Supplementary Material for
A Multi-step Inertial Forward–Backward Splitting Method for Non-convex Optimization

Jingwei Liang and Jalal M. Fadili
Normandie Univ, ENSICAEN, CNRS, GREYC
{Jingwei.Liang,Jalal.Fadili}@greyc.ensicaen.fr

Gabriel Peyré
CNRS, DMA, ENS Paris
Gabriel.Peyre@ens.fr

1 Proof of Theorem 2.2

Lemma 1. *Let $\{d_k\}_{k\in\mathbb{N}}, \{\varepsilon_k\}_{k\in\mathbb{N}}$ be two non-negative sequences, and $\{\omega_i\}_{i\in I} \in \mathbb{R}^s$ such that*

$$d_{k+1} \le \sum_{i\in I} \omega_i d_{k-i} + \varepsilon_k, \tag{1}$$

for all $k \ge s$. If $\sum_i \omega_i \in [0,1[$ and $\sum_{k\in\mathbb{N}} \varepsilon_k < +\infty$, then

$$\sum_{k\in\mathbb{N}} d_k < +\infty.$$

Remark 2. Lemma 1 is an extension of [3, Lemma 3]. It should be noted that in our case, non-negativity is *not* imposed to the weight ω_i's, but only the sum of them. In fact, we can even afford all ω_i's to be negative, as long as $\sum_{i\in I} \omega_i d_{k-i} + \varepsilon_k$ is positive for all $k \in \mathbb{N}$.

Proof. From (1), suppose that $d_{-1} = d_{-2} = d_{-s+1} = 0$, then sum up for both sides from $k = 0$,

$$\begin{aligned}\sum_{k\in\mathbb{N}} d_{k+1} \le \sum_{k\in\mathbb{N}}\sum_{i\in I} \omega_i d_{k-i} + \sum_{k\in\mathbb{N}} \varepsilon_k &\Longrightarrow \sum_{k\in\mathbb{N}} d_k \le d_0 + \sum_{i\in I}\omega_i \sum_{k\in\mathbb{N}} d_k + \sum_{k\in\mathbb{N}} \varepsilon_k \\ &\Longrightarrow \Big(1 - \sum_{i\in I}\omega_i\Big)\sum_{k\in\mathbb{N}} d_k \le d_0 + \sum_{k\in\mathbb{N}} \varepsilon_k.\end{aligned}$$

Since we assume $\sum_{i\in I}\omega_i < 1$ and ε_k is summable, then we have

$$\sum_{k\in\mathbb{N}} d_k \le \Big(1 - \sum_{i\in I}\omega_i\Big)^{-1}\Big(d_0 + \sum_{k\in\mathbb{N}} \varepsilon_k\Big) < +\infty,$$

which concludes the proof. □

Define $\Delta_k \stackrel{\text{def}}{=} \|x_k - x_{k-1}\|$.

Lemma 3. *For the update of x_{k+1} in (1.7), given any $k \in \mathbb{N}$, define*

$$g_{k+1} \stackrel{\text{def}}{=} \frac{1}{\gamma_k}(y_{a,k} - x_{k+1}) - \nabla F(y_{b,k}) + \nabla F(x_{k+1}).$$

We have $g_{k+1} \in \partial\Phi(x_{k+1})$, and moreover,

$$\|g_{k+1}\| \le \Big(\frac{1}{\underline{\gamma}} + L\Big)\Delta_{k+1} + \sum_{i\in I}\Big(\frac{|a_{i,k}|}{\underline{\gamma}} + |b_{i,k}|\Big)\Delta_{k-i}. \tag{2}$$

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Proof. From the definition of the proximity operator and the update of x_{k+1} (1.7), we have $y_{a,k} - \gamma_k \nabla F(y_{b,k}) - x_{k+1} \in \gamma_k \partial R(x_{k+1})$. Adding $\gamma_k \nabla F(x_{k+1})$ to both sides, we get

$$g_{k+1} = \frac{y_{a,k} - \gamma_k \nabla F(y_{b,k}) - x_{k+1} + \gamma_k \nabla F(x_{k+1})}{\gamma_k} \in \partial \Phi(x_{k+1}).$$

Now, applying the triangle inequality and using Lipschitz continuity of ∇F, we get

$$\begin{aligned}
\|g_{k+1}\| &= \|\tfrac{1}{\gamma_k}(y_{a,k} - x_{k+1}) - \nabla F(y_{b,k}) + \nabla F(x_{k+1})\| \\
&\leq \frac{1}{\gamma_k}\|y_{a,k} - x_{k+1}\| + L\|y_{b,k} - x_{k+1}\| \\
&\leq \frac{1}{\gamma_k}\Big(\Delta_{k+1} + \sum_{i\in I}|a_{i,k}|\Delta_{k-i}\Big) + L\Big(\Delta_{k+1} + \sum_{i\in I}|b_{i,k}|\Delta_{k-i}\Big) \\
&\leq \Big(\frac{1}{\underline{\gamma}} + L\Big)\Delta_{k+1} + \sum_{i\in I}\Big(\frac{|a_{i,k}|}{\underline{\gamma}} + |b_{i,k}|\Big)\Delta_{k-i},
\end{aligned}$$

which concludes the proof. □

Lemma 4. *For Algorithm 1, given the parameters $\gamma_k, a_{i,k}, b_{i,k}$, the following inequality holds*

$$\Phi(x_{k+1}) + \underline{\beta}\Delta_{k+1}^2 \leq \Phi(x_k) + \sum_{i\in I}\overline{\alpha}_i\Delta_{k-i}^2. \tag{3}$$

Proof. Define the function

$$\mathcal{L}_k(x) = \gamma_k R(x) + \frac{1}{2}\|x - y_{a,k}\|^2 + \gamma_k\langle x,\, \nabla F(y_{b,k})\rangle.$$

It can be shown that the update of x_{k+1} in (1.7) is equivalent to

$$x_{k+1} \in \operatorname{Argmin}_{x\in\mathbb{R}^n}\mathcal{L}_k(x), \tag{4}$$

which means that $\mathcal{L}_k(x_{k+1}) \leq \mathcal{L}_k(x_k)$, and

$$R(x_{k+1}) + \frac{1}{2\gamma_k}\|x_{k+1} - y_{a,k}\|^2 + \langle x_{k+1},\, \nabla F(y_{b,k})\rangle \leq R(x_k) + \frac{1}{2\gamma_k}\|x_k - y_{a,k}\|^2 + \langle x_k,\, \nabla F(y_{b,k})\rangle,$$

which in turn leads to,

$$\begin{aligned}
R(x_k) &\geq R(x_{k+1}) + \frac{1}{2\gamma_k}\|x_{k+1} - y_{a,k}\|^2 + \langle x_{k+1} - x_k,\, \nabla F(y_{b,k})\rangle - \frac{1}{2\gamma_k}\|x_k - y_{a,k}\|^2 \\
&= R(x_{k+1}) + \langle x_{k+1} - x_k,\, \nabla F(x_k)\rangle + \frac{1}{2\gamma_k}\Delta_{k+1}^2 \\
&\quad + \frac{1}{\gamma_k}\langle x_k - x_{k+1},\, \textstyle\sum_{i\in I}a_{i,k}(x_{k-i} - x_{k-i-1})\rangle + \langle x_{k+1} - x_k,\, \nabla F(y_{b,k}) - \nabla F(x_k)\rangle.
\end{aligned} \tag{5}$$

Since ∇F is L-Lipschitz, we have the classical inequality

$$\langle \nabla F(x_k),\, x_{k+1} - x_k\rangle \geq F(x_{k+1}) - F(x_k) - \frac{L}{2}\Delta_{k+1}^2.$$

Applying Young's inequality, we obtain

$$\begin{aligned}
\langle x_k - x_{k+1},\, \textstyle\sum_{i\in I}a_{i,k}(x_{k-i} - x_{k-i-1})\rangle &\geq -\Big(\frac{\mu}{2}\Delta_{k+1}^2 + \frac{1}{2\mu}\|\textstyle\sum_{i\in I}a_{i,k}(x_{k-i} - x_{k-i-1})\|^2\Big) \\
&\geq -\Big(\frac{\mu}{2}\Delta_{k+1}^2 + \sum_{i\in I}\frac{sa_{i,k}^2}{2\mu}\Delta_{k-i}^2\Big),
\end{aligned} \tag{6}$$

where $\mu > 0$. Similarly, for $\nu > 0$, we have

$$\begin{aligned}
\langle x_{k+1} - x_k,\, \nabla F(y_{b,k}) - \nabla F(x_k)\rangle &\geq -\Big(\frac{\nu}{2}\Delta_{k+1}^2 + \frac{1}{2\nu}\|\nabla F(y_{b,k}) - \nabla F(x_k)\|^2\Big) \\
&\geq -\Big(\frac{\nu}{2}\Delta_{k+1}^2 + \sum_{i\in I}\frac{sb_{i,k}^2L^2}{2\nu}\Delta_{k-i}^2\Big).
\end{aligned} \tag{7}$$

Combining (5), (6) and (7) leads to

$$\Phi(x_{k+1}) + \beta_k\Delta_{k+1}^2 \leq \Phi(x_k) + \sum_{i\in I}\Big(\frac{sa_{i,k}^2}{2\gamma_k\mu} + \frac{sb_{i,k}^2L^2}{2\nu}\Big)\Delta_{k-i}^2 = \Phi(x_k) + \sum_{i\in I}\alpha_{k,i}\Delta_{k-i}^2. \tag{8}$$

Therefore, we obtain

$$\Phi(x_{k+1}) + \underline{\beta}\Delta_{k+1}^2 \le \Phi(x_{k+1}) + \beta_k\Delta_{k+1}^2 \le \Phi(x_k) + \sum_{i\in I}\alpha_{k,i}\Delta_{k-i}^2 \le \Phi(x_k) + \sum_{i\in I}\bar{\alpha}_i\Delta_{k-i}^2,$$

which concludes the proof. $\square$

Define $\mathbb{R}_s^n$ the product space $\mathbb{R}_s^n \stackrel{\text{def}}{=} \underbrace{\mathbb{R}^n \times \cdots \times \mathbb{R}^n}_{s \text{ times}}$ and $z_k = (x_k, x_{k-1}, ..., x_{k-s+1}) \in \mathbb{R}_s^n$. Then given z_k, define the function

$$\Psi(z_k) = \Phi(x_k) + \sum_{i\in I}\sum_{j=i}^{s-1}\bar{\alpha}_j\Delta_{k-i}^2,$$

which is is a KL function if Φ is. Denote $\mathcal{C}_{x_k}, \mathcal{C}_{z_k}$ the set of cluster points of $\{x_k\}_{k\in\mathbb{N}}$ and $\{z_k\}_{k\in\mathbb{N}}$ respectively, and $\mathrm{crit}(\Psi) = \{z = (x, ..., x) \in \mathbb{R}_s^n : x \in \mathrm{crit}(\Phi)\}$.

Lemma 5. *For Algorithm 1, choose $\mu, \nu, \gamma_k, a_{i,k}, b_{i,k}$ such that* (2.3) *holds. If Φ is bounded from below, then*

(i) $\sum_{k\in\mathbb{N}}\Delta_k^2 < +\infty$*;*
(ii) *The sequence $\Psi(z_k)$ is monotonically decreasing and convergent;*
(iii) *The sequence $\Phi(x_k)$ is convergent.*

Proof. Define

$$\delta = \underline{\beta} - \sum_{i\in I}\bar{\alpha}_i > 0.$$

From the Lemma 4, we have

$$\delta\Delta_{k+1}^2 \le \big(\Phi(x_k) - \Phi(x_{k+1})\big) + \sum_{i\in I}\bar{\alpha}_i(\Delta_{k-i}^2 - \Delta_{k+1}^2).$$

Since we let $x_{1-s} = ... = x_0 = x_1$, for the above inequality, sum over k we get

$$\begin{aligned}
\delta\sum_{k\in\mathbb{N}}\Delta_{k+1}^2 &\le \sum_{k\in\mathbb{N}}\big(\Phi(x_k) - \Phi(x_{k+1})\big) + \sum_{k\in\mathbb{N}}\sum_{i\in I}\bar{\alpha}_i(\Delta_{k-i}^2 - \Delta_{k+1}^2) \\
&\le \Phi(x_0) + \sum_{i\in I}\bar{\alpha}_i\sum_{k\in\mathbb{N}}(\Delta_{k-i}^2 - \Delta_{k+1}^2) \\
&= \Phi(x_0) + \sum_{i\in I}\bar{\alpha}_i\sum_{j=1-i}^{1}\Delta_j^2 = \Phi(x_0),
\end{aligned}$$

which means, as $\Phi(x_0)$ is bounded,

$$\sum_{k\in\mathbb{N}}\Delta_{k+1}^2 \le \frac{\Phi(x_0)}{\delta} < +\infty.$$

From Lemma 4, by pairing terms on both sides of (3), we get

$$\Psi(z_{k+1}) + \Big(\underline{\beta} - \sum_{i\in I}\bar{\alpha}_i\Big)\Delta_{k+1}^2 \le \Psi(z_k).$$

Since we assume $\underline{\beta} - \sum_{i\in I}\bar{\alpha}_i > 0$, hence $\Psi(z_k)$ is monotonically non-increasing. The convergence of $\Phi(x_k)$ is straightforward. $\square$

Lemma 6. *For Algorithm 1, choose $\mu, \nu, \gamma_k, a_{i,k}, b_{i,k}$ such that* (2.3) *holds. If Φ is bounded from below and $\{x_k\}_{k\in\mathbb{N}}$ is bounded, then x_k converges to a critical point of Φ.*

Proof. Since $\{x_k\}_{k\in\mathbb{N}}$ is bounded, there exists a subsequence $\{x_{k_j}\}_{j\in\mathbb{N}}$ and cluster point $\bar{x}$ such that $x_{k_j} \to \bar{x}$ as $j \to \infty$. Next we show that $\Phi(x_{k_j}) \to \Phi(\bar{x})$ and that $\bar{x}$ is a critical point of Φ.

Since R is lsc, then $\liminf_{j\to\infty} R(x_{k_j}) \geq R(\bar{x})$. From (4), we have $\mathcal{L}_{k_j-1}(x_{k_j}) \leq \mathcal{L}_{k_j-1}(\bar{x})$,

$$
\begin{aligned}
R(\bar{x}) &\geq R(x_{k_j}) + \frac{1}{2\gamma_{k_j-1}}\|x_{k_j} - y_{a,k_j-1}\|^2 + \langle x_{k_j} - \bar{x},\, \nabla F(y_{b,k_j-1})\rangle - \frac{1}{2\gamma_{k_j-1}}\|\bar{x} - y_{a,k_j-1}\|^2 \\
&= R(x_{k_j}) + \frac{1}{2\gamma_{k_j-1}}\big(\|x_{k_j} - \bar{x}\|^2 + 2\langle x_{k_j} - \bar{x},\, \bar{x} - y_{a,k_j-1}\rangle\big) + \langle x_{k_j} - \bar{x},\, \nabla F(y_{b,k_j-1})\rangle
\end{aligned}
$$

Since $\Delta_k^2 \to 0$ and $x_{k_j} \to \bar{x}$, then passing to the limit in the inequality we obtain $\limsup_{j\to\infty} R(x_{k_j}) \leq R(\bar{x})$. As a result, $\lim_{k\to\infty} R(x_{k_j}) = R(\bar{x})$. Since F is continuous, then $F(x_{k_j}) \to F(\bar{x})$, hence $\Phi(x_{k_j}) \to \Phi(\bar{x})$.

Furthermore, owing to Lemma 3, $g_{k_j} \in \partial\Phi(x_{k_j})$, and (i) of Lemma 5 we have $g_{k_j} \to 0$ as $k \to \infty$. As a consequence,

$$
g_{k_j} \in \partial\Phi(x_{k_j}),\ (x_{k_j}, g_{k_j}) \to (\bar{x}, 0) \ \text{ and } \ \Phi(x_{k_j}) \to \Phi(\bar{x}),
$$

as $j \to \infty$. Hence $0 \in \partial\Phi(\bar{x})$, *i.e.* $\bar{x}$ is a critical point. □

Now we present the proof of Theorem 2.2.

Proof of Theorem 2.2. Putting together the above lemmas, we draw the following useful conclusions:

- **(R.1)** Denote $\delta = \underline{\beta} - \sum_{i\in I}\bar{\alpha}_i$, then $\Psi(z_{k+1}) + \delta\Delta_{k+1}^2 \leq \Psi(z_k)$;
- **(R.2)** Define
$$
w_{k+1} \stackrel{\text{def}}{=} \begin{pmatrix} g_{k+1} + 2\sum_{i=0}^{s-1}\bar{\alpha}_i(x_{k+1} - x_k) \\ 2\sum_{i=0}^{s-1}\bar{\alpha}_i(x_k - x_{k+1}) + 2\sum_{i=1}^{s-1}\bar{\alpha}_i(x_k - x_{k-1}) \\ \vdots \\ 2\bar{\alpha}_{s-1}(x_{k+2-s} - x_{k+1-s}) \end{pmatrix},
$$
then we have $w_{k+1} \in \partial\Psi(z_{k+1})$. Owing to Lemma 3, there exists a $\sigma > 0$ such that $\|w_{k+1}\| \leq \sigma\sum_{j=k+2-s}^{k+1}\Delta_j$;
- **(R.3)** if x_{k_j} is a subsequence such that $x_{k_j} \to \bar{x}$, then $\Psi(z_k) \to \Psi(\bar{z})$ where $\bar{z} = (\bar{x}, ..., \bar{x})$.
- **(R.4)** $\mathcal{C}_{z_k} \subseteq \mathrm{crit}(\Psi)$;
- **(R.5)** $\lim_{k\to\infty}\mathrm{dist}(z_k, \mathcal{C}_{z_k}) = 0$;
- **(R.6)** $\mathcal{C}_{z_k}$ is non-empty, compact and connected;
- **(R.7)** Ψ is finite and constant on $\mathcal{C}_{z_k}$.

Next we prove the claims of Theorem 2.2.

(i) Consider a critical point of Φ, $\bar{x} \in \mathrm{crit}(\Phi)$, such that $\bar{z} = (\bar{x}, ..., \bar{x}) \in \mathcal{C}_{z_k}$, then owing to (**R.3**), we have $\Psi(z_k) \to \Psi(\bar{z})$.
Suppose there exists K such that $\Psi(z_K) = \Psi(\bar{z})$, then the descent property (**R.1**) implies that $\Psi(z_k) = \Psi(\bar{z})$ holds for all $k \geq K$. Then z_k is constant for $k \geq K$, hence has finite length.
On the other hand, let $\Psi(z_k) > \Psi(\bar{z})$, denote $\psi_k = \Psi(z_k) - \Psi(\bar{z})$. Owing to (**R.6**), (**R.7**) and Definition 2.1, the KL property of Ψ means that there exist ϵ, η and a concave function φ, and

$$
\mathcal{U} \stackrel{\text{def}}{=} \big\{u \in \mathbb{R}_s^n : \mathrm{dist}(u, \mathcal{C}_{z_k}) < \epsilon\big\} \bigcap \big[\Psi(\bar{z}) < \Psi(u) < \Psi(\bar{z}) + \eta\big], \tag{9}
$$

sucht ath for all $z \in \mathcal{U}$,

$$
\varphi'\big(\Psi(z) - \Psi(\bar{z})\big)\mathrm{dist}\big(0, \partial\Psi(z)\big) \geq 1. \tag{10}
$$

Let $k_1 \in \mathbb{N}$ be such that $\Psi(z_k) < \Psi(\bar{z}) + \eta$ holds for all $k \geq k_1$. Owing to (**R.5**), there exists another $k_2 \in \mathbb{N}$ such that $\mathrm{dist}(z_k, \mathcal{C}_{z_k}) < \epsilon$ holds for all $k \geq k_2$. Let $K = \max\{k_1, k_2\}$, then $z_k \in \mathcal{U}$ holds for all $k \geq K$. Then from (10), we have for $k \geq K$

$$
\varphi'(\psi_k)\mathrm{dist}\big(0, \partial\Psi(z_k)\big) \geq 1.
$$

Since φ is concave, φ' is decreasing, and $\Psi(z_k)$ is decreasing, we have

$$
\varphi(\psi_k) - \varphi(\psi_{k+1}) \geq \varphi'(\psi_k)\big(\Psi(z_k) - \Psi(z_{k+1})\big) \geq \frac{\Psi(z_k) - \Psi(z_{k+1})}{\mathrm{dist}(0, \partial\Psi(z_k))}.
$$

From (**R.1**), since $\mathrm{dist}(0, \partial\Psi(z_k)) \leq \|w_k\|$, then

$$\varphi(\psi_k) - \varphi(\psi_{k+1}) \geq \frac{\Psi(z_k) - \Psi(z_{k+1})}{\|w_k\|} \geq \frac{\Psi(z_k) - \Psi(z_{k+1})}{\sigma \sum_{j=k+1-s}^{k} \Delta_j}.$$

Moreover, $\Psi(z_k) - \Psi(z_{k+1}) \geq \delta\Delta_{k+1}^2$ from (**R.2**), therefore we get

$$\varphi(\psi_k) - \varphi(\psi_{k+1}) \geq \frac{\delta\Delta_{k+1}^2}{\sigma \sum_{j=k+1-s}^{k} \Delta_j},$$

which yields

$$\Delta_{k+1}^2 \leq \big(\tfrac{\sigma}{\delta}(\varphi(\psi_k) - \varphi(\psi_{k+1}))\big) \textstyle\sum_{j=k+1-s}^{k} \Delta_j. \tag{11}$$

Taking the square root of both sides and applying Young's inequality with $\kappa > 0$, we further obtain

$$\begin{aligned} 2\Delta_{k+1} &\leq \frac{1}{\kappa} \textstyle\sum_{j=k+1-s}^{k} \Delta_j + \frac{\kappa\sigma}{\delta}(\varphi(\psi_k) - \varphi(\psi_{k+1})) \\ (\kappa = s) \ &\leq \frac{1}{s} \textstyle\sum_{j=k+1-s}^{k} \Delta_j + \frac{s\sigma}{\delta}(\varphi(\psi_k) - \varphi(\psi_{k+1})). \end{aligned} \tag{12}$$

Summing up both sides over k, and since $x_0 = \ldots = x_{-s}$, we get

$$\ell \overset{\text{def}}{=} \textstyle\sum_{k\in\mathbb{N}} \Delta_k \leq \Delta_1 + \frac{s\sigma}{\delta}\varphi(\psi_1) < +\infty,$$

which concludes the finite length property of x_k.

(ii) Then the convergence of the sequence follows from the fact that $\{x_k\}_{k\in\mathbb{N}}$ is a Cauchy sequence, hence convergent. Owing to Lemma 6, there exists a critical point $x^\star \in \mathrm{crit}(\Phi)$ such that $\lim_{k\to\infty} x_k = x^\star$.

(iii) We now turn to proving local convergence to a global minimmizer. Note that if $x^\star$ is a global minimizer of Φ, then $z^\star$ is a global minimizer of Ψ. Let $r > \rho > 0$ such that $\mathbb{B}_r(z^\star) \subset \mathcal{U}$ and $\eta < \delta(r-\rho)^2$. Suppose that the initial point x_0 is chosen such that following conditions hold,

$$\Psi(z^\star) \leq \Psi(z_0) < \Psi(z^\star) + \eta \tag{13}$$

$$\|x_0 - x^\star\| + \ell(s-1) + 2\sqrt{\frac{\Psi(z_0) - \Psi(z^\star)}{\delta}} + \frac{\sigma}{\delta}\varphi(\psi_0) < \rho. \tag{14}$$

The descent property (**R.1**) of Ψ together with (13) imply that for any $k \in \mathbb{N}$, $\Psi(z^\star) \leq \Psi(z_{k+1}) \leq \Psi(z_k) \leq \Psi(z_0) < \Psi(z^\star) + \eta$, and

$$\|x_{k+1} - x_k\| \leq \sqrt{\frac{\Psi(z_k) - \Psi(z_{k+1})}{\delta}} \leq \sqrt{\frac{\Psi(z_k) - \Psi(z^\star)}{\delta}}. \tag{15}$$

Therefore, given any $k \in \mathbb{N}$, if we have $x_k \in \mathbb{B}_\rho(x^\star)$, then

$$\begin{aligned} \|x_{k+1} - x^\star\| &\leq \|x_k - x^\star\| + \|x_{k+1} - x_k\| \leq \|x_k - x^\star\| + \sqrt{\frac{\Psi(z_k) - \Psi(z^\star)}{\delta}} \\ &\leq \rho + (r-\rho) = r, \end{aligned} \tag{16}$$

which means that $x_{k+1} \in \mathbb{B}_r(x^\star)$.
For any $k \in \mathbb{N}$, define the following partial sum

$$p_k = \textstyle\sum_{j=k+1-s}^{k-1} \sum_{i=1}^{j} \Delta_i.$$

Note that $p_k = 0$ for $k = 1$, and $\lim_{k\to\infty} p_k = \ell(s-1)$. Next we prove the following claims through induction: for $k \in \mathbb{N}$

$$x_k \in \mathbb{B}_\rho(x^\star) \tag{17}$$

$$\textstyle\sum_{j=1}^{k} \Delta_{j+1} + \Delta_{k+1} \leq \Delta_1 + p_k + \frac{\sigma}{\delta}(\varphi(\psi_1) - \varphi(\psi_{k+1})). \tag{18}$$

From (15) we have

$$\|x_1 - x_0\| \leq \sqrt{\frac{\Psi(z_0) - \Psi(z^\star)}{\delta}}. \tag{19}$$

Applying the triangle inequality we then obtain

$$\|x_1 - x^\star\| \leq \|x_0 - x^\star\| + \|x_1 - x_0\| \leq \|x_0 - x^\star\| + \sqrt{\frac{\Psi(z_0) - \Psi(z^\star)}{\delta}} < \rho,$$

which means $x_1 \in \mathbb{B}_\rho(x^\star)$. Now, taking $\kappa = 1$ in (12) yields, for any $k \in \mathbb{N}$,

$$2\Delta_{k+1} \leq \sum\nolimits_{j=k+1-s}^{k} \Delta_j + \frac{\sigma}{\delta}(\varphi(\psi_k) - \varphi(\psi_{k+1})). \tag{20}$$

Let $k = 1$. Since $x_0 = \ldots = x_{-s}$, we have

$$2\Delta_2 \leq \Delta_1 + \frac{\sigma}{\delta}(\varphi(\psi_1) - \varphi(\psi_2)).$$

Therefore, (17) and (18) hold for $k = 1$.
Now assume that they hold for some $k > 1$. Using the triangle inequality and (18),

$$\begin{aligned}
\|x_{k+1} - x^\star\| &\leq \|x_0 - x^\star\| + \Delta_1 + \sum\nolimits_{j=1}^{k} \Delta_j \\
&\leq \|x_0 - x^\star\| + 2\Delta_1 + p_k + \frac{\sigma}{\delta}(\varphi(\psi_1) - \varphi(\psi_{k+1})) \\
&\leq \|x_0 - x^\star\| + 2\Delta_1 + (s-1)\ell + \frac{\sigma}{\delta}(\varphi(\psi_1) - \varphi(\psi_{k+1})) \\
(19) &\leq \|x_0 - x^\star\| + 2\sqrt{\frac{\Psi(z_0) - \Psi(z^\star)}{\delta}} + (s-1)\ell + \frac{\sigma}{\delta}(\varphi(\psi_1) - \varphi(\psi_{k+1})).
\end{aligned}$$

As $\varphi(\psi) \geq 0$ and $\varphi'(\psi) > 0$ for $\psi \in]0, \eta[$, and in view of (14), we arrive at

$$\|x_{k+1} - x^\star\| \leq \|x_0 - x^\star\| + 2\sqrt{\frac{\Psi(z_0) - \Psi(z^\star)}{\delta}} + (s-1)\ell + \frac{\sigma}{\delta}\varphi(\psi_0) < \rho$$

whence we deduce that (17) holds at $k+1$. Now, taking (20) at $k+1$ gives

$$\begin{aligned}
2\Delta_{k+2} &\leq \sum\nolimits_{j=k+2-s}^{k+1} \Delta_j + \frac{\sigma}{\delta}(\varphi(\psi_{k+1}) - \varphi(\psi_{k+2})) \\
&\leq \Delta_{k+1} + \sum\nolimits_{j=k+2-s}^{k} \Delta_j + \frac{\sigma}{\delta}(\varphi(\psi_{k+1}) - \varphi(\psi_{(k+2)})).
\end{aligned} \tag{21}$$

Adding both sides of (21) and (18) we get

$$\begin{aligned}
\sum\nolimits_{j=1}^{k+1} \Delta_{j+1} + \Delta_{k+2} &\leq \Delta_1 + p_k + \sum\nolimits_{j=k+2-s}^{k} \Delta_j + \frac{\sigma}{\delta}(\varphi(\psi_1) - \varphi(\psi_{k+2})) \\
&= \Delta_1 + p_{k+1} + \frac{\sigma}{\delta}(\varphi(\psi_1) - \varphi(\psi_{k+2})),
\end{aligned}$$

meaning that (18) holds at $k+1$. This concludes the induction proof.
In summary, the above result shows that if we start close enough from $x^\star$ (so that (13)-(14) hold), then the sequence $\{x_k\}_{k \in \mathbb{N}}$ will remain in the neighbourhood $\mathbb{B}_\rho(x^\star)$ and thus converges to a critical point $\bar{x}$ owing to Lemma 6. Moreover, $\Psi(z_k) \to \Psi(\bar{z}) \geq \Psi(z^\star)$ by virtue of (**R.3**). Now we need to show that $\Psi(\bar{z}) = \Psi(z^\star)$. Suppose that $\Psi(\bar{z}) > \Psi(z^\star)$. As Ψ has the KL property at $z^\star$, we have

$$\varphi'\big(\Psi(\bar{z}) - \Psi(z^\star)\big)\mathrm{dist}\big(0, \partial\Psi(\bar{z})\big) \geq 1.$$

But this is impossible since $\varphi'(s) > 0$ for $s \in]0, \eta[$, and $\mathrm{dist}\big(0, \partial\Psi(\bar{z})\big) = 0$ as $\bar{z}$ is a critical point. Hence we have $\Psi(\bar{z}) = \Psi(z^\star)$, which means $\Phi(\bar{x}) = \Phi(x^\star)$, *i.e.* the cluster point $\bar{x}$ is actually a global minimizer. This concludes the proof.

□

2 Proof of Theorem 3.2

Proof of Theorem 3.2. Under the conditions of Theorem 2.2, there exists a critical point $x^\star \in \mathrm{crit}(\Phi)$ such that $x_k \to x^\star$, $R(x_k) \to R(x^\star)$ and $\Phi(x_k) \to \Phi(x^\star)$ (see the proof of Lemma 6).

Convergence properties of $\{x_k\}_{k \in \mathbb{N}}$ (Theorem 2.2) entails $\|y_{a,k} - x_k\| \to 0$ and $\|y_{b,k} - x^\star\| \to 0$. In turn,

$$\mathrm{dist}\big(-\nabla F(x^\star), \partial R(x_{k+1})\big) \leq \frac{1}{\underline{\gamma}}\|y_{a,k} - x_{k+1}\| + \frac{1}{\beta}\|y_{b,k} - x^\star\| \to 0.$$

Altogether, this shows that the conditions of [10, Theorem 4.10] or [6, Proposition 10.12] are fulfilled on R at $x^\star$ for $-\nabla F(x^\star)$, and the identification result follows.

□

3 Proof of Theorem 3.4

Before presenting the proofs, we need some extra result from partial smoothness, and also Riemannian geometry.

3.1 Partial smoothness and Riemannian geometry

From the sharpness in Definition 3.1, Proposition 2.10 of [9] allows to prove the following fact.

Fact 7 (Local normal sharpness). If $R \in \mathrm{PSF}_x(\mathcal{M})$, then all $x' \in \mathcal{M}$ near x satisfy $\mathcal{T}_{\mathcal{M}}(x') = T_{x'}$. In particular, when $\mathcal{M}$ is affine or linear, then $T_{x'} = T_x$.

We now give expressions of the Riemannian gradient and Hessian (see Section 3.2 for definitions) for the case of partly smooth functions relative to a C^2 submanifold. This is summarized in the following fact which follows by combining (23), (24), Definition 3.1, Fact 7 and [5, Proposition 17] (or [13, Lemma 2.4]).

Fact 8. If $R \in \mathrm{PSF}_x(\mathcal{M})$, then for any $x' \in \mathcal{M}$ near x

$$\nabla_{\mathcal{M}} R(x') = \mathrm{P}_{T_{x'}}(\partial R(x')),$$

and this does not depend on the smooth representation of R on $\mathcal{M}$. In turn, for all $h \in T_{x'}$

$$\nabla^2_{\mathcal{M}} G(x')h = \mathrm{P}_{T_{x'}} \nabla^2 \widetilde{R}(x')h + \mathfrak{W}_{x'}\big(h, \mathrm{P}_{T_{x'}^{\perp}} \nabla \widetilde{R}(x')\big),$$

where $\widetilde{R}$ is a smooth extension (representative) of R on $\mathcal{M}$, and $\mathfrak{W}_x(\cdot,\cdot) : T_x \times T_x^{\perp} \to T_x$ is the Weingarten map of $\mathcal{M}$ at x (see Section 3.2 below for definitions).

3.2 Riemannian Geometry

Let $\mathcal{M}$ be a C^2-smooth embedded submanifold of $\mathbb{R}^n$ around a point x. With some abuse of terminology, we shall state C^2-manifold instead of C^2-smooth embedded submanifold of $\mathbb{R}^n$. The natural embedding of a submanifold $\mathcal{M}$ into $\mathbb{R}^n$ permits to define a Riemannian structure and to introduce geodesics on $\mathcal{M}$, and we simply say $\mathcal{M}$ is a Riemannian manifold. We denote respectively $\mathcal{T}_{\mathcal{M}}(x)$ and $\mathcal{N}_{\mathcal{M}}(x)$ the tangent and normal space of $\mathcal{M}$ at point near x in $\mathcal{M}$.

Exponential map Geodesics generalize the concept of straight lines in $\mathbb{R}^n$, preserving the zero acceleration characteristic, to manifolds. Roughly speaking, a geodesic is locally the shortest path between two points on $\mathcal{M}$. We denote by $\mathfrak{g}(t; x, h)$ the value at $t \in \mathbb{R}$ of the geodesic starting at $\mathfrak{g}(0; x, h) = x \in \mathcal{M}$ with velocity $\dot{\mathfrak{g}}(t; x, h) = \frac{d\mathfrak{g}}{dt}(t; x, h) = h \in \mathcal{T}_{\mathcal{M}}(x)$ (which is uniquely defined). For every $h \in \mathcal{T}_{\mathcal{M}}(x)$, there exists an interval I around 0 and a unique geodesic $\mathfrak{g}(t; x, h) : I \to \mathcal{M}$ such that $\mathfrak{g}(0; x, h) = x$ and $\dot{\mathfrak{g}}(0; x, h) = h$. The mapping

$$\mathrm{Exp}_x : \mathcal{T}_{\mathcal{M}}(x) \to \mathcal{M}, \ \ h \mapsto \mathrm{Exp}_x(h) = \mathfrak{g}(1; x, h),$$

is called *Exponential map*. Given $x, x' \in \mathcal{M}$, the direction $h \in \mathcal{T}_{\mathcal{M}}(x)$ we are interested in is such that

$$\mathrm{Exp}_x(h) = x' = \mathfrak{g}(1; x, h).$$

Parallel translation Given two points $x, x' \in \mathcal{M}$, let $\mathcal{T}_{\mathcal{M}}(x), \mathcal{T}_{\mathcal{M}}(x')$ be their corresponding tangent spaces. Define

$$\tau : \mathcal{T}_{\mathcal{M}}(x) \to \mathcal{T}_{\mathcal{M}}(x'),$$

the parallel translation along the unique geodesic joining x to x', which is isomorphism and isometry w.r.t. the Riemannian metric.

Riemannian gradient and Hessian For a vector $v \in \mathcal{N}_{\mathcal{M}}(x)$, the Weingarten map of $\mathcal{M}$ at x is the operator $\mathfrak{W}_x(\cdot, v) : \mathcal{T}_{\mathcal{M}}(x) \to \mathcal{T}_{\mathcal{M}}(x)$ defined by

$$\mathfrak{W}_x(\cdot, v) = -\mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)} \mathrm{d}V[h],$$

where V is any local extension of v to a normal vector field on $\mathcal{M}$. The definition is independent of the choice of the extension V, and $\mathfrak{W}_x(\cdot, v)$ is a symmetric linear operator which is closely tied to the second fundamental form of $\mathcal{M}$, see [4, Proposition II.2.1].

Let G be a real-valued function which is C^2 along the $\mathcal{M}$ around x. The covariant gradient of G at $x' \in \mathcal{M}$ is the vector $\nabla_{\mathcal{M}} G(x') \in \mathcal{T}_{\mathcal{M}}(x')$ defined by

$$\langle \nabla_{\mathcal{M}} G(x'),\, h\rangle = \frac{d}{dt} G\big(\mathrm{P}_{\mathcal{M}}(x'+th)\big)\big|_{t=0},\ \forall h \in \mathcal{T}_{\mathcal{M}}(x'),$$

where $\mathrm{P}_{\mathcal{M}}$ is the projection operator onto $\mathcal{M}$. The covariant Hessian of G at x' is the symmetric linear mapping $\nabla^2_{\mathcal{M}} G(x')$ from $\mathcal{T}_{\mathcal{M}}(x')$ to itself which is defined as

$$\langle \nabla^2_{\mathcal{M}} G(x')h,\, h\rangle = \frac{d^2}{dt^2} G\big(\mathrm{P}_{\mathcal{M}}(x'+th)\big)\big|_{t=0},\ \forall h \in \mathcal{T}_{\mathcal{M}}(x'). \tag{22}$$

This definition agrees with the usual definition using geodesics or connections [13]. Now assume that $\mathcal{M}$ is a Riemannian embedded submanifold of $\mathbb{R}^n$, and that a function G has a C^2-smooth restriction on $\mathcal{M}$. This can be characterized by the existence of a C^2-smooth extension (representative) of G, *i.e.* a C^2-smooth function $\widetilde{G}$ on $\mathbb{R}^n$ such that $\widetilde{G}$ agrees with G on $\mathcal{M}$. Thus, the Riemannian gradient $\nabla_{\mathcal{M}} G(x')$ is also given by

$$\nabla_{\mathcal{M}} G(x') = \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')} \nabla \widetilde{G}(x'), \tag{23}$$

and $\forall h \in \mathcal{T}_{\mathcal{M}}(x')$, the Riemannian Hessian reads

$$\begin{aligned} \nabla^2_{\mathcal{M}} G(x')h &= \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')} \mathrm{d}(\nabla_{\mathcal{M}} G)(x')[h] = \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')} \mathrm{d}\big(x' \mapsto \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')} \nabla_{\mathcal{M}} \widetilde{G}\big)[h] \\ &= \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')} \nabla^2 \widetilde{G}(x')h + \mathfrak{W}_{x'}\big(h, \mathrm{P}_{\mathcal{N}_{\mathcal{M}}(x')} \nabla \widetilde{G}(x')\big), \end{aligned} \tag{24}$$

where the last equality comes from [1, Theorem 1]. When $\mathcal{M}$ is an affine or linear subspace of $\mathbb{R}^n$, then obviously $\mathcal{M} = x + \mathcal{T}_{\mathcal{M}}(x)$, and $\mathfrak{W}_{x'}(h, \mathrm{P}_{\mathcal{N}_{\mathcal{M}}(x')} \nabla \widetilde{G}(x')) = 0$, hence (24) reduces to

$$\nabla^2_{\mathcal{M}} G(x') = \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')} \nabla^2 \widetilde{G}(x') \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')}.$$

See [8, 4] for more materials on differential and Riemannian manifolds.

The following lemmas summarize two key properties that we will need throughout.

Lemma 9. *Let $x \in \mathcal{M}$, and x_k a sequence converging to x in $\mathcal{M}$. Denote $\tau_k : \mathcal{T}_{\mathcal{M}}(x) \to \mathcal{T}_{\mathcal{M}}(x_k)$ be the parallel translation along the unique geodesic joining x to x_k. Then, for any bounded vector $u \in \mathbb{R}^n$, we have*

$$(\tau_k^{-1} \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x_k)} - \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)})u = o(\|u\|).$$

Proof. See Lemma B.1 of [12]. □

Lemma 10. *Let x, x' be two close points in $\mathcal{M}$, denote $\tau : \mathcal{T}_{\mathcal{M}}(x) \to \mathcal{T}_{\mathcal{M}}(x')$ the parallel translation along the unique geodesic joining x to x'. The Riemannian Taylor expansion of $\Phi \in C^2(\mathcal{M})$ around x reads,*

$$\tau^{-1} \nabla_{\mathcal{M}} \Phi(x') = \nabla_{\mathcal{M}} \Phi(x) + \nabla^2_{\mathcal{M}} \Phi(x) \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)}(x'-x) + o(\|x'-x\|).$$

Proof. See Lemma B.2 of [12]. □

3.3 Proof of Theorem 3.4

The proof of Theorem 1 consists of several steps, first we prove that under the required setting, we can obtain (3.5), *i.e.* the linearized fixed-point iteration.

Proposition 11 (Locally linearized iteration). *For Algorithm 1, suppose that the conditions in Theorem 2.2 hold so that the generated sequence $\{x_k\}_{k\in\mathbb{N}}$ converges to a critical point $x^\star \in \mathrm{crit}(\Phi)$ such that Theorem 3.2 and condition* (3.2) *and* (3.3) *hold. Then for all k large enough, we have*

$$d_{k+1} = M d_k + o(\|d_k\|). \tag{25}$$

The term $o(\cdot)$ vanishes if R is polyhedral around $x^\star$ and $(\gamma_k, a_{i,k}, b_{i,k})$ are chosen constant.

Define the iteration-dependent versions of the matrices in (3.1) and (3.4), *i.e.*

$$
\begin{aligned}
&H_k \stackrel{\text{def}}{=} \gamma_k \mathrm{P}_{T_{x^\star}} \nabla^2 F(x^\star) \mathrm{P}_{T_{x^\star}},\ G_k \stackrel{\text{def}}{=} \mathrm{Id} - H_k,\ Q_k \stackrel{\text{def}}{=} \gamma_k \nabla^2_{\mathcal{M}_{x^\star}} \Phi(x^\star) \mathrm{P}_{T_{x^\star}} - H_k,\\
&M_{k,0} \stackrel{\text{def}}{=} (a_{k,0} - b_{k,0})P + (1 + b_{k,0})PG,\ M_{k,s} \stackrel{\text{def}}{=} -(a_{k,s-1} - b_{k,s-1})P - b_{k,s-1}PG,\\
&M_{k,i} \stackrel{\text{def}}{=} -\big((a_{k,i-1} - a_{k,i}) - (b_{k,i-1} - b_{k,i})\big)P - (b_{k,i-1} - b_{k,i})PG,\ i = 1, ..., s-1,\\
&M_k \stackrel{\text{def}}{=} \begin{bmatrix} M_{k,0} & M_{k,1} & \cdots & M_{k,s-1} & M_{k,s} \\ \mathrm{Id} & 0 & \cdots & 0 & 0 \\ 0 & \mathrm{Id} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathrm{Id} & 0 \end{bmatrix}.
\end{aligned}
\tag{26}
$$

After the finite identification of $\mathcal{M}_{x^\star}$, we have $x_k \in \mathcal{M}_{x^\star}$ for x_k close enough to $x^\star$. Let T_{x_k} be their corresponding tangent spaces, and define $\tau_k : T_{x^\star} \to T_{x_k}$ the parallel translation along the unique geodesic joining from x_k to $x^\star$.

Before proving Proposition 11, we first establish the following intermediate result which provides useful estimates.

Proposition 12. *Under the assumptions of Proposition 11, we have*

$$
\begin{aligned}
&\|y_{a,k} - x^\star\| = O(\|d_k\|),\ \|y_{b,k} - x^\star\| = O(\|d_k\|),\ \|r_{k+1}\| = O(\|d_k\|),\\
&(\tau_{k+1}^{-1} \mathrm{P}_{T_{x_{k+1}}} - \mathrm{P}_{T_{x^\star}})\big(\nabla F(y_{b,k}) - \nabla F(x_{k+1})\big) = o(\|d_k\|).
\end{aligned}
\tag{27}
$$

and

$$
\|P(Q_k - Q)r_{k+1}\| = o(\|d_k\|),\ \|(M_k - M)d_k\| = o(\|d_k\|). \tag{28}
$$

Proof. Since $|a_{i,k}| \le 1$, then

$$
\begin{aligned}
\|y_{a,k} - x^\star\| &= \|x_k + \textstyle\sum_{i\in I} a_{i,k}(x_{k-i} - x_{k-i-1}) - x^\star + \sum_{i\in I} a_{i,k}(x^\star - x^\star)\|\\
&\le \|x_k - x^\star\| + \sum_{i\in I} a_{i,k}(\|x_{k-i} - x^\star\| + \|x_{k-i-1} - x^\star\|)\\
&\le 2\sum_{i\in I} \|r_{k-i}\| \le 2\sqrt{s+1}\|d_k\|,
\end{aligned}
\tag{29}
$$

hence we get the first and second estimates. From prox-regularity of R at $x^\star$ for $-\nabla F(x^\star)$, invoking [15, Proposition 13.37], we have that there exists $\bar{r} > 0$ such that for all $\gamma_k \in]0, \min(\bar{\gamma}, \bar{r})[$, there exists a neighbourhood U of $x^\star - \gamma_k \nabla F(x^\star)$ on which $\mathrm{Prox}_{\gamma_k R}$ is single-valued and l-Lipschitz continuous with $l = \bar{r}/(\bar{r} - \gamma_k)$. Since ∇F is continuous and $x_k \to x^\star$, we have $y_{a,k} - \gamma_k \nabla F(y_{b,k}) \to x^\star - \gamma_k \nabla F(x^\star)$. In turn, $y_{a,k} - \gamma_k \nabla F(y_{b,k}) \in U$ for all k sufficiently large. Thus, we obtain

$$
\begin{aligned}
\|r_{k+1}\| &= \|\mathrm{Prox}_{\gamma_k R}\big(y_{a,k} - \gamma_k \nabla F(y_{b,k})\big) - \mathrm{Prox}_{\gamma_k R}\big(x^\star - \gamma_k \nabla F(x^\star)\big)\|\\
&\le l\|(y_{a,k} - x^\star) - \gamma_k\big(\nabla F(y_{b,k}) - \nabla F(x^\star)\big)\|\\
&\le l\big(\|y_{a,k} - x^\star\| + \gamma_k L\|y_{b,k} - x^\star\|\big)\\
&\le 2\sqrt{s+1}(1 + \gamma_k L)\|d_k\| \le 4\sqrt{s+1}\|d_k\|,
\end{aligned}
\tag{30}
$$

which yields the third estimate. Combining Lemma 9, (29) and (30), we get

$$
\begin{aligned}
(\tau_{k+1}^{-1} \mathrm{P}_{T_{x_{k+1}}} - \mathrm{P}_{T_{x^\star}})\big(\nabla F(y_{b,k}) - \nabla F(x_{k+1})\big) &= o(\|\nabla F(y_{b,k}) - \nabla F(x_{k+1})\|)\\
&= o(\|y_{b,k} - x^\star\|) + o(\|r_{k+1}\|) = o(\|d_k\|).
\end{aligned}
$$

Let's now turn to (28). First, define the function $\overline{R}(x) \stackrel{\text{def}}{=} R(x) + \langle x,\, \nabla F(x^\star)\rangle$. From the smooth perturbation rule of partial smoothness [9, Corollary 4.7], $\overline{R} \in \mathrm{PSF}_{x^\star}(\mathcal{M}_{x^\star})$. Moreover, from Fact 8 and normal sharpness, the Riemannian Hessian of $\overline{R}$ at $x^\star$ is such that, $\forall h \in T_{x^\star}$,

$$
\begin{aligned}
\gamma \nabla^2_{\mathcal{M}_{x^\star}} \overline{R}(x^\star)h &= \gamma \mathrm{P}_{T_{x^\star}} \nabla^2 \widetilde{\overline{R}}(x^\star)h + \gamma \mathfrak{W}_{x^\star}\big(h, \mathrm{P}_{T_{x^\star}^\perp} \nabla \widetilde{\overline{R}}(x^\star)\big)\\
&= \gamma \mathrm{P}_{T_{x^\star}} \nabla^2 \widetilde{R}(x^\star)h + \gamma \mathfrak{W}_{x^\star}\big(h, \mathrm{P}_{T_{x^\star}^\perp} \nabla \widetilde{\Phi}(x^\star)\big)\\
&= \gamma \nabla^2_{\mathcal{M}_{x^\star}} \Phi(x^\star) \mathrm{P}_{T_{x^\star}} h - Hh = Qh,
\end{aligned}
$$

where $\widetilde{\cdot}$ is the smooth representative of the corresponding function. We have

$$\begin{aligned}\lim_{k\to\infty}\frac{\|P(Q_k-Q)r_{k+1}\|}{\|r_{k+1}\|}&=\lim_{k\to\infty}\frac{\|P(\gamma_k-\gamma)\nabla^2_{\mathcal{M}_{x^\star}}\overline{R}(x^\star)\mathrm{P}_{T_{x^\star}}r_{k+1}\|}{\|r_{k+1}\|}\\&\leq\lim_{k\to\infty}|\gamma_k-\gamma|\|P\|\|\nabla^2_{\mathcal{M}_{x^\star}}\overline{R}(x^\star)\mathrm{P}_{T_{x^\star}}\|=0,\end{aligned}$$

which entails $\|P(Q_k-Q)r_{k+1}\|=o(\|r_{k+1}\|)=o(\|d_k\|)$. Similarly, since H is Lipschitz, we have

$$\lim_{k\to\infty}\frac{\|P(G_k-G)r_k\|}{\|r_k\|}=\lim_{k\to\infty}\frac{\|P(\gamma_k-\gamma)Hr_k\|}{\|r_k\|}\leq\lim_{k\to\infty}|\gamma_k-\gamma|L\|P\|=0. \tag{31}$$

Now, let's consider $(M_k-M)d_k$

$$M_k-M=\begin{bmatrix}M_{k,0}-M_0 & M_{k,1}-M_1 & \cdots & M_{k,s-1}-M_{s-1} & M_{k,s}-M_s\\ 0 & 0 & \cdots & 0 & 0\\ 0 & 0 & \cdots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \cdots & 0 & 0\end{bmatrix}.$$

Take $(M_{k,0}-M_0)r_k$, we have

$$\begin{aligned}&(M_{k,0}-M_0)r_k\\&=\big((a_{k,0}-b_{k,0})P+(1+b_{k,0})PG_k\big)r_k-\big((a_0-b_0)P+(1+b_0)PG\big)r_k\\&=\big((a_{k,0}-b_{k,0})-(a_0-b_0)\big)Pr_k+(1+b_{k,0})P(G_k-G)r_k+(b_{k,0}-b_0)PGr_k.\end{aligned}$$

Since we assume that $a_{i,k}\to a_i, b_{i,k}\to b_i, i=0,1$ and $\gamma_k\to\gamma$, plus (31), it can be shown that

$$\begin{aligned}&\lim_{k\to\infty}\frac{\|(M_{k,0}-M_0)r_k\|}{\|r_k\|}\\&\leq\lim_{k\to\infty}|(a_{k,0}-b_{k,0})-(a_0-b_0)|\|P\|+|1+b_{k,0}||\gamma_k-\gamma|L\|P\|+|b_{k,0}-b_0|\|P\|\|G\|=0,\end{aligned}$$

that is $\|(M_{k,0}-M_0)r_k\|=o(\|r_k\|)$. Using the same arguments, we can show that

$$\|(M_{k,i}-M_i)r_{k-i}\|=o(\|r_{k-i}\|),\ i=1,...,s-1\ \text{ and }\ \|(M_{k,s}-M_s)r_{k,s}\|=o(\|r_{k,s}\|).$$

Assemble them together, we obtain

$$\|(M_k-M)d_k\|=o(\|d_k\|),$$

which concludes the proof. ☐

Proof of Proposition 11. From the update (1.7) and the condition for a critical point $x^\star$ of problem $(\mathcal{P})$, we have

$$\begin{aligned}y_{a,k}-x_{k+1}-\gamma_k\big(\nabla F(y_{b,k})-\nabla F(x_{k+1})\big)&\in\gamma_k\partial\Phi(x_{k+1})\\0&\in\gamma_k\partial\Phi(x^\star).\end{aligned}$$

Projecting into $T_{x_{k+1}}$ and $T_{x^\star}$, respectively, and using Fact 8, leads to

$$\begin{aligned}\gamma_k\tau_{k+1}^{-1}\nabla_{\mathcal{M}_{x^\star}}\Phi(x_{k+1})&=\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}\big(y_{a,k}-x_{k+1}-\gamma_k\big(\nabla F(y_{b,k})-\nabla F(x_{k+1})\big)\big)\\\gamma_k\nabla_{\mathcal{M}_{x^\star}}\Phi(x^\star)&=0.\end{aligned}$$

Adding both identities, and subtracting $\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}x^\star$ on both sides, we arrive at

$$\begin{aligned}&\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}r_{k+1}+\gamma_k\big(\tau_{k+1}^{-1}\nabla_{\mathcal{M}_{x^\star}}\Phi(x_{k+1})-\nabla_{\mathcal{M}_{x^\star}}\Phi(x^\star)\big)\\&=\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}(y_{a,k}-x^\star)-\gamma_k\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}\big(\nabla F(y_{b,k})-\nabla F(x_{k+1})\big).\end{aligned} \tag{32}$$

By virtue of Lemma 9, we get

$$\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}r_{k+1}=\mathrm{P}_{T_{x^\star}}r_{k+1}+(\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}-\mathrm{P}_{T_{x^\star}})r_{k+1}=\mathrm{P}_{T_{x^\star}}r_{k+1}+o(\|r_{k+1}\|).$$

Using [11, Lemma 5.1], we also have

$$r_{k+1}=\mathrm{P}_{T_{x^\star}}r_{k+1}+o(\|r_{k+1}\|),$$

and thus

$$\tau_{k+1}^{-1} \mathrm{P}_{T_{x_{k+1}}} r_{k+1} = r_{k+1} + o(\|r_{k+1}\|) = r_{k+1} + o(\|d_k\|), \tag{33}$$

where we also used (27). Similarly

$$\begin{aligned}
&\tau_{k+1}^{-1} \mathrm{P}_{T_{x_{k+1}}} (y_{a,k} - x^\star) \\
&= \mathrm{P}_{T_{x^\star}} (y_{a,k} - x^\star) + (\tau_{k+1}^{-1} \mathrm{P}_{T_{x_{k+1}}} - \mathrm{P}_{T_{x^\star}})(y_{a,k} - x^\star) \\
&= \mathrm{P}_{T_{x^\star}} (y_{a,k} - x^\star) + o(\|y_{a,k} - x^\star\|) = \mathrm{P}_{T_{x^\star}} (y_{a,k} - x^\star) + o(\|d_k\|) \\
&= \mathrm{P}_{T_{x^\star}} (x_k - x^\star) + \sum_{i \in I} a_{i,k} \mathrm{P}_{T_{x^\star}} \big((x_{k-i} - x^\star) - (x_{k-i-1} - x^\star)\big) + o(\|d_k\|) \\
&= r_k + o(\|r_k\|) + \sum_{i \in I} a_{i,k} \big(r_{k-i} - r_{k-i-1} + o(\|r_{k-i}\|) + o(\|r_{k-i-1}\|)\big) + o(\|d_k\|) \\
&= r_k + \sum_{i \in I} a_{i,k} (r_{k-i} - r_{k-i-1}) + \sum_{i \in I \cup \{s\}} o(\|r_{k-i}\|) + o(\|d_k\|) \\
&= (y_{a,k} - x^\star) + o(\|d_k\|).
\end{aligned} \tag{34}$$

Moreover owing to Lemma 10 and (27),

$$\begin{aligned}
\tau^{-1} \nabla_{\mathcal{M}_{x^\star}} \Phi(x_{k+1}) - \nabla_{\mathcal{M}_{x^\star}} \Phi(x^\star) &= \nabla^2_{\mathcal{M}_{x^\star}} \Phi(x^\star) \mathrm{P}_{T_{x^\star}} r_{k+1} + o(\|r_{k+1}\|) \\
&= \nabla^2_{\mathcal{M}_{x^\star}} \Phi(x^\star) \mathrm{P}_{T_{x^\star}} r_{k+1} + o(\|d_k\|).
\end{aligned} \tag{35}$$

Therefore, inserting (33), (34) and (35) into (32), we obtain

$$\begin{aligned}
&\big(\mathrm{Id} + \gamma_k \nabla^2_{\mathcal{M}_{x^\star}} \Phi(x^\star) \mathrm{P}_{T_{x^\star}}\big) r_{k+1} \\
&= (y_{a,k} - x^\star) - \gamma_k \tau_{k+1}^{-1} \mathrm{P}_{T_{x_{k+1}}} \big(\nabla F(y_{b,k}) - \nabla F(x_{k+1})\big) + o(\|d_k\|).
\end{aligned} \tag{36}$$

Owing to (27) and local C^2-smoothness of F, we have

$$\begin{aligned}
&\tau_{k+1}^{-1} \mathrm{P}_{T_{x_{k+1}}} \big(\nabla F(y_{b,k}) - \nabla F(x_{k+1})\big) \\
&= \mathrm{P}_{T_{x^\star}} \big(\nabla F(y_{b,k}) - \nabla F(x_{k+1})\big) + o(\|d_k\|) \\
&= \mathrm{P}_{T_{x^\star}} \big(\nabla F(y_{b,k}) - \nabla F(x^\star)\big) - \mathrm{P}_{T_{x^\star}} \big(\nabla F(x_{k+1}) - \nabla F(x^\star)\big) + o(\|d_k\|) \\
&= \mathrm{P}_{T_{x^\star}} \nabla^2 F(x^\star)(y_{b,k} - x^\star) + o(\|y_{b,k} - x^\star\|) - \mathrm{P}_{T_{x^\star}} \nabla^2 F(x^\star) r_{k+1} + o(\|r_{k+1}\|) + o(\|d_k\|) \\
&= \mathrm{P}_{T_{x^\star}} \nabla^2 F(x^\star) \mathrm{P}_{T_{x^\star}} (y_{b,k} - x^\star) - \mathrm{P}_{T_{x^\star}} \nabla^2 F(x^\star) \mathrm{P}_{T_{x^\star}} (x_{k+1} - x^\star) + o(\|d_k\|).
\end{aligned} \tag{37}$$

Injecting (37) in (36), we get

$$\begin{aligned}
&\big(\mathrm{Id} + \gamma_k \nabla^2_{\mathcal{M}_{x^\star}} \Phi(x^\star) \mathrm{P}_{T_{x^\star}} - \gamma_k \mathrm{P}_{T_{x^\star}} \nabla^2 F(x^\star) \mathrm{P}_{T_{x^\star}}\big) r_{k+1} \\
&= (\mathrm{Id} + Q_k) r_{k+1} = (y_{a,k} - x^\star) - H_k (y_{b,k} - x^\star) + o(\|d_k\|),
\end{aligned} \tag{38}$$

which can be further written as, recall that $H_k = \mathrm{Id} - G_k$,

$$
\begin{aligned}
&(\mathrm{Id} + Q_k)r_{k+1} \\
&= (\mathrm{Id} + Q)r_{k+1} + (Q_k - Q)r_{k+1} \\
&= (y_{a,k} - x^\star) - H_k(y_{b,k} - x^\star) + o(\|d_k\|) \\
&= r_k + \sum_{i\in I} a_{i,k}(r_{k-i} - r_{k-i-1}) - H_k\Big(r_k + \sum_{i\in I} b_{i,k}(r_{k-i} - r_{k-i-1})\Big) + o(\|d_k\|) \\
&= (1 + a_{k,0})r_k - \sum_{i=1}^{s-1}(a_{k,i-1} - a_{k,i})r_{k-i} - a_{k,s-1}r_{k-s} \\
&\qquad - H_k\Big((1 + b_{k,0})r_k - \sum_{i=1}^{s-1}(b_{k,i-1} - b_{k,i})r_{k-i} - b_{k,s-1}r_{k-s}\Big) + o(\|d_k\|) \\
&= (1 + a_{k,0})r_k - \sum_{i=1}^{s-1}(a_{k,i-1} - a_{k,i})r_{k-i} - a_{k,s-1}r_{k-s} \\
&\qquad - (1 + b_{k,0})H_k r_k + H_k\sum_{i=1}^{s-1}(b_{k,i-1} - b_{k,i})r_{k-i} + H_k b_{k,s-1}r_{k-s} + o(\|d_k\|) \\
&= \big((1 + a_{k,0})\mathrm{Id} - (1 + b_{k,0})H_k\big)r_k - (a_{k,s-1}\mathrm{Id} - b_{k,s-1}H_k)r_{k-s} \\
&\qquad - \sum_{i=1}^{s-1}\big((a_{k,i-1} - a_{k,i})\mathrm{Id} - (b_{k,i-1} - b_{k,i})H_k\big)r_{k-i} + o(\|d_k\|) \\
&= \big((a_{k,0} - b_{k,0})\mathrm{Id} + (1 + b_{k,0})G_k\big)r_k - \big((a_{k,s-1} - b_{k,s-1})\mathrm{Id} + b_{k,s-1}G_k\big)r_{k-s} \\
&\qquad - \sum_{i=1}^{s-1}\big((a_{k,i-1} - a_{k,i})\mathrm{Id} - (b_{k,i-1} - b_{k,i})\mathrm{Id} + (b_{k,i-1} - b_{k,i})G_k\big)r_{k-i} + o(\|d_k\|).
\end{aligned}
$$

Inverting $\mathrm{Id} + Q$ (which is possible thanks to assumption (3.2)), we obtain

$$
\begin{aligned}
&r_{k+1} + P(Q_k - Q)r_{k+1} \\
&= \big((a_{k,0} - b_{k,0})P + (1 + b_{k,0})PG_k\big)r_k - \big((a_{k,s-1} - b_{k,s-1})P + b_{k,s-1}PG_k\big)r_{k-s} \\
&\qquad - \sum_{i=1}^{s-1}\big((a_{k,i-1} - a_{k,i})P - (b_{k,i-1} - b_{k,i})P + (b_{k,i-1} - b_{k,i})PG_k\big)r_{k-i} + o(\|d_k\|) \\
&= M_{k,0}r_k + M_{k,s}r_{k-s} + \sum_{i=1}^{s-1} M_{k,i}r_{k-i} + o(\|d_k\|).
\end{aligned}
$$

Using the estimates (28), we get

$$
d_{k+1} = (M + (M_k - M))d_k + o(\|d_k\|) = Md_k + o(\|d_k\|). \qquad \square
$$

With the above result, we are able to prove the claim (3.6), hence Theorem 3.4.

Proof of Theorem 3.4. Since $\rho(M) < 1$, then we have M is convergent with $\lim_{k\to\infty} M^k = 0$. Define $\psi_k = o(d_k)$, suppose after $K > 0$ iterations, (3.5) holds, then for $k \geq K$

$$
d_{k+1} = M^{k+1-K}d_K + \sum_{j=K}^{k} M^{k-j}\psi_j \tag{39}
$$

Since the spectral radius $\rho(M) < 1$, then from the spectral radius formula, given any $\rho \in]\rho(M), 1[$, there exists a constant C such that, for any $k \in \mathbb{N}$

$$
\|M^k\| \leq \|M\|^k \leq C\rho^k.
$$

Therefore, from (39), we get

$$\begin{aligned}\|d_{k+1}\| &\leq \|M^{k+1-K} d_K + \sum_{j=K}^{k} M^{k-j} \psi_j\| \\ &\leq \|M\|^{k+1-K} \|d_K\| + \sum_{j=K}^{k} \|M\|^{k-j} \|\psi_j\| \\ &\leq C\rho^{k+1-K} \|d_K\| + C \sum_{j=K}^{k} \rho^{k-j} \|\psi_j\|.\end{aligned}$$

Together with the fact that $\psi_j = o(\|d_j\|)$ leads to the claimed result. See also the result of [14, Section 2.1.2, Theorem 1]. $\square$

References

[1] P-A. Absil, R. Mahony, and J. Trumpf. An extrinsic look at the Riemannian Hessian. In *Geometric Science of Information*, pages 361–368. Springer, 2013.

[2] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, Forward–Backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

[3] R. I. Boţ, E. R. Csetnek, and S. C. László. An inertial Forward–Backward algorithm for the minimization of the sum of two nonconvex functions. *EURO Journal on Computational Optimization*, pages 1–23, 2014.

[4] I. Chavel. *Riemannian geometry: a modern introduction*, volume 98. Cambridge University Press, 2006.

[5] A. Daniilidis, W. Hare, and J. Malick. Geometrical interpretation of the predictor-corrector type algorithms in structured optimization problems. *Optimization: A Journal of Mathematical Programming & Operations Research*, 55(5-6):482–503, 2009.

[6] D. Drusvyatskiy and A. S. Lewis. Optimality, identifiability, and sensitivity. *Mathematical Programming*, pages 1–32, 2013.

[7] P. Frankel, G. Garrigos, and J. Peypouquet. Splitting methods with variable metric for kurdyka–łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165(3):874–900, 2015.

[8] J. M. Lee. *Smooth manifolds*. Springer, 2003.

[9] A. S. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3):702–725, 2003.

[10] A. S. Lewis and S. Zhang. Partial smoothness, tilt stability, and generalized Hessians. *SIAM Journal on Optimization*, 23(1):74–94, 2013.

[11] J. Liang, J. Fadili, and G. Peyré. Local linear convergence of Forward–Backward under partial smoothness. In *Advances in Neural Information Processing Systems*, pages 1970–1978, 2014.

[12] J. Liang, M. J. Fadili, and G. Peyré. Activity identification and local linear convergence of Forward–Backward-type methods. arXiv:1503.03703, 2015.

[13] S. A. Miller and J. Malick. Newton methods for nonsmooth convex minimization: connections among-Lagrangian, Riemannian Newton and SQP methods. *Mathematical programming*, 104(2-3):609–633, 2005.

[14] B. T. Polyak. *Introduction to optimization*. Optimization Software, 1987.

[15] R. T. Rockafellar and R. Wets. *Variational analysis*, volume 317. Springer Verlag, 1998.