SUPPLEMENTARY MATERIALS

Proof of Theorem 1

In this Appendix we rename the sequence of points (x_k^{md}, x_k^t, x_k) (see listing of the Algorithm 1) to $(\tilde{x}_k, y_k, x_k)$. We use the following definition to simplify calculations.

Definition 3. Let $(\varphi_{\delta,L_\varphi}(x), \nabla\varphi_{\delta,L_\varphi}(x))$ be a (δ, L_φ) - oracle of function φ at a point x, then $\Omega_{1,\delta,L_\varphi}(\varphi, z, x)$ is the following linear function of z:

$$\Omega_{1,\delta,L_\varphi}(\varphi, x, z) = \varphi_{\delta,L_\varphi}(x) + \langle \nabla\varphi_{\delta,L_\varphi}(x), z - x\rangle \tag{115}$$

To prove the Theorem 1, we need the following Theorem 9, which is based on Theorem 2.1 from [Bubeck et al., 2019].

Theorem 9. Let $(y_k)_{k\geq 1}$ — be a sequence in $\mathbb{R}^d$, and $(\lambda_k)_{k\geq 1}$ — a sequence in $\mathbb{R}_+$. Define $(a_k)_{k\geq 1}$ such that $\lambda_k A_k = a_k^2$ and $A_k = \sum_{i=1}^k a_i$. Define also for any $k \geq 0, x_k = x_0 - \sum_{i=1}^k a_i(\nabla\varphi_{\delta,L_\varphi}(y_i) + \nabla\psi_{\delta,L_\psi}(y_i))$ and $\tilde{x}_k := \frac{a_{k+1}}{A_{k+1}} x_k + \frac{A_k}{A_{k+1}} y_k$. Finally assume if for some $\sigma \in [0, 1]$

$$\|y_{k+1} - (\tilde{x}_k - \lambda_{k+1}(\nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1})))\| \leq \sigma \cdot \|y_{k+1} - \tilde{x}_k\|, \tag{116}$$

then one has for any $x \in \mathbb{R}^d$,

$$F(y_k) - F(x) \leq \frac{\|x - x_0\|^2}{2A_k} + 2\left(\sum_{i=1}^k A_i\right)\delta_2/A_k + \delta_1 + \left(\sum_{i=1}^{k-1} A_i\right)\delta_1/A_k, \tag{117}$$

To prove this Theorem we introduce auxiliaries Lemmas based on lemmas 2.2-2.5 and 3.1 from [Bubeck et al., 2019].

Consider a linear combination of gradients:

$$x_k = x_0 - \sum_{i=1}^k a_i(\nabla\varphi_{\delta,L_\varphi}(y_i) + \nabla\psi_{\delta,L_\psi}(y_i))$$

where coefficients $(a_i)_{i\geq 1} \geq 0$ and points $(y_i)_{i\geq 1}$ is not defined yet. A key observation for such a linear combination of gradients is that it minimizes the approximate lower bound of F.

Lemma 5. Let $\xi_0(x) = \frac{1}{2}\|x - x_0\|^2$ and define by induction $\xi_k(x) = \xi_{k-1}(x) + a_k\left(\Omega_{1,\delta,L_\varphi}(\varphi, y_k, x) + \Omega_{1,\delta,L_\psi}(\psi, y_k, x)\right) = \xi_{k-1}(x) + a_k\Omega_{1,2\delta,L_\varphi+L_\psi}(F, y_k, x)$. Then $x_k = x_0 - \sum_{i=1}^k a_i(\nabla\varphi_{\delta,L_\varphi}(y_i) + \nabla\psi_{\delta,L_\psi}(y_i))$ is the minimizer of ξ_k, and $\xi_k(x) \leq A_k F(x) + \frac{1}{2}\|x - x_0\|^2 + A_k\delta_1$, where $A_k = \sum_{i=1}^k a_i$.

Доказательство. Since $\xi_k(x)$ is strongly convex and smooth then expression

$$\nabla\xi_k(x) = 0 \tag{118}$$

is the criterion of minimum.

The sequence x_k is satisfied

$$\nabla\xi_k(x_k) = \nabla\left(\left[\sum_{i=1}^k a_i\Omega_{1,2\delta,L_\varphi+L_\psi}(F, y_k, x)\right] + \frac{1}{2}\|x_k - x_0\|^2\right) = \tag{119}$$

$$= \left[\sum_{i=1}^k a_i\left(\nabla\varphi_{\delta,L_\varphi}(y_i) + \nabla\psi_{\delta,L_\psi}(y_i)\right)\right] + x_k - x_0 = 0. \tag{120}$$

Therefore, x_k is a minimizer of the function ξ_k. Let us prove now that

$$\Omega_{1,2\delta,L_\varphi+L_\psi}(F,y_k,x)\le F(x)+\delta_1. \tag{121}$$

From the definition of $\Omega_{1,2\delta,L_\varphi+L_\psi}(F,y_k,x)$ we obtain

$$\Omega_{1,2\delta,L_\varphi+L_\psi}(F,y_k,x)=F_{2\delta,L_\varphi+L_\psi}(y_i)+\langle\nabla F_{2\delta,L_\varphi+L_\psi}(y_i),x-y_i\rangle\le F(x)+\delta_1. \tag{122}$$

Using $\xi_k(x)=\left[\sum_{i=1}^k a_i\Omega_{1,2\delta,L_\varphi+L_\psi}(F,y_k,x)\right]+\frac{1}{2}\|x-x_0\|^2$ we obtain the statement of the theorem. □

The next idea is to produce a control sequence $(z_k)_{k\ge1}$ demonstrating that ξ_k is not too far below A_kF. From this we can directly yield a convergence rate for z_k.

Lemma 6. Let (z_k) be a sequence such that

$$\xi_k(x_k)-A_kF(z_k)\ge-2\left(\sum_{i=1}^k A_i\right)\delta_2-\left(\sum_{i=1}^{k-1}A_i\right)\delta_1\,. \tag{123}$$

Then one has for any x,

$$F(z_k)\le F(x)+\frac{\|x-x_0\|^2}{2A_k}+2\left(\sum_{i=1}^k A_i\right)\delta_2/A_k+\delta_1+\left(\sum_{i=1}^{k-1}A_i\right)\delta_1/A_k\,. \tag{124}$$

Доказательство. Using Lemma 5 we obtain

$$A_kF(z_k)\le\xi_k(x_k)+2\left(\sum_{i=1}^k A_i\right)\delta_2+\left(\sum_{i=1}^{k-1}A_i\right)\delta_1\le\xi_k(x)+2\left(\sum_{i=1}^k A_i\right)\delta_2+\left(\sum_{i=1}^{k-1}A_i\right)\delta_1 \tag{125}$$

$$\le A_kF(x)+\frac{1}{2}\|x-x_0\|^2+2\left(\sum_{i=1}^k A_i\right)\delta_2+\left(\sum_{i=1}^{k-1}A_i\right)\delta_1+A_k\delta_1\,. \tag{126}$$

□

Our aim now to get sequences (a_k,y_k,z_k), satisfying (123).

Lemma 7. One has for any $x,z_k\in\mathbb{R}^d$ and $k\in\mathbb{N}$

$$\begin{aligned}&\xi_{k+1}(x)-A_{k+1}F(y_{k+1})-(\xi_k(x_k)-A_kF(z_k))\\&\ge A_{k+1}\langle\nabla\varphi_{\delta,L_\varphi}(y_{k+1})+\nabla\psi_{\delta,L_\psi}(y_{k+1}),\frac{a_{k+1}}{A_{k+1}}x+\frac{A_k}{A_{k+1}}z_k-y_{k+1}\rangle+\frac{1}{2}\|x-x_k\|^2-2A_{k+1}\delta_2-A_k\delta_1\,.\end{aligned}$$

Доказательство. Firstly from $H(\xi_k)=I$ using that x_k is a minimizer of $\xi_k(x)$ we get

$$\xi_k(x)=\xi_k(x_k)+\frac{1}{2}\|x-x_k\|^2,$$

and

$$\xi_{k+1}(x)=\xi_k(x_k)+\frac{1}{2}\|x-x_k\|^2+a_{k+1}\Omega_{1,2\delta,L_\varphi+L_\psi}(F,y_{k+1},x)\,,$$

we can rewrite this as follows

$$\xi_{k+1}(x) - \xi_k(x_k) = a_{k+1}\Omega_{1,2\delta,L_\varphi+L_\psi}(F, y_{k+1}, x) + \frac{1}{2}\|x - x_k\|^2 . \tag{127}$$

Now using (14):

$$\Omega_{1,2\delta,L_\varphi+L_\psi}(F, y_{k+1}, z_k) = F_{2\delta,L_\varphi+L_\psi}(y_{k+1}) + \langle \nabla F_{2\delta,L_\varphi+L_\psi}(y_{k+1}), z_k - y_{k+1}\rangle \le F(z_k) + \delta_1 \tag{128}$$

we obtain:

$$\begin{array}{cl}
 & a_{k+1}\Omega_{1,2\delta,L_\varphi+L_\psi}(F, y_{k+1}, x) = A_{k+1}\Omega_{1,2\delta,L_\varphi+L_\psi}(F, y_{k+1}, x) \\
- & A_k\Omega_{1,2\delta,L_\varphi+L_\psi}(F, y_{k+1}, x) = A_{k+1}\Omega_{1,2\delta,L_\varphi+L_\psi}(F, y_{k+1}, x) \\
- & A_k\langle \nabla F_{2\delta,L_\varphi+L_\psi}(y_{k+1}), x - z_k\rangle - A_k\Omega_{1,2\delta,L_\varphi+L_\psi}(F, y_{k+1}, z_k) \\
= & A_{k+1}\Omega_{1,2\delta,L_\varphi+L_\psi}\left(F, y_{k+1}, x - \dfrac{A_k}{A_{k+1}}(x - z_k)\right) - A_k\Omega_{1,2\delta,L_\varphi+L_\psi}(F, y_{k+1}, z_k) \\
= & A_{k+1}F_{2\delta,L_\varphi+L_\psi}(y_{k+1}) + A_{k+1}\langle \nabla F_{2\delta,L_\varphi+L_\psi}(y_{k+1}), \left(x - \dfrac{A_k}{A_{k+1}}(x - z_k)\right) - y_{k+1}\rangle \\
- & A_k\Omega_{1,2\delta,L_\varphi+L_\psi}(F, y_{k+1}, z_k) \overset{(128)}{\ge} A_{k+1}F_{2\delta,L_\varphi+L_\psi}(y_{k+1}) - A_kF(z_k) - A_k\delta_1 \\
+ & A_{k+1}\langle \nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1}), \dfrac{a_{k+1}}{A_{k+1}}x + \dfrac{A_k}{A_{k+1}}z_k - y_{k+1}\rangle \\
\overset{(14)}{\ge} & A_{k+1}F(y_{k+1}) - 2A_{k+1}\delta_2 - A_kF(z_k) - A_k\delta_1 \\
+ & A_{k+1}\langle \nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1}), \dfrac{a_{k+1}}{A_{k+1}}x + \dfrac{A_k}{A_{k+1}}z_k - y_{k+1}\rangle ,
\end{array}$$

which concludes the proof. □

Lemma 8. Denoting

$$\lambda_{k+1} := \frac{a_{k+1}^2}{A_{k+1}} \tag{129}$$

and $\tilde{x}_k := \frac{a_{k+1}}{A_{k+1}}x_k + \frac{A_k}{A_{k+1}}y_k$, one has:

$$\xi_{k+1}(x_{k+1}) - A_{k+1}F(y_{k+1}) - (\xi_k(x_k) - A_kF(y_k)) \ge$$
$$\frac{A_{k+1}}{2\lambda_{k+1}}\left(\|y_{k+1} - \tilde{x}_k\|^2 - \|y_{k+1} - (\tilde{x}_k - \lambda_{k+1}(\nabla\varphi_{\delta,L_\varphi}(y_{k+1})) + \nabla\psi_{\delta,L_\psi}(y_{k+1}))\|^2\right) - 2A_{k+1}\delta_2 - A_k\delta_1 .$$

In particular, we have in light of (116)

$$\xi_k(x_k) - A_kF(y_k) \ge \frac{1-\sigma^2}{2}\sum_{i=1}^{k}\frac{A_i}{\lambda_i}\|y_i - \tilde{x}_{i-1}\|^2 - 2\left(\sum_{i=1}^{k}A_i\right)\delta_2 - \left(\sum_{i=1}^{k-1}A_i\right)\delta_1.$$

Доказательство. We apply Lemma 7 with $z_k = y_k$ and $x = x_{k+1}$, and note that (with $\tilde{\zeta} := \frac{a_{k+1}}{A_{k+1}}x + \frac{A_k}{A_{k+1}}y_k$):

$$\langle \nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1}), \frac{a_{k+1}}{A_{k+1}}x + \frac{A_k}{A_{k+1}}y_k - y_{k+1}\rangle + \frac{1}{2A_{k+1}}\|x - x_k\|^2$$
$$= \langle \nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1}), \tilde{\zeta} - y_{k+1}\rangle + \frac{1}{2A_{k+1}}\left\|\frac{A_{k+1}}{a_{k+1}}\left(\tilde{\zeta} - \frac{A_k}{A_{k+1}}y_k\right) - x_k\right\|^2$$
$$= \langle \nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1}), \tilde{\zeta} - y_{k+1}\rangle + \frac{A_{k+1}}{2a_{k+1}^2}\left\|\tilde{\zeta} - \left(\frac{a_{k+1}}{A_{k+1}}x_k + \frac{A_k}{A_{k+1}}y_k\right)\right\|^2 .$$

This yields, using (129):

$$\begin{aligned}
&\xi_{k+1}(x_{k+1}) - A_{k+1}F(y_{k+1}) - (\xi_k(x_k) - A_kF(y_k))\\
&\geq A_{k+1}\cdot\langle\nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1})), \tilde{\zeta} - y_{k+1})\rangle + \frac{A_{k+1}}{2\lambda_{k+1}}\|\tilde{\zeta} - \tilde{x}_k\|^2 - 2A_{k+1}\delta_2 - A_k\delta_1\\
&\geq A_{k+1}\cdot\min_{x\in\mathbb{R}^d}\left\{\langle\nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1}), x - y_{k+1}\rangle + \frac{1}{2\lambda_{k+1}}\|x - \tilde{x}_k\|^2\right\} - 2A_{k+1}\delta_2 - A_k\delta_1\,.
\end{aligned}$$

The value of the minimum is easy to compute. Due to the strong convexity of the minimized function and its continuous differentiability, achieving a minimum is equivalent to the condition

$$\begin{aligned}
0 &= \nabla\left[\langle\nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1}), x - y_{k+1}\rangle + \frac{1}{2\lambda_{k+1}}\|x - \tilde{x}_k\|^2\right] =\\
&= (\nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1})) + \frac{1}{\lambda_{k+1}}(x - \tilde{x}_k)
\end{aligned}$$

Then

$$x_* = \tilde{x}_k - \lambda_{k+1}(\nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1}))$$

Substituting into the last inequality we obtain the statement of the theorem. □

Proof of the Theorem 9

Using Lemma 8 we get

$$\begin{aligned}
\xi_k(x_k) - A_kF(y_k) &\geq \frac{1-\sigma^2}{2}\sum_{i=1}^{k}\frac{A_i}{\lambda_i}\|y_i - \tilde{x}_{i-1}\|^2 - 2\left(\sum_{i=1}^{k}A_i\right)\delta_2 - \left(\sum_{i=1}^{k-1}A_i\right)\delta_1\\
&\geq -2\left(\sum_{i=1}^{k}A_i\right)\delta_2 - \left(\sum_{i=1}^{k-1}A_i\right)\delta_1.
\end{aligned}$$

Applying Lemma 6 for $z_k = y_k$ one has for any $x\in\mathbb{R}^d$:

$$F(y_k) - F(x) \leq \frac{\|x - x_0\|^2}{2A_k} + 2\left(\sum_{i=1}^{k}A_i\right)\delta_2/A_k + \delta_1 + \left(\sum_{i=1}^{k-1}A_i\right)\delta_1/A_k\,, \tag{130}$$

□

that conclude the proof.

Now one will formulate the sufficient condition (16) for the accuracy of solving auxiliary problems (15). Let us assume, that auxiliary problems (15) can not be solved exactly. Let the algorithm only have an inaccurate solution y_{k+1} satisfying

$$\begin{aligned}
(16) :&\|\nabla\left(\Omega_{1,\delta,L_\varphi}(\varphi, \tilde{x}_k, y_{k+1}) + \frac{H}{2}\|y_{k+1} - \tilde{x}_k\|^2\right) + \nabla\psi_{\delta,L_\psi}(y_{k+1})\|\\
&\leq \frac{H}{4}\|y_{k+1} - \tilde{x}_k\| - 2\sqrt{2\delta_2L_\varphi}
\end{aligned}$$

in this case:

Lemma 9. Assume that $\varphi(x)$ has (δ, L_φ) -oracle, $\psi(x)$ has (δ, L_ψ) -oracle and the auxiliary subproblem (15) is solved inexactly in such a way that the inequality (16) holds. If

$$H \geq 2L_\varphi$$

then equation (116) holds true with $\sigma = 7/8$ for (15). In the case $p = 1$ one can consider $\lambda_{k+1} = \lambda = \frac{1}{2H}$.

Доказательство.

Using that φ is equipped with a (δ, L_φ) -oracle and Corollary 4.2. from [Devolder, 2013] one obtains:

$$\|\nabla\varphi_{\delta,L_\varphi}(y) - \nabla_y\Omega_{1,\delta,L_\varphi}(\varphi, x, y)\| \leq L_\varphi\|y - x\| + 2\sqrt{2L_\varphi\delta_2}\,. \tag{131}$$

By (16) and (131) we can get next inequalities:

$$\begin{aligned}
&\|y_{k+1} - (\tilde{x}_k - \lambda_{k+1}(\nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1})))\| \\
&= \|y_{k+1} \pm \lambda_{k+1}\nabla_y\Omega_{1,\delta,L_\varphi}(\varphi, \tilde{x}_k, y_{k+1}) \pm H\lambda_{k+1}(y_{k+1} - \tilde{x}_k) \\
&- (\tilde{x}_k - \lambda_{k+1}(\nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1})))\| \leq (1 - H\lambda_{k+1})\|(y_{k+1} - \tilde{x}_k)\| \\
&+ \lambda_{k+1}\|\nabla_y\Omega_{1,\delta,L_\varphi}(\varphi, \tilde{x}_k, y_{k+1}) + H(y_{k+1} - \tilde{x}_k) + \nabla\psi_{\delta,L_\psi}(y_{k+1})\| \\
&+ \lambda_{k+1}\|\nabla\varphi_{\delta,L_\varphi}(y_{k+1}) - \nabla_y\Omega_{1,\delta,L_\varphi}(\varphi, \tilde{x}_k, y_{k+1})\| \overset{(16),(131)}{\leq} (1 - H\lambda_{k+1})\|y_{k+1} - \tilde{x}_k\| \\
&+ \lambda_{k+1}\frac{H}{4}\|y_{k+1} - \tilde{x}_k\| - 2\lambda_{k+1}\sqrt{2L_\varphi\delta_2} + \lambda_{k+1}\left(L_\varphi\|y_{k+1} - \tilde{x}_k\| + 2\sqrt{2L_\varphi\delta_2}\right) \\
&\leq \left(\frac{5}{8} + \frac{L_\varphi}{2H}\right)\|y_{k+1} - \tilde{x}_k\| \leq \frac{7}{8}\|y_{k+1} - \tilde{x}_k\|
\end{aligned}$$

that ends the proof. □

Recall from Lemma 6 that the rate of convergence of AM-1 is $\|x_0 - x^*\|/A_k + 2\left(\sum_{i=1}^k A_i\right)\delta_2/A_k + \delta_1 + \left(\sum_{i=1}^{k-1} A_i\right)\delta_1/A_k$. We now finally give an estimate of A_k:

Lemma 10. Suppose $H \geq 2L_\varphi$. Then one has, with $c_1 = 4$,

$$A_k \geq \frac{k^2}{c_1 H} \tag{132}$$

Доказательство. In case, when $p = 1$ λ_{k+1} are defined as

$$\lambda_{k+1} = \frac{1}{2H}.$$

Inequality (132) holds when $k = 1$.

Let us proof that if (132) holds for k then it holds for $k + 1$. Using the definition of A_k

$$a_{k+1} = \frac{\lambda_{k+1} + \sqrt{\lambda_{k+1}^2 + 4\lambda_{k+1}A_k}}{2}\ , \ A_{k+1} = A_k + a_{k+1},$$

we obtain

$$A_{k+1} \geq \frac{k^2}{2c_1L_\varphi} + \frac{1}{8L_\varphi}\left(1 + \sqrt{1 + \frac{16k^2}{c_1}}\right) \geq \frac{(k+1)^2}{2c_1L_\varphi}.$$

□

Proof of Theorem 1

To prove the Theorem 1 it suffices to combine Lemmas 9,10 with Theorem 9. □

Proof of Theorem 2

In this Appendix we rename the sequence of points (x_k^{md}, x_k^t, x_k) (see listing of the Algorithm 1) to $(\tilde{x}_k, y_k, x_k)$.

Доказательство. Firstly, let us choose δ according to (19):

$$\forall k : \delta_1 + \delta_2 + 2\left(\sum_{i=1}^{k} A_i\right)\delta_2/A_k + 2\left(\sum_{i=1}^{k-1} A_i\right)\delta_1/A_k \le \frac{\varepsilon}{2},$$

where ε is solution accuracy in terms of $F(x) - F(x_*) \le \varepsilon$.

Then, as $\varepsilon/2 \le c_1 HR^2/k^2$ with $c_1 = 4$, next inequality holds true

$$\forall k : \delta_1 + \delta_2 + 2\left(\sum_{i=1}^{k} A_i\right)\delta_2/A_k + 2\left(\sum_{i=1}^{k-1} A_i\right)\delta_1/A_k \le \frac{c_1 HR^2}{k^2}.$$

From (δ, L, μ) - oracle definition (7) we get

$$\begin{aligned} \frac{\mu}{2}\|z - x_*\|^2 - \delta_1 \le F(z) - (F_{\delta,L,\mu}(x_*) + \langle \nabla F_{\delta,L,\mu}(x_*), z - x_*\rangle) = \\ = (F(z) - F_{\delta,L,\mu}(x_*)) + \langle \partial F(x_*) - \nabla F_{\delta,L,\mu}(x_*), z - x_*\rangle - \langle \partial F(x_*), z - x_*\rangle \le \\ \le (F(z) - F(x_*) + \delta_2) + \sqrt{2\delta_2 L}\|z - x_*\|. \end{aligned} \tag{133}$$

Therefore

$$\frac{\mu}{2}\|z - x_*\|^2 - \sqrt{2\delta_2 L}\|z - x_*\| \le (F(z) - F(x_*) + \delta_1 + \delta_2).$$

If δ_2 is small enough such that

$$\frac{4\sqrt{2\delta_2 L}}{\mu} \le \varepsilon/2,$$

then taking into account that $\forall k : \varepsilon/2 \le \|z_{k+1} - x_*\|$ we obtain

$$\forall k : \frac{\mu}{4}\|z_{k+1} - x_*\|^2 \ge \sqrt{2\delta_2 L}\|z_{k+1} - x_*\|, \tag{134}$$

which implies the following inequality

$$\frac{\mu}{4}\|z_k - x_*\|^2 \le (F(z_k) - F(x_*) + \delta_1 + \delta_2). \tag{135}$$

Finally, we can conclude that R_k decreases as a geometric progression:

$$\begin{aligned} R_{k+1} = \|z_{k+1} - x_*\| \overset{(135)}{\le} \left(\frac{4\left(F(z_{k+1}) - F(x_*) + \delta_1 + \delta_2\right)}{\mu}\right)^{\frac{1}{2}} \\ \overset{(17)}{\le} \left(\frac{4\left(\frac{c_1 HR_k^2}{N_k^2} + 2\left(\sum_{i=1}^{k} A_i\right)\delta_1/A_k + 2\left(\sum_{i=1}^{k-1} A_i\right)\delta_2/A_k + \delta_1 + \delta_2\right)}{\mu}\right)^{\frac{1}{2}} \\ \overset{(19)}{\le} \left(\frac{4\left(\frac{2c_1 HR_k^2}{N_k^2}\right)}{\mu}\right)^{\frac{1}{2}} = \left(\frac{8c_1 HR_k^2}{\mu N_k^2}\right)^{\frac{1}{2}} \overset{(18)}{\le} \left(\frac{R_k^2}{2^2}\right)^{\frac{1}{2}} = \frac{R_k}{2}. \end{aligned}$$

Which in turn guarantees that

$$F(z_K) - F(x_*) \leq \frac{\mu R_0^2}{4 \cdot 4^K}. \tag{136}$$

It is sufficient to choose $K = 2\log_2 \frac{\mu R_0^2}{4\varepsilon}$ in order that $F(z_k) - F(x_*) \leq \varepsilon$.

Now we compute the total number of AM steps.

$$\sum_{k=0}^{K} N_k \leq \sum_{k=0}^{K} \left(\frac{32c_1H}{\mu}\right)^{\frac{1}{2}} + K \leq \sum_{k=0}^{K} \left(\frac{32c_1H}{\mu}\right)^{\frac{1}{2}} + K$$
$$= \left(\frac{32c_1H}{\mu}\right)^{\frac{1}{2}} K + K = \left(\sqrt{\frac{128H}{\mu}} + 1\right) \cdot 2\log_2 \frac{\mu R_0^2}{4\varepsilon} \leq \left(16\sqrt{2}\sqrt{\frac{H}{\mu}} + 2\right) \log_2 \frac{\mu R_0^2}{\varepsilon}$$

□

Proof of Theorem 3 and Theorem 4

The Theorem 3 show that the fulfillment of condition (16) keep the linear rate of convergence when solving the auxiliary problems (15). Also in this Appendix we rename the sequence of points (x_k^{md}, x_k^t, x_k) (see listing of the Algorithm 1) to $(\tilde{x}_k, y_k, x_k)$.
Firstly, based on (16) we try to relate the accuracy $\tilde{\varepsilon}$ we need to solve (15) in terms of the following criteria:

$$\|\nabla\left(\Omega_{1,\delta,L_\varphi}(\varphi, \tilde{x}_k, y_{k+1}) + \frac{H}{2}\|y_{k+1} - \tilde{x}_k\|^2\right) + \nabla\psi_{\delta,L_\psi}(y_{k+1})\| \leq \tilde{\varepsilon}. \tag{137}$$

For this we prove the auxiliary lemma for (δ, L_φ) -oracle of φ and (δ, L_ψ) -oracle of ψ, that is based on the Lemma 2.1 from [Grapiglia, Nesterov, 2020] .

Lemma 11. Let $\tilde{x}_k \in \mathbb{R}^d, H, \Theta > 0$.
Assume that $\varphi(x)$ admits (δ, L_φ) -oracle, $\psi(x)$ admits (δ, L_ψ) -oracle. If inquality

$$\|\nabla\left(\Omega_{1,\delta,L_\varphi}(\varphi, \tilde{x}_k, y_{k+1}) + \frac{H}{2}\|y_{k+1} - \tilde{x}_k\|^2\right) + \nabla\psi_{\delta,L_\psi}(y_{k+1})\| \tag{138}$$
$$\leq \min\left\{\frac{1}{2}, \frac{\Theta}{2[L_\varphi + H]}\right\}\left(\|\nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1})\|\right) \tag{139}$$

holds true, then y_{k+1} satisfies

$$\|\nabla\left(\Omega_{1,\delta,L_\varphi}(\varphi, \tilde{x}_k, y_{k+1}) + \frac{H}{2}\|y_{k+1} - \tilde{x}_k\|^2\right) + \nabla\psi_{\delta,L_\psi}(y_{k+1})\| \tag{140}$$
$$\leq \Theta\|y_{k+1} - \tilde{x}_k\| + \frac{2\Theta}{[L_\varphi + H]}\sqrt{2L_\varphi\delta_2} \tag{141}$$

Доказательство. Using that φ is equipped with a (δ, L_φ) -oracle and Corollary 4.2. from [Devolder, 2013] one obtains:

$$\|\nabla\varphi_{\delta,L_\varphi}(y) - \nabla_y\Omega_{1,\delta,L_\varphi}(\varphi, x, y)\| \leq L_\varphi\|y - x\| + 2\sqrt{2L_\varphi\delta_2}\,. \tag{142}$$

Combining (138) and (142) we obtain

$$\begin{aligned}
&\|\nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1})\| \\
&\leq \|\nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1}) - \nabla\psi_{\delta,L_\psi}(y_{k+1}) - \nabla\Omega_{1,\delta,L_\varphi}(\varphi,\tilde{x}_k,y_{k+1})\| \\
&+ \|\nabla\Omega_{1,\delta,L_\varphi}(\varphi,\tilde{x}_k,y_{k+1}) \pm \nabla\psi_{\delta,L_\psi}(y_{k+1}) - \nabla\left(\Omega_{1,\delta,L_\varphi}(\varphi,\tilde{x}_k,y_{k+1}) + \frac{H}{2}\|y_{k+1}-\tilde{x}_k\|^2\right)\| \\
&+ \|\nabla\left(\Omega_{1,\delta,L_\varphi}(\varphi,\tilde{x}_k,y_{k+1}) + \frac{H}{2}\|y_{k+1}-\tilde{x}_k\|^2\right) + \nabla\psi_{\delta,L_\psi}(y_{k+1})\| \\
&\overset{(142),(138)}{\leq} \left(L_\varphi\|y_{k+1}-\tilde{x}_k\| + 2\sqrt{2L_\varphi\delta_2}\right) + H\|y_{k+1}-\tilde{x}_k\| + \frac{1}{2}\|\nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1})\|.
\end{aligned}$$

Thus,

$$\frac{\|\nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1})\|}{2} \leq [L_\varphi + H]\,\|y_{k+1}-\tilde{x}_k\| + 2\sqrt{2L_\varphi\delta_2} \tag{143}$$

which gives

$$\frac{\Theta}{2\,[L_\varphi + H]}\|\nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1})\| \leq \Theta\|y_{k+1}-\tilde{x}_k\| + \frac{2\Theta}{[L_\varphi + H]}\sqrt{2L_\varphi\delta_2} \tag{144}$$

Finally, (140) follows directly from the (138) and (144). □

Lemma 12. Assume that $H \geq 2L_\varphi$, $\varphi(x)$ admits (δ, L_φ)-oracle, $\psi(x)$ admits (δ, L_ψ)-oracle, $F(x)$ admits $(2\delta, L_\varphi + L_\psi, \mu)$-oracle; $y_{k+1}, \tilde{x}_k \in \mathbb{R}^d$ and $\varepsilon \in (0,1)$. If inequalities

$$\varepsilon \leq F_{2\delta,L_\varphi+L_\psi,\mu}(y) - \min_{x\in Q_f} F(x) \tag{145}$$

$$\delta_2 \leq \frac{\varepsilon\mu}{64^2\cdot L_\varphi}, \tag{146}$$

are satisfied then inequality (16) holds true if one solve the auxiliary problem (15) with the accuracy

$$\tilde{\varepsilon} = \frac{\sqrt{\varepsilon\mu}}{72} \tag{147}$$

in terms of criteria (137).

Доказательство. According to the conditions of the lemma, the problem (15) is solved with the accuracy

$$(137):\ \|\nabla\left(\Omega_{1,\delta,L_\varphi}(\varphi,\tilde{x}_k,y_{k+1}) + \frac{H}{2}\|y_{k+1}-\tilde{x}_k\|^2\right) + \nabla\psi_{\delta,L_\psi}(y_{k+1})\| \leq \tilde{\varepsilon}$$

To prove the lemma, it suffices to show the following chain of inequalities

$$\begin{aligned}
&\tilde{\varepsilon} \leq \min\{\frac{1}{2}, \frac{H}{8\,[L_\varphi + H]}\}\left(\|\nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1})\|\right) - \left(2 + \frac{H}{2\,[L_\varphi + H]}\right)\sqrt{2\delta_2 L_\varphi} \\
&\leq \frac{H}{4}\|y_{k+1}-\tilde{x}_k\| - 2\sqrt{2\delta_2 L_\varphi}
\end{aligned} \tag{148}$$

Lemma 11 for $\Theta = \frac{H}{4}$ guarantee that if the next inequality holds true

$$\begin{aligned}
&\|\nabla\left(\Omega_{1,\delta,L_\varphi}(\varphi,\tilde{x}_k,y_{k+1}) + \frac{H}{2}\|y_{k+1}-\tilde{x}_k\|^2\right) + \nabla\psi_{\delta,L_\psi}(y_{k+1})\| \\
&\leq \min\{\frac{1}{2}, \frac{H}{8\,[L_\varphi + H]}\}\left(\|\nabla\varphi_{\delta,L_\varphi}(y_{k+1}) + \nabla\psi_{\delta,L_\psi}(y_{k+1})\|\right) - \left(2 + \frac{H}{2\,[L_\varphi + H]}\right)\sqrt{2\delta_2 L_\varphi}
\end{aligned} \tag{149}$$

then the equation (16) is satisfied.
If (149) is sufficient condition for (16), it means that right-hand sides of (149) less the right-hand sides of (16). From this consequence the next inequality

$$\frac{1}{12}\left(\|\nabla F_{2\delta,L_\varphi+L_\psi}(y_{k+1})\|\right)-\frac{5}{2}\sqrt{2\delta_2 L_\varphi}=\frac{1}{12}\left(\|\nabla\varphi_{\delta,L_\varphi}(y_{k+1})+\nabla\psi_{\delta,L_\psi}\left(y_{k+1}\right)\|\right)-\frac{5}{2}\sqrt{2\delta_2 L_\varphi} \tag{150}$$

$$\leq \min\{\frac{1}{2},\frac{H}{8\left[L_\varphi+H\right]}\}\left(\|\nabla\varphi_{\delta,L_\varphi}(y_{k+1})+\nabla\psi_{\delta,L_\psi}\left(y_{k+1}\right)\|\right)-\left(2+\frac{H}{2\left[L_\varphi+H\right]}\right)\sqrt{2\delta_2 L_\varphi} \tag{151}$$

$$\leq \frac{H}{4}\|y_{k+1}-\tilde{x}_k\|-2\sqrt{2\delta_2 L_\varphi} \tag{152}$$

The second inequality of the equation (148) is satisfied, let us prove the first one.
The fact that F has $(2\delta, L_\varphi+L_\psi,\mu)$ -oracle guarantee

$$\frac{\mu}{2}\|x-y\|^2+\left(F_{2\delta,L_\varphi+L_\psi,\mu}(y)+\left\langle\nabla F_{2\delta,L_\varphi+L_\psi,\mu}(y),x-y\right\rangle\right)\leq F(x) \text{ for all } x\in Q_f \tag{153}$$

Let us minimize the right-hand and left-hand sides of (153) with respect to x independently

$$F^*=\min_{x\in Q_f}F(x)\overset{(153)}{\geq}F_{2\delta,L_\varphi+L_\psi,\mu}(y)+\min_{x\in Q_f}\left\{\frac{\mu}{2}\|x-y\|^2+\left\langle\nabla F_{2\delta,L_\varphi+L_\psi,\mu}(y),x-y\right\rangle\right\}$$
$$=F_{2\delta,L_\varphi+L_\psi,\mu}(y)-\frac{1}{2\mu}\|\nabla F_{2\delta,L_\varphi+L_\psi,\mu}(y)\|^2$$

Then obtain

$$\varepsilon\overset{(145)}{\leq}F_{2\delta,L_\varphi+L_\psi,\mu}(y)-F^*\leq\frac{1}{2\mu}\|\nabla F_{2\delta,L_\varphi+L_\psi,\mu}(y)\|^2 \tag{154}$$

Inequality (154) guarantee that

$$\frac{1}{2}\sqrt{\frac{\varepsilon\mu}{18}}-\frac{5}{2}\sqrt{2\delta_2 L_\varphi}\leq\frac{1}{12}\|\nabla F_{2\delta,L_\varphi+L_\psi,\mu}(y)\|-\frac{5}{2}\sqrt{2\delta_2 L_\varphi} \tag{155}$$

In case of (146) inequality (155) give us guarantees that the first inequality of the equation (148) holds true

$$\tilde{\varepsilon}=\frac{\sqrt{\varepsilon\mu}}{72}\overset{(146)}{\leq}\frac{1}{2}\sqrt{\frac{\varepsilon\mu}{18}}-\frac{5}{2}\sqrt{2\delta_2 L_\varphi}\overset{(155)}{\leq}\frac{1}{12}\|\nabla F_{2\delta,L_\varphi+L_\psi,\mu}(y)\|-\frac{5}{2}\sqrt{2\delta_2 L_\varphi} \tag{156}$$

Finally, combine the equations (152) and (156) obtain the required chain of inequalities (148).

□

Let us prove the Theorem 3 using Lemma 12:

Доказательство. Firstly, let us collect all restrictions on δ_1,δ_2 and auxiliary problem precision for obtaining convergence of outer Algorithm-2 and fulfillment of the Lemma 12 together:

$$(137):\|\nabla\left(\Omega_{1,\delta,L_\varphi}\left(\varphi,\tilde{x}_k,y_{k+1}\right)+\frac{H}{2}\left\|y_{k+1}-\tilde{x}_k\right\|^2\right)+\nabla\psi_{\delta,L_\psi}\left(y_{k+1}\right)\|\leq\tilde{\varepsilon},$$

$$(146):\delta_2\leq\frac{\varepsilon\mu}{64^2\cdot L_\varphi},$$

$$(19):\forall k:\delta_1+\delta_2+2\left(\sum_{i=1}^{k}A_i\right)\delta_2/A_k+\left(\sum_{i=1}^{k-1}A_i\right)\delta_1/A_k\leq\frac{\varepsilon}{2},$$

$$(20):\frac{4\sqrt{2\delta_2 L}}{\mu}\leq\varepsilon/2.$$

Let us have a look at (137). We need obtain the sufficient condition for it in terms of the criterion (22).

$$\begin{aligned}
&\|\nabla\left(\Omega_{1,\delta,L_\varphi}(\varphi,\tilde{x}_k,y_{k+1})+\frac{H}{2}\|y_{k+1}-\tilde{x}_k\|^2\right)+\nabla\psi_{\delta,L_\psi}(y_{k+1})\|\\
\le&\|\nabla\Omega_{1,\delta,L_\varphi}(\varphi,\tilde{x}_k,y_{k+1})\pm\partial\varphi(\tilde{x}_*)\|+H\|y_{k+1}-\tilde{x}_k\|+\|\nabla\psi_{\delta,L_\psi}(y_{k+1})\pm\partial\psi(\tilde{x}_*)\|\\
\le&L_\varphi\|\tilde{x}_k-x_*\|+2\sqrt{L_\varphi\delta_2}+H\|y_{k+1}-\tilde{x}_k\|+L_\psi\|y_{k+1}-x_*\|+2\sqrt{L_\psi\delta_2}\\
\le&(L_\varphi+L_\psi+H)\max\left\{\|\tilde{x}_k-x_*\|,\|y_{k+1}-x_*\|\right\}+2\sqrt{L_\varphi\delta_2}+2\sqrt{L_\psi\delta_2}\\
\overset{(135)}{\le}&(L_\varphi+L_\psi+H)\sqrt{\frac{4(\tilde{\varepsilon}_f+\delta_1+\delta_2)}{\mu}}+2\sqrt{L_\varphi\delta_2}+2\sqrt{L_\psi\delta_2}.
\end{aligned}$$

Then, according to (137), the sufficient condition for (16) holds true is

$$(L_\varphi+L_\psi+H)\sqrt{\frac{4(\tilde{\varepsilon}_f+\delta_1+\delta_2)}{\mu}}+2\sqrt{L_\varphi\delta_2}+2\sqrt{L_\psi\delta_2}\le\frac{\sqrt{\varepsilon\mu}}{72}.$$

Next, under the assumption $\delta_1\le\delta_2$, (19) is converting into more simple sufficient condition

$$\delta_2\le\frac{\varepsilon}{2\left(1+4N\right)}\le\frac{\varepsilon}{2\left(1+4\left(\sum_{i=1}^k A_i\right)/A_k\right)} \tag{157}$$

where N is the number of outer steps. There was used the fact that $A_i\le A_{i+1}$. Finally, if δ_2 satisfies the inequality

$$\delta_2\le\frac{\varepsilon^{3/2}}{5\sqrt{2c_1HR^2}}$$

then (157) holds true.

If we choose $\delta_1,\delta_2,\tilde{\varepsilon}_f$ such that:

$$\begin{aligned}
&\delta_1,\delta_2=\min\left\{\frac{\varepsilon\mu}{864^2L_\varphi},\frac{\varepsilon\mu}{864^2L_\psi},\frac{\varepsilon\mu^2}{864^2(L_\varphi+L_\psi+H)^2},\frac{\varepsilon^{3/2}}{5\sqrt{2c_1HR^2}}\right\},\\
&\tilde{\varepsilon}_f\le\frac{\varepsilon\mu^2}{864^2(L_\varphi+L_\psi+H)^2},
\end{aligned}$$

then all required inequalities are satisfied:

$$\begin{aligned}
&(L_\varphi+L_\psi+H)\sqrt{\frac{4(\tilde{\varepsilon}_f+\delta_1+\delta_2)}{\mu}}+2\sqrt{L_\varphi\delta_2}+2\sqrt{L_\psi\delta_2}\le\frac{\sqrt{\varepsilon\mu}}{72},\\
&\delta_2\le\frac{\varepsilon\mu}{64^2\cdot L_\varphi},\\
&\delta_2\le\frac{\varepsilon^{3/2}}{5\sqrt{2c_1HR^2}},\\
&\frac{4\sqrt{2\delta_2L}}{\mu}\le\varepsilon/2.
\end{aligned}$$

Also dependences $\delta_1(\varepsilon),\delta_2(\varepsilon),\tilde{\varepsilon}_f(\varepsilon)$ are polynomial. □

Let's prove the Theorem 4 using Theorem 3:

Доказательство. Suppose that at each iteration of the Algorithm 2 one have:

1. inexact $(\delta, \sigma_0, \mu_\varphi, L_\varphi), (\delta, \sigma_0, \mu_\psi, L_\psi)$-oracles of φ, ψ;
2. the (ε, σ_0)-solution of auxiliary problem.

Let us estimate the probability $\mathbb{P}$ with which inexact $(\delta, \mu_\varphi, L_\varphi), (\delta, \mu_\psi, L_\psi)$-oracles of φ, ψ and the ε-solution of auxiliary problem will be available at all iterations (36) of the Algorithm 2

$$\mathbb{P} = (1-\sigma_0)^{N(\varepsilon)}(1-\tilde{\sigma})^{N(\varepsilon)} \geq 1 - N(\varepsilon)\left(\sigma_0 + \tilde{\sigma}\right) \overset{(32),(35),(36)}{\geq} 1 - \sigma \tag{158}$$

Hence with probability (158) the conditions of the Theorem 3 are satisfied which ends the proof. □

L-SVRG

In this Appendix we reformulate the convergence results of Algorithm L-SVRG from [Morin, Giselsson, 2020] in terms of large deviations.

Lemma 13. (Corollary 5.6 from [Morin, Giselsson, 2020]) We consider the problem

$$\min_{x \in \mathbb{R}^d} F(x) = \varphi(x) + \psi(x) \tag{159}$$

where φ is of finite sum form

$$\varphi(x) = \frac{1}{n}\sum_{i=1}^{n} \varphi_i(x)$$

and ψ is L_ψ-smooth, convex and prox-friendly. The function φ_i is convex and L_i-smooth for all $i = 1, \ldots, n$. The function φ is convex, L-smooth with $L \leq \frac{1}{n}\sum_{i=1}^{n} L_i$ an μ-strongly convex. Then L-SVRG [Morin, Giselsson, 2020] achieves an (ε, σ)-solution of (159), i.e.

$$\mathbb{P}\{F(x_k) - F(x_*) \geq \varepsilon\} \leq \sigma \tag{160}$$

within

$$O\left(\left(\sqrt{n} + \sqrt{2D_L\frac{\bar{L}}{\mu}}\right)^2 \log\frac{1}{\epsilon\sigma}\right)$$

iterations where $\bar{L} = \frac{1}{n}\sum_{i=1}^{n} L_i$, $D_L = 4 - 3\frac{\mu}{L}$ and x_* is solution of (159). We note that $1 \leq D_L \leq 4$.

Доказательство. According to Corollary 5.6 from [Morin, Giselsson, 2020] we obtain that after $O\left(\left(\sqrt{n} + \sqrt{2D_L\frac{\bar{L}}{\mu}}\right)^2 \log\frac{1}{\epsilon'}\right)$ steps L-SVRG [Morin, Giselsson, 2020] gives ε' accurate solution, i.e.

$$\mathbb{E}\left\|x^k - x_*\right\|^2 \leq \epsilon', \tag{161}$$

holds true. For arbitrary $\varepsilon, \sigma > 0$ let us take x_k, $\varepsilon' = 2\varepsilon\sigma/L_F$ accurate solution in terms of (161). Then from $L_F = L + L_\psi$-smoothness of F we have

$$\mathbb{E}[F(x_k) - F(x_*)] \leq \mathbb{E}\left[\frac{L_F}{2}\|x_k - x_*\|^2\right] \leq \varepsilon\sigma. \tag{162}$$

Using Markov inequality and (162) we obtain that

$$\mathbb{P}\{F(x_k) - F(x_*) \geq \varepsilon\} \leq \frac{\mathbb{E}[F(x_k) - F(x_*)]}{\varepsilon} \leq \sigma. \tag{163}$$

In other words, after

$$O\left(\left(\sqrt{n} + \sqrt{2D_L\frac{\bar{L}}{\mu}}\right)^2 \log\frac{1}{\epsilon'}\right) = O\left(\left(\sqrt{n} + \sqrt{2D_L\frac{\bar{L}}{\mu}}\right)^2 \log\frac{1}{\epsilon\sigma}\right)$$

Algorithm L-SVRG from [Morin, Giselsson, 2020] gives random point x_k such as (163) holds true. In other words, x_k is (ε, σ)-solution of (159). □

A Variant of Accelerated Framework for Saddle-Point Problems.

In this appendix we consider saddle-point problem under the same assumptions as in Section 1. We describe in detail the structure of a general framework for solving such problems which consists of three inner-outer loops. The only difference compared with the general framework in Section 1 is that the order of the Loop 2 and Loop 3 has been reversed. We also summarize the steps of the algorithm in Table 5. In each loop we apply Algorithm 2 with different value of parameter H which defines its complexity. In the subsection after description of the loops we carefully choose the value of this parameter in each level of the loops. Later, in the next Appendix we use this general framework in the proof of Theorems 7 and 8 with complexity estimates for problem (95) under Assumption 5, as well as Corollary 3 with complexity estimates for problem (71) with $m_h = 1$.

Main loops of the framework

In each of the three loops of the general framework we have a target accuracy ε and a confidence level σ which define the required quality of the solution to an optimization problem in this loop. These quantities define the inexactness of the oracle in this loop via inequalities (31) and (32) and the target accuracy and confidence level for the optimization problem in the next loop via (34), (35). Due to inexact strong convexity provided by (δ, σ, L, μ)-oracle, Algorithm 2 has logarithmic dependence of the complexity on the target accuracy and confidence level (see Theorem 4). Since the dependencies on the target accuracy and confidence level in (31), (32), (34) and (35) are polynomial, we obtain that the dependency of the complexity in each loop on the target accuracy and confidence level in the first loop, i.e. target accuracy and confidence level for the solution to problem (37), is logarithmic. We hide such logarithmic factors in $\widetilde{O}$ notation.

For convenience, we summarize the main details of the loops in Table 5.

Loop 1

The goal of Loop 1 is to find an (ε, σ)-solution of problem (39), which is considered as a minimization problem in y with the objective given in the form of auxiliary maximization problem in x. Finding an (ε, σ)-solution of this minimization problem gives an approximate solution to the saddle-point problem (37).

To solve problem (39), we would like to apply Algorithm 2 with

$$\varphi = 0, \quad \psi = h(y) + \max_{x \in \mathbb{R}^{d_x}} \{-G(x, y) - f(x)\}. \tag{164}$$

The function φ is, clearly, convex and is known exactly. What makes solving problem (39) not straightforward is that the exact value of ψ is not available. At the same time we can construct an inexact oracle for this function. First, the function h is μ_y-strongly convex, L_h-smooth and its exact gradient is available. Second, thanks to Assumption 3, it is possible to construct a $\left(\delta^{(1)}(\varepsilon), \sigma_0^{(1)}(\varepsilon, \sigma), 2L_G + 4\frac{L_G^2}{\mu_x}\right)$-oracle for the function $r(y) = \max_{x\in\mathbb{R}^{d_x}}\{-f(x) - G(x,y)\}$ for any $\delta^{(1)}(\varepsilon) = \mathbf{poly}(\varepsilon)$ and $\sigma_0^{(1)}(\varepsilon,\sigma) = \mathbf{poly}(\varepsilon,\sigma)$. Combining these two parts and using Lemma 1, we obtain that we can construct a $\left(\delta^{(1)}(\varepsilon), \sigma_0^{(1)}(\varepsilon,\sigma), L_h + 2L_G + 4\frac{L_G^2}{\mu_x}, \mu_y\right)$-oracle for ψ. Thus, we can apply Algorithm 2 with parameter $H = H_1$, which will be chosen later, to solve problem (39). Moreover, since Assumption 3 requires $\delta^{(1)}(\varepsilon) = \mathbf{poly}(\varepsilon)$ and $\sigma_0^{(1)}(\varepsilon,\sigma) = \mathbf{poly}(\varepsilon,\sigma)$, which holds for the dependencies in (31) and (32), we can choose $\delta^{(1)}(\varepsilon)$ and $\sigma_0^{(1)}(\varepsilon,\sigma)$ such that (31) and (32) hold. So, the first main assumption of Theorem 4 holds. At the same time, according to Assumptions 1 and 3, constructing inexact oracle for ψ requires τ_h calls of the basic oracle for h, τ_G calls of the basic oracle of $G(x,\cdot)$, $\mathcal{N}_G^x(\tau_G)\mathcal{K}_G^x(\varepsilon,\sigma)$ calls of the basic oracle for $G(\cdot,y)$, $\mathcal{N}_f(\tau_f)\mathcal{K}_f(\varepsilon,\sigma)$ calls of the basic oracle for f.

Let us discuss the second main assumption of Theorem 4. To ensure that this assumption holds, we need in each iteration of Algorithm 1, used as a building block in Algorithm 2, to find an $\left(\tilde{\varepsilon}_f^{(1)}(\varepsilon), \tilde{\sigma}^{(1)}(\varepsilon,\sigma)\right)$-solution to the auxiliary problem (15), where $\tilde{\sigma}^{(1)}(\varepsilon,\sigma), \tilde{\varepsilon}_f^{(1)}(\varepsilon)$ satisfy inequalities (34), (35). For the particular definitions of φ, ψ (164) in this Loop, this problem has the following form:

$$y_{k+1}^t = \arg\min_{y\in\mathbb{R}^{d_y}}\left\{h(y) + \max_{x\in\mathbb{R}^{d_x}}\{-G(x,y) - f(x)\} + \frac{H_1}{2}\|y - y_k^{md}\|^2\right\}. \tag{165}$$

Below, in the next paragraph "Loop 2 we explain how to solve this auxiliary problem to obtain its $\left(\tilde{\varepsilon}_f^{(1)}(\varepsilon), \tilde{\sigma}^{(1)}(\varepsilon,\sigma)\right)$-solution. To summarize Loop 1, both main assumptions of Theorem 4 hold and we can use it to guarantee that we obtain an (ε,σ)-solution of problem (39). This requires $\widetilde{O}\left(1 + \left(\frac{H_1}{\mu_\varphi+\mu_\psi}\right)^{\frac{1}{2}}\right) = \widetilde{O}\left(1 + \left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)$ calls to the inexact oracles for φ and for ψ, and the same number of times solving the auxiliary problem (165). Combining this oracle complexity with the cost of calculating inexact oracles for φ and for ψ, we obtain that solving problem (39) requires $\widetilde{O}\left(1 + \left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)\tau_h$ calls of the basic oracle for h, $\widetilde{O}\left(1 + \left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)\tau_G$ calls of the basic oracle of $G(x,\cdot)$, $\widetilde{O}\left(1 + \left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)\mathcal{N}_G^x(\tau_G)\mathcal{K}_G^x(\varepsilon,\sigma)$ calls of the basic oracle for $G(\cdot,y)$, $\widetilde{O}\left(1 + \left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)\mathcal{N}_f(\tau_f)\mathcal{K}_f(\varepsilon,\sigma)$ calls of the basic oracle for f. The only remaining thing is to provide an inexact solution to problem (165) and, next, we move to the Loop 2 to explain how to guarantee this. Note that we need to solve problem (165) $\widetilde{O}\left(1 + \left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)$ times.

Loop 2

As mentioned in the previous Loop 1, in each iteration of Algorithm 2 in Loop 1 we need many times to find an $(\varepsilon_2', \sigma_2')$-solution of the auxiliary problem (165), where we denoted for simplicity $\sigma_2' = \tilde{\sigma}^{(1)}(\varepsilon,\sigma)$ and $\varepsilon_2' = \tilde{\varepsilon}_f^{(1)}(\varepsilon)$. To do this, we reformulate problem (165) by changing the order of minimization and maximization as follows:

$$\min_{y\in\mathbb{R}^{d_y}}\left\{h(y)+\frac{H_1}{2}\|y-y_k^{md}\|^2+\max_{x\in\mathbb{R}^{d_x}}\{-G(x,y)-f(x)\}\right\} \tag{166}$$

$$=\min_{y\in\mathbb{R}^{d_y}}\max_{x\in\mathbb{R}^{d_x}}\left\{h(y)-G(x,y)-f(x)+\frac{H_1}{2}\|y-y_k^{md}\|^2\right\} \tag{167}$$

$$=\max_{x\in\mathbb{R}^{d_x}}\min_{y\in\mathbb{R}^{d_y}}\left\{h(y)-G(x,y)-f(x)+\frac{H_1}{2}\|y-y_k^{md}\|^2\right\} \tag{168}$$

$$=-\min_{x\in\mathbb{R}^{d_x}}\left\{f(x)+\max_{y\in\mathbb{R}^{d_y}}\left\{G(x,y)-h(y)-\frac{H_1}{2}\|y-y_k^{md}\|^2\right\}\right\} \tag{169}$$

and obtain an $(\varepsilon_2',\sigma_2')$-solution of the problem (165) by solving minimization problem (169). Assume that we can find an (ε_2,σ_2)-solution $\hat{x}$ of the minimization problem (169). Then, according to Assumption 2, we can also obtain a point $\hat{y}$ which is $(\bar{\delta}(\varepsilon_2)/2,\bar{\sigma}_0(\sigma_2))$-solution to the problem

$$\max_{y\in\mathbb{R}^{d_y}}\left\{G(x,y)-h(y)-\frac{H_1}{2}\|y-y_k^{md}\|^2\right\}, \tag{170}$$

where $\bar{\delta}(\varepsilon_2),\bar{\sigma}_0(\sigma_2)$ satisfy the following polynomial dependencies

$$\bar{\delta}(\varepsilon_2)\le\frac{H_1+\mu_y}{4\mu_x\left(\frac{H_1+\mu_y}{4L_G}\right)^2}\varepsilon_2,\quad \bar{\sigma}_0(\sigma_2)\le\sigma_2. \tag{171}$$

If we choose $\varepsilon_2,\sigma_2,\bar{\delta}(\varepsilon_2),\bar{\sigma}_0(\sigma_2)$ satisfying

$$\varepsilon_2\le\left(\frac{H_1+\mu_y}{4L_G}\right)^2\frac{\mu_x}{L_h+H_1+L_G+\frac{2L_G^2}{\mu_x}}\varepsilon_2', \tag{172}$$

$$\sigma_2\le\frac{\sigma_2'}{2}, \tag{173}$$

$$\bar{\sigma}_0(\sigma_2)\overset{(171)}{\le}\sigma_2\le\frac{\sigma_2'}{2},\quad \bar{\delta}(\varepsilon_2)\le\frac{H_1+\mu_y}{4\mu_x\left(\frac{H_1+\mu_y}{4L_G}\right)^2}\varepsilon_2\overset{(171)}{\le}\frac{H_1+\mu_y}{4L_h+4H_1+4L_G+\frac{8L_G^2}{\mu_x}}\varepsilon_2', \tag{174}$$

then

$$2\frac{L_h+H_1+L_G+\frac{2L_G^2}{\mu_x}}{H_1+\mu_y}\bar{\delta}(\varepsilon_2)+8\left(\frac{L_G}{H_1+\mu_y}\right)^2\frac{L_h+H_1+L_G+\frac{2L_G^2}{\mu_x}}{\mu_x}\varepsilon_2\le\varepsilon_2', \tag{175}$$

$$\sigma_2+\bar{\sigma}_0(\sigma_2)\le\sigma_2'. \tag{176}$$

Thus, applying Corollary 1 to minimization problem (169) with $F(x,y)=G(x,y)$, $w(y)=h(y)+$ $+\frac{H_1}{2}\|y-y_k^{md}\|^2$, $\varepsilon_x=\varepsilon_2$, $\sigma_x=\sigma_2$, $\varepsilon_y=\bar{\delta}(\varepsilon_2)$, $\sigma_y=\bar{\sigma}_0(\sigma_2)$ we obtain (see (46), (48)) that $\hat{y}$ satisfies inequality

$$h(\hat{y})+\frac{H_1}{2}\|\hat{y}-y_k^{md}\|^2+\max_{x\in\mathbb{R}^{d_x}}\{-G(x,\hat{y})-f(x)\}$$
$$-\min_{y\in\mathbb{R}^{d_y}}\max_{x\in\mathbb{R}^{d_x}}\{h(y)+\frac{H_1}{2}\|y-y_k^{md}\|^2-G(x,y)-f(x)\}\le\varepsilon_2'$$

with probability σ_2'. Thus, it is an $(\varepsilon_2', \sigma_2')$-solution of the problem (165). By Assumption 2, calculation of $\hat{y}$ requires $\mathcal{N}_G^y\left(\tau_G, H\right)\mathcal{K}_G^y\left(\varepsilon_2, \sigma_2\right)$ calls of the basic oracle O_G^y of $G(x, \cdot)$, τ_G calls of the basic oracle O_G^x of $G(\cdot, y)$ and $\mathcal{N}_h\left(\tau_h, H\right)\mathcal{K}_h\left(\varepsilon_2, \sigma_2\right)$ calls of the basic oracle O_h of h.

Our next step is to provide an $(\varepsilon_2, \sigma_2)$-solution to minimization problem (169), for which we again apply Algorithm 2, but this time with

$$\varphi = f(x), \quad \psi = \max_{y \in \mathbb{R}^{d_y}} \left\{ G(x,y) - h(y) - \frac{H_1}{2}\|y - y_k^{md}\|^2 \right\}. \tag{177}$$

The function φ is μ_x-strongly convex, L_f-smooth and its exact gradient is available. What makes solving problem (169) not straightforward is that the exact value of ψ is not available. At the same time we can construct an inexact oracle for this function. Thanks to Assumption 2, it is possible to construct a $\left(\delta^{(2)}\left(\varepsilon_2\right), \sigma_0^{(2)}\left(\varepsilon_2, \sigma_2\right), 2L_G + 4\frac{L_G^2}{H_1+\mu_y}\right)$-oracle for the function ψ for any $\delta^{(2)}\left(\varepsilon_2\right) = \operatorname{poly}\left(\varepsilon_2\right)$ and $\sigma_0^{(2)}\left(\varepsilon_2, \sigma_2\right) = \operatorname{poly}\left(\varepsilon_2, \sigma_2\right)$. Using Lemma 1, we obtain that we can construct
a $\left(\delta^{(2)}\left(\varepsilon_2\right), \sigma_0^{(2)}\left(\varepsilon_2, \sigma_2\right), L_f + 2L_G + 4\frac{L_G^2}{H_1+\mu_y}, \mu_x\right)$-oracle for the function $\varphi + \psi$. Thus, we can apply Algorithm 2 with parameter $H = H_2 \geq 2L_f$, which will be chosen later, to solve the problem (169). Moreover, since Assumption 2 requires $\delta^{(2)}\left(\varepsilon_2\right) = \operatorname{poly}\left(\varepsilon_2\right)$ and $\sigma_0^{(2)}\left(\varepsilon_2, \sigma_2\right) = = \operatorname{poly}\left(\varepsilon_2, \sigma_2\right)$, which holds for the dependencies in (31) and (32), we can choose $\delta^{(2)}\left(\varepsilon_2\right)$ and $\sigma_0^{(2)}\left(\varepsilon_2, \sigma_2\right)$ such that (31) and (32) hold. So, the first main assumption of Theorem 4 holds. At the same time, according to Assumptions 1 and 2, constructing inexact oracle for ψ requires $\mathcal{N}_G^y\left(\tau_G, H_1\right)\mathcal{K}_G^y\left(\varepsilon_2, \sigma_2\right)$ calls of the basic oracle for $G(x, \cdot)$, τ_G calls of the basic oracle for $G(\cdot, y)$, $\mathcal{N}_h\left(\tau_h, H_1\right)\mathcal{K}_h\left(\varepsilon_2, \sigma_2\right)$ calls of the basic oracle for h, and constructing exact oracle for $\varphi = f$ requires τ_f calls of the basic oracle for f.

Let us discuss the second main assumption of Theorem 4. To ensure that this assumption holds, we need in each iteration of Algorithm 1, used as a building block in Algorithm 2, to find $\left(\tilde{\varepsilon}_f^{(2)}\left(\varepsilon_2\right), \tilde{\sigma}^{(2)}\left(\varepsilon_2, \sigma_2\right)\right)$-solution to the auxiliary problem (15), where $\tilde{\sigma}^{(2)}\left(\varepsilon_2, \sigma_2\right), \tilde{\varepsilon}_f^{(2)}\left(\varepsilon_2\right)$ satisfy inequalities (34), (35). For the particular definitions of φ, ψ (177) in this Loop, this problem has the following form:

$$\begin{aligned} x_{l+1}^t = \arg\min_{x \in \mathbb{R}^{d_x}} \Big\{ & \langle \nabla f(x_l^{md}), x - x_l^{md} \rangle \\ & + \max_{y \in \mathbb{R}^{d_y}} \left\{ G(x,y) + h(y) - \frac{H_1}{2}\|y - y_k^{md}\|^2 \right\} + \frac{H_2}{2}\|x - x_l^{md}\|^2 \Big\}, \end{aligned} \tag{178}$$

Below, in the next paragraph "Loop 3 we explain how to solve this auxiliary problem to obtain its
$\left(\tilde{\varepsilon}_f^{(2)}\left(\varepsilon_2\right), \tilde{\sigma}^{(2)}\left(\varepsilon_2, \sigma_2\right)\right)$-solution.

To summarize Loop 2, both main assumptions of Theorem 4 hold and we can use it to guarantee that we obtain an $(\varepsilon_2', \sigma_2')$-solution of the auxiliary problem (165). This requires one time to solve the problem (170), which, by Assumption 2 has the same cost as evaluating inexact oracle for the function ψ. Further, we need $O\left(\left(1 + \left(\frac{H_2}{\mu_\varphi + \mu_\psi}\right)^{\frac{1}{2}}\right)\log \varepsilon_2^{-1}\right) = = O\left(\left(1 + \left(\frac{H_2}{\mu_x}\right)^{\frac{1}{2}}\right)\log \varepsilon_2^{-1}\right)$ calls to the inexact oracles for φ and for ψ, and the same number of times solving the auxiliary problem (178). Combining this oracle complexity with the cost of calculating inexact oracles for φ and for ψ, we obtain that solving problem (169) requires $O\left(\left(1 + \left(\frac{H_2}{\mu_x}\right)^{\frac{1}{2}}\right)\log \varepsilon_2^{-1}\right)\tau_f$ calls of the basic oracle for

f, $O\left(\left(1+\left(\frac{H_2}{\mu_x}\right)^{\frac{1}{2}}\right)\log\varepsilon_2^{-1}\right)\mathcal{N}_G^y(\tau_G,H_1)\,\mathcal{K}_G^y(\varepsilon_2,\sigma_2)$ calls of the basic oracle for $G(x,\cdot)$, $O\left(\left(1+\left(\frac{H_2}{\mu_x}\right)^{\frac{1}{2}}\right)\log\varepsilon_2^{-1}\right)\tau_G$ calls of the basic oracle for $G(\cdot,y)$, $O\left(\left(1+\left(\frac{H_2}{\mu_x}\right)^{\frac{1}{2}}\right)\log\varepsilon_2^{-1}\right)\mathcal{N}_h(\tau_h,H_1)\,\mathcal{K}_h(\varepsilon_2,\sigma_2)$ calls of the basic oracle for h. The only remaining thing is to provide an inexact solution to problem (178) and, next, we move to Loop 3 to explain how to guarantee this. Note that we need to solve problem (178) $O\left(\left(1+\left(\frac{H_2}{\mu_x}\right)^{\frac{1}{2}}\right)\log\varepsilon_2^{-1}\right)$ times.

Loop 3

As mentioned in the previous Loop 2, in each iteration of Algorithm 2 in Loop 2 we need to find many times an (ε_3,σ_3)-solution of the auxiliary problem (178), where we denoted for simplicity $\sigma_3=\tilde{\sigma}^{(2)}(\varepsilon_2,\sigma_2)$ and $\varepsilon_3=\tilde{\varepsilon}_f^{(2)}(\varepsilon_2)$. To solve problem (178), we would like to apply Algorithm 2 with

$$\varphi=\max_{y\in\mathbb{R}^{d_y}}\left\{G(x,y)-h(y)-\frac{H_1}{2}\|y-y_k^{md}\|^2\right\},\quad \psi=\langle\nabla f(x_l^{md}),x-x_l^{md}\rangle+\frac{H_2}{2}\|x-x_l^{md}\|^2. \tag{179}$$

The function ψ is, clearly, H_2-strongly convex, H_2-smooth and its exact gradient is available. What makes solving problem (178) not straightforward is that the exact value of φ is not available. At the same time, we can construct an inexact oracle for this function. Thanks to Assumption 2, it is possible to construct a $\left(\delta^{(3)}(\varepsilon_3),\sigma_0^{(3)}(\varepsilon_3,\sigma_3),2L_G+4\frac{L_G^2}{H_1+\mu_y}\right)$-oracle for the function φ for any $\delta^{(3)}(\varepsilon_3)=\text{poly}(\varepsilon_3)$ and $\sigma_0^{(3)}(\varepsilon_3,\sigma_3)=\text{poly}(\varepsilon_3,\sigma_3)$. Using Lemma 1, we obtain that we can construct a $\left(\delta^{(3)}(\varepsilon_3),\sigma_0^{(3)}(\varepsilon_3,\sigma_3),H_2+2L_G+4\frac{L_G^2}{H_1+\mu_x},H_2\right)$-oracle for the function $\varphi+\psi$. Thus, we can apply Algorithm 2 with parameter $H=H_3\geq 2L_G+4\frac{L_G^2}{H_1+\mu_y}$, which will be chosen later, to solve problem (178). Moreover, since Assumption 2 requires $\delta^{(3)}(\varepsilon_3)=\mathbf{poly}(\boldsymbol{\varepsilon_3})$ and $\sigma_0^{(3)}(\varepsilon_3,\sigma_3)=$ $=\mathbf{poly}(\boldsymbol{\varepsilon_3},\boldsymbol{\sigma_3})$, which holds for the dependencies in (31) and (32), we can choose $\delta^{(3)}(\varepsilon_3)$ and $\sigma_0^{(3)}(\varepsilon_3,\sigma_3)$ such that (31) and (32) hold. So, the first main assumption of Theorem 4 holds. At the same time, according to Assumptions 1 and 2, constructing inexact oracle for φ requires $\mathcal{N}_G^y(\tau_G,H_1)\,\mathcal{K}_G^y(\varepsilon_3,\sigma_3)$ calls of the basic oracle for $G(x,\cdot)$, τ_G calls of the basic oracle for $G(\cdot,y)$, $\mathcal{N}_h(\tau_h,H_1)\,\mathcal{K}_h(\varepsilon_3,\sigma_3)$ calls of the basic oracle for h. At the same time, no calls to the oracle for f are needed.

Let us discuss the second main assumption of Theorem 4. To ensure that this assumption holds, we need in each iteration of Algorithm 1, used as a building block in Algorithm 2, to find $\left(\tilde{\varepsilon}_f^{(3)}(\varepsilon_3),\tilde{\sigma}^{(3)}(\varepsilon_3,\sigma_3)\right)$-solution to the auxiliary problem (15), where $\tilde{\sigma}^{(3)}(\varepsilon_3,\sigma_3),\tilde{\varepsilon}_f^{(3)}(\varepsilon_3)$ satisfy inequalities (34), (35). For the particular definitions of φ, ψ in (179) in this Loop, this

	Goal	φ, ψ	μ in Th.4	Iteration number of Algorithm 1 (Th. 4)	Each iteration requires
Loop 1	(ε, σ)-solution of problem (39)	(164)	μ_y	$\widetilde{O}\left(1+\sqrt{H_1/\mu_y}\right)$	Find $(\varepsilon_1, \sigma_1)$-solution of (165) and calculate $\left(\delta^{(1)}, L_\psi\right)$-oracle of $\psi(y)$
Loop 2	$(\varepsilon_1, \sigma_1)$-solution of problem (169)	(177)	μ_x	$\widetilde{O}(1+\sqrt{H_2/\mu_x})$	Find $(\varepsilon_2, \sigma_2)$-solution of (178) and calculate $\left(\delta^{(2)}, L_\psi\right)$-oracle of $\psi(x)$
Loop 3	$(\varepsilon_2, \sigma_2)$-solution of problem (178)	(179)	H_2	$\widetilde{O}(1+\sqrt{H_3/H_2})$	Find $(\varepsilon_3, \sigma_3)$-solution of (180) and calculate $\left(\delta^{(3)}, L_\varphi\right)$-oracle of $\varphi(x)$

Table 5. Summary of the three loops of the framework described in this Appendix.

problem has the following form:

$$u^t_{m+1} = \arg\min_{u\in\mathbb{R}^{d_x}} \{\langle \nabla\varphi_{\delta^{(3)},2L_\varphi}(u^{md}_m), u-u^{md}_m\rangle + \psi(u) + \frac{H_3}{2}\|u-u^{md}_m\|_2^2\}$$

$$= \arg\min_{u\in\mathbb{R}^{d_x}} \{\langle \nabla\varphi_{\delta^{(3)},2L_\varphi}(u^{md}_m), u-u^{md}_m\rangle + \langle\nabla f(x^{md}_l), u-x^{md}_l\rangle + \frac{H_2}{2}\|u-x^{md}_l\|_2^2 + \frac{H_3}{2}\|u-u^{md}_m\|_2^2\}, \tag{180}$$

where $L_\varphi = L_G + \frac{L_G^2}{H_1+\mu_x}$. This quadratic auxiliary problem (180) can be solved explicitly and exactly since at the point it needs to be solved, $\nabla\varphi_{\delta^{(3)},2L_\varphi}(u^{md}_m)$ is already calculated. Thus, the second main assumption of Theorem 4 is satisfied with $\tilde{\sigma}^{(3)}(\varepsilon_3, \sigma_3) = 0$ and $\tilde{\varepsilon}^{(3)}_f(\varepsilon_3) = 0$, which clearly satisfy (31) and (32).

To summarize Loop 3, both main assumptions of Theorem 4 hold and we can use it to guarantee that we obtain an $(\varepsilon_3, \sigma_3)$-solution of the auxiliary problem (178). This requires $O\left(\left(1+\left(\frac{H_3}{\mu_\varphi+\mu_\psi}\right)^{\frac{1}{2}}\right)\log\varepsilon_3^{-1}\right) = O\left(\left(1+\left(\frac{H_3}{H_2}\right)^{\frac{1}{2}}\right)\log\varepsilon_3^{-1}\right)$ calls to the inexact oracles for φ and for ψ, and the same number of times solving the auxiliary problem (180). Combining this oracle complexity with the cost of calculating inexact oracles for φ and for ψ, we obtain that solving problem (178) requires $O\left(\left(1+\left(\frac{H_3}{H_2}\right)^{\frac{1}{2}}\right)\log\varepsilon_3^{-1}\right)\mathcal{N}^y_G(\tau_G, H_1)\,\mathcal{K}^y_G(\varepsilon_3, \sigma_3)$ calls of the basic oracle for $G(x,\cdot)$, $O\left(\left(1+\left(\frac{H_3}{H_2}\right)^{\frac{1}{2}}\right)\log\varepsilon_3^{-1}\right)\tau_G$ calls of the basic oracle for $G(\cdot,y)$ and $O\left(\left(1+\left(\frac{H_3}{H_2}\right)^{\frac{1}{2}}\right)\log\varepsilon_3^{-1}\right)\mathcal{N}_h(\tau_h, H_1)\,\mathcal{K}_h(\varepsilon_3, \sigma_3)$ calls of the basic oracle for h.

Complexity of the framework

Below we formally finalize in Theorem 10 the analysis of the framework by carefully combining the bounds obtained in Loop 1 - Loop 3 to obtain the final bounds for the total number of oracle calls for each part f, G, h of the objective in problem (37). In the next Appendix, we apply Theorem 10 to obtain complexity bounds for our framework applied to problem (71) in the case $m_h = 1$.

Theorem 10. Let Assumptions 1, 2, 3 hold. Then, execution of the optimization framework described in Loop 1 - Loop 3 with

$$H_1 = 2L_G, H_2 = 2L_f, H_3 = 2\left(L_G + \frac{2L_G^2}{\mu_y + H_1}\right)$$

generates an (ε, σ)-solution to the problem (37) i.e., satisfy 10. Moreover, for the number of basic oracle calls it holds that

Number of calls of basic oracle O_f for f is :

$$\widetilde{O}\left(\left(1+\sqrt{\frac{L_G}{\mu_y}}\right)\left(\mathcal{N}_f\left(\tau_f\right)+\left(1+\sqrt{\frac{L_f}{\mu_x}}\right)\cdot\tau_f\right)\right), \tag{181}$$

Number of calls of basic oracle O_h for h is :

$$\widetilde{O}\left(\left(1+\sqrt{\frac{L_G}{\mu_y}}\right)\left(\tau_h+\left(1+\sqrt{\frac{L_f}{\mu_x}}\right)\left(1+\sqrt{\frac{L_G}{L_f}}\right)\mathcal{N}_h\left(\tau_h, 2L_G\right)\right)\right), \tag{182}$$

Number of calls of basic oracle O_G^x for $G(\cdot, y)$ is :

$$\widetilde{O}\left(\left(1+\sqrt{\frac{L_G}{\mu_y}}\right)\left(\mathcal{N}_G^x\left(\tau_G\right)+\left(1+\sqrt{\frac{L_f}{\mu_x}}\right)\left(1+\sqrt{\frac{L_G}{L_f}}\right)\tau_G\right)\right), \tag{183}$$

Number of calls of basic oracle O_G^y for $G(x, \cdot)$ is :

$$\widetilde{O}\left(\left(1+\sqrt{\frac{L_G}{\mu_y}}\right)\left(\tau_G+\left(1+\sqrt{\frac{L_f}{\mu_x}}\right)\left(1+\sqrt{\frac{L_G}{L_f}}\right)\mathcal{N}_G^y\left(\tau_G, 2L_G\right)\right)\right). \tag{184}$$

Доказательство. By construction, as an output of Loop 1 we obtain an (ε, σ)-solution to the problem (37).

We prove the estimates of for the numbers of oracle calls in two steps. The first step is to formally prove that in each loop the dependence of the number of oracle calls on the target accuracy ε and a confidence level σ is logarithmic. The second step is to multiply the estimates for the number of oracle calls between loops and choose the parameters H_1, H_2, H_3.

Step 1. Polynomial dependence. Proof of this part is equivalent to the proof of the Theorem 5.

Step 2. Final estimates.

We have already counted the number of oracles calls for each oracle in each loop Loop 1 - Loop 3, see the last paragraph of the description of each loop. We start with the number of basic oracle calls of f, which is called in each step of Loop 1 and Loop 2. Thus, the total number is

$$\begin{aligned} &\# \text{ of calls in Loop1} + (\# \text{ of steps in Loop 1})\cdot(\# \text{ of calls in Loop 2}) \\ &= \widetilde{O}\left(1+\left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)\mathcal{N}_f\left(\tau_f\right)\mathcal{K}_f\left(\varepsilon,\sigma\right)+\widetilde{O}\left(1+\left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)\cdot\left(\widetilde{O}\left(1+\left(\frac{H_2}{\mu_x}\right)^{\frac{1}{2}}\right)\tau_f\right) \\ &= \widetilde{O}\left(\left(1+\sqrt{\frac{H_1}{\mu_y}}\right)\left(\mathcal{N}_f\left(\tau_f\right)+\left(1+\sqrt{\frac{H_2}{\mu_x}}\right)\cdot\tau_f\right)\right), \end{aligned}$$

where we used that $\mathcal{K}_f\left(\varepsilon,\sigma\right)=\widetilde{O}(1)$.

The basic oracle of h is called in each step of all the three loops. Thus, the total number is

$$\begin{aligned}
&\# \text{ of calls in Loop1 } + (\# \text{ of steps in Loop 1})\cdot(\# \text{ of calls in Loop 2})\\
&+ (\# \text{ of steps in Loop 1})\cdot(\# \text{ of steps in Loop 2})\cdot(\# \text{ of calls in Loop 3})\\
&=\widetilde{O}\left(1+\left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)\tau_h+\widetilde{O}\left(1+\left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)\cdot\left(\widetilde{O}\left(1+\left(\frac{H_2}{\mu_x}\right)^{\frac{1}{2}}\right)\mathcal{N}_h\left(\tau_h,H_1\right)\mathcal{K}_h\left(\varepsilon_2,\sigma_2\right)\right)\\
&+\widetilde{O}\left(1+\left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)\cdot\left(\widetilde{O}\left(1+\left(\frac{H_2}{\mu_x}\right)^{\frac{1}{2}}\right)\right)\cdot\left(\widetilde{O}\left(1+\left(\frac{H_3}{H_2}\right)^{\frac{1}{2}}\right)\mathcal{N}_h\left(\tau_h,H_1\right)\mathcal{K}_h\left(\varepsilon_3,\sigma_3\right)\right)\\
&=\widetilde{O}\left(\left(1+\sqrt{\frac{H_1}{\mu_y}}\right)\left(\tau_h+\left(1+\sqrt{\frac{H_2}{\mu_x}}\right)\left(\mathcal{N}_h\left(\tau_h,H_1\right)+\left(1+\sqrt{\frac{H_3}{H_2}}\right)\cdot\mathcal{N}_h\left(\tau_h,H_1\right)\right)\right)\right)\\
&=\widetilde{O}\left(\left(1+\sqrt{\frac{H_1}{\mu_y}}\right)\left(\tau_h+\left(1+\sqrt{\frac{H_2}{\mu_x}}\right)\left(1+\sqrt{\frac{H_3}{H_2}}\right)\mathcal{N}_h\left(\tau_h,H_1\right)\right)\right),
\end{aligned}$$

where we used that $\mathcal{K}_h\left(\varepsilon,\sigma\right)=\widetilde{O}(1)$.

The basic oracle of $G(\cdot,y)$ is called in each step of all the three loops. Thus, the total number is

$$\begin{aligned}
&\# \text{ of calls in Loop1 } + (\# \text{ of steps in Loop 1})\cdot(\# \text{ of calls in Loop 2})\\
&+ (\# \text{ of steps in Loop 1})\cdot(\# \text{ of steps in Loop 2})\cdot(\# \text{ of calls in Loop 3})\\
&=\widetilde{O}\left(1+\left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)\mathcal{N}_G^x\left(\tau_G\right)\mathcal{K}_G^x\left(\varepsilon,\sigma\right)+\widetilde{O}\left(1+\left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)\cdot\left(\widetilde{O}\left(1+\left(\frac{H_2}{\mu_x}\right)^{\frac{1}{2}}\right)\tau_G\right)\\
&+\widetilde{O}\left(1+\left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)\cdot\left(\widetilde{O}\left(1+\left(\frac{H_2}{\mu_x}\right)^{\frac{1}{2}}\right)\right)\cdot\left(\widetilde{O}\left(1+\left(\frac{H_3}{H_2}\right)^{\frac{1}{2}}\right)\tau_G\right)\\
&=\widetilde{O}\left(\left(1+\sqrt{\frac{H_1}{\mu_y}}\right)\left(\mathcal{N}_G^x\left(\tau_G\right)+\left(1+\sqrt{\frac{H_2}{\mu_x}}\right)\left(\tau_G+\left(1+\sqrt{\frac{H_3}{H_2}}\right)\tau_G\right)\right)\right)\\
&=\widetilde{O}\left(\left(1+\sqrt{\frac{H_1}{\mu_y}}\right)\left(\mathcal{N}_G^x\left(\tau_G\right)+\left(1+\sqrt{\frac{H_2}{\mu_x}}\right)\left(1+\sqrt{\frac{H_3}{H_2}}\right)\tau_G\right)\right),
\end{aligned}$$

where we used that $\mathcal{K}_G^x\left(\varepsilon,\sigma\right)=\widetilde{O}(1)$.

Finally, the basic oracle of $G(x,\cdot)$ is called in each step of all the three loops. Thus, the total number is

$$
\begin{aligned}
&\# \text{ of calls in Loop1} + (\# \text{ of steps in Loop 1})\cdot(\# \text{ of calls in Loop 2})\\
&+ (\# \text{ of steps in Loop 1})\cdot(\# \text{ of steps in Loop 2})\cdot(\# \text{ of calls in Loop 3})\\
&= \widetilde{O}\left(1+\left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)\tau_G + \widetilde{O}\left(1+\left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)\cdot\left(\widetilde{O}\left(1+\left(\frac{H_2}{\mu_x}\right)^{\frac{1}{2}}\right)\mathcal{N}_G^y\left(\tau_G, H_1\right)\mathcal{K}_G^y\left(\varepsilon_2,\sigma_2\right)\right)\\
&+\widetilde{O}\left(1+\left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)\cdot\left(\widetilde{O}\left(1+\left(\frac{H_2}{\mu_x}\right)^{\frac{1}{2}}\right)\right)\cdot\left(\widetilde{O}\left(1+\left(\frac{H_3}{H_2}\right)^{\frac{1}{2}}\right)\mathcal{N}_G^y\left(\tau_G, H_1\right)\mathcal{K}_G^y\left(\varepsilon_2,\sigma_2\right)\right)\\
&= \widetilde{O}\left(\left(1+\sqrt{\frac{H_1}{\mu_y}}\right)\left(\tau_G+\left(1+\sqrt{\frac{H_2}{\mu_x}}\right)\left(\mathcal{N}_G^y\left(\tau_G,H_1\right)+\left(1+\sqrt{\frac{H_3}{H_2}}\right)\mathcal{N}_G^y\left(\tau_G,H_1\right)\right)\right)\right)\\
&= \widetilde{O}\left(\left(1+\sqrt{\frac{H_1}{\mu_y}}\right)\left(\tau_G+\left(1+\sqrt{\frac{H_2}{\mu_x}}\right)\left(1+\sqrt{\frac{H_3}{H_2}}\right)\mathcal{N}_G^y\left(\tau_G,H_1\right)\right)\right),
\end{aligned}
$$

where we used that $\mathcal{K}_G^y\left(\varepsilon_2,\sigma_2\right)=\widetilde{O}(1)$.

The final estimates are obtained by substituting the constants H_1, H_2, H_3 given by

$$
H_1 = 2L_G, H_2 = 2L_f, H_3 = 2\left(L_G + \frac{2L_G^2}{\mu_y + H_1}\right) \le 2\left(L_G + \frac{2L_G^2}{H_1}\right) = 4L_G.
$$

Proof of Theorem 5

Let us prove the Theorem 5.

Доказательство. The proof of the Theorem 5 consists of three parts: firstly, we explicitly formulate that after "Loop 1 "Loop 4"the algorithm get a solution to the saddle problem. Then we prove a technical statement about the polynomial dependence

$$
\begin{aligned}
\sigma^{(k)}\left(\varepsilon,\sigma\right) = \mathbf{poly}\left(\varepsilon,\sigma\right), \tilde{\sigma}^{(\mathbf{k})}\left(\varepsilon,\sigma\right) = \mathbf{poly}\left(\varepsilon,\sigma\right), \sigma_{\mathbf{0}}^{(\mathbf{k})}\left(\varepsilon,\sigma\right)\\
= \mathbf{poly}\left(\varepsilon,\sigma\right), \tilde{\varepsilon}_{\mathbf{f}}^{(\mathbf{k})}\left(\varepsilon\right) = \mathbf{poly}\left(\varepsilon\right), \delta^{(\mathbf{k})}\left(\varepsilon\right) = \mathbf{poly}\left(\varepsilon\right).
\end{aligned}
$$

Finally, using the last statement, we show how to get the final estimates on the number of oracle calls.

Solution obtained Let us show that the random point $\hat{y}$ obtained after $\widetilde{O}\left(\left(\frac{H_1}{\mu_y}\right)^{1/2}\right)$ iterations of the "Loop 1"satisfies 10 of (ε,σ)-solution of the saddle problem. As mentioned in the "Loop 1 after N_1 iteration we receive an (ε,σ)-solution for function $h(y)+\max_{x\in\mathbb{R}^{d_x}} -G(x,y)-f(x)$, i.e. inequality

$$
h(\hat{y}) + \max_{x\in\mathbb{R}^{d_x}}\left\{-G(x,\hat{y})-f(x)\right\} - \min_{y\in\mathbb{R}^{d_y}}\max_{x\in\mathbb{R}^{d_x}}\left\{h(y)-G(x,y)-f(x)\right\} \le \varepsilon
$$

holds True with probability $1-\sigma$. Then, it is an (ε,σ)-solution for saddle problems.

Polynomial dependence Before obtaining the final estimates, it is important to prove that for all $i=1,2,3$ dependences

$$
\begin{aligned}
\sigma^{(i)}\left(\varepsilon,\sigma\right) = \mathbf{poly}\left(\varepsilon,\sigma\right), \tilde{\sigma}^{(\mathbf{i})}\left(\varepsilon,\sigma\right) = \mathbf{poly}\left(\varepsilon,\sigma\right), \sigma_{\mathbf{0}}^{(\mathbf{i})}\left(\varepsilon,\sigma\right)\\
= \mathbf{poly}\left(\varepsilon,\sigma\right), \tilde{\varepsilon}_{\mathbf{f}}^{(\mathbf{i})}\left(\varepsilon\right) = \mathbf{poly}\left(\varepsilon\right), \delta^{(\mathbf{i})}\left(\varepsilon\right) = \mathbf{poly}\left(\varepsilon\right) \qquad (185)
\end{aligned}
$$

are polynomial. This can be proved by induction: base holds true for $i = 1$. Then let us suppose that for $k \in \{1, 2\}$ we have

$$\sigma^{(k)}(\varepsilon, \sigma) = \mathbf{poly}(\varepsilon, \sigma), \tilde{\sigma}^{(\mathbf{k})}(\varepsilon, \sigma) = \mathbf{poly}(\varepsilon, \sigma), \sigma_{\mathbf{0}}^{(\mathbf{k})}(\varepsilon, \sigma) \\ = \mathbf{poly}(\varepsilon, \sigma), \tilde{\varepsilon}_{\mathbf{f}}^{(\mathbf{k})}(\varepsilon) = \mathbf{poly}(\varepsilon), \delta^{(\mathbf{k})}(\varepsilon) = \mathbf{poly}(\varepsilon).$$

According to paragraphs "Loop 1 "Loop 4"$\sigma^{(k+1)}, \tilde{\sigma}^{(k+1)}, \sigma_0^{(k+1)}, \tilde{\varepsilon}_f^{(k+1)}, \delta^{(k+1)}$ are chosen such that (34), (35) and $\sigma^{(k+1)} = \tilde{\sigma}^{(k)}$ hold true. These equations guarantee a polynomial dependence

$$\sigma^{(k+1)}(\varepsilon, \sigma) = \mathbf{poly}(\varepsilon, \sigma), \tilde{\sigma}^{(\mathbf{k+1})}(\varepsilon, \sigma) = \mathbf{poly}(\varepsilon, \sigma), \sigma_{\mathbf{0}}^{(\mathbf{k+1})}(\varepsilon, \sigma) = \mathbf{poly}(\varepsilon, \sigma), \tilde{\varepsilon}_{\mathbf{f}}^{(\mathbf{k+1})}(\varepsilon) = \mathbf{poly}(\varepsilon), \delta^{(\mathbf{k+1})}($$

Which finishes the proof of polynomial dependence. According last statement (185), we can use notation $\widetilde{O}(\cdot)$ at all levels "Loop 1 "Loop 4 implying that the logarithmic part depends on the initial ε, σ.
Final estimates The only thing left to finish proof of the Theorem 5 is an accurately count the number of oracle calls at each loop "Loop 1 "Loop 4"of the general scheme. At each step we must take into account only three places in which the oracle can be called:

- calculation of inexact oracles of functions φ, ψ,
- searching the solution to the auxiliary problem (15),
- step along the gradient (5) of the functions φ, ψ in the Algorithms 1,2;

The technique for counting the number of oracle calls is the same for $\nabla f, \nabla h, \nabla_x G, \nabla_y G$. Below we give only an example of calculation the number of oracle calls of ∇h.

According to the Table 4 and paragraphs "Loop 1 "Loop 4"we have

- ∇h is called at each of $\widetilde{O}(\sqrt{H_1/\mu_y})$ iteration of "Loop 1"when:
  - Algorithm 2 do the step along the gradient (5), it costs $\widetilde{O}(\tau_h)$ calls,
  - auxiliary problem (52) is solved; ∇h is called at each of $\widetilde{O}(\sqrt{H_2/\mu_x})$ iteration of "Loop 2"when:
    - an inexact model of ψ (62) is determined, it costs $\widetilde{O}(\mathcal{N}_h(\tau_h, H_1))$ calls,
    - auxiliary problem (63) is solved: ∇h is called at each of $\widetilde{O}(\sqrt{H_3/H_2})$ iteration of "Loop 3"when an inexact model of φ (65) is determined, it costs $\widetilde{O}(\mathcal{N}_h(\tau_h, H_1))$ calls;

Thus, we obtain that the estimate for the number of oracle calls ∇h has an nested structure of the form

$$\nabla h \text{ - oracle calls} : \widetilde{O}\left(\sqrt{\frac{H_1}{\mu_y}}\left(\tau_h + \sqrt{\frac{H_2}{\mu_x}}\left(\mathcal{N}_h(\tau_h, H_1) + \sqrt{\frac{H_3}{H_2}} \cdot \mathcal{N}_h(\tau_h, H_1)\right)\right)\right).$$

The remaining estimates are obtained similarly. Finally, for obtaining an (ε, σ)-solution of problem (39) it is sufficient to do the next number of oracle calls

$$\nabla f \text{ - oracle calls} : \widetilde{O}\left(\sqrt{\frac{H_1}{\mu_y}}\left(\mathcal{N}_f\left(\tau_f\right)+\sqrt{\frac{H_2}{\mu_x}}\cdot\tau_f\right)\right),$$

$$\nabla h \text{ - oracle calls} : \widetilde{O}\left(\sqrt{\frac{H_1}{\mu_y}}\left(\tau_h+\sqrt{\frac{H_2}{\mu_x}}\left(\mathcal{N}_h\left(\tau_h, H_1\right)+\sqrt{\frac{H_3}{H_2}}\cdot\mathcal{N}_h\left(\tau_h, H_1\right)\right)\right)\right),$$

$$\nabla_x G \text{ - oracle calls} : \widetilde{O}\left(\sqrt{\frac{H_1}{\mu_y}}\left(\mathcal{N}_G^x\left(\tau_G\right)+\sqrt{\frac{H_2}{\mu_x}}\left(\tau_G+\sqrt{\frac{H_3}{H_2}}\cdot\tau_G\right)\right)\right),$$

$$\nabla_y G \text{ - oracle calls} : \widetilde{O}\left(\sqrt{\frac{H_1}{\mu_y}}\left(\tau_G+\sqrt{\frac{H_2}{\mu_x}}\left(\mathcal{N}_G^y\left(\tau_G, H_1\right)+\sqrt{\frac{H_3}{H_2}}\cdot\mathcal{N}_G^y\left(\tau_G, H_1\right)\right)\right)\right).$$

Final estimates can be obtained by choosing the constants H_1, H_2, H_3 in the following way

$$H_1 = 2L_G, H_2 = 2L_f, H_3 = 2\left(L_G + \frac{L_G^2}{\mu_y + H_1}\right).$$

□

Proof of Theorem 7 and Theorem 8

In this appendix we prove Theorems 7, 8 and Corollary 3 and construct algorithms for problem (95) using the results of Section 1, in particular, Theorem 5, for the case $L_f \geq L_G$, and the results of the previous Appendix, in particular, Theorem 10. To use these theorems we need to satisfy Assumptions 2, 3, which is done in the first subsection. Then, in the next subsections, we combine the building blocks to obitan the final results.

Algorithms to guarantee Assumptions 2, 3

We start with two auxiliary results, that show how to satisfy Assumptions 2, 3 algorithmically. The first lemma provides complexity for inexact solution of the maximization problem (49) and the complexity of finding an inexact oracle for function g defined in the same equation, thereby proving that Assumption 2 holds. We underline that the algorithm which guarantees Assumption 2 depends on whether $L_h \geq L_G$ or $L_h \leq L_G$. After that we provide a simple corollary to show that Assumption 3 also holds.

Lemma 14. Let the function g be defined via maximization problem in (49), i.e.

$$g(x) = \max_{y\in\mathbb{R}^{d_y}}\left\{G(x,y) - h(y) - \frac{H}{2}\|y-y_0\|^2\right\}, \qquad (186)$$

where $G(x, y)$, $h(y)$ are according to (95) and satisfy Assumption 5.1,2,3(a), $y_0 \in \mathbb{R}^{d_y}$. Then, for each of two cases $L_h \geq L_G$ and $L_h \leq L_G$ we organize computations in two loops and apply Algorithm 2, so that Assumption 2 holds with τ_G basic oracle calls for $G(\cdot, y)$ and the following estimates for the number of basic oracle calls for $G(x, \cdot)$ and h respectively

$$\mathcal{N}_G^y\left(\tau_G, H\right) = O\left(\tau_G + \tau_G\sqrt{L_G/(H+\mu_y)}\right), \qquad (187)$$

$$\mathcal{N}_h\left(\tau_h, H\right) = O\left(\tau_h + \tau_h\sqrt{L_h/(H+\mu_y)}\right). \qquad (188)$$

We name these algorithms "Sliding $L_h \geq L_G$"and "Sliding $L_h \leq L_G$".

Доказательство. To satisfy Assumption 2 we need to provide an $(\delta(\varepsilon)/2, \sigma_0(\varepsilon, \sigma))$-solution to the problem (186) and $(\delta(\varepsilon), \sigma_0(\varepsilon, \sigma), 2L_g)$-oracle of g in (186), where $L_g = L_G + 2L_G^2/(\mu_y + H)$.
By Lemma 2 with $F(x,y) = G(x,y)$, $w(y) = h(y) + \frac{H}{2}\|y - y_0\|^2$, $\delta = \delta(\varepsilon)$ and $\sigma_0 = \sigma_0(\varepsilon, \sigma)$ applied to the problem (186), if we find a $(\delta/2, \sigma_0)$-solution $\tilde{y}_{\delta/2}(x)$ of the problem (186), then $\nabla_x G\left(x, \tilde{y}_{\delta/2}(x)\right)$ is $(\delta, \sigma_0, 2L_g)$-oracle of g and its calculation requires τ_G calls of the oracle $\nabla_x G(\cdot, y)$. To finish the proof, we now focus on obtaining a $(\delta/2, \sigma_0)$-solution $\tilde{y}_{\delta/2}(x)$ of the problem (186). For this we consider two cases $L_h \geq L_G$ and $L_h \leq L_G$ and for each one we construct a two-loop procedure described below. We begin with the case $L_h \geq L_G$.

Sliding for $L_h \geq L_G$, Loop 1

The goal of Loop 1 is to find an $(\delta(\varepsilon)/2, \sigma_0(\varepsilon, \sigma))$-solution of problem (186) as a maximization problem in y. To obtain such an approximate solution, we change the sign of this optimization problem and apply Algorithm 2 with

$$\varphi = -G(x,y), \quad \psi = h(y) + \frac{H}{2}\|y - y_0\|^2. \tag{189}$$

Function φ is convex and has L_G-Lipschitz continuous gradient, function ψ is $H + \mu_y$-strongly convex and has $L_h + H$-Lipschitz continuous gradient. Thus, we can apply Algorithm 2 with exact oracles and parameter $H_1 \geq 2L_G$, which will be chosen later, to solve problem (186). To satisfy the conditions of Theorem 4, which gives the complexity of Algorithm 2, we, first, observe that the oracles of φ and ψ are exact and, second, observe that we need in each iteration of Algorithm 1, used as a building block in Algorithm 2, to find an $\left(\tilde{\varepsilon}_f^{(1)}(\delta/2), \tilde{\sigma}^{(1)}(\delta/2, \sigma_0)\right)$-solution to the auxiliary problem (15), which in this case has the following form:

$$\begin{aligned} z_{k+1}^t &= \arg\min_{z \in \mathbb{R}^{d_y}} \{\langle \nabla\varphi(z_k^{md}), z - z_k^{md}\rangle + \psi(z) + \frac{H_1}{2}\|z - z_k^{md}\|_2^2\} \\ &= \arg\min_{z \in \mathbb{R}^{d_y}} \{-\langle \nabla_z G(x, z_k^{md}), z - z_k^{md}\rangle + h(z) + \frac{H}{2}\|z - y_0\|^2 + \frac{H_1}{2}\|z - z_k^{md}\|_2^2\}, \end{aligned} \tag{190}$$

where $\tilde{\sigma}^{(1)}(\delta/2, \sigma_0), \tilde{\varepsilon}_f^{(1)}(\delta/2)$ need to satisfy inequalities (34), (35). Below, in the Loop 2, we explain how to solve this auxiliary problem in such a way that these inequalities hold.

To summarize Loop 1, both main assumptions of Theorem 4 hold and we can use it to guarantee that we obtain an $(\delta/2, \sigma_0)$-solution of problem (186). Due to polynomial dependencies $\delta(\varepsilon) = \mathbf{poly}(\varepsilon)$, $\sigma_0(\varepsilon, \sigma) = \mathbf{poly}(\varepsilon, \sigma)$ this requires $\widetilde{O}\left(1 + \left(\frac{H_1}{\mu_\varphi + \mu_\psi}\right)^{\frac{1}{2}}\right) = \widetilde{O}\left(1 + \left(\frac{H_1}{\mu_y + H}\right)^{\frac{1}{2}}\right)$ calls to the (exact) oracles for φ and for ψ, and the same number of times solving the auxiliary problem (190). Combining this oracle complexity with the cost of calculating (exact) oracles for φ and for ψ, we obtain that solving problem (74) requires $\widetilde{O}\left(1 + \left(\frac{H_1}{\mu_y + H}\right)^{\frac{1}{2}}\right)\tau_G$ calls of the basic oracle for $G(x, \cdot)$ and $\widetilde{O}\left(\tau_h\left(1 + \left(\frac{H_1}{\mu_y + H}\right)^{\frac{1}{2}}\right)\right)$ of the basic oracles for h. The only remaining thing is to provide an inexact solution to problem (190) and, next, we move to Loop 2 to explain how to guarantee this. Note that we need to solve problem (190) $\widetilde{O}\left(1 + \left(\frac{H_1}{\mu_y + H}\right)^{\frac{1}{2}}\right)$ times.

Sliding for $L_h \geq L_G$, Loop 2

As mentioned in the previous Loop 1, in each iteration of Algorithm 2 in Loop 1 we need to find many times an $(\varepsilon_2, \sigma_2)$-solution of the auxiliary problem (190), where we denoted for simplicity $\sigma_2 = \tilde{\sigma}^{(1)}(\delta/2, \sigma_0)$ and $\varepsilon_2 = \tilde{\varepsilon}_f^{(1)}(\delta/2)$. To solve problem (190), we would like to apply Algorithm 2 with

$$\varphi = h(z), \quad \psi = -\langle \nabla_z G(x, z_k^{md}), z - z_k^{md} \rangle + \frac{H}{2}\|z - y_0\|^2 + \frac{H_1}{2}\|z - z_k^{md}\|_2^2. \tag{191}$$

Function φ is μ_y-strongly convex and has L_h-Lipschitz continuous gradient, function ψ is $H + H_1$-strongly convex and has $H + H_1$-Lipschitz continuous gradient. Thus, we can apply Algorithm 2 with exact oracles and parameter $H_2 \geq 2L_h$, which will be chosen later, to solve problem (190). To satisfy the conditions of Theorem 4, which gives the complexity of Algorithm 2, we, first, observe that the oracles of φ and ψ are exact and, second, observe that we need in each iteration of Algorithm 1, used as a building block in Algorithm 2, to find an $\left(\tilde{\varepsilon}_f^{(2)}(\varepsilon_2), \tilde{\sigma}^{(2)}(\varepsilon_2, \sigma_2)\right)$-solution to the auxiliary problem (15), which in this case has the following form:

$$\begin{aligned} u_{m+1}^t &= \arg\min_{u \in \mathbb{R}^{d_x}} \{\langle \nabla\varphi(u_m^{md}), u - u_m^{md} \rangle + \psi(u) + \frac{H_2}{2}\|u - u_m^{md}\|_2^2\} \\ &= \arg\min_{u \in \mathbb{R}^{d_x}} \{\langle \nabla h(u_m^{md}), u - u_m^{md} \rangle - \langle \nabla_z G(x, z_k^{md}), u - z_k^{md} \rangle \\ &+ \frac{H}{2}\|u - y_0\|^2 + \frac{H_1}{2}\|u - z_k^{md}\|_2^2 + \frac{H_2}{2}\|u - u_m^{md}\|_2^2\}. \end{aligned} \tag{192}$$

This quadratic auxiliary problem (192) can be solved explicitly and exactly. Thus, the second main assumption of Theorem 4 is satisfied with $\tilde{\sigma}^{(2)}(\varepsilon_2, \sigma_2) = 0, \tilde{\varepsilon}_f^{(2)}(\varepsilon_2) = 0$, which clearly satisfy (31) and (32).

To summarize Loop 2, both main assumptions of Theorem 4 hold and we can use it to guarantee that we obtain an $(\varepsilon_2, \sigma_2)$-solution of the auxiliary problem (190). This requires $O\left(\left(1 + \left(\frac{H_2}{\mu_\varphi + \mu_\psi}\right)^{\frac{1}{2}}\right)\log \varepsilon_2^{-1}\right) = O\left(\left(1 + \left(\frac{H_2}{\mu_y + H + H_1}\right)^{\frac{1}{2}}\right)\log \varepsilon_2^{-1}\right)$ calls to the (exact) oracles for φ and for ψ, and the same number of times solving the auxiliary problem (192). Combining this oracle complexity with the cost of calculating (exact) oracles for φ and for ψ, we obtain that solving problem (190) requires $O\left(\tau_h\left(1 + \left(\frac{H_2}{\mu_y + H + H_1}\right)^{\frac{1}{2}}\right)\log \varepsilon_2^{-1}\right)$ calls of the basic oracle for h. Also according to the polynomial dependencies (34), (35) we obtain that

$$\sigma_2 = \tilde{\sigma}^{(1)}(\delta/2, \sigma_0) = \mathrm{poly}(\delta/2, \sigma_0), \quad \varepsilon_2 = \tilde{\varepsilon}_f^{(1)}(\delta/2, \sigma_0) = \mathrm{poly}(\delta/2, \sigma_0).$$

Using conditions $\delta(\varepsilon) = \mathbf{poly}(\varepsilon)$, $\sigma_0(\varepsilon, \sigma) = \mathbf{poly}(\varepsilon, \sigma)$ in the formulation of Assumption 2 we obtain that the dependencies

$$\sigma_2(\varepsilon, \sigma), \tilde{\sigma}^{(1)}(\varepsilon, \sigma), \varepsilon_2(\varepsilon, \sigma), \tilde{\varepsilon}_f^{(1)}(\varepsilon, \sigma)$$

are polynomial. Then, we can use notation $\widetilde{O}(\cdot)$ without specifying what precision we mean and implying that the logarithmic part depends on the initial ε, σ.

Sliding $L_h \geq L_G$, combining the estimates of both loops

Combining the estimates of the above Loop 1 and Loop 2 we see that, finding a point $\tilde{y}_{\delta/2}(x)$ that is a $(\delta(\varepsilon)/2, \sigma_0(\varepsilon,\sigma))$-solution to the problem (186) requires the following number of calls of the basic oracles of $G(x,\cdot)$ and h respectively

$$\widetilde{O}\left(\tau_G + \tau_G\sqrt{H_1/(H+\mu_y)}\right), \tag{193}$$

$$\widetilde{O}\left(\tau_h\left(1+\sqrt{H_1/(H+\mu_y)}\right) + \left(1+\sqrt{H_1/(H+\mu_y)}\right)\tau_h\left(1+\sqrt{\frac{H_2}{\mu_y+H+H_1}}\right)\right). \tag{194}$$

Finding $(\delta(\varepsilon), \sigma_0(\varepsilon,\sigma), 2L_g)$-oracle of g by calculating $\nabla_x G\left(x, \tilde{y}_{\delta/2}(x)\right)$ requires additionally $\tau_G = m_G$ calls of the basic oracle for $G(\cdot,y)$. Since in Assumption 2 we denote the dependence on the target accuracy ε and confidence level σ by a separate quantities denoted by $\mathcal{K}(\varepsilon,\sigma)$ and in this case it is logarithmic, choosing $H_1 = 2L_G$ and $H_2 = 2L_h$ we get the final estimates for $\mathcal{N}_G^y$ and $\mathcal{N}_h$ to guarantee that Assumption 2 holds:

$$\begin{aligned}
\mathcal{N}_G^y &= O\left(\tau_G + \tau_G\sqrt{\frac{L_G}{\mu_y+H}}\right),\\
\mathcal{N}_h &= O\left(\tau_h\left(1+\sqrt{2L_G/(H+\mu_y)}\right)\left(1+\sqrt{\frac{2L_h}{\mu_y+H+2L_G}}\right)\right)\\
&= O\left(1+\sqrt{\frac{2L_G}{\mu_y+H}}+\sqrt{\frac{2L_h}{\mu_y+H}}+\sqrt{\frac{2L_G}{H+\mu_y}}\sqrt{\frac{2L_h}{\mu_y+H+2L_G}}\right)\tau_h\\
&= O\left(\tau_h\left(1+\sqrt{\frac{L_h}{\mu_y+H}}\right)\right),
\end{aligned}$$

where we used that $L_h \geq L_G$

Our aim now is to obtain the same estimates on $\mathcal{N}_G^y$ and $\mathcal{N}_h$ for the case when $L_h \leq L_G$. We do this by changing the order of Loop 1 and Loop 2 in the construction of previous Algorithm.

Sliding for $L_h \leq L_G$, Loop 1

The goal of Loop 1 is to find an $(\delta(\varepsilon)/2, \sigma_0(\varepsilon,\sigma))$-solution of problem (186) as a maximization problem in y. To obtain such an approximate solution, we change the sign of this optimization problem and apply Algorithm 2 with

$$\varphi = h(y), \quad \psi = -G(x,y) + \frac{H}{2}\|y-y_0\|^2. \tag{195}$$

Function φ is μ_y-strongly convex and has L_h-Lipschitz continuous gradient, function ψ is H-strongly convex and has L_h+H-Lipschitz continuous gradient. Thus, we can apply Algorithm 2 with exact oracles and parameter $H_1 \geq 2L_h$, which will be chosen later, to solve problem (186). To satisfy the conditions of Theorem 4, which gives the complexity of Algorithm 2, we, first, observe that the oracles of φ and ψ are exact and, second, observe that we need in each iteration of Algorithm 1, used as a building block in Algorithm 2, to find an $\left(\tilde{\varepsilon}_f^{(1)}\left(\delta/2\right), \tilde{\sigma}^{(1)}\left(\delta/2, \sigma_0\right)\right)$-solution to the auxiliary problem (15), which in this case has the following form:

$$\begin{aligned} z_{k+1}^t &= \arg\min_{z\in\mathbb{R}^{d_y}} \{\langle\nabla\varphi(z_k^{md}), z - z_k^{md}\rangle + \psi(z) + \frac{H_1}{2}\|z - z_k^{md}\|_2^2\} \\ &= \arg\min_{z\in\mathbb{R}^{d_y}} \{\langle\nabla_z h(z_k^{md}), z - z_k^{md}\rangle - G(x, z) + \frac{H}{2}\|z - y_0\|^2 + \frac{H_1}{2}\|z - z_k^{md}\|_2^2\}, \end{aligned} \tag{196}$$

where $\tilde{\sigma}^{(1)}\left(\delta/2, \sigma_0\right), \tilde{\varepsilon}_f^{(1)}\left(\delta/2\right)$ need to satisfy inequalities (34), (35). Below, in the Loop 2, we explain how to solve this auxiliary problem in such a way that these inequalities hold.

To summarize Loop 1, both main assumptions of Theorem 4 hold and we can use it to guarantee that we obtain an $(\delta/2, \sigma_0)$-solution of problem (186). Due to polynomial dependencies $\delta\left(\varepsilon\right) = \mathbf{poly}\left(\varepsilon\right)$, $\sigma_0\left(\varepsilon, \sigma\right) = \mathbf{poly}\left(\varepsilon, \sigma\right)$ this requires $\widetilde{O}\left(1 + \left(\frac{H_1}{\mu_\varphi + \mu_\psi}\right)^{\frac{1}{2}}\right) = \widetilde{O}\left(1 + \left(\frac{H_1}{\mu_y + H}\right)^{\frac{1}{2}}\right)$ calls to the (exact) oracles for φ and for ψ, and the same number of times solving the auxiliary problem (196). Combining this oracle complexity with the cost of calculating (exact) oracles for φ and for ψ, we obtain that solving problem (74) requires $\widetilde{O}\left(1 + \left(\frac{H_1}{\mu_y + H}\right)^{\frac{1}{2}}\right)\tau_G$ calls of the basic oracle for $G(x, \cdot)$ and $\widetilde{O}\left(1 + \left(\frac{H_1}{\mu_y + H}\right)^{\frac{1}{2}}\right)\tau_h$ of the basic oracles for h. The only remaining thing is to provide an inexact solution to problem (196) and, next, we move to Loop 2 to explain how to guarantee this. Note that we need to solve problem (196) $\widetilde{O}\left(1 + \left(\frac{H_1}{\mu_y + H}\right)^{\frac{1}{2}}\right)$ times.

Sliding for $L_h \leq L_G$, Loop 2

As mentioned in the previous Loop 1, in each iteration of Algorithm 2 in Loop 2 we need to find many times an $(\varepsilon_2, \sigma_2)$-solution of the auxiliary problem (196), where we denoted for simplicity $\sigma_2 = \tilde{\sigma}^{(1)}\left(\delta/2, \sigma_0\right)$ and $\varepsilon_2 = \tilde{\varepsilon}_f^{(1)}\left(\delta/2\right)$. To solve problem (196), we would like to apply Algorithm 2 with

$$\varphi = -G(x, z), \quad \psi = \langle\nabla h(z_k^{md}), z - z_k^{md}\rangle + \frac{H}{2}\|z - y_0\|^2 + \frac{H_1}{2}\|z - z_k^{md}\|_2^2. \tag{197}$$

Function φ is convex and has L_G-Lipschitz continuous gradient, function ψ is $H+H_1+\mu_y$-strongly convex and has $H+H_1$-Lipschitz continuous gradient. Thus, we can apply Algorithm 2 with exact oracles and parameter $H_2 \geq 2L_G$, which will be chosen later, to solve problem (196). To satisfy the conditions of Theorem 4, which gives the complexity of Algorithm 2, we, first, observe that the oracles of φ and ψ are exact and, second, observe that we need in each iteration of Algorithm 1, used as a building block in Algorithm 2, to find an $\left(\tilde{\varepsilon}_f^{(2)}\left(\varepsilon_2\right), \tilde{\sigma}^{(2)}\left(\varepsilon_2, \sigma_2\right)\right)$-solution to the auxiliary problem (15), which in this case has the following form:

$$\begin{aligned} u_{m+1}^t &= \arg\min_{u\in\mathbb{R}^{d_x}} \{\langle\nabla\varphi(u_m^{md}), u - u_m^{md}\rangle + \psi(u) + \frac{H_2}{2}\|u - u_m^{md}\|_2^2\} \\ &= \arg\min_{u\in\mathbb{R}^{d_x}} \{-\langle\nabla_u G(x, u_m^{md}), u - u_m^{md}\rangle + \langle\nabla h(z_k^{md}), u - z_k^{md}\rangle \qquad (198) \\ &+ \frac{H}{2}\|u - y_0\|^2 + \frac{H_1}{2}\|u - z_k^{md}\|_2^2 + \frac{H_2}{2}\|u - u_m^{md}\|_2^2\}. \qquad (199) \end{aligned}$$

This quadratic auxiliary problem (199) can be solved explicitly and exactly. Thus, the second main assumption of Theorem 4 is satisfied with $\tilde{\sigma}^{(2)}\left(\varepsilon_2,\sigma_2\right)=0, \tilde{\varepsilon}_f^{(2)}\left(\varepsilon_2\right)=0$, which clearly satisfy (31) and (32).

To summarize Loop 2, both main assumptions of Theorem 4 hold and we can use it to guarantee that we obtain an (ε_2,σ_2)-solution of the auxiliary problem (196). This requires $O\left(\left(1+\left(\frac{H_2}{\mu_\varphi+\mu_\psi}\right)^{\frac{1}{2}}\right)\log\varepsilon_2^{-1}\right)=O\left(\left(1+\left(\frac{H_2}{\mu_y+H+H_1}\right)^{\frac{1}{2}}\right)\log\varepsilon_2^{-1}\right)$ calls to the (exact) oracles for φ and for ψ, and the same number of times solving the auxiliary problem (199). Combining this oracle complexity with the cost of calculating (exact) oracles for φ and for ψ, we obtain that solving problem (196) requires $O\left(\left(1+\left(\frac{H_2}{\mu_y+H+H_1}\right)^{\frac{1}{2}}\right)\log\varepsilon_2^{-1}\right)\tau_G$ calls of the basic oracle for $G(x,\cdot)$. Also according to the polynomial dependences (34), (35) we obtain that

$$\sigma_2=\tilde{\sigma}^{(1)}\left(\delta/2,\sigma_0\right)=\mathrm{poly}(\delta/2,\sigma_0),\;\;\varepsilon_2=\tilde{\varepsilon}_f^{(1)}\left(\delta/2,\sigma_0\right)=\mathrm{poly}(\delta/2,\sigma_0).$$

Using conditions $\delta\left(\varepsilon\right)=\mathbf{poly}\left(\varepsilon\right)$, $\sigma_0\left(\varepsilon,\sigma\right)=\mathbf{poly}\left(\varepsilon,\sigma\right)$ in the formulation of Asumption 2 we obtain that the dependencies

$$\sigma_2\left(\varepsilon,\sigma\right),\tilde{\sigma}^{(1)}\left(\varepsilon,\sigma\right),\varepsilon_2\left(\varepsilon,\sigma\right),\tilde{\varepsilon}_f^{(1)}\left(\varepsilon,\sigma\right)$$

are polynomial. Then, we can use notation $\widetilde{O}(\cdot)$ without specifying what precision we mean and implying that the logarithmic part depends on the initial ε,σ.

Sliding for $L_h\le L_G$, combining the estimates of both loops

Combining the estimates of the above Loop 1 and Loop 2 we see that, finding a point $\tilde{y}_{\delta/2}(x)$ which is an $\left(\delta\left(\varepsilon\right)/2,\sigma_0\left(\varepsilon,\sigma\right)\right)$-solution to the problem (186) requires the following number of calls of the basic oracles of h and $G(x,\cdot)$ respectively

$$\widetilde{O}\left(1+\sqrt{H_1/(H+\mu_y)}\right)\tau_h, \tag{200}$$

$$\widetilde{O}\left(\tau_G+\tau_G\sqrt{H_1/(H+\mu_y)}+\left(1+\sqrt{H_1/(H+\mu_y)}\right)\left(\tau_G+\tau_G\sqrt{\frac{H_2}{\mu_y+H+H_1}}\right)\right). \tag{201}$$

Finding $\left(\delta\left(\varepsilon\right),\sigma_0\left(\varepsilon,\sigma\right),2L_g\right)$-oracle of g by calculating $\nabla_x G\left(x,\tilde{y}_{\delta/2}(x)\right)$ requires additionally τ_G calls of the basic oracle for $G(\cdot,y)$. Since in Assumption 2 we denote the dependence on the target accuracy ε and confidence level σ by a separate quantities denoted by $\mathcal{K}(\varepsilon,\sigma)$ and in this case it is logarithmic, choosing $H_1=2L_h$ and $H_2=2L_G$ we get the final estimates for $\mathcal{N}_G^y$ and $\mathcal{N}_h$ to guarantee that Assumption 2 holds:

$$\mathcal{N}_G^y=O\left(\left(1+\sqrt{2L_h/(H+\mu_y)}\right)\left(1+\sqrt{\frac{2L_G}{\mu_y+H+2L_h}}\right)\right)\tau_G$$

$$=O\left(1+\sqrt{\frac{2L_h}{\mu_y+H}}+\sqrt{\frac{2L_G}{\mu_y+H}}+\sqrt{\frac{2L_h}{\mu_y+H}}\sqrt{\frac{2L_G}{\mu_y+H+2L_h}}\right)\tau_G=O\left(\tau_G+\tau_G\sqrt{\frac{L_G}{\mu_y+H}}\right),$$

$$\mathcal{N}_h=O\left(1+\sqrt{\frac{L_h}{\mu_y+H}}\right)\tau_h,$$

where for the first bound we used that $L_h \leq L_G$.

It is important to note that the estimates on $\mathcal{N}_G^y$ and $\mathcal{N}_h$ obtained in both cases $L_h \geq L_G$ and $L_h \leq L_G$ are exactly the same. Thus, regardless of the relation between L_h and L_G, we obtain the estimates in the statement of the Lemma. Yet, we underline that the algorithm actually depends on whether $L_h \geq L_G$ or $L_h \leq L_G$. □

We now obtain a simple counterpart of the previous Lemma for the case when Assumption 5.3(b) holds instead of Assumption 5.3(a). In this case h is prox-friendly and there is no need to consider different cases and just one Loop is enough since the auxiliary problem (190) in Loop 1 can be solved explicitly.

Lemma 15. Let the function g be defined via maximization problem in (49), i.e.

$$g(x) = \max_{y \in \mathbb{R}^{d_y}} \left\{ G(x,y) - h(y) - \frac{H}{2}\|y - y_0\|^2 \right\}, \tag{202}$$

where $G(x,y)$, $h(y)$ are according to (95) and satisfy Assumption 5.1,2,3(b), $y_0 \in \mathbb{R}^{d_y}$. Then, applying Algorithm 2 to this problem, we guarantee that Assumption 2 holds with τ_G basic oracle calls for $G(\cdot, y)$ and the following estimates for the number of basic oracle calls for $G(x, \cdot)$ and h respectively

$$\mathcal{N}_G^y(\tau_G, H) = O\left(\tau_G + \tau_G\sqrt{L_G/(H + \mu_y)}\right), \tag{203}$$

$$\mathcal{N}_h(\tau_h, H) = 0. \tag{204}$$

Доказательство. The proof is similar to the proof for the case "Sliding $L_h \geq L_G$"in the proof of Lemma 14 with the only change that the auxiliary problem (190) is solved explicitly thanks to h being prox-friendly. □

By changing the variables x and y in Lemma 14 and choosing $H = 0$ we obtain the following simple corollary that ensures Assumption 3.

Corollar 4. Let the function r be defined via maximization problem in (50), i.e.

$$r(y) = \min_{x \in \mathbb{R}^{d_x}} \{G(x,y) + f(x)\}, \tag{205}$$

where $G(x,y), f(y)$ are according to (95) and satisfy Assumption 5.1,2,3(a).Then, for each of two cases $L_f \geq L_G$ and $L_f \leq L_G$ we organize computations in two loops and apply Algorithm 2, so that Assumption 3 holds with τ_G basic oracle calls for $G(x, \cdot)$ and the following estimates for the number of basic oracle calls for $G(\cdot, y)$ and f respectively

$$\mathcal{N}_G^x(\tau_G) = O\left(\tau_G + \tau_G\sqrt{L_G/\mu_x}\right), \tag{206}$$

$$\mathcal{N}_f(\tau_f) = O\left(\tau_f + \tau_f\sqrt{L_f/\mu_x}\right). \tag{207}$$

We name these algorithms "Sliding $L_f \geq L_G$"and "Sliding $L_f \leq L_G$".

Different regimes	$L_h \geq L_G$	$L_h \leq L_G$
$L_f \leq L_G$	Framework from Appendix (Theorem 10) + Sliding for $L_h \geq L_G$ (Lemma 14) + Sliding for $L_f \leq L_G$ (Corollary 4)	Framework from Appendix (Theorem 10) + Sliding for $L_h \leq L_G$ (Lemma 14) + Sliding for $L_f \leq L_G$ (Corollary 4)
$L_f \geq L_G$	General Framework (Theorem 5) + Sliding for $L_h \geq L_G$ (Lemma 14) + Sliding for $L_f \geq L_G$ (Corollary 4)	General Framework (Theorem 5) + Sliding for $L_h \leq L_G$ (Lemma 14) + Sliding for $L_f \geq L_G$ (Corollary 4)

Table 6. Summary of the proof of Theorem 3. For each regime we apply the algorithms described in the proofs of the corresponding results listed in the table to obtain the complexity estimates (103)-(106) for the number of basic oracle calls for each part of the objective f, h, and G.

Proof of Theorem 7

Finally, we prove Theorem 7 for problem (95) by combining the building blocks depending on the relation between L_f and L_G and relation between L_h and L_G. If $L_f \geq L_G$ we use the general framework from the main text (see Section 1 and Theorem 5). In the opposite case we apply the variation of this framework described in Appendix (see Theorem 10). In both cases we use Lemma 14 and Corollary 4 to ensure Assumptions 2, 3, but with different order of the loops described inside these Lemma and Corollary depending on the relation between L_h and L_G, i.e. we use either sliding $L_h \geq L_G$ or sliding $L_h \leq L_G$ in Lemma 14 and either sliding $L_f \geq L_G$ or sliding $L_f \leq L_G$ in Corollary 4. For convenience, we summarize which results are used in which case in Table 6.

Proof of Theorem 7

Assumption 5.1,2,3(a) with (97) guarantee that Assumption 1 holds. Further, the choice $H = 2L_G$ in Lemma 14 guarantee that Assumption 2 holds with the number of oracle calls given by (187) and (188). Corollary 4 guarantee that Assumption 3 holds with the number of oracle calls given by (206) and (207). We consider two cases $L_f \geq L_G$ and $L_f \leq L_G$ and, for each case, apply either the general framework from the main text or from the previous appendix. We show that in both cases the estimates are the same and are equal to the ones in the statement of the theorem. In each case we make the derivations with $\sigma = 0$ since all the algorithms are deterministic in this case.

We begin with the case $L_f \geq L_G$.

Case $L_f \geq L_G$

Applying Theorem 5 with $\tau_f = \tau_h = 1$ and $\tau_G = m_G$, Lemma 14 with $H = 2L_G$, Corollary 4 and combining the complexity estimates in these results, we obtain the following final complexity bounds.

Number of basic oracle calls of f:

$$\begin{aligned}
&\widetilde{O}\left(\left(1+\sqrt{\frac{L_G}{\mu_y}}\right)\left(1+\sqrt{\frac{L_f}{\mu_x}}+\left(1+\sqrt{\frac{L_G}{\mu_x}}\right)\left(1+\sqrt{\frac{L_f}{L_G}}\right)\right)\right)\\
&=\widetilde{O}\left(\left(\sqrt{\frac{L_G}{\mu_y}}\right)\left(\sqrt{\frac{L_f}{\mu_x}}+\left(\sqrt{\frac{L_G}{\mu_x}}\right)\left(\sqrt{\frac{L_f}{L_G}}\right)\right)\right)\\
&=\widetilde{O}\left(\left(\sqrt{\frac{L_f L_G}{\mu_x\mu_y}}\right)\right),
\end{aligned}$$

where we used that, $L_G \leq L_f$ and, by the assumptions of this Theorem, $1 \leq L_G/\mu_y$, $1 \leq L_G/\mu_x$, $1 \leq L_f/\mu_x$.

Number of basic oracle calls of h:

$$\begin{aligned}
&\widetilde{O}\left(\left(1+\sqrt{\frac{L_G}{\mu_y}}\right)\left(1+\left(1+\sqrt{\frac{L_G}{\mu_x}}\right)\left(1+\sqrt{\frac{L_h}{2L_G+\mu_y}}\right)\right)\right)\\
&=\widetilde{O}\left(\left(\sqrt{\frac{L_G}{\mu_y}}\right)\left(1+\left(\sqrt{\frac{L_G}{\mu_x}}\right)\left(1+\sqrt{\frac{L_h}{L_G}}\right)\right)\right)\\
&=\widetilde{O}\left(\max\left\{\sqrt{\frac{L_G L_h}{\mu_x\mu_y}},\sqrt{\frac{L_G^2}{\mu_x\mu_y}}\right\}\right),
\end{aligned}$$

where we used that $H = 2L_G$ in Lemma 14 and, by the assumptions of this Theorem, $1 \leq L_G/\mu_y$, $1 \leq L_G/\mu_x$.

Number of basic oracle calls of $G(\cdot, y)$:

$$\widetilde{O}\left(\left(1+\sqrt{\frac{L_G}{\mu_y}}\right)\left(m_G+m_G\sqrt{\frac{L_G}{\mu_x}}+m_G\left(1+\sqrt{\frac{L_G}{\mu_x}}\right)\right)\right)=\widetilde{O}\left(m_G\sqrt{\frac{L_G^2}{\mu_x\mu_y}}\right),$$

where we used that, by the assumptions of this Theorem, $1 \leq L_G/\mu_y$ and $1 \leq L_G/\mu_x$.

Number of basic oracle calls of $G(x, \cdot)$:

$$\begin{aligned}
&\widetilde{O}\left(\left(1+\sqrt{\frac{L_G}{\mu_y}}\right)\left(m_G+m_G\left(1+\sqrt{\frac{L_G}{\mu_x}}\right)\left(1+\sqrt{\frac{L_G}{2L_G+\mu_y}}\right)\right)\right)\\
&=\widetilde{O}\left(m_G\left(1+\sqrt{\frac{L_G}{\mu_y}}\right)\left(1+\sqrt{\frac{L_G}{\mu_x}}\right)\right)=\widetilde{O}\left(m_G\sqrt{\frac{L_G^2}{\mu_x\mu_y}}\right),
\end{aligned}$$

where we used that $H = 2L_G$ in Lemma 14 and, by the assumptions of this Theorem, $1 \leq L_G/\mu_y$ and $1 \leq L_G/\mu_x$.

Case $L_f \leq L_G$

Applying Theorem 10 with $\tau_f = \tau_h = 1$ and $\tau_G = m_G$, Lemma 14 with $H = 2L_G$, Corollary 4 and combining the, complexity estimates we obtain the final complexity bounds as follows.

Number of basic oracle calls of f:

$$\widetilde{O}\left(\left(1+\sqrt{\frac{L_G}{\mu_y}}\right)\left(1+\sqrt{\frac{L_f}{\mu_x}}+\left(1+\sqrt{\frac{L_f}{\mu_x}}\right)\right)\right)=\widetilde{O}\left(\left(\sqrt{\frac{L_G}{\mu_y}}\right)\left(\sqrt{\frac{L_f}{\mu_x}}+\left(\sqrt{\frac{L_f}{\mu_x}}\right)\right)\right)$$
$$=\widetilde{O}\left(\left(\sqrt{\frac{L_fL_G}{\mu_x\mu_y}}\right)\right),$$

where we used that, by the assumptions of this Theorem, $1 \le L_G/\mu_y$ and $1 \le L_f/\mu_x$.

Number of basic oracle calls of h:

$$\widetilde{O}\left(\left(1+\sqrt{\frac{L_G}{\mu_y}}\right)\left(1+\left(1+\sqrt{\frac{L_f}{\mu_x}}\right)\left(1+\sqrt{\frac{L_G}{L_f}}\right)\left(1+\sqrt{\frac{L_h}{2L_G+\mu_y}}\right)\right)\right)$$
$$=\widetilde{O}\left(\left(\sqrt{\frac{L_G}{\mu_y}}\right)\left(1+\left(\sqrt{\frac{L_f}{\mu_x}}\right)\cdot\left(\sqrt{\frac{L_G}{L_f}}\right)\cdot\left(1+\sqrt{\frac{L_h}{L_G}}\right)\right)\right)=\widetilde{O}\left(\max\left\{\sqrt{\frac{L_G^2}{\mu_x\mu_y}},\sqrt{\frac{L_GL_h}{\mu_x\mu_y}}\right\}\right),$$

where we used that $H = 2L_G$ in Lemma 14, $L_G \ge L_f$ and, by the assumptions of this Theorem, $1 \le L_G/\mu_y$, $1 \le L_f/\mu_x$.

Number of basic oracle calls of $G(\cdot, y)$:

$$\widetilde{O}\left(\left(1+\sqrt{\frac{L_G}{\mu_y}}\right)\left(m_G+m_G\sqrt{\frac{L_G}{\mu_x}}+m_G\left(1+\sqrt{\frac{L_f}{\mu_x}}\right)\left(1+\sqrt{\frac{L_G}{L_f}}\right)\right)\right)$$
$$=\widetilde{O}\left(m_G\left(\sqrt{\frac{L_G}{\mu_y}}\right)\left(\sqrt{\frac{L_G}{\mu_x}}+\left(\sqrt{\frac{L_f}{\mu_x}}\right)\cdot\left(\sqrt{\frac{L_G}{L_f}}\right)\right)\right)=\widetilde{O}\left(m_G\left(\sqrt{\frac{L_G}{\mu_y}}\right)\left(\sqrt{\frac{L_G}{\mu_x}}\right)\right)$$
$$=\widetilde{O}\left(m_G\sqrt{\frac{L_G^2}{\mu_x\mu_y}}\right),$$

where we used that $L_G \ge L_f$ and, by the assumptions of this Theorem, $1 \le L_G/\mu_y$, $1 \le L_G/\mu_x$, $1 \le L_f/\mu_x$.

Number of basic oracle calls of $G(x, \cdot)$:

$$\widetilde{O}\left(\left(1+\sqrt{\frac{L_G}{\mu_y}}\right)\left(m_G+\left(1+\sqrt{\frac{L_f}{\mu_x}}\right)\left(m_G+m_G\sqrt{\frac{L_G}{2L_G+\mu_y}}+\sqrt{\frac{L_G}{L_f}}\cdot m_G\left(1+\sqrt{\frac{L_G}{2L_G+\mu_y}}\right)\right)\right)\right)$$
$$=\widetilde{O}\left(m_G\left(\sqrt{\frac{L_G}{\mu_y}}\right)\left(1+\left(\sqrt{\frac{L_f}{\mu_x}}\right)\cdot\left(\sqrt{\frac{L_G}{L_f}}\right)\right)\right)=\widetilde{O}\left(\max\left\{m_G\sqrt{\frac{L_G}{\mu_y}},m_G\sqrt{\frac{L_G^2}{\mu_x\mu_y}}\right\}\right)$$
$$=\widetilde{O}\left(m_G\sqrt{\frac{L_G^2}{\mu_x\mu_y}}\right),$$

where we used that $H = 2L_G$ in Lemma 14, $L_G \ge L_f$ and, by the assumptions of this Theorem, $1 \le L_G/\mu_y$, $1 \le L_G/\mu_x$, $1 \le L_f/\mu_x$.

□

Proof of Theorem 8

The only difference in the proof of Theorem 8 from the proof of Theorem 7 is the use of Lemma 15 instead of Lemma 14 to satisfy Assumption 2. Thus, applying expressions (203), (204) for $\mathcal{N}_G^y$ and $\mathcal{N}_h$ and following the proof of Theorem 7 without any changes we obtain the same estimates for the number of basic oracle calls of $f, G(\cdot, y), G(x, \cdot)$. Considering $\mathcal{N}_h = 0$ and using that, by the assumptions of this Theorem, $1 \leq L_G/\mu_y$, we obtain that the number of basic oracle calls of h is

$$\widetilde{O}\left(\sqrt{\frac{L_G}{\mu_y}}\right).$$

□

Proof of Lemma 1 and Lemma 2

Let us proof Lemma 1

Доказательство. Using the Deffinition 2 for function φ and ψ, we can obtain:

$$\frac{\mu_\varphi}{2}\|x-y\|^2 \leq \varphi(x) - \left(\varphi_{\delta_\varphi, L_\varphi, \mu_\varphi}(y) + \left\langle \nabla\varphi_{\delta_\varphi, L_\varphi, \mu_\varphi}(y), x-y \right\rangle\right) \leq \frac{L_\varphi}{2}\|x-y\|^2 + \delta_\varphi \quad \text{w.p. } 1-\sigma_\varphi \tag{208}$$

$$\frac{\mu_\psi}{2}\|x-y\|^2 \leq \psi(x) - \left(\psi_{\delta_\psi, L_\psi, \mu_\psi}(y) + \left\langle \nabla\psi_{\delta_\psi, L_\psi, \mu_\psi}(y), x-y \right\rangle\right) \leq \frac{L_\psi}{2}\|x-y\|^2 + \delta_\psi \quad \text{w.p. } 1-\sigma_\psi \tag{209}$$

Let us sum this equations:

$$\frac{\mu_\varphi}{2}\|x-y\|^2 + \frac{\mu_\psi}{2}\|x-y\|^2 \leq \varphi(x) + \psi(x) - \varphi_{\delta_\varphi, L_\varphi, \mu_\varphi}(y) - \psi_{\delta_\psi, L_\psi, \mu_\psi}(y) \tag{210}$$

$$- \left\langle \nabla\varphi_{\delta_\varphi, L_\varphi, \mu_\varphi}(y) - \nabla\psi_{\delta_\psi, L_\psi, \mu_\psi}(y), x-y \right\rangle$$

$$\leq \frac{L_\varphi}{2}\|x-y\|^2 + \frac{L_\psi}{2}\|x-y\|^2 + \delta_\varphi + \delta_\psi \quad \text{w.p. } 1-\sigma_\varphi-\sigma_\psi \tag{211}$$

The equation (211) means that the pair $\left(\varphi_{\delta_\varphi, L_\varphi, \mu_\varphi}(y) + \psi_{\delta_\psi, L_\psi, \mu_\psi}(y),\right.$ $\left.\nabla\varphi_{\delta_\varphi, L_\varphi, \mu_\varphi}(y) + \nabla\psi_{\delta_\psi, L_\psi, \mu_\psi}(y)\right)$ is $(\delta_\varphi + \delta_\psi, \sigma_\varphi + \sigma_\psi, L_\varphi + L_\psi, \mu_\varphi + \mu_\psi)$-oracle for $\varphi + \psi$. □

Let us proof Lemma 2

Доказательство. The function $\hat{S}(x, \cdot)$ is μ_y-strongly concave, and $\hat{S}(\cdot, y)$ is differentiable. Therefore, by Demyanov–Danskin's theorem, for any $x \in \mathbb{R}^{d_x}$, we have

$$\nabla g(x) = \nabla_x \tilde{S}(x, y^*(x)) = \nabla_x F(x, y^*(x)). \tag{A1}$$

To prove that $g(\cdot)$ has an L–Lipschitz gradient for $L = L_F + \frac{2L_F^2}{\mu_y}$, let us prove the Lipschitz condition for $y^*(\cdot)$ with a constant, the function y^* is defined as:

$$y^*(x) := \arg\max_{y \in \mathbb{R}^{d_y}} \hat{S}(x, y) \quad \forall x \in \mathbb{R}^{d_x}, \tag{212}$$

Since $\hat{S}(x_1, \cdot)$ is μ_y-strongly concave, for arbitrary $x_1, x_2 \in \mathbb{R}^{d_x}$:

$$\|y^*(x_1) - y^*(x_2)\|_2^2 \leq \frac{2}{\mu_y}\left(\hat{S}(x_1, y^*(x_1)) - \hat{S}(x_1, y^*(x_2))\right). \tag{A2}$$

On the other hand, $\hat{S}(x_2, y^*(x_1)) - \hat{S}(x_2, y^*(x_2)) \leq 0$, since $y^*(x_2)$ affords the maximum to $\hat{S}(x_2, .)$ on $\mathbb{R}^{d_y}$. We have

$$\begin{aligned}
&\hat{S}(x_1, y^*(x_1)) - \hat{S}(x_1, y^*(x_2)) \leq \hat{S}(x_1, y^*(x_1)) - \hat{S}(x_1, y^*(x_2)) - \hat{S}(x_2, y^*(x_1)) + \hat{S}(x_2, y^*(x_2)) = \\
&\overset{\text{from (28)}}{=} (F(x_1, y^*(x_1)) - F(x_1, y^*(x_2))) - (F(x_2, y^*(x_1)) - F(x_2, y^*(x_2))) = \\
&= \int_0^1 \langle \nabla_x F(x_1 + t(x_2 - x_1), y^*(x_1)) - \nabla_x F(x_1 + t(x_2 - x_1), y^*(x_2)), x_2 - x_1 \rangle dt \leq \\
&\leq \|\nabla_x F(x_1 + t(x_2 - x_1), y^*(x_1)) - \nabla_x F(x_1 + t(x_2 - x_1), y^*(x_1))\|_2 \cdot \|x_2 - x_1\|_2 \leq \\
&\leq L_F \|y^*(x_1) - y^*(x_2)\|_2 \cdot \|x_2 - x_1\|_2.
\end{aligned} \tag{A3}$$

Thus, (A2) and (A3) imply the inequality

$$\|y^*(x_2) - y^*(x_1)\|_2 \leq \frac{2L_F}{\mu_y} \|x_2 - x_1\|_2, \tag{A4}$$

i.e., the function $y^*(\cdot)$ satisfies the Lipschitz condition with a constant $\frac{2L_F}{\mu_y}$. Next, from (A1), we obtain

$$\begin{aligned}
&\|\nabla g(x_1) - \nabla g(x_2)\|_2 = \|\nabla_x F(x_1, y^*(x_1)) - \nabla_x F(x_2, y^*(x_2))\|_2 = \\
&= \|\nabla_x F(x_1, y^*(x_1)) - \nabla_x F(x_1, y^*(x_2)) + \nabla_x F(x_1, y^*(x_2)) - \nabla_x F(x_2, y^*(x_2))\|_2 \leq \\
&\leq \|\nabla_x F(x_1, y^*(x_1)) - \nabla_x F(x_1, y^*(x_2))\|_2 + \|\nabla_x F(x_1, y^*(x_2)) - \nabla_x F(x_2, y^*(x_2))\|_2 \leq \\
&\leq L_F \|y^*(x_1) - y^*(x_2)\|_2 + L_F \|x_2 - x_1\|_2 = \\
&\overset{\text{from (A4)}}{=} \left(L_F + \frac{2L_F^2}{\mu_y} \right) \|x_2 - x_1\|_2.
\end{aligned}$$

This means that $g(\cdot)$ has an L–Lipschitz gradient with $L = L_F + \frac{2L_F^2}{\mu_y}$.

Let us now prove that $\nabla_x F\left(x, \tilde{y}_{\delta/2}(x)\right)$ is $(\delta, 2L_g)$-oracle of g, i.e.:

$$0 \leq g(z) - \left[\left\{ F(x, \tilde{y}_{\delta/2}(x)) - w(\tilde{y}_{\delta/2}(x)) \right\} + \langle \nabla_x F(x, \tilde{y}_{\delta/2}(x)), z - x \rangle \right] \leq \frac{2L}{2} \|z - x\|_2^2 + \delta, \tag{213}$$

First, we prove that, for any $\delta \geq 0$ and $x \in \mathbb{R}^{d_x}$

$$\|\nabla_x \hat{S}(x, \tilde{y}_{\delta/2}(x)) - \nabla g(x)\|_2 \leq L_F \sqrt{\frac{\delta}{\mu_y}}. \tag{A5}$$

For any $x \in \mathbb{R}^{d_x}$, , it is true that $\nabla_x \hat{S}(x, \tilde{y}_{\delta/2}(x)) = \nabla_x F(x, \tilde{y}_{\delta/2}(x))$. Then,

$$\begin{aligned}
\|\nabla_x \hat{S}(x, \tilde{y}_{\delta/2}(x)) - \nabla g(x)\|_2^2 &= \|\nabla_x F(x, \tilde{y}_{\delta/2}(x)) - \nabla_x F(x, y^*(x))\|_2^2 \leq \\
&\leq L_F^2 \|y^*(x) - \tilde{y}_{\delta/2}(x)\|_2^2 \leq \\
&\overset{\text{from (A2)}}{\leq} \frac{2L_F^2}{\mu_y} \left(\hat{S}(x, y^*(x)) - \hat{S}(x, \tilde{y}_{\delta/2}(x)) \right) \leq \\
&\overset{\text{from (30)}}{\leq} \frac{\delta L_F^2}{\mu_y},
\end{aligned}$$

which justifies inequality (A5).

Now , due to the μ_x-strong convexity of $\hat{S}(\cdot, \tilde{y}_{\delta/2}(x))$ on $\mathbb{R}^{d_x}$, for arbitrary $x, z \in \mathbb{R}^{d_x}$ it is true that

$$g(z) \overset{\text{from (28)}}{\geq} \hat{S}(z, \tilde{y}_{\delta/2}(x)) \geq \hat{S}(x, \tilde{y}_{\delta/2}(x)) + \langle \nabla_x \hat{S}(x, \tilde{y}_{\delta/2}(x)), z - x \rangle.$$

Thus,

$$0 \geq \hat{S}(x, \tilde{y}_{\delta/2}(x)) - g(z) + \langle \nabla_x \hat{S}(x, \tilde{y}_{\delta/2}(x)), z - x \rangle,$$

which proves the left-hand side of (213). To prove the right-hand side of (213), note that g is convex and has an L–Lipschitz gradient on $\mathbb{R}^{d_x}$. Therefore, for arbitrary $x, z \in \mathbb{R}^{d_x}$, we have

$$\begin{aligned}
&g(z) \leq g(x) + \langle \nabla g(x), z - x \rangle + \frac{L}{2}\|z - x\|_2^2 \leq \\
&\overset{\text{from (30)}}{\leq} \hat{S}(x, \tilde{y}_{\delta/2}(x)) + \delta/2 + \frac{L}{2}\|z - x\|_2^2 + \langle \nabla g(x), z - x \rangle + \langle \nabla_x \hat{S}(x, \tilde{y}_{\delta/2}(x)), x - z \rangle - \\
&- \langle \nabla_x \hat{S}(x, \tilde{y}_{\delta/2}(x)), x - z \rangle = \\
&= \hat{S}(x, \tilde{y}_{\delta/2}(x)) + \delta/2 + \langle \nabla_x \hat{S}(x, \tilde{y}_{\delta/2}(x)), z - x \rangle + \langle \nabla_x \hat{S}(x, \tilde{y}_{\delta/2}(x)) - \nabla g(x), x - z \rangle + \\
&+ \frac{L}{2}\|z - x\|_2^2 \leq \\
&\overset{\text{from (A5)}}{\leq} \hat{S}(x, \tilde{y}_{\delta/2}(x)) + \delta/2 + \langle \nabla_x \hat{S}(x, \tilde{y}_{\delta/2}(x)), z - x \rangle + L_F \sqrt{\frac{\delta}{\mu_y}} \cdot \|z - x\|_2 + \frac{L}{2}\|z - x\|_2^2.
\end{aligned}$$

However,

$$L_F \sqrt{\frac{\delta}{\mu_y}} \cdot \|z - x\|_2 = \sqrt{\frac{L_F^2}{\mu_y}\|z - x\|_2^2 \cdot \delta} \leq \frac{L_F^2}{2\mu_y}\|z - x\|_2^2 + \delta/2$$

due to the classical inequality between the arithmetic and geometric mean. Therefore,

$$g(z) \leq \hat{S}(x, \tilde{y}_{\delta/2}(x)) + \delta + \langle \nabla_x \hat{S}(x, \tilde{y}_{\delta/2}(x)), z - x \rangle + \frac{L_F^2}{\mu_y}\|z - x\|_2^2 + \frac{L}{2}\|z - x\|_2^2,$$

and since $L = L_F + \frac{2L_F^2}{\mu_y}$, we have $\frac{L_F^2}{\mu_y} \leq \frac{L}{2}$;therefore,

$$g(z) \leq \hat{S}(x, \tilde{y}_{\delta/2}(x)) + \langle \nabla_x \hat{S}(x, \tilde{y}_{\delta/2}(x)), z - x \rangle + \delta + L\|z - x\|_2^2.$$

Thus, we have

$$g(z) - \hat{S}(x, \tilde{y}_{\delta/2}(x)) - \langle \nabla_x \hat{S}(x, \tilde{y}_{\delta/2}(x)), z - x \rangle \leq L\|z - x\|_2^2 + \delta,$$

which implies the left-hand side of inequality (213).

In the statement of Lemma 2 only $(\delta/2, \sigma)$-solution to (28) is available. In this case the inequality (213) will be satisfied with probability $1-\sigma$. Then $\nabla_x F\left(x, \tilde{y}_{\delta/2}(x)\right)$ is $(\delta, \sigma, 2L_g)$-oracle of g. □

Proof of Lemma 3

We let $\Phi(y) = \max_{x\in\mathbb{R}^{d_x}} \{h(y) - G(x, y) - f(x)\}$ and note that $\Phi(y)$ is μ_y-strongly convex. Under Condition 1 the function $h(y) - G(x, y) - f(x)$ has unique saddle point (x_*, y_*). Then, with probability $1 - \sigma_y$ we have

$$\|\hat{y} - y_*\|^2 \le \frac{2}{\mu_y}\left(\max_{x\in\mathbb{R}^{d_x}} \{h(\hat{y}) - G(x, \hat{y}) - f(x)\} - \min_{y\in\mathbb{R}^{d_y}} \max_{x\in\mathbb{R}^{d_x}} \{h(y) - G(x, y) - f(x)\}\right) \le \frac{2\varepsilon_y}{\mu_y}. \tag{214}$$

We denote $x^*(\hat{y}) = \arg\max_{x\in\mathbb{R}^{d_x}} \{h(\hat{y}) - G(x, \hat{y}) - f(x)\}$, then according to Lemma 2 $x^*(y)$ is $2L_G/\mu_x$ Lipschitz continuous. Since $\{h(\hat{y}) - G(x, \hat{y}) - f(x)\}$ is μ_x-strongly concave, we obtain that the inequality

$$\|\hat{x} - x_*\|^2 \le 2\|\hat{x} - x_*(\hat{y})\|^2 + 2\|x_*(\hat{y}) - x_*(y_*)\|^2 \le \frac{4\varepsilon_x}{\mu_x} + 8\left(\frac{L_G}{\mu_x}\right)^2 \|\hat{y} - y_*\|^2 \tag{215}$$

holds true with probability $1-\sigma_x-\sigma_y$. By consecutive application of Lemma 1 and Lemma 2 we can obtain that $\Psi(x) = \min_{y\in\mathbb{R}^{d_y}} \{h(y) - G(x, y) - f(x)\}$ is concave and $L_f + L_G + \frac{2L_G^2}{\mu_y}$-smooth. Whence,

$$\begin{aligned}\max_{x\in\mathbb{R}^{d_x}} \min_{y\in\mathbb{R}^{d_y}} \{h(y) - G(x, y) - f(x)\} - \min_{y\in\mathbb{R}^{d_y}} \{h(y) - G(\hat{x}, y) - f(\hat{x})\} = \Psi(x^*) - \Psi(\hat{x}) \\ \le \frac{L_f + L_G + \frac{2L_G^2}{\mu_y}}{2}\|\hat{x} - x^*\|^2 \le 2\frac{L_f + L_G + \frac{2L_G^2}{\mu_y}}{\mu_x}\varepsilon_x + 8\left(\frac{L_G}{\mu_x}\right)^2 \frac{L_f + L_G + \frac{2L_G^2}{\mu_y}}{\mu_y}\varepsilon_y,\end{aligned} \tag{216}$$

with probability $1 - \sigma_x - \sigma_y$.In the first inequality we used that x_* is the optimal point, and, hence, $\nabla\Psi(x_*) = 0$. □

Proof of Theorem 5

By construction, as an output of Loop 1 we obtain an (ε, σ)-solution to the problem (37) satisfy (10).

We prove the estimates for the numbers of oracle calls in two steps. The first step is to formally prove that in each loop the dependence of the number of oracle calls on the target accuracy ε and a confidence level σ is logarithmic. The second step is to multiply the estimates for the number of oracle calls between loops and choose the parameters H_1, H_2, H_3.

Step 1. Polynomial dependence. The goal of this technical step is to prove that

$$\begin{aligned}&\varepsilon_i(\varepsilon) = \mathrm{poly}\,(\varepsilon)\,, \sigma_i\,(\varepsilon, \sigma) = \mathrm{poly}\,(\varepsilon, \sigma)\,, \tilde{\sigma}^{(i)}\,(\varepsilon, \sigma)\\ &= \mathrm{poly}\,(\varepsilon, \sigma)\,, \sigma_0^{(i)}\,(\varepsilon, \sigma) = \mathrm{poly}\,(\varepsilon, \sigma)\,,\\ &\tilde{\varepsilon}_f^{(i)}\,(\varepsilon) = \mathrm{poly}\,(\varepsilon)\,, \delta^{(i)}\,(\varepsilon) = \mathrm{poly}\,(\varepsilon)\,, \varepsilon_2' = \mathrm{poly}\,(\varepsilon)\,, \sigma_2'\\ &= \mathrm{poly}\,(\varepsilon, \sigma)\,, \bar{\delta}(\varepsilon_2) = \mathrm{poly}\,(\varepsilon)\,, \bar{\sigma}_0(\sigma_2) = \mathrm{poly}\,(\varepsilon, \sigma)\end{aligned} \tag{217}$$

where $i = 1, 2, 3$. For $i = 1$, according the polynomial dependencies (31), (32), (34), (35) we obtain the polynomial dependencies

$$\begin{gathered}\varepsilon_1(\varepsilon) = \mathrm{poly}\,(\varepsilon)\,, \sigma_1\,(\varepsilon, \sigma) = \mathrm{poly}\,(\varepsilon, \sigma)\,, \tilde{\sigma}^{(1)}\,(\varepsilon, \sigma) = \mathrm{poly}\,(\varepsilon, \sigma)\,, \sigma_0^{(1)}\,(\varepsilon, \sigma) = \mathrm{poly}\,(\varepsilon, \sigma)\,,\\ \tilde{\varepsilon}_f^{(1)}\,(\varepsilon) = \mathrm{poly}\,(\varepsilon)\,, \delta^{(1)}\,(\varepsilon) = \mathrm{poly}\,(\varepsilon)\,.\end{gathered}$$

Now using that $\varepsilon_2' = \tilde{\varepsilon}_f^{(1)}$, $\sigma_2' = \tilde{\sigma}^{(1)}$ and (56), (57) we have that $\varepsilon_2(\varepsilon) = \operatorname{poly}(\varepsilon), \sigma_2(\varepsilon,\sigma) = = \operatorname{poly}(\varepsilon,\sigma)$. Further, by (55), $\bar{\delta}(\varepsilon) = \operatorname{poly}(\varepsilon), \bar{\sigma}_0(\varepsilon,\sigma) = \operatorname{poly}(\varepsilon,\sigma)$. Using the same argument as for $i = 1$, according the polynomial dependencies (31), (32), (34), (35) we obtain the polynomial dependencies

$$\varepsilon_2(\varepsilon) = \operatorname{poly}(\varepsilon), \sigma_2(\varepsilon,\sigma) = \operatorname{poly}(\varepsilon,\sigma), \tilde{\sigma}^{(2)}(\varepsilon,\sigma) = \operatorname{poly}(\varepsilon,\sigma), \tilde{\varepsilon}_f^{(2)}(\varepsilon) = \operatorname{poly}(\varepsilon),$$
$$\delta^{(2)}(\varepsilon) = \operatorname{poly}(\varepsilon), \sigma_0^{(2)}(\varepsilon,\sigma) = \operatorname{poly}(\varepsilon,\sigma).$$

Taking into account that $\varepsilon_3 = \tilde{\varepsilon}_f^{(2)}, \sigma_3 = \tilde{\sigma}^{(2)}$, the polynomial dependencies (31), (32), (34),(35) we obtain

$$\varepsilon_3(\varepsilon) = \operatorname{poly}(\varepsilon), \sigma_3(\varepsilon,\sigma) = \operatorname{poly}(\varepsilon,\sigma), \tilde{\sigma}^{(3)}(\varepsilon,\sigma) = \operatorname{poly}(\varepsilon,\sigma), \sigma_0^{(3)}(\varepsilon,\sigma) = \operatorname{poly}(\varepsilon,\sigma),$$
$$\tilde{\varepsilon}_f^{(3)}(\varepsilon) = \operatorname{poly}(\varepsilon), \delta^{(3)}(\varepsilon) = \operatorname{poly}(\varepsilon).$$

This finishes the proof of polynomial dependence. Thus, due to (217) in each loop when Assumptions 2, 3 are applied, the dependencies $\mathcal{K}_G^y, \mathcal{K}_h, \mathcal{K}_G^x, \mathcal{K}_f$ have only logarithmic dependence on the target accuracy ε and confidence level σ, i.e.

$$\mathcal{K}_G^y(\varepsilon,\sigma) = \widetilde{O}(1),\ \mathcal{K}_h(\varepsilon,\sigma) = \widetilde{O}(1),\ \mathcal{K}_G^x(\varepsilon,\sigma) = \widetilde{O}(1),\ \mathcal{K}_f(\varepsilon,\sigma) = \widetilde{O}(1),$$
$$O(\log\varepsilon_1^{-1}) = \widetilde{O}(1),\ O(\log\varepsilon_2^{-1}) = \widetilde{O}(1),\ O(\log\varepsilon_3^{-1}) = \widetilde{O}(1).$$

Step 2. Final estimates. We have already counted the number of oracles calls for each oracle in each loops, see the last paragraph of the description of each loop. We start with the number of basic oracle calls of f, which is called in each step of all the three loops. Thus, the total number is

$$\begin{gathered}
\#\text{ of calls in Loop1} + (\#\text{ of steps in Loop 1})\cdot(\#\text{ of calls in Loop 2})\\
+ (\#\text{ of steps in Loop 1})\cdot(\#\text{ of steps in Loop 2})\cdot(\#\text{ of calls in Loop 3})\\
= \widetilde{O}\left(1+\left(\frac{H_1}{\mu_y}\right)^{\frac12}\right)\mathcal{N}_f(\tau_f)\,\mathcal{K}_f(\varepsilon,\sigma) + \widetilde{O}\left(1+\left(\frac{H_1}{\mu_y}\right)^{\frac12}\right)\cdot\left(\widetilde{O}\left(1+\left(\frac{H_2}{\mu_x}\right)^{\frac12}\right)\tau_f\right)\\
+\widetilde{O}\left(1+\left(\frac{H_1}{\mu_y}\right)^{\frac12}\right)\cdot\left(\widetilde{O}\left(1+\left(\frac{H_2}{\mu_x}\right)^{\frac12}\right)\right)\cdot\left(\widetilde{O}\left(1+\left(\frac{H_3}{H_2}\right)^{\frac12}\right)\tau_f\right)\\
= \widetilde{O}\left(\left(1+\sqrt{\frac{H_1}{\mu_y}}\right)\left(\mathcal{N}_f(\tau_f) + \left(1+\sqrt{\frac{H_2}{\mu_x}}\right)\left(1+\sqrt{\frac{H_3}{H_2}}\right)\cdot\tau_f\right)\right),
\end{gathered}$$

where we used that $\mathcal{K}_f(\varepsilon,\sigma) = \widetilde{O}(1)$.

The basic oracle of h is called in each step of "Loop 1"and "Loop 2". Thus, the total number is

$$\begin{gathered}
\#\text{ of calls in Loop1} + (\#\text{ of steps in Loop 1})\cdot(\#\text{ of calls in Loop 2})\\
= \widetilde{O}\left(1+\left(\frac{H_1}{\mu_y}\right)^{\frac12}\right)\tau_h + \widetilde{O}\left(1+\left(\frac{H_1}{\mu_y}\right)^{\frac12}\right)\cdot\left(\widetilde{O}\left(1+\left(\frac{H_2}{\mu_x}\right)^{\frac12}\right)\mathcal{N}_h(\tau_h,H_1)\,\mathcal{K}_h(\varepsilon_2,\sigma_2)\right)\\
= \widetilde{O}\left(\left(1+\sqrt{\frac{H_1}{\mu_y}}\right)\left(\tau_h + \left(1+\sqrt{\frac{H_2}{\mu_x}}\right)\mathcal{N}_h(\tau_h,H_1)\right)\right),
\end{gathered}$$

where we used that $\mathcal{K}_h(\varepsilon,\sigma) = \widetilde{O}(1)$.

The basic oracle of $G(\cdot,y)$ is called in each step of "Loop 1"and "Loop 2". Thus, the total number is

$$\begin{aligned}&\#\text{ of calls in Loop1} + (\#\text{ of steps in Loop 1})\cdot(\#\text{ of calls in Loop 2})\\ &= \widetilde{O}\left(1+\left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)\mathcal{N}_G^x(\tau_G)\,\mathcal{K}_G^x(\varepsilon,\sigma) + \widetilde{O}\left(1+\left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)\cdot\left(\widetilde{O}\left(1+\left(\frac{H_2}{\mu_x}\right)^{\frac{1}{2}}\right)\tau_G\right)\\ &= \widetilde{O}\left(\left(1+\sqrt{\frac{H_1}{\mu_y}}\right)\left(\mathcal{N}_G^x(\tau_G) + \left(1+\sqrt{\frac{H_2}{\mu_x}}\right)\tau_G\right)\right),\end{aligned}$$

where we used that $\mathcal{K}_G^x(\varepsilon,\sigma) = \widetilde{O}(1)$.

Finally, the basic oracle of $G(x,\cdot)$ is called in each step of "Loop 1"and "Loop 2". Thus, the total number is

$$\begin{aligned}&\#\text{ of calls in Loop1} + (\#\text{ of steps in Loop 1})\cdot(\#\text{ of calls in Loop 2})\\ &= \widetilde{O}\left(1+\left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)\tau_G + \widetilde{O}\left(1+\left(\frac{H_1}{\mu_y}\right)^{\frac{1}{2}}\right)\cdot\left(\widetilde{O}\left(1+\left(\frac{H_2}{\mu_x}\right)^{\frac{1}{2}}\right)\mathcal{N}_G^y(\tau_G,H_1)\,\mathcal{K}_G^y(\varepsilon_2,\sigma_2)\right)\\ &= \widetilde{O}\left(\left(1+\sqrt{\frac{H_1}{\mu_y}}\right)\left(\tau_G + \left(1+\sqrt{\frac{H_2}{\mu_x}}\right)\mathcal{N}_G^y(\tau_G,H_1)\right)\right),\end{aligned}$$

where we used that $\mathcal{K}_G^y(\varepsilon_2,\sigma_2) = \widetilde{O}(1)$.

The final estimates are obtained by substituting the constants H_1, H_2, H_3 given by

$$H_1 = 2L_G, H_2 = 2\left(L_G + \frac{2L_G^2}{\mu_y + H_1}\right) \le 2\left(L_G + \frac{2L_G^2}{H_1}\right) = 4L_G, H_3 = 2L_f.$$

□

Proof of Theorem 6

Condition 4 with (73) guarantee that Condition 1 holds. Further, assumption $\mu_y \le L_G$ and the choice $H = 2L_G$ guarantee that $\mu_y + H \le 4L_G$. This inequality, assumption that $m_h(4L_G + \mu_y) \le L_h$ and the choice $H = 2L_G$ allow to apply Lemma 4 and conclude that Condition 2 holds with the number of oracle calls given by (75) and (76).

Assumptions $2L_G + \mu_x \le L_f$ and $\mu_x \le L_G$ by Corollary 2 guarantee that Assumption 3 holds with the number of oracle calls given by (85) and (86). Applying Theorem 5 and combining its complexity estimates, we obtain the final complexity bounds as follows.

Number of basic oracle calls of f:

$$\begin{aligned}&\widetilde{O}\left(\left(1+\sqrt{\frac{L_G}{\mu_y}}\right)\left(\sqrt{\frac{L_f}{\mu_x}} + \left(1+\sqrt{\frac{L_G}{\mu_x}}\right)\left(1+\sqrt{\frac{L_f}{L_G}}\right)\right)\right)\\ &\quad= \widetilde{O}\left(\left(\sqrt{\frac{L_G}{\mu_y}}\right)\left(\sqrt{\frac{L_f}{\mu_x}} + \left(\sqrt{\frac{L_G}{\mu_x}}\right)\left(\sqrt{\frac{L_f}{L_G}}\right)\right)\right) = \widetilde{O}\left(\left(\sqrt{\frac{L_fL_G}{\mu_x\mu_y}}\right)\right),\end{aligned}$$

where we used that, by the assumptions of this Theorem, $1 \le L_G/\mu_y$, $1 \le L_G/\mu_x$ and $1 \le L_f/L_G$.

Number of basic oracle calls of h:

$$\begin{aligned}
&\widetilde{O}\left(\left(1+\sqrt{\frac{L_G}{\mu_y}}\right)\left(m_h+\left(1+\sqrt{\frac{L_G}{\mu_x}}\right)\sqrt{\frac{m_hL_h}{2L_G+\mu_y}}\right)\right)\\
&\quad=\widetilde{O}\left(\left(\sqrt{\frac{L_G}{\mu_y}}\right)\left(m_h+\left(\sqrt{\frac{L_G}{\mu_x}}\right)\left(\sqrt{\frac{m_hL_h}{2L_G+\mu_y}}\right)\right)\right)\\
&\quad=\widetilde{O}\left(\max\left\{\underbrace{m_h\sqrt{\frac{L_G}{\mu_y}}}_{=\widetilde{O}\left(\sqrt{m_hL_h/\mu_y}\right)}, \sqrt{\frac{m_hL_GL_h}{\mu_x\mu_y}}\right\}\right)=\widetilde{O}\left(\sqrt{\frac{m_hL_GL_h}{\mu_x\mu_y}}\right),
\end{aligned}$$

where we used that, by the assumptions of this Theorem, $1 \le L_G/\mu_y$, $1 \le L_G/\mu_x$ and

$$m_h(4L_G+\mu_y) \le L_h \Rightarrow \sqrt{m_hL_G} \le \sqrt{L_h}$$

Number of basic oracle calls of $G(\cdot, y)$:

$$\widetilde{O}\left(\left(1+\sqrt{\frac{L_G}{\mu_y}}\right)\left(1+\sqrt{\frac{L_G}{\mu_x}}+\left(1+\sqrt{\frac{L_G}{\mu_x}}\right)\right)\right)=\widetilde{O}\left(\sqrt{\frac{L_G^2}{\mu_x\mu_y}}\right),$$

where we used that, by the assumptions of this Theorem, $1 \le L_G/\mu_y$ and $1 \le L_G/\mu_x$.

Number of basic oracle calls of $G(x, \cdot)$:

$$\begin{aligned}
&\widetilde{O}\left(\left(1+\sqrt{\frac{L_G}{\mu_y}}\right)\left(1+\left(1+\sqrt{\frac{L_G}{\mu_x}}\right)\left(1+\sqrt{\frac{L_G}{2L_G+\mu_y}}\right)\right)\right)\\
&\quad=\widetilde{O}\left(\left(1+\sqrt{\frac{L_G}{\mu_y}}\right)\left(1+\sqrt{\frac{L_G}{\mu_x}}\right)\right)=\widetilde{O}\left(\sqrt{\frac{L_G^2}{\mu_x\mu_y}}\right),
\end{aligned}$$

where we used that, by the assumptions of this Theorem, $1 \le L_G/\mu_y$ and $1 \le L_G/\mu_x$. □