Primal-Dual Lagrangian Transformation method for Convex Optimization

Abstract. A class $\Psi$ of strongly concave and smooth functions $\psi: \mathbb{R} \to \mathbb{R}$ with specific properties is used to transform the terms of the classical Lagrangian associated with the constraints. The transformation is scaled by a positive vector of scaling parameters, one for each constraint.

Each step of the Lagrangian Transformation (LT) method alternates unconstrained minimization of LT in primal space with both Lagrange multipliers and scaling parameters update.

Our main focus is on the primal-dual LT method. We introduce the primal-dual LT method and show that under the standard second order optimality condition the method generates a primal-dual sequence that globally converges to the primal-dual solution with asymptotic quadratic rate.

Key words. Lagrangian transformation – duality – interior quadratic prox – primal-dual LT method

1. Introduction

We replace the unconstrained minimization of the LT in the primal space and the Lagrange multipliers update by solving the primal-dual (PD) system of equations. The application of Newton method to the PD system leads to the primal-dual LT (PDLT) method, which is our main focus.

By solving the PD system a given vector of Lagrange multipliers is mapped into a new primal-dual pair, while the scaling parameters vector remains fixed. The contractibility properties of the corresponding map are critical for both the convergence and the rate of convergence. To understand the conditions under which the corresponding map is contractive and to find the contractibility bounds one has to analyze the primal-dual maps (see [18]–[20]). It should be emphasized that neither the primal LT sequence nor the dual sequence generated by the Interior Quadratic Prox method provides sufficient information for this
analysis. Only the PD system, solving which is equivalent to one LT step, has all necessary components for such analysis. This reflects the important observation that for any multipliers method neither the primal nor the dual sequences control the computational process. The numerical process is governed rather by the PD system. The importance of the PD system associated with nonlinear rescaling methods has been recognized for quite some time (see [18],[19]).

Recently, the corresponding PD systems were used to develop globally convergent primal-dual nonlinear rescaling methods with an up to 1.5-Q superlinear rate (see [23], [24] and references therein).

In this paper we introduce a general primal-dual LT method. The method generates a globally convergent primal-dual sequence, which, under the standard second order optimality condition converges to the primal-dual solution with asymptotic quadratic rate. This is our main contribution.

In the initial phase, the PDLT works as the Newton LT method, i.e. Newton method for LT minimization followed by the Lagrange multipliers and the scaling parameters update. At some point, the Newton LT method automatically turns into the Newton method for the Lagrange system of equations corresponding to the active constraints. We would like to emphasize that the PDLT is not a mechanical combination of two different methods. It is rather a universal procedure, which computes at each step a primal-dual direction. Then, dependent on the reduction rate of the merit function, it uses either the primal direction to minimize the LT in primal space or the primal-dual direction to reduce the merit function value. In many aspects it is reminiscent of the Newton method with step-length for smooth unconstrained optimization.

This similarity is due to:

1) the special properties of $\psi \in \Psi$;

2) the structure of the LT method, in particular, the way in which the Lagrange multipliers and scaling parameters are updated at each step;

3) the fact that the Lagrange multipliers corresponding to the passive constraints converge to zero with at least quadratic rate;

4) the way in which we use the merit functions for updating the penalty parameter;

5) the way in which we employ the merit function for the Lagrangian regularization.
It should be emphasized that the PDLT is free from any stringent conditions on accepting the Newton step, which are typical for constrained optimization problems.

There are three important features that make Newton method for the primal-dual LT system free from such restrictions.

First, the LT is defined on the entire primal space.

Second, after a few Lagrange multipliers updates the terms of the LT corresponding to the passive constraints become negligibly small due to the at least quadratic convergence to zero of the Lagrange multipliers corresponding to the passive constraints. Therefore, on the one hand, these terms became irrelevant for finding the Newton direction. On the other hand, there is no need to enforce their nonnegativity.

Third, the LT multipliers method is, generally speaking, an exterior point method in the primal space. Therefore there is no need to enforce the nonnegativity of the slack variables for the active constraints as takes place in the Interior Point Methods (see [28]).

After a few Lagrange multipliers updates, the primal-dual LT direction becomes practically identical to the Newton direction for the Lagrange system of equations corresponding to the active constraints. This makes it possible to prove the asymptotic quadratic convergence of the primal-dual sequence.

The paper is organized as follows. In the next section we state the problem and introduce the basic assumptions on the input data. In section 3 we formulate the general LT method and describe some convergence results, which will be used later. In section 4 we introduce the Primal-Dual LT method and prove its local quadratic convergence under the standard second order optimality conditions. In section 5 we consider the globally convergent primal-dual LT method and show that the primal-dual sequence converges with an asymptotic quadratic rate. We conclude the paper with some remarks concerning future research.
2. Statement of the problem and basic assumptions

Let \( f : \mathbb{R}^n \to \mathbb{R}^1 \) be convex and all \( c_i : \mathbb{R}^n \to \mathbb{R}^1, i = 1, \ldots, q \), be concave and smooth functions. We consider the following convex optimization problem:

\[
x^* \in X^* = \text{Argmin}\{f(x) \mid x \in \Omega\},
\]

where \( \Omega = \{x : c_i(x) \geq 0, i = 1, \ldots, q\} \). We assume that:

**A:** The optimal set \( X^* \) is nonempty and bounded.

**B:** Slater’s condition holds, i.e. there exists \( \hat{x} : c_i(\hat{x}) > 0, i = 1, \ldots, q \).

Let us consider the Lagrangian \( L(x, \lambda) = f(x) - \sum_{i=1}^{q} \lambda_i c_i(x) \), the dual function

\[
d(\lambda) = \inf_{x \in \mathbb{R}^n} L(x, \lambda)
\]

and the dual problem

\[
\lambda \in L^* = \text{Argmax}\{d(\lambda) \mid \lambda \in \mathbb{R}^q_+\}. \quad (D)
\]

Due to assumption **B**, the Karush-Kuhn-Tucker (KKT) conditions hold true, i.e. there exists a vector \( \lambda^* = (\lambda_1^*, \ldots, \lambda_q^*) \in \mathbb{R}^q_+ \) such that

\[
\nabla_x L(x^*, \lambda^*) = \nabla f(x^*) - \sum_{i=1}^{q} \lambda_i^* \nabla c_i(x^*) = 0 \quad (1)
\]

and the complementary slackness conditions

\[
\lambda_i^* c_i(x^*) = 0, \quad i = 1, \ldots, q \quad (2)
\]

are satisfied.

We assume that the active constraint set at \( x^* \) is \( I^* = \{i : c_i(x^*) = 0\} = \{1, \ldots, r\}, r < n \). Let us consider the vector-functions \( c^T(x) = (c_1(x), \ldots, c_q(x)), e^T(r)_i(x) = (c_1(x), \ldots, c_r(x)) \) and their Jacobians \( \nabla c(x) = J(c(x)) \) and \( \nabla c(r)(x) = J(c(r)(x)) \).

The sufficient regularity condition

\[
\text{rank}\nabla c(r)(x^*) = r, \quad \lambda_i^* > 0, \quad i \in I^* \quad (3)
\]
together with the sufficient condition for the minimum $x^*$ to be isolated

$$\nabla^2_{xx}L(x^*, \lambda^*) y, y \geq \mu(y, y), \quad \mu > 0 \quad \forall y \neq 0: \nabla c_i(x^*) y = 0$$

(4)

comprise the standard second order optimality condition.

3. Lagrangian transformation method.

We consider a class $\Psi$ of twice continuously differentiable functions $\psi: (-\infty, \infty) \to \mathbb{R}$ with the following properties (see [25]):

1. $\psi(0) = 0$;

2. a) $\psi'(t) > 0$; b) $\psi'(0) = 1$; c) $\psi'(t) \leq at^{-1}$; d) $|\psi''(t)| \leq bt^{-2}$ \(\forall t \in [1, \infty), a > 0, b > 0\);

3. $-m^{-1} \leq \psi''(t) < 0 \quad \forall t \in (-\infty, \infty)$;

4. $\psi''(t) \leq -M^{-1} \quad \forall t \in (-\infty, 0]$ and $0 < m < M < \infty$;

5. $-\psi''(t) \geq 0.5t^{-1}\psi'(t) \quad \forall t \in [1, \infty)$.

The Lagrangian Transformation (LT) $L: \mathbb{R}^n \times \mathbb{R}^q_{++} \times \mathbb{R}^q_{++} \to \mathbb{R}$ we define by the following formula:

$$L(x, \lambda, k) = f(x) - \sum_{i=1}^{q} k^{-1}_i \psi(k_i \lambda_i c_i(x)),$$

(5)

and assume that $k_i \lambda_i^2 = k > 0, i = 1, \ldots, q$. Due to concavity of $\psi(t)$, convexity of $f(x)$ and concavity of $c_i(x), i = 1, \ldots, q$ the LT $L(x, \lambda, k)$ is a convex function in $x$ for any fixed $\lambda \in \mathbb{R}_{++}^q$ and $k \in \mathbb{R}_{++}^q$. Also due to property 4, assumption A and convexity of $f(x)$ and all $-c_i(x)$ for any given $\lambda \in \mathbb{R}_{++}^q$ and $k = (k_1, \ldots, k_q) \in \mathbb{R}_{++}^q$, the minimizer

$$\hat{x} \equiv \hat{x}(\lambda, k) = \text{argmin}\{L(x, \lambda, k) \mid x \in \mathbb{R}^n\}$$

(6)

exists. This can be proved using arguments similar to those in [1]. Due to the complementarity condition (2) and properties 1 and 2(b) for any KKT pair $(x^*, \lambda^*)$ and any $k \in \mathbb{R}_{++}^q$ we have

1) $L(x^*, \lambda^*, k) = f(x^*)$,

2) $\nabla L(x^*, \lambda^*, k) = \nabla x L(x^*, \lambda^*) = 0$,

3) $\nabla^2_{xx} L(x^*, \lambda^*, k) = \nabla^2_{xx} L(x^*, \lambda^*) + \psi''(0) \nabla c_i(x^*) K_{(r)} A_{(r)}^{-2} \nabla c_i(x^*)$,

where $K_{(r)} = \text{diag}(k_i)_{i=1} \Lambda_{(r)} = \text{diag}(\lambda_i)_{i=1}$. Therefore, for $K_{(r)} = k \Lambda_{(r)}^{-2}$ we have

$$\nabla^2_{xx} L(x^*, \lambda^*, k) = \nabla^2_{xx} L(x^*, \lambda^*) - k\psi''(0) \nabla c_i(x^*) \nabla c_i(x^*).$$

(7)
It follows from (7) and 3° that the LT Hessian $\nabla_{xx} \mathcal{L}(x^*, \lambda^*, k)$ is practically identical to the Hessian of the Quadratic Augmented Lagrangian (see [12], [26]–[28]) corresponding to the active constraints set.

Moreover, due to the Debreu lemma, (see [7]) under the standard second order optimality condition (3)–(4), the LT Hessian $\nabla^2_{xx} \mathcal{L}(x^*, \lambda^*, k)$ is positive definite for all $k \geq k_0$ if $k_0 > 0$ is large enough, whether $f(x)$ and all $-c_i(x)$ are convex or not. This is another important property of the LT $\mathcal{L}(x, \lambda, k)$, allowing us to extend some of the known results for nonconvex optimization problems.

The LT $\mathcal{L}(x, \lambda, k)$ along with the penalty parameter $k > 0$ has two extra tools: the vector of the Lagrange multipliers $\lambda \in \mathbb{R}_q^+$ and the scaling parameter vector $k \in \mathbb{R}_q^+$. However, using all three tools properly, it becomes possible to develop a globally convergent primal-dual LT method, which generates a primal-dual sequence that converges to the primal-dual solution with asymptotic quadratic rate. This is the main purpose of the paper.

First, we introduce the LT method.

Let $x^0 \in \mathbb{R}^n$, $\lambda^0 \in \mathbb{R}_q^+$, $k > 0$ and $k^0 = (k^0_i = k(\lambda^0_i)^{-2}, i = 1, \ldots, q)$. The LT multipliers method maps the triple $(x^s, \lambda^s, k^s)$ into $(x^{s+1}, \lambda^{s+1}, k^{s+1})$ defined by the following formulas:

$$x^{s+1} = \arg\min_{x \in \mathbb{R}^n} \{\mathcal{L}(x, \lambda^s, k^s)\}, \quad (8)$$

$$\lambda^{s+1}_i = \lambda^s_i \psi'(k^s_i \lambda^s_i c_i(x^{s+1})) = \lambda^s_i \psi'(k(\lambda^s_i)^{-1} c_i(x^{s+1})), \quad i = 1, \ldots, q, \quad (9)$$

$$k^{s+1}_i = k(\lambda^{s+1}_i)^{-2}, \quad i = 1, \ldots, q. \quad (10)$$

The minimizer $x^{s+1}$ in (8) exists for any $\lambda^* \in \mathbb{R}_q^+$ and any $k^* \in \mathbb{R}_q^+$ due to the boundness of $X^*$, convexity of $f$, concavity of $c_i$ and property 4° of $\psi \in \Psi$. It can be proven using considerations similar to those in [1], i.e. the LT method is well-defined.

Using arguments similar to those we used to prove Theorem 3.1 in [25], we can show that

1) the LT method (8)–(10) is equivalent to the Quadratic Prox for the dual problem in the rescaled dual space from step to step, i.e.

$$\lambda^{s+1} = \arg\max_{\lambda} \{d(\lambda) - \frac{1}{2} k^{-1} ||\lambda - \lambda^s||^2_{R_\lambda} | \lambda \in \mathbb{R}^q\}, \quad (11)$$

where $||\lambda||^2_{R_\lambda} = \lambda^T R_\lambda \lambda$ with $R_\lambda = (-\psi''_{[s]}(\cdot))^{-1}$, $\psi''_{[s]}(\cdot) = \text{diag}(\psi''_{[s]}(\cdot))_{i=1}^q$, $\psi''_{[s]}(\cdot) = \psi''(\theta^* \lambda^s_i c_i(x^{s+1}))$ and $0 < \theta^* < 1$, and
2) the LT method (8)–(10) is equivalent to the following Prox method

\[
\lambda^{s+1} = \arg \max \{ d(\lambda) - k^{-1} D(\lambda, \lambda^s) \mid \lambda \in \mathbb{R}^q \},
\]

where \(D : \mathbb{R}^q_+ \times \mathbb{R}^q_+ \to \mathbb{R}_+\), which is given by the formula \(D(u, v) = \sum_{i=1}^q v_i^2 \varphi(u_i/v_i)\), is the second order \(\varphi\)-divergence distance with the kernel \(\varphi = -\psi^* : \mathbb{R}^q_+ \to \mathbb{R}_+\), where \(\psi^*\) is the Fenchel transform of \(\psi\).

Let \(\Phi = \{ \varphi : \varphi = -\psi^*, \psi \in \Psi \}\) be the class of kernels, which corresponds to the class \(\Psi\) of transformations \(\psi\). Due to (20c) any kernel \(\varphi \in \Phi\) is a barrier type function, therefore the distance function \(D(u, v)\) is an interior distance and the method (12) is an Interior Prox method.

The method (11) reminds the classical quadratic prox (see [3], [6], [10], [14], [15], [28]). The difference is that the quadratic distance is computed in the rescaled from step to step dual space. Also \(\lambda^s \in \mathbb{R}^q_+\), therefore the method (12) is in fact an Interior Quadratic Prox. On the other hand, the LT method (8)–(10) is an exterior-point method in the primal space.

For a variety of particular kernels \(\varphi\), the Interior Prox method (12) has been studied in [2], [3], [4], [22], [30]. For the general class \(\Phi\) the Interior Prox method (12) has been analyzed in [25].

The equivalence of (8)–(10) to (11) and (12), along with properties of the dual kernel \(\varphi\) are critical for the convergence of the LT method (8)–(10).

The properties of the kernels \(\varphi\) induced by properties 1\textsuperscript{0}–5\textsuperscript{0} of the original transformation \(\psi \in \Psi\) were established in the following theorem.

**Theorem 1** [25] *The kernels \(\varphi \in \Phi\) are convex twice continuously differentiable and possess the following properties:

1) \(\varphi(s) \geq 0 \forall s \in (0, \infty)\) and \(\min_{s \geq 0} \varphi(s) = \varphi(1) = 0\);

2) a) \(\lim_{s \to 0^+} \varphi'(s) = -\infty\); b) \(\varphi'(s)\) is monotone increasing; c) \(\varphi'(1) = 0\);

3) a) \(\varphi''(s) \geq m > 0 \forall s \in (0, \infty)\); b) \(\varphi''(s) \leq M < \infty \forall s \in [1, \infty)\).

Unfortunately several well known transformations including exponential \(\psi_1(t) = 1 - e^{-t}\) [13], logarithmic \(\psi_2(t) = \ln(t + 1)\) and hyperbolic \(\psi_3(t) = t(t + 1)^{-1}\) MBF [18] as well as log-sigmoid \(\psi_4(t) = 2(\ln 2 + t - \ln(1 + e^t))\) and modified CHKS transformation \(\psi_5(t) = t - \sqrt{t^2 + 4 \eta} + 2 \sqrt{\eta}, \eta > 0\) ([22]) do not satisfy 1\textsuperscript{0}–5\textsuperscript{0}. Transformations \(\psi_1-\psi_3\) do not satisfy property 3\textsuperscript{0} \((m = 0)\), while for \(\psi_4\) and \(\psi_5\) the
property $4^0$ is violated ($M = \infty$). This can be fixed (see [25]) by using the quadratic extrapolation idea, which was first applied in [5] to modify the logarithmic MBF transformation $\psi_2$.

In [2] to guarantee 3a) the authors regularized the logarithmic MBF kernel $\psi_2(s) = s - \ln s - 1$. The regularized logarithmic MBF kernel $\bar{\psi}_2(s) = 0.5\nu(s - 1)^2 + \mu(s - \ln s - 1)$, $\mu > 0$, $\nu > 0$ has some very interesting properties allowing the authors to prove the global convergence of the dual sequence $\{\lambda^*\}$ generated by the Interior Prox method (12) to the dual solution $\lambda^*$ with $O((ks)^{-1})$ rate. The Fenchel transform $\bar{\phi}_2^*$ of the kernel $\bar{\phi}_2$ leads to the primal transformation $\bar{\psi}_2 = -\bar{\phi}_2^*$, which satisfies properties 1$^0$–5$^0$, therefore such a transformation, along with those given in [25], can be used in the framework of the primal-dual LT method, which we develop in sections 4 and 5. The regularization, which provides the strong convexity of the dual kernel, was an important element in the analysis given in [2]. Unfortunately, such a modification of the dual kernel in some instances leads to substantial difficulties when it comes to finding the primal transformation, which is a Fenchel conjugate for the dual kernel. For example, in case of the exponential transformation, it leads to solving a transcendental equation. Therefore, the results of [2] cannot be applied for the exponential multiplier method [30].

We would like to point out that in case of $k_1 = k_2 = \cdots = k_q = k$ there is a fundamental difference between the LT $L_1(x, \lambda, k) = f(x) - \sum_{i=1}^q k^{-1}\psi(k\lambda_i c_i(x))$ and the Lagrangian for the equivalent problem $L_2(x, \lambda, k) = f(x) - \sum k^{-1}\lambda_i \psi(kc_i(x))$, which is used in the Nonlinear Rescaling methods with a single scaling-penalty parameter (see [21], [23], [24]). The differences between $L_1(x, \lambda, k)$ and $L_2(x, \lambda, k)$ leads to substantial difference between corresponding multipliers methods and their dual equivalents. To the best of our knowledge, the multipliers method based on LT $L_1(x, \lambda, k)$ and its dual equivalent, which is an Interior Prox method, are far from being well understood so far. On the other hand, it seems that the LT $L_1(x, \lambda, k)$ is more suitable for semi-definite programming then the Lagrangian $L_2(x, \lambda, k)$ (see [8]).

In the remaining part of this section we just mention some convergence results, which will be used later. The results are taking place for any transformation $\psi \in \Psi$.

The LT method (8)–(10) requires finding an unconstrained minimizer $x^*$ at each step, which is unpractical. On the other hand, the convergence of the LT method (8)–(10) is the foundation for the primal-dual LT method. Therefore we just briefly mention the convergence results for the exact LT method (8)–(10).
For a bounded close set $Y \subset \mathbb{R}^n$ and $y_0 \notin Y$ we consider the distance $\rho(y_0, Y) = \min\{\|y_0 - y\| : y \in Y\}$ from $y_0$ to $Y$. We also introduce the set of indices $I^-(x) = \{i : c_i(x) < 0\}$. The following theorem can be proven the same way as Theorem 4.1 in [25].

**Theorem 2** If the assumptions A and B are satisfied then

1) the primal-dual sequence $\{x^s, \lambda^s\}$ is bounded, and the following estimations hold:

$$d(\lambda^{s+1}) - d(\lambda^s) \geq mk^{-1}\|\lambda^{s+1} - \lambda^s\|^2,$$
$$d(\lambda^{s+1}) - d(\lambda^s) \geq kmM^{-2} \sum_{i \in I^-(x^{s+1})} c_i^2(x^{s+1});$$

2) for the maximum constraint violation $v_s$ and the duality gap $d_s$ the following bounds hold:

$$v_s = O((ks)^{-0.5}), \quad d_s = O((ks)^{-0.5});$$

3) the primal-dual sequence $\{x^s, \lambda^s\}$ converges to the primal-dual solution in value, i.e.

$$f(x^*) = \lim_{s \to \infty} f(x^s) = \lim_{s \to \infty} d(\lambda^s) = d(\lambda^*),$$

and

$$\lim_{s \to \infty} \rho(x^s, X^*) = 0, \quad \lim_{s \to \infty} \rho(\lambda^s, L^*) = 0;$$

besides, any converging primal-dual subsequence has the primal-dual solution as a limit point.

**Remark 1.** It follows from Theorem 2 (see second bound in 1)) that for any $\tau < 0$ and any $i = 1, \ldots, q$ the inequality $c_i(x^{s+1}) \leq \tau$ is possible only for a finite number of steps. Therefore from some point on only original transformations $\psi_1^5$ can be used in the LT method. In fact, for $k > 0$ large enough the quadratic branch can be used just once. Therefore, the asymptotic analysis and the numerical performance of the LT method (8)–(10) and its dual equivalent (11) or (12) depends only on the properties of the original transformations $\psi_1^5$ and the corresponding original dual kernels $\varphi_1^5$. The transformations $\psi_1^5$ for $t \geq \tau$ are infinite time differentiable and so is the LT $\mathcal{L}(x, \lambda, k)$ if the input data has the corresponding property. This allows us to use the Newton method for solving the primal-dual system, which is equivalent to (8)–(9). We will concentrate on it in section 4.
Each second-order $\varphi$–divergence distance function $D_i(u,v) = \sum_{i=1}^q v_i^2 \varphi_i(u_i/v_i)$ leads to a corresponding Interior Prox method (12) for finding maximum of a concave function on $\mathbb{R}_+^q$.

Sometimes the origin of the function $d(\lambda)$ is irrelevant for the convergence analysis of the method (12) (see [3]). However, when $d(\lambda)$ is the dual function originated from the Lagrangian duality, such analysis can produce only limited results, because neither the primal nor the dual sequence controls the LT method. The LT method is controlled rather by the PD system, solving which is equivalent to LT step.

The PD system is defined by the primal-dual map similar to those we used to establish the rate of convergence of nonlinear rescaling methods (see [18]–[20]). Using the corresponding primal-dual map we can strengthen the convergence results of Theorem 2 by assuming the standard second order optimality conditions.

From (10) we have $\lim_{s \to \infty} k_i^s = k(\lambda_*^i)^{-2}$, $i = 1, \ldots, r$, i.e. the scaling parameters corresponding to the active constraints grow linearly with $k > 0$. Therefore the technique we used in [18] and [20] can be applied for the asymptotic analysis of the method (8)–(10).

For a given small enough $\delta > 0$, we define the following set:

$$D(\lambda^*, k, \delta) = \{(\lambda, k) \in \mathbb{R}_+^q \times \mathbb{R}_{++}^q : \lambda_i \geq \delta, |\lambda_i - \lambda_*^i| \leq \delta k, \quad i = 1, \ldots, r, \quad 0 \leq \lambda_i \leq k \delta, \quad k \geq k_0, \quad i = r + 1, \ldots, q; \quad k_i = k \lambda_*^{-2}, \quad i = 1, \ldots, q \}.$$}

The following theorem is similar to Theorem 6.2 in [20] and can be proven using the same technique.

**Theorem 3** If $f(x), c_i(x) \in C^2$ and the standard second order optimality conditions (3)–(4) hold, then there exists sufficiently small $\delta > 0$ and large enough $k_0 > 0$ that for any $(\lambda, k) \in D(\cdot)$ we have:

1) There exists $\hat{x} = \hat{x}(\lambda, k) = \arg\min\{L(x, \lambda, k) | x \in \mathbb{R}^n\}$ such that

$$\nabla_x L(\hat{x}, \lambda, k) = 0$$

and

$$\hat{\lambda}_i = \lambda_i \psi'(k(\lambda_*)^{-1} c_i(\hat{x})), \quad \hat{k}_i = k \hat{\lambda}_i^{-2}, \quad i = 1, \ldots, q.$$}

2) For the pair $(\hat{x}, \hat{\lambda})$ the bound

$$\max\{||\hat{x} - x^*||, ||\hat{\lambda} - \lambda_*||\} \leq c k^{-1} ||\lambda - \lambda_*||$$
holds and \( c > 0 \) is independent on \( k \geq k_0 \).

3) The LT \( \mathcal{L}(x, \lambda, k) \) is strongly convex in the neighborhood of \( \hat{x} \).

**Remark 2.** All results of Theorem 3 do not require convexity of \( f(x) \) and all \(-c_i(x), i = 1, \ldots, q\). Therefore the LT method can be used for solving nonconvex optimization problems as long as one can find an approximation for \( \hat{x}(\lambda, k) \) just for any \( k > k_0 \), where \( k_0 > 0 \) is large enough. Then LT requires at each step finding an approximation for the minimum of a strongly convex function at each step. To find an approximation for the first unconstrained minimizer for a wide class of nonconvex functions one can use the very interesting cubic regularization of Newton’s method recently developed in [16].

Finding \( x^{s+1} \) requires solving an unconstrained minimization problem (8), which is generally speaking, an infinite procedure. The following stopping criteria (see [20]) allows to replace \( x^{s+1} \) by an approximation \( \tilde{x}^{s+1} \), which can be found in a finite number of Newton steps by minimizing \( \mathcal{L}(x, \tilde{\lambda}^s, \tilde{k}^s) \) in \( x \in \mathbb{R}^n \). If \( \tilde{x}^{s+1} \) is used instead of \( x^{s+1} \) in the formula (9) for the Lagrange multipliers update, then bounds similar to those established in 2) of Theorem 3 remain true.

For a given \( \sigma > 0 \) let us consider the sequence \( \{ \tilde{x}^s, \tilde{\lambda}^s, \tilde{k}^s \} \) which is generated by the following formulas:

\[
\tilde{x}^{s+1}: \| \nabla \mathcal{L}(\tilde{x}^{s+1}, \tilde{\lambda}^s, \tilde{k}^s) \| \leq \sigma k^{-1} \| \Psi'(k(\tilde{\lambda}^s)^{-1} c(\tilde{x}^{s+1})) \lambda^s - \tilde{\lambda}^s \| ,
\]

\[
\tilde{\lambda}^{s+1} = \Psi'(k(\tilde{\lambda}^s)^{-1} c(\tilde{x}^{s+1})) \lambda^s ,
\]

where

\[
\Psi'(k(\tilde{\lambda}^s)^{-1} c(\tilde{x}^{s+1})) = \text{diag} \{ \psi'(k(\tilde{\lambda}^s_i)^{-1} c_i(\tilde{x}^{s+1})) \}_{i=1}^q
\]

and

\[
\tilde{k}^s = (\tilde{k}^s_i = k(\tilde{\lambda}^s_i)^{-2}) , \quad i = 1, \ldots, q.
\]

The following theorem can be proven the same way we proved Theorem 7.1 in [20].

**Theorem 4** If the standard second order optimality conditions (3)–(4) hold and the Hessians \( \nabla^2 f(x) \) and \( \nabla^2 c_i(x) \), \( i = 1, \ldots, m \) satisfy the Lipschitz conditions

\[
\| \nabla^2 f(x) - \nabla^2 f(y) \| \leq L_0 \| x - y \| , \quad \| \nabla^2 c_i(x) - \nabla^2 c_i(y) \| \leq L_i \| x - y \|
\]

(15)
then there is $k_0 > 0$ large enough such that for the primal-dual sequence $\{\bar{x}^s, \bar{\lambda}^s\}$ generated by the formulas (13) and (14) the following estimations hold true and $c > 0$ is independent of $k \geq k_0$ for $s \geq 0$:

$$
\|\bar{x}^{s+1} - x^*\| \leq c(1 + \sigma)k^{-1}\|\bar{x}^s - x^*\|, \quad \|\bar{\lambda}^{s+1} - \lambda^*\| \leq c(1 + \sigma)k^{-1}\|\bar{\lambda}^s - \lambda^*\|.
$$  (16)

To find an approximation $\bar{x}^{s+1}$ one can use Newton’s method with step-length for minimization $\mathcal{L}(x, \bar{\lambda}^s, k^s)$ in $x$. It requires generally speaking several Newton steps to find $\bar{x}^{s+1}$. Then we update the vector of Lagrange multipliers $\bar{\lambda}^s$ and the scaling parameters vector $\bar{k}^s$ using $\bar{x}^{s+1}$ instead of $x^{s+1}$ in (9) and $\bar{\lambda}^{s+1}$ instead of $\lambda^{s+1}$ in (10).

In the following section we develop a different approach. Instead of finding $\bar{x}^{s+1}$ and then updating the Lagrange multipliers we consider a primal-dual system, finding an approximate solution for which is equivalent to finding $\bar{x}^{s+1}$ and $\bar{\lambda}^{s+1}$. The Newton method for solving the primal-dual system, which is equivalent to one LT step (8)–(9), leads to the Primal-Dual LT method.

4. Local Primal-Dual LT method

In this section we describe the PDLT method and prove local quadratic rate of convergence of the primal-dual sequence to the primal-dual solution under the standard second order optimality condition. One step of the LT method (8)–(10) maps the given triple $(x, \lambda, k) \in \mathbb{R}^n \times \mathbb{R}_+^q \times \mathbb{R}_+^q$ into a triple $(\hat{x}, \hat{\lambda}, \hat{k}) \in \mathbb{R}^n \times \mathbb{R}_+^q \times \mathbb{R}_+^q$ by formulas

$$
\hat{x} : \nabla_x L(\hat{x}, \lambda, k) = \nabla f(\hat{x}) - \sum_{i=1}^q \lambda_i \psi'(k\lambda_i^{-1}c(\hat{x})) \frac{\partial}{\partial x} c_i(\hat{x}) = 0,
$$  (17)

$$
\hat{\lambda} : \hat{\lambda}_i = \lambda_i \psi'(k\lambda_i^{-1}c(\hat{x})), \quad i = 1, \ldots, q,
$$  (18)

$$
\hat{k} : \hat{k}_i = k\lambda_i^{-2}, \quad \ldots, q.
$$  (19)

By removing the scaling vector update formula (19) from the system (17)–(19), we obtain the primal-dual LT system

$$
\nabla_x L(\hat{x}, \hat{\lambda}) = \nabla f(\hat{x}) - \sum_{i=1}^q \hat{\lambda}_i \frac{\partial}{\partial x} c_i(\hat{x}) = 0,
$$  (20)

$$
\hat{\lambda} = \psi'(k\lambda^{-1}c(\hat{x})) \lambda,
$$  (21)
where $\Psi'(k\lambda^{-1}c(\bar{x})) = \text{diag}(\psi'(k_i\lambda^{-1}_i c_i(\bar{x})))_{i=1}^q$.

From the standard second order optimality condition (3)–(4) follows the uniqueness of $x^*$ and $\lambda^*$.

Also there is $\tau > 0$ that

\[ a) \min\{c_i(x^*) \mid r + 1 \leq i \leq q\} \geq \tau^* \text{ and } \min\{\lambda_i^* \mid 1 \leq i \leq r\} \geq \tau^*. \]

Therefore, due to (16), there is $k_0 > 0$ large enough that for any $k \geq k_0$ and $s \geq 1$

\[ a) \min\{c_i(x^*) \mid r + 1 \leq i \leq q\} \geq 0.5\tau^* \text{ and } \min\{\lambda_i^* \mid 1 \leq i \leq r\} \geq 0.5\tau^*. \] (22)

Using formula (14) and the property $2^{th}$ c) we have

\[ \lambda_i^{s+1} = \psi'(k(\lambda_i^*)^{-1}c_i(x^{s+1}))\hat{\lambda}_i^s \leq 2a(k\tau^*)^{-1}(\lambda_i^*)^2, \quad s \geq 1. \]

Hence for any fixed $k > \max\{k_0, 2a(\tau^*)^{-1}\}$ we have

\[ \hat{\lambda}_i^{s+1} \leq (\lambda_i^*)^2, \quad s \geq 1, \quad r + 1 \leq i \leq q. \]

So for a given accuracy $0 < \varepsilon << 1$, in at most $s = O(\ln \ln s^{-1})$ Lagrange multipliers updates, the Lagrange multipliers for the passive constraints will be of the order $a(\varepsilon^2)$. From this point onwards, terms related to the passive constraints will be automatically ignored in the further calculations. Therefore the primal-dual system (20)–(21) will be actually reduced to the following system for $\hat{x}$ and $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_r)$:

\[ \nabla_x L(\hat{x}, \hat{\lambda}) = \nabla f(\hat{x}) - \sum_{i=1}^r \hat{\lambda}_i \nabla c_i(\hat{x}) = 0, \] (23)

\[ \hat{\lambda}_i = \psi'(k(\lambda_i)^{-1}c_i(\bar{x}))\lambda_i, \quad i = 1, \ldots, r. \] (24)

To simplify notation, we use $L(x, \lambda)$ for the truncated Lagrangian i. e. $L(x, \lambda) = f(x) - \sum_{i=1}^r \lambda_i c_i(x)$ and $c(x)$ for the active constraints vector-function, i. e. $c^T(x) = (c_1(x), \ldots, c_r(x))$.

We use the vector norm $\|x\| = \max_{1 \leq i \leq n} |x_i|$ and the matrix $A : \mathbb{R}^n \to \mathbb{R}^n$ norm $\|A\| = \max_{1 \leq i \leq n} (\sum_{j=1}^n |a_{ij}|)$.

For a given $\varepsilon_0 > 0$ we define the $\varepsilon_0$-neighborhood $\Omega_{\varepsilon_0} = \{ y = (x, \lambda) \in \mathbb{R}^n \times \mathbb{R}^r_{++} : \|y - y^*\| \leq \varepsilon_0 \}$ of the primal-dual solution $y^* = (x^*, \lambda^*)$.

We will measure the distance between the current approximation $y = (x, \lambda)$ and the solution $y^*$ using the following merit function:

\[ \nu(y) = \nu(x, \lambda) = \max\{\|\nabla_x L(x, \lambda)\|, \min_{1 \leq i \leq q} c_i(x), \sum_{i=1}^q |\lambda_i| c_i(x)|, \min_{1 \leq i \leq r} \lambda_i\}. \]
assuming that the input data is properly normalized. It follows from the KKT conditions (1) and (2) that

$$\nu(x, \lambda) = 0 \iff x = x^*, \lambda = \lambda^*.$$  

Later we will use the following lemma.

**Lemma 1.** [24] If the standard second order optimality condition (3)–(4) and Lipschitz condition (15) are satisfied, then there exists $0 < m_0 < M_0 < \infty$ and $\varepsilon_0 > 0$ small enough that

$$m_0\|y - y^*\| \leq \nu(y) \leq M_0\|y - y^*\|, \quad \forall y \in \Omega_{\varepsilon_0}. \quad (25)$$

It follows from (25) that in the neighbourhood $\Omega_{\varepsilon_0}$ the merit function $\nu(y)$ is similar to $\|\nabla f(x)\|$ for the unconstrained optimization problem $\min \{f(x) \mid x \in \mathbb{R}^n\}$, when $f(x)$ is strongly convex and $\nabla f(x)$ satisfies the Lipschitz condition. The merit function $\nu(y)$ will be used

1) to update the penalty parameter $k > 0$;

2) to control accuracy at each step as well as for the overall stopping criteria;

3) to identify “small” and “large” Lagrange multipliers at each PDLT step;

4) to decide whether the primal or primal-dual direction has to be used at the current step.

First we consider the Newton method for solving system (23)–(24) and show its local quadratic convergence. To find the Newton direction $(\Delta x, \Delta \lambda)$ we have to linearize the system (23)–(24) at $y = (x, \lambda)$.

We start with system (24). Due to $3^0$ the inverse $\psi^{-1}$ exists. Therefore, using the identity $\psi^{-1} = \psi^*$ and keeping in mind $\varphi = -\psi^*$, we can rewrite (24) as follows:

$$c_i(\hat{x}) = k^{-1}\lambda_i\psi^{-1}(\hat{\lambda}_i/\lambda_i) = k^{-1}\lambda_i\psi^*(\hat{\lambda}_i/\lambda_i) = -k^{-1}\lambda_i\varphi'(\hat{\lambda}_i/\lambda_i).$$

Assuming $\hat{x} = x + \Delta x$ and $\hat{\lambda} = \lambda + \Delta \lambda$, keeping in mind $\varphi'(1) = 0$ and ignoring terms of the second and higher order we obtain

$$c_i(\hat{x}) = c_i(x) + \nabla c_i(x)\Delta x = -k^{-1}\lambda_i\varphi'((\lambda_i + \Delta \lambda_i)/\lambda_i)$$

$$= -k^{-1}\lambda_i\varphi'(1 + \Delta \lambda_i/\lambda_i) = -k^{-1}\varphi''(1)\Delta \lambda_i, \quad i = 1, \ldots, r,$$

or

$$c_i(x) + \nabla c_i(x)\Delta x + k^{-1}\varphi''(1)\Delta \lambda_i = 0, \quad i = 1, \ldots, r.$$
Now we linearize the system (23) at $y = (x, \lambda)$. We have

$$\nabla f(x) + \nabla^2 f(x) \Delta x - \sum_{i=1}^{r} (\lambda_i + \Delta \lambda_i)(\nabla c_i(x) + \nabla^2 c_i(x) \Delta x) = 0.$$  

Again, ignoring terms of the second and higher orders, we obtain the following linearization of the PD system (23)–(24):

$$\nabla^2_{xx} L(x, \lambda) \Delta x - \nabla c^T(x) \Delta \lambda = -\nabla_x L(x, \lambda),$$  

$$\nabla c(x) \Delta x + k^{-1} \varphi''(1) I_r \Delta \lambda = -c(x),$$  

where $I_r$ is the identical matrix in $R^r$ and $\nabla c(x) = J(c(x))$ is the Jacobian of the vector-function $c(x)$.

Let us introduce the matrix

$$N_k(x, \lambda) = N_k(y) = N_k(\cdot) = \begin{bmatrix} \nabla^2_{xx} L(\cdot) & -\nabla c^T(\cdot) \\ \nabla c(\cdot) & k^{-1} \varphi''(1) I_r \end{bmatrix}.$$  

Then the system (26)–(27) can be written as follows:

$$N_k(\cdot) \begin{bmatrix} \Delta x \\ \Delta \lambda \end{bmatrix} = \begin{bmatrix} -\nabla_x L(\cdot) \\ -c(\cdot) \end{bmatrix}.$$  

The local PDLT method consists of the following operations:

1. Find the primal-dual Newton direction $\Delta y = (\Delta x, \Delta \lambda)$ from the system

$$N_k(\cdot) \begin{bmatrix} \Delta x \\ \Delta \lambda \end{bmatrix} = \begin{bmatrix} -\nabla_x L(\cdot) \\ -c(\cdot) \end{bmatrix}.$$  

2. Find the new primal-dual vector $\tilde{y} = (\tilde{x}, \tilde{\lambda})$ by formulas

$$\tilde{x} := x + \Delta x, \quad \tilde{\lambda} := \lambda + \Delta \lambda.$$  

3. Update the scaling parameter

$$\tilde{k} = (\nu(\tilde{y}))^{-1}.$$  

Along with the matrix $N_k(\cdot)$ we consider the matrix

$$N_\infty(y) = N_\infty(\cdot) = \begin{bmatrix} \nabla^2 L(\cdot) - \nabla c^T(\cdot) \\ \nabla c(\cdot) & 0 \end{bmatrix}.$$  

We will use the following technical Lemmas later.
Lemma 2. [24] Let $A : \mathbb{R}^n \to \mathbb{R}^n$ be an invertible matrix and $\|A^{-1}\| \leq c_0$. Then for small enough $\varepsilon > 0$ and any $B : \mathbb{R}^n \to \mathbb{R}^n$ such that $\|A - B\| \leq \varepsilon$, the matrix $B$ is invertible and the following bounds hold:

\[
\begin{align*}
\text{a)} & \quad \|B^{-1}\| \leq 2c_0 \quad \text{and} \quad \text{b)} & \quad \|A^{-1} - B^{-1}\| \leq 2c_0^2 \varepsilon.
\end{align*}
\] (31)

Lemma 3. If the standard second order optimality conditions (3)–(4) and the Lipschitz conditions (15) are satisfied then there exists small enough $\varepsilon_0 > 0$ and large enough $k_0 > 0$ so that both matrices $N_{\infty}(y)$ and $N_k(y)$ are non-singular and there is $c_0 > 0$ independent of $y \in \Omega_{\varepsilon_0}$ and $k \geq k_0$ so that

\[
\max\{\|N_{\infty}^{-1}(y)\||N_{k}^{-1}(y)\|\} \leq 2c_0 \quad \forall y \in \Omega_{\varepsilon_0} \quad \text{and} \quad \forall k \geq k_0.
\] (32)

Proof. It is well known (see Lemma 2, Chapter 8 [17]) that under the standard second order optimality conditions (3)–(4) the matrix

\[
N_{\infty}(x^*, \lambda^*) = \begin{bmatrix}
\nabla^2 L(x^*, \lambda^*) - \nabla c^T(x^*) \\
\n\nabla c(x^*) \\
\0
\end{bmatrix}
\]

is non-singular, hence there exists $c_0 > 0$ that $\|N_{\infty}^{-1}(y^*)\| \leq c_0$. Due to the Lipschitz condition (15) there exists $L > 0$ that $\|N_k(y) - N_{\infty}(y^*)\| \leq L\|y - y^*\| + k^{-1}\varphi''(1)$ and $\|N_{\infty}(y) - N_{\infty}(y^*)\| \leq L\|y - y^*\|$. Therefore for any small enough $\varepsilon > 0$ there exist small $\varepsilon_0 > 0$ and large $k_0 > 0$ so that

\[
\max\{\|N_k(y) - N_{\infty}(y^*)\|, \|N_{\infty}(y) - N_{\infty}(y^*)\|\} \leq \varepsilon \quad \forall y \in \Omega_{\varepsilon_0}, \forall k \geq k_0.
\]

Applying Lemma 2 first with $A = N_{\infty}(y^*)$ and $B = N_k(y)$ and then with $A = N_{k}(y^*)$ and $B = N_{k}(y)$ we obtain (32).

The following theorem establishes the local quadratic convergence of the PDLT method.

Theorem 5 If the standard second order optimality conditions (3)–(4) and the Lipschitz condition (15) are satisfied then there exists $\varepsilon_0 > 0$ small enough that for any primal-dual pair $y = (x, \lambda) \in \Omega_{\varepsilon_0}$ the PDLT methods (28)–(30) generate the primal-dual sequence that converges to the primal-dual solution with quadratic rate, i. e., the following bound holds:

\[
\|\hat{y} - y^*\| \leq c\|y - y^*\|^2 \quad \forall y \in \Omega_{\varepsilon_0},
\]

and $c > 0$ is independent on $y \in \Omega_{\varepsilon_0}$. 

Proof. The primal-dual Newton direction \( \Delta y = (\Delta x, \Delta \lambda) \) we find from the system

\[
N_k(y) \Delta y = b(y),
\]

where

\[
b(y) = \begin{bmatrix}
-\nabla_x L(x, \lambda) \\
-c(x)
\end{bmatrix}.
\]

Along with the primal-dual system (26)–(27), we consider the Lagrange system of equations, which corresponds to the active constraints at the same point \( y = (x, \lambda) \):

\[
\nabla_x L(x, \lambda) = \nabla f(x) - \nabla c(x)^T \lambda = 0,
\]

\[
c(x) = 0.
\]

We apply Newton method for the system (34)–(35) from the same starting point \( y = (x, \lambda) \). The Newton directions \( \Delta \bar{y} = (\Delta \bar{x}, \Delta \bar{\lambda}) \) for the system (34)–(35) we find from the following system of linear equations:

\[
N_\infty(y) \Delta \bar{y} = b(y).
\]

The new approximation for the system (34)–(35) we obtain by formulas

\[
\bar{x} = x + \Delta \bar{x}, \quad \bar{\lambda} = \lambda + \Delta \bar{\lambda} \quad \text{or} \quad \bar{y} = y + \Delta \bar{y}.
\]

Under standard second order optimality conditions (3) and (4) and the Lipschitz conditions (15) there is \( c_1 > 0 \) independent on \( y \in \Omega_{\varepsilon_0} \) so that the following bound holds (see Theorem 9, Ch. 8 [17]):

\[
\| \bar{y} - y^* \| \leq c_1 \| y - y^* \|^2.
\]

Now we can prove the similar bound for \( \| \bar{y} - y^* \| \). We have

\[
\| \bar{y} - y^* \| = \| y + \Delta y - y^* \| = \| y + \Delta \bar{y} + \Delta y - \Delta \bar{y} - y^* \|
\]

\[
\leq \| \bar{y} - y^* \| + \| \Delta y - \Delta \bar{y} \|.
\]

For \( \| \Delta y - \Delta \bar{y} \| \) we obtain

\[
\| \Delta y - \Delta \bar{y} \| = \| (N_k^{-1}(y) - N_\infty^{-1}(y)) b(y) \| \leq \| N_k^{-1}(y) - N_\infty^{-1}(y) \| \| b(y) \|.\]
From Lemma 3 we have \( \max\{\|N_k^{-1}(y)\|, \|N_k^{-\infty}(y)\|\} \leq 2c_0 \). Besides, \( \|N_k(y) - N_{\infty}(y)\| = k^{-1}\phi''(1) \), therefore using Lemma 2 with \( A = N_k(y), B = N_{\infty}(y) \) we obtain

\[
\|\Delta y - \Delta \hat{y}\| \leq 2k^{-1}\phi''(1)c_0^2\|b(y)\|. \tag{37}
\]

In view of \( \nabla_x L(x^*, \lambda^*) = 0, c(x^*) = 0 \) and the Lipschitz condition (15) we have

\[
\|b(y)\| \leq L\|y - y^*\| \quad \forall y \in \Omega_{\epsilon_0}.
\]

Using (25), (30) and (37) we obtain

\[
\|\Delta y - \Delta \hat{y}\| \leq 2\phi''(1)c_0^2\nu(y)L\|y - y^*\| \\
\leq 2\phi''(1)c_0^2M_0L\|y - y^*\|^2.
\]

Therefore for \( c_2 = 2\phi''(1)c_0^2M_0L \), which is independent of \( y \in \Omega_{\epsilon_0} \), we have

\[
\|\Delta y - \Delta \hat{y}\| \leq c_2\|y - y^*\|^2. \tag{38}
\]

Using (36) and (38) for \( c = 2\max\{c_1, c_2\} \) we obtain

\[
\|\hat{y} - y^*\| \leq \|\hat{y} - y^*\| + \|\Delta y - \Delta \hat{y}\| \leq c\|y - y^*\|^2 \quad \forall y \in \Omega_{\epsilon_0}
\]

and \( c > 0 \) is independent of \( y \in \Omega_{\epsilon_0} \). We completed the proof.

5. Primal-Dual LT method

In this section we describe the globally convergent PDLT method. The globally convergent PDLT method, roughly speaking, works as the LT multipliers method (13)–(14) in the initial phase and as the primal-dual LT method (28)–(30) in the final phase of the computational process.

It is worthwhile to emphasize again that PDLT is not just a mechanical combination of two different methods. It is a unified procedure, each step of which consists of finding the primal-dual direction \( \Delta y = (\Delta x, \Delta \lambda) \) by solving the linearized primal-dual system (26)–(27). Then we use either the primal-dual Newton direction \( \Delta y \) to find a new primal-dual vector \( \hat{y} \) or the primal Newton direction \( \Delta x \) to minimize \( \mathcal{L}(x, \lambda, k) \) in \( x \).
The choice at each step depends on the merit function $\nu(y)$ value and how the value changes after one step. If the primal-dual step produces quadratic reduction of the merit function then the primal-dual step is accepted, otherwise we use the primal direction $\Delta x$ to minimize $L(x, \lambda, k)$ in $x$.

The important part of the method is the way the system (20)–(21) is linearized. Let us start with $y = (x, \lambda)$ and compute $\nu(y)$. By linearizing the system (20) we obtain

$$\nabla^2_{xx} L(x, \lambda) \Delta x - \nabla c^T(x) \Delta \lambda = -\nabla_x L(x, \lambda). \quad (39)$$

The system (21) we split into two sub-systems. The first is associated with the set $I_+(y) = \{i : \lambda_i > \nu(y)\}$ of “big” Lagrange multipliers, while the second is associated with the set $I_0(y) = \{i : \lambda_i \leq \nu(y)\}$ of “small” Lagrange multipliers. Therefore, $I_+(y) \cap I_0(y) = \emptyset$ and $I_+(y) \cup I_0(y) = \{1, \ldots, q\}$. We consider two sub-systems:

$$\tilde{\lambda}_i = \psi'(k\lambda_i^{-1} c_i(\tilde{x})) \lambda_i, \quad i \in I_+(y), \quad (40)$$

$$\tilde{\lambda}_i = \psi'(k\lambda_i^{-1} c_i(\tilde{x})) \lambda_i, \quad i \in I_0(y). \quad (41)$$

The equations (40) can be rewritten as follows:

$$k\lambda_i^{-1} c_i(\tilde{x}) = \psi^{-1}(\tilde{\lambda}_i / \lambda_i) = -\varphi'(\tilde{\lambda}_i / \lambda_i).$$

Let $\tilde{x} = x + \Delta x$ and $\tilde{\lambda} = \lambda + \Delta \lambda$, then

$$c_i(x) + \nabla c_i(x) \Delta x = -k^{-1} \lambda_i \varphi'(1 + \Delta \lambda_i / \lambda_i), \quad i \in I_+(y).$$

Taking into account $\varphi'(1) = 0$ and ignoring terms of second and higher order we obtain

$$c_i(x) + \nabla c_i(x) \Delta x = -k^{-1} \varphi''(1) \Delta \lambda_i, \quad i \in I_+(y). \quad (42)$$

Let $c_+(x)$ be the vector-function associated with “big” Lagrange multipliers, i.e. $c_+(x) = (c_i(x), \ i \in I_+(y))$, $\nabla c_+(x) = J(c_+(x))$ is the correspondent Jacobian and $\Delta \lambda_+ = (\Delta \lambda_i, \ i \in I_+(y))$ is the dual Newton direction associated with “big” Lagrange multipliers. Then the system (42) can be rewritten as follows:

$$\nabla c_+(x) \Delta x + k^{-1} \varphi''(1) \Delta \lambda_+ = -c_+(x). \quad (43)$$
Now let us linearize the system (41). Ignoring terms of second and higher order we obtain

\[ \hat{\lambda}_i = \lambda_i + \Delta \lambda_i = \psi'(k\lambda_i^{-1}c_i(x) + \nabla c_i(x)\Delta x)\lambda_i \]

\[ = \psi'(k\lambda_i^{-1}c_i(x))\lambda_i + k\psi''(k\lambda_i^{-1}c_i(x))\Delta c_i(x) \Delta x \]

\[ = \lambda_i + k\psi''(k\lambda_i^{-1}c_i(x))\nabla c_i(x) \Delta x, \quad i \in I_0(y). \] (44)

Let \( c_0(x) \) be the vector-function associated with “small” Lagrange multiplier, \( \nabla c_0(x) = J(c_0(x)) \) be the corresponding Jacobian, \( \lambda_0 = (\lambda_i, i \in I_0(y)) \) the vector of “small” Lagrange multipliers and \( \Delta \lambda_0 = (\lambda_i, i \in I_0(y)) \) the corresponding dual Newton direction. Then (44) can be rewritten as follows:

\[ -k\Psi''(k\lambda_0^{-1}c_0(x))\Delta c_0(x) \Delta x + \Delta \lambda_0 = \bar{\lambda}_0 - \lambda_0, \] (45)

where

\[ \bar{\lambda}_0 = \Psi'(k\lambda_0^{-1}c_0(x))\lambda_0, \]

\[ \Psi'(k\lambda_0^{-1}c_0(x)) = \text{diag}(\psi'(k\lambda_i^{-1}c_i(x)))_{i \in I_0(y)}, \]

\[ \Psi''(k\lambda_0^{-1}c_0(x)) = \text{diag}(\psi''(k\lambda_i^{-1}c_i(x)))_{i \in I_0(y)}. \]

Combining (39), (44), and (45) we obtain the following system for finding the primal-dual direction \( \Delta y = (\Delta x, \Delta \lambda) \), where \( \Delta \lambda = (\Delta \lambda_+, \Delta \lambda_0) \) and \( I_B \) and \( I_S \) are identity matrices in spaces of “big” and “small” Lagrange multipliers:

\[
M(x, \lambda) \Delta y = 
\begin{bmatrix}
\nabla^2_{xx} L(x, \lambda) & -\nabla c^T(x) & -\nabla c^T_0(x) \\
\nabla c_1(x) & k^{-1}\phi''(1)I_B & 0 \\
-k\Psi''(k\lambda_0^{-1}c_0(x))\nabla c_0(x) & 0 & I_S \\
\n\nabla x L(x, \lambda) & -c_1(x) & \bar{\lambda}_0 - \lambda_0 \\
\end{bmatrix}
\begin{bmatrix}
\Delta x \\
\Delta \lambda_+ \\
\Delta \lambda_0 \\
\end{bmatrix}
\]

(46)
To guarantee the existence of the primal-dual LT direction \(\Delta y\) for any \((x, \lambda) \in \mathbb{R}^n \times \mathbb{R}_+^q\) we replace the system (46) by the following regularized system where \(I^n\) is the identity matrix in \(\mathbb{R}^n\):

\[
M_k(x, \lambda)\Delta y = \begin{bmatrix}
\nabla^2_x L(x, \lambda) + k^{-1} I^n & -\nabla c_+^T(x) & -\nabla c_0^T(x) \\
-\nabla c_+(x) & k^{-1} \varphi''(1) I_B & 0 \\
-k\Psi''(k \lambda^{-1} c_0(x)) & 0 & I_S
\end{bmatrix}
\begin{bmatrix}
\Delta x \\
\Delta \lambda_+ \\
\Delta \lambda_0
\end{bmatrix}
\]

\[
= \begin{bmatrix}
-\nabla_x L(x, \lambda) \\
-c_+(x) \\
\hat{\lambda}_0 - \lambda_0
\end{bmatrix}.
\]

(47)

Finding the primal-dual direction \(\nabla y\) from the system (47) we call the PDLTD\((x, \lambda)\) procedure.

Now we are ready to describe the PDLT method.

**Step 1:** Initialization: We choose an initial primal approximation \(x^0 \in \mathbb{R}^n\), Lagrange multipliers vector \(\lambda^0 = (1, \ldots, 1) \in \mathbb{R}^q\), penalty parameter \(k > 0\) and vector of scaling parameters \(k^0 = k \lambda^0\). Let \(\varepsilon > 0\) be the overall accuracy. We choose parameters \(\alpha > 1, 0 < \eta < 0.5, 0 < \sigma \leq 1\) and \(0 < \theta < 0.25\). Set \(x := x^0, \lambda = \lambda^0, \nu := \nu(x, \lambda), \lambda_c := \lambda^0, k := k^0\).

**Step 2:** If \(\nu \leq \varepsilon\) then stop. **Output:** \(x, \lambda\).

**Step 3:** Find direction: \((\Delta x, \Delta \lambda) := \text{PDLTD}(x, \lambda, \nu(x, \lambda), \lambda_c)\). Set \(\hat{x} := x + \Delta x, \hat{\lambda} := \lambda + \Delta \lambda\).

**Step 4:** If \(\nu(\hat{x}, \hat{\lambda}) \leq \min\{\nu^2 - \theta, (1 - \theta) \nu\}\), set \(x := \hat{x}, \lambda := \hat{\lambda}, \nu := \nu(x, \lambda), k := \max\{\nu^{-1}, k\}\), Goto Step 2.

**Step 5:** Decrease \(0 < t \leq 1\) until \(L(x + t \Delta x, \lambda_c, k) - L(x, \lambda_c, k) \leq \eta t (\Delta L(x, \lambda_c, k), \Delta x)\).

**Step 6:** Set \(x := x + t \Delta x, \hat{\lambda} := \Psi'(k \lambda^{-1} c(x)) \lambda_c\).

**Step 7:** If \(\|\Delta x L(x, \lambda_c, k)\| \leq \frac{\eta}{\sigma} \|\hat{\lambda} - \lambda_c\|\), Goto Step 9.

**Step 8:** Find direction: \((\Delta x, \Delta \lambda) := \text{PDLTD}(x, \lambda, \nu(x, \lambda))\), Goto Step 5.

**Step 9:** If \(\nu(x, \hat{\lambda}) \leq \nu^2 - \theta\), set \(\lambda_c := \hat{\lambda}, \lambda := \lambda_c, \nu := \nu(x, \lambda), k := \max\{\nu^{-1}, k\}, k := (k_i = k \lambda^{-2}_i, i = 1, \ldots, q)\), Goto Step 2.

**Step 10:** Set \(k := k\alpha\), Goto step 8.

The following theorem proves the global convergence of the PDLT method and establishes the asymptotic quadratic rate.
Theorem 6 \textit{If the standard second order optimality conditions (3)-(4) and the Lipschitz condition (15) are satisfied, then PDLT method generates a globally convergent primal-dual sequence that converges to the primal-dual solution with asymptotic quadratic rate.}

\textbf{Proof.} The matrix $M_k(y)\equiv M_k(x, \lambda)$ is nonsingular for any $(x, \lambda)\in \mathbb{R}^n \times \mathbb{R}^q_+$, $\lambda \in \mathbb{R}^q_+$ and any $k > 0$. Let us consider a vector $w = (u, v_+, v_0)$. Keeping in mind $\psi''(t) < 0$, convexity $f(x)$, concavity $c_i(x)$ and the regularization term $k^{-1} P_t$, it is easy to see that $M_k(y)w = 0 \Rightarrow w = 0$. Therefore $M_k^{-1}(y)$ exists and the primal-dual LT direction $\Delta y$ can be found for any $y = (x, \lambda)\in \mathbb{R}^n \times \mathbb{R}^q_+$ and any $k > 0$.

It follows from (25) that for $\forall y \in \Omega_{\leq 0}$ there is $\tau > 0$ that $\nu(y) \geq \tau$, therefore from (30) we have $k^{-1} = \nu(y) \geq \tau$.

After finding $\Delta \lambda_+$ and $\Delta \lambda_0$ from the second and third system in (47) and substituting their values into the first system we obtain

$$P_k(y) \Delta x \equiv P_k(x, \lambda) \Delta x = -\nabla_x L(x, \lambda) = -\nabla_x \mathcal{L}(x, \lambda, k),$$

where

$$P_k(y) = \nabla_x^2 L(x, \lambda) + k^{-1} P_t$$

$$+ k(\psi''(1))^{-1} \nabla c_+^T(x) \nabla c_+ (x) - k \nabla c_0^T(x) \psi''(k \lambda_0^{-1} c_0(x)) \nabla c_0(x)$$

and $\lambda = (\bar{\lambda}_+, \bar{\lambda}_0)$, where $\bar{\lambda}_+ = \lambda_+ - k(\psi''(1))^{-1} c_+ (x)$, $\bar{\lambda}_0 = (\lambda_i = \psi'(k \lambda_0^{-1} c_i(x)) \lambda_i$, $i \in I_0(y))$.

Using arguments similar to those we used in case of $M_k(y)$ we conclude that the symmetric matrix $P_k(y)$ is positive definite. Moreover due to $k^{-1} \geq \tau > 0$ the matrix $P_k(y)$ has, uniformly bounded from below, mineigval $P_k(y) \geq \tau > 0 \ \forall y \in \Omega_{\leq 0}$. On the other hand, for any $y \in \Omega_{\leq 0}$ the mineigvalue $P_k(y) \geq \rho > 0$ due to Debreu’s lemma [7], the standard second order optimality condition (3)-(4) and the Lipschitz condition (15). Therefore the primal Newton direction $\Delta x$ defined by (47) or (48) is a descent direction for minimization $\mathcal{L}(x, \lambda, k)$ in $x$. Therefore for $0 < \eta \leq 0.5$ we can find $t \geq t_0 > 0$ that

$$\mathcal{L}(x + t \Delta x, \lambda, k) - \mathcal{L}(x, \lambda, k) \leq \eta t (\nabla_x \mathcal{L}(x, \lambda, k), \Delta x) \leq -\tau t \eta \|\Delta x\|_2^2.$$  

Due to the boundedness of the primal-dual sequence and the Lipschitz conditions (15) there exists such $M > 0$ that $\|\nabla_x \mathcal{L}(x, \lambda, k)\| \leq M$. Hence, the primal sequence generated by Newton method $x := x + t \Delta x$ with $t > 0$ defined from (49) converges to $\hat{x} = \hat{x}(\lambda, k)$.
\[ \nabla_x L(\hat{x}, \lambda, k) = \nabla_x L(\hat{x}, \hat{\lambda}) = 0. \]

Keeping in mind the standard second order optimality condition (3)–(4) and the Lipschitz condition (15), it follows from Theorem 4 that for the approximation \((\hat{x}^{s+1}, \hat{x}^{s+1})\) the estimate (16) holds. It requires finite number of Newton step to find \(\hat{x}^{s+1}\). Therefore, after \(s_0 = O(\ln \varepsilon^{-1})\) Lagrange multipliers and scaling parameters updates, we find the primal-dual approximation \(y \in \Omega_{\varepsilon_0}\).

Let \(0 < \varepsilon << \varepsilon_0 < 1\) be the desired accuracy.

Keeping in mind properties 2\(^0\)(c), 2\(^0\)(d) of the transformation \(\psi \in \Psi\) as well as (22), (25) and (30) after \(s_1 = O(\ln \ln \varepsilon^{-1})\) updates we obtain

\[
\max\{||k\psi''(k\lambda_0^{-1}c_0(x))||, ||\hat{x}_0 - \lambda_0||\} = o(\varepsilon^2), \quad i \in I_0. \tag{50}
\]

For any \(y \in \Omega_{\varepsilon_0}\) the term \(\|\nabla c_0(x)\|\) is bounded. The boundedness of \(\|\Delta x\|\) follows from boundedness of \(\|\nabla_x L(x, \lambda)\|\) and the fact that \(P_k(y)\) has a mineigenvalue bounded from below by a positive number uniformly in \(y\).

Let us consider the third part of the system (47), that is associated with the “small” Lagrange multipliers

\[ k\psi''(k\lambda_0^{-1}c_0(x))\nabla c_0(x)\Delta x + \Delta \lambda_0 = \hat{x}_0 - \lambda_0. \]

It follows from (50) that \(\|\Delta \lambda_0\| = o(\varepsilon^2)\). This means that after \(s = \max\{s_0, s_1\}\) updates the part of the system (47) associated with “small” Lagrange multipliers becomes irrelevant for the calculation of a Newton direction from (47). In fact, the system (47) reduces to the following system:

\[
\bar{M}_k(x, \lambda)\Delta \bar{y} = \bar{b}(x, \lambda), \tag{51}
\]

where \(\Delta \bar{y}^T = (\Delta x, \Delta \lambda)\), \(\bar{b}(x, \lambda) = (-\nabla_x L(x, \lambda) - c_+(x))\), and

\[
\bar{M}_k(x, \lambda) = \begin{bmatrix}
\nabla^2_{xx} L(x, \lambda) + k^{-1} I^n - \nabla c_{(+)}^T(x) \\
\nabla c_{(+)}(x) & k^{-1} \varphi''_{(1)} I_B
\end{bmatrix}.
\]

At this point we have \(y \in \Omega_{\varepsilon_0}\), therefore it follows from (25) that \(\nu(y) \leq M_0 \varepsilon_0\). Hence for small enough \(\varepsilon_0 > 0\) from \(|\lambda_i - \lambda^*_i| \leq \varepsilon_0\) we obtain \(\lambda_i \geq \nu(y), i \in I^*\). On the other hand we have \(\nu(y) > \lambda_i = O(\varepsilon^2), i \in I_0\), otherwise we obtain \(\nu(y) \leq O(\varepsilon^2)\) and from (25) follows \(\|y - y^*\| = o(\varepsilon^2)\). So, if after \(s = \max\{s_0, s_1\}\)
Lagrange multipliers updates we have not solved the problem with a given accuracy \( \varepsilon > 0 \) then \( I_\varepsilon(y) = I^* \) and \( I_0(y) = I_0^0 = \{ r + 1, \ldots, q \} \) and we continue to perform the PDLT (28)–(30) using

\[
\bar{M}_k(y) = \bar{M}_k(x, \lambda) = \begin{bmatrix}
\nabla^2_x L(x, \lambda) + k^{-1} I^n - \nabla c_i^T(x) \\
\nabla c_i(x) & k^{-1} \varphi''(1) I^c
\end{bmatrix}
\]

instead of \( N_k(\cdot) \) where \( L(x, \lambda) = f(x) - \sum_{i=1}^r \lambda_i c_i(x) \).

Therefore we have

\[
||\Delta y - \Delta \bar{y}|| = ||(\bar{M}_k^{-1}(y) - N_k^{-1}(y))b(y)|| \leq ||\bar{M}_k^{-1}(y) - N_k^{-1}(y)||b(y)||.
\]

On the other hand \( ||\bar{M}_k(y) - N_k(y)|| \leq k^{-1}(1 + \varphi''(1)) \). From Lemma 3 we have max \( \{ ||\bar{M}_k^{-1}(y)||, ||N_k^{-1}(y)|| \} \leq 2c_0 \).

Keeping in mind (25), (30) (15) we obtain the following estimate:

\[
||\Delta y - \Delta \bar{y}|| \leq 2c_0^2 k^{-1}(1 + \varphi''(1)) ||\bar{b}(y)|| = 2c_0^2 \nu(y)(1 + \varphi''(1)) ||\bar{b}(y)||
\]

\[
\leq 2(1 + \varphi''(1))c_0^2 M_0 L ||y - y^*||^2 = c_3 ||y - y^*||^2, \tag{52}
\]

where \( c_3 > 0 \) is independent of \( y \in S(y^*, \varepsilon_0) \).

By using the bound (52) instead of (38) and repeating the consideration used in Theorem 5, we conclude that the primal-dual sequence generated by PDLT converges to the primal-dual solution \((x^*, \lambda^*)\) with asymptotic quadratic rate.

\[\square\]

6. Concluding remarks

The PDLT method is fundamentally different from the Newton LT multipliers method. The distinct characteristic of the PDLT method is its global convergence with an asymptotic quadratic rate. The PDLT method combines the best features of the Newton LT method and the Newton method for the Lagrange system of equations corresponding to the active constraints. At the same time the PDLT method is free from their critical drawbacks. In the initial phase the PDLT method is similar to Newton’s method for LT minimization followed by Lagrange multipliers and scaling parameters update. Such a method (see Theorem 4) converges under a fixed penalty parameter. This allows us to reach the \( \varepsilon_0 \)-neighborhood of
the primal-dual solution in $O(\ln \varepsilon_0^{-1})$ Lagrange multipliers updates without compromising the condition number of the LT Hessian.

On the other hand, once in the neighborhood of the primal-dual solution, the penalty parameter, which is inversely proportional to the merit function, grows extremely fast. Again, the unbounded increase of the scaling parameter at this point does not compromise the numerical stability, because instead of unconstrained minimization, the PDLT solves the primal-dual LT system. Moreover, the primal-dual direction $\Delta y$ becomes very close to the Newton direction (see (52)) for the Lagrange system of equations corresponding to the active constraints. This guarantees the asymptotic quadratic convergence.

The situation recalls the Newton method with step-length for unconstrained smooth convex optimization.

The important part of the PDLT is the way we regularize the Hessian of the Lagrangian (see (47)). This allows us to guarantee the global convergence, without compromising the asymptotic quadratic rate of convergence.

Several issues remain for future research.

First, the neighborhood of the primal-dual solution where the quadratic rate of convergence occurs needs to be characterized through parameters which measure the non-degeneracy of the constrained optimization problems.

Second, the value of the scaling parameter $k_0 > 0$ is a priori unknown and depends on the condition number (see [18]) which can be expressed through parameters of a constrained optimization problem at the solution. These parameters are obviously unknown. Therefore it is important to find an efficient way to change the penalty parameter $k > 0$ using the merit function value.

Third, it is important to understand to what extent the PDLT method can be used in the non-convex case. In this regard recent results from [16] together with local convexity properties of the LT that follow from Debreu’s lemma [7] may play an important role.

Fourth, numerical experiments using various versions of the primal-dual NR methods produce very encouraging results (see [11], [23] and [24]). On the other hand the PDLT method has certain specific features that requires more numerical work to understand better its practical efficiency.
Acknowledgements. The author would like to thank Professor Kurt Anstreicher and two anonymous referees for their remarks, which helped to improve the paper.

The research was supported by NSF Grant CCF-0324999.

References


