Supplementary material

Quoc Tran Dinh QUOC.TRANDINH@EPFL.CH
Anastasios Kyrillidis ANASTASIOS.KYRILLIDIS@EPFL.CH
Volkan Cevher VOLKAN.CEVHER@EPFL.CH
LIONS lab, École Polytechnique Fédérale de Lausanne, Switzerland

A. The proofs of technical statements

A.1. The proof of Theorem 3.2

Proof. Let $\mathbf{x}^k \in \mathrm{dom}(F)$, we define

$$P_k^g := (\nabla^2 f(\mathbf{x}^k) + \partial g)^{-1},$$
$$S_k(\mathbf{z}) := \nabla^2 f(\mathbf{x}^k)\mathbf{z} - \nabla f(\mathbf{z}).$$

and

$$\mathbf{e}_k \equiv \mathbf{e}_k(\mathbf{x}) := [\nabla^2 f(\mathbf{x}^k) - \nabla^2 f(\mathbf{x})]\mathbf{d}^k).$$

It follows from the optimality condition (7) in the main text that

$$\mathbf{0} \in \partial g(\mathbf{x}^{k+1}) + \nabla f(\mathbf{x}^k) + \nabla^2 f(\mathbf{x}^k)(\mathbf{x}^{k+1} - \mathbf{x}^k).$$

This condition can be written equivalently to

$$S_k(\mathbf{x}^k) + \mathbf{e}_k(\mathbf{x}^k) \in \nabla^2 f(\mathbf{x}^k)\mathbf{x}^{k+1} + \partial g(\mathbf{x}^{k+1}).$$

Therefore, the last relation leads to

$$\mathbf{x}^{k+1} = P_k^g(S_k(\mathbf{x}^k) + \mathbf{e}_k). \tag{1}$$

If we define $\mathbf{d}^k := \mathbf{x}^{k+1} - \mathbf{x}^k$ then

$$\mathbf{d}^k = P_k^g(S_k(\mathbf{x}^k) + \mathbf{e}_k) - \mathbf{x}^k.$$

Consequently, we also have

$$\mathbf{d}_{k+1} = P_k^g(S_k(\mathbf{x}^{k+1}) + \mathbf{e}_{k+1}) - \mathbf{x}^{k+1}. \tag{2}$$

We consider the norm $\lambda_k^1 := \left\|\mathbf{d}^{k+1}\right\|_{\mathbf{x^k}}$. By using the nonexpansive property of P_k^g, it follows from (1) and (2) that

$$\begin{aligned}
\lambda_k^1 &= \left\|\mathbf{d}^{k+1}\right\|_{\mathbf{x}^k}\\
&= \left\|P_k^g\left(S_k(\mathbf{x}^{k+1}) + \mathbf{e}_{k+1}\right) - P_k^g\left(S_k(\mathbf{x}^k) + \mathbf{e}_k\right)\right\|_{\mathbf{x}^k}\\
&\overset{(5)}{\leq} \left\|S_k(\mathbf{x}^{k+1}) + \mathbf{e}_{k+1} - S_k(\mathbf{x}^k) - \mathbf{e}_k\right\|_{\mathbf{x}^k}^*\\
&\leq \left\|\nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k) - \nabla^2 f(\mathbf{x}^k)(\mathbf{x}^{k+1} - \mathbf{x}^k)\right\|_{\mathbf{x}^k}^*\\
&+ \left\|\mathbf{e}_{k+1} - \mathbf{e}_k\right\|_{\mathbf{x}^k}^*\\
&= \left[\left\|\int_0^1 [\nabla^2 f(\mathbf{x}_\tau^k) - \nabla^2 f(\mathbf{x}^k)](\mathbf{x}^{k+1} - \mathbf{x}^k)d\tau\right\|_{\mathbf{x}^k}^*\right]_{[1]}\\
&+ \left[\left\|\mathbf{e}_{k+1} - \mathbf{e}_k\right\|_{\mathbf{x}^k}^*\right]_{[2]},
\end{aligned} \tag{3}$$

where $\mathbf{x}_\tau^k := \mathbf{x}^k + \tau(\mathbf{x}^{k+1} - \mathbf{x}^k)$. First, we estimate the first term in the last line of (3) which we denote by $[\cdot]_{[1]}$. Now, we define

$$\mathbf{M}_k := \int_0^1 [\nabla^2 f(\mathbf{x}^k + \tau(\mathbf{x}^{k+1} - \mathbf{x}^k)) - \nabla^2 f(\mathbf{x}^k)]d\tau,$$

and

$$\mathbf{N}_k := \nabla^2 f(\mathbf{x}^k)^{-1/2}\mathbf{M}_k\nabla^2 f(\mathbf{x}^k)^{-1/2}.$$

Similar to the proof of Theorem 4.1.14 in (Nesterov, 2004), we can show that $\|\mathbf{N}_k\| \leq (1 - \left\|\mathbf{d}^k\right\|_{\mathbf{x}^k})^{-1}\left\|\mathbf{d}^k\right\|_{\mathbf{x}^k}$. Combining this inequality and (3) we deduce

$$\begin{aligned}
[\cdot]_{[1]} &= \left\|\mathbf{M}_k\mathbf{d}^k\right\|_{\mathbf{x}^k}^* \leq \|\mathbf{N}_k\|\left\|\mathbf{d}^k\right\|_{\mathbf{x^k}}\\
&= (1 - \lambda_k)^{-1}\lambda_k^2.
\end{aligned} \tag{4}$$

Next, we estimate the second term of (3) which is denoted by $[\cdot]_{[2]}$. We note that $\mathbf{e}_k = \mathbf{e}_k(\mathbf{x}^k) = 0$ and

$$\mathbf{e}_{k+1} = \mathbf{e}_{k+1}(\mathbf{x}^{k+1}) = [\nabla^2 f(\mathbf{x}^k) - \nabla^2 f(\mathbf{x}^{k+1})]\mathbf{d}^{k+1}.$$

Let

$$\mathbf{P}_k := \nabla^2 f(\mathbf{x}^k)^{-1/2}[\nabla^2 f(\mathbf{x}^{k+1}) - \nabla^2 f(\mathbf{x}^k)]\nabla^2 f(\mathbf{x}^k)^{-1/2}.$$

By applying Theorem 4.1.6 in (Nesterov, 2004), we can estimate $\|\mathbf{P}_k\|$ as

$$\begin{aligned}
\|\mathbf{P}_k\| &\leq \max\left\{1 - (1 - \left\|\mathbf{d}^k\right\|_{\mathbf{x}^k})^2, \frac{1}{(1 - \left\|\mathbf{d}^k\right\|_{\mathbf{x}^k})^2} - 1\right\}\\
&= \frac{2\lambda_k - \lambda_k^2}{(1 - \lambda_k)^2}.
\end{aligned} \tag{5}$$

Therefore, from the definition of $[\cdot]_{[2]}$ we have

$$\begin{aligned}
[\cdot]_{[2]}^2 &= [\|\mathbf{e}_{k+1} - \mathbf{e}_k\|_{\mathbf{x}^k}^*]^2\\
&= (\mathbf{e}_{k+1} - \mathbf{e}_k)^T\nabla^2 f(\mathbf{x}^k)^{-1}(\mathbf{e}_{k+1} - \mathbf{e}_k)\\
&= (\mathbf{d}^{k+1})^T\nabla^2 f(\mathbf{x}^k)^{1/2}\mathbf{P}_k^2\nabla^2 f(\mathbf{x}^k)^{1/2}\mathbf{d}^{k+1}\\
&\leq \|\mathbf{P}_k\|^2\left\|\mathbf{d}^{k+1}\right\|_{\mathbf{x}^k}^2.
\end{aligned} \tag{6}$$

By substituting (5) into (6) we obtain

$$[\cdot]_{[2]} \leq \frac{2\lambda_k - \lambda_k^2}{(1-\lambda_k)^2}\lambda_k^1. \tag{7}$$

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

$$\lambda_k^1 \leq \left[\frac{1-\lambda_k}{1-4\lambda_k+2\lambda_k^2}\right]\lambda_k^2. \tag{8}$$

On the other hand, by applying Theorem 4.1.6 in (Nesterov, 2004), we can easily show that

$$\lambda_{k+1} = \left\|\mathbf{d}^{k+1}\right\|_{\mathbf{x}^{k+1}} \leq \frac{\left\|\mathbf{d}^{k+1}\right\|_{\mathbf{x}^k}}{1-\left\|\mathbf{d}^k\right\|_{\mathbf{x}^k}} = \frac{\lambda_k^1}{1-\lambda_k}. \tag{9}$$

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 $\left\{\mathbf{x}^k\right\}_{k\geq 0}$ generated by (9) in the main text. From (11) in the main text, we have

$$\begin{aligned}\lambda_1 &\leq (1-4\lambda_0+2\lambda_0^2)^{-1}\lambda_0^2\\ &\leq (1-4\sigma+2\sigma^2)^{-1}\sigma^2\\ &\leq \sigma\end{aligned}$$

provided that $0<\sigma\leq \frac{5-\sqrt{17}}{4} \approx 0.219224$. By induction, we can conclude that $\lambda_k \leq \beta$ for all $k\geq 0$. It follows from (11) in the main text that

$$\lambda_{k+1} \leq\leq (1-4\sigma+2\sigma^2)^{-1}\lambda_k^2$$

for all k, which shows that $\left\{\left\|\mathbf{x}^k-\mathbf{x}^*\right\|_{\mathbf{x}^k}\right\}$ converges to zero at a quadratic rate. $\square$

A.2. The proof of Theorem 3.5

Proof. First, we note that

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k\mathbf{d}^k = \mathbf{x}^k + (1+\lambda_k)^{-1}\mathbf{x}^k.$$

Hence, we can estimate $\mathbf{d}^{k+1}$ as

$$\lambda_{k+1} = \left\|\mathbf{d}^{k+1}\right\|_{\mathbf{x}^{k+1}} \leq \frac{\left\|\mathbf{d}^{k+1}\right\|_{\mathbf{x}^k}}{1-\alpha_k\lambda_k} = (1+\lambda_k)\left\|\mathbf{d}^{k+1}\right\|_{\mathbf{x}^k}.$$

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:

$$\omega(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}) + f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) \leq f(\mathbf{y}).$$

On the other hand, since g is convex we have

$$g(\mathbf{y}) \geq g(\mathbf{x}) + \mathbf{v}^T(\mathbf{y}-\mathbf{x})$$

for any $\mathbf{v}\in\partial g(\mathbf{x})$. Hence,

$$\begin{aligned}F(\mathbf{y}) &\geq F(\mathbf{x}) + [\nabla f(\mathbf{x})+\mathbf{v}]^T(\mathbf{y}-\mathbf{x}) + \omega(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})\\ &\geq F(\mathbf{x}) - \lambda(\mathbf{x})\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} + \omega(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}),\end{aligned}$$

where $\lambda(\mathbf{x}) := \|\nabla f(\mathbf{x})+\mathbf{v}\|_{\mathbf{x}}^*$. Let:

$$\mathcal{L}_F(F(\mathbf{x})) := \{\mathbf{y}\in\mathbb{R}^n \mid F(\mathbf{y})\leq F(\mathbf{x})\}$$

be a sublevel set of F. For any $y\in\mathcal{L}_F(F(\mathbf{x}))$ we have $F(\mathbf{y})\leq F(\mathbf{x})$ which leads to:

$$\lambda(\mathbf{x})\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} \geq \omega(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})$$

due to the previous inequality. Note that ω is a convex and strictly increasing, the equation $\lambda(\mathbf{x})t = \omega(t)$ has unique solution $\bar{t}>0$ if $\lambda(\mathbf{x})<1$. Therefore, for any $0\leq t\leq\bar{t}$ we have $\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\leq\bar{t}$. This implies that $\mathcal{L}_F(F(\mathbf{x}))$ is bounded. Hence, $\mathbf{x}^*$ exists. The uniqueness of $\mathbf{x}^*$ follows from the increase of ω. $\square$

Algorithm 1 (*Fast-projected-gradient algorithm*)

Input: The current iteration $\mathbf{\Theta}_i$ and a given tolerance $\varepsilon_{\rm in} > 0$.

Output: An approximate solution $\mathbf{U}_k$ of (25) in the main text.

Initialization: Compute a Lipschitz constant L and find a starting point $\mathbf{U}_0 \succ 0$.

Set $\mathbf{V}_0 := \mathbf{U}_0$, $t_0 := 1$.

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

1. $\mathbf{V}_{k+1} := \mathtt{clip}_1\left(\mathbf{U}_k - \frac{1}{L}\left[\mathbf{\Theta}_i(\mathbf{U}_k + \frac{1}{\rho}\hat{\Sigma})\mathbf{\Theta}_i - \frac{2}{\rho}\mathbf{\Theta}_i\right]\right)$.
2. If $\|\mathbf{V}_{k+1} - \mathbf{V}_k\|_{\rm Fro} \le \varepsilon_{\rm in}\max\{1, \|\mathbf{V}_k\|_{\rm Fro}\}$ 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. $\mathbf{U}_{k+1} := \mathbf{V}_{k+1} + \beta_k(\mathbf{V}_{k+1} - \mathbf{V}_k)$.

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 $\mathtt{clip}_r(X) := \mathrm{sign}(X)\min\{|X|, r\}$ (a point-wise operator). The algorithm is presented as follows.

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