A Primal–Dual Penalty Method via Rounded Weighted-ℓ_1 Lagrangian Duality

R. S. Burachik* C. Y. Kaya* C. J. Price†

May 4, 2020

Abstract We propose a new duality scheme based on a sequence of smooth minorants of the weighted-ℓ_1 penalty function, interpreted as a parametrized sequence of augmented Lagrangians, to solve nonconvex and nonsmooth constrained optimization problems. For the induced sequence of dual problems, we establish strong asymptotic duality properties. Namely, we show that (i) the sequence of dual problems are convex and (ii) the dual values monotonically increase converging to the optimal primal value. We use these properties to devise a subgradient based primal–dual method, and show that the generated primal sequence accumulates at a solution of the original problem. We illustrate the performance of the new method with three different types of test problems: A polynomial nonconvex problem, large instances of the celebrated kissing number problem, and the Markov–Dubins problem. Our numerical experiments demonstrate that, when compared with the traditional implementation of a well-known smooth solver, our new method (using the same solver in its subproblem) can find better quality solutions, i.e., "deeper" local minima, or solutions closer to the global minimum. Moreover, our method seems to be more time efficient, especially when the problem has a large number of constraints.

Key words: Nonconvex optimization; Nonsmooth optimization; Subgradient methods; Duality schemes; Primal–dual methods; Penalty function methods; ℓ_1-penalty function; Kissing number problem; Markov–Dubins problem.

Mathematical Subject Classification: 49M29; 90C26; 90C90

1 Introduction

Let X be a compact metric space, and $f : X \to \mathbb{R}_\infty := \mathbb{R} \cup \{\infty\}$ be a lower semicontinuous function. Assume that $h : X \to \mathbb{R}^m$ and $g : X \to \mathbb{R}^r$ are continuous and define $X_0 := \{x \in X : h(x) = 0,\, g(x) \le 0\}$. We consider the minimization problem:

$$(P) \qquad\qquad \text{minimize } f(x) \;\text{ subject to }\; x \text{ in } X_0$$

where X_0 is assumed to be a non-empty proper subset of X (i.e., $\emptyset \subsetneq X_0 \subsetneq X$). It is well-known that when (P) is convex, a classical Lagrangian can be used to obtain a dual problem, and zero

*Mathematics, UniSA STEM, University of South Australia, Australia; {regina.burachik} or {yalcin.kaya} @unisa.edu.au .

†School of Mathematics and Statistics, University of Canterbury, New Zealand; chrisj.price@canterbury.ac.nz .

duality gap will hold under mild constraint qualifications. When (P) is not convex, augmented Lagrangian functions can provide a duality scheme with zero duality gap. These functions combine the objective and the constraints of (P) in a suitable way, and this combination is obtained by means of penalty parameters. Theoretical studies ensuring zero duality gap for (P) using augmented Lagrangians can be found in [8, 9, 11, 17, 19–21, 37, 38]. These works consider a fixed augmented Lagrangian function, and hence a fixed dual problem.

However, the above-mentioned duality schemes have disadvantages, such as the lack of smoothness of the Lagrangian at points close to the solution set, even when the problem data is smooth. This motivated the authors of [27] to analyse suitable perturbations of a given augmented Lagrangian function for nonlinear semidefinite programming problems. The work in [27] was later extended to more general parametrized families of augmented Lagrangians in [16]. More recently, [5] has analyzed asymptotic duality properties for primal-dual pairs induced by a sequence (L_k) of Lagrangian functions, which, in turn, induce a sequence of dual problems (D_k).

Our aim in the current paper is three-fold:

1. *Propose and analyze a primal–dual scheme in which we have, as in [5], a sequence (L_k) of Lagrangian functions and a corresponding sequence of dual problems (D_k).* Unlike [5], which establishes zero duality gap under specific assumptions on the sequence (L_k), our analysis exploits a type of Lagrangian function recently used in [33]. The latter work defines a smooth, or rounded, approximation of a weighted-ℓ_1 penalty function, which is parametrized by a positive scalar w. In this way, given a sequence (w_k) of decreasing positive numbers, we obtain a sequence of dual problems that approach the weighted-ℓ_1 penalty dual problem. The properties of the Lagrangian we use ensure that when the problem is smooth the augmented Lagrangian function is also smooth.

   For our novel duality framework, we show that the dual functions are concave and that the sequence of dual optimal values is increasing and converges to the optimal value of (P). Our proof is inspired by [5] but adapted to our particular type of Lagrangian (see Corollary 3.4).

2. *Devise a subgradient-type method adapted to our sequence of (convex) dual functions, and use the duality properties to develop a primal–dual method in which the primal sequence is easier to compute (thanks to smoothness) and accumulates at a solution of (P).* Indeed, we show that the primal sequence generated by our method accumulates at a solution of (P) for two possible choices of the step size.

3. *Illustrate the computational advantages of our method by means of challenging smooth optimization problems.* Since our augmented Lagrangian is smooth, it makes sense to study its performance for smooth problems. Our experiments demonstrate that our method is competitive when compared with existing differentiable solvers, both in terms of the quality of the solutions and the computational speed. Our method becomes particularly advantageous when Problem (P) has a large number of constraints.

Many efficient numerical methods for nonconvex problems have been developed using the duality properties induced by augmented Lagrangian functions, for a fixed dual problem. Indeed, when the augmented Lagrangian is the one introduced by Rockafellar and Wets in [34, Chapter 11], the dual problem happens to be convex. This makes it possible to use subgradient-type steps for improving the current value of the dual function. This approach has since given rise to the so-called deflected subgradient (DSG) methods [6, 7, 9, 13–15, 23, 24].

DSG methods can generate primal sequences that accumulate at a solution of (P). Even though the convergence theory of DSG methods requires the global minimization of a nonsmooth and nonconvex problem in each iteration, it has been observed in computational practice, for example in [6, 14], that local minimization of the subproblems will still lead to a local minimizer of the original problem.

A further virtue of DSG methods, which has been demonstrated in [13, 14], is that they can provide meaningful updates of the penalty parameter when the augmented Lagrangian is the ℓ_1 penalty function. In the present paper, inspired by [13, 14], we update the "weights," or the "penalty parameters," of the rounded weighted-ℓ_1 penalty function. Therefore we refer to the new method we propose here as the *primal–dual (P–D) penalty method*, and provide an associated computational procedure in Algorithm 1.

Since the augmented Lagrangians we propose are smooth, our subproblems are also smooth when the data of the problem is smooth. There is a plethora of methods and software for smooth optimization problems—see e.g. [1–3, 25, 36]. Over the last few decades, these methods have been demonstrated to be successful and efficient in finding a local minimizer, or at least a stationary point, of a wide range of challenging problems. Despite this reported success, the problem of finding a global minimizer of such challenging problems is extremely difficult. Especially for smooth problems that are relatively large-scale and have a large number of local minima, these popular methods can at most promise to find a local minimum, which is not necessarily near a global minimum.

Even though our theoretical analysis covers general nonsmooth problems with variables in any metric compact space, our numerical experiments focus on some challenging instances of smooth problems. These experiments show that our proposed method is competitive when compared with the existing differentiable methods in two major aspects:

- In the presence of many local minima, the P–D penalty method seems more likely to find a local minimum close to the global minimum and find it in a shorter time;
- If there is a large number of constraints in the problem, the P–D penalty method seems to take a shorter CPU time in finding a local minimizer.

For the purpose of illustrating these aspects, we consider three problems for running Algorithm 1: (i) a nonconvex quartic polynomial optimization problem in $\mathbb{R}^3$ with polynomial equality constraints [29], (ii) several large-scale instances of the celebrated kissing number problem [14, 28, 32] and (iii) the Markov–Dubins problem [30]. We use the optimization modelling language software AMPL [22] paired up with the popular commercial constrained optimization software Knitro [2] in the experiments.

For the comparisons, we solve a given problem by using the AMPL–Knitro suite in two different ways: (i) We code the objective function and the constraints in AMPL in the usual way and (ii) we code the constraints as embedded in the rounded weighted-ℓ_1 penalty function and perform the penalty parameter updates in a loop, also in AMPL. In other words, we effectively compare Knitro "with itself," by running it (i) on its own and (ii) as part of the P–D penalty method in Step 2 of Algorithm 1. We carry out many thousands of runs with randomized initial guesses for each example so as to obtain reliable information on the percentages of the time a minimum is found. This information in turn provides an idea as to how close in general the local minima that were found are close to the global one, as well as the computational time that each run took on the average, for each approach.

The paper is organized as follows. In Section 2 we give the basic definitions, including the definition of the Lagrangian, as used in [33]. In Section 3 we present the duality setting, show that the dual problem is convex for each fixed parameter $w > 0$, establish the zero duality gap property and obtain results regarding the structure of the set of dual solutions. In this section we provide the theoretical basis for the search direction and the stopping criteria to be used later in Algorithm 1. In Section 4 we describe the subgradient method for improving the dual values and show that every accumulation point of the primal sequence is a solution of the primal problem. In Section 5, we present a computational algorithm and the numerical experiments. Finally, Section 6 concludes the paper.

2 Basic Facts, Definitions and Assumptions

For $m \in \mathbb{N}$, denote by $\mathbb{R}^m_{++} := \{y \in \mathbb{R}^m \,:\, y_i > 0 \text{ for all } i = 1, \ldots, m\}$ the positive orthant of $\mathbb{R}^m$ and by $\mathbb{R}^m_{+} := \{y \in \mathbb{R}^m \,:\, y_i \geq 0 \text{ for all } i = 1, \ldots, m\}$ the nonnegative orthant of $\mathbb{R}^m$. For $p \in \mathbb{N}$ and $z, z' \in \mathbb{R}^p$, we write $z \leq z'$ when $z_l \leq z'_l$ for every coordinate $l = 1, \ldots, p$. We denote by $\|\cdot\|_2$ the ℓ_2-norm and by $\|\cdot\|_\infty$ the ℓ_∞-norm. For problem (P), denote by S^* the (nonempty) set of solutions of (P) and by

$$M_P := \inf_{x \in X_0} f(x),$$

the optimal value of the problem (P). Our analysis needs to take care of both equality and inequality constraints. This will be achieved by means of the following two auxiliary functions from [33].

Definition 2.1 *Let $w \geq 0$. Define $\eta : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$ as*

$$\eta(t, w) := \begin{cases} \dfrac{t^2}{2w} & \text{if } |t| < w, \\[2ex] |t| - \dfrac{w}{2} & \text{if } |t| \geq w. \end{cases} \tag{1}$$

Define $\gamma : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$ as

$$\gamma(t, w) := \begin{cases} \dfrac{t^2}{2w}\,, & \text{if } \ 0 < t < w\,, \\[2ex] t - \dfrac{w}{2}\,, & \text{if } \ t \geq w\,, \\[2ex] 0\,, & \text{if } \ t \leq 0\,. \end{cases} \tag{2}$$

Illustrations of η and γ appear in Figure 1.

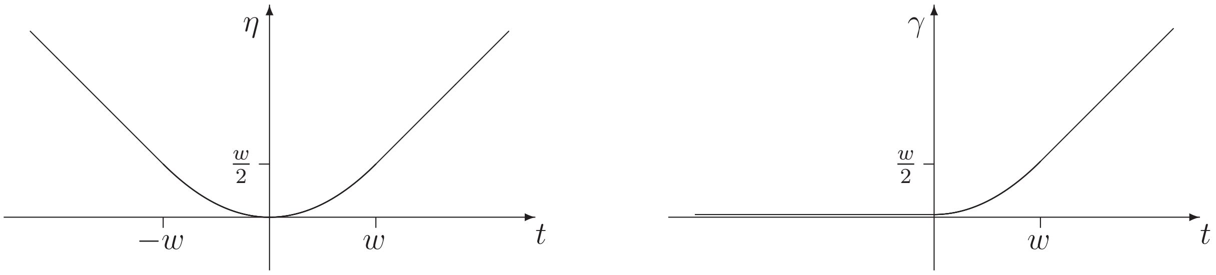


Figure 1: Graphs of $\eta(t)$ (left image) and $\gamma(t)$ showing the rounding regions.

Remark 2.1 From (1) we see that η is quadratic over the interval $(-w, w)$, and hence this interval can be seen as a *rounding region* of 'width' w. Outside the rounding region, η behaves like an ℓ_1 penalty term. Note that the function η is a smooth (i.e., continuously differentiable) lower minorant of the function $\bar{g} := |\cdot|$. Hence, for small values of $w > 0$, it can be seen as a smooth approximation of $\bar{g}$. In a similar way, the function γ can be seen as a smooth approximation of the function $[\cdot]_+ := \max\{\cdot, 0\}$ for small values of $w > 0$.

The following lemma collects several useful properties of the functions η and γ.

Lemma 2.1 *Let η and γ be as defined in (1) and (2), respectively. The following properties hold.*

(a) $\eta(t, w) \geq 0$ *and* $\gamma(t, w) \geq 0$ *for all* $(t, w) \in \mathbb{R} \times \mathbb{R}_{++}$.

(b) For all $(t, w) \in \mathbb{R} \times \mathbb{R}_{++}$,

$$\frac{\partial \eta}{\partial w}(t, w) = \begin{cases} -\dfrac{t^2}{2w^2} & \text{if } |t| < w \\ -\dfrac{1}{2} & \text{if } |t| \geq w \end{cases} \quad \text{and} \quad \frac{\partial \gamma}{\partial w}(t, w) = \begin{cases} -\dfrac{t^2}{2w^2} & \text{if } 0 < t < w \\ -\dfrac{1}{2} & \text{if } t \geq w \\ 0 & \text{if } t \leq 0 \end{cases}$$

(c) $|t| - w/2 \leq \eta(t, w) \leq |t|$ *and* $[t]_+ - w/2 \leq \gamma(t, w) \leq [t]_+$, *for all* $(t, w) \in \mathbb{R} \times \mathbb{R}_+$.

(d) For all $t_1, t_2 \in \mathbb{R}$ *and* $w \in \mathbb{R}_{++}$, $|t_1| < |t_2|$ *implies* $\eta(t_1, w) < \eta(t_2, w)$.

(e) For all $t_1, t_2 \in \mathbb{R}$ *and* $w \in \mathbb{R}_{++}$, $0 < t_1 < t_2$ *or* $t_1 < 0 < t_2$ *implies* $\gamma(t_1, w) < \gamma(t_2, w)$.

Proof Item (a) follows immediately from Definitions (1) and (2). Item (b) follows from differentiating (1) and (2) with respect to w. Item (c) for η holds trivially when $|t| \geq w$ and in particular when $w = 0$. Otherwise, write $\eta(t, w) = |t|^2/(2w)$, and the right hand inequality in (c) follows from

$$|t| < w \quad \Rightarrow \quad |t|^2 \leq w|t| \leq 2w|t| \quad \Rightarrow \quad \eta(t, w) = \frac{|t|^2}{2w} \leq |t|.$$

For the left hand inequality in (c)

$$(|t|-w)^2 = |t|^2 - 2w|t| + w^2 \geq 0 \qquad \forall (t,w) \in \mathbb{R} \times \mathbb{R}_{++}$$

which implies

$$2w|t| - w^2 \leq |t|^2 \qquad \Rightarrow \qquad |t| - \frac{w}{2} \leq \frac{|t|^2}{2w} = \eta(t,w)\,.$$

Let us now check item (c) for γ. If $t \leq 0$ then $\gamma(t,w) = 0$ so, in particular, $[t]_+ - w/2 = -w/2 \leq \gamma(t,w) \leq [t]_+ = 0$. If $t > 0$ then we consider two cases. If $t \in (0,w)$ then a calculation similar to the one for η above yields

$$0 < t < w \qquad \Rightarrow \qquad t^2 \leq wt \leq 2wt \qquad \Rightarrow \qquad \gamma(t,w) = \frac{t^2}{2w} \leq t.$$

As above, the inequality $t - (1/2w) \leq t^2/2w$ is always true, and the result follows a similar argument, mutatis mutandis, as the one in the proof of (c) for η. For proving (d), we first note that $\eta(t,w) = \eta(|t|,w)$ for all $(t,w) \in \mathbb{R} \times \mathbb{R}_{++}$. The cases when $|t_1| < |t_2| < w$ and $w \leq |t_1| < |t_2|$ are obvious from the definition of η. Otherwise, for $|t_1| < w \leq |t_2|$ we have

$$\eta(t_1,w) = \frac{|t_1|^2}{2w} < \frac{w^2}{2w} = w - \frac{w}{2} \leq |t_2| - \frac{w}{2} = \eta(t_2,w).$$

Finally, we prove (e). As in (d), the claim is trivial if $0 < t_1 < t_2 < w$ or $0 < w \leq t_1 < t_2$. Assume that $0 < t_1 < w \leq t_2$. We have

$$\gamma(t_1,w) = \frac{(t_1)^2}{2w} < \frac{w^2}{2w} = w - \frac{w}{2} \leq t_2 - \frac{w}{2} = \gamma(t_2,w).$$

If $t_1 < 0 < w \leq t_2$ then

$$\gamma(t_1,w) = 0 < w - \frac{w}{2} \leq t_2 - \frac{w}{2} = \gamma(t_2,w),$$

which completes the proof of (e). □

The following technical lemma will be important in the establishment of zero duality gap.

Lemma 2.2 *Assume that $(w_k) \subset \mathbb{R}_+$ is a sequence bounded above by $\bar{w} > 0$, and let $(t_k) \subset \mathbb{R}$. The following properties hold.*

(a) If $\lim_{k\to\infty} \eta(t_k,w_k) = 0$ then $\lim_{k\to\infty} |t_k| = 0$.

(b) If $\lim_{k\to\infty} \gamma(t_k,w_k) = 0$ then $\lim_{k\to\infty} [t_k]_+ = 0$.

Proof (a) Assume that (a) is not true, which means that there exists an infinite set $\mathcal{K} \subset \mathbb{N}$ and a $\delta > 0$ such that $|t_k| > \delta$ for all $k \in \mathcal{K}$. We claim that in this case there exists $k_0 \in \mathcal{K}$ such that $|t_k| \geq w_k$ for all $k \geq k_0$, $k \in \mathcal{K}$. Indeed, assume that the claim is not true, i.e., there exists an infinite subset $\mathcal{K}_1 \subset \mathcal{K}$ such that $|t_k| < w_k$ for all $k \in \mathcal{K}_1$. Using the definition of η we can write

$$0 = \lim_{\substack{k \to \infty \\ k \in \mathcal{K}_1}} \eta(t_k,w_k) = \lim_{\substack{k \to \infty \\ k \in \mathcal{K}_1}} \frac{t_k^2}{2w_k} \geq \lim_{\substack{k \to \infty \\ k \in \mathcal{K}_1}} \frac{\delta^2}{2w_k} \geq \frac{\delta^2}{2\bar{w}} > 0\,,$$

a contradiction. Therefore, the claim is true and there exists $k_0 \in \mathcal{K}$ such that $|t_k| \geq w_k$ for all $k \geq k_0$, $k \in \mathcal{K}$. This implies that for all $k \geq k_0$, $k \in \mathcal{K}$, we have

$$0 = \lim_{\substack{k \to \infty \\ k \in \mathcal{K}}} \eta(t_k, w_k) = \lim_{\substack{k \to \infty \\ k \in \mathcal{K}}} |t_k| - w_k/2 \geq \lim_{\substack{k \to \infty \\ k \in \mathcal{K}}} w_k/2 \geq 0,$$

which in turn implies that $w_k \to 0$ for $k \in \mathcal{K}$ tending to ∞. Hence,

$$0 = \lim_{\substack{k \to \infty \\ k \in \mathcal{K}}} \eta(t_k, w_k) + w_k/2 = \lim_{\substack{k \to \infty \\ k \in \mathcal{K}}} |t_k| \geq \delta > 0\,,$$

a contradiction. Altogether, we deduce that $\lim_{k\to\infty} |t_k| = 0$. An identical proof works for (b), with $|t_k|$ replaced everywhere by $[t_k]_+$. □

2.1 A Lagrangian function for (P)

Using the functions η and γ given in Definition 2.1, a Lagrangian for problem (P) can be defined as follows.

Definition 2.2 *Fix $w \geq 0$. For $h : X \to \mathbb{R}^m$ and $g : X \to \mathbb{R}^r$ as in problem (P), consider the function $L_w : X \times \mathbb{R}^m_{++} \times \mathbb{R}^r_{++} \to \mathbb{R}$ defined by*

$$L_w(x, u, v) := f(x) + \sum_{i=1}^{m} u_i\, \eta(h_i(x), w) + \sum_{j=1}^{r} v_j\, \gamma(g_j(x), w). \tag{3}$$

Our Lagrangian approximates the weighted-ℓ_1 penalty function for small values of $w > 0$. To make this statement precise, we recall next the (weighted) ℓ_1 penalty function in the context of Problem (P).

Definition 2.3 *For u and h as in Definition 2.2, the (weighted) ℓ_1 penalty function for problem (P) is defined as the function $\varphi_{\ell_1} : X \times \mathbb{R}^m_{++} \times \mathbb{R}^r_{++} \to \mathbb{R}$ given by*

$$\varphi_{\ell_1} := f(x) + \sum_{i=1}^{m} u_i |h_i(x)| + \sum_{j=1}^{r} v_j\, [g_j(x)]_+\,. \tag{4}$$

Remark 2.2 In (3) and (4), the coordinates u_i for $i = 1, \ldots, m$ and v_j for $j = 1, \ldots, r$ act as penalty parameters. The advantage of (3) over (4) is that the terms $u_i \eta(h_i(x), w)$ and $v_j\, \gamma(g_j(x), w)$ are quadratic when $|h_i(x)| < w$ and $g_j(x) \in (0, w)$, respectively. Everywhere else (3) mimics the ℓ_1 penalty function defined in (4). The Lagrangian L_w can thus be viewed as an inexact penalty function for Problem (P) with the property that the exact ℓ_1 penalty function is recovered when $w = 0$. Note that the parameter $w \geq 0$ gives a measure of the proximity between L_w and φ_{ℓ_1}. Indeed, it follows readily from Lemma 2.1(c) that for all $w \geq 0$ we can write

$$|L_w(x, u, v) - \varphi_{\ell_1}(x, u, v)| \leq \frac{w}{2}(\|u\|_1 + \|[v]_+\|_1),$$

where $\|\cdot\|_1$ is the ℓ_1-norm in $\mathbb{R}^m$, and for $z \in \mathbb{R}^p$ we define $([z]_+)_j := [z_j]_+$ for all $j = 1, \ldots, p$. The above inequality clearly gives $L_0 = \varphi_{\ell_1}$.

3 Primal–Dual Setting

In our analysis, we may consider a decreasing sequence $(w_k) \subset \mathbb{R}_+$ and generate a sequence of Lagrangians $(L_{w_k}(\cdot,\cdot,\cdot))$, defined as in (3) for $w := w_k$. For simplicity we may use the notation $L_k := L_{w_k}(\cdot,\cdot,\cdot)$. We now define the dual problem, by means of the Lagrangian L_w given in Definition 2.2. Fix $w \geq 0$ and let $q_w : \mathbb{R}^m \times \mathbb{R}^r \to \mathbb{R}_{-\infty}$ be defined as

$$q_w(u,v) := \begin{cases} \inf_{x\in X} L_w(x,u,v), & \text{if } (u,v) \in \mathbb{R}^m_{++} \times \mathbb{R}^r_{++}\,, \\ \\ -\infty & \text{c.c.}\,, \end{cases} \tag{5}$$

The dual problem, denoted by (D_w) is given by

$$\sup_{(u,v)\in\mathbb{R}^m_{++}\times\mathbb{R}^r_{++}} q_w(u,v). \tag{6}$$

We denote its supremum by M_{D_w}.

Proposition 3.1 *Fix $(u,v,w) \in \mathbb{R}^m_{++} \times \mathbb{R}^r_{++} \times \mathbb{R}_+$. The dual function q_w is concave and everywhere continuous.*

Proof The proof is standard, we include it for completeness. By Definition 5 and compactness of X, we have that $\text{dom}(q_w) := \{(u,v) : q_w(u,v) > -\infty\} = \mathbb{R}^m_{++} \times \mathbb{R}^r_{++}$, a convex set. Hence, it is enough to check concavity over the domain set $\mathbb{R}^m_{++} \times \mathbb{R}^r_{++}$. Indeed, for $(u,v) \in \mathbb{R}^m_{++} \times \mathbb{R}^r_{++}$ we have that

$$q_w(u,v) = \inf_{x\in X}\left[f(x) + \sum_{i=1}^m u_i \eta(h_i(x),w) + \sum_{j=1}^r v_j \gamma(g_j(x),w)\right],$$

where the function between square brackets is affine in the variable (u,v). The infimum of affine functions is concave. The continuity follows from the fact that $q_w(u,v)$ (with w fixed) is a concave function of (u,v) and is everywhere finite (because X is compact). □

The following result, which can be found in [12, Proposition 3.1.15] shows that the infimum in (5) is always achieved for some point in X.

Lemma 3.3 *If ϕ is a lower semi-continuous function mapping from a compact set K into $\mathbb{R}$, then there exists $x_0 \in K$ such that $\phi(x) \geq \phi(x_0)$ for all $x \in K$.*

Corollary 3.1 *For all $(u,v,w) \in \mathbb{R}^m_{++} \times \mathbb{R}^r_{++} \times \mathbb{R}_+$ there exists an $x^\sharp(u,v,w) \in X$ such that*

$$q_w(u,v) = L_w\left(x^\sharp(u,v,w),u,v\right)$$

Proof Via Lemma 3.3 with $\phi := L_w(\cdot,u,v)$ and $K := X$. Indeed, our assumptions on (P) imply that, for every fixed $(u,v,w) \in \mathbb{R}^m_{++} \times \mathbb{R}^r_{++} \times \mathbb{R}_+$, the function $L_w(\cdot,u,v)$ is lower-semicontinuous (note that $\eta(\cdot,w)$, $\gamma(\cdot,w)$, $h(\cdot)$ and $g(\cdot)$ are continuous). Since X is compact, the claim follows from Lemma 3.3. □

The corollary above motivates the next definition.

Definition 3.4 *Given $(u, v, w) \in \mathbb{R}^m_{++} \times \mathbb{R}^r_{++} \times \mathbb{R}_+$, consider the set*

$$T(u, v, w) := \operatorname*{argmin}_{x \in X} L_w(x, u, v),$$

of all minimizers of the Lagrangian induced by (u, v, w).

Corollary 3.1 implies that $T(u, v, w) \neq \emptyset$ for every $(u, v, w) \in \mathbb{R}^m_{++} \times \mathbb{R}^r_{++} \times \mathbb{R}_+$.

Corollary 3.2 *There exists $F_L \in \mathbb{R}$ such that $f(x) \geq F_L$ for all $x \in X$.*

Proof Via Lemma 3.3 with $\phi := f$ and $K := X$. □

We will be concerned with establishing zero duality gap properties in the situation in which the parameter w is decreasing. More precisely, from now on we assume that we have a sequence $(w_k) \subset \mathbb{R}_+$ such that $w_{k+1} \leq w_k$ for all k. For each k we have a dual problem $(D_k) := (D_{w_k})$, a Lagrangian $L_k := L_{w_k}$ and a dual function $q_{w_k} := q_k$.

Lemma 3.4 (weak duality) *Take $(w_k) \subset \mathbb{R}_+$ and consider problem (D_k). For every $(u, v) \in \mathbb{R}^m_+ \times \mathbb{R}^r_+$ we have*

$$q_k(u, v) \leq M_P \quad \forall\, k \in \mathbb{N}. \tag{7}$$

Consequently, $M_{D_k} \leq M_P$ for all $k \in \mathbb{N}$.

Proof For any $x \in X_0$ we must have $h_i(x) = 0$ and $g_j(x) \leq 0$ for every $i = 1, \ldots, m$ and every $j = 1, \ldots, r$. Hence $\eta(h_i(x), w_k) = 0 = \gamma(g_j(x), w_k)$ for all k, which means $L_k(x, u, v) = f(x)$ for all $(x, u, v) \in X_0 \times \mathbb{R}^m_+ \times \mathbb{R}^r_+$. Now,

$$q_k(u, v) = \min_{x \in X} L_k(x, u, v) \leq \min_{x \in X_0} L_k(x, u, v) = \min_{x \in X_0} f(x) = M_P$$

where the inequality is because $X_0 \subseteq X$. Since no additional restrictions were placed on u, v or w_k, this holds for all $u, v \geq 0$ and $w_k \geq 0$. The last statement follows by taking the supremum of q_k over all $u, v \geq 0$. □

Lemma 3.5 *Let $(w_k) \subset \mathbb{R}_+$ be a decreasing sequence and let (L_k) be defined as in (3) for $w := w_k$. Fix $(x, u, v) \in X \times \mathbb{R}^m_{++} \times \mathbb{R}^r_{++}$. The following properties hold.*

(a) $L_k(x, u, v) \leq L_{k+1}(x, u, v) \leq \varphi_{\ell_1}(x, u, v)$, *for every $k \in \mathbb{N}$.*

(b) The dual optimal values verify $M_{D_k} \leq M_{D_{k+1}} \leq M_P$ for every $k \in \mathbb{N}$.

(c) In the particular case in which $(w_k) \downarrow 0$, we have

$$\lim_{k \to \infty} L_k(x, u, v) = \varphi_{\ell_1}(x, u, v).$$

Proof For part (a), we use Lemma 2.1(b). Indeed, because $0 \le w_{k+1} \le w_k$ and the partial derivatives w.r.t. w of both $\eta(t,\cdot)$ and $\gamma(t,\cdot)$ are non-positive, we have that

$$|\cdot| \ge \eta(\cdot, w_{k+1}) \ge \eta(\cdot, w_k),$$

and

$$[\cdot]_+ \ge \gamma(\cdot, w_{k+1}) \ge \gamma(\cdot, w_k).$$

Hence,

$$\begin{aligned} L_k(x,u,v) &= f(x) + \textstyle\sum_{i=1}^m u_i\, \eta(h_i(x), w_k) + \sum_{j=1}^m v_j\, \gamma(g_j(x), w_k) \\ &\le f(x) + \textstyle\sum_{i=1}^m u_i\, \eta(h_i(x), w_{k+1}) + \sum_{j=1}^m v_j\, \gamma(g_j(x), w_{k+1}) \\ &= L_{k+1}(x,u,v) \le f(x) + \textstyle\sum_{i=1}^m u_i\, |h_i(x)| + \sum_{j=1}^m v_j\, [g_j(x)]_+ = \varphi_{\ell_1}(x,u,v), \end{aligned}$$

where we also used Lemma 2.1(c). This proves part (a). We proceed to prove part (b). Using (a) and the definition of the dual function, write

$$q_k(u,v) \le L_k(x,u,v) \le L_{k+1}(x,u,v),$$

for every $(x,u,v) \in X \times \mathbb{R}^m_{++} \times \mathbb{R}^r_{++}$. The above expression gives $q_k(u,v) \le L_{k+1}(x,u,v)$ for every $x \in X$. Taking infimum over $x \in X$ in the right hand side we derive

$$q_k(u,v) \le q_{k+1}(u,v) \le M_{D_{k+1}},$$

for every $(u,v) \in \mathbb{R}^m_{++} \times \mathbb{R}^r_{++}$. Using the leftmost and rightmost sides of the last expression we deduce that $M_{D_k} \le M_{D_{k+1}}$. The inequality $M_{D_{k+1}} \le M_P$ follows from Lemma 3.4. We proceed to prove (c). Assume that $(w_k) \downarrow 0$. We will use Remark 2.2, which in our case becomes

$$\lim_{k\to\infty} |L_k(x,u,v) - \varphi_{\ell_1}(x,u,v)| \le \lim_{k\to\infty} \frac{w_k}{2}(\|u\|_1 + \|[v]_+\|_1) = 0,$$

as desired. □

The following property establishes our (asymptotic) weak duality result.

Corollary 3.3 (Asymptotic weak duality) *Let (L_k) be as in (3) for a decreasing sequence $(w_k) \subset \mathbb{R}_+$. Then*

$$\lim_{k\to\infty} M_{D_k} \le M_P.$$

Proof The statement follows directly from Lemma 3.5(b). □

The next lemma is crucial in showing that the limit in the last corollary is precisely M_P.

Lemma 3.6 *Assume that the sequence $(w_k) \subset \mathbb{R}_+$ is such that $w_{k+1} \le w_k$ for all k. Then for every $k \in \mathbb{N}$ we have that*

$$M_P \le \sup_{u,v\ge 0} q_k(u,v) = M_{D_k}.$$

Proof Take a sequence $(u^k, v^k) \subset \mathbb{R}^m_{++} \times \mathbb{R}^r_{++}$ with coordinates u_i^k for $i = 1, \dots, m$ and v_j^k for $j = 1, \dots, r$ respectively. Assume that

$$\lim_{k\to\infty} u_i^k = \lim_{k\to\infty} v_j^k = +\infty, \tag{8}$$

for every (i, j). For each k, define (for notational simplicity),

$$x_k := x^\sharp(u^k, v^k, w_k) \in T(u^k, v^k, w_k),$$

where the set T is as in Definition 3.4. Corollaries 3.1 and 3.2 together with Lemma 3.4 imply

$$\begin{aligned} &F_L + \textstyle\sum_{i=1}^m u_i^k \eta(h_i(x_k), w_k) + \sum_{j=1}^r v_j^k\, \gamma(g_j(x_k), w_k) \le \\ &f(x_k) + \textstyle\sum_{i=1}^m u_i^k \eta(h_i(x_k), w_k) + \sum_{j=1}^r v_j^k\, \gamma(g_j(x_k), w_k) = q_k(u^k, v^k) \le M_P \end{aligned} \tag{9}$$

Inequality (9) and (8) yield

$$\lim_{k\to\infty} \eta\left(h_i(x_k), w_k\right) = \lim_{k\to\infty} \gamma\left(g_j(x_k), w_k\right) = 0,$$

for every i, j. By Lemma 2.2, we deduce that $|h(x_k)| \to 0$ and $[g(x_k)]_+ \to 0$ as $k \to \infty$.

Compactness of X implies that the sequence (x_k) contains a convergent subsequence $(x_k)_{k\in\mathcal{K}}$ with its limit x in X. Continuity of h and g then yields $h(x) = 0$ and $g(x) \le 0$, and thus $x \in X_0$. Lower semi-continuity of f on X implies

$$\liminf_{k\to\infty,\, k\in\mathcal{K}} f(x_k) \ge f(x) \ge M_P.$$

The inequality above and the definition of x_k yield

$$q_k(u^k, v^k) = L_k(x_k, u^k, v^k) \ge f(x_k) \qquad \forall k \in \mathcal{K},$$

which gives

$$M_{D_k} = \sup_{u,v\ge 0} q_k(u, v) \ge q_k(u^k, v^k) \ge f(x_k).$$

Hence,

$$M_{D_k} = \sup_{u,v\ge 0} q_k(u, v) \ge \liminf_{k\to\infty,\, k\in\mathcal{K}} f(x_k) \ge f(x) \ge M_P,$$

as claimed. □

Lemmas 3.4 and 3.6 imply that, asymptotically, there is zero duality gap. This is formally stated in the corollary below.

Corollary 3.4 *Assume that the sequence $(w_k) \subset \mathbb{R}_+$ is such that $w_{k+1} \le w_k$ for all k. Then,*

$$M_P = \lim_{k\to\infty} M_{D_k} = \sup_{k\in\mathbb{N}} M_{D_k}.$$

Proof The first equality follows directly from Corollary 3.3 and Lemma 3.6, while the second equality follows from Lemma 3.5(b). □

3.1 Properties of the dual solution set

Definition 3.5 *Fix $j \in \{1,\dots,p\}$ and $z \in \mathbb{R}^p$. Denote by $z_{-j} \in \mathbb{R}^{p-1}$ the vector obtained from z by extracting the coordinate j. Namely, for every $j \in \{1,\dots,p\}$, $z_{-j} := (z_1,\dots,z_{j-1},z_{j+1},\dots,z_p)$. For $j \in \{1,\dots,m\}$ and the Lagrangian L_w as in Definition 2.2, consider the function*

$$\hat{L}_w[x,u_{-j},v](\cdot) := L_w(x,u_1,\dots,u_{j-1},u_{j+1},\dots,u_m,v)(\cdot) : \mathbb{R}_+ \to \mathbb{R},$$

obtained by fixing in L_w all variables with the exception of u_j. In a similar way, define $\hat{L}_w[x,u,v_{-l}](\cdot)$ for $l \in \{1,\dots,r\}$, obtained by fixing in L_w all variables with the exception of v_l.

Lemma 3.7 *For all $(x,u,v,w) \in X \times \mathbb{R}^m_{++} \times \mathbb{R}^r_{++} \times \mathbb{R}_{++}$, every $i \in \{1,\dots,m\}$ and every $j \in \{1,\dots,r\}$, the following hold.*

(i) $\hat{L}_w[x,u_{-i},v](\cdot)$ is a non-decreasing function on $\mathbb{R}_{++}$when x,v and u_{-i} are fixed.

(ii) $\hat{L}_w[x,u,v_{-j}](\cdot)$ is a non-decreasing function on $\mathbb{R}_{++}$ when x,u and v_{-j} are fixed.

(iii) If $x \notin X_0$, then $L_{(\cdot)}(x,u,v)$ is a strictly decreasing function on $\mathbb{R}_{++}$ when x,u and v are fixed.

Proof (i) For fixed x,v and w the Lagrangian $L_w(x,\cdot,v)$ is differentiable with respect to $u \in \mathbb{R}^m_{++}$. Direct calculation gives

$$\frac{\partial \hat{L}_w[x,u_{-i},v](\cdot)}{\partial u_i} = \eta(h_i(x),w) \geq 0.$$

The mean value theorem now yields, for every $t>0$, and some θ between u_i and u_i+t,

$$\hat{L}_w[x,u_{-i}](u_i+t) - \hat{L}_w[x,u_{-i}](u_i) = \frac{\partial \hat{L}_w[x,u_{-i}](\theta)}{\partial u_i} t \geq 0,$$

so the non-decreasing property is established. For proving (ii), use an identical argument for $\hat{L}_w[x,u,v_{-j}](\cdot)$, because also in this case the partial derivative is non-negative. For (iii), assume that $x \notin X_0$. In this case, there exists i_0 such that $|h_{i_0}(x)| > 0$ or j_0 such that $g_{j_0}(x) > 0$. In the first case, we have by Lemma 2.1(b) that $\dfrac{\partial \eta(h_{i_0}(x),w)}{\partial w} < 0$. In the second case, again by 2.1(b) we obtain $\dfrac{\partial \gamma(g_{j_0}(x),w)}{\partial w} < 0$. Altogether, we can write

$$\frac{\partial L_{(\cdot)}(x,u,v)}{\partial w} = \sum_{i=1}^m u_i \frac{\partial \eta(h_i(x),w)}{\partial w} + \sum_{j=1}^r v_j \frac{\partial \gamma(g_j(x),w)}{\partial w} < 0,$$

because at least one term is strictly negative and all others are non-positive due to Lemma 2.1(b) and the fact that $u,v>0$. □

Definition 3.6 *Define the set*

$$S(D) := \{(u,v,w) \in \mathbb{R}^m_+ \times \mathbb{R}^r_+ \times \mathbb{R}_+ \ : \ q_w(u,v) = M_P\},$$

of all possible dual solutions (considering the parameter w as a dual variable).

The following lemma, which is a direct consequence of the previous one, shows an important property of the set of dual solutions.

Lemma 3.8 *If $(u, v) \geq (\hat{u}, \hat{v}) \geq 0$ and $0 \leq w \leq \hat{w}$ then $q_w(u, v) - q_{\hat{w}}(\hat{u}, \hat{v}) \geq 0$. In particular, if $(u^*, v^*, w^*) \in S(D)$, then $(u, v, w) \in S(D)$ whenever $(u, v) \geq (u^*, v^*)$ and $0 \leq w \leq w^*$.*

Proof For a fixed $w > 0$, we can write for $(u, v) \geq (\hat{u}, \hat{v})$

$$q_w(u, v) = \min_{x \in X} L_w(x, u, v) \geq \min_{x \in X} L_w(x, \hat{u}, \hat{v}) = q_w(\hat{u}, \hat{v}),$$

because $\eta, \gamma \geq 0$. On the other hand, if $w \leq \hat{w}$ by Lemma 3.7(iii) we have $L_w(x, \hat{u}, \hat{v}) \geq L_{\hat{w}}(x, \hat{u}, \hat{v})$, so we have

$$q_w(\hat{u}, \hat{v}) = \min_{x \in X} L_w(x, \hat{u}, \hat{v}) \geq \min_{x \in X} L_{\hat{w}}(x, \hat{u}, \hat{v}) = q_{\hat{w}}(\hat{u}, \hat{v}).$$

Altogether,

$$q_w(u, v) - q_{\hat{w}}(\hat{u}, \hat{v}) = [q_w(u, v) - q_w(\hat{u}, \hat{v})] + [q_w(\hat{u}, \hat{v}) - q_{\hat{w}}(\hat{u}, \hat{v})] \geq 0,$$

as wanted. The last statement in the lemma directly follows from the inequality above. Indeed, assume that $q_{w^*}(u^*, v^*) = M_P$ and take $(u, v) \geq (u^*, v^*)$ and $0 \leq w \leq w^*$, then

$$M_P \geq q_w(u, v) \geq q_{w^*}(u^*, v^*) = M_P,$$

where we used Lemma 3.4 in the leftmost inequality. □

Remark 3.3 Lemma 3.8 shows that if zero duality gap is achieved for a given (u^*, v^*, w^*), then zero duality gap holds for all (u, v, w) satisfying $(u, v) \geq (u^*, v^*)$ and $w \leq w^*$. From a theoretical point of view this means the coordinates of u, v, as well as $1/w$ can be increased with impunity. Practically speaking, doing so might make the task of calculating $q_w(u, v)$ via (5) increasingly difficult. By Lemma 3.5(c), when $w_k \downarrow 0$, L_{w_k} converges to the ℓ_1 penalty function. In many cases, for sufficiently large but finite (u, v), the minimizer(s) of $\varphi_{\ell_1}(\cdot, u, v)$ will lie in X_0. Alternatively, if the penalty parameters (u, v) become arbitrarily large, $L_w(\cdot, u, v)$ approaches the extreme barrier function given by $B(x) := f(x)$ when $x \in X_0$, and $B(x) := \infty$ otherwise. Given that X_0 is defined by equality and inequality constraints, the extreme barrier function is likely to be much more difficult to work with than the ℓ_1 exact penalty function. Hence our preferred strategy will be to increase (u, v) where necessary, and decrease w to zero in the limit.

Definition 3.7 *Fix $(u, v, w) \in \mathbb{R}^m_+ \times \mathbb{R}^r_+ \times \mathbb{R}_+$, and take $x \in T(u, v, w)$ (see Definition 3.4). Define the vectors:*

$$p_1(x, w) := \sum_{i=1}^{m} \eta(h_i(x), w) e^i \in \mathbb{R}^m_+, \quad p_2(x, w) := \sum_{j=1}^{r} \gamma(g_j(x), w) \hat{e}^j \in \mathbb{R}^r_+, \tag{10}$$

where e^i is the i-th canonical vector in $\mathbb{R}^m$, and $\hat{e}^j$ is the j-th canonical vector in $\mathbb{R}^r$. Write $p(x, w) := (p_1(x, w), p_2(x, w))$.

Remark 3.4 It follows directly from the definition of η and γ that

$$p(x,w)=0 \Longleftrightarrow x\in X_0,$$

for every $w\geq 0$.

Definition 3.8 *Recall that, for $r\geq 0$ and a given concave function $c:\mathbb{R}^p\to\mathbb{R}_{-\infty}$, the r-* supergradient of c at $z\in\mathrm{dom}(c)=\{z'\in\mathbb{R}^p\,:\,c(z')>-\infty\}$, *denoted as $\partial_r c(z)$, is the set*

$$\partial_r c(z):=\{\xi\in\mathbb{R}^p\,:\,c(z')\leq c(z)+\langle z'-z,\xi\rangle+r,\,\forall z'\in\mathbb{R}^p\}.$$

Proposition 3.2 *Fix $(u,v,w)\in\mathbb{R}^m_+\times\mathbb{R}^r_+\times\mathbb{R}_+$ and $x\in T(u,v,w)$. The following properties hold. Write $z:=(u,v)$ and fix $\pi_0\in\mathbb{R}^{m+r}_+$. Take $p(x,w)$ as in Definition 3.7. Then, we have that*

$$p(x,w)+\pi_0\in\partial_r q_w(z),$$

where $r\geq\langle z,\pi_0\rangle$. In particular, $p(x,w)\in\partial q_w(z)$.

Proof The proof is straightforward. Indeed, let $z':=(u',v')\in\mathbb{R}^m_+\times\mathbb{R}^r_+$ so we can write

$$\begin{aligned}
q_w(z') &\leq f(x)+\textstyle\sum_{i=1}^m u_i'\,\eta(h_i(x),w)+\sum_{j=1}^r v_j'\gamma(g_j(x),w)\\
&\leq f(x)+\textstyle\sum_{i=1}^m u_i'\,(\eta(h_i(x),w)+\pi_{0,i})+\sum_{j=1}^r v_j'\,(\gamma(g_j(x),w)+\pi_{0,j})\\
&= f(x)+\textstyle\sum_{i=1}^m(u_i'-u_i)\,(\eta(h_i(x),w)+\pi_{0,i})+\sum_{j=1}^r(v_j'-v_j)\,(\gamma(g_j(x),w)+\pi_{0,j})\\
&\qquad+\textstyle\sum_{i=1}^m u_i\,(\eta(h_i(x),w)+\pi_{0,i})+\sum_{j=1}^r v_j(\gamma(g_j(x),w)+\pi_{0,j})\\
&= f(x)+\textstyle\sum_{i=1}^m u_i\,\eta(h_i(x),w)+\sum_{j=1}^r v_j\gamma(g_j(x),w)\\
&\qquad+\textstyle\sum_{i=1}^m(u_i'-u_i)\,(\eta(h_i(x),w)+\pi_{0,i})\\
&\qquad+\textstyle\sum_{j=1}^r(v_j'-v_j)\,(\gamma(g_j(x),w)+\pi_{0,j})+\langle z,\pi_0\rangle\\
&= q_w(z)+\langle z'-z,p(x,w)+\pi_0\rangle+\langle z,\pi_0\rangle\\
&\leq q_w(z)+\langle z'-z,p(x,w)+\pi_0\rangle+r,
\end{aligned}$$

where the first inequality follows from the definition of q_w, the second inequality holds because $\pi_0\geq 0$, and the last equality follows from the definition of x. The last inequality above implies that $p(x,w)+\pi_0\in\partial_r q_w(z)$. The last statement follows by taking $\pi_0:=0$ and $r:=0$. □

Proposition 3.3 *Fix $(u,v,w)\in\mathbb{R}^m_+\times\mathbb{R}^r_+\times\mathbb{R}_+$. The following statements are equivalent.*

(a) $x\in T(u,v,w)\cap X_0$.

(b) $x\in S^*$ *and* $(u,v,w)\in S(D)$.

Proof If $x \in T(u,v,w) \cap X_0$, we can write

$$M_P \geq q_w(u,v) = L_w(x,u,v) = f(x) \geq M_P,$$

where we used Lemma 3.4 in the first inequality, the definition of x in the first equality, and the fact that x is feasible in the last equality and in the rightmost inequality. This shows that $x \in S^*$ and $(u,v,w) \in S(D)$, which is (b). Call $z := (u,v)$. If (b) holds with $(z,w) \in S(D)$, then $x \in X_0$, and $f(x) = M_P$ so we can write

$$\inf_{x' \in X} L_w(x',z) = q_w(z) = M_P = f(x) + \langle p(x,w), z \rangle,$$

where we used the fact that $(z,w) \in S(D)$ in the second equality, and Remark 3.4 in the third equality. The above expression yields $x \in T(z,w)$. Since we have $x \in S^* \subset X_0$, we deduce that $x \in T(z,w) \cap X_0$. □

4 A Primal–Dual Subgradient Algorithm

Note that Proposition 3.2 provides a search direction for the updates of the dual variables that is tailored to the algorithm's progress, while Proposition 3.3 provides a stopping criteria.

Definition 4.9 *Let $(u^k, v^k, w_k) \in \mathbb{R}^m_{++} \times \mathbb{R}^r_{++} \times \mathbb{R}_{++}$ be given, and set $x_k \in T(u^k, v^k, w_k)$. Consider the vector $p^k := p(x_k, w_k) = (p_1(x_k, w_k), p_2(x_k, w_k)) \in \mathbb{R}^m_+ \times \mathbb{R}^r_+$ as defined in (10). Given $z^k := (u^k, v^k)$, define*

$$z^{k+1} := z^k + s_k p^k \in \mathbb{R}^m_{++} \times \mathbb{R}^r_{++},$$

where $s_k > 0$. The vector $z^{k+1} = (u^{k+1}, v^{k+1})$ is called a subgradient step *from z^k. The sequence (z^k) generated in this way is called the* dual sequence*. A sequence (x_k) generated by minimizing $L_k(\cdot, u^k, v^k)$ over the set X is called the* primal sequence*.*

4.1 A Primal–Dual Algorithm

Next an algorithm that performs a subgradient step for each q_k is described.

Step 1 Set $k = 1$. Select $(z_1, w_1) \in \mathbb{R}^m_{++} \times \mathbb{R}^r_{++} \times \mathbb{R}_{++}$.

Step 2 Find a global minimizer of $L_k(\cdot, z^k)$ over X, label it x_k and compute p^k as in (10). If $p^k = 0$, STOP. Otherwise, go to Step 3.

Step 3 Set $z^{k+1} = z^k + s_k\,(p^k + d^k)$, where $s_k > 0$ and $d^k \geq 0$. Set $w_{k+1} \leq w_k$, increment k and go to Step 2.

Remark 4.5 If the algorithm stops in Step 2 for a certain k, then by Proposition 3.3, $x_k \in S^*$ and $(z^k, w_k) \in S(D)$. This fact justifies the stopping criteria.

4.2 Convergence Analysis

We first show some simple properties of our algorithm.

Lemma 4.9 *The following properties hold for the sequences (z^k) and (x_k).*

(i) The sequence (z^k) converges to some $z \in \mathbb{R}^m_{++} \times \mathbb{R}^r_{++}$ if and only if $\sum_{k=1}^{\infty} s_k(p^k + d^k) < \infty$. This situation happens if and only if (z^k) is bounded.

(ii) Write $r_k := \langle z^k, d^k\rangle \geq 0$. The search direction $\Delta_k := p^k + d^k \in \partial_{r_k} q_k(z^k)$.

(iii) The sequence $(q_k(z^k))$ is non-decreasing.

(iv) Assume that $d^k_i \geq \hat{d} > 0$ for all $i = 1, \dots, m+r$ and all k. If for some k we have $(z^k, w_k) \in S(D)$ then either $p^k = 0$ or $p^{k+1} = 0$. Consequently, if for some k we have $(z^k, w_k) \in S(D)$ then the generated sequence is finite.

(v) Take $\bar{d} > \hat{d} > 0$ and assume that $d^k_i \in [\hat{d}, \bar{d}\,]$ for all $i = 1, \dots, m+r$, and all k. The sequence (z^k) is unbounded if and only if (z^k_i) is unbounded for all $i = 1, \dots, m+r$.

Proof Since (z^k) is an increasing sequence, its convergence is equivalent to its boundedness. This establishes the last statement in (i). Let us prove the first statement in (i). From Step 3 of the algorithm we have, for every $k \in \mathbb{N}$,

$$z^k - z^1 = \sum_{t=1}^{k-1} s_t\,(p^t + d^t).$$

Hence,

$$\lim_{k\to\infty} z^k = z^1 + \sum_{t=1}^{\infty} s_t\,(p^t + d^t), \tag{11}$$

which establishes (i). Part (ii) follows directly from Proposition 3.2 and the fact that $d^k \geq 0$. Part (iii) is a direct consequence of Lemma 3.8 and the fact that $z^{k+1} \geq z^k$ and $w_{k+1} \leq w_k$. Let us prove part (iv). Note that the last statement in (iv) follows directly from Step 2 of the algorithm. Let us show the first statement in (iv). Assume that for some k we have $(z^k, w_k) \in S(D)$. This means that $q_k(z^k) = M_P$. If $p^k = 0$ we are done, so assume that $p^{k+1} \neq 0 \neq p^k$. This means that we perform the step $k+1$. Using Proposition 3.2 for $\pi_0 := 0$, the supergradient inequality and the assumption $(z^k, w_k) \in S(D)$, we can write

$$M_P = q_k(z^k) \leq q_{k+1}(z^k) \leq q_{k+1}(z^{k+1}) + \langle z^k - z^{k+1}, p^{k+1}\rangle, \tag{12}$$

where $p^{k+1} = p(x_{k+1}, w_{k+1})$ with $x_{k+1} \in T(z^{k+1}, w_{k+1})$. Part (iii) of this lemma, Lemma 3.4 and the assumption $(z^k, w_k) \in S(D)$ yield

$$M_P = q_k(z^k) \leq q_{k+1}(z^{k+1}) \leq M_P,$$

so $q_{k+1}(z^{k+1}) = M_P$. Combine this fact with the supergradient inequality (12) and the definition of the subgradient step to write

$$\begin{aligned} M_P &\leq q_{k+1}(z^{k+1}) + \langle z^k - z^{k+1}, p^{k+1}\rangle, \\ &= M_P - s_k\langle p^k + d^k, p^{k+1}\rangle \\ &\leq M_P - s_k\langle p^k + \hat{d}, p^{k+1}\rangle < M_P, \end{aligned}$$

where the second inequality holds because $p^k + d^k \geq p^k + \hat{d}$ and the third inequality holds because $p^k + \hat{d} \geq \hat{d} > 0$ and $p^{k+1} \neq 0$ (recall that $p^k \geq 0$). The above expression entails a contradiction, so we must have $p^{k+1} = 0$ as claimed.

Let us prove (v). It is enough to prove the necessity, because the sufficiency is trivial. Assume that (z^k) is unbounded above. Since (z^k) is increasing this implies that there exists $l \in \{1, \ldots, m + r\}$ such that $z_l^k \uparrow \infty$. Using (11) for the coordinate l of the sequence, we can write

$$z_l^{k+1} = z_l^1 + \sum_{t=1}^{k} s_t\,(p_l^t + d_l^t) \leq z_l^1 + \sum_{t=1}^{k} s_t\,p_l^t + \bar{d}\sum_{t=1}^{k} s_t.$$

Since the function $p(\cdot,\cdot)$ is continuous over the compact set $X \times [0, w_1]$ there exists $\bar{P}$ such that $\|p(x', w')\|_\infty \leq \bar{P}$ for all $(x', w') \in X \times [0, w_1]$. Altogether,

$$z_l^{k+1} \leq z_l^1 + (\bar{P} + \bar{d})\sum_{t=1}^{k} s_t.$$

Since $z_l^k \uparrow \infty$, we must have $\sum_{t=1}^{\infty} s_t = \infty$. Now take any coordinate $i \neq l$. Using (11) for the coordinate i we obtain

$$z_i^{k+1} = z_i^1 + \sum_{t=1}^{k} s_t\,(p_i^t + d_i^t) \geq z_i^1 + \hat{d}\sum_{t=1}^{k} s_t.$$

We have just proved that the rightmost term tends to infinity, so we deduce that $z_i^k \uparrow \infty$, as wanted. □

Remark 4.6 By inspecting part (v) in the last result, we see that, when $0 < \hat{d} \leq d_i^k \leq \bar{d}$ for all $i = 1, \ldots, m + r$, and all k, boundedness of (z^k) is equivalent to $\sum_{t=1}^{\infty} s_t < \infty$.

4.3 Convergence under two choices of Step-size

In our analysis, we will consider the following step sizes.

(a) $s_k := \dfrac{1}{\|p^k\|_2}$, or

(b) $s_k := \|p^k\|_2$,

for all k. To avoid trivial situations, we assume that $X \supsetneq X_0$. Choice (a) above, together with the compactness of X, implies that s_k is bounded below by a positive constant. Choice (b) implies that s_k is bounded above by a positive constant. In our numerical experiments, we only use (a). Our main convergence result, which is Theorem 4.1, can also be established (with trivial modifications) for $s_k := s > 0$ constant.

Theorem 4.1 *Assume that the step-size s_k verifies (a) or (b) above, and assume that the sequence (w_k) is decreasing with limit $w := \lim_{k\to\infty} w_k$. The following properties hold for the sequences (z^k) and (x_k).*

(I) Every accumulation point of (x_k) is in S^* and $q_k \uparrow M_P$.

(II) If (z^k) is bounded, then its limit $z \in \mathbb{R}^m_{++} \times \mathbb{R}^r_{++}$ verifies $(z, w) \in S(D)$.

(III) If $d^k \geq \hat{d} > 0$ for all k, then $(q_k(z^k))$ is a strictly increasing sequence.

Proof Let us prove (I). The compactness of X implies the sequence (x_k) of iterates has one or more accumulation points in X. Let x^* be an arbitrary accumulation point. By replacing (x_k) with a proper subsequence of itself if necessary, let (x_k) converge to x^*. First, we show that $x^* \in X_0$ by contradiction. Assume $x^* \notin X_0$. This can happen in two possible ways. Either (a) there exists i such that $|h_i(x^*)| = 2H > 0$, or (b) there exists j such that $g_j(x^*) = 2H > 0$. We assume first that (a) holds. Case (b) can be dealt with similarly. If there exists i such that $|h_i(x^*)| = 2H > 0$, the continuity of h_i and the convergence of (x_k) to x^* imply

$$\exists K \in \mathbb{N} \quad \text{such that} \quad |h_i(x_k)| > H \quad \forall k > K.$$

This, together with Lemma 2.1(d) and $u_i^k > 0$ implies

$$u_i^k \eta(h_i(x_k), w_k) \geq u_i^k \eta(H, w_k) \quad \forall k > K.$$

The updating strategy for w means $w_k \leq w_1$ for all k. Lemma 2.1(b) shows that $\eta(H, \cdot)$ is a decreasing function (of w) for fixed H, so

$$u_i^k \eta(h_i(x_k), w_k) \geq u_i^k \eta(H, w_k) \geq u_i^k \eta(H, w_1) \quad \forall k > K.$$

This yields

$$q_k(u^k, v^k) = L_k(x_k, u^k, v^k) \geq F_L + u_i^k \eta(H, w_1) \quad \forall k > K,$$

where we used the definition of x_k in the equality, and Corollary 3.2 in the inequality. We claim that, in this situation, we must have $u_i^k \uparrow \infty$ as $k \to \infty$. Indeed, assume this is not the case. Since the sequence (u_i^k) is increasing and we assume it is not tending to infinity, there exists $u_i^* \geq 0$ such that $\lim_{k\to\infty} u_i^k = u_i^*$. Equality (11) restricted to the sequence (u_i^k) becomes

$$u_i^* = \lim_{k\to\infty} u_i^k = u_i^1 + \sum_{t=1}^{\infty} s_t \left(\eta(h_i(x_t), w_t) + d_i^t\right).$$

This implies that

$$\lim_{t\to\infty} s_t(\eta(h_i(x_t), w_t) + d_i^t) = 0. \tag{13}$$

We now consider two cases. In case 1, we assume that $s_t \not\to 0$. In case 2, we assume that $s_t \to 0$. If we are in case 1, then by (13) we must have $\lim_{t\to\infty} \eta(h_i(x_t), w_t) + d_i^t = 0$. Since $d^t \geq 0$ this yields $\lim_{t\to\infty} \eta(h_i(x_t), w_t) = 0$. By Lemma 2.2(a) we obtain that $\lim_{t\to\infty} |h_i(x_t)| = 0$. The latter contradicts our assumption that $|h_i(x^*)| = 2H > 0$. Hence this case cannot happen and we are left with case 2, i.e., when $s_t \to 0$. We claim in this case we cannot have s_t is as in (a). Indeed, if s_t is as in (a) and tends to 0 then we must have $\|p^t\|_2 \to \infty$. Since X is compact, the continuous function $\theta : X \times [0, w_1] \to \mathbb{R}_+$ defined by $\theta(\cdot,\cdot) := \|p(\cdot,\cdot)\|_2$ must be bounded above by a constant $\bar{\theta}$ on the compact set $X \times [0, w_1]$. Since $X \supsetneq X_0$, Remark 3.4 implies that $\bar{\theta} > 0$. Hence,

$$0 = \lim_{t\to\infty} s_t = \lim_{t\to\infty} 1/\theta(x_t, w_t) \geq 1/\bar{\theta} > 0,$$

a contradiction. So our claim is true and if $s_t \to 0$ we must have s_t is as in (b). Thus $\|p^t\|_2 \to 0$ which in particular again yields $\lim_{t\to\infty} \eta(h_i(x_t), w_t) = 0$. We again obtain a contradiction as in case 1.

The contradiction in both cases means that the claim on (u_i^k) is true and we must have $u_i^k \uparrow \infty$ as $k \to \infty$. Hence $F_L + u_i^k \eta(H, w_1) \to \infty$ as $k \to \infty$, contradicting the fact that $q_k(u^k, v^k) \le M_P$ (by Lemma 3.4). This shows that (a) cannot happen and hence for every x^* accumulation point of (x_k) we must have $h(x^*) = 0$.

The remaining case (b), in which there exists j such that $g_j(x^*) = 2H > 0$ is dealt with in exactly the same way, but using γ, (v^k), and Lemma 2.2(b) instead. So we deduce that every accumulation point belongs to X_0. Second, we show $f(x^*) = M_P$. The definition of M_P immediately gives $f(x^*) \ge M_P$. The definition of x_k and L_k imply that $f(x_k) \le q_k(u^k, v^k) \le M_P$ by Lemma 3.4. The lower semi-continuity of f on X then gives $f(x^*) \le M_P$, as required. We have shown that every accumulation point of (x_k) is in S^*. Let us show that $q_k(z^k) \uparrow M_P$. We know from Lemma 4.9(iii) that $(q_k(z^k))$ is non-decreasing. By Lemma 3.4, $q_k(z^k) \le M_P$. Altogether,

$$M_P = \lim_{k\to\infty} f(x_k) \le \lim_{k\to\infty} f(x_k) + \langle z^k, p^k\rangle = \lim_{k\to\infty} q_k(z^k) \le M_P,$$

where the first equality follows from part (I), the first inequality follows from the fact that $z^k \ge 0$ and $p^k \ge 0$, the second equality uses the definition of x_k, and the last inequality follows from Lemma 3.4. This proves that $q_k(z^k) \uparrow M_P$.

Let us prove (II). Note that (w_k) is a decreasing sequence which is bounded below by w. Part (I) shows that

$$q_k(z^k) \le q_{k+1}(z^{k+1}) \le M_P,$$

so it is an increasing sequence which is bounded above. By Lemma 3.4 and the definition of x_k, we can write

$$\lim_{k\to\infty} q_k(z^k) = \lim_{k\to\infty} f(x_k) + \langle z^k, p^k\rangle.$$

By part (I), we know that $\lim_{k\to\infty} f(x_k) = M_P$. Assume (z^k) is bounded and call its limit z. We can write

$$\lim_{k\to\infty} q_k(z^k) = M_P + \langle z, p(x^*, w)\rangle,$$

where $p(x^*, w) = \lim_{k\to\infty} p(x_k, w_k) = 0$ because $x^* \in S^*$. Altogether, we obtain

$$\lim_{k\to\infty} q_k(z^k) = M_P = q_w(z),$$

so $(z, w) \in S(D)$ as wanted.

Let us prove (III). We have two possibilities: either the algorithm stops at a finite value k_F, or it generates an infinite sequence. If it stops, by Remark 3.4 and Proposition 3.2, it does so at a primal–dual solution pair (x_{k_F}, z_{k_F}) such that $x_{k_F} \in S^*$ and $(z_{k_F}, w_{k_F}) \in S(D)$. Otherwise, we have an infinite sequence and by Lemma 4.9(iv) we must have $(z^k, w_k) \notin S(D)$ for all k. As we recalled in part (I), for a fixed x the functions $\eta(h_i(x), \cdot)$ and $\gamma(g_j(x), \cdot)$ are decreasing for all i, j. Since $w_k \ge w_{k+1}$, this implies that, for a fixed x, we must have $p(x, w_k) \le p(x, w_{k+1})$.

Using the latter fact and the definition of q_{k+1}, we can write

$$\begin{aligned}
q_{k+1}(z^{k+1}) &= \min_{x\in X} f(x) + \langle z^{k+1}, p(x, w_{k+1})\rangle \\
&\geq \min_{x\in X} f(x) + \langle z^{k+1}, p(x, w_k)\rangle \\
&= q_k(z^{k+1}) = f(\hat{x}_k) + \langle z^{k+1}, p(\hat{x}_k, w_k)\rangle \\
&= f(\hat{x}_k) + \langle z^k + s_k(p^k + d^k), p(\hat{x}_k, w_k)\rangle \\
&= f(\hat{x}_k) + \langle z^k, p(\hat{x}_k, w_k)\rangle + s_k\langle (p^k + d^k), p(\hat{x}_k, w_k)\rangle, \\
&\geq f(\hat{x}_k) + \langle z^k, p(\hat{x}_k, w_k)\rangle + s_k\langle d^k, p(\hat{x}_k, w_k)\rangle,
\end{aligned}$$

for $\hat{x}_k \in T(z^{k+1}, w_k)$. In the last inequality, we used the fact that $\langle p^k, p(\hat{x}_k, w_k)\rangle \geq 0$. We consider now two cases. Case 1: $\hat{x}_k \notin X_0$, and Case 2: $\hat{x}_k \in X_0$. In Case 1, by Remark 3.4 we must have $p(\hat{x}_k, w_k) \neq 0$. Using also the fact that $\hat{d} > 0$, in the expression above, we find

$$\begin{aligned}
q_{k+1}(z^{k+1}) &\geq f(\hat{x}_k) + \langle z^k, p(\hat{x}_k, w_k)\rangle + s_k\langle \hat{d}, p(\hat{x}_k, w_k)\rangle \\
&> f(\hat{x}_k) + \langle z^k, p(\hat{x}_k, w_k)\rangle \\
&\geq f(x_k) + \langle z^k, p(x_k, w_k)\rangle = q_k(z^k),
\end{aligned}$$

where we used the definition of x_k in the last inequality. Hence, $q_{k+1}(z^{k+1}) > q_k(z^k)$ in this case. Case 2: $\hat{x}_k \in X_0$. This implies that $\hat{x}_k \in X_0 \cap T(z^{k+1}, w_k)$ and hence $p(\hat{x}_k, w_k) = 0$. By Proposition 3.2 this means that $\hat{x}_k \in S^*$ and $(z^{k+1}, w_k) \in S(D)$. Altogether, from the above expression we deduce

$$q_{k+1}(z^{k+1}) \geq f(\hat{x}_k) + \langle z^k, p(\hat{x}_k, w_k)\rangle = M_P > q_k(z^k),$$

where the equality follows from the fact that $\hat{x}_k \in S^*$ and $p(\hat{x}_k, w_k) = 0$. The last inequality uses the fact that $(z^k, w_k) \notin S(D)$. Hence, also in this case we obtain strict increase of the sequence $(q_k(z^k))$. □

5 Numerical Implementation and Experiments

5.1 Numerical implementation

We consider constrained problems of the form

$$\text{(Pc)} \left\{ \begin{array}{rl}
\displaystyle\min_{x\in X} & f(x) \\
\text{subject to} & h_i(x) = 0\,, \quad i = 1, \ldots, m\,, \\
& g_j(x) \leq 0\,, \quad j = 1, \ldots, r\,,
\end{array} \right.$$

where $X \subset \mathbb{R}^n$.

Recall the functions η and γ defined in (1)–(2), which can be used to model the equality and inequality constraints in (Pc), respectively. In what follows we present a numerical algorithm for implementing the primal–dual penalty method.

Algorithm 1 (The Primal–Dual Penalty Method)

Step 1 (*Initialization*) Set the penalty parameters $u_i^0 > 0$, $i = 1, \ldots, m$, and $v_j^0 > 0$, $j = 1, \ldots, r$. Set the smoothing parameter $w_0 = 1$, the tolerance $\varepsilon > 0$, and the index $k = 0$.

Step 2 (*Find a minimizer of the penalty function*) Solve

$$x_k \in \operatorname*{argmin}_{x\in X} f(x) + \sum_{i=1}^{m} u_i^k\, \eta(h_i(x), w_k) + \sum_{j=1}^{r} v_j^k\, \gamma(g_j(x), w_k)\,.$$

Step 3 (*Check stopping criterion*) If $\max\{\max_i |h_i(x_k)|, \max_j [g_j(x_k)]_+\} < \varepsilon$ then stop.

Step 4 (*Update penalty parameters*)

$$\begin{aligned} u_i^{k+1} &:= u_i^k + s_k\, \eta(h_i(x_k), w_k)\,, \quad i = 1, \ldots, m\,,\\ v_j^{k+1} &:= v_j^k + s_k\, \gamma(g_j(x_k), w_k)\,, \quad j = 1, \ldots, r\,, \end{aligned}$$

for some $s_k > 0$. Set $k = k + 1$, $w_{k+1} < w_k$, and go to Step 2.

Note that this implementation uses $d^k = 0$ for all k. The selection of a step size in primal–dual schemes has previously been studied in the context of the deflected subgradient method and its associated penalty methods in [6–11, 13–15]. We note the following three possible choices for s^k in Step 4 of Algorithm 1: one can choose

(i) $s^k = \dfrac{1}{\|p(x_k, w_k)\|_2}\,,$

(ii) $s^k = \|p(x_k, w_k)\|_2\,,$

(iii) $s^k = \dfrac{M_P - q_k(u^k, v^k)}{\|p(x_k, w_k)\|_2^2}\,,$

where $p(x_k, w_k)$ is as in Definition 3.7. The step size expressed in (iii) above is referred to as the *Polyak step size*. Although the Polyak step size is well-established to be effective, it has the disadvantage that it requires the optimal value of the problem, M_P, which usually is not available. In solving all of the test problems in Subsection 5.2 by employing Algorithm 1, we will use the step size in (i) for its ease of implementation as well as its relatively smaller size in early iterations. We point that, if the expression in (ii) is used, then, in the case of a large initial feasibility error, s^k could become large right at the beginning and cause numerical instabilities.

In solving the subproblem in Step 2 of Algorithm 1, we have paired up the optimization modelling language AMPL [22] and the well-reputed finite-dimensional optimization software

Knitro [2]. It should be noted that one might as well use, instead of Knitro, other popular optimization software, such as Algencan [1,3], which implements augmented Lagrangian techniques, or Ipopt [36], which implements an interior point method, or SNOPT [25], which implements a sequential quadratic programming algorithm.

The AMPL–Knitro suite was run on a 13-inch 2018 model MacBook Pro, with the operating system macOS Mojave (version 10.14.6), the processor 2.7 GHz Intel Core i7 and the memory 16 GB 2133 MHz LPDDR3. Within our AMPL codes, we have used the Knitro options `alg=0`, `maxit=5000`, `feastol=1e-8` and `opttol=1e-8`. In Step 3 of Algorithm 1, we have taken $\varepsilon = 10^{-7}$, for all of the test problems that we studied in Subsection 5.2. The CPU times are reported by using the values of the AMPL built-in parameter `_total_solve_time`.

5.2 Numerical experiments

In the numerical experiments with the three test problems presented in the subsequent subsections, we show that the primal–dual penalty method is viable, i.e., it is successful in solving a variety of equality- and inequality-constrained nonconvex optimization problems. Moreover, we observe that in the presence of multiple local minima, the primal–dual penalty method can be useful in finding deeper minima. We demonstrate that in some cases even the global minimum itself can be obtained more often than the case when using a local optimization solver conventionally (on its own), as observed with the simple example studied in Section 5.2.1.

We make comparisons with the case when Problem (Pc) is modelled directly within AMPL and paired up with Knitro in order to solve Problem (Pc), in the conventional/usual way. We refer to this case especially in the tables we present as "Knitro on its own." However, when we use Algorithm 1 and effectively implement the primal–dual penalty method, also modelled in AMPL in the way it is instructed in Step 2 of Algorithm 1, we refer to this case in the tables as "Algorithm 1." We note that in the "Knitro on its own" case Knitro performs constrained optimization. However, in the case of Algorithm 1, Knitro either performs unconstrained optimization (when $X = \mathbb{R}^n$), or box-constrained optimization (when X is a box).

5.2.1 An equality-constrained problem

We consider a test problem from [29], Problem 79, Classification PPR-P1-5. It is a small-scale problem, with just five variables and three constraints.

$$
\text{(P1)}\quad\left\{\begin{array}{rl}
\min & (x_1-1)^2+(x_1-x_2)^2+(x_2-x_3)^2+(x_3-x_4)^4+(x_4-x_5)^4 \\[2mm]
\text{subject to} & x_1+x_2^2+x_3^3-2-3\sqrt{2}=0\,, \\[1mm]
 & x_2-x_3^2+x_4+2-2\sqrt{2}=0\,, \\[1mm]
 & x_1\,x_5-2=0\,.
\end{array}\right.
$$

For Algorithm 1, we have used $w^{k+1} = 1/(k+1)^6$. Via numerical experiments, we have identified six isolated local minimizers of Problem (P1). These local minimizers along with the locally optimal values are listed in Table 1. We refer to Solution 1 listed in the table as the global minimizer of Problem (P1). (Otherwise, we have no certificate to conclude that Solution 1 is indeed a global minimizer.)

	Solution 1	Solution 2	Solution 3	Solution 4	Solution 5	Solution 6
x^*	1.191127	2.717678	−0.766173	−1.246781	0.949471	−2.702207
	1.362603	2.033384	2.666726	2.422242	−2.266633	−2.989944
	1.472818	−0.847948	−0.468170	1.174983	0.537796	0.171917
	1.635017	−0.485941	−1.619116	−0.213229	3.384285	3.847927
	1.679081	0.735922	−2.610377	−1.604131	2.106436	−0.740136
$f(x^*)$	0.0787768	13.9668249	27.4520041	27.5219615	86.5275397	649.5048650
Knitro on its own	45%	12%	7%	11%	18%	7%
Algorithm 1	100%	0%	0%	0%	0%	0%

Table 1: Problem (P1) – Local optimal solutions and the percentage of the times Knitro on its own or Algorithm 1 finds a certain solution, after 30,000 runs of each approach with random initial guesses.

For both implementations of Knitro, i.e., on its own as well as in Step 2 of Algorithm 1, we have generated the coordinates of the initial guesses $x^{(0)}$ at random uniformly in the interval $[-4, 4]$, namely that as an initial guess we have chosen

$$x_i^{(0)} \in \mathcal{U}[-4, 4], \quad i = 1, \ldots, 5\,,$$

each time an implementation is run. In each run of Algorithm 1, we have set $u_i^0 = 0.3$ for $i = 1, 2, 3$, and used a random guess as above only for $x^{(0)}$. Otherwise, $x^{(k-1)}$ that was computed in Step 2 constituted an initial guess for the subsequent $x^{(k)}$, $k = 1, 2, \ldots$. We have run Knitro on its own, as well as Algorithm 1 with Knitro implemented in its Step 2, 30,000 times. Algorithm 1 has always found the global minimizer. On the other hand Knitro, implemented on its own, has found the global minimizer only 45% of the time.

While the Knitro-on its own approach has taken about 0.023 seconds for each run for solving Problem (P1) (averaged over 30,000 runs), Algorithm 1 has taken 0.12 seconds, which is about five times longer. However, it should be kept in mind that Algorithm 1 can be made slightly faster by solving the subproblems in Step 2 by imposing coarser feasibility and optimality tolerances initially, for example with 10^{-3} or 10^{-4}, and then refining them down to 10^{-8}, rather than keeping them the same at 10^{-8} in solving every subproblem using Knitro.

It should be noted that, if the aim is to find the global minimum, Knitro on its own will be “expected” to be run more than twice, given the percentage of the times, or the probability 0.45, for it to find the global solution. On the other hand, it will suffice running Algorithm 1 only once to get the global solution.

We have solved a small-scale problem in this subsection. As we will see in the next subsection, for the kissing number problem, which can have a large number of variables and constraints, Algorithm 1 still produces good quality solutions more often but also in much shorter CPU times than Knitro on its own.

5.2.2 Kissing number problem

A *kissing number* is defined as the maximum number of spheres of radius r that touch another given sphere of radius r, without overlapping. The *kissing number problem* seeks to find the maximum number κ_n of non-overlapping spheres of radius r in $\mathbb{R}^n$ that can simultaneously touch (kiss) a central sphere of the same radius. We present and formulate the mathematical model for this problem as in [14].

The kissing number problem, also referred to as touching spheres, hard spheres, or kissing balls problem, has been a challenge for researchers since the time of Newton—for a more detailed account, see [32]. The problem can be formulated as a nonconvex optimization problem, where the minimum pairwise distance between the centres of p spheres touching a central sphere in $\mathbb{R}^n$ is to be maximized:

$$\text{(PKN)}\ \left\{\begin{array}{rl} \max & \displaystyle\min_{j>i} \|y_i - y_j\| \\ \text{subject to} & \|y_k\| = 1\,,\quad k = 1,\ldots p\,. \end{array}\right.$$

Here, $y_k \in \mathbb{R}^n$ is the vector of coordinates of the kth sphere's center, and without loss of generality the radii of the spheres have been taken as $1/2$. The maximum p, for which the optimal value of Problem (PKN) is greater than (or equal to) one, is then nothing but κ_n. Problem (PKN) has in practice a large number of stationary points, which makes the task of finding κ_n even for relatively small dimensions quite difficult. For dimensions $n = 1, 2, 3$ and 4, it is known that $\kappa_n = 2, 6, 12$ and 24, respectively. On the other hand, currently only some lower and upper bounds of κ_n are known for dimensions $n = 5, 6$ and 7, namely

$$40 \le \kappa_5 \le 44\,,\quad 72 \le \kappa_6 \le 78 \quad\text{and}\quad 126 \le \kappa_7 \le 134\,.$$

Our aim is not to improve these bounds, as this is extremely difficult, if not impossible, given the resources allocated to the writing of the current paper. Problem (PKN) has been used as a test problem for various optimization techniques, with $p \le \kappa_n$; see e.g. [6, 14, 28, 32]. Our aim is to compare our proposed primal–dual penalty approach with the classical approach using Knitro on its own.

Problem (PKN) can be re-formulated, as in [14], as a smooth problem:

$$\text{(P2)}\ \left\{\begin{array}{rl} \max & \alpha \\ \text{subject to} & \|y_k\|^2 = 1\,,\quad k = 1,\ldots p\,, \\ & \|y_i - y_j\|^2 \ge \alpha^2\,,\quad i,j = 1,\ldots,p,\quad j > i\,, \end{array}\right.$$

where α is a new optimization variable. Problem (P2), as opposed to Problem (P1), not only has equality but also inequality constraints. It has $(n\,p+1)$ variables, p equality constraints and $(p\,(p-1)/2)$ inequality constraints.

We have generated initial guesses for y_is at random, uniformly in each run of each method, in a similar fashion as we did for the example in Section 5.2.1. Namely we have chosen

$$y_{i,j} \in \mathcal{U}[-2, 2]\,,\quad i = 1,\ldots,p\,,\quad j = 1,\ldots,n\,,$$

each time an implementation is run, 1,000 times for each method. We have also imposed the box constraints

$$-2 \le y_{i,j} \le 2\,,\quad i = 1,\ldots,p\,,\quad j = 1,\ldots,n\,,$$

in the implementation of both methods. For Algorithm 1, we set $u_i^0 = 0.05$ for $i = 1,\ldots,p$, $v_j^0 = 0.05$ for $j = 1,\ldots,(p\,(p-1)/2)$. We use the rule $w_{k+1} = 1/(k+1)^4$. Numerical results for Problem (P2) are summarised in Table 2. Furthermore, a histogram of the local optimal values of α are depicted in Figure 2.

For the first kissing number problem instance where $(n, p) = (5, 38)$, of the 1,000 locally optimal solutions found by Knitro on its own, only seven solutions had 38 spheres fit on/kiss a

Problem size				Max of min dist. betw. two spheres				
$\begin{bmatrix} n \\ p \end{bmatrix}$	# of var.	# of constr.	Approach	α^*_{min}	α^*_{ave}	α^*_{max}	Percentage $\alpha^* > 1$	CPU time [sec]
$\begin{bmatrix} 5 \\ 38 \end{bmatrix}$	191	741	Knitro on its own	0.9807810	0.9927489	1.0019176	0.7%	0.89
			Algorithm 1	0.9912749	0.9968095	1.0037513	16.7%	0.92
$\begin{bmatrix} 6 \\ 62 \end{bmatrix}$	373	1953	Knitro on its own	0.9804186	0.9890666	0.9961310	0.0%	8.8
			Algorithm 1	0.9875817	0.9938880	1.0041174	1.1%	4.7
$\begin{bmatrix} 7 \\ 92 \end{bmatrix}$	645	4278	Knitro on its own	0.9886343	0.9953140	0.9991757	0.0%	65
			Algorithm 1	0.9946906	0.9987890	1.0021734	13.2%	25

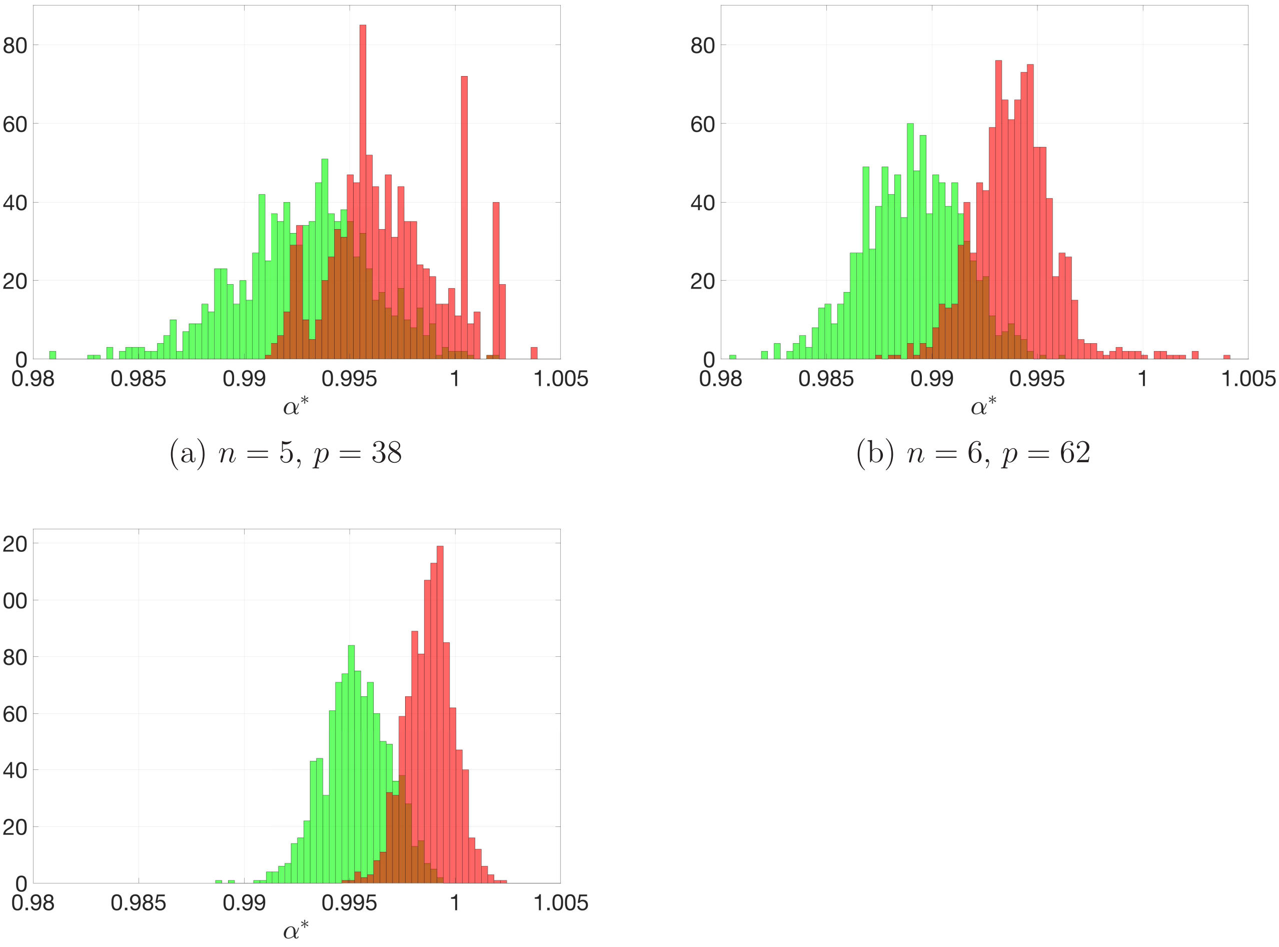
Table 2: Problem (P2) – Numerical results with Knitro on its own and Algorithm 1, after 1,000 locally optimal solutions (with values α^*) are found by each approach with random initial guesses.

central sphere without overlapping each other (i.e., with optimal $\alpha^* > 1$). On the other hand, the primal–dual penalty method, or Algorithm 1, has found 167 such solutions. A distribution of all local minima found by each method is depicted in Figure 2(a). It is clearly seen that the distribution of the maxima found by Algorithm 1 is skewed to the right, providing evidence to the method's ability/inclination to find higher maxima.

The ability of the proposed method in finding local optima closer to the global one is accentuated further in the kissing number instances $(n, p) = (6, 62)$ and $(n, p) = (7, 92)$, as can be seen in Figures 2(b) and 2(c), respectively. The distributions of the local maxima α^* are shifted further to the right compared to those found by Knitro on its own. Table 2 corroborates these observations by listing 11 solutions with $\alpha^* > 1$ for the second instance and 132 solutions with $\alpha^* > 1$ for the third, where respectively 62 and 92 spheres in $\mathbb{R}^6$ and $\mathbb{R}^7$ kiss a central ball without overlapping each other. For these instances, not even a single solution (with non-overlapping spheres) can be provided by Knitro on its own.

The quality of the solutions put aside, while the averaged CPU times of both methods are similar in the first instance, in the second and third instances our proposed method requires on the average roughly a half and a third of the CPU times Knitro on its own requires, respectively. We note that in the second and third instances of Problem (P2), the number of variables are 373 and 645, and the number of constraints 1953 and 4278, respectively.

Let us now interpret the results in Table 2 from a slightly different view point. Suppose that we want to estimate the CPU time required to find a *viable solution*, in which the spheres fit on the central sphere without overlapping. Then, to get such a solution for the instance $(n, p) = (5, 38)$, given the percentages (or probabilities) a solution with $\alpha^* > 1$ is found, it will be necessary to run Knitro on its own about 140 times, while it will be enough to run Algorithm 1 just about six times. In other words, while Knitro would be expected to find a solution with $\alpha^* > 1$ in about 125 seconds, Algorithm 1 would find a similar solution in about six seconds, for that instance. Moreover, for the second and third instances of Problem (P2), while Algorithm 1 is expected to find a viable solution respectively in about 430 and 1900 seconds (7.2 and 31.7 minutes), Knitro is expected to fail to find any.



(a) $n = 5$, $p = 38$

(b) $n = 6$, $p = 62$

(c) $n = 7$, $p = 92$

Figure 2: Problem (P2) – Histogram of local optima found by Knitro on its own (in green) and Algorithm 1 (in red), depicting $1,000$ locally optimal solutions (with values α^*) found by each approach. Overlaps are displayed in brown.

5.2.3 Markov–Dubins path

Markov–Dubins path is the shortest planar curve of constrained curvature joining two points with prescribed tangents. Although Markov posed the problem of finding such a shortest path and studied some special instances in 1889, it was Dubins who solved the problem fully in 1957—see [30] and the references therein. Suppose that a circular arc is represented by C and a straight line segment by S. Dubins' elegant result asserts that the sequence of concatenated arcs in such a shortest path can be of type CSC, CCC, or a subset thereof. Note that a circular arc C might be a "left turn" arc, denoted by L, "wrapping" a part of a circle in the counter-clockwise direction, or a "right turn" arc, denoted by R, wrapping a part of a circle in the clockwise direction. Examples to curves of these types can be found in Figure 3.

In [30], first, the problem is formulated as an optimal control problem, which is an infinite-dimensional optimization problem, and then, by using the maximum principle, it is shown that the problem can be reduced to a finite-dimensional optimization problem, denoted (Ps) below. The shortest path, of the types stated above, can be posed as a subsequence of the concatenation

	Soln 1	Soln 2	Soln 3	Soln 4	Soln 5	Soln 6	Soln 7
ξ^*	0.0000	0.0000	0.7096	0.8627	0.0000	1.6036	0.0000
	1.5822	1.7354	0.0000	0.3064	0.3818	1.7880	1.6781
	0.5914	0.0000	0.5914	0.0000	0.0000	0.0000	1.0093
	0.0000	0.3063	1.5594	1.7126	1.7880	0.3590	1.8526
	0.3376	0.4908	0.0000	0.0000	1.2317	0.0000	0.0000
ℓ^*	2.5113	2.5326	2.8603	2.8817	3.4015	3.7506	4.5401
Knitro on its own	8%	0%	3%	0%	40%	24%	25%
Algorithm 1	36%	0%	62%	0%	2%	0%	0%

Table 3: Problem (P3) – Local optimal solutions, and the percentage of the times Knitro on its own or Algorithm 1 finds a certain solution curve of length ℓ^*, after 20,000 runs of each approach with random initial guesses.

$L_{\xi_1}R_{\xi_2}S_{\xi_3}L_{\xi_4}R_{\xi_5}$, where ξ_j, $j = 1,\dots,5$, are the lengths of the respective arcs in the sequence. Let $\xi := (\xi_1,\dots,\xi_5)$. Suppose that the initial position and angle, i.e. the initial *oriented point*, is prescribed as $((x_0, y_0), \theta_0)$, and similarly the terminal oriented point is specified as $((x_f, y_f), \theta_f)$, where $(x_0, y_0), (x_f, y_f) \in \mathbb{R}^2$ and $\theta_0, \theta_f \in [-\pi, \pi]$. Suppose that the curvature is required to be equal to $a > 0$. Then a circular arc in a Markov–Dubins path will have the radius $1/a$. Problem (Ps) can now be written as in [30] as follows.

$$\text{(Ps)}\left\{\begin{array}{ll} \min & \ell = \displaystyle\sum_{j=1}^{5} \xi_j \\ \text{s.t.} & x_0 - x_f + \dfrac{1}{a}\,(-\sin\theta_0 + 2\,\sin\theta_1 - 2\,\sin\theta_2 + 2\,\sin\theta_4 - \sin\theta_f) + \xi_3\,\cos\theta_2 = 0\,, \\ & y_0 - y_f + \dfrac{1}{a}\,(\cos\theta_0 - 2\,\cos\theta_1 + 2\,\cos\theta_2 - 2\,\cos\theta_4 + \cos\theta_f) + \xi_3\,\sin\theta_2 = 0\,, \\ & \sin\theta_f = \sin\theta_5\,, \quad \cos\theta_f = \cos\theta_5\,, \\ & \xi_j \ge 0\,, \quad \text{for } j = 1,\dots,5\,, \end{array}\right.$$

where

$$\theta_1 = \theta_0 + a\,\xi_1\,, \qquad \theta_2 = \theta_1 - a\,\xi_2\,, \qquad \theta_4 = \theta_2 + a\,\xi_4\,, \qquad \theta_5 = \theta_4 - a\,\xi_5\,. \tag{14}$$

After substituting the relations in (14) into the constraints, problem (Ps) is expressed only in terms of the five variables, ξ_j, $j = 1,\dots,5$, and four equality constraints.

Next, consider the instance of the Markov–Dubins problem with $((x_0, y_0), \theta_0) = ((0, 0), -\pi/3)$, $((x_f, y_f), \theta_f) = ((0.4, 0.4), -\pi/6)$, and $a = 3$. In [30], seven stationary solutions are reported for this instance, which are listed in Table 3. Note that the solution subarcs of length zero must be excluded leaving at most three subarcs of the required types.

For this instance, we have solved Problem (Ps) with Knitro on its own as well as with P–D penalty method using Knitro, with initial guesses for ξ drawn from the uniform distribution $\mathcal{U}[0, \pi]$. To aid convergence, we have imposed the box constraints $2 \le \ell \le 4.6$ for both approaches. We have also imposed the following box constraints: (i) $0 \le \xi_j \le (2\,\pi/a) - \hat{\varepsilon}$, $j = 1, 2, 4, 5$, with $\hat{\varepsilon} = 10^{-5}$, in order to avoid stationary solutions with circular subarcs longer than a full circle and (ii) from the geometry, $0 \le \xi_3 \le \sqrt{(x_f - x_0)^2 + (y_f - y_0)^2} + 2/a + \hat{\varepsilon}$, for the length of the straight line segment to help convergence. In Algorithm 1, we have set the coordinates of $u_i^0 = 0.1$ for $i = 1,\dots,4$, and used the rule $w_{k+1} = 1/(k+1)^6$.

The results are summarized in Table 3: While Knitro on its own found 89% of the time Solutions 5, 6 and 7, which are of length greater than 3.4 units (see Figure 3(c)–(e)), Algorithm 1 found 98% of the time Solutions 1 and 3, which are of length less than 2.9 units (see Figure 3(a)–(b)). In particular, the shortest length solution in Figure 3(a) was found by Algorithm 1 36% of the time, while Knitro on its own found the same solution only 8% of the time.

In the numerical runs, the computational time Knitro on its own required was on the average 0.043 seconds per run, which is about eight times less than 0.34 seconds per run needed by Algorithm 1. It should however be pointed out that, in Step 2 of Algorithm 1, the optimality and feasibility tolerances were kept at 10^{-8} in each iteration. It is conceivable to think that lower tolerances, say at 10^{-3} or 10^{-4}, in the earlier iterations of Algorithm 1, are likely to speed up the computations with Algorithm 1 considerably.

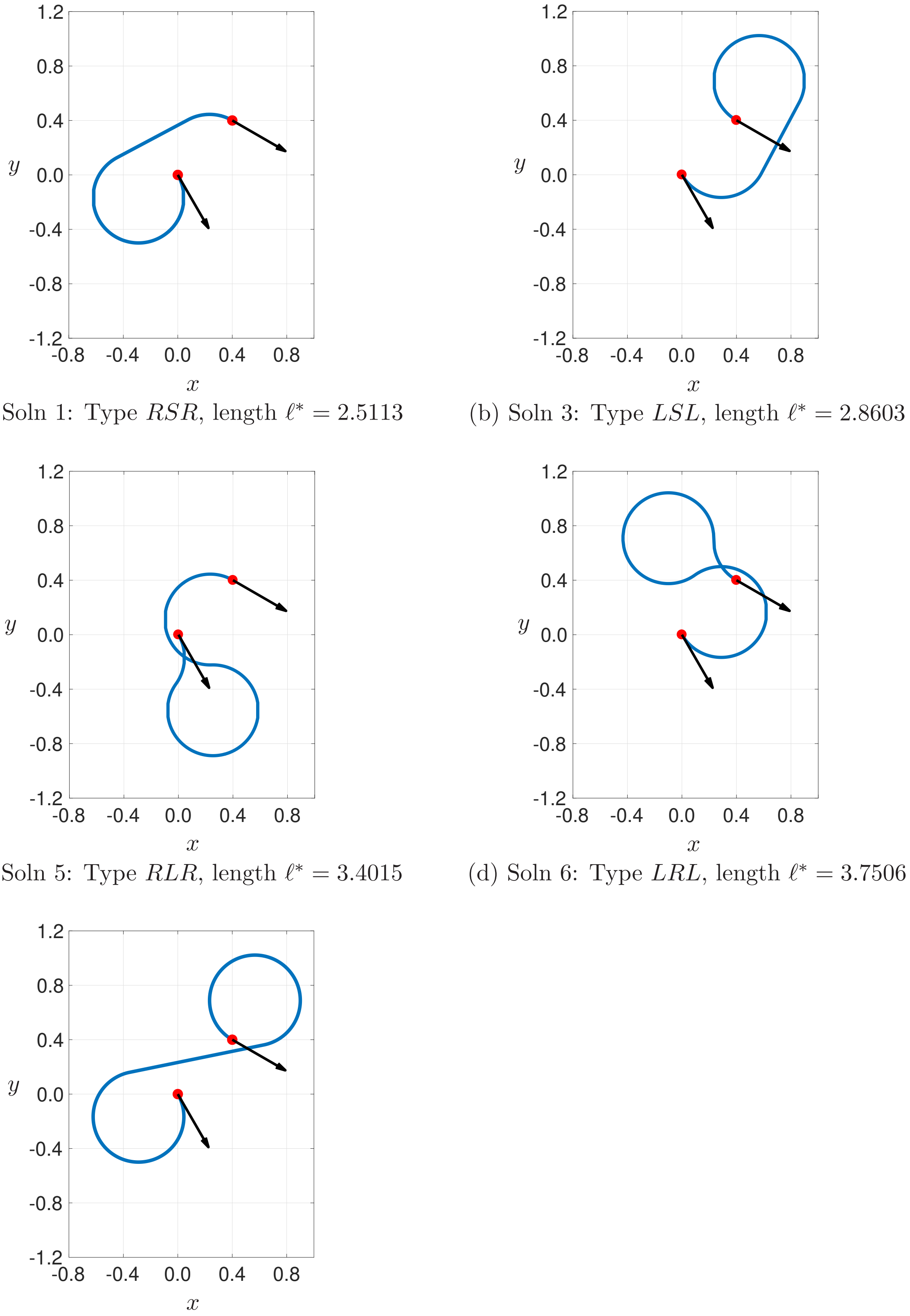
Despite the higher computational expense in this particular case, Algorithm 1 is superior in finding "better quality" solutions, given the fact that it finds curves with length less than 2.9 units 98% of the time, while Knitro on its own can find the same solutions only 11% of the time. This also amounts to saying that while we would expect to find a good quality solution (Solutions 1 and 3) by running Algorithm 1 just once, in 0.3 sec, even without making the economy with lower tolerances in earlier iterations mentioned above. Knitro on its own, on the other hand, is expected to be run about nine times (in 0.4 sec) to get the same quality solution.

As was observed in the case of the kissing number problem studied in Section 5.2.2, with the increasing size of the problem, in particular with the growing number of constraints, Algorithm 1 seems to become more time-efficient than Knitro on its own—see the CPU times in Table 2. Therefore, Algorithm 1 might prove even more useful in certain extensions of the Markov–Dubins problem, such as the problem of finding Markov-Dubins interpolating curves, where the shortest path is required to pass through a number of given intermediate points (see [31]), in turn increasing the size of the problem.

6 Conclusion

We study a general nonsmooth and nonconvex problem with equality and inequality constraints, defined in a compact metric space. A C^1-Lagrangian function based on rounding the weighted-ℓ_1 exact penalty function has been given, and its use in solving constrained optimization problems has been examined. The Lagrangian is indexed by a parameter w which governs the size of the rounding region. By taking a sequence (w_k) of decreasing positive parameters, we have obtained a sequence of augmented Lagrangians that can be seen as smooth approximations of the classical ℓ_1-penalty function. For the induced sequence of dual problems, we have shown that the sequence of dual optimal values is increasing and converges to the primal optimal value.

Furnished with this strong duality result, we have defined a primal–dual subgradient algorithm which reduces w_k in each iteration. In doing so, the Lagrangian approaches the (nonsmooth) weighted-ℓ_1 exact penalty function. We have shown that the primal sequence generated by the method accumulates at a solution of the primal problem. Moreover, if the dual variables happen to be bounded, they converge to a dual solution.



(a) Soln 1: Type RSR, length $\ell^* = 2.5113$

(b) Soln 3: Type LSL, length $\ell^* = 2.8603$

(c) Soln 5: Type RLR, length $\ell^* = 3.4015$

(d) Soln 6: Type LRL, length $\ell^* = 3.7506$

(e) Soln 7: Type RSL, length $\ell^* = 4.5401$

Figure 3: (a) Markov-Dubins path from $((0, 0), -\pi/3)$ to $((0.4, 0.4), -\pi/6)$ with a maximum curvature of 3 units and (b)–(e) some other stationary curves between the same oriented points. Solutions in (a)–(b) are found by Algorithm 1 98% of the time, and by Knitro on its own 11% of the time—see Table 3 for more details.

A specific implementation of the algorithm was numerically tested on challenging test problems, including the kissing number problem of various sizes, and the Markov-Dubins problem. These results have shown Algorithm 1 is viable and performs well. Compared with the conventional approach of using Knitro on its own, we have observed that Algorithm 1 generates near globally optimal solutions more often. We have also observed via the kissing number problem that it requires far less computational time when the number of constraints is rather large.

Algorithm 1 is in fact also applicable to nonsmooth problems. It would therefore be interesting to carry out experiments with nonsmooth and nonconvex constrained optimization problems, as part of future work.

References

[1] R. Andreani, E. G. Birgin, J. M. Martínez, and M. L. Schuverdt, On augmented Lagrangian methods with general lower-level constraints, *SIAM J. Optim.*, 18(4), 1286–1309, 2007.

[2] Artelys Knitro – Nonlinear optimization solver, https://www.artelys.com/knitro.

[3] E. G. Birgin, and J. M. Martínez, *Practical Augmented Lagrangian Methods for Constrained Optimization*, SIAM Publications, 2014.

[4] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, *Numerical Optimization*, Springer-Verlag, Berlin, 2003.

[5] R. S. Burachik, On asymptotic Lagrangian duality for nonsmooth optimization. *ANZIAM journal*, **58**, C93–C123, 2017.

[6] R. S. Burachik, W. P. Freire, and C. Y. Kaya, Interior epigraph directions method for nonsmooth and nonconvex optimization via generalized augmented Lagrangian duality. *Journal of Global Optimization*, **60**(3), 501–529, 2014.

[7] R. S. Burachik, R. N. Gasimov, N. A. Ismayilova, and C. Y. Kaya, On a modified subgradient algorithm for dual problems via sharp augmented Lagrangian. *Journal of Global Optimization*, **34**(1), 55–78, 2006.

[8] R. S. Burachik, A. N. Iusem, and J. G. Melo, The exact penalty map for nonsmooth and nonconvex optimization. *Optimization*, **64**(4): 717–738, 2015.

[9] R. S. Burachik, A. N. Iusem, and J. G. Melo, An inexact modified subgradient algorithm for primal-dual problems via augmented Lagrangians. *Journal of Optimization Theory and Applications*, **157**(1), 108–131, 2013.

[10] R. S. Burachik, A. N. Iusem, and J. G. Melo, A primal dual modified subgradient algorithm with sharp Lagrangian. *Journal of Global Optimization*, **46**(3), 347–361, 2010.

[11] R. S. Burachik, A. N. Iusem, and J. G. Melo, Duality and exact penalization for general augmented Lagrangians, *Journal of Optimization Theory and Applications*, **147**(1), 125–140, 2010.

[12] R. S. Burachik, and A. N. Iusem, *Set-Valued Mappings and Enlargements of Monotone Operators*, Springer-Verlag, Series Optimization and its Applications, Vol 8, New York, 2008.

[13] R. S. Burachik, and C. Y. Kaya, An update rule and a convergence result for a penalty function method. *Journal of Industrial Management and Optimization*, **3**(2), 381–398, 2007.

[14] R. S. Burachik, and C. Y. Kaya, An augmented penalty function method with penalty parameter updates for nonconvex optimization, *Nonlinear Analysis: Theory, Methods & Applications*, **75**(3), 1158–1167, 2012.

[15] R. S. Burachik, C. Y. Kaya, and M. Mammadov, An inexact modified subgradient algorithm for nonconvex optimization. *Computational Optimization and Applications*, **45**(1), 1–24, 2010.

[16] R. S. Burachik, and X. Q. Yang, Asymptotic strong duality, *Numerical Algebra, Control and Optimization*, **1**(3), 539–548, 2011.

[17] R. S. Burachik, and A. M. Rubinov, Abstract convexity and augmented Lagrangians. *SIAM Journal on Optimization*, **18**(2), 413–436, 2007.

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

[19] Dolgopolik, M. V., A unifying theory of exactness of linear penalty functions II: parametric penalty functions, *Optimization*, **66**(10), 1577–1622, 2017.

[20] Dolgopolik, M. V., A unified approach to the global exactness of penalty and augmented Lagrangian functions I: Parametric exactness, *Journal of Optimization Theory and Applications*, **176**(3), 728–744, 2018.

[21] Dolgopolik, M. V., A unified approach to the global exactness of penalty and augmented Lagrangian functions II: Extended exactness, *Journal of Optimization Theory and Applications*, **176**(3), 745–762, 2018.

[22] R. Fourer, D. M. Gay, and B. W. Kernighan, *AMPL: A Modeling Language for Mathematical Programming, Second Edition.* Brooks/Cole Publishing Company/Cengage Learning, 2003.

[23] R. N. Gasimov, Augmented Lagrangian duality and nondifferentiable optimization methods in nonconvex programming, *Journal of Global Optimization*, **24**(2), 187–203, 2002.

[24] R. N. Gasimov, and A. M. Rubinov, On augmented Lagrangians for optimization problems with a single constraint. *Journal of Global Optimization*, **28**(2), 153–173, 2004.

[25] P. E. Gill, W. Murray, and M. A. Saunders, SNOPT: an SQP algorithm for large-scale constrained optimization, *SIAM Rev.*, 47, 99–131, 2005.

[26] O. Güler, *Foundations of Optimization*, Graduate Texts in Mathematics 258, Springer-Verlag, Berlin, 2010.

[27] X. X. Huang, K. L. Teo and X. Q. Yang, Approximate augmented Lagrangian functions and nonlinear semidefinite programs, *Acta Mathematica Sinica*, English Series, **22**, 1283-–1296, 2006.

[28] E. S. Helou, S. A. Santos, L. E. Simoes, A new sequential optimality condition for constrained nonsmooth optimization. *SIAM Journal on Optimization*, 2020.

[29] W. Hock, and K. Schittkowski, *Test Examples for Nonlinear Programming Codes*, Lecture Notes in Economics and Mathematical Systems, Vol. 187, Springer-Verlag, Berlin Heidelberg, Germany, 1981.

[30] C. Y. Kaya, Markov–Dubins path via optimal control theory. *Comput. Optim. Appl.*, 68(3), 719–747, 2017.

[31] C. Y. Kaya, Markov–Dubins interpolating curves. *Comput. Optim. Appl.*, 73(2), 647–677, 2019.

[32] N. Krejić, J. M. Martínez, M. Mello, and E. A. Pilotta, Validation of an augmented Lagrangian algorithm with a Gauss-Newton Hessian approximation using a set of hard-spheres problems, *Comput. Optim. Appl.*, 16, 247–263, 2000.

[33] C. J. Price, Nonsmooth constrained optimization via rounded ℓ_1 penalty functions. *Optim. Methods Softw.*, appeared online: https://doi.org/10.1080/10556788.2020.1746961, 2020.

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

[35] A. M. Rubinov, X. X. Huang, and X. Q. Yang, The zero duality gap property and lower semicontinuity of the perturbation function, *Mathematics of Operations Research*, **27**, 775–791, 2002.

[36] A. Wächter, and L. T. Biegler, On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. *Math. Progr.*, 106, 25–57, 2006.

[37] C. Y. Wang, X. Q. Yang, and X. M. Yang, Nonlinear Augmented Lagrangian and Duality Theory, *Mathematics of Operations Research* **38**(4) 740–760, 2013.

[38] C. Y. Wang, X. Q. Yang, and X. M. Yang, A unified nonlinear Lagrangian approach to duality and optimal paths, *Journal of Optimization Theory and Applications* **135**, 85–100, 2007.