Supplementary material

Quoc Tran Dinh  
Anastasios Kyrillidis  
Volkan Cevher  
LIONS lab, École Polytechnique Fédérale de Lausanne, Switzerland

A. The proofs of technical statements

A.1. The proof of Theorem 3.2

Proof. Let \( x^k \in \text{dom}(F) \), we define
\[
P^g_k := (\nabla^2 f(x^1) + \partial g)^{-1},
\]
\[
S_k(z) := \nabla^2 f(x^k)z - \nabla f(z).
\]
and
\[
e_k := e_k(x) := [\nabla^2 f(x^k) - \nabla^2 f(x)]d^k.
\]

It follows from the optimality condition (7) in the main text
\[
0 \in \partial g(x^{k+1}) + \nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k).
\]
This condition can be written equivalently to
\[
S_k(x^k) + e_k(x^k) \in \nabla^2 f(x^k)x^{k+1} + \partial g(x^{k+1}).
\]
Therefore, the last relation leads to
\[
x^{k+1} = P^g_k(S_k(x^k) + e_k).
\]
(1)

If we define \( d^k := x^{k+1} - x^k \) then
\[
d^k = P^g_k(S_k(x^k) + e_k) - x^k.
\]
Consequently, we also have
\[
d_{k+1} = P^g_k(S_k(x^{k+1}) + e_{k+1}) - x^{k+1}.
\]
(2)

We consider the norm \( \lambda^g_k(1) := \|d^{k+1}\|_{x^k} \). By using the nonexpansive property of \( P^g_k \), it follows from (1) and (2) that
\[
\lambda^g_k(1) = \|d^{k+1}\|_{x^k} = \|P^g_k(S_k(x^{k+1}) + e_{k+1}) - P^g_k(S_k(x^k) + e_k)\|_{x^k} \\
\quad \leq \|S_k(x^{k+1}) + e_{k+1} - S_k(x^k) - e_k\|_{x^k} \\
\quad \leq \|\nabla f(x^{k+1}) - \nabla f(x^k) - \nabla^2 f(x^k)(x^{k+1} - x^k)\|_{x^k} \\
\quad + \|e_{k+1} - e_k\|_{x^k} \\
\quad = \left[ \int_0^1 [\nabla^2 f(x^\tau) - \nabla^2 f(x^k)](x^{k+1} - x^k) d\tau \right]_{x^k}^{*} \\
\quad + \|e_{k+1} - e_k\|_{x^k}^{*}.
\]
(3)

where \( x^k := x^k + \tau(x^{k+1} - x^k) \). First, we estimate the first term in the last line of (3) which we denote by \( [\cdot]_1 \). Now, we define
\[
M_k := \int_0^1 |\nabla^2 f(x^k + \tau(x^{k+1} - x^k)) - \nabla^2 f(x^k)| d\tau,
\]
and
\[
N_k := \nabla^2 f(x^k)^{-1/2}M_k \nabla^2 f(x^k)^{-1/2}.
\]
Similar to the proof of Theorem 4.14 in (Nesterov, 2004), we can show that \( \|N_k\| \leq (1 - \|d^k\|_{x^k}^{-1})^{-1}\|d^k\|_{x^k} \). Combining this inequality and (3) we deduce
\[
[\cdot]_1 = \|M_k d^k\|_{x^k}^{*} \leq \|N_k\| \|d^k\|_{x^k}^{*} \\
= (1 - \lambda^g_k)^{-1}\lambda^2_k.
\]
(4)

Next, we estimate the second term of (3) which is denoted by \([\cdot]_2\). We note that \( e_k = e_k(x^k) = 0 \) and
\[
e_{k+1} = e_{k+1}(x^{k+1}) = [\nabla^2 f(x^k) - \nabla^2 f(x^{k+1})]d^{k+1}.
\]
Let
\[
P_k := \nabla^2 f(x^k)^{-1/2}[\nabla^2 f(x^{k+1}) - \nabla^2 f(x^k)]\nabla^2 f(x^k)^{-1/2}.
\]
By applying Theorem 4.16 in (Nesterov, 2004), we can estimate \( \|P_k\| \) as
\[
\|P_k\| \leq \max \left\{ 1 - (1 - \|d^k\|_{x^k}^{-2})^2, \frac{1}{1 - \|d^k\|_{x^k}^{-2}} - 1 \right\} \\
= \frac{2\lambda_k - \lambda^2_k}{(1 - \lambda^g_k)^2}.
\]
(5)

Therefore, from the definition of \([\cdot]_2\) we have
\[
[\cdot]_2 = \|e_{k+1} - e_k\|_{x^k}^{*} \\
= (e_{k+1} - e_k)^T \nabla^2 f(x^k)^{-1}(e_{k+1} - e_k) \\
= (d^{k+1})^T \nabla^2 f(x^k)^{-1/2}P_k^2 \nabla^2 f(x^k)^{-1/2}d^{k+1} \\
\leq \|P_k\| \|d^{k+1}\|_{x^k}^{2}.
\]
(6)
By substituting (5) into (6) we obtain
\[ [\parallel x_k^{k+1} \parallel_{x_{k+1}}] \leq \frac{2\lambda_k - \lambda_k^2}{(1 - \lambda_k)^2} \lambda_k^1. \] (7)

Substituting (4) and (7) into (3) we obtain
\[ \lambda_k^1 \leq \frac{\lambda_k^2}{1 - \lambda_k} + \frac{2\lambda_k - \lambda_k^2}{(1 - \lambda_k)^2} \lambda_k^1. \]

By rearrange this inequality we obtain
\[ \lambda_k^1 \leq \left[ \frac{1 - \lambda_k}{4\lambda_k^2 + 2\lambda_k} \right] \lambda_k^2. \] (8)

On the other hand, by applying Theorem 4.1.6 in (Nesterov, 2004), we can easily show that
\[ \lambda_{k+1} = \left\| d^{k+1} \right\|_{x_{k+1}} \leq \frac{\left\| d^{k+1} \right\|_{x_k}}{1 - \left\| d^k \right\|_{x_k}} = \lambda_k^1. \] (9)

Combining (8) and (9) we obtain
\[ \lambda_{k+1} \leq \frac{\lambda_k^2}{1 - 4\lambda_k + 2\lambda_k^2}, \]
which is (11) in the main text. Finally, we consider the sequence \( \{x^k\}_{k \geq 0} \) generated by (9) in the main text.
From (11) in the main text, we have
\[
\lambda_1 \leq (1 - 4\lambda_0 + 2\lambda_0^2)^{-1} \lambda_0^2 \\
\leq (1 - 4\sigma + 2\sigma^2)^{-1} 2^2 \\
\leq \sigma
\]
provided that \( 0 < \sigma \leq \frac{5 - \sqrt{27}}{4} \approx 0.219224 \). By induction, we can conclude that \( \lambda_k \leq \beta \) for all \( k \geq 0 \). It follows from (11) in the main text that
\[ \lambda_{k+1} \leq (1 - 4\sigma + 2\sigma^2)^{-1} \lambda_k^2 \]
for all \( k \), which shows that \( \{\left\| x^k - x^* \right\|_{x^k} \} \) converges to zero at a quadratic rate. \( \square \)

A.2. The proof of Theorem 3.5

Proof. First, we note that
\[ x^{k+1} = x^k + \alpha_k d^k = x^k + (1 + \lambda_k)^{-1} x^k. \]

Hence, we can estimate \( d^{k+1} \) as
\[ \lambda_{k+1} = \left\| d^{k+1} \right\|_{x_{k+1}} \leq \frac{\left\| d^{k+1} \right\|_{x_k}}{1 - \alpha_k \lambda_k} = (1 + \lambda_k) \left\| d^{k+1} \right\|_{x_k}. \]

Combining this inequality and (8) we obtain (19) in the main text.

In order to prove the quadratic convergence, we first show that if \( \lambda_k \leq \sigma \) then \( \lambda_{k+1} \leq \sigma \) for all \( k \geq 0 \).
Indeed, we note that the function:
\[ \varphi(t) := (1 - t^2)(1 - 4t + 2t^2) \]
is increasing in \([0, 1 - 1/\sqrt{2}]\). Let \( \lambda_0 \leq \sigma \). From (19) we have:
\[ \lambda_1 \leq (1 - \sigma^2)\sigma^2(1 - 4\sigma + 2\sigma^2). \]

Therefore, if
\[ (1 - \sigma^2)\sigma^2(1 - 4\sigma + 2\sigma^2) \leq \sigma, \]
then \( \lambda_1 \leq \sigma \). The last requirement leads to \( 0 < \sigma \leq \bar{\sigma} := 0.22187616 \). From this argument, we conclude that if \( \sigma \in (0, \bar{\sigma}] \) then if \( \lambda_0 \leq \sigma \) then \( \lambda_1 \leq \sigma \). By induction, we have \( \lambda_k \leq \sigma \) for \( k \geq 0 \). If we define
\[ c := (1 - \sigma^2)(1 - 4\sigma + 2\sigma^2) \]
then \( c > 0 \) and (19) implies \( \lambda_{k+1} \leq c\lambda^2 \) which shows that the sequence \( \{\lambda_k\}_{k \geq 0} \) locally converges to 0 at a quadratic rate. \( \square \)

A.3. The proof of Lemma 2.2.

Proof. From the self-concordance of \( f \) we have:
\[ \omega(\|y - x\|_x) + f(x) + \nabla f(x)^T (y - x) \leq f(y). \]

On the other hand, since \( g \) is convex we have
\[ g(y) \geq g(x) + v^T (y - x) \]
for any \( v \in \partial g(x) \). Hence,
\[ F(y) \geq F(x) + (\nabla f(x) + v)^T (y - x) + \omega(\|y - x\|_x) \]
\[ \geq F(x) - \lambda(\|y - x\|_x) + \omega(\|y - x\|_x), \]
where \( \lambda(x) := \|\nabla f(x) + v\|^* \). Let:
\[ \mathcal{L}_F(F(x)) := \{ y \in \mathbb{R}^n \mid F(y) \leq F(x) \} \]
be a sublevel set of \( F \). For any \( y \in \mathcal{L}_F(F(x)) \) we have \( F(y) \leq F(x) \) which leads to:
\[ \lambda(x) \|y - x\|_x \geq \omega(\|y - x\|_x) \]
due to the previous inequality. Note that \( \omega \) is a convex and strictly increasing, the equation \( \lambda(x)t = \omega(t) \) has unique solution \( t > 0 \) if \( \lambda(x) < 1 \). Therefore, for any \( 0 \leq t \leq t' \) we have \( |y - x| \leq t \). This implies that \( \mathcal{L}_F(F(x)) \) is bounded. Hence, \( x^* \) exists. The uniqueness of \( x^* \) follows from the increase of \( \omega \). \( \square \)
Algorithm 1 (Fast-projected-gradient algorithm)

**Input:** The current iteration $\Theta_i$ and a given tolerance $\varepsilon_{in} > 0$.

**Output:** An approximate solution $U_k$ of (25) in the main text.

**Initialization:** Compute a Lipschitz constant $L$ and find a starting point $U_0 > 0$.

Set $V_0 := U_0$, $t_0 := 1$.

**for** $k = 0$ to $k_{max}$ **do**

1. $V_{k+1} := \text{clip}_1 \left( U_k - \frac{1}{7} \left[ \Theta_i (U_k + \frac{1}{\rho} \hat{\Sigma}) \Theta_i - \frac{2}{\rho} \Theta_i \right] \right)$.

2. If $\|V_{k+1} - V_k\|_{\text{Fro}} \leq \varepsilon_{in} \max\{1, \|V_k\|_{\text{Fro}}\}$ then terminate.

3. $t_{k+1} := 0.5(1 + \sqrt{1 + 4t_k^2})$ and $\beta_k := \frac{t_{k-1}}{t_{k+1}}$.

4. $U_{k+1} := V_{k+1} + \beta_k (V_{k+1} - V_k)$.

**end for**

B. A fast projected gradient algorithm

For completeness, we provide here a variant of the fast-projected gradient method for solving the dual subproblem (25) in the main text. Let us recall that $\text{clip}_r(X) := \text{sign}(X) \min\{|X|, r\}$ (a point-wise operator). The algorithm is presented as follows.

The main operator in Algorithm 1 is $\Theta_i U_k \Theta_i$ at Step 2, where $\Theta_i$ and $U_k$ are symmetric and $\Theta_i$ may be sparse. This operator requires twice matrix-matrix multiplications. The worst-case complexity of Algorithm 1 is typically $O \left( \sqrt{\frac{L}{\varepsilon_{in}}} \right)$ which is sublinear. If $\mu = \lambda_{\min}(\Theta_i)$, the smallest eigenvalue of $\Theta_i$, is available, we can set $\beta_k := \frac{\sqrt{T} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$ and we get a linear convergence rate.