Noname manuscript No.
(will be inserted by the editor)

Dual Fast Projected Gradient Method for Quadratic Programming

Roman A. Polyak · **James Costa** · **Saba Neyshabouri**

Received: date / Accepted: date

Abstract The application of the fast gradient method to the dual QP leads to the Dual Fast Projected Gradient (DFPG) method. The DFPG converges with $O\left(k^{-2}\right)$ rate, where $k > 0$ is the number of steps. At each step, it requires $O(nm)$ operations. Therefore for a given $\varepsilon > 0$ an ε-approximation to the optimal dual function value one achieves in $O\left(nm\varepsilon^{-\frac{1}{2}}\right)$ operations.

We present numerical results which strongly corroborate the theory. In particular, we demonstrate high efficiency of the DFPG for large scale QP.

Keywords: Quadratic Programming . Dual Fast Gradient method . Dual Problem . Convergence Rate . Complexity .

1 Introduction

The purpose of the paper is to introduce and analyze two Projected Gradient (PG) methods for solving Quadratic Programming (QP) problems.

The PG method was introduced in the 60's (see [4,6,3]). Its efficiency depends on the projection operation. In many instances the PG has only theoretical value because the projection on the feasible set requires solving at each step a constrained optimization problem as difficult as the initial one.

Both the Dual Projected Gradient (DPG) and the Dual Fast Projected Gradient (DFPG) are free from this fundamental drawback.

They are designed for solving the dual QP with the feasible set $\mathbb{R}^m_+$, the projection operation on which is very simple.

The origin of the DPG one can trace back to B. Pschenichny's linearization method [13] in the early 70's. On the other hand the DPG is a Quadratic Prox type method for the dual QP, which is important for the convergence analysis.

Department of Mathematical Sciences and SEOR Department
George Mason University
E-mail: rpolyak@gmu.edu

The DPG requires $O(mn)$ operations per step and converges with $O(k^{-1})$ rate where k is the number of steps.

Therefore the ε-approximation to the optimal dual function value one can find in $O(mn\varepsilon^{-1})$ operations.

The origin of the DFPG one can trace back to Yu Nesterov's gradient mapping approach (see [8,9]) in the early 80's. The DFPG is closely related to FISTA algorithm developed and analyzed by A. Beck and M. Teboulle [1,2] (see also [5] and references there in).

The DFPG requires the same $O(mn)$ operations per step while the convergence rate is $O(k^{-2})$, therefore the ε-approximation to the optimal function value one can find in $O(mn\varepsilon^{-\frac{1}{2}})$ operations.

The DFPG not only improves both convergence rate and complexity of DPG, but in a number of instances for large scale QP can be considered as an alternative for interior point methods. Moreover, for very large QP using interior point methods can be very difficult or even impossible, because they require solving large scale linear systems at each step (see [11] and references there in). The DFPG requires at each step a matrix by vector multiplication, which can be easily done in parallel. In this respect, DFPG can be suitable for very large scale QP.

Using the dual QP for developing DPG and DFPG is critical, because application of the PG methods to the primal QP leads to solving at each step a QP, which is practically as difficult as the original problem.

Both DPG and DFPG can be used for solving dual problems originated from general nonlinear optimization problem as long as the gradient of the dual function exists, satisfies Lipschitz condition, and can be relatively easily computed.

Both methods have been numerically tested on a number of randomly generated QP's. The results obtained strongly corroborate the theory, and demonstrate the high efficiency of the DFPG method for large scale QP.

2 Problem formulation and basic assumptions

We consider a symmetric negative definite matrix $Q : \mathbb{R}^n \longrightarrow \mathbb{R}^n$and a matrix $A : \mathbb{R}^n \longrightarrow \mathbb{R}^m$. The QP consists of finding :

$$\text{(P)} \qquad p(x^*) = \max\left\{ p(x) = \frac{1}{2}\langle Qx, x\rangle + \langle c, x\rangle \mid x \in \Omega \right\}$$

where $c \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$ and the primal feasible set $\Omega = \{x : Ax - b \leq 0\}$ has a nonempty interior, i.e. int $\Omega \neq \varnothing$. We also assume that

$$\hat{x} = \operatorname{argmax}\{p(x) \mid x \in \mathbb{R}^n\} = -Q^{-1}c \notin \Omega$$

Therefore, $x^* \in \partial\Omega$ and there is a vector of Lagrange multipliers $\lambda^* \in \mathbb{R}^m_+$, such that:

$$\nabla_x \mathscr{L}(x^*, \lambda^*) = Qx^* + c - A^T\lambda^* = 0$$

where $\mathscr{L}(x, \lambda) = \frac{1}{2}\langle Qx, x\rangle + \langle c, x\rangle + \langle \lambda, b - Ax\rangle$ is the Lagrangian for the primal QP (P).

Also, for $\lambda^* \in \mathbb{R}^m_+$and the residual vector $r(x^*) = b - Ax^* \geq 0$ the complementarity condition, $\langle \lambda^*, r(\lambda^*) \rangle = 0$, holds true.

Let's consider the dual problem

$$\text{(D)} \qquad d(\lambda^*) = \min\left\{d(\lambda) \mid \lambda \in \mathbb{R}^m_+\right\}$$

where

$$d(\lambda) = p(x(\lambda)) + \langle \lambda, b - Ax(\lambda) \rangle = max\{\mathscr{L}(x,\lambda) \mid x \in \mathbb{R}^n\}$$

and

$$x(\lambda) = Q^{-1}(A^T\lambda - c) \tag{1}$$

Due to the uniqueness of the primal maximizer $x(\lambda)$, the dual function has a smooth gradient. Keeping in mind $\nabla_x \mathscr{L}(x(\lambda),\lambda) = 0$, we obtain

$$\nabla d(\lambda) = \nabla_x \mathscr{L}(x(\lambda),\lambda)\nabla_\lambda(x(\lambda)) + \nabla_\lambda \mathscr{L}(x(\lambda),\lambda)$$

$$= \nabla_\lambda \mathscr{L}(x(\lambda),\lambda) = r(x(\lambda)) = b - Ax(\lambda) \tag{2}$$

From (1) we have $\nabla_\lambda x(\lambda) = Q^{-1}A^T$ where $\nabla_\lambda x(\lambda)$ is the Jacobian of vector function $x(\lambda)^T = (x_1(\lambda), ..., x_n(\lambda))$. Let's consider the Hessian $\nabla^2_{\lambda\lambda} d(\lambda)$ of the dual function. We obtain

$$\nabla^2_{\lambda\lambda} d(\lambda) = \nabla_\lambda(\nabla d(\lambda)) = \nabla_\lambda(r(x(\lambda))) \cdot \nabla_\lambda(x(\lambda)) = -AQ^{-1}A^T = B$$

In view of negative definiteness of Q , the matrix B is nonnegative definite.

It is well known (see for example Lemma 1.2.2 in [9]) that for a twice differentiable function $d(\lambda)$ the gradient $\nabla d(\lambda)$ satisfies Lipschitz condition with constant L if and only if

$$\left\|\nabla^2 d(\lambda)\right\| \leq L$$

Let $L \geq \sqrt{maxeigval(B)}$, then for any $\lambda_1, \lambda_2 \in \mathbb{R}^m_+$ we have

$$\|\nabla d(\lambda_1) - \nabla d(\lambda_2)\| \leq L\|\lambda_1 - \lambda_2\| \tag{3}$$

3 Dual Projected Gradient Method

In this section we first consider the Dual Projected Gradient (DPG) method for solving the dual problem (D), then we establish rate of convergence as well as complexity bound for the DPG method.

Due to the Slater condition the dual optimal set, $\varLambda^* = \text{Argmin}\left\{d(\lambda) \mid \lambda \in \mathbb{R}^m_+\right\}$, is bounded.

The optimality condition for $\lambda^* \in \varLambda^*$is given by the following inequality.

$$\langle \nabla d(\lambda^*), \varLambda - \lambda^* \rangle \geq 0 \quad \forall \varLambda \in \mathbb{R}^m_+$$

To formulate the DPG method, let's consider the quadratic approximation of $d(\lambda)$. Keeping in mind the Lipschitz condition (3) for the gradient $\nabla d(\lambda)$ the quadratic approximation $\psi_L : \mathbb{R}^m_+ \longrightarrow \mathbb{R}$ at the point $\lambda \in \mathbb{R}^m_+$ we define by the following formula.

$$\psi_L(\Lambda,\lambda) = d(\lambda) + \langle \Lambda - \lambda, \nabla d(\lambda)\rangle + \frac{L}{2}\|\Lambda - \lambda\|^2$$

For a given $\lambda \in \mathbb{R}^m_+$, there exists a unique minimizer

$$\lambda^L_+ \equiv \lambda^L_+(\lambda) = \operatorname{argmin}\left\{\psi_L(\Lambda,\lambda) \mid \Lambda \in \mathbb{R}^m_+\right\} \tag{4}$$

Let's fix $\lambda \in \mathbb{R}^m_+$, then the optimality criteria for $\lambda^L_+ \in \mathbb{R}^m_+$ is given by the following inequality.

$$\nabla_\Lambda \psi_L(\lambda^L_+,\lambda) = \nabla d(\lambda) + L(\lambda^L_+ - \lambda) \geq 0 \tag{5}$$

and the complementarity condition

$$\left\langle \lambda^L_+, \nabla_\Lambda \psi_L(\lambda^L_+,\lambda)\right\rangle = 0 \tag{6}$$

The optimality conditions (5)-(6) yield the following closed form solution for the problem (4)

$$\lambda^L_+ = \left[\lambda - L^{-1}\nabla d(\lambda)\right]_+ \tag{7}$$

where $[a]_+ = ([a_i]_+, i = 1,...,m)$ and

$$[a_i]_+ = \begin{cases} a_i & a_i \geq 0 \\ 0 & a_i < 0 \end{cases}$$

In other words, $\lambda^L_+$is the projection of $(\lambda - L^{-1}\nabla d(\lambda))$on $\mathbb{R}^m_+$.

The solution (7) for the problem (4) leads to the following Dual Projected Gradient (DPG) method

$$\lambda_{s+1} = \left[\lambda_s - L^{-1}\nabla d(\lambda_s)\right]_+ \tag{8}$$

which is in fact a PG method for the dual QP.

On the other hand,

$$\lambda^L_+ = \operatorname{argmin}\left\{\langle \Lambda - \lambda, \nabla d(\lambda)\rangle + \frac{L}{2}\|\Lambda - \lambda\|^2 \mid \Lambda \in \mathbb{R}^m_+\right\} \tag{9}$$

therefore (8) has the flavor of a quadratic prox method.

Note that application of the PG [4,6] (see also [3]) method to the primal leads at each step to finding

$$P_\Omega(x_s - t\nabla p(x_s)) = \operatorname{argmin}\{\|y - (x_s + t\nabla p(x_s))\| \mid y \in \Omega\}$$

which is a problem similar to the original QP.

Let's establish the convergence properties of the DPG method. Due to the Lipschitz condition (3) for a convex function $d : \mathbb{R}^m_+ \longrightarrow \mathbb{R}$ the following bound holds.

$$d(\Lambda) - d(\lambda) - \langle \Lambda - \lambda, \nabla d(\lambda) \rangle \leq \frac{L}{2} \|\Lambda - \lambda\|^2$$

Therefore for any pair $(\Lambda, \lambda) \in \mathbb{R}^m_+ \times \mathbb{R}^m_+$ we have

$$d(\Lambda) \leq \psi_L(\Lambda, \lambda) = d(\lambda) + \langle \Lambda - \lambda, \nabla d(\lambda) \rangle + \frac{L}{2} \|\Lambda - \lambda\|^2$$

The following lemma, which is similar to Lemma 2.3 in [2], is taking place.

Lemma 1 *For any given $\lambda \in \mathbb{R}^m_+$and $L > 0$ such that*

$$d(\lambda^L_+) \leq \psi_L(\lambda^L_+, \lambda) \tag{10}$$

the following inequality holds for any $\Lambda \in \mathbb{R}^m_+$.

$$d(\Lambda) - d(\lambda^L_+) \geq \frac{L}{2} \left\|\lambda^L_+ - \lambda\right\|^2 + L \left\langle \lambda - \Lambda, \lambda^L_+ - \lambda \right\rangle \tag{11}$$

Proof From (10) and the convexity of $d(\Lambda)$ we have

$$\begin{aligned}
d(\Lambda) - d(\lambda^L_+) &\geq d(\Lambda) - \psi(\lambda^L_+, \lambda) \\
&= d(\Lambda) - d(\lambda) - \left\langle \lambda^L_+ - \lambda, \nabla d(\lambda) \right\rangle - \frac{L}{2} \left\|\lambda^L_+ - \lambda\right\|^2 \\
&\geq d(\lambda) + \langle \nabla d(\lambda), \Lambda - \lambda \rangle - d(\lambda) - \left\langle \lambda^L_+ - \lambda, \nabla d(\lambda) \right\rangle - \frac{L}{2} \left\|\lambda^L_+ - \lambda\right\|^2 \\
&= \left\langle \nabla d(\lambda), \Lambda - \lambda^L_+ \right\rangle - \frac{L}{2} \left\|\lambda^L_+ - \lambda\right\|^2 + L \left\|\lambda^L_+ - \lambda\right\|^2 - L \left\|\lambda^L_+ - \lambda\right\|^2 \\
&= \frac{L}{2} \left\|\lambda^L_+ - \lambda\right\|^2 + \left\langle \nabla d(\lambda), \Lambda - \lambda^L_+ \right\rangle - L \left\|\lambda^L_+ - \lambda\right\|^2
\end{aligned}$$

Let's consider the optimality criteria for

$$\lambda^L_+ = \lambda^L_+(\lambda) = \arg\min \left\{ \psi_L(\Lambda, \lambda) \mid \Lambda \in \mathbb{R}^m_+ \right\}$$

we obtain

$$\left\langle \nabla d(\lambda) + L(\lambda^L_+ - \lambda), \Lambda - \lambda^L_+ \right\rangle \geq 0 \qquad \forall \Lambda \in \mathbb{R}^m_+$$

i.e. we have

$$\left\langle \nabla d(\lambda), \Lambda - \lambda^L_+ \right\rangle \geq -L \left\langle \lambda^L_+ - \lambda, \Lambda - \lambda^L_+ \right\rangle \qquad \forall \Lambda \in \mathbb{R}^m_+$$

Therefore

$$\begin{aligned}
d(\Lambda) - d(\lambda^L_+) &\geq \frac{L}{2} \left\|\lambda^L_+ - \lambda\right\|^2 - L \left\langle \lambda^L_+ - \lambda, \Lambda - \lambda^L_+ \right\rangle - L \left\langle \lambda^L_+ - \lambda, \lambda^L_+ - \lambda \right\rangle \\
&= \frac{L}{2} \left\|\lambda^L_+ - \lambda\right\|^2 + L \left\langle \lambda^L_+ - \lambda, \lambda - \Lambda \right\rangle
\end{aligned}$$

□

The DPG method (8) one can view as a linearization method [13] where the quadratic regularization term in (9) is used to normalize the gradient direction.

On the other hand DPG has the features of a quadratic prox method, which plays an important role.

The following theorem establishes the convergence properties of the DPG method (8).

Theorem 1 *For the dual sequence $\{\lambda_s\}$ generated by the DPG method (8) the following bound holds.*

$$\Delta_k = d(\lambda_k) - d(\lambda^*) \leq \frac{L}{2k}\|\lambda_0 - \lambda^*\|^2$$

Proof Let's use (10) with $\Lambda = \lambda^*, \lambda = \lambda_s$and $\lambda_+^L = \lambda_{s+1}$. We obtain

$$\begin{aligned}
\frac{2}{L}(d(\lambda^*) - d(\lambda_{s+1})) &\geq \|\lambda_{s+1} - \lambda_s\|^2 + 2\langle\lambda_s - \lambda^*, \lambda_{s+1} - \lambda_s\rangle \\
&= \langle\lambda_{s+1}, \lambda_{s+1}\rangle - 2\langle\lambda_{s+1}, \lambda_s\rangle + \langle\lambda_s, \lambda_s\rangle + 2\langle\lambda_s, \lambda_{s+1}\rangle \\
&\quad - 2\langle\lambda^*, \lambda_{s+1}\rangle - 2\langle\lambda_s, \lambda_s\rangle + 2\langle\lambda^*, \lambda_s\rangle + \langle\lambda^*, \lambda^*\rangle - \langle\lambda^*, \lambda^*\rangle \\
&= \|\lambda_{s+1} - \lambda^*\|^2 - \|\lambda_s - \lambda^*\|^2
\end{aligned}$$

Summing up the last inequality from $s = 0$ to $s = k - 1$ we obtain

$$\frac{2}{L}(kd(\lambda^*) - \sum_{s=0}^{k-1} d(\lambda_{s+1})) \geq \|\lambda^* - \lambda_k\|^2 - \|\lambda^* - \lambda_0\|^2 \tag{12}$$

Using (10) with $\Lambda = \lambda = \lambda_s$, $\lambda_+^L = \lambda_{s+1}$we obtain

$$\frac{2}{L}(d(\lambda_s) - d(\lambda_{s+1})) \geq \|\lambda_{s+1} - \lambda_s\|^2$$

or

$$sd(\lambda_s) - sd(\lambda_{s+1}) \geq \frac{L}{2}s\|\lambda_{s+1} - \lambda_s\|^2$$

Therefore

$$sd(\lambda_s) - (s+1)d(\lambda_{s+1}) + d(\lambda_{s+1}) \geq \frac{L}{2}s\|\lambda_{s+1} - \lambda_s\|^2$$

summing up the last inequality from $s = 0$ to $s = k - 1$, we obtain

$$-kd(\lambda_k) + \sum_{s=0}^{k-1} d(\lambda_{s+1}) \geq \frac{L}{2}\sum_{s=0}^{k-1} s\|\lambda_{s+1} - \lambda_s\|^2$$

From (12) we have

$$kd(\lambda^*) - \sum_{s=0}^{k-1} d(\lambda_{s+1}) \geq \frac{L}{2}\left[\|\lambda^* - \lambda_k\|^2 - \|\lambda^* - \lambda_0\|^2\right]$$

By adding last two inequalities we obtain

$$k(d(\lambda^*) - d(\lambda_k)) \geq \frac{L}{2}\left[\sum_{s=0}^{k-1} s\|\lambda_{s+1} - \lambda_s\|^2 + \|\lambda^* - \lambda_k\|^2 - \|\lambda^* - \lambda_0\|^2\right]$$

i.e.

$$\Delta_k = d(\lambda_k) - d(\lambda^*) \leq \frac{L}{2k} \|\lambda^* - \lambda_0\|^2 \tag{13}$$

It follows from from (13) that for a given accuracy $\varepsilon > 0$ it takes $k = \frac{L\|\lambda^* - \lambda_0\|^2}{2\varepsilon}$ steps of the DPG method to get $\Delta_k \leq \varepsilon$.

It follows from (9) that each step requires computing $\nabla d(\lambda) = r(x(\lambda)) = b - Ax(\lambda) = b - AQ^{-1}(A^T\lambda - c) = -AQ^{-1}A\lambda + (b + AQ^{-1}c)$. In other words, at each step one updates the first term only which requires $O(mn)$ operations, therefore the complexity of the DPG method is given by the following formula.

$$Comp\ DPG = O(\varepsilon^{-1} mnL \|\lambda^* - \lambda_0\|^2)$$

In the following section we consider the Dual Fast Projected Gradient (DFPG) method, which improves substantially both the convergence rate and complexity of the DPG.

4 Dual Fast Projected Gradient Method

The DFPG is based on Yu. Nesterov gradient mapping [8,9] and closely related to the FISTA algorithm by A. Beck and M. Teboulle [2] (see also [1,5] and references there in).

The DFPG generates an auxiliary sequence $\{\lambda_k\}_{k=0}^{\infty}$ and the main sequence $\{\Lambda_k\}_{k=0}^{\infty}$. Vector λ_k one can view as a predictor while vector Λ_k is the corrector, i.e. the approximation at the step k.

DFPG Method

```
1. Input:
   – $L$ - the upper bound for the Lipschitz constant of the gradient $\nabla d(\lambda)$
   – $\lambda_1 = \lambda_0 \in \mathbb{R}^m_{++}$
   – $t_1 = 1$
2. Step $k$
   (a) Find $\Lambda_k = \lambda_+^L(\lambda_k) = \arg\min\{\psi(\Lambda, \lambda_k) \mid \Lambda \in \mathbb{R}^m_+\}$
   (b) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$
   (c) $\lambda_{k+1} = \Lambda_k + (\frac{t_k - 1}{t_{k+1}})(\Lambda_k - \Lambda_{k-1})$
```

Each step of the DFPG method consists of two phases. The prediction phase (c) generates the predictor λ_{k+1} by extrapolating two previous approximations Λ_{k-1} and Λ_k. The correction phase (a) produces the new approximation

$$\Lambda_{k+1} = argmin\left\{\langle \Lambda - \lambda_{k+1}, \nabla d(\lambda_{k+1})\rangle + \frac{L}{2}\|\Lambda - \lambda_{k+1}\|^2 \mid \Lambda \in \mathbb{R}^m_+\right\} \tag{14}$$

From the optimality criteria for Λ_{k+1} follows the closed form solution for the problem (14)

$$\Lambda_{k+1} = [\lambda_{k+1} - \frac{1}{L}\nabla d(\lambda_{k+1})]_+ \tag{15}$$

It means that the correction phase of DFPG is just a projected gradient step for the dual problem with step-length L^{-1}.

First off all, it follows from (b) that $t_k \geq \frac{1}{2}(k+1), \quad \forall k \geq 1$. In fact, it is true for $k=1$. Assuming that $t_k \geq \frac{1}{2}(k+1)$ from (b) we have

$$t_{k+1} = \frac{1}{2}(1+\sqrt{1+4t_k^2}) \geq \frac{1}{2}(1+\sqrt{1+(k+1)^2}) = \frac{1}{2}(k+2)$$

Let $\Delta_k = d(\Lambda_k) - d(\lambda^*)$, $v_k = t\Lambda_k + (t_k - 1)\Lambda_{k-1} - \lambda^*$.

Theorem 2 *For the sequence $\{\Lambda_k\}_{k=1}^{\infty}$ generated by DFPG method (a)-(c), the following bound holds.*

$$\Delta_{k+1} \leq \frac{2L\|\lambda_0 - \lambda^*\|^2}{(k+2)^2} \tag{16}$$

Proof We first establish the following inequality.

$$\frac{2}{L}(t_k^2 \Delta_k - t_{k+1}^2 \Delta_{k+1}) \geq \|v_{k+1}\|^2 - \|v_k\|^2 \tag{17}$$

From the basic inequality (11) for $\Lambda = \lambda^*$, $\lambda = \lambda_{k+1}$ and $\lambda_+^L = \Lambda_{k+1}$ follows

$$d(\lambda^*) - d(\Lambda_{k+1}) \geq \frac{L}{2}\|\Lambda_{k+1} - \lambda_{k+1}\|^2 + L\langle \lambda_{k+1} - \lambda^*, \Lambda_{k+1} - \lambda_{k+1}\rangle \tag{18}$$

By taking $\Lambda = \Lambda_k$, $\lambda = \lambda_{k+1}$ and $\lambda_+^L = \Lambda_{k+1}$ from (11) we obtain

$$d(\Lambda_k) - d(\Lambda_{k+1}) \geq \frac{L}{2}\|\Lambda_{k+1} - \lambda_{k+1}\|^2 + L\langle \lambda_{k+1} - \Lambda_k, \Lambda_{k+1} - \lambda_{k+1}\rangle \tag{19}$$

or

$$\begin{aligned} \frac{2}{L}(\triangle_k - \triangle_{k+1}) &= \frac{2}{L}\left[d(\Lambda_k) - d(\lambda^*) - (d(\Lambda_{k+1}) - d(\lambda *))\right] \\ &\geq \|\Lambda_{k+1} - \lambda_{k+1}\|^2 + 2\langle \lambda_{k+1} - \Lambda_k, \Lambda_{k+1} - \lambda_{k+1}\rangle \end{aligned} \tag{20}$$

From (18) we have

$$-\frac{2}{L}\Delta_{k+1} \geq \|\Lambda_{k+1} - \lambda_{k+1}\|^2 + 2\langle \lambda_{k+1} - \lambda^*, \Lambda_{k+1} - \lambda_{k+1}\rangle \tag{21}$$

After multiplying (20) by $(t_{k+1} - 1) > 0$ we obtain

$$\begin{aligned} \frac{2}{L}(t_{k+1} - 1)\Delta_k - \frac{2}{L}(t_{k+1} - 1)\Delta_{k+1} \geq \\ (t_{k+1} - 1)\left[\|\Lambda_{k+1} - \lambda_{k+1}\|^2 + 2\langle \lambda_{k+1} - \Lambda_k, \Lambda_{k+1} - \lambda_{k+1}\rangle\right] \end{aligned}$$

By adding the last inequality to (21) we have

$$\begin{aligned}\frac{2}{L}\left[(t_{k+1}-1)\Delta_k - t_{k+1}\Delta_{k+1}\right] \geq\; & t_{k+1}\left\|\Lambda_{k+1}-\lambda_{k+1}\right\|^2 \\ & + 2\left\langle \Lambda_{k+1}-\lambda_{k+1}, (t_{k+1}-1)(\lambda_{k+1}-\Lambda_k)+\lambda_{k+1}-\lambda^*\right\rangle \\ =\; & t_{k+1}\left\|\Lambda_{k+1}-\lambda_{k+1}\right\|^2 \\ & + 2\left\langle \Lambda_{k+1}-\lambda_{k+1}, t_{k+1}(\lambda_{k+1}-\Lambda_k)+\Lambda_k-\lambda^*\right\rangle \\ =\; & t_{k+1}\left\|\Lambda_{k+1}-\lambda_{k+1}\right\|^2 \\ & + 2\left\langle \Lambda_{k+1}-\lambda_{k+1}, t_{k+1}\lambda_{k+1}-(t_{k+1}-1)\Lambda_k-\lambda^*\right\rangle\end{aligned} \tag{22}$$

Note that from (b) of the DFPG method follows $2t_{k+1}-1=\sqrt{1+4t_k^2}$ or

$$t_k^2 = t_{k+1}^2 - t_{k+1} = t_{k+1}(t_{k+1}-1) \tag{23}$$

Multiplying (22) by t_{k+1} and keeping in mind (23) we obtain

$$\begin{aligned}\frac{2}{L}\left[(t_{k+1}-1)t_{k+1}\Delta_k - t_{k+1}^2\Delta_{k+1}\right] =\; & \frac{2}{L}\left[t_k^2\Delta_k - t_{k+1}^2\Delta_{k+1}\right] \\ \geq\; & \left\|t_{k+1}(\Lambda_{k+1}-\lambda_{k+1})\right\|^2 \\ & + 2t_{k+1}\left\langle \Lambda_{k+1}-\lambda_{k+1}, t_{k+1}\lambda_{k+1}-(t_{k+1}-1)\Lambda_k-\lambda^*\right\rangle\end{aligned} \tag{24}$$

Using the three vectors identity

$$\|b-a\|^2 + 2\langle b-a, a-c\rangle = \|b-c\|^2 - \|a-c\|^2 \tag{25}$$

with $a=t_{k+1}\lambda_{k+1}, b=t_{k+1}\Lambda_{k+1}$ and $c=(t_{k+1}-1)\Lambda_k+\lambda^*$ from (24) we obtain

$$\begin{aligned}&\frac{2}{L}(t_k^2\Delta_k - t_{k+1}^2\Delta_{k+1}) \geq \\ &\qquad \|t_{k+1}\Lambda_{k+1}-(t_{k+1}-1)\Lambda_k-\lambda^*\|^2 - \|t_{k+1}\lambda_{k+1}-(t_{k+1}-1)\Lambda_k-\lambda^*\|^2\end{aligned}$$

Using (c) of the DFPG method we obtain

$$t_{k+1}\lambda_{k+1} = t_{k+1}\Lambda_k + (t_k-1)(\Lambda_k-\Lambda_{k-1}) \tag{26}$$

Keeping in mind $v_k = t_k\Lambda_k + (t_k-1)\Lambda_{k-1} - \lambda^*$ from (25) and (26) follows

$$\begin{aligned}\frac{2}{L}(t_k^2\Delta_k - t_{k+1}^2\Delta_{k+1}) \geq\; & \|v_{k+1}\|^2 - \|t_{k+1}\Lambda_k + (t_k-1)(\Lambda_k-\Lambda_{k-1}) - (t_{k+1}-1)\Lambda_k - \lambda^*\|^2 \\ =\; & \|v_{k+1}\|^2 - \|t_k\Lambda_k - (t_k-1)\Lambda_{k-1} - \lambda^*\| \\ =\; & \|v_{k+1}\|^2 - \|v_k\|^2\end{aligned}$$

i.e. (17) holds. It follows from (17) that

$$t_k^2\Delta_k - t_{k+1}^2\Delta_{k+1} \geq \frac{L}{2}\left[\|v_{k+1}\|^2 - \|v_k\|^2\right]$$

Therefore,

$$\begin{aligned} t_{k+1}^2 \Delta_{k+1} + \frac{L}{2}\|v_{k+1}\|^2 &\leq t_k^2 \Delta_k + \frac{L}{2}\|v_k\|^2 \\ &\leq t_{k-1}^2 \Delta_{k-1} + \frac{L}{2}\|v_{k-1}\|^2 \\ &\quad \vdots \\ &\leq t_1^2 \Delta_1 + \frac{L}{2}\|v_1\|^2 \end{aligned} \tag{27}$$

Hence,

$$\begin{aligned} t_{k+1}^2 \Delta_{k+1} &\leq t_1^2 \Delta_1 + \frac{L}{2}\|v_1\|^2 - \frac{L}{2}\|v_{k+1}\|^2 \\ &\leq (d(\Lambda_1) - d(\lambda^*)) + \frac{L}{2}\|\Lambda_1 - \lambda^*\|^2 \end{aligned} \tag{28}$$

For $\Lambda = \lambda^*$, $\lambda_+^L = \Lambda_1$ and $\lambda = \lambda_0$ it follows from (11) that

$$\begin{aligned} d(\lambda^*) - d(\Lambda_1) &\geq \frac{L}{2}\|\Lambda_1 - \lambda_0\|^2 + L\langle \lambda_0 - \lambda^*, \Lambda_1 - \lambda_0 \rangle \\ &= \frac{L}{2}\left[\|\Lambda_1 - \lambda^*\|^2 - \|\lambda_0 - \lambda^*\|^2\right] \end{aligned}$$

Therefore,

$$d(\Lambda_1) - d(\lambda^*) \leq \frac{L}{2}\left[\|\lambda_0 - \lambda^*\|^2 - \|\Lambda_1 - \lambda^*\|^2\right] \tag{29}$$

From (28) and (29) we have

$$t_{k+1}^2 \Delta_{k+1} \leq \frac{L}{2}\|\lambda_0 - \lambda^*\|^2$$

Keeping in mind $t_{k+1} \geq \frac{1}{2}(k+2)$ we obtain (16). □

We completed the proof of Theorem 2.

So the DFPG practically requires numerical effort similar to DPG method per step, but has a much better convergence rate.

It follows from (16) that for a given accuracy $\varepsilon > 0$ it takes

$$k = \frac{\sqrt{2L}\|\lambda_0 - \lambda^*\|}{\sqrt{\varepsilon}}$$

steps of DFPG to get $\Delta_k \leq \varepsilon$, therefore the overall complexity of the DFPG is

$$Comp\, DFPG = O(mn\frac{\sqrt{L}\|\lambda_0 - \lambda^*\|}{\sqrt{\varepsilon}})$$

The interior point methods have a better complexity bound, but for large scale QP, they require solving a large linear system of equations at each step which can be difficult to say the least.

The main operation at each step of DFPG is matrix by vector multiplication which can be done in parallel. In this regard, the DFPG is uniquely suitable for solving large scale QP.

In the following section we describe numerical results obtained with both DPG and DFPG.

5 Numerical results

Both the DPG and DFPG methods were developed and tested using MATLAB. We also designed a random generator, which generates QP with an *apriori* given size and solution vector. All matrices associated with the randomly generated QP are dense.

The randomly generated QP's were solved by DPG and DFPG. The gap $\Delta_s = \frac{d(\lambda_s)-d(\lambda^*)}{d(\lambda^*)}$ defines the progress achieved in a given number of steps.

Figures (1) and (2) below show convergence of both DFPG and DPG methods for two QP of different sizes. On the X-axis we show the number of steps and on the Y-axis we show the gap Δ_s.

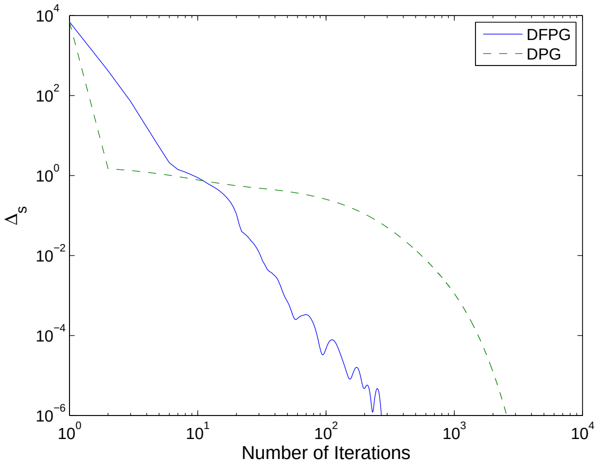

Fig. 1 DFPG (solid) vs. DPG (dashed) for $n = 100$, $m = 50$

As it can be seen from the graphs, DFPG out performs DPG in terms of number of iterations to get the required accuracy ($\varepsilon = 10^{-6}$).

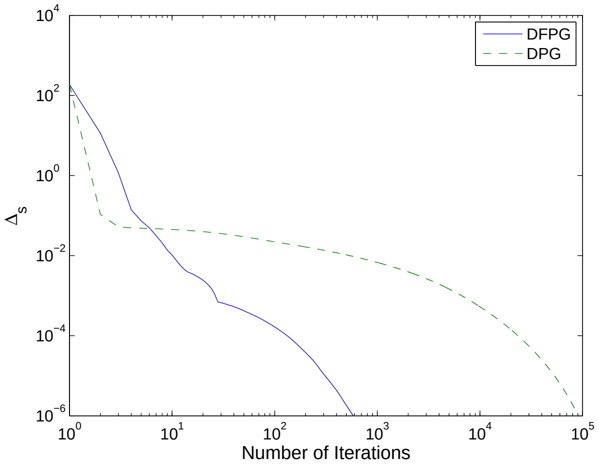

Fig. 2 DFPG (solid) vs. DPG (dashed) for $n = 1000$, $m = 500$

In order to compare the performances of these methods, QPs of varying size were generated and solved by DFPG and DPG. Note that all problems generated had completely dense A matrices. The maximum number of iterations was set to 120,000 iterations and six digits of accuracy was used as the stopping criteria for both algorithms. Both algorithms were tested on an ordinary MacBook Pro laptop. The results obtained are presented in the following table.

Variables	Constraints	DFPG Method		DPG Method	
n	m	Iterations	Time (sec)	Iterations	Time (sec)
100	50	329	0.02357	4337	0.30292
		280	0.02038	9184	0.59667
		271	0.01837	8582	0.56209
		278	0.02176	2762	0.17936
200	100	534	0.06932	31477	3.46527
		391	0.04623	24132	2.68583
		402	0.04651	5447	0.60237
		356	0.05333	7689	0.85084
400	200	424	0.11949	8734	2.20834
		526	0.13622	13603	3.4149
		813	0.20794	73766	18.56511
		457	0.11871	13549	3.39471
800	400	681	0.60041	32783	27.77011
		514	0.43933	76704	63.62423
		758	0.65507	26355	22.27011
		1037	0.87214	70217	58.25214
1600	800	726	4.73007	110836	715.59573
		553	3.48822	94567	606.5642
3200	1600	698	17.65899	120000	3008.12890
		1851	45.03016	120000	2924.31770

As one can see, the DFPG has substantial advantages as compared to DPG in both computation time and number of iterations. In order to show the performance of the DFPG, another set of QP with sizes $1600\times 3200 \div 4000\times 10000$ was generated and solved to six digits of accuracy. The results are presented in the following table.

Variables	Constraints	*L* Constant	DFPG Method	
n	*m*	*L*	Iterations	Time (sec)
3200	1600	5.72411	785	25.03843
		3.97624	751	23.32618
		2.78926	791	24.06845
5000	2000	7.73193	962	56.92800
		2.54017	609	35.77332
		3.35125	768	45.09735
6400	3200	99.81955	2542	301.94020
		4.44271	815	96.38931
		5.03076	782	92.15438
8000	4000	4.30588	878	160.35110
		4.04671	898	164.83670
		4.38200	895	163.06810
10000	4000	7.72731	1120	255.01110
		9.09082	1287	293.89160
		7.13478	1166	265.70370

Our results show that DFPG efficiently handles large-scale QP problems with completely dense matrix A. It is important to note that the matrix algebra is the biggest issue when the size of the problem grows. Implementation can be enhanced by taking advantage of the problem structure and managing memory efficiently using parallel processing.

6 Concluding Remarks

Both theoretical and numerical advances of the method are not due to improvement in number of steps, which is impossible because the method in this regard is optimal (see [6]), but due to low complexity per step.

The numerical results strongly corroborate the theory. A number of QPs of varying size were solved. The results presented in section 5 show that DFPG has the potential to become an efficient tool in particular for large scale QP when IPM can't be efficient due to the necessity to solve a very large linear system at each step (see [10]).

Both methods can be applied for solving a dual problem

$$d(\lambda^*) = min\{d(\lambda) \mid \lambda \in \mathbb{R}^m\}$$

arising from a nonlinear optimization problem as long as $d(\lambda)$ is smooth and the gradient $\nabla d(\lambda)$ satisfies Lipschitz condition.

For the choice of L see the end of section 2. It follows from (1) that $\|x(\lambda) - x(\lambda^*)\| \leq \left\|Q^{-1}A^T\right\| \|\lambda - \lambda^*\|$. It follows from theorem 3 that the Housdorff distance between

L^* and ∂L_k converges to zero, where $\partial L_k = \{\lambda : d(\lambda) = d(\Lambda_k)\}$ therefore $x(\Lambda_k) \rightarrow x(\lambda^*)$.

Another issue is finding the ε-approximation for the primal solution. It is enough to apply, for example, one step of MBF method [12] with an appropriate scaling parameter using the dual ε-approximation vector obtained by DFPG as the dual starting point.

The formula (15) can be viewed as a decomposition method for the dual problem. The components of Λ_k can be computed in parallel, which is important for very large QP.

In our opinion, with appropriate parallelization the method can be uniquely suitable for solving large scale QP arising, in particular, in statistical as well as support vector machine applications (see [7,16,15,14]).

References

1. Beck, A., and Teboulle, M. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *Trans. Img. Proc. 18* (November 2009), 2419–2434.
2. Beck, A., and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences 2*, 1 (2009), 183–202.
3. Bertsekas, D. On the goldstein-levitin-polyak gradient projection method. *Automatic Control IEEE Transactions 21*, 2 (1976), 174–184.
4. Goldstein, A. Convex programming in hilbert space. *Bull. Amer. Mat. Soc. 70* (1964).
5. Guler, O. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization 2*, 4 (1992), 649–664.
6. Levitin, E. S., and Polyak, B. T. Constrained minimization methods. *Zh. Vychisl. Mat. i Mat. Fiz. 6*, 5 (1966), 787–823.
7. Migdalas, A., Pardalos, P., and Storoy, S. *Parallel Computing in Optimization*. Kluwer Academic Publishers, 1997.
8. Nesterov, Y. A method for solving convex programming problems with convergence rate $o(\frac{1}{k^2})$. *Dokl. Akad. Navk. SSSR 269*, 3 (1983), 543–547.
9. Nesterov, Y. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, 2004.
10. Pardalos, P., and Wolkowicz, H. Topics in semidefinite and interior-point methods. *Fields Institute Communications Series 18, American Mathematical Society* (1998).
11. Pardalos, P., Ye, Y., and Han, G. Computational aspects of an interior point algorithm for quadratic problems with box constraints. *Large-Scale Numerical Optimization, SIAM Philadelphia* (1990), 92–112.
12. Polyak, R. Modified barrier functions (theory and method). *Mathematical Programming 54* (92), 177–222.
13. Pschenichny, B. Algorithms for general problems of mathematical programming. *Kibernetica 6* (1970), 120–125.
14. Vapnik, V. *The Nature of Statistical Learning Theory*. Springer Verlag, 1995.
15. Vapnik, V. *Statistical Learning Theory*. John Wiley & Sons, 1998.
16. Vapnik, V. *Estimation of Dependences Based on Empirical Data*. Springer Verlag, 2006.