A Additional Background

Definition A.1 (Convex conjugate)**.** Given a convex function $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, its convex conjugate ψ^* is defined by:

$$(\forall \mathbf{z} \in \mathbb{R}^d) : \quad \psi^*(\mathbf{z}) = \sup_{\mathbf{x} \in \mathbb{R}^d} \{\langle \mathbf{z}, \mathbf{x}\rangle - \psi(\mathbf{x})\}.$$

The following standard fact can be derived using Fenchel-Young inequality $\forall \mathbf{x}, \mathbf{z} \in \mathbb{R}^d : \psi(\mathbf{x}) + \psi^*(\mathbf{z}) \geq \langle \mathbf{z}, \mathbf{x}\rangle$, and it is a simple corollary of Danskin's theorem (see, e.g., Bertsekas (1971); Bertsekas et al. (2003)).

Fact A.2. *Let $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a closed convex proper function and let ψ^* be its convex conjugate. Then, $\forall \mathbf{g} \in \partial\psi^*(\mathbf{z})$,*

$$\mathbf{g} \in \operatorname*{argsup}_{\mathbf{x} \in \mathbb{R}^d} \{\langle \mathbf{z}, \mathbf{x}\rangle - \psi(\mathbf{x})\},$$

where $\partial\psi^*(\mathbf{z})$ is the subdifferential set (the set of all subgradients) of ψ^* at point $\mathbf{z}$. In particular, if ψ^* is differentiable, then $\operatorname{argsup}_{\mathbf{x} \in \mathbb{R}^d} \{\langle \mathbf{z}, \mathbf{x}\rangle - \psi(\mathbf{x})\}$ is a singleton set and $\nabla\psi^*(\mathbf{z})$ is its only element.

Proposition 2.3. *Given, $\mathbf{z}, \mathbf{u} \in \mathbb{R}^d$, $p \in (1, \infty)$ and $q \in \{p, 2\}$, let*

$$\mathbf{w} = \operatorname*{argmin}_{\mathbf{v} \in \mathbb{R}^d} \Big\{ \langle \mathbf{z}, \mathbf{v}\rangle + \frac{1}{q}\|\mathbf{u} - \mathbf{v}\|_p^q \Big\}.$$

Then, for $p^* = \frac{p}{p-1}$, $q^* = \frac{q}{q-1}$:

$$\mathbf{w} = \mathbf{u} - \nabla\Big(\frac{1}{q^*}\|\mathbf{z}\|_{p^*}^{q^*}\Big) \quad \text{and} \quad \frac{1}{q}\|\mathbf{w} - \mathbf{u}\|_p^q = \frac{1}{q}\|\mathbf{z}\|_{p^*}^{q^*}.$$

Proof. The statements in the proposition are simple corollaries of conjugacy of the functions $\psi(\mathbf{u}) = \frac{1}{q}\|\mathbf{u}\|_p^q$ and $\psi^*(\mathbf{z}) = \frac{1}{q^*}\|\mathbf{z}\|_{p^*}^{q^*}$. In particular, the first part follows from

$$\psi^*(\mathbf{z}) = \sup_{\mathbf{v} \in \mathbb{R}^d} \{\langle \mathbf{z}, \mathbf{v}\rangle - \psi(\mathbf{v})\},$$

by the definition of a convex conjugate and using that $\frac{1}{q}\|\mathbf{u}\|_p^q$ and $\frac{1}{q^*}\|\mathbf{z}\|_{p^*}^{q^*}$ are conjugates of each other, which are standard exercises in convex analysis for $q \in \{p, 2\}$ (see, e.g., (Borwein and Zhu, 2004, Exercise 4.4.2) and (Boyd et al., 2004, Example 3.27)).

The second part follows by $\nabla\psi^*(\mathbf{z}) = \arg\sup_{\mathbf{v} \in \mathbb{R}^d} \{\langle \mathbf{z}, \mathbf{v}\rangle - \psi(\mathbf{v})\}$, due to Fact A.2 ($\psi$ and ψ^* are both continuously differentiable for $p \in (1, \infty)$). Lastly, $\frac{1}{q}\|\mathbf{w} - \mathbf{u}\|_p^q = \frac{1}{q}\|\mathbf{z}\|_{p^*}^{q^*}$ can be verified by setting $\mathbf{w} = \mathbf{u} - \nabla\big(\frac{1}{q^*}\|\mathbf{z}\|_{p^*}^{q^*}\big)$. □

Proposition 2.4. *For any $L > 0$, $\kappa > 0$, $q \geq \kappa$, $t \geq 0$, and $\delta > 0$,*

$$\frac{L}{\kappa}t^\kappa \leq \frac{\Lambda}{q}t^q + \frac{\delta}{2},$$

where $\Lambda = \big(\frac{2(q-\kappa)}{\delta q \kappa}\big)^{\frac{q-\kappa}{\kappa}} L^{q/\kappa}$.

Proof. The proof is based on the Fenchel-Young inequality and the conjugacy of functions $\frac{|x|^r}{r}$ and $\frac{|y|^s}{s}$ for $r, s \geq 1$, $\frac{1}{r} + \frac{1}{s} = 1$, which implies $xy \leq \frac{x^r}{r} + \frac{y^s}{s}$, $\forall x, y \geq 0$. In particular, setting $r = q/\kappa$, $s = q/(q - \kappa)$, and $x = t^\kappa$, we have

$$\frac{L}{\kappa}t^\kappa \leq \frac{Lt^q}{qy} + \frac{L(q-\kappa)}{q\kappa} y^{\frac{\kappa}{q-\kappa}}.$$

It remains to set $\frac{\delta}{2} = \frac{L(q-\kappa)}{q\kappa} y^{\frac{\kappa}{q-\kappa}}$, which, solving for y, gives $y = \big(\frac{\delta q \kappa}{2L(q-\kappa)}\big)^{q-\kappa}$, and verify that, under this choice, $\Lambda = \frac{Lt^q}{qy}$. □

B Omitted Proofs from Section 3

Lemma 3.1. *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be an arbitrary L-Lipschitz operator that satisfies Assumption 1 for some $\mathbf{u}^* \in \mathcal{U}^*$. Given an arbitrary initial point $\mathbf{u}_0$, let the sequences of points $\{\mathbf{u}_i\}_{i\geq 1}$, $\{\bar{\mathbf{u}}_i\}_{i\geq 0}$ evolve according to (EG+) for some $\beta \in (0, 1]$ and positive step sizes $\{a_i\}_{i\geq 0}$. Then, for any $\gamma > 0$ and any $k \geq 0$, we have:*

$$
\begin{aligned}
h_k \leq\; & \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_k\|^2 - \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|^2 \\
& + \frac{a_k}{2}\big(\rho - a_k(1-\beta)\big)\|F(\bar{\mathbf{u}}_k)\|^2 \\
& + \frac{{a_k}^2}{2\beta^2}\big(a_k L\gamma - \beta\big)\|F(\mathbf{u}_k)\|^2 \\
& + \frac{1}{2}\Big(\frac{a_k L}{\gamma} - \beta\Big)\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|^2,
\end{aligned}
\tag{3.2}
$$

where h_k is defined as in Eq. (3.1).

Proof. Fix any $k \geq 0$ and write h_k equivalently as

$$
\begin{aligned}
h_k = a_k \langle F(\bar{\mathbf{u}}_k), \mathbf{u}_{k+1} - \mathbf{u}^*\rangle + a_k \langle F(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle \\
+ a_k \langle F(\bar{\mathbf{u}}_k) - F(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle + a_k \frac{\rho}{2}\|F(\bar{\mathbf{u}}_k)\|^2.
\end{aligned}
\tag{B.1}
$$

The proof proceeds by bounding above individual terms on the right-hand side of Eq. (B.1). For the first term, the first-order optimality in the definition of $\mathbf{u}_{k+1}$ gives:

$$
a_k F(\bar{\mathbf{u}}_k) + \mathbf{u}_{k+1} - \mathbf{u}_k = \mathbf{0}.
$$

Thus, we have

$$
\begin{aligned}
a_k \langle F(\bar{\mathbf{u}}_k), \mathbf{u}_{k+1} - \mathbf{u}^*\rangle &= -\langle \mathbf{u}_{k+1} - \mathbf{u}_k, \mathbf{u}_{k+1} - \mathbf{u}^*\rangle \\
&= \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_k\|^2 - \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|^2 - \frac{1}{2}\|\mathbf{u}_k - \mathbf{u}_{k+1}\|^2.
\end{aligned}
\tag{B.2}
$$

For the second term on the right-hand side of Eq. (B.1), the first-order optimality in the definition of $\bar{\mathbf{u}}_k$ implies:

$$
\frac{a_k}{\beta}\langle F(\mathbf{u}_k) + \bar{\mathbf{u}}_k - \mathbf{u}_k, \mathbf{u}_{k+1} - \bar{\mathbf{u}}_k\rangle = 0,
$$

which, similarly as for the first term, leads to:

$$
a_k \langle F(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle = \frac{\beta}{2}\|\mathbf{u}_k - \mathbf{u}_{k+1}\|^2 - \frac{\beta}{2}\|\mathbf{u}_k - \bar{\mathbf{u}}_k\|^2 - \frac{\beta}{2}\|\mathbf{u}_{k+1} - \bar{\mathbf{u}}_k\|^2.
\tag{B.3}
$$

For the third term on the right-hand side of Eq. (B.1), applying Cauchy-Schwarz inequality, L-Lipschitzness of F, and Young's inequality, respectively, we have:

$$
\begin{aligned}
a_k \langle F(\bar{\mathbf{u}}_k) - F(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle &\leq a_k\|F(\bar{\mathbf{u}}_k) - F(\mathbf{u}_k)\|\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\| \\
&\leq a_k L\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\| \\
&\leq \frac{a_k L\gamma}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|^2 + \frac{a_k L}{2\gamma}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|^2,
\end{aligned}
\tag{B.4}
$$

where the last inequality holds for any $\gamma > 0$.

Using that $\bar{\mathbf{u}}_k - \mathbf{u}_k = -\frac{a_k}{\beta}F(\mathbf{u}_k)$, $\mathbf{u}_{k+1} - \mathbf{u}_k = -a_k F(\bar{\mathbf{u}}_k)$ and combining Eqs. (B.2)-(B.4) with Eq. (B.1), we have:

$$
\begin{aligned}
h_k \leq\; & \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_k\|^2 - \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|^2 + \frac{a_k}{2}\big(\rho - a_k(1-\beta)\big)\|F(\bar{\mathbf{u}}_k)\|^2 \\
& + \frac{{a_k}^2}{2\beta^2}\big(a_k L\gamma - \beta\big)\|F(\mathbf{u}_k)\|^2 + \frac{1}{2}\Big(\frac{a_k L}{\gamma} - \beta\Big)\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|^2,
\end{aligned}
$$

as claimed. ∎

C Omitted Proofs from Section 4

We start by first proving the following lemma that holds for generic choices of algorithm parameters a_k and β. We will then use this lemma to deduce the convergence bounds for different choices of $p > 1$ and both the deterministic and the stochastic oracle access to F.

Lemma C.1. *Let $p > 1$ and let $F : \mathbb{R}^d \to \mathbb{R}^d$ be an arbitrary L-Lipschitz operator w.r.t. $\|\cdot\|_p$ that satisfies Assumption 1 for some $\mathbf{u}^* \in \mathcal{U}^*$. Given an arbitrary initial point $\mathbf{u}_0$, let the sequences of points $\{\mathbf{u}_i\}_{i\geq 1}$, $\{\bar{\mathbf{u}}_i\}_{i\geq 0}$ evolve according to (EG$_p$+) for some $\beta \in (0, 1]$ and positive step sizes $\{a_i\}_{i\geq 0}$. Then, for any $\gamma > 0$ and any $k \geq 0$:*

$$\begin{aligned} h_k \leq & -a_k \langle \bar{\boldsymbol{\eta}}_k, \bar{\mathbf{u}}_k - \mathbf{u}^* \rangle - a_k \langle \bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \rangle + \frac{a_k \rho}{2} \|F(\bar{\mathbf{u}}_k)\|_{p^*}^2 \\ & + \phi_p(\mathbf{u}^*, \mathbf{u}_k) - \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1}) + \frac{\beta - m_p}{q} \|\mathbf{u}_{k+1} - \mathbf{u}_k\|_p^q \\ & + \frac{a_k \Lambda_k \gamma - \beta}{q} \|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^q + \frac{a_k \Lambda_k/\gamma - \beta m_p}{q} \|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^q + a_k \delta_k, \end{aligned}$$

where h_k is defined as in Eq. (4.7), δ_k is any positive number, and $\Lambda_k = \left(\frac{q-2}{\delta_k q}\right)^{\frac{q-2}{2}} L^{q/2}$. When $q = 2$, the statement also holds with $\delta_k = 0$ and $\Lambda_k = L$.

Proof. We begin the proof by writing h_k equivalently as:

$$\begin{aligned} h_k &= a_k \left\langle \tilde{F}(\bar{\mathbf{u}}_k), \bar{\mathbf{u}}_k - \mathbf{u}^* \right\rangle - a_k \langle \bar{\boldsymbol{\eta}}_k, \bar{\mathbf{u}}_k - \mathbf{u}^* \rangle + \frac{a_k \rho}{2} \|F(\bar{\mathbf{u}}_k)\|_{p^*}^2 \\ &= a_k \left\langle \tilde{F}(\bar{\mathbf{u}}_k), \mathbf{u}_{k+1} - \mathbf{u}^* \right\rangle + a_k \left\langle \tilde{F}(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \right\rangle \\ &\quad + a_k \left\langle \tilde{F}(\bar{\mathbf{u}}_k) - \tilde{F}(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \right\rangle - a_k \langle \bar{\boldsymbol{\eta}}_k, \bar{\mathbf{u}}_k - \mathbf{u}^* \rangle + \frac{a_k \rho}{2} \|F(\bar{\mathbf{u}}_k)\|_{p^*}^2. \end{aligned} \tag{C.1}$$

The proof now proceeds by bounding individual terms on the right-hand side of the last equality.

Let $M_{k+1}(\mathbf{u}) = a_k \left\langle \nabla \tilde{F}(\bar{\mathbf{u}}_k), \mathbf{u} - \mathbf{u}_k \right\rangle + \phi_p(\mathbf{u}, \mathbf{u}_k)$, so that $\mathbf{u}_{k+1} = \mathrm{argmin}_{\mathbf{u}\in\mathbb{R}^d} M_{k+1}(\mathbf{u})$. By the definition of Bregman divergence of M_{k+1} :

$$M_{k+1}(\mathbf{u}^*) = M_{k+1}(\mathbf{u}_{k+1}) + \langle \nabla M_{k+1}(\mathbf{u}_{k+1}), \mathbf{u}^* - \mathbf{u}_{k+1} \rangle + D_{M_{k+1}}(\mathbf{u}^*, \mathbf{u}_{k+1}).$$

As $\mathbf{u}_{k+1} = \mathrm{argmin}_{\mathbf{u}\in\mathbb{R}^d} M_{k+1}(\mathbf{u})$, we have $\nabla M_{k+1}(\mathbf{u}_{k+1}) = \mathbf{0}$. Further, $D_{M_{k+1}}(\mathbf{u}^*, \mathbf{u}_{k+1}) = D_{\phi_p(\cdot, \mathbf{u}_k)}(\mathbf{u}^*, \mathbf{u}_{k+1})$. When $p \leq 2$, ϕ_p itself is a Bregman divergence, and we have $D_{M_{k+1}}(\mathbf{u}^*, \mathbf{u}_{k+1}) = \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1})$. When $p > 2$, $\phi_p(\mathbf{u}, \mathbf{u}_k) = \frac{1}{p}\|\mathbf{u} - \mathbf{u}_k\|_p^p$, and as ϕ_p is p-uniformly convex with constant 1, it follows that $D_{M_{k+1}}(\mathbf{u}^*, \mathbf{u}_{k+1}) \geq \frac{1}{p}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|_p^p = \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1})$. Thus:

$$M_{k+1}(\mathbf{u}^*) \geq M_{k+1}(\mathbf{u}_{k+1}) + \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1}).$$

Equivalently, applying the definition of $M_{k+1}(\cdot)$ to the last inequality:

$$\begin{aligned} a_k \left\langle \nabla \tilde{F}(\bar{\mathbf{u}}_k), \mathbf{u}_{k+1} - \mathbf{u}^* \right\rangle &\leq \phi_p(\mathbf{u}^*, \mathbf{u}_k) - \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1}) - \phi_p(\mathbf{u}_{k+1,\mathbf{u}_k}) \\ &\leq \phi_p(\mathbf{u}^*, \mathbf{u}_k) - \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1}) - \frac{m_p}{q} \|\mathbf{u}_{k+1} - \mathbf{u}_k\|_p^q, \end{aligned} \tag{C.2}$$

where the last inequality follows from Eq. (4.6).

Now let $\bar{M}_k(\mathbf{u}) = \frac{a_k}{\beta} \left\langle \tilde{F}(\mathbf{u}_k), \mathbf{u} - \mathbf{u}_k \right\rangle + \frac{1}{q}\|\mathbf{u} - \mathbf{u}_k\|_p^q$ so that $\bar{\mathbf{u}}_k = \mathrm{argmin}_{\mathbf{u}\in\mathbb{R}^d} \bar{M}_k(\mathbf{u})$. By similar arguments as above,

$$\begin{aligned} \bar{M}_k(\mathbf{u}_{k+1}) &= \bar{M}_k(\bar{\mathbf{u}}_k) + \left\langle \nabla \bar{M}_k(\bar{\mathbf{u}}_k), \mathbf{u}_{k+1} - \bar{\mathbf{u}}_k \right\rangle + D_{M_k}(\mathbf{u}_{k+1}, \bar{\mathbf{u}}_k) \\ &\geq \bar{M}_k(\bar{\mathbf{u}}_k) + \frac{m_p}{q} \|\mathbf{u}_{k+1} - \bar{\mathbf{u}}_k\|_p^q, \end{aligned}$$

where the inequality is by $\nabla \bar{M}_k(\bar{\mathbf{u}}_k) = \mathbf{0}$ and the fact that $\frac{1}{q}\|\cdot\|_p^q$ is q-uniformly convex w.r.t. $\|\cdot\|_p$ with constant m_p, by the choice of q from Eq. (4.3). Applying the definition of $\bar{M}_k(\mathbf{u})$ to the last inequality:

$$a_k \left\langle \tilde{F}(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \right\rangle \leq \frac{\beta}{q}\big(\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_p^q - \|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^q - m_p\|\mathbf{u}_{k+1} - \bar{\mathbf{u}}_k\|_p^q\big). \tag{C.3}$$

The remaining term to bound is $\left\langle \tilde{F}(\bar{\mathbf{u}}_k) - \tilde{F}(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \right\rangle$. Using the definitions of $\bar{\boldsymbol{\eta}}_k, \boldsymbol{\eta}_k$, we have:

$$\begin{aligned}
\left\langle \tilde{F}(\bar{\mathbf{u}}_k) - \tilde{F}(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \right\rangle &= \langle F(\bar{\mathbf{u}}_k) - F(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle - \langle \bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle \\
&\overset{(i)}{\leq} - \langle \bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle + \|F(\bar{\mathbf{u}}_k) - F(\mathbf{u}_k)\|_{p^*}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p \\
&\overset{(ii)}{\leq} - \langle \bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle + L\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p \\
&\overset{(iii)}{\leq} - \langle \bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle + \frac{L\gamma}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^2 + \frac{L}{2\gamma}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^2,
\end{aligned}$$

where (i) is by Hölder's inequality, (ii) is by L-Lipschitzness of F, and (iii) is by Young's inequality, which holds for any $\gamma > 0$. Now, let $\delta_k > 0$ and $\Lambda_k = \left(\frac{2(q-\kappa)}{\delta_k q\kappa}\right)^{\frac{q-\kappa}{\kappa}} L^{q/\kappa}$. Then, applying Proposition 2.4 to the last two terms in the last inequality:

$$\begin{aligned}
\left\langle \tilde{F}(\bar{\mathbf{u}}_k) - \tilde{F}(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \right\rangle \leq & - \langle \bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle \\
& + \frac{\Lambda_k\gamma}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^q + \frac{\Lambda_k}{q\gamma}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^q + \delta_k.
\end{aligned} \tag{C.4}$$

Observe that when $q = 2$, there is no need to apply Proposition 2.4, and the last inequality is satisfied with $\delta_k = 0$ and $\Lambda_k = L$.

Combining Eqs. (C.2)-(C.4) with Eq. (C.1), we have:

$$\begin{aligned}
h_k \leq & - a_k\langle \bar{\boldsymbol{\eta}}_k, \bar{\mathbf{u}}_k - \mathbf{u}^*\rangle - a_k\langle \bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle + \frac{a_k\rho}{2}\|F(\bar{\mathbf{u}}_k)\|_{p^*}^2 \\
& + \phi_p(\mathbf{u}^*, \mathbf{u}_k) - \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1}) + \frac{\beta - m_p}{q}\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_p^q \\
& + \frac{a_k\Lambda_k\gamma - \beta}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^q + \frac{a_k\Lambda_k/\gamma - \beta m_p}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^q + a_k\delta_k,
\end{aligned}$$

as claimed. □

We are now ready to state and prove the main convergence bounds. For simplicity, we first start with the case of exact oracle access to F. We then show that we can build on this result by separately bounding the error terms due to the variance of the stochastic estimates $\tilde{F}$.

Deterministic Oracle Access. The main result is summarized in the following theorem.

Theorem 4.1. *Let $p > 1$ and let $F : \mathbb{R}^d \to \mathbb{R}^d$ be an arbitrary L-Lipschitz operator w.r.t. $\|\cdot\|_p$ that satisfies Assumption 1 with $\rho = 0$ for some $\mathbf{u}^* \in \mathcal{U}^*$. Assume that we are given oracle access to the exact evaluations of F, i.e., $\bar{\boldsymbol{\eta}}_i = \boldsymbol{\eta}_i = \mathbf{0}$, $\forall i$. Given an arbitrary initial point $\mathbf{u}_0 \in \mathbb{R}^d$, let the sequences of points $\{\mathbf{u}_i\}_{i\geq 1}$, $\{\bar{\mathbf{u}}_i\}_{i\geq 0}$ evolve according to (EG_p+) for $\beta \in (0, 1]$ and step sizes $\{a_i\}_{i\geq 0}$ specified below. Then, we have:*

(i) Let $p \in (1, 2]$. If $\beta = m_p = p - 1$, $a_k = \frac{{m_p}^{3/2}}{2L}$, then all accumulation points of $\{\mathbf{u}_k\}_{k\geq 0}$ are in $\mathcal{U}^*$, and, furthermore $\forall k \geq 0$:

$$\begin{aligned}
\frac{1}{k+1}\sum_{i=0}^{k}\|F(\mathbf{u}_i)\|_{p^*}^2 &\leq \frac{16L^2\phi_p(\mathbf{u}^*, \mathbf{u}_0)}{{m_p}^2(k+1)} \\
&= O\Big(\frac{L^2\|\mathbf{u}^* - \mathbf{u}_0\|_p^2}{(p-1)^2(k+1)}\Big).
\end{aligned}$$

In particular, within $k = O\big(\frac{L^2\|\mathbf{u}^* - \mathbf{u}_0\|_p^2}{(p-1)^2\epsilon^2}\big)$ *iterations* EG$_p$+ *can output a point* $\mathbf{u}$ *with* $\|F(\mathbf{u})\|_{p^*} \leq \epsilon$.

(ii) Let $p \in (2, \infty)$*. If* $\beta = \frac{1}{2}$, $\delta_k = \delta > 0$, $\Lambda = \big(\frac{q-2}{\delta q}\big)^{\frac{q-2}{2}} L^{q/2}$*, and* $a_k = \frac{1}{2\Lambda} = a$*, then,* $\forall k \geq 0$*:*

$$\frac{1}{k+1}\sum_{i=0}^{k} \|F(\bar{\mathbf{u}}_i)\|_{p^*}^{p^*} \leq \frac{2\|\mathbf{u}^* - \mathbf{u}_0\|_p^p}{a^{p^*}(k+1)} + \frac{2p\delta}{a^{p^*-1}}.$$

In particular, for any $\epsilon > 0$*, there is a choice of* $\delta = \frac{\epsilon^2}{C_p L}$*, where* C_p *is a constant that only depends on* p*, such that* EG$_p$+ *can output a point* $\mathbf{u}$ *with* $\|F(\mathbf{u})\|_{p^*} \leq \epsilon$ *in at most*

$$k = O_p\bigg(\Big(\frac{L\|\mathbf{u}^* - \mathbf{u}_0\|_p}{\epsilon}\Big)^p\bigg)$$

iterations. Here, the O_p *notation hides constants that only depend on* p*.*

Proof. Observe that, as $\bar{\boldsymbol{\eta}}_i = \boldsymbol{\eta}_i = \mathbf{0}$, $\forall i \geq 0$ and $\rho = 0$, Lemma C.1 and the definition of h_k give:

$$\begin{aligned} 0 \leq h_k \leq \phi_p(\mathbf{u}^*, \mathbf{u}_k) - \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1}) + \frac{\beta - m_p}{q}\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_p^q \\ + \frac{a_k\Lambda_k\gamma - \beta}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^q + \frac{a_k\Lambda_k/\gamma - \beta m_p}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^q + a_k\delta_k, \end{aligned} \tag{C.5}$$

Proof of Part (i). In this case, we can set $\delta_k = 0$ (see Lemma C.1), $\Lambda_k = L$, and $q = 2$. Therefore, setting $\beta = m_p$, $a_k = \frac{{m_p}^{3/2}}{2L}$, and $\gamma = \frac{1}{\sqrt{m_p}}$ we get from Eq. (C.5) that

$$\phi_p(\mathbf{u}^*, \mathbf{u}_{k+1}) \leq \phi_p(\mathbf{u}^*, \mathbf{u}_k) - \frac{m_p}{4}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^2. \tag{C.6}$$

It follows that $\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^2$ converges to zero as $k \to \infty$. By the definition of $\bar{\mathbf{u}}_k$ and Proposition 2.3, $\frac{1}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^2 = \frac{{a_k}^2}{2\beta^2}\|F(\mathbf{u}_k)\|_{p^*}^2$, and, so, $\|F(\mathbf{u}_k)\|_{p^*}$ converges to zero as $k \to \infty$. Further, as $\phi_p(\mathbf{u}^*, \mathbf{u}_k) \leq \phi_p(\mathbf{u}^*, \mathbf{u}_0) < \infty$ and $\phi_p(\mathbf{u}^*, \mathbf{u}_k) \geq \frac{m_p}{2}\|\mathbf{u}^* - \mathbf{u}_k\|_p^2$, $m_p > 0$, it follows that $\|\mathbf{u}^* - \mathbf{u}_k\|_p$ is bounded, and, thus, $\{\mathbf{u}_k\}_{k\geq 0}$ is a bounded sequence. The proof that all accumulation points of $\{\mathbf{u}_k\}_{k\geq 0}$ are in $\mathcal{U}^*$ is standard and omitted (see the proof of Theorem 3.2 for a similar argument).

To bound $\frac{1}{k+1}\sum_{i=0}^{k}\|F(\mathbf{u}_i)\|_{p^*}^2$, we telescope the inequality from Eq. (C.6) to get:

$$m_p\sum_{i=0}^{k}\|\bar{\mathbf{u}}_i - \mathbf{u}_i\|_p^2 \leq 4(\phi_p(\mathbf{u}^*, \mathbf{u}_0) - \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1})) \leq 4\phi_p(\mathbf{u}^*, \mathbf{u}_0).$$

To complete the proof of this part, it remains to use that $\|\bar{\mathbf{u}}_i - \mathbf{u}_i\|_p^2 = \frac{{a_k}^2}{\beta^2}\|F(\mathbf{u}_i)\|_{p^*}^2$ (already argued above), the definitions of a_k and β, and $m_p = p - 1$. The bound on $\phi_p(\mathbf{u}^*, \mathbf{u}_0)$ follows from the definition of ϕ_p in this case. In particular, if we denote $\psi(\mathbf{u}) = \frac{1}{2}\|\mathbf{u} - \mathbf{u}_0\|_p^2$, then $\phi_p(\mathbf{u}^*, \mathbf{u}_0) = D_\psi(\mathbf{u}^*, \mathbf{u}_0)$. Using the definition of Bregman divergence and the fact that, for this choice of ψ, we have $\|\nabla\psi(\mathbf{u})\|_{p^*} = \|\mathbf{u} - \mathbf{u}_0\|_p$, $\forall \mathbf{u} \in \mathbb{R}^d$, (see the last part of Proposition 2.3) it follows that:

$$\phi_p(\mathbf{u}^*, \mathbf{u}_0) = \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_0\|_p^2 - \frac{1}{2}\|\mathbf{u}_0 - \mathbf{u}_0\|_p^2 - \Big\langle \nabla_{\mathbf{u}}\Big(\frac{1}{2}\|\mathbf{u} - \mathbf{u}_0\|_p^2\Big)\Big|_{\mathbf{u}=\mathbf{u}_0}, \mathbf{u}^* - \mathbf{u}_0\Big\rangle \qquad = \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_0\|_p^2.$$

Proof of Part (ii). In this case, $q = p$, $\phi_p(\mathbf{u}, \mathbf{v}) = \frac{1}{p}\|\mathbf{u} - \mathbf{v}\|_p^p$, and $m_p = 1$. Using Proposition 2.3, $\|\mathbf{u}_k - \bar{\mathbf{u}}_k\|_p^p = \frac{{a_k}^{p^*}}{\beta^{p^*}}\|F(\mathbf{u}_k)\|_{p^*}^{p^*}$ and $\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_p^p = {a_k}^{p^*}\|F(\bar{\mathbf{u}}_k)\|_{p^*}^{p^*}$. Combining with Eq. (C.5), we have:

$$\begin{aligned} 0 \leq \frac{1}{p}\|\mathbf{u}^* - \mathbf{u}_k\|_p^p - \frac{1}{p}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|_p^p + \frac{(\beta - 1){a_k}^{p^*}}{p}\|F(\bar{\mathbf{u}}_k)\|_{p^*}^{p^*} \\ + \frac{(a_k\Lambda_k\gamma - \beta){a_k}^{p^*}}{p\beta^{p^*}}\|F(\mathbf{u}_k)\|_{p^*}^{p^*} + \frac{a_k\Lambda_k/\gamma - \beta}{p}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^p + a_k\delta_k. \end{aligned} \tag{C.7}$$

Now let $\gamma = 1$, $\beta = \frac{1}{2}$, $\delta_k = \delta > 0$, and $a_k = \frac{1}{2\Lambda_k} = \frac{1}{2\Lambda} = a$. Then $a_k\Lambda_k\gamma - \beta = a_k\Lambda_k/\gamma - \beta = 0$ and Eq. (C.7) simplifies to:

$$\frac{a^{p^*}}{2p}\|F(\bar{\mathbf{u}}_k)\|_{p^*}^{p^*} \leq \frac{1}{p}\|\mathbf{u}^* - \mathbf{u}_k\|_p^p - \frac{1}{p}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|_p^p + a\delta.$$

Telescoping the last inequality and then dividing it by $\frac{{a_k}^{p^*}(k+1)}{2p}$, we have:

$$\frac{1}{k+1}\sum_{i=0}^{k}\|F(\bar{\mathbf{u}}_i)\|_{p^*}^{p^*} \leq \frac{2\|\mathbf{u}^* - \mathbf{u}_0\|_p^p}{a^{p^*}(k+1)} + \frac{2p\delta}{a^{p^*-1}}. \tag{C.8}$$

Now, for $\textsc{eg}_p+$ to be able to output a point $\mathbf{u}$ with $\|F(\mathbf{u})\|_{p^*} \leq \epsilon$, it suffices to show that for some choice of δ and k we can make the right-hand side of Eq. (C.8) at most ϵ^{p^*}. This is true because then $\textsc{eg}_p+$ can output the point $\bar{\mathbf{u}}_i = \mathrm{argmin}_{0\leq i\leq k}\|F(\bar{\mathbf{u}}_i)\|_{p^*}$. For stochastic setups, the guarantee would be in expectation, and $\textsc{eg}_p+$ could output a point $\bar{\mathbf{u}}_i$ with i chosen uniformly at random from $\{0, \dots, k\}$, similarly as discussed in the proof of Theorem 3.2.

Observe first that, as $\Lambda = \big(\frac{p-2}{p\delta}\big)^{\frac{p-2}{2}}L^{p/2}$ and $p^* = \frac{p}{p-1}$, we have that:

$$\begin{aligned}\frac{\delta}{a^{p^*-1}} &= \delta(2\Lambda)^{p^*-1} = \delta 2^{\frac{1}{p-1}}\Lambda^{\frac{1}{p-1}}\\ &= 2^{\frac{1}{p-1}}\delta^{\frac{p}{2(p-1)}}\Big(\frac{p-2}{p}\Big)^{\frac{p-2}{2(p-1)}}L^{\frac{p}{2(p-1)}}.\end{aligned}$$

Setting $\frac{2p\delta}{a^{p^*-1}} \leq \frac{\epsilon^{p^*}}{2}$, recalling that $p^* = \frac{p}{p-1}$, and rearranging, we have:

$$\delta^{\frac{p^*}{2}} \leq \frac{\epsilon^{p^*}}{2^{\frac{2p-1}{p}}p}\Big(\frac{p}{p-2}\Big)^{\frac{p-2}{2p}p^*}L^{-p^*/2}.$$

Equivalently:

$$\delta \leq \frac{\epsilon^2}{L\cdot 2^{\frac{2(2p-1)}{p}}p^{\frac{2(p-1)}{p}}(\frac{p-2}{p})^{\frac{p-2}{p}}}.$$

It can be verified numerically that $(\frac{p-2}{p})^{\frac{p-2}{p}}$ is a constant between $\frac{1}{e}$ and 1, while it is clear that $2^{\frac{2(2p-1)}{p}}p^{\frac{2(p-1)}{p}} = O(p^2)$ is a constant that only depends on p. Hence, it suffices to set $\delta = \frac{\epsilon^2}{C_pL}$, where $C_p = 2^{\frac{2(2p-1)}{p}}p^{\frac{2(p-1)}{p}}$.

It remains to bound the number of iterations k so that $\frac{2\|\mathbf{u}^*-\mathbf{u}_0\|_p^p}{a^{p^*}(k+1)} \leq \frac{\epsilon^{p^*}}{2}$. Equivalently, we need $k+1 \geq \frac{4\|\mathbf{u}^*-\mathbf{u}_0\|_p^p}{a^{p^*}\epsilon^{p^*}}$. Plugging $\delta = \frac{\epsilon^2}{C_pL}$ into the definition of Λ, using that $p^* = \frac{p}{p-1}$, and simplifying, we have:

$$\begin{aligned}a^{p^*} &= (2\Lambda)^{p^*} = 2^{\frac{p}{p-1}}\Big(\frac{p-2}{p\delta}\Big)^{\frac{p-2}{2}\cdot\frac{p}{p-1}}L^{\frac{p}{2}\cdot\frac{p}{p-1}}\\ &= O_p\bigg(\Big(\frac{1}{\epsilon}\Big)^{\frac{p(p-2)}{p-1}}L^p\bigg).\end{aligned}$$

Thus,

$$k = O_p\bigg(\Big(\frac{1}{\epsilon}\Big)^{\frac{p(p-2)}{p-1}+\frac{p}{p-1}}L^p\|\mathbf{u}^* - \mathbf{u}_0\|_p^p\bigg) = O_p\bigg(\Big(\frac{L\|\mathbf{u}^* - \mathbf{u}_0\|_p}{\epsilon}\Big)^p\bigg),$$

as claimed. □

Stochastic Oracle Access. To obtain results for stochastic oracle access to F, we only need to bound the terms $\mathcal{E}^s \stackrel{\text{def}}{=} -a_k\langle\bar{\boldsymbol{\eta}}_k, \bar{\mathbf{u}}_k - \mathbf{u}^*\rangle - a_k\langle\bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle$ from Lemma C.1 corresponding to the stochastic error in expectation, while for the rest of the analysis we can appeal to the results for the deterministic oracle access to F. In the case of $p = 2$, there is one additional term that appears in h_k due to replacing $F(\bar{\mathbf{u}}_k)$ with $\tilde{F}(\bar{\mathbf{u}}_k)$. This term is simply equal to:

$$\frac{a_k\rho}{2}\mathbb{E}[\|\tilde{F}(\bar{\mathbf{u}}_k)\|_2^2 - \|F(\bar{\mathbf{u}}_k)\|_2^2|\bar{\mathcal{F}}_k] = \frac{a_k\rho}{2}\mathbb{E}[\|F(\bar{\mathbf{u}}_k) + \bar{\boldsymbol{\eta}}_k\|_2^2 - \|F(\bar{\mathbf{u}}_k)\|_2^2|\bar{\mathcal{F}}_k] = \frac{a_k\rho}{2}\bar{\sigma}_k^2. \tag{C.9}$$

We start by bounding the stochastic error $\mathcal{E}^s$ in expectation.

Lemma 4.3. *Let $\mathcal{E}^s = -a_k\langle\bar{\boldsymbol{\eta}}_k, \bar{\mathbf{u}}_k - \mathbf{u}^*\rangle - a_k\langle\bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle$, where $\bar{\boldsymbol{\eta}}_k$ and $\boldsymbol{\eta}_k$ are defined as in Eq. (4.2) and all the assumptions of Theorem 4.4 below apply. Then, for q defined by Eq. (4.3) and any $\tau > 0$:*

$$\mathbb{E}[\mathcal{E}^s] \leq \frac{2^{q^*/2}{a_k}^{q^*}({\sigma_k}^2 + \bar{\sigma}_k^2)^{q^*/2}}{q^*\tau^{q^*}} + \mathbb{E}\Big[\frac{\tau^q}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^q\Big],$$

where the expectation is w.r.t. all the randomness in the algorithm.

Proof. Let us start by bounding $-a_k\langle\bar{\boldsymbol{\eta}}_k, \bar{\mathbf{u}}_k - \mathbf{u}^*\rangle$ first. Conditioning on $\bar{\mathcal{F}}_k$, $\bar{\boldsymbol{\eta}}_k$ is independent of $\bar{\mathbf{u}}_k$ and $\mathbf{u}^*$, and, thus:

$$\mathbb{E}[-a_k\langle\bar{\boldsymbol{\eta}}_k, \bar{\mathbf{u}}_k - \mathbf{u}^*\rangle] = \mathbb{E}\big[\mathbb{E}[-a_k\langle\bar{\boldsymbol{\eta}}_k, \bar{\mathbf{u}}_k - \mathbf{u}^*\rangle\,|\bar{\mathcal{F}}_k]\big] = 0.$$

The second term, $-a_k\langle\bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle$, can be bounded using Hölder's inequality and Young's inequality as follows:

$$\begin{aligned}\mathbb{E}\big[-a_k\langle\bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle\big] &\leq \mathbb{E}\big[a_k\|\bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k\|_{p^*}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p\big]\\ &\leq \mathbb{E}\Big[\frac{{a_k}^{q^*}\|\bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k\|_{p^*}^{q^*}}{q^*\tau^{q^*}}\Big] + \mathbb{E}\Big[\frac{\tau^q}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^q\Big].\end{aligned}$$

It remains to bound $\mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k\|_{p^*}^{q^*}\big]$. Using triangle inequality,

$$\begin{aligned}\mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k\|_{p^*}^{q^*}\big] &\leq \mathbb{E}\Big[\big(\|\bar{\boldsymbol{\eta}}_k\|_{p^*} + \|\boldsymbol{\eta}_k\|_{p^*}\big)^{q^*}\Big]\\ &= \mathbb{E}\Big[\big(\big(\|\bar{\boldsymbol{\eta}}_k\|_{p^*} + \|\boldsymbol{\eta}_k\|_{p^*}\big)^2\big)^{q^*/2}\Big]\\ &\leq \Big(\mathbb{E}\big[\big(\|\bar{\boldsymbol{\eta}}_k\|_{p^*} + \|\boldsymbol{\eta}_k\|_{p^*}\big)^2\big]\Big)^{q^*/2},\end{aligned}$$

where the last line is by Jensen's inequality, as $q^* \in (1, 2]$, and so $(\cdot)^{q^*/2}$ is concave. Using Young's inequality and linearity of expectation:

$$\begin{aligned}\mathbb{E}\big[\big(\|\bar{\boldsymbol{\eta}}_k\|_{p^*} + \|\boldsymbol{\eta}_k\|_{p^*}\big)^2\big] &\leq 2\Big(\mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_k\|_{p^*}^2\big] + \mathbb{E}\big[\|\boldsymbol{\eta}_k\|_{p^*}^2\big]\Big)\\ &\leq 2({\sigma_k}^2 + \bar{\sigma}_k^2).\end{aligned}$$

Putting everything together:

$$\mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k\|_{p^*}^{q^*}\big] \leq 2^{q^*/2}({\sigma_k}^2 + \bar{\sigma}_k^2)^{q^*/2}$$

and

$$\begin{aligned}\mathbb{E}[\mathcal{E}^s] &= \mathbb{E}\big[-a_k\langle\bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle\big]\\ &\leq \frac{2^{q^*/2}{a_k}^{q^*}({\sigma_k}^2 + \bar{\sigma}_k^2)^{q^*/2}}{q^*\tau^{q^*}} + \mathbb{E}\Big[\frac{\tau^q}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^q\Big],\end{aligned}$$

as claimed. ∎

We are now ready to bound the total oracle complexity of EG$_p$+ (and its special case EG+), as follows.

Theorem 4.4. *Let $p > 1$ and let $F : \mathbb{R}^d \to \mathbb{R}^d$ be an arbitrary L-Lipschitz operator w.r.t. $\|\cdot\|_p$ that satisfies Assumption 1 for some $\mathbf{u}^* \in \mathcal{U}^*$. Given an arbitrary initial point $\mathbf{u}_0 \in \mathbb{R}^d$, let the sequences of points $\{\mathbf{u}_i\}_{i\geq 1}$, $\{\bar{\mathbf{u}}_i\}_{i\geq 0}$ evolve according to (EG_p+) for some $\beta \in (0,1]$ and positive step sizes $\{a_i\}_{i\geq 0}$. Let the variance of a single query to the stochastic oracle $\tilde{F}$ be bounded by some $\sigma^2 < \infty$.*

(i) Let $p = 2$ and $\rho \in [0, \bar{\rho})$, where $\bar{\rho} = \frac{1}{4\sqrt{2}L}$. If $\beta = \frac{1}{2}$ and $a_k = \frac{1}{2\sqrt{2}L}$, then EG_p+ can output a point $\mathbf{u}$ with $\mathbb{E}[\|\tilde{F}(\mathbf{u})\|_2] \leq \epsilon$ with at most

$$O\Big(\frac{L\|\mathbf{u}^* - \mathbf{u}_0\|_2^2}{\epsilon^2(\bar{\rho}-\rho)}\Big(1 + \frac{\sigma^2}{L\epsilon^2(\bar{\rho}-\rho)}\Big)\Big)$$

oracle queries to $\tilde{F}$.

(ii) Let $p \in (1, 2]$ and $\rho = 0$. If $a_k = \frac{{m_p}^{3/2}}{2L}$ and $\beta = m_p$, then EG_p+ can output a point $\mathbf{u}$ with $\mathbb{E}[\|\tilde{F}(\mathbf{u})\|_{p^*}] \leq \epsilon$ with at most

$$O\Big(\frac{L^2\|\mathbf{u}^* - \mathbf{u}_0\|_p^2}{{m_p}^2\epsilon^2}\Big(1 + \frac{\sigma^2}{m_p\epsilon^2}\Big)\Big)$$

oracle queries to $\tilde{F}$, where $m_p = p - 1$.

(iii) Let $p > 2$ and $\rho = 0$. If $\beta = \frac{1}{2}$ and $a_k = a = \frac{1}{4\Lambda}$, then EG_p+ can output a point $\mathbf{u}$ with $\mathbb{E}[\|\tilde{F}(\mathbf{u})\|_{p^*}] \leq \epsilon$ with at most

$$O_p\Big(\Big(\frac{L\|\mathbf{u}^* - \mathbf{u}_0\|_p}{\epsilon}\Big)^p\Big(1 + \Big(\frac{\sigma}{\epsilon}\Big)^{p^*}\Big)\Big)$$

oracle queries to $\tilde{F}$, where $p^* = \frac{p}{p-1}$.

Proof. Combining Lemmas C.1 and 4.3, we have, $\forall k \geq 0$:

$$\begin{aligned} 0 \leq \mathbb{E}[h_k] \leq\; & \frac{2^{q^*/2}{a_k}^{q^*}({\sigma_k}^2 + \bar{\sigma}_k^2)^{q^*/2}}{q^*\tau^{q^*}} + \mathbb{E}\Big[\frac{\tau^q}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^q\Big] + \frac{a_k\rho\bar{\sigma_k}^2}{2} + \mathbb{E}\Big[\frac{a_k\rho}{2}\|\tilde{F}(\bar{\mathbf{u}}_k)\|_{p^*}^2\Big] \\ & + \mathbb{E}\Big[\phi_p(\mathbf{u}^*, \mathbf{u}_k) - \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1}) + \frac{\beta - m_p}{q}\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_p^q\Big] \\ & + \mathbb{E}\Big[\frac{a_k\Lambda_k\gamma - \beta}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^q + \frac{a_k\Lambda_k/\gamma - \beta m_p}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^q + a_k\delta_k\Big], \end{aligned} \tag{C.10}$$

Proof of Part (i). In this case, $q = 2$, $m_p = 1$, $\delta = 0$, $\Lambda_k = L$, and $\phi_p(\mathbf{u}^*, \mathbf{u}) = \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}\|_2^2$, and, further, $\mathbf{u}_{k+1} - \mathbf{u}_k = -a_k F(\bar{\mathbf{u}}_k)$, so Eq. (C.10) simplifies to

$$\begin{aligned} 0 \leq \mathbb{E}[h_k] \leq\; & \frac{2{a_k}^2(\bar{\sigma_k}^2 + \bar{\sigma}_k^2)}{2\tau^2} + \frac{a_k\rho{\sigma_k}^2}{2} \\ & + \mathbb{E}\Big[\frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_k\|_2^2 - \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|_2^2 + \frac{{a_k}^2(\beta - 1) + a_k\rho}{2}\|\tilde{F}(\bar{\mathbf{u}}_k)\|_2^2\Big] \\ & + \mathbb{E}\Big[\frac{a_kL\gamma - \beta}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_2^2 + \frac{a_kL/\gamma - \beta + \tau^2}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_2^2\Big], \end{aligned}$$

Taking $\beta = \frac{1}{2}$, $\tau^2 = \frac{1}{4}$, $\gamma = \sqrt{2}$, and $a_k = \frac{1}{2\sqrt{2}L}$, and recalling that $\bar{\rho} = \frac{1}{4\sqrt{2}L}$, we have:

$$a_k(\bar{\rho} - \rho)\mathbb{E}\big[\|\tilde{F}(\bar{\mathbf{u}}_k)\|_2^2\big] \leq \mathbb{E}\big[\|\mathbf{u}^* - \mathbf{u}_k\|_2^2 - \|\mathbf{u}^* - \mathbf{u}_{k+1}\|_2^2\big] + 4{a_k}^2({\sigma_k}^2 + \bar{\sigma}_k)^2 + \frac{a_k\rho\bar{\sigma_k}^2}{2}.$$

Telescoping the last inequality and dividing both sides by $a_k(\bar{\rho} - \rho)(k + 1)$, we get:

$$\frac{1}{k+1}\sum_{i=0}^{k}\mathbb{E}\big[\|\tilde{F}(\bar{\mathbf{u}}_i)\|_2^2\big] \leq \frac{2\sqrt{2}L\|\mathbf{u}^* - \mathbf{u}_0\|_2^2}{(k+1)(\bar{\rho} - \rho)} + \frac{\sqrt{2}\sum_{i=0}^{k}({\sigma_i}^2 + \bar{\sigma}_i^2)}{L(\bar{\rho} - \rho)(k+1)} + \frac{\rho\sum_{i=0}^{k}\bar{\sigma}_i^2}{2(k+1)(\bar{\rho} - \rho)}.$$

In particular, if variance of a single sample of $\tilde{F}$ evaluated at an arbitrary point is σ^2 and we take n samples of $\tilde{F}$ in each iteration, then:

$$\frac{1}{k+1}\sum_{i=0}^{k}\mathbb{E}\big[\|\tilde{F}(\bar{\mathbf{u}}_i)\|_2^2\big] \le \frac{2\sqrt{2}L\|\mathbf{u}^* - \mathbf{u}_0\|_2^2}{(k+1)(\bar{\rho}-\rho)} + \frac{\sigma^2(4\sqrt{2}/L+\rho)}{2n(\bar{\rho}-\rho)}.$$

To finish the proof of this part, we require that both terms on the right-hand side of the last inequality are bounded by $\frac{\epsilon^2}{2}$. For the first term, this leads to:

$$k = \left\lceil \frac{4\sqrt{2}L\|\mathbf{u}^* - \mathbf{u}_0\|_2^2}{\epsilon^2(\bar{\rho}-\rho)} - 1 \right\rceil = O\Big(\frac{L\|\mathbf{u}^* - \mathbf{u}_0\|_2^2}{\epsilon^2(\bar{\rho}-\rho)}\Big).$$

For the second term, the bound is:

$$n = \left\lceil \frac{2\sigma^2(4\sqrt{2}/L+\rho)}{\epsilon^2(\bar{\rho}-\rho)} \right\rceil = O\Big(\frac{\sigma^2}{L\epsilon^2(\bar{\rho}-\rho)}\Big).$$

Thus, the total number of required oracle queries to $\tilde{F}$ is bounded by:

$$k(1+n) = O\Big(\frac{L\|\mathbf{u}^* - \mathbf{u}_0\|_2^2}{\epsilon^2(\bar{\rho}-\rho)}\Big(1 + \frac{\sigma^2}{L\epsilon^2(\bar{\rho}-\rho)}\Big)\Big).$$

As discussed before, $\bar{\mathbf{u}}_i$ with i chosen uniformly at random from $\{0,\dots,k\}$ will satisfy $\|F(\bar{\mathbf{u}}_i)\|_2 \le \epsilon$ in expectation.

Proof of Part (ii). In this case, $q=2$, $m_p = p-1$, $\delta = 0$, $\Lambda_k = L$, and $\rho = 0$. Thus, Eq. (C.10) simplifies to:

$$\begin{aligned} 0 \le \mathbb{E}[h_k] \le\ & \frac{2{a_k}^2({\sigma_k}^2 + \bar{\sigma}_k^2)}{2\tau^2} \\ & + \mathbb{E}\Big[\phi_p(\mathbf{u}^*, \mathbf{u}_k) - \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1}) + \frac{\beta - m_p}{2}\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_p^2\Big] \\ & + \mathbb{E}\Big[\frac{a_k L\gamma - \beta}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^2 + \frac{a_k L/\gamma - \beta m_p + \tau^2}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^2\Big]. \end{aligned}$$

In this case, the same choices for a_k and β as in the deterministic case suffice. In particular, let $a_k = \frac{{m_p}^{3/2}}{2L}$, $\beta = m_p$, $\gamma = \frac{1}{\sqrt{m_p}}$, and $\tau^2 = \frac{{m_p}^2}{2}$. Then, using that, from Proposition 2.3, $\frac{1}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^2 = \frac{{a_k}^2}{2\beta^2}\|\tilde{F}(\mathbf{u}_k)\|_{p^*}^2$, we have

$$\frac{{a_k}^2 m_p}{4\beta^2}\mathbb{E}\big[\|\tilde{F}(\mathbf{u}_k)\|_{p^*}^2\big] \le \mathbb{E}\Big[\phi_p(\mathbf{u}^*, \mathbf{u}_k) - \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1})\Big] + \frac{{a_k}^2({\sigma_k}^2 + \bar{\sigma}_k^2)}{\tau}.$$

Telescoping the last inequality and dividing both sides by $(k+1)\frac{{a_k}^2 m_p}{4\beta^2}$, we have:

$$\frac{1}{k+1}\sum_{i=0}^{k}\mathbb{E}\big[\|\tilde{F}(\mathbf{u}_i)\|_{p^*}^2\big] \le \frac{16L^2\phi_p(\mathbf{u}^*, \mathbf{u}_0)}{(k+1){m_p}^2} + \frac{8\sum_{i=0}^{k}({\sigma_i}^2 + \bar{\sigma_i}^2)}{(k+1)m_p}. \tag{C.11}$$

Now let ${\sigma_i}^2 = \bar{\sigma}_i^2 = \sigma^2/n$, where σ^2 is the variance of a single sample of $\tilde{F}$ and n is the number of samples taken per iteration. Then, similarly as in Part (i), to bound the total number of samples, it suffices to bound each term on the right-hand side of Eq. (C.11) by $\frac{\epsilon^2}{2}$. The first term was already bounded in Theorem 4.1, and it leads to:

$$k = O\Big(\frac{L^2\|\mathbf{u}^* - \mathbf{u}_0\|_p^2}{{m_p}^2\epsilon^2}\Big).$$

For the second term, it suffices that:

$$n = O\Big(\frac{\sigma^2}{m_p\epsilon^2}\Big),$$

and the bound on the total number of samples follows.

Proof of Part (iii). In this case, $q = p$, $m_p = 1$, $\rho = 0$, $\phi_p(\mathbf{u}^*, \mathbf{u}) = \frac{1}{p}\|\mathbf{u}^* - \mathbf{u}\|_p^p$, and we take $\delta_k = \delta > 0$, $\Lambda_k = \Lambda = \big(\frac{p-2}{p\delta}\big)^{\frac{p-2}{2}} L^{\frac{p}{2}}$. Eq. (C.10) now simplifies to:

$$
\begin{aligned}
0 \leq \mathbb{E}[h_k] \leq\ & \frac{2^{p^*/2} {a_k}^{p^*} ({\sigma_k}^2 + \bar{\sigma}_k^2)^{p^*/2}}{p^* \tau^{p^*}} \\
&+ \mathbb{E}\Big[\frac{1}{p}\|\mathbf{u}^* - \mathbf{u}_k\|_p^p - \frac{1}{p}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|_p^p + \frac{\beta - 1}{p}\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_p^p\Big] \\
&+ \mathbb{E}\Big[\frac{a_k \Lambda \gamma - \beta}{p}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^p + \frac{a_k\Lambda/\gamma + \tau^p - \beta}{p}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^p + a_k\delta\Big].
\end{aligned} \tag{C.12}
$$

Recall that, by Proposition 2.3, $\frac{1}{p}\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_p^p = \frac{a^{p^*}}{p}\|\tilde{F}(\bar{\mathbf{u}}_k)\|_{p^*}^{p^*}$. Let $\beta = \frac{1}{2}$, $a_k = a = \frac{1}{4\Lambda}$, $\tau^p = \frac{1}{4}$, and $\gamma = 1$. Then $\beta - 1 = -\frac{1}{2}$, $a_k\Lambda\gamma - \beta = -\frac{1}{4} < 0$, and $a_k\Lambda/\gamma + \tau^p - \beta = 0$, and Eq. (C.12) leads to:

$$
\frac{a^{p^*}}{2p}\mathbb{E}\big[\|\tilde{F}(\bar{\mathbf{u}}_k)\|_{p^*}^{p^*}\big] \leq \mathbb{E}\Big[\frac{1}{p}\|\mathbf{u}^* - \mathbf{u}_k\|_p^p - \frac{1}{p}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|_p^p\Big] + \frac{2^{\frac{4+p}{2(p-1)}} a^{p^*}({\sigma_k}^2 + \bar{\sigma}_k^2)^{p^*/2}}{p^*} + a\delta.
$$

Telescoping the last inequality and then dividing both sides by $\frac{a^{p^*}}{2p}(k+1)$, we have:

$$
\frac{1}{k+1}\sum_{i=0}^{k}\mathbb{E}\big[\|\tilde{F}(\bar{\mathbf{u}}_i)\|_{p^*}^{p^*}\big] \leq \frac{2\|\mathbf{u}^* - \mathbf{u}_0\|_p^p}{a^{p^*}(k+1)} + \frac{2^{\frac{3p+2}{2(p-1)}} p \sum_{i=0}^{k}({\sigma_i}^2 + \bar{\sigma}_i^2)^{p^*/2}}{p^*(k+1)} + \frac{2p\delta}{a^{p^*} - 1}.
$$

Now let σ^2 be the variance of a single sample of $\tilde{F}$ and suppose that in each iteration we take n samples to estimate $F(\bar{\mathbf{u}}_i)$ and $F(\mathbf{u}_i)$. Then ${\sigma_i}^2 = {\bar{\sigma}_i}^2 = \frac{\sigma^2}{n}$, and the last equation simplifies to

$$
\frac{1}{k+1}\sum_{i=0}^{k}\mathbb{E}\big[\|\tilde{F}(\bar{\mathbf{u}}_i)\|_{p^*}^{p^*}\big] \leq \frac{2\|\mathbf{u}^* - \mathbf{u}_0\|_p^p}{a^{p^*}(k+1)} + \frac{2^{\frac{p+2}{p-1}} p\sigma^{p^*}}{p^* n} + \frac{2p\delta}{a^{p^*} - 1}.
$$

To complete the proof, similarly as before, it suffices to show that we can choose k and n so that $\frac{2p\|\mathbf{u}^* - \mathbf{u}_0\|_p^p}{a^{p^*}(k+1)} + \frac{2p\delta}{a^{p^*} - 1} \leq \frac{\epsilon^{p^*}}{2}$ and $\frac{2^{\frac{p+2}{p-1}} p\sigma^{p^*}}{p^* n} \leq \frac{\epsilon^{p^*}}{2}$. For the former, following the same argument as in the proof of Theorem 4.1, Part (ii), it suffices to choose $\delta = O_p(\frac{\epsilon^2}{L})$, which leads to:

$$
k = O_p\Big(\Big(\frac{L\|\mathbf{u}^* - \mathbf{u}_0\|_p}{\epsilon}\Big)^p\Big).
$$

For the latter, it suffices to choose:

$$
n = \frac{2^{\frac{p+2}{p-1}+1} p\sigma^{p^*}}{p^* \epsilon^{p^*}} = O\Big(\frac{p\sigma^{p^*}}{\epsilon^{p^*}}\Big).
$$

The total number of queries to the stochastic oracle is then bounded by $k(1+n)$. □