \bar{y}_2) \in S(\bar{x}, \bar{y})$ be fixed for the entire proof.

We start with the inclusion "$\subset$". Fix any $u \in D^*(F_1 + F_2)(\bar{x}, \bar{y})(v)$. By definition, this means

$$(u, -v) \in N((\bar{x}, \bar{y}); \operatorname{gph}(F_1 + F_2)).$$

Define the sets $\Omega_1$ and $\Omega_2$ as follows:

$$\Omega_1 = \{(x, y_1, y_2) \in X \times Y \times Y \mid y_1 \in F_1(x)\},$$

$$\Omega_2 = \{(x, y_1, y_2) \in X \times Y \times Y \mid y_2 \in F_2(x)\}.$$

It is easy to check that $(u, -v, -v) \in N((\bar{x}, \bar{y}_1, \bar{y}_2); \Omega_1 \cap \Omega_2)$. By construction, assumption (i) in the statement of the theorem implies $\operatorname{int}(\Omega_1) \cap \Omega_2 \neq \emptyset$ and assumption (ii) implies $\Omega_1 \cap \operatorname{int}(\Omega_2) \neq \emptyset$. So we can apply Proposition 2.2.1 to get

$$(u, -v, -v) \in N((\bar{x}, \bar{y}_1, \bar{y}_2); \Omega_1) + N((\bar{x}, \bar{y}_1, \bar{y}_2); \Omega_2).$$

This gives us the representation $(u, -v, -v) = (u_1, -v, 0) + (u_2, 0, -v)$ where

$$(u_1, -v) \in N((\bar{x}, \bar{y}_1); \operatorname{gph} F_1) \qquad \text{and} \qquad (u_2, -v) \in N((\bar{x}, \bar{y}_2); \operatorname{gph} F_2).$$

Thus we have $u = u_1 + u_2 \in D^*F_1(\bar{x}, \bar{y}_1)(v) + D^*F_2(\bar{x}, \bar{y}_2)(v)$, which proves the desired inclusion.

For the opposite inclusion, let $w \in D^*F_1(\bar{x}, \bar{y}_1)(v) + D^*F_2(\bar{x}, \bar{y}_2)(v)$ so $w = w_1 + w_2$ where $w_1 \in D^*F_1(\bar{x}, \bar{y}_1)(v)$ and $w_2 \in D^*F_2(\bar{x}, \bar{y}_2)(v)$. This means

$$\begin{aligned} &\langle (w_1, -v), (x_1, y_1) - (\bar{x}, \bar{y}_1) \rangle \leq 0 && \text{for all } y_1 \in F_1(x_1) \text{ and} \\ &\langle (w_2, -v), (x_2, y_2) - (\bar{x}, \bar{y}_2) \rangle \leq 0 && \text{for all } y_2 \in F_2(x_2). \end{aligned}$$

This implies

$$\langle w_1, x_1 - \bar{x} \rangle + \langle w_2, x_2 - \bar{x} \rangle - \langle v, y_1 - \bar{y}_1 \rangle - \langle v, y_2 - \bar{y}_2 \rangle \leq 0$$
$$\text{for all } y_1 \in F_1(x_1) \text{ and all } y_2 \in F_2(x_2).$$

Setting $x_1 = x_2 = x$ then implies the more restrictive statement

$$\langle w, x - \bar{x} \rangle - \langle v, y_1 - \bar{y}_1 \rangle - \langle v, y_2 - \bar{y}_2 \rangle \leq 0 \qquad \text{for all } y_1 + y_2 \in (F_1 + F_2)(x).$$

Setting $y = y_1 + y_2$, we then get

$$\langle w, x - \bar{x} \rangle - \langle v, y - \bar{y} \rangle \leq 0 \qquad \text{for all } (x, y) \in \operatorname{gph}(F_1 + F_2).$$

Thus $w \in D^*(F_1 + F_2)(\bar{x}, \bar{y})(v)$. □

Next we define composition for set-valued maps and present a theorem for coderivatives defined by these types of maps.

If $F\colon X \rightrightarrows Y$ and $G\colon Y \rightrightarrows Z$ are set-valued maps, then we define the composition $(G \circ F)\colon X \rightrightarrows Z$ by

$$(G \circ F)(x) = \bigcup_{y \in F(x)} G(y).$$

It is easy to check that if $F$ and $G$ are convex, then $(G \circ F)$ is convex.

For any $\bar{z} \in (G \circ F)(\bar{x})$, define $T(\bar{x}, \bar{z}) = F(\bar{x}) \cap G^{-1}(\bar{z})$.

**Theorem 2.2.3.** *Let $F\colon X \rightrightarrows Y$ and $G\colon Y \rightrightarrows Z$ be convex set-valued maps. Assume that one of the following holds:*

**(i)** *there exists $(a, b) \in \operatorname{int}(\operatorname{gph} F)$ with $b \in \operatorname{dom} G$; or*

**(ii)** *there exists $(b, c) \in \operatorname{int}(\operatorname{gph} G)$ with $b \in F(a)$ for some $a \in X$.*

*Then for all $(\bar{x}, \bar{z}) \in \operatorname{gph}(G \circ F)$ and for all $w \in Z^*$ we have*

$$D^*(G \circ F)(\bar{x}, \bar{z})(w) = [D^*F(\bar{x}, \bar{y}) \circ D^*G(\bar{y}, \bar{z})](w)$$

*for each $\bar{y} \in T(\bar{x}, \bar{z})$.*

**Proof.** Let $(\bar{x}, \bar{z}) \in \operatorname{gph}(G \circ F)$ and $w \in Z^*$ and $\bar{y} \in T(\bar{x}, \bar{z})$ be fixed for the entire proof.

The "$\subset$" inclusion is proven as follows. Let $u \in D^*(G \circ F)(\bar{x}, \bar{z})(w)$. By definition, this means that

$$\langle u, x - \bar{x} \rangle - \langle w, z - \bar{z} \rangle \leq 0 \qquad \text{for all } (x, z) \in \operatorname{gph}(g \circ F).$$

Define the following sets:

$$\Omega_1 = \{(x, y, z) \in X \times Y \times Z \mid (x, y) \in \operatorname{gph}(F)\}$$

$$\Omega_2 = \{(x, y, z) \in X \times Y \times Z \mid (y, z) \in \operatorname{gph}(G)\}.$$

Since $\Omega_1 \cap \Omega_2$ contains $\operatorname{gph}(F \circ G)$ as its first and third entries, we have $(u, 0, -w) \in N((\bar{x}, \bar{y}, \bar{z}); \Omega_1 \cap \Omega_2)$. Assumption (i) indicates that $\operatorname{int}(\Omega_1) \cap \Omega_2 \neq \emptyset$ and assumption (ii) indicates that $\Omega_1 \cap \operatorname{int}(\Omega_2) \neq \emptyset$, so by 2.2.1 we have

$$N((\bar{x}, \bar{y}, \bar{z}); \Omega_1 \cap \Omega_2) = N((\bar{x}, \bar{y}, \bar{z}); \Omega_1) + N((\bar{x}, \bar{y}, \bar{z}); \Omega_2).$$

By the construction of $\Omega_1$ and $\Omega_2$, we have

$$\begin{aligned} N((\bar{x}, \bar{y}, \bar{z}); \Omega_1) =& N((\bar{x}, \bar{y}); \operatorname{gph} F) \times \{0\} \qquad \text{and} \\ N((\bar{x}, \bar{y}, \bar{z}); \Omega_2) =& \{0\} \times N((\bar{y}, \bar{z}); \operatorname{gph} G). \end{aligned}$$

So we can write $(u, 0, -w) = (u, -v, 0) + (0, v, -w)$ where $(u, -v) \in N((\bar{x}, \bar{y}); \operatorname{gph}(F))$ and $(v, -w) \in N((\bar{y}, \bar{z}); \operatorname{gph}(G))$. This means that $v \in D^*G(\bar{y}, \bar{z})(w)$ and $u \in D^*F(\bar{x}, \bar{y})(v)$ so we can conclude $u \in [D^*F(\bar{x}, \bar{y}) \circ D^*G(\bar{y}, \bar{z})](w)$.

For the opposite inclusion, let $u \in [D^*F(\bar{x}, \bar{y}) \circ D^*G(\bar{y}, \bar{z})](w)$. This means there exists $v \in D^*G(\bar{y}, \bar{z})(w)$ so that $u \in F(\bar{x}, \bar{y})(v)$. This means that

$$\begin{aligned} \langle u, x - \bar{x} \rangle - \langle v, y_1 - \bar{y} \rangle \leq& 0 \qquad \text{for all } y_1 \in F(x) \\ \langle v, y_2 - \bar{y} \rangle - \langle w, z - \bar{z} \rangle \leq& 0 \qquad \text{for all } z \in G(y_2). \end{aligned}$$

Summing these two inequalities and restricting to the case when $y_1 = y_2 = y$, we get

$$\langle u, x - \bar{x} \rangle - \langle w, z - \bar{z} \rangle \leq 0 \qquad \text{for all } z \in (G \circ F)(x).$$

Thus, $u \in D^*(G \circ F)(\bar{x}, \bar{z})(w)$ and the set equality is established and completes the proof. □

Next we discuss a formula for the coderivative of an *intersection* of set-valued mappings. Let $F_1, F_2 \colon X \rightrightarrows Y$ be set-valued mappings. Then $(F_1 \cap F_2) \colon X \rightrightarrows Y$ is a

set-valued map defined by

$$(F_1 \cap F_2)(x) = F_1(x) \cap F_2(x).$$

First we note the set equality $\operatorname{gph}(F_1 \cap F_2) = (\operatorname{gph} F_1) \cap (\operatorname{gph} F_2)$. Thus, if $F_1$ and $F_2$ are convex, then so is $(F_1 \cap F_2)$. Then we have the following formula for the coderivative of an intersection of set-valued maps.

**Theorem 2.2.4.** *Let $F_1, F_2 \colon X \rightrightarrows Y$ be set-valued maps with convex graphs. If $\operatorname{int}(\operatorname{gph} F_1) \cap \operatorname{gph} F_2 \neq \emptyset$ or $\operatorname{gph} F_1 \cap \operatorname{int}(\operatorname{gph} F_2) \neq \emptyset$, then for any $\bar{y} \in (F_1 \cap F_2)(\bar{x})$ and any $v \in Y^*$, we have*

$$D^*(F_1 \cap F_2)(\bar{x}, \bar{y})(v) = \bigcup_{v_1 + v_2 = v} [D^*F_1(\bar{x}, \bar{y})(v_1) + D^*F_2(\bar{x}, \bar{y})(v_2)].$$

**Proof.** Let $\bar{y} \in (F_1 \cap F_2)(\bar{x})$ and $v \in Y^*$.

First we prove the inclusion "$\subset$". Fix any $u \in D^*(F_1 \cap F_2)(\bar{x}, \bar{y})(v)$. Since $\operatorname{int}(\operatorname{gph} F_1) \cap \operatorname{gph} F_2 \neq \emptyset$ or $\operatorname{gph} F_1 \cap \operatorname{int}(\operatorname{gph} F_2) \neq \emptyset$, we can apply 2.2.1 to have

$$(u, -v) \in N((\bar{x}, \bar{y}); \operatorname{gph}(F_1 \cap F_2)) = N((\bar{x}, \bar{y}); \operatorname{gph} F_1) + N((\bar{x}, \bar{y}); \operatorname{gph} F_2).$$

Thus $(u, -v) = (u_1, -v_1) + (u_2, -v_2)$ where $(u_1, -v_1) \in N((\bar{x}, \bar{y}); \operatorname{gph} F_1)$ and $(u_2, -v_2) \in N((\bar{x}, \bar{y}); \operatorname{gph} F_2)$. So $u \in D^*F_1(\bar{x}, \bar{y})(v_1) + D^*F_2(\bar{x}, \bar{y})(v_2)$ where $v = v_1 + v_2$, which proves the desired inclusion.

For the inclusion "$\supset$", let $v_1, v_2 \in Y^*$ with $v_1 + v_2 = v$. Let $u \in D^*F_1(\bar{x}, \bar{y})(v_1) + D^*F_2(\bar{x}, \bar{y})(v_2)$. This means $u = u_1 + u_2$ where $u_1 \in D^*F_1(\bar{x}, \bar{y})(v_1)$ and

$u_2 \in D^*F_2(\bar{x},\bar{y})(v_2)$. So

$$(u,-v) = (u_1,-v_1) + (u_2,-v_2) \in N((\bar{x},\bar{y});\operatorname{gph} F_1) + N((\bar{x},\bar{y});\operatorname{gph} F_2)$$
$$= N((\bar{x},\bar{y});\operatorname{gph}(F_1 \cap F_2)).$$

Thus $u \in D^*(F_1 \cap F_2)(\bar{x},\bar{y})(v)$, which completes the proof. □

Next we derive a formula for the solution mapping of a generalized equation, defined below. We start with a proposition that is used in the proof of the theorem.

**Proposition 2.2.5.** *Let $F\colon X \rightrightarrows Y$ be a set-valued map with convex graph. Given $\bar{x} \in \operatorname{dom} F$, we have*

$$N(\bar{x};\operatorname{dom} F) = D^*F(\bar{x},\bar{y})(0)$$

*for any $\bar{y} \in F(\bar{x})$.*

**Proof.** Since $(x^*,0) \in N((\bar{x},\bar{y});\operatorname{gph} F) \iff x^* \in N(\bar{x};\operatorname{dom} F)$, the proof follows directly from the definition. □

If $F,G\colon X \times Y \rightrightarrows Z$ are set-valued mappings with convex graphs, we define their *generalized equation* to be

$$0 \in F(x,y) + G(x,y).$$

The *solution map* associated with this generalized equation is $S\colon X \rightrightarrows Y$ defined by

$$S(x) = \{y \in Y \mid 0 \in F(x,y) + G(x,y)\}.$$

The following theorem gives us a formula for the coderivative of $S$.

**Theorem 2.2.6.** *Let $F, G\colon X \times Y \rightrightarrows Z$ be convex set-valued maps and $S\colon X \rightrightarrows Y$ be the associated solution map, as defined above. Assume that* $\operatorname{int}(\operatorname{gph} F) \cap (\operatorname{int}(-\operatorname{gph} G)) \neq \emptyset$. *Then*

$$D^*S(\bar{x},\bar{y})(v) = \bigcup_{w\in Z^*} \left\{u \in X^* \;\middle|\; (u,-v) \in [D^*F((\bar{x},\bar{y}),\bar{z})(w) + D^*G((\bar{x},\bar{y}),-\bar{z})(-w)]\right\}$$

*for every $v \in Y^*$ and every $\bar{z} \in F(\bar{x},\bar{y}) \cap [-G(\bar{x},\bar{y})]$.*

**Proof.** Let $v \in Y^*$ and $\bar{z} \in F(\bar{x},\bar{y}) \cap [-G(\bar{x},\bar{y})]$. First, notice that $\operatorname{gph} S = \operatorname{dom}[F \cap (-G)]$. Then the proof can be completed with a string of logical equivalences as follows.

Let $u \in D^*S(\bar{x},\bar{y})(v)$. By definition, this means $(u,-v) \in N((\bar{x},\bar{y}); \operatorname{dom}[F \cap (-G)])$. By Proposition 2.2.5 we have

$$N((\bar{x},\bar{y}); \operatorname{dom}[F \cap (-G)]) = D^*[F \cap (-G)]((\bar{x},\bar{y}),\bar{x})(0),$$

and so by Theorem 2.2.4 we can write

$$(u,-v) \in \bigcup_{w\in Z^*} [D^*F((\bar{x},\bar{y}),\bar{z})(w) + D^*(-G)((\bar{x},\bar{y}),\bar{z})(-w)].$$

Moving a minus sign in the second coderivative then allows us to rewrite as

$$u \in \bigcup_{w\in Z^*} \left\{u \in X^* \;\middle|\; (u,-v) \in [D^*F((\bar{x},\bar{y}),\bar{z})(w) + D^*G((\bar{x},\bar{y}),-\bar{z})(-w)]\right\},$$

which completes the proof. □

# 3

## Applications to Facility Location Problems

Facility location problems include a broad range of optimization problems with many generalizations. The term *facility location*, true to its name, refers to the problem of finding the optimal location for a facility (typically thought of as some sort of product distribution center) to serve a fixed set of demand centers (typically thought of as customers, often referred to as targets). These problems have intrinsic geometric appeal and attracted interest from the likes of Descartes, Fermat, and Gauss (see some history below), but also stand as one of the oldest classes of problems in operations research due to their obvious practical applications. For example, facility location models have been applied to choose locations for emergency medical services [**62**], fast-food restaurants [**36**], and vehicle inspection stations [**23**].

Many generalized versions of the location problem have been stated and studied over the years; see, for example, [**4, 40, 42, 43, 44, 47, 48, 78**]. In this thesis, we consider two different generalizations, both concerning an arbitrary number of targets in $\mathbb{R}^n$, and both leading to nonsmooth optimization problems. In particular, the second generalization we consider leads to a non-convex optimization problem. Both generalizations are explained in the two sections below.

## 3.1. Introduction to the Fermat-Torricelli Problem and Nesterov's Method

In 1638, Descartes wrote a letter to Fermat which discussed curves in the plane whose points had a constant sum of distances to a given set of four points. These curves came to be know as multifocal ellipses or polyellipses. Five years later, possibly prompted by this discussion, Fermat posed the following problem: Find the point that minimizes the sum of distances to three given points in the plane. Early contributions and solutions to this problem were given by Torricelli, Cavalieri, and Viviani. The problem became know as the *Fermat-Torricelli* problem, and its solution the *Fermat-Torricelli point*. [**31**]

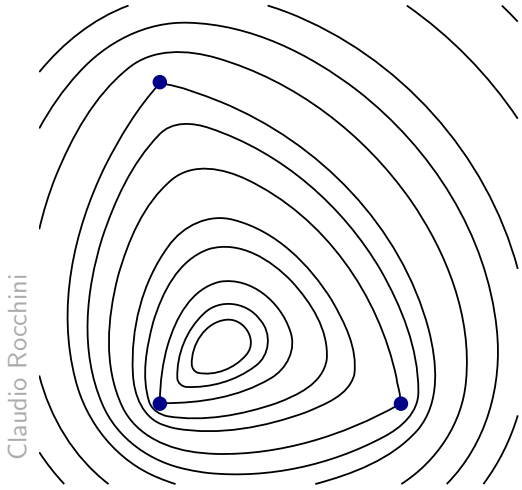



FIGURE 3.1. Polyellipses with three foci.

The original solutions to this problem were based on ruler and compass constructions. But it can be shown via Galois theory that no such construction exists when the number of points is greater than or equal to five. That is to say that "no exact algorithms under computational models with arithmetic operations and extractions

of $k$th roots can be used, leaving only numerical or symbolic approximation methods for more than four given points."[**31**]

3.1.0.4. *An Early Computational Solution: The Weiszfeld Algorithm.* In 1937, Endre Weiszfeld (also known as Andrew Vázsonyi) presented an algorithm to numerically solve the Fermat-Torricelli problem for $m$ points. For clarity, we state the problem here.

Let $a_1, ..., a_m$ be given points in $\mathbb{R}^n$, and let $\|\cdot\|$ be the Euclidean norm in $\mathbb{R}^n$. Then the Fermat-Torricelli problem is:

$$\text{minimize } \sum_{j=1}^{m} \|x - a_j\|,\ x \in \mathbb{R}^n.$$

The Weiszfeld algorithm is given by

$$x_{k+1} := \frac{\left(\sum_{j=1}^{m} \frac{a_j}{\|x_k - a_j\|}\right)}{\left(\sum_{j=1}^{m} \frac{1}{\|x_k - a_j\|}\right)}.$$

In this way, the algorithm makes a weighted average where the weight associated with each point $a_j$ is inversely proportional to its distance to the current estimate. But, as can be seen, the algorithm fails when an estimate $x_k$ coincides with one of the given points $a_j$. A modification to the algorithm was given in the year 2000 by [**81**] that converges for all initial points.

3.1.0.5. *Methodology.* In general, a natural approach for solving nonsmooth optimization problems is to approximate the original nonsmooth problem by a smooth one and apply a smooth optimization scheme to the smooth approximation. One successful implementation of this idea was provided by Nesterov. In his seminal papers

[**51, 52**], Nesterov introduced a fast first-order method for solving convex smooth optimization problems in which the cost functions have Lipschitz gradient. In contrast to the convergence rate of $O(1/k)$ when applying the classical gradient method to this class of problems, Nesterov's accelerated gradient method gives a convergence rate of $O(1/k^2)$. In Nesterov's nonsmooth optimization scheme, an original nonsmooth function of a particular form is approximated by a smooth convex function with Lipschitz gradient. Then the accelerated gradient method is applied to solve the smooth approximation. This method is considered to be a highly significant contribution to the field of convex optimization, and has sparked a vast array of applications and related research.

**3.1.1. Nesterov's Smoothing Technique.** We introduce Nesterov's smoothing technique in its full generality. The class of functions under consideration for Nesterov's smoothing technique are those given by

$$f(x) := \max\{\langle Ax, u\rangle - \phi(u) \mid u \in Q\},\ x \in \mathbb{R}^n,$$

where $A$ is an $m \times n$ matrix, $Q$ is a nonempty compact convex subset of $\mathbb{R}^m$, and $\phi$ is a continuous convex function on $Q$.

**Example 3.1.1.** *A simple example is as follows. If* $Q = \{u \in \mathbb{R}^n \mid \|u\|_1 \leq 1\}$, $A = I$, *and* $\phi \equiv 0$, *then we have* $f(x) = \|x\|_\infty$.

Let $d$ be a continuous strongly convex function on $Q$ with parameter $\sigma > 0$. This means that $d(x) - \frac{\sigma}{2}\|x\|^2$ is convex on $Q$. The function $d$ is called a *prox-function*; it is used to define the smooth approximation of $f$ presented below. Since $d$ is strongly convex on $Q$, it has a unique optimal solution on this set. Denote

$$\bar{u} := \arg\min\{d(u) \mid u \in Q\}.$$

Without loss of generality, we assume that $d(\bar{u}) = 0$. From the strong convexity of $d$, we also have

$$d(u) \geq \frac{\sigma}{2}\|u - \bar{u}\|^2 \text{ for all } u \in Q.$$

Throughout the chapter we will work mainly with the choice of $d(u) = \frac{1}{2}\|u - \bar{u}\|^2$.

Let $\mu$ be a positive number which we will call a *smooth parameter*. Define

$$f_\mu(x) := \max\{\langle Ax, u\rangle - \phi(u) - \mu d(u) \mid u \in Q\}. \tag{3.1.13}$$

The function $f_\mu$ will be the *Nesterov smooth approximation* of $f$. The forthcoming theorem provides a detailed understanding of this $f_\mu$ as an approximation of $f$.

For an $m \times n$ matrix $A$, define

$$\|A\| := \max\{\|Ax\| \mid \|x\| \leq 1\}. \tag{3.1.14}$$

The definition gives us

$$\|Ax\| \leq \|A\| \, \|x\| \text{ for all } x \in \mathbb{R}^n.$$

We also recall the definition of the *Euclidean projection* from point $x \in \mathbb{R}^n$ to a nonempty closed convex subset $\Omega$ of $\mathbb{R}^n$:

$$\pi(x; \Omega) := \{w \in \Omega \mid d(x; \Omega) = \|x - w\|\},$$

where $d(\cdot; \cdot)$ is the *distance function*

$$d(x; \Omega) := \inf\{\|x - w\| \mid w \in \Omega\}. \tag{3.1.15}$$

The theorem below is a simplified version of [**52**, Theorem 1] that involves the usual inner product of $\mathbb{R}^n$. In the paper by An, Nam, Rector and Sun [**46**], a new detailed proof is provided for the convenience of the reader.

**Theorem 3.1.2.** *Consider the function $f$ given by*

$$f(x) := \max\{\langle Ax, u\rangle - \langle b, u\rangle \mid u \in Q\},\ x \in \mathbb{R}^n,$$

*where $A$ is an $m \times n$ matrix and $Q$ is a compact subset of $\mathbb{R}^m$. Let $d(u) = \dfrac{1}{2}\|u - \bar{u}\|^2$ with $\bar{u} \in Q$.*

*Then the function $f_\mu$ in* (3.1.13) *has the explicit representation*

$$f_\mu(x) = \frac{\|Ax - b\|^2}{2\mu} + \langle Ax - b, \bar{u}\rangle - \frac{\mu}{2}\left[d\left(\bar{u} + \frac{Ax - b}{\mu}; Q\right)\right]^2$$

*and is continuously differentiable on $\mathbb{R}^n$ with its gradient given by*

$$\nabla f_\mu(x) = A^\top u_\mu(x),$$

*where $u_\mu$ can be expressed in terms of the Euclidean projection*

$$u_\mu(x) = \pi\left(\bar{u} + \frac{Ax - b}{\mu}; Q\right).$$

*The gradient $\nabla f_\mu$ is a Lipschitz function with constant*

$$\ell_\mu = \frac{1}{\mu}\|A\|^2.$$

*Moreover,*

$$f_\mu(x) \le f(x) \le f_\mu(x) + \frac{\mu}{2}[D(\bar{u}; Q)]^2 \text{ for all } x \in \mathbb{R}^n, \tag{3.1.16}$$

*where $D(\bar{u}; Q)$ is the farthest distance from $\bar{u}$ to $Q$ defined by*

$$D(\bar{u}; Q) := \sup\{\|\bar{u} - u\| \mid u \in Q\}.$$

The following examples illustrate this result.

**Example 3.1.3.** Let $\|\cdot\|_{X_1}$ and $\|\cdot\|_{X_2}$ be two norms in $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively, and let $\|\cdot\|_{X_1^*}$ and $\|\cdot\|_{X_2^*}$ be the corresponding dual norms, i.e.,

$$\|x\|_{X_i^*} := \max\{\langle x, u\rangle \mid \|u\|_{X_i} \leq 1\},\ i = 1, 2.$$

Denote $\mathbb{B}_{X_1} := \{u \in \mathbb{R}^m \mid \|u\|_{X_1} \leq 1\}$ and $\mathbb{B}_{X_2} := \{u \in \mathbb{R}^n \mid \|u\|_{X_2} \leq 1\}$. Consider the function $g : \mathbb{R}^n \to \mathbb{R}$ defined by

$$g(x) := \|Ax - b\|_{X_1^*} + \lambda\|x\|_{X_2^*},$$

where $A$ is an $m \times n$ matrix, $b \in \mathbb{R}^m$, and $\lambda > 0$. Using the prox-function $d(u) = \frac{1}{2}\|u\|^2$, one finds the Nesterov smooth approximation of $g$ as:

$$g_\mu(x) = \frac{\|Ax - b\|^2}{2\mu} - \frac{\mu}{2}\left[d\left(\frac{Ax - b}{\mu}; \mathbb{B}_{X_1}\right)\right]^2 + \lambda\left(\frac{\|x\|^2}{2\mu} - \frac{\mu}{2}\left[d\left(\frac{x}{\mu}; \mathbb{B}_{X_2}\right)\right]^2\right).$$

The gradient of $g_\mu$ is

$$\nabla g_\mu(x) = A^\top \pi\left(\frac{Ax - b}{\mu}; \mathbb{B}_{X_1}\right) + \lambda\pi\left(\frac{x}{\mu}; \mathbb{B}_{X_2}\right),$$

and its Lipschitz constant is

$$L_\mu = \frac{\|A\|^2 + \lambda}{\mu}.$$

Moreover,

$$g_\mu(x) \leq g(x) \leq g_\mu(x) + \frac{\mu}{2}([D(0; \mathbb{B}_{X_1})]^2 + [D(0; \mathbb{B}_{X_2})]^2) \text{ for all } x \in \mathbb{R}^n.$$

For example, if $\|\cdot\|_{X_1}$ is the Euclidean norm, and $\|\cdot\|_{X_2}$ is the $\ell^\infty-$norm on $\mathbb{R}^n$, then

$$\nabla g_\mu(x) = A^\top \frac{Ax-b}{\max\{\|Ax-b\|,\mu\}} + \lambda \text{median}\left(\frac{x}{\mu}, e, -e\right),$$

where $e = [1,\dots,1]^\top \in \mathbb{R}^n$.

Let us provide another example involving *support vector machines*, well-known from machine learning. Our approach simplifies and improves the results in [**86**].

**Example 3.1.4.** Let $S := \{(X_i, y_i)\}_{i=1}^m$ be a training set, where $X_i \in \mathbb{R}^p$ is the $i$th row of a matrix $X$ and $y_i \in \{-1, 1\}$. The corresponding linear support vector machine problem can be reduced to solving the following problem:

$$\text{minimize } g(w) := \frac{1}{2}\|w\|^2 + \lambda \sum_{i=1}^m \ell_i(w), w \in \mathbb{R}^p,$$

where $\ell_i(w) = \max\{0, 1 - y_i X_i w\}$, $\lambda > 0$.

Let $Q := \{u \in \mathbb{R}^m \mid 0 \le u_i \le 1\}$ and define

$$f(w) := \sum_{i=1}^m \ell_i(w) = \max_{u\in Q} \langle e - YXw, u\rangle,$$

where $e = [1,\dots,1]^\top$ and $Y = \text{diag}(y)$ with $y = [y_1,\dots,y_m]^\top$.

Using the prox-function $d(u) = \frac{1}{2}\|u\|^2$, one has

$$f_\mu(w) = \max_{u\in Q}[\langle e - YXw, u\rangle - \mu d(u)].$$

Then

$$u_\mu(w) = \pi\left(\frac{e-YXw}{\mu}; Q\right) = \left\{u \in \mathbb{R}^m \;\middle|\; u_i = \text{median}\left(\frac{1-y_i X_i w}{\mu}, 0, 1\right)\right\}.$$

The gradient of $f_\mu$ is given by

$$\nabla f_\mu(w) = -(YX)^\top u_\mu(w),$$

and its Lipschitz constant is $\ell_\mu = \dfrac{\|YX\|^2}{\mu}$, where the matrix norm is defined in (3.1.14).

Moreover,

$$f_\mu(w) \le f(w) \le f_\mu(w) + \frac{m\mu}{2} \text{ for all } w \in \mathbb{R}^p.$$

Then we use the following smooth approximation of the original objective function $g$:

$$g_\mu(w) := \frac{1}{2}\|w\|^2 + \lambda f_\mu(w), w \in \mathbb{R}^p.$$

Obviously,

$$\nabla g_\mu(w) = w + \lambda \nabla f_\mu(w),$$

and a Lipschitz constant is

$$L_\mu = 1 + \lambda \frac{\|YX\|^2}{\mu}.$$

**3.1.2. Nesterov's Accelerated Gradient Method.** The smooth approximation obtained above is convenient for applying Nesterov's accelerated gradient method, presented as follows. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable convex function with Lipschitz gradient. This means that there exists $\ell > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \le \ell \|x - y\| \text{ for all } x, y \in \mathbb{R}^n.$$

Let $\Omega$ be a nonempty closed convex set. In his paper [**52**], Nesterov considered the optimization problem

$$\text{minimize } f(x) \text{ subject to } x \in \Omega.$$

For $x \in \mathbb{R}^n$, define

$$T_\Omega(x) := \arg\min \left\{ \langle \nabla f(x), y - x \rangle + \frac{\ell}{2}\|x - y\|^2 \;\middle|\; y \in \Omega \right\}.$$

Let $\rho : \mathbb{R}^n \to \mathbb{R}$ be a strongly convex function with parameter $\sigma > 0$ and let $x_0 \in \mathbb{R}^n$ such that

$$x_0 := \arg\min \{\rho(x) \mid x \in \Omega\}.$$

Further, assume that $\rho(x_0) = 0$. Then Nesterov's accelerated gradient method is outlined as follows.

**Algorithm 1. Nesterov's Accelerated Gradient Method**.

INPUT: $f$, $\ell$.

INITIALIZE: Choose $x_0 \in \Omega$.

Set $k = 0$

**Repeat the following**

Find $y_k := T_\Omega(x_k)$.

Find $z_k := \arg\min \left\{ \frac{\ell}{\sigma}\rho(x) + \sum_{i=0}^{k} \frac{i+1}{2}[f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] \;\middle|\; x \in \Omega \right\}$.

Set $x_{k+1} := \frac{2}{k+3} z_k + \frac{k+1}{k+3} y_k$.

Set $k := k + 1$.

**until a stopping criterion is satisfied.**

OUTPUT: $y_k$.

For simplicity, we choose $\rho(x) = \frac{\sigma}{2}\|x - x_0\|^2$, where $x_0 \in \Omega$ and $\sigma = 1$. Then the terms $y_k$ and $z_k$ in Algorithm 1 take on the following closed-form expressions. Following the proof of Theorem 3.1.2 (as provided in [**46**]), it is not hard to see that

$$y_k = T_\Omega(x_k) = \pi\left(x_k - \frac{\nabla f(x_k)}{\ell}; \Omega\right).$$

Moreover,

$$z_k = \pi\left(x_0 - \frac{1}{\ell}\sum_{i=0}^{k}\frac{i+1}{2}\nabla f(x_i);\Omega\right).$$

These closed-form expressions are essential to the practical implementation of the algorithms presented in the next section.

### 3.2. Generalized Fermat-Torricelli Problems Involving Points

Here we present our generalized version of the Fermat-Torricelli problem. Our aim is to apply Nesterov's smoothing technique and accelerated gradient method to this nonsmooth optimization problem. Our version of the problem is associated with the generalized notion of distance as generated by a Minkowski gauge.

Let $K$ be a nonempty, closed, bounded, and convex set in $\mathbb{R}^n$ containing the origin in its interior. Define the *Minkowski gauge* associated with $K$ by

$$\rho_K(x) := \inf\{t > 0 \mid x \in tK\}.$$

Note that, if $K$ is the closed unit ball in $\mathbb{R}^n$, then $\rho_K(x) = \|x\|$.

Given a nonempty and bounded set $F$, the *support function* associated with $F$ is given by

$$\sigma_F(x) := \sup\{\langle x, y\rangle \mid y \in F\}.$$

It follows from the definition of the Minkowski gauge (see, e.g., [**20**, Proposition 2.1]) that $\rho_F(x) = \sigma_{F^\circ}(x)$ and $\rho_{F^\circ}(x) = \sigma_F(x)$, where $F^\circ$ is the *polar* of $F$ defined by

$$F^\circ := \{y \in \mathbb{R}^n \mid \langle x, y\rangle \le 1 \text{ for all } x \in F\}.$$

Under these general notions of distance, we pursue the following generalized version of the Fermat-Torricelli problem. Let $\Omega$ be a nonempty closed convex subset of $\mathbb{R}^n$

and let $a_i \in \mathbb{R}^n$ for $i = 1, \ldots, m$. Our generalized version of the problem is:

$$\text{minimize } H(x) := \sum_{i=1}^{m} \sigma_F(x - a_i) \text{ subject to } x \in \Omega. \tag{3.2.17}$$

Let us start with some properties of the function $\sigma_F$. The following proposition can be proven easily.

**Proposition 3.2.1.** *For the function $\sigma_F$ defined as part of* (3.2.17)*, the following properties hold for all $u, v \in \mathbb{R}^n$ and $\lambda \geq 0$:*

**(i)** $|\sigma_F(u) - \sigma_F(v)| \leq \|F\|\|u - v\|$, *where* $\|F\| := \sup\{\|f\| \mid f \in F\}$.

**(ii)** $\sigma_F(u + v) \leq \sigma_F(u) + \sigma_F(v)$.

**(iii)** $\sigma_F(\lambda u) = \lambda \sigma_F(u)$, *and* $\sigma_F(u) = 0$ *if and only if* $u = 0$.

**(iv)** $\sigma_F$ *is a norm if we assume additionally that $F$ is symmetric, i.e., $F = -F$.*

**(v)** $\gamma\|u\| \leq \sigma_F(u)$, *where* $\gamma := \sup\{r > 0 \mid \mathbb{B}(0; r) \subset F\}$.

Let $\Theta$ be a nonempty closed convex subset of $\mathbb{R}^n$ and let $\bar{x} \in \Theta$. The *normal cone* in the sense of convex analysis to $\Theta$ at $\bar{x}$ is defined by

$$N(\bar{x}; \Theta) := \{v \in \mathbb{R}^n \mid \langle v, x - \bar{x} \rangle \leq 0 \text{ for all } x \in \Theta\}.$$

It follows from the definition that the normal cone mapping $N(\cdot; \Theta)$ has a closed graph in the sense that for any sequence $x_k \to \bar{x}$ and $v_k \to \bar{v}$ where $v_k \in N(x_k; \Theta)$, one has that $\bar{v} \in N(\bar{x}; \Theta)$.

Given an element $v \in \mathbb{R}^n$, we also define $\operatorname{cone}\{v\} := \{\lambda v \mid \lambda \geq 0\}$.

In what follows, we study the existence and uniqueness of the optimal solution of problem (3.2.17). The following definition and the proposition afterward are important for this purpose. We use $\operatorname{bd} F$ to denote the boundary of $F$.

**Definition 3.2.2.** *We say that $F$ is* normally smooth *if for every $x \in \operatorname{bd} F$ there exists $a_x \in \mathbb{R}^n$ such that $N(x;F) = \operatorname{cone}\{a_x\}$.*

**Example 3.2.3.** *For a given a positive definite matrix $A$, let*

$$\|x\|_A := \sqrt{x^\top Ax}.$$

*It is not hard to see that the set $F := \{x \in \mathbb{R}^n \mid \|x\|_A \leq 1\}$ is normally smooth. Indeed, $N(x;F) = \operatorname{cone}\{Ax\}$ if $\|x\|_A = 1$; see* [**41**, Proposition 2.48].

Define the set

$$\mathbb{B}_F^* := \{u \in \mathbb{R}^n \mid \sigma_F(u) \leq 1\}$$

and recall that a convex subset $\Theta$ of $\mathbb{R}^n$ is said to be *strictly convex* if $tu + (1-t)v \in \operatorname{int}\Theta$ whenever $u, v \in \Theta$, $u \neq v$, and $t \in (0,1)$.

**Proposition 3.2.4.** *We have that $F$ is normally smooth if and only if $\mathbb{B}_F^*$ is strictly convex.*

**Proof.** Suppose that $F$ is normally smooth. Fix any $u, v \in \mathbb{B}_F^*$ with $u \neq v$ and $t \in (0,1)$. Let us show that $tu+(1-t)v \in \operatorname{int}\mathbb{B}_F^*$, or equivalently, $\sigma_F(tu+(1-t)v) < 1$. We only need to consider the case where $\sigma_F(u) = \sigma_F(v) = 1$. Fix $\bar{x}, \bar{y} \in F$ such that

$$\langle u, \bar{x}\rangle = \sigma_F(u) = 1 \text{ and } \langle v, \bar{y}\rangle = \sigma_F(v) = 1,$$

and fix $e \in F$ such that

$$\langle tu + (1-t)v, e\rangle = \sigma_F(tu + (1-t)v).$$

It is obvious that $\sigma_F(tu + (1-t)v) \leq 1$. By contradiction, suppose that $\sigma_F(tu + (1-t)v) = 1$. Then

$$1 = \langle tu + (1-t)v, e\rangle = t\langle u, e\rangle + (1-t)\langle v, e\rangle \leq t\langle u, \bar{x}\rangle + (1-t)\langle v, \bar{y}\rangle = 1.$$

This implies $\langle u, e\rangle = \langle u, \bar{x}\rangle = 1 = \sigma_F(u)$ and $\langle v, e\rangle = \langle v, \bar{y}\rangle = 1 = \sigma_F(v)$. Then

$$\langle u, x\rangle \leq \langle u, e\rangle \text{ for all } x \in F,$$

which implies $u \in N(e; F)$. Similarly, $v \in N(e; F)$. Since $F$ is normally smooth, $u = \lambda v$, where $\lambda > 0$. Thus,

$$1 = \langle u, e\rangle = \langle \lambda v, e\rangle = \lambda \langle v, e\rangle = \lambda.$$

Hence $\lambda = 1$ and $u = v$, a contradiction.

Now suppose that $\mathbb{B}_F^*$ is strictly convex. Fix $\bar{x} \in \operatorname{bd} F$ and fix any $u, v \in N(\bar{x}; F)$ with $u, v \neq 0$. Let $\alpha := \sigma_F(u)$ and $\beta := \sigma_F(v)$. Then

$$\langle u, x\rangle \leq \langle u, \bar{x}\rangle \text{ for all } x \in F$$

and

$$\langle v, x\rangle \leq \langle v, \bar{x}\rangle \text{ for all } x \in F.$$

It follows that $\langle u, \bar{x}\rangle = \alpha$ and $\langle v, \bar{x}\rangle = \beta$. Moreover,

$$\sigma_F(u + v) \geq \langle u, \bar{x}\rangle + \langle v, \bar{x}\rangle = \alpha + \beta = \sigma_F(u) + \sigma_F(v),$$

and hence $\sigma_F(u + v) = \sigma_F(u) + \sigma_F(v)$. We have $u/\alpha, v/\beta \in \mathbb{B}_F^*$ and

$$\sigma_F\left(\frac{u}{\alpha}\frac{\alpha}{\alpha+\beta} + \frac{v}{\beta}\frac{\beta}{\alpha+\beta}\right) = 1.$$

Since $\mathbb{B}_F^*$ is strictly convex, one has $\dfrac{u}{\alpha} = \dfrac{v}{\beta}$, and hence $u = \lambda v$, where $\lambda := \alpha/\beta > 0$. The proof is now complete. □

**Remark 3.2.5.** Suppose that $F$ is normally smooth. It follows from the proof of Proposition 3.2.4 that for $u, v \in \mathbb{R}^n$ with $u, v \neq 0$, one has that $\sigma_F(u+v) = \sigma_F(u) + \sigma_F(v)$ if and only if $u = \lambda v$ for some $\lambda > 0$.

The proposition below gives sufficient conditions that guarantee the uniqueness of an optimal solution of (3.2.17).

**Proposition 3.2.6.** *Suppose that $F$ is normally smooth. If for any $x, y \in \Omega$ with $x \neq y$, the line connecting $x$ and $y$, $\mathcal{L}(x, y)$, does not contain at least one of the points $a_i$ for $i = 1, \ldots, m$, then problem* (3.2.17) *has a unique optimal solution.*

**Proof.** It is not hard to see that for any $\alpha \in \mathbb{R}$, the set

$$\mathcal{L}_\alpha := \{x \in \Omega \mid H(x) \leq \alpha\}$$

is compact, and so (3.2.17) has an optimal solution since $H$ is continuous. Let us show that the assumptions made guarantee that $H$ is strictly convex on $\Omega$, and hence (3.2.17) has a unique optimal solution.

By contradiction, suppose that there exist $\bar{x}, \bar{y} \in \Omega$ with $\bar{x} \neq \bar{y}$ and $t \in (0, 1)$ such that

$$H(t\bar{x} + (1-t)\bar{y}) = tH(\bar{x}) + (1-t)H(\bar{y}).$$

Then

$$\begin{aligned}\sigma_F(t(\bar{x} - a_i) + (1-t)(\bar{y} - a_i)) &= t\sigma_F(\bar{x} - a_i) + (1-t)\sigma_F(\bar{y} - a_i) \\ &= \sigma_F(t(\bar{x} - a_i)) + \sigma_F((1-t)(\bar{y} - a_i))\end{aligned}$$

for all $i = 1, \ldots, m$.

If $\bar{x} = a_i$ or $\bar{y} = a_i$, then $a_i$ is obviously contained in $\mathcal{L}(\bar{x}, \bar{y})$. Otherwise, by Remark 3.2.5, there exists $\lambda_i > 0$ such that

$$t(\bar{x} - a_i) = \lambda_i(1 - t)(\bar{y} - a_i).$$

This also implies that $a_i \in \mathcal{L}(\bar{x}, \bar{y})$. We have seen that $a_i \in \mathcal{L}(\bar{x}, \bar{y})$ for all $i = 1, \ldots, m$. This contradiction shows that (3.2.17) has a unique optimal solution. □

Let us consider the smooth approximation function given by

$$H_\mu(x) := \sum_{i=1}^{m} \left( \frac{\|x - a_i\|^2}{2\mu} + \langle x - a_i, \bar{u} \rangle - \frac{\mu}{2} \left[ d\left( \bar{u} + \frac{x - a_i}{\mu}; F \right) \right]^2 \right), \tag{3.2.18}$$

where $\bar{u} \in F$.

**Proposition 3.2.7.** *The function $H_\mu$ defined by* (3.2.18) *is continuously differentiable on $\mathbb{R}^n$ with its gradient given by*

$$\nabla H_\mu(x) = \sum_{i=1}^{m} \pi \left( \bar{u} + \frac{x - a_i}{\mu}; F \right).$$

*The gradient $\nabla H_\mu$ is a Lipschitz function with constant*

$$\mathcal{L}_\mu = \frac{m}{\mu}.$$

*Moreover, one has the following estimate*

$$H_\mu(x) \leq H(x) \leq H_\mu(x) + m\frac{\mu}{2}[D(\bar{u}; F)]^2 \text{ for all } x \in \mathbb{R}^n.$$

**Proof.** Given $b \in \mathbb{R}^n$, define the function on $\mathbb{R}^n$ given by

$$f(x) := \sigma_F(x - b) = \max\{\langle x - b, u \rangle \mid u \in F\}, x \in \mathbb{R}^n.$$

Consider the prox-function

$$d(u) := \frac{1}{2}\|u - \bar{u}\|^2.$$

Applying Theorem 3.1.2, one has that the function $f_\mu$ is continuously differentiable on $\mathbb{R}^n$ with its gradient given by

$$\nabla f_\mu(x) = u_\mu(x) = \pi\left(\bar{u} + \frac{x-b}{\mu}; F\right).$$

Moreover, the gradient $\nabla f_\mu$ is a Lipschitz function with constant

$$\ell_\mu = \frac{1}{\mu}.$$

The explicit formula for $f_\mu$ is

$$f_\mu(x) = \frac{\|x-b\|^2}{2\mu} + \langle x - b, \bar{u}\rangle - \frac{\mu}{2}\left[d\left(\bar{u} + \frac{x-b}{\mu}; F\right)\right]^2.$$

The conclusions then follow easily. □

We are now ready to write a pseudocode for solving the Fermat-Torricelli problem (3.2.17) via Nesterov's smoothing technique and accelerated gradient method.

**Algorithm 2**.

INPUT: $a_i$ for $i = 1, \ldots, m$, $\mu$.

INITIALIZE: Choose $x_0 \in \Omega$ and set $\ell = \dfrac{m}{\mu}$.

Set $k = 0$

**Repeat the following**

Compute $\nabla H_\mu(x_k) = \sum_{i=1}^{m} \pi\left(\bar{u} + \dfrac{x_k - a_i}{\mu}; F\right)$.

Find $y_k := \pi(x_k - \frac{1}{\ell}\nabla H_\mu(x_k); \Omega)$.

Find $z_k := \pi(x_0 - \frac{1}{\ell}\sum_{i=0}^{k} \dfrac{i+1}{2}\nabla H_\mu(x_i); \Omega)$.

Set $x_{k+1} := \dfrac{2}{k+3} z_k + \dfrac{k+1}{k+3} y_k$.

**until a stopping criterion is satisfied.**

**Remark 3.2.8.** When implementing Nesterov's accelerated gradient method, in order to get a more effective algorithm, instead of using a fixed smoothing parameter $\mu$, we often change $\mu$ during the iteration. The general optimization scheme is:

INITIALIZE: $x_0 \in \Omega$, $\mu_0 > 0$, $\mu_* > 0$, $\sigma \in (0, 1)$.

Set $k = 0$.

**Repeat the following**

Apply Nesterov's accelerated gradient method with $\mu = \mu_k$ and starting point $x_k$ to obtain an approximate solution $x_{k+1}$.

Update $\mu_{k+1} = \sigma\mu_k$.

**until $\mu \leq \mu_*$.**

**Example 3.2.9.** In the case where $F$ is the closed unit Euclidean ball, $\sigma_F(x) = \|x\|$ is the Euclidean norm and

$$\pi(x; F) = \begin{cases} \left\{\dfrac{x}{\|x\|}\right\}, & \|x\| > 1 \\ \{x\}, & \|x\| \leq 1. \end{cases}$$

Consider the $\ell_1$-norm on $\mathbb{R}^n$. For any $x \in \mathbb{R}^n$, one has

$$\|x\|_1 = \max\{\langle x, u\rangle \mid \|u\|_\infty \leq 1\},$$

In this case,

$$F = \{x \in \mathbb{R}^n \mid |x_i| \leq 1 \text{ for all } i = 1, \ldots, n\}.$$

The smooth approximation of the function $f(x) := \|x\|_1$ depends on the Euclidean projection to the set $F$, which can be found explicitly. In fact, for any $u \in \mathbb{R}^n$, one has

$$\pi(u; F) = \{v \in \mathbb{R}^n \mid v_i = \text{median } \{u_i, 1, -1\}\}.$$

Now we consider the $\ell_\infty$-norm in $\mathbb{R}^n$. For any $x \in \mathbb{R}^n$, one has

$$\|x\|_\infty = \max\{\langle x, u\rangle \mid \|u\|_1 \leq 1\}.$$

In this case,

$$F = \{x \in \mathbb{R}^n \mid \|x\|_1 \leq 1\}.$$

It is straightforward to find the Euclidean projection of a point to $F$ in two and three dimensions. In the case of high dimensions, there are available algorithms to find an approximation of the projection; see, e.g., [**17**].

**3.2.1. Numerical Examples.** To demonstrate the presented methods, let us consider a numerical examples below.

**Example 3.2.10.** The latitude/longitude coordinates in decimal format of 1217 US cities are recorded at `http://www.realestate3d.com/gps/uslatlongdegmin.htm`. We convert the longitudes provided by the website above from positive to negative to match with the real data. Our goal is to find a point that minimizes the sum of the distances to the given points representing the cities.

If we consider the case where $\sigma_F(x) = \|x\|$, the Euclidean norm, Algorithm 2 allows us to find an approximate optimal value $V^* \approx 23409.33$ and an approximate optimal solution $x^* \approx (38.63, -97.35)$. Similarly, if $\sigma_F(x) = \|x\|_1$, an approximate optimal value is $V^* \approx 28724.68$ and an approximate optimal solution is $x^* \approx (39.48, -97.22)$. With the same problem setup but considering the $\ell_\infty$-norm, an approximate optimal value is $V^* \approx 21987.76$ and an approximate optimal solution is $x^* \approx (37.54, -97.54)$.

The graph below shows the relation between the number of iterations $k$ and the optimal value $V_k = H(y_k)$ generated by different norms.

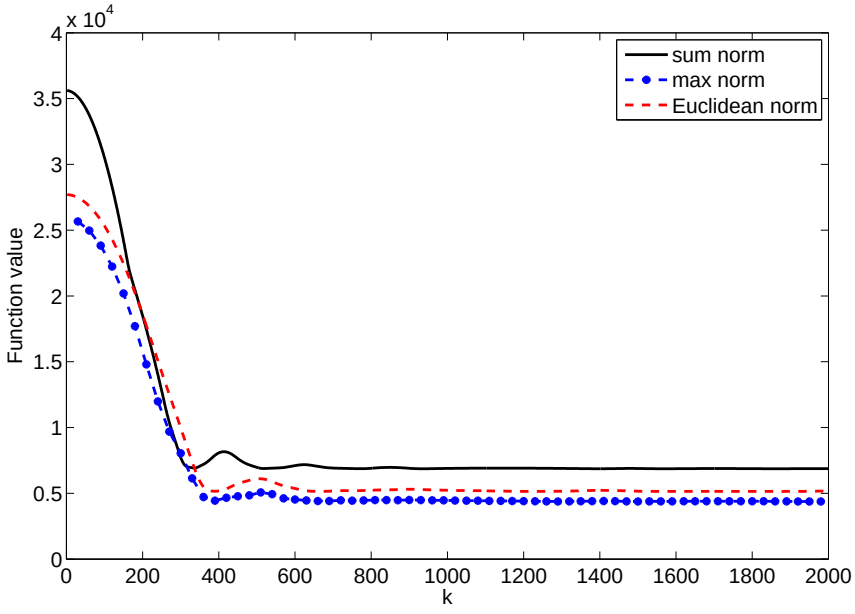



FIGURE 3.2. Generalized Fermat-Torricelli problems with different norms.

**3.2.2. Additional Work: Location Problems involving Sets.** The above techniques can be extended to apply to location problems where the targets are *sets*, as opposed to points. Such models may be appropriate when the targets have non-negligible sizes. This extension depends on an application of the minimization

majorization (MM) principle, which is described below. A detailed presentation of the algorithms and additional discussion can be found in [**46**].

3.2.2.1. *MM Principle.* This section describes an important tool of convex optimization and computational statistics called the *MM Principle (minimization majorization)*; see [**15, 25, 32**] and the references therein. Here we provide a more general version (as we apply in [**46**]) for generalized Fermat-Torricelli problems when the targets are sets rather than points. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and let $\Omega$ be a nonempty closed convex subset of $\mathbb{R}^n$. Consider the optimization problem

$$\text{minimize } f(x) \text{ subject to } x \in \Omega. \tag{3.2.19}$$

Let $\mathcal{M} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ and let $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^p$ be a set-valued mapping with nonempty values such that the following properties hold for all $x, y \in \mathbb{R}^n$:

$$f(x) \le \mathcal{M}(x, z) \; \forall z \in F(y), \text{ and } f(x) = \mathcal{M}(x, z) \; \forall z \in F(x).$$

Given $x_0 \in \Omega$, the MM algorithm to solve (3.2.19) is given by

$$\text{Choose } z_k \in F(x_k) \text{ and find } x_{k+1} \in \arg\min\{\mathcal{M}(x, z_k) \mid x \in \Omega\}.$$

Then

$$f(x_{k+1}) \le \mathcal{M}(x_{k+1}, z_k) \le \mathcal{M}(x_k, z_k) = f(x_k).$$

Finding an appropriate majorization is an important step in this algorithm. It has been shown in [**14**] that the MM Principle provides an effective tool for solving the generalized Fermat-Torricelli problem. Figure 3.3 below illustrates the MM principle as applied to the generalized Fermat-Torricelli problem with sets $\Omega_1, \Omega_2$, and $\Omega_3$.

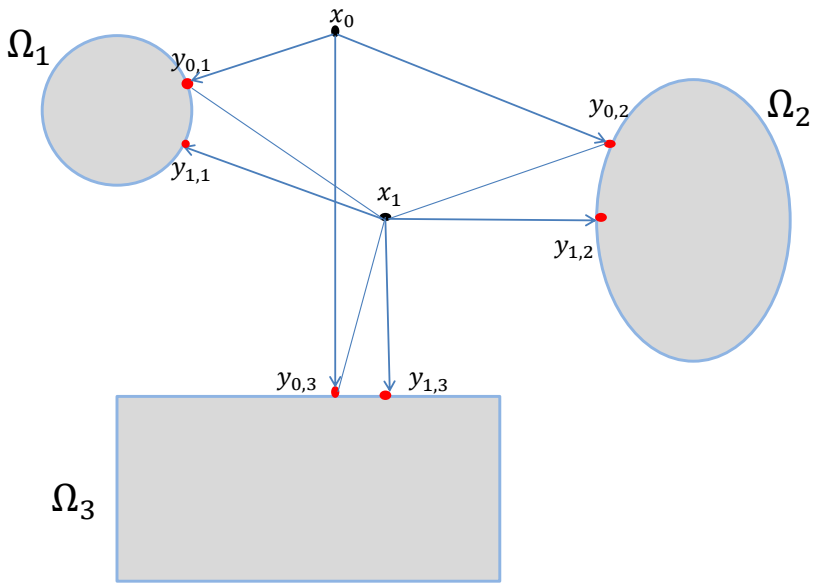



FIGURE 3.3. The first steps of an application of the MM principle for a generalized Fermat-Torricelli problem with sets. The initial guess $x_0$ is projected onto the sets and the Fermat-Torricelli problem is solved using those points as the targets, resulting in the next iterate $x_1$.

### 3.3. Multifacility Location Problems and Non-convex Optimization

In this section we develop algorithms to solve a version of the generalized Fermat-Torricelli problem that is both nonsmooth and non-convex. Specifically, we consider a *multifacility* location problem (i.e. a facility location problem with multiple centers) where the targets can take on both positive and negative weights and where distance is determined by Minkowski gauges. Using the Nesterov smoothing technique and an algorithm for minimizing differences of convex functions introduced by Tao and An, effective algorithms are developed for solving these problems.

**3.3.1. Introduction to Multifacility Location.** To solve the classical Fermat-Torricelli problem, one finds a point that minimizes the sum of the Euclidean distances to three points in the plane. This problem was introduced by Pierre de Fermat in the 17th century and originally solved by Evangelista Torricelli. A more general model asks for a point that minimizes the sum of the Euclidean distances to a finite number of points in a finite dimensional Euclidean space. In spite of its simplicity, this problem has been a topic of extensive research due to both its mathematical beauty and practical applications in the field of *facility location*. The first algorithm for solving the Fermat-Torricelli problem was introduced in 1937 by Weiszfeld in [**83**]. This algorithm was studied in depth by Kuhn in [**29**]. The Fermat-Torricelli problem and Weiszeld's algorithm have been revisited and further studied by many authors; see, e.g., [**10, 16, 30, 81**] and the references therein.

Several generalized models for the Fermat-Torricelli problem have been introduced and studied in the literature. The Fermat-Torricelli problem in general normed spaces was considered in [**35**]. The generalized Fermat-Torricelli problems involving Minkowski gauges and distances to convex sets were the topics of [**26, 40, 46, 48**]. In particular, the recent paper by Nam, An, Rector and Sun [**46**] focused on numerical algorithms with the use of the Nesterov smoothing technique and accelerated gradient method to study these problems.

Given the locations of a finite number of "customers", the *multifacility location problem* asks for the optimal locations of a finite number of "facilities" (also known as centroids) to serve these customers, where each customer is assigned to the nearest facility. The multifacility location problem has a close relationship to clustering problems. A recent paper by An, Belghiti, and Tao [**2**] uses the so-called DCA (Difference of Convex functions Algorithm) to solve a clustering problem that involves

squared Euclidean distances. Their method shows robustness, efficiency, and superiority compared with the well-known $K-$means algorithm, when applied to a number of real-world data sets. The DCA was introduced by Tao in 1986, and then extensively developed in the works of An, Tao, and others; see [**79, 80**] and the references therein. An important feature of the DCA is its simplicity compared with other methods, while still being very effective for many applications. In fact, the DCA is one of the most successful algorithms for solving non-convex optimization problems.

We consider the weighted Fermat-Torricelli problem with both positive and negative weights. Additionally, we consider a continuous multifacility location problem, which involves distance measurements generated by Minkowski gauges. Considering Minkowski gauges, it is possible to unify the problems generated by arbitrary norms and even more generalized notions of distances; see [**26, 40, 46**] and the references therein. Our approach is based on the Nesterov smoothing technique [**52**] and the DCA. We also propose a method to solve a new model of clustering called *set clustering*. This model involves squared Euclidean distances to convex sets, and hence coincides with the model considered in [**2**] when the sets reduce to singletons. Using sets instead of points allows us to classify objects with non-negligible sizes.

The remainder of this section is organized as follows. In Section 3.3.2, we give an accessible presentation of DC programming and the DCA. Section 3.3.3 is devoted to developing algorithms to solve generalized weighted Fermat-Torricelli problems involving possibly negative weights and Minkowski gauges. Algorithms for solving multifacility location problems with Minkowski gauges are presented in Section 3.3.4. We demonstrate the effectiveness of our algorithms through a variety of numerical examples in Section 3.3.5. In Section 3.3.6, we introduce and develop an algorithm to solve the set clustering model.

**3.3.2. Tools of DC Programming.** This section provides background on DC programming and the DCA for the convenience of the reader. Most of the results in this section can be found in [**79, 80**], although the present presentation is tailored to the algorithms presented in the subsequent sections.

Consider the problem

$$\text{minimize } f(x) := g(x) - h(x), x \in \mathbb{R}^n, \tag{3.3.20}$$

where $g\colon \mathbb{R}^n \to (-\infty, +\infty]$ and $h\colon \mathbb{R}^n \to \mathbb{R}$ are convex functions. The function $f$ in (3.3.20) is called a *DC function* and $g - h$ is called a *DC decomposition* of $f$.

For a convex function $\varphi\colon \mathbb{R}^n \to (-\infty, +\infty]$, the *Fenchel conjugate* of $\varphi$ is defined by

$$\varphi^*(y) := \sup\{\langle y, x\rangle - \varphi(x) \mid x \in \mathbb{R}^n\}.$$

Note that, if $\varphi$ is *proper*, i.e. $\text{dom}\,(\varphi) := \{x \in \mathbb{R}^n \mid \varphi(x) < +\infty\} \neq \emptyset$, then $\varphi^*\colon \mathbb{R}^n \to (-\infty, +\infty]$ is also a convex function. Given $\bar{x} \in \text{dom}\,(\varphi)$, an element $v \in \mathbb{R}^n$ is called a *subgradient* of $\varphi$ at $\bar{x}$ if

$$\langle v, x - \bar{x}\rangle \leq \varphi(x) - \varphi(\bar{x}) \text{ for all } x \in \mathbb{R}^n.$$

The collection of all subgradients of $\varphi$ at $\bar{x}$ is called the *subdifferential* of $\varphi$ at this point and is denoted by $\partial\varphi(\bar{x})$. If $\varphi$ is proper and lower semicontinuous, then $v \in \partial\varphi^*(y)$ if and only if $y \in \partial\varphi(v)$; see, e.g., [**22, 41, 64**].

Introduced by Tao and An [**79, 80**], the DCA is a simple but effective optimization scheme for minimizing differences of convex functions. Although the algorithm is used for non-convex optimization problems, the convexity of the functions involved still plays a crucial role with the presence of elements of convex analysis such as subgradients and Fenchel conjugates. The algorithm is summarized below, as applied

to problem (3.3.20). As can be seen, a "zig-zag" approach is taken between the primal variables $x_k$ and the dual variables $y_k$. The result is a sequence $x_k$ that descends in value of the objective function $f$ at each step.

**DC Algorithm 1**.

INPUT: $x_1 \in \mathbb{R}^n$, $N \in I\!N$.

**for** $k = 1, \ldots, N$ **do**

Find $y_k \in \partial h(x_k)$.

Find $x_{k+1} \in \partial g^*(y_k)$.

**end for**

OUTPUT: $x_{N+1}$.

The convergence properties of this algorithm are discussed in Theorem 3.3.1.

Let $g, h \colon \mathbb{R}^n \to (-\infty, +\infty]$ be proper, lower semicontinuous, and convex functions. It is well-known that $v \in \partial g^*(y)$ if and only if

$$v \in \operatorname{argmin} \left\{ g(x) - \langle y, x \rangle \mid x \in \mathbb{R}^n \right\}. \tag{3.3.21}$$

Moreover, $w \in \partial h(x)$ if and only if

$$w \in \operatorname{argmin} \left\{ h^*(y) - \langle y, x \rangle \mid y \in \mathbb{R}^n \right\}. \tag{3.3.22}$$

Thus, in the case where we cannot find $y_k$ or $x_{k+1}$ exactly in Algorithm 1, we can find them approximately by solving two convex optimization problems in each iteration, as in the algorithm below.

**DC Algorithm 2**.

INPUT: $x_1 \in \mathbb{R}^n, N \in I\!N$.

**for** $k = 1, \ldots, N$ **do**

Find $y_k \in \partial h(x_k)$ or find $y_k$ approximately by solving the problem:

$$\text{minimize } \psi_k(y) := h^*(y) - \langle y, x_k \rangle, \ y \in \mathbb{R}^n.$$

Find $x_{k+1} \in \partial g^*(y_k)$ or find $x_{k+1}$ approximately by solving the problem:

$$\text{minimize } \phi_k(x) := g(x) - \langle y_k, x \rangle, \ x \in \mathbb{R}^n.$$

**end for**

OUTPUT: $x_{N+1}$.

Let us now discuss the convergence of the DCA. Recall that a function $h\colon \mathbb{R}^n \to (-\infty, +\infty]$ is called $\gamma$-*convex* ($\gamma \geq 0$) if the function defined by $k(x) := h(x) - \frac{\gamma}{2}\|x\|^2$, $x \in \mathbb{R}^n$, is convex. As mentioned previously, if there exists $\gamma > 0$ such that $h$ is $\gamma-$convex, then $h$ is called *strongly convex*.

We say that an element $\bar{x} \in \mathbb{R}^n$ is a *critical point* of the function $f$ from (3.3.20) if

$$\partial g(\bar{x}) \cap \partial h(\bar{x}) \neq \emptyset.$$

Obviously, in the case where both $g$ and $h$ are differentiable, $\bar{x}$ is a critical point of $f$ if and only if $\bar{x}$ satisfies the Fermat rule $\nabla f(\bar{x}) = 0$.

The theorem below provides a convergence result for the DCA. The result can be derived directly from [**80**, Theorem 3.7].

**Theorem 3.3.1.** *Consider the function $f$ defined in* (3.3.20) *and the sequence $\{x_k\}$ generated by Algorithm 1. Then the following properties are valid:*

**(i)** *If $g$ is $\gamma_1$-convex and $h$ is $\gamma_2$-convex, then*

$$f(x_k) - f(x_{k+1}) \geq \frac{\gamma_1 + \gamma_2}{2}\|x_{k+1} - x_k\|^2 \text{ for all } k \in \mathbb{N}. \tag{3.3.23}$$

**(ii)** *The sequence $\{f(x_k)\}$ is monotone decreasing.*

**(iii)** *If $f$ is bounded from below, $g$ is lower semicontinuous, $g$ is $\gamma_1$-convex and $h$ is $\gamma_2$-convex with $\gamma_1 + \gamma_2 > 0$, and $\{x_k\}$ is bounded, then every subsequential limit of the sequence $\{x_k\}$ is a critical point of $f$.*

The following propositions serve as a discussion for the constructibility of the sequence $\{x_k\}$, which give sufficient conditions for [**80**, Lemma 3.6].

**Proposition 3.3.2.** *Let $g\colon \mathbb{R}^n \to (-\infty, +\infty]$ be a proper, lower semicontinuous, and convex function. Then*

$$\partial g(\mathbb{R}^n) := \bigcup_{x \in \mathbb{R}^n} \partial g(x) = \operatorname{dom} \partial g^* := \{y \in \mathbb{R}^n \mid \partial g^*(y) \neq \emptyset\}.$$

**Proof.** Let $x \in \mathbb{R}^n$ and $y \in \partial g(x)$. Then $x \in \partial g^*(y)$, which implies $\partial g^*(y) \neq \emptyset$, and so $y \in \operatorname{dom} \partial g^*$. The opposite inclusion follows by a similar argument. □

We say that a function $g\colon \mathbb{R}^n \to (-\infty, +\infty]$ is *coercive of superior order* if

$$\lim_{\|x\| \to +\infty} \frac{g(x)}{\|x\|} = +\infty.$$

**Proposition 3.3.3.** *Let $g\colon \mathbb{R}^n \to (-\infty, +\infty]$ be a proper, coercive of superior order, and convex function. Then* $\operatorname{dom}(\partial g^*) = \mathbb{R}^n$.

**Proof.** By [**22**, Proposition 1.3.8] and the fact that $f$ is proper, the Fenchel conjugate $g^*$ is a finite convex function. Therefore, $\partial g^*(y)$ is nonempty for all $y \in \mathbb{R}^n$, which completes the proof. □

**3.3.3. The DCA for a Generalized Multifacility Location Problem.** In this section we develop algorithms for solving weighted Fermat-Torricelli problems involving Minkowski gauges. In particular, the algorithms developed here are applicable to solving the unweighted version introduced and studied in [**26**]. The present method is based on the Nesterov smoothing technique and the DCA. This approach allows us to solve generalized versions of the Fermat-Torricelli problem generated by different norms and generalized distances.

Let $F$ be a nonempty, closed, bounded, and convex set in $\mathbb{R}^n$ containing the origin in its interior. Recall that the Minkowski gauge associated with $F$ is defined by

$$\rho_F(x) := \inf\{t > 0 \mid x \in tF\}.$$

Note that, if $F$ is the closed unit ball in $\mathbb{R}^n$, then $\rho_F(x) = \|x\|$.

Given a nonempty and bounded set $K$, recall that the support function associated with $K$ is given by

$$\sigma_K(x) := \sup\{\langle x, y\rangle \mid y \in K\}.$$

It follows from the definition of the Minkowski function (see, e.g., [**20**, Proposition 2.1]) that $\rho_F(x) = \sigma_{F^\circ}(x)$, where

$$F^\circ := \{y \in \mathbb{R}^n \mid \langle x, y\rangle \leq 1 \text{ for all } x \in F\}.$$

Let us present below a direct consequence of the Nesterov smoothing technique given in [**52**]. In the proposition below, $d(x;\Omega)$ denotes the Euclidean distance and $P(x;\Omega)$ denotes the Euclidean projection from a point $x$ to a nonempty, closed, and convex set $\Omega$ in $\mathbb{R}^n$.

**Proposition 3.3.4.** *Given any $a \in \mathbb{R}^n$ and $\mu > 0$, a Nesterov smoothing approximation of $\varphi(x) := \rho_F(x-a)$ has the representation*

$$\varphi_\mu(x) = \frac{1}{2\mu}\|x-a\|^2 - \frac{\mu}{2}\left[d\left(\frac{x-a}{\mu}; F^\circ\right)\right]^2.$$

*Moreover, $\nabla\varphi_\mu(x) = P\left(\frac{x-a}{\mu}; F^\circ\right)$ and*

$$\varphi_\mu(x) \le \varphi(x) \le \varphi_\mu(x) + \frac{\mu}{2}\|F^\circ\|^2, \tag{3.3.24}$$

*where $\|F^\circ\| := \sup\{\|u\| \mid u \in F^\circ\}$.*

**Proof.** The function $\varphi$ can be represented as

$$\varphi(x) = \sigma_{F^\circ}(x-a) = \sup\{\langle x-a, u\rangle \mid u \in F^\circ\}.$$

Using the prox-function $d(x) = \frac{1}{2}\|x\|^2$ in [**52**], one obtains a smooth approximation of $\varphi$ given by

$$\begin{aligned}
\varphi_\mu(x) &:= \sup\left\{\langle x-a, u\rangle - \frac{\mu}{2}\|u\|^2 \mid u \in F^\circ\right\} \\
&= \sup\left\{-\frac{\mu}{2}\left(\|u\|^2 - \frac{2}{\mu}\langle x-a, u\rangle\right) \;\middle|\; u \in F^\circ\right\} \\
&= \sup\left\{-\frac{\mu}{2}\left\|u - \frac{1}{\mu}(x-a)\right\|^2 + \frac{1}{2\mu}\|x-a\|^2 \;\middle|\; u \in F^\circ\right\} \\
&= \frac{1}{2\mu}\|x-a\|^2 - \frac{\mu}{2}\inf\left\{\left\|u - \frac{1}{\mu}(x-a)\right\|^2 \;\middle|\; u \in F^\circ\right\} \\
&= \frac{1}{2\mu}\|x-a\|^2 - \frac{\mu}{2}\left[d\left(\frac{x-a}{\mu}; F^\circ\right)\right]^2.
\end{aligned}$$

The formula for computing the gradient of $\varphi_\mu$ follows from the well-known gradient formulas for the squared Euclidean norm and the squared distance function generated

by a nonempty, closed, and convex set: $\nabla d^2(x;\Omega) = 2[x - P(x;\Omega)]$; see, e.g., [**41**, Exercise 3.2]. The estimate (3.3.24) can be proved directly. □

Let $a^i \in \mathbb{R}^n$ for $i = 1, \ldots, m$ and let $c_i \neq 0$ for $i = 1, \ldots, m$ be real numbers. In this setup, the points $a^i$ will represent the targets and the numbers $c_i$ represent the weights. For the remainder of this section, we study the following generalized version of the Fermat-Torricelli problem:

$$\text{minimize } f(x) := \sum_{i=1}^{m} c_i \rho_F(x - a^i),\ x \in \mathbb{R}^n. \tag{3.3.25}$$

The function $f$ in (3.3.25) can be written as

$$f(x) = \sum_{c_i > 0} c_i \rho_F(x - a^i) - \sum_{c_i < 0} (-c_i) \rho_F(x - a^i).$$

Let $I := \{i : c_i > 0\}$ and $J := \{i \mid c_i < 0\}$ with $\alpha_i = c_i$ if $i \in I$, and $\beta_i = -c_i$ if $i \in J$. Then

$$f(x) = \sum_{i \in I} \alpha_i \rho_F(x - a^i) - \sum_{j \in J} \beta_j \rho_F(x - a^j). \tag{3.3.26}$$

An essential step in applying DC Algorithm 1 for minimizing a function $f$ represented as the difference of two convex functions $g$ and $h$ is to find subgradients of $g^*$. The function $f$ given in (3.3.26) has the obvious DC decomposition $f = g - h$, where

$$g(x) := \sum_{i \in I} \alpha_i \rho_F(x - a^i) \qquad \text{and} \qquad h(x) := \sum_{j \in J} \beta_j \rho_F(x - a^j).$$

However, there is no explicit formula for subgradients of this $g^*$, and hence we cannot apply DC Algorithm 1. The following Proposition 3.3.5 gives a Nesterov-type approximation for the function $f$, which is more favorable for applying this algorithm.

**Proposition 3.3.5.** *Consider the function $f$ defined in* (3.3.26). *Given any $\mu > 0$, an approximation of the function $f$ has the following DC decomposition:*

$$f_\mu(x) := g_\mu(x) - h_\mu(x),\ x \in \mathbb{R}^n, \tag{3.3.27}$$

*where*

$$g_\mu(x) := \sum_{i\in I} \frac{\alpha_i}{2\mu}\|x - a^i\|^2,$$
$$h_\mu(x) := \sum_{i\in I} \frac{\mu\alpha_i}{2}\left[d\left(\frac{x-a^i}{\mu}; F^\circ\right)\right]^2 + \sum_{j\in J}\beta_j\rho_F(x-a^j).$$

*Moreover, $f_\mu(x) \le f(x) \le f_\mu(x) + \frac{\mu\|F^\circ\|^2}{2}\sum_{i\in I}\alpha_i$ for all $x \in \mathbb{R}^n$.*

**Proof.** By Proposition 3.3.4,

$$\begin{aligned} f_\mu(x) &= \sum_{i\in I}\left[\frac{\alpha_i}{2\mu}\|x-a^i\|^2 - \frac{\mu\alpha_i}{2}\left[d\left(\frac{x-a^i}{\mu}; F^\circ\right)\right]^2\right] - \sum_{j\in J}\beta_j\rho_F(x-a_j) \\ &= \sum_{i\in I}\frac{\alpha_i}{2\mu}\|x-a^i\|^2 - \left[\sum_{i\in I}\frac{\mu\alpha_i}{2}\left[d\left(\frac{x-a^i}{\mu}; F^\circ\right)\right]^2 + \sum_{j\in J}\beta_j\rho_F(x-a^j)\right]. \end{aligned}$$

The inequality estimate follows directly from (3.3.24). □

**Proposition 3.3.6.** *Let $\gamma_1 := \sup\{r > 0 \mid B(0;r) \subset F\}$ and $\gamma_2 := \inf\{r > 0 \mid F \subset B(0;r)\}$. Suppose that*

$$\gamma_1 \sum_{i\in I}\alpha_i > \gamma_2 \sum_{j\in J}\beta_j.$$

*Then the function $f$ defined in* (3.3.26) *and its approximation $f_\mu$ defined in* (3.3.27) *have absolute minima.*

**Proof.** Fix any $r > 0$ such that $B(0;r) \subset F$. By the definition, for any $x \in \mathbb{R}^n$,

$$\begin{aligned}\rho_F(x) &= \inf\{t > 0 \mid t^{-1}x \in F\} \leq \inf\{t > 0 \mid t^{-1}x \in B(0;r)\} \\ &= \inf\{t > 0 \mid r^{-1}\|x\| < t\} = r^{-1}\|x\|.\end{aligned}$$

This implies $\rho_F(x) \leq \gamma_1^{-1}\|x\|$. Similarly, $\rho_F(x) \geq \gamma_2^{-1}\|x\|$.

Then

$$\begin{aligned}\sum_{i\in I} \alpha_i \rho_F(x - a^i) &\geq \gamma_2^{-1} \sum_{i\in I} \alpha_i \|x - a^i\| \geq \gamma_2^{-1} \sum_{i\in I} \alpha_i \left(\|x\| - \|a^i\|\right), \\ \sum_{j\in J} \beta_j \rho_F(x - a^j) &\leq \gamma_1^{-1} \sum_{j\in J} \beta_j \left(\|x\| + \|a^j\|\right).\end{aligned}$$

It follows that

$$f(x) \geq \left[(\gamma_2)^{-1} \sum_{i\in I} \alpha_i - (\gamma_1)^{-1} \sum_{j\in J} \beta_j\right] \|x\| - c,$$

where $c := \gamma_2^{-1} \sum_{i\in I} \alpha_i \|a^i\| + \gamma_1^{-1} \sum_{j\in J} \beta_j \|a^j\|$.

The assumption guarantees that $\lim_{\|x\|\to+\infty} f(x) = +\infty$, and so $f$ has an absolute minimum.

By Proposition 3.3.5,

$$f(x) \leq f_\mu(x) + \frac{\mu\|F^\circ\|^2}{2} \sum_{i\in I} \alpha_i.$$

This implies that $\lim_{\|x\|\to+\infty} f_\mu(x) = +\infty$, and so $f_\mu$ has an absolute minimum as well. □

**Remark 3.3.7.** We see that $h_\mu$ from (3.3.27) is the sum of a smooth function and a nonsmooth function. We can calculate the gradient of the smooth term as follows. Define

$$h_\mu^1(x) := \sum_{i\in I} \frac{\mu\alpha_i}{2} \left[d\left(\frac{x - a^i}{\mu}; F^\circ\right)\right]^2, \qquad h_\mu^2(x) := \sum_{j\in J} \beta_j \rho_F(x - a^j).$$

Then $h_\mu = h_\mu^1 + h_\mu^2$ and $h_\mu^1$ is differentiable with

$$\nabla h_\mu^1(x) = \sum_{i \in I} \alpha_i \left[ \frac{x - a^i}{\mu} - P\left( \frac{x - a^i}{\mu}; F^\circ \right) \right].$$

The next proposition gives us a formula for the gradient of $g_\mu^*$, as per (3.3.21).

**Proposition 3.3.8.** *Consider the function $g_\mu$ defined in Proposition 3.3.5. For any $y \in \mathbb{R}^n$, the function*

$$\phi_\mu(x) := g_\mu(x) - \langle y, x \rangle, x \in \mathbb{R}^n,$$

*has a unique minimizer given by*

$$x = \frac{y + \sum_{i \in I} \alpha_i a^i / \mu}{\sum_{i \in I} \alpha_i / \mu}.$$

**Proof.** The gradient of the convex function $\phi_\mu$ is given by

$$\nabla \phi_\mu(x) = \sum_{i \in I} \frac{\alpha_i}{\mu}(x - a^i) - y.$$

The result then follows by solving $\nabla \phi_\mu(x) = 0$. □

Based on DC Algorithm 1, the DC decomposition (3.3.27), Remark 3.3.7, and (3.3.21), we present the following DC Algorithm 3 to solve the generalized Fermat-Torricelli problem (3.3.25). As can be seen by the use of the subdifferential, this algorithm still retains some of the nonsmooth structure of the original problem, but the calculation of subgradients of the original $g^*$ is avoided, as intended. Following DC Algorithm 3, we develop another algorithm which does not require the use of the subdifferential.

**DC Algorithm 3**.

INPUT: $\mu > 0$, $x_1 \in \mathbb{R}^n$, $N \in I\!N$, $F$, $a^1, \ldots, a^m \in \mathbb{R}^n$, $c_1, \ldots, c_m \in \mathbb{R}$.

**for** $k = 1, \ldots, N$ **do**

Find $y_k = u_k + v_k$, where

$$u_k := \sum_{i \in I} \alpha_i \left[ \frac{x_k - a^i}{\mu} - P\left( \frac{x_k - a^i}{\mu}; F^\circ \right) \right],$$

$$v_k \in \sum_{j \in J} \beta_j \partial \rho_F (x_k - a^j).$$

Find $x_{k+1} = \frac{y_k + \sum_{i \in I} \alpha_i a^i / \mu}{\sum_{i \in I} \alpha_i / \mu}$.

OUTPUT: $x_{N+1}$.

Let us introduce another algorithm to solve the problem. This algorithm is obtained by using the Nesterov smoothing method for all functions involved in the problem and in the following proposition. The proof of the proposition follows directly from Proposition 3.3.4, as in the proof of Proposition 3.3.5.

**Proposition 3.3.9.** *Consider the function $f$ defined in* (3.3.26). *Given any $\mu > 0$, a smooth approximation of the function $f$ has the following DC decomposition:*

$$f_\mu(x) := g_\mu(x) - h_\mu(x), \ x \in \mathbb{R}^n,$$

*where*

$$g_\mu(x) := \sum_{i \in I} \frac{\alpha_i}{2\mu} \|x - a^i\|^2,$$

$$h_\mu(x) := \sum_{j \in J} \frac{\beta_j}{2\mu} \|x - a^j\|^2 - \sum_{j \in J} \frac{\mu \beta_j}{2} \left[ d\left( \frac{x - a^j}{\mu}; F^\circ \right) \right]^2 + \sum_{i \in I} \frac{\mu \alpha_i}{2} \left[ d\left( \frac{x - a^i}{\mu}; F^\circ \right) \right]^2.$$

*Moreover,*

$$f_\mu(x) - \frac{\mu \|F^\circ\|^2}{2} \sum_{i \in I} \beta_i \leq f(x) \leq f_\mu(x) + \frac{\mu \|F^\circ\|^2}{2} \sum_{i \in I} \alpha_i$$

*for all* $x \in \mathbb{R}^n$.

Note that both functions $g_\mu$ and $h_\mu$ in Proposition 3.3.9 are smooth with the gradients given by

$$\begin{aligned}
\nabla g_\mu(x) &= \sum_{i\in I} \frac{\alpha_i}{\mu}(x-a^i),\\
\nabla h_\mu(x) &= \sum_{j\in J} \frac{\beta_j}{\mu}(x-a^j) - \sum_{j\in J}\beta_j\left[\frac{x-a^j}{\mu} - P\left(\frac{x-a_j}{\mu};F^\circ\right)\right]\\
&\quad + \sum_{i\in I}\alpha_i\left[\frac{x-a^i}{\mu} - P\left(\frac{x-a^i}{\mu};F^\circ\right)\right]\\
&= \sum_{j\in J}\beta_j\left[P\left(\frac{x-a^j}{\mu};F^\circ\right)\right] + \sum_{i\in I}\alpha_i\left[\frac{x-a^i}{\mu} - P\left(\frac{x-a^i}{\mu};F^\circ\right)\right].
\end{aligned}$$

Based on the same approach as in DC Algorithm 3, we obtain another algorithm for solving problem (3.3.25).

**DC Algorithm 4**.

INPUT: $\mu > 0$, $x_1 \in \mathbb{R}^n$, $N \in I\!N$, $F$, $a^1,\ldots,a^m \in \mathbb{R}^n$, $c_1,\ldots,c_m \in \mathbb{R}$.

**for** $k = 1,\ldots,N$ **do**

Find $y_k = u_k + v_k$, where

$$u_k := \textstyle\sum_{i\in I}\alpha_i\left[\frac{x_k-a^i}{\mu} - P\left(\frac{x_k-a^i}{\mu};F^\circ\right)\right],$$

$$v_k := \textstyle\sum_{j\in J}\beta_j\left[P\left(\frac{x_k-a^j}{\mu};F^\circ\right)\right],$$

Find $x_{k+1} = \frac{y_k+\sum_{i\in I}\alpha_i a^i/\mu}{\sum_{i\in I}\alpha_i/\mu}$.

OUTPUT: $x_{N+1}$.

**Remark 3.3.10.** When implementing Algorithm 3 and Algorithm 4, instead of using a fixed smoothing parameter $\mu$, we often change $\mu$ during the iteration. The general optimization scheme is

INITIALIZE: $x_1 \in \mathbb{R}^n$, $\mu_0 > 0$, $\mu_* > 0$, $0 < \sigma < 1$.

Set $k = 1$.

**Repeat the following**

Apply DC Algorithm 3 (or DC Algorithm 4) with $\mu = \mu_k$ and starting point $x_k$ to obtain an approximate solution $x_{k+1}$.

Update $\mu_{k+1} = \sigma\mu_k$.

**Until** $\mu_k \leq \mu_*$.

**3.3.4. Multifacility Location.** In this section, we consider multifacility location problems involving distances generated by Minkowski gauges. Given $a^i \in \mathbb{R}^n$ for $i = 1, \ldots, m$, we need to choose $x^\ell$ for $\ell = 1, \ldots, k$ in $\mathbb{R}^n$ as centroids and assign each member $a^i$ to its closest centroid. The objective function to be minimized is the sum of the assignment distances:

(3.3.28)

$$\text{minimize } f(x^1, \ldots, x^k) = \sum_{i=1}^{m} \min_{\ell=1,\ldots,k} \rho_F(x^\ell - a^i), \qquad x^\ell \in \mathbb{R}^n, \ell = 1, \ldots, k.$$

Let us first discuss the existence of an optimal solution.

**Proposition 3.3.11.** *The optimization problem* (3.3.28) *admits a global optimal solution* $(x^1, \ldots, x^k) \in (\mathbb{R}^n)^k$.

**Proof.** We only need to consider the case where $k < m$ because otherwise a global solution can be found by setting $x^\ell = a^\ell$ for $\ell = 1, \ldots, m$, and $x^{\ell+1} = \cdots = x^k = a^m$. Choose $r > 0$ such that

$$r > \max\{\rho_F(a^i) \mid i = 1, \ldots, m\} + \max\{\rho_F(a^i - a^j) \mid i \neq j\}.$$

Define

$$\Omega := \{(x^1, \ldots, x^k) \in (\mathbb{R}^n)^k \mid \rho_F(x^i) \leq r \text{ for all } i = 1, \ldots, k\}.$$

Then $\Omega$ is a compact set. It suffices to show that

$$\inf\{f(x^1,\dots,x^k) \mid (x^1,\dots,x^k) \in \Omega\} = \inf\{f(x^1,\dots,x^k) \mid (x^1,\dots,x^k) \in (\mathbb{R}^n)^k\}.$$

Fix any $(x^1,\dots,x^k) \in (\mathbb{R}^n)^k$. Suppose without loss of generality that $\rho_F(x^i) > r$ for all $i = 1,\dots,p$, where $p \leq k$, and $\rho_F(x^i) \leq r$ for all $i = p+1,\dots,k$. Since $\rho_F$ is subadditive,

$$\rho_F(x^\ell - a^i) \geq \rho_F(x^\ell) - \rho_F(a^i) > r - \rho_F(a^i) \geq \rho_F(a^\ell - a^i),$$

for all $\ell = 1,\dots,p, i = 1,\dots,m$.

Therefore,

$$\begin{aligned} f(x^1,x^2,\dots,x^k) &= \sum_{i=1}^{m} \min_{\ell=1,\dots,k} \rho_F(x^\ell - a^i) \\ &\geq f(a^1,a^2,\dots,a^p,x^{p+1},\dots,x^k) \\ &\geq \inf\{f(x^1,\dots,x^k) : (x^1,\dots,x^k) \in \Omega\}. \end{aligned}$$

Thus,

$$\inf\{f(x^1,\dots,x^k) \mid (x^1,\dots,x^k) \in \Omega\} \leq \inf\{f(x^1,\dots,x^k) \mid (x^1,\dots,x^k) \in (\mathbb{R}^n)^k\},$$

which completes the proof. □

For our DC decomposition, we start with the following formula:

$$\min_{\ell=1,\dots,k} \rho_F(x^\ell - a^i) = \sum_{\ell=1}^{k} \rho_F(x^\ell - a^i) - \max_{r=1,\dots,k} \sum_{\ell=1,\ell\neq r}^{k} \rho_F(x^\ell - a^i).$$

Then

$$f(x^1,\dots,x^k)=\sum_{i=1}^{m}\left[\sum_{\ell=1}^{k}\rho_F(x^\ell-a^i)\right]-\sum_{i=1}^{m}\max_{r=1,\dots,k}\left[\sum_{\ell=1,\ell\neq r}^{k}\rho_F(x^\ell-a^i)\right].$$

Similar to the situation with minimizing the function $f$ in (3.3.26), this DC decomposition is not favorable for applying the DCA from Algorithm 1. Our approach here is to apply the Nesterov smoothing technique to obtain an approximation of the objective function favorable for applying the DCA.

By Proposition 3.3.4, the objective function $f$ then has the following approximation:

$$\begin{aligned}&f_\mu(x^1,\dots,x^k)\\&\quad=\frac{1}{2\mu}\sum_{i=1}^{m}\sum_{\ell=1}^{k}\|x^\ell-a^i\|^2\\&\qquad-\left[\frac{\mu}{2}\sum_{i=1}^{m}\sum_{\ell=1}^{k}\left[d\left(\frac{x^\ell-a^i}{\mu};F^\circ\right)\right]^2+\sum_{i=1}^{m}\max_{r=1,\dots,k}\sum_{\ell=1,\ell\neq r}^{k}\rho_F(x^\ell-a^i)\right].\end{aligned}$$

Thus, $f_\mu(x^1,\dots,x^k)=g_\mu(x^1,\dots,x^k)-h_\mu(x^1,\dots,x^k)$ is a DC decomposition of the function $f_\mu$, where $g_\mu$ and $h_\mu$ are convex functions defined by

$$g_\mu(x^1,\dots,x^k):=\frac{1}{2\mu}\sum_{i=1}^{m}\sum_{\ell=1}^{k}\|x^\ell-a^i\|^2 \text{ and}$$

$$h_\mu(x^1,\dots,x^k):=\frac{\mu}{2}\sum_{i=1}^{m}\sum_{\ell=1}^{k}\left[d\left(\frac{x^\ell-a^i}{\mu};F^\circ\right)\right]^2+\sum_{i=1}^{m}\max_{r=1,\dots,k}\sum_{\ell=1,\ell\neq r}^{k}\rho_F(x^\ell-a^i).$$

Let $X$ be the $k\times n$-matrix whose rows are $x^1,\dots,x^k$. We consider the inner product space $\mathcal{M}$ of all $k\times n$ matrices with the inner product of $A,B\in\mathcal{M}$ given by

$$\langle A,B\rangle:=\operatorname{trace}(AB^T)=\sum_{i=1}^{k}\sum_{j=1}^{n}a_{ij}b_{ij}.$$

The norm induced by this inner product is the Frobenius norm.

Then define

$$\begin{aligned} G_\mu(X) := g_\mu(x^1, \dots, x^k) &= \frac{1}{2\mu} \sum_{\ell=1}^{k} \sum_{i=1}^{m} \left( \|x^\ell\|^2 - 2 \left\langle x^\ell, a^i \right\rangle + \|a^i\|^2 \right) \\ &= \frac{1}{2\mu} \left( m\|X\|^2 - 2 \langle X, B \rangle + k\|A\|^2 \right) \\ &= \frac{m}{2\mu}\|X\|^2 - \frac{1}{\mu} \langle X, B \rangle + \frac{k}{2\mu}\|A\|^2, \end{aligned}$$

where $A$ is the $m \times n$-matrix whose rows are $a^1, \dots, a^m$ and $B$ is the $k \times n$-matrix with $a := \sum_{i=1}^{m} a^i$ for every row.

Then the function $G_\mu$ is differentiable with gradient given by

$$\nabla G_\mu(X) = \frac{m}{\mu} X - \frac{1}{\mu} B.$$

From the relation $X = \nabla G_\mu^*(Y)$ if and only if $Y = \nabla G_\mu(X)$, one has

$$\nabla G_\mu^*(Y) = \frac{1}{m}(B + \mu Y).$$

Let us now provide a formula to compute the subdifferential of $H_\mu$ (defined below) at $X$.

First, consider the function

$$\begin{aligned} H^1_\mu(X) :&= \frac{\mu}{2}\sum_{i=1}^{m}\sum_{\ell=1}^{k}\left[d\left(\frac{x^\ell - a^i}{\mu}; F^\circ\right)\right]^2 \\ &= \frac{\mu}{2}\left\{\left[d\left(\frac{x^1 - a^1}{\mu}; F^\circ\right)\right]^2 + \cdots + \left[d(\frac{x^1 - a^m}{\mu}; F^\circ)\right]^2\right\} \\ &+ \cdots \\ &+ \frac{\mu}{2}\left\{\left[d(\frac{x^k - a^1}{\mu}; F^\circ)\right]^2 + \cdots + \left[d(\frac{x^k - a^m}{\mu}; F^\circ)\right]^2\right\}. \end{aligned}$$

The partial derivatives of $H^1_\mu$ are given by

$$\begin{aligned} \frac{\partial H^1_\mu}{\partial x^1}(X) &= \frac{x^1 - a^1}{\mu} - P\left(\frac{x^1 - a^1}{\mu}; F^\circ\right) + \cdots + \frac{x^1 - a^m}{\mu} - P\left(\frac{x^1 - a^m}{\mu}; F^\circ\right) \\ &= \sum_{i=1}^{m}\left[\frac{x^1 - a^i}{\mu} - P\left(\frac{x^1 - a^i}{\mu}; F^\circ\right)\right], \\ &\vdots \\ \frac{\partial H^1_\mu}{\partial x^k}(X) &= \frac{x^k - a^1}{\mu} - P\left(\frac{x^k - a^1}{\mu}; F^\circ\right) + \cdots + \frac{x^k - a^m}{\mu} - P\left(\frac{x^k - a^m}{\mu}; F^\circ\right) \\ &= \sum_{i=1}^{m}\left[\frac{x^k - a^i}{\mu} - P\left(\frac{x^k - a^i}{\mu}; F^\circ\right)\right]. \end{aligned}$$

The gradient $\nabla H^1_\mu(X)$ is the $k \times n$-matrix whose rows are $\frac{\partial H^1_\mu}{\partial x^1}(X), \ldots, \frac{\partial H^1_\mu}{\partial x^k}(X)$.

Let $H_\mu(X) := h_\mu(x^1, \ldots, x^k)$. Then $H_\mu = H^1_\mu + H^2$, where

$$H^2(X) := \sum_{i=1}^{m}\max_{r=1,\ldots,k}\sum_{\ell=1,\ell\neq r}^{k}\rho_F(x^\ell - a^i).$$

In what follows, we provide a formula to find a subgradient of $H^2$ at $X$.

Define the function

$$F^{i,r}(X) := \sum_{\ell=1,\ell\neq r}^{k} \rho_F(x^\ell - a^i).$$

Choose the row vector $v^{i,\ell} \in \partial\rho_F(x^\ell - a^i)$ if $\ell \neq r$ and $v^{i,r} = 0$. Then the $k \times n$-matrix formed by the rows $v^{i,r}$ for $i = 1, \ldots, k$ is a subgradient of $F^{i,r}$ at $X$.

Define

$$F^i(X) := \max_{r=1,\ldots,k} F^{i,r}(X).$$

In order to find a subgradient of $F^i$ at $X$, we first find an index $r \in I_i(X)$, where

$$I^i(X) := \{r = 1, \ldots, k \mid F^i(X) = F^{i,r}(X)\}.$$

Then, choose $V_i \in \partial F^{i,r}(X)$ and we have that $\sum_{i=1}^m V_i$ is a subgradient of the function $H^2$ at $X$. This results in our first algorithm for the multifacility location problem.

**DC Algorithm 5**.

INPUT: $X_1 \in \mathcal{M}$, $N \in I\!N$, $F$, $a^1, \ldots, a^m \in \mathbb{R}^n$.

**for** $k = 1, \ldots, N$ **do**

Find $Y_k = U_k + V_k$, where

$$U_k := \nabla H^1_\mu(X_k),\ V_k \in \partial H^2(X_k).$$

Find $X_{k+1} = \frac{1}{m}(B + \mu Y_k)$.

OUTPUT: $X_{N+1}$.

Let us now present the second algorithm for solving the multifacility problem. By Proposition 3.3.4, the function $F^{i,r}(X) := \sum_{\ell=1,\ell\neq r}^k \rho_F(x^\ell - a^i)$ has the following smooth approximation:

$$F^{i,r}_\mu(X) = \sum_{\ell=1,\ell\neq r}^{k} \left[\frac{1}{2\mu}\|x^\ell - a^i\|^2 - \frac{\mu}{2}\left[d(\frac{x^\ell - a^i}{\mu}; F^\circ)\right]^2\right].$$

For a fixed $r$, define the row vectors $v^{i,\ell} = P(\frac{x^\ell - a^i}{\mu}; F^\circ)$ if $\ell \neq r$ and $v^{i,r} = 0$. Then $\nabla F_\mu^{i,r}(X)$ is the $k \times n$ matrix $V_{i,r}$ formed by these rows.

Now we define the function $F_\mu^i(X) := \max_{r=1,\ldots,k} F_\mu^{i,r}(X)$. This is an approximation of the function

$$F^i(X) := \max_{r=1,\ldots,k} \sum_{\ell=1,\ell\neq r}^{k} \rho_F(x^\ell - a^i).$$

As a result, $H_\mu^2 := \sum_{i=1}^m F_\mu^i$ is an approximation of the function $H^2$.

Define the *active index set*

$$I_\mu^i(X) := \{r = 1, \ldots, k \mid F_\mu^i(X) = F_\mu^{i,r}(X)\}.$$

Choose $r \in I_\mu^i(X)$ and calculate $V_i = \nabla F_\mu^{i,r}(X)$. Then $V := \sum_{i=1}^m V_i$ is a subgradient of the function $H_\mu^2$ at $X$.

**DC Algorithm 6**.

INPUT: $X_1 \in \mathcal{M}$, $N \in I\!N$, $F$, $a^1, \ldots, a^m \in \mathbb{R}^n$.

**for** $k = 1, \ldots, N$ **do**

Find $Y_k = U_k + V_k$, where

$$U_k := \nabla H_\mu^1(X_k),\ V_k \in \partial H_\mu^2(X_k).$$

Find $X_{k+1} = \frac{1}{m}(B + \mu Y_k)$.

OUTPUT: $X_{N+1}$.

**Remark 3.3.12.** Similar to the case of DC Algorithm 3 and DC Algorithm 4, when implementing DC Algorithm 5 and DC Algorithm 6, instead of using a fixed smoothing parameter $\mu$, we often change $\mu$ during the iteration.

**3.3.5. Numerical Implementation.** We demonstrate the above DC algorithms on several problems. All code was written in *MATLAB*. Unless otherwise stated, we use the closed Euclidean unit ball for the set $F$ associated with the Minkowski gauge. In accordance with Remark 3.3.10, we use $\mu_* = 10^{-6}$, decreasing $\mu$ over 3 implementations, each of which runs until $\sum_{\ell=1}^{k} d(x_j^\ell, x_{j-1}^\ell) < k \cdot 10^{-6}$, where $k$ is the number of centers and $j$ is the iteration counter. The starting value $\mu_0$ is specified in each example.

**Example 3.3.13.** In this example we implement DC Algorithms 3 and 4 to solve the generalized Fermat-Torricelli problem under the $\ell_1$ norm with randomly generated points as shown in Figure 3.4 . This synthetic data set has 10,000 points with weight $c_i = 1$ and three points with weight $c_i = -1000$. For the smoothing parameter, we use an initial $\mu_0 = 0.1$. Both algorithms converge to an optimal solution of $x \approx (17.29, 122.46)$. The convergence rate is shown in Figure 3.5 .

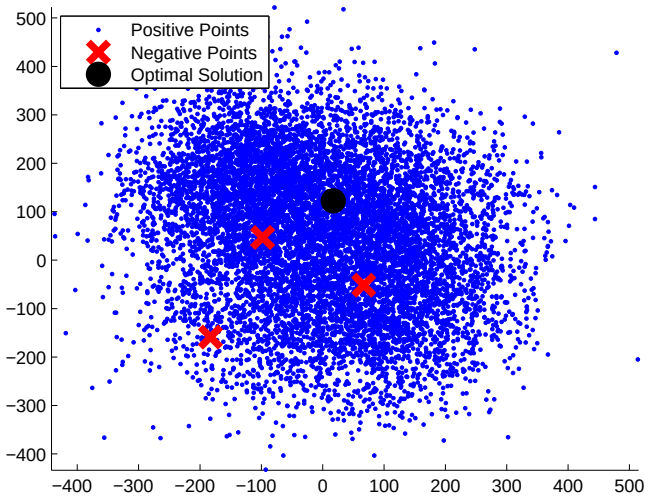



FIGURE 3.4. A generalized Fermat-Torricelli problem in $\mathbb{R}^2$. Each negative point has weight of -1000; each positive point has a weight of 1; the optimal solution is represented by $\bullet$ for the $\ell_1$ norm.

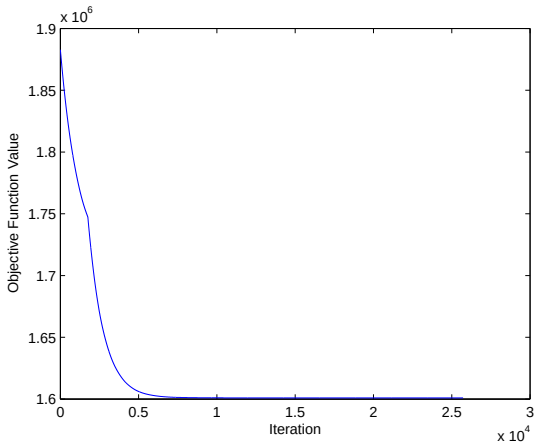



FIGURE 3.5. The objective function values for Algorithm 4 for the generalized Fermat-Torricelli problem under the $\ell_1$ norm shown in Figure 3.4.

**Example 3.3.14.** We implement DC Algorithm 5 to solve multifacility location problems given by function (3.3.28). We use the following six real data sets[1]: *WINE* contains 178 instances of $k = 3$ wine cultivars in $\mathbb{R}^{13}$. The classical *IRIS* data set contains 150 observations in $\mathbb{R}^4$, describing $k = 3$ varieties of Iris flower. The *PIMA* data set contains 768 observations, each with 8 features describing the medical history of adults of Pima American-Indian heritage. *IONOSPHERE* contains data on 351 radar observations in $\mathbb{R}^{34}$ of free electrons in the ionosphere. *USCity*[2] contains the latitude and longitude of 1217 US cities; we use $k = 3$ centroids (See Figure 3.6).

Reported values are as follows: $m$ is the number of points in the data set; $n$ is the dimension; $k$ is the number of centers; $\mu_0$ is the starting value for the smoothing parameter $\mu$, as discussed in Remark 3.3.10 (in each case, $\sigma$ is chosen so that $\mu$ decreases to $\mu_*$ in three iterations); $Iter$ is the number of iterations until convergence; $CPU$ is the computation time in seconds; $Objval$ is the final value of the true objective

[1]Available at *https://archive.ics.uci.edu/ml/datasets.html*
[2]http:/www.realestate3d.com/gps/uslatlongdegmin.htm

| | $m$ | $n$ | $k$ | $\mu_0$ | Iter | CPU | Objval |
|---|---|---|---|---|---|---|---|
| *WINE* | 178 | 13 | 3 | 10 | 690 | 1.86 | $1.62922 \cdot 10^4$ |
| *IRIS* | 150 | 4 | 3 | 0.1 | 314 | 0.66 | 96.6565 |
| *PIMA* | 768 | 8 | 2 | 10 | 267 | 2.22 | $4.75611 \cdot 10^4$ |
| *IONOSPHERE* | 351 | 34 | 2 | 0.1 | 391 | 1.68 | $7.93712 \cdot 10^2$ |
| *USCity* | 1217 | 2 | 3 | 1 | 940 | 16.0 | $1.14211 \cdot 10^4$ |

TABLE 3.1. Results for Example 6.3, the performance of Algorithm 5 on real data sets.

function (3.3.28), not the smoothed version $f_\mu$. Implementations of Algorithm 6 produced nearly identical results on each example and thus are not reported.

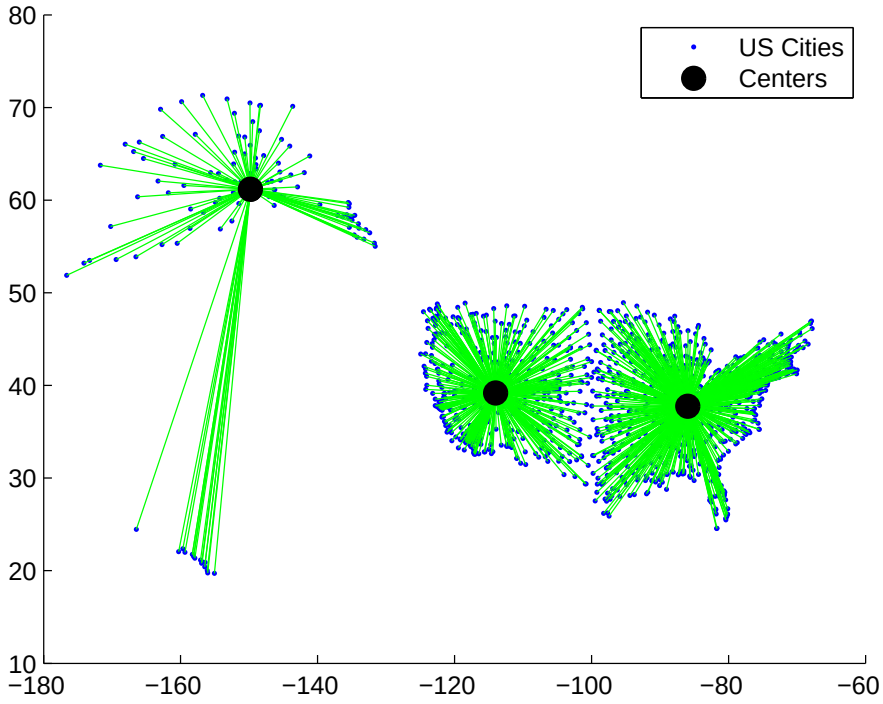



FIGURE 3.6. The solution to the multifacility location problem with three centers and Euclidean distance to 1217 US Cities. A line connects each city with its closest center.

**3.3.6. Additional Work: Set Clustering.** Here we provide a brief discussion of how the above work can be extended to location problems where the targets are *sets*, as opposed to points. Applications include location problems where the targets have non-negligible sizes and set clustering problems. More details and additional discussion can be found in [**50**].

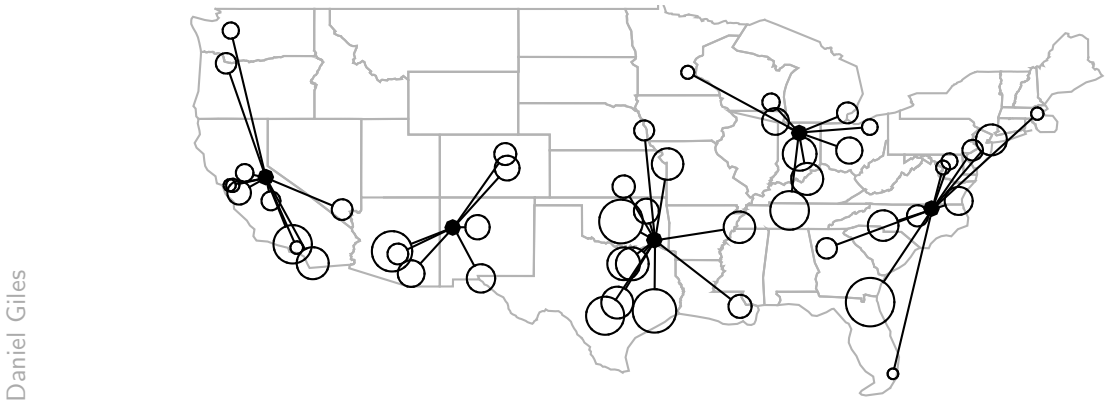



FIGURE 3.7. The fifty most populous US cities, approximated by a ball proportional to their area. Each city is assigned to the closest of five centroids (•), which are the optimal facilities.

3.3.6.1. *Set Clustering.* In this section, we study the problem of *set clustering*, where the objects being classified are *sets* rather than points. Given a nonempty, closed, and convex set $\Omega \subset \mathbb{R}^n$, observe that

$$\begin{aligned}
[d(x;\Omega)]^2 &= \inf\{\|x-w\|^2 \mid w \in \Omega\} \\
&= \inf\{\|x\|^2 - 2\langle x, w\rangle + \|w\|^2 \mid w \in \Omega\} \\
&= \|x\|^2 + \inf\{\|w\|^2 - 2\langle x, w\rangle \mid w \in \Omega\} \\
&= \|x\|^2 - \sup\{\langle 2x, w\rangle - \|w\|^2 \mid w \in \Omega\}.
\end{aligned}$$

**Proposition 3.3.15.** *Let $\Omega$ be a nonempty, closed, and convex set in $\mathbb{R}^n$. Define the function*

$$\varphi_\Omega(x) := \sup\{\langle 2x, w\rangle - \|w\|^2 \mid w \in \Omega\} = 2\sup\{\langle x, w\rangle - \frac{1}{2}\|w\|^2 \mid w \in \Omega\}.$$

*Then $\varphi$ is convex and differentiable with $\nabla\varphi_\Omega(x) = 2P(x;\Omega)$.*

**Proof.** It follows from the representation of $[d(x;\Omega)]^2$ above that

$$\varphi_\Omega(x) = \|x\|^2 - [d(x;\Omega)]^2.$$

Note that the function $\psi(x) := [d(x;\Omega)]^2$ is differentiable with $\nabla\psi(x) = 2[x - P(x;\Omega)]$; see, e.g., [**41**, Exercise 3.2]. Then the function $\varphi_\Omega$ is differentiable with

$$\nabla\varphi_\Omega(x) = 2x - 2[x - P(x;\Omega)] = 2P(x;\Omega),$$

which completes the proof. □

Let $\Omega^i$ for $i = 1, \ldots, m$ be nonempty, closed, and convex sets in $\mathbb{R}^n$. We need to choose $x^\ell$ for $\ell = 1, \ldots, k$ in $\mathbb{R}^n$ as centroids and assign each member $\Omega^i$ to its closest centroid. The objective function to be minimized is the sum of these distances.

Then we have to solve the optimization problem:

(3.3.29)

$$\text{minimize } f(x^1, \ldots, x^k) := \sum_{i=1}^{m} \min_{\ell=1,\ldots,k} [d(x^\ell; \Omega^i)]^2, \qquad x^\ell \in \mathbb{R}^n, \ell = 1, \ldots, k.$$

**Proposition 3.3.16.** *Suppose that the convex sets $\Omega_i$ for $i = 1, \ldots, m$ are nonempty, closed, and bounded. Then* (3.3.29) *has a global optimal solution.*

**Proof.** Choose $r > 0$ such that $\Omega^i \subset B(0;r)$ for all $i = 1, \ldots, m$. Fix $a^i \in \Omega^i$ for $i = 1, \ldots, m$. Define

$$S := \{(x^1, \ldots, x^k) \in (\mathbb{R}^n)^k \mid \|x^i\| \leq 6r \text{ for } i = 1, \ldots, k\}.$$

Let us show that

$$\inf\{f(x^1, \ldots, x^k) \mid (x^1, \ldots, x^k) \in (\mathbb{R}^n)^k\} = \inf\{f(x^1, \ldots, x^k) \mid (x^1, \ldots, x^k) \in S\}.$$

Fix any $(x^1, \ldots, x^k) \in (\mathbb{R}^n)^k$. Without loss of generality, suppose that $k < m$ and $\|x^\ell\| > 6r$ for $\ell = 1, \ldots, p$, and $\|x^{p+1}\| \leq 6r$, $\ldots$, $\|x^k\| \leq 6r$, where $p \leq k$. Let $p^{\ell,i} := P(x^\ell; \Omega^i)$. Then for $\ell = 1, \ldots, p$, we have

$$\begin{aligned}
[d(x^\ell; \Omega^i)]^2 &= \|x^\ell - p^{\ell,i}\|^2 \\
&= \|x^\ell\|^2 - 2\left\langle x^\ell, p^{\ell,i} \right\rangle + \|p^{\ell,i}\|^2 \\
&\geq \|x^\ell\|^2 - 2\|x^\ell\| \, \|p^{\ell,i}\| \\
&= \|x^\ell\|(\|x^\ell\| - 2\|p^{\ell,i}\|) \geq \|x^\ell\|(6r - 2\|p^{\ell,i}\|) \geq 4r\|x^\ell\| \geq 4r^2.
\end{aligned}$$

In addition, for all $\ell = 1, ..., m$, we have

$$[d(a^\ell; \Omega^i)]^2 \leq \|a^\ell - a^i\|^2 \leq 4r^2 \leq [d(x^\ell; \Omega^i)]^2.$$

It follows that

$$\begin{aligned}
f(x^1, \ldots, x^k) &= \sum_{i=1}^{m} \min_{\ell=1,\ldots,k} [d(x^\ell; \Omega^i)]^2 \\
&\geq f(a^1, \ldots, a^p, x^{p+1}, x^\ell) \\
&\geq \inf\{f(x^1, \ldots, x^k) : (x^1, \ldots, x^k) \in S\}.
\end{aligned}$$

The rest of the proof follows from the proof of Proposition 3.3.11. □

For our DC decomposition, we use the following formula

$$\min_{\ell=1,\dots,k} [d(x^\ell;\Omega^i)]^2 = \sum_{\ell=1}^{k} [d(x^\ell;\Omega^i)]^2 - \max_{r=1,\dots,k} \sum_{\ell=1,\ell\neq r}^{k} [d(x^\ell;\Omega^i)]^2.$$

Then

$$\begin{aligned} f(x^1,\dots,x^k) &= \sum_{i=1}^{m}\sum_{\ell=1}^{k}[d(x^\ell;\Omega^i)]^2 - \left[\sum_{i=1}^{m} \max_{r=1,\dots,k} \sum_{\ell=1,\ell\neq r}^{k} [d(x^\ell;\Omega^i)]^2\right] \\ &= \sum_{i=1}^{m}\sum_{\ell=1}^{k} \|x^\ell\|^2 - \Big[\sum_{i=1}^{m}\sum_{\ell=1}^{k} \varphi_{\Omega^i}(x^\ell) + \sum_{i=1}^{m} \max_{r=1,\dots,k} \sum_{\ell=1,\ell\neq r}^{k} [d(x^\ell;\Omega^i)]^2\Big]. \end{aligned}$$

Define

$$\begin{aligned} g(x^1,\dots,x^k) &:= \sum_{i=1}^{m}\sum_{\ell=1}^{k} \|x^\ell\|^2 \\ h(x^1,\dots,x^k) &:= \sum_{i=1}^{m}\sum_{\ell=1}^{k} \varphi_{\Omega^i}(x^\ell) + \sum_{i=1}^{m} \max_{r=1,\dots,k} \sum_{\ell=1,\ell\neq r}^{k} [d(x^\ell;\Omega^i)]^2. \end{aligned}$$

Then we have the DC decomposition $f = g - h$.

For $X \in \mathcal{M}$, define

$$G(X) := \sum_{i=1}^{m}\sum_{\ell=1}^{k} \|x^\ell\|^2 = m\|X\|^2.$$

Thus, $\nabla G^*(X) = \frac{1}{2m}(X)$.

Define

$$H^1(X) := \sum_{i=1}^{m}\sum_{\ell=1}^{k} \varphi_{\Omega^i}(x^\ell).$$

Then

$$\frac{\partial H^1}{\partial x^1} = 2P(x^1;\Omega^1) + \cdots + 2P(x^1;\Omega^m),$$
$$\vdots$$
$$\frac{\partial H^1}{\partial x^k} = 2P(x^k;\Omega^1) + \cdots + 2P(x^k;\Omega^m).$$

Then $\nabla H^1(X)$ is the $k \times n$ matrix whose rows are $\frac{\partial H^1}{\partial x^i}$ for $i = 1, \ldots, k$.

Let us now present a formula to compute a subgradient of the function

$$H^2(X) = \sum_{i=1}^{m} \max_{r=1,\ldots,k} \sum_{\ell=1,\ell\neq r}^{k} [d(x^\ell;\Omega^i)]^2.$$

Define

$$H_2^i(X) := \max_{r=1,\ldots,k} \sum_{\ell=1,\ell\neq r}^{k} [d(x^\ell;\Omega^i)]^2 = \max_{r=1,\ldots,k} H_2^{i,r},$$

where

$$H_2^{i,r} := \sum_{\ell=1,\ell\neq r}^{k} [d(x^\ell;\Omega^i)]^2.$$

Consider the following row vectors

$$v_{i,\ell} := 2(x^\ell - P(x^\ell;\Omega^i)) \text{ if } \ell \neq r,$$
$$v_{i,r} := 0.$$

Then $\nabla H_2^{i,r}$ is the $k \times n$ matrix whose rows are these vectors.

Define the active index set

$$I^i(X) := \{r = 1, \ldots, k : H_2^{i,r}(X) = H_2^i(X)\}.$$

Choose $r \in I^i(X)$ and let $V_i := \nabla H_2^{i,r}(X)$. Then $V := \sum_{i=1}^m V_i$ is a subgradient of $H^2$ at $X$. This leads to our algorithm for solving the set clustering problem (3.3.29).

**Algorithm 7**.

INPUT: $X_1 \in \mathcal{M}$, $N \in I\!N$, $\Omega^1, \ldots, \Omega^m \in \mathbb{R}^n$.

**for** $k = 1, \ldots, N$ **do**

Find $Y_k = U_k + V_k$, where

$$U_k := \nabla H^1(X_k),\ V_k \in \partial H^2(X_k).$$

Find $X_{k+1} = \frac{1}{2m}(Y_k)$.

OUTPUT: $X_{N+1}$.

# 4

## Applications to Electric Power Systems

Historically, the electric grid was designed for a one-way flow of both power and control signals from central generation and control centers down toward end-users. Today, two significant problems facing the electric grid are (1) the large scale integration of renewable (i.e. intermittent and non-dispatchable) generation and (2) the massive control problem posed by a large number of Distributed Energy Resources[1] (DERs). These electric grid modernization problems are being approached with *distributed control* techniques that require individual participants in the grid to carry out their own optimization [**60, 77**]. This paradigm creates fertile ground for inter-disciplinary work and applications of optimization. One such distributed control technique, called *transactive control* or *transactive energy*, seeks to control grid connected devices through price or value signals. Background information on transactive energy is presented in this thesis in section 4.2.1

In this thesis, we identify one particular component of the electric grid that would require optimization in a transactive control setting: a smart solar inverter with battery storage. A smart solar inverter is a device that converts direct current generated by photovoltaic (PV) panels into the alternating current synchronized with the electric grid. The battery allows the electricity generated by the PV panels to either be sold immediately to the grid or to be saved for later. The inverter also allows for electricity to be purchased from the grid and used to charge the battery.

[1]Distributed Energy Resources are include, for example, residential and rooftop solar, in-home battery storage, commercial cogeneration facilities, and demand-response technologies.

## 4.1. Introduction

Distributed solar and storage have presented themselves as an integral part of the future electric grid. As installed capacity continues to increase, challenges have emerged regarding the coordination and control of such distributed energy resources [**60, 77**]. In a smart grid, a utility or other DER resource owner will seek to maximize economic value and optimize the operation of the DER asset(s) while considering other generation and demand response resources, environmental conditions, and external market forces. A *transactive control* system is one potential method for coordinating and controlling a smart grid efficiently in real-time [**13, 55**].

In this paper we develop an optimal control scheme for the smart solar inverter and battery storage system in terms of providing real power sales, reactive power support, and spinning reserve capacity in a transactive energy market setting. The method considers transactive market prices, the state-of-charge of the storage resource, weather, and solar-with-storage system dynamics to model revenue earned. The value of DER services provided to the grid in a transactive market varies in time and reflects real-time conditions at a single location of the electric grid. In such a setting, consumers and DER devices are expected to respond to transactive price signals in a way that maximizes their own best interest. The prices are set (or "settled on") so that this response, in turn, maximizes benefit to the greater electric grid.

The principal technology under consideration here is the smart solar inverter. This device converts direct current (DC) from the storage system and PV panels into alternating current (AC) synchronized with the electric grid. The simplified diagram in Figure 4.1 shows the system topology. This system also allows for electricity to be purchased from the grid and used to charge the battery. Additionally, the smart solar inverter can assist in the management of reactive power. Thus the inverter, coupled

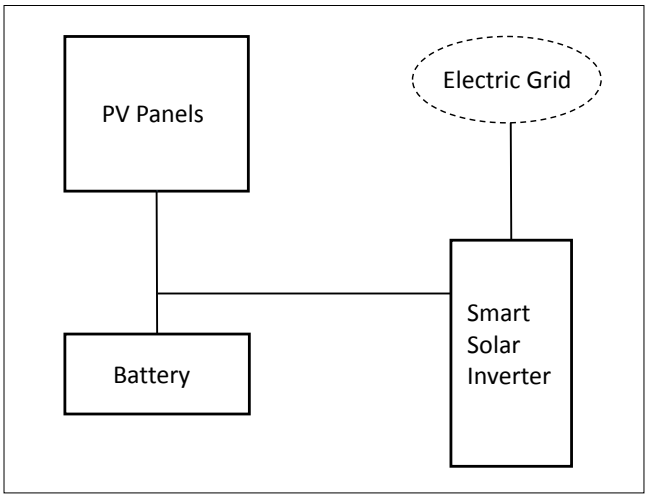



FIGURE 4.1. PV Panel, Battery, and Smart Solar Inverter diagram.

with battery storage, can offer reactive power support and spinning reserve capacity as ancillary services.

Various methods exist for incorporating real-time pricing and weather forecasts into DER decision making, but none of these methods incorporate all three revenue streams: real power sales, reactive power support, and spinning reserve capacity. Bilevel optimization methods have been used [**11, 87, 69, 82**] to coordinate DERs and help retailers determine price-setting strategies while considering the response from end-users. These methods contain explicit lower-level objective functions and constraints that model a general end-user's response to price signals, but are not formulated for the specific solar inverter with storage system we describe and do not take into account reactive power support or spinning reserve capacity.

Much work has been done from a technical standpoint to develop control strategies for enlisting DERs (and microgrids) in the management of reactive power (for example, [**7, 67**]). These perspectives originate from a centralized "grid operator" point of view and may be useful for determining how prices are set for providing reactive

power support. The works in [**8, 76**] propose strategies for management of energy storage connected to PV panels but is not formulated to maximize revenue for the DER owner. The work in [**3**] minimizes operation cost and energy import to a system comprised of PV and energy storage. The comprehensive paper [**63**], which seeks to optimize power flow management in a grid-connected PV system with storage, incorporates onsite loads, but does not include the possibility that revenue could be earned from providing reactive power support or spinning reserve capacity.

**4.1.1. Chapter Organization.** The remainder of this chapter is organized as follows. In section 4.2 we describe the problem setting. This includes brief discussions of transactive energy itself, how our method enables transactive control, and the physical system constraints. In section 4.3, we state the optimization problem and develop our solution. In section 4.4, we provide the results of a numerical experiment using historical data from the Midcontinent Independent System Operator (MISO) to simulate a transactive market. We compare the resulting operations profile of our control scheme to a business-as-usual control scheme in financial terms. Finally, we offer some concluding remarks and directions for further research.

## 4.2. System Overview

**4.2.1. Transactive Energy Systems.** Transactive energy provides a means for embedding economic control signals into the operation of the electric grid. It is especially well-suited for the management of DERs and the integration of renewable energy sources. The ideas behind transactive energy originated in 1981 with a paper [**72**] by Schweppe and other researchers at MIT. At the same time, work was being done by Smith at Stanford [**74**] to develop a high-level protocol for communication among

nodes in a distributed problem solver. Nearly 25 years later, the term *transactive control* was introduced in a report by Pacific Northwest National Laboratory [**13**], bringing together the ideas of Smith and others. Additional background on transactive control and related ideas can be found in [**12, 19, 24, 27, 28, 53, 54, 55, 59, 60, 77**]. A discussion of the difference between the terms *transactive energy* and *transactive control* can be found in [**5**].

In practice, there are several methods currently claiming recognition as transactive energy systems, including but not limited to extended financial market systems [**6**], double-auction type systems [**68**], and hierarchical economic control systems [**19, 55**]. Here, we apply concepts from [**55**], where there are two signals exchanged in the transactive system: a *value signal* and a *demand signal*. The value signal contains the predicted price for future time periods, called transactive intervals. The demand signal contains the predicted load in response to that set of prices. These signals are exchanged frequently as prices are "settled on" so that the cost or need for a service is reflected by the price at that specific location of the grid at that specific time. Thus any device participating in a transactive energy system must not only decide its own operation during the current transactive interval, but also must forecast its operation for all upcoming transactive intervals, based on the current value signal. The transactive intervals themselves may be defined by any length of time, although the transactive value signal used in [**55**] was composed of three days of forward prices consisting of minute, hourly, and daily average values.

The value signal and the demand signal are communicated through a network of nodes. A node is defined as a physical point in the grid where demand may be aggregated and predicted. The nodes can be classified in a hierarchical structure as described by Hammerstorm, et al. in [**19**]. The five different classes of nodes

are: generation, transmission, distribution substations, distribution transformers, and sites. In a transactive energy system, each node must have some method by which it determines its optimal response to the given value signal. The work presented in this thesis provides such a method specifically for generation-level nodes consisting of PV panels, a smart inverter and energy storage. The method is easy to use and simple enough that the optimal response to a given value signal can be determined in real-time.

**4.2.2. Economic Model.** In our transactive control setting, a smart solar inverter and storage system (as shown in Figure 4.1), known from now on simply as a *smart inverter system*, can earn revenue for providing the following services to the grid: real power sales, reactive power support, and spinning reserve capacity. As discussed above, the prices for these services vary in real-time and are forecast for the upcoming transactive intervals by the transactive value signal. As time increments, the value signal is updated so that it always contains the most up-to-date information. We also assume that our smart inverter system has access (via some external information feed) to forecasted solar irradiance values (that is, the expected generation of the PV panels) for each of the upcoming transactive intervals. Again, as time increments, these forecasted values are updated so that they always contain the most up-to-date information.

**4.2.3. System Constraints.** The smart inverter system is subject to certain efficiencies and limitations posed by the physical constraints of the inverter and battery as well as the system dynamics. These are:

(1) *Maximum Rated Capacity of the Inverter.* The inverter cannot import or export more power than rated over a given time period.

(2) *Battery Capacity.* The storage device is limited by a maximum and minimum amount of energy it can hold.

(3) *Battery Discharge Rate Limitation.* The battery capacity is effectively diminished under a faster discharge rate.

(4) *Constraints on Reserve Capacity.* The amount of energy sold as reserve capacity must not exceed the amount available.

(5) *Battery Efficiency.* A certain efficiency must be applied when charging and discharging the battery.

(6) *Power Factor.* The amount of reactive power support provided is subject to power factor limitations of the inverter.

Mathematical formulations of these parameters and constraints are given in the next section.

## 4.3. The Optimization Problem

**4.3.1. Problem Statement.** We maximize revenue earned by the smart inverter system in a transactive control setting for providing real power sales, reactive power support and spinning reserve capacity, subject to the constraints described in section 4.2.3. This process involves determining a sequence of control decisions over the $n$ transactive intervals given the forecasted value signal and expected solar irradiance over the same $n$ intervals.

**4.3.2. Problem Solution.** In the following sections, we proceed to solve this general optimization problem by developing an objective function and formulating the system constraints so that optimal inputs may be obtained by off-the-shelf optimization software. This process involves making precise statements of the problem

variables, parameters and constraints, and various problem-simplifications leading to the final objective function.

**4.3.3. Variables and Parameters.** We group the variables into four categories: decision variables, problem variables, state variables, and parameters. The *decision variables* model the behavior of the inverter, such as selling or purchasing energy from the grid. The *problem variables* inform the decisions made by the inverter. They are inputs into the function and assumed to be given at the beginning of each optimization cycle. The *state variable* reflects the system's state in each time period. It is a function of the problem variables and decision variables. The *parameters* define the performance limits of the smart inverter system. We write each decision and problem variable as an $n$-dimensional vector, representing the values during each of the $n$ transactive time periods of equal length.

Decision Variables

(1) $s = (s_1, ..., s_n)$: The amount of energy to sell to the grid (MVAh); represents the combination of real (MWh) and reactive power support (MVARh) sold to the grid

(2) $r = (r_1, ..., r_n)$: The amount of energy to be sold as spinning reserve capacity (MWh)

(3) $t = (t_1, ..., t_n)$: The amount of energy purchased from the grid (MWh)

Problem Variables

(4) $a = (a_1, ..., a_n)$: The price of real energy for sales to the grid ($/MWh)

(5) $b = (b_1, ..., b_n)$: The price of reactive energy for sales to the grid ($/MVARh)

(6) $c = (c_1, ..., c_n)$: The cost of energy purchased from the grid ($/MWh)

(7) $d = (d_1, ..., d_n)$: The price of energy for reserve capacity ($/MWh)

(8) $e = (e_1, ..., e_n)$: The solar iradiance during that time period (MWh)

State Variables

(9) $l = (l_2, ..., l_{n+1})$: The charge in the battery at the beginning of that time period (kWh). (Note: the value of $l_1$ is input into the function. The remainder of the values $l_2, ..., l_{n+1}$ are a result of the choices for the values in $s, r$, and $t$, as formulated in 4.3.7.)

Parameters

(10) $F$: The limit on power factor for power supplied by the inverter (%)

(11) $M$: The maximum rated power output of the inverter (MW)

(12) $\gamma$: Accounts for the discharge rate limitations on the battery (%)

(13) $\eta$: One-way efficiency loss due to charging or discharging the battery (%)

(14) $L$: The maximum charge capacity of the battery (MWh)

(15) $l_1$: The initial charge in the battery (MWh)

**4.3.4. Objective Function Intuition.** Initially, we denote our objective function by $g$. We model the revenue earned as a stock and flow such that the revenue earned is the "money coming in minus the money going out":

*Revenue earned = (Revenue from energy sales) + (Revenue from reserve capacity) – (Cost of energy purchased).*

An objective function $g$ that defines the revenue earned during transactive interval $i$ can then be stated as

$$g(s_i, r_i, t_i \ ; \ a_i, b_i, c_i, d_i, e_i) = h(s_i \ ; \ a_i, b_i) + (d_i)(r_i) - (c_i)(t_i), \tag{4.3.30}$$

where the function $h(s_i \ ; \ a_i, b_i)$ equals the revenue earned from real and reactive power sales given the values of $s_i, a_i$, and $b_i$. (We use the semi-colon ";" in $g$ and $h$ to separate the decision variables from the problem variables.)

**4.3.5. Details on the energy sales revenue function $h$.** As denoted, the function $h(s_i \ ; \ a_i, b_i)$ should return the maximum possible revenue earned from selling real and reactive power given $s_i, a_i$, and $b_i$. That is, the function $h$ returns the *function value* from the following optimization problem:

$$\text{Maximize} \quad \tilde{f}(p, q \mid a_i, b_i) := a_i p + b_i q \qquad \text{over } p, q, \tag{4.3.31}$$

where $p, q$ represent the real and reactive energy sold to the grid, respectively, with the restrictions

$$p^2 + q^2 = s_i^2, \qquad p \geq F s_i.$$

The constant $F$ is the power factor limit on the inverter.[2] We use the constraint $p^2 + q^2 = s_i^2 \Rightarrow q = \sqrt{s_i^2 - p^2}$ to write a new function $f$ in place of $\tilde{f}$, thus rewriting (4.3.31) as

$$\text{Maximize} \quad f(p \ ; \ a_i, b_i, s_i) = a_i p + b_i \sqrt{s_i^2 - p^2} \qquad \text{over } p, \tag{4.3.32}$$

[2]This formulation of the constraint represents a *symmetric* limitation on the power factor. In practice this may not be the case, but could easily be adopted into this scheme. A typical value for $F$ may be 0.8, although some modern inverters may be able to operate with a power factor as low as $F = 0$.

where $p$ represents the real energy sold to the grid, with the constraint

$$p \geq F s_i.$$

Figure 4.2 below shows the graph of the function $f(p\ ;\ a_i, b_i, s_i)$ for hypothetical values of $a_i$, $b_i$, and $s_i$. In the figure, we can see how the amount of revenue earned is related to the power factor of the inverter for the given prices $a_i$ and $b_i$.

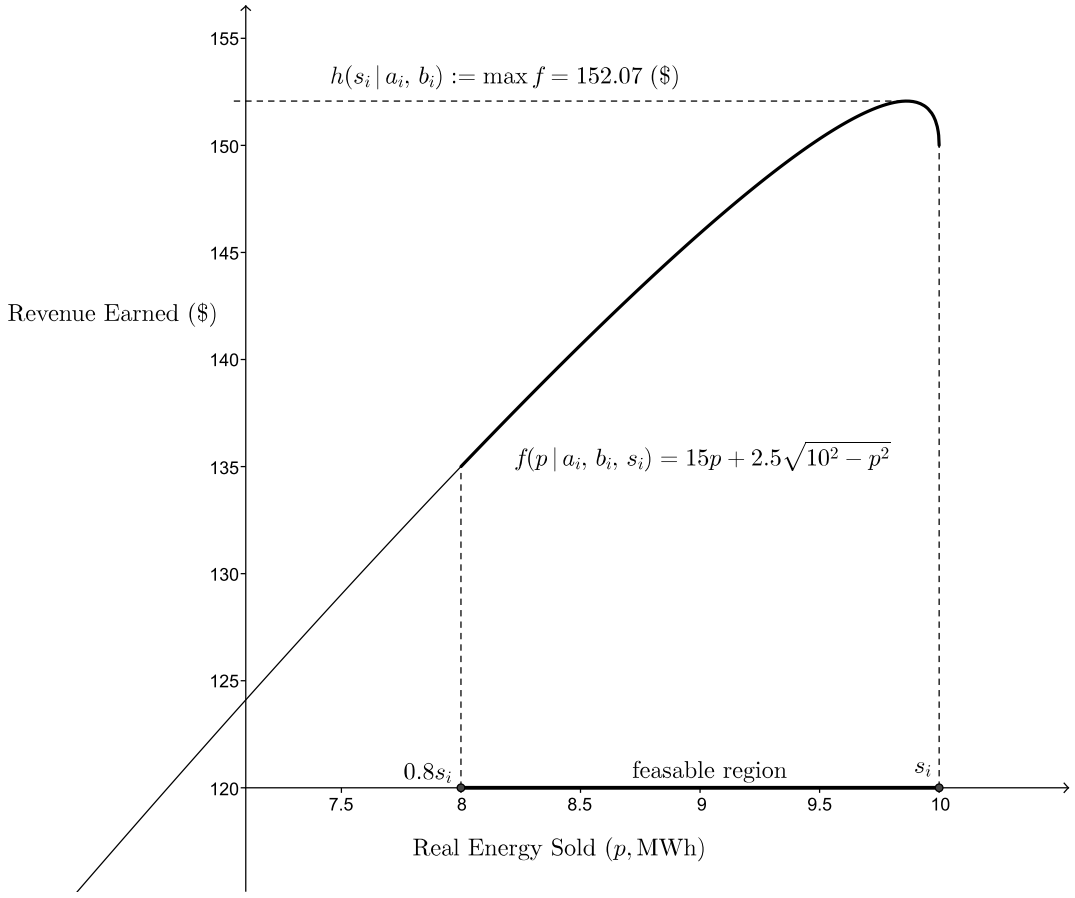



FIGURE 4.2. The energy sales revenue function $h$ is the maximum of $f$ over the feasible region. The graph of $f$ is shown with the example input values $a_i = 15$ ($/MWh), $b_i = 2.5$ ($/MVARh), $s_i = 10$ (MVAh), and constant $L = 0.8$.

**4.3.6. An analytic solution for $h$.** Considering the earlier formulation (4.3.31), we can see that a maximum occurs for some combination of p and q for values on the circle $p^2 + q^2 = s_i^2$ (with the additional constraint $p \geq F s_i$). Let $\theta_i$ be the phase angle of the power supplied in transactive interval $i$, so $\cos\theta_i = p/s_i$. Then $\tilde{f}$ from (4.3.31) can be equivalently written as

$$\tilde{\tilde{f}}(\theta_i \ ; \ a_i, b_i) := a_i(\cos\theta_i)s_i + b_i(\sin\theta_i)s_i. \tag{4.3.33}$$

The goal is to find the value of $\theta_i$ that maximizes this quantity, so we can disregard the $s_i$ terms. Taking the derivative with respect to $\theta_i$ and setting that equal to zero yields

$$\theta_i = \tan^{-1}\left(\frac{b_i}{a_i}\right).$$

The restriction $p \geq F s_i$ means that we must have $\theta_i \leq \cos^{-1}(F)$. Thus we can write the analytic solution for $h$ as

$$h(s_i \ ; \ a_i, b_i) = a_i \cos(\overline{\theta}_i)s_i + b_i \sin(\overline{\theta}_i)s_i, \tag{4.3.34}$$

where

$$\overline{\theta}_i = \min\left\{\tan^{-1}\left(\frac{b_i}{a_i}\right), \ \cos^{-1}(F)\right\}.$$

This makes $h$ a *linear* function of $s_i$, for fixed $a_i, b_i$. That is, $h(s_i \ ; \ a_i, b_i) = (a_i \cos(\overline{\theta}_i) + b_i \sin(\overline{\theta}_i))s_i$. In this formulation, the value of $\overline{\theta}_i$ is completely determined by $a_i$ and $b_i$ (and $L$). Thus, the quantity $a_i \cos(\overline{\theta}_i) + b_i \sin(\overline{\theta}_i)$ (and hence $a_i$ and $b_i$ themselves) can be replaced by a single value, which we call $p_i$, completely determined by $a_i$, $b_i$, and $F$.

We note that we have posed the problem here all in terms of a *positive* $\theta_i$, corresponding to reactive power $q$ with positive phase angle. If the grid requested $q$ with

negative phase angle (still with price $b_i$), these calculations would follow in exactly the same way; the calculated value of $\overline{\theta}_i$ should take on the same sign as the phase angle of the requested reactive power.

**4.3.7. Problem Constraints.** In this section we discuss the incorporation of the battery charge $l_i$ and the solar irradiance $e_i$ into the revenue function $g$. Notably, $g$ may be simplified by setting the reserve capacity sold $r_i$ in terms of other variables, thereby eliminating it as an independent "controllable" variable.

4.3.7.1. *Constraints on Reserve Capacity.* The reserve capacity sold must be the smaller of the two:

- The charge in the battery minus the planned real energy sales plus the expected solar irradiance, with the appropriate efficiency corrections, or
- The maximum rated capacity of the inverter, $M$, minus the planned real power sales.

That is,

$$r_i \leq \min\{\gamma l_i - \cos(\overline{\theta}_i)s_i + e_i,\ M - \cos(\overline{\theta}_i)s_i\}$$

where $\gamma$ is a constant that accounts for the discharge rate limitations on the battery[3]. But in order to maximize revenue, we assume that the value of $r_i$ should be as large as possible. That is, we replace "$\leq$" in the above expression with "$=$" to obtain:

$$\begin{aligned} r_i &= \min\{\gamma l_i - \cos(\overline{\theta}_i)s_i + e_i,\ M - \cos(\overline{\theta}_i)s_i\} \\ &= \min\{\gamma l_i + e_i,\ M\} - \cos(\overline{\theta}_i)s_i \end{aligned}$$

[3]In general, a faster discharge rate in a battery results in a smaller effective capacity. We approximate this phenomenon with a constant coefficient $\gamma < 1$. This means that the amount of energy actually available is proportional to the charge in the battery. In practice, this "discharge efficiency" follows a non-linear curve, but we approximate it with the linear term $\gamma$.

This has the benefit, as mentioned above, of removing $r_i$ as a "controllable variable", leaving us with only two controllable variables, $s_i$ and $t_i$. We henceforth consider $g$ under this new formulation of $r_i$.

4.3.7.2. *Incorporating the Battery Charge $l_i$.* A certain efficiency must be applied when discharging (or charging) the battery [4]. We denote the loss associated with that efficiency by $\eta$. The level of battery charge in the next transactive interval must equal the current charge level plus the energy purchased minus the energy sold plus the irradiance, while accounting for $\eta$. This leads to the following formula for updating the charge level $l_i$:

$$\begin{aligned} l_{i+1} =& l_i + (t_i)(1-\eta) \\ & - (\underbrace{\max\{\cos(\overline{\theta}_i)s_i - e_i, 0\}}_{\substack{\text{Real energy sold} \\ \text{from battery}}})(1+\eta) + (\underbrace{\max\{e_i - \cos(\overline{\theta}_i)s_i, 0\}}_{\substack{\text{Solar applied} \\ \text{to battery charge}}})(1-\eta), \end{aligned} \tag{4.3.35}$$

where $L$ is the maximum charge the battery can hold[5].

This can be rewritten as

$$l_{i+1} = l_i + (t_i)(1-\eta) + (e_i - \cos(\overline{\theta}_i)s_i) - \eta|\cos(\overline{\theta}_i)s_i - e_i| \tag{4.3.36}$$

4.3.7.3. *Constraints on Energy Sold and Purchased.* The energy sold ($s_i$) and purchased ($t_i$) in the transactive market must be greater than or equal to zero and less than or equal to the maximum rated capacity of the inverter $M$. In addition, the amount of energy purchased or sold in transactive interval $i$ cannot result in the battery having charge greater than its capacity $L$ or less than zero in transactive interval $i+1$.

[4]This efficiency is different than the discharge rate limitation $\gamma$.
[5]Many batteries also have a lower limit on the charge they hold (that is, they must maintain some minimum level of charge). Without loss of generality, we assume that the lower limit is 0.

Thus we enforce the following effective constraints on $s_i$ and $t_i$:

$$\begin{aligned} 0 \leq & s_i \leq M, \\ 0 \leq & t_i \leq M, \\ 0 \leq & l_{i+1} \leq L. \end{aligned}$$

It is notable that these choices of constraints allow for energy to be purchased and sold to the grid in the same transactive interval. Such a situation could arise in a traditional market if it were cheaper to purchase electricity than sell it. We keep this possibility to add flexibility to our objective function in consideration of traditional market interaction. In any case, the energy sold (or purchased) cannot exceed the rated capacity of the inverter.

**4.3.8. Optimization Problem Statement.** We restate the function $g$ (without $r_i$) as follows. Compare with the initial formulation (4.3.30):

$$g(s_i, t_i \ ; \ p_i, c_i, d_i, l_i, e_i) = (p_i)(s_i) + (d_i)(\min\{\gamma l_i + e_i, \ M\} - \cos(\overline{\theta}_i)s_i) - (c_i)(t_i),$$

where the term $p_i$ is given by

$$p_i = a_i \cos\overline{\theta}_i + b_i \sin\overline{\theta}_i$$

with $\overline{\theta}_i = \min\left\{\tan^{-1}\left(\frac{b_i}{a_i}\right), \ \cos^{-1}(F)\right\}$, as per the formulation (4.3.34).

Since the terms $p_i, c_i, d_i, e_i$ are fixed at the beginning of the problem, we can absorb them into the function $g$ itself, now writing $g_i$ instead of $g$ and no longer listing them as function inputs. Rearranging terms then leads to the formulation:

$$g_i(s_i, t_i \ ; \ l_i) = (p_i - d_i\cos(\overline{\theta}_i))(s_i) - (c_i)(t_i) + (d_i)(\min\{\gamma l_i + e_i, \ M\}). \tag{4.3.37}$$

This will be the term under the summation in the objective function.

Also for notational convenience, let us denote our update rule (4.3.36) by $f_i$. That is, $l_{i+1} = f_i(s_i, t_i; l_i)$ where

$$f_i(s_i, t_i; l_i) := l_i + (t_i)(1-\eta) + (e_i - \cos(\overline{\theta}_i)s_i) - \eta|\cos(\overline{\theta}_i)s_i - e_i|. \tag{4.3.38}$$

Finally, some value must be assigned to the level of charge $l_{n+1}$ left in the battery after the $n$ intervals have passed. We denote this final price as $p_f$; the user may determine the process by which it is calculated. The total revenue earned by the inverter can then be modeled as the sum of the $g_i$ function over the $i = 1, ..., n$ time periods plus the product $p_f l_{n+1}$.

We complete our objective function formulation and optimization problem statement as follows:

Given $L, \eta, \gamma, M, F, l_1$ and $a_i, b_i, c_i, d_i, e_i$ for $i = 1, ..., n$,

(4.3.39)

$$\text{maximize} \quad \left(\sum_{i=1}^{n}(p_i - d_i\cos(\overline{\theta}_i))(s_i) - (c_i)(t_i) + (d_i)(\min\{\gamma l_i + e_i,\ M\})\right) + p_f l_{n+1}$$

$$\text{over } s_1, t_1, ..., s_n, t_n$$

subject to

$$0 \leq s_i \leq M,$$

$$0 \leq t_i \leq M,$$

$$0 \leq l_{i+1} \leq L \qquad \text{for all } i = 1, ..., n,$$

where $l_2, ..., l_{n+1}$ are given by

$$l_{i+1} = l_i + (t_i)(1-\eta) + (e_i - \cos(\overline{\theta}_i)s_i) - \eta|\cos(\overline{\theta}_i)s_i - e_i|,$$

and where $p_1, ..., p_n$ are given by

$$p_i = a_i \cos\overline{\theta}_i + b_i \sin\overline{\theta}_i$$

with $\overline{\theta}_i = \min\left\{\tan^{-1}\left(\frac{b_i}{a_i}\right),\ \cos^{-1}(F)\right\}$.

**4.3.9. Implementation Considerations.** This section is dedicated to discussion of some specific issues related to implementing the method.

4.3.9.1. *Time Horizon.* Determining the optimal values $s_1, t_1, ..., s_n, t_n$ in (4.3.39) provides the control decisions $s_1, t_1$ for the current time period and the *best guess* at the control decisions for time periods $2, ..., n$. After time period 1 has passed, the value signal and irradiance forecasts for the *next* $n$ time periods are updated and the algorithm is run again. In a transactive energy system, the number $n$ of transactive intervals is fixed and is the same for all participants. Our model has the added flexibility of allowing the user to determine the $p_f$ value, thus reflecting any information or assumptions she or he has about prices beyond the $n^{th}$ transactive interval. The time periods in our formulation are all of equal length, but the method is extensible to time periods of varying lengths.

4.3.9.2. *Battery Degradation.* The current model provides a tool for determining the amount of revenue that could be earned by a smart inverter system (for investment decision purposes), or to determine optimal operation once the equipment is already in place. If the user wanted to associate a cost with the charge or discharge of the battery, this could be done via the efficiency parameter $\eta$.

4.3.9.3. *Inverter Power Electronics.* The inverter can provide reactive power without depleting the energy reserves of the battery. In this way, there is "nothing to optimize" regarding the ratio of real to reactive power provided, *except* when the inverter is being asked to provide power at its rated capacity $M$. In this case, there is a trade-off between providing real and reactive power, which is taken into account in the present model via the energy revenue function $h$, as discussed previously. In any case, the amount of real energy actually discharged from the battery is taken into account by the $\cos(\overline{\theta}_i)$ term in (4.3.38).

4.3.9.4. *Concavity of the Objective Function.* The objective function in (4.3.39) is concave as long as the terminal price $p_f$ is not negative, as explained in the Appendix.

4.3.9.5. *Uncertainty.* It is well-known that electricity prices in traditional markets are uncertain. In a transactive market, that uncertainty is accounted for in the process by which the value signal is determined. Additionally, the value signal is updated with each time step so it always reflects the most accurate forecast. From an operation standpoint, the only decision variable values that are actually of consequence are $s_1$ and $t_1$, which depend mostly on the problem variables $a_1, b_1, c_1, d_1$ and $l_1$, which are known with certainty.

The solar irradiance forecast $e = (e_1, ..., e_n)$ is updated with each real-time step, and naturally contains the most accurate values for the nearest time periods.

4.3.9.6. *Settings other than transactive control.* Although the method is formulated in a transactive control setting, it can easily be applied under other rate or compensatory structures. For example, if the market does not offer compensation for reactive power support or spinning reserve capacity, those prices can simply be set to zero. Fixed time-of-use or even flat rates can be input for the energy price $a = (a_1, ..., a_n)$ or cost $c = (c_1, ..., c_n)$.

## 4.4. Numerical Experiment.

This section describes a numerical experiment in which we use the objective function to develop the optimal control scheme $s_1, t_1, ..., s_n, t_n$ for the smart solar inverter and battery storage system. The input data for this experiment was carefully selected from the Midcontinent Independent System Operator (MISO) historical market data and is intended to closely represent a transactive value signal consisting of $n = 24$ transactive intervals, all one hour in length.

As discussed previously, the solution $s_1, t_1, ..., s_{24}, t_{24}$ to the optimization problem (4.3.39) provides the values of the decision variables $s_1$ and $t_1$ for the current time period and a prediction of the values $s_2, t_2, ..., s_{24}, t_{24}$ for the upcoming time periods.

After the algorithm is executed to find $s_1, t_1$ for the current time period, those values are recorded, and the information $a_2, b_2, c_2, d_2, e_2, ..., a_{25}, b_{25}, c_{25}, d_{25}, e_{25}$ is updated with the best available values. The problem (4.3.39) is then solved again but using $l_2$ and $a_2, b_2, c_2, d_2, e_2, ..., a_{25}, b_{25}, c_{25}, d_{25}, e_{25}$ as inputs to find $s_2$ and $t_2$. This process continues over the one-month time span of the numerical experiment.

**4.4.1. Input data.** To demonstrate the use of our objective function, we utilize price data from a Locational Marginal Price (LMP) node in Clinton, Illinois, (AMIL.CLINTO51) located in MISO territory. The cost to buy and price to sell energy are provided by the publicly available LMP historical data [**56**]. Specifically, we use the "Day-Ahead ExAnte Market LMPs", providing $/MWh hourly data. The Day-Ahead data simulates the predicted prices that would be used to inform DERs in a transactive energy market. The same value was used for both the cost to purchase ($c_i$) and the price to sell ($a_i$) real power. In the absence of a real-time market for reactive power, we estimate a price for reactive power ($b_i$) using the MISO Historical

Rate Information [**57**], June 2015, Schedule 2 for Reactive Supply and Voltage Control under Ameren Illinois, the electric utility that serves Clinton, IL. This gives us $b_i = \$0.86$ during on-peak (6 a.m. – 10 p.m.) and $b_i = \$0.41$ during off-peak. For the terminal price $p_f$ we use the average of the last five prices $p_f = \frac{1}{5}\sum_{i=n-4}^{n} p_i$.

The price for spinning reserve is provided by the Market Clearing Price (MCP) for operating reserves historical data [**56**]. Specifically, we use the report "ASM Day-Ahead Market ExAnte MCPs", providing $/MWh hourly data. Both the LMP and MCP Day-Ahead ExAnte values are calcualted by MISO using a Security Constrained Economic Dispatch process. For more information, please see the MISO Pricing Reports Reader's Guide [**58**].

The solar irradiance data is provided by the Surface Radiation Budget Network (SURFRAD) weather station in Bondville, IL., approximately 30 miles east of Clinton, IL. The SURFRAD network was established through the National Oceanic & Atmospheric Administration (NOAA). These solar irradiance datasets are publicly available via FTP download [**61**].

The data used are from the one-month period covering the month of June, 2015. Since the algorithm "looks ahead" at each iteration for the upcoming 24 hours, data from July 1, 2015 were used as well. The average price of energy over this time period is 23.72 $/MWh, with a high of 49.26 $/MWh and a low of -13.92 $/MWh. Negative LMP values occur when there is an *over-supply* of electricity or voltage on the grid, so that devices are actually *paid* to consume real power, or have *to pay* to produce it. During this time, these negative (day-ahead) LMP values occur twice, on June 6 from 2-5am and on June 7 from 2-6am. The ability of our objective function to handle negative price signals is highly important to its usefulness to DERs operating

in such a system. The average price for spinning reserve over this time period was 2.41 $/MWh, with a high of 12.98 $/MWh and a low of 0.20 $/MWh.

The following parameters were used for the numerical experiment:

- $F = 0.8$: The limit on power factor for power supplied by the inverter (%)
- $M = 10$: The maximum rated power output of the inverter (MW)
- $\gamma = 0.9$: Accounts for the discharge rate limitations on the battery (%)
- $\eta = 0.05$: One-way efficiency loss from charging/discharging the battery (%)
- $L = 50$: The maximum charge capacity of the battery (MWh)
- $l_1 = 20$: The starting charge in the battery (MWh)

**4.4.2. Numerical Results.** With these input values, the inverter earns $70,184.25 over the 30-day period using our optimal control scheme. The total energy generated by the PV system over the 30-day period is 1,596 MWh. Thus, our control method allows the inverter to earn $43.98 per MWh generated by the PV panel, on average.

We compare our results to a second control method called *business-as-usual* (BAU). Using this method, the inverter is programmed to sell exactly the amount of real power generated by the PV panels to the market for all periods where production is available. This method does not utilize the battery capability. In our BAU method we set the controller to sell zero energy during any negative LMP periods. Using the BAU method, the inverter would earn $45,810.98, or $28.70 per MWh. Thus, our algorithm, together with the battery and smart inverter capability, offers a $15.28 per MWh, or 53%, increase in revenue earned.

The figures below show the input variables and inverter behavior over a typical 3-day period. Figure 4.3 shows that the energy sold is not always exactly what is generated

by the PV panels. Figure 4.4 shows that the inverter tends to purchase energy in the early morning when prices are the lowest, but the amount purchased depends on the expected solar irradiance and prices. Moreover, when exceptionally high prices are anticipated, the inverter retains energy in the battery to be available for sale at that time. Figure 4.5 shows the price of energy and spinning reserve.

## 4.5. Chapter Conclusion

This work represents a contribution to the implementation of a transactive control system, demonstrating optimal control of a smart solar inverter with battery backup based on price signals. As the electric grid modernizes and becomes more responsive, a deep understanding of how DER devices may respond to price changes will be necessary. This understanding is key to enabling the distributed decision making necessary and desirable for the grid of the future. Further work in this direction is suggested below.

4.5.0.1. *Control on a large scale.* What are the appropriate prices to encourage DERs to optimally support the grid? How should these prices be set? Transactive energy alludes to such a method but other, more deterministic options may be explored. Additionally, the incorporation of an energy sales "smoothing parameter" may be explored, to prevent violent swings in energy sales based on price.

4.5.0.2. *Optimization methods.* We use the built-in optimization tool *fmincon* in MATLAB to obtain our optimal values. A more efficient or elegant solution may be explored. Fast optimization approaches are needed as electricity markets become more and more granular. A constrained DC programming approach may be applied to the problem, since the objective function can be expressed as a difference of convex functions when the terminal price $p_f$ is negative, as explained in the Appendix. A

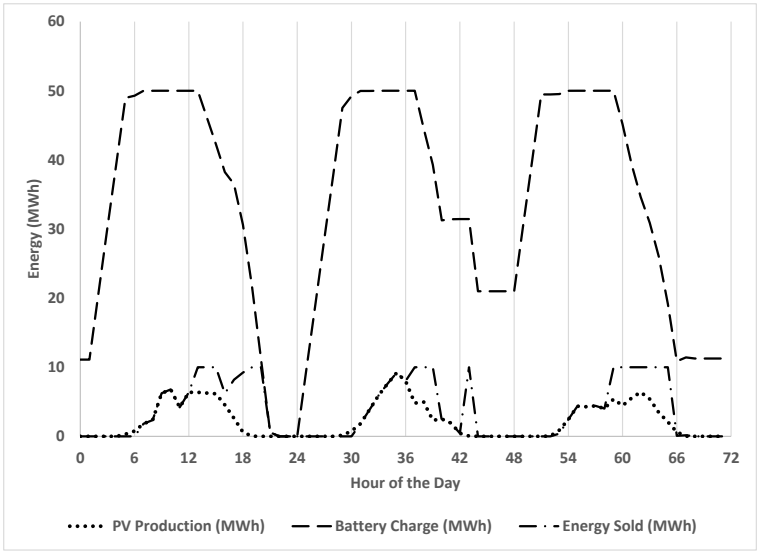



FIGURE 4.3. Typical charge and discharge behavior can be seen over the 3-day period June 20-22, 2015.

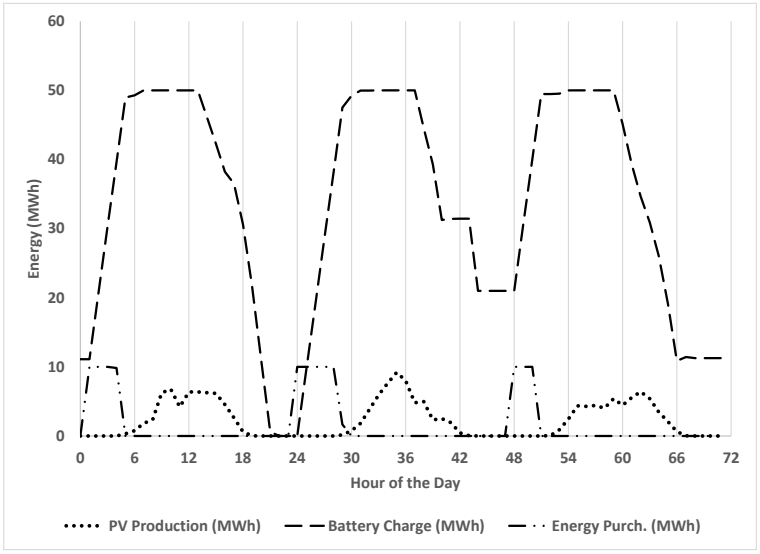



FIGURE 4.4. Battery charge, solar irradiance, and energy purchased over the 3-day period June 20-22, 2015.

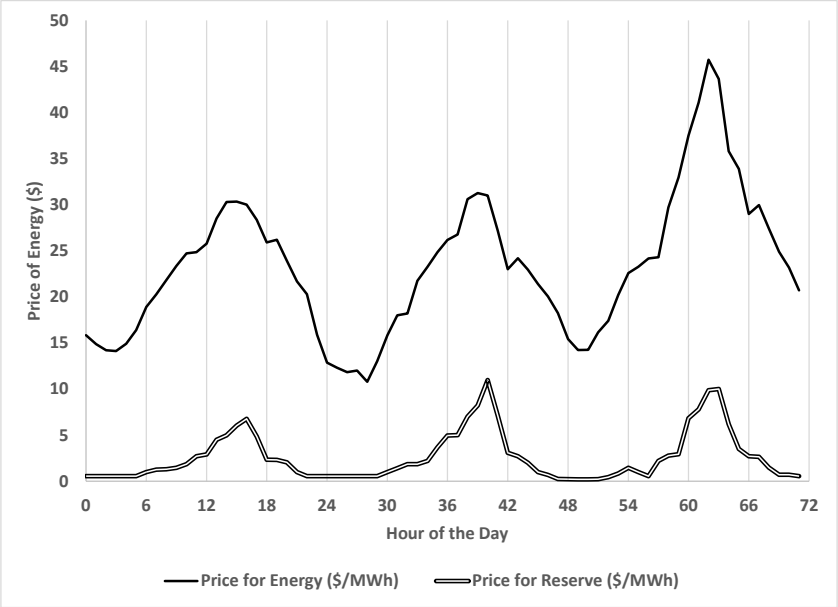



FIGURE 4.5. Price of energy and spinning reserve over the 3-day period June 20-22, 2015.

version of stochastic gradient descent may also be fruitful, since the objective function can be expressed as a sum of convex functions (with the additional restriction that $p_f \geq 0$). A method known as *dynamic programming* may also be useful in these types of problems. This method has already been used to approach various problems within smart grid [**63, 3, 84, 73, 85, 34**], and is well-known in economics [**71**].

4.5.0.3. *Valuation of the final charge.* The valuation of the final charge in the battery may significantly effect the accuracy of the $s_2, t_2, ..., s_n, t_n$ control estimates, especially those closer to $n$. Further work could provide a more accurate estimate. Also, the discharge rate limitation could be more accurately modeled with some non-linear function, rather than the linear estimate $\gamma$.

**4.5.1. Significance of this Research.** As residential-based energy storage and distributed PV generation become even more prevalent, customers and utilities alike will require a method for deciding when to use the PV generated energy to charge the battery, when to sell to the grid, and when to purchase energy from the grid. In certain cases, customers are dissuaded from purchasing energy from the grid to charge their onsite batteries since enacting such capabilities negates the Income Tax Credit (ITC) benefit from installing solar panels. A method such as the one presented here enables customers to perform an evaluation to see if the additional money earned is worth the loss from forfeiting the ITC. In general, enabling an optimal control scheme for smart solar inverters from a revenue standpoint contributes toward a more efficient electricity market.

# 5

# Conclusion

This thesis contains contributions in three areas, all of which are related to optimization. The work in Chapter 2 demonstrates the usefulness of the geometric approach in variational analysis. We see how this approach provides new proofs to fundamental results in convex analysis. These new proofs are such an improvement, in fact, that these arguments can now be presented to a beginning graduate or even undergraduate audience. This represents an important step forward for the field of convex analysis and for variational analysis as a whole. In addition, this chapter contains new theorems giving practitioners new facility with coderivative mappings. These theorems represent improvements to those previously found in the literature, and represent the first time these formulas have been proven in an infinite dimensional setting.

The work in Chapter 3 represents a significant contribution in the application of fast first-order methods and DC programming to convex and non-convex problems in location science. As demonstrated, this type of distance minimization has direct applications to data science including machine learning (support vector machines) and clustering, as well as the practical applications of facility location. Some of the work in this chapter also utilizes the theoretical findings on the subdifferential presented in the previous chapter.

Finally, in Chapter 4 we see an application of optimization to a control problem in the modern energy landscape. As variable generation becomes more prevalent, individual resources must be equipped to respond to real-time market signals in a

way that maximizes their own best interest and hence provides optimal support to the electric grid. In such a transactive type system, resources settle on prices for services locally, so that the price reflects the value of that service to that specific location of the grid at that time. The work provided in this chapter provides valuable insight into how distributed resources may respond to price signals in such a system. This work also solves the very practical optimization problem of solar plus storage resource providing energy and spinning reserve in a wholesale market. In addition, the presented model provides the additional flexibility that revenue could be earned by providing reactive power support, a forward-looking market feature.

## References


[1] D. An and N. Yen, *Differential stability of convex optimization problems under inclusion constraints*, Applicable Analysis, 94 (2015).

[2] L. An, M. Belghiti, and P. Tao, *A new efficient algorithm based on DC programming and DCA for clustering*, J. Global Optim., 37 (2007), pp. 593–608.

[3] L. An and T. Quoc-Tran, *Optimal energy management for grid connected microgrid by using dynamic programming method*, in IEEE Power & Energy Society General Meeting, July 2015.

[4] N. An, D. Giles, N. Nam, and R. Rector, *Log-exponential smoothing technique and Nesterov's accelerated gradient method for generalized Sylvester problems*, Journal of Optimization Theory and Applications, 168 (2016), p. 559.

[5] A. Annaswamy and T. Nudell, *Transactive control – what's in a name?*, in IEEE Smart Grid Newsletter, September 2015.

[6] S. Barrager and E. Cazalet, *Transactive Energy: A Sustainable Business and Regulatory Model for Electricity*, Baker Street Publishing, San Francisco, 2014.

[7] M. Bayat, K. Sheshyekani, and A. Rezazadeh, *A unified framework for participation of responsive end-user devices in voltage and frequency control of the smart grid*, IEEE Trans. Power Syst., 30 (2015).

[8] M. Biabani, M. A. Golkar, A. Johar, and M. Johar, *Propose a home demand-side-management algorithm for smart nano-grid*, in 4th Power Electronics, Drive Systems & Technologies Conference (PEDSTC2013), February 2013.

[9] J. Borwein and Q. Zhu, *Techniques of Variational Analysis*, CMS Books in Mathematics, Springer, New York, 2005.

[10] J. Brimberg, *The Fermat Weber location problem revisited*, Math. Program., 71 (1995), pp. 71–76.

[11] M. Carrión, J. M. Arroyo, and A. J. Conejo, *A bilevel stochastic programming approach for retailer futures market trading*, IEEE Trans. Power Syst., 24 (2009).

[12] S. Chandler and J. Hughes, *Smart grid distribution prediction and control using computational intelligence*, in Conference on Technologies for Sustainability, 2013.

[13] D. Chassin, N. Lu, J. Malard, S. Katipamula, C. Posse, J. Mallow, and A. Gangopadhyaya, *Modeling power systems as complex adaptive systems*, Tech. Rep. Contract DE-AC05-76RL01830, Pacific Northwest National Laboratory, Richland, WA, December 2004.

[14] E. Chi and K. Lange, *A look at the generalized Heron problem through the lens of majorization-minimization*, Amer. Math. Monthly, 121 (2014).

[15] E. Chi, H. Zhou, and K. Lange, *Distance majorization and its applications*, Mathematical Programming, Series A, 146 (2014), pp. 409–436.

[16] Z. Drezner, *On the convergence of the generalized Weiszfeld algorithm*, Ann. Oper. Res., 167 (2009), pp. 327–336.

[17] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, *Efficient projections onto the $\ell_1$-ball for learning in high dimensions*, in Proceedings of the 25th international conference on Machine learning. ACM, pp. 272–279., 2008.

[18] W. Fenchel, *Convex cones, sets and functions, lecture notes, Princeton University*, 1951.

[19] D. Hammerstorm, T. Oliver, R. Melton, and R. Ambrosio, *Standardization of a hierarchical transactive control system*, in Grid-Interop Forum, November 2007.

[20] Y. He and K. F. Ng, *Subdifferentials of a minimum time function in Banach spaces*, J. Math. Anal. Appl., 321 (2006), pp. 896–910.

[21] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms I*, Springer-Verlag, 1993.

[22] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*, Springer-Verlag, 2001.

[23] J. Hodgson, K. Rosing, and J. Zhang, *Locating vehicle inspection stations to protect a transportation network*, Geog. Anal., 28 (1996), pp. 299–314.

[24] P. Huang, J. Kalagnanam, R. Natarajan, M. M. Sharma, R. Ambrosio, D. Hammerstrom, and R. Melton, *Analytics and transactive control design for the Pacific Northwest smart grid demonstration project*, in Proc. First IEEE Intl Conf. on Smart Grid Communications, October 2010.

[25] D. Hunter and K. Lange, *Tutorial on MM algorithms*, Amer. Statist., 58, pp. 30–37.

[26] T. Jahn, Y. Kupitz, H. Martini, and C. Richter, *Minsum location extended to gauges and to convex sets*, J. Optim. Theory Appl., 166 (2015), pp. 711–746.

[27] S. Katipamula, D. Chassin, D. Hatley, R. Pratt, and D. Hammerstorm, *Transactive controls: Market-based gridwise controls for building systems*, Tech. Rep. Contract DE-AC05-76RL01830, Pacific Northwest National Laboratory, Richland, WA, July 2006.

[28] J. Kosowatz, *Making the grid really smart.* https://www.asme.org/engineering-topics/articles/energy/making-the-grid-really-smart, February 2016.

[29] H. Kuhn, *A note on Fermat-Torricelli problem*, Math. Program., 4 (1973), pp. 98–107.

[30] Y. Kupitz and H. Martini, *Geometric aspects of the generalized Fermat-Torricelli problem*, in Intuitive Geometry, Bolyai Society of Mathematical Studies, I. Barany and K. Boroczky, eds., vol. 6, 1997, pp. 55–127.

[31] Y. Kupitz, H. Martini, and M. Spirova, *The Fermat-Torricelli problem, part i: A discrete gradient-method approach*, J. Optim. Theory Appl., 158 (2013), pp. 305–327.

[32] K. Lange, D. Hunter, and I. Yang, *Optimization transfer using surrogate objective functions*, J. Comput. Graph. Statist., 9 (2000), pp. 1–59.

[33] B. Lemarie, *Applications of a subdifferential of a convex composite functional to optimal control in variational inequalities*, in Nondifferentiable Optimization: Motivations and Applications, Sopron, Hungary, 1984.

[34] D. Maly and K. Kwan, *Optimal battery energy storage system (BESS) charge scheduling with dynamic programming*, in IEE Proc.-Sci. Meas. Technol., November 1995.

[35] H. Martini, K. Swanepoel, and G. Weiss, *The Fermat-Torricelli problem in normed planes and spaces*, J. Optim. Theory Appl., 115 (2002), pp. 283–314.

[36] H. Min, *A multiobjective retail service location model for fast food resteraunts*, Omega, 15 (1987), pp. 429–441.

[37] H. Minkowski, *Theorie der Konvexen Körper, Insebesondere Begründung inhres Ober Flächenbegriffs*, Gesammelte Abhandlungen, G. Teubner, Leipzig, 1911.

[38] B. Mordukhovich, *Variational Analysis and Generalized Differentiation*, Springer, Berlin, 2006.

[39] B. Mordukhovich, M. Nam, R. Rector, and T. Tran, *Variational geometric approach to generalized differential and conjugate calculi in convex analysis*, submitted, (2016).

[40] B. Mordukhovich and N. Nam, *Applications of variational analysis to a generalized Fermat-Torricelli problem*, J. Optim. Theory Appl., 148 (2011), pp. 431–454.

[41] ——, *An Easy Path to Convex Analysis and Applications*, Synthesis Lectures on Mathematics and Statistics: book series edited by S.G. Krantz, Morgan & Claypool Publishers, 2014.

[42] B. Mordukhovich, N. Nam, and J. Salinas, *Applications of variational analysis to a generalized Heron problem*, Appl. Anal., 91 (2012), pp. 1915–1942.

[43] ——, *Solving a generalized Heron problem by means of convex analysis*, Amer. Math. Monthly, 119 (2012), pp. 87–99.

[44] B. Mordukhovich, N. Nam, and C. Villalobos, *The smallest enclosing ball problem and the smallest intersecting ball problem: existence and uniqueness of optimal solutions*, Optim. Lett., 7 (2013), pp. 839–853.

[45] J.-J. Moreau, *Fonctionelles sous-différentiables*, C.R.Acad.Sci., 257 (1963), pp. 4117–4119.

[46] N. Nam, N. An, R. Rector, and J. Sun, *Nonsmooth algorithms and Nesterov's smoothing technique for generalized Fermat-Torricelli problems*, SIAM J. Optim., 24 (2014), pp. 1815–1839.

[47] N. Nam, N. An, and C. Villalobos, *Minimal time functions and the smallest intersecting ball problem generated by unbounded dynamics*, J. Optim. Theory Appl., 154 (2012), pp. 768–791.

[48] N. Nam and N. Hoang, *A generalized Sylvester problem and a generalized Fermat-Torricelli problem*, J. Convex Analysis, 20 (2013), pp. 669–687.

[49] N. Nam, N. Hoang, and R. Rector, *A unified approach to convex and convexified generalized differentiation of nonsmooth functions and set-valued mappings*, Vietnam Journal of Mathematics, 42 (2014), pp. 479–497.

[50] N. Nam, R. Rector, and D. Giles, *Minimizing differences of convex functions with applications to facility location and clustering*, J. Optim. Theory Appl., 173 (2017), pp. 255–278.

[51] Y. Nesterov, *A method for unconstrained convex minimization problem with the rate of convergence* $o(\frac{1}{k^2})$, Doklady AN SSSR (translated as Soviet Math. Docl.), 269 (1983), pp. 543–547.

[52] ——, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 127–152.

[53] W. Parks and R. Risser, *The role of transactive energy in grid modernization: Leveraging building technologies*, in NASEO Energy Outlook Conference, February 2014.

[54] Bonneville Power Administration, *Pacific Northwest smart grid demonstration project, technology partners*. `http://www.bpa.gov/energy/n/Smart_Grid-Demand_Response/\\Smart_Grid/PNWSGDP/TechnologyPartners.cfm`.

[55] GridWise Architecture Council, *Gridwise transactive energy framework version 1.0*. `http://www.gridwiseac.org/pdfs/te_framework_report_pnnl-22946.pdf`.

[56] Midcontinent Independent System Operator, *Market reports*. `https://www.misoenergy.org/Library/MarketReports/Pages/MarketReports.aspx`.

[57] ——, *MISO historical rate information*. `http://www.oasis.oati.com/woa/docs/MISO/MISOdocs/Historical_Rate.html`.

[58] ——, *MISO pricing reports readers' guide*. https://www.misoenergy.org/Library/Repository/Report/Readers\\\%20Guide/MISO\%20Pricing\%20Reports\%20Readers\%20Guide.pdf.

[59] Quality Logic, *What is transactive control (revision 1.0)*. http://www.qualitylogic.com/tuneup/uploads/docfiles/\\What-Is-Transactive-Control.pdf, February 2013.

[60] Resnick Institute, *Grid 2020: Towards a policy of renewable and distributed energy resources*, tech. rep., Resnick Institute, September 2012.

[61] Surface Radiation Network (SURFRAD). aftp.cmdl.noaa.gov.

[62] C. ReVelle, D. Bigman, and D. Schilling, *Facility location: a review of context-free and EMS models*, Health Serv. Res., 12 (1980), pp. 129–146.

[63] Y. Riffonneau, S. Bacha, F. Barruel, and S. Ploix, *Optimal power flow management for grid connected PV systems with batteries*, IEEE Trans. Sustain. Energy, 2 (2011).

[64] R. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, New Jersy, 1970.

[65] R. Rockafellar and J.-B. Wets, *Variational Analysis*, Springer, Berlin, 1998.

[66] R. T. Rockafellar, *Convex Functions and Dual Extremum Problems*, PhD thesis, Department of Mathematics, Harvard University, 1963.

[67] K. Rogers, R. Klump, H. Khurana, A. A. Aquino-Lugo, and T. Overbye, *An authenticated control framework for distributed voltage support on the smart grid*, IEEE Trans. Smart Grid, 1 (2010).

[68] S. Widergren, *et al.*, *AEP Ohio gridSMART demonstration project real-time pricing demonstration analysis*, Tech. Rep. Contract DE-AC05-76RL01830, Pacific Northwest National Laboratory, Richland, WA, February 2014.

[69] A. Safdarian, M. Fotuhi-Firuzabad, and M. Lehtonen, *A distributed algorithm for managing residential demand response in smart grids*, IEEE Trans. Ind. Informat., 10 (2014).

[70] S. Saks, *Theory of the Integral*, Hafner Publishing Co., New York, 1937.

[71] T. J. Sargent, *Dynamic Macroeconomic Theory*, Harvard University Press, 1987.

[72] F. C. Schweppe, R. D. Tabors, and J. L. Kirtley, *Homeostatic control : the utility customer marketplace for electric power*. Energy Laboratory report (Massachusetts Institute of Technology. Energy Laboratory) no. MIT-EL 81-033., 1981.

[73] R. Sioshansi, S. Madaeni, and P. Denholm, *A dynamic programming approach to estimate the capacity value of energy storage*, IEEE Trans. Power Syst, 29 (2014).

[74] R. G. Smith, *The contract net protocol: High-level communication and control in a distributed problem solver*, IEEE Trans. on Computers, C-29 (1980).

[75] S. Sobolev, *Applications of Functional Analysis in Mathematical Physics*, American Mathematical Society, Providence, Rhode Island, 1963.

[76] D. Stimoniaris, D. Tsiamitros, and E. Dialynas, *Improved energy storage management and PV-active power control infrastructure and strategies for microgrids*, IEEE Trans. Power Syst., 31 (2016).

[77] J. Taft and P. D. Martini, *Ultra large-scale power system control architecture*. `http://www.cisco.com/web/strategy/docs/energy/\\control_architecture.pdf`, October 2012.

[78] T. Tan, *An extension of the Fermat-Torricelli problem*, J. Optim Theory Appl., 146 (2010), pp. 735–744.

[79] P. Tao and L. An, *Convex analysis approach to d.c. programming: Theory, algorithms and applications*, ACTA Mathematica Vietnamica, 22 (1997), pp. 289–355.

[80] ——, *A d.c. optimization algorithm for solving the trust-region subproblem*, SIAM J. Optim., 8 (1998), pp. 476–505.

[81] Y. Vardi and C.-H. Zhang, *The multivariate $l_1$-median and associated data depth*, Proceedings of the National Acadamey of Sciences of the United States of America, 97 (2000), pp. 1423–1426.

[82] G. Wang, M. Negrete-Pincetic, A. Kowli, E. Shafieepoorfard, S. Meyn, and U. V. Shanbhag, *Real-time prices in an entropic grid*, in Conference on Decision and Control, 2011.

[83] E. Weiszfeld, *Sur le point pour lequel la somme des distances de n points donnés est minimum*, Tohoku Math. J, 43 (1937), pp. 355–386.

[84] Y. Xu, W. Zhang, and W. Liu, *Distributed dynamic programming approach for economic dispatch in smart grids*, IEEE Trans. Ind. Informat., 11 (2015).

[85] L. Zhang and Y. Li, *Optimal energy management of hybrid power system with two-scale dynamic programming*, in American Control Conference, June 2011.

[86] T. Zhou, D. Tao, and X. Wu, *NESVM, a fast gradient method for support vector machines*, in IEEE International Conference on Data Mining (ICDM), 2010.

[87] M. Zugno, J. Morales, P. Pinson, and H. Madsen, *A bilevel model for elecrticity retailers' participation in a demand response market environment*, Energy Economics, 36 (2013).

## Appendix

# Concavity of the Smart Solar Inverter Objective Function.

The purpose of this Appendix is to discuss the concavity of the objective function presented in (4.3.39). Recall that a function $f$ is *concave* if $-f$ is convex. As we will see, this objective function is concave when the terminal price $p_f$ is not negative.

Let $s = (s_1, ..., s_n)$ and $t = (t_1, ..., t_n)$ be the decision variables and $p = (p_1, ..., p_n)$ and $\overline{\theta} = (\overline{\theta}_1, ..., \overline{\theta}_n)$ be the price and phase angle as formulated in section 4.3.6. Then the objective function is

$$\sum_{i=1}^{n} \Big[(p_i - d_i \cos(\overline{\theta}_i))(s_i) - (c_i)(t_i) + (d_i)(\min\{\gamma l_i + e_i, M\})\Big] + p_f l_{n+1}, \tag{0.0.40}$$

where

$$l_{i+1} = f_i(s_i, t_i; l_i) = l_i + (t_i)(1-\eta) + (e_i - \cos(\overline{\theta}_i)s_i) - \eta|\cos(\overline{\theta}_i)s_i - e_i|. \tag{0.0.41}$$

Note that each function $f_i$ is concave in $s_i$, $t_i$ since the first part, $l_i + (t_i)(1-\eta) + (e_i - \cos(\overline{\theta}_i)s_i)$, is affine, and the second part, $-\eta|\cos(\overline{\theta}_i)s_i - e_i|$, is concave. Also note that we can rewrite equation (0.0.41) as

$$\begin{aligned} l_{i+1} &= f_i(s_i, t_i; l_i) \\ &= l_1 + \sum_{j=1}^{i} \big[(t_j)(1-\eta) + (e_j - \cos(\overline{\theta}_j)s_j) - \eta|\cos(\overline{\theta}_j)s_j - e_j|\big]. \end{aligned} \tag{0.0.42}$$

This shows us how $l_{i+1}$ can be expressed as a function of $l_1$ and $s_1, ..., s_i$ and $t_1, ..., t_i$. Since each piece under the sum in (0.0.42) is concave, the entire sum is concave, and so the function $f_i$ can be seen as a concave function of the (entire) vectors $s$ and $t$. To reflect all this, we now write $l_{i+1} = f_i(s,t)$ for $i = 1, ..., n$ with the formulation as in (0.0.42). We define $f_0(s,t) := l_1$.

The entire sum (0.0.40) can then be rewritten as

$$\underbrace{\sum_{i=1}^{n}\left[(p_i - d_i\cos(\overline{\theta}_i))(s_i)\right] - \sum_{i=1}^{n}\left[(c_i)(t_i)\right]}_{Part\ A} + \underbrace{\sum_{i=1}^{n}\left[(d_i)(\min\{\gamma f_{i-1}(s,t) + e_i, M\})\right]}_{Part\ B} + \underbrace{p_f l_{n+1}}_{Part\ C}. \tag{0.0.43}$$

We see that $Part\ A$ of (0.0.43) is linear in $s, t$. Let us inspect $Part\ B$. Consider a single member of the sum:

$$(d_j)(\min\{\gamma f_{j-1}(s,t) + e_j, M\}).$$

As discussed above, the function $f_{j-1}$ is concave in $s, t$. Since $\gamma$ is not negative, the function $\gamma f_{j-1}(s,t) + e_j$ is also concave in $s, t$. The minimum of concave functions is concave, so $\min\{\gamma f_{j-1}(s,t) + e_j, M\}$ is concave. Finally, since $d_j \geq 0$ (there would never be a negative price for reserve capacity), we can see that the entire expression is concave in $s, t$. So the sum in $Part\ B$ is concave. Finally, we look at $Part\ C$. Since $l_{n+1} = f_n(s,t)$ is concave, we can see that $p_f l_{n+1}$ is concave as long as the price $p_f$ is not negative.

Thus the expression (0.0.43) is concave as long as the terminal price $p_f$ is not negative.

**Numerical Testing of the Concavity of the Objective Function.**

We randomly draw $s, t$ and $\alpha$ to confirm the concavity of the objective function. For notational convenience, denote the objective function (0.0.40) as $g(x)$ where $x = [s; t]$ is a single vector containing $s$ and $t$. The following is a summary of the process used for testing. First, we fix the number of time periods $n \in \mathbb{N}$.

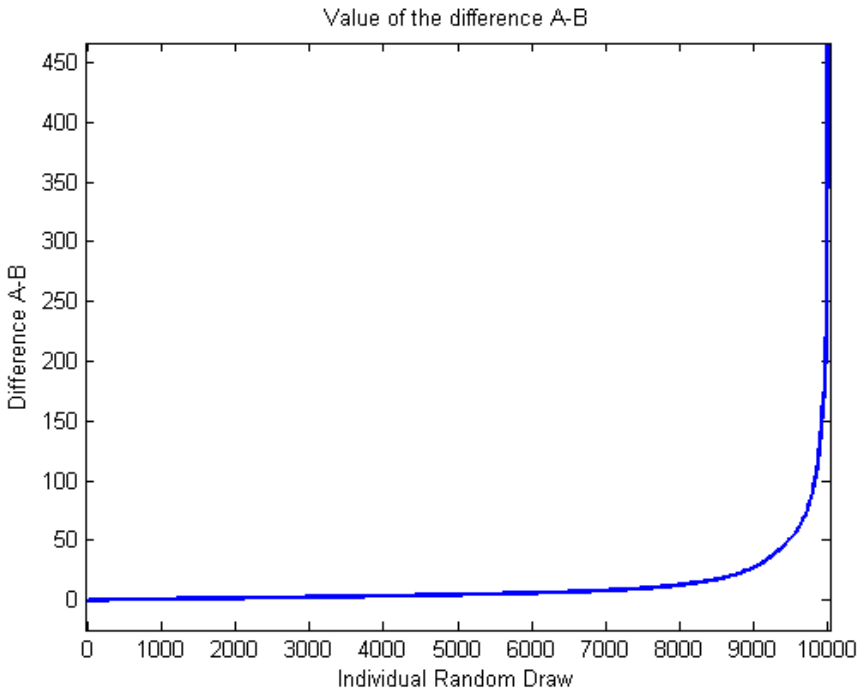
**Step 1.** Choose uniformly at random $x_1 \in [0, M]^{2n}$, $x_2 \in [0, M]^{2n}$, and $\alpha \in [0, 1]$.

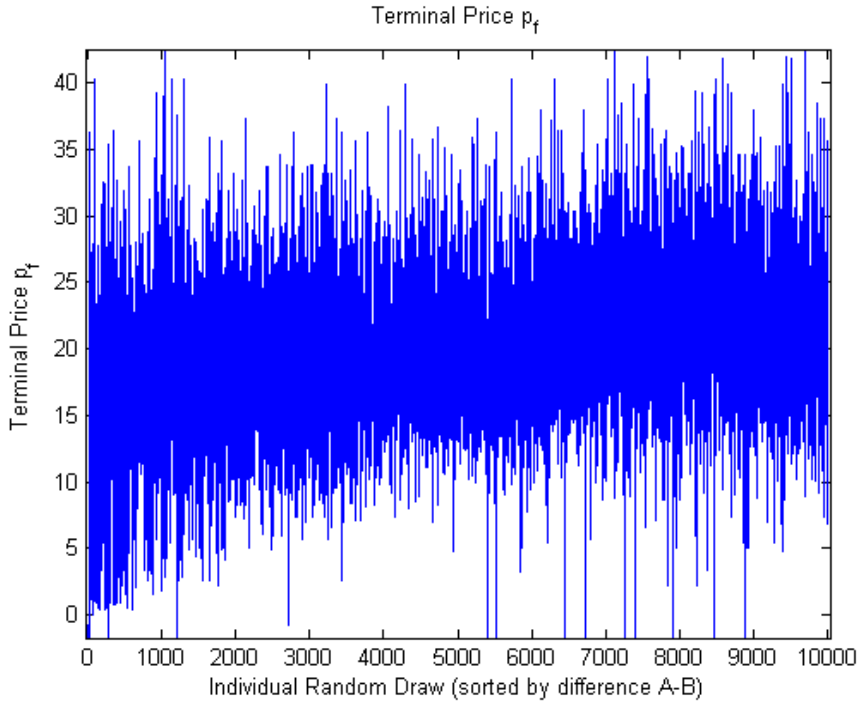
**Step 2.** Calculate $A := g(\alpha x_1 + (1 - \alpha)x_2)$ and $B := \alpha g(x_1) + (1 - \alpha)g(x_2)$.

**Step 3.** Calculate the difference $A - B$. If $g$ is concave, we expect $A - B \geq 0$.

For our test we use $n = 24$ as in section 4.4. The input data used to define $g$ for each execution of Step 2 are chosen at random as follows. For each execution, randomly choose an hour in the March-July MISO data to be the starting hour for the $n = 24$ contiguous time periods of input data. Vary the input data starting place at random in the available March-July time frame for each execution of the steps above. The remainder of the function parameters are the same as in the numerical experiment in section 4.4.

In one such experiment, we executed these steps 10,000 times. Out of these 10,000 draws, the concavity of $g$ was violated 25 times. In each of these 25 occurrences, the terminal price $p_f$ was negative, which confirms our assertion that the objective function is concave as long as the terminal price is not negative. These 10,000 draws are shown in the plots below. The plot titled *Value of the difference A-B* gives a sense of how "strongly" concave the function $g$ is. The plot titled *Terminal Price* $p_f$ shows the value of the terminal price $p_f$, arranged in the same order as the other plot.


Value of the difference A-B
Difference A-B
Individual Random Draw
0
50
100
150
200
250
300
350
400
450
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000


Terminal Price $p_f$
Terminal Price $p_f$
Individual Random Draw (sorted by difference A-B)
0
5
10
15
20
25
30
35
40
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000

The plots for another run of the same experiment with 500 draws are shown below. Out of these 500 draws, concavity was violated 3 times, all having negative terminal price, as can be seen below.