Comput Optim Appl
DOI 10.1007/s10589-008-9208-3

Extended duality for nonlinear programming

Yixin Chen · Minmin Chen

Received: 29 May 2007 / Revised: 7 May 2008

Abstract Duality is an important notion for nonlinear programming (NLP). It provides a theoretical foundation for many optimization algorithms. Duality can be used to directly solve NLPs as well as to derive lower bounds of the solution quality which have wide use in other high-level search techniques such as branch and bound. However, the conventional duality theory has the fundamental limit that it leads to duality gaps for nonconvex problems, including discrete and mixed-integer problems where the feasible sets are generally nonconvex.

In this paper, we propose an extended duality theory for nonlinear optimization in order to overcome 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 spaces under mild conditions. Comparing to recent developments in nonlinear Lagrangian functions and exact penalty functions, the proposed theory always requires lesser penalty to achieve zero duality. This is very desirable as the lower function value leads to smoother search terrains and alleviates the ill conditioning of dual optimization.

Based on the extended duality theory, we develop a general search framework for global optimization. Experimental results on engineering benchmarks and a sensor-network optimization application show that our algorithm achieves better performance than searches based on conventional duality and Lagrangian theory.

Keywords Nonlinear programming · Global optimization · Duality gap · Extended duality

This research is supported by Department of Energy Early Career Principal Investigator grant ER25737 and Microsoft Research New Faculty Fellowship.

Y. Chen (✉) · M. Chen
Department of Computer Science and Engineering, Washington University, St. Louis, MO 63130, USA
e-mail: chen@cse.wustl.edu

M. Chen
e-mail: mc15@cec.wustl.edu

1 Introduction

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

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

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 and 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 (DNLP) when there are only discrete variables y.

Duality is an important notion for mathematical programming. An important issue of duality is the existence of the *duality gap*, which is 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 generally nonzero for nonconvex problems, and may be large for some problems, in which case the dual approach is not useful. Moreover, for discrete and mixed-integer problems, the duality gap may be nonzero even if the functions are convex. Nonzero duality gaps make the direct dual methods fail and the global optimization algorithms such as branch and bound less effective.

There are a number of previous efforts that aim at removing the duality gap for nonconvex optimization. Examples include the generalized augmented Lagrangian functions [3, 17], nonlinear Lagrangian functions [19, 26], and ℓ_1-penalty functions [16]. Extensive study has been performed to establish the duality results for new dual problems where those new functions are used in place of the traditional Lagrangian function. However, there are some limitations of the previous work that motivate this research.

First, we found through theoretical analysis and experimentation that some previous dual functions require large penalty values to ensure zero duality gap for nonconvex problems. The large penalty values result in ill-conditioned unconstrained optimization problems and increase the difficulty of search. Second, some previous zero-gap results are developed for continuous and differentiable problems. Often the parameters for achieving zero-duality is related the Lagrange multipliers, which do not exist for discrete problems. Such results cannot be applied to discrete or mixed-integer problems, or problems with non-differentiable constraints. Most other results that can handle discrete problems are based on geometric approaches. There is a lack of analytical approaches to discrete and mixed-integer duality. Third, most previous work only give theoretical results without experimental investigation. They provide interesting theoretical results that shed insights into the analytical properties and geometrical structures of various dual functions. However, there is little empirical study on the efficiency and effectiveness of implementations of the new theory.

In this paper, we propose another duality theory to characterize the constrained global optimal solutions of nonlinear programming problems (NLPs) in discrete, continuous, and mixed spaces. Based on a new ℓ_1^m-penalty formulation, a generalization of the ℓ_1-penalty function, our theory transforms an NLP to an equivalent extended dual problem that has no duality gap for nonconvex NLPs under weak assumptions. The new dual function requires smaller penalty values than the traditional ℓ_1-penalty function in order to achieve a zero duality gap. The smaller penalty can address the ill conditioning of unconstrained optimization, making the search much easier.

We have developed a dual search algorithm that implements the proposed theory. We have evaluated the performance of the algorithm on standard NLP benchmarks and a sensor-network optimization application. Experimental results show that our solver can optimally solve many nonconvex NLPs that are difficult to solve by other existing solvers. The empirical results also demonstrate the benefit of the penalty reduction effect of the proposed theory.

2 Related previous work

In this section, we review some related previous work and discuss their limitations and differences to the proposed work.

2.1 Duality

Duality is an important notion for mathematical programming. We consider the following continuous nonlinear programming (CNLP) problem.

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

where X is a compact subset of $\mathbb{R}^n$, f, g, h are lower bounded and continuous, but not necessarily differentiable.

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:

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

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

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

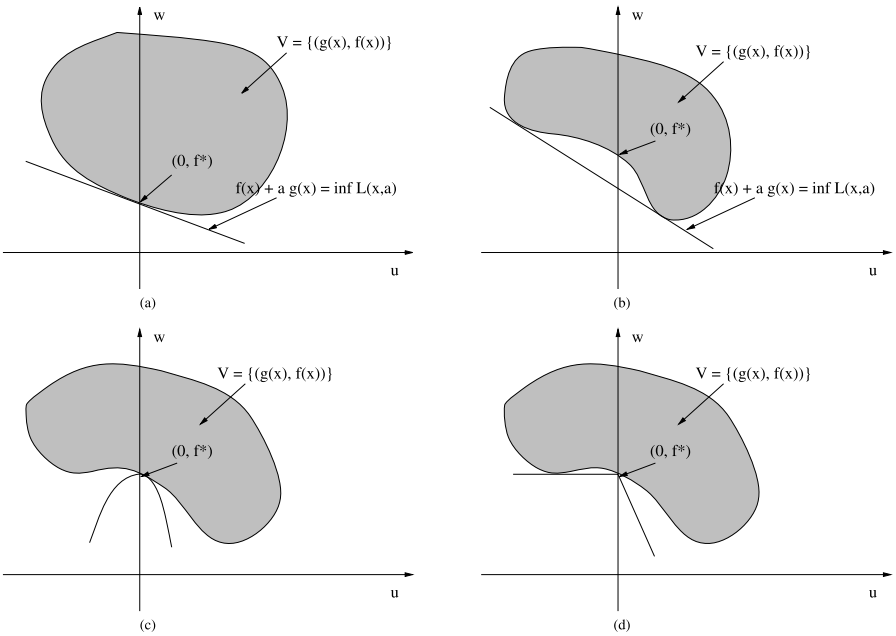

Fig. 1 Geometric interpretation of duality. (**a**) The zero duality gap achieved by the Lagrangian function for convex problems. (**b**) The nonzero duality gap of the Lagrangian function for nonconvex problems. (**c**) Using an augmented Lagrangian function can remove the duality gap. (**d**) Using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

The main results of the duality 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. $q^* \leq f^*$. Second, for CNLPs with a convex objective function and a convex feasible set, there is no duality gap under very general conditions.

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 have indicated that the duality gap can be reduced when a problem is decomposed or has certain special structures [1, 2, 21].

It is well known that the existence of the duality gap is closely related to the geometric problem of finding the hyperplane supporting the set V of constraint-objective pairs (cf. Sect. 5, [4]). 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 (Fig. 1a). For nonconvex problems, there is a gap between the two intercepts (Fig. 1b).

To remove the duality gaps for nonconvex problems, augmented Lagrangian functions [3, 17] 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 (Fig. 1c). 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 [6, 7]. The geometric interpretation of exact penalty functions is visualized in Fig. 1d.

Rubinov et al. [19, 26] 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:

$$l_\gamma(x,c) = \left[f^\gamma(x) + c\left(\sum_{i=1}^{m} |h_i(x)|^\gamma + \sum_{j=1}^{r} g_j^+(x)^\gamma \right) \right]^{1/\gamma}, \tag{6}$$

where $\gamma > 0$ is a parameter.

Luo et al. [14] have proposed a nonconvex and nonsmooth penalty function with zero duality gap based on the following formulation, where $\gamma > 0$:

$$l_\gamma(x,c) = f(x) + c\left(\sum_{i=1}^{m} |h_i(x)| + \sum_{j=1}^{r} g_j^+(x) \right)^\gamma. \tag{7}$$

An exact penalty function with zero duality gap under certain assumptions is proposed by Pang [16] as follows:

$$l_\gamma(x,c) = f(x) + c\left[\max\left\{ |h_1(x)|, \ldots, |h_m(x)|, g_1^+(x), \ldots, g_r^+(x) \right\} \right]^\gamma. \tag{8}$$

Recently, there are a number of efforts to provide unified frameworks that accommodate both augmented Lagrangian functions and exact penalty functions with zero duality gap for nonconvex continuous optimization [5–7, 12, 14, 15, 18–20, 26].

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 [12, 15]:

$$l(x,\lambda,\mu,c) = f(x) + \tau(\lambda,\mu,h,g) + c\sigma(h,g) \tag{9}$$

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.

Rockafellar and Wets [18] 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 [5]. A recent work by Nedić and Ozdaglar [15] 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 [12] 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.

We see that most previous methods for removing the duality gaps use a single penalty multiplier c before the augmenting term. However, a suitable c is often hard to locate and control. In practice, a common problem is that the single c is often too large, which makes the unconstrained optimization difficult. In this paper, we propose to use multiple penalty multipliers which can effectively reduce the penalty values needed for ensuring a zero duality gap.

In theory, all the new augmented Lagrangian function and exact penalty functions proposed above can be used to derive new dual methods to solve NLPs that satisfy their corresponding assumptions. However, there is limited study on the experimental evaluation. In this paper, we experimentally evaluate the performance of the new dual algorithms and apply it to a sensor network management application.

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 the CNLP 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:

$$L_m(x, \alpha, \beta) = f(x) + \alpha^T |h(x)| + \beta^T g^+(x), \tag{10}$$

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:

$$q(\alpha, \beta) = \min_{x \in X} L_m(x, \alpha, \beta). \tag{11}$$

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:

$$\begin{aligned} &\text{maximize } q(\alpha, \beta) \\ &\text{subject to} \quad \alpha \geq 0 \quad \text{and} \quad \beta \geq 0, \end{aligned} \tag{12}$$

and the *optimal extended dual value* as:

$$q^* = \max_{\alpha \geq 0, \beta \geq 0} q(\alpha, \beta). \tag{13}$$

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:

$$f'(x; p) = \lim_{\epsilon \to 0} \frac{f(x + \epsilon p) - f(x)}{\epsilon}. \tag{14}$$

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 the objective is non-zero but the directional derivatives of continuous equality and continuous active inequality constraints are all zero. That is, $\nexists\, p \in \mathbb{R}^n$ such that

$$\begin{aligned} &f'(x; p) \neq 0, \quad h_i'(x; p) = 0 \quad \text{and} \quad g_j'(x; p) = 0, \\ &\forall i \in C_h \text{ and } j \in C_g, \end{aligned} \tag{15}$$

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 (10) 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 (10) at x with respect to $x + p$. To ensure a local minimum of (10) at x, the above scenario must not be true for any p at x.

We compare our constraint-qualification condition to the well-known regularity condition in KKT condition that requires the linear independence of gradients of active constraint functions. Of course, the regularity condition only applies to differentiable problems, while our constraint-qualification condition does not require differentiability.

Proposition 3.1 *If $f(x)$, $h(x)$ and $g(x)$ in P_c are differentiable, if a CGM$_c$ $x^* \in X$ of P_c is regular, then it satisfies the constraint qualification.*

Proof If x^* is a regular point, then we have that all the vectors in the set $\{\nabla h_i(x^*), i \in C_h\} \cup \{\nabla g_j(x^*), j \in C_g\}$ are linearly independent. Since x^* is a CGM_c and it is regular, it satisfies the KKT condition. There exist $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^r$ such that

$$\nabla f(x^*) + \lambda^* \nabla h(x^*) + \mu^* \nabla g(x^*) = 0, \quad \text{and} \quad \mu_j^* = 0, \quad \forall j \notin C_g. \tag{16}$$

Suppose there is a direction $p \in \mathbb{R}^n$ such that $h_i'(x^*; p) = 0$ and $g_j'(x^*; p) = 0$, $\forall i \in C_h$ and $j \in C_g$. When $h(x^*)$ and $g(x^*)$ in P_c are differentiable, we have for any direction $p \in \mathbb{R}^n$, $h_i'(x^*; p) = \nabla h_i(x^*)^T p$. From (16), we get:

$$\begin{aligned} 0 &= \nabla f(x^*)^T p + \lambda^* \nabla h(x^*)^T p + \sum_{j \in C_g} \mu_j^* \nabla g_j(x^*)^T p + \sum_{j \notin C_g} \mu_j^* \nabla g_j(x^*)^T p \\ &= f'(x^*; p) + \lambda^* h_i'(x^*; p) + \sum_{j \in C_g} \mu_j^* g_j'(x^*; p) = f'(x^*; p). \end{aligned} \tag{17}$$

Thus, there exists no direction satisfying (15) and x^* must meet the constraint qualification. □

Proposition 3.1 shows that our constraint qualification is always less restrictive than the regularity condition. This is true because, if x^* is regular, then it must also satisfy the constraint qualification. One the other hand, a point x^* satisfying the constraint qualification may not be a regular point. For example, when $\nabla f(x^*) = (1, 1, 0)^T$, $\nabla h_1(x^*) = (0, 1, 0)^T$, $\nabla g_1(x^*) = (0, -1, 0)^T$, and $\nabla g_2(x^*) = (-1, 0, 0)^T$, x^* is not regular since $\nabla h_1(x^*)$, $\nabla g_1(x^*)$, and $\nabla g_2(x^*)$ are not linearly independent, but x^* satisfies the constraint qualification. Further, unlike the regularity condition, our definition is applicable to non-differentiable problems as well.

Definition 3.5 (Feasible set and ϵ-extension) Let the set of all feasible points of P_c be:

$$\mathcal{F} = \{x \mid x \in X, h(x) = 0, g(x) \leq 0\}, \tag{18}$$

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

$$\mathcal{F}_\epsilon^+ = \left\{x \mid x \in X, \left(\min_{y \in \mathcal{F}} \|y - x\|\right) \leq \epsilon\right\}. \tag{19}$$

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*:

$$\|h(x)\|^2 + \|g^+(x)\|^2 \geq \xi, \quad \textit{for any } x \in X - \mathcal{F}_\epsilon^+. \tag{20}$$

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

$$\|h(x_i)\|^2 + \|g^+(x_i)\|^2 \leq \xi_i. \tag{21}$$

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:

$$\|h(x)\|^2 + \|g^+(x)\|^2 = \lim_{i\to\infty} \|h(x_i)\|^2 + \|g^+(x_i)\|^2 \le \lim_{i\to\infty} \xi_i = 0, \tag{22}$$

which implies that $\|h(x)\|^2 + \|g^+(x)\|^2 = 0$. Thus, we must have $h(x) = 0$ and $g(x) \le 0$, 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 exist finite $\alpha^* \ge 0$ and $\beta^* \ge 0$ such that*[1]

$$f(x^*) = \min_{x\in X} L_m(x, \alpha^{**}, \beta^{**}), \quad \textit{for any } \alpha^{**} \ge \alpha^*, \beta^{**} \ge \beta^*. \tag{23}$$

Proof Since we have:

$$L_m(x^*, \alpha^{**}, \beta^{**}) = f(x^*) + \sum_{i=1}^{m} \alpha_i^{**} |h_i(x)| + \sum_{j=1}^{r} \beta_j^{**} g_j^+(x) = f(x^*),$$
$$\forall \alpha^{**} \ge 0, \beta^{**} \ge 0, \tag{24}$$

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

$$f(x^*) \le L_m(x, \alpha^{**}, \beta^{**}), \quad \text{for any } x \in X, \alpha^{**} > \alpha^*, \beta^{**} > \beta^*. \tag{25}$$

We prove (25) in three parts. First, we prove that (25) is true for any point x in the feasible set $\mathcal{F}$. Then, we show that (25) is true for any point x within $\mathcal{F}_{\epsilon_{min}}^+$ for a small $\epsilon_{min} > 0$. Last, we prove (25) for the points in $X - \mathcal{F}_{\epsilon_{min}}^+$. For simplicity, we assume that x^* is the only CGM_c in X. The case of more than one CGM_c can be proved similarly.

Part (a) For every feasible point $x' \in \mathcal{F}$, (25) is true for any $\alpha^{**} \ge 0$ and $\beta^{**} \ge 0$ since

$$L_m(x', \alpha^{**}, \beta^{**}) = f(x') \ge f(x^*), \tag{26}$$

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

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

Part (b) We show that (25) 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 (25).

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 $x = x' + \epsilon p$, $p \in \mathbb{R}^n$ is a unit-length direction vector with $\|p\| = 1$ and $\epsilon = \|x - x'\|$. When ϵ is small enough, we have:

$$\begin{aligned} L_m(x, \alpha^{**}, \beta^{**}) &= f(x) + \sum_{i=1}^{m} \alpha_i^{**} |h_i(x)| + \sum_{j=1}^{r} \beta_j^{**} g_j^+(x) \\ &\geq f(x) = f(x') + \epsilon f'(x'; p) + o(\epsilon^2) \\ &\geq f(x^*) + \xi + \epsilon f'(x'; p) + o(\epsilon^2) \geq f(x^*). \end{aligned} \tag{27}$$

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 (25) is true. We consider the following two cases:

(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,

$$L_m(x, \alpha^{**}, \beta^{**}) = f(x) + \sum_{i=1}^{m} \alpha_i^{**} |h_i(x)| + \sum_{j=1}^{r} \beta_j^{**} g_j^+(x) = f(x) \geq f(x^*). \tag{28}$$

(2) In this case, other than inactive inequality constraints, if there are equality or active inequality constraints at x^*. According to the constraint-qualification condition, unless $f'(x^*; p) = 0$, in which case x^* minimizes $L_m(x^*, \alpha^{**}, \beta^{**})$ in $\mathcal{F}_\epsilon^+$ for small enough ϵ, 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 $\alpha_k^{**} > \frac{|f'(x^*; p)|}{|h_k'(x^*; p)|}$ and ϵ small enough, then:

$$\begin{aligned} L_m(x, \alpha^{**}, \beta^{**}) &= f(x) + \sum_{i=1}^{m} \alpha_i^{**} |h_i(x)| + \sum_{j=1}^{r} \beta_j^{**} g_j^+(x) \\ &\geq f(x) + \alpha_k^{**} |h_k(x)| \\ &\geq f(x^*) + \epsilon f'(x^*; p) + o(\epsilon^2) + \alpha_k^{**} \epsilon |h_k'(x^*; p)| \\ &\geq f(x^*) + \epsilon(\alpha_k^{**} |h_k'(x^*; p)| - |f'(x^*; p)|) + o(\epsilon^2) \\ &\geq f(x^*). \end{aligned} \tag{29}$$

Combining the results in parts (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 (25) is true for any point $x \in X$ in $\mathcal{F}_{\epsilon_{min}}^+$, the ϵ_{min}-extension of $\mathcal{F}$.

Part (c) Parts (a) and (b) have proved that (25) is true for any point $x \in \mathcal{F}^+_{\epsilon_{min}}$. We now prove that (25) 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

$$\|h(x)\|^2 + \|g^+(x)\|^2 \geq \xi. \tag{30}$$

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

$$\alpha_i^* = \frac{f(x^*) - f_{min}}{\xi} |h_i(x)|, \quad i = 1, \ldots, m, \tag{31}$$

and

$$\beta_j^* = \frac{f(x^*) - f_{min}}{\xi} g_j^+(x), \quad j = 1, \ldots, r. \tag{32}$$

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

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

$$\begin{aligned} L_m(x, \alpha^{**}, \beta^{**}) &= f(x) + \sum_{i=1}^{m} \alpha_i^{**} |h_i(x)| + \sum_{j=1}^{r} \beta_j^{**} g_j^+(x) \\ &\geq f(x) + \frac{f(x^*) - f_{min}}{\xi} (\|h(x)\|^2 + \|g^+(x)\|^2) \\ &\geq f(x) + f(x^*) - f_{min} \quad \text{(according to (30))} \\ &\geq f(x^*). \end{aligned} \tag{33}$$

Equation (25) 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* (13), *i.e. $q^* = f(x^*)$.*

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

$$\begin{aligned} q^* &= \max_{\alpha \geq 0, \beta \geq 0} q(\alpha, \beta) = \max_{\alpha \geq 0, \beta \geq 0} \left(\min_{x \in X} L_m(x, \alpha, \beta) \right) \\ &\leq \max_{\alpha \geq 0, \beta \geq 0} L_m(x^*, \alpha, \beta) = \max_{\alpha \geq 0, \beta \geq 0} f(x^*) = f(x^*). \end{aligned} \tag{34}$$

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

$$q^* = \max_{\alpha \geq 0, \beta \geq 0} q(\alpha, \beta) \geq q(\alpha^{**}, \beta^{**}) = f(x^*). \tag{35}$$

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

$$\begin{aligned}(P_d)\quad &\min_{y} f(y), \quad \text{where } y=(y_1,\ldots,y_w)^T \in Y \\ &\text{subject to} \quad h(y)=0 \quad \text{and} \quad g(y)\le 0,\end{aligned} \tag{36}$$

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 $f(y^*) \le f(y)$ for all feasible $y \in Y$.

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

$$L_m(y,\alpha,\beta) = f(y) + \alpha^T |h(y)| + \beta^T g^+(y), \tag{37}$$

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 exist finite $\alpha^* \ge 0$ and $\beta^* \ge 0$ such that*

$$f(y^*) = \min_{y\in Y} L_m(y,\alpha^{**},\beta^{**}), \quad \textit{for any } \alpha^{**} \ge \alpha^*, \beta^{**} \ge \beta^*. \tag{38}$$

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

$$f(y^*) \le L_m(y,\alpha^{**},\beta^{**}), \quad \text{for any } \alpha^{**} \ge \alpha^*, \beta^{**} \ge \beta^*, \tag{39}$$

for any $y \in Y$.

We set the following α^* and β^*:

$$\alpha_i^* = \max_{y\in Y, |h_i(y)|>0} \left\{ \frac{f(y^*)-f(y)}{|h_i(y)|} \right\}, \quad i=1,\ldots,m, \tag{40}$$

$$\beta_j^* = \max_{y\in Y, g_j(y)>0} \left\{ \frac{f(y^*)-f(y)}{g_j(y)} \right\}, \quad j=1,\ldots,r. \tag{41}$$

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

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

$$L_m(y,\alpha^{**},\beta^{**}) = f(y) \ge f(y^*). \tag{42}$$

For an infeasible point $y \in Y$, if there is at least one equality constraint $h_i(y)$ that is not satisfied ($|h_i(y)| > 0$), we have:

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

$$\geq f(y) + \frac{f(y^*) - f(y)}{|h_i(y)|} |h_i(y)| = f(y^*). \tag{43}$$

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

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

$$\geq f(y) + \frac{f(y^*) - f(y)}{g_j(y)} g_j(y) = f(y^*). \tag{44}$$

Equation (39) is proved after combining (42), (43), and (44). □

The *extended dual problem* for P_d is the same as (11) to (13) 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 in Theorem 3.1 is not needed for Theorem 3.3.

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:

$$L_m(z, \alpha, \beta) = f(z) + \alpha^T |h(z)| + \beta^T g^+(z), \tag{45}$$

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*

$$f(z^*) = \min_{z \in X \times Y} L_m(z, \alpha^{**}, \beta^{**}), \quad \text{for any } \alpha^{**} \geq \alpha^*, \beta^{**} \geq \beta^*. \tag{46}$$

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 any $z = (x, y) \in X \times Y$, there exist finite $\alpha^* \geq 0$ and $\beta^* \geq 0$ such that

$$f(x^*, y^*) \leq L_m(x, y, \alpha^{**}, \beta^{**}), \quad \text{for any } \alpha^{**} \geq \alpha^*, \beta^{**} \geq \beta^*. \tag{47}$$

Define:

$$V(y) = \min_{x' \in X} \left(\|h(x', y)\|^2 + \|g^+(x', y)\|^2 \right). \tag{48}$$

We consider two cases.

(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:

$$\alpha_i^* = \frac{f(x^*, y^*) - f_{min}|_y}{V(y)} |h_i(x)|, \quad i = 1, \ldots, m, \tag{49}$$

and

$$\beta_j^* = \frac{f(x^*, y^*) - f_{min}|_y}{V(y)} g_j^+(x), \quad j = 1, \ldots, r. \tag{50}$$

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

$$\begin{aligned} L_m(x, y, \alpha^{**}, \beta^{**}) &= f(x, y) + \sum_{i=1}^{m} \alpha_i^{**} |h_i(x, y)| + \sum_{j=1}^{r} \beta_j^{**} g_j^+(x, y) \\ &\geq f(x, y) + \frac{f(x^*, y^*) - f_{min}|_y}{V(y)} (\|h(x, y)\|^2 + \|g^+(x, y)\|^2) \\ &\geq f(x, y) + f(x^*, y^*) - f_{min}|_y \quad \text{(according to (48))} \\ &\geq f(x^*, y^*) \quad \text{(since } f_{min}|_y \leq f(x, y)). \end{aligned} \tag{51}$$

(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,

$$x^*|_y = \operatorname*{argmin}_{x \in X} f(x, y) \quad \text{subject to} \quad h(x, y) = 0, \quad g(x, y) \leq 0. \tag{52}$$

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 $\beta^{**} \geq \beta^*$:

$$f(x^*|_y, y) \leq L_m(x, y, \alpha^{**}, \beta^{**}). \tag{53}$$

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,

$$f(x^*|_y, y) \geq f(x^*, y^*). \tag{54}$$

Combining (53) and (54), we have, for any $\alpha^{**} \geq \alpha^*$ and $\beta^{**} \geq \beta^*$:

$$f(x^*, y^*) \leq f(x^*|_y, y) \leq L_m(x, y, \alpha^{**}, \beta^{**}). \tag{55}$$

The theorem is proved after combining the two cases. □

The *extended dual problem* for P_m is the same as (11) to (13) 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. $q^* = f(z^*)$.*

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 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 Sect. 2. Most previous results on ℓ_1-penalty are proved under continuity and differentiability assumptions. For example, the finiteness of ℓ_1-penalty functions [4, 11] 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.

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 ([4], p. 497):

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

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:

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

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

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:

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

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:

$$q(\alpha) = \min_{x\geq 0} L_m(x_1, x_2, \alpha) = \min_{x\geq 0}(x_1 + x_2 + \alpha|x_1 x_2 - 1|).$$

It is easy to validate that, for $\alpha^{**} \geq \alpha^* = 2$, $q(\alpha) = \min_{x\geq 0} L_m(x_1, x_2, \alpha) = 2$. Therefore, we have $q^* = f^* = 2$ and there is no duality gap for the extended duality approach.

3.5 Penalty-reduction effect of ℓ_1^m-penalty

A 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 (7) with $\gamma = 1$. An important advantage of ℓ_1^m-penalty is that, to achieve zero duality, it requires smaller penalties than the single c required by the ℓ_1-penalty function.

Theorem 3.7 *For P_m where $z^* \in Z$ is the CGM$_m$, let c^* be the minimum value of c such that $\inf_{z\in Z} l_1(z, c^*) = f(z^*)$, and α^*, β^* be the minimum value of α, β such that $\inf_{z\in Z} L_m(z, \alpha^*, \beta^*) = f(z^*)$, we have $L_m(z, \alpha^*, \beta^*) \leq l_1(z, c^*)$, for all $z \in Z$.*

Theorem 3.7 can be seen from the fact that if $\inf_{z\in Z} l_1(z, c^*) = f(z^*)$, then $\inf_{z\in Z} L_m(z, c^*, c^*) = f(z^*)$. Therefore, we have that every component of α^* and β^* is no larger than c^*. Theorem 3.7 can then be shown by using Theorem 3.5.

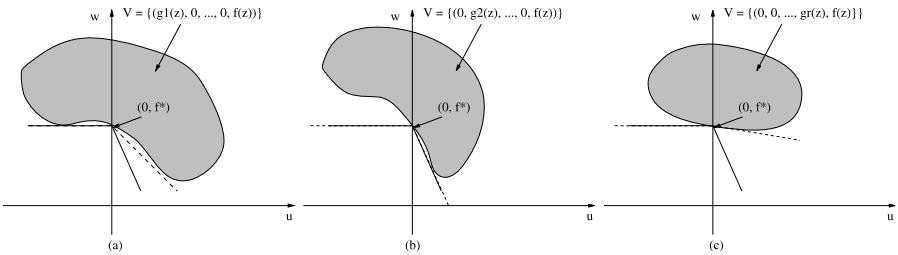

Fig. 2 Illustration of the penalty-reduction effect of ℓ_1^m-penalty. The supporting hyperplanes formed by ℓ_1-penalty and ℓ_1^m-penalty are shown in solid and dotted lines, respectively

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

$$V = \{(g_1(z), \ldots, g_r(z), f(z)) \mid z \in X \times Y\}. \tag{56}$$

Figure 2 plots the region of V around the feasible axis $\{(0, 0, \ldots, 0, f(z)) | z \in X \times Y, g_i(z) \leq 0, i = 1, \ldots, r\}$, sliced along different dimensions corresponding to different constraints. The w-axis corresponds to $f(z)$, while the u-axis corresponds to various 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 (solid lines) 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 (dotted lines) of the supporting hyperplane for different dimensions of u. Therefore, 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.

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 $\alpha_i^* \leq c^*, i = 1, \ldots, m$ and $\beta_j^* \leq c^*, j = 1, \ldots, r$, we have $L_m(x, \alpha^*, \beta^*) \leq l_1(x, c^*), \ \forall x \in X$ since:

$$\begin{aligned} L_m(x, \alpha^*, \beta^*) &= f(x) + \sum_{i=1}^{m} \alpha_i^* |h_i(x)| + \sum_{j=1}^{r} \beta_h^* g_j^+(x) \\ &\leq f(x) + \sum_{i=1}^{m} c^* |h_i(x)| + \sum_{j=1}^{r} c^* g_j^+(x) \\ &= f(x) + c^* \left(\sum_{i=1}^{m} |h_i(x)| + \sum_{j=1}^{r} g_j^+(x) \right) = l_1(x, c^*), \end{aligned} \tag{57}$$

where $l_1(x, c)$ is the ℓ_1-penalty function defined in (7). Therefore, using ℓ_1^m-penalty always leads to a smaller function value everywhere in the search space. Extensive empirical experiences have shown that penalty reduction makes the optimization easier in practice.

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

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

Obviously the optimal solution is $x^* = (5, -1)$ with $f(x^*) = 5$. Consider the ℓ_1-penalty function

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

We can see $l_1(x, c^*)$ is minimized at x^* when $c^* \ge 10$.

Now we consider the ℓ_1^m-penalty function

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

It can be easily verified that $\beta_1^* \ge 1, \beta_2^* \ge 10$ is sufficient to make $L_m(x, \beta^*)$ minimized at x^*.

Thus, for this problem, the ℓ_1^m-penalty function requires less severe penalty than the ℓ_1-penalty function to achieve zero duality gap.

4 Experimental results

In this section, we develop and test a general extended duality search (EDS) framework that directly implements the theory, and show that, by eliminating the duality gap, EDS can successfully solve many nonconvex constrained optimization problems. We also apply EDS to solve a complex NLP from sensor network management.

4.1 Extended duality search

Based on the extended duality theory, we derive a search framework that looks for the constrained global minimum points satisfying the extended duality condition. The search framework provides a general and unified framework for solving CNLPs, DNLPs, and MINLPs.

An important feature of extended duality over the original duality is that, instead of finding unique geometric multipliers λ^* and μ^*, it suffices to find sufficiently large α^{**} and β^{**} such that $\alpha^{**} > \alpha^* \ge 0$ and $\beta^{**} > \beta^* \ge 0$. However, in practice, large penalties leads to ill-conditioned unconstrained functions that are difficult to optimize. For this reason, it is a standard practice to start with small penalties and gradually increase them until a zero duality gap is achieved.

procedure Extended_Duality_Search(P_m, x, $\alpha^{\max}$, $\beta^{\max}$);
1. $\alpha \leftarrow 0$; $\beta \leftarrow 0$;
2. **repeat**
3. **for** $(i = 1, \ldots, m)$ **if** $(h_i(x) \neq 0$ **and** $\alpha_i < \alpha_i^{\max})$ **then** increase α_i;
4. **for** $(j = 1, \ldots, r)$ **if** $(g_j(x) \not\leq 0$ **and** $\beta_j < \beta_j^{\max})$ **then** increase β_j;
5. solve the dual problem and set $z \leftarrow \text{argmin}_{z \in X \times Y} L_m(z, \alpha, \beta)$;
6. **if** (penalty decrease condition satisfied) scale down α and β;
7. **until** $(\alpha_i > \alpha_i^{\max}$ for all $h_i(x) \neq 0$ **and** $\beta_j > \beta_j^{\max}$ for all $g_j(x) \not\leq 0)$
or a CGM_m of P_m is found;
8. **return** CGM_m if found;
9. **end_procedure**

Fig. 3 Iterative procedures to look for CGM_m of P_m. The bounds on α and β, α^{max} and β^{max}, are user-provided

Figure 3 shows the pseudo code of the EDS search algorithm which solves P_m by looking for x^*, y^*, α^{**}, and β^{**} that satisfy Theorem 3.5. According to the extended duality theory, solving a *constrained* optimization problem can be decomposed into a series of *unconstrained* optimization problems, each with fixed penalty vectors α^{**} and β^{**}. This observation allows us to solve P_m by an iterative search in Fig. 3. (The algorithms for solving P_c and P_d are similar and not shown.) Assuming $\alpha^{**} \geq \alpha^*$ and $\beta^{**} \geq \beta^*$ have been found, the inner loop looks for a minimum of $L_m(x, y, \alpha^{**}, \beta^{**})$ in order to find x^* and y^*. If a feasible solution to P_m is not found at the point minimizing $L_m(x, y, \alpha^{**}, \beta^{**})$, the penalties corresponding to the violated constraints are increased. The process is repeated until a CGM_m is found or when α^{**} (*resp.* β^{**}) is larger than the user-provided maximum bound α^{max} (*resp.* β^{max}), where α^{max} (*resp.* β^{max}) is chosen to be so large that it exceeds α^* (*resp.* β^*).

We now describe the implementation details of the EDS algorithm. A key component of EDS is the strategy for updating the ℓ_1^m-penalties. In our experiments, after each unconstrained optimization in Line 5 of Fig. 3, we use the following strategy to update the penalty values α and β the correspond to violated constraints (Lines 3 and 4):

$$\alpha \longleftarrow \alpha + \rho^T |h(z)|, \qquad \beta \longleftarrow \beta + \varrho^T g^+(z), \tag{58}$$

where ρ and ϱ are vectors for controlling the rate of updating α and β. Thus, after initialization, the update of each element of α and β is proportional to the violation of the corresponding constraint and stops when it is satisfied.

We update each element of ρ and ϱ dynamically until the corresponding constraint is satisfied. For each constraint h_i, $i = 1, \ldots, m$, we use c_i to count the number of consecutive iterations in which h_i is violated since the last update of ρ_i. After an iteration, we increase c_i by 1 if h_i is violated; if c_i reaches threshold K, which means that h_i has not been satisfied in K consecutive iterations, we increase ρ_i by:

$$\rho_i \longleftarrow \rho_i \cdot \delta, \quad \text{where } \delta > 1, \tag{59}$$

and reset c_i to 0. If h_i is satisfied, we reset ρ_i to ρ_0 and c_i to 0. In our implementation, we choose $\rho_0 = 0.01$, $K = 2$ and $\delta = 2.25$. We update ϱ in the same manner.

From experimental studies, we have found that monotonically increasing α and β may lead to excessively large penalties values and it is beneficial to periodically scale down the penalty values to ease the unconstrained optimization. In our implementation, we scale down α and β by multiplying each penalty by a random value between 0.7 to 0.95 if we cannot decrease the maximum violation of constraints after solving five consecutive subproblems.

In our experiments, we modify a simulated annealing package SIMMAN [22] to perform unconstrained global optimization of $L_m(x, y, \alpha^{**}, \beta^{**})$. Simulated annealing (SA) is a popular global optimization algorithm that ensures asymptotic convergence to a global minimum when a logarithmic cooling schedule is used. However, the logarithmic schedule is very slow in practice and the SA package uses a geometric cooling schedule which reduces the temperature T by $T = \xi T, 0 < \xi < 1$ in each major iteration. In our experiments, we have used the default setting of the SA package and set $\xi = 0.8$.

4.2 Results on engineering benchmarks

We report experimental results on using the EDS algorithm to solve some nonconvex NLPs. We use two test sets from engineering applications. Problems G1-G10 [13] were originally developed for testing and tuning various constraint handling techniques in evolutionary algorithms. The second set of benchmarks [9] was collected by Floudas and Pardalos and was derived from practical engineering applications.

In order to illustrate the benefit of the penalty reduction effects of the proposed ℓ_1^m-penalty function, we have also implemented an ℓ_1-PM method which is the same as EDS except that the unconstrained minimization minimizes an ℓ_1-penalty function with a single penalty multiplier c. We have used and tuned a similar strategy for updating c.

We show our results for G1–G10 in Table 1 and the results for Floudas and Pardalos's problems in Table 2. In both tables, the first three columns show the problem ID, number of variables (n_v), and number of constraints (n_c), respectively. The fourth and fifth columns show the quality (Q) and solution time (T) of SNOPT [10], a sequential quadratic programming solver based on the Lagrangian theory. The sixth and seventh columns show the results of LANCELOT [8], an augmented Lagrangian method that uses a single penalty value for all constraints.

Since EDS is a stochastic algorithm using a random seed, we make 100 runs of EDS for each problem. Columns 8–11 show four metrics for EDS. These metrics are Q_{best}, the best solution quality found in the 100 runs, Q_{avg}, the average quality of the 100 runs, T_{avg}, the average solution time of the 100 runs, and P_{succ}, the percentage of successful runs where a run is successful if it finds a feasible solution. We list the same metrics for the ℓ_1-PM method in columns 12–15.

From Table 1 and 2, we can see that EDS can find the optimal solution for many nonconvex problems that other methods have difficulty to deal with. For most problems, EDS can achieve the best solution quality. Since SNOPT and LANCELOT are suboptimal solvers based on local optimality conditions, they are generally faster but have worse quality. EDS provides an effective method for optimally solving nonlinear constrained problems.

Table 1 Results on solving NLP benchmarks G1 to G10. All runs are made on a PC workstation with Intel Xeon 2.4 GHZ CPU and 2 GB memory with a time limit of 3600 seconds. All times are in seconds. 100 runs of EDS and ℓ_1-PM are made for each problem. "–" denotes that no feasible solution is found. Where there are differences, we use bold to highlight the best solution found among the four algorithms, and the better Q_{avg} between EDS and ℓ_1-PM

Test problem			SNOPT		LANCELOT		EDS				ℓ_1-PM			
ID	n_v	n_c	Q	T	Q	T	Q_{best}	Q_{avg}	T_{avg}	P_{succ}	Q_{best}	Q_{avg}	T_{avg}	P_{succ}
G1	13	9	–	–	−12.66	0.00	**−15.00**	**−15.00**	1.67	100%	**−15.00**	−13.84	3.24	5%
G2	20	2	–	–	–	–	**−0.80**	**−0.77**	8.95	100%	–	–	–	0%
G3	20	1	–	–	–	–	**−1.00**	**−1.00**	1.53	100%	−0.96	−0.03	2.95	57%
G4	5	6	−30665.54	0.02	−30665.54	0.03	−30665.54	**−30665.54**	0.57	100%	−30665.54	−30637.92	0.27	13%
G5	4	5	**4221.96**	0.05	–	–	**4221.96**	**4221.96**	0.68	100%	–	–	–	0%
G6	2	2	−6961.81	0.01	−6961.81	0.01	−6961.81	**−6961.81**	7.44	100%	−6961.81	−3573.06	15.7	100%
G7	10	8	24.31	0.02	24.31	0.01	24.31	**24.31**	5.03	100%	–	–	–	0%
G8	2	2	0.00	0.01	0.00	0.01	**−0.096**	**−0.096**	0.26	100%	−0.095	−0.014	3.02	15%
G9	7	4	680.63	0.02	680.63	0.01	680.63	**680.63**	1.40	100%	–	–	–	0%
G10	8	6	–	–	2.58	0.12	**1.00**	**1.00**	3.97	100%	–	–	–	0%

Table 2 Results on solving Floudas and Pardalos's NLP benchmarks. All runs are made on a PC workstation with Intel Xeon 2.4 GHZ CPU and 2 GB memory with a time limit of 6000 seconds. All times are in seconds. 100 runs of EDS and ℓ_1-PM are made for each problem. "–" denotes that no feasible solution is found. Where there are differences, we use bold to highlight the best solution found among the four algorithms, and the better Q_{avg} between EDS and ℓ_1-PM

Test problem			SNOPT		LANCELOT		EDS				ℓ_1-PM			
ID	n_v	n_c	Q	T	Q	T	Q_{best}	Q_{avg}	T_{avg}	P_{succ}	Q_{best}	Q_{avg}	T_{avg}	P_{succ}
2.1	5	1	–	–	−16.0	0.01	**−17.0**	**−17.0**	0.12	100%	**−17.0**	−13.4	0.52	100%
2.2	6	2	–	–	−213.0	0.01	−213.0	**−213.0**	0.18	100%	−213.0	−198.3	0.35	95%
2.3	13	9	–	–	−12.7	0.01	**−15.0**	**−15.0**	1.65	100%	**−15.00**	−13.27	3.22	4%
2.4	6	5	–	–	−11.0	0.01	−11.0	−11.0	0.37	100%	−11.0	−11.0	0.86	100%
2.5	10	11	–	–	**−268.0**	0.01	**−268.0**	**−268.0**	2.13	100%	–	–	–	0%
2.6	10	5	–	–	−29.0	0.01	**−39.0**	**−39.0**	0.46	100%	−38.9	−31.8	0.90	100%
2.7.1	20	10	–	–	−183.9	0.01	**−394.8**	**−394.8**	16.2	100%	–	–	–	0%
2.7.2	20	10	–	–	−673.9	0.01	**−884.8**	**−884.7**	10.9	100%	–	–	–	0%
2.7.3	20	10	–	–	−5478.0	0.01	**−8695.0**	**−6703.2**	12.4	100%	–	–	–	0%
2.7.4	20	10	–	–	−633.4	0.01	**−754.8**	**−754.8**	10.2	100%	–	–	–	0%
2.7.5	20	10	–	–	−4105.3	0.01	**−4150.4**	**−4150.4**	28.9	100%	–	–	–	0%
3.3	6	6	−298.0	0.01	−298.0	0.01	**−310.0**	**−310.0**	0.29	100%	−308.0	−304.5	0.21	100%
3.4	3	3	−4.0	0.01	−4.0	0.01	−4.0	−4.0	0.13	100%	−4.0	−4.0	0.22	100%
4.3	4	3	−4.5	0.01	−4.5	0.01	−4.5	−4.5	0.22	100%	−4.5	−4.5	0.19	100%
4.4	4	3	−2.2	0.01	−2.2	0.01	−2.2	−2.2	0.41	100%	−2.2	−2.2	0.35	100%
4.5	6	6	−13.4	0.01	−13.4	0.01	−13.4	−13.4	0.88	100%	−13.4	−13.4	1.26	100%
4.6	2	2	−4.4	0.01	**−5.5**	0.01	**−5.5**	**−5.5**	0.09	100%	**−5.5**	−5.2	0.15	100%
4.7	2	1	−16.7	0.01	−16.7	0.01	−16.7	−16.7	0.07	100%	−16.7	−16.7	0.12	100%

Table 2 (*Continued*)

Test problem			SNOPT		LANCELOT		EDS				ℓ_1-PM			
ID	n_v	n_c	Q	T	Q	T	Q_{best}	Q_{avg}	T_{avg}	P_{succ}	Q_{best}	Q_{avg}	T_{avg}	P_{succ}
5.2	46	36	1.6	0.12	1.6	0.29	1.6	**1.8**	4423.0	90%	–	–	–	0%
5.4	32	26	2.1	0.01	**1.9**	0.29	**1.9**	**2.1**	1924.6	100%	–	–	–	0%
6.2	9	6	−400.0	0.01	−400.0	0.01	−400.0	**−400.0**	3.42	100%	−400.0	−245.3	4.57	42%
6.3	9	6	**−600.0**	0.01	−120.0	0.01	**−600.0**	**−600.0**	31.25	100%	−240.0	−176.0	23.33	24%
6.4	9	6	−750.0	0.01	−750.0	0.01	−750.0	−750.0	0.92	100%	−750.0	−750.0	0.78	100%
7.3	27	19	**1.0**	0.01	1.3	0.64	1.1	**1.3**	5068.7	10%	–	–	–	0%
7.4	38	23	–	–	**1.1**	0.85	**1.1**	**1.2**	3191.2	100%	–	–	–	0%

The comparison between EDS and ℓ_1-PM shows the benefit of the penalty reduction effect of extended duality. It is evident that EDS has much better performance than ℓ_1-PM in terms of both success rate and solution quality. This is due to the fact that ℓ_1-PM, with a single penalty term, often leads to exceedingly large penalty, making the unconstrained optimization ill conditioned and difficult. The results show that extended duality can effectively address this issue.

It should be noted that, as a sampling-based method, EDS takes very long to solve larger problems such as 5.2 and 5.4. A faster unconstrained solver will improve the speed. The non-differentiability of the penalty function can be removed by a simple transformation. Also, we are studying decomposition methods to further improve the efficiency.

4.3 Application to wireless sensor network optimization

As an application of the proposed approach, we study wireless sensor network (WSN) optimization. Our goal is to address the critical challenge of placement and operation management for real-world WSNs in order to achieve desired quality of sensing. Most traditional WSN management schemes are based on a deterministic model [23, 24] which assumes that there is a definite coverage boundary of a sensor. In practice, sensor signals are *probabilistic* since they are affected by signal attenuation and environment noise [25]. Furthermore, many WSNs use distributed *data fusion* techniques to significantly improve their sensing performance [25]. However, considering probabilistic models and data fusion leads to difficult NLPs.

We have found that many optimization problems based on the probabilistic model are suitable to be solved by the EDS algorithm in Fig. 3. Here, we report results on a representative problem. The problem is to find out the locations of N sensors that minimize the *false alarm rate* while maintaining a *minimum detection probability* for every point in a 2-D region A [25]:

$$\text{minimize } P_F \quad \text{subject to} \quad P_D(x, y) \geq \beta, \quad \forall (x, y) \in A, \tag{60}$$

where $P_D(x, y)$ denotes the detection probability of a location (x, y) and P_F denotes the false alarm rate. To compute P_D and P_F, we need to first compute the *local* detection probability P_{D_i} and *local* false alarm rate P_{F_i} for each sensor $i, i = 1, \ldots, N$ as follows:

$$\begin{aligned} P_{F_i} &= \int_{b_i}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{z^2}{2\sigma^2} \right) dz, \\ P_{D_i}(x, y) &= \int_{b_i}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{z - \sqrt{\frac{e}{d((x,y),(x_i,y_i))^a}}}{2\sigma^2} \right) dz. \end{aligned} \tag{61}$$

The probabilistic model is based on a Gaussian noise assumption, where b_i, σ, a, and e are constants, and (x_i, y_i) is the coordinates of the ith sensor. After all local decisions are sent to a *fusion center*, it will find that an event happens at (x, y) if a

majority of the sensors report so. Therefore, we have:

$$\begin{aligned} P_D(x, y) &= \sum_{|S_1|>|S_0|} \prod_{i \in S_0} (1 - P_{D_i}(x, y)) \prod_{j \in S_1} P_{D_j}(x, y), \\ P_F &= \sum_{|S_1|>|S_0|} \prod_{i \in S_0} (1 - P_{F_i}) \prod_{j \in S_1} P_{F_j}, \end{aligned} \tag{62}$$

where S_0 and S_1 denote the set of nodes who detect or do not detect an event, respectively.

The above representative problem manifests the following characteristics of WSN optimization. First, the number of variables is not high. There are only $2N$ variables denoting the locations of N sensors. Second, the problem is nonconvex. Third, the functions are very expensive to evaluate. In fact, the cost for computing $P_D(x, y)$ is $\Theta(2^n)$ since we need to consider *all* combinations of S_0 and S_1. The cost is so expensive that it is *impossible* to directly compute $P_D(x, y)$ or its derivatives. Instead, a Monte-Carlo simulation is typically used in order to estimate $P_D(x, y)$ within reasonable time [25]. Thus, with no closed form, it is difficult to apply existing optimization packages. Finally, in some other scenarios, such as power management, it is required to use discrete variables denoting the on/off status of a sensor. Therefore, the problem may become a DNLP or MINLP.

Due to these difficulties, most existing optimization solvers cannot be applied. Previous work in WSN have used some greedy heuristic methods which are ad-hoc and suboptimal [25]. However, in WSN applications, plan quality is important for economic and security reasons, and global optimality is desirable.

We have applied EDS to solve the problem in (60) to (62). We have found that, based on the extended duality which eliminates the duality gap for nonconvex problems, EDS can yield much better solution than the greedy heuristic methods [25]. For example, in a 30×30 grid, EDS can find a sensor placement with only 13 sensors to meet the given thresholds of $P_D \geq 90\%$ and $P_F \leq 5\%$ while a previous heuristic method needs 18 sensors. In a 120×120 grid, EDS can find a sensor placement with only 202 sensors to meet the same constraints while the heuristic method needs 226 sensors. The minimum number is obtained by a binary search that solves (60) multiple times using EDS. Since the number of variables is moderate, the optimization can be done in a reasonable time. It takes 6 minutes on a PC workstation to solve a 120×120 grid.

EDS provides an attractive and practical solution to the above problem. The extended duality theory allows us to optimally solve this nonconvex and nondifferentiable problem. Since the number of variables is not high, the time cost is moderate. It allows us to add into the sensor-network model other complex constraints, such as communication constraints.

5 Conclusions

In this paper, we have proposed the theory of extended duality for nonlinear optimization. The theory overcomes the limitations of conventional duality theory by provid-

ing a duality condition that leads to zero duality gap for general nonconvex optimization problems in discrete, continuous and mixed spaces. Based on an ℓ_1^m-penalty function, the proposed theory requires less penalty values to achieve zero duality gap comparing to previous efforts for removing the duality gap, thus alleviating the ill conditioning of dual functions. A search algorithm based on extended duality has optimally solved many nonlinear constrained problems and shown computational advantages over augmented Lagrangian and ℓ_1-penalty methods. It also solves a complex, nonconvex optimization problem in sensor network management.

References

1. Aubin, J.P., Ekeland, I.: Estimates of the duality gap in nonconvex optimization. Math. Oper. Res. **1**, 225–245 (1976)
2. Ben-Tal, A., Eiger, G., Gershovitz, V.: Global optimization by reducing the duality gap. Math. Program. **63**, 193–212 (1994)
3. Bertsekas, D.P.: Distributed dynamic programming. Trans. Autom. Control **AC-27**(3), 610–616 (1982)
4. Bertsekas, D.P.: Nonlinear Programming. Athena Scientific, Belmont (1999)
5. Burachik, R.S., Rubinov, A.: On the absence of duality gap for Lagrange-type functions. J. Indust. Manag. Optim. **1**(1), 33–38 (2005)
6. Burke, J.V.: Calmness and exact penalization. SIAM J. Control Optim. **29**, 493–497 (1991)
7. Burke, J.V.: An exact penalization viewpoint of constrained optimization. SIAM J. Control Optim. **29**, 968–998 (1991)
8. Conn, A.R., Gould, N.I.M., Toint, Ph.L.: LANCELOT: A Fortran Package for Large-Scale Nonlinear Optimization. Heidelberg, Springer (1992)
9. Floudas, C.A., Pardalos, P.M.: A Collection of Test Problems for Constrained Global Optimization Algorithms. Lecture Notes in Computer Science, vol. 455. Springer, Berlin (1990)
10. Gill, P.E., Murray, W., Saunders, M.: SNOPT: An SQP algorithm for large-scale constrained optimization. SIAM J. Optim. **12**, 979–1006 (2002)
11. Gould, N.I.M., Orban, D., Toint, P.L.: An interior-point ℓ_1-penalty method for nonlinear optimization. Technical Report RAL-TR-2003-022, Rutherford Appleton Laboratory Chilton, Oxfordshire, UK, November (2003)
12. Huang, X.X., Yang, X.Q.: A unified augmented Lagrangian approach to duality and exact penalization. Math. Oper. Res. **28**(3), 533–552 (2003)
13. Koziel, S., Michalewics, Z.: Evolutionary algorithms, homomorphous mappings, and constrained parameter optimization. Evolut. Comput. **7**(1), 19–44 (1999)
14. Luo, Z.Q., Pang, J.S.: Error bounds in mathematical programming. Math. Program. Ser. B, **88**(2) (2000)
15. Nedić, A., Ozdaglar, A.: A geometric framework for nonconvex optimization duality using augmented Lagrangian functions. J. Glob. Optim. **40**(4), 545–573 (2008)
16. Pang, J.S.: Error bounds in mathematical programming. Math. Program. **79**, 299–332 (1997)
17. Rockafellar, R.T.: Augmented Lagrangian multiplier functions and duality in nonconvex programming. SIAM J. Control Optim. **12**, 268–285 (1974)
18. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Springer, Berlin (1998)
19. Rubinov, A.M., Glover, B.M., Yang, X.Q.: Decreasing functions with applications to penalization. SIAM J. Optim. **10**, 289–313 (1999)
20. Rubinov, A.M., Glover, B.M., Yang, X.Q.: Modified Lagrangian and penalty functions in continuous optimization. Optimization **46**, 327–351 (1999)
21. Tuy, H.: On solving nonconvex optimization problems by reducing the duality gap. J. Glob. Optim. **32**, 349–365 (2005)
22. Ferrier, G.D., Goffe, W.L., Rogers, J.: Global optimization of statistical functions with simulated annealing. J. Econ. **60**(1), 65–99 (1994)
23. Wang, X., Xing, G., Zhang, Y., Lu, C., Pless, R., Gill, C.: Integrated coverage and connectivity configura-tion in wireless sensor networks. In: Proc. First ACM Conference on Embedded Networked Sensor Systems (2003)

24. Xing, G., Lu, C., Pless, R., Huang, Q.: On greedy geographic routing algorithms in sensing-covered networks. In: Proc. ACM International Symposium on Mobile Ad Hoc Networking and Computing (2004)
25. Xing, G., Lu, C., Pless, R., O'Sullivan, J.A.: Co-Grid: An efficient coverage maintenance protocol for distributed sensor networks. In: Proc. International Symposium on Information Processing in Sensor Networks (2004)
26. Yang, X.Q., Huang, X.X.: A nonlinear Lagrangian approach to constraint optimization problems. SIAM J. Optim. **11**, 1119–1144 (2001)