Universidade Federal do Rio de Janeiro

Convexification by Averages

Iago Leal de Freitas

Rio de Janeiro

Novembro de 2019

Universidade Federal do Rio de Janeiro

Convexification by Averages

Iago Leal de Freitas

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Matemática, Instituto de Matemática da Universidade Federal do Rio de Janeiro (UFRJ), como parte dos requisitos necessários à obtenção do título de Mestre em Matemática.

Advisor: Bernardo Freitas Paulo da Costa

Rio de Janeiro

Novembro de 2019

Leal de Freitas, Iago
Convexification by Averages / Iago Leal de Freitas. - Rio de Janeiro: UFRJ/IM, 2019.
xvi, 143f.: il.; 29,7cm.
Orientador: Bernardo Freitas Paulo da Costa
Dissertação (Mestrado) - UFRJ/IM/Programa de Pós-Graduação em Matemática, 2019.
Referências Bibliográficas: f. 141-143
a. Stochastic Programming. b. Cutting-plane method. c. Convex Analysis. d. Theory of Distributions. I. Freitas Paulo da Costa, Bernardo. II. Universidade Federal do Rio de Janeiro, Programa de Pós-Graduação em Matemática. III. Título.

Convexification by Averages

Iago Leal de Freitas

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Matemática, Instituto de Matemática da Universidade Federal do Rio de Janeiro (UFRJ), como parte dos requisitos necessários à obtenção do título de Mestre em Matemática.

Aprovada por:

Prof. Bernardo Freitas Paulo da Costa (Orientador)

Dr. Sérgio Granville

Prof. Vincent Gérard Yannick Guigues

Prof. César Javier Niche Mazzeo

Prof. Alexander Shapiro

Rio de Janeiro

Novembro de 2019

Acknowledgments

Although the content of this work is written entirely in English, the following words are written in Portuguese, my native language. This is done with the hope that the acknowledged people may know how important they were for the process of writing this thesis.

Gostaria de começar agradecendo aos meus pais Giselle e Danrlei e minha irmã Maria Bela. Se eu cheguei até aqui, com certeza foi graças ao seu suporte. Também agradeço aos meus avós Clézio, Sedalice, Marly e Danrlei e aos meus tios Carol, Luciano e Alexandre por todo apoio e suporte. Especialmente, quero agradecer a Giselle e Clézio, pois com vocês aprendi a beleza que existe em se criar algo, seja concreto ou abstrato.

Agradeço ao meu orientador Bernardo Costa por ter sido um ponto ótimo no espaço de orientadores. Obrigado por todo o seu suporte e pelo tanto que acreditou em mim. Nesses dois últimos anos nós debatemos *muita* matemática mas também houve diversos concertos na Sala Cecília Meireles e até planos de seguir de bicicleta do Leme ao Pontal. Além de um grande orientador e matemático, você também é um grande amigo.

Sinceramente agradeço aos professores Sérgio Granville, Vicent Guigues, César Niche e Alexander Shapiro por aceitarem fazer parte dessa banca e se darem o tempo de ler e comentar esta dissertação.

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo fomento recebido através de uma bolsa de mestrado durante os últimos dois anos. Também agradeço à Fundação Coppetec e ao Operador Nacional do Sistema Elétrico (ONS). Durante o período de mestrado, participei do projeto IM-21780 de colaboração entre a UFRJ e o ONS, o que serviu de motivação para muito do que está escrito nesta dissertação.

Gostaria de agradecer ao professor João Paixão por ter me mostrado o quão interessante a matemática pode ser. Com certeza aquelas aulas sobre números primos e grupos finitos, quando eu era apenas um calouro que mal sabia integrar polinômios, foram o que pela primeira vez me mostrou que matemática não é só sobre fazer contas mas também, principalmente, sobre a beleza que existe em encontrar padrões no mundo.

Tenho que agradecer a Alexandre Moreira, Patrick Oliveira e Pedro Aragão pelas infindáveis discussões sobre matemática (e a vida) que tivemos nestes

últimos anos. É difícil contar a quantos resultados interessantes eu cheguei graças à ajuda de vocês. Obrigado por passarem tanto tempo me ouvindo tagarelar sobre espaços de funções, distribuições, funções convexas e autovalores de medidas matriciais.

Agradeço a Rafael Klausner por esses últimos meses, fosse me acompanhando nas reuniões ou prontamente me enviando algum arquivo com políticas e cortes do SDDP. Sem você essa dissertação não teria ficado pronta, literalmente.

Agradeço a Filipe Cabral e Joari Costa. Quero que saibam que são inspirações para mim e que foi um prazer trabalhar com vocês nesses últimos anos. Também gostaria de estender esses agradecimentos a Débora Jardim que esteve presente nesses últimos meses de projeto.

Menciono aqui todos os amigos que fiz na Matemática Aplicada ou Computação e todos os momentos que passamos juntos, para mim vocês são como uma segunda família. Cito aqui Bruno Lima Netto, Cynthia Herkenhoff, Danilo Naiff, Felipe Pagginelli, Gabriel Picanço, Gabriela Lewenfus, Gustavo Martins, Ivani Ivanova, Jessica Nascimento, Leonardo Gama, Marcelo Carneiro, Maria Luiza Avlis, Matheus Fontoura, Pedro Xavier, Ricardo Turano, Rodrigo Lima, Thiago Holleben, Vítor Luiz e todos os outros que fizeram parte desta jornada junto comigo. Estendo esses agradecimentos aos professores Alejandro Cabrera, Fábio Ramos, Felipe Acker, Hugo Carvalho e aos outros professores do Instituto de Matemática.

Também cito o pessoal do Mononoke Hime: Erick Pires, Diego Tertuliano, Gustavo Moreira e Tetsuo Shiino (além dos previamente citados Alexandre e Pedro). Até hoje guardo no coração todas as infindáveis conversas que tivemos. Vocês certamente forem essenciais para me manter de pé nos momentos mais complicados da minha vida acadêmica.

Quero citar meus amigos André Sales, Carolina Pierre e Guilherme Gurgel (além da Alice, é claro!) que mesmo estudando áreas totalmente distintas estiveram aqui junto comigo desde os longínquos temos do ensino médio. Na verdade, estendo esse agradecimento a todos os amigos dos tempos do CEFET com quem tenho o prazer de ainda manter contato.

Por fim, agradeço a Nathalie Deziderio por todo apoio durante a escrita dessa dissertação. É difícil dizer o quão importante você foi, seja lendo e comentando as versões preliminares, seja me ajudando a não enlouquecer com a cabeça perdida em convoluções e aproximações poliedrais.

Resumo

Convexification by Averages

Iago Leal de Freitas

Orientador: Bernardo Freitas Paulo da Costa

Essa dissertação é dedicada ao estudo de funções de valor ótimo para programas estocásticos não convexos e como estas funções podem ser aproximadas por cortes exatos. São estudadas duas maneiras de se medir não convexidade: o *gap* entre uma função e sua relaxação convexa e a parte negativa da segunda derivada de uma função. Baseando-nos nas semelhanças destes dois operadores, introduzimos o conceito de medida de não convexidade. Estas tem a propriedade de sempre serem reduzidas pela operação de tomar médias. Diversos destes resultados também continuam valendo para o *gap* ao substituir-se a média por uma medida de risco coerente qualquer. Esses resultados são aplicados à aproximação de programas estocásticos multiestágio não convexos por cortes válidos ao considerarmos a diferença entre encontrar um corte médio através de uma formulação decomposta ou conjunta para os cenários.

Palavras-chave: Stochastic Programming, Cutting-plane method, Convex Analysis, Theory of Distributions.

Abstract

Convexification by Averages

Iago Leal de Freitas

Advisor: Bernardo Freitas Paulo da Costa

This dissertation is dedicated to the study of optimal value function for non-convex stochastic programs and how these functions can be approximated by tight cuts. Two ways to measure non-convexity are studied in this work: the gap between a function and its convex relaxation, and the negative part of a function's second derivative. Influenced by the similarities between these two operators, we introduce the concept of a non-convexity measure. These have the property of being reduced by the operation of taking averages. Many of these results also hold for the gap when considering an arbitrary risk measure instead of the expected value. Theses results can be applied for approximating non-convex multi-stage stochastic programs by tighter cuts by considering the difference between calculating an average cut via a decomposed or a linked formulation for the scenarios.

Keywords: Stochastic Programming, Cutting-plane method, Convex Analysis, Theory of Distributions.

Contents

List of Figures **xv**

1 Introduction **1**

2 Fundamentals of Convex Analysis **5**
2.1 Convex sets . . . 6
2.1.1 Convexity-preserving operations . . . 7
2.2 Convex functions . . . 11
2.2.1 Convexity-preserving operations . . . 14
2.2.2 Convex relaxation . . . 16
2.2.3 Modes of convergence . . . 18
2.2.3.1 Lower semi-continuity . . . 19
2.2.3.2 Epi-convergence . . . 22
2.2.4 Fenchel conjugate . . . 22
2.2.4.1 The biconjugate . . . 24
2.3 Cones and inequalities . . . 25
2.3.1 Convexity in relation to a cone . . . 27
2.3.1.1 Composition of K-convex functions . . . 28
2.3.1.2 Dual cones and K-monotone functions . . . 29
2.3.1.3 Valid cuts for K-convex functions . . . 30

3 Optimization **31**
3.1 Optimization problems . . . 32
3.1.1 Lagrangian relaxation . . . 33
3.2 Optimal value functions . . . 35
3.2.1 Characterizations of optimal value functions . . . 36
3.2.2 Approximation by cuts . . . 40
3.2.2.1 Benders cuts . . . 44
3.2.2.2 Strengthened Benders cuts . . . 46
3.2.2.3 Lagrangian cuts . . . 49
3.2.2.4 Comparison between cut types . . . 50
3.3 Multi-stage optimization . . . 51
3.3.1 Two-stage problems . . . 53
3.3.2 Multi-stage problems . . . 54
3.4 Optimization under uncertainty . . . 55
3.4.1 Notation for random functions . . . 55

3.4.2 Risk-neutral optimization . . . 56
3.4.2.1 Decomposed formulation . . . 57
3.4.2.2 Linked formulation . . . 58
3.4.3 Risk-averse optimization . . . 61
3.4.3.1 Coherent risk measures . . . 61
3.4.3.2 Risk-averse stochastic programs . . . 63
3.4.3.3 Conditional value-at-risk . . . 66
3.4.4 Multi-stage stochastic programming . . . 67

4 Measures and Distributions 69
4.1 Measures . . . 70
4.1.1 The Hahn-Jordan decomposition . . . 73
4.1.1.1 Total variation norm . . . 75
4.1.2 Vector and matrix valued measures . . . 76
4.2 Distributions . . . 76
4.2.1 Distributions . . . 77
4.2.1.1 Jump formula . . . 79
4.2.2 Regularity results . . . 81
4.2.3 Convolution of distributions . . . 83
4.2.3.1 Convolutions and convexity . . . 85

5 Convexification by Averages 87
5.1 A pictorial discussion . . . 88
5.2 The gap as a measure of non-convexity . . . 93
5.2.1 Quantifying the non-convexity . . . 95
5.2.2 Additive noise . . . 98
5.2.2.1 Uniform norm and asymptotic behavior . . . 100
5.3 Second derivative's negative part . . . 103
5.3.1 Additive noise . . . 105
5.3.2 Optimal convexification . . . 106
5.4 Non-convexity measures . . . 112
5.4.1 Other examples of non-convexity measures . . . 115

6 Convexification via Risk Measures 119
6.1 Gap function and risk measures . . . 119
6.2 Additive noise . . . 121

7 Expected Cost-to-go 125
7.1 Stochastic dual dynamic programming . . . 126
7.2 Computational environment . . . 128
7.3 Unidimensional control problem . . . 129
7.3.1 Convex case . . . 129
7.3.2 Non-convex case . . . 132
7.4 Hydrothermal operational planning . . . 136
7.4.1 Computational simulations with 2 subsystems . . . 137

Bibliography **141**

List of Figures

2.1 Example of an affine set. 7
2.2 Examples of convex and non-convex sets. 8
2.3 Example of a non-convex set and its convex hull. 8
2.4 Union and intersection of convex sets. 9
2.5 Graph of a typical convex function. 12
2.6 The epigraph of a function is the set of points above its graph. . . 13
2.7 Example of a function f and its convex relaxation $\check{f}$. 17
2.8 A function is lower semi-continuous if and only if its epigraph is a closed set. 20
2.9 The conjugate f^* represents the vertical axis intercept of the largest affine function that is everywhere less than f. 23
2.10 Example of a cone. 26
2.11 Any valid cut for a non-decreasing ($\mathbb{R}_+$-monotone) convex function has non-negative slope. 29

3.1 A non-convex optimal value function f, its relaxation f_c and a Benders cut. 45
3.2 The strengthened Benders cut is parallel to the Benders cut but tight for $\check{f}$ at some point. 47
3.3 A Lagrangian cut is as tight as possible at the chosen point. . . . 49
3.4 The three types of cuts calculated for the same optimal value function f at a chosen point b. 50
3.5 For non-convex problems, the average of cuts that are tight for each scenario may not be tight for the expected cost-to-go. 59
3.6 Any valid cut for a convex non-decreasing function has non-negative slope. 63

4.1 An illustration of Lebesgue measure as a covering by rectangles. . 73

5.1 The function W and its decomposition as minimum of absolute values. 88
5.2 On the left, the function W and the expected function when the uncertainty is $\xi \sim U(-1, 1)$ and, on the right, the expected function when the uncertainty is $\xi \sim N(0, 1)$. 89
5.3 Example of convexification for different uniform distributions. . . 90
5.4 Example of convexification for different Normal distributions. . . . 91

5.5 Example of convexification for different discrete distributions, approximating a uniform on the interval $[-1, 1]$. 92
5.6 Example of gap function for some non-convex function. 93
5.7 Example of two non-convex functions were f is less non-convex than g. 94
5.8 Comparison between $\mathbb{E}[\check{Q}]$ and $\widetilde{\mathbb{E}[Q]}$ as underapproximations of $\mathbb{E}[Q]$. The function Q is the that from (5.2) with an uncertainty $\xi \sim U(-0.5, 0.5)$. 95
5.9 The function f is intuitively more non-convex than g but their gap functions are not comparable. 96
5.10 An illustration of $\|\mathrm{gap}(f)\|_1$ and $\|\mathrm{gap}(f)\|_\infty$. 97
5.11 On this image, we perceive both functions as being equally non-convex. Translation invariant norms capture this intuitive notion. 99
5.12 Graph of the minimum of two quadratic functions. 107
5.13 A discrete noise can totally convexify a polyhedral function. . . . 109
5.14 A discrete noise can only spread the corners of a minimum of convex functions, without totally convexifying it. 111

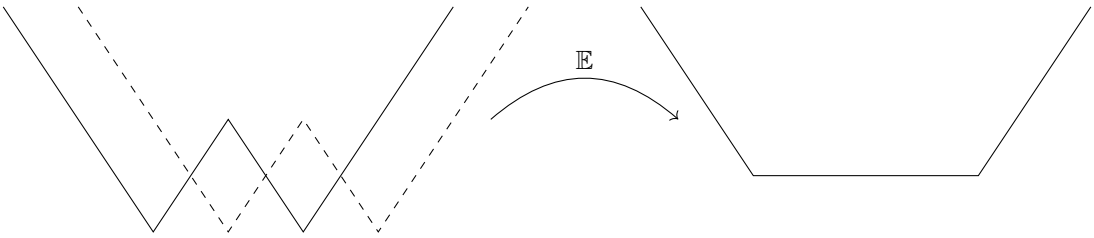
6.1 The left image is the graph of f superimposed over the graph of its convex relaxation $\check{f}$. The right image is the graph of the supremum $\rho(\tau_\xi f)$ superimposed over the graph of its convex relaxation $\widetilde{\rho(\tau_\xi f)}$. Notice the difference on their gaps. 124

7.1 Expected cost-to-go for convex problem. 130
7.2 Expected cost-to-go for a convex problem with two equally probable scenarios. 131
7.3 Expected cost-to-go for a convex problem with four scenarios. . . 132
7.4 Expected cost-to-go for non-convex problem. 133
7.5 Expected cost-to-go for a non-convex problem with two scenarios. 134
7.6 Expected cost-to-go for a non-convex problem with four scenarios. 135

Introduction 1

Convexity plays a major role in optimization. There are significant differences on the available methods to solve convex or non-convex problems. In fact, many methods for solving non-convex optimization problems, such as cutting planes algorithms, consist on iteratively solving finer and finer convex approximations to the original problem. This discrepancy becomes even larger in the context of stochastic programming, where decisions must take many possible scenarios into account. In this case, we minimize the average of the cost for each scenario, resulting in problems that may be much more complicated to solve.

Although stochastic programs are usually harder to solve than deterministic ones, we have observed an interesting phenomenon while dealing with non-convex problems: when we take the pointwise average of non-convex functions, the wrinkles on their graphs tend to cancel out and the resulting function can be much less non-convex. As an example consider the two W-shaped functions below, which are slightly offset.



The function on the right is the average between both functions on the left and is actually convex.

The main goal of of this work consists in rigorously defining what is meant by a function being less non-convex than another. A natural way to do this is to consider the gap $f - \check{f}$ between a function f and the largest convex function below it, called f's convex relaxation or $\check{f}$. In this context, the convexification can be summarized by saying that if Q is a random function, then the gap of its average is everywhere below the average of each realization's gap, or in other words:

$$\mathbb{E}[Q] - \widecheck{\mathbb{E}[Q]} \leqslant \mathbb{E}[Q] - \mathbb{E}[\check{Q}]. \tag{1.1}$$

An analogous result also holds in the risk-averse setting. In this case, instead of minimizing the expectation of multiple scenarios, the cost is given by a coherent risk measure ρ applied to the optimal value of all scenarios and the gap also decreases:

$$\rho(Q) - \widetilde{\rho(Q)} \leqslant \rho(Q) - \rho(\check{Q}). \tag{1.2}$$

Other reasonable way to measure the non-convexity of a function is to look at its second derivative. A continuous function is convex if and only if its weak second derivative is non-negative. Therefore, we can measure how non-convex f is by looking at $[f'']_-$, the negative part of its second derivative. Interestingly, a result similar to that for the gap holds,

$$\mathbb{E}\left[[Q'']_-\right] \leqslant [\mathbb{E}\left[Q\right]'']_-. \tag{1.3}$$

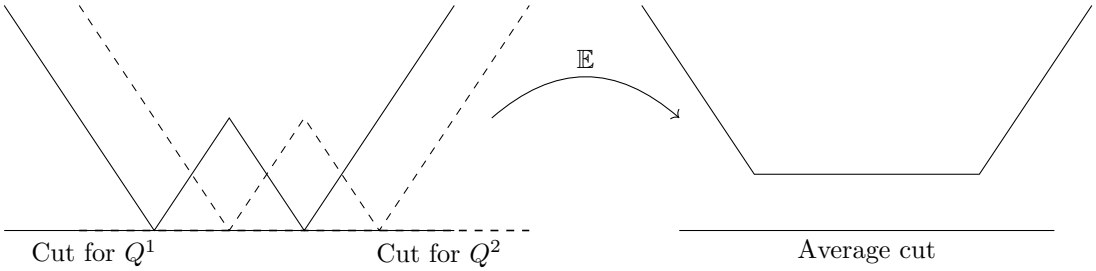
Moreover, a parallel can be made between the convexification results for the gap and those concerning the second derivative, showing that both non-convexity measures must be equivalent in some sense.

The relation between these two ways to measure the non-convexity is made clear if we notice that both Equation (1.1) and (1.3) are generalized instances of Jensen's inequality and saying that taking averages convexify a random function amounts to saying that these operators are convex. Thus, we can use the theory of functions that are convex in relation to a cone to properly define what is expected of a *non-convexity measure*. In order to generalize the previous examples and maintain their important properties, we will say that a non-convexity measure must satisfy three properties:

- Be zero if and only if the function is convex;
- Be non-negative with respect to a certain cone;
- Be convex with respect to this same cone.

Despite being much more abstract than the other ways we used to measure the non-convexity of a function, this definition has the advantage of encompassing the gap and second derivative as well as many other possible non-convexity measures that are only applicable for some restricted class of functions.

The knowledge of this non-convexity reduction can be used to approximate the expected cost-to-go of multi-stage non-convex programs through cuts that are *tight* for it. In the usual decomposition coming from the dynamic programming formulation, we calculate a cut for each possible realization of the next stage uncertainty separately and the average cut is guaranteed to be a valid underapproximation of the expected cost-to-go. This procedure has the disadvantage that even if the cuts are tight for the cost-to-go for each scenario, their average is at best tight for $\mathbb{E}\left[\check{Q}\right]$, as illustrated in the following figure:



A way to assure that the calculated cuts are tight consists in linking all scenarios into a single optimization problem that directly calculates the expected cost-to-go. Then, the usual techniques for calculating cuts for an optimal value function through its dual problem can be used to calculate cuts that are tight for the convex relaxation $\widetilde{\mathbb{E}[Q]}$. In the favorable case when the uncertainty actually turns $\mathbb{E}[Q]$ into a convex function, these cuts approximate the true expected cost-to-go as if the original problem was convex.

Fundamentals of Convex Analysis 2

This chapter is dedicated to establishing some important results concerning convex functions that will be widely used in the following chapters.

In Section 2.1, we present the theory of convex sets, which are special subsets of a real vector space with the property of being "closed by taking averages". An emphasis is put on the setwise operations that, when applied to convex sets, also result in a convex set. These sets play an important role in the methods of optimization, which are discussed in Chapter 3.

In Section 2.2, we introduce extended real functions and discuss the advantages of allowing the functions considered to take the values of $\pm\infty$. Then, we introduce convex functions as those functions whose graph is always below their secants. The operations that preserve convexity are studied in Section 2.2.1. Section 2.2.2 introduces one of the most thoroughly used concepts in this work, the convex relaxation of a function f. It is denoted $\check{f}$ and is defined as the largest convex function that is everywhere below f. This concept forms the base for approximations by cuts in Chapters 3 and 7 and will be fundamental in Chapters 5 and 6 when we discuss convexification. In Section 2.2.3, we study the convergence of convex functions and introduce *lower semi-continuous functions*. Later, in Section 2.2.4, we discuss some elements of duality theory and the conjugate of a function.

Section 2.3 is dedicated to a special type of convex set called a *convex cone*. Each cone induces an order compatible with the vector space structure, called a conic inequality, that can be used to extend the notions of convexity and monotonicity to functions whose image is not the real line, but an arbitrary vector space. These will allow us some flexibility in Chapter 3 when representing constraints of optimization problems and can be used to show that some familiar operators such as the Hahn-Jordan decomposition in Chapter 4 can be regarded as convex. Convexity in relation to a cone will a fundamental aspect of non-convexity measures in Section 5.4.

With the exception of Sections 2.2.3 and 2.2.4, dealing with convergence, we will mostly phrase this chapter's results in terms of an abstract vector space V instead of $\mathbb{R}^n$. We do this because in many occasions we will need to consider

convex operators whose domain is an infinite-dimensional function space.

This chapter is mainly intended as a reference to later chapters and its content is denser than the rest of this work. The most important notions that the reader must know from here are the definitions of *convex set* and *convex function* as being "well-behaved" in relation to taking averages and the fact that a convex, proper and lower semi-continuous function can always be represented as the supremum of the affine functions everywhere less than it.

2.1 Convex sets

One of the fundamental objects in linear algebra are the subspaces of a vector space. These are characterized by the fact that they are closed by linear combinations. That is, for any two elements x, y of a subspace W of V,

$$\alpha x + \beta y \in W, \ \forall \alpha, \beta \in \mathbb{R}. \tag{2.1}$$

Working with subspaces is, in general, too restrictive because Equation (2.1) requires that they pass through the origin. In what follows, we will consider translations of subspaces, called *affine sets*.

Definition 2.1 (Affine set)**.** A subset A of a vector space V is *affine* if there is a subspace W of V and a point $b \in V$ such that

$$A = b + W := \{b + w \mid w \in W\}.$$

Remark 2.1. While a subspace W can be represented as the solution set of a homogeneous linear system $Tx = 0$, the affine sets act as their inhomogeneous counterparts. That is, a set A is affine if and only if it is the solution set of a linear system $Tx = b$, for some fixed $b \in V$.

Affine sets can also be characterized by a geometric property similar to Equation (2.1). Given two points x, y in a vector space, we can form the line passing through them by

$$\{\alpha x + \beta y \mid \alpha + \beta = 1\} \tag{2.2}$$

and a set A is affine if and only if it contains every line passing through its points. This is the same as the definition of a linear subspace with the additional constraint that the coefficients must sum to 1. Noticing that we could write $\beta = 1 - \alpha$ in Equation (2.2), we obtain another, more typical, characterization of an affine set.

Theorem 2.2. *A subset A of a vector space is affine if and only if it contains the line passing through each pair of points in it. In other words, given x, $y \in A$, the point $\lambda x + (1 - \lambda)y$ is also in A for every $\lambda \in \mathbb{R}$.*

The collection of affine sets is too restrictive for many applications. For that reason, we will mostly work with *convex sets*, which are the sets that contain the

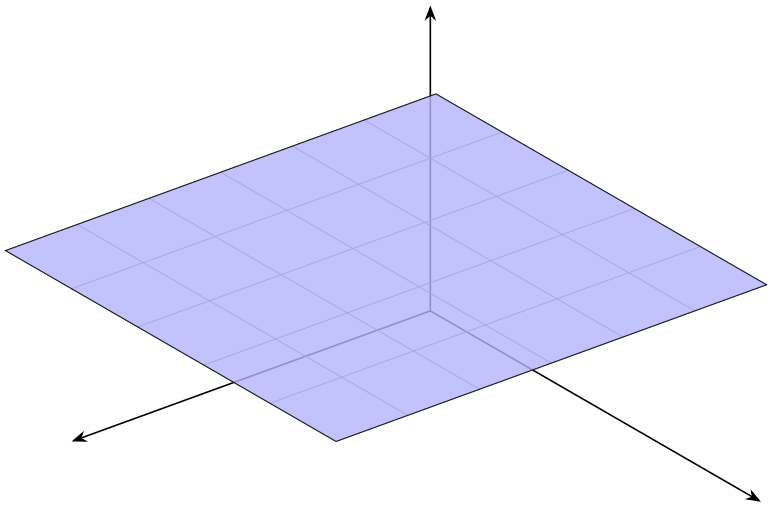


Figure 2.1: Example of an affine set.

line segment between each pair of points in it. A convex and a non-convex set are illustrated in Figure 2.2.

Definition 2.3 (Convex set)**.** A subset C of a vector space V is *convex* if the line segment between any two points x, $y \in C$ is entirely contained in C. That is, given x and $y \in C$, the point $\lambda x + (1 - \lambda)y$ is also in C for every $\lambda \in [0, 1]$.

Any point of the form $\lambda x+(1-\lambda)y$ with $\lambda \in [0, 1]$ is called a *convex combination* of x and y. In general, we can consider a convex combination with any finite amount of terms as long as the coefficients are non-negative and sum to one. Using this, we can define the *convex hull* of an arbitrary set X to be the set of all convex combinations of its elements. As expected, a set equals its convex hull if and only if it is convex.

Definition 2.4 (Convex hull)**.** The *convex hull* of a subset X of a vector space is the set of all finite convex combinations of its elements,

$$\operatorname{conv}(X) = \left\{ \sum_{i=1}^{k} \lambda_i x_i \mid k \in \mathbb{N},\ \lambda_i \geqslant 0,\ \sum_{i=1}^{k} \lambda_i = 1 \right\}$$

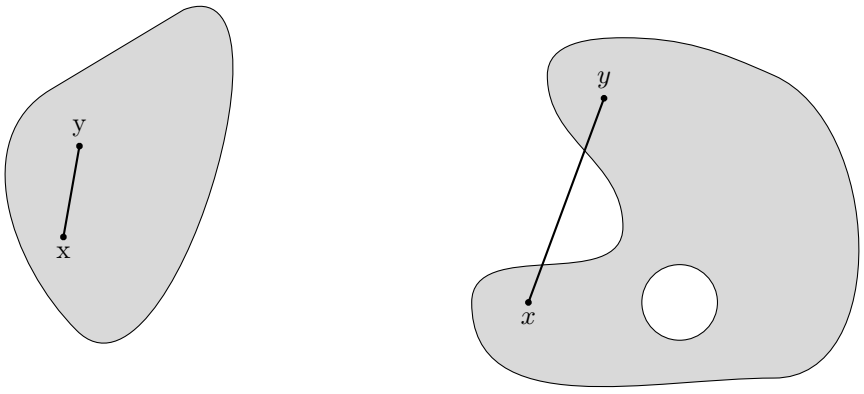
2.1.1 Convexity-preserving operations

In this section, we summarize some operations that preserve the convexity of sets. The first of these is the intersection of a family of convex sets.

Theorem 2.5 (Intersection of convex sets)**.** *If C_i is a family of convex sets indexed by $i \in I$, then their intersection*

$$C = \bigcap_{i \in I} C_i$$

is also a convex set.



(a) Convex set. (b) Non-convex set.

Figure 2.2: Examples of convex and non-convex sets.

Proof. If x, $y \in C$, then they are in C_i for every $i \in I$. Since each C_i is convex, the line segment between x and y is also contained in each C_i. Therefore, it is contained in C. This implies that C is convex. □

Remark 2.2. Although the intersection of convex sets is also convex, the same cannot always be said of the union of convex sets. An example where A and B are convex sets but $A \cup B$ is not can be seen in Figure 2.4.

Using Theorem 2.5, we can properly talk about the smallest convex set containing a set. This allows us to give an alternative representation of the convex hull of X as the smallest set containing it,

$$\operatorname{conv} X = \bigcap \{C \mid C \text{ is convex, } X \subset C\}. \tag{2.3}$$

This definition is equivalent to the one given before in 2.4 and a proof can be found at [Lucchetti, 2005].

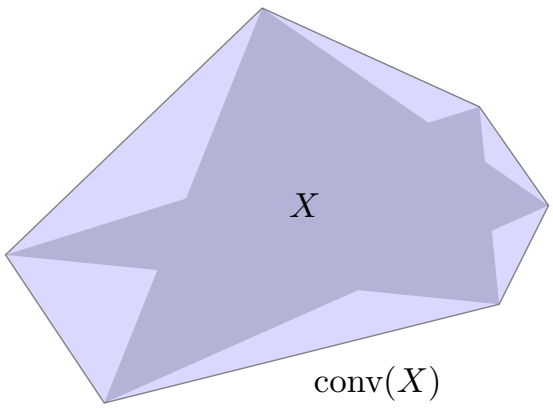


Figure 2.3: Example of a non-convex set and its convex hull.

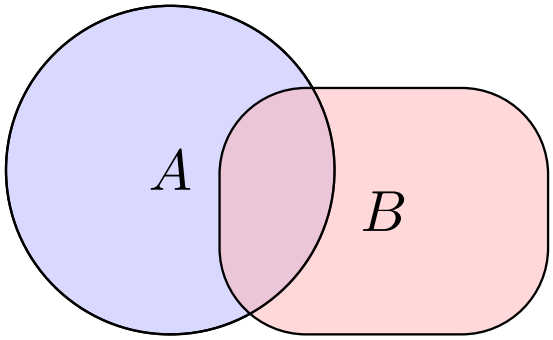


Figure 2.4: Union and intersection of convex sets.

The cartesian product of a family of vector spaces is also a vector space via componentwise addition and scalar multiplication. If we take a convex set in each of these vector spaces, their cartesian product will also be a convex set on the product space.

Theorem 2.6 (Cartesian product of convex sets)**.** *If C_i is a family of convex sets indexed by $i \in I$, then their Cartesian product*

$$C = \prod_{i \in I} C_i$$

is also a convex set.

Proof. As each component of C is convex and convex combinations are taken componentwisely, C is convex. □

Other operations which also preserve convexity are setwise addition and scalar multiplication, defined shortly.

Definition 2.7 (Minkowski sum of sets)**.** The *Minkowski sum* of two subsets A and B of a vector space is defined by

$$A + B = \{a + b \mid a \in A, b \in B\}.$$

Definition 2.8 (Multiplication by scalar for sets)**.** The multiplication of a set X by a scalar λ is the set

$$\lambda X = \{\lambda x \mid x \in X\}.$$

Remark 2.3. Although the notation of the definitions above is made to resemble the notation for addition and scalar multiplication of vectors, these setwise operations do not possess the same properties.

One of the main properties of the Minkowski sum is that it commutes with convex hulls, that is,

$$\operatorname{conv}(A + B) = \operatorname{conv}(A) + \operatorname{conv}(B). \tag{2.4}$$

If we take A and B to be convex on the equation above, we have that they are equal to their convex hulls, which implies

$$A + B = \text{conv}(A + B).$$

Therefore the sum of convex sets is also convex.

In a similar manner, the scalar multiplication of a set satisfies

$$\text{conv}(\lambda X) = \lambda \,\text{conv}(X),$$

which implies that λX is convex whenever X is.

Now we consider the image of sets by functions that preserve convexity. Taking a linear transformation $T\colon V \to W$ and a point $b \in W$, we can define the function

$$f(x) = Tx + b,$$

called an *affine function* from T to W. As the following theorems show, both linear and affine transformations preserve the convexity of sets.

Theorem 2.9 (Image of a convex set by a linear function)**.** *If $T\colon V \to W$ is a linear transformation then its image and pre-image preserve convexity. That is, for any convex $A \subset V$,*

$$T(A) = \{T(x) \mid x \in A\}$$

is convex and for any $B \subset W$,

$$T^{-1}(B) = \{x \in V \mid T(x) \in B\}$$

is also convex.

Proof. For the image, recall that any element of $T(A)$ can be written as Tx, with $x \in A$. Thus, a convex combination of Tx and Ty satisfies

$$\lambda Tx + (1 - \lambda)Ty = T(\lambda x + (1 - \lambda)y) \in T(A),$$

since $\lambda x + (1 - \lambda)y \in A$.

For the pre-image, take two elements x, $y \in V$ such that $T(x)$, $T(y) \in B$. Then any convex combination $z := \lambda x + (1 - \lambda)y$ satisfies

$$T(\lambda x + (1 - \lambda)y) = \lambda T(x) + (1 - \lambda)T(y) \in B,$$

since B is convex. □

Corollary 2.9.1 (Image of a convex set by an affine function). If $f\colon V \to W$ is an affine function then its image and pre-image preserve convexity.

Proof. The set $f(A)$ equals $T(A) + \{b\}$, which is a Minkowski sum of convex sets, hence convex.

Similarly,

$$f^{-1}(B) = \{x \in V \mid Tx + b \in B\} = \{x \in V \mid Tx \in B - \{b\}\},$$

which is convex. □

2.2 Convex functions

Throughout this work, it is convenient to allow the functions to take the values $\pm\infty$.

Definition 2.10 (Extended real function)**.** An *extended real function* is a function $f\colon V \to [-\infty, +\infty]$.

This work focuses on minimization, hence, it will be natural to allow a certain asymmetry between positive and negative infinite values. In what follows, we will mostly work with *proper* functions that are finite for at least some point and nowhere equal minus infinity.

For maximization problems, the most useful definition requires that we exchange the role played by $+\infty$ and $-\infty$.

Definition 2.11 (Proper function)**.** An extended real function $f\colon V \to [-\infty, +\infty]$ is *proper* if $f(x) > -\infty$ for all $x \in V$ and there is at least one point $y \in V$ such that $f(y) < \infty$.

Definition 2.12 (Domain of a extended real function)**.** The *domain* of $f\colon V \to [-\infty, +\infty]$ is the set

$$\operatorname{dom}(f) = \{x \in V \mid |f(x)| < +\infty\}.$$

If C is a non-empty subset of V, every function $f\colon C \to \mathbb{R}$ has a unique extension to a proper extended real function $\tilde{f}$ defined over V such that $\operatorname{dom}(\tilde{f}) = C$, given by

$$\tilde{f}(x) = \begin{cases} f(x), & x \in C \\ +\infty, & x \notin C. \end{cases} \tag{2.5}$$

In view of this, no confusion shall arise if we denote the extension of f by the same symbol 'f'. By considering only extended real functions defined on V, the presentation will be cleaner, since we do not need to worry about domain restrictions.

It will also be useful, hereafter, to consider functions which are everywhere less than some other. This is a partial order on the set of extended real functions, and will also be denoted by "$\leqslant$".

Definition 2.13 (Functional inequalities)**.** Given functions $f\colon X \to [-\infty, +\infty]$ and $g\colon X \to [-\infty, +\infty]$, we say that f *is less or equal than* g, denoted by $f \leqslant g$, if $f(x) \leqslant g(x)$, $\forall x \in X$.

We now proceed to introduce the main theme of this work: *convex functions*. These can be intuitively thought as the functions whose graphs are always below any of their chords, as exemplified in Figure 2.5

Definition 2.14 (Convex function)**.** A function $f\colon V \to [-\infty, +\infty]$ is said to be *convex* if its domain is a convex set and for any pair of points x, y on its domain

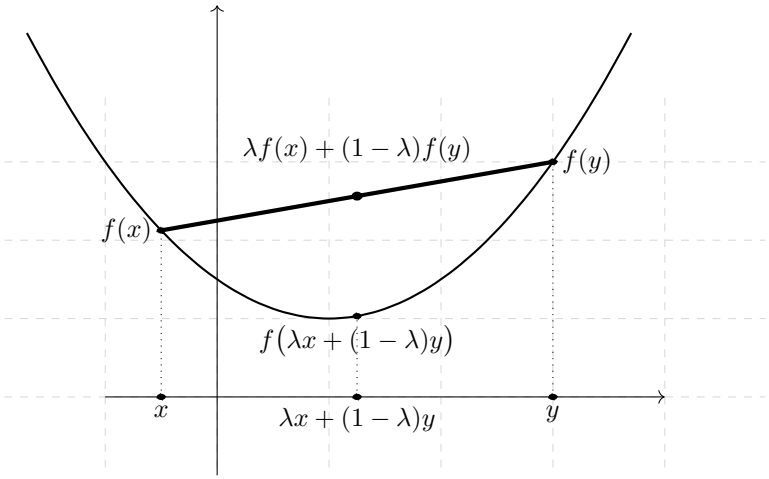


Figure 2.5: Graph of a typical convex function.

and any parameter $\lambda \in [0, 1]$,

$$f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda)f(y).$$

If $-f$ is a convex function, we say that f is a *concave function*. Since $\max f = -\min -f$, most of what we will develop for the minima of convex functions has an equivalent formulation for the maxima of concave functions.

The definition of a convex function only depends on how it behaves when restricted to each line intersecting its domain, so we obtain the following characterization of convex functions:

Theorem 2.15. *A function $f\colon V \to [-\infty, +\infty]$ is convex if and only if for each $x \in \operatorname{dom}(f)$ and all $v \in V$, the unidimensional function $g\colon \mathbb{R} \to [-\infty, +\infty]$, given by*

$$g(t) = f(x + tv),$$

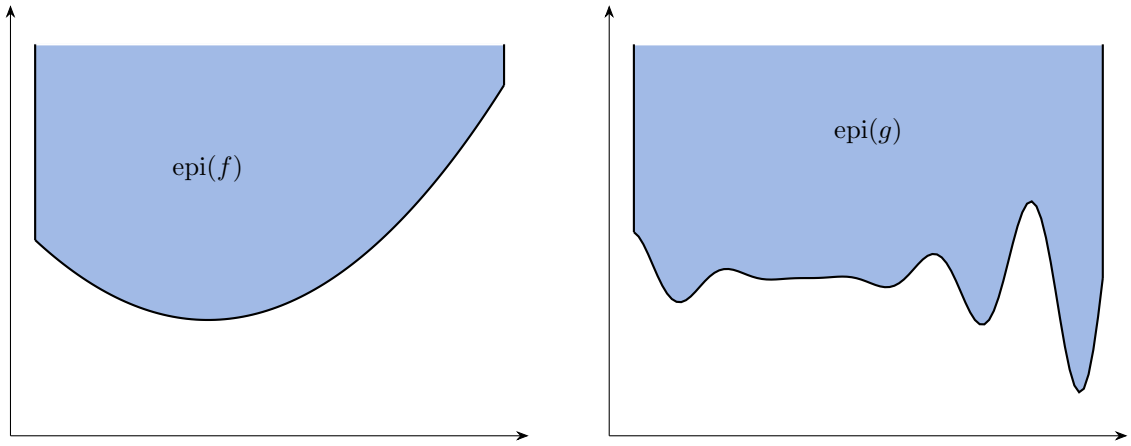
is convex.

The defining property of a convex function is called *Jensen's inequality*. We can think about the fact that $f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda)f(y)$ as saying that if z is a weighted average with parameter λ between x and y, that is, z lies on the line between x and y, a convex function evaluated at z will always be less than this same weighted average applied to $f(x)$ and $f(y)$. Geometrically, this means that any secant to a convex function lies above its graph, as illustrated on Figure 2.5. A remarkable fact is that this property can be extended for any probability distribution defined on V.

Theorem 2.16 (Jensen's inequality)**.** *A function f is convex if and only if it satisfies*

$$f(\mathbb{E}[X]) \leqslant \mathbb{E}[f(X)]$$

for any random variable X whose support is contained on $\operatorname{dom}(f)$.



(a) Convex function. (b) Non-convex function.

Figure 2.6: The epigraph of a function is the set of points above its graph.

Affine and linear functions can be characterized by their graphs being respectively affine or linear subspaces of $V \times \mathbb{R}$. The graph of a convex function is generally not convex but a similar characterization can be made in terms of the set of points above its graph, called its *epigraph*.

Definition 2.17 (Epigraph)**.** The *epigraph* of a function $f \colon V \to [-\infty, +\infty]$ is the subset of $V \times \mathbb{R}$ defined by

$$\operatorname{epi}(f) = \{(x, t) \in V \times \mathbb{R} \mid f(x) \leqslant t\}.$$

Theorem 2.18 (Epigraph of convex function)**.** *A function $f \colon V \to [-\infty, +\infty]$ is convex if and only if its epigraph is a convex set.*

For a function f we can define its α-sublevel set as the set where the value of f is below α:

$$S_{f,\alpha} = \{x \in V \mid f(x) \leqslant \alpha\}.$$

Any sublevel set of a convex function f is convex, since it is the projection on V of intersection between $\operatorname{epi}(f)$ and the set $\{(x, t) \in V \times \mathbb{R} \mid t = a\}$.

The converse is not true. There are non-convex functions whose all sublevel sets are convex. An example is given by any non-convex increasing function from $\mathbb{R}$ to $\mathbb{R}$ such as $\tanh(x)$, for which all sublevels are of the form $(-\infty, f^{-1}(\alpha)]$.

Remark 2.4. Notice that a function being less than another also has an interpretation in terms of their epigraphs:

$$f \leqslant g \iff \operatorname{epi}(g) \subset \operatorname{epi}(f).$$

Remark 2.5. A proper function also has a characterization in terms of its epigraph. In this case, saying that f is proper is equivalent to saying that its epigraph is non-empty nor contains any vertical line.

Local minima of convex functions In the case of vector spaces with some norm $\|\cdot\|$, we consider the concept of a *local minima and maxima* of a function.

Definition 2.19 (Local minimum)**.** If $(V, \|\cdot\|)$ is a normed vector space, a point $y \in V$ is called a *local minimum* of a function f from V to $\mathbb{R}$ if there is some $\epsilon > 0$ such that $f(y) \leqslant f(y + v)$ for every v with $\|v\| < \epsilon$.

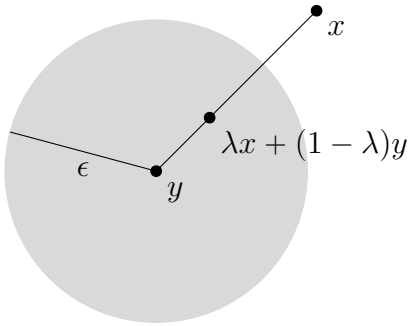
A local maximum is analogously defined as the point for which $f(y) \geqslant f(y+v)$ for every v in some ball around the origin.

Convex functions have an important property which says that if y is a local minimum, then it is in fact a global minimum.

Theorem 2.20 (Local minima of convex functions are global)**.** *If f is convex and y is a local minimum, then*

$$f(y) \leqslant f(x), \forall x.$$

Proof. Fix some point x. If $\|x - y\| < \epsilon$, the result follows from the definition of local minimum. If $\|x - y\| \geqslant \epsilon$, we can find a convex combination of x and y which lies inside the ball of radius ϵ around y by choosing $\lambda = \epsilon/(2\|x - y\|) \leqslant 1$.



Rewriting $\lambda x + (1 - \lambda)y$ as $y + \lambda(x - y)$, we can apply f to get

$$f(y + \lambda(x - y)) = f(\lambda x + (1 - \lambda)y) \leqslant \lambda f(x) + (1 - \lambda)f(y).$$

Since $\|\lambda(x - y)\| < \epsilon$, the fact that y is a local minimum implies that

$$f(y) \leqslant f(y + \lambda(x - y)) \leqslant \lambda f(x) + (1 - \lambda)f(y).$$

Which implies $f(y) \leqslant f(x)$. Since x was arbitrary, the proof is complete. □

2.2.1 Convexity-preserving operations

Here we discuss some operations on families of convex functions that result in a function that is also convex. Some of these are direct consequences of the results on Section 2.1.1 about convexity-preserving operations on *convex sets* when applied to the epigraphs of the functions while others, such as the rules for addition and composition, only have an interpretation for functions.

Theorem 2.21 (Sum of convex functions)**.** *If f and g are convex functions, so is $\alpha f + \beta g$ for any α, $\beta \geqslant 0$.*

Corollary 2.21.1. For any finite collection of convex functions f_i and non-negative scalars λ_i, the function $\sum_{i=1}^{m} \lambda_i f_i$ is convex.

Remark 2.6. The above theorem implies that the set of convex functions is itself a convex set. As we will see in Section 2.3, it is in fact a special type of convex set called a convex cone.

For the composition, we have in fact two different results. Theorem 2.22 says that first passing through an affine function and then through a convex function is a convex operation and Theorem 2.23 says that the composition of two convex functions is convex as long as the second one is non-decreasing.

Theorem 2.22 (Composition of convex and affine function)**.** *If f is a convex function, then $g(x) = f(Ax + b)$ is also convex for any linear operator A and vector b.*

Theorem 2.23 (Composition of convex functions)**.** *If $f\colon V \to (-\infty, +\infty]$ is convex and $g\colon \mathbb{R} \to (-\infty, +\infty]$ is convex and non-decreasing, with the convention that $g(+\infty) = +\infty$, then the composition $g \circ f\colon V \to (-\infty, +\infty]$ is convex.*

Proof. From the convexity of f, for any points x, y and $\lambda \in [0, 1]$,

$$f(\lambda x + (1 - \lambda)y) \leqslant \lambda f(x) + (1 - \lambda) f(y).$$

Since g is non-decreasing, it preserves the inequality above

$$g(f(\lambda x + (1 - \lambda)y)) \leqslant g(\lambda f(x) + (1 - \lambda) f(y)).$$

and the convexity of g implies that

$$g(\lambda f(x) + (1 - \lambda) f(y)) \leqslant \lambda g(f(x)) + (1 - \lambda) g(f(y)).$$

Taking it all together, we see that $g \circ f$ is convex. □

Now we see how a convex functions behaves with respect to partial maximization or minimization. As we will see, the maximum of any family of convex functions is always convex while the minimum is only convex when certain special conditions are met.

Theorem 2.24 (Supremum of convex functions)**.** *If f_α is an arbitrary collection of convex functions indexed by $\alpha \in A$, their pointwise supremum $g(x) = \sup_{\alpha \in A} f_\alpha(x)$ is also convex.*

Proof. The epigraph of g is the set

$$\begin{aligned}\operatorname{epi}(g) &= \{(x,t) \mid \sup_{\alpha\in A} f_\alpha(x) \leqslant t\} \\ &= \{(x,t) \mid f_\alpha(x) \leqslant t,\ \forall \alpha \in A\} \\ &= \bigcap_{\alpha\in A} \{(x,t) \mid f_\alpha(x) \leqslant t\} \\ &= \bigcap_{\alpha\in A} \operatorname{epi}(f_\alpha),\end{aligned}$$

which is the intersection of convex sets and therefore also convex. □

The supremum of convex functions is convex because the intersection of convex epigraphs is also a convex set. On the other side, the epigraph of $h(x) = \inf_\alpha f_\alpha(x)$ is the *union* of convex sets and, therefore, not convex in general. A special case is when we do the partial minimization of a convex function over some convex set. In this case, as we will see in Theorem 2.25, the infimum of convex functions can be in fact convex.

Theorem 2.25 (Infimum of convex functions)**.** *Suppose $f\colon X \times Y \to [-\infty, +\infty]$ is a proper convex function and C is a non-empty* convex *set. Then the function*

$$h(x) = \inf_{y\in C} f(x,y)$$

is convex provided that $h(x) > -\infty$ for all $x \in X$.

A proof to this fact can be found at [Boyd and Vandenberghe, 2004].

2.2.2 Convex relaxation

As seen on Section 2.1, any set X has a convex hull $\operatorname{conv}(X)$ defined as the smallest convex set containing X. If we apply this operation to the epigraph of a function f, we obtain a convex set which is the epigraph of the largest convex function below f, called its *convex relaxation*.

Definition 2.26 (Convex relaxation of a function)**.** The *convex relaxation* of a function f, denoted by $\operatorname{conv}(f)$ or $\check{f}$, is the largest convex function which is always below f.

We have that

$$\operatorname{epi}(\check{f}) = \operatorname{conv} \operatorname{epi}(f)$$

and, since the set on the right can be written as the intersection of all convex sets containing $\operatorname{epi}(f)$, it can be seen as the intersection of the epigraphs of all convex functions that are underneath f. This allows us to write $\check{f}(x)$ as the maximum value a convex function everywhere less than f can attain at the point x. That is,

$$\begin{aligned}\check{f}(x) = \sup_g \quad & g(x) \\ \text{s.t.} \quad & g(y) \leqslant f(y),\ \forall y \in V, \\ & g \text{ convex.}\end{aligned} \tag{2.6}$$

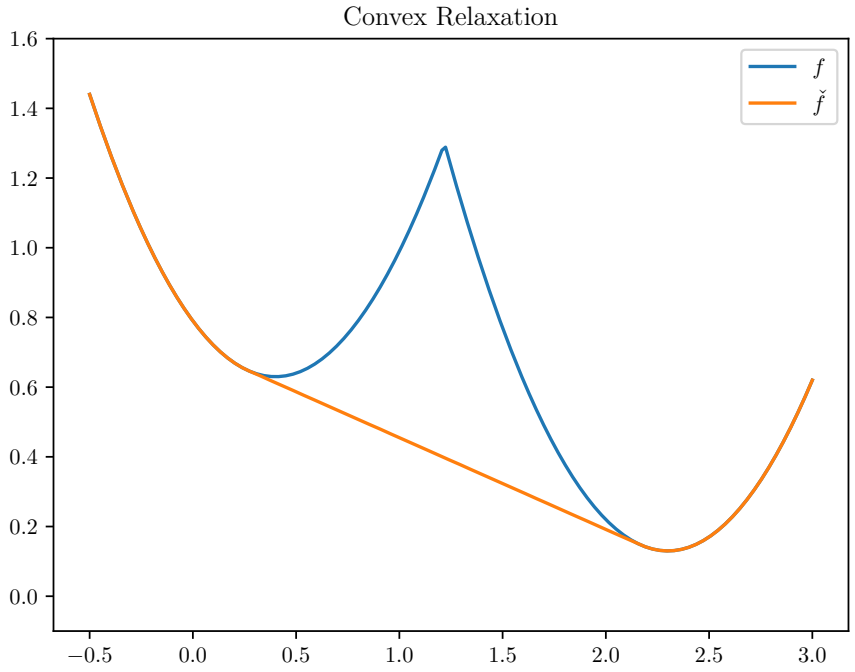


Figure 2.7: Example of a function f and its convex relaxation $\check{f}$.

This representation can be further refined to consider only *affine functions* below f. For this, we need an equivalent version of the Hahn-Banach theorem, which can be found as proposition (HB4) in Section 12.31 of [Schechter, 1997].

Theorem 2.27 (Hahn-Banach Convex Support Theorem)**.** *Any finite valued convex function $f\colon V \to \mathbb{R}$ is the pointwise maximum of the affine functions below it. That is, for each $x_0 \in V$ there exists an affine functional ϕ such that $\phi(x) \leqslant f(x)$ for all $x \in V$ and $\phi(x_0) = f(x_0)$.*

For any function f, an affine function which is everywhere below f is called a *feasible cut* to f and a feasible cut that equals f on at least a point is called a *subgradient* to f. Theorem 2.27 says that a convex function has a subgradient at any point. Although this result requires f to be convex, we will see in Theorem 2.39 that the same holds for a extended real function provided that it is *lower semi-continuous*.

Representing a convex function via its feasible cuts is an important method in convex optimization as we will discuss in Chapter 3. Now, we see that a non-convex function cannot be represented only by feasible cuts. The best we can do is represent its convex relaxation via cuts.

Since affine functionals are convex and $\check{f}$ is the largest convex function below f, for any affine function ϕ,

$$\phi \leqslant f \iff \phi \leqslant \check{f}.$$

Using this, we can represent the convex relaxation of a function f as the pointwise

maximum of affine functions which are everywhere below f.

$$\begin{aligned} \check{f}(x) = \sup_{a,b} \quad & \langle a, x\rangle + b \\ \text{s.t.} \quad & \langle a, y\rangle + b \leqslant f(y),\ \forall y \in V, \\ & a \in V^\star,\ b \in \mathbb{R}. \end{aligned} \tag{2.7}$$

2.2.3 Modes of convergence

In this section we discuss modes of convergence for some classes of functions, with an special emphasis on how convexity and infima behave in relation to these. We start by the concept of pointwise convergence, that is, $f_n \to f$ if for each point x, the evaluations $f_n(x)$ converge to $f(x)$.

Definition 2.28 (Pointwise convergence)**.** A sequence f_n of functions is said to *converge pointwisely* to a function f on a set C if for each fixed $x \in C$,

$$\lim_{n\to\infty} f_n(x) = f(x).$$

The set of convex functions is closed with relation to pointwise convergence, that is, pointwise limits of sequences of convex functions are also convex.

Theorem 2.29. *Let f_n be a sequence of convex functions which converges pointwisely to f. Then f is also convex.*

Proof. Fix points x and y and some $\lambda \in [0, 1]$. From the convexity of f_n,

$$f_n(\lambda x + (1-\lambda)y) \leqslant \lambda f_n(x) + (1-\lambda) f_n(y).$$

Since limits preserve inequalities and $\lim_{n\to\infty} f_n(a) = f(a)$ for each point,

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &= \lim_{n\to\infty} f_n(\lambda x + (1-\lambda)y) \\ &\leqslant \lim_{n\to\infty} \lambda f_n(x) + (1-\lambda) f_n(y) \\ &= \lambda \lim_{n\to\infty} f_n(x) + (1-\lambda) \lim_{n\to\infty} f_n(y) \\ &= \lambda f(x) + (1-\lambda) f(y). \end{aligned}$$

Therefore f is convex. □

Although it preserves convexity, pointwise convergence is not well suited to work with optimization problems because, if $f_n \to f$ pointwisely, we cannot guarantee that $\inf f_n \to \inf f$.

A notion of convergence that works well with optimal values is that of *uniform convergence*.

Definition 2.30 (Uniform Convergence)**.** A sequence of functions f_n is said to *converge uniformly* to a function f on a set C if

$$\lim_{n\to\infty} \sup_{x\in C} |f_n(x) - f(x)| = 0.$$

Uniform convergence has the property that any space of continuous functions is closed in relation to it. Additionally, it is stronger than pointwise convergence it the sense that if a sequence f_n converges uniformly to f, it also converges pointwisely. If the functions f_n are all convex and defined on an open subset of $\mathbb{R}^n$, an almost converse result holds; if a sequence of convex functions converges pointwisely on an open convex subset Ω of $\mathbb{R}^n$, then it converges uniformly when restricted to any compact subset of Ω.

Theorem 2.31 (Pointwise convergence of convex functions is uniform on compact sets)**.** *Let Ω be a open convex subset of $\mathbb{R}^n$ and let f_n be a sequence of convex functions converging pointwisely to f on Ω. Then, for every compact subset K of Ω, f_n converges uniformly to f on K.*

A proof to this theorem can be found at Theorem 10.8 of [Rockafellar, 1996].

An important consequence of uniform convergence is that the supremum and the infimum of a uniformly convergent sequence converge to the supremum and infimum of the limit.

Theorem 2.32 (Optima of uniformly convergent sequence)**.** *Suppose f_n converges uniformly to f on a set K. Then*

$$\inf_{x \in K} f_n(x) \to \inf x \in K f(x) \tag{2.8}$$

$$\sup_{x \in K} f_n(x) \to \sup_{x \in K} f(x). \tag{2.9}$$

Albeit these nice properties, uniform converge is too restrictive when working with possibly infinite-valued functions. In this case, if f equals ∞ at a single point, no sequence of finite valued functions can converge uniformly to f. A similar problem happens with continuous functions because a function f cannot be continuous outside of its domain.

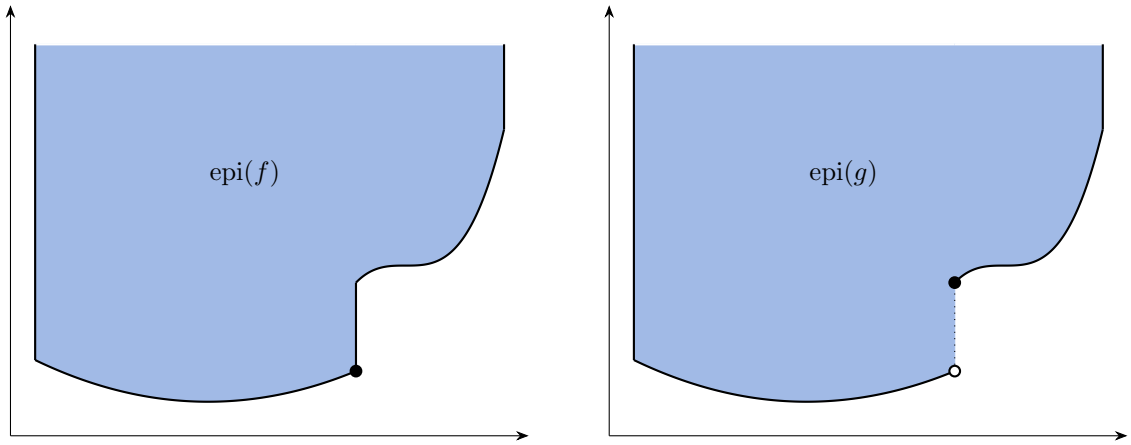
On the following, we introduce weaker notions of continuity and convergence that are more appropriate to work with extended real functions. First, we will discuss lower semi-continuous functions, which are precisely those functions whose epigraph is closed. Then, we will discuss the notion of epi-convergence, that can be seem as a weaker form of uniform convergence that also preserves infima.

2.2.3.1 Lower semi-continuity

Convex functions on $\mathbb{R}^n$ have the property that they are continuous on the interior of their domains, a result whose proof can be found at Corollary 2.1.3 of [Lucchetti, 2005].

Theorem 2.33 (Continuity of convex functions)**.** *Any proper convex function f is continuous on the interior of its domain.*

This theorem guarantees that a finite valued convex function is everywhere continuous. When working with extended value functions, we also must consider



(a) Lower semi-continuous function. (b) Non lower semi-continuous function.

Figure 2.8: A function is lower semi-continuous if and only if its epigraph is a closed set.

what happens at the boundary points of the function's domain. Asking for continuity at these points is too strong for if f is infinite at any point, it cannot be continuous. A solution arises by separating continuity in two different properties: lower and upper semi-continuity. Here we will focus only on the lower, since, as we will shortly see, if f is lower semi-continuous, $-f$ is upper semicontinuous.

Definition 2.34 (Lower semi-continuity)**.** A function f is *lower semi-continuous* at a point x if

$$f(x) \leqslant \liminf_{k\to\infty} f(x_k)$$

for every sequence x_k which converges to x. We say that f is *lower semi-continuous* if it is lower semi-continuous for every point in its domain.

Analogously, we say that a function is *upper semi-continuous* at x if

$$f(x) \geqslant \limsup_{k\to\infty} f(x_k)$$

for every sequence x_k which converges to x.

Since a continuous function takes sequences to sequences, it is always lower and upper semi-continuous. A converse result also holds.

Theorem 2.35. *A function f is continuous if and only if it is both lower and upper semi-continuous.*

Similarly to properness and convexity, lower semi-continuity also has an interpretation in terms of the epigraph of f.

Theorem 2.36 (Epigraph of lower semi-continuous function)**.** *A function $f\colon X \to [-\infty, +\infty]$ is lower semi-continuous if and only if its epigraph is a closed set in $X \times \mathbb{R}$.*

Notice that since the intersection of closed sets is also closed, we have an analogous to Theorem 2.24 for the supremum of lower semi-continuous functions.

Theorem 2.37. *If f_α is an arbitrary collection of lower semi-continuous functions indexed by $\alpha \in A$, their pointwise supremum $g(x) = \sup_{\alpha \in A} f_\alpha(x)$ is also lower semi-continuous.*

The closure of a subset A of $\mathbb{R}^n$ is the smallest closed set containing it, defined by

$$\operatorname{cl}(A) = \bigcap \{C \mid C \text{ is closed}, A \subset C\}. \tag{2.10}$$

This definition is similar to the characterization of the convex hull of a set given in Equation (2.3). In the same manner of Definition 2.26, we can apply it to the convex hull of a function to obtain the largest lower semi-continuous function below it.

Definition 2.38 (Lower semi-continuous regularization)**.** Given a function f, its *lower semi-continuous regularization*, denoted $\operatorname{cl}(f)$, is the function defined by

$$\operatorname{cl}(f)(x) = \inf \{a \in \mathbb{R} \mid (x, a) \in \operatorname{epi}(f)\}.$$

It is possible to check that $\operatorname{cl}(f)$ is in fact lower semi-continuous and its epigraph is given by

$$\operatorname{epi} \operatorname{cl}(f) = \operatorname{cl} \operatorname{epi}(f).$$

We are mostly interested on functions that are proper, lower-semicontinuous and convex. That is, functions whose epigraphs are non-empty, closed, convex and do not contain vertical lines. For these, some results regarding continuity and approximation by cuts exist, including a sharper version of Equation (2.7) which does not requires f to be finite valued.

In Theorem 2.27, we saw that any finite valued convex function can be represented as the maximum of its feasible cuts. The same result also holds for extended valued functions that besides being convex are also proper and lower semi-continuous.

Theorem 2.39 (Representation by cuts)**.** *If f is a convex, proper and lower semi-continuous function, then it can be represented as the supremum of the continuous affine functionals that are everywhere less than f. That is,*

$$\begin{aligned} f(x) = \sup_{a,b} \quad & \langle a, x \rangle + b \\ s.t. \quad & \langle a, y \rangle + b \leqslant f(y),\ \forall y \in V, \\ & a \in \mathbb{R}^n,\ b \in \mathbb{R}. \end{aligned}$$

A proof to this theorem is out of scope and can be found at 2.2.8 and Theorem 2.2.21 of [Lucchetti, 2005].

2.2.3.2 Epi-convergence

Now we introduce the notion of epi-convergence of functions. This can be intuitively thought as the epigraph of f_n converging to the epigraph of f. That last statement can, in fact, be made precise using convergence of sets but pursuing that would go out of the scope of this work. More about epi-convergence and convergences of sets can be found at Chapter 7 of [Rockafellar and Wets, 2011] for functions over $\mathbb{R}^n$ or at [Lucchetti, 2005], [Artstein and Wets, 1995] and [Attouch, 1984] for infinite-dimensional or topological spaces.

Definition 2.40 (Epi-convergence)**.** A sequence f_n of functions *epi-converges* to f if for all $x \in \mathbb{R}^n$,

1. For each sequence $x_n \to x$, $\liminf f_n(x_n) \geqslant f(x)$;
2. There is at least one sequence $y_n \to x$ such that $\lim f_n(y_n) = f(x)$.

An important consequence of the definition above is that if f_n epi-converges to f, then f must be lower semi-continuous. Furthermore, we can guarantee a result similar to 2.28 saying that a sequence of convex functions can only converge to a convex function.

Theorem 2.41 (Epi-limit of convex functions)**.** *If a sequence f_n of convex functions epi-converges to a function f, then f is also convex.*

Previously, we said that epi-convergence can be seem as a weaker form of uniform convergence. This comes from the fact that, a uniformly convergent sequence of lower semi-continuous functions also epi-converges.

Theorem 2.42 (Uniform and epi-convergence)**.** *If a sequence f_n of lower semi-continuous functions converges uniformly to f, it also epi-converges to f.*

With respect to minimization, the following theorem says that under some regularity conditions epi-convergence implies in convergence of the infimum. The proof to this can be found at Theorem 7.33 of [Rockafellar and Wets, 2011].

Theorem 2.43 (Epi-convergence and infimum)**.** *Suppose f_n is a sequence of proper lower semi-continuous functions such that eventually the its subsets $S_\alpha(f_n)$ are bounded. If f_n epi-converges to a proper lower-semicontinuous function f, then*

$$\inf f_n(x) \to \inf f(x).$$

2.2.4 Fenchel conjugate

In this section, we discuss some elements of duality theory for extended real functions over $\mathbb{R}^n$. This will be important on Chapter 3 when we discuss the Lagrangian relaxations of optimal value functions. As we will see, for a function f, there is a deep relation between the convex relaxation $\check{f}$ and its biconjugate f^{**}.

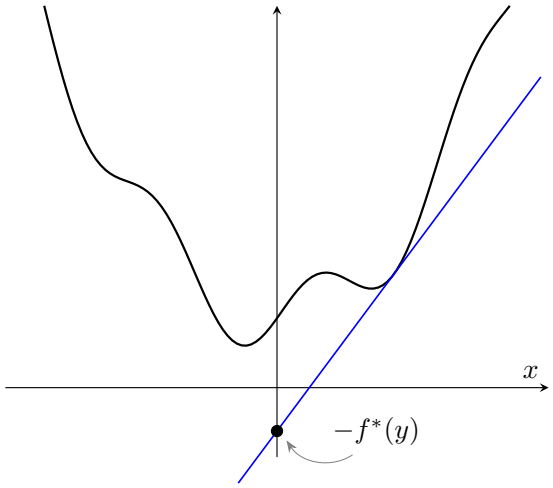


Figure 2.9: The conjugate f^* represents the vertical axis intercept of the largest affine function that is everywhere less than f.

Definition 2.44 (Fenchel conjugate)**.** The *Fenchel conjugate* of a function f is the function f^* defined by

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

For each point x, $y \mapsto \langle y, x, -\rangle f(x)$ is a continuous linear function on y. Since linear functions are convex, their supremum is a convex function by Theorem 2.24. Moreover, the supremum of lower semi-continuous functions is also lower semi-continuous by Theorem 2.37. This implies that f^* is always convex and lower semi-continuous.

For an arbitrary function f, its conjugate may be rewritten as

$$f^*(y) = -\sup\left\{b \mid b + \langle y, x\rangle \leqslant f(x), \forall x \in X\right\} \tag{2.11}$$

which means, geometrically, that $-f^*(y)$ is the maximum intercept with the vertical axis such that an affine function with inclination y is everywhere less than f as illustrated in Figure 2.9. Notice that $f^*(y) = \infty$ means that no cut with inclination y is valid for f. This interpretation will be useful again in Section 3.2.2 where we will use affine functions to underapproximate optimal value functions.

Hereby, we summarize some properties of the conjugate function.

1. The value of f^* at zero is minus the infimum of f.
2. Conjugation reverses inequalities, that is, if $f \leqslant g$ then $f^* \geqslant g^*$.
3. The conjugate of an infimum is the supremum of the conjugates:

$$(\inf_\nu f_\nu)^* = \sup_\nu f_\nu^*.$$

4. But the conjugate of a supremum is only less than the infimum of the conjugate:
$$(\sup_{\nu} f_\nu)^* \leqslant \inf_{\nu} f_\nu^*.$$

5. Summed constants leave the conjugate with reversed sign. That is, if $c \in \mathbb{R}$,
$$(f + c)^* = f^* - c.$$

6. Scaling by $k > 0$ becomes applying a perspective function:
$$(kf)^*(y) = kf^*\left(\frac{y}{k}\right).$$

7. Argument translations is the sum of a linear function. That is, if $g(x) = f(x - a)$,
$$g^*(y) = f^*(y) + \langle y, a, . \rangle$$

The definition of conjugate function also implies the following inequality relating f to its conjugate:
$$f(x) + f^*(y) \geqslant \langle y, x \rangle \tag{2.12}$$
for any $x \in X$ and $y \in X'$. This is known as *Fenchel's inequality* and is useful for taking estimates about f based on its conjugate.

2.2.4.1 The biconjugate

The conjugate function can be taken again with respect to f^* to get a new function f^{**}.

Definition 2.45 (Biconjugate)**.** The *biconjugate* of a function f is the function f^{**} defined by
$$f^{**}(x) = \sup_{y \in \mathbb{R}^n} \{\langle y, x \rangle - f^*(y)\} = \sup_{y \in \mathbb{R}^n} \inf_{z \in \mathbb{R}^n} \{f(z) + \langle y, x - z \rangle\}.$$

It always holds that
$$f^{**} \leqslant f, \tag{2.13}$$
since, by Fenchel's inequality (2.12), $\langle y, x \rangle - f^*(y) \leqslant f(x)$ and f^{**} is the supremum of the left-hand side over y.

If f is convex, proper and lower semi-continuous, we have that, in fact, $f = f^{**}$. This is known as the Fenchel-Moreau theorem.

Theorem 2.46 (Fenchel-Moreau)**.** *A function f equals its biconjugate f^{**} if and only if one of the following is true:*

1. *f is convex, proper and lower semi-continuous,*

2. f is identically equal to $+\infty$, or

3. f is identically equal to $-\infty$.

Additionally, under mildly regularly conditions, we can guarantee that when f is not convex, its biconjugate f^{**} is the largest convex lower semi-continuous function everywhere less than f.

Theorem 2.47 (Biconjugate and convex relaxation)**.** *If $\check{f}$ is proper, then*

$$\operatorname{epi}(f^{**}) = \operatorname{cl}\operatorname{conv}\operatorname{epi}(f).$$

In other words,

$$f^{**} = \operatorname{cl}(\check{f}).$$

2.3 Cones and inequalities

Although we can define linear or affine operators between any two vector spaces, the notion of convex function depends on its codomain being the real line $\mathbb{R}$. In this section, we use a special type of convex set called a *convex cone* to induce a partial order on a vector space. Thus, fixing cones on vector spaces allows us to extend the definitions of monotone and convex functions with respect to their induced orders.

Definition 2.48. A subset K of a vector space V is a *convex cone* if it is closed under non-negative linear combinations. That is, if x, $y \in K$ and t, $s \geqslant 0$ imply that $tx + sy \in K$.

Notice the similarity with the definition of a convex set. The difference in here is that the linear combinations can be made using any non-negative scalars, without the requirement of summing one. Of course, any convex cone is a convex set.

Remark 2.7. In most of literature, the term *cone* refers to a non-negative homogeneous set, that is, to a set K such that for any $\lambda \geqslant 0$,

$$x \in K \implies \lambda x \in K$$

and a convex cone is a cone that is also convex. Since only convex cones appear throughout this work, no confusion shall arise if we refer to them simply as cones. Therefore, in what follows, a cone will always mean a convex cone.

Definition 2.49. A convex cone K is *pointed* if it does not contain any line. In other words, K is pointed if for any non-zero $x \in K$, its negative $-x \notin K$.

Some examples of convex cones include

1. Any subspace is a cone, since it is closed by arbitrary linear combinations.
2. The cone $\mathbb{R}^n_+$ of vectors in $\mathbb{R}^n$ whose components are non-negative.

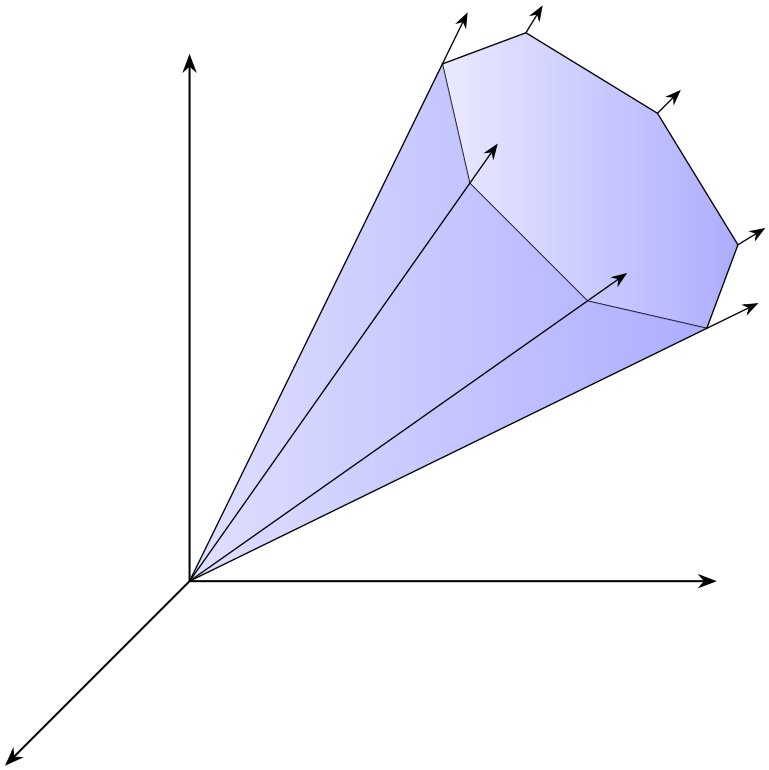


Figure 2.10: Example of a cone.

3. More generally, if V is a function space, the set V_+ of non-negative functions in V is a pointed convex cone.

4. Also, the set of convex functions in a function space V is a convex cone. Notice that it is not necessarily pointed, since it may contain affine functions (which are both convex and concave).

5. The set of $n \times n$ symmetric matrices is a vector space. An example of pointed convex cone on it is the set of all positive semi-definite matrices.

Some of the convexity-preserving operations from Section 2.1.1 also preserve a set being a cone.

1. The cartesian product $K_1 \times K_2$ of two cones is also a cone.

2. The intersection $K_1 \cap K_2$ of two cones is also a cone.

3. The Minkowski sum $K_1 + K_2$ of two cones is also a cone.

4. The image and preimage via a linear operator T are also cones.

The cartesian product and intersection also preserve being pointed.

A cone K on a vector space V induces a preorder on V by

$$x \leqslant_K y \iff x - y \in K. \tag{2.14}$$

Notice that if K is the cone of positive real numbers, this is the usual order on the reals. Analogously to case in $\mathbb{R}$, the vectors greater than zero are precisely those in K:

$$x \geqslant_K 0 \iff x \in K.$$

This preorder has some nice properties, summarized on Theorem 2.50 including the fact that $\leqslant_K$ is a partial order if and only if K is pointed.

Theorem 2.50. *A preorder $\leqslant_K$ induced by a cone $K \subset V$ satisfies*

1. *$x \leqslant_K x$ for all $x \in V$.*
2. *If $x \leqslant_K y$ and $y \leqslant_K z$ then $x \leqslant_K z$.*
3. *If $x \leqslant_K y$ then $-x \geqslant_K y$.*
4. *If $\lambda \geqslant 0$ and $x \leqslant_K y$ then $\lambda x \leqslant_K \lambda y$.*
5. *If $x \leqslant_K y$ and $v \leqslant_K w$ then $x + v \leqslant_K y + w$.*

Additionally, if K is pointed, $\leqslant_K$ is antisymmetric. That is, $x \leqslant_K y$ and $y \leqslant_K x$ implies $x = y$.

More information about conic inequalities can be found at [Rockafellar and Wets, 2011] and [Boyd and Vandenberghe, 2004], including the converse to Theorem 2.50 which says that any order satisfying these properties must arise from some cone.

2.3.1 Convexity in relation to a cone

We can use the orders generated by cones to extend the definitions of monotone and convex functions to maps between two vector spaces endued with cones.

Definition 2.51. Given two cones $K \subset V$ and $L \subset W$, a function $f\colon V \to W$ is said to be (K, L)-*monotone* or *monotone in relation to K and L* if

$$x \leqslant_K y \implies f(x) \leqslant_L f(y).$$

If a function $f\colon V \to \mathbb{R}$ is monotone in relation to K and R_+ we will simply say that it is K-monotone.

Definition 2.52. Given a cone $K \subset W$, a function $f\colon X \to W$ whose domain is a convex set is K-*convex* or *convex in relation to K* if for any pair of point x, $y \in V$ and $\lambda \in [0, 1]$:

$$f(\lambda x + (1 - \lambda)y) \leqslant_K \lambda f(x) + (1 - \lambda) f(y).$$

Remark 2.8. A function is convex in the usual sense if it is convex in relation to the cone $\mathbb{R}_+$ of non-negative real numbers.

Dually to Definition 2.52, we say that a function f is *K-concave* if $-f$ is K-convex.

A curious example is the map $f \mapsto \check{f}$, which is itself *concave* in relation to the cone of non-negative functions. To see this, consider two functions f, g and some $\lambda \in [0,1]$. The function $\lambda\check{f} + (1-\lambda)\check{g}$ is convex and always below $\lambda f + (1-\lambda)g$. Therefore, the definition of convex relaxation implies that

$$\operatorname{conv}(\lambda f + (1-\lambda)g) \geqslant \lambda \operatorname{conv}(f) + (1-\lambda)\operatorname{conv}(g). \tag{2.15}$$

Remember that convexity of a function $f\colon X \to \mathbb{R}$ can be characterized by the convexity of its epigraph. Similarly, we can associate to each function $f\colon X \to W$ its K-epigraph in $X \times W$. This set is convex if and only if f is K-convex.

Definition 2.53. The K-epigraph of a function $f\colon X \to W$ is the set

$$\operatorname{epi}_K(f) = \{(x,t) \in V \times W \mid f(x) \leqslant_K t\}.$$

Theorem 2.54. *A function f is K-convex if and only if $\operatorname{epi}_K(f)$ is a convex set.*

Proof. If f is K-convex then it is always true that

$$f(\lambda x + (1-\lambda)y) \leqslant_K \lambda f(x) + (1-\lambda)f(y)$$

and if (x,t), (y,s) are elements of $\operatorname{epi}_K(f)$,

$$\lambda f(x) + (1-\lambda)f(y) \leqslant_K \lambda t + (1-\lambda)s.$$

These two inequalities imply that for any $\lambda \in [0,1]$, $\lambda(x,t) + (1-\lambda)(y,s)$ is also in $\operatorname{epi}_K(f)$. In other words: it is a convex set.

Now assume that $\operatorname{epi}_K(f)$ is a convex set. As the points $(x, f(x))$ and $(y, f(y))$ are always elements of it, the convexity tells us that any convex combination

$$\lambda(x, f(x)) + (1-\lambda)(y, f(y)) = (\lambda x + (1-\lambda)y,\ \lambda f(y) + (1-\lambda)f(y))$$

is also an element of $\operatorname{epi}_K(f)$. Which is equivalent to say that

$$f(\lambda x + (1-\lambda)y) \leqslant_K \lambda f(x) + (1-\lambda)f(y)$$

for any $x, y \in V$ and $\lambda \in [0,1]$, which is the Definition of K-convexity. □

2.3.1.1 Composition of K-convex functions

As in the scalar case, the K-convex functions are closed by addition and multiplication by non-negative scalars. Therefore they also constitute a cone.

Theorem 2.55. *If K is a cone in W and V an arbitrary vector space, the set $\{\, f\colon V \to W \mid f$ is K-convex$\}$ is a cone.*

We can also generalize Theorem 2.23 about composition of convex functions to deal with compositions of conic convex functions.

Theorem 2.56. *If $f\colon X \to Y$ is K-convex and $g\colon Y \to Z$ is L-convex and monotone with respect to K and L, then the composition $g \circ f$ is L-convex.*

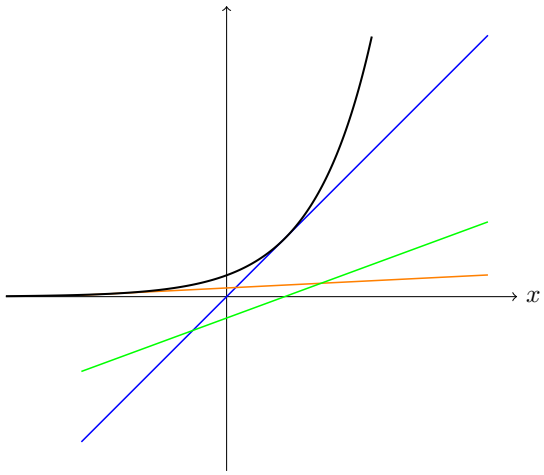


Figure 2.11: Any valid cut for a non-decreasing ($\mathbb{R}_+$-monotone) convex function has non-negative slope.

2.3.1.2 Dual cones and K-monotone functions

Here, we give a characterization of K-monotone functions via the inclinations of their valid cuts. This theorem is a generalization of a theorem needed at Section 3.4, where we will prove an *almost Jensen's inequality* for convex monotone functions and coherent risk measures. In Figure 2.11, we see an example applied to a 1-dimensional function.

Definition 2.57. If K is a cone in V, there is a cone $K^\star$ in its dual space V^* defined by

$$K^* = \{\omega \in V^* \mid \langle \omega, x \rangle \geqslant 0,\ \forall x \in K\}$$

and called the *dual cone* to K.

Theorem 2.58. *Let f be a real valued convex function. Then f is K-monotone if and only if any of its valid cuts $\langle a, x\rangle + b$ satisfies $a \in K^\star$.*

Proof. First, assume that f is K-monotone. Since f is convex, we may use Theorem 2.27 and assume without loss of generality that this cut equals f at some point x_0,

$$\langle a, x\rangle + b = \langle a, x - x_0\rangle + f(x_0).$$

Given an element $k \in K$, this cut satisfies

$$f(x_0 - k) \geqslant f(x_0) + \langle a, -k\rangle$$

and the K-monotonicity of f implies that

$$x_0 - k \leqslant_K x_0 \implies f(x_0 - k) \leqslant f(x_0).$$

Taking both of those inequalities together,

$$\langle a, k\rangle \geqslant f(x_0) - f(x_0 - k) \geqslant 0.$$

Since $k \in K$ was arbitrary, we conclude that $a \in K^*$.

Now assume that all valid cuts to f have inclination $a \in K^*$. Taking a tight cut at the point x, we have for all y that

$$f(y) \geqslant f(x) + \langle a, y - x \rangle.$$

If $y \geqslant_K x$, the inner product on the previous expression is non-negative and the left-hand side is greater than $f(x)$,

$$y \geqslant_K x \implies f(y) \geqslant f(x) + \langle a, y - x \rangle \geqslant f(x).$$

Therefore, f is K-monotone. □

2.3.1.3 Valid cuts for K-convex functions

In this section, we show that if the order $\leqslant_K$ satisfies some regularity properties, there is an analogous of theorem 2.27 for K-convex functions. The theorem in this section is a consequence of the Hahn-Banach theorem and a proof to it can be found as proposition (VHB4) in Section 12.34 of [Schechter, 1997]. We begin with some definitions.

Definition 2.59. A non-empty subset A of V is *bounded above in* $\leqslant_k$ if there is a $y \in V$ such that for all $x \in A$, $x \leqslant_K y$.

Definition 2.60. The order in V induced by a pointed convex cone K is *Dedekind complete* if any non-empty subset A of V which is bounded above has a least upper bound. In other words, there exists a $w \in V$ such that if $x \leqslant_K y$ for all $x \in A$ then $w \leqslant_K y$.

If the order induced by a cone is Dedekind complete, any K-convex function can be written as the pointwise maximum (in relation to this order) of the affine functionals that are everywhere less than it.

Theorem 2.61 (K-convex Support Theorem)**.** *Suppose $\leqslant_K$ is a Dedekind complete order on W. If $f\colon V \to W$ is K-convex, for each point $x_0 \in V$ there exists an affine functional ϕ such that $\phi(x) \leqslant_K f(x)$ for all $x \in V$ and $\phi(x_0) = f(x_0)$.*

Optimization

3

In this chapter we study many forms of optimization problems in a general setting. Despite this generality, most of the motivation for the techniques described here comes from mixed integer programs and is readily applicable to these problems. Hence, the reader is welcome to think about these problems as being composed of a finite number of real or integer decision variables, and that most of the structure of the problem's constraints is linear or convex.

As we will see, the theory of convex sets and functions, developed in Chapter 2, will play a major role in the way we solve optimization problems. This happens because convex functions are "easy" to minimize when compared to the general case, thanks to results like Theorem 2.20, which ensures that every local minimum of a convex function is global; or Theorem 2.27, which says that they can be approximated by affine functions that lie below them.

We start in Section 3.1 by defining and studying the properties of both general and convex optimization problems. In Section 3.2 we study parameterized optimization problems and their *optimal value functions*, which give the optimal value of the problem as a function of the parameters. We begin by developing the properties in a general setting and proceed to deduce finer results in the special cases with the most importance. In Section 3.3 we look at the properties of multi-stage optimization problems, which deal with sequences of problems where each one depends on the optimal value functions of the previous. In Section 3.4 we study what happens when we consider that some parameters of our problems are random. It is composed of a study of uncertainty in two-stage problems as well as a discussion about risk averse optimization.

Through this entire chapter, a strong emphasis is put into approximating optimal value functions by affine underestimators, called *valid cuts*. We do this because many important algorithms to solve multistage and stochastic programs rely on iteratively approximating the cost-to-go functions by cuts. For examples, the interested reader can see Benders Decomposition for mixed integer problems in [Benders, 1962], Stochastic Dual Dynamic Programming (SDDP) for convex multi-stage stochastic programs in [Pereira and Pinto, 1991] or Stochastic Dual Dynamic integer Programming (SDDiP) for mixed integer multi-stage stochastic programs in [Zou et al., 2018].

3.1 Optimization problems

Definition 3.1. An *optimization problem* consists of minimizing a function c subject to its arguments being inside a set X. Throughout this work, these problems will be denoted as

$$\begin{aligned} \min_{x} \quad & c(x) \\ \text{s.t.} \quad & x \in X. \end{aligned}$$

The function c is called the problem's *objective function* and the set X its *feasible set*, while any point $x \in X$ is a *feasible point*. Sometimes, the expression $x \in X$ itself will be called a *constraint* for the problem. The term "min" on a optimization problem stands for *minimize* and "s.t." is an abbreviation for *subject to*, meaning that an optimization problem is read as "minimize $c(x)$ subject to $x \in X$. A problem's optimal value, often denoted p^*, corresponds the infimum of $c(x)$ on the set X,

$$p^* = \inf_{x \in X} c(x).$$

On the following, specially when dealing with parameterized optimization problems, we will write the optimization problem as denoting its optimal value. No confusion should arise from this but it is important to notice that despite writing minimize, the optimal value is the problem's infimum and may not be attained.

If $X = \varnothing$, the problem is called *infeasible* and we will apply the usual convention of setting $\inf_{x \in \varnothing} c(x) = +\infty$ no matter the objective function c.

An analogous definition could be made for maximization problems by exchanging *minimize* for *maximize* and defining the optimal value to be the problem's supremum. Since $\max c = -\min -c$, the theory can be entirely developed for minimization without loss of generality.

When minimizing a convex function over a convex set, the problem in Definition 3.1 is called a *convex optimization problem*.

Defining the indicator function of a set X,

$$I_X(x) = \begin{cases} 0, & x \in X \\ +\infty, & x \notin X, \end{cases} \tag{3.1}$$

any optimization problem can be written as an unconstrained one, since

$$\min_{x} \; c(x) + I_X(x) \;=\; \begin{aligned} \min_{x} \quad & c(x) \\ \text{s.t.} \quad & x \in X \end{aligned} \tag{3.2}$$

and the minimum is attained at the same point $x^\star$. Notice that the function I_X is convex if and only if the set X is convex.

In practice, the constraint set in Definition 3.1 may be too abstract to work with. Thus, we will generally restrict ourselves to constraints which are sublevel

or level sets of some function. That is, problems in the form

$$\begin{aligned} \min_x \quad & c(x) \\ \text{s.t.} \quad & g(x) \leqslant_K 0 \\ & h(x) = 0 \end{aligned} \tag{3.3}$$

where K is some cone, as defined in Section 2.3. An important case is when K is the non-negative cone on $\mathbb{R}^l$, denoted $\mathbb{R}^l_+$. Then, $g \leqslant_{\mathbb{R}^l_+} 0$ means that each component g_i of g is less or equal than zero and the problem can be equivalently formulated as

$$\begin{aligned} \min_x \quad & c(x) \\ \text{s.t.} \quad & g_i(x) \leqslant 0, \text{ for } i = 1, \ldots, l \\ & h_i(x) = 0, \text{ for } i = 1, \ldots, k \end{aligned}$$

If g is a K-convex function and h is an affine function, they constrain the problem's decision variable to be inside a convex set and, therefore, problem (3.3) is convex if the objective function is convex.

Remark 3.1. Notice that $x \in X$ if and only if $I_X(x) \leqslant 0$. Thus, without loss of generality, any optimization problem can be written using only sublevel constraints.

3.1.1 Lagrangian relaxation

On Equation (3.2), we showed a way in which an optimization problem's constraints can be written as part of the objective function. This method may have some problems, such as the objective function c being everywhere differentiable but the function $c + I_X$ not.

Other problem that may arise is that the constraints may be described via inequalities, such as in Equation (3.3), which are infeasible. In this case, it is better to substitute the constraint by a term that approximates the problem by a another, which is feasible. When we substitute an equality or inequality constraint by a linear term on the objective, it is called a *Lagrangian relaxation* of the problem.

Definition 3.2 (Lagrangian Relaxation)**.** A *Lagrangian relaxation* to an optimization problem with the form

$$\begin{aligned} \min_x \quad & c(x) \\ \text{s.t.} \quad & g(x) \leqslant_K 0 \\ & h(x) = 0 \\ & x \in X \end{aligned}$$

is another optimization problem

$$\begin{aligned} \min_x \quad & c(x) + \langle \lambda, g(x) \rangle + \langle \nu, h(x) \rangle \\ \text{s.t.} \quad & x \in X \end{aligned}$$

where $\lambda \in K^*$, that is, $\langle \lambda, k \rangle \geqslant 0$ for all $k \in K$.

The function

$$L(x, \lambda, \nu) = c(x) + \langle \lambda, g(x) \rangle + \langle \nu, h(x) \rangle \tag{3.4}$$

is called the problem's *Lagrangian* and the parameters λ and ν, the *Lagrange multipliers*. Notice that since $\langle \nu, h(x) \rangle = 0$ and $\langle \lambda, g(x) \rangle \leqslant 0$, the Lagrangian always satisfies

$$L(x, \lambda, \nu) \leqslant c(x) + I_{\{g(x) \leqslant_K 0\}} + I_{\{h(x)=0\}}. \tag{3.5}$$

In particular, for each feasible point x of the original problem, $L(x, \lambda, \nu) \leqslant c(x)$.

The name *Lagrangian relaxation* comes from the fact that the Lagrangian is an underapproximation of the original objective function on the original feasible points but is well-defined on a larger feasible set.

When forming the Lagrangian relaxation of an optimization problem, we get different results according to the parameters we choose. This means that we can define a function

$$d(\lambda, \nu) = \begin{cases} \inf_{x \in X} L(x, \lambda, \nu), & \lambda \geqslant_{K*} 0 \\ -\infty, & \lambda \not\geqslant_{K*} 0 \end{cases} \tag{3.6}$$

called the problem's *dual function*. The function dual function d gives the value of the relaxed problem as a function of the Lagrange multipliers; this an instance of an optimal value function as we will study in Section 3.2. Notice that the dual function is *concave*, by Theorem 2.24, since the Lagrangian is linear on the Lagrange multipliers and, therefore, d is the minimum of linear functions.

By calling p^* the optimal value of a minimization problem,

$$\begin{aligned} p^* = \min_x \quad & c(x) \\ \text{s.t.} \quad & g(x) \leqslant_K 0 \\ & h(x) = 0 \\ & x \in X, \end{aligned} \tag{3.7}$$

the dual function to this problem will always be below it for any value of λ, ν,

$$d(\lambda, \nu) \leqslant p^*. \tag{3.8}$$

This means that we can construct a maximization problem in terms of the dual function which is always below the original problem's optimal value, called the *dual problem*.

Definition 3.3 (Dual problem)**.** Given an optimization problem such as in (3.7), we define its *dual problem* as the maximum of all its Lagrangian relaxations,

$$\begin{aligned} d^* = \max_{\lambda, \nu} \quad & d(\lambda, \nu) \\ & \lambda \geqslant_{K*} 0. \end{aligned}$$

Remark 3.2. Notice that the dual problem is always a convex problem, even if the original problem is not.

By taking the supremum on inequality 3.8, we always have that

$$d^* \leqslant p^*,$$

a property known as *weak duality*. In terms of the Lagrangian, weak duality says that

$$\sup_{\lambda \in K^*} \inf_{x \in X} L(x, \lambda, \nu) \leqslant \inf_{x \in X} \sup_{\lambda \in K^*} L(x, \lambda, \nu),$$

which is an usual property of infima and suprema.

When $d^* = p^*$, we say that *strong duality* holds and the difference $p^* - d^*$ is called the problem's *duality gap*. There are conditions on the objective function and constraints of the primal problem that guarantee strong duality. One such condition is Slater's condition, whose proof can be found at Chapter 5 of [Boyd and Vandenberghe, 2004].

Theorem 3.4 (Slater's condition)**.** *Given a* convex *optimization problem*

$$\begin{array}{ll} \min\limits_{x} & c(x) \\ s.t. & g_i(x) \leqslant_K 0, \text{ for } i = 1, \ldots, k \\ & Ax = b, \end{array}$$

and suppose that there is a feasible point y such that $g_i(y) <_{K_i} 0$ whenever g_i is not affine. That is, $-g_i(y)$ is in the interior of K_i for all non-affine g_i. Then strong duality holds for this problem.

In practical applications, this condition is enough to imply that in a convex optimization problem, strong duality normally holds. Furthermore, Slater's condition is equivalent to the problem being feasible if all its constraints are affine such as in a linear program.

Corollary 3.4.1. If an optimization problem

$$\begin{array}{ll} \min\limits_{x} & c(x) \\ \text{s.t.} & Gx \leqslant h, \\ & Ax = b, \end{array}$$

has at least one feasible point, then strong duality holds for it.

Thereafter we shall always assume that strong duality holds for convex problems, but not for non-convex problems.

3.2 Optimal value functions

It is common for us to want to solve not a single optimization problem, but a family of related problems in which we change certain parameters, as the right-hand side of constraints or the objective function. In this section we study the

properties of the so called *optimal value function* of a parameterized family of optimization problems. This function tells us how the problem's optimal value changes when we modify the problem's parameters.

We begin in Section 3.2.1 by studying some regularity and characterization results for optimal value functions. First, we show that the optimal value of any optimization problem is convex with respect to varying its objective function. Afterwards, we proceed to study optimal value functions that vary the right-hand side of equality and inequality constraints. We will begin by showing that the optimal value functions of convex problems are also convex functions. Then we will particularize for linear programs and, subsequently, use these results to characterize the optimal value functions of mixed integer problems.

In Section 3.2.2, we use the tools from Section 3.1.1 to study ways to approximate an optimal value function by collection of affine functions that are everywhere below it. These are called *cuts* and will be of great importance in Chapter 7.

3.2.1 Characterizations of optimal value functions

Depending on the structure of a parameterized optimization problem, its optimal value function may posses some simple characterization or can be even guaranteed to be convex. This section discusses some special cases of parameterized problems whose optimal value functions are either convex or piecewise convex. References to these results include [Fiacco and Kyparisis, 1986] for a compendium of convexity properties and [Hassanzadeh and Ralphs, 2014; Bank et al., 1984] for mixed integer problems.

As a first case, let us consider a fixed feasible set X. Then, given a set of functions, we may look at the optimal value function $f(c) = \inf_{x \in X} c(x)$ which, for each objective function c, returns the smallest value of c over X. Theorem 3.5, proved shortly, says that independently of our choice for the set X or the function space where c lies, this function is always *concave*.

Theorem 3.5. *For any fixed set X, the function defined by*

$$f(c) = \begin{array}{ll} \min\limits_{x} & c(x) \\ s.t. & x \in X \end{array}$$

is concave *in c.*

Proof. For each fixed $x \in X$, the function $f_x(c) = c(x)$ is linear in c. The result follows from the fact that the infimum of linear functions is concave. □

An example of this result was the dual function to an optimization problem, from Equation (3.6). This function is always concave, no matter the original problem. As a corollary to Theorem 3.5, the optimal value function which varies the objective function of a maximization problem is always *convex*.

Corollary 3.5.1. For any fixed set X, the function defined by

$$\begin{aligned} f(c) = \max_{x} \quad & c(x) \\ \text{s.t.} \quad & x \in X \end{aligned}$$

is *convex* in c.

We now proceed to study the optimal value functions that vary the right-hand side of the problem's constraints. That is, functions with the form

$$\begin{aligned} f(a,b) = \min_{x} \quad & c(x) \\ \text{s.t.} \quad & g(x) \leqslant_K a \\ & h(x) = b \\ & x \in X. \end{aligned} \tag{3.9}$$

Theorem 3.6 (Optimal value function for a convex problem)**.** *Let f be the optimal value function defined by* (3.9) *and suppose that the represented problems are convex, that is, c is convex, g is K-convex, h is affine, and X is a convex set. Then f is a convex function.*

Proof. We will show that f satisfies the Jensen's inequality for any $\lambda \in [0,1]$. If the problem represented by f is infeasible at some point (a,b), the inequality follows directly. Then, we can consider only the case when $f(a,b)$ has at least a feasible point.

Suppose x_1 is feasible for the problem $f(a_1,b_1)$ and x_2 is feasible for the problem $f(a_2,b_2)$. Then, for any $\lambda \in [0,1]$, the point $\lambda x_1 + (1-\lambda)x_2$ is feasible for the problem with the average parameters, $f(\lambda a_1 + (1-\lambda)a_2, \lambda b_1 + (1-\lambda)b_2)$, because

$$\begin{aligned} &g(\lambda x_1 + (1-\lambda)x_2) \leqslant_K \lambda g(x_1) + (1-\lambda)g(x_2) \leqslant_K \lambda a_1 + (1-\lambda)a_2, \\ &h(\lambda x_1 + (1-\lambda)x_2) = \lambda h(x_1) + (1-\lambda)h(x_2) = \lambda b_1 + (1-\lambda)b_2. \end{aligned}$$

Using the convexity of the objective function,

$$c(\lambda x_1 + (1-\lambda)x_2) \leqslant \lambda c(x_1) + (1-\lambda)c(x_2).$$

By minimizing the left-hand side over all the feasible points for the average parameters,

$$f(\lambda a_1 + (1-\lambda)a_2, \lambda b_1 + (1-\lambda)b_2) \leqslant \lambda c(x_1) + (1-\lambda)c(x_2),$$

and by minimizing the right-hand side over all x_1 feasible for (a_1,b_1) and all x_2 feasible for (a_2,b_2):

$$f(\lambda a_1 + (1-\lambda)a_2, \lambda b_1 + (1-\lambda)b_2) \leqslant \lambda f(a_1,b_1) + (1-\lambda)f(a_2,b_2). \qquad \square$$

We now proceed to discuss optimal value functions of linear programs, that is, problems of the form

$$\begin{aligned} f(a,b) = \min_{x} \quad & \langle c, x \rangle \\ \text{s.t.} \quad & Dx \leqslant a \\ & Ax = b \end{aligned} \tag{3.10}$$

where c is a vector and A, D are matrices. Since linear functions are convex, a linear problem is always convex. Furthermore, the only condition for strong duality to hold for a linear problem it that it or its dual must be feasible.

Any linear program can be put into a *standard form* where all decision variables are non-negative and the only restriction is of the form $Ax = b$, as can be found in [Dantzig, 1963, sec 3-8, pg 60]. This means that we can restrict our attentions to problems of the form

$$\begin{aligned} \min_{x} \quad & \langle c, x \rangle \\ \text{s.t.} \quad & Ax = b \\ & x \geqslant 0. \end{aligned} \tag{3.11}$$

The optimal value function that varies the right-hand side of the equality constraint on standard form linear programs can be characterized as a *polyhedral function*, that is, a convex piecewise linear function, as can be found in [Hassanzadeh and Ralphs, 2014, prop 1, pg 6] or [Blair and Jeroslow, 1977, prop 3.1, pg 131].

Theorem 3.7 (Optimal value function for a linear program)**.** *The optimal value function of a standard form linear program, defined by*

$$\begin{aligned} f(b) = \min_{x} \quad & \langle c, x \rangle \\ s.t. \quad & Ax = b, \\ & x \geqslant 0 \end{aligned}$$

is a polyhedral function.

Remark 3.3. Notice that in Theorem 3.6, the optimal value function $f(a, b)$ varied the right-hand side of both the inequality and equality constraints while in the Theorem 3.7 only the equality constraints are being varied by $f(b)$.

These results can be used to characterize the optimal value functions of a class of non-convex problems called *mixed integer linear programs*, or *MILP* for short. These are linear programs with the additional constraint that some of its decision variables must be integer. In standard form:

$$\begin{aligned} \min_{x,z} \quad & \langle c_1, x \rangle + \langle c_2, z \rangle \\ \text{s.t.} \quad & Ax + Bz = b, \\ & x, z \geqslant 0, \\ & x \in \mathbb{R}^n,\ z \in \mathbb{Z}^k. \end{aligned} \tag{3.12}$$

Theorem 3.8 (Optimal value function for a mixed integer linear program)**.** *Assume that A, B are matrices with rational coefficients. The optimal value function of a mixed integer linear program defined by*

$$\begin{aligned} f(b) = \min_{x,z} \ & \langle c_1, x\rangle + \langle c_2, z\rangle \\ s.t. \ & Ax + Bz = b, \\ & x, z \geqslant 0, \\ & x \in \mathbb{R}^n,\ z \in \mathbb{Z}^k. \end{aligned}$$

is a piecewise polyhedral function.

See [Hassanzadeh and Ralphs, 2014] for further discussion of this type of optimal value function. Moreover, we can give a further characterization of this function as the minimum of a countable collection of translations of a polyhedral function ϕ. To do this, let f be as in Theorem 3.8 and set ϕ to be its restriction to continuous variables,

$$\begin{aligned} \phi(b) = \min_{x,z} \ & \langle c_1, x\rangle \\ \text{s.t.} \ & Ax = b, \\ & x \geqslant 0. \end{aligned}$$

That is, ϕ represents the same problem as f when the integer decision variables are all zero. Then, f is the minimum of a countable amount of translations of ϕ, which is polyhedral,

$$\begin{aligned} f(b) = \min_{x,z} \ & \langle c_1, x\rangle + \langle c_2, z\rangle \\ \text{s.t.} \ & Ax + Bz = b, \\ & x, z \geqslant 0, \\ & x \in \mathbb{R}^n,\ z \in \mathbb{Z}^k \end{aligned} = \min_{\substack{z\geqslant 0,\\ z\in\mathbb{Z}^k}} \left\{ \langle c_2, z\rangle + \min_{x} \begin{aligned} & \langle c_1, x\rangle \\ \text{s.t.} \ & Ax = b - Bz, \\ & x \geqslant 0 \end{aligned} \right\}$$

$$= \min_{\substack{z\geqslant 0,\\ z\in\mathbb{Z}^k}} \langle c_2, z\rangle + \phi(b - Bz). \tag{3.13}$$

On the equation above, the minimum is taken over all possible integer values that the decision variable z can take. In [Hassanzadeh and Ralphs, 2014], it is shown that this minimum can be taken over a smaller family of integer points and this minimal family is studied thoroughly.

This same idea can be applied to study the optimal value functions of optimization problems that are convex except for some variables being integer. Before proceeding, we notice that any optimization problem is equivalent to another one with a linear objective function. This is done by going from the objective c to its epigraph:

$$\begin{aligned} \min_{x} \ & c(x) \\ \text{s.t.} \ & x \in X \end{aligned} = \begin{aligned} \min_{x,t} \ & t \\ \text{s.t.} \ & (x,t) \in \operatorname{epi}(c) \\ & x \in X \end{aligned} = \begin{aligned} \min_{x,t} \ & t \\ \text{s.t.} \ & c(x) \leqslant t \\ & x \in X. \end{aligned} \tag{3.14}$$

By virtue of this, we will consider only the case with linear objective functions on the following theorem.

Theorem 3.9 (Optimal value function for a mixed integer convex program)**.** *Given a convex set X a compact subset Z of $\mathbb{Z}^k$, and rational matrices A and B, the optimal value function*

$$\begin{aligned} f(b) = \min_{x,z} \quad & \langle c_1, x\rangle + \langle c_2, z\rangle \\ s.t. \quad & Ax + Bz = b, \\ & x \in X,\ z \in Z \end{aligned}$$

is piecewise convex*.*

Proof. The convex restriction

$$\begin{aligned} \phi(b) = \min_{x,z} \quad & \langle c_1, x\rangle \\ \text{s.t.} \quad & Ax = b, \\ & x \in X \end{aligned}$$

is a convex function, by Theorem 3.6. The function f can be written as the minimum over translations of ϕ by

$$\begin{aligned} f(b) = \min_{x,z} \quad & \langle c_1, x\rangle + \langle c_2, z\rangle \\ \text{s.t.} \quad & Ax + Bz = b, \\ & x \in X,\ z \in Z \end{aligned} = \min_{z\in Z} \langle c_2, z\rangle + \min_{x} \quad \begin{aligned} & \langle c_1, x\rangle \\ \text{s.t.} \quad & Ax = b - Bz, \\ & x \in X \end{aligned}$$

$$= \min_{z\in Z} \langle c_2, z\rangle + \phi(b - Bz).$$ □

3.2.2 Approximation by cuts

An affine function ψ everywhere less than f is called a *valid cut* for f, and this cut is said to be *tight* for f if there is at least one point a where $\psi(a) = f(a)$.

Remember from formula (2.7) that the convex relaxation of a function f can be calculated as the maximum of all affine functions everywhere less than it. If f is an optimal value function, we can use the duality theory from Section 3.1.1 to estimate valid cuts for $\check{f}$ from the optimal Lagrange multipliers of the problem represented by f. As we will see, in this context, strong duality is equivalent to these cuts being tight for f at some point.

This way to underapproximate an optimal value function is an essential part of many algorithms to solve multi-stage stochastic programs such as [Pereira and Pinto, 1991] for convex problems or [Zou et al., 2018] for mixed integer problems and will be important in Chapter 7 where we will apply the theoretical results from Chapters 5 and 6 to better estimate cuts for stochastic optimization problems.

In this section, we will always consider an optimal value function f that varies the constraints of a problem, as defined in Equation (3.9). We will assume that both this optimal value function and its convex relaxation $\check{f}$ are proper and lower semi-continuous. In this context, properness amounts to saying that there

is at least one parameter (a, b) such that the problem represented by $f(a, b)$ is feasible and that for all parameters the objective function c is bounded below on the feasible set. The lower semi-continuity assumptions will be important to guarantee some results related to duality, as we will shortly see.

We can form the Lagrangian relaxation of the problem represented by $f(a, b)$,

$$d_{a,b}(\lambda, \nu) = \begin{cases} \inf\limits_{x \in X} c(x) + \langle \lambda, g(x) - a \rangle + \langle \nu, h(x) - b \rangle & , \lambda \geqslant_{K*} 0 \\ -\infty & , \lambda \not\geqslant_{K*} 0 \end{cases} \tag{3.15}$$

and the dual optimal value function

$$\begin{aligned} d(a, b) = \max_{\lambda, \nu} \quad & d_{a,b}(\lambda, \nu) \\ \text{s.t.} \quad & \lambda \geqslant_{K*} 0. \end{aligned} \tag{3.16}$$

is, by weak duality, always below the original optimal value function,

$$d(a, b) \leqslant f(a, b). \tag{3.17}$$

The duality gap for optimal value functions is a function $\text{gap}(f) = f - d$ of the parameters. Supposing that the maximum is attained in the dual problem, we can also define the optimal Lagrange multipliers $\lambda^*_{a,b}$ and $\nu^*_{a,b}$ satisfying $d_{a,b}(\lambda^*_{a,b}, \nu^*_{a,b}) = d(a, b)$.

The dual functions have an interpretation in terms of the Fenchel conjugate, from Section 2.2.4, of the optimal value function f. In fact, the optimal dual function d equals the biconjugate of f. To see this, we must write the Lagrangian relaxation in a slightly different way. By adding decision variables $\bar{a}$ and $\bar{b}$, the optimal value f can be equivalently formulated as

$$\begin{aligned} f(a, b) = \min_{x, \bar{a}, \bar{b}} \quad & c(x) \\ \text{s.t.} \quad & g(x) \leqslant_K \bar{a} \\ & h(x) = \bar{b} \\ & a = \bar{a}, b = \bar{b} \\ & x \in X. \end{aligned}$$

Relaxing only the constraints $a = \bar{a}$, $b = \bar{b}$, we get another formulation of the dual function that is, nevertheless, equivalent to the one at (3.15):

$$\begin{aligned} d_{a,b}(\lambda, \nu) = \min_{x, \bar{a}, \bar{b}} \quad & c(x) + \langle \lambda, \bar{a} - a \rangle + \langle \nu, \bar{b} - b \rangle \\ \text{s.t.} \quad & g(x) \leqslant_K \bar{a} \\ & h(x) = \bar{b} \\ & x \in X. \end{aligned} \tag{3.18}$$

Separating the terms that depend on x or $\bar{a}, \bar{b}$, this expression can be rewritten

using the conjugate of f as

$$\begin{aligned} d_{a,b}(\lambda,\nu) &= \inf_{\bar{a},\bar{b}} \begin{pmatrix} \min\limits_{x} & c(x) \\ \text{s.t.} & g(x) \leqslant_K \bar{a} \\ & h(x) = \bar{b} \\ & x \in X. \end{pmatrix} + \langle \lambda, \bar{a} - a \rangle + \langle \nu, \bar{b} - b \rangle \\ &= \inf_{\bar{a},\bar{b}} \big\{ f(\bar{a},\bar{b}) + \langle \lambda, \bar{a} - a \rangle + \langle \nu, \bar{b} - b \rangle \big\} \\ &= \inf_{\bar{a},\bar{b}} \big\{ f(\bar{a},\bar{b}) + \langle \lambda, \bar{a} \rangle + \langle \nu, \bar{b} \rangle \big\} - \langle \lambda, a \rangle - \langle \nu, b \rangle \\ &= -\sup_{\bar{a},\bar{b}} \big\{ \langle -\lambda, \bar{a} \rangle + \langle -\nu, \bar{b} \rangle - f(\bar{a},\bar{b}) \big\} - \langle \lambda, a \rangle - \langle \nu, b \rangle \\ &= -f^*(-\lambda,-\nu) - \langle \lambda, a \rangle - \langle \nu, b \rangle. \end{aligned}$$

Finally, we can take the supremum over all possible Lagrange multipliers to get the dual optimal value function. Notice that, by construction, $d_{a,b}(\lambda,\nu)$ equals $-\infty$ when $\lambda \not\geqslant_{K^*} 0$, therefore the supremum will not be attained at these points,

$$\begin{aligned} d(a,b) &= \sup_{\lambda \geqslant_{K^*} 0,\, \nu} d_{a,b}(\lambda,\nu) = \sup_{\lambda,\,\nu} d_{a,b}(\lambda,\nu) \\ &= \sup_{\lambda,\,\nu} \big\{ -f^*(-\lambda,-\nu) - \langle \lambda, a \rangle - \langle \nu, b \rangle \big\} \\ &= f^{**}(a,b). \end{aligned}$$

Under our hypotheses of the convex relaxation $\check{f}$ being proper and lower semi-continuous, it follows from Theorem 2.47 that

$$d = f^{**} = \check{f}. \tag{3.19}$$

Therefore, weak duality always holds because the convex relaxation of a function is everywhere below it. The points (a,b) where strong duality holds are precisely the points where the optimal value function f equals its convex relaxation and the duality gap may be written as $\operatorname{gap}(f) = f - \check{f}$. On what follows, equation 3.19 will be extensively use. In particular, we will always denote the optimal dual value by $\check{f}$.

Since $\check{f}$ is convex and lower semi-continuous, it can be written as the supremum of the affine functions that are everywhere below it, from Theorem 2.39. If strong duality holds at a point (α,β), $\check{f}(\alpha,\beta) = f(\alpha,\beta)$, and the dual problem's optimal Lagrange multipliers represent cuts to f that are *tight* at the point (α,β).

Theorem 3.10 (Cuts from strong duality)**.** *Let f be an optimal value function with the form* (3.9) *and (α,β) a point where strong duality holds, that is, $\check{f}(\alpha,\beta) = f(\alpha,\beta)$. Then for all (a,b),*

$$f(a,b) \geqslant f(\alpha,\beta) - \langle \lambda^*_{\alpha,\beta}, a - \alpha \rangle - \langle \nu^*_{\alpha,\beta}, b - \beta \rangle.$$

Proof. By strong duality and the definition of the dual problem,

$$\begin{aligned} f(\alpha,\beta) = \check{f}(\alpha,\beta) &= d_{\alpha,\beta}(\lambda^*_{\alpha,\beta}, \nu^*_{\alpha,\beta}) \\ &\leqslant c(x) + \langle \lambda^*_{\alpha,\beta}, g(x) - \alpha \rangle + \langle \nu^*_{\alpha,\beta}, h(x) - \beta \rangle. \end{aligned}$$

Taking a point x that is feasible for the problem represented by $f(a,b)$, we have $g(x) \leqslant a$ and $h(x) \leqslant b$. Therefore, it follows from $\lambda^*_{\alpha,\beta} \geqslant_{K^*} 0$ that

$$f(\alpha,\beta) \leqslant c(x) + \langle \lambda^*_{\alpha,\beta}, a - \alpha \rangle + \langle \nu^*_{\alpha,\beta}, b - \beta \rangle.$$

Taking the infimum on the feasible set for $f(a,b)$ we get that

$$f(\alpha,\beta) \leqslant f(a,b) + \langle \lambda^*_{\alpha,\beta}, a - \alpha \rangle + \langle \nu^*_{\alpha,\beta}, b - \beta \rangle.$$

The result follows from reorganizing this expression. □

The cut constructed in this theorem is a global minimizer of the function f. Hence, it can be used to estimate how much the optimal value f varies as the parameters (a,b) are changed. This argument can be made rigorous by relating the optimal Lagrange multipliers to the derivatives of f.

Theorem 3.11. *Let f be an optimal value function with the form* (3.9) *and assume that f is differentiable at (α,β). Then,*

$$\nabla_a f(\alpha,\beta) = -\lambda^*_{\alpha,\beta}, \qquad \nabla_b f(\alpha,\beta) = -\nu^*_{\alpha,\beta}.$$

Proof. Let v be a unit vector and $t \in \mathbb{R}$. Theorem 3.10 implies that

$$f(\alpha + tv, \beta) \geqslant f(\alpha,\beta) - \langle \lambda^*_{\alpha,\beta}, tv \rangle.$$

If $t > 0$, this expression may be reorganized as

$$\frac{f(\alpha + tv, \beta) - f(\alpha,\beta)}{t} \geqslant \langle -\lambda^*_{\alpha,\beta}, v \rangle$$

and taking the limit as $t \to 0$ we get that $\nabla_a f(\alpha,\beta) \geqslant \langle -\lambda^*_{\alpha,\beta}, v \rangle$. Similarly, if $t < 0$, we have the opposite inequality

$$\frac{f(\alpha + tv, \beta) - f(\alpha,\beta)}{t} \leqslant \langle -\lambda^*_{\alpha,\beta}, v \rangle$$

implying that in the limit as $t \to 0$, $\nabla_a f(\alpha,\beta) \leqslant \langle -\lambda^*_{\alpha,\beta}, v \rangle$. Taking both these results together we get to the desired result,

$$\nabla_a f(\alpha,\beta) = \langle -\lambda^*_{\alpha,\beta}, v \rangle.$$

The proof for the other multipliers is equivalent. □

Notice that the cut constructed in Theorem 3.10 is equal to the original function on the point (α, β). This is a consequence of the strong duality assumption, which is unusual to hold when the optimization problems considered are non-convex, since in these cases the optimal value function f is not necessarily convex. Hereupon, we study some methods to calculate cuts for mixed integer problems, which are problems with convex structure besides the fact that some variables are constrained to be integer, as in Theorem 3.9. These cuts will be constructed by applying Theorem 3.10 to some convex approximation to the problem in question. For convex problems, all these methods should be equivalent as long as f equals $\check{f}$.

3.2.2.1 Benders cuts

Suppose f is the optimal value function of a mixed integer linear program

$$\begin{aligned} f(b) = \min_{x,z} \quad & c(x,z) \\ \text{s.t.} \quad & Ax + Bz = b \\ & (x,z) \in X \\ & x \in \mathbb{R}^n, z \in \mathbb{Z}^k. \end{aligned}$$

A simple and generally inexpensive way to underapproximate f by a convex function f_c is by relaxing the integrality constraints to allow the decision variable z to take any real value.

$$\begin{aligned} f_c(b) = \min_{x,z} \quad & c(x,z) \\ \text{s.t.} \quad & Ax + Bz = b \\ & (x,z) \in X \\ & x \in \mathbb{R}^n, z \in \mathbb{R}^k \end{aligned}$$

This is called the *continuous relaxation* of f and is a convex function, by Theorem 3.6. Since the feasible set of the relaxed problem always contains that of the original problem, it satisfies $f_c \leqslant f$.

Since, for all b, the constraints of the problem represented by $f_c(b)$ are all linear, Slater's condition (Theorem 3.4) says that strong duality holds for it whenever f_c is not infinite. Then, we can use Theorem 3.10 to calculate cuts for f_c. Given a right-hand side β, let ν_B be an optimal Lagrange multiplier for the problem $f_c(\beta)$. Then

$$f_c(b) \geqslant f_c(\beta) - \langle \mu_B, b - \beta \rangle, \ \forall b.$$

Using that f is always greater than f_c, we can apply this same cut for f and obtain

$$f(b) \geqslant f_c(\beta) - \langle \mu_B, b - \beta \rangle, \ \forall b. \tag{3.20}$$

This is called a *Benders cut* for f.

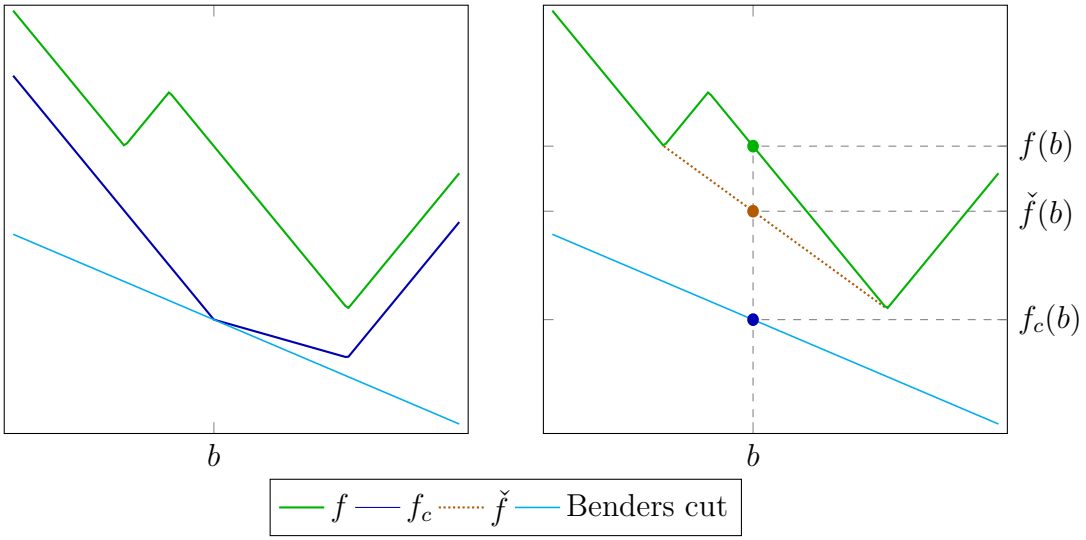


Figure 3.1: A non-convex optimal value function f, its relaxation f_c and a Benders cut.

Notice that although the Benders cuts are valid for f, they are not necessarily tight at any given point, since it is possible that $f(b) > f_c(b)$ and the cut obtained is everywhere strictly below f. This is exemplified in Figure 3.1.

The procedure to calculate Benders cuts can be generalized to the optimal value functions of any convex problem with integrality constraints. That is, suppose

$$\begin{aligned} f(a,b) = \min_x \quad & c(x) \\ \text{s.t.} \quad & g(x) \leqslant_K a \\ & h(x) = b \\ & x \in X \\ & x \in \mathbb{R}^n \times \mathbb{Z}^k \end{aligned} \tag{3.21}$$

where c is convex, g is K-convex, h is affine and X is convex. We can again define a continuous relaxation of f by removing the integrality constraints.

Definition 3.12 (Continuous relaxation)**.** The *continuous relaxation* of an optimal value function as in Equation (3.21) is the optimal value function

$$\begin{aligned} f_c(a,b) = \min_x \quad & c(x) \\ \text{s.t.} \quad & g(x) \leqslant_K a \\ & h(x) = b \\ & x \in X \\ & x \in \mathbb{R}^n \times \mathbb{R}^k. \end{aligned}$$

By Theorem 3.6, the continuous relaxation is a convex function and, by definition, it always satisfies

$$f_c \leqslant \check{f} \leqslant f. \tag{3.22}$$

Since f_c is convex, it is easier for strong duality to hold at a given point (α, β) in view of Equation (3.19). A Benders cut for f is again a cut for the continuous relaxation f_c applied to f.

Definition 3.13 (Benders cut)**.** Let f be an optimal value function as in Equation (3.21). Calculating a *Benders cut* for f at a point (α, β) consists in solving the convex relaxation $f_c(\alpha, \beta)$ to get optimal Lagrange multipliers λ_B and ν_B. Assuming that strong duality holds, these Lagrange multipliers satisfy

$$f_c(a, b) \geqslant f_c(\alpha, \beta) - \langle \lambda^*_{\alpha,\beta}, a - \alpha \rangle - \langle \nu^*_{\alpha,\beta}, b - \beta \rangle, \ \forall (a, b)$$

by virtue of Theorem 3.10. Since $f \geqslant f_c$, this cut can be applied to f resulting in

$$f(a, b) \geqslant f_c(\alpha, \beta) - \langle \lambda^*_{\alpha,\beta}, a - \alpha \rangle - \langle \nu^*_{\alpha,\beta}, b - \beta \rangle, \ \forall (a, b).$$

As discussed before, these cuts in most cases do not touch the graph f at any given point. Nevertheless, calculating Benders cuts consists in a way to find under-approximations to a non-convex optimal value function f whose computational cost is much less expensive than the other methods we will present.

3.2.2.2 Strengthened Benders cuts

After calculating a Benders cut for an optimal value function f, we can use the Lagrangian relaxation of f to *strengthen* this cut. That is, to find another cut that is parallel to the Benders one but that is guaranteed to be tight for $\check{f}$. We begin by showing the procedure to strengthen an arbitrary valid cut for f. Later, in Definition 3.14, we apply this method to Benders cuts to get what we call a *strengthened Benders cut* to f.

Suppose f is the optimal value function

$$\begin{aligned} f(a, b) = \ & \min_{x} && c(x) \\ & \text{s.t.} && g(x) \leqslant_K a \\ & && h(x) = b \\ & && x \in X \end{aligned}$$

and that we have a valid cut for f given by

$$f(a, b) \geqslant \gamma - \langle \lambda, a - \alpha \rangle - \langle \nu, b - \beta \rangle, \ \forall (a, b).$$

Knowing that a valid cut exists, we can define the set

$$C_{\lambda,\nu} = \left\{ q \mid f(a, b) \geqslant q - \langle \lambda, a - \alpha \rangle - \langle \nu, b - \beta \rangle, \ \forall (a, b) \right\}$$

of all intercepts with the same inclination that still give a valid cut. As we assumed f to be proper, this is a bounded subset of $\mathbb{R}$, because if we choose any parameters $(a, b) \in \operatorname{dom}(f)$, it holds for all $q \in C_{\lambda,\nu}$ that

$$q \leqslant f(a, b) + \langle \lambda, a - \alpha \rangle + \langle \nu, b - \beta \rangle.$$

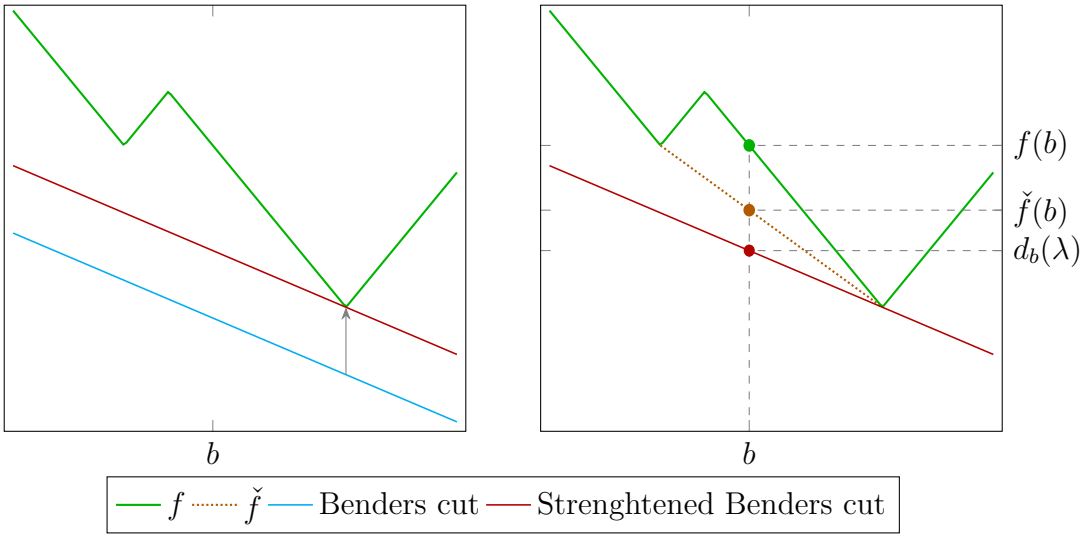


Figure 3.2: The strengthened Benders cut is parallel to the Benders cut but tight for $\check{f}$ at some point.

This means that the supremum of $C_{\lambda,\nu}$ is finite. Denoting by $q^\star = \sup C_{\lambda,\nu}$, we have that this is the largest intercept such that a cut with inclination (λ, ν) is still valid for f. Notice that this cut must be tight for $\check{f}$, by maximality.

The optimal intercept q^* can be calculated using the Lagrangian relaxation of $f(\alpha, \beta)$. First notice that for all (a, b),

$$q^* \leqslant f(a,b) + \langle \lambda, a - \alpha \rangle + \langle \nu, b - \beta \rangle = \begin{array}{ll} \min\limits_{x} & c(x) + \langle \lambda, a - \alpha \rangle + \langle \nu, b - \beta \rangle \\ \text{s.t.} & g(x) \leqslant_K a \\ & h(x) = b \\ & x \in X. \end{array}$$

The right-hand side of the above inequality is formulation 3.18 for the dual problem of an optimal value function on (α, β). Therefore the Lagrangian relaxation is an upper bound for q_*, that is,

$$q^* \leqslant d_{\alpha,\beta}(\lambda, \nu).$$

Now we proceed to show that this is in fact an equality.

Consider the usual definition of the Lagrangian relaxation,

$$d_{\alpha,\beta}(\lambda, \nu) = \inf_{x \in X} c(x) + \langle \lambda, g(x) - \alpha \rangle + \langle \nu, h(x) - \beta \rangle.$$

By rearranging all terms that are independent of the variable x to the left-hand side,

$$d_{\alpha,\beta}(\lambda, \nu) - \langle \lambda, -\alpha \rangle - \langle \nu, -\beta \rangle = \inf_{x \in X} c(x) + \langle \lambda, g(x) \rangle + \langle \nu, h(x) \rangle.$$

For any parameters (a, b), we can sum $\langle \lambda, a \rangle + \langle \nu, b \rangle$ to both sides and reorganize the inner products to obtain

$$\begin{aligned} d_{\alpha,\beta}(\lambda, \nu) - \langle \lambda, a - \alpha \rangle - \langle \nu, b - \beta \rangle &= \inf_{x \in X} c(x) + \langle \lambda, g(x) - a \rangle + \langle \nu, h(x) - b \rangle \\ &= d_{a,b}(\lambda, \nu). \end{aligned}$$

We can maximize the right-hand side over all possible Lagrange multipliers and obtain a valid cut for the dual function

$$d_{\alpha,\beta}(\lambda, \nu) - \langle \lambda, a - \alpha \rangle - \langle \nu, b - \beta \rangle \leqslant \max_{\lambda,\nu} d_{a,b}(\lambda, \nu) = \check{f}(a, b) \leqslant f(a, b).$$

Since q^* was the supremum of all intercepts for which the cut was still valid, this means that $q^* \geqslant d_{\alpha,\beta}(\lambda, \nu)$. Therefore, we conclude that the optimal intercept is attained on the Lagrangian relaxation

$$q^* = d_{\alpha,\beta}(\lambda, \nu).$$

The above discussion means that there is a point $(\tilde{a}, \tilde{b})$ for which (λ, ν) is dual optimal and the cut obtained is *tight* for the convex relaxation of f,

$$d_{\alpha,\beta}(\lambda, \nu) - \langle \lambda, \tilde{a} - \alpha \rangle - \langle \nu, \tilde{b} - \beta \rangle = d_{\tilde{a},\tilde{b}}(\lambda, \nu) = \check{f}(a, b).$$

Notice that this point $(\tilde{a}, \tilde{b})$ may differ from the original point (α, β). Therefore, we can only guarantee that a strengthened cut is tight somewhere but not on the point we were originally considering.

Definition 3.14 (Strenghtened Benders cut)**.** Calculating a *strengthened Benders cut* for an optimal value function f at (α, β) consists of two steps:

1. Calculating a Benders cut for the relaxation $f_c(\alpha, \beta)$ to get the Benders optimal Lagrange multipliers (λ_B, ν_B);
2. Solving the Lagrangian relaxation

$$d_{\alpha,\beta}(\lambda_B, \nu_B) = \inf_{x \in X} c(x) + \langle \lambda_B, g(x) - \alpha \rangle + \langle \nu_B, h(x) - \beta \rangle$$

   to get a new intercept with the same inclination.

The strengthened Benders cut for f is then

$$f(a, b) \geqslant d_{\alpha,\beta}(\lambda_B, \nu_B) - \langle \lambda_B, a - \alpha \rangle - \langle \nu_B, b - \beta \rangle.$$

Finding a strengthened Benders cut for an optimal value function f requires solving a convex optimization problem to find the Lagrange multipliers and, afterwards, a non-convex problem to find the optimal value of the Lagrangian relaxation. In virtue of this second step, strengthening a Benders cut can take reasonably longer than only finding a Benders cut. Nevertheless, the fact that strengthened cuts are tight for the convex relaxation $\check{f}$ makes them much more effective. This difference will be crucial on Chapter 7 where we will calculate cuts for stochastic programs.

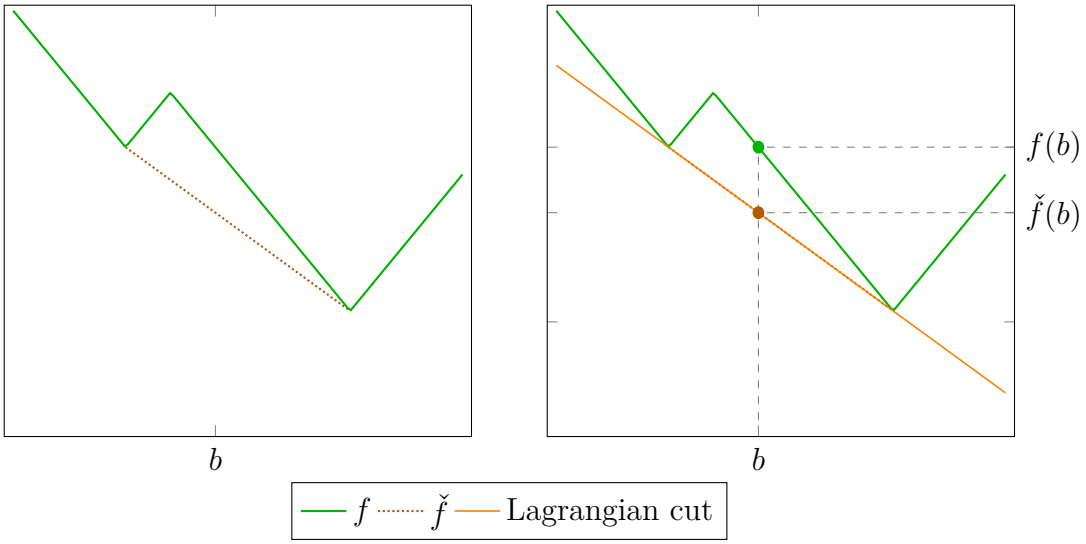


Figure 3.3: A Lagrangian cut is as tight as possible at the chosen point.

3.2.2.3 Lagrangian cuts

When calculating a strengthened Benders cut, we cannot choose in which point it will be tight. In the following, we show how, given a point (α, β), we can use the fact that $\check{f}(\alpha, \beta)$ equals the dual optimal value to calculate a cut for $\check{f}$ that is tight at (α, β). This is called a *Lagrangian cut*.

Consider an optimal value function f with the same form as before and suppose that at the point (α, β) its dual optimal is attained. That is, by solving the dual problem $\max d_{\alpha,\beta}(\lambda, \nu)$ we get optimal Lagrange multipliers $(\lambda^*_{\alpha,\beta}, \nu^*_{\alpha,\beta})$ satisfying

$$\check{f}(\alpha, \beta) = \inf_{x \in X} c(x) + \langle \lambda^*_{\alpha,\beta}, g(x) - \alpha \rangle + \langle \nu^*_{\alpha,\beta}, h(x) - \beta \rangle.$$

The terms independent of the variable x can be all passed to the left-hand side,

$$\check{f}(\alpha, \beta) - \langle \lambda^*_{\alpha,\beta}, -\alpha \rangle - \langle \nu^*_{\alpha,\beta}, -\beta \rangle = \inf_{x \in X} c(x) + \langle \lambda^*_{\alpha,\beta}, g(x) \rangle + \langle \nu^*_{\alpha,\beta}, h(x) \rangle$$

and by summing $\langle \lambda^*_{\alpha,\beta}, a \rangle + \langle \nu^*_{\alpha,\beta}, b \rangle$ to both sides and reorganizing the inner products we get

$$\check{f}(\alpha, \beta) - \langle \lambda^*_{\alpha,\beta}, a - \alpha \rangle - \langle \nu^*_{\alpha,\beta}, b - \beta \rangle = \inf_{x \in X} c(x) + \langle \lambda^*_{\alpha,\beta}, g(x) - a \rangle + \langle \nu^*_{\alpha,\beta}, h(x) - b \rangle.$$

By maximizing the right-hand side with respect to all feasible Lagrange multipliers, we retrieve the dual optimal value function evaluated at (a, b),

$$\begin{aligned} \check{f}(\alpha, \beta) - \langle \lambda^*_{\alpha,\beta}, a - \alpha \rangle - \langle \nu^*_{\alpha,\beta}, b - \beta \rangle &\leqslant \sup_{\substack{\lambda \in K^*, \\ \nu}} \inf_{x \in X} c(x) + \langle \lambda, g(x) - a \rangle + \langle \nu, h(x) - b \rangle \\ &= \check{f}(a, b) \leqslant f(a, b). \end{aligned}$$

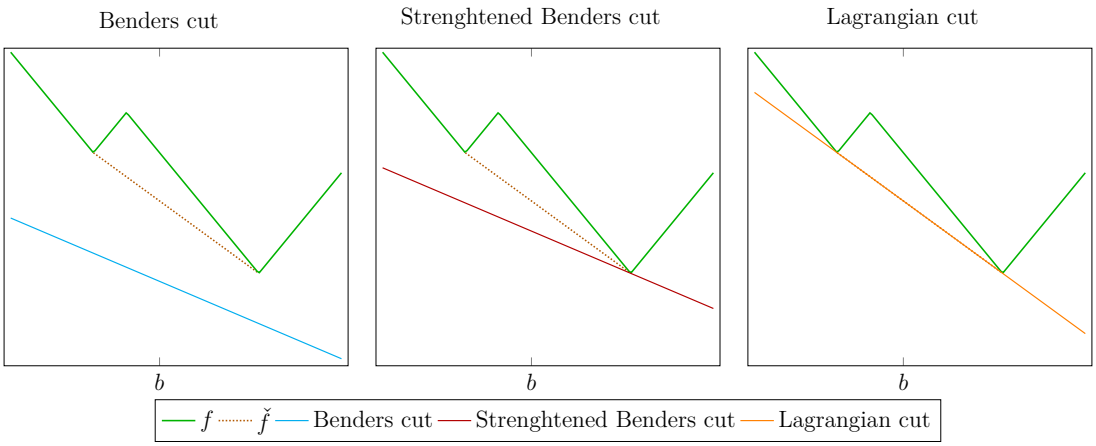


Figure 3.4: The three types of cuts calculated for the same optimal value function f at a chosen point b.

From this, we get a cut for f that is tight at its convex relaxation at the chosen point (α, β).

Definition 3.15 (Lagrangian Cut)**.** Calculating a *Lagrangian cut* for an optimal value function f at a point (α, β) consists of solving the dual problem

$$\check{f}(\alpha, \beta) = \sup_{\substack{\lambda \in K^*, \\ \nu}} \inf_{x \in X} c(x) + \langle \lambda, g(x) - a \rangle + \langle \nu, h(x) - b \rangle$$

to get optimal Lagrange multipliers $(\lambda^*_{\alpha,\beta}, \nu^*_{\alpha,\beta})$. The *Lagrangian cut* is the affine function

$$f(a, b) \geqslant \check{f}(\alpha, \beta) - \langle \lambda^*_{\alpha,\beta}, a - \alpha \rangle - \langle \nu^*_{\alpha,\beta}, b - \beta \rangle, \ \forall (a, b),$$

which is tight for $\check{f}$ at the chosen point (α, β).

Although the Lagrangian cuts are the most precise type of cut, their computational cost is, in general, extremely expensive. Since, for fixed (α, β), the Lagrangian relaxation $d_{\alpha,\beta}$ is a concave function, maximizing it is a convex optimization problem. However, each evaluation $d_{\alpha,\beta}(\lambda, \nu)$ requires solving a non-convex optimization problem and any iterative method to solve the dual problem will require successive evaluations of the Lagrangian relaxation.

3.2.2.4 Comparison between cut types

If the optimization problems represented by f are all convex, the continuous relaxation f_c from Definition (3.12) and the convex relaxation $\check{f}$ are both equal to f,

$$f = f_c = \check{f}.$$

Therefore, for convex problems, the three types of cuts previously described are equal and equivalent to the cut obtained in Theorem 3.10. Particularly, they

are all tight for f at the point in which they are being calculated. This means that there is no advantage in using the more computationally expensive cuts to approximate the optimal value function of a convex problem.

When the problems represented by f are not convex, there is a trade-off between computational cost and precision for each type of cut. This is summarized in Table 3.1.

Table 3.1: Comparison between the different kinds of cuts for non-convex problems.

Cut type	Computational cost	Tight for $\check{f}$?
Benders	Solve one convex problem	Not in general
Strengthened Benders	Solve one convex and one non-convex problem	Yes, at an arbitrary point
Lagrangian	Repeatedly solve a non-convex problem	Yes, at the chosen point

3.3 Multi-stage optimization

Many real world problems require taking not a single decision but a series of sequential decisions over time. Although these problems can be modeled as a single optimization problem, this is, in general, not computationally efficient. In this section we formulate a multi-stage optimization problem as a control problem; that is, as a series of optimization problems linked together.

Let's consider a concrete example of a two-stage problem which will be useful to fix notation and give some intuition before we properly define two-stage problems in Definition 3.16. On what follows we will model a simple hydrothermal scheduling problem for a two months schedule.

In a hydrothermal schedule, energy generation is has two sources: hydroelectric and thermoelectric power plants. Our problem is to minimize the cost of energy generation subject to attending the energy demand and respecting the physical constraints such as non-negativity and maximum energy generation. We will begin by considering the schedule for a single month.

Let us denote by x_1 the decision variable saying how much energy was came from hydroelectric power plants and by u_1 the decision variable saying how much energy came from thermoelectric power plants. Let c_H denote the cost of producing one unit of energy at a hydroelectric and c_T denote the cost of producing one unit of energy at a thermoelectric power plant. The demand will be denoted by d_1, the maximum thermoelectric generation by M_T and the maximum hydroelectric generation, which equals the total stored energy at the

hydroelectric power plants, by M_H. This optimization problem is

$$\begin{aligned} \min_{x_1,u_1} \quad & c_H x_1 + c_T u_1 \\ \text{s.t.} \quad & 0 \leqslant x_1 \leqslant M_H \\ & 0 \leqslant u_1 \leqslant M_T \\ & x_1 + u_1 = d_1. \end{aligned}$$

To extend this problem to two-stages, we must notice that the total stored energy on the second stage is no longer M_H but $M_h - x_1$. Writing x_2 and u_2 for the second stage decision variables and d_2 for the demand at the second stage, the two month decision schedule is

$$\begin{aligned} \min_{x_1,u_1,x_2,u_2} \quad & c_H x_1 + c_T u_1 + c_H x_2 + c_T u_2 \\ \text{s.t.} \quad & 0 \leqslant x_1 \leqslant M_H \\ & 0 \leqslant u_1 \leqslant M_T \\ & x_1 + u_1 = d_1 \\ & 0 \leqslant x_2 \leqslant M_T - x_1 \\ & 0 \leqslant u_2 \leqslant M_T \\ & x_2 + u_2 = d_2. \end{aligned}$$

By noting that most of the variables from the first and second stage are decoupled, this problem can be rewritten as

$$\begin{aligned} \min_{x_1,u_1} \quad & c_H x_1 + c_T u_1 + \\ \text{s.t.} \quad & 0 \leqslant x_1 \leqslant M_H \\ & 0 \leqslant u_1 \leqslant M_T \\ & x_1 + u_1 = d_1 \end{aligned} \qquad \begin{aligned} \min_{x_2,u_2} \quad & c_H x_2 + c_T u_2 \\ \text{s.t.} \quad & 0 \leqslant x_2 \leqslant M_T - x_1 \\ & 0 \leqslant u_2 \leqslant M_T \\ & x_2 + u_2 = d_2 \end{aligned}$$

The second minimum can be represented as an optimal value function that gives the optimal cost of the second stage depending on how much stored energy is left from the first stage,

$$\begin{aligned} Q(x_1) = \min_{x_2,u_2} \quad & c_H x_2 + c_T u_2 \\ \text{s.t.} \quad & 0 \leqslant x_2 \leqslant M_T - x_1 \\ & 0 \leqslant u_2 \leqslant M_T \\ & x_2 + u_2 = d_2, \end{aligned}$$

and the original problem can be expressed as depending only on the first stage variables as

$$\begin{aligned} \min_{x_1,u_1} \quad & c_H x_1 + c_T u_1 + Q(x_1) \\ \text{s.t.} \quad & 0 \leqslant x_1 \leqslant M_H \\ & 0 \leqslant u_1 \leqslant M_T \\ & x_1 + u_1 = d_1. \end{aligned}$$

The function Q is called the problem's *cost-to-go* and depends, directly, only on the first stage's hydroelectric generation x_1.

Mimicking this example, we get to the general definition of a two-stage problem.

3.3.1 Two-stage problems

Definition 3.16 (Two-stage optimization problem)**.** A *two-stage optimization problem* is an optimization problem with the form

$$\begin{array}{ll} \min\limits_{x_1,u_1} & c_1(x_1,u_1)+Q(x_1) \\ \text{s.t.} & (x_1,u_1)\in X_1 \end{array}$$

where Q is the optimal value function

$$\begin{array}{ll} \min\limits_{x_2,u_2} & c_2(x_2,u_2) \\ \text{s.t.} & (x_1,x_2,u_2)\in X_2. \end{array}$$

Letting $t=1,2$, the variables x_t are called the problem's *state variables* because they describe the problems state as one goes from a stage to another and the variables u_t are the problem's *local or control variables* since they are only "seem" in their respective stages. The functions c_t are the problem's *present costs* and the optimal value function Q is the problem's *cost-to-go*.

When solving a two-stage problem, each evaluation of the objective function requires solving another optimization problem to evaluate $Q(x_1)$. Since most computational methods to solve optimization problems iteratively evaluate the objective function, this can become inefficient.

A solution to this problem is to approximate Q by cuts using the methods of Section 3.2.2 and incorporate these approximations as restrictions of the first-stage problem. This will give an underapproximation to the original problem that becomes increasingly more precise as more cuts are added.

Let us start with an example using a single cut

$$Q(x_1)\geqslant q-\langle\lambda,x_1\rangle$$

and denote by p^* the problem's solution,

$$\begin{array}{rlll} p^*= & \min\limits_{x_1,u_1} & c_1(x_1,u_1)+Q(x_1) & = \min\limits_{x_1,u_1,\alpha} \quad c_1(x_1,u_1)+\alpha \\ & \text{s.t.} & (x_1,u_1)\in X_1 & \text{s.t.} \quad (x_1,u_1)\in X_1 \\ & & & \qquad\quad Q(x_1)\leqslant\alpha, \end{array}$$

where the second equality comes from passing the function Q to epigraph form as in Equation (3.14). Since the cut is everywhere less than Q,

$$Q(x_1)\leqslant\alpha \implies q-\langle\lambda,x_1\rangle\leqslant\alpha$$

and we can form another less constrained problem

$$\begin{array}{rll} \tilde{p}= & \min\limits_{x_1,u_1,\alpha} & c_1(x_1,u_1)+\alpha \\ & \text{s.t.} & (x_1,u_1)\in X_1 \\ & & q-\langle\lambda,x_1\rangle\leqslant\alpha. \end{array}$$

It always holds that $\tilde{p} \leqslant p^*$.

If instead of a single cut we have a family of cuts satisfying

$$Q(x_1) \geqslant q_i - \langle \lambda_i, x_1 \rangle, \ \forall i$$

we can take its maximum over any finite set of k cuts to obtain a polyhedral underapproximation to Q,

$$\mathfrak{Q}^{(k)}(x_1) = \max_{i=1,\dots,k} \big\{ q_i - \langle \lambda_i, x_1 \rangle \big\}, \tag{3.23}$$

and produce the optimization problem

$$\begin{array}{rcl} p^{(k)} = & \min\limits_{x_1,u_1,\alpha} & c_1(x_1,u_1) + \alpha \\ & \text{s.t.} & (x_1,u_1) \in X_1 \\ & & \mathfrak{Q}^{(k)}(x_1) \leqslant \alpha, \end{array} \quad = \quad \begin{array}{ll} \min\limits_{x_1,u_1,\alpha} & c_1(x_1,u_1) + \alpha \\ \text{s.t.} & (x_1,u_1) \in X_1 \\ & q_i - \langle \lambda_i, x_1 \rangle \leqslant \alpha \text{ for } i = 1,\dots,k. \end{array}$$

Notice that since $\mathfrak{Q}^{(k)} \leqslant \mathfrak{Q}^{(k+1)} \leqslant Q$, it also holds that $p^{(k)} \leqslant p^{(k+1)} \leqslant p^*$. In particular, if Q is itself polyhedral, as is the case for the optimal value functions of linear problems (Theorem 3.7), there exists a finite family of cuts such that the approximation $\mathfrak{Q}$ equals Q.

3.3.2 Multi-stage problems

In the Definition 3.16, it is possible that the cost-to-go is itself the optimal value function of another two-stage problem. In fact, we could have T linked optimization problems.

Definition 3.17 (Multi-stage optimization problem)**.** An optimization problem with T*-stages* is a problem of the form

$$\begin{array}{ll} \min\limits_{x_1,u_1,\dots,x_n,y_n} & \sum_{i=1}^{T} c_t(x_t,u_t) \\ \text{s.t.} & (x_1,u_1) \in X_1 \\ & (x_{t-1},x_t,u_t) \in X_i, \text{ for } i = 1,\dots,T. \end{array}$$

Alternatively, a T-stage problem can be written in *dynamic programming formulation* as

$$\begin{array}{rcl} Q_t(x_{t-1}) = & \min\limits_{x_t,u_t} & c_t(x_t,u_t) + Q_{t+1}(x_t) \\ & \text{s.t.} & (x_{t-1},x_t,u_t) \in X_t, \end{array} \tag{3.24}$$

where $Q_t = 0$ when t equals T and we assume that x_0 is constant. In this formulation, the function Q_t is the cost-to-go function from the $(t-1)$-th stage to the t-th stage.

For large problems, it is in general computationally more efficient to solve a multi-stage problem using formulation (3.24). Notice also that, although the first stage problem looks exactly the same to the two-stage formulation in Definition 3.16, in that case the cost-to-go function Q was an arbitrary optimal value function while in the multi-stage case it is the optimal value function of a problem with $T-1$ stages.

3.4 Optimization under uncertainty

When modeling the hydrothermal scheduling example in the beginning of Section 3.3, there is a detail that we glossed over: *we cannot predict the future with certainty*. More specifically, when passing from the first to the second stage, the total amount of stored energy should not be $M_H - x_1$, the initial amount of energy minus how much hydroelectric energy was produced.

To make a more realistic model, it is necessary to consider how much rain occurred between the two stages producing a certain gain on the total stored energy. That is, if we denote by ξ the *random variable* representing the energy gain that occurred between both stages because of rainfall, the total stored energy at the second stage is $M_H - x_1 + \xi$ and the problem can be written as

$$\begin{array}{ll} \min\limits_{x_1,u_1} & c_H x_1 + c_T u_1 + Q(x_1,\xi) \\ \text{s.t.} & 0 \leqslant x_1 \leqslant M_H \\ & 0 \leqslant u_1 \leqslant M_T \\ & x_1 + u_1 = d_1 \end{array} \quad , \quad Q(x_1,\xi) = \begin{array}{ll} \min\limits_{x_2,u_2} & c_H x_2 + c_T u_2 \\ \text{s.t.} & 0 \leqslant x_2 \leqslant M_T - x_1 + \xi \\ & 0 \leqslant u_2 \leqslant M_T \\ & x_2 + u_2 = d_2. \end{array}$$

A *stochastic program* is a optimization problem where the cost-to-go functions depends on a random variable. This makes the first-stage optimal value itself a random variable

$$\begin{array}{rll} p^*(\xi) = & \min\limits_{x_1,u_1} & c_1(x_1,u_1) + Q(x_1,\xi) \\ & \text{s.t.} & (x_1,u_1) \in X_1. \end{array} \tag{3.25}$$

To find a solution to this problem, we would need to know in advance what the random variable's realization will be. In practice, it is better to minimize the cost-to-go's average, $\mathbb{E}\left[Q(x_1,\xi)\right]$, over all possible scenarios or even another risk measure applied to the cost-to-go Q.

In Section 3.4.1 we will fix some notation that will make the presentation of random functions cleaner. Then, in Section 3.4.2 we will present *risk neutral stochastic programs*, were we use expected cost-to-go to solve optimization problems. In Section 3.4.3, we introduce *coherent risk measures* and show how can they be used in place of the expected value to model *risk averse stochastic programs*.

In this section, we will only cover the necessary concepts to motivate and explain the later results of Chapters 5, 6 and 7. A much more thorough explanation of stochastic programming can be found in the book [Shapiro et al., 2014].

3.4.1 Notation for random functions

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a random variable ξ and a function $Q(x,\xi)$. We can think of Q as a function $x \mapsto Q(x,\xi)$ that for each value x returns a random variable. In this case, we can speak of the expected value of Q as the function

$$\mathbb{E}\left[Q\right](x) = \mathbb{E}^{\xi}\left[Q(x,\xi)\right], \tag{3.26}$$

which is a (non-random) function of only x.

Alternatively, we could think of Q as random variable on the space of functions. That is, for each realization of ξ, we have a function $Q_\xi(x) = Q(x, \xi)$. In this case, we define the convex relaxation of Q as acting on the function Q_ξ for each realization of ξ,

$$\check{Q}(x, \xi) = \check{Q}_\xi(x) \tag{3.27}$$

and, similarly, the Fenchel conjugate of Q as the function

$$Q^*(x, \xi) = {Q_\xi}^*(x). \tag{3.28}$$

During the remainder of this work, we will many times use an operator acting on a random function as acting on only one of its variables. If this operator acts on random variables, the definition will always be the same as that of equation (3.26) and if the operator acts on functions, the definition will always be the same as that of equation (3.27).

It is important to notice that in general the order we take these operations do not commute. In fact, much of Chapter 5 can be thought as studying the difference between the functions $\mathbb{E}[\check{Q}]$ and $\widecheck{\mathbb{E}[Q]}$.

3.4.2 Risk-neutral optimization

In this section we formulate stochastic programs in a manner that is *neutral* towards risk. This is done by optimizing the expected valued of the cost-to-go function instead of considering the cost-to-go as a random variable function.

Definition 3.18 (Two-stage stochastic program)**.** A *two-stage stochastic program* is an optimization problem with the form

$$\begin{aligned} \min_{x_1, u_1} \quad & c_1(x_1, u_1) + \mathbb{E}[Q(x_1, \xi)] \\ \text{s.t.} \quad & (x_1, u_1) \in X_1 \end{aligned}$$

where $Q(x_1, \xi)$ is the optimization problem

$$\begin{aligned} \min_{x_2, u_2} \quad & c_2(\xi)(x_2, u_2) \\ \text{s.t.} \quad & (x_1, x_2, u_2) \in X_2(\xi). \end{aligned}$$

That is Q is a random optimal value function on x_1 with both the objective function c_2 and the feasible set depending on the realization of the random variable ξ.

The function $\mathbb{E}[Q]$ is called the problem's *expected cost-to-go*. It depends only on the first stage state variable x_1 and is, therefore, not random. Each possible outcome $\xi = \xi^i$ from the random variable is called a *scenario* for the problem. The function that takes $Q(\cdot, \xi^i)$ is the cost-to-go for scenario the ξ^i.

We now proceed to discuss the methods to evaluate the expected cost-to-go of a stochastic program and how to calculate valid cuts for it. For simplicity, we will assume that the uncertainty ξ has finite support. That is, there exists a finite number of scenarios $\xi_1, \ldots, \xi_N$ and numbers p_i such that

$$\mathbb{P}(\xi = \xi^i) = p_i.$$

Each scenario ξ^i stands for a objective function c_2^i and a feasible set X^i. To help unclutter notation, we we will sometimes write $Q^i(x_1) = Q(x_1, \xi^i)$ for the cost-to-go of scenario ξ^i. In this setting the expected cost-to-go can be written as a convex combination of the cost-to-go of each scenario

$$\mathbb{E}\left[Q(x_1, \xi)\right] = \sum_{i=1}^{N} p_i Q^i(x_1). \tag{3.29}$$

This sum can be evaluated via a *decomposed formulation* or via a *linked formulation*. On the following, we discuss the advantages and disadvantages of each one.

3.4.2.1 Decomposed formulation

Definition 3.19 (Decomposed formulation)**.** The *decomposed formulation* of the expected cost-to-go consists in solving the N optimization problems $Q_i(x_1)$ separately and, afterwards, evaluating the sum in Equation (3.29) to get the value of $\mathbb{E}\left[Q(x_1, \xi)\right]$.

The decomposed formulation can also be used to calculate cuts for the expected cost-to-go from cuts obtained for each scenario. To do this, suppose we calculated cuts at the point $x_1 = y$ for each scenario using any of the methods in Section 3.2.2. This gives us N affine functions satisfying

$$Q^i(x_1) \geqslant q^i - \langle \lambda^i, x_1 - y \rangle, \ \forall i.$$

By taking the averages of the q^i and λ^i, we get an *average cut* that is valid for the expected cost-to-go,

$$\sum_{i=1}^{N} p_i Q^i(x_1) \geqslant \sum_{i=1}^{N} p_i \left(q^i - \langle \lambda^i, x_1 - y \rangle\right) = \sum_{i=1}^{N} p_i q^i - \langle \sum_{i=1}^{N} p_i \lambda^i, x_1 - y \rangle.$$

Setting

$$\bar{q} = \sum_{i=1}^{N} p_i q^i, \qquad \bar{\lambda} = \sum_{i=1}^{N} p_i \lambda^i \tag{3.30}$$

we get an expression for the average cut as

$$\mathbb{E}\left[Q(x_1, \xi)\right] \geqslant \bar{q} - \langle \bar{\lambda}, x_1 - y \rangle. \tag{3.31}$$

A question that may arise is if tightness of the cuts is preserved by this procedure. For convex problems, the answer is in general affirmative. As we will

shortly see, this requires the cut to be tight for the cost-to-go of all scenarios at the same point, something that is normally not possible when the cost-to-go functions are non-convex.

Theorem 3.20 (Tightness of average cut)**.** *Suppose that we have cuts*

$$Q^i(x_1) \geqslant q^i - \langle \lambda^i, x_1 - y \rangle$$

that are tight for each Q^i at the point $x_1 = y$. That is, $q^i = Q^i(x_1)$. Then the average cut from Equation (3.30) *is also tight at $x_1 = y$,*

$$\bar{q} = \mathbb{E}\left[Q(y, \xi)\right].$$

Proof. By evaluating the average cut at $x_1 = y$, we get

$$\bar{q} = \sum_{i=1}^{N} p_i q^i = \sum_{i=1}^{N} p_i Q^i(x_1) = \mathbb{E}\left[Q(y, \xi)\right]. \qquad \square$$

An important thing to notice about this theorem is that it requires the cuts to be tight for the cost-to-go functions Q^i of each scenario, not only for their convex relaxations $\check{Q}^i$. In other words, the average cut is only guaranteed to be tight at y if strong duality holds for the cost-to-go functions of all scenarios.

When solving a non-convex stochastic program, this requirement is rather impractical because the best cuts we can calculate are only tight for $\check{Q}^i$. To see this, suppose that we have calculated a cut for each scenario that is tight on $\check{Q}^i$ at the point $x_1 = y$. This could be done via Lagrangian cuts (Definition 3.15), for example. Then we can use the fact that the convex relaxation is itself a *concave* function with respect to the cone of non-negative functions to obtain

$$\bar{q} = \sum_{i=1}^{N} p_i \check{Q}^i(y) = \mathbb{E}\left[\check{Q}(y, \xi)\right] \leqslant \mathbb{E}\left[\widetilde{Q(y, \xi)}\right].$$

This means that although in each scenario there was no *gap* between the cut and the convex relaxation of the cost-to-go, the process of taking the averages may produce a gap between the *average cut* and the expected cost-to-go.

3.4.2.2 Linked formulation

To calculate cuts that are tight for $\widetilde{\mathbb{E}\left[Q\right]}$, we must formulate the expected cost-to-go in such a way that all scenarios are *linked* in a single optimization problem. We begin by writing the decision variables of the scenario ξ^i as (x_2^i, u_2^i) and the cost-to-go as

$$\begin{aligned} Q^i(x_1) = \quad & \min_{x_2^i, u_2^i} \quad c_2^i(x_2^i, u_2^i) \\ & \text{s.t.} \quad (x_1, x_2^i, u_2^i) \in X^i. \end{aligned}$$

The expected cost-to-go the average of these optimization problems. Considering the decision variables to be different when the scenarios are different, the expect

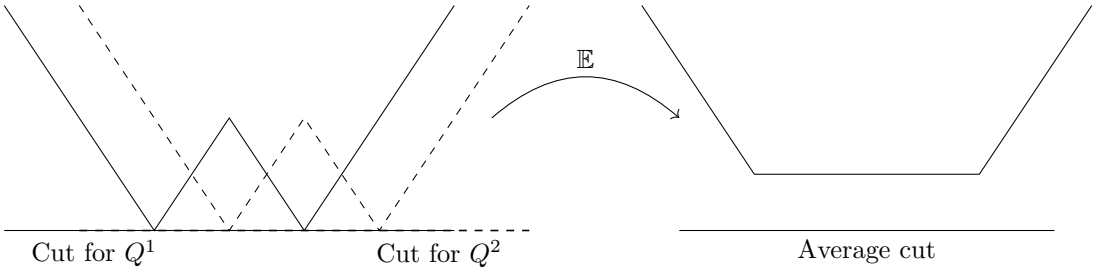


Figure 3.5: For non-convex problems, the average of cuts that are tight for each scenario may not be tight for the expected cost-to-go.

cost-to-go can be written as a single optimization with all the variables at the same time. This works because the sum of minima over different variables can always be written as the minimum of the sum over all variables,

$$\begin{aligned}\mathbb{E}\left[Q(x_1,\xi)\right] &= \sum_{i=1}^{N} p_i Q^i(x_1) = \sum_{i=1}^{N} p_i \left(\begin{array}{ll} \min\limits_{x_2^i,u_2^i} & c_2^i(x_2^i,u_2^i) \\ \text{s.t.} & (x_1,x_2^i,u_2^i)\in X^i \end{array} \right) \\ &= \min_{\substack{x_2^i,u_2^i \\ \text{for } i=1,\ldots,N}} \quad \sum_{i=1}^{N} p_i\, c_2^i(x_2^i,u_2^i) \\ &\qquad \text{s.t.} \quad (x_1,x_2^i,u_2^i)\in X^i, \text{ for } i=1,\ldots,N.\end{aligned} \tag{3.32}$$

This allows us to write the expected cost-to-go as a single optimal value function considers all scenarios at the same time.

Definition 3.21 (Linked Formulation)**.** The *linked formulation* for the expected cost-to-go consists of the optimization problem

$$\begin{aligned}\mathbb{E}\left[Q(x_1,\xi)\right] = \min_{\substack{x_2^i,u_2^i \\ \text{for } i=1,\ldots,N}} \quad & \sum_{i=1}^{N} p_i\, c_2^i(x_2^i,u_2^i) \\ \text{s.t.} \quad & (x_1,x_2^i,u_2^i)\in X^i, \text{ for } i=1,\ldots,N.\end{aligned}$$

To evaluate the expected cost-to-go using the decomposed formulation amounts to solving N separate optimization problems while the linked formulation requires solving a single problems with N times the number of variables. Since solving an optimization problems generally requires a computational cost that is above liner on the number of variables, the decomposed formulation is faster to calculate.

Nevertheless, the advantage of the linked formulation lies in the fact that it can be used to calculate better cuts when the expected cost-to-go is non-convex. This occurs because we are representing the expected cost-to-go as a single optimal value function and, therefore, strengthened Benders and Lagrangian cuts calculated for the linked formulation problem will always be tight for the convex relaxation $\widetilde{\mathbb{E}\left[Q\right]}$, while they were in general not tight when calculated via the decomposed formulation. As occurred for the different types of cuts in

Section 3.2.2, we again have a trade-off between computational cost and precision when choosing a formulation for the expected cost-to-go.

We close this section with an important remark regarding Benders cuts.As we will shortly see, there is no difference between taking the continuous relaxation of the linked formulation problem and making the linked formulation of the continuous relaxation for each scenario. As a consequence, calculating a Benders cut using the decomposed or the linked formulation is the same. This means that, for computational reasons, it is always better to calculate Benders cuts via the decomposed formulation.

Theorem 3.22 (Linked formulation of continuous relaxation)**.** *Suppose the cost-to-go for each scenario Q^i is a convex problem except for some integrality constraints. Then, the linked formulation satisfies*

$$\mathbb{E}\left[Q_c\right] = \mathbb{E}\left[Q\right]_c.$$

Proof. Instead of separating the variables in state and control, let us separate them in continuous and integer. That is, call y^i the continuous variables and z^i the integer variables in scenario ξ^i and write the cost-to-go and its continuous relaxation as

$$Q^i(x_1) = \begin{array}{ll} \min\limits_{y^i,z^i} & c_2^i(y^i,z^i) \\ \text{s.t.} & (x_1,y^i,z^i) \in X^i \\ & y^i \in \mathbb{R}^{n_i},\ z^i \in \mathbb{Z}^{k_i} \end{array}, \quad Q_c^i(x_1) = \begin{array}{ll} \min\limits_{y^i,z^i} & c_2^i(y^i,z^i) \\ \text{s.t.} & (x_1,y^i,z^i) \in X^i \\ & y^i \in \mathbb{R}^{n_i},\ z^i \in \mathbb{R}^{k_i}. \end{array}$$

The linked formulation for the expected cost-to-go is the optimal value function

$$\mathbb{E}\left[Q\right](x_1) = \begin{array}{ll} \min\limits_{y^i,z^i} & \sum_{i=1}^N p_i c_2^i(y^i,z^i) \\ \text{s.t.} & (x_1,y^i,z^i) \in X^i,\ \text{for } i = 1,\dots,N \\ & y^i \in \mathbb{R}^{n_i},\ z^i \in \mathbb{Z}^{k_i},\ \text{for } i = 1,\dots,N \end{array}$$

whose continuous relaxation is the optimal value function obtained by relaxing all integrality constraints

$$\mathbb{E}\left[Q\right]_c(x_1) = \begin{array}{ll} \min\limits_{y^i,z^i} & \sum_{i=1}^N p_i c_2^i(y^i,z^i) \\ \text{s.t.} & (x_1,y^i,z^i) \in X^i,\ \text{for} i = 1,\dots,N \\ & y^i \in \mathbb{R}^{n_i},\ z^i \in \mathbb{R}^{k_i},\ \text{for} i = 1,\dots,N. \end{array}$$

But looking at it, we see that it is the same optimal value function we would obtain by writing the linked formulation for the continuous relaxations of the cost-to-go Q^i for each scenario. □

Corollary 3.22.1. It is equivalent to calculate a Benders cut using the linked or the decomposed formulation.

Proof. The functions Q_c^i are convex, meaning that we can calculate cuts that are tight for them at a point x_1. By Theorem 3.20, these cuts are tight for $\mathbb{E}[Q_c]$ at x_1. Using Theorem 3.22, these cuts will also be tight at x_1 for the continuous relaxation of the linked formulation $\mathbb{E}[Q]_c$, meaning that they are Benders cuts for the linked formulation. □

3.4.3 Risk-averse optimization

In this section we extend the notion of stochastic program to also encompass risk measures different from the average value applied to the cost-to-go functions. We will introduce *coherent risk measures*, firstly introduced in [Artzner et al., 1999], which are functionals over random variables that generalize concepts such as the average and the supremum still maintaining most of their important properties.

A *risk-averse stochastic program*, definition 3.27, is the same as a risk neutral stochastic program where we exchange all the averages by another coherent risk measure.

To motivate this new concept, consider again the hydrothermal scheduling example viewed as a stochastic program. In this case the cost-to-go is a random function depending on the rainfall ξ between the stages and the first-stage problem is

$$\begin{aligned} \min_{x_1,u_1} \quad & c_H x_1 + c_T u_1 + \mathbb{E}[Q(x_1,\xi)] \\ \text{s.t.} \quad & 0 \leqslant x_1 \leqslant M_H \\ & 0 \leqslant u_1 \leqslant M_T \\ & x_1 + u_1 = d_1. \end{aligned}$$

This problem makes the first stage decision considering the expected cost-to-go as the second stage's cost.

Suppose now that we are rather paranoid and do no think the expected cost-to-go represents well the second stage. Instead, we want to take a first stage decision considering that the worst case scenario will happen in the second stage. That is, considering that the second stage's cost is $\sup_\xi Q(x_1,\xi)$. In this case, the problem becomes

$$\begin{aligned} \min_{x_1,u_1} \quad & c_H x_1 + c_T u_1 + \sup_\xi Q(x_1,\xi) \\ \text{s.t.} \quad & 0 \leqslant x_1 \leqslant M_H \\ & 0 \leqslant u_1 \leqslant M_T \\ & x_1 + u_1 = d_1. \end{aligned}$$

This problem can be see as another formulation of our example stochastic program that is much more averse to risk.

3.4.3.1 Coherent risk measures

Both the expected value $\mathbb{E}$ and the supremum over all scenarios $\sup_\xi$ are examples of *coherent risk measures*. These permit us to model stochastic programs with

different degrees of risk aversion, what explains why in Section 3.4.2 we called optimization using the expected cost-to-go risk-neutral. The formulation we use throughout this book is based on that found in [Shapiro et al., 2014], with the definition of coherent risk measures coming from [Artzner et al., 1999].

Definition 3.23. A function $\rho\colon V \to [-\infty, +\infty]$ is called a *coherent risk measure* if it satisfies

1. Monotonicity: If $X \leqslant Y$, then $\rho(X) \leqslant \rho(Y)$;
2. Translation equivariance: For any $a \in \mathbb{R}$, $\rho(X + a) = \rho(X) + a$;
3. Convexity: The risk measure ρ is convex. That is, for any $\lambda \in [0, 1]$,
$$\rho(\lambda X + (1 - \lambda)Y) \leqslant \lambda\rho(X) + (1 - \lambda)\rho(Y);$$
4. Positive homogeneity: If $t \geqslant 0$, then $\rho(tX) = t\rho(X)$.

If V is a space of random variables, both the expected value and the supremum over all realizations satisfy the properties in Definition 3.23.

Theorem 3.24. *Let V be a space of random variables. Then $X \mapsto \mathbb{E}[X]$ and $X \mapsto \sup X$ are coherent risk measures.*

When using coherent risk measures instead of the expected value, a natural question to ask is how it relates to convex functions. As we will shortly prove, any convex non-decreasing function satisfies an analogous of Jensen's inequality in relation to any coherent risk measure.

Theorem 3.25 (Almost Jensen's inequality)**.** *Let $g\colon \mathbb{R} \to \mathbb{R}$ be a convex and non-decreasing function and ρ a coherent risk measure. Then, for any random variable X:*
$$g(\rho(X)) \leqslant \rho(g(X)).$$

Proof. Denote by $c = \rho(X)$. Since g is convex, Theorem 2.27 says that there exists an $a \in \mathbb{R}$ such that for all values of X
$$g(X) \geqslant g(c) + a(X - c).$$
This is a tight cut for g and as a consequence of Theorem 2.58, the multiplier a must be non-negative because g is non-decreasing.

From the monotonicity of the risk measure ρ,
$$\rho(g(x)) \geqslant \rho\big(g(c) + a(X - c)\big)$$
and we can use the positive homogeneity and translation equivariance to rewrite this last expression as
$$\begin{aligned}\rho(g(x)) &\geqslant g(c) + a\big(\rho(X) - c\big)\\ \rho(g(x)) &\geqslant g(\rho(X)) + a\big(\rho(X) - \rho(X)\big) = g(\rho(X)).\end{aligned}$$
□

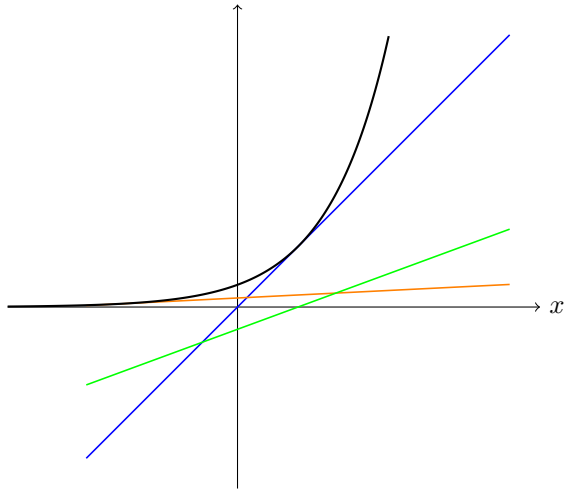


Figure 3.6: Any valid cut for a convex non-decreasing function has non-negative slope.

If we select a family of probabilities $\mathcal{P}$, taking the expected value of a random variable with respect to each of them and choosing their supremum defines a coherent risk measure

$$\rho(X) = \sup_{\mu \in \mathcal{P}} \mathbb{E}^{\mu}[X].$$

As we will see, in fact any coherent risk measure that is proper and lower semi-continuous can be represented in such a manner.

Theorem 3.26 (Dual representation of coherent risk measures)**.** *Any coherent risk measure ρ, defined over a space V of random variables, that is proper and lower semi-continuous may be written as*

$$\rho(X) = \sup_{\mu \in \mathcal{P}} \mathbb{E}^{\mu}[X]$$

where the supremum is taken over the convex *set*

$$\mathcal{P} = \{\mu \text{ is a probability} \mid \mathbb{E}^{\mu}[X] \leqslant \rho(X), \forall X \in V\}.$$

3.4.3.2 Risk-averse stochastic programs

Definition 3.27 (Two-stage risk-averse stochastic program)**.** Let ρ be a coherent risk measure, a *risk-averse stochastic program* with respect to ρ is an optimization problem with the form

$$\begin{aligned} \min_{x_1, u_1} \quad & c_1(x_1, u_1) + \rho(Q(x_1, \xi)) \\ \text{s.t.} \quad & (x_1, u_1) \in X_1 \end{aligned}$$

where $Q(x_1, \xi)$ is the optimization problem

$$\begin{aligned} \min_{x_2, u_2} \quad & c_2(\xi)(x_2, u_2) \\ \text{s.t.} \quad & (x_1, x_2, u_2) \in X_2(\xi). \end{aligned}$$

That is Q is a random optimal value function on x_1 with both the objective function c_2 and the feasible set depending on the realization of the random variable ξ.

Notice that this is the same as Definition 3.18 if we take ρ to be the expected value $\mathbb{E}$. The function $\rho(Q)$ is the problem's *risk-averse cost-to-go* or ρCTG, for short.

The optimization of coherent risk measures is a rich topic and will not be pursued in depth in here. In what follows, we will restrict ourselves to the differences between the decomposed and linked formulation for the risk-averse cost-to-go and its consequences for calculating $\rho(Q)(x_1)$ and valid cuts for it. The reader who is interested in the details theory's details can find them in Chapter 6 of [Shapiro et al., 2014].

Suppose, as in Section 3.4.2 that the sample space is finite. That is, given a random variable ξ there is only a finite number of possible outcomes $\xi^1, \ldots, \xi^N$. Remember also the notation $Q^i(x_1) = Q(x_1, \xi_i)$. In this context, the risk measure ρ can be seem as a function from $\mathbb{R}^N$ to $\mathbb{R}$. Assuming that it is finite valued, the fact that it is convex implies that it is also lower semi-continuous. Therefore, writing $p = (p_1, \ldots, p_N)$ for a vector of probabilities, the dual representation becomes

$$\rho(Q)(x_1) = \sup_{p \in \mathcal{P}} \sum_{i=1}^{N} p_i Q^i(x_1) \tag{3.33}$$

where the set $\mathcal{P}$ is a subset of the probability simples on $\mathbb{R}^n$ which uniquely determines the risk measure ρ. From this, we can write the decomposed formulation for the risk-averse cost-to-go.

Decomposed formulation

Definition 3.28 (Decomposed formulation for ρCTG)**.** The *decomposed formulation* for the risk-averse cost-to-go consists in solving the N optimization problems $Q_i(x_1)$ separately and, afterwards, evaluating the risk measure applied to the vector $(Q^i(x_1), \ldots, Q^N(x_1)) \in \mathbb{R}^N$ using the optimization problem of Equation (3.33) to get the value of $\rho(Q)(x_1)$.

Similarly to the risk-neutral case, the decomposed formulation can be used to calculate a valid cut for $\rho(Q)$ from cuts defined for each scenario. Assume that we have calculated cuts at the point $x_1 = y$ for each scenario and, therefore, have N affine functions satisfying

$$Q^i(x_1) \geqslant q^i - \langle \lambda^i, x_1 - y \rangle.$$

Assume also that the maximum is attained when calculating $\rho(Q)(y)$ using the decomposed formulation. That is, there is a probability vector $p^* = (p_1^*, \ldots, p_N^*)$ such that

$$\rho(Q)(y) = \sup_{p \in \mathcal{P}} \sum_{i=1}^{N} p_i Q^i(y) = \sum_{i=1}^{N} p_i^* Q^i(y).$$

Notice that p^* is feasible for calculating $\rho(Q)(x_1)$ for any value of the variable x_1. Hence, we get a valid cut for the risk-averse cost-to-go by

$$\rho(Q)(x_1) \geqslant \sum_{i=1}^{N} p_i^* Q^i(x_1) \geqslant \sum_{i=1}^{N} p_i^* \left(q^i - \langle \lambda^i, x_1 - y \rangle\right) = \sum_{i=1}^{N} p_i^* q^i - \langle \sum_{i=1}^{N} p_i^* \lambda^i, x_1 - y \rangle.$$

Setting

$$\bar{q} = \sum_{i=1}^{N} p_i^* q^i, \qquad \bar{\lambda} = \sum_{i=1}^{N} p_i^* \lambda^i \tag{3.34}$$

we get can express this new cut as

$$\rho(Q)(x_1) \geqslant \bar{q} - \langle \bar{\lambda}, x_1 - y \rangle. \tag{3.35}$$

An equivalent result to Theorem 3.20 holds for the risk-averse cost-to-go. That is, we the cuts are tight for each scenario at $x_1 = y$, the cut constructed above we also be tight for the risk-averse cost-to-go at this same point.

Theorem 3.29 (Tightness of average cut)**.** *Suppose that we have cuts*

$$Q^i(x_1) \geqslant q^i - \langle \lambda^i, x_1 - y \rangle$$

that are tight for each Q^i at the point $x_1 = y$. That is, $q^i = Q^i(x_1)$. Then the cut constructed in Equation (3.34) *is also tight at $x_1 = y$,*

$$\bar{q} = \rho(Q)(y).$$

Proof. By evaluating the average cut at $x_1 = y$, we get

$$\bar{q} = \sum_{i=1}^{N} p_i^* q^i = \sum_{i=1}^{N} p_i^* Q^i(x_1) = \sup_{p \in \mathcal{P}} \sum_{i=1}^{N} p_i^* Q^i(x_1) = \rho(Q)(x_1). \qquad \square$$

In Section 3.4.2, we discussed the fact that this tightness result requires that in each scenario the cuts are tight for Q^i. It is, in general, no sufficient to have cuts that are tight for the convex relaxations $\check{Q}^i$. The same happens for risk-averse problems. To see this, notice that the function $x_1 \mapsto (\check{Q}^1(x_1), \ldots, \check{Q}^N(x_1))$ is convex in relation to the non-negative cone and that ρ is convex and non-decreasing, by the definition of coherent risk measure. Thus, the map $x_1 \mapsto \rho(\check{Q})(x_1)$ is a convex function. By monotonicity, we know that $\check{Q} \leqslant Q \implies \rho(\check{Q}) \leqslant \rho(Q)$. But $\rho(\check{Q})$ is a convex function below $\rho(Q)$ and by the definition of convex relaxation,

$$\rho(\check{Q}) \leqslant \widetilde{\rho(Q)} \leqslant \rho(Q). \tag{3.36}$$

Therefore, calculating cuts that are tight for $\check{Q}^i$ is not enough to calculate cuts that are tight for the convex relaxation of the entire risk-averse cost-to-go.

Linked formulation Remember from Equation (3.32) that we can write the average of all scenarios as single minimization problem. Since a coherent risk measure is representable by the maximum of a set of averages the linked formulation for risk-averse problems consists in solving this maxmin problem.

Definition 3.30 (Linked formulation for ρCTG)**.** The *linked formulation* for the risk-averse cost-to-go consists of the optimization problem

$$\begin{aligned}\rho(Q)(x_1) = \max_{p\in\mathcal{P}} \quad & \min_{\substack{x_2^i,u_2^i \\ \text{for } i=1,\dots,N}} \quad \sum_{i=1}^N p_i\, c_2^i(x_2^i,u_2^i) \\ & \text{s.t.} \quad (x_1,x_2^i,u_2^i)\in X^i, \text{ for } i=1,\dots,N.\end{aligned}$$

For an arbitrary risk measure, this linked formulation may be a rather complicated problem. If the cost-to-go functions are convex, the minimization for each fixed p is a convex problem and the maximization is taken for a concave function of p over a convex set, therefore is also a convex problem. This problem can be used to calculate cuts that are tight for the convex relaxation $\widetilde{\rho(Q)}$.

We give next an example of a coherent risk measure that can be written as a minimization problem and, hence, whose linked formulation can be written as a single optimal value function.

3.4.3.3 Conditional value-at-risk

The first example of coherent risk measure that we introduced consisted in the maximum of a random variable. A middle ground between it and the risk-neutral approach is given by the *conditional value at risk*, defined using the formulation from [Rockafellar and Uryasev, 2000, thm. 1, pg. 5].

Definition 3.31 (Conditional Value-at-Risk)**.** The α *conditional value-at-risk* is the function defined by

$$\texttt{CVaR}_\alpha[X] = \inf_{z\in\mathbb{R}}\; z + \frac{1}{1-\alpha}\mathbb{E}\big[[X-z]_+\big]$$

where $[x]_+ = \max\{x,0\}$ is the positive part of x.

The $\texttt{CVaR}_\alpha$ is a coherent risk measure [Shapiro et al., 2014, ex 6.16, pg 272] that is always between the expected value and the supremum for any α. Moreover it approaches the supremum when $\alpha\to 0$ and approaches the expected value when $\alpha\to 1$. In order to express a linked formulation for the $\texttt{CVaR}_\alpha$ of a cost-to-go function, we will use the formulation given below.

Supposing that the random variable X has finite support, the conditional value-at-risk can be rewritten as the optimal value function of a linear program. To see this, call $\mathbb{P}[X=x_i]=p_i$. Then the formulation may be rewritten as

$$\texttt{CVaR}_\alpha[X] = \min_{z\in\mathbb{R}}\; z + \frac{1}{1-\alpha}\sum_{i=1}^N p_i[x_i-z]_+.$$

By rewriting each term $[x_i - z]_+$ of the sum in epigraph form, this yields

$$\begin{aligned} \texttt{CVaR}_\alpha[X] = \min_{z,t} \;& z + \tfrac{1}{1-\alpha}\textstyle\sum_{i=1}^N p_i t_i \\ \text{s.t.} \;& \max\{x_i - z,\, 0\} \leqslant t_i, \\ & z \in \mathbb{R} \end{aligned} \quad = \quad \begin{aligned} \min_{z,t} \;& z + \tfrac{1}{1-\alpha}\textstyle\sum_{i=1}^N p_i t_i \\ \text{s.t.} \;& x_i \leqslant t_i + z, \\ & t \geqslant 0,\, z \in \mathbb{R}, \end{aligned} \tag{3.37}$$

which is the optimal value function of a linear program.

The representation (3.37) for the conditional value-at-risk can be used instead of Definition 3.30 to express the linked formulation for $\texttt{CVaR}_\alpha$ in a manner that is more suitable for computational applications. To do that, recall that the cost-to-go for scenario ξ^i is

$$\begin{aligned} Q^i(x_1) = \min_{x_2^i, u_2^i} \;& c^i(x_2^i, u_2^i) \\ \text{s.t.} \;& (x_1, x_2^i, u_2^i) \in X^i. \end{aligned}$$

Since for each value x_1, the cost-to-go is a random variable whose possible realizations are $Q^i(x_1)$ for $i = 1, \ldots, N$, the linked formulation for $\texttt{CVaR}_\alpha[Q](x_1)$ can be written as

$$\begin{aligned} \texttt{CVaR}_\alpha[Q](x_1) = \min_{z,t} \;& z + \tfrac{1}{1-\alpha}\textstyle\sum_{i=1}^N p_i t_i \\ \text{s.t.} \;& Q^i(x_1) \leqslant t_i + z, \text{ for } i = 1, \ldots, N \\ & t \geqslant 0, z \in \mathbb{R}. \end{aligned} \tag{3.38}$$

If we introduce additional variables x_2^i, u_2^i, the expression above may be equivalently rewritten as

$$\begin{aligned} \texttt{CVaR}_\alpha[Q](x_1) = \min_{z,t,x_2,u_2} \;& z + \tfrac{1}{1-\alpha}\textstyle\sum_{i=1}^N p_i t_i \\ \text{s.t.} \;& c^i(x_2^i, u_2^i) \leqslant t_i + z, \text{ for } i = 1, \ldots, N \\ & (x_1, x_2^i, u_2^i) \in X^i, \text{ for } i = 1, \ldots, N \\ & t \geqslant 0, z \in \mathbb{R}. \end{aligned} \tag{3.39}$$

This optimal value function is the *linked formulation for* $\texttt{CVaR}_\alpha[Q]$.

To see that in fact the problems in Equations (3.38) and (3.39) are equivalent, fix x_1 and let z^*, t^* be optimal solutions of (3.38) and $({x_2^i}^*, {u_2^i}^*)$ be optimal solutions for the cost-to-go $Q^i(x_1)$ for each stage. Then the point (z^*, t^*, x_2^*, u_2^*) is feasible for the problem in (3.39). Furthermore, since $c^i(x_2^*, u_2^*) = Q^i(x_1)$ is the minimum possible value that c^i may attain at a feasible point, this problem has the same optimal value as (3.38) and, therefore, also equals $\texttt{CVaR}_\alpha[Q](x_1)$.

3.4.4 Multi-stage stochastic programming

So far, we have only discussed two-stage stochastic programs, but multi-stage problems from Section 3.17 can also be stochastic. For a T-stage problem, the uncertainty is given by a stochastic process ξ_t with $t = 2, \ldots, T$ and the problem at stage t is assumed to depend on the entire past but not on the future realizations of

the uncertainty. In this work, multi-stage stochastic programs will only reappear in Chapter 7 with the assumption that ξ is stagewise independent. That is, the random variable ξ_t is independent of all previous uncertainties $\xi_2, \ldots, \xi_{t-1}$.

Definition 3.32 (Multi-stage stochastic program)**.** A multi-stage stochastic program is an optimization problem with the form

$$\min_{(x_1,u_1)\in X_1} c_1(x_1,u_1) + \rho_2\Bigg(\min_{(x_1,x_2,u_2)\in X_2} c_2(x_2,u_2) + \rho_3\Big(\min_{(x_3,x_3,u_3)\in X_3} c_3(x_3,u_3) + \cdots \\ \cdots + \rho_T\Big(\min_{(x_{T-1},x_T,u_T)\in X_T} c_T(x_T,u_T)\Big)\Big)\Bigg)$$

where ξ_t is a stochastic process and ρ_t are coherent risk measures.

In the most common case where all ρ_t equal the expected value $\mathbb{E}$, the problem is said to be *risk-neutral*.

Suppose that the uncertainty ξ is stagewise independent. Then, this problem may be given a dynamic programming formulation similar to Equation (3.24) as

$$\begin{aligned} Q_t(x_{t-1},\xi_t) = \min_{x_t,u_t} \quad & c_t(x_t,u_t) + \bar{Q}_{t+1}(x_t) \\ \text{s.t.} \quad & (x_{t-1},x_t,u_t) \in X_t \end{aligned} \tag{3.40}$$

where (c_t, X_t) are the random data representing the random variable ξ_t and the function

$$\bar{Q}_{t+1}(x_t) = \begin{cases} \rho_{t+1}\big(Q_{t+1}(x_t,\xi_{t+1})\big), & t = 1,\ldots,T-1 \\ 0, & t = T \end{cases} \tag{3.41}$$

is the expected cost-to-go for stage $t+1$. Since we supposed the uncertainty to be stagewise independent, this formulation is indeed equivalent to Definition 3.32 because the expected cost-to-go for stage $t+1$ is independent from the realizations of the uncertainty at the previous stages and we can rewrite this expression as a sum over all stages. For a deeper discussion of the formulations for multi-stage stochastic programs, see [Shapiro et al., 2014, sec 3.1].

Measures and Distributions

4

For smooth functions on some open subset Ω of $\mathbb{R}^n$, the usual definition of convexity is equivalent to saying that its Hessian is always positive semidefinite. That is,

$$f \text{ is convex in } \Omega \iff D^2 f(x) \geqslant 0,\ \forall x \in \Omega. \tag{4.1}$$

Remembering from Theorem 2.33 that a convex function on $\mathbb{R}^n$ is always continuous in the interior of its domain, a question arises: Is there an analogous of Equation (4.1) for functions which are only *continuous*?

The answer is affirmative in the context of the Theory of Distributions, developed by Laurent Schwartz in [Schwartz, 1966], and explained in this work throughout Section 4.2. In this theory, we consider not just functions in the strict sense but also other more general objects for which we can extend the notion of derivative. The main result in this section is theorem 4.27 which says that a distribution is a convex function if and only if its generalized derivative is positive semidefinite in a sense that will be analogous to Equation (4.1).

Besides distributions, this chapter also deals with two other topics that are closely related to it: measures and convolutions. A measure is a special type of function that generalizes the notions of volume and probability. They are introduced in Section 4.1 where we discuss decompositions of measures, which will later be used to study the generalized derivatives of continuous functions. The development of measure theory will not pursued in any depth and we only focus on those definitions and results that essential for the later applications. In particular, we will not discuss integration theory here. The interested reader may consult [Folland, 1999] for a more thorough presentation of measure theory. Convolutions are introduced in Section 4.2.3 and are a suitable tool to represent additive noises, as will be done in Chapter 5. The convolution of two functions is a kind of product that can be interpreted as a moving average of one function in relation to the other. It is of great importance thanks to its smoothing properties and the fact that it preserves convexity.

Throughout this chapter, we put a special emphasis in the role played by *convexity*, both of functions and of sets, when dealing with measures and distributions.

4.1 Measures

The study of collections of *measurable sets* and *measures* over them is called *measure theory*. Besides generalizing and better explaining the ideas of length, area, and volume, it is also serves as a foundation to probability and is of great importance in many applications of analysis.

This section introduces some results from measure theory that will be important afterwards. Since this subject is only tangential to the main topics of this work, the treatment given here is in no way comprehensive. Thus, we recommend the books [Folland, 1999] and [Tao, 2011] as references on Measure Theory.

Given a set X, we wish to define a proper way to measure its subsets. It is not always possible to do for every subset of X. Hence, we restrict ourselves to study families of subsets called σ-algebras, defined below.

Definition 4.1. A σ-algebra on a set X is a subset $\mathcal{F}$ of $\mathcal{P}(X)$ satisfying

- $\varnothing \in \mathcal{F}$,
- Closure under complements. That is, $A \in \mathcal{F} \implies A^c \in \mathcal{F}$,
- Closure under countable unions. That is, if $A_1, A_2, \ldots$ is a countable sequence of elements in $\mathcal{F}$, then $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{F}$.

The axioms for a σ-algebra also imply that it is closed under countable intersections, since

$$\bigcap_{n\in\mathbb{N}} A_n = \Big(\bigcup_{n\in\mathbb{N}} A_n^c\Big)^c \tag{4.2}$$

and that it contains the entire space X, since $\varnothing^c = X$.

The pair $(X, \mathcal{F})$ is called *measurable space* and the elements of $\mathcal{F}$ are called *measurable sets*. Given some family C of subsets of X, it is often useful to consider the smallest σ-algebra containing it, denoted by $\sigma(C)$. This is always well-defined since $\mathcal{P}(X)$ is a σ-algebra and an arbitrary intersection of σ-algebras is also a σ-algebra.

Definition 4.2. The σ-algebra generated by a set $C \subset \mathcal{P}(X)$ is the intersection of all the σ-algebras containing C.

One important case of σ-algebra that often arises is the *Borel σ-algebra* of a topological space X, which is the σ-algebra generated by the open sets of that space and denoted $\mathscr{B}(X)$. Thanks to the closure under complements, this σ-algebra also contains the closed sets as well as every countable union and intersection of open and closed sets. The example that we will encounter the most is $\mathscr{B}(\mathbb{R}^n)$.

After establishing what we can measure (the measurable sets), we want to establish how we can measure them. This is accomplished via Definition 4.3 of measures over a σ-algebra.

Definition 4.3. A *measure* over a measurable space $(X, \mathcal{F})$ is a function $\mu\colon \mathcal{F} \to \mathbb{R}$ satisfying

- $\mu(\varnothing) = 0$,
- If $A_1, A_2, \ldots$ is a sequence of *disjoint* elements of $\mathcal{F}$,

$$\mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} \mu(A_n).$$

Remark 4.1. Definition above is sometimes called a *finite signed* measure to distinguish it from the other cases below.

The last property in Definition 4.3 is called *countably additivity* and is the distinguishing characteristic of a measure. If we look back at the definition of σ-algebra, we see that its properties are defined in such a way that a measure is always well-defined over them. The pair $(X, \mathcal{F}, \mu)$ is called a *measure space*.

As we did for ordinary functions in Chapter 2, it is also very useful to consider extended real-valued measures, that is, measures $\mu\colon \mathcal{F} \to [-\infty, +\infty]$. A difference in this context is that we can extend the codomain only in one direction, for additivity implies that if there are disjoint sets P and N such that $\mu(P) = +\infty$ and $\mu(N) = -\infty$, then

$$\mu(P \cup N) = \mu(P) + \mu(N) = \infty - \infty,$$

that is an expression from which we cannot make sense. As we already gave some preference to $+\infty$ when dealing with convex functions, we will always assume that a measure can only take positive infinite values.

An important case consists of the *non-negative measures*, which, as the name implies, are measures whose image is always non-negative. In other words, A *non-negative measure* over a measurable space $(X, \mathcal{F})$ is a measure $\mu\colon \mathcal{F} \to [0, \infty]$ which only takes non-negative values. If, besides non-negativity, it also holds that $\mu(X) = 1$, the measure μ is called a *probability* which is certainly the most important kind of measure throughout the remainder of this dissertation.

Notice that the set of signed measures over some measurable space is a vector space with addition and multiplication by scalar defined pointwisely while the non-negative measures form a convex cone and probabilities form a convex set.The cone of non-negative measures induces a natural ordering on the signed measures, which we will always use thereafter.

Definition 4.4 (Measure Ordering)**.** If $\mu, \nu\colon \mathcal{F} \to \mathbb{R}$ are measures, we say that $\mu \leqslant \nu$ if

$$\mu(A) \leqslant \nu(A),\ \forall A \in \mathcal{F}.$$

An example of measure is the point mass at a $a \in X$. It is well-defined for

any σ-algebra of X by the formula

$$\delta_a(B) = \begin{cases} 1, & a \in B \\ 0, & a \notin B. \end{cases} \tag{4.3}$$

These measures will play an important role later in this work because they form a basis for the space of measures with finite support. One consequence of this is that any probability p with support in $\{x_1, x_2, \ldots, x_N\}$ and assigning $p(x_k) = p_k$ can be written as convex combinations of the δ_{x_k} as

$$p = \sum_{k=1}^{N} p_k \delta_{x_k}.$$

Lebesgue measure The main example of a measure is the Lebesgue Measure on $\mathbb{R}^n$, which generalizes the notions of length, areas and volumes for every set on the Borel σ-algebra of $\mathbb{R}^n$, and even to a larger collection of sets called the Lebesgue σ-algebra of $\mathbb{R}^n$ and denoted by $\mathcal{L}(\mathbb{R}^n)$.

It is the only translation invariant measure in $\mathbb{R}^n$ which is equivalent to the usual notion of volume in a cube. Defining the volume of a cube to be the product of its sides,

$$\mathrm{Vol}([a_1, b_1] \times \cdots \times [a_n, b_n]) = \prod_{i=1}^{n} |b_i - a_i|,$$

the Lebesgue measure of a set A is the smallest "volume" of a cover by open cubes that a set can have:

$$\lambda(A) = \inf \Big\{ \sum_{k \in \mathbb{N}} \mathrm{Vol}(C_k) \,\Big|\, C_1, C_2, \ldots \text{ such that } A \subset \bigcup_{k \in \mathbb{N}} C_k \Big\}. \tag{4.4}$$

The Lebesgue measure of a set always coincides with the volume defined via Riemann integration or other classical methods, when these are defined for A. The proof that (4.4) is truly a measure on a σ-algebra containing $\mathscr{B}(\mathbb{R}^n)$ is done via Carathéodory's method, which is out of scope for this dissertation but whose details can be found at [Folland, 1999].

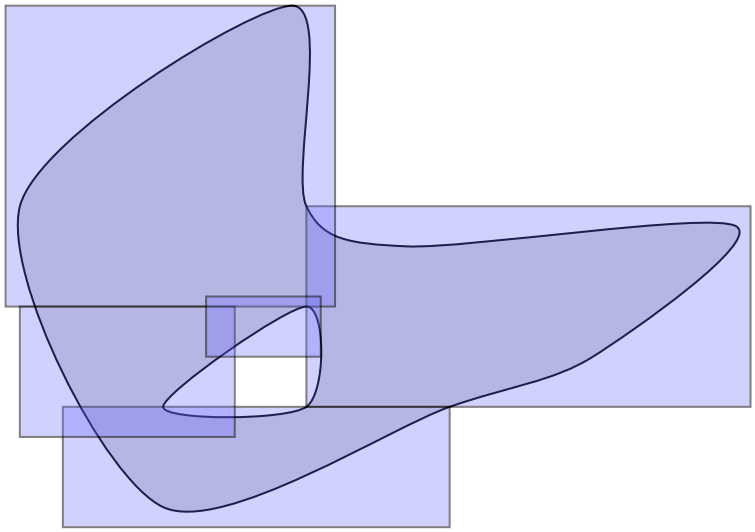


Figure 4.1: An illustration of Lebesgue measure as a covering by rectangles.

4.1.1 The Hahn-Jordan decomposition

Many times, when dealing with a function f from a set X to $\mathbb{R}$, it is useful to consider its decomposition into its positive and negative part.

Definition 4.5. The positive part of a function f is defined by

$$f_+(x) = \max\{f(x), 0\},$$

and its negative part by

$$f_-(x) = -\min\{f(x), 0\}.$$

Both f_+ and f_- are non-negative functions and the function f can be reconstructed from them as $f = f_+ - f_-$. These two functions also satisfy a minimality property because of their definition as optimal value functions: they are below any other pair of functions whose difference is f.

Theorem 4.6. *If $g, h\colon X \to \mathbb{R}$ are non-negative functions such that $f = g - h$, then*

$$f_+ \leqslant g \quad \textit{and} \quad f_- \leqslant h.$$

Measures also admit a similar decomposition as a difference of non-negative measures. It is not as simple as the one in Definition 4.5 since even if $\mu\colon \mathcal{F} \to \mathbb{R}$ is a measure the function $A \mapsto \max\{\mu(A), 0\}$, in general, is not. In what follows we will construct the *Hahn-Jordan decomposition* of a measure μ as the difference of two non-negative measures. We begin with Theorem 4.7, which says that any measure space can be partitioned into a union of a positive and a negative set.

Theorem 4.7 (Hahn decomposition)**.** *Let $\mathcal{F}$ be a σ-algebra of the subsets of X and $\mu\colon \mathcal{F} \to (-\infty, \infty]$ be an extended measure over $\mathcal{F}$. Then there are disjoint measurable sets P and N such that $X = P \cup N$ and*

$$\begin{aligned} A \subset P &\implies \mu(A) \geqslant 0, \\ B \subset N &\implies \mu(B) \leqslant 0. \end{aligned}$$

A proof of theorem 4.7 can be found at [Folland, 1999] or [Durrett, 2017]. This decomposition is not unique, since we can transfer any set such that all its subsets have zero measure from P to N or vice versa. Nevertheless, any such decomposition can be used to write μ as a difference of two non-negative measures and the result will be the same.

Definition 4.8. A pair of measures $\mu_1, \mu_2\colon \mathcal{F} \to (-\infty, +\infty]$ are *mutually singular* if it exists a measurable set S such that

$$\begin{aligned} A \subset S &\implies \mu_1(A) = 0, \\ A \subset S^c &\implies \mu_2(A) = 0. \end{aligned}$$

This definition says that the space X can be decomposed into two sets S and S^c such that μ_1 is zero for any subset of S and μ_2 is zero for for any subset of S^c.

Theorem 4.9 (Hahn-Jordan decomposition)**.** *Let $\mu\colon \mathcal{F} \to (-\infty, +\infty]$ be a signed measure. Then, there exists a unique decomposition*

$$\mu = \mu_+ - \mu_-$$

where μ_+ and μ_- are mutually singular non-negative measures.

Similarly to the decomposition of ordinary functions in Definition 4.5, the Hahn-Jordan decomposition of a measure μ also admits a characterization as a optimal value function. Furthermore, the fact that they are mutually singular also provides a minimality property, analogous to Theorem 4.6.

Corollary 4.9.1. If $\mu\colon \mathcal{F} \to \mathbb{R}$ is a finite measure, then the components of its Hahn-Jordan decomposition satisfy

$$\mu_+(A) = \begin{array}{ll} \sup & \mu(E) \\ \text{s.t.} & E \subset A, \\ & E \in \mathcal{F} \end{array}, \qquad \mu_-(A) = \begin{array}{ll} -\inf & \mu(E) \\ \text{s.t.} & E \subset A, \\ & E \in \mathcal{F} \end{array}$$

Corollary 4.9.2 (Minimality of the Hahn-Jordan decomposition). Any other decomposition of a signed measure μ has components greater than the ones given by the Hahn-Jordan decomposition. That is, if there are positive measures λ and ν such that $\mu = \lambda - \nu$ then

$$\lambda \geqslant \mu_+, \quad \nu \geqslant \mu_-.$$

A proof to both of these corollaries can be found at [Fischer, 2012]. In the subsequent sections, we will denote by $[\cdot]_+$ and $[\cdot]_-$, the functions that take a measure to its positive or negative part. An interesting fact, that will be used in Section 5.3, is that both these functions are convex with relation to the cone of non-negative measures.

Theorem 4.10. *Both $[\cdot]_+$ and $[\cdot]_-$ are convex in relation to the cone of non-negative measures. That is, if μ and ν are measures, then for any $t \in [0,1]$:*

$$[t\mu + (1-\tau)\nu]_+ \leqslant t[\mu]_+ + (1-t)[\nu]_+,$$
$$[t\mu + (1-\tau)\nu]_- \leqslant t[\mu]_- + (1-t)[\nu]_-.$$

Proof. From Corollary 4.9.1, for each fixed $A \in \mathcal{F}$, we can represent $\mu \mapsto [\mu]_+(A)$ as the optimal value function that varies the objective function of a maximization problem. As we saw in Theorem 3.5, this is a convex function in the measure μ. Therefore,

$$[t\mu + (1-t)\nu]_+(A) \leqslant t[\mu]_+(A) + (1-t)[\nu]_+, \ \forall A \in \mathcal{F}.$$

The same argument holds for $\mu \mapsto [\mu]_-$. □

4.1.1.1 Total variation norm

The Hahn-Jordan decomposition permits us to write any measure as $\mu = \mu_+ - \mu_-$. If we sum these components, we get a non-negative measure called the *total variation* of μ.

Definition 4.11. The *total variation* of a measure μ is the non-negative measure

$$|\mu| = \mu_+ + \mu_-,$$

where μ_+ and μ_- are the components of the Hahn-Jordan decomposition of μ.

As the notation hints, the total variation of a measure plays an analogous role to the absolute value of a function. As an example, it always holds that $\mu \leqslant |\mu|$ and the function $\mu \mapsto |\mu|$ is convex in relation to the cone of non-negative measures.

On a measurable space $(X, \mathcal{F})$, the set of all measures such that $|\mu|(X)$ is finite forms a vector space. Moreover, the total variation applied to X is a norm which turns it into a Banach space.

Theorem 4.12. *Given a measurable space $(X, \mathcal{F})$, the set of all measures such that $|\mu|(X)$ is finite is a Banach space with norm $\|\mu\| = |\mu|(X)$.*

Remark 4.2. Notice that, since $|\mu|$ is non-negative, additivity says that

$$A \subset B \implies \mu(B) = \mu(A) + \mu(B \backslash A) \geqslant \mu(A).$$

Therefore the maximum of $|\mu|$ is attained when it is applied to the entire X:

$$|\mu|(X) = \sup_{E \in \mathcal{F}} |\mu|(E).$$

4.1.2 Vector and matrix valued measures

The countably additivity property of measures only requires that we can sum elements of the measure's codomain and that there is some notion of convergence on these elements. Therefore, it is possible to extend this notion to a finite-dimensional real vector space.

Definition 4.13. Given a measurable space $(X, \mathcal{F})$, a *vector measure* is a function $\mu\colon \mathcal{F} \to \mathbb{R}^n$ satisfying

- $\mu(\varnothing) = 0$,
- If $A_1, A_2, \ldots$ is a sequence of *disjoint* elements of $\mathcal{F}$,

$$\mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} \mu(A_n).$$

Vector measures will be useful in Section 4.2, where we will see that they can be used to represent the gradient and Hessian of a continuous function.

The interested reader can find the theory of vector measures in the books [Diestel and Uhl, 1977] and [Rao, 2011] as well as the article [Robertson and Rosenberg, 1968], where the theory of matrix-valued measures is studied.

4.2 Distributions

A classical result for smooth functions, as can be found in [Boyd and Vandenberghe, 2004], says that $f \in C^2(\mathbb{R}^n)$ is convex if and only if its Hessian is always positive semi-definite.

The aim of this section is to build a generalization to this theorem for a general convex function whose domain is open. To achieve this, we will define distributions as the continuous linear functionals over the space of infinitely differentiable, compactly supported functions which is a space "large" enough to contain all the locally integrable functions as well as sufficiently regular measures. Then, we show that the notion of derivative can be properly generalized to any distribution in a way that is compatible with the usual sense when applied to a differentiable function. In fact, theorem 4.19 says that, in this generalized sense, any distribution has infinitely many derivatives.

In Section 4.2.2 we present some regularity results concerning the derivatives of distributions. That is, theorems saying that if a distribution's derivative is regular enough then the distribution is indeed a function. Specially, theorem 4.27 says that a distribution is a convex function if and only if its generalized Hessian is a positive semi-definite measure, in the sense of Section 4.1.2.

Later, in Section 4.2.3, we present some results regarding convolutions of distributions, which will be a useful tool in order to represent additive noises

throughout Chapter 5. The notion of convexity will also reappear in this context through Titchmarsh theorem, which says that

$$\operatorname{conv}\operatorname{supp}(f * g) = \operatorname{conv}\operatorname{supp}(f) + \operatorname{conv}\operatorname{supp}(g). \tag{4.5}$$

More thorough treatments of these subjects can be found on the original treatise [Schwartz, 1966] or on a more modern language in [Hörmander, 2003].

4.2.1 Distributions

Our main motivation to work with distributions is to extend the algebraic rules of calculus that work well for smooth functions to more general objects. This can be done via duality on a suitable space of *test functions*. One of the most regular classes of functions is that of infinitely differentiable functions whose support is a compact set. These will be our test functions.

Throughout what follows, we will work with functions defined on some open subset Ω of $\mathbb{R}^n$.

Definition 4.14. The *support* of a function $f\colon \Omega \to \mathbb{R}$ is the closure of the set where it does not equal zero,

$$\operatorname{supp}(f) = \operatorname{cl}\{x \in \Omega \mid f(x) \neq 0\}.$$

Definition 4.15 (Test functions)**.** a *test function* on Ω is an element of $C_c^\infty(\Omega)$, the space of infinitely differentiable functions with compact support on Ω.

As an example of a test function on $\mathbb{R}^n$, we can define

$$f(x) = \begin{cases} \exp\left(\frac{1}{1-\|x\|_2^2}\right), & \|x\|_2 \leqslant 1 \\ 0, & \text{otherwise} \end{cases}$$

where $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$ is the usual euclidean norm on $\mathbb{R}^n$.

In order to better handle the notation of multiple derivatives, we will use throughout this section the *multi-index* notation.

Definition 4.16 (Multi-index notation)**.** If $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$ is a tuple of non-negative integers, we write

$$|\alpha| = \sum_{i=1}^n \alpha_i, \qquad \partial^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}}.$$

A distribution is then defined to be a linear functional over the test functions satisfying some regularity conditions.

Definition 4.17 (Distributions)**.** A *distribution* f in Ω is a linear function over $C_c^\infty(\Omega)$ such that for each fixed compact subset K of Ω there are constants C and l such that for all $\phi \in C_c^\infty(\Omega)$,

$$|f(\phi)| \leqslant C \sum_{|\alpha| \leqslant l} \sup_{x \in \Omega} |\partial^\alpha \phi(x)| .$$

Although Definition 4.17 is rather technical, it can be seem as the condition that the distribution must be continuous for a certain topology on the test functions. This point of view will not be pursued here but can be found by the interested reader at [Schwartz, 1966].

We now proceed to present some examples of distributions, showing that the definition applies to a large family of objects.

- Any locally integrable function defines a distribution. That is, if f is a function such that $\int_K |f|$ is finite for every compact set K, then we can define a distribution T_f by

$$T_f(\phi) = \int f(x)\phi(x)\, dx.$$

  This class of functions encompasses all $L^p(\Omega)$ functions as well as the continuous functions, including the test function themselves.

- Any measure μ on the Borel sets of $\mathbb{R}^n$ which is finite on every compact set defines a distribution T_μ by the formula

$$T_\mu(\phi) = \int \phi\, d\mu.$$

- The linear functionals δ^α defined by

$$\delta^\alpha(\phi) = \partial^\alpha \phi(0).$$

On the previous examples, we saw that functions and measures can be used to defined distributions through integration. Since this representation is well-defined and injective, no confusion shall arise if we talk about a distribution being a function or a measure. In other words, we will refer to f and T_f interchangeably.

In order to define the derivative of a distribution, notice that any single variable test function ψ defines a distribution and that its derivative ψ' is also a test function. Thus, we can consider ψ' as a distribution and apply it to another test function ϕ. From integration by parts and the fact that the support of ϕ is compact:

$$\int \psi'(x)\phi(x)\, dx = [\psi\phi]_{-\infty}^{+\infty} - \int \psi(x)\phi'(x)\, dx = -\int \psi(x)\phi'(x)\, dx.$$

Thus, we get that $\psi'(\phi) = -\psi(\phi')$. If we iterate this process, we get a formula to the k-th derivative by $\psi^{(k)}(\phi) = (-1)^k \psi(\phi^{(k)})$. This process can be used to define the partial derivatives of an arbitrary distribution.

Definition 4.18 (Distributional derivative)**.** If T is a distribution and α some multi-index, we define the α-th distributional derivative of T as

$$\partial^{|\alpha|} T(\phi) = (-1)^{\alpha} T(\partial^{\alpha} \phi).$$

The derivative of a distribution is also a linear functional over the test functions satisfying the conditions of definition 4.17. Therefore, it is also a distribution. From this we see that every distribution is infinitely differentiable on the sense of definition 4.18. Furthermore, it can be proved that if f is actually a differentiable function, its distributional derivative equals the usual one These facts are proved in [Hörmander, 2003].

Theorem 4.19. *For any distribution f, $\partial^{\alpha} f$ is also a distribution. Furthermore, if f is a differentiable function, its distributional derivative coincides with its usual one.*

4.2.1.1 Jump formula

We proceed to derive formulas for the derivatives of single variable piecewise smooth functions. The main interest of this procedure is that the optimal value functions of mixed integer linear programs are piecewise linear. We give the name of *jump formula* to these results.

Here, two distributions will play a major role. The point mass measures from Equation (4.3), which when viewed as distributions act as

$$\delta_a(\phi) = \phi(a), \tag{4.6}$$

and the Heaviside function, defined shortly.

Definition 4.20. The *Heaviside function* H is defined by

$$H(x) = \begin{cases} 1, & x > 0 \\ 0, & x < 0. \end{cases}$$

The Heaviside function is constant by parts with a jump on zero. Its derivative is a point mass at zero as proved in [Schwartz, 1966]:

$$H' = \delta_0. \tag{4.7}$$

This result can be seem as saying that H has zero derivative at all its points of differentiability while the "jump" at zero is represented by the δ_0 on the derivative.

This same intuition can be applied to an arbitrary piecewise smooth function wielding the general *jump formula*.

Theorem 4.21 (Jump formula)**.** *Let f_1 and f_2 be differentiable functions and define, for $a \in \mathbb{R}$, the function*

$$f(x) = f_1(x) \cdot H(x-a) + f_2(x) \cdot H(a-x) = \begin{cases} f_1(x), & x < a \\ f_2(x), & x \geqslant a \end{cases}.$$

Then the distributional derivative of f is

$$f' = f_1' \cdot H(x-a) + f_2' \cdot H(a-x) + [f_2(a^+) - f_1(a^-)]\delta_a.$$

This theorem can be interpreted as saying that if f is differentiable in all point except for a, then its distributional derivative is the expected one at these points and at a it is a point mass proportional to the jump of f at this point. As an example, we will apply this formula to the absolute value $|x|$, which is differentiable except when $x = 0$. Since $|0^+| = |0^-|$,

$$|x|' = \begin{cases} 1, & x \geqslant 0 \\ -1, & x < 0 \end{cases}.$$

Theorem 4.21 can also be extended to when f is a function whose points of non-differentiability are all isolated.

Corollary 4.21.1. Let $f = \sum f_i \cdot \mathbb{1}_{(a_{i-1}, a_i)}$ where each f_i is differentiable in (a_{i-1}, a_i). The its derivative is given by

$$f' = \sum_i f_i' \cdot \mathbb{1}_{(a_{i-1}, a_i)} + \sum_i \left[f_{i+1}(a_i^+) - f_i(a_i^-)\right] \delta_{a_i}.$$

We can also specialize this result to the optimal value functions of mixed integer programs. These are the minimum of convex functions, which can be represented as piecewise convex functions.

Corollary 4.21.2 (Minimum of convex functions). Let $f = \min_i f_i$ where each f_i is a convex function. The function f is continuous and there are intervals (a_{i-1}, a_i) such that

$$f = \sum_i f_i \cdot \mathbb{1}_{(a_{i-1}, a_i)}.$$

The first and second derivatives of f are

$$f' = \sum_i f_i' \cdot \mathbb{1}_{(a_{i-1}, a_i)},$$
$$f'' = \sum_i f_i'' \cdot \mathbb{1}_{(a_{i-1}, a_i)} - \sum_i \left[f_i'(a_i^-) - f_{i+1}'(a_i^+)\right] \delta_{a_i},$$

where each f_i'' is a non-negative measure (by Theorem 4.27.1) and each jump is non-positive.

If the f_i on Corollary 4.21.2 are polyhedral, their second derivatives are non-negative linear combinations of point masses. In this case, f'' is a sum of point masses which are positive in the points of differentiability of each f_i and negative on the points where the minimum changes from one f_i to another.

4.2.2 Regularity results

In this section we discuss some regularity results concerning distributions. That is, theorems saying that if a distribution f satisfies some given property, then it must be a measure or a function.

Our first result is that the only distributions whose derivatives are continuous functions are the continuously differentiable functions.

Theorem 4.22. *A distribution f is a continuously differentiable function if and only if $\partial_i f$ is a continuous function for each i from 1 to n.*

Analogous results exist for when the partial derivatives of a distribution are locally continuous functions or measures. Together, these theorems show that the distributional derivative reduces the regularity in a similar way to the usual definition of derivative, taking differentiable to continuous functions, continuous to locally integrable functions and locally integrable functions to measures.

Theorem 4.23. *A distribution f is a continuous function if and only if $\partial_i f$ is a locally integrable function for each i from 1 to n.*

Theorem 4.24. *A distribution f is a locally integrable function if and only if $\partial_i f$ is a measure for each i from 1 to n.*

Proofs to Theorems 4.22, 4.23, and 4.24 can be respectively found at theorem VII of Chapter II and theorem XVIII of Chapter VI of [Schwartz, 1966].

Since distributions are linear functionals over $C_c^\infty(\Omega)$, we can define the cone of non-negative distributions as the dual cone to the non-negative test functions.

Definition 4.25. A distribution f on Ω is *non-negative* if for all $\phi \in C_c^\infty(\Omega)$ such that $\phi \geqslant 0$,

$$f(\phi) \geqslant 0.$$

An important aspect of this cone is that it coincides with the cone of non-negative measures.

Theorem 4.26. *If f is a non-negative distribution, it is in fact a non-negative measure.*

A proof to this theorem can be found at theorem 2.1.7 of [Hörmander, 2003].

The main result of this section is Theorem 4.27 together with its Corollary 4.27.1, which characterize convex functions in terms of their second derivatives.

Given a distribution f, we can write its distributional Hessian as the matrix

$$(D^2 f)_{ij} = \partial_i \partial_j f$$

whose components are the second partial derivatives of f. This generalizes the usual notion of a function's Hessian and defines a quadratic form from $\mathbb{R}^n$ to the distributions. That is, if a and b are vectors in $\mathbb{R}^n$, they give a distribution by

$$\langle a, (D^2 f)b\rangle = \sum_{i,j} a_i b_j \partial_i \partial_j f.$$

The next theorem says that a distribution is a convex function if and only if its Hessian defines a positive semi-definite quadratic form.

Theorem 4.27. *A distribution f is a convex function if and only if*

$$\langle c, (D^2 f)c\rangle = \sum_{i,j} c_i c_j \partial_i \partial_j f \geqslant 0$$

for any choice of $c \in \mathbb{R}^n$.

The proof to this theorem can be found at Theorem 4.1.7 of [Hörmander, 2003].

From Theorems 4.23 and 4.24 we see that a distribution f is a continuous function if and only if $\partial_i \partial_j f$ is a measure for each $i, j = 1, \ldots, n$. Thus, the distributional Hessian of f,

$$(D^2 f)_{ij} = \partial_i \partial_j f,$$

is a symmetric matrix measure. Remembering that a convex function is always continuous on the interior of its domain, theorem 4.27.1 says that f is convex if and only if $D^2 f$ is positive semi-definite.

Corollary 4.27.1. A distribution f is a convex function if and only if its distributional Hessian $D^2 f$ is a positive semi-definite matrix measure. That is, for each Borel set A, $D^2 f(A)$ is a positive semi-definite matrix.

Proof. If f is a convex function, Theorems 4.23 and 4.24 imply that $\partial_i \partial_j f$ are all measures. By Theorem 4.27, for each vector $c \in \mathbb{R}^n$, $\langle c, (D^2 f)c\rangle$ is non-negative. Thus $D^2 f$ is positive semi-definite.

If $D^2 f$ is a positive semi-definite matrix measure, then for each vector $c \in \mathbb{R}^n$, $\langle c, (D^2 f)c\rangle$ is a non-negative measure and Theorem 4.27.1 says that f is a convex function. □

Remark 4.3. Notice that for distributions over $\mathbb{R}$, the theorems above are equivalent to saying that f is a convex function if and only if $[f'']_- = 0$, where $[\cdot]_-$ is the negative part of the Hahn-Jordan decomposition 4.9. This point of view will be retaken in Section 5.3 where use $[f'']_-$ as a way to measure the non-convexity of a continuous function f.

4.2.3 Convolution of distributions

Definition 4.28. Given test functions $f, g \in C_c^\infty(\mathbb{R}^n)$, their *convolution* is the function

$$(f * g)(x) = \int f(x-y)g(y)\,dy.$$

The convolution of two test functions is also a test function. Furthermore, it satisfy some properties that turn it into a "product" on the space of test functions.

Theorem 4.29. *The convolution between test functions satisfies*

- *Commutativity:* $f * g = g * f$;
- *Associativity:* $f * (g * h) = (f * g) * h$;
- *Linearity: For all* $\lambda \in \mathbb{R}$, $(\lambda f + g) * h = \lambda(f * h) + g * h$.

Unlike the usual product of functions, when taking the derivative of a convolution, we can chose which term we want to derive.

Theorem 4.30. *The derivatives of the convolution* $f * g$ *satisfy*

$$\partial^\alpha(f * g) = (\partial^\alpha f) * g = f * (\partial^\alpha g).$$

Analogously to the derivative, definition 4.28 can be extended to distributions by observing what happens when we look at the convolution of test functions $\phi * \psi$ as a distribution and apply it to another test function h.

$$(\phi * \psi)(h) = \int (\phi * \psi)(x)h(x)\,dx = \int \left(\int \phi(y)\psi(x-y)\,dy \right) h(x)\,dx.$$

From Fubini's theorem, we can exchange the order of integration,

$$(\phi * \psi)(h) = \int \phi(y) \left(\int \psi(x-y)h(x)\,dx \right) dy = \phi(\tilde{\psi} * h)$$

where $\tilde{\psi}(x) = \psi(-x)$.

Thus, we can define the convolution of a distribution with a test function in this way.

Definition 4.31. Let f be a distribution and ψ a test function. Its *convolution* is the distribution $f * \psi$ defined by

$$(f * \psi)(\phi) = f(\tilde{\psi} * \phi)$$

where $\tilde{\psi}(x) = \psi(-x)$.

Notice that the convolution between a distribution and a test function is well-defined because $\tilde{\psi} * \phi$ is always an element of $C_c^\infty(\mathbb{R}^n)$.

Theorem 4.32. *If f is a distribution and ψ a test function, their convolution $f * \psi$ is an infinitely differentiable function.*

Now we proceed to extend the convolution to the case when both terms are distributions. Before doing this we first need to define the support of a distribution.

We say that a distribution f *vanishes* at an open set Ω if $\text{supp}(\phi) \subset \Omega$ implies that $f(\phi) = 0$. That is, if when restricted to $C_c^\infty(\Omega)$, f equals zero. In a way analogous to Definition 4.14 for functions, we define the support of a distribution as the complement of the largest open set where it vanishes.

Definition 4.33. The *support of a distribution* f is the complement of the largest open set Ω such that f vanishes at Ω.

Notice that if f is a function, definitions 4.14 and 4.33 coincide.

From Theorem 4.32, the convolution of a distribution g with a test function ψ is an infinitely differentiable function. This function also has compact support whenever f is compactly supported. Thus, if we define $\tilde{g}$ by $\tilde{g}(\phi) = g(\tilde{\phi})$, we can extend the convolution to compactly supported distributions in a way that is compatible with Definition 4.31 when g is a test function.

Definition 4.34 (Convolution of distributions)**.** If f and g are distributions and at least one of them is compactly supported, their *convolution* is the distribution defined by

$$(f * g)(\phi) = f(\tilde{g} * \phi)$$

Although this rather complicated definition, the intuition of Definition 4.28 must be kept in mind. When we use convolutions of distributions in Chapter 5, it should always be seem as a moving average.

The convolution of distributions can also be defined in some cases when neither of the distributions is compactly supported. An interesting example, as can be found in page 104 of [Hörmander, 2003], is that if the supports of both f and g are contained in a *closed pointed convex cone* K, their convolution $f * g$ is well-defined and its support is also contained in K.

Both Theorem 4.29 and Theorem 4.30 are also valid for the convolution of distributions. We also have some results saying that the support of the convolution $f * g$ is contained in the sum of the supports of f and g.

Theorem 4.35. *If f and g are distributions such that $f * g$ is well-defined, then*

$$\text{supp}(f * g) \subset \text{supp}(f) + \text{supp}(g).$$

When both f and g are compactly supported, equality is attained in Theorem 4.35 if instead of the supports, we consider their convex hulls.

Theorem 4.36 (Titchmarsch theorem)**.** *If f and g are compactly supported distributions,*

$$\operatorname{conv}\operatorname{supp}(f * g) = \operatorname{conv}\operatorname{supp}(f) + \operatorname{conv}\operatorname{supp}(g).$$

The proof to this theorem can be found at the original article [Lions, 1951] or at Theorem 4.3.3 of [Hörmander, 2003].

4.2.3.1 Convolutions and convexity

We end this section with a result saying that the convolution of a convex function f with a non-negative distribution μ is also a convex function.

Remembering that every non-negative distribution is in fact a non-negative measure, it is useful to think about the case when μ is a probability to gain some intuition about this result. In this case, the convolution $f * \mu$ can be thought as exchanging the value of $f(x)$ by the average of f around x with respect to μ. Thus, theorem 4.37 says that taking these averages preserves the convexity of f.

In Chapter 5, we define ways to measure the non-convexity of a function g and the theorems from Sections 5.2.2 and 5.3.1 can be viewed as extensions of Theorem 4.37 which say that $g * \mu$ is always less non-convex than g.

Theorem 4.37. *Let f be a convex function and μ non-negative. Then $f * \mu$ is also convex.*

Proof. We begin by showing that if $\mu \geqslant 0$ and $\nu \geqslant 0$, then $\nu * \mu \geqslant 0$. To do this, notice that for any distribution h,

$$h \geqslant 0 \iff \tilde{h} \geqslant 0$$

and that if ϕ and ψ are non-negative test functions, then Definition 4.28 implies that their convolution $\phi * \psi$ is also non-negative.

Then if μ is a non-negative distribution, $\mu * \phi$ is also non-negative for any non-negative test function ϕ because,

$$(\mu * \phi)(\psi) = g(\tilde{\phi} * \psi) \geqslant 0,$$

for any $\psi \geqslant 0$.

If ν is another non-negative distribution,

$$(\nu * \mu)(\phi) = \nu(\tilde{\mu} * \phi) \geqslant 0$$

for any $\phi \geqslant 0$. Therefore,

$$\mu \geqslant 0 \text{ and } \nu \geqslant 0 \implies \mu * \nu \geqslant 0$$

Now suppose f is a convex function and μ is non-negative. From Theorem 4.27, for any $c \in \mathbb{R}^n$, the derivatives of f satisfy

$$\sum_{i,j} c_i c_j \partial_i \partial_j f \geqslant 0.$$

By the linearity of the convolution,

$$\sum_{i,j} c_i c_j (\partial_i \partial_j f) * \mu \geqslant 0.$$

Since, by Theorem 4.30, $(\partial_i \partial_j f) * \mu = \partial_i \partial_j (f * \mu)$, we get that $f * \mu$ is a convex function. □

As a corollary, we see that if f is a possible non-convex function and there is a distribution μ that convexifies f, that is, $f * \mu$ is convex. Then there is an infinite family of distributions that convexify f.

Corollary 4.37.1. Let f be a function and μ a non-negative distribution such that $f * \mu$ is convex. Then, for every other non-negative distribution ν, $f * (\mu * \nu)$ is also convex.

Proof. By the associativity of the convolution, $f * (\mu * \nu) = (f * \mu) * \nu$. Since $f * \mu$ is convex and ν is non-negative, the result follows from Theorem 4.37. □

Convexification by Averages

5

When solving a non-convex stochastic program, it is possible that the expected cost-to-go is much less non-convex than the cost-to-go for each stage. This chapter is dedicated to study this process of convexification and to rigorously define what we mean by a function being less non-convex than another. As we will see, there are many different ways to measure the non-convexity of a function that, nevertheless, share some important properties, including the fact that they are reduced by the operation of taking averages.

Since these notions work not just for cost-to-go functions but for any random function, we will mostly discuss ways of measuring the non-convexity of an arbitrary function. Nevertheless, the intuition of optimal value functions and their polyhedral approximations should be often invoked and will act as a guide throughout this chapter.

We open this chapter in Section 5.1 with some illustrations of families of non-convex functions that become convex when averaged.

Section 5.2 introduces the gap of a function, a concept closely related to the duality gap from Section 3.2.2, and demonstrate how it can be naturally used to measure a function's non-convexity. This section's main result is Theorem 5.3, stating that for a random function Q, the gap of the average $\mathbb{E}[Q]$ is always the average of the gap on all scenarios. Then these concepts are applied for the special case of additive noises where sharper results can be estimated.

In Section 5.3, we remember the results of Chapter 4 to introduce another natural way to measure a continuous function's non-convexity: the negative part of its second distributional derivative. It is noted that an interesting similarity exists between the results in this section and those already deduced for the gap.

Section 5.4 is dedicated to the similarities between the two ways to measure a function's non-convexity that were previously introduced. It is shown that both are examples of a certain type of cone convex functions that we will call *non-convexity measures*. From this generalized point of view, the results from the previous sections can be seem as applications of the convexity of these operators.

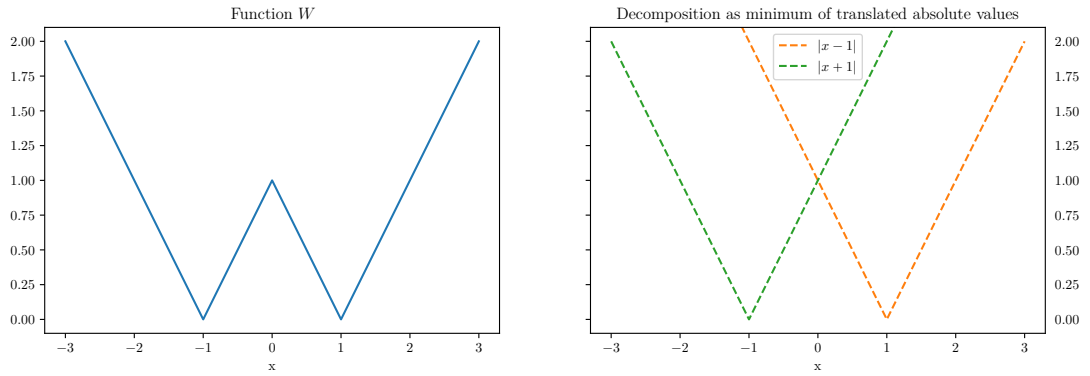


Figure 5.1: The function W and its decomposition as minimum of absolute values.

5.1 A pictorial discussion

In this section, we illustrate how a non-convex unidimensional function can become convex when subject to noise. The approach is rather informal and focus on visualizing the functions. For now, every time we say that a function becomes less non-convex, it is should be interpreted as saying that its graph resembles more that of a convex function. Afterwards, we will rigorously define how to measure the non-convexity of a function and explain everything that is illustrated here.

For this discussion, we will work with the function

$$W(x) = \min\left\{\,|x+1|\,,|x-1|\,\right\} \tag{5.1}$$

that is represented in Figure 5.1 and could also be realized as the optimal value function of a mixed integer linear program as

$$\begin{aligned} W(x) = \min_{z,y,t} \quad & t \\ \text{s.t.} \quad & y = 2z + x - 1 \\ & y \leqslant t \\ & -y \leqslant t \\ & z \in \{0,1\}. \end{aligned}$$

The noise considered will be a random variable ξ which represents the fact that we may not know the argument with certainty. Thus, we will look at a random cost-to-go function

$$\begin{aligned} Q(x,\xi) = W(x-\xi) = \min_{z,y,t} \quad & t \\ \text{s.t.} \quad & y = 2z + x - \xi - 1 \\ & y \leqslant t \\ & -y \leqslant t \\ & z \in \{0,1\} \end{aligned} \tag{5.2}$$

that for each realization of ξ, translates the argument of the function W.

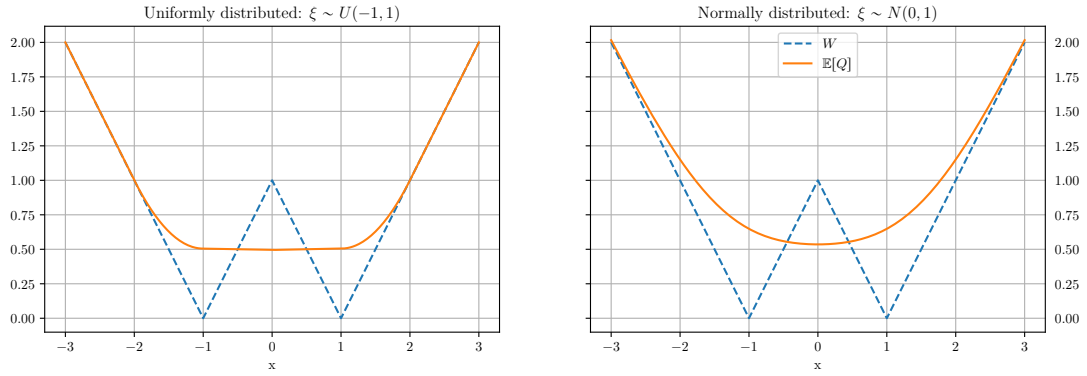


Figure 5.2: On the left, the function W and the expected function when the uncertainty is $\xi \sim U(-1, 1)$ and, on the right, the expected function when the uncertainty is $\xi \sim N(0, 1)$.

As a first example, let us consider that ξ is uniformly distributed in the interval $[-1, 1]$. In this case, the expected cost-to-go $\mathbb{E}[Q](x)$ equals the average of the function W on the interval $[x - 1, x + 1]$,

$$\mathbb{E}[Q(x, \xi)] = \mathbb{E}[W(x - \xi)] = \frac{1}{2} \int_{x-1}^{x+1} W(y)\, dy.$$

The non-convexities cancel out and the expected function is convex. This is illustrated on the left image of Figure 5.2. Informally, we can think that by taking this average, the peak at $x = 0$ is lowered while the valleys at $x = \pm 1$ rise in such a way that the region of non-convexity of the function W disappears.

Another typical uncertainty to consider is when the random variable is distributed as standard normal, $\xi \sim N(0, 1)$. Again, the convexification is total. The expected cost-to-go is a infinitely differentiable convex function, as seem on the right image of Figure 5.2.

Of course, it is not guaranteed that an arbitrary uncertainty can convexify a function. As we will see throughout this chapter, there is a certain relation between how "disperse" is the distribution of ξ and how much it convexifies when averaged over. This is illustrated in Figures 5.3 and 5.4, where we progressively augment the uncertainty for a family of uniformly and normally distributed random variables, respectively. In both these pictures, we see that distributions that are too concentrated will reduce the function's non-convexity but do not totally convexify it, while distributions that are disperse enough tend to generate for a better convexification.

Finally, in Figure 5.5, we show how the average over a discrete distribution supported on the interval $[-1, 1]$ gets less non-convex as these distributions become denser.

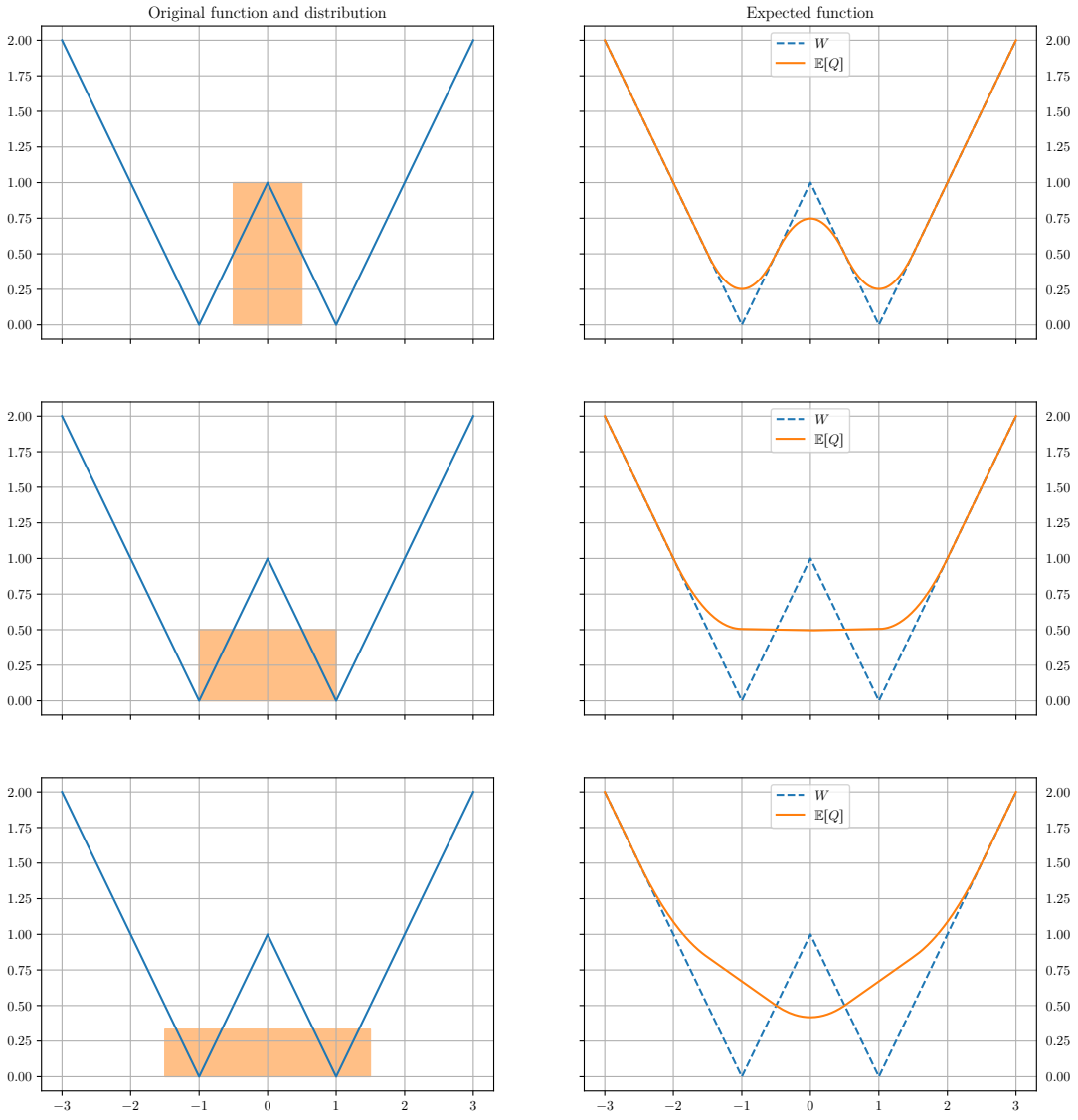


Figure 5.3: Example of convexification for different uniform distributions.

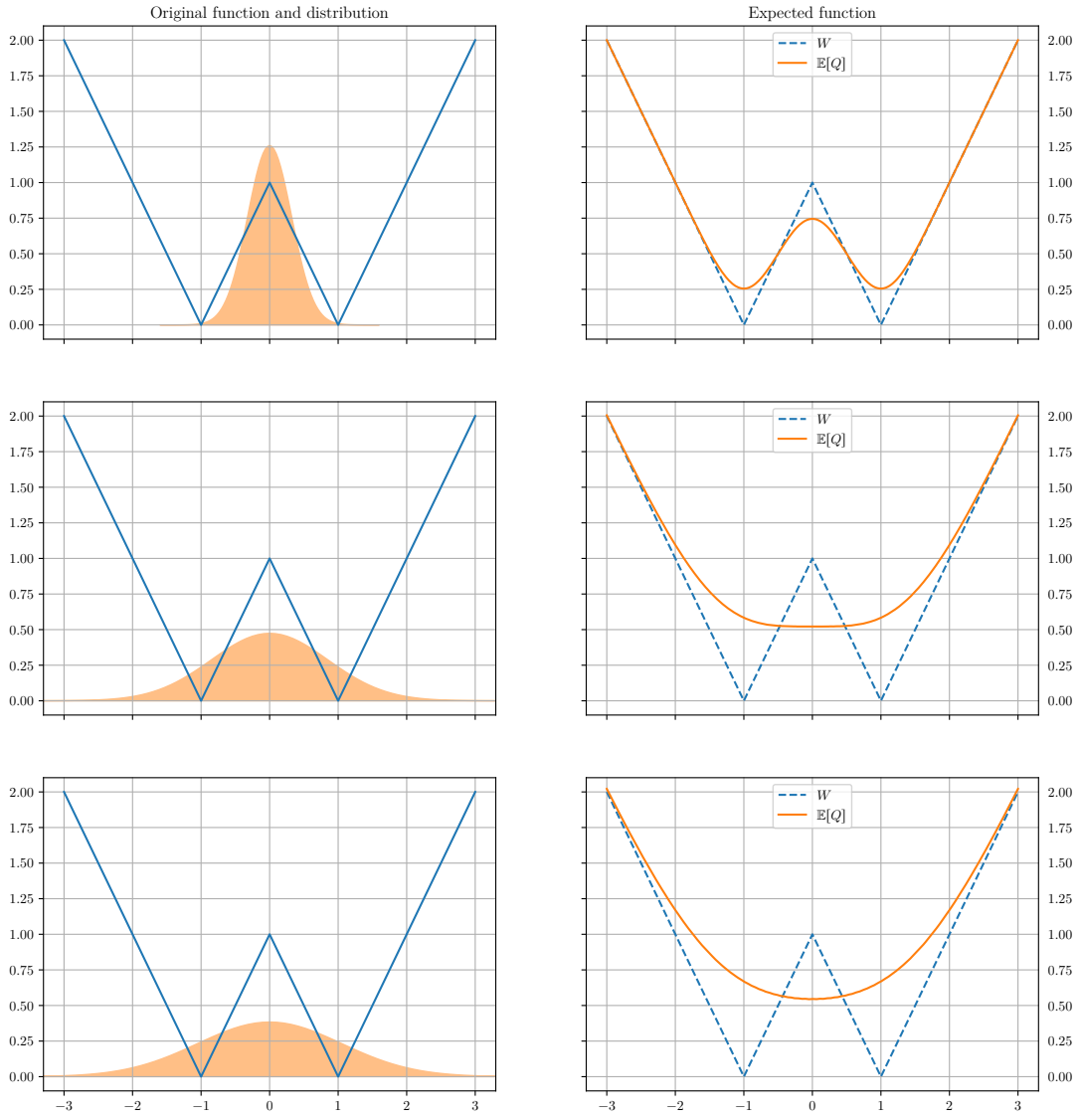


Figure 5.4: Example of convexification for different Normal distributions.

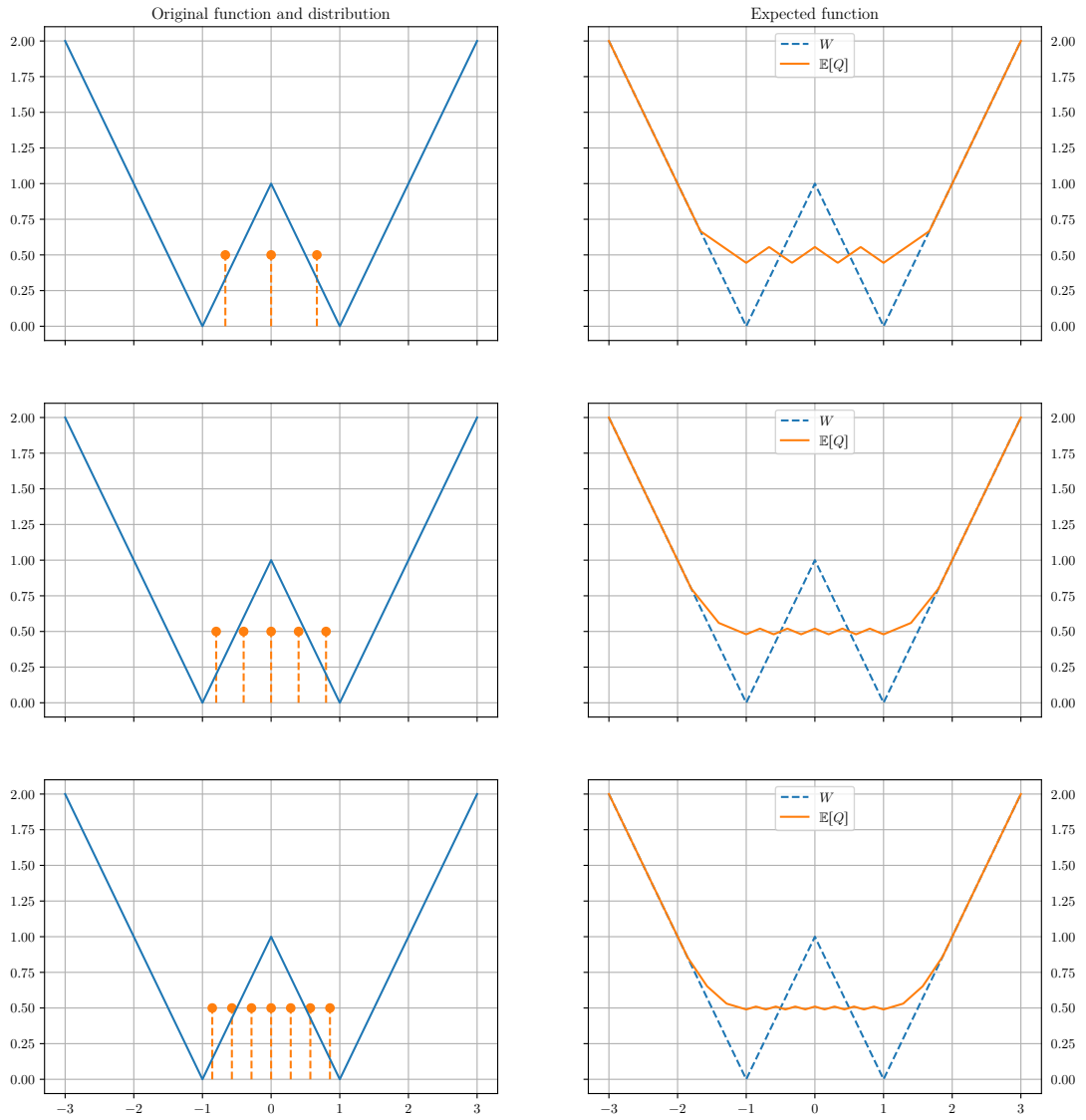


Figure 5.5: Example of convexification for different discrete distributions, approximating a uniform on the interval $[-1, 1]$.

5.2 The gap as a measure of non-convexity

For any extended real function f, we may consider the gap between it and its convex relaxation. Besides its pictorial interpretation as $\operatorname{epi} \check{f} \setminus \operatorname{epi} f$, we may look at it as the deviation of $f(x)$ from $\check{f}(x)$ for each fixed point f. This motivates the definition of the gap as

$$\operatorname{gap}(f) = f - \check{f}.$$

Although it works when both f and $\check{f}$ are finite, this expression is ill-defined if both $\check{f}$ and f are $\pm\infty$ at the same time. Since in this case both functions are equal, the geometrical interpretation of the gap prompts us to consider the gap to be zero at these points. This may be intuitively thought as "focalizing" in the region where f and $\check{f}$ differ. Definition 5.1 takes this technicality into account for defining the gap of a function.

Definition 5.1. The *gap of* f is the function defined by

$$\operatorname{gap}(f)(x) = \begin{cases} f(x) - \check{f}(x), & \text{if } f(x) \neq \check{f}(x) \\ 0, & \text{otherwise.} \end{cases}$$

The gap has two rather simple yet useful properties that turn it into a good metric to measure the non-convexity of a function. One is that $\operatorname{gap}(f)$ is zero if and only if f is a convex function, by the very definition of the convex relaxation. The other is that for any function f, $\operatorname{gap}(f)$ is a non-negative function, since $\check{f}$ is always below f.

Definition 5.2. We shall say that a function f *is less non-convex* than a function g if

$$\operatorname{gap}(f) \leqslant \operatorname{gap}(g).$$

For a function f to be less non-convex than a function g, it must be closer to its convex relaxation than g for each fixed point on their domain. Figure 5.7 shows an example where a function f is less non-convex than g together with their respective gaps.

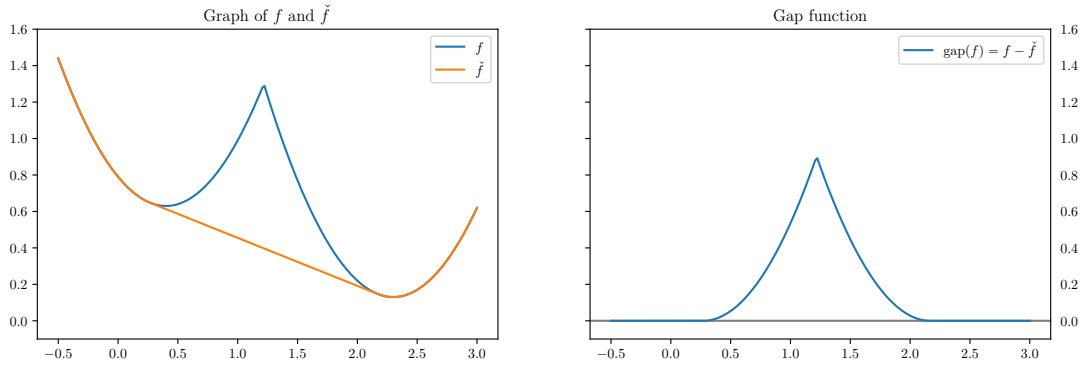


Figure 5.6: Example of gap function for some non-convex function.

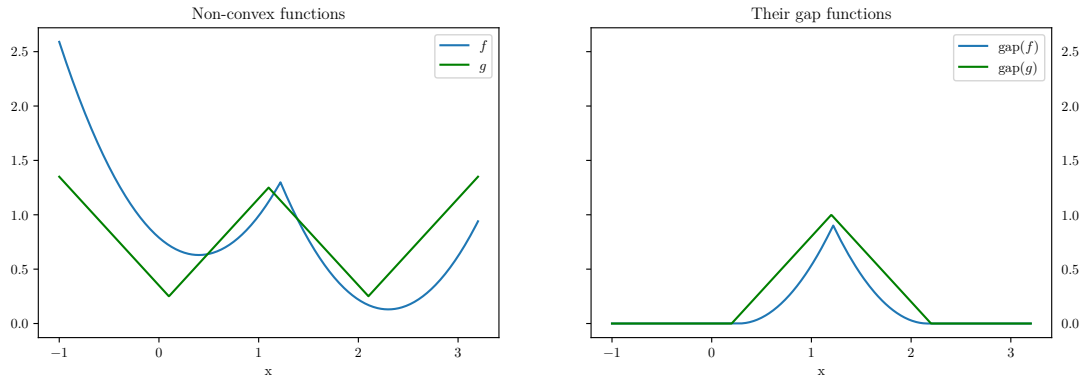


Figure 5.7: Example of two non-convex functions were f is less non-convex than g.

Consider a random function Q. In Section 3.4.1, we introduced the notation $\check{Q}$ to be the random function whose realizations are the convex relaxation of each realization of Q and the notation $\mathbb{E}\left[Q\right]$ to be the function $x \mapsto \mathbb{E}^{\xi}\left[Q(x,\xi)\right]$. These notations will be thoroughly used in this chapter.

Recall the discussion from Section 3.4.2 about the distinction between the decomposed and the linked formulation for the expected cost-to-go of a stochastic program. Theorem 5.3 is a version of it for an arbitrary random function. When calculation cuts, the difference difference of the two formulations lies in the fact that decomposed cuts are at best tight for $\mathbb{E}\left[\check{Q}\right]$ while linked cuts can be tight for $\widetilde{\mathbb{E}\left[Q\right]}$.

Theorem 5.3. *For any random function Q, the following functional inequality holds:*

$$\mathbb{E}\left[Q\right] - \widetilde{\mathbb{E}\left[Q\right]} \leqslant \mathbb{E}\left[Q\right] - \mathbb{E}\left[\check{Q}\right].$$

Proof. Notice that since for each value ξ,

$$\check{Q}(\cdot,\xi) \leqslant Q(\cdot,\xi),$$

the average preserves this inequality:

$$\mathbb{E}[\check{Q}] \leqslant \mathbb{E}\left[Q\right].$$

The average of convex functions is again a convex function, hence $\mathbb{E}[\check{Q}]$ is convex. By the Definition 2.26 of the convex relaxation, $\widetilde{\mathbb{E}\left[Q\right]}$ is greater than every convex function below $\mathbb{E}\left[Q\right]$, including $\mathbb{E}\left[\check{Q}\right]$. That is,

$$\mathbb{E}\left[\check{Q}\right] \leqslant \widetilde{\mathbb{E}\left[Q\right]} \leqslant \mathbb{E}\left[Q\right],$$

which implies

$$\mathbb{E}\left[Q\right] - \widetilde{\mathbb{E}\left[Q\right]} \leqslant \mathbb{E}\left[Q\right] - \mathbb{E}\left[\check{Q}\right]. \qquad \square$$

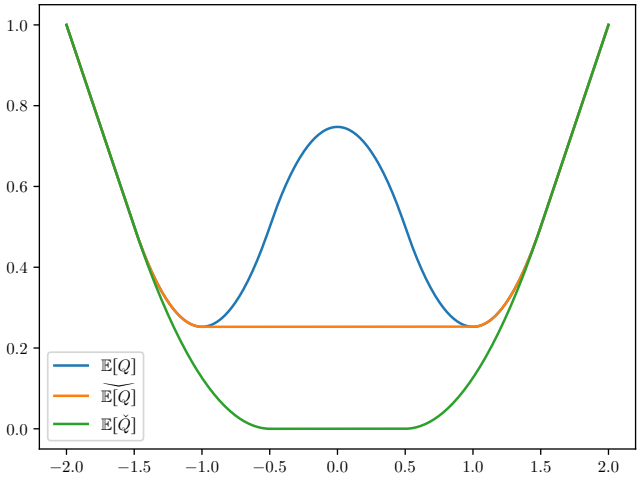


Figure 5.8: Comparison between $\mathbb{E}[\check{Q}]$ and $\widetilde{\mathbb{E}[Q]}$ as underapproximations of $\mathbb{E}[Q]$. The function Q is the that from (5.2) with an uncertainty $\xi \sim U(-0.5, 0.5)$.

In terms of the gap function, the result of this theorem is expressed as

$$\operatorname{gap}(\mathbb{E}[Q]) \leqslant \mathbb{E}[\operatorname{gap}(Q)]. \tag{5.3}$$

That is, the gap of the average of random functions is below the average of the gap of all possible realizations.

Notice that this form of the theorem says that the gap is a convex function in relation to the cone of non-negative functions. Now that we have this in mind, we may give another equivalent proof of Theorem 5.3 which is much more succinct, though more abstract.

We begin by remembering that the function $f \mapsto \check{f}$ is concave in relation to the non-negative cone, which we will denote by K. As the product of cones is itself a cone, this means that the function $f \mapsto (f, -\check{f})$ is K^2-convex. The gap may be decomposed as $(+) \circ (f \mapsto (f, -\check{f}))$, which is the composition of a K^2-convex function with an affine and (K^2, K)-monotone function, therefore K-convex.

5.2.1 Quantifying the non-convexity

Equation (5.3) tells us that the gap of the average function is below the average gap for each fixed argument. This is a *pointwise* theorem about convexification. In general, pointwise comparisons as that of Definition (5.2) are too rigid, since they must hold for every point, which implies that many functions may not be comparable, even if one is intuitively more convex than the other.

To solve this problem, we need a way to *globally* quantify how much non-convex a given function is. That is, we have to project the gap of a function on a totally ordered set in a manner that is compatible with the local inequalities. A

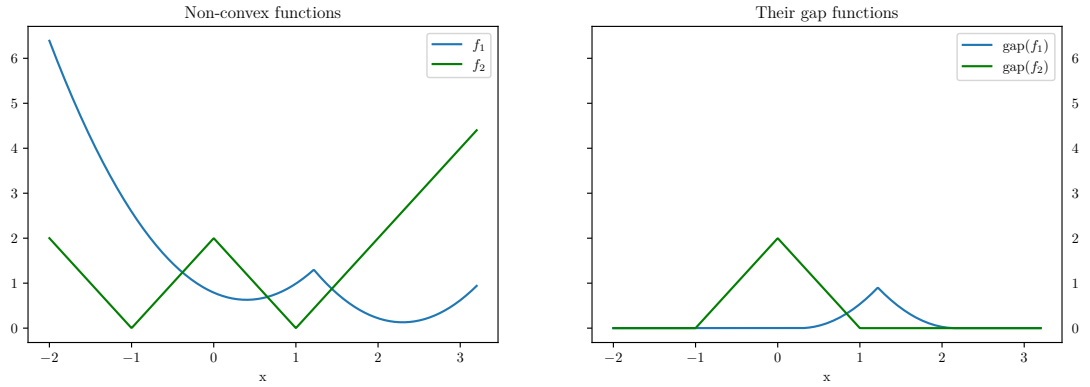


Figure 5.9: The function f is intuitively more non-convex than g but their gap functions are not comparable.

way to accomplish this is by considering monotone functionals. In what follows we will use *monotone norms* as a way to project the non-convexity of a function into the non-negative real numbers.

Definition 5.4. A function norm $\|\cdot\|$ is said to be *monotone* if for any pair of non-negative functions $g, h \geqslant 0$,

$$g \leqslant h \implies \|g\| \leqslant \|h\| .$$

Remark 5.1. When working with norms there is a technical aspect we must consider: the gap function is well-defined and non-negative for any arbitrary function f but, generally, the norm $\|\cdot\|$ is only defined on some smaller subspace, on which $\operatorname{gap}(f)$ may not lie on. This may occur because of measurability or boundedness issues, for example. In this case we will say that $\|\operatorname{gap}(f))\| = \infty$. In applications, we generally work with smaller function spaces, such as continuous or integrable functions. Since the intersection of the non-negative cone with a function space is the non-negative cone in that space, no problems related to $\|\operatorname{gap}(f)\|$ being infinite should occur.

Since monotone norms respect functional inequalities between non-negative functions, that is,

$$\operatorname{gap}(f) \leqslant \operatorname{gap}(g) \implies \|\operatorname{gap}(f)\| \leqslant \|\operatorname{gap}(g)\| , \tag{5.4}$$

they preserve our previous notion, from Definition 5.2, of a function being more non-convex than another.

A special application is to project the result of Theorem 5.3, which tells us that for any monotone norm

$$\left\|\mathbb{E}[Q] - \widetilde{\mathbb{E}[Q]}\right\| \leqslant \|\mathbb{E}[Q] - \mathbb{E}[\check{Q}]\| . \tag{5.5}$$

As we will see below, there are geometrical relations between the graphs of $\mathbb{E}[Q]$, $\widetilde{\mathbb{E}[Q]}$ and $\mathbb{E}[\check{Q}]$ which may be interpreted as applications of these norms.

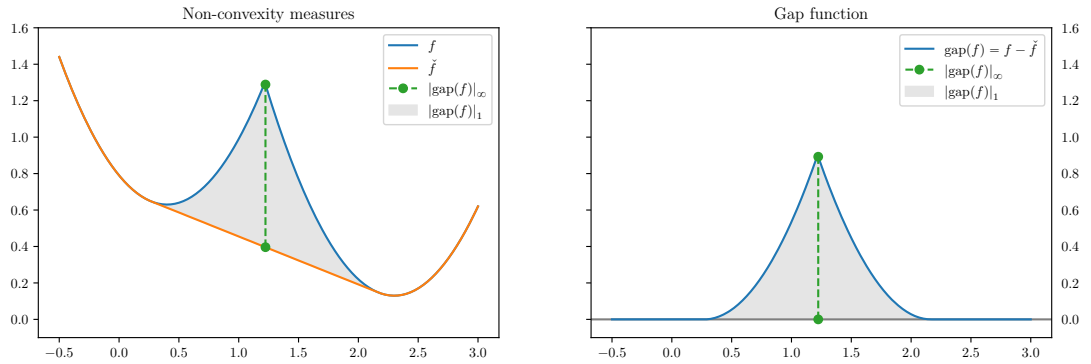


Figure 5.10: An illustration of $\|\mathrm{gap}(f)\|_1$ and $\|\mathrm{gap}(f)\|_\infty$.

When properly applicable, we may use monotone norms to represent certain characteristics of the region between the graphs of two functions. The first example is when we have a function f and some lower approximation $h \leqslant f$. Then, the volume of the region $\mathrm{epi}(h) \setminus \mathrm{epi}(f)$ may be calculated as

$$\mathrm{Vol}(\mathrm{epi}(h) \setminus \mathrm{epi}(f)) = \int |f(x) - h(x)| \, dx = \|f - h\|_1 \tag{5.6}$$

provided that the function $f - h$ is integrable, of course. This is a monotone norm on the space of integrable functions and, therefore, preserves functional inequalities as well as theorem 5.3. In this context, we have that $\mathbb{E}\left[\check{Q}\right]$ is the convex under-approximation closest to $\mathbb{E}\left[Q\right]$ and, consequently, the one which minimizes the volume of the region between them both.

The norm $\|\cdot\|_1$ is a special case of the L^p-norms, defined by

$$\|f\|_p := \left(\int |f(x)|^p \, dx \right)^{\frac{1}{p}}, \; p \in [1, +\infty). \tag{5.7}$$

All these are monotone norms and play a significant role in many results and applications of functional analysis.

Another monotone norm that is useful to consider is the *uniform norm*, defined by

$$\|g\|_\infty := \sup_x |g(x)| \, . \tag{5.8}$$

If $h \leqslant f$ is a lower approximation of f, we can interpret $\|f - h\|_\infty$ as the worst error h commits on their entire domain.

In general a norm which depends only on the evaluations of f at given points will be monotone. An example of a non-monotone norm is given by the C^1 norm, defined by

$$\|f\| = \|f\|_\infty + \|f'\|_\infty \, ,$$

which equals the supremum of a function added to the supremum of its derivative. Since a function may be arbitrarily small while its derivative is arbitrarily large, this norm is not monotone.

5.2.2 Additive noise

Until now, we had no major hypothesis on the manner the uncertainty operated upon our random function. In what follows, we will focus our attention on the case where we have some base function and the uncertainty acts as translations of its argument. In other words, there is a function f and a random variable ξ such that

$$Q(x, \xi) = f(x - \xi). \tag{5.9}$$

For these problems, some finer results may be deduced.

To properly take advantage of the structure implied by equation (5.9) we will make one more assumption about the way we quantify the non-convexity of a function. In conjunction with monotonicity, we will also require that the $\|\cdot\|$ is translation invariant, as we define shortly in 5.6. Notice that all the L^p-norms of (5.7) are translation invariant, as well as the uniform norm in (5.8)

Definition 5.5. The *translation operator by a* is the function τ_a that takes a function f and returns its shift by a, that is,

$$(\tau_a f)(x) = f(x - a).$$

Definition 5.6. A function norm $\|\cdot\|$ is said to be *translation invariant* if for any function $f\colon V \to [-\infty, +\infty]$ and any $a \in V$,

$$\|f\| = \|\tau_a f\|.$$

Since the gap function does not change by the addition of a constant,

$$\operatorname{gap}(f + a) = (f + a) - \widetilde{f + a} = f + a - (\check{f} + a) = \operatorname{gap}(f),$$

we already have translation invariance on the function's image. This gives an interesting geometrical interpretation to the norm of the gap function. If $\|\cdot\|$ is translation invariant, $\|\operatorname{gap}(f)\|$ is a way to quantify the non-convexity of f that only depends on the geometric properties of the set $\operatorname{conv}(\operatorname{epi}(f)) \setminus \operatorname{epi} f$ and not on where in the space it is, which agrees with an intuitive idea about determining how non-convex a function is. Figure 5.11 illustrates this. Notice that this reasoning only applies to *translations* of the region, because since the epigraphs have a special direction, rotations and other transformations should not preserve how non-convex a function is considered to be.

This section's main result is Theorem 5.7 below, which says that if we consider a function f subjected to additive noise, then the gap of its average $\mathbb{E}[f]$ is always less than or equal to the original gap when quantified through a translation invariant monotone norm.

Theorem 5.7. *Let $\|\cdot\|$ be a translation invariant monotone norm. Then, for any function f subjected to an additive noise ξ,*

$$\left\|\mathbb{E}[\tau_\xi f] - \widetilde{\mathbb{E}[\tau_\xi f]}\right\| \leqslant \|\mathbb{E}[\tau_\xi f] - \mathbb{E}[\tau_\xi \check{f}]\| \leqslant \|f - \check{f}\|.$$

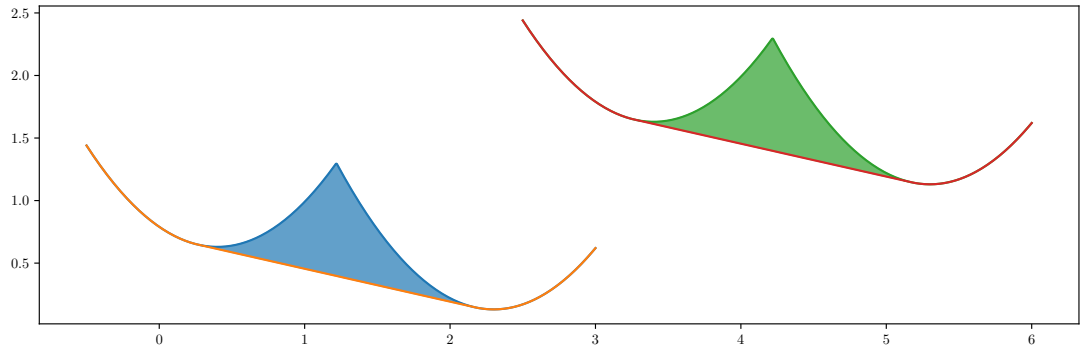


Figure 5.11: On this image, we perceive both functions as being equally non-convex. Translation invariant norms capture this intuitive notion.

In other words, if Q is a random function of the form $Q(x, \xi) = f(x - \xi)$, then

$$\left\|\mathbb{E}\left[Q\right] - \widetilde{\mathbb{E}\left[Q\right]}\right\| \leqslant \left\|\mathbb{E}\left[Q\right] - \mathbb{E}\left[\check{Q}\right]\right\| \leqslant \left\|f - \check{f}\right\|.$$

Proof. Since $\|\cdot\|$ is monotone and $\widetilde{\tau_\xi(f)} = \tau_\xi \check{f}$, the first inequality is a special case of Theorem 5.3 and Equation (5.5). The second inequality will follow as a consequence of the norm's translation invariance. For this, notice that, by linearity of $\mathbb{E}$ and τ_ξ,

$$\left\|\mathbb{E}\left[\tau_\xi f\right] - \mathbb{E}\left[\tau_\xi \check{f}\right]\right\| = \left\|\mathbb{E}\left[\tau_\xi (f - \check{f})\right]\right\|.$$

Since any norm is a convex function, the norm of the average is always below the average of the norm by Jensen's inequality (Theorem 2.18). Therefore,

$$\left\|\mathbb{E}\left[\tau_\xi (f - \check{f})\right]\right\| \leqslant \mathbb{E}\left[\left\|\tau_\xi (f - \check{f})\right\|\right] = \mathbb{E}\left[\left\|f - \check{f}\right\|\right] = \left\|f - \check{f}\right\|.$$

Putting all the inequalities together, we conclude the theorem. □

Representing additive noise through convolutions Given a random variable ξ, call μ its probability density. We can characterize the average function $\mathbb{E}\left[\tau_\xi f\right]$ as the convolution between f and μ.

Recall the results from Section 4.2.3. We may look at both the function f and the measure μ as distributions. Using that taking the average of a function of random variable ξ consists of integrating it with respect to the probability density μ, which gives precisely the expression of a convolution,

$$\mathbb{E}\left[f(x - \xi)\right] = \int f(x - y) d\mu(y) = f * \mu(x). \tag{5.10}$$

As a first use, we will rewrite the result of Theorem 5.7 in terms of convolutions. That is, given a function $f \colon V \to [-\infty, +\infty]$ and a probability measure μ over V, the following inequalities hold for any translation invariant monotone norm:

$$\left\|f * \mu - \widetilde{f * \mu}\right\| \leqslant \left\|f * \mu - \check{f} * \mu\right\| \leqslant \left\|f - \check{f}\right\|. \tag{5.11}$$

5.2.2.1 Uniform norm and asymptotic behavior

A question that arises about the inequalities in theorem 5.7 is whether they are *strict* for a given probability distribution μ. The answer depends on which norm we choose to quantify the non-convexity but is affirmative when we use the uniform norm $\|\cdot\|_\infty$, under some assumptions on μ.

As a counterexample, we begin with the integral norm $\|\cdot\|_1$. In this case, the second inequality in Theorem 5.7 is always an equality for any integrable function f, no matter the distribution μ,

$$\|f * \mu - \check{f} * \mu\|_1 = \|f - \check{f}\|_1 . \tag{5.12}$$

To see this, let's develop the expression for the norm on the left:

$$\|f * \mu - \check{f} * \mu\|_1 = \int (f - \check{f}) * \mu(x)\, dx = \int\int f(x-y) - \check{f}(x-y)\, d\mu(y)\, dx.$$

By Fubini's theorem we may change the order of the integrals on the right. From this and the fact that $f - \check{f}$ is non-negative:

$$\begin{aligned}
\|f * \mu - \check{f} * \mu\|_1 &= \int\int f(x-y) - \check{f}(x-y)\, dx\, d\mu(y) \\
&= \int \left(\int |f(x) - \check{f}(x)|\ dx \right)\ d\mu(y) \\
&= \int \|f - \check{f}\|_1\ d\mu(y) \\
&= \|f - \check{f}\|_1 .
\end{aligned}$$

Remark 5.2. Even though Equation (5.12) says that the second inequality in Theorem 5.7 is in fact an equality, convexification may still occur due to Theorem 5.3.

For the uniform norm $\|\cdot\|$, the result of Theorem 5.8 below tells us that there is a better bound on theorem 5.7 when the image of the random function ξ is sparse enough in relation to the support of the gap function.

Theorem 5.8. *Let f be a function subject to an additive noise with probability density μ. If we call $K = \operatorname{supp}(f - \check{f})$ (the region where $f - \check{f}$ is not zero) and let*

$$\kappa = \sup_{x \in V} \mu(x - K) \leqslant 1,$$

then

$$\left\| f * \mu - \widetilde{f * \mu} \right\|_\infty \leqslant \|f * \mu - \check{f} * \mu\|_\infty \leqslant \kappa \|f - \check{f}\|_\infty .$$

Proof. Because of the definition of support, the function $f - \check{f}$ is equal to $(f - \check{f}) \cdot \mathbb{1}_K$. This means that integrating $f - \check{f}$ on the entire space is equal to integrating

it only on K. Therefore, for any point x fixed,

$$\begin{aligned}(f-\check{f})*\mu(x) &= \int f(y)-\check{f}(y)\,d\mu(x-y) = \int (f(y)-\check{f}(y))\cdot \mathbb{1}_K(y)\,d\mu(x-y)\\ &= \int_K f(y)-\check{f}(y)\,d\mu(x-y)\\ &\leqslant \int_K \|f-\check{f}\|_\infty\,d\mu(x-y) = \|f-\check{f}\|_\infty \int_K d\mu(x-y)\\ &\leqslant \|f-\check{f}\|_\infty \sup_{x\in V}\mu(x-K).\end{aligned}$$

Since $f-\check{f}$ is non-negative, $\|f-\check{f}\|_\infty$ equals the supremum of $f-\check{f}$. This proves the theorem, for

$$\|(f-\check{f})*\mu\|_\infty \leqslant \|f-\check{f}\|_\infty \sup_{x\in V}\mu(x-K). \qquad \square$$

Theorem 5.8 is specially useful when f only differs from its convex relaxation inside of a compact set. In this case, any probability distribution whose support is sparse enough will have an associated κ smaller than one, since no translation of $\operatorname{supp} f-\check{f}$ is going to have total probability.

If we have a sequence of probabilities μ_n that are representable by bounded functions, we can use Hölder's inequality to deduce uniform bounds on the non-convexity on $f*\mu_n$ as n tends to infinity. This is the content of Theorem 5.9, together with its corollaries, which apply this idea to some commonly found distributions.

Theorem 5.9. *Suppose $\mu_k \in L^\infty(\mathbb{R}^n)$ for each $k\in\mathbb{N}$ and $f\colon \mathbb{R}^n \to [-\infty,+\infty]$ is a function whose gap is integrable. Then*

$$\|f*\mu_k - \check{f}*\mu_k\|_\infty \leqslant \|f-\check{f}\|_1\,\|\mu_k\|_\infty$$

and $\operatorname{gap}(f*\mu_k)$ *converges uniformly to zero if* $\|\mu_k\|_\infty \to 0$.

Proof. For each fixed $x\in\mathbb{R}^n$,

$$\operatorname{gap}(f)*\mu_k(x) = \int_{\mathbb{R}^n} (f-\check{f})(y)\mu_k(x-y)\,dy.$$

Applying Hölder's inequality to the integral on the right,

$$\operatorname{gap}(f)*\mu_k(x) \leqslant \|f-\check{f}\|_1\,\|\mu_k\|_\infty\,.$$

Since the uniform norm of a non-negative function is its supremum, the equation above gives a uniform bound:

$$\|\operatorname{gap}(f)*\mu_k\|_\infty \leqslant \|f-\check{f}\|_1\,\|\mu_k\|_\infty\,.$$

Furthermore, if $\|\mu_k\|_\infty$ converges to zero, this equation together with the gap inequality from Theorem 5.3 says that

$$0 \leqslant \lim_{k\to\infty} \|\operatorname{gap}(f * \mu_k)\|_\infty \leqslant \lim_{k\to\infty} \|\operatorname{gap}(f) * \mu_k\|_\infty \leqslant 0.$$

This proves the uniform convergence. □

Corollary 5.9.1. Let μ_k be the uniform distributions on a sequence C_k of subsets of $\mathbb{R}^n$ such that $\operatorname{Vol}(C_k)$ converges to infinity. Then, if f is a function whose gap is integrable, the sequence $\operatorname{gap}(f * \mu_k)$ converges uniformly to zero.

Proof. We have $\|\mu_k\|_\infty = \operatorname{Vol}(C_k)^{-1}$, which converges to zero. □

Corollary 5.9.2. Let ξ_k be a sequence of gaussian random variables with density μ_k and covariance matrix Σ_k. Then, if $f\colon \mathbb{R}^n \to [-\infty, +\infty]$ is a function with integrable gap,

$$\|f * \mu_k - \check{f} * \mu_k\|_\infty \leqslant \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma_k}} \|f - \check{f}\|_1 .$$

In special, if $\det \Sigma_k$ converges to infinity, $\operatorname{gap}(f * \mu_k)$ converges uniformly to zero.

Proof. The density of a multivariate Gaussian with mean m_k and covariance matrix Σ_k is

$$\mu_k(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma_k}} \exp\left(-\frac{1}{2}(x - m_k)^\top {\Sigma_k}^{-1}(x - m_k)\right).$$

Its maximum is attained when $x = m_k$ and equals

$$\|\mu_k\|_\infty = \mu_k(m_k) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma_k}}.$$

The remnant follows from Theorem 5.9. □

At first it may seem that the results of theorem 5.9 and its corollaries could be applied for both f and $\check{f}$ separately, implying that the gap converges to zero because both $f * \mu_k$ and $\check{f} * \mu_k$ converge to zero. However, it is possible for the gap of f to be integrable while both f and $\check{f}$ are not.

Since a convex function is integrable on the entire $\mathbb{R}^n$ if and only if it is constant, the function $\check{f}$, in general, cannot be integrated. If the gap of f is finite and compactly supported, f and $\check{f}$ differ only in a compact set and, therefore, f also cannot be integrated while $\operatorname{gap}(f)$ is certainly integrable. If we look back to the graphics in Figure 5.6, we see an example of a non-integrable function whose gap is integrable.

5.3 Second derivative's negative part

Throughout Section 5.2, we measured the non-convexity of a function f via its gap function $\operatorname{gap}(f) = f - \check{f}$. Since the convex relaxation is well-defined for any function, the results in that section are appropriate in the general case.

In this section, we restrict ourselves to consider only the space $C^0(\mathbb{R})$ of continuous functions of a single real variable. This restriction allows us to take another route about how to measure the non-convexity of a function: the negative part of its second derivative. Initially, it may seem too restrictive to work only with single variable functions but the importance of this section lies in the analogies between the theorems in here and those of Section 5.2. As we will see, every theorem we proved for $\operatorname{gap}(f)$ has a corresponding one for the negative part of f''. Later, in Section 5.4, we will explain the relation of between the gap and the second derivative's negative part, which will lead to the definition of a general *non-convexity measure*. In that section we will also show equivalent results for the Hessian of a multivariate function.

We start by recalling some theorems from Chapter 4. Theorems 4.23 and 4.24 said that a distribution f is a continuous function if and only if its second derivative f'' (in the distributional sense) is represented by a measure and Theorem 4.27 says that f is in fact a convex function if and only if f'' is a *non-negative measure*. Finally, the Hahn-Jordan decomposition from Theorem 4.9 allows us to write the second derivative of any continuous function as a difference between two singular non-negative measures:

$$f'' = [f'']_+ - [f'']_-. \tag{5.13}$$

Since f'' is a non-negative measure if and only if its negative part is zero, we see that

$$f \text{ is convex} \iff [f'']_- = 0.$$

The discussion above motivates us to look at $[f'']_-$ as a tool to evaluate how non-convex a continuous function is. Interestingly, our main result is theorem 5.10, which says that for any random continuous function, the negative part of the second derivative of its average is always below the average of the negative part of each realization's second derivative. Notice the parallel between this and Theorem 5.3, which states the same for the gap function.

Theorem 5.10. *Let $Q\colon \mathbb{R} \times \Omega \to \mathbb{R}$ be a* continuous *random function. Then, the following measure inequality holds:*

$$\left[\mathbb{E}\,[Q]''\right]_- \leqslant \mathbb{E}\,[[Q'']_-]$$

Proof. For each fixed ω, $Q(\cdot, \omega)$ is a continuous function, which implies that its second derivative is a measure. This means that Q'' is a random measure and, for each fixed ω, we may apply Theorem 4.9 and write $Q''(\cdot, \omega)$ as its Hahn-Jordan decomposition. That is, as differences of mutually singular non-negative measures. Besides that, the function $\mathbb{E}^\omega\,[Q(\cdot, \omega)]$ is also continuous so its second

derivative admits a Hahn-Jordan decomposition of its own. We write each of these decompositions as

$$Q(\cdot,\omega)'' = \lambda_+(\omega) - \lambda_-(\omega),\ \omega \in \Omega$$
$$\mathbb{E}\left[Q\right]'' = \nu_+ - \nu_-.$$

Besides these, we can find another decomposition for $\mathbb{E}\left[Q\right]''$ passing the derivative inside the expectation

$$\mathbb{E}\left[Q\right]'' = \mathbb{E}\left[Q''\right] = \mathbb{E}\left[\lambda_+ - \lambda_-\right] = \mathbb{E}\left[\lambda_+\right] - \mathbb{E}\left[\lambda_-\right].$$

By the minimality property of the Hahn-Jordan decomposition (Theorem 4.9.2), the measures ν_+ and ν_- are smaller than the components of any other decomposition of $\mathbb{E}\left[Q\right]''$. In particular,

$$\nu_- \leqslant \mathbb{E}\left[\lambda_-\right].$$

That is,

$$\left[\mathbb{E}\left[Q\right]''\right]_- \leqslant \mathbb{E}\left[\left[Q''\right]_-\right]. \qquad \square$$

Remark 5.3. Since the objects we deal with in here are not ordinary functions from $\mathbb{R}$ to $\mathbb{R}$, the inequality of Theorem 5.10 can be interpreted in two different but equivalent ways. As a measure inequality, it says that for any Borel subset B of $\mathbb{R}$,

$$\left[\mathbb{E}\left[Q\right]''\right]_-(B) \leqslant \mathbb{E}\left[\left[Q''\right]_-\right](B). \tag{5.14}$$

As a distributional inequality, it says that for any *positive*, infinitely differentiable and compactly supported function ϕ,

$$\left\langle\left[\mathbb{E}\left[Q\right]''\right]_-, \phi\right\rangle \leqslant \left\langle\mathbb{E}\left[\left[Q''\right]_-\right], \phi\right\rangle. \tag{5.15}$$

Even though these inequalities are equivalent, they are defined with respect to the non-negative cones of different spaces.

Remember Definition 4.11 where we saw that there is a natural norm for signed measures which turns the set of all finite signed measures into a Banach space. It is called the total variation norm and given by

$$\left\|\mu\right\|_1 = \left|\mu\right|(\mathbb{R}). \tag{5.16}$$

Notice that, if μ has a density which is a function, this norm is equal to the integral norm of (5.6). The norm in Equation (5.16) is monotone and therefore preserves the inequality of Theorem 5.10. In the general case, any monotone norm will preserve that inequality (Corollary 5.10.1 below), thus if we know any additional information about $\left[Q''\right]_-$, it may be more suitable to consider other norms.

Corollary 5.10.1. Let $\|\cdot\|$ be monotone in relation to the cone of non-negative measures (or distributions). Then, for any random function Q:

$$\left\|\left[\mathbb{E}\left[Q\right]''\right]_-\right\| \leqslant \left\|\mathbb{E}\left[\left[Q''\right]_-\right]\right\|.$$

5.3.1 Additive noise

Continuing our parallel with the results of Section 5.2, we will consider the case of additive noise, in the manner of Section 5.2.2. In what follows, the random function will always be of the form $Q(x,\omega) = f(x-\xi(\omega))$, for some fixed continuous function f and random variable ξ.

Here, as we are working with measures and distributions, it is natural to use the notation of (5.10) to represent an additive noise as a convolution with the probability density of the random variable. We start to use it on Theorem 5.11 below, which, analogously to Theorem 5.7, says that the negative part of $(f*\mu)''$ is always below that of f'', when quantified via a translation invariant norm.

Theorem 5.11. *Let $\|\cdot\|$ be a translation invariant monotone norm, $f\colon\mathbb{R}\to\mathbb{R}$ a continuous function and μ a probability measure over $\mathbb{R}$. Then*

$$\left\|[(f*\mu)'']_-\right\| \leqslant \|[f'']_- * \mu\| \leqslant \|[f'']_-\|\,.$$

Proof. The monotonicity of the norm $\|\cdot\|$ implies we can use Theorem 5.10.1 to get the first inequality. The invariance by translation, together with the fact that μ is a probability measure, gives

$$\|[f'']_- * \mu\| = \left\|\mathbb{E}^{\xi}\left[\tau_\xi[f'']_-\right]\right\| \leqslant \mathbb{E}^{\xi}\left[\|\tau_\xi[f'']_-\|\right] = \mathbb{E}^{\xi}\left[\|[f'']_-\|\right] = \|[f'']_-\|\,.$$

Putting the inequalities together, we conclude the theorem. □

Based on our previous analogies between $[f'']_-$ and $f-\check{f}$, it is expected that if we follow the same path of Section 5.2 and consider the uniform norm as a way of quantifying the non-convexity of a function, a result sharper than Theorem 5.11 is possible. As envisioned, theorem 5.12 below is a quantitative version of the second inequality in theorem 5.7 with the exact same constant found in Theorem 5.8.

Theorem 5.12. *Let $f\colon\mathbb{R}\to\mathbb{R}$ be a continuous function such that the negative part of its second derivative $[f'']_-$ is an essentially bounded function, that is, $\|[f'']_-\|_\infty < \infty$, and call $K = \operatorname{supp}[f'']_-$. Then, for any probability measure μ we define*

$$\kappa = \sup_{x\in\mathbb{R}} \mu(x-K) \leqslant 1,$$

which gives the bound

$$\left\|[(f*\mu)'']_-\right\|_\infty \leqslant \|[f'']_- * \mu\|_\infty \leqslant \kappa\,\|[f'']_-\|_\infty\,.$$

Proof. Analogous to that of Theorem 5.8 exchanging $f-\check{f}$ with $[f'']_-$. □

Once more we see that if $[f'']_-$ has compact support, sparser probability distributions will guarantee a better convexification. Still, these bounds are too conservative and may not reflect how much a function is really convexified.

Fortunately, in this simpler context of single variable functions, we can explicitly describe the probabilities μ for which $f * \mu$ is a *convex function*. As we will see below, the problem of finding a probability distribution that convexifies a fixed function f can be written as linear feasibility problem in the space of signed measures.

5.3.2 Optimal convexification

In the course of Section 5.1 we saw examples of uncertainties that not only reduced the non-convexity of the function but actually made $\mathbb{E}[f]$ convex. In our context of additive noises of $C^0(\mathbb{R})$ functions, this is equivalent to

$$(f * \mu)'' \geqslant 0,$$

which is a conic inequality in the space of measures and, therefore, describes a convex set.

Theorem 5.13. *For any function $f \in C^0(\mathbb{R})$, the set of probability measures μ which make $f * \mu$ a convex function is a* convex set *defined by the system*

$$\begin{cases} \mu \geqslant 0, \\ \mu(\mathbb{R}) = 1, \\ f'' * \mu \geqslant 0. \end{cases}$$

Proof. Considering the space of all signed measures, the condition that $\mu \colon \mathcal{F} \to [-\infty, +\infty]$ is a probability consists of being non-negative,

$$\mu(B) \geqslant 0, \forall B \in \mathcal{F},$$

and having total variation equal to one, which for non-negative measures is equivalent to $\mu(\mathbb{R}) = 1$.

For the last inequality, remember that $f * \mu$ is convex if and only if its second derivative is a non-negative measure. From Theorem 4.30, we know that taking the derivative of the convolution is the same as taking the derivative of the first term and the making the convolution,

$$f'' * \mu = (f * \mu)'' \geqslant 0.$$

This can be interpreted as set of linear inequalities on μ that must hold for any borelian of $\mathbb{R}$.

Since this set is described only by linear equalities and inequalities on μ, it is convex. □

As an example of these tools, we will use Theorem 5.13 to find a description of the probabilities that convexify a minimum of two translated quadratics, defined by

$$g(x) = \min\left\{(x - a_1)^2, (x - a_2)^2\right\}, \ a_1 \neq a_2,$$

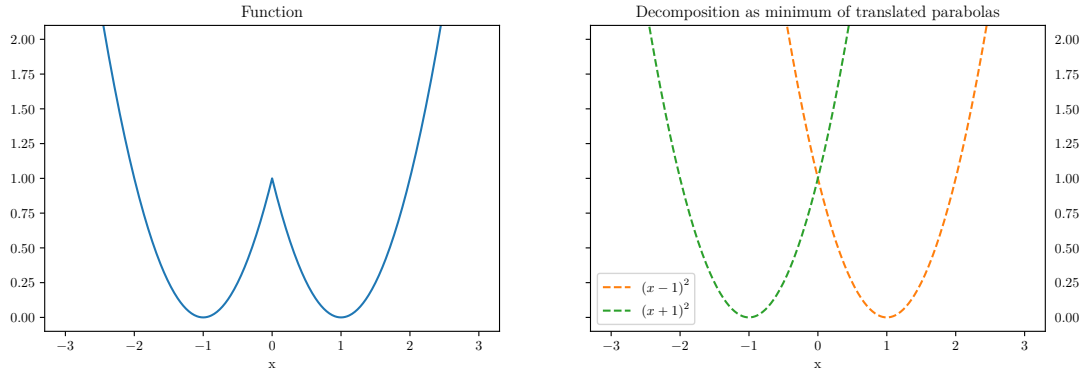


Figure 5.12: Graph of the minimum of two quadratic functions.

which is a piecewise convex function, illustrated in Figure 5.12.

Since the minimum is symmetric in its arguments, we may assume $a_1 < a_2$ without loss of generality. Then, an alternative description of this function is

$$g(x) = \begin{cases} (x-a_1)^2, & x < \frac{a_1+a_2}{2} \\ (x-a_2)^2, & x \geqslant \frac{a_1+a_2}{2}. \end{cases}$$

Or, with the help of the Heaviside function 4.20,

$$g(x) = (x-a_1)^2 \cdot H\left(\frac{a_1+a_2}{2} - x\right) + (x-a_2)^2 \cdot H\left(x - \frac{a_1+a_2}{2}\right).$$

Applying the jump formula (Theorem 4.21), we have a closed expression for both the first and second derivatives of g:

$$\begin{aligned} g'(x) &= 2(x-a_1) \cdot H\left(\frac{a_1+a_2}{2} - x\right) + 2(x-a_2) \cdot H\left(x - \frac{a_1+a_2}{2}\right) \\ g''(x) &= 2 - 2(a_2-a_1)\delta_{\frac{a_1+a_2}{2}}. \end{aligned}$$

Therefore, if μ is a measure over $\mathbb{R}$, the condition for convexification is

$$\begin{aligned} g'' * \mu = 2 * \mu - 2(a_2-a_1)\delta_{\frac{a_1+a_2}{2}} * \mu = \int 2 \, d\mu - 2(a_2-a_1)\tau_{\frac{a_1+a_2}{2}}\mu \geqslant 0 \\ \iff \tau_{\frac{a_1+a_2}{2}}\mu \leqslant \frac{1}{a_2-a_1}. \end{aligned}$$

Since this is a distributional inequality, this bound must hold for *any* positive, smooth, compactly supported function. This means that the translation on the left hand side is irrelevant, because the any such function can be written as the translation of another one. Thus, the condition for convexification is

$$\mu \leqslant \frac{1}{a_2-a_1},$$

where the right hand side must be interpreted as the measure $B \mapsto \int_B \frac{1}{a_2-a_1}\, d\lambda$, where λ is the Lebesgue measure. This means that the set of probabilities that convexify g is described by

$$\begin{cases} 0 \leqslant \mu \leqslant \frac{1}{a_2-a_1} \cdot \lambda, \\ \mu(\mathbb{R}) = 1. \end{cases} \tag{5.17}$$

To satisfy these constraints, the measure μ cannot be too concentrated, which is in accordance with the result of Theorems 5.8 and 5.12.

We can narrow this description even more by noticing that (5.17) implies that μ must be representable by a bounded function. This follows from the fact that for any integrable function h,

$$\int h\, d\mu \leqslant \frac{1}{a_2 - a_1} \int |h|\,(x)\, d\lambda \leqslant \frac{1}{a_2 - a_1} \|h\|_1 ,$$

which means that μ defines a continuous functional on $L^1(\mathbb{R})$ and is consequently representable by an element of $L^\infty(\mathbb{R})$. In view of this, the condition for convexification becomes

$$\mu \in L^\infty(\mathbb{R}) \text{ and } \|\mu\|_\infty \leqslant \frac{1}{a_2 - a_1}.$$

This allows us to restrict (5.17) to the space of bounded functions:

$$\begin{cases} 0 \leqslant \mu(x) \leqslant \frac{1}{a_2-a_1}, \quad \forall x \in \mathbb{R}, \\ \int \mu(x) dx = 1. \end{cases} \tag{5.18}$$

Let's apply constraints (5.18) for two elucidating examples: a uniform and a gaussian distribution. If μ is the uniform distribution in a bounded set B these constraints imply that $f * \mu$ is convex if and only if

$$\|\mu\|_\infty = \frac{1}{\lambda(B)} \leqslant \frac{1}{a_2 - a_1} \iff \lambda(B) \geqslant a_2 - a_1.$$

Compare this result with the one of Theorem 5.12. In there we only knew that the uniform distribution in B reduces the non-convexity of f in proportion to the length of B. Here, even if only for this example, we know that if $\lambda(B)$ is large enough, the function $g * \mu$ is in fact convex.

In the case of a gaussian distribution with mean y and variance σ^2, its maximum is attained at $\mu(y)$, thus

$$\|\mu\|_\infty = \mu(y) = \frac{1}{\sqrt{2\pi}\sigma} \leqslant \frac{1}{a_2 - a_2} \iff \sigma \geqslant \frac{a_2 - a_1}{\sqrt{2\pi}}.$$

This conditions says that if the standard deviation of a gaussian distribution μ is large enough, the function $g * \mu$ is convex.

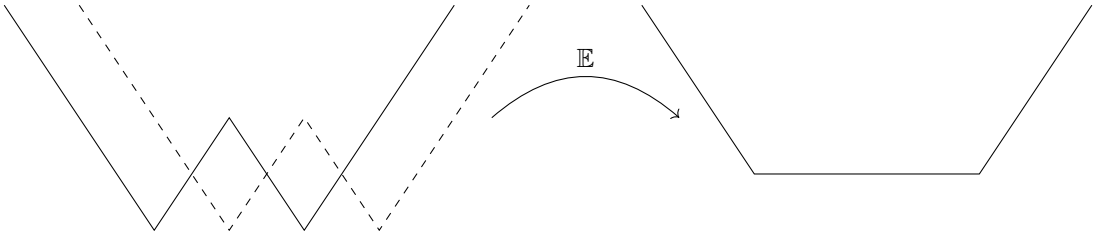


Figure 5.13: A discrete noise can totally convexify a polyhedral function.

An important property of this example is that total convexification is only attained when the additive noise has a probability distribution which is absolutely continuous in relation to Lebesgue measure. This is not a coincidence, as we will see in Theorem 5.14. If f is the minimum of a finite number of convex differentiable functions, there is no discrete probability which can totally convexify f. This contrasts with the polyhedral case, where the finite minimum of polyhedral functions (in particular, the optimal value functions of mixed integer linear problems) can be convexified by discrete noises. Although it was already shown in Section 5.1, we will calculate below the second derivative of

$$W(x) = \min\{|x+1|, |x-1|\}$$

subject to an uncertainty given by a Bernoulli distribution. The density of this random variable is expressible using the Dirac delta functions from Equation (4.3) as

$$\mu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1.$$

The second derivative W'' equals $2\delta_{-1} - 2\delta_0 + 2\delta_1$, and its convolution with μ is

$$\begin{aligned}(W * \mu)'' = W'' * \mu &= (2\delta_{-1} - 2\delta_0 + 2\delta_1) * \left(\frac{1}{2}\delta_0 + \frac{1}{2}\delta_1\right) \\ &= (\delta_{-1} * \delta_0 + \delta_{-1} * \delta_1) - (\delta_0 * \delta_0 + \delta_0 * \delta_1) + (\delta_1 * \delta_0 + \delta_1 * \delta_1) \\ &= \delta_{-1} + \delta_0 - \delta_0 - \delta_1 + \delta_1 + \delta_2 \\ &= \delta_1 + \delta_2 \geqslant 0.\end{aligned}$$

Hence, the function $W * \mu$ is convex.

Despite the unpleasant algebra of the derivation above, this example can be thought geometrically as is illustrated in Figure 5.13. Looking at it, we see that the convexification happens because the noise perfectly aligns the non-convexities such that they cancel each other. This is a special property of this tailored example, since, in practice, it is really hard that such a simple noise can convexify the function.

As a last theorem, we will show that there are minima of convex functions that cannot be convexified by using discrete noise. For this, let f_j be convex functions and define $f(x) = \min_j f_j(x)$. We will suppose that, as in the previous

examples, this function has only a finite number of components. That is, there is a finite collection of points $a_1, \ldots, a_{N-1}$ such that

$$f = \sum_{i=1}^{N} f_i \cdot \mathbb{1}_{(a_{i-1},a_i)}, \tag{5.19}$$

with $f_i \neq f_{i+1}$. On the expression above, we denote $a_0 = -\infty$, $a_N = +\infty$ and $\mathbb{1}_{(a_{i-1},a_i)}$ is zero-one indicator of the open interval (a_{i-1}, a_i). In Theorem 5.14, we will prove that if the f_i are differentiable, no discrete probability is capable of convexifying such a function.

Theorem 5.14. *Let f be as in Equation* (5.19) *and suppose that all $f_i \in C^1(\mathbb{R})$. If f is not convex, there is no discrete probability $\mu = \sum c_j \delta_{x_j}$ for which $(f * \mu)'' \geqslant 0$.*

Proof. Notice that f is continuous, because it is the minimum of continuous functions. From this we can use the jump formula 4.21 to express the derivatives of f in terms of the f_i. Since f is continuous, its first derivative is representable by a discontinuous function and its second derivative is a sum of functions and point masses:

$$\begin{aligned} f' &= \sum f_i' \cdot \mathbb{1}_{(a_{i-1},a_i)} \\ f'' &= \sum f_i'' \cdot \mathbb{1}_{(a_{i-1},a_i)} - \sum \left[f_i'(a_i^-) - f_{i+1}'(a_i^+) \right] \cdot \delta_{a_i} \end{aligned}$$

where the f_i'' are locally integrable functions, as a consequence of Theorems 4.22 and 4.23 and non-negative because the f_i are all convex. Under our assumptions, we also have that the coefficients $\left[f_i'(a_i^-) - f_{i+1}'(a_i^+) \right]$ multiplying the δ_{a_i} are all non-negative because any point of discontinuity a_i is a point where the piecewise representation changes from one f_i to f_{i+1}. That is, $f_i(a_i) = f_{i+1}(a_i)$ and there is $\epsilon_i > 0$ such that for all $0 < \epsilon < \epsilon_i$,

$$f_i(a_i - \epsilon) \leqslant f_{i+1}(a_i - \epsilon) \tag{5.20}$$

$$f_i(a_i + \epsilon) \geqslant f_{i+1}(a_i + \epsilon). \tag{5.21}$$

Since the functions are differentiable, this implies that $f_i'(a_i) \geqslant f_{i+1}'(a_i)$. Thus, for any discrete probability $\mu = \sum c_j \delta_{x_j}$,

$$\begin{aligned} (f * \mu)'' = f'' * \mu &= \sum f_i'' \cdot \mathbb{1}_{(a_{i-1},a_i)} * \mu + \left(\sum \left[f_i'(a_i^-) - f_{i+1}'(a_i^+) \right] \cdot \delta_{a_i} \right) * \mu \\ &= \sum f_i'' \cdot \mathbb{1}_{(a_{i-1},a_i)} * \left(\sum c_j \delta_{x_j} \right) + \left(\sum \left[f_i'(a_i^-) - f_{i+1}'(a_i^+) \right] \cdot \delta_{a_i} \right) * \left(\sum c_j \delta_{x_j} \right) \\ &= \underbrace{\sum_{i,j} \tau_{x_j} f_i'' \cdot \mathbb{1}_{(a_{i-1}+x_j, a_i+x_j)}}_{\text{integrable part}} - \underbrace{\sum_{i,j} c_j \left[f_i'(a_i^-) - f_{i+1}'(a_i^+) \right] \cdot \delta_{a_i+x_j}}_{\text{discrete part}} \end{aligned}$$

In the last line of the expression above, the second derivative $(f * \mu)''$ is decomposed as a difference between an integrable and a discrete part. Since each f_i

is convex, their second derivatives f_i'' are non-negative functions, meaning that the integrable part defines a non-negative measure. Similarly, by the previous discussion and the fact that $c_j \geqslant 0$, every term on the discrete part must be non-negative. Furthermore, integrable and discrete measures are mutually singular, implying that the expression above is the Hahn-Jordan decomposition of $(f * \mu)''$. This means that if we take the negative part of the expression above, only the discrete part remains,

$$[(f * \mu)'']_- = \sum_{i,j} c_j \left[f_i'(a_i^-) - f_{i+1}'(a_i^+)\right] \cdot \delta_{a_i + x_j}.$$

As we assumed that f is not convex, there must be at least one index k such that $f_k'(a_k) < f_{k+1}'(a_k)$ strictly. Therefore, we can apply the negative part of $(f * \mu)''$ to the set $\{a_k + x_1\}$ and the result must be strictly positive, since a sum of non-negative terms is greater than any of its individual terms,

$$\begin{aligned}[(f * \mu)'']_-(\{a_k + x_1\}) &= \sum_{i,j} c_j \left[f_i'(a_i^-) - f_{i+1}'(a_i^+)\right] \cdot \delta_{a_i + x_j}(\{a_k + x_1\}) \\ &\geqslant c_1 \left[f_k'(a_k^-) - f_{k+1}'(a_k^+)\right] \cdot \delta_{a_k + x_1}(\{a_k + x_1\}) > 0\end{aligned}$$

From this, we conclude that $f * \mu$ is not convex. □

It is instructive to graphically compare the result of this theorem with the polyhedral function in Figure 5.13. In that figure, an average between discrete shifts could be convex because the function had hard corners pointing both up and down, yielding δ's with positive and negative coefficients in its second derivative. Therefore a discrete noise was capable of aligning these points in a way that made the second derivative non-negative. On the other side, for a minimum of convex differentiable functions, all hard corners are pointing up and a discrete noise may scatter these corners but is incapable of smoothing them out. See Figure 5.14 for an example.

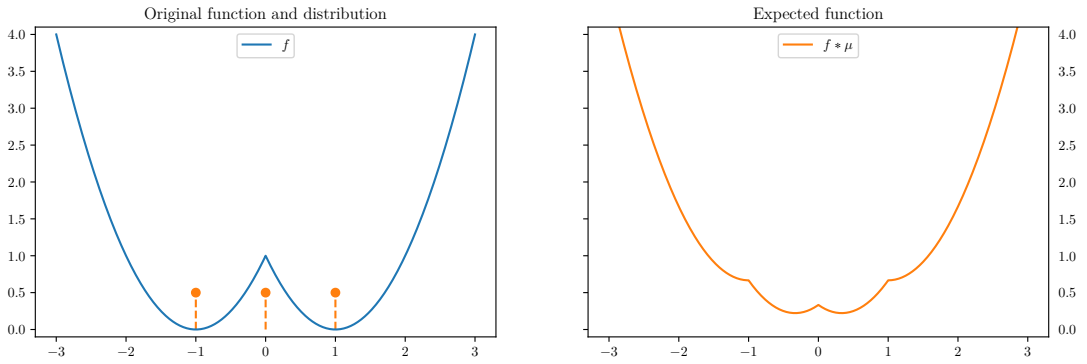


Figure 5.14: A discrete noise can only spread the corners of a minimum of convex functions, without totally convexifying it.

5.4 Non-convexity measures

In Section 5.2, the first method we encountered to measure the non-convexity of a function was its gap function, $\operatorname{gap}(f) = f - \check{f}$, which is well-defined for any function over a vector space and whose image is contained in the set of non-negative functions. Afterwards, in Section 5.3 we restricted ourselves to the study of non-convexity of single variable continuous functions. In this case, our tool to measure the non-convexity of a function was the negative part of its second derivative, $[f'']_-$, which maps a continuous function into a non-negative measure. Interestingly enough, most of the theorems of Section 5.2 have some equivalent formulation in the context of Section 5.3.

The present section analyzes the commonalities between these two ways of measuring non-convexity, generalizing them so that this chapter's previous theorems still hold.

Definition 5.15. A *non-convexity measure* on a convex set of functions X is a function $\mathcal{M}\colon X \to K$ satisfying

1. K is a convex cone,
2. $\mathcal{M}(f) = 0$ if and only if f is convex,
3. $\mathcal{M}$ is convex in relation to K.

Notice that the definition of K-convex function requires that the set X must be convex. Notice also that from the discussion in the previous sections, we know that both the gap function satisfies these properties with respect to the cone of non-negative functions and the second derivative's negative part satisfies with respect to the cone of non-negative measures.

From this point of view, theorems 5.3 and 5.10 are particular cases of Theorem 5.16 below. This tells us that these theorems, which at first seemed complicated functional equations, are in fact simple applications of Jensen's inequality for the operators gap and $f \mapsto [f'']_-$.

Theorem 5.16 (Jensen's inequality)**.** *Given a random function Q, any non-convexity measure $\mathcal{M}$ satisfies*

$$\mathcal{M}(\mathbb{E}[Q]) \leqslant_K \mathbb{E}[\mathcal{M}(Q)].$$

Proof. Because $\mathcal{M}$ is K-convex, it satisfies Jensen's inequality for any probability measure. □

Calling to mind the discussion of Section 5.2.1, we remember that the way we found to globally quantify the non-convexity of a function was through projecting via monotone norms, since it preserved functional inequalities. As we found out, this was specially helpful, for these norms preserved the result of theorem 5.3. As we expose in Theorem 5.17, this happens because for any non-convexity measure $\mathcal{M}$, $\|\cdot\| \circ \mathcal{M}$ is also a non-convexity measure if the norm is K-monotone.

Theorem 5.17. *If $\mathcal{M}\colon X \to K$ is a non-convexity measure and $\|\cdot\|$ is a K-monotone norm, then $\|\cdot\| \circ \mathcal{M}\colon X \to [0,\infty]$ is also a non-convexity measure on the set X.*

Proof. We prove that $\|\cdot\| \circ \mathcal{M}$ satisfies all the properties of Definition 5.15.

1. The set $[0,\infty]$ is closed for sums and multiplication by positive scalars, therefore it is a cone.
2. Since the norm of a vector is zero if and only if it is zero, we have for a function f that
$$\|\mathcal{M}(f)\| = 0 \iff \mathcal{M}(f) = 0 \iff f \text{ is convex.}$$
3. The function $\mathcal{M}$ is K-convex and the norm $\|\cdot\|$ is convex (by the definition of a norm) and K-monotone. Thus, their composition is a convex function. □

The results of Sections 5.2.2 and 5.3.1 depend on an additional hypothesis: that $\mathcal{M}$ is *translation invariant*. Notice that in the previous sections we were using translation invariant norms because both the gap and the negative part of the second derivative are only translation *equivariant*, that is, they commute with translation operators, so we needed some translation invariant way to project them on the non-negative reals.

In general, we may consider non-convexity measures which satisfy

$$\mathcal{M}(\tau_a f) = \mathcal{M}(f) \tag{5.22}$$

for any f and a. These are *translation invariant non-convexity measures* and satisfy Theorem 5.18 for any additive noise.

Theorem 5.18. *If a non-convexity measure $\mathcal{M}\colon X \to K$ is translation invariant, then for any function $f \in X$ and any random variable ξ,*

$$\mathcal{M}(\mathbb{E}\left[\tau_\xi f\right]) \leqslant_K \mathbb{E}\left[\mathcal{M}(\tau_\xi f)\right] \leqslant_K \mathcal{M}(f).$$

Proof. From Jensen's inequality (Theorem 5.16), we get that

$$\mathcal{M}(\mathbb{E}\left[\tau_\xi f\right]) \leqslant_K \mathbb{E}\left[\mathcal{M}(\tau_\xi f)\right]$$

and from the translation invariance of $\mathcal{M}$:

$$\mathbb{E}\left[\mathcal{M}(\tau_\xi f)\right] = \mathbb{E}\left[\mathcal{M}(f)\right] = \mathcal{M}(f). \qquad \square$$

In the alternative notation with convolutions, we wrote $\mathbb{E}\left[\tau_\xi f\right]$ as $f * \mu$, where μ is the probability density of the random variable ξ. Using this, theorem 5.18 may be written in a rather elegant form:

$$\mathcal{M}(f * \mu) \leqslant_K \mathcal{M}(f) * \mu \leqslant_K \mathcal{M}(f). \tag{5.23}$$

The results of Section 5.2.2.1 deal with asymptotic convexification on the uniform norm. To generalize these theorems, the only thing we must ask of $\mathcal{M}$ is that its image must be contained in a set of measurable functions, such that we can make sense of both integration and the essential supremum for $\mathcal{M}(f)$. The following theorems recall classical results regarding convolutions, such as Hölder's inequality, adapted to the context of non-convexity measures.

Theorem 5.19. *Let $\mathcal{M}\colon X \to K$ be a non-convexity measure and suppose that K is contained in the set of measurable functions. Then, for any probability measure μ,*

$$\|\mathcal{M}(f) * \mu\|_\infty \leqslant \gamma \|\mathcal{M}(f)\|_\infty ,$$

where

$$\gamma := \sup_x \mathbb{P}\left[x - \xi \in \operatorname{supp} \mathcal{M}(f)\right] = \sup_x \mu\big(x - \operatorname{supp} \mathcal{M}(f)\big) \leqslant 1.$$

Proof. Any function is equal to itself multiplied by the indicator of its support, thus $\mathcal{M}(f) = \mathcal{M}(f) \cdot \mathbb{1}_S$, where we denoted the support of $\mathcal{M}(f)$ by S. Therefore, for any point x fixed,

$$\begin{aligned}
|\mathcal{M}(f) * \mu(x)| &= \left| \int \mathcal{M}(f)(y)\, d\mu(x-y) \right| = \left| \int \mathcal{M}(f)(y) \cdot \mathbb{1}_S(y)\, d\mu(x-y) \right| \\
&= \left| \int_S \mathcal{M}(f)(y)\, d\mu(x-y) \right| \leqslant \int_S |\mathcal{M}(f)(y)|\, d\mu(x-y) \\
&\leqslant \int_S \|\mathcal{M}(f)\|_\infty \, d\mu(x-y) = \|\mathcal{M}(f)\|_\infty \int_S d\mu(x-y) \\
&\leqslant \|\mathcal{M}(f)\|_\infty \sup_x \mu(x - S).
\end{aligned}$$

Since the bound on the last equation is uniform, it is also valid for the supremum. Hence,

$$\|\mathcal{M}(f) * \mu\|_\infty = \sup_x |\mathcal{M}(f) * \mu(x)| \leqslant \|\mathcal{M}(f)\|_\infty \sup_x \mu(x - S). \qquad \square$$

Theorem 5.20. *Let $\mathcal{M}\colon X \to K$ be a non-convexity measure such that K is contained in the set of measurable functions. Then, if μ_n is a sequence of bounded functions,*

$$\|\mathcal{M}(f * \mu_n)\|_\infty \leqslant \|\mu_n\|_\infty \|\mathcal{M}(f)\|_1$$

and $\mathcal{M}(f * \mu_n)$ converges uniformly to zero if $\|\mathcal{M}(f)\|_1$ is finite and $\|\mu_n\|_\infty$ converges to zero.

Proof. For each fixed x, we write

$$|\mathcal{M}(f * \mu_n)(x)| = \left| \int \mathcal{M}(f)(y) \mu_n(x-y)\, dy \right| \leqslant \int |\mathcal{M}(f)(y)|\, \mu_n(x-y)\, dy.$$

By Hölder's inequality, the right-hand side is below the product of the norms in its integrand. Therefore,

$$|\mathcal{M}(f * \mu_n)(x)| \leqslant \|\mathcal{M}(f)\|_1 \, \|\mu_n\|_\infty .$$

Taking the supremum on both sides gives the uniform norm bound,

$$\|\mathcal{M}(f * \mu_n)\|_\infty \leqslant \|\mathcal{M}(f)\|_1 \, \|\mu_n\|_\infty .$$

The uniform convergence is then a consequence of the above equation. □

5.4.1 Other examples of non-convexity measures

Now we present two other examples of non-convexity measures satisfying the properties on Definition 5.15. The first example generalizes the second derivative' negative part discussed in 5.3 for multivariate functions while the second example gives a non-convexity measure appropriate for the study of piecewise linear or non-convex Lipschitz functions.

We begin by generalizing the second derivative's negative part from Section 5.3 to multivariate functions. If Ω is an open subset of $\mathbb{R}^n$, call

$$X = \{\, f \colon \mathbb{R}^n \to (-\infty, \infty] \mid \operatorname{dom}(f) = \Omega,\ f \text{ continuous on } \Omega \,\}$$

the set of functions whose domain equals Ω. The set X is a convex set of functions that can be seen as a subset of the distributions in Ω. From Theorem 4.27.1, an element $f \in X$ is a convex function if and only if its distributional Hessian $D^2 f$ defines a positive semi-definite matrix measure. That is, the components $\partial_i \partial_j f$ of $D^2 f$ are all measures and for all vectors $v \in \mathbb{R}^n$, the measure $\sum v_i v_j \partial_i \partial_j f$ is non-negative. Using the notation from Chapter 4, we write this condition as

$$\langle v, (D^2 f) v \rangle \geqslant 0,\ \forall v \in \mathbb{R}^n. \tag{5.24}$$

A natural way to measure the non-convexity of a function in X is by considering how negative this expression can be if the argument v is a unit vector,

$$\Lambda(f) = \inf_{\|v\|_2 = 1} \langle v, (D^2 f) v \rangle. \tag{5.25}$$

Notice that by the expression above we mean that $\Lambda(f)$ is the function that for each Borel subset B of Ω, applies this expression for the matrix $(D^2 f)(B)$.

Since we want our non-convexity measure to be non-negative, we will consider the negative part of $\Lambda(f)$, using the variational characterization for the Hahn-Jordan decomposition in Theorem 4.9.1:

$$\mathcal{R}(f) = [\Lambda(f)]_-. \tag{5.26}$$

Notice that $\Lambda(f)$ is not guaranteed to be a measure and the operator $[\cdot]_-$ above uses the same definition but may not maintain the same properties from the

Hahn-Jordan decomposition for measures. Using that both the negative part and Λ are optimal value functions, we can find a more suitable expression for $\mathcal{R}$ in terms of the negative parts of a family of measures. For each Borel set A, we use the fact that infima commute to get

$$\begin{aligned}
\mathcal{R}(f)(A) &= -\inf_{\substack{E\subset A\\ E\in\mathscr{B}(\Omega)}} \inf_{\|v\|_2=1} \langle v, (D^2 f(E))\, v\rangle \\
&= -\inf_{\|v\|_2=1} \inf_{\substack{E\subset A\\ E\in\mathscr{B}(\Omega)}} \langle v, (D^2 f(E))\, v\rangle \\
&= \sup_{\|v\|_2=1} \left\{ -\inf_{\substack{E\subset A\\ E\in\mathscr{B}(\Omega)}} \langle v, (D^2 f(E))\, v\rangle \right\} \\
&= \sup_{\|v\|_2=1} \left\{ [\langle v, (D^2 f)\, v\rangle]_-(A) \right\}
\end{aligned}$$

Since for each $v \in \mathbb{R}^n$ the inner products $\langle v, (D^2 f)\, v\rangle$ are indeed measures, this characterizes $\mathcal{R}(f)$ as the largest value that the negative part of the quadratic form induced by the Hessian of f may attain for a unit vector. For unidimensional functions, $\mathcal{R}$ is precisely the non-convexity measure $f \mapsto [f'']_-$ from Section 5.3. We now proceed to show that $\mathcal{R}$ is a non-convexity measure on X.

Remark 5.4. A famous result [Lax, 1997, thm 10, pg 116] says that the expression for $\Lambda(f)$ is in fact a characterization of the smallest eigenvalue of the matrix $D^2 f$. Since this is a function, the non-convexity measure $\mathcal{R}(f)$ may be seen as returning the negative part of the smallest eigenvalue of $D^2 f$.

Theorem 5.21 (Negative part of the Hessian's smallest eigenvalue)**.** *Let Ω be an open subset of $\mathbb{R}^n$, X be the set of continuous functions on Ω, and $K = \{\, T\colon \mathscr{B}(\Omega) \to [0, +\infty] \,\}$ the cone of non-negative functions over the Borel sets of Ω. Then the operator $\mathcal{R}\colon X \to K$ defined by*

$$\mathcal{R}(f) = \sup_{\|v\|_2=1} [\langle v, (D^2 f)\, v\rangle]_-$$

is a non-convexity measure.

Proof. We will show that $\mathcal{R}$ satisfies the three properties on Definition 5.15. That K is a convex cone follows from the fact that the sum and non-negative scalar multiplication of non-negative functions is also a non-negative function.

To show that $\mathcal{R}(f)$ equals zero if and only if f is convex, we notice that the following propositions are equivalent

$$\begin{aligned}
f \text{ is convex} &\iff \langle v, (D^2 f)\, v\rangle \geqslant 0,\ \forall v \in \mathbb{R}^n \\
&\iff [\langle v, (D^2 f)\, v\rangle]_- = 0,\ \forall v \in \mathbb{R}^n \\
&\iff \mathcal{R}(f) = 0.
\end{aligned}$$

To show that $\mathcal{R}$ is convex in relation to K, we notice that for each fixed Borel set A, the function

$$\mathcal{R}(f)(A) = \sup_{\|v\|_2=1} \left\{ [\langle v, (D^2 f)\, v\rangle]_-(A) \right\}$$

is an optimal value function varying the objective. From Theorem 3.5.1, we know that it is convex. Since this function is convex for each fixed argument and K is a cone of non-negative functions, it follows by definition that $\mathcal{R}$ is convex in relation to K. This concludes the proof. □

Another non-convexity measure arises from the reverse norm cuts from [Ahmed et al., 2019]. From Theorem 2.39, a proper lower semi-continuous convex function can always be underapproximated by affine functions, which we previously called valid cuts. An extension to this setting consists in considering cuts that also have a reverse norm component,

$$f(y) \geqslant f(x) + \langle \mu_x, y - x \rangle - L \|y - x\|, \ \forall y, \tag{5.27}$$

where $\|\cdot\|$ is some norm and L is non-negative. Notice that if f is a convex function, the constant L in 5.27 may be taken to be zero and for each $x \in \mathbb{R}^n$ there is an inclination μ_x such that the inequality still holds. For simplicity,

Let X be the set of all Lipschitz functions over $\mathbb{R}^n$,

$$X = \{f \colon \mathbb{R}^n \to \mathbb{R} \mid \exists L \geqslant 0 \text{ such that } |f(y) - f(x)| \leqslant L \|y - x\|\}.$$

Notice that Lipschitz functions are finite valued and everywhere continuous and, therefore, are proper lower semi-continuous functions. As a way to measure the non-convexity of $f \in X$, we will take the smallest constant L such that for all $x \in \mathbb{R}^n$, there is an inclination μ_x such that inequality (5.27) holds. More formally,

$$\begin{aligned} \mathcal{L}(f) = \min_{L} \quad & L \\ \text{s.t.} \quad & \forall x, \exists \mu_x, \forall y, \ f(y) \geqslant f(x) + \langle \mu_x, y - x \rangle - L \|y - x\| \\ & L \geqslant 0. \end{aligned} \tag{5.28}$$

Theorem 5.22 (Minimum Lipschitz constant)**.** *Let X be the set of Lipschitz functions on $\mathbb{R}^n$. Then the operator $\mathcal{L} \colon X \to [0, \infty]$ defined in* (5.28) *is a non-convexity measure.*

Proof. The set $K = [0, \infty]$ is a convex cone, since it is closed by addition and multiplication.

If $f \in X$ is a convex function, since it finite everywhere, it equals the supremum of its valid cuts and Equation (5.27) still holds with $L = 0$. Thus, $\mathcal{L}(f) = 0$. Conversely, if $\mathcal{L}(f) = 0$, we have that for all x there is an inclination μ_x such that

$$f(y) \geqslant f(x) + \langle \mu_x, y - x \rangle.$$

Since these are tight cuts, it follows that $f = \check{f}$, which implies that f is convex.

To prove that $\mathcal{L}$ is convex, we will explicitly show that it satisfies Jensen's inequality. Let f, $g \in X$. Then for each $x \in \mathbb{R}^n$ there are inclinations $\mu_{f,x}$ and $\mu_{g,x}$ such that

$$f(y) \geqslant f(x) + \langle \mu_{f,x}, y - x \rangle - \mathcal{L}(f) \|y - x\|, \ \forall y,$$
$$g(y) \geqslant g(x) + \langle \mu_{g,x}, y - x \rangle - \mathcal{L}(g) \|y - x\|, \ \forall y.$$

By taking any $\lambda \in [0, 1]$, call $\mu_x = \lambda \mu_{f,x} + (1 - \lambda)\mu_{g,x}$. Then, for all x, the convex combination $\lambda f + (1 - \lambda)g$ satisfies, for all y,

$$\big(\lambda f + (1-\lambda)g\big)(y) \geqslant \big(\lambda f + (1-\lambda)g\big)(x) + \langle \mu_x, y - x \rangle - \big(\lambda \mathcal{L}(f) + (1-\lambda)\mathcal{L}(g)\big) \|y - x\|.$$

This means that $\lambda \mathcal{L}(f) + (1 - \lambda)\mathcal{L}(g)$ is a feasible point for the optimization represented by $\mathcal{L}(\lambda f + (1 - \lambda)g)$. Since it is the smallest feasible constant,

$$\mathcal{L}(\lambda f + (1 - \lambda)g) \leqslant \lambda \mathcal{L}(f) + (1 - \lambda)\mathcal{L}(g),$$

concluding the proof. □

Convexification via Risk Measures 6

In Section 5.2 we defined the gap of a function f, $\operatorname{gap}(f) = f - \check{f}$, as a measure of its non-convexity and saw that the gap of an average of functions is always less than the average gap,

$$\operatorname{gap}(\mathbb{E}[Q]) \leqslant \mathbb{E}[\operatorname{gap}(Q)]. \tag{6.1}$$

This result can be readily applied to risk-neutral stochastic programs since it says that the gap of the expected cost-to-go is always below the average of the gap of all scenarios.

The concept of gap function was later generalized in Section 5.4 to a larger family of operators called *non-convexity measures* that also satisfy an inequality similar to (6.1).

In this chapter, we take another route and see how the gap behaves when we substitute the average value $\mathbb{E}$ by an arbitrary coherent risk measure ρ. As we will shortly see, some assertions still hold. Results such as Theorem 6.1 ensure that an analogous result still holds, that is, $\widetilde{\rho(Q)}$ is closer to $\rho(Q)$ than $\widetilde{\rho(\check{Q})}$. This means that in risk-averse optimization problems, calculating cuts using $\widetilde{\rho(Q)}$ gives us sharper results than the usual approach.

6.1 Gap function and risk measures

An important characteristic of the expectation is its linearity. Distinctively, it means that we can write the expectation of the gap in two equivalent ways,

$$\mathbb{E}[Q - \check{Q}] = \mathbb{E}[Q] - \mathbb{E}[\check{Q}].$$

Since, for an arbitrary risk measure ρ, $\rho(Q - \check{Q})$ in general dos not coincide with $\rho(Q) - \rho(\check{Q})$, we must choose which of these expressions we want to generalize. These two formulations are related, as will be shown in Theorem 6.3.

We begin with a slightly more general result. Let G be any operator that takes a random function and returns a deterministic function. If G is monotone and preserves convexity, a result equivalent to Theorem 5.3 holds.

Theorem 6.1. *Let G be a monotone operator that takes take convex random functions to convex functions. Then, for any random function Q,*

$$G(Q) - \widetilde{G(Q)} \leqslant G(Q) - G(\check{Q}).$$

Proof. As G preserves convexity, the function $G(\check{Q})$ is convex. From the monotonicity of G,

$$\check{Q}(\cdot, \omega) \leqslant Q(\cdot, \omega) \implies G(\check{Q}) \leqslant G(Q).$$

Using that the convex relaxation $\widetilde{G(Q)}$ is above any other convex function bounded by $G(Q)$,

$$G(\check{Q}) \leqslant \widetilde{G(Q)} \leqslant G(Q)$$

which implies the desired result:

$$G(Q) - \widetilde{G(Q)} \leqslant G(Q) - G(\check{Q}). \qquad \square$$

Any coherent risk measure is monotone and preserves convexity, therefore an important corollary to Theorem 6.1 is that it may be applied to any coherent risk measure. A special case which deserves to be highlighted is when the risk measure is the supremum of Q over some set.

Corollary 6.1.1. Let Q be a random function and ρ a coherent risk measure. Then

$$\rho(Q) - \widetilde{\rho(Q)} \leqslant \rho(Q) - \rho(\check{Q}).$$

Corollary 6.1.2. Let Q be a random function. Then

$$\sup_{\xi} Q(x, \xi) - \operatorname{conv}(\sup_{\xi} Q(x, \xi)) \leqslant \sup_{\xi} Q(x, \xi) - \sup_{\xi} \check{Q}(x, \xi), \ \forall x.$$

The importance of Corollary 6.1.1 for risk-averse optimization problems lies in the fact that we can calculate cuts for both $\rho(\check{Q})$ or $\widetilde{\rho(Q)}$. From the corollary we see that for any random function,

$$\rho(\check{Q}) \leqslant \widetilde{\rho(Q)} \leqslant \rho(Q). \tag{6.2}$$

Therefore sharper results can be obtained if we calculate cuts to $\widetilde{\rho(Q)}$.

Previously, we argued that since a coherent risk measure is not necessarily linear, there is a difference between taking the gap of the risk or the risk of the gap. Theorem 6.2 says how these notions are related when we change this order.

Theorem 6.2. *If ρ is a coherent risk measure,*

$$\rho(Q) - \rho(\check{Q}) \leqslant \rho(Q - \check{Q}).$$

Proof. From the subadditivity of ρ, we have that

$$\rho(Q) = \rho(Q - \check{Q} + \check{Q}) \leqslant \rho(Q - \check{Q}) + \rho(\check{Q}).$$

Reorganizing the terms, we get that

$$\rho(Q) - \rho(\check{Q}) \leqslant \rho(Q - \check{Q}). \qquad \square$$

Combining the previous results, we can relate what happens when we change the order of evaluating the risk measure ρ and the convexification.

Theorem 6.3. *Let ρ a coherent risk measure. For any random function Q,*

$$\rho(Q) - \widetilde{\rho(Q)} \leqslant \rho(Q) - \rho(\check{Q}) \leqslant \rho(Q - \check{Q}).$$

Proof. Follows directly by combining Theorems 6.1.1 and 6.2. $\square$

6.2 Additive noise

In the same manner of Section 5.2.2, we restrict our attention to random functions of the form

$$Q(x, \xi) = (\tau_\xi f)(x) = f(x - \xi).$$

The main result of that section was theorem 5.7 which said that for any translation invariant monotone norm $\|\cdot\|$,

$$\|\mathbb{E}\left[\tau_\xi(f - \check{f})\right]\| \leqslant \|f - \check{f}\|.$$

When working with an arbitrary coherent risk measure ρ, however, there is no analogous result encompassing all translation invariant monotone norms.

In what follows, we prove Theorem 6.4 which says that for the uniform norm, an analogue of Theorem 5.7 holds for coherent risk measures. As in Sections 5.2.2 and 5.3.1, the uniform norm again seems to possess stronger bounds for the gap reduction. Afterwards, we display an example of a combination of a coherent risk measure ρ and a translation invariant norm $\|\cdot\|$ such that $\left\|\rho(\tau_\xi f) - \widetilde{\rho(\tau_\xi f)}\right\|$ is strictly greater than $\|f - \check{f}\|$.

Theorem 6.4. *Let ρ be a proper, lower semi-continuous coherent risk measure. For any function f subjected to an additive noise ξ,*

$$\left\|\rho(\tau_\xi f) - \widetilde{\rho(\tau_\xi f)}\right\|_\infty \leqslant \|\rho(\tau_\xi f) - \rho(\tau_\xi \check{f})\|_\infty \leqslant \|\rho(\tau_\xi f - \tau_\xi \check{f})\|_\infty \leqslant \|f - \check{f}\|_\infty.$$

Proof. The first two inequalities come from Theorem 6.3 applied to $Q(x, \xi) = f(x - \xi)$.

For the last inequality, notice that since $f - \check{f}$ is non-negative, the monotonicity and non-negative homogeneity of ρ imply that

$$\tau_\xi(f - \check{f}) \geqslant 0 \implies \rho(\tau_\xi(f - \check{f})) \geqslant \rho(0) = 0.$$

Since the uniform norm of a non-negative function is simply its supremum over x,

$$\|\rho(\tau_\xi f - \tau_\xi \check{f})\|_\infty = \sup_x \rho(\tau_\xi f - \tau_\xi \check{f}).$$

Remember that any proper and lower semi-continuous coherent risk measure has a dual representation as

$$\rho(\xi) = \sup_{\mu \in \mathcal{P}} \mathbb{E}^\mu [\xi]$$

for some family of probabilities $\mathcal{P}$, as seen in Theorem 3.26. Therefore, since suprema commute,

$$\begin{aligned}
\|\rho(\tau_\xi f - \tau_\xi \check{f})\|_\infty &= \sup_x \sup_{\mu \in \mathcal{P}} \mathbb{E}^\mu [\tau_\xi f - \tau_\xi \check{f}] \\
&= \sup_{\mu \in \mathcal{P}} \sup_x \mathbb{E}^\mu [\tau_\xi f - \tau_\xi \check{f}] \\
&= \sup_{\mu \in \mathcal{P}} \|\mathbb{E}^\mu [\tau_\xi f - \tau_\xi \check{f}]\|_\infty .
\end{aligned}$$

The norm $\|\cdot\|_\infty$ is convex, which means that we can apply Jensen's inequality to the expected value inside it. It is also translation invariant, as discussed in Section 5.2.2. From these two facts,

$$\begin{aligned}
\|\rho(\tau_\xi f - \tau_\xi \check{f})\|_\infty &\leqslant \sup_{\mu \in \mathcal{P}} \mathbb{E}^\mu \left[\|\tau_\xi f - \tau_\xi \check{f}\|_\infty\right] \\
&= \sup_{\mu \in \mathcal{P}} \mathbb{E}^\mu \left[\|f - \check{f}\|_\infty\right].
\end{aligned}$$

As $\|f - \check{f}\|_\infty$ is a constant,

$$\|\rho(\tau_\xi f - \tau_\xi \check{f})\|_\infty \leqslant \|f - \check{f}\|_\infty . \qquad \square$$

To prove this theorem we had to restrict ourselves to the uniform norm $\|\cdot\|_\infty$. Now we give an example using the integral norm where an analogous to that inequality does not hold. For this, choose as a risk measure $\rho = \max_{\xi \in \{-\frac{1}{2}, \frac{1}{2}\}}$ and define the function

$$f(x) = \min\left\{(x+2)^2,\, (x-2)^2\right\} = \begin{cases} (x+2)^2, & x \leqslant 0 \\ (x-2)^2, & x > 0. \end{cases} \tag{6.3}$$

This is a continuous, piecewise convex function that can be realized as the optimal value function of the following mixed integer program with convex objective

$$\begin{aligned}
f(x) = \min_{y,z} \quad & y^2 \\
\text{s.t.} \quad & y - 4z = x + 2 \\
& x, y \in \mathbb{R} \\
& z \in \{0, 1\}
\end{aligned}$$

and whose convex relaxation is the function

$$\check{f}(x) = \begin{cases} (x+2)^2, & x < -2 \\ 0, & x \in (-2, 2) \\ (x-2)^2, & x > 2. \end{cases} \tag{6.4}$$

Applying the risk measure ρ to $\tau_\xi f$, we get the following piecewise convex function as a result

$$\begin{aligned} \rho(\tau_\xi f)(x) &= \max_{\xi \in \{-\frac{1}{2}, \frac{1}{2}\}} \min\left\{(x-\xi+2)^2,\, (x-\xi-2)^2\right\} \\ &= \max\left\{ \min\{(x+2.5)^2,\, (x-1.5)^2\},\, \min\{(x+1.5)^2,\, (x-2.5)^2\}\right\}. \end{aligned}$$

After some calculations, the function $\rho(\tau_\xi f)$ can be written piecewisely as

$$\rho(\tau_\xi f) = \begin{cases} (x+1.5)^2, & x < -2 \\ (x+2.5)^2, & x \in (-2, -0.5) \\ (x-1.5)^2, & x \in (-0.5, 0) \\ (x+1.5)^2, & x \in (0, 0.5) \\ (x-2.5)^2, & x \in (0.5, 2) \\ (x-1.5)^2, & x \in x > 2 \end{cases} \tag{6.5}$$

and its convex relaxation $\widetilde{\rho(\tau_\xi f)}$ can be written piecewisely as

$$\widetilde{\rho(\tau_\xi f)} = \begin{cases} (x+1.5)^2, & x < -2 \\ 0.25, & x \in (-2, 2) \\ (x-1.5)^2, & x > 2. \end{cases} \tag{6.6}$$

The graphs of the functions f, $\check{f}$, $\rho(\tau_\xi f)$ and $\widetilde{\rho(\tau_\xi f)}$ can be visualized in Figure 6.1.

We now proceed to calculate the integral norm for the gap of both these functions. For f, the volume of the gap is

$$\begin{aligned} \|\mathrm{gap}(f)\|_1 &= \int f - \check{f} = \int_{-2}^{0} (x+2)^2\, dx + \int_{0}^{2} (x-2)^2\, dx \\ &= \frac{1}{3}\left(2^3 - 0^2\right) + \frac{1}{3}\left(0^3 - (-2)^3\right) = \frac{16}{3}. \end{aligned}$$

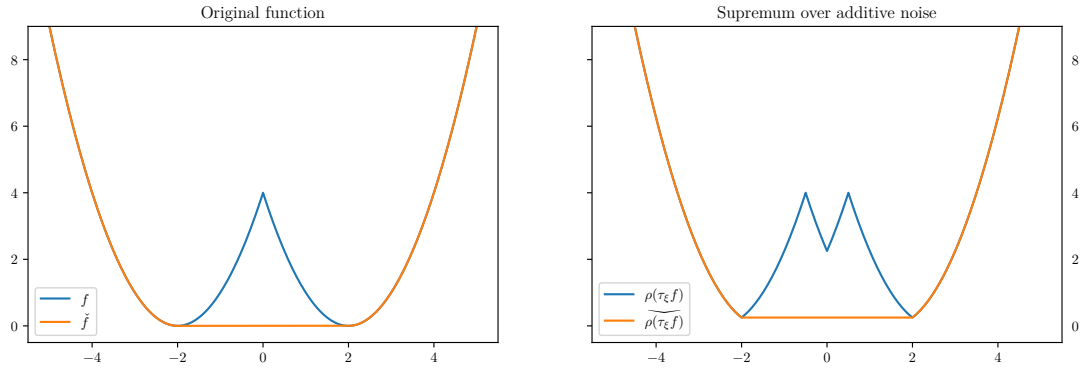


Figure 6.1: The left image is the graph of f superimposed over the graph of its convex relaxation $\check{f}$. The right image is the graph of the supremum $\rho(\tau_\xi f)$ superimposed over the graph of its convex relaxation $\widetilde{\rho(\tau_\xi f)}$. Notice the difference on their gaps.

For $\rho(\tau_\xi f)$, the volume of the gap is

$$\begin{aligned}
\|\mathrm{gap}(\rho(\tau_\xi f))\|_1 = \int \rho(\tau_x f) - \widetilde{\rho(\tau_\xi f)} &= \int_{-2}^{-\frac{1}{2}} (x+2.5)^2 - \frac{1}{4}\,dx + \int_{-\frac{1}{2}}^{0} (x-1.5)^2 - \frac{1}{4}\,dx \\
&\quad + \int_{0}^{\frac{1}{2}} (x+1.5)^2 - \frac{1}{4}\,dx + \int_{\frac{1}{2}}^{2} (x-2.5)^2 - \frac{1}{4}\,dx \\
&= 4\left(-\frac{1}{4}\right) + \frac{1}{3}\left[2^3 - \left(\frac{1}{2}\right)^3 + \left(-\frac{3}{2}\right)^3 - (-2)^3 + 2^3 - \left(\frac{3}{2}\right)^3 \right. \\
&\quad \left. + \left(-\frac{1}{2}\right)^3 - (-2)^3\right] \\
&= -1 + \frac{25}{3} = \frac{22}{3}
\end{aligned}$$

Therefore we see that, for this example,

$$\|\mathrm{gap}(\rho(\tau_\xi f))\|_1 > \|\mathrm{gap}\, f\|_1 ,$$

showing that an analogous of Theorem 6.4 cannot hold for an arbitrary translation invariant monotone norm.

Remark 6.1. Although this counterexample was constructed for the maximum over the uncertainty, this also happens for an infinite family of coherent risk measures. To see this, consider the risk measures defined by

$$\rho_\epsilon = \max_\xi + \epsilon\mathbb{E}.$$

These are strictly monotone coherent risk measures and, for $\epsilon > 0$ small enough, the inequality in the counterexample still holds.

Expected Cost-to-go

7

A consequence of Theorem 5.3 is that if Q is the cost-to-go function of a stochastic program,

$$\mathbb{E}\left[\check{Q}\right] \leqslant \widetilde{\mathbb{E}\left[Q\right]} \leqslant \mathbb{E}\left[Q\right].$$

Moreover, it is possible that $\mathbb{E}\left[Q\right]$ is in fact a convex function even if each realization $Q(\cdot, \xi)$ is non-convex. This means that if our goal is to approximate $\mathbb{E}\left[Q\right]$ by cuts, there may be a difference in tightness between calculating cuts for $\mathbb{E}\left[\check{Q}\right]$ or directly for $\widetilde{\mathbb{E}\left[Q\right]}$.

Remember from Section 3.4.2 that there are two formulations for representing the expected cost-to-go of a stochastic program. Assume that there exists a finite number of possible scenarios $\xi^1, \ldots, \xi^N$ with probabilities $\mathbb{P}(\xi = \xi^i) = p_i$, we define the cost-to-go functions $Q^i(x) = Q(x, \xi^i)$ for each scenario. The *decomposed formulation* for $\mathbb{E}\left[Q\right]$ (Definition 3.19) consists in solving each $Q^i(x)$ separately and evaluating the average

$$\mathbb{E}\left[Q\right](x) = \sum_{i=1}^{N} p_i Q^i(x).$$

If we also calculate cuts for each scenario,

$$Q^i(x) \geqslant q^i - \langle \lambda, x \rangle,$$

the average cut gotten by defining $\bar{q} = \mathbb{E}\left[q^i\right]$, $\bar{\lambda} = \mathbb{E}\left[\lambda^i\right]$ is valid for $\mathbb{E}\left[Q\right]$ but is at best tight for $\mathbb{E}\left[\check{Q}\right]$.

The *linked formulation* for $\mathbb{E}\left[Q\right]$ (Definition 3.21) consists in forming a large problem that considers all scenarios at the same time. This formulation is more computationally expensive but allows us to directly calculate cuts that are tight for $\widetilde{\mathbb{E}\left[Q\right]}$ and, in the possible cases that the uncertainty totally convexifies the expected cost-to-go, these cuts will also be tight for $\mathbb{E}\left[Q\right]$. These will be called *ECTG* cuts, since they are calculated taking the entire expected cost-to-go into account.

This chapter is dedicated to make a proof of concept for the advantages and disadvantages between calculating cuts using the decomposed and the linked formulation for non-convex multi-stage stochastic programs. In Section 7.1, we give

a brief description of the *Stochastic Dual Dynamic Programming* method, used in the remainder of this chapter to computationally solve multi-stage stochastic programs. Section 7.2 presents the computational and programming environment used to solve the problems of the following sections. In Section 7.3, we will solve a simple multi-stage stochastic program with only one state variable per stage to allow a better visualization of the cost-to-go functions and their approximation by cuts using each formulation. This section should be seem as a visual motivation in a similar manner to Section 5.1. In Section 7.4, we apply both the linked and decomposed formulation for a hydrothermal operational planning problem with disjunctive constraints. This is a mixed integer linear program where the cost-to-go for each scenario is non-convex.

7.1 Stochastic dual dynamic programming

The computational experiments from this chapter were made using the *Stochastic Dual Dynamic Programming* (SDDP) algorithm, originated in [Pereira and Pinto, 1991], to solve multi-stage stochastic programs. In this section we briefly discuss how this method iteratively calculates cuts to construct polyhedral underapproximations to the each stage's expected cost-to-go. Since our focus is comparing the different methods for calculating cuts, we in no sense try to give a detailed explanation of the SDDP algorithm. The interested reader may find it on the original [Pereira and Pinto, 1991] or in Chapter 4 of [Cabral, 2018]. For a discussion of this algorithm's convergence properties, see [Shapiro, 2011].

Recall that a risk-neutral multi-stage stochastic program may be written in dynamic programming form as

$$\begin{aligned} Q_t(x_{t-1}, \xi_t) = \min_{x_t, u_t} \quad & c_t(x_t, u_t) + \bar{Q}_{t+1}(x_t) \\ \text{s.t.} \quad & (x_{t-1}, x_t, u_t) \in X_t \end{aligned} \tag{7.1}$$

where $\xi_t = (c_t, X_t)$ is a stochastic process representing the problem's random data and

$$\bar{Q}_{t+1}(x_t) = \begin{cases} \mathbb{E}\left[Q_{t+1}(x_t, \xi_{t+1})\right], & t = 1, \dots, T-1 \\ 0, & t = T \end{cases} \tag{7.2}$$

is the expected cost-to-go for each stage. Throughout this discussion, we assume that the stochastic process is *stagewise independent*, that is, the random variable ξ_{t+1} is independent of all previous uncertainties $\xi_1, \dots, \xi_t$. As in Section 3.4, we also assume that each ξ_t has finite support. The SDDP algorithm consists in iteratively estimating the functions $\bar{Q}_t$ by families of valid cuts.

The procedure of calculating the cuts is called the *backward step* and is similar to the approximations discussed in Section 3.3 for deterministic two-stage problems. At the beginning, it considers the cost-to-go for the last stage

$$\begin{aligned} Q_T(x_{T-1}, \xi_T) = \min_{x_T, u_T} \quad & c_T(x_T, u_T) \\ \text{s.t.} \quad & (x_{T-1}, x_T, u_T) \in X_T. \end{aligned}$$

This may be viewed as the cost-to-go of a two-stage problem and, therefore, any of the techniques in 3.4 can be applied to calculate a valid cut $\phi_T^{(1)}$ for the expected cost-to-go $\bar{Q}_T$. Set $\mathfrak{Q}_T^{(1)} = \phi_T^{(1)}$. This yields an approximation for the expected cost-to-go at the penultimate stage given by

$$\tilde{Q}_{T-1}^{(1)}(x_{T-2}) = \mathbb{E}\left[\begin{array}{rl} \min\limits_{x_{T-1}, u_{T-1}} & c_{T-1}(x_{T-1}, u_{T-1}) + \mathfrak{Q}_T^{(1)}(x_{T-1}) \\ \text{s.t.} & (x_{T-2}, x_{T-1}, u_{T-1}) \in X_{T-1} \end{array}\right] \tag{7.3}$$

which is everywhere below the real expected cost-to-go $\bar{Q}_{T-1}$, implying that any valid cut for the approximation $\tilde{Q}_{T-1}^{(1)}$ is also valid for $\bar{Q}_{T-1}$. Thus, we may use it to calculate a valid cut $\phi_{T-1}^{(1)}$ and set a polyhedral approximation $\mathfrak{Q}_{T-1}^{(1)} = \phi_{T-1}^{(1)}$ for the penultimate stage problem. This process is repeated backwards until we have found an approximation $\phi_1^{(1)}$ for $\bar{Q}_1$ at the first stage.

This same process is repeated iteratively to yield finer and finer approximations for the expected costs-to-go. On the k-th iteration, we repeat this same procedure calculating new cuts $\phi_t^{(k)}$ for each stage and setting the polyhedral approximations to be

$$\mathfrak{Q}_t^{(k)} = \max\{\mathfrak{Q}_t^{(k-1)}, \phi_t^{(k)}\}. \tag{7.4}$$

These can be used to define approximated optimization problems

$$\tilde{Q}_t^{(k)}(x_{t-1}) = \mathbb{E}\left[\begin{array}{rl} \min\limits_{x_t, u_t} & c_t(x_t, u_t) + \mathfrak{Q}_{t+1}^{(k)}(x_t) \\ \text{s.t.} & (x_{t-1}, x_t, u_t) \in X_t \end{array}\right] \tag{7.5}$$

that are used to further calculate valid cuts for the previous stages. Because of the way that they are defined, these approximations are monotone in the number of iterations,

$$\mathfrak{Q}_t^{(k-1)} \leqslant \mathfrak{Q}_t^{(k)}, \quad \tilde{Q}_t^{(k-1)} \leqslant \tilde{Q}_t^{(k)}.$$

Besides that, they serve as underapproximations of the true solution, since it always holds by construction that

$$\mathfrak{Q}_t^{(k)} \leqslant \bar{Q}_t \text{ and } \tilde{Q}_t^{(k)} \leqslant \bar{Q}_t.$$

The *forward step* of the SDDP algorithm consists in using the approximations $\tilde{Q}_t$ previously calculated to estimate an upper bound for the true solution of the problem. Begin by considering a realization of the stochastic process $\xi_t = (c_t, X_t)$ and let x_t^* be optimal solutions of the problems

$$\begin{array}{rl} \min\limits_{x_t, u_t} & c_t(x_t, u_t) + \mathfrak{Q}_{t+1}^{(k)}(x_t) \\ \text{s.t.} & (x_{t-1}, x_t, u_t) \in X_t. \end{array}$$

Since the solution of each stage depends on the solution of the previous, this process requires that we solve the problem by going forward from one stage to the next. The calculated solution $x^* = (x_1^*, \ldots, x_T^*)$ satisfies all the constraints

and is therefore feasible for the real problem given the realization of ξ_t. This means that by sampling M realizations of the stochastic process and setting x^i to be the optimal solution for the approximated problem for scenario ξ^i, the sample average

$$\frac{1}{M}\sum_{i=1}^{M}\left(\sum_{t=1}^{T} c_t^i(x_t^i)\right)$$

gives an upper bound on the true optimal value conditioned to the stochastic process' sample. Forward steps may also be used to select the trial points used for calculating cuts on the backward step. Choosing these points is required to ensure the algorithm convergence, see [Shapiro, 2011, prop 3.1], [Philpott and Guan, 2008], [Guigues, 2016]. Therefore, it is usual to alternate between forward steps with a small sample size and backward steps for a fixed number of iterations followed by a larger forward step for truly estimating the upper bound.

7.2 Computational environment

All simulations in this chapter were done in a computer consisting in a Intel i7-8700 processor with 6 cores, 12 threads and 4.60 GHz of clock frequency, with 32 GB of RAM, DDR, 2666 Mhz. The operating system is a Ubuntu Linux 18.04.2. All models were written in the Julia programming language [Bezanson et al., 2017], version 1.1.1, using the implementation of the package `SDDP.jl` [Dowson and Kapelevich, 2017] for the Stochastic Dual Dynamic Programming method. All the convex and mixed integer linear programs were solved by Gurobi [Gurobi Optimization, 2016], version 8.1.

With the aim of reproducibility, we give in Table 7.1 the version and git commit number of all Julia packages used while running these simulations. The packages `DisjHTPlan` and `SDDP_SB` were developed in the setting of a technical collaboration between UFRJ and the Brazilian Independent System Operator (ONS), for the project IM-21780. They consist, respectively, of an implementation in `Julia` of the model for hydrothermal operational planning with disjunctive constraints described in Section 7.4, and of a module adding the functionality of strengthened Benders cuts for the package `SDDP.jl`. The version of `SDDP.jl` used is also not the original one but a version edited for compatibility with the developed packages.

Table 7.1: Julia packages used for the simulations.

Package	Version	commit (SHA-1)
DisjHTPlan	0.1.0	5c30496e
SDDP_SB	0.1.0	59a45430
SDDP	0.0.0	f4570300
ConfParser	0.1.0	88353bc9
Revise	2.1.6	295af30f
Documenter	0.22.4	e30172f5
GLPK	0.10.0	60bf3e95
GZip	0.5.0	92fee26a
Gurobi	0.6.0	2e9cd046
JuMP	0.19.2	4076af6c
Libz	1.0.0	2ec943e9
MathOptFormat	0.1.1	f4570300
MathOptInterface	0.8.4	b8f27783
NPZ	0.4.0	15e1cf62
OSQP	0.5.2	ab2f91bb
PyPlot	2.8.1	d330b81b
Libdl		8f399da3
Random		9a3f8284

7.3 Unidimensional control problem

In this section we illustrate the difference between the different formulations for approximating the expected cost-to-go by cuts. We will consider simple examples with only a unidimensional state variable per stage. These examples are simple enough to have their cost-to-go functions analytically calculated and it is easy to graphically visualize the difference between the different approximations and the real expected cost-to-go functions. In this sense, these results serve as a "sanity check" for the more complicated model considered in Section 7.4.

7.3.1 Convex case

We begin by considering a convex multi-stage stochastic program, whose cost-to-go functions satisfy the dynamic programming relation

$$\begin{aligned} Q_{t-1}(x_{t-1}, \xi_t) = \min_{x_t, u_t} \quad & |x_t| + \mathbb{E}\left[Q_t(x_t, \xi_{t+1})\right] \\ \text{s.t.} \quad & x_t = x_{t-1} + u_t + \xi_t \\ & u_t \in [-1, 1]. \end{aligned} \tag{7.6}$$

Here, u_t denotes the control variable on stage t and can be chosen on the entire interval $[-1, 1]$. The state variable x_t is constrained to equal the previous state plus the control and the uncertainty. The uncertainty is supposed to be discrete and *stagewise independent* with only two scenarios $\xi_t \in \{-0.5, 0.5\}$ having

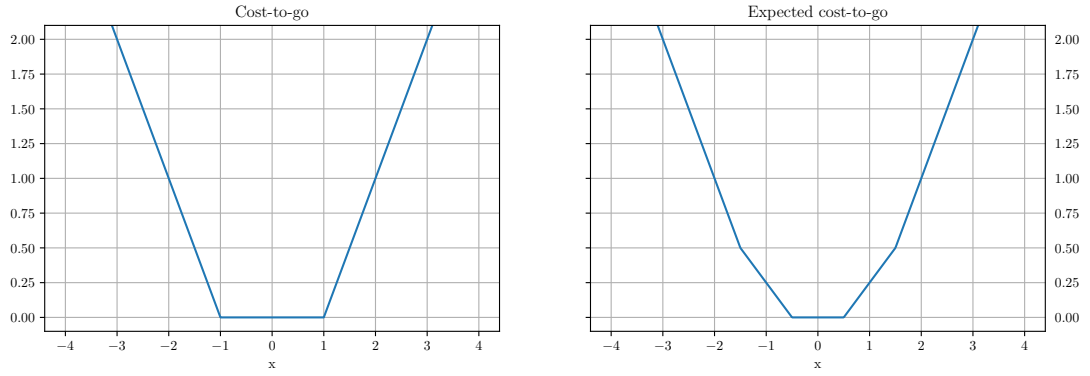


Figure 7.1: Expected cost-to-go for convex problem.

probabilities

$$\mathbb{P}(\xi_t = -0.5) = \frac{1}{2},\ \mathbb{P}(\xi_t = 0.5) = \frac{1}{2}.$$

Figure 7.1, shows the graph of the expected cost-to-go for the last stage together with total cost for a given realization of ξ_T.

Since this problem is convex with only linear constraints, the calculated cuts are guaranteed to be tight. Therefore, we have

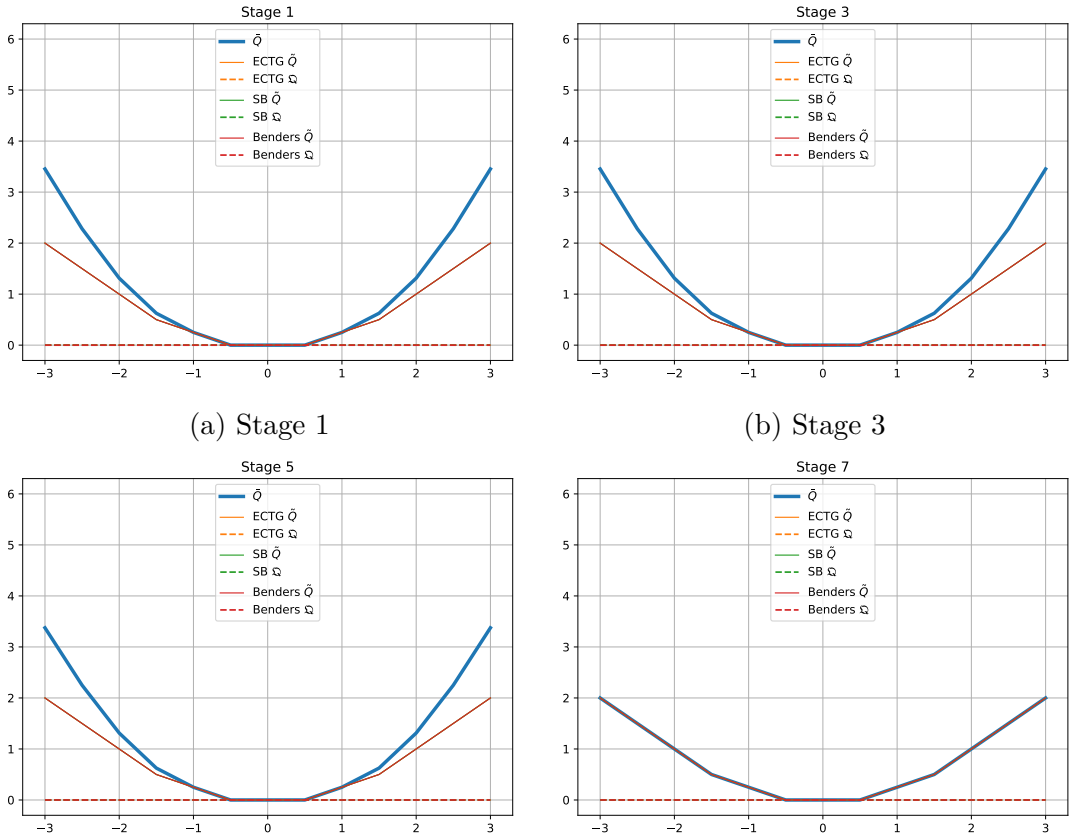
$$\mathfrak{Q}_t^{(k)}(x_t^k) = \tilde{Q}_t^{(k)}(x_t^k)$$

for each point x_t^k chosen on the k-th iteration. Notice that at each iteration, these cuts are only guaranteed to be tight for $\tilde{Q}_t^k$ and not for the true expected cost-to-go $\bar{Q}_t$. In Figure 7.2, we present the approximations $\mathfrak{Q}_t$ and $\tilde{Q}_t$ through different methods together with the true, analytically calculated, expected cost-to-go functions $\bar{Q}_t$ for stages 1, 3, 5 and 7. We denote by *Benders* and *SB* the cost-to-go functions calculated using, respectively, Benders and strengthened Benders with decomposed formulation. The cuts denoted by *ECTG* are strengthened Benders cuts calculated directly for the linked formulation. As expected for a convex problem, the approximations constructed using the three methods are essentially equivalent and find the minimum of $\bar{Q}_t$. The Lagrangian cuts were not used for these simulations because of the huge amount of time needed for calculating them. Calculating strengthened Benders cuts for the linked formulation already provides cuts that are tight for the expected cost-to-go.

For a matter of comparison with next section's plots of non-convex problems, we also consider the expected cost-to-go of the same problem in equation (7.6) but with a more complicated uncertainty with four scenarios $\xi_t \in \{-1.5, -0.5, 0.5, 1.5\}$ having probabilities

$$\mathbb{P}(\xi_t = -1.5) = \frac{1}{6},\ \mathbb{P}(\xi_t = -0.5) = \frac{1}{3},\ \mathbb{P}(\xi_t = 0.5) = \frac{1}{3},\ \mathbb{P}(\xi_t = 1.5) = \frac{1}{6}.$$

The calculated approximations to this problem are shown in Figure 7.3. Since the problem is convex, the approximations $\tilde{Q}_t$ built with each method are essentially



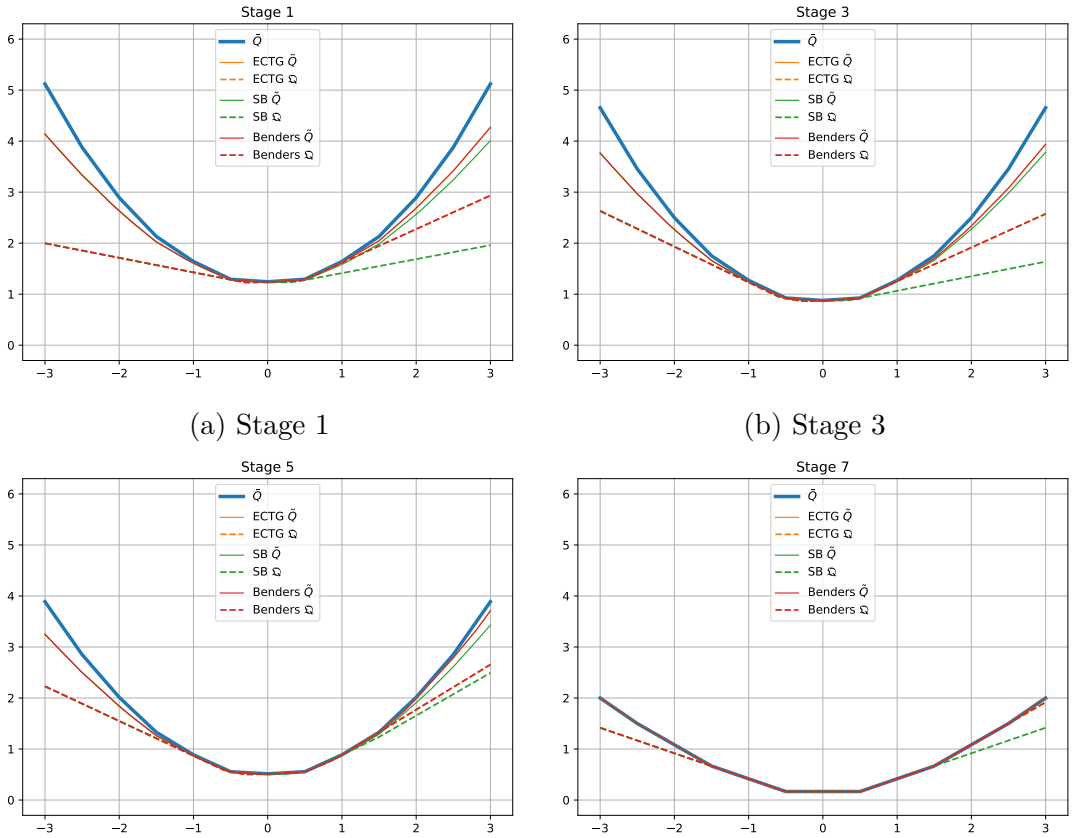
(a) Stage 1

(b) Stage 3

(c) Stage 5

(d) Stage 7

Figure 7.2: Expected cost-to-go for a convex problem with two equally probable scenarios.



(a) Stage 1

(b) Stage 3

(c) Stage 5

(d) Stage 7

Figure 7.3: Expected cost-to-go for a convex problem with four scenarios.

the same. Notice that, when $t = 7$, the expected cost-to-go is known exactly and all approximations overlap, but at the previous stages their quality increasingly decreases.

7.3.2 Non-convex case

Consider the same stochastic program as the one discussed on last section but with the control variables constrained to only be able to take the values -1 or 1,

$$\begin{aligned} Q_{t-1}(x_{t-1}, \xi_t) = \min_{x_t, u_t} \quad & |x_t| + \mathbb{E}\left[Q_t(x_t, \xi_{t+1})\right] \\ \text{s.t.} \quad & x_t = x_{t-1} + u_t + \xi_t \\ & u_t \in \{-1, 1\}. \end{aligned} \tag{7.7}$$

This is a mixed integer program, meaning that we can only guarantee that the expected cost-to-go functions are piecewise convex. In this case, the cuts calculated using the different formulations are expected to behave differently. Figure 7.4 shows the graph of last stage's cost given a realization of ξ_t together with the expected cost-to-go for the last stage.

In Figure 7.5, we see the approximations calculated by the SDDP algorithm

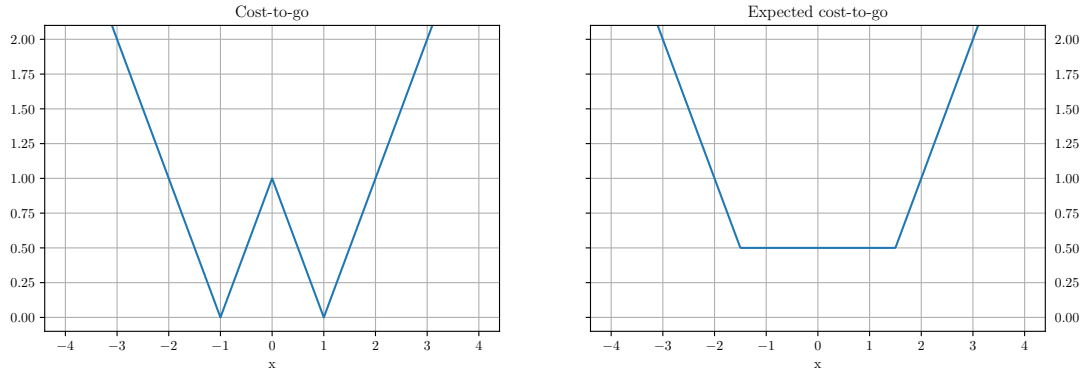


Figure 7.4: Expected cost-to-go for non-convex problem.

for the case with two scenarios

$$\mathbb{P}(\xi_t = -0.5) = \frac{1}{2},\ \mathbb{P}(\xi_t = 0.5) = \frac{1}{2},$$

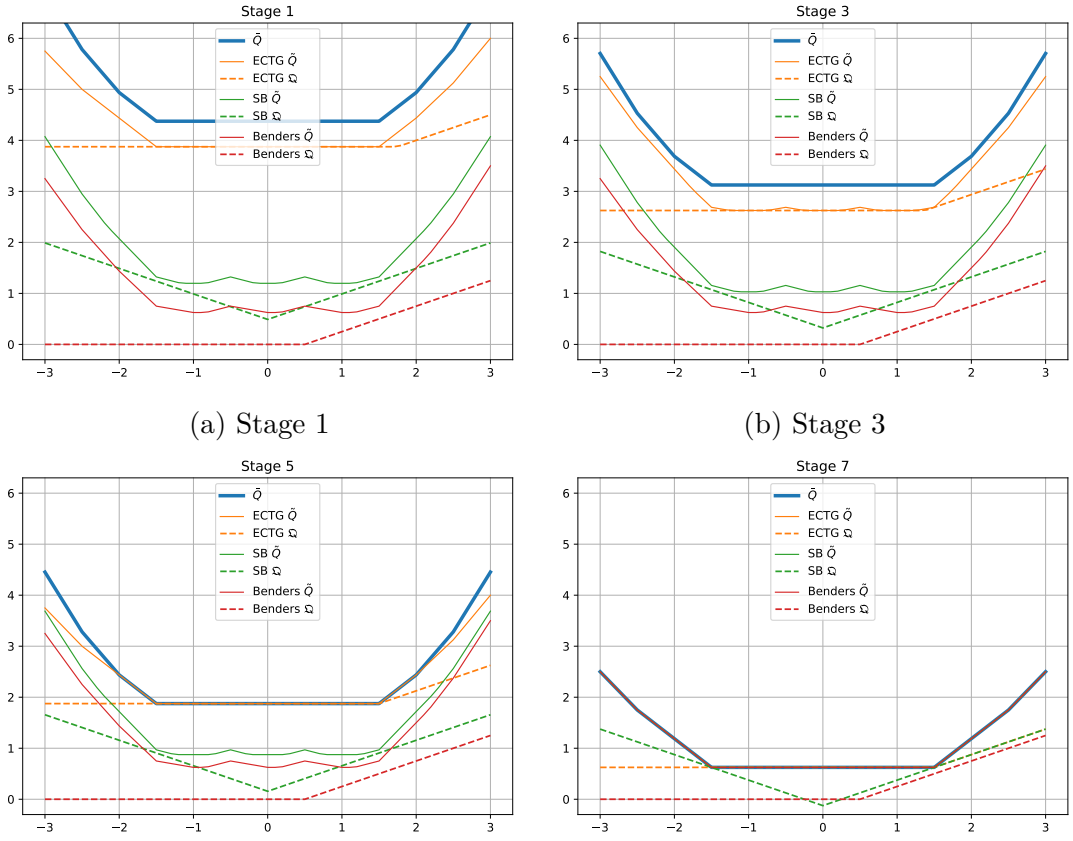
and in Figure 7.6, we see those for the case with four scenarios

$$\mathbb{P}(\xi_t = -1.5) = \frac{1}{6},\ \mathbb{P}(\xi_t = -0.5) = \frac{1}{3},\ \mathbb{P}(\xi_t = 0.5) = \frac{1}{3},\ \mathbb{P}(\xi_t = 1.5) = \frac{1}{6}.$$

Notice that in both cases the true expected cost-to-go is in fact convex.

We start analyzing these figures by the seventh stage, since at the last stage $\bar{Q}_8 = 0$ and, thus, the true expected cost-to-go $\bar{Q}_t$ is calculated exactly when $t = 7$. From the figures for both stages, we see that there is a gap between the cuts calculated using the decomposed formulation and $\tilde{Q}$ that is absent for the ECTG cuts calculated using the linked formulation. This happens because decomposed cuts can be, at best, tight for $\mathbb{E}[\check{Q}]$ while the ECTG cuts can be tight for the true convexification $\widetilde{\mathbb{E}[Q]}$. In these cases, since the true expected cost-to-go is convex, we have $\widetilde{\mathbb{E}[Q]} = \mathbb{E}[Q]$ and the ECTG cuts are tight for $\mathbb{E}[Q]$ while the decomposed cuts will always have a gap.

In the other stages, we calculate the cuts only for $\tilde{Q}_t$, meaning that the approximations are worse. The gap between these approximations and the true expected cost-to-go $\bar{Q}_t$ is propagated through the backward step implying in even worse approximations on the earlier stages. As we can see in the figures, the approximations $\tilde{Q}_t$ calculated using Benders cuts for the decomposed formulation appear to be "frozen" from the first to the fifth stage and those calculated with decomposed strengthened Benders increase only slightly. In comparison, the ECTG approximations closely follow the true expected cost-to-go.



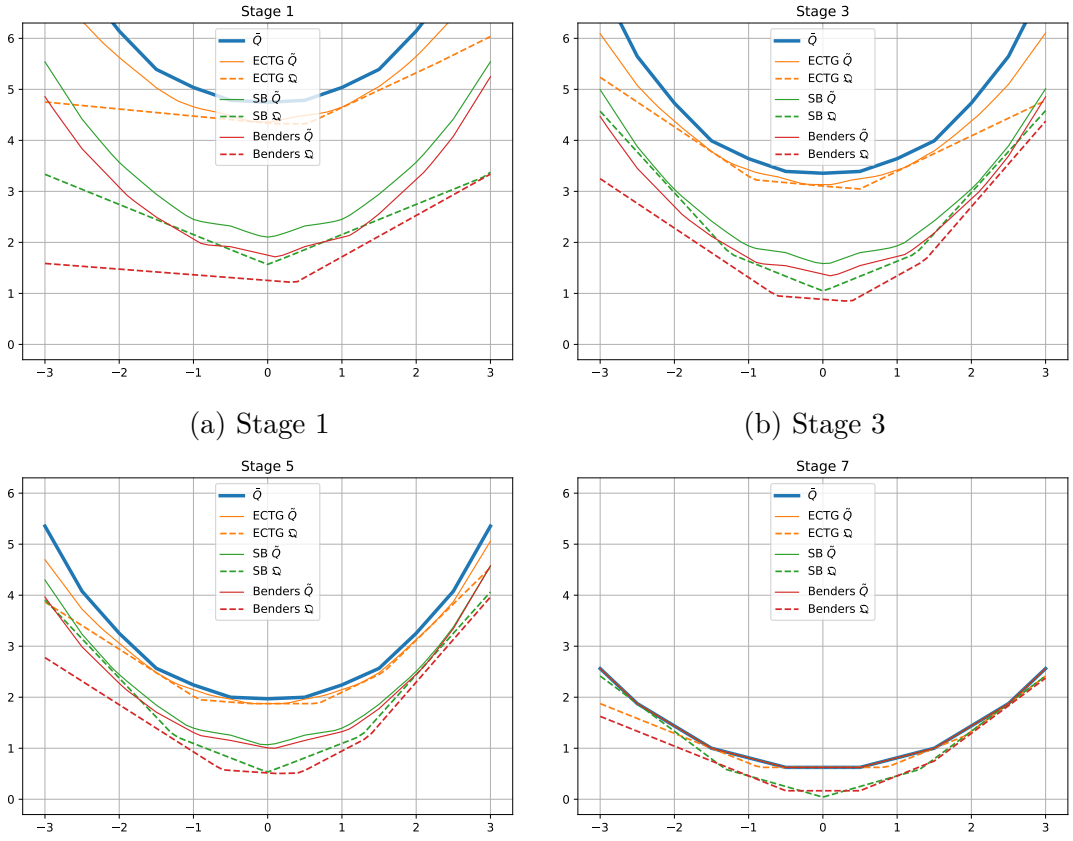
(a) Stage 1

(b) Stage 3

(c) Stage 5

(d) Stage 7

Figure 7.5: Expected cost-to-go for a non-convex problem with two scenarios.



(a) Stage 1

(b) Stage 3

(c) Stage 5

(d) Stage 7

Figure 7.6: Expected cost-to-go for a non-convex problem with four scenarios.

7.4 Hydrothermal operational planning

This section presents a simplified long-term hydrothermal operational planning model with disjunctive constraints and shows approximated solutions to this problem calculated using the decomposed formulation or the ECTG cuts from the linked formulation. This model represents a energy planning system with a fixed number of hydro and thermoelectric subsystems and energy interchange lines between some of these systems. The objective function is to minimize the amount of thermal generation as well as the possible energy deficit.

We begin by considering a *convex model*, whose dynamic programming formulation is

$$\begin{aligned} Q_t(v_{t-1}, \xi_t) = \min_{v_t, q_t, s_t, g_t, df_t, f_t} \quad & \langle c, g_t \rangle + \langle g_{df}, df_t \rangle + \beta \bar{Q}_{t+1}(v_t) \\ \text{s.t.} \quad & v_t = v_{t-1} + \xi_t - q_t - s_t, \\ & q_t + M_I g_t + df_t + M_D f_t = d_t, \\ & 0 \leqslant v_t \leqslant \bar{v}, \quad 0 \leqslant q_t \leqslant \bar{q}, \quad 0 \leqslant s_t \\ & 0 \leqslant g_t \leqslant \bar{g}, \quad 0 \leqslant f_t \leqslant \bar{f}, \quad 0 \leqslant df_t, \end{aligned} \tag{7.8}$$

where, as usual, the expected cost-to-go is

$$\bar{Q}_{t+1}(v_t) = \begin{cases} \mathbb{E}\left[Q_{t+1}(v_t, \xi_{t+1})\right], & t = 1, \ldots, T-1 \\ 0, & t = T. \end{cases}$$

In each stage, the decision variable is given by the vector $x_t = (v_t, q_t, s_t, g_t, df_t, f_t)$. The state variable is the vector v_t and represents the stored energy at the equivalent reservoir of each subsystem at the end of stage t. The control variables are: a vector q_t representing the turbined energy during the stage, a vector s_t representing the spilled energy during the stage, a vector g_t whose components are the amounts of thermal generation, a vector df_t for the amount of deficit on each subsystem, and a vector f_t for the amount of energy interchange between the subsystems. The uncertainty is the stochastic process ξ_t, representing the energy inflows for each subsystem at the beginning of the stage, assumed to be stagewise independent. All other parameters are deterministic.

Besides the constraints giving the upper and lower bounds for the decision variables, there are two more equations relating the variables. Equation

$$v_t = v_{t-1} + \xi_t - q_t - s_t \tag{7.9}$$

is called the *hydro balance equation* and says that the total amount of water at the end of stage t equals the amount of water at the beginning of the stage plus the stochastic inflow minus how much water was turbined or spilled during the stage. The equation

$$q_t + M_I g_t + df_t + M_D f_t = d_t \tag{7.10}$$

is the *load balance equation* and says that the total generated energy must equal the demand at each stage for all subsystems. The parameter d_t is a vector whose

components are the energy demand at stage t for each subsystem, M_I is an indicator matrix with zeros and ones associating each component of the thermal generation vector to its corresponding subsystem and, M_D is a matrix providing the correct sign for the energy interchange at stage t, that is, the components $(M_D)_{ij}$ equals zero if there is no connection between subsystem i and j and equals ± 1 depending if subsystem i receives or sends energy to subsystem j.

The objective function consists of a vector c with the unitary costs of thermal generation times the amount g_t of thermal generation, the unitary cost of deficit c_{df} multiplied by the energy deficit df_t for each subsystem and, the expected cost-to-go $\bar{Q}$ correct by a discount factor β.

Since the model discussed above is convex, all the previously introduced ways to calculate cuts are equivalent. Therefore, we will add some non-convex constraints that model the policy of *minimum stored operational energy* using the technique of disjunctive constraints [Balas, 2011]. This is done by establishing a vector v_{MinOp} representing a desired minimum value for the stored energy of each subsystem. If a component of v_t is below the same component of v_{MinOp}, we say that the corresponding reservoir is below the operational minimum and at least a given amount g_0 of thermal generation must be dispatched. This "if-else" constraint cannot be modeled using only convex equalities and inequalities but can be written as linear constraints involving binary decision variables:

$$\begin{aligned} v_t &\geqslant (1 - z_t)\, v_{\text{MinOp}} \\ g_t &\geqslant z_t\, g_0 \\ z_t &\in \{0,1\}^n, \end{aligned}$$

where n denotes the number of subsystems. The values v_{MinOp} and g_0 are parameters of the model while the z_t are binary decision variables. We arrived at the *non-convex model* for hydrothermal planning, defined by

$$\begin{aligned} Q_t(v_{t-1}, \xi_t) = \min_{v_t, q_t, s_t, g_t, df_t, f_t, z_t} \quad & \langle c, g_t \rangle + \langle g_{df}, df_t \rangle + \beta \bar{Q}_{t+1}(v_t) \\ \text{s.t.} \quad & v_t = v_{t-1} + \xi_t - q_t - s_t, \\ & q_t + M_I g_t + df_t + M_D f_t = d_t, \\ & 0 \leqslant v_t \leqslant \bar{v}, \quad 0 \leqslant q_t \leqslant \bar{q}, \quad 0 \leqslant s_t \\ & 0 \leqslant g_t \leqslant \bar{g}, \quad 0 \leqslant f_t \leqslant \bar{f}, \quad 0 \leqslant df_t, \\ & v_t \geqslant (1 - z_t)\, v_{\text{MinOp}}, \\ & g_t \geqslant z_t\, g_0, \\ & z_t \in \{0,1\}^n. \end{aligned} \tag{7.11}$$

7.4.1 Computational simulations with 2 subsystems

This section is dedicated to compare the solutions obtained for the convex model in Equation (7.8) using the SDDP algorithm with different cut types. This models consists of two subsystems with one interchange line between them and three thermal power plants. A total of 500 valid cuts were constructed for a planning

horizon of 12 stages, were each stage consists of a month. This took 500 backward and forward iterations of the SDDP algorithm.

The simulations were run for decomposed Benders and strengthened Benders cuts as well as for ECTG cuts, that is, strengthened Benders cuts for the linked formulation. Table 7.2 summarizes, for each cut type, the total time and maximum RAM memory needed for calculating the 500 cuts, which may be viewed as measures of computational cost. It also contains the total *calculated cost*, a lower bound for the true policy cost calculated from the polyhedral approximations during the optimization process and the *simulated cost*, an upper bound estimate, calculated through 500 simulations after the policy was decided. These two values give an estimative of the problem's duality gap, which is also given on the table as the ratio between the absolute gap and the smaller simulated cost for all cut types.

Table 7.2: Results for the convex model with 2 subsystems.

	Cut types		
	Benders	SB	ECTG
`Time (seconds)`	30	142	758
`Memory (GB)`	0.394	0.396	1.005
`Calculated cost (Bi R$)`	8.668	8.671	8.670
`Simulated cost (Bi R$)`	9.135	9.153	9.128
`Gap (%)`	5.04	5.01	5.02

Both the calculated and simulated costs are approximately the same for all cut types. An implication is that all methods find almost the same estimative for the gap. This is expected for a convex problem, since, besides for numerical reasons, all methods should calculate a cut that is tight for the expected cost-to-go at a chosen point. Nevertheless, the computational cost for the Benders cuts is drastically smaller than these of the other types of cuts, probably due to the simplicity of this method. The time required by the ECTG cuts is more than 5 times larger than that of the Benders cuts, showing that there is no advantage in using the linked formulation if the problem is convex.

Now we proceed to compare the results obtained by the SDDP algorithm for the non-convex model in Equation (7.11). Besides the addition of the disjunctive constraints for the minimum stored operational energy, everything is as in the previous section, including the 500 scenarios used for calculating the *simulated cost*. The results are summarized in Table 7.3.

Since this problem contains binary variables, there is a visible difference between the different cut types. The calculated cost for the ECTG cuts is about 10% higher in relation to both decomposed Benders and strengthened Benders cuts. This means that while the relative gap using the decomposed formulation is no less than 14.44%, the linked formulation can reduce the gap to 8.19%.

Table 7.3: Results for the non-convex model with 2 subsystems.

	Cut types		
	Benders	SB	ECTG
`Time (seconds)`	41	759	16854
`Memory (GB)`	0.422	0.418	1.089
`Calculated cost (Bi R$)`	10.222	10.410	11.170
`Simulated cost (Bi R$)`	12.231	12.227	12.209
`Gap (%)`	15.98	14.44	8.19

This is analogous to the phenomenon observed in Figures 7.5 and 7.6, where the decomposed cuts could not "touch" the real expected cost-to-go everywhere. Concerning the computational cost, the memory usage is about the same of the convex case but there huge differences on the elapsed time. The decomposed Benders cuts required a little more time, probably due to the need of solving a mixed integer program during the forward step. The time required by the decomposed strengthened Benders cuts was about 5 times higher than the convex case because for calculating tight cuts, we also need to solve the Lagrangian relaxation of mixed integer programs during the backward steps. Despite this increase in the required time, these cuts provide an increase in the lower bound given by the calculated cost in comparison to the Benders cuts, provoking a reduction of about 1.5% on the estimated gap. The ECTG cuts take about 22 times longer to calculate in the mixed integer setting than for the convex problem because it requires solving a large mixed integer program during all backward steps. In comparison with the other cut types, the ECTG cuts take about 400 times longer than the Benders cuts and 22 times longer than the decomposed strengthened Benders cuts. This increase in computational cost is counterbalanced by a large reduction on the estimated gap.

Since calculating cuts with the linked formulation takes much longer than with the decomposed formulation, a question that may arise is how much the gap could be reduced if we let the algorithm calculate decomposed cuts for the same time required to calculate 500 ECTG cuts. In Table 7.4, we show the result of calculating 4000 decomposed strengthened Benders cuts, which takes a little longer than the 500 ECTG cuts.

As we can see, the additional 3500 decomposed cuts do not make a great difference for the estimated gap, reducing it by only 0.85%. Nonetheless, the ECTG were capable, with the same amount of time, of reducing the estimated gap by 6.25% with respect to the original 500 strengthened Benders cuts.

Table 7.4: Comparison between decomposed and linked strengthened Benders cuts for approximately the same running time.

	Cut types	
	SB	ECTG
`Time (seconds)`	18154	16854
`Memory (GB)`	0.474	1.089
`Calculated cost (Bi R$)`	10.498	11.170
`Simulated cost (Bi R$)`	12.141	12.209
`Gap (%)`	13.79	8.19

Bibliography

Ahmed, S., Cabral, F. G., and da Costa, B. F. P. (2019). Stochastic Lipschitz dynamic programming. preprint avaiable on webpage at `http://www.optimization-online.org/DB_HTML/2019/05/7193.html`.

Artstein, Z. and Wets, R. J.-B. (1995). *Consistency of Minimizers and the SLLN for Stochastic Programs*, volume 2. Heldermann Verlag.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3):203–228.

Attouch, H. (1984). *Variational convergence for functions and operators*. Pitman Advanced Pub. Program.

Balas, E. (2011). Disjunctive programming. *Wiley Encyclopedia of Operations Research and Management Science*.

Bank, B., Guddat, J., Klatte, D., Kummer, B., and Tammer, K. (1984). Non-linear parametric optimization. *SIAM Review*, 26(4):594–595.

Benders, J. F. (1962). Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik*, 4(1):238–252.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98.

Blair, C. and Jeroslow, R. (1977). The value function of a mixed integer program: I. *Discrete Mathematics*, 19(2):121–138.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, New York, NY, USA.

Cabral, F. G. (2018). The role of extreme points for convex hull operations. Master's thesis, Instituto de Matemática – Universidade Federal do Rio de Janeiro, Rio de Janeiro.

Dantzig, G. B. (1963). *Linear Programming and Extensions*. Princeton University Press.

Diestel, J. and Uhl, J. J. (1977). *Vector measures*. American Mathematical Society, Providence, R.I.

Dowson, O. and Kapelevich, L. (2017). SDDP.jl: a Julia package for stochastic dual dynamic programming. *Optimization Online*.

Durrett, R. (2017). *Probability: Theory and examples*. Cambridge Universy Press.

Fiacco, A. V. and Kyparisis, J. (1986). Convexity and concavity properties of the optimal value function in parametric nonlinear programming. *Journal of Optimization Theory and Applications*, 48(1):95–126.

Fischer, T. (2012). Existence, uniqueness, and minimality of the Jordan measure decomposition. *arXiv e-prints*, page arXiv:1206.5449.

Folland, G. B. (1999). *Real analysis: Modern Techniques And Their Applications*. Wiley, New York.

Guigues, V. (2016). Convergence analysis of sampling-based decomposition methods for risk-averse multistage stochastic convex programs. *SIAM Journal on Optimization*, 26(4):2468–2494.

Gurobi Optimization, I. (2016). Gurobi optimizer reference manual.

Hassanzadeh, A. and Ralphs, T. (2014). On the value function of a mixed integer linear optimization problem and an algorithm for its construction. Technical report, COR@L Laboratory Report 14T-004, Lehigh University.

Hörmander, L. (2003). *The Analysis of Linear Partial Differential Operators I*. Springer Berlin Heidelberg.

Lax, P. D. (1997). *Linear Algebra and its applications*. John Wiley & Sons, Inc.

Lions, J.-L. (1951). Supports de produits de composition. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences*, 232:1530–1532.

Lucchetti, R. (2005). *Convexity and Well-Posed Problems*. CMS Books in Mathematics. Springer, 1st edition.

Pereira, M. V. F. and Pinto, L. M. V. G. (1991). Multi-stage stochastic optimization applied to energy planning. *Mathematical Programming*, 52(1-3):359–375.

Philpott, A. and Guan, Z. (2008). On the convergence of stochastic dual dynamic programming and related methods. *Operations Research Letters*, 36(4):450–455.

Rao, M. M. (2011). *Random and Vector Measures*. World Scientific Pub Co Inc.

Robertson, J. B. and Rosenberg, M. (1968). The decomposition of matrix-valued measures. *The Michigan Mathematical Journal*, 15(3):353–368.

Rockafellar, R. T. (1996). *Convex Analysis*. Princeton University Press.

Rockafellar, R. T. and Uryasev, S. (2000). Optimization of conditional value-at-risk. *The Journal of Risk*, 2(3):21–41.

Rockafellar, R. T. and Wets, R. J.-B. (2011). *Variational Analysis*. Springer.

Schechter, E. (1997). *Handbook of Analysis and Its Foundations*. Academic Press.

Schwartz, L. (1966). *Théorie des distributions*. Hermann, Paris, 3rd edition.

Shapiro, A. (2011). Analysis of stochastic dual dynamic programming method. *European Journal of Operational Research*, 209(1):63–72.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. P. (2014). *Lectures on Stochastic Programming: Modeling and Theory*. Society for Industrial and Applied Mathematics, 2nd edition.

Tao, T. (2011). *An introduction to measure theory*. American Mathematical Society, Providence, Rhode Island.

Zou, J., Ahmed, S., and Sun, X. A. (2018). Stochastic dual dynamic integer programming. *Mathematical Programming*, 175(1-2):461–502.