Report Number: WUCSE-2007-12

2007

A Duality Theory with Zero Duality Gap for Nonlinear Programming

Yixin Chen

Duality is an important notion for constrained optimization which provides a theoretical foundation for a number of constraint decomposition schemes such as separable programming and for deriving lower bounds in space decomposition algorithms such as branch and bound. However, the conventional duality theory has the fundamental limit that it leads to duality gaps for nonconvex optimization problems, especially discrete and mixed-integer problems where the feasible sets are nonconvex. In this paper, we propose a novel extended duality theory for nonlinear optimization that overcomes some limitations of previous dual methods. Based on a new dual function, the extended duality theory... Read complete abstract on page 2.

A Duality Theory with Zero Duality Gap for Nonlinear Programming

Yixin Chen

Complete Abstract:

Duality is an important notion for constrained optimization which provides a theoretical foundation for a number of constraint decomposition schemes such as separable programming and for deriving lower bounds in space decomposition algorithms such as branch and bound. However, the conventional duality theory has the fundamental limit that it leads to duality gaps for nonconvex optimization problems, especially discrete and mixed-integer problems where the feasible sets are nonconvex. In this paper, we propose a novel extended duality theory for nonlinear optimization that overcomes some limitations of previous dual methods. Based on a new dual function, the extended duality theory leads to zero duality gap for general nonconvex problems defined in discrete, continuous, and mixed-integer spaces under mild conditions.

2007-12

A Duality Theory with Zero Duality Gap for Nonlinear Programming

Authors: Yixin Chen

Corresponding Author: chen@cse.wustl.edu

Web Page: http://www.cse.wustl.edu/~chen/

Abstract: Duality is an important notion for constrained optimization which provides a theoretical foundation for a number of constraint decomposition schemes such as separable programming and for deriving lower bounds in space decomposition algorithms such as branch and bound. However, the conventional duality theory has the fundamental limit that it leads to duality gaps for nonconvex optimization problems, especially discrete and mixed-integer problems where the feasible sets are nonconvex. In this paper, we propose a novel extended duality theory for nonlinear optimization that overcomes some limitations of previous dual methods. Based on a new dual
function, the extended duality theory leads to zero duality gap for general nonconvex problems defined in discrete, continuous,
and mixed-integer spaces under mild conditions.

Type of Report: Other

A Duality Theory with Zero Duality Gap for Nonlinear Programming

Yixin Chen

Department of Computer Science and Engineering
Washington University in St. Louis
Saint Louis, MO 63130, USA
Email: chen@cse.wustl.edu

Abstract

Duality is an important notion for constrained optimization which provides a theoretical foundation for a number of constraint decomposition schemes such as separable programming and for deriving lower bounds in space decomposition algorithms such as branch and bound. However, the conventional duality theory has the fundamental limit that it leads to duality gaps for nonconvex optimization problems, especially discrete and mixed-integer problems where the feasible sets are nonconvex. In this paper, we propose a novel extended duality theory for nonlinear optimization that overcomes some limitations of previous dual methods. Based on a new dual function, the extended duality theory leads to zero duality gap for general nonconvex problems defined in discrete, continuous, and mixed-integer spaces under mild conditions.

1 Introduction

In this paper, we study solving the general nonlinear programming problem (NLP) of the following form:

$$(P_m): \quad \min_{z} \quad f(z), \qquad \text{subject to} \quad h(z) = 0 \ \text{ and } \ g(z) \leq 0, \tag{1}$$

where variable $z = (x, y)$, $x \in X$ is the continuous part, where X is a compact subset of $\mathbb{R}^n$, and $y \in Y$ is the discrete part, where Y is a finite discrete set of k-element integer vectors. We assume that the objective function f is lower bounded and is continuous and differentiable with respect to x, whereas the constraint functions $g = (g_1, \ldots, g_r)^T$ and $h = (h_1, \ldots, h_m)^T$ are continuous in the continuous subspace X for any given $y \in Y$.

The NLP defined in (1) cover a large class of nonlinear optimization problems. When both x are y present in z, it is a mixed-integer NLP (MINLP). It becomes a continuous NLP (CNLP) when there are only continuous variables x, and a discrete NLP (CNLPs) when there are only discrete variables y.

Duality is an important notion for mathematical programming and provides a rich theory for global optimization of NLPs. Duality can be used to directly solve NLPs as well as to derive lower bounds of the solution quality which is the key to many global optimization algorithms such as branch and bound.

An important issue is the existence of the *duality gap*, i.e. the difference between the optimal solution quality of the original problem and the lower bound obtained by solving the dual problem. The duality gap is often nonzero for nonconvex problems, and may be large for some problems, in which case the duality approach is not useful. Moreover, the duality theory has greater difficulty with discrete and mixed-integer problems, for which the duality gap may be nonzero even if the functions are convex.

The rest of the paper is organized as follows. In Section 2, we briefly review related existing work. In Section 3, we present the proposed theory of extended duality.

2 Related Previous Work

In this section, we review some related previous work and discuss their limitations and differences to the proposed approach. We first overview the duality theory and previous decompositions methods based on duality. Then, we review previous work for reducing or removing the duality gap.

2.1 Duality

Duality is an important notion for mathematical programming and provides a rich theory for global optimization. Many theoretical results of duality are developed for continuous nonlinear programming (CNLP) problems defined as the following.

$$\begin{aligned} (P_c): \qquad & \min_x \quad f(x) \qquad \text{where } x = (x_1, \ldots, x_n)^T \in X \\ & \text{subject to} \quad h(x) = (h_1(x), \ldots, h_m(x))^T = 0 \quad \text{and} \quad g(x) = (g_1(x), \ldots, g_r(x))^T \leq 0, \end{aligned} \tag{2}$$

where X is a compact subset of $\mathbb{R}^n$, f is lower bounded, continuous and differentiable, and g and h are continuous.

The duality theory is based on a Lagrangian function of the form:

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{r} \mu_j g_j(x) \tag{3}$$

Dual methods transform the original problem into a dual problem defined as follows:

$$(P_{dual}): \quad \begin{array}{ll} \text{maximize} & q(\lambda, \mu) \\ \text{subject to} & \lambda \in \mathbb{R}^m \text{ and } \mu \geq 0, \end{array} \tag{4}$$

where the dual function $q(\lambda, \mu)$ is defined as:

$$q(\lambda, \mu) = \inf_{x \in X} L(x, \lambda, \mu) = \inf_{x \in X} \left[f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{r} \mu_j g_j(x) \right]. \tag{5}$$

The main results of the dual theory are the following.

First, the objective value q^* obtained from solving the dual problem (P_{dual}) is a lower bound to the optimal objective value f^* of the original problem, i.e. $q^* \leq f^*$. Namely, the solution to the dual problem is a *lower bound* of the objective value of the original problem. The difference between q^* and f^* is called the *duality gap*.

Second, for CNLPs with a convex objective function and convex feasible sets, there is no duality gap under very general conditions. Therefore, for these problems, solving the original problem is equivalent to solving the dual problem, which is much easier in many cases. Usually, for problems without duality gap, a dual method carries out a two-level search that tries to find μ to maximize $q(\mu)$ at the top level and look for z to minimize $L(z, \mu)$ at the lower level. The dual method is most powerful when there is no duality gap and when the minimization of $L(z, \mu)$ can be done in closed form or is relatively simple.

A major benefit of using a dual formulation is that, when the problem is well structured, the solution of the the dual problem can be made faster by using decomposition. For example, the *separable programming* [5, 2, 12, 11, 23, 22, 13] solves the following problem, where variables x has m components $x_1, \cdots, x_m$ of dimension $n_1, \cdots, n_m$, respectively:

$$\begin{array}{ll} \text{minimize} & \displaystyle\sum_{i=1}^{m} f_i(x_i) \\ \text{subject to} & \displaystyle\sum_{i=1}^{m} g_{ij}(x_i) \leq 0, j = 1, \cdots, r, \\ & x_i \in X_i, i = 1, \cdots, m. \end{array} \tag{6}$$

Here f_i and g_{ij} are continuous and differentiable functions, and X_i is a given subset in R^{n_i}. Note that if the constraints $\sum_{i=1}^{m} g_{ij} \leq 0$ were not present in (6), then it would be straightforward to decompose this problem into m independent subproblems. However, the constraints link all the subproblems together and create possibly global inconsistencies.

Separable programming methods consider the following dual problem of (6):

$$\text{maximize} \qquad q(\mu) \tag{7}$$

$$\text{subject to} \qquad \mu \geq 0, \tag{8}$$

where μ is the vector of Lagrange multipliers and the dual function $q(\mu)$ is formulated as:

$$q(\mu) = \inf_{x_i \in X_i, i=1..m} \left\{ \sum_{i=1}^{m} \left(f_i(x_i) + \sum_{j=1}^{r} \mu_j g_{ij}(x_i) \right) \right\} = \sum_{i=1}^{m} q_i(\mu) \tag{9}$$

and

$$q_i(\mu) = \inf_{x_i \in X_i} \left\{ f_i(x_i) + \sum_{j=1}^{r} \mu_j g_{ij}(x_i) \right\}, \quad i = 1, \cdots, m. \tag{10}$$

Therefore, the minimization involved in computing the dual function $q(\mu)$ in (9) can be decomposed into m simpler subproblems in (10). These minimizations on the subproblems can be done efficiently when the functions in the subproblems are convex or linear, which lead to efficient computation of the overall dual function.

In addition to separable programming, the *Dantzig-Wolfe decomposition* [9] can also be viewed as a method that decomposes a piecewise linear approximation to the dual function for linealy constrained problems with a separable constraint structure [5].

The dual theory has some limitations. The direct dual methods work only for convex problems with linear or convex constraints and cannot solve general NLPs with nonconvex functions due to the duality gap. This greatly restricts its applicability. This limitation is particularly restrictive for discrete nonlinear programming (DNLP) problems and mixed-integer nonlinear programming (MINLP) problems, since their variable spaces are usually nonconvex. Therefore, for DNLPs and MINLPs, there can be duality gaps even when the functions are linear or convex. As a consequence of the above limitations, existing duality-based decomposition methods, such as separable programming and Dantzig-Wolfe decomposition, work only for convex or linear problems and require continuity and differentiability of the functions. For nonconvex CNLPs, DNLPs, and MINLPs, although they often have the separable constraint structure, these decomposition methods cannot be applied due to the duality gap. It is the objective of this paper to develop a new duality theory that leads to no duality gap and supports decompositions for these general CNLPs, DNLPs, and MINLPs with separable structure.

Note that duality is also widely used for DNLPs and MINLPs in some space decomposition methods such as branch and bound (B&B) and generalized Bender's decomposition (GBD). However, in these methods, duality is used to derive lower bounds of subproblems rather than to

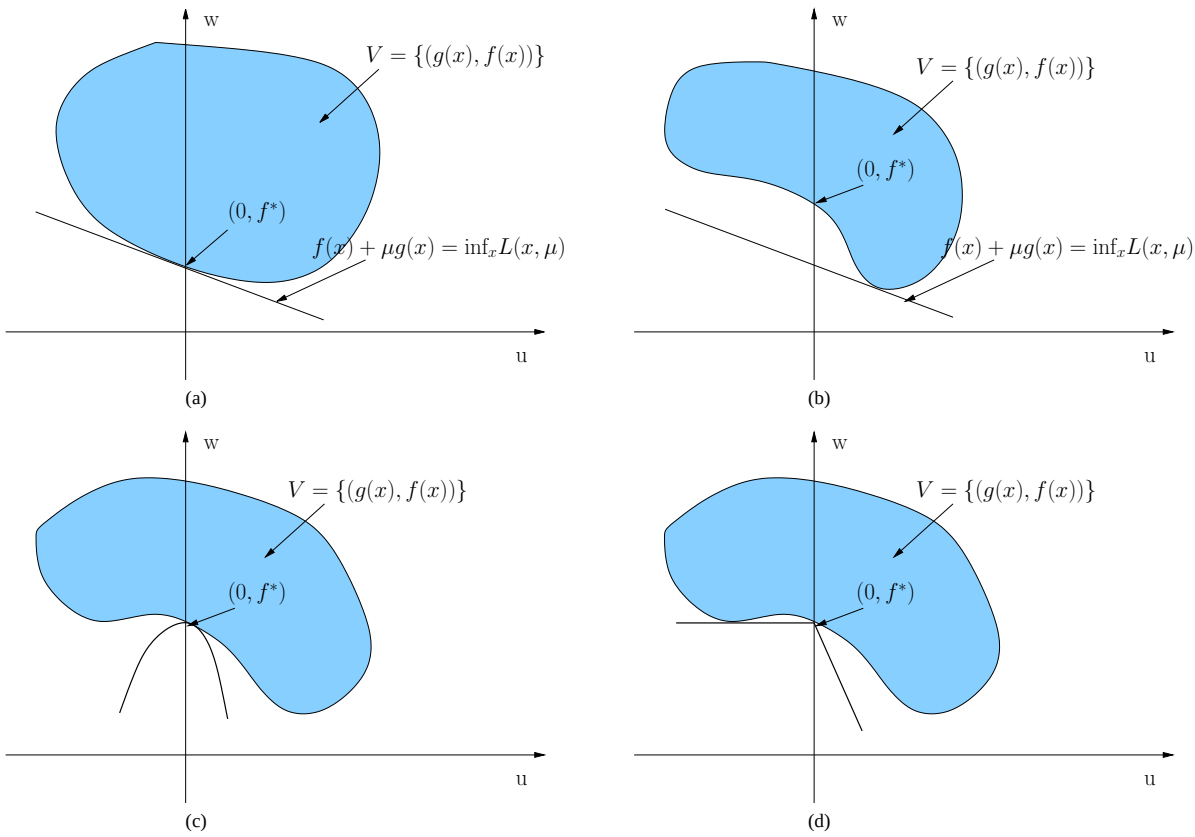

Figure 1: Geometric interpretation of duality. (a) The zero duality gap achieved by the Lagrangian function for convex problems. (b) The nonzero duality gap of the Lagrangian function for nonconvex problems. (c) Using an augmented Lagrangian function can remove the duality gap. (d) Using an exact penalty function can remove the duality gap.

decompose the problem at the high level. In fact, in these methods, the problem is decomposed by the variable space instead of the separable constraints.

In this research, we propose a new duality theory that has no duality gap for general, nonconvex problems, and that works for discrete, continuous, and mixed problems in a unified fashion.

2.2 Removing the duality gap for nonconvex optimization

There have been extensive previous studies aiming at reducing or eliminating the duality gap. A number of previous work has indicated that the duality gap can be reduced when a problem is decomposed or has certain special structures [1, 3, 24].

It is well known that the existence of duality gap is closely related to the geometric problem of finding the hyperplane supporting the set V of constraint-objective pairs (c.f. Section 5, [5]). Figure 1 visualizes this relation for inequality-constrained problems. The primal problem can be

visualized as finding the minimum intercept of V with the w-axis, while the dual problem can be visualized as finding the maximum point of interception from all the hyperplanes supporting V from below. It can be seen that for convex problems, the minimum intercept of V and the maximum intercept of the supporting hyperplane are identical (Figure 1.a). For nonconvex problems, there is a gap between the two intercepts (Figure 1.b).

To remove the duality gaps for nonconvex problems, augmented Lagrangian functions [18, 4] were introduced for continuous NLPs. The idea can be visualized as penetrating the dent at the bottom of V by introducing a nonlinear augmenting function (Figure 1.c). It has also been shown that, instead of using the classical Lagrangian function, using an ℓ_1-penalty function can lead to zero duality gap for nonconvex problems under mild conditions [7, 8]. The geometric interpretation of exact penalty functions is visualized as the solid line in Figure 1.d.

Recently, it has been an active research topic on developing general dual functions with zero duality gap for nonconvex continuous optimiztion [6, 14, 16, 7, 8, 15, 20, 21, 25, 19] that accommodate both augmented Lagrangian functions and exact penalty functions.

For a continuous problem in (2), most of the existing augmented Lagrangian functions and exact penalty functions that achieve zero duality gap for nonconvex problems fit into the following general function [14, 16]:

$$l(x, \lambda, \mu, c) = f(x) + \tau(\lambda, \mu, h, g) + c\sigma(h, g) \tag{11}$$

where λ, μ are the Lagrange-multiplier vector, $\tau(\lambda, \mu, h, g)$ is a nonlinear Lagrangian term, $c \geq 0$ is a penalty parameter, and $\sigma(h, g)$ is an augmenting function. When λ and μ are 0, $l(x, \lambda, \mu, c)$ becomes a penalty function; when c is 0, $l(x, \lambda, \mu, c)$ becomes a nonlinear Lagrangian function; and when c is 0 and $\tau(\lambda, \mu, h, g) = \lambda^T g(x) + \mu^T g(x)$, $l(x, \lambda, \mu, c)$ becomes the Lagrangian function.

Rubinov et al. [20, 25] have extended the ℓ_1- penalty function to a class of nonlinear penalty functions with zero duality gap, where the functions take the following form:

$$l_\gamma(x, c) = \left[f^\gamma(x) + c\left(\sum_{i=1}^{m} |h_i(x)|^\gamma + \sum_{j=1}^{r} g_j^+(x)^\gamma \right) \right]^{1/\gamma}, \tag{12}$$

where $\gamma > 0$ is a parameter.

Luo et al. [15] have proposed a nonconvex and nonsmooth penalty function with zero duality gap based on the following formulation, where $\gamma > 0$:

$$l_\gamma(x, c) = f(x) + c\left(\sum_{i=1}^{m} |h_i(x)| + \sum_{j=1}^{r} g_j^+(x) \right)^\gamma. \tag{13}$$

An exact penalty function with zero duality gap under certain assumptions is proposed by Pang [17] as follows:

$$l_\gamma(x,c) = f(x) + c\left[\max\left\{|h_1(x)|,\cdots,|h_m(x)|,,g_1^+(x),\cdots,g_r^+(x)\right\}\right]^\gamma. \tag{14}$$

There are a number of efforts to provide unified frameworks to characterize the augmented Lagrangian functions and exact penalty functions with zero duality gaps for nonconvex problems. Rockafellar and Wets [19] have proposed a class of augmented Lagrangian functions with a convex, nonnegative augmenting term, which lead to zero duality gap for constrained optimization problems under coercivity assumptions. A general framework that provides a unified treatment for a family of Lagrange-type functions and conditions for achieving zero duality gap is given by Burachik and Rubinov [6]. A recent work by Nedić and Ozdaglar [16] developes necessary and sufficient conditions for $l(x,\lambda,\mu,c)$ to have zero duality gaps based on a geometric analysis, which considers the *geometric primal problem* of finding the minimum intercept of the epigraph V and the *geometric dual problem* of finding the maximum intercept of the supporting hyperplanes of V. Huang and Yang [14] have proposed a generalized augmented Lagrangian function, which includes many previous work as special cases, and proved the zero duality gap and exact penalization for this function.

Remarks. Several observations on the limitations of previous work motivate our work in this paper. Most results are developed for continuous or semi-continuous problems. This is partly due to the fact that discrete and mixed problems often have nonconvex feasible sets. The results we develop in this paper provide a unified theory for continuous, discrete, and mixed problems. Further, as we can see from (11) to (14), that all the previous methods for removing the duality gaps use a single penalty multiplier c. However, a suitable c (and the associated *unique* Lagrange multipliers, if used) is often large to locate and control. In practice, a popular problem is that the single c is often too large, which makes the search difficult. In this paper, we propose to use multiple penalty multiples which can effectively lead to smaller penalty values for ensuring a zero duality gap.

3 Theory of Extended Duality

We describe in this section our theory of extended duality in discrete, continuous, and mixed spaces based on an ℓ_1^m-penalty function. Since the result for MINLPs is derived based on the results for continuous and discrete NLPs, we will first develop the theory for continuous and discrete problems before presenting a unified theory for mixed problems.

3.1 Extended duality for continuous optimization

We first develop our results for continuous nonlinear programming problems (CNLPs) defined as P_c in (2).

Definition 3.1 ***(Constrained Global Minimum of** P_c**)*** *A point $x^* \in X$ is a CGM_c, a constrained global minimum of P_c, if x^* is feasible and $f(x^*) \leq f(x)$ for all feasible $x \in X$.*

Definition 3.2 *The ℓ_1^m-penalty function for P_c in (2) is defined as follows:*

$$L_m(x, \alpha, \beta) = f(x) + \alpha^T |h(x)| + \beta^T g^+(x), \tag{15}$$

where $|h(x)| = (|h_1(x)|, \ldots, |h_m(x)|)^T$ and $g^+(x) = (g_1^+(x), \ldots, g_r^+(x))^T$, where we define $\phi^+(x) = \max(0, \phi(x))$ for a function ϕ, and $\alpha \in \mathbb{R}^m$ and $\beta \in \mathbb{R}^r$ are penalty multipliers.

In the term *ℓ_1^m-penalty*, the subscript 1 denotes the fact that L_m uses an ℓ_1 transformation of the constraints, while the superscript m denotes the fact that L_m has multiple penalty multipliers as opposed to the single penalty multiplier used by the conventional ℓ_1-penalty.

We consider the *extended dual function* defined for $\alpha \in \mathbb{R}^m$ and $\beta \in \mathbb{R}^r$ as:

$$q(\alpha, \beta) = \min_{x \in X} L_m(x, \alpha, \beta). \tag{16}$$

It is straightforward to show that the dual function $q(\alpha, \beta)$ is concave over $\alpha \geq 0$ and $\beta \geq 0$. We define the *extended dual problem* as:

$$\begin{aligned} \text{maximize} \quad & q(\alpha, \beta) \\ \text{subject to} \quad & \alpha \geq 0, \quad \text{and} \quad \beta \geq 0, \end{aligned} \tag{17}$$

and the *optimal extended dual value* as:

$$q^* = \max_{\alpha \geq 0, \beta \geq 0} q(\alpha, \beta). \tag{18}$$

For continuous problems, we need the following constraint-qualification condition in order to rule out the special case in which all continuous constraints have zero derivative along a direction.

Definition 3.3 *The directional derivative of a function $f : \mathbb{R}^n \mapsto \mathbb{R}$ at a point $x \in \mathbb{R}^n$ along a direction $p \in \mathbb{R}^n$ is:*

$$f'(x; p) = \lim_{\epsilon \to 0} \frac{f(x + \epsilon p) - f(x)}{\epsilon}. \tag{19}$$

Definition 3.4 Constraint-qualification condition. *A point $x \in X$ of P_c meets the constraint qualification if there exists no direction $p \in \mathbb{R}^n$ along which the directional derivatives of continuous equality and continuous active inequality constraints are all zero. That is,*

$$\nexists\, p \in \mathbb{R}^n \text{ such that } h_i'(x;p) = 0 \text{ and } g_j'(x;p) = 0, \quad \forall i \in C_h \text{ and } j \in C_g, \tag{20}$$

where C_h and C_g are, respectively, the sets of indices of continuous equality and continuous active inequality constraints. The constraint qualification is satisfied if both C_h and C_g are empty.

Intuitively, constraint qualification at x ensures the existence of finite α and β that lead to a local minimum of (15) at x. Consider a neighboring point $x + p$ infinitely close to x, where the objective function f at x decreases along p and all active constraints at x have zero directional derivative along p. In this case, all the active constraints at $x + p$ are close to zero, and it will be impossible to find finite α and β in order to establish a local minimum of (15) at x with respect to $x + p$. To ensure a local minimum of (15) at x, the above scenario must not be true for any p at x.

Definition 3.5 ***Feasible Set and ϵ-Extension.*** *Let the set of all feasible points of P_c be:*

$$\mathcal{F} = \left\{ x \;\middle|\; x \in X, h(x) = 0, g(x) \leq 0 \right\}, \tag{21}$$

the ϵ-extension of $\mathcal{F}$, where $\epsilon > 0$ is a scalar value, is:

$$\mathcal{F}_\epsilon^+ = \left\{ x \;\middle|\; x \in X, \quad (\min_{y \in \mathcal{F}} \|y - x\|) \leq \epsilon \right\}. \tag{22}$$

Namely, $\mathcal{F}_\epsilon^+$ includes the points in $\mathcal{F}$ and all those points whose projection distance to $\mathcal{F}$ is within ϵ. Here, $\|\cdot\|$ denotes the Euclidean norm.

Lemma 3.1 *For any constant $\epsilon > 0$, there exists a finite scalar value $\xi > 0$ such that:*

$$\|h(x)\|^2 + \|g^+(x)\|^2 \geq \xi, \quad \text{for any } x \in X - \mathcal{F}_\epsilon^+. \tag{23}$$

Proof. We prove by contradiction. Suppose we cannot find such a ξ, then for a sequence $\{\xi_1, \xi_2, \cdots\}$ where $\lim_{i\to\infty} \xi_i = 0$, there exists a sequence $\{x_1, x_2, \cdots,\}, x_i \in X - \mathcal{S}_\epsilon^+, i = 1, 2, \cdots$, such that:

$$\|h(x_i)\|^2 + \|g^+(x_i)\|^2 \leq \xi_i. \tag{24}$$

Since $X - \mathcal{F}_\epsilon^+$ is bounded, the $\{x_i\}$ sequence has at least one limit point x. Since $X - \mathcal{F}_\epsilon^+$ is closed, x belongs to $X - \mathcal{F}_\epsilon^+$. From the continuity of $h(x)$ and $g(x)$, we have:

$$\|h(x)\|^2 + \|g^+(x)\|^2 = \lim_{i\to\infty} \|h(x_i)\|^2 + \|g^+(x_i)\|^2 \leq \lim_{i\to\infty} \xi_i = 0, \tag{25}$$

which implies that $\|h(x)\|^2 + \|g^+(x)\|^2 = 0$. Thus, we must have $h(x) = 0$ and $g(x) \leq 0$, which means that x is feasible and contradicts to the assumption that $x \in X - \mathcal{F}_\epsilon^+$ is outside of the feasible set.

The following theorems state the main results of extended duality.

Theorem 3.1 *Suppose $x^* \in X$ is a CGM_c to P_c and x^* satisfies the constraint qualification, then there exists finite $\alpha^* \geq 0$ and $\beta^* \geq 0$ such that* [1]

$$f(x^*) = \min_{x \in X} L_m(x, \alpha^{**}, \beta^{**}), \text{ for any } \alpha^{**} \geq \alpha^*, \beta^{**} \geq \beta^*. \tag{26}$$

Proof. Since we have:

$$L_m(x^*, \alpha^{**}, \beta^{**}) = f(x^*) + \sum_{i=1}^{m} \alpha_i^{**} |h_i(x)| + \sum_{j=1}^{r} \beta_j^{**} g_j^+(x) = f(x^*), \ \forall \alpha^{**} \geq 0, \beta^{**} \geq 0, \tag{27}$$

it is equivalent to show that there exist finite $\alpha^* \geq 0$ and $\beta^* \geq 0$ such that

$$f(x^*) \leq L_m(x, \alpha^{**}, \beta^{**}), \text{ for any } \alpha^{**} > \alpha^*, \beta^{**} > \beta^*, \tag{28}$$

for any $x \in X$. We prove (28) in three parts. First, we prove that (28) is true for any point x in the feasible set $\mathcal{F}$. Then, we show that (28) is true for any point x within $\mathcal{F}_{\epsilon_{min}}^+$ for a small $\epsilon_{min} > 0$. Last, we prove (28) for the the points in $X - \mathcal{F}_{\epsilon_{min}}^+$. For simplicity, we assume that x^* is the only CGM_c in X. The case of multiple CGM_c can be proved similarly.

Part a). For every feasible point $x' \in \mathcal{F}$, (28) is true for any $\alpha^{**} \geq 0$ and $\beta^{**} \geq 0$ since

$$L_m(x', \alpha^{**}, \beta^{**}) = f(x') \geq f(x^*), \tag{29}$$

noting that $h(x') = 0$ and $g(x') \leq 0$, and $f(x') \geq f(x^*)$ by the definition of CGM_c.

Part b). We show that (28) is satisfied in $\mathcal{F}_\epsilon^+$ when ϵ is small enough. To this end, we show that for each feasible point $x' \in \mathcal{F}$, any point x in the close neighborhood of x' satisfies (28).

For any feasible $x' \in \mathcal{F}$ that is not in the neighborhood of x^*, we have $f(x') - f(x^*) \geq \xi > 0$ for a finite positive ξ since x^* is the only CGM_c. Let $x = x' + \epsilon p$, $p \in \mathbb{R}^n, \|p\| = 1$ is a unit-length direction vector and $\epsilon = \|x - x'\|$. When ϵ is small enough, we have:

$$\begin{aligned} L_m(x, \alpha^{**}, \beta^{**}) &= f(x) + \sum_{i=1}^{m} \alpha_i^{**} |h_i(x)| + \sum_{j=1}^{r} \beta_j^{**} g_j^+(x) \\ &\geq f(x) = f(x') + \epsilon \nabla_x f(x')^T p + o(\epsilon^2) \\ &\geq f(x^*) + \xi + \epsilon \nabla_x f(x')^T p + o(\epsilon^2) \geq f(x^*). \end{aligned} \tag{30}$$

[1] Given two vectors a and b of the same size n, we say that $a \geq b$ if $a_i \geq b_i$ for $i = 1, \cdots, n$.

For any point x in the neighborhood of x^*, let $x = x^* + \epsilon p$, where $p \in \mathbb{R}^n$, $\|p\| = 1$ is a unit-length direction vector and $\epsilon = \|x - x^*\|$. We show that when ϵ is small enough, there always exist finite α^* and β^* such that (28) is true. We consider the following two cases:

Case 1) If at x^* all the constraints are inactive inequality constraints, then when ϵ is small enough, x is also a feasible point. Hence, x^* being a CGM_c implies that $f(x) \geq f(x^*)$ and, regardless the choice of the penalties,

$$L_m(x, \alpha^{**}, \beta^{**}) = f(x) + \sum_{i=1}^{m} \alpha_i^{**} |h_i(x)| + \sum_{j=1}^{r} \beta_j^{**} g_j^+(x) = f(x) \geq f(x^*). \tag{31}$$

Case 2) Other than inactive inequality constraints, if there are equality or active inequality constraints at x^*, then according to the constraint-qualification condition, there must exist an equality constraint or an active inequality constraint that has non-zero derivative along p. Suppose there exists an equality constraint h_k that has non-zero derivative along p (the case with an active inequality constraint is similar), which means $|h_k'(x^*; p)| > 0$. If we set $\alpha_k^{**} > \frac{|\nabla_x f(x^*)^T p|}{|h_k'(x^*;p)|}$ and ϵ small enough, then:

$$\begin{aligned}
L_m(x, \alpha^{**}, \beta^{**}) &= f(x) + \sum_{i=1}^{m} \alpha_i^{**} |h_i(x)| + \sum_{j=1}^{r} \beta_j^{**} g_j^+(x) \\
&\geq f(x) + \alpha_k^{**} |h_k(x)| \quad \geq \quad f(x^*) + \epsilon \nabla_x f(x^*)^T p + o(\epsilon^2) + \alpha_k^{**} \epsilon |h_k'(x^*; p)| \\
&\geq f(x^*) + \epsilon \left(\alpha_k^{**} \left| h_k'(x^*; p) \right| - \left| \nabla_x f(x^*)^T p \right| \right) + o(\epsilon^2) \\
&\geq f(x^*).
\end{aligned} \tag{32}$$

Combining the results in part a) and b), and taking the minimum of the sufficiently small ϵ over all $x \in \mathcal{F}$, we have shown that there exists a finite $\epsilon_{min} > 0$ such that (28) is true for any point $x \in X$ in $\mathcal{F}_{\epsilon_{min}}^+$, the ϵ_{min}-extension of $\mathcal{F}$.

Part c). Part a) and b) have proved that (28) is true for any point $x \in \mathcal{F}_{\epsilon_{min}}^+$. We now prove that (28) is true for any point $x \in X - \mathcal{F}_{\epsilon_{min}}^+$.

For a point $x \in X - \mathcal{F}_{\epsilon_{min}}^+$, according to Lemma 3.1, there exists finite $\xi > 0$ such that

$$\|h(x)\|^2 + \|g^+(x)\|^2 \geq \xi. \tag{33}$$

Let $f_{min} = \min_{x \in X} f(x)$. Since $f(x)$ is lower bounded, f_{min} is finite. We set:

$$\alpha_i^* = \frac{f(x^*) - f_{min}}{\xi} \left| h_i(x) \right|, \qquad i = 1, \cdots, m, \tag{34}$$

$$\text{and} \qquad \beta_j^* = \frac{f(x^*) - f_{min}}{\xi} g_j^+(x), \qquad j = 1, \cdots, r. \tag{35}$$

Note that $\alpha^* \geq 0$ and $\beta^* \geq 0$ since $f(x^*) \geq f_{min}$.

We have, for any $\alpha^{**} \geq \alpha^*$, $\beta^{**} \geq \beta^*$:

$$\begin{aligned} L_m(x, \alpha^{**}, \beta^{**}) &= f(x) + \sum_{i=1}^{m} \alpha_i^{**} |h_i(x)| + \sum_{j=1}^{r} \beta_j^{**} g_j^+(x) \\ &\geq f(x) + \frac{f(x^*) - f_{min}}{\xi} \left(\|h(x)\|^2 + \|g^+(x)\|^2 \right) \\ &\geq f(x) + f(x^*) - f_{min} \qquad \text{(according to (33))} \\ &\geq f(x^*). \end{aligned} \tag{36}$$

(28) is shown after combining the three parts, thus completing the proof. ■

Given the above result, now we can show that the ℓ_1^m-penalty function leads to zero duality gap for a general CNLP defined as P_c.

Theorem 3.2 ***(Extended Duality Theorem for Continuous Nonlinear Programming)*** *Suppose $x^* \in X$ is a CGM$_c$ to P_c and x^* satisfies the constraint qualification, then there is no duality gap for the extended dual problem defined in (18), i.e. $q^* = f(x^*)$.*

Proof. First, we have $q^* \leq f(x^*)$ since

$$\begin{aligned} q^* &= \max_{\alpha \geq 0, \beta \geq 0} q(\alpha, \beta) = \max_{\alpha \geq 0, \beta \geq 0} \left(\min_{x \in X} L_m(x, \alpha, \beta) \right) \\ &\leq \max_{\alpha \geq 0, \beta \geq 0} L_m(x^*, \alpha, \beta) = \max_{\alpha \geq 0, \beta \geq 0} f(x^*) = f(x^*). \end{aligned} \tag{37}$$

Also, according to Theorem 3.1, there are $\alpha^{**} \geq 0$ and $\beta^{**} \geq 0$ such that $q(\alpha^{**}, \beta^{**}) = f(x^*)$, we have:

$$q^* = \max_{\alpha \geq 0, \beta \geq 0} q(\alpha, \beta) \geq q(\alpha^{**}, \beta^{**}) = f(x^*). \tag{38}$$

Since $q^* \leq f(x^*)$ and $q^* \geq f(x^*)$, we have $q^* = f(x^*)$. ■

3.2 Extended duality for discrete optimization

Consider the following DNLP

$$\begin{aligned} (P_d): \qquad \min_{y} \quad & f(y) \qquad \text{where } y = (y_1, \ldots, y_w)^T \in Y \\ \text{subject to} \quad & h(y) = 0 \quad \text{and} \quad g(y) \leq 0. \end{aligned} \tag{39}$$

whose f is lower bounded, Y is a finite discrete set, and f, g and h are not necessarily continuous and differentiable with respect to y.

Definition 3.6 ***(Constrained Global Minimum of*** P_d***)*** *A point* $y^* \in Y$ *is a* CGM_d*, a constrained global minimum of* P_d*, if* y^* *is feasible and* $f(y^*) \leq f(y)$ *for all feasible* $y \in Y$*.*

Definition 3.7 *The* ℓ_1^m*-penalty function for* P_d *is defined as follows:*

$$L_m(y, \alpha, \beta) = f(y) + \alpha^T |h(y)| + \beta^T g^+(y), \tag{40}$$

where $\alpha \in \mathcal{R}^m$ *and* $\beta \in \mathcal{R}^r$*.*

Theorem 3.3 *Let* $y^* \in Y$ *be a* CGM_d *to* P_d*, there exists finite* $\alpha^* \geq 0$ *and* $\beta^* \geq 0$ *such that*

$$f(y^*) = \min_{y \in Y} L_m(y, \alpha^{**}, \beta^{**}), \text{ for any } \alpha^{**} \geq \alpha^*, \beta^{**} \geq \beta^*. \tag{41}$$

Proof. Given y^*, since $L_m(y^*, \alpha^{**}, \beta^{**}) = f(y^*)$ for any $\alpha^{**} \geq 0$ and $\beta^{**} \geq 0$, we need to prove that there exist finite $\alpha^* \geq 0$ and $\beta^* \geq 0$ such that

$$f(y^*) \leq L_m(y, \alpha^{**}, \beta^{**}), \text{ for any } \alpha^{**} \geq \alpha^*, \beta^{**} \geq \beta^*, \tag{42}$$

for any $y \in Y$.

We set the following α^* and β^*:

$$\alpha_i^* = \max_{y \in Y, |h_i(y)| > 0} \left\{ \frac{f(y^*) - f(y)}{|h_i(y)|} \right\}, \qquad i = 1, \cdots, m, \tag{43}$$

$$\beta_j^* = \max_{y \in Y, g_j(y) > 0} \left\{ \frac{f(y^*) - f(y)}{g_j(y)} \right\}, \qquad j = 1, \cdots, r. \tag{44}$$

Next, we show that $f(y^*) \leq L_m(y, \alpha^{**}, \beta^{**})$ for any $y \in Y$, $\alpha^{**} \geq \alpha^*$, and $\beta^{**} \geq \beta^*$.

For a feasible point $y \in Y$, since $h(y) = 0$ and $g(y) \leq 0$, we have:

$$L_m(y, \alpha^{**}, \beta^{**}) = f(y) \geq f(y^*). \tag{45}$$

For an infeasible point $y \in Y$, if there is at least one equality constraint $h_i(y)$ that is not satisfied, we have:

$$\begin{aligned} L_m(y, \alpha^{**}, \beta^{**}) &= f(y) + \sum_{i=1}^{m} \alpha_i^{**} |h_i(y)| + \sum_{j=1}^{r} \beta_j^{**} g_j^+(y) \geq f(y) + \alpha_i^{**} |h_i(y)| \\ &\geq f(y) + \frac{f(y^*) - f(y)}{|h_i(y)|} |h_i(y)| = f(y^*) \end{aligned} \tag{46}$$

If there is at least one inequality constraint $g_j(y)$ that is not satisfied ($g_j(y) > 0$), we have:

$$\begin{aligned} L_m(y, \alpha^{**}, \beta^{**}) &= f(y) + \sum_{i=1}^{m} \alpha_i^{**} |h_i(y)| + \sum_{j=1}^{r} \beta_j^{**} g_j^+(y) \geq f(y) + \beta_i^{**} g_j(y) \\ &\geq f(y) + \frac{f(y^*) - f(y)}{g_j(y)} g_j(y) = f(y^*) \end{aligned} \tag{47}$$

(42) is proved after combining (45), (46), and (47). ■

The *extended dual problem* for P_d is the same as (16) to (18) defined for P_c, except that the variable space is Y instead of X. Based on Theorem 3.3, we have the following result for discrete-space extended duality, which can be proved in the same way as the proof to Theorem 3.2.

Theorem 3.4 ***(Extended Duality Theorem for Discrete Nonlinear Programming)*** *Suppose $y^* \in Y$ is a CGM_d to P_d, then there is no duality gap for the extended dual problem, i.e. $q^* = f(y^*)$.*

Note that the constraint-qualification condition in Theorem 3.1 is not needed in Theorem 3.3 because constraint functions are not changing continuously in discrete problems.

3.3 Extended duality for mixed optimization

Last, we present the extended duality results for the MINLP problem P_m defined in (1).

Definition 3.8 ***(Constrained Global Minimum of P_m)*** *A point $z^* = (x^*, y^*) \in X \times Y$ is a CGM_m, a constrained global minimum of P_m, if z^* is feasible and $f(z^*) \leq f(z)$ for all feasible $z \in X \times Y$.*

Definition 3.9 *The ℓ_1^m-penalty function for P_m is defined as follows:*

$$L_m(z, \alpha, \beta) = f(z) + \alpha^T |h(z)| + \beta^T g^+(z), \tag{48}$$

where $\alpha \in \mathcal{R}^m$ and $\beta \in \mathcal{R}^r$.

Theorem 3.5 *Let $z^* \in X \times Y$ be a CGM_m to P_m, there exist finite $\alpha^* \geq 0$ and $\beta^* \geq 0$ such that*

$$f(z^*) = \min_{z \in X \times Y} L_m(z, \alpha^{**}, \beta^{**}), \text{ for any } \alpha^{**} \geq \alpha^*, \beta^{**} \geq \beta^*. \tag{49}$$

Proof. Given $z^* = (x^*, y^*)$, since $L_m(z^*, \alpha^{**}, \beta^{**}) = f(z^*)$ for any $\alpha^{**} \geq 0$ and $\beta^{**} \geq 0$, we need to prove that, for each $z = (x, y) \in X \times Y$, there exist finite $\alpha^* \geq 0$ and $\beta^* \geq 0$ such that

$$f(x^*, y^*) \leq L_m(x, y, \alpha^{**}, \beta^{**}), \text{ for any } \alpha^{**} \geq \alpha^*, \beta^{**} \geq \beta^*. \tag{50}$$

Define:

$$V(y) = \min_{x' \in X} \left(\|h(x', y)\|^2 + \|g^+(x', y)\|^2 \right) \tag{51}$$

We consider two cases.

Case 1) Suppose we have $V(y) > 0$, then there is no feasible solution when the discrete part is fixed at y. Let $f_{min}|_y = \min_{x' \in X} f(x', y)$, we set:

$$\alpha_i^* = \frac{f(x^*, y^*) - f_{min}|_y}{V(y)} \Big| h_i(x) \Big|, \qquad i = 1, \cdots, m, \tag{52}$$

$$\text{and} \qquad \beta_j^* = \frac{f(x^*, y^*) - f_{min}|_y}{V(y)} g_j^+(x), \qquad j = 1, \cdots, r. \tag{53}$$

We have, for any $\alpha^{**} \geq \alpha^*$, $\beta^{**} \geq \beta^*$:

$$\begin{aligned} L_m(x, y, \alpha^{**}, \beta^{**}) &= f(x, y) + \sum_{i=1}^{m} \alpha_i^{**} |h_i(x, y)| + \sum_{j=1}^{r} \beta_j^{**} g_j^+(x, y) \\ &\geq f(x, y) + \frac{f(x^*, y^*) - f_{min}|_y}{V(y)} \left(\|h(x, y)\|^2 + \|g^+(x, y)\|^2 \right) \\ &\geq f(x, y) + f(x^*, y^*) - f_{min}|_y \qquad \text{(according to (51))} \\ &\geq f(x^*, y^*) \qquad \text{(since } f_{min}|_y \leq f(x, y)\text{)}. \end{aligned} \tag{54}$$

Case 2) Suppose we have $V(y) = 0$, then there exists feasible solutions when y is fixed. If we fix the discrete part of z as y and regard x as the variables, then P_m becomes a continuous CNLP. Let $x^*|_y$ be the CGM_c to this CNLP. Namely,

$$x^*|_y = \text{argmin}_{x \in X} f(x, y) \qquad \text{subject to:} \qquad h(x, y) = 0, g(x, y) \leq 0. \tag{55}$$

Since $x^*|_y$ is the CGM_c to the CNLP, according to Theorem 3.1, there exist finite α^* and β^* such that, for any $\alpha^{**} \geq \alpha^*$ and $\beta^{**} \geq \beta^*$:

$$f(x^*|_y, y) \leq L_m(x, y, \alpha^{**}, \beta^{**}) \tag{56}$$

One the other hand, since $(x^*|_y, y)$ is a feasible solution to P_m and (x^*, y^*) is the CGM_c to P_m,

$$f(x^*|_y, y) \geq f(x^*, y^*). \tag{57}$$

Combining (56) and (57), we have, for any $\alpha^{**} \geq \alpha^*$ and $\beta^{**} \geq \beta^*$:

$$f(x^*, y^*) \leq f(x^*|_y, y) \leq L_m(x, y, \alpha^{**}, \beta^{**}) \tag{58}$$

The theorem is proved after combining the two cases. ■

The *extended dual problem* for P_m is the same as (16) to (18) defined for P_c, except that the variable space is Z instead of X. Based on Theorem 3.5, we have the following result for mixed-space extended duality.

Theorem 3.6 ***(Extended Duality Theorem for Mixed Nonlinear Programming)*** *Suppose $z^* \in X \times Y$ is a CGM_m to P_m, then there is no duality gap for the extended dual problem, i.e. $q^* = f(z^*)$.*

3.4 Illustrative Examples

We discuss two examples to illustrate the difference between original duality theory and the proposed extended duality.

Example 3.1 We illustrate a discrete problem where there is a duality gap for the original duality theory but not for the proposed extended duality theory. Consider the following DNLP ([5], *p497*):

$$\begin{aligned} \min \quad & f(y) = -y \\ \text{subject to:} \quad & g(y) = y - 1/2 \leq 0, \quad y \in Y = \{0, 1\}, \end{aligned}$$

whose optimal value is $f^* = 0$ at $y^* = 0$. For the original duality, we have:

$$q(\mu) = \min_{y \in \{0,1\}} \{-y + \mu(y - 1/2)\} = \min\{-\mu/2, \mu/2 - 1\}$$

The maximum of $q(\mu)$ is $-1/2$ at $\mu = 1$. The duality gap is $f^* - q^* = 1/2$.

For the extended duality theory, we have:

$$q_e(\beta) = \min_{y \in \{0,1\}} \{-y + \beta(y - 1/2)^+\} = \min\{0, -1 + \beta/2\},$$

and the maximum $q_e^* = 0$ is achieved for any $\beta^{**} \geq \beta^* = 2$. There is no gap for extended duality since $f^* = q_e^* = 0$. ■

Example 3.2 We illustrate a continuous problem where there is a duality gap for the original duality theory but not for the proposed extended duality theory. Consider the following CNLP:

$$\begin{aligned} \min_{x \in \mathbb{R}^2, x \geq 0} \quad & f(x) = x_1 + x_2 \\ \text{subject to:} \quad & h(x) = x_1 x_2 - 1 = 0. \end{aligned}$$

It is obvious that $f^* = 2$ at $(x_1^*, x_2^*) = (1, 1)$.

For the original duality, the dual function is:

$$q(\mu) = \min_{x \geq 0} L(x_1, x_2, \mu) = \min_{x \geq 0} \Big(x_1 + x_2 + \mu(x_1 x_2 - 1) \Big).$$

Consider three cases.

- If $\mu = 0$, then $q(\mu) = \min_{x \geq 0}(x_1 + x_2) = 0$.
- If $\mu > 0$, then $q(\mu) \leq L(0, 0, \mu) = -\mu \leq 0$.

- If $\mu < 0$, then $q(\mu) = -\infty$ since $L(x_1, x_2, \mu)$ is minimized at $(x_1, x_2) = (\infty, \infty)$.

Combining the three cases, we can see that, for any $\mu \in \mathbb{R}$, $q(\mu) \leq 0$. Therefore, we have $q^* = \max_{\mu \in \mathbb{R}} q(\mu) \leq 0$. As a result, there is a nonzero duality gap since $f^* - q^* \geq 2 - 0 = 2$.

In contrast, using the extended duality theory, the extended dual function is:

$$q_e(\alpha) = \min_{x \geq 0} L_e(x_1, x_2, \alpha) = \min_{x \geq 0} \Big(x_1 + x_2 + \alpha |x_1 x_2 - 1| \Big)$$

When $x_1 x_2 - 1 \geq 0$, let $y = 1/x_1$, we have that $x_2 \geq y$. Therefore, for any $\alpha \geq 0$,

$$L_e(x_1, x_2, \alpha) = x_1 + x_2 + \alpha(x_1 x_2 - 1) \geq x_1 + y + \alpha(x_1 y - 1) = x_1 + 1/x_1 \geq 2.$$

Since $L_e(1, 1, \alpha) = 2$, we have that, for any $\alpha \geq 0$,:

$$\min_{\{x \geq 0 | x_1 x_2 - 1 \geq 0\}} L_e(x_1, x_2, \alpha) = L_e(1, 1, \alpha) = 2 \tag{59}$$

When $x_1 x_2 - 1 < 0, x_1, x_2 \geq 0$, we have:

$$L_e(x_1, x_2, \alpha) = x_1 + x_2 + \alpha(1 - x_1 x_2)$$

We can see that for $\alpha^* = 2$,

$$q_e(\alpha^*) = \min_{\{x \geq 0 | x_1 x_2 - 1 < 0\}} L_e(x_1, x_2, 2) = \min_{\{x \geq 0 | x_1 x_2 - 1 < 0\}} 2 + x_1 + x_2 - 2 x_1 x_2 \geq 2$$

When $(x_1 x_2 - 1) < 0, x_1, x_2 \geq 0$, we also have $L_e(x_1, x_2, \alpha^{**}) > L_e(x_1, x_2, \alpha^*)$ when $\alpha^{**} > \alpha^*$. Therefore, for any $\alpha^{**} \geq \alpha^* = 2$:

$$\min_{\{x \geq 0 | x_1 x_2 - 1 < 0\}} L_e(x_1, x_2, \alpha^{**}) \geq \min_{\{x \geq 0 | x_1 x_2 - 1 < 0\}} L_e(x_1, x_2, \alpha^*) \geq 2 \tag{60}$$

Combining (59) and (60) together, it follows that, for $\alpha^{**} \geq \alpha^* = 2$,

$$q_e(\alpha) = \min_{x \geq 0} L_e(x_1, x_2, \alpha) = 2. \tag{61}$$

Therefore, we have $q_e^* = f^* = 2$ and there is no duality gap for the extended duality approach.

■

3.5 Penalty-reduction effect of ℓ_1^m-penalty

In summary, we have presented in this section a new dual function and dual problem which have zero duality gap for continuous, discrete, and mixed NLPs without assuming convexity. The similarity

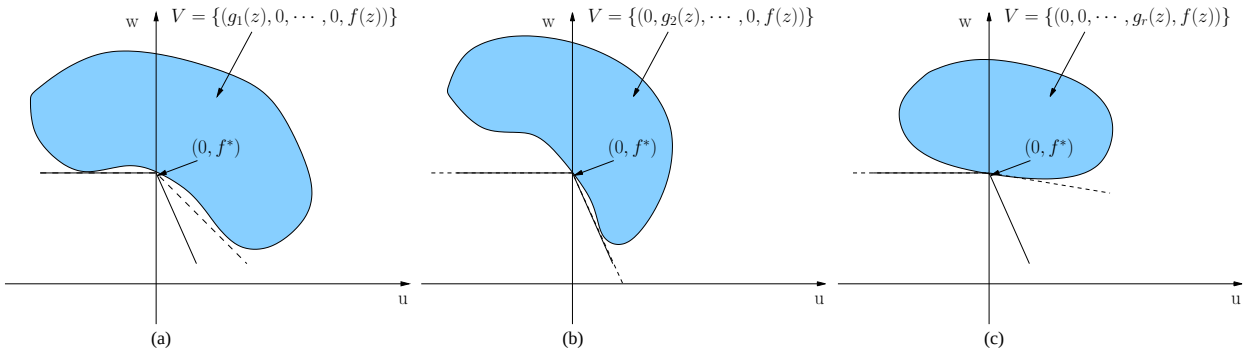

Figure 2: Illustration of the penalty-reduction effect of ℓ_1^m-penalty. The supporting hyperplanes formed by ℓ_1-penalty and ℓ_1^m-penalty are shown in solid and dotted lines, respectively.

of the conditions for the three types of search spaces allows problems in these three classes to be solved in a unified fashion.

The ℓ_1^m-penalty function is different from augmented Lagrangian function and the ℓ_1-penalty function discussed in Chapter 2. Most previous results are proved under continuity and differentiability assumptions. For example, the finiteness of ℓ_1-penalty functions [5, 10] has been proved by relating the penalty value c to the Lagrange multipliers λ^*, whose existence requires the continuity and differentiability of functions. Our development, in contrast, does not rely on the existence of Lagrange multipliers and is general for discrete and mixed-integer problems.

Another salient feature of our theory is that the penalty function uses multiple penalty multipliers, one for each constraints. This feature is the main difference between the proposed ℓ_1^m-penalty and the conventional ℓ_1-penalty function defined in (13) with $\gamma = 1$. An important advantage of ℓ_1^m-penalty is that, to achieve zero duality, it requires much smaller penalty multipliers than the single c required by the the ℓ_1-penalty function.

A geometric explanation of this improvement is illustrated in Figure 2. Consider an inequality-constrained MINLP in (1) with $g(z)$ only, define the set V in $\mathbb{R}^r \times \mathbb{R}$ as:

$$V = \left\{ (g_1(z), \cdots, g_r(z), f(z)) \middle| z \in X \times Y \right\} \tag{62}$$

Figure 2 plots the region of V around the feasible axis $\{(0, 0, \cdots, 0, f(z)) | z \in X \times Y, g_i(z) \leq 0, i = 1, \cdots, r\}$, sliced along different dimensions corresponding to different constraints. We also show the supporting hyperplanes formed by the ℓ_1-penalty and ℓ_1^m-penalty in solid and dotted lines, respectively.

A steeper slope of the hyperplane in the region $u \geq 0$ corresponds to a larger penalty value. For the ℓ_1-penalty, since a single c is used, the slopes are **uniform** for all the dimensions of u.

Therefore, we need to take the **maximum** of the required slopes that support V from below, for all the dimensions of u. For this reason, the ℓ_1-penalty requires a relatively large c. In other words, the maximum c is only necessary for one dimension, and is unnecessary for all the other dimensions. For problems with a large number of constraints, such waste can be huge and unnecessarily increase the difficulty of optimization. The multiple penalty multipliers in the ℓ_1^m-penalty, in contrast, allow **non-uniform** slopes of the supporting hyperplane for different dimensions of u. Therefore, as shown by the dotted line, the hyperplane of the ℓ_1^m-penalty function *closely* supports V from below at each dimension of u, which leads to penalty multipliers smaller than c.

Excessively large penalty multipliers lead to large function values and rugged search terrain, which often make the problem ill-conditioned and difficult to solve. This is the main reason for increasing the penalty gradually instead of setting a large penalty at the beginning in most implementations of penalty methods. Suppose c^* is required for the ℓ_1-penalty function to achieve zero duality gap, and α^* and β^* are required for the ℓ_1^m-penalty function. Since $\alpha_i^* \leq c^*, i = 1, \cdots, m$ and $\beta_j^* \leq c^*, j = 1, \cdots, r$, we can see that:

$$\begin{aligned} L_m(x, \alpha^*, \beta^*) &= f(x) + \sum_{i=1}^{m} \alpha_i^* |h_i(x)| + \sum_{j=1}^{r} \beta_h^* g_j^+(x) \\ &\leq f(x) + \sum_{i=1}^{m} c^* |h_i(x)| + \sum_{j=1}^{r} c^* g_j^+(x) \\ &= f(x) + c^* \left(\sum_{i=1}^{m} |h_i(x)| + \sum_{j=1}^{r} g_j^+(x) \right) = l_1(x, c^*), \ \forall x \in X, \end{aligned} \tag{63}$$

where $l_1(x, c)$ is the ℓ_1-penalty function defined in (13). Therefore, using ℓ_1^m-penalty always leads to a smaller function value everywhere in the search space. Extensive empirical experiences have shown that the difficulty in minimizing the penalty function increases as the penalty multipliers increase. Reduction in penalty value makes the optimization easier in practice.

Example 3.3 We illustrate the penalty-reduction effect using an example. Consider the following problem:

$$\begin{aligned} \text{minimize} \quad & f(x) = -x_1 - 10x_2 \\ \text{subject to} \quad & g_1(x) = x_1 - 5 \leq 0 \ , \\ \text{and} \quad & g_2(x) = x_2 + 1 \leq 0 \ . \end{aligned}$$

Obviously the optimal solution is $x^* = (5, -1)$ with $f(x^*) = 5$.

Consider the ℓ_1-penalty function

$$l_1(x,c) = -x_1 - 10x_2 + c((x_1-5)^+ + (x_2+1)^+)$$

To find c^* such that$l_1(x,c^*)$ is minimized at x^*, we need to have, $\forall x_1 > 5, x_2 > -1$,

$$\begin{aligned} l_1(x,c^*) &= -x_1 - 10x_2 + c^*(x_1 - 5 + x_2 + 1) \\ &= (c^*-1)(x_1-5) + (c^*-10)(x_2+1) + 5 \geq f(x^*) = 5, \end{aligned}$$

which leads to $c^* \geq 10$.

Now we consider the ℓ_1^m-penalty function

$$L_m(x,\beta) = -x_1 - 10x_2 + \beta_1(x_1-5)^+ + \beta_2(x_2+1)^+$$

To find β^* such that $L_m(x,\beta^*)$ is minimized at x^*, we need to have, $\forall x_1 > 5, x_2 > -1$,

$$\begin{aligned} L_m(x,\beta^*) &= -x_1 - 10x_2 + \beta_1(x_1 - 5) + \beta_2(x_2+1) \\ &= (\beta_1^*-1)(x_1-5) + (\beta_2^*-10)(x_2+1) + 5 \geq f(x^*) = 5, \end{aligned}$$

which leads to $\beta_1^* \geq 1, \beta_2^* \geq 10$. It can be easily verified that $\beta_1^* \geq 1, \beta_2^* \geq 10$ is sufficient to make $L_m(x,\beta^*)$ greater than $f(x^*)$ for other regions of x.

In conclusion, we have:

$$\begin{aligned} L_m(x,\beta^*) &= -x_1 - 10x_2 + (x_1-5)^+ + 10(x_2+1)^+ \\ &\leq -x_1 - 10x_2 + 10((x_1-5)^+ + (x_2+1)^+) = l_1(x,c^*) \end{aligned}$$

Thus, for this problem, the ℓ_1^m-penalty function requires less severe penalty than the ℓ_1-penalty function. ■

References

[1] J. P. Aubin and I. Ekeland. Estimates of the duality gap in nonconvex optimization. *Math. Operations Research*, 1:225–245, 1976.

[2] A. M. Bagirov and J. Ugon. Piecewise partially separable functions and a derivative-free algorithm for large scale nonsmooth optimization. *Journal of Global Optimization*, 35(2):163–195, 2006.

[3] A. Ben-Tal, G. Eiger, and V. Gershovitz. Global optimization by reducing the duality gap. *Mathematical Programming*, 63:193–212, 1994.

[4] D. P. Bertsekas. Distributed dynamic programming. *Trans. on Automatic Control*, AC-27(3):610–616, June 1982.

[5] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, Massachusetts, 1999.

[6] R. S. Burachik and A. Rubinov. On the absence of duality gap for Lagrange-type functions. *Journal of Industrial and Management Optimization*, 1(1):33–38, 2005.

[7] J. V. Burke. Calmness and exact penalization. *SIAM J. Control and Optimization*, 29:493–497, 1991.

[8] J. V. Burke. An exact penalization viewpoint of constrained optimization. *SIAM J. Control and Optimization*, 29:968–998, 1991.

[9] G. B. Dantzig and P. Wolfe. Decomposition principle for linear programming. *Operations Research*, 8:101–111, 1960.

[10] N. I. M. Gould, D. Orban, and P. L. Toint. An interior-point ℓ_1-penalty method for nonlinear optimization. *Technical Report RAL-TR-2003-022, Rutherford Appleton Laboratory Chilton, Oxfordshire, UK*, November 2003.

[11] F. Granot and J. Skorin-Kapov. Some proximity and sensitivity results in quadratic integer programming. *Mathematical Programming*, 47(1-3):259–268, 1990.

[12] F. Güder and J. G. Morris. Optimal objective function approximation for separable convex quadratic programming. *Mathematical Programming*, 67(1-3):133–142, 1994.

[13] D. S. Hochbaum and S.P. Hong. About strongly polynomial time algorithms for quadratic optimization over submodular constraints. *Mathematical Programming*, 69(1-3):269–309, 1995.

[14] X. X. Huang and X. Q. Yang. A unified augmented Lagrangian approach to duality and exact penalization. *Mathematics of Operations Research*, 28(3):533–552, 2003.

[15] Z. Q. Luo and J. S. Pang. Error bounds in mathematical programming. *Math. Programming Ser. B*, 88(2), 2000.

[16] A. Nedić and A. Ozdaglar. A geometric framework for nonconvex optimization duality using augmented lagrangian functions. *Journal of Global Optimization*, accpeted, 2006.

[17] J. S. Pang. Error bounds in mathematical programming. *Math. Programming*, 79:299–332, 1997.

[18] R. T. Rockafellar. Augmented Lagrangian multiplier functions and duality in nonconvex programming. *SIAM J. Control and Optimization*, 12:268–285, 1974.

[19] R. T. Rockafellar and R. J.-B Wets. *Variational Analysis*. Springer-Verlag, Berlin, 1998.

[20] A. M. Rubinov, B. M. Glover, and X. Q. Yang. Decreasing functions with applications to penalization. *SIAM J. Optim.*, 10:289–313, 1999.

[21] A. M. Rubinov, B. M. Glover, and X. Q. Yang. Modified Lagrangian and penalty functions in continuous optimization. *Optim.*, 46:327–351, 1999.

[22] J. E. Spingarn. Applications of the method of partial inverses to convex programming: Decomposition. *Mathematical Programming*, 32(2):199–223, 1985.

[23] P. Tseng. Dual coordinate ascent methods for non-strictly convex minimization. *Mathematical Programming*, 59(1-3):231–247, 1993.

[24] H. Tuy. On solving nonconvex optimization problems by reducing the duality gap. *Journal of Global Optimization*, 32:349–365, 2005.

[25] X. Q. Yang and X. X. Huang. A nonlinear Lagrangian approach to constraint optimization problems. *SIAM J. Optim.*, 11:1119–1144, 2001.