Appendix

A Continuity

Proof of Theorem 4.4 We will prove the theorem by induction. Let $f \in C^1(\mathbb{R}^d)$ with a set of global minima X^*. For $h \in \mathcal{F}_{X^*}$, let $g = f + h \in f + \mathcal{F}_{X^*}$. For $n = 0$, $x_0(\cdot, \mathcal{A}_\theta, x_0) \equiv x_0$ is clearly continuous (in the sense of $\|\cdot\|_*$) for any fixed initial point $x_0 \in \mathbb{R}^d$; assume that the continuity property is verified up to some $n \in \mathbb{N}$: i.e. $\forall\, \epsilon > 0\ \forall\, i = 0, \dots, n$, $\exists\, \eta = \eta(\epsilon, i, \mathcal{K}) > 0$ such that for $g \in f + \mathcal{F}_{X^*}$, if $\|f - g\|_* < \eta$, then $\forall\, x_0 \in \mathcal{K}$, $\|x_i(f, \mathcal{A}_\theta, x_0) - x_i(g, \mathcal{A}_\theta, x_0)\|_2 < \epsilon$.

Let $\epsilon > 0$ and

$$x_{n+1}(f, \mathcal{A}_\theta, x_0) = \mathcal{A}_\theta\Big(\{x_i\}_{i=0\dots n}, \{f(x_i)\}_{i=0\dots n}, \{\nabla f(x_i)\}_{i=0\dots n}\Big);$$

$\mathcal{A}_\theta$ being a continuous FOA implies that given $\epsilon > 0$, there exists $\delta > 0$ such that if $\forall\, i = 0, \dots, n$

$$\|x_i(f, \mathcal{A}_\theta, x_0) - x_i(g, \mathcal{A}_\theta, x_0)\|_2 < \delta \tag{7}$$
$$\|f(x_i(f, \mathcal{A}_\theta, x_0)) - g(x_i(g, \mathcal{A}_\theta, x_0))\|_2 < \delta \tag{8}$$
$$\|\nabla f(x_i(f, \mathcal{A}_\theta, x_0)) - \nabla g(x_i(g, \mathcal{A}_\theta, x_0))\|_2 < \delta \tag{9}$$

for $f \in C^1(\mathbb{R}^d)$, $g \in f + \mathcal{F}_{X^*}$, then the claim follows

$$\|x_{n+1}(f, \mathcal{A}_\theta, x_0) - x_{n+1}(g, \mathcal{A}_\theta, x_0)\|_2 < \epsilon.$$

The idea now is to quantify how "close" f and g need to be (in $\|\cdot\|_*$-norm) in order to ensure that the above inequalities are satisfied.

For equation (7): by recurrence hypothesis ($n \in \mathbb{N}$ is finite), given $\delta > 0$ there exists $\eta_1 = \eta_1(\delta)$ (simply consider $= \min_{i=1,\dots,n}\{\eta(\delta, i, \mathcal{K})\} > 0$) such that $\forall\, g \in f + \mathcal{F}_{X^*}$, $\|f - g\|_* < \eta_1$, then $\forall\, x_0 \in \mathcal{K}$, $\|x_i(f, \mathcal{A}_\theta, x_0) - x_i(g, \mathcal{A}_\theta, x_0)\|_2 < \delta$, $\forall\, i = 1, \dots, n$.

For equation (8):

$$\begin{aligned} &\|\nabla f(x_i(f, \mathcal{A}_\theta, x_0)) - \nabla g(x_i(g, \mathcal{A}_\theta, x_0))\|_2 \\ \leq\ &\|\nabla f(x_i(f, \mathcal{A}_\theta, x_0)) - \nabla f(x_i(g, \mathcal{A}_\theta, x_0))\|_2 + \|\nabla f(x_i(g, \mathcal{A}_\theta, x_0)) - \nabla g(x_i(g, \mathcal{A}_\theta, x_0))\|_2 \end{aligned} \tag{10}$$

The first term can be easily estimated: since $f \in C^1(\mathbb{R}^d)$, given $\delta > 0$ there exists $\rho = \rho(\delta) > 0$ such that $\forall\, x, y \in \mathbb{R}^d$ with $\|x - y\|_2 < \rho$, then $\|\nabla f(x) - \nabla f(y)\|_2 < \frac{\delta}{2}$ and $\|f(x) - f(y)\|_2 < \frac{\delta}{2}$. In particular, for such a $\rho > 0$, $\exists\, \eta_2 = \eta_2(\rho, \delta) > 0$ such that for $\|f - g\|_* < \min\{\eta_1, \eta_2\}$, then $\|x_i(f, \mathcal{A}_\theta, x_0) - x_i(g, \mathcal{A}_\theta, x_0)\|_2 < \rho$ $\forall\, i = 0, \dots, n$. Therefore,

$$\|\nabla f(x_i(f, \mathcal{A}_\theta, x_0)) - \nabla f(x_i(g, \mathcal{A}_\theta, x_0))\|_2 < \frac{\delta}{2} \qquad \forall\, i = 0, \dots, n. \tag{11}$$

Regarding the second term, we first introduce the quantity

$$R_{f,n} := \max_{i=0,\dots,n} \left\{ \sup_{x_0 \in \mathcal{K}} d\left(x_i(f, \mathcal{A}_\theta, x_0), X^*\right) \right\}$$

$x_i(f, \mathcal{A}_\theta, \cdot)$ is a finite composition of continuous functions and is therefore continuous (in x_0). This ensures that the image of $\mathcal{K}$ is a compact and thus that $R_{f,n}$ is indeed finite.

$$\begin{aligned}&\|\nabla f(x_i(g, \mathcal{A}_\theta, x_0)) - \nabla g(x_i(g, \mathcal{A}_\theta, x_0))\|_2 \\ &= \frac{\|\nabla f(x_i(g, \mathcal{A}_\theta, x_0)) - \nabla g(x_i(g, \mathcal{A}_\theta, x_0))\|_2}{d\left(x_i(g, \mathcal{A}_\theta, x_0), X^*\right)} d\left(x_i(g, \mathcal{A}_\theta, x_0), X^*\right) \leq \|f - g\|_* R_{g,n} \end{aligned} \tag{12}$$

We want to claim that if $\|f - g\|_*$ is small enough (for $g \in f + \mathcal{F}_{X^*}$), then $R_{g,n} < R_{f,n} + \delta$: indeed, if $g \in f + \mathcal{F}_{X^*}$ is such that $\|f - g\|_* < \min\{\eta_1, \eta_2\}$, by recurrence hypothesis we have $\forall\, i = 0, \dots, n$

$$\begin{aligned} d\left(x_i(g, \mathcal{A}_\theta, x_0), X^*\right) &= \inf_{x^* \in X^*} \|x_i(g, \mathcal{A}_\theta, x_0) - x^*\|_2 \\ &\leq \inf_{x^* \in X^*} \{\|x_i(g, \mathcal{A}_\theta, x_0) - x_i(f, \mathcal{A}_\theta, x_0)\|_2 + \|x_i(f, \mathcal{A}_\theta, x_0) - x^*\|_2\} \\ &= \|x_i(g, \mathcal{A}_\theta, x_0) - x_i(f, \mathcal{A}_\theta, x_0)\|_2 + \inf_{x^* \in X^*} \|x_i(f, \mathcal{A}_\theta, x_0) - x^*\|_2 \\ &= \|x_i(g, \mathcal{A}_\theta, x_0) - x_i(f, \mathcal{A}_\theta, x_0)\|_2 + d\left(x_i(f, \mathcal{A}_\theta, x_0), X^*\right) \\ &< \delta + R_{f,n}. \end{aligned} \tag{13}$$

Then,

$$\|\nabla f(x_i(g, \mathcal{A}_\theta, x_0)) - \nabla g(x_i(g, \mathcal{A}_\theta, x_0))\|_2 \leq \|f - g\|_*(R_{f,n} + \delta) < \frac{\delta}{2} \tag{14}$$

as long as $\|f - g\|_* < \min\{\eta_1, \eta_2, \frac{\delta}{2(R_{f,n}+\delta)}\}$.

In conclusion, $\forall\, i = 0, \dots, n$

$$\begin{aligned}&\|\nabla f(x_i(f, \mathcal{A}_\theta, x_0)) - \nabla g(x_i(g, \mathcal{A}_\theta, x_0))\|_2 \\ &\leq \|\nabla f(x_i(f, \mathcal{A}_\theta, x_0)) - \nabla f(x_i(g, \mathcal{A}_\theta, x_0))\|_2 + \|\nabla f(x_i(g, \mathcal{A}_\theta, x_0)) - \nabla g(x_i(g, \mathcal{A}_\theta, x_0))\|_2 \\ &\leq \frac{\delta}{2} + \frac{\delta}{2} = \delta \end{aligned} \tag{15}$$

For equation (9): similarly, we have

$$\begin{aligned}&\|f(x_i(f, \mathcal{A}_\theta, x_0)) - g(x_i(g, \mathcal{A}_\theta, x_0))\|_2 \\ &\leq \|f(x_i(f, \mathcal{A}_\theta, x_0)) - f(x_i(g, \mathcal{A}_\theta, x_0))\|_2 + \|f(x_i(g, \mathcal{A}_\theta, x_0)) - g(x_i(g, \mathcal{A}_\theta, x_0))\|_2 \end{aligned} \tag{16}$$

The first term is bounded by $\delta/2$ thanks the same argument as in (11). The second term is bounded in the following way: call $\bar{x} = x_i(g, \mathcal{A}_\theta, x_0)$ and let $\bar{x}_p^* \in X^*$ the projection of $\bar{x}$ on X^*. Note that $\forall\, t \in [0, 1]$, we have $d\left(\bar{x}_p^* + t(\bar{x} - \bar{x}_p^*), X^*\right) \leq t \left\|\bar{x} - \bar{x}_p^*\right\|_2$ since $\bar{x}_p^* \in X^*$. It follows that

$$\begin{aligned} |f(\bar{x}) - g(\bar{x})| &= |f(\bar{x}) - g(\bar{x}) - (f^* - g^*)| \\ &= \left| \int_0^1 \langle \nabla(f - g)(\bar{x}_p^* + t(\bar{x} - \bar{x}_p^*)), \bar{x}_p^* - \bar{x} \rangle \, \mathrm{d}t \right| \\ &\leq \int_0^1 \left\|\nabla(f - g)(\bar{x}_p^* + t(\bar{x} - \bar{x}_p^*))\right\|_2 \left\|\bar{x}_p^* - \bar{x}\right\|_2 \, \mathrm{d}t \\ &= \int_0^1 \frac{\left\|\nabla(f - g)(\bar{x}_p^* + t(\bar{x} - \bar{x}_p^*))\right\|_2}{d\left(\bar{x}_p^* + t(\bar{x} - \bar{x}_p^*), X^*\right)} d\left(\bar{x}_p^* + t(\bar{x} - \bar{x}_p^*), X^*\right) \left\|\bar{x}_p^* - \bar{x}\right\|_2 \, \mathrm{d}t \\ &\leq \int_0^1 \frac{\left\|\nabla(f - g)(\bar{x}_p^* + t(\bar{x} - \bar{x}_p^*))\right\|_2}{d\left(\bar{x}_p^* + t(\bar{x} - \bar{x}_p^*), X^*\right)} t \left\|\bar{x}_p^* - \bar{x}\right\|_2^2 \, \mathrm{d}t \\ &\leq \|f - g\|_* d\left(\bar{x}, X^*\right)^2 \int_0^1 t \, \mathrm{d}t \\ &\leq \|f - g\|_* \frac{(R_{f,n} + \delta)^2}{2} \\ &< \frac{\delta}{2} \end{aligned} \tag{17}$$

as long as $\|f-g\|_* < \min\left\{\eta_1, \eta_2, \frac{\delta}{2(R_{f,n}+\delta)}, \frac{\delta}{(R_{f,n}+\delta)^2}\right\}$.

In conclusion, $\forall\, i = 0, \dots, n$

$$\begin{aligned} &\|f(x_i(f, \mathcal{A}_\theta, x_0)) - g(x_i(g, \mathcal{A}_\theta, x_0))\|_2 \\ \leq\ &\|f(x_i(f, \mathcal{A}_\theta, x_0)) - f(x_i(g, \mathcal{A}_\theta, x_0))\|_2 + \|f(x_i(g, \mathcal{A}_\theta, x_0)) - g(x_i(g, \mathcal{A}_\theta, x_0))\|_2 \\ <\ &\frac{\delta}{2} + \frac{\delta}{2} = \delta. \end{aligned} \tag{18}$$

□

Proof of Corollary 4.5 Let $N_{\mathcal{K}} = \sup_{x_0 \in \mathcal{K}} N_{x_0}$ (note that $N_{\mathcal{K}} < +\infty$, since $\mathcal{K}$ is compact). We will note $g = f + h$ for $h \in \mathcal{F}_{X^*}$.

For $x_0 \in \mathcal{K}$, we have

$$x_{N_{x_0}-1}(f, \mathcal{A}_\theta, x_0) \notin \mathcal{B}(X^*, \varepsilon) \quad \text{and} \quad x_{N_{x_0}}(f, \mathcal{A}_\theta, x_0) \in \mathcal{B}(X^*, \varepsilon)$$

and thanks to Theorem 4.4, there exists $\eta > 0$ such that for any $g \in f + \mathcal{F}_{X^*}$, if $||f-g||_* \leq \eta$ then $\forall\, i \leq N_{\mathcal{K}}$, $\forall\, x_0 \in \mathcal{K}$, $\|x_i(f, \mathcal{A}_\theta, x_0) - x_i(g, \mathcal{A}_\theta, x_0)\|_2 \leq \delta$. Therefore,

$$x_{N_{x_0}-1}(g, \mathcal{A}_\theta, x_0) \notin \mathcal{B}(X^*, \varepsilon - \delta) \quad \text{and} \quad x_{N_{x_0}}(g, \mathcal{A}_\theta, x_0) \in \mathcal{B}(X^*, \varepsilon + \delta).$$

□

Proof of Proposition 4.6 f is a piecewise quadratic with second derivative $f''(x) = (2 + \frac{2}{\varepsilon})$ for $x \in [1, 1+\varepsilon^2]$ and $f''(x) = 2$ elsewhere. Therefore, the optimal μ of strong convexity is 2 and the optimal L of smoothness is $2 + \frac{2}{\varepsilon}$: $f \in \mathrm{SC}^-(2) \cap \mathrm{SC}^+(2 + \frac{2}{\varepsilon})$.

Consider the gradient descent update rule with step size $\alpha = \frac{1}{2}$:

$$x - \frac{1}{2} f'(x) = \begin{cases} 0 & x \leq 1 \\ \frac{x-1}{\varepsilon} & 1 \leq x \leq 1+\varepsilon^2 \\ -\varepsilon & x \geq 1+\varepsilon^2 \end{cases}$$

It is easy to see that $|x - \frac{1}{2} f'(x)| \leq \varepsilon |x|$, which proves the linear convergence rate of f_ε with tuning $\alpha = \frac{1}{2}$; in fact, for $\varepsilon \leq 1$, the algorithm can converge to $x^* = 0$ in at most two steps.

Let us now assume we use the standard tuning based on strong convexity and smoothness

$$\alpha = \frac{2}{\mu_\varepsilon + L_\varepsilon} = \frac{\varepsilon}{2\varepsilon + 1}.$$

We then have

$$x - \alpha f'(x) = \begin{cases} \frac{x}{2\varepsilon+1} & x \leq 1 \\ \frac{2-x}{2\varepsilon+1} & 1 \leq x \leq 1+\varepsilon^2 \\ \frac{x-2\varepsilon^2}{2\varepsilon+1} & x \geq 1+\varepsilon^2 \end{cases}$$

which leads to

$$\forall\, x \in \mathbb{R}, \quad |x - \alpha f'(x)| \geq \frac{1-\varepsilon^2}{(2\varepsilon+1)(1+\varepsilon^2)} |x|.$$

□

Proof of Theorem 4.10 Let f a $\bar{L}$-smooth and $\bar{\mu}$-strongly convex function with a set of minima $X^* \subseteq \mathbb{R}^d$. Note that strong convexity implies $X^* = \{x^*\}$.

Let $\varepsilon > 0$. We define the function $\omega_\varepsilon \in C^1(\mathbb{R})$ by $\omega_\varepsilon(0) = 0$ and its derivative:

$$\omega'_\varepsilon(t) = \begin{cases} 0 & t \leq 1-\varepsilon^2 \\ \frac{1-t-\varepsilon^2}{\varepsilon} & 1-\varepsilon^2 \leq t \leq 1 \\ \frac{t-\varepsilon^2-1}{\varepsilon} & 1 \leq t \leq 1+\varepsilon^2 \\ 0 & 1+\varepsilon^2 \leq t \end{cases}$$

It is easy to see that $|\omega'_\varepsilon(t)| \leq \varepsilon$, $\forall\, t \in \mathbb{R}$.

Let $z \in \mathbb{R}^d \setminus \{x^*\}$ and define

$$\phi(x) := \frac{\langle x - x^*, z - x^* \rangle}{\|z - x^*\|_2}, \qquad x \in \mathbb{R}^d \tag{19}$$

$$f_\varepsilon(x) := f(x) + \omega_\varepsilon \circ \phi(x) \tag{20}$$

Note that $\omega_\varepsilon \circ \phi(x^*) = 0$ and

$$\nabla(f - f_\varepsilon)(x) = \nabla\left(\omega_\varepsilon \circ \phi\right)(x) = \frac{z - x^*}{\|z - x^*\|_2} \omega'_\varepsilon \circ \phi(x); \tag{21}$$

for $x \in \mathbb{R}^d$,

$$\text{if } \phi(x) \leq 1-\varepsilon^2, \text{then } ||\nabla(f - f_\varepsilon)(x)||_2 = 0 \tag{22}$$

$$\text{if } \phi(x) \geq 1-\varepsilon^2, \text{then } ||\nabla(f - f_\varepsilon)(x)||_2 \leq \varepsilon \leq \frac{\varepsilon}{1-\varepsilon^2} ||x - x^*||_2 \tag{23}$$

since $\phi(x) \geq 1 - \varepsilon^2$ implies $\|x - x^*\|_2 \geq 1 - \varepsilon^2$. Therefore, $f - f_\varepsilon \in \mathcal{F}_{X^*}$ and $\|f - f_\varepsilon\|_* \leq \frac{\varepsilon}{1-\varepsilon^2} \to 0$ when $\varepsilon \to 0$.

Let $L, \mu > 0$. We now want to prove that for ε sufficiently small, f_ϵ is not L-smooth and not μ-strong convex. Consider

$$x = x^* + (1-\varepsilon^2)\frac{z - x^*}{\|z - x^*\|_2}$$

$$y = x^* + \frac{z - x^*}{\|z - x^*\|_2}$$

so that we have $\phi(x) = 1 - \varepsilon^2$, $\phi(y) = 1$, and $y - x = \varepsilon^2 \frac{z - x^*}{\|z - x^*\|_2}$. Since f is L_f-smooth, $\forall\, \varepsilon > 0$ we have

$$\begin{aligned} & f_\varepsilon(y) - f_\varepsilon(x) - \langle \nabla f_\varepsilon(x), y - x \rangle \\ = {} & f(y) - f(x) - \langle \nabla f(x), y - x \rangle + \omega_\varepsilon \circ \phi(y) - \omega_\varepsilon \circ \phi(x) - \langle \nabla \omega_\varepsilon \circ \phi(x), y - x \rangle \\ \leq {} & \frac{L_f}{2} ||x - y||_2^2 + \omega_\varepsilon(1) - \omega_\varepsilon(1-\varepsilon^2) - \varepsilon^2 \omega'_\varepsilon(1-\varepsilon^2) \\ = {} & \frac{L_f}{2} ||x - y||_2^2 - \frac{\varepsilon^3}{2} = \left(\frac{L_f}{2} - \frac{1}{2\varepsilon}\right) \|x - y\|_2^2; \end{aligned} \tag{24}$$

therefore, if we pick ε such that $\frac{1}{\varepsilon} > L_f - \mu$, then f_ε is not μ-strong convex.

Similarly, consider

$$x = x^* + \frac{z - x^*}{\|z - x^*\|_2}$$

$$y = x^* + (1+\varepsilon^2)\frac{z - x^*}{\|z - x^*\|_2}$$

So that we have $\phi(x) = 1$, $\phi(y) = 1 + \varepsilon^2$, and $y - x = \varepsilon^2 \frac{z - x^*}{\|z - x^*\|_2}$. Since f is μ_f-strong convex, $\forall\, \varepsilon > 0$ we have

$$\begin{aligned} & f_\varepsilon(y) - f_\varepsilon(x) - \langle \nabla f_\varepsilon(x), y - x \rangle \\ = {} & f(y) - f(x) - \langle \nabla f(x), y - x \rangle + \omega_\varepsilon \circ \phi(y) - \omega_\varepsilon \circ \phi(x) - \langle \nabla \omega_\varepsilon \circ \phi(x), y - x \rangle \\ \geq {} & \frac{\mu_f}{2} \|x - y\|_2^2 + \omega_\varepsilon(1+\varepsilon^2) - \omega_\varepsilon(1) - \varepsilon^2 \omega'_\varepsilon(1) \\ = {} & \frac{\mu_f}{2} \|x - y\|_2^2 + \frac{\varepsilon^3}{2} = \left(\frac{\mu_f}{2} + \frac{1}{2\varepsilon}\right) \|x - y\|_2^2; \end{aligned} \tag{25}$$

therefore, if we pick ε such that $\frac{1}{\varepsilon} > L - \mu_f$, then f_ε is not L-smooth.
Finally, for any $\varepsilon \leq \min\{\frac{1}{\max\{1, L_f - \mu\}}, \frac{1}{\max\{1, L - \mu_f\}}\}$, f_ε is not L-smooth and not μ-strong convex, which concludes the proof. □

B Proof of Theorem 5.5

Note that all lower conditions listed in the theorem are continuous without any additional constraint; on the other hand, the upper conditions require the assumption of the objective function f to belong to $\mathrm{QG}^-(\mu)$ (for some $\mu > 0$) in order to be continuous.

We stress that this extra condition is a mild adding, since tuning of a FOA usually requires f to satisfy both an upper and a lower condition (and $\mathrm{QG}^-(\mu)$ is the weakest among the conditions we proposed). On the other hand, this is necessary to guarantee that the set of minimizer for the original f and the perturbed $f + h$, $h \in \mathcal{F}_{X^*}$, are the same.

Continuity of ${}^*\mathrm{SC}^-$ and ${}^*\mathrm{SC}^+$: Given $f \in {}^*\mathrm{SC}^+(L)$: $f^* \leq f(x) + \langle \nabla f(x), x_p^* - x \rangle + \frac{L}{2}\|x - x_p^*\|_2^2$, $\forall\, x \in \mathbb{R}^d$ (with $x_p^* \in X^*$ the corresponding projection point onto X^*). Consider $g = f + h$, $h \in \mathcal{F}_{X^*}$ with $\|f - g\|_* = \|h\|_* = \sup \frac{\|\nabla f(x) - \nabla g(x)\|_2}{d(x, X^*)} = \sup \frac{\|\nabla h(x)\|_2}{d(x, X^*)} \leq \frac{\epsilon}{3}$, then

$$g^* = f^* \leq f(x) + \langle \nabla f(x), x_p^* - x \rangle + \frac{L}{2}\|x - x_p^*\|_2^2, \tag{26}$$

where g^* is the value of g at each point of X^*. Note that $\forall\, x \in \mathbb{R}^d \setminus X^*$

$$\begin{aligned}
\langle \nabla f(x), x_p^* - x \rangle &= \langle \nabla f(x) - \nabla g(x), x_p^* - x \rangle + \langle \nabla g(x), x_p^* - x \rangle \\
&\leq \|\nabla f(x) - \nabla g(x)\|_2 \left\|x - x_p^*\right\|_2 + \langle \nabla g(x), x_p^* - x \rangle \\
&\leq \|f - g\|_* \left\|x - x_p^*\right\|_2^2 + \langle \nabla g(x), x_p^* - x \rangle \\
&\leq \frac{\epsilon}{3} \left\|x - x_p^*\right\|_2^2 + \langle \nabla g(x), x_p^* - x \rangle
\end{aligned} \tag{27}$$

and

$$\begin{aligned}
0 = h^* &= \omega(x) + \int_0^1 \langle \nabla h(x + t(x_p^* - x)), x_p^* - x \rangle \, \mathrm{d}t \\
&\leq h(x) + \int_0^1 \left\|\nabla h(x + t(x_p^* - x))\right\|_2 \|x - x_p^*\|_2 \, \mathrm{d}t \\
&= h(x) + \int_0^1 \frac{\left\|\nabla h(x + t(x_p^* - x))\right\|_2}{d(x + t(x_p^* - x), X^*)} \, d(x + t(x_p^* - x), X^*) \|x - x_p^*\|_2 \, \mathrm{d}t \\
&\leq h(x) + \|h\|_* \|x - x_p^*\|_2^2 \int_0^1 1 - t \, \mathrm{d}t = h(x) + \frac{1}{2} \|f - g\|_* \|x - x_p^*\|_2^2 \\
&\leq h(x) + \frac{\epsilon}{6} \|x - x^*\|_2^2
\end{aligned} \tag{28}$$

where we used $d(x + t(x_p^* - x), X^*) = (1 - t)\|x - x_p^*\|_2$, $\forall\, t \in [0, 1]$ (indeed any point lying on the line segment $x + t(x_p^* - x)$ has projection onto X^* equal to x_p^*).

Therefore, $\forall\, x \in \mathbb{R}^d \setminus X^*$

$$\begin{aligned}
g^* &\leq f(x) + \left[\frac{\epsilon}{3}\|x - x_p^*\|_2^2 + \langle \nabla g(x), x_p^* - x \rangle\right] + \frac{L}{2}\|x - x_p^*\|_2^2 \\
&\leq f(x) + h(x) + \frac{\epsilon}{6}\|x - x_p^*\|_2^2 + \frac{\epsilon}{3}\|x - x_p^*\|_2^2 + \langle \nabla g(x), x_p^* - x \rangle + \frac{L}{2}\|x - x^*\|_2^2 \\
&= g(x) + \langle \nabla g(x), x_p^* - x \rangle + \frac{L + \epsilon}{2}\|x - x_p^*\|_2^2
\end{aligned} \tag{29}$$

(for $x = x^* \in X^*$ the inequality is trivial), i.e. $g \in {}^*\mathrm{SC}(L + \epsilon)$.

Similarly, given $f \in {}^*\mathrm{SC}^-(\mu)$: $f^* \geq f(x) + \langle \nabla f(x), x_p^* - x\rangle + \frac{\mu}{2}\|x - x_p^*\|_2^2$, $\forall\, x \in \mathbb{R}^d$. Consider $g = f + h$, $h \in \mathcal{F}_{X^*}$ with $\|f - g\|_* = \|h\|_* = \sup \frac{\|\nabla f(x) - \nabla g(x)\|_2}{d(x,X^*)} = \sup \frac{\|\nabla h(x)\|_2}{d(x,X^*)} \leq \frac{\epsilon}{3} < \mu$, then

$$g^* = f^* \geq f(x) + \langle \nabla f(x), x_p^* - x\rangle + \frac{\mu}{2}\|x - x_p^*\|_2^2. \tag{30}$$

$\forall\, x \in \mathbb{R}^d \setminus X^*$

$$\begin{aligned}
\langle \nabla f(x), x_p^* - x\rangle &= \langle \nabla f(x) - \nabla g(x), x_p^* - x\rangle + \langle \nabla g(x), x_p^* - x\rangle \\
&\geq -\|\nabla f(x) - \nabla g(x)\|_2 \left\|x - x_p^*\right\|_2 + \langle \nabla g(x), x_p^* - x \\
&\geq -\|f - g\|_* \left\|x - x_p^*\right\|_2^2 + \langle \nabla g(x), x_p^* - x\rangle \qquad (31)\\
&\geq -\frac{\epsilon}{3}\left\|x - x_p^*\right\|_2^2 + \langle \nabla g(x), x_p^* - x\rangle \qquad (32)
\end{aligned}$$

and

$$\begin{aligned}
0 = h^* &= \omega(x) + \int_0^1 \langle \nabla h(x + t(x_p^* - x)), x_p^* - x\rangle \,\mathrm{d}t \\
&\geq h(x) - \int_0^1 \left\|\nabla h(x + t(x_p^* - x))\right\|_2 \|x - x_p^*\|_2 \,\mathrm{d}t \\
&\geq h(x) - \frac{1}{2}\|f - g\|_* \|x - x_p^*\|_2^2 \\
&\geq h(x) - \frac{\epsilon}{6}\|x - x_p^*\|_2^2 \qquad (33)
\end{aligned}$$

Therefore, $\forall\, x \in \mathbb{R}^d \setminus X^*$, $g^* \geq g(x) + \langle \nabla g(x), x^* - x\rangle + \frac{\mu-\epsilon}{2}\|x - x^*\|_2^2$ and for $x = x^* \in X^*$ the inequality is trivial: $g \in {}^*\mathrm{SC}(\mu - \epsilon)$.

Continuity of RSI^- and RSI^+: Given $f \in \mathrm{RSI}^+(L)$: $\forall\, x \in \mathbb{R}^d$, $\langle \nabla f(x), x - x_p^*\rangle \leq L\left\|x - x_p^*\right\|_2^2$. Consider $g = f + h$ with $h \in \mathcal{F}_{X^*}$ such that $\|h\|_* = \sup_{x \in \mathbb{R}^d \setminus \{X^*\}} \frac{\|\nabla h(x)\|_2}{d(x,X^*)} \leq \epsilon$, then we have $\forall\, x \in \mathbb{R}^d \setminus X^*$

$$\begin{aligned}
\langle \nabla g(x), x - x_p^*\rangle &= \langle \nabla f(x) + \nabla h(x), x - x_p^*\rangle = \langle \nabla f(x), x - x_p^*\rangle + \langle \nabla h(x), x - x_p^*\rangle \\
&\leq L\left\|x - x_p^*\right\|_2^2 + \|\nabla h(x)\|_2 \left\|x - x_p^*\right\|_2 \\
&\leq L\left\|x - x_p^*\right\|_2^2 + \epsilon\left\|x - x_p^*\right\|_2^2 = (L + \epsilon)\left\|x - x_p^*\right\|_2^2 \qquad (34)
\end{aligned}$$

(for $x = x^* \in X^*$ it is trivial and we have an equality), i.e. $g \in \mathrm{RSI}^+(L + \epsilon)$.

Similarly, if $f \in \mathrm{RSI}^-(\mu)$, i.e. $\forall\, x \in \mathbb{R}^d$, $\langle \nabla f(x), x - x_p^*\rangle \geq \mu\left\|x - x_p^*\right\|_2^2$, consider $g = f + h$ with $h \in \mathcal{F}_{X^*}$, $\|h\|_* = \sup_{x \in \mathbb{R}^d \setminus X^*} \frac{\|\nabla h(x)\|_2}{d(x,X^*)} < \epsilon < \mu$, then we have $\forall\, x \in \mathbb{R}^d \setminus X^*$

$$\begin{aligned}
\langle \nabla g(x), x - x_p^*\rangle &= \langle \nabla f(x) + \nabla h(x), x - x_p^*\rangle = \langle \nabla f(x), x - x_p^*\rangle + \langle \nabla h(x), x - x_p^*\rangle \\
&\geq \mu\left\|x - x_p^*\right\|_2^2 - \|\nabla h(x)\|_2 \left\|x - x_p^*\right\|_2 \\
&\geq \mu\left\|x - x_p^*\right\|_2^2 - \epsilon\left\|x - x_p^*\right\|_2^2 = (\mu - \epsilon)\left\|x - x_p^*\right\|_2^2 \qquad (35)
\end{aligned}$$

(for $x = x^* \in X^*$ it is trivial and we have an equality), i.e. $g \in \mathrm{RSI}^-(\mu - \epsilon)$.

Continuity of EB^- and EB^+: Given $f \in \mathrm{EB}^+(L)$: $\forall\, x \in \mathbb{R}^d$, $\|\nabla f(x)\|_2 \leq L\, d(x, X^*) = L\left\|x - x_p^*\right\|_2$, with $x_p^* \in X^*$ the unique projection of x on X^*; this implies

$$\sup_{x \in \mathbb{R}^d \setminus X^*} \frac{\|\nabla f(x)\|_2}{d(x, X^*)} \leq L. \tag{36}$$

Given $\epsilon > 0$, consider $g \in f + \mathcal{F}_{X^*}$, such that $\|f - g\|_* < \epsilon$: then,

$$\sup_{x \in \mathbb{R}^d \setminus X^*} \frac{\|\nabla g(x)\|_2}{d(x, X^*)} \leq \sup_{x \in \mathbb{R}^d \setminus X^*} \frac{\|\nabla g(x) - \nabla f(x)\|_2}{d(x, X^*)} + \sup_{x \in \mathbb{R}^d \setminus X^*} \frac{\|\nabla f(x)\|_2}{d(x, X^*)} \leq \epsilon + L \tag{37}$$

Additionally, since $g \in f + \mathcal{F}_{x^*}$, $\nabla g(x^*) = 0 \; \forall \, x^* \in X^*$, therefore

$$\|\nabla g(x)\|_2 \leq (L+\epsilon) d(x, X^*), \qquad \forall \, x \in \mathbb{R}^d \tag{38}$$

i.e. $g \in \mathrm{EB}^+(L+\epsilon)$.

Given $f \in \mathrm{EB}^-(\mu)$: $\forall \, x \in \mathbb{R}^d \;\; \|\nabla f(x)\|_2 \geq \mu \, d(x, X^*) = \mu \|x - x_p^*\|_2$. Fix $\epsilon > 0$ and consider $g \in f + \mathcal{F}_{x^*}$, such that $\|f - g\|_* < \epsilon < \mu$; in particular $\forall \, x \in \mathbb{R}^d \setminus X^*$, $\|\nabla f(x) - \nabla g(x)\|_2 < \epsilon \, d(x, X^*)$. Then, $\forall \, x \in \mathbb{R}^d \setminus X^*$

$$\begin{aligned} 0 < (\mu - \epsilon) \, d(x, X^*) &\leq \|\nabla f(x)\|_2 - \epsilon d(x, X^*) \\ &\leq \|\nabla f(x)\|_2 - \|\nabla f(x) - \nabla g(x)\|_2 \\ &\leq \|\nabla f(x) - \nabla f(x) + \nabla g(x)\|_2 = \|\nabla g(x)\|_2 \end{aligned} \tag{39}$$

(for $x = x^* \in X^*$ the inequality is trivial), i.e. $g \in \mathrm{EB}^-(\mu - \epsilon)$.

Continuity of PL^- and PL^+:

Let $f \in \mathrm{PL}^-(\mu)$ and $\epsilon > 0$. From Figure 1 we have $f \in \mathrm{QG}^-(\mu)$. Given $g \in f + \mathcal{F}_{X^*}$ such that $\|f - g\|_* < \mu$, for any $x \in \mathbb{R}^d$ with projection x_p^* onto X^* we have:

$$\|\nabla f(x) - \nabla g(x)\|_2 \leq \|f - g\|_* d \, (x, X^*) \tag{40}$$

additionally for $t \in [0, 1]$, $d(x_p^* + t(x - x_p^*), X^*) = t \left\| x - x_p^* \right\|_2$ since $x_p^* \in X^*$ and

$$\begin{aligned} |f(x) - g(x)| = |f(x) - g(x) - (f(x_p^*) - g(x_p^*))| &= \left| \int_0^1 \langle \nabla (f - g)(x_p^* + t(x - x_p^*)), x - x_p^* \rangle \, \mathrm{d}t \right| \\ &\leq \int_0^1 \|f - g\|_* d(x_p^* + t(x - x_p^*), X^*) \|x - x_p^*\|_2 \, \mathrm{d}t \\ &\leq \|f - g\|_* \|x - x_p^*\|_2^2 \int_0^1 t \, \mathrm{d}t \\ &\leq \frac{\|f - g\|_*}{2} \, d \, (x, X^*)^2 \end{aligned} \tag{41}$$

Since $f \in \mathrm{QG}^-(\mu)$ and $\|f - g\|_* < \mu$:

$$\begin{aligned} g(x) - g(x_p^*) &\geq f(x) - f^* - |f(x) - g(x) - (f(x_p^*) - g(x_p^*))| \\ &\geq \frac{\mu - \|f - g\|_*}{2} \, d(x, X^*)^2 \geq 0 \end{aligned} \tag{42}$$

Thus g admits a minimum value g^* which is attained at any $x^* \in X^*$. Therefore, $\forall \, x \in \mathbb{R}^d$

$$g(x) - g^* \geq \frac{\mu - \|f - g\|_*}{2} \, d(x, X^*)^2. \tag{43}$$

Since $f \in \mathrm{PL}^-(\mu)$, we have

$$\begin{aligned} \|\nabla g(x)\|_2^2 &= \|\nabla g(x) - \nabla f(x)\|_2^2 + \|\nabla f(x)\|_2 + 2\langle \nabla g(x) - \nabla f(x), \nabla f(x) \rangle \\ &\geq 0 + 2\mu(f(x) - f^*) - 2\|\nabla g(x) - \nabla f(x)\|_2 \sqrt{2\mu(f(x) - f^*)} \\ &\geq 2\mu(g(x) - g^*) - 2\mu |f(x) - g(x) - (f^* - g^*)| \\ &\quad - 2\|f - g\|_* \, d(x, X^*) \sqrt{2\mu(g(x) - g^* + |f(x) - g(x) - (f^* - g^*)|)} \end{aligned} \tag{44}$$

The second term can be easily bounded by (41) and 43; the third term can be bounded as follows

$$\begin{aligned}\sqrt{g(x)-g^*+|f(x)-g(x)-(f^*-g^*)|} &\leq \sqrt{(g(x)-g^*)}+\sqrt{|f(x)-g(x)-(f^*-g^*)|}\\ &\leq \sqrt{(g(x)-g^*)}+\sqrt{\frac{\|f-g\|_*}{2}d(x,X^*)^2}\\ &\leq \sqrt{(g(x)-g^*)}+\sqrt{\frac{\|f-g\|_*}{\mu-\|f-g\|_*}(g(x)-g^*)}\\ &= \left(1+\sqrt{\frac{\|f-g\|_*}{\mu-\|f-g\|_*}}\right)\sqrt{(g(x)-g^*)} \end{aligned} \tag{45}$$

where we applied again (41) and (43). Finally we get:

$$\begin{aligned}\|\nabla g(x)\|_2^2 &\geq 2\left[\mu-\frac{\mu\|f-g\|_*}{\mu-\|f-g\|_*}-\|f-g\|_*\sqrt{\frac{2}{\mu-\|f-g\|_*}}\left(1+\sqrt{\frac{\|f-g\|_*}{\mu-\|f-g\|_*}}\right)\right](g(x)-g^*)\\ &\geq 2(\mu-\epsilon)(g(x)-g^*)\end{aligned} \tag{46}$$

provided that $\|f-g\|_*$ is small enough. Indeed, the quantity

$$0\leq \frac{\mu\|f-g\|_*}{\mu-\|f-g\|_*}+\|f-g\|_*\sqrt{\frac{2}{\mu-\|f-g\|_*}}\left(1+\sqrt{\frac{\|f-g\|_*}{\mu-\|f-g\|_*}}\right)\to 0,\quad \text{as } \|f-g\|_*\to 0,$$

therefore $\forall\,\epsilon>0$, $\exists\,\delta>0$ such that for $\|f-g\|_*\leq\delta$, we have

$$\frac{\mu\|f-g\|_*}{\mu-\|f-g\|_*}+\|f-g\|_*\sqrt{\frac{2}{\mu-\|f-g\|_*}}\left(1+\sqrt{\frac{\|f-g\|_*}{\mu-\|f-g\|_*}}\right)\leq\epsilon.$$

In conclusion, $g\in \mathrm{PL}^-(\mu-\epsilon)$.

Let us now consider $f\in \mathrm{PL}^+(L)\cap \mathrm{QG}^-(\mu)$, and $g\in f+\mathcal{F}_{X^*}$ such that $||f-g||_*<\mu$.

$$\|\nabla g(x)\|_2^2=\|\nabla g(x)-\nabla f(x)\|_2^2+\|\nabla f(x)\|_2^2+2\langle\nabla g(x)-\nabla f(x),\nabla f(x)\rangle \tag{47}$$

The second term can be estimated thanks to (41) and (43):

$$\begin{aligned}\|\nabla f(x)\|_2^2 &\leq 2L(f(x)-g(x)-(f^*-g^*))+2L(g(x)-g^*)\\ &\leq 2L\left(\frac{||f-g||_*}{\mu-||f-g||_*}+1\right)(g(x)-g^*);\end{aligned} \tag{48}$$

and similarly the third term:

$$\begin{aligned}\langle\nabla g(x)-\nabla f(x),\nabla f(x)\rangle &\leq \|\nabla g(x)-\nabla f(x)\|_2\,\|\nabla f(x)\|_2\\ &\leq ||f-g||_* d(x,X^*)\sqrt{2L\left(\frac{||f-g||_*}{\mu-||f-g||_*}+1\right)(g(x)-g^*)}\\ &\leq ||f-g||_*\sqrt{\frac{4L}{\mu-||f-g||_*}\left(\frac{||f-g||_*}{\mu-||f-g||_*}+1\right)}(g(x)-g^*).\end{aligned} \tag{49}$$

This finally leads to:

$$\|\nabla g(x)\|_2^2\leq 2(L+K)(g(x)-g^*)\leq 2(L+\epsilon)(g(x)-g^*) \tag{50}$$

where

$$K=\left[\frac{\|f-g\|_*}{\mu-\|f-g\|_*}+\frac{L\|f-g\|_*}{\mu-\|f-g\|_*}+\|f-g\|_*\sqrt{\frac{4L}{\mu-\|f-g\|_*}\left(\frac{\|f-g\|_*}{\mu-\|f-g\|_*}+1\right)}\right],$$

provided that $\|f-g\|_*$ is small enough. Following a similar argument as before, we can easily see that $K \geq 0$ and $K \to 0$ as $\|f-g\|_* \to 0$, therefore $\forall\, \epsilon > 0$, $\exists\, \delta > 0$ such that if $\|f-g\|_* \leq \delta$, then $K \leq \epsilon$. Therefore, $g \in \mathrm{PL}^+(L+\epsilon)$.

Continuity of QG^- and QG^+:

Given $f \in \mathrm{QG}^+(L)$: $f(x) - f^* \leq \frac{L}{2}\, d(x, X^*)^2$, $\forall\, x \in \mathbb{R}^d$. Consider $g = f + h$, $h \in \mathcal{F}_{X^*}$ with $\|h\|_* = \sup \frac{\|\nabla f(x) - \nabla g(x)\|_2}{d(x,X^*)} = \sup \frac{\|\nabla h(x)\|_2}{d(x,X^*)} \leq \epsilon$, then $\forall\, x \in \mathbb{R}^d$, with $x_p^* \in X^*$ the corresponding projection on X^*,

$$\begin{aligned}
g(x) - g^* &= f(x) + h(x) - (f^* + h^*) = f(x) - f^* + h(x) - h^* \\
&\leq \frac{L}{2} d(x, X^*)^2 + \int_0^1 \langle \nabla h(x_p^* + t(x - x_p^*)), x - x_p^* \rangle \,\mathrm{d}t \\
&\leq \frac{L}{2} d(x, X^*)^2 + \int_0^1 \left\| \nabla h(x_p^* + t(x - x_p^*)) \right\|_2 \|x - x_p^*\|_2 \,\mathrm{d}t \\
&\leq \frac{L}{2} d(x, X^*)^2 + \int_0^1 \frac{\left\| \nabla h(x_p^* + t(x - x_p^*)) \right\|_2}{\|x_p^* + t(x - x_p^*) - X^*\|_2} \|x_p^* + t(x - x_p^*) - X^*\|_2 \|x - x_p^*\|_2 \,\mathrm{d}t \\
&\leq \frac{L}{2} d(x, X^*)^2 + \|h\|_* \|x - x_p^*\|_2^2 \int_0^1 t \,\mathrm{d}t \\
&\leq \frac{L+\epsilon}{2} d(x, X^*)^2
\end{aligned} \tag{51}$$

where we used $\|x_p^* + t(x - x_p^*) - X^*\|_2 \leq t\|x - x_p^*\|_2$, $\forall\, t \in [0,1]$; as before, for $x = x^* \in X^*$ the inequality is trivial. Therefore, $g \in \mathrm{QG}^+(L+\epsilon)$.

The proof that $g \in \mathrm{QG}^-(\mu - \epsilon)$ if $f \in \mathrm{QG}^-(\mu)$ for $g \in f + \mathcal{F}_{X^*}$, $\|f-g\|_* \leq \epsilon < \mu$, follows the same argument.

C Graph of lower assumptions

$\mathrm{SC}^-(\mu) \to {}^*\mathrm{SC}^-(\mu)$**:** Immediate by taking $y = x_p^*$ (the projection of $x \in \mathbb{R}^d$ onto X^*) in the definition of strong convexity.

${}^*\mathrm{SC}^-(\mu) \to \mathrm{PL}^-(\mu)$**:** Assume $f \in {}^*\mathrm{SC}^-(\mu)$: $f^* \geq f(x) + \langle \nabla f(x), x_p^* - x \rangle + \frac{\mu}{2} \left\| x_p^* - x \right\|_2^2$, $\forall\, x \in \mathbb{R}^d$. Hence,

$$f^* - f(x) \geq -\frac{1}{2\mu} \|\nabla f(x)\|_2^2 + \frac{1}{2\mu} \left\| \nabla f(x) + \mu \left(x_p^* - x \right) \right\|_2^2 \geq -\frac{1}{2\mu} \|\nabla f(x)\|_2^2 \tag{52}$$

i.e. $\|\nabla f(x)\|_2^2 \geq 2\mu(f - f^*)$. Therefore, $f \in \mathrm{PL}^-(\mu)$.

$\mathrm{PL}^-(\mu) \to \mathrm{QG}^-(\mu)$**:** The claim was originally proven in Karimi et al. (2016), following some arguments from Bolte et al. (2017) and Zhang (2017) and we will report it here for the sake of completeness.

Consider the gradient flow of $g(x) = \sqrt{f(x) - f^*}$: $x'(t) = -\nabla g(x(t))$. Note the $f \in \mathrm{PL}^-(\mu)$ implies that $\|\nabla g(x)\|_2^2 \geq \frac{\mu}{2} > 0\ \forall\, x \in \mathbb{R}^d$; in particular, despite the fact that g attains its minimum on the set X^*, ∇g may not be defined on X^* and the gradient flow equation ceases to be defined once X^* is reached. We then study the path of a gradient flow of g until it hits X^*: $\forall\, x_0 \in \mathbb{R}^d$, $\forall\, T > 0$ for which the flow is defined,

$$\begin{aligned}
g(x_0) \geq g(x_0) - g(x_T) &= -\int_0^T \langle \nabla g(x(t)), x'(t) \rangle \,\mathrm{d}t = \int_0^T \|\nabla g(x(t))\|_2^2 \,\mathrm{d}t \\
&\geq \int_0^T \frac{\mu}{2} \,\mathrm{d}t = \frac{\mu}{2} T,
\end{aligned} \tag{53}$$

where the first inequality follows from the fact that g is non-negative and the second inequality follows from the $\mathrm{PL}^-(\mu)$ property. This proves the existence of $T^* = T^*(x_0)$ such that $x_{T^*} \in X^*$.

Therefore, $\forall\, x_0 \in \mathbb{R}^d$

$$\begin{aligned} g(x_0) =& g(x_0) - g(x_{T^*}) = \int_0^{T^*} \|\nabla g(x(t))\|_2^2 \,\mathrm{d}t \\ \geq& \sqrt{\frac{\mu}{2}} \int_0^{T^*} \|\nabla g(x(t))\|_2 \,\mathrm{d}t = \sqrt{\frac{\mu}{2}} \int_0^{T^*} \|x'(t)\|_2 \,\mathrm{d}t \\ \geq& \sqrt{\frac{\mu}{2}} \left\| \int_0^{T^*} x'(t) \,\mathrm{d}t \right\|_2 = \sqrt{\frac{\mu}{2}} \|x_0 - x_{T^*}\|_2 \\ \geq& \sqrt{\frac{\mu}{2}} d(x_0, X^*) \end{aligned} \tag{54}$$

and, by squaring on both sides,

$$f(x) - f^* = g(x)^2 \geq \frac{\mu}{2} d(x, X^*)^2; \tag{55}$$

i.e. $f \in \mathrm{QG}^-(\mu)$.

$\underline{{}^*\mathrm{SC}^-(\mu_1) \text{ and } \mathrm{QG}^-(\mu_2) \to \mathrm{RSI}^-\left(\frac{\mu_1+\mu_2}{2}\right)\textbf{:}}$ For $f \in {}^*\mathrm{SC}^-(\mu_1) \cap \mathrm{QG}^-(\mu_2)$, we have

$$\langle \nabla f(x), x - x_p^* \rangle \geq f(x) - f^* + \frac{\mu_1}{2} \left\| x_p^* - x \right\|_2^2 \geq \frac{\mu_1 + \mu_2}{2} \left\| x_p^* - x \right\|_2^2 \tag{56}$$

i.e. $f \in \mathrm{RSI}^-\left(\frac{\mu_1+\mu_2}{2}\right)$. Note that this holds also for non positive μ_1. In particular, if $f \in \mathrm{QG}^-(\mu)$ and f is *-convex ($\mu_1 = 0$), then $f \in \mathrm{RSI}^-(\frac{\mu}{2})$.

$\underline{{}^*\mathrm{SC}^-(\mu) \to \mathrm{RSI}^-(\mu)\textbf{:}}$ This follows directly from the three previous results. Indeed, ${}^*\mathrm{SC}^-(\mu) \subseteq \mathrm{PL}^-(\mu) \subseteq \mathrm{QG}^-(\mu)$ and ${}^*\mathrm{SC}^-(\mu) \cap \mathrm{QG}^-(\mu) \subseteq \mathrm{RSI}^-(\mu)$.

$\underline{\mathrm{RSI}^-(\mu) \to \mathrm{QG}^-(\mu)\textbf{:}}$ For every $x \in \mathbb{R}^d$ consider the line segment $x(t) = x_p^* + t(x - x_p^*)$, $t \in [0,1]$, with $x_p^* \in X^*$ the projection of x onto X^*. It is clear that $\forall\, t \in [0,1]$ the projection of $x(t)$ onto X^* is still x_p^*. Since $f \in \mathrm{RSI}^-(\mu)$, $\forall\, x \in \mathbb{R}^d$

$$\langle \nabla f(x_p^* + t(x - x_p^*)), t(x - x_p^*) \rangle \geq \mu \|t(x - x_p^*)\|_2^2 = \mu t^2 \|(x - x_p^*)\|_2^2, \tag{57}$$

therefore

$$f(x) - f^* = \int_0^1 \langle \nabla f(x_p^* + t(x - x_p^*)), x - x_p^* \rangle \,\mathrm{d}t \geq \int_0^1 \mu t \left\| x - x_p^* \right\|_2^2 \,\mathrm{d}t = \frac{\mu}{2} \left\| x - x_p^* \right\|_2^2, \tag{58}$$

implying that $f \in \mathrm{QG}^-(\mu)$.

$\underline{\mathrm{RSI}^-(\mu) \to \mathrm{EB}^-(\mu)\textbf{:}}$ It follows from Cauchy-Schwartz inequality.

$\underline{\mathrm{PL}^-(\mu_1) \cap \mathrm{QG}^-(\mu_2) \to \mathrm{EB}^-\left(\sqrt{\mu_1\mu_2}\right)\textbf{:}}$ Assume $f \in \mathrm{PL}^-(\mu_1) \cap \mathrm{QG}^-(\mu_2)$:

$$\frac{1}{2} \|\nabla f(x)\|_2^2 \geq \mu_1 \left(f(x) - f^*\right) \geq \frac{\mu_1 \mu_2}{2} \|x - x_p^*\|_2^2 \tag{59}$$

i.e. $\|\nabla f(x)\|_2 \geq \sqrt{\mu_1\mu_2} \|x - x_p^*\|_2$.

Hence, $f \in \mathrm{EB}^-\left(\sqrt{\mu_1\mu_2}\right)$. Note that $\mathrm{PL}^-(\mu) \subseteq \mathrm{QG}^-(\mu)$, therefore $\mathrm{PL}^-(\mu) \subseteq \mathrm{EB}^-(\mu)$ (set $\mu_1 = \mu_2 = \mu$).

$\underline{\mathrm{EB}^-(\mu) \cap \mathrm{QG}^+(L) \to \mathrm{PL}^-(\mu^2/L)\textbf{:}}$ Given $f \in \mathrm{EB}^-(\mu) \cap \mathrm{QG}^+(L)$, $\forall\, x \in \mathbb{R}^d$

$$\|\nabla f(x)\|_2^2 \geq \mu^2 \|x - x_p^*\|_2^2 \geq \frac{2\mu^2}{L} \left(f(x) - f^*\right) \tag{60}$$

i.e. $f \in \mathrm{PL}^-(\mu^2/L)$.

D Graph of upper assumptions

$\underline{\mathrm{SC}^+(L) \to \mathrm{PL}^+(L)}$**:** Assume $f \in \mathrm{SC}^+(L)$, hence $\forall\, x, y \in R^d$

$$\begin{aligned} f(y) &\leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2 \\ &= f(x) - \frac{1}{2L} \|\nabla f(x)\|_2^2 + \frac{1}{2L} \|\nabla f(x) + L(y - x)\|_2^2 \end{aligned} \tag{61}$$

In particular, $\forall\, x, y \in \mathbb{R}^d$

$$f^* \leq f(y) \leq f(x) - \frac{1}{2L} \|\nabla f(x)\|_2^2 + \frac{1}{2L} \|\nabla f(x) + L(y - x)\|_2^2 \tag{62}$$

and by choosing $y = x - \frac{\nabla f(x)}{L}$, we have

$$f^* - f(x) \leq -\frac{1}{2L} \|\nabla f(x)\|_2^2, \quad \text{i.e.} \quad \frac{1}{2} \|\nabla f(x)\|_2^2 \leq L(f(x) - f^*) \tag{63}$$

Hence, $f \in \mathrm{PL}^+(L)$.

$\underline{\mathrm{PL}^+(L) \to {}^*\mathrm{SC}^+(L)}$**:** Assume $f \in \mathrm{PL}^+(L)$, hence

$$\begin{aligned} f^* - f(x) \leq & -\frac{1}{2L} \|\nabla f(x)\|_2^2 \\ \leq & -\frac{1}{2L} \|\nabla f(x)\|_2^2 + \frac{1}{2L} \left\|\nabla f(x) + L\left(x_p^* - x\right)\right\|_2^2 \\ f^* \leq & f(x) + \langle \nabla f(x), x_p^* - x \rangle + \frac{L}{2} \left\|x_p^* - x\right\|_2^2 \end{aligned} \tag{64}$$

Hence, $f \in {}^*\mathrm{SC}^+(L)$.

$\underline{\mathrm{PL}^+(L) \to \mathrm{QG}^+(L)}$**:** Assume $f \in \mathrm{PL}^+(L)$ and consider the function $g(x) = \sqrt{f(x) - f^*}$: since $f \in \mathrm{PL}^+(L)$, we have $\|\nabla g(x)\|_2^2 \leq \frac{L}{2}$, $\forall\, x \in \mathbb{R}^d$. Then,

$$\begin{aligned} g(x) =& g(x) - g(x_p^*) = \int_0^1 \langle \nabla g(x_p^* + t(x - x_p^*)), x - x_p^* \rangle \, \mathrm{d}t \\ \leq & \int_0^1 \left\|\nabla g(x_p^* + t(x - x_p^*))\right\|_2 \left\|(x - x_p^*)\right\|_2 \, \mathrm{d}t \\ \leq & \int_0^1 \sqrt{\frac{L}{2}} \left\|x - x_p^*\right\|_2 \, \mathrm{d}t \leq \sqrt{\frac{L}{2}} \left\|x - x_p^*\right\|_2 \end{aligned} \tag{65}$$

Therefore, by squaring on both sides,

$$f(x) - f^* \leq \frac{L}{2} \|x - x_p^*\|_2^2. \tag{66}$$

Note: this result is not explicit in the graph as it can be recover by following the existing edges. However we needed to prove it here for the following result.

$\underline{\mathrm{PL}^+(L) \to \mathrm{EB}^+(L)}$**:** Assume $f \in \mathrm{PL}^+(L_1) \cap \mathrm{QG}^+(L_2)$, then

$$\|\nabla f(x)\|_2^2 \leq 2L_1(f(x) - f^*) \leq L_1 L_2 \|x - x_p^*\|_2^2, \tag{67}$$

hence $f \in \mathrm{EB}^+(\sqrt{L_1 L_2})$. In particular, from the previous result we have that if $f \in \mathrm{PL}^+(L)$, then $f \in \mathrm{QG}^+(L)$, hence $f \in \mathrm{EB}^+(L)$ (take $L_1 = L_2 = L$).

$\underline{\mathrm{EB}^+(L) \to \mathrm{RSI}^+(L)}$**:** Given $f \in \mathrm{EB}^+(L)$,

$$\langle \nabla f(x), x - x_p^* \rangle \leq \|\nabla f(x)\|_2 \cdot \|x - x_p^*\|_2 \leq L\|x - x_p^*\|_2^2, \tag{68}$$

therefore $f \in \mathrm{RSI}^+(L)$.

$\underline{{}^*\mathrm{SC}^+(L) \to \mathrm{QG}^+(L)}$**:** For each $x \in \mathbb{R}^d$, with $x_p^* \in X^*$ its projection onto X^*, define

$$g(t) = \frac{\frac{L}{2}\|t(x - x_p^*)\|_2^2 - \big(f(x_p^* + t(x - x_p^*)) - f^*\big)}{t}, \qquad t \in (0, +\infty).$$

We verify that

$$g'(t) = \frac{\frac{L}{2}\left\|t(x - x_p^*)\right\|_2^2 - \langle \nabla f(x_p^* + t(x - x_p^*)), x - x_p^* \rangle + \big(f(x^* + t(x - x_p^*)) - f^*\big)}{t^2} \geq 0 \tag{69}$$

since $f \in {}^*\mathrm{SC}^+(L)$. Therefore, g is monotonically increasing on $(0, +\infty)$. Additionally, g can be continuously extended in $t = 0$ by l'Hôpital's rule:

$$\lim_{t \to 0_+} g(t) = \lim_{t \to 0_+} Lt\|(x - x_p^*)\|_2^2 - \langle \nabla f(x_p^* + t(x - x_p^*)), x - x_p^* \rangle = 0.$$

Therefore,

$$g(1) = \frac{L}{2}\|x - x_p^*\|_2^2 - (f(x) - f^*) \geq g(0) = 0$$

i.e. $f(x) - f^* \leq \frac{L}{2}\|x - x_p^*\|_2^2$: $f \in \mathrm{QG}^+(L)$.

$\underline{{}^*\mathrm{SC}^+(L) \to \mathrm{RSI}^+(L)}$**:** Let $f \in {}^*\mathrm{SC}^+(L_1) \cap \mathrm{QG}^+(L_2)$:

$$\langle \nabla f(x), x - x_p^* \rangle \leq f(x) - f^* + \frac{L_1}{2}\left\|x - x_p^*\right\|_2^2 \leq \frac{L_1 + L_2}{2}\left\|x - x_p^*\right\|_2^2, \tag{70}$$

therefore $f \in \mathrm{RSI}^+(\frac{L_1 + L_2}{2})$. In particular, since ${}^*\mathrm{SC}^+(L) \subseteq \mathrm{QG}^+(L)$, then ${}^*\mathrm{SC}^+(L) \subseteq \mathrm{RSI}^+(L)$.

$\underline{\mathrm{RSI}^+(L) \to {}^*\mathrm{SC}^+(2L)}$**:** For $f \in \mathrm{RSI}^+(L)$, we have

$$\langle \nabla f(x), x - x_p^* \rangle \leq L\left\|x - x_p^*\right\|_2^2 \leq f(x) - f^* + L\left\|x - x_p^*\right\|_2^2 \tag{71}$$

i.e. $f \in {}^*\mathrm{SC}^+(2L)$.

$\underline{\mathrm{RSI}^+(L) \to \mathrm{QG}^+(L)}$**:** For every $x \in \mathbb{R}^d$ consider the line segment $x(t) = x_p^* + t(x - x_p^*)$, $t \in [0, 1]$; recall that $\forall\, t \in [0, 1]$ the projection of $x(t)$ onto X^* is still x_p^*. Since $f \in \mathrm{RSI}^+(L)$, $\forall\, x \in \mathbb{R}^d$

$$\langle \nabla f(x_p^* + t(x - x_p^*)), t(x - x_p^*) \rangle \leq L\|t(x - x_p^*)\|_2^2 = Lt^2\|x - x_p^*\|_2^2, \tag{72}$$

Therefore, $f \in \mathrm{QG}^+(L)$:

$$f(x) - f^* = \int_0^1 \langle \nabla f(x^* + t(x - x_p^*)), x - x_p^* \rangle \, \mathrm{d}t \leq \int_0^1 Lt\left\|x - x_p^*\right\|_2^2 \, \mathrm{d}t = \frac{L}{2}\left\|x - x_p^*\right\|_2^2. \tag{73}$$

$\underline{\mathrm{SC}^-(\mu) \text{ and } \mathrm{QG}^+(L) \to \mathrm{EB}^+\left(L + \sqrt{L(L-\mu)}\right)}$**:** Assume $f \in \mathrm{SC}^-(\mu) \cap \mathrm{QG}^+(L)$, with $\mu < L$, and μ can be non positive (we recall that $f \in \mathrm{SC}^-(0)$ is convex). The case $\mu \geq L$ is trivial as it implies $f(x) - f^* = \frac{L}{2}\|x - x_p^*\|^2$ $\forall\, x \in \mathbb{R}^d$.

We have by definition: $\forall\, x, y \in \mathbb{R}^d$

$$f(x) - f^* + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2 \overset{\mathrm{SC}^-}{\leq} f(y) - f^* \overset{\mathrm{QG}^+}{\leq} \frac{L}{2}\left\|y - y_p^*\right\|_2^2 \leq \frac{L}{2}\left\|y - x_p^*\right\|_2^2; \tag{74}$$

in particular,

$$f(x) - f^* + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_2^2 \leq \frac{L}{2} \left\|y - x_p^*\right\|_2^2 \tag{75}$$

and by choosing $y = \frac{Lx_p^* - \mu x + \nabla f(x)}{L - \mu}$ we have

$$L\mu \left\|x - x_p^*\right\|_2^2 + \|\nabla f(x)\|_2^2 + 2L\langle \nabla f(x), x_p^* - x \rangle \leq 2(L - \mu) \cdot (f^* - f(x)) \tag{76}$$

The RHS is non positive, then by removing it and factoring the LHS

$$\left\|\nabla f(x) + L(x_p^* - x)\right\|_2^2 \leq L(L - \mu) \left\|x - x_p^*\right\|_2^2; \tag{77}$$

finally, by triangle inequality,

$$\|\nabla f(x)\|_2 - L \left\|x_p^* - x\right\|_2 \leq \sqrt{L(L - \mu)} \left\|x - x_p^*\right\|_2 \tag{78}$$

$$\|\nabla f(x)\|_2 \leq \left(L + \sqrt{L(L - \mu)}\right) \left\|x - x_p^*\right\|_2 \tag{79}$$

Hence, $f \in \mathrm{EB}^+\left(L + \sqrt{L(L-\mu)}\right)$. Note that for $\mu = 0$ (i.e. f is convex), we have $\mathrm{QG}^+(L) \to \mathrm{EB}^+(2L)$.

$\mathrm{SC}^-(\mu)$ and ${}^*\mathrm{SC}^+(L) \to \mathrm{EB}^+\left(L + 2\max\{-\mu, 0\}\right)$**:** Assume $f \in \mathrm{SC}^-(\mu) \cap {}^*\mathrm{SC}^+(L)$. In particular $f \in \mathrm{QG}^+(L)$, then all the previous results still hold. From (76) we have

$$\begin{aligned} L\mu \left\|x - x_p^*\right\|_2^2 + \|\nabla f(x)\|_2^2 + 2L\langle \nabla f(x), x_p^* - x \rangle &\leq 2(L - \mu) \cdot (f^* - f(x)) \\ &\leq 2(L - \mu) \cdot \left[\langle \nabla f(x), x_p^* - x \rangle + \frac{L}{2} \|x - x_2^*\|_2^2\right] \end{aligned} \tag{80}$$

thanks to $f \in {}^*\mathrm{SC}^+(L)$, i.e.

$$\|\nabla f(x)\|_2^2 + 2\mu\langle \nabla f(x), x_p^* - x \rangle \leq L(L - 2\mu) \left\|x - x_p^*\right\|_2^2.$$

After rearranging the terms, we obtain $\left\|\nabla f(x) + \mu(x_p^* - x)\right\|_2^2 \leq (L - \mu)^2 \left\|x - x_p^*\right\|_2^2$ and by triangle inequality

$$\|\nabla f(x)\|_2 - |\mu| \left\|x_p^* - x\right\|_2 \leq (L - \mu) \left\|x - x_p^*\right\|_2, \tag{81}$$

i.e.

$$\|\nabla f(x)\|_2 \leq (L + 2\max\{-\mu, 0\}) \left\|x - x_p^*\right\|_2.$$

Finally $f \in \mathrm{EB}^+\left(L + 2\max\{-\mu, 0\}\right)$. In particular, under convex assumption ($\mu = 0$), ${}^*\mathrm{SC}^+(L) \to \mathrm{EB}^+(L)$.

$\mathrm{QG}^-(\mu)$ and $\mathrm{EB}^+(L) \to \mathrm{PL}^+\left(\frac{L^2}{\mu}\right)$**:** Let $f \in \mathrm{QG}^-(\mu) \cap \mathrm{EB}^+(L)$, we have:

$$\frac{1}{2} \|\nabla f(x)\|_2^2 \leq \frac{1}{2} L^2 \left\|x - x_p^*\right\|_2^2 \leq \frac{1}{2} L^2 \frac{2}{\mu} (f(x) - f^*) = \frac{L^2}{\mu} (f(x) - f^*), \tag{82}$$

therefore, $f \in \mathrm{PL}^+\left(\frac{L^2}{\mu}\right)$.

E Rates of convergence

Under $\mathrm{SC}^-(\mu)$ and $\mathrm{SC}^+(L)$ This is a known result and we refer to the proof in Section 3.4.2 in Bubeck (2015). Let's assume $f \in \mathrm{SC}^-(\mu) \cap \mathrm{SC}^+(L)$ with $L > \mu$ (the other case is trivial): $\forall\, x, y, z \in \mathbb{R}^d$

$$f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|_2^2 \overset{\mathrm{SC}^-(\mu)}{\leq} f(x) \overset{\mathrm{SC}^+(L)}{\leq} f(z) + \langle \nabla f(z), x - z \rangle + \frac{L}{2} \|x - z\|_2^2 \tag{83}$$

i.e. $\forall\, x, y, z \in \mathbb{R}^d$

$$f(z) - f(y) + \langle \nabla f(z), x - z\rangle - \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x - z\|_2^2 - \frac{\mu}{2}\|x - y\|_2^2 \geq 0. \tag{84}$$

By minimizing the left hand side of the above expression with respect to the variable x, we find that for

$$x = \frac{Lz - \mu y + \nabla f(y) - \nabla f(z)}{L - \mu} \tag{85}$$

the inequality becomes

$$f(y) - f(z) \leq \frac{1}{L-\mu}\left[\langle z - y, \mu\nabla f(z) - L\nabla f(y)\rangle - \frac{1}{2}\|\nabla f(y) - \nabla f(z)\|_2^2 - \frac{L\mu}{2}\|y - z\|_2^2\right] \tag{86}$$

$\forall\, y, z \in \mathbb{R}^d$. By swapping the roles of y and z, summing, and rearranging terms, we obtain the well-known inequality (see, e.g. Nesterov (2004)): $\forall\, y, z \in \mathbb{R}^d$

$$\langle z - y, \nabla f(z) - \nabla f(y)\rangle \geq \frac{1}{L+\mu}\left(\|\nabla f(y) - \nabla f(z)\|^2 + L\mu\|y - z\|^2\right). \tag{87}$$

Note that $f \in \mathrm{SC}^-(\mu)$ implies that $X^* = \{x^*\}$. In conclusion,

$$\begin{aligned}
\|x_{n+1} - x^*\|_2^2 =\ & \|x_n - x^* - \alpha\nabla f(x_n)\|_2^2 \\
& \|x_n - x^*\|_2^2 - 2\alpha\langle \nabla f(x_n), x_n - x^*\rangle + \alpha^2\|\nabla f(x_n)\|_2^2 \\
\leq\ & \left(1 - \frac{2\alpha L\mu}{L+\mu}\right)\|x_n - x^*\|_2^2 + \alpha\left(\alpha - \frac{2}{L+\mu}\right)\|\nabla f(x_n)\|_2^2 \\
=\ & \left(\frac{\kappa - 1}{\kappa + 1}\right)^2\|x_n - x^*\|_2^2
\end{aligned} \tag{88}$$

for $\alpha = \frac{2}{L+\mu}$.

Under $\mathrm{PL}^-(\mu)$ and $\mathrm{SC}^+(L)$ Let's assume $f \in \mathrm{PL}^-(\mu) \cap \mathrm{SC}^+(L)$. Then,

$$\begin{aligned}
f(x_{n+1}) - f^* &\leq f(x_n) - f^* - \alpha\left(1 - \frac{L\alpha}{2}\right)\|\nabla f(x_n)\|_2^2 \\
&\leq f(x_n) - f^* - 2\mu\alpha\left(1 - \frac{L\alpha}{2}\right)(f(x_n) - f^*) \\
&= \left(1 - \frac{1}{\kappa}\right)(f(x_n) - f^*)
\end{aligned} \tag{89}$$

for $\alpha = \frac{1}{L}$.

Under ${}^*\mathrm{SC}^-(\mu)$ and $\mathrm{PL}^+(L)$ Assume $f \in {}^*\mathrm{SC}^-(\mu) \cap \mathrm{PL}^+(L)$. $\forall\ n \in \mathbb{N}$, let $x_{n,p}^*$ be the projection of x_n on X^*. Then,

$$\begin{aligned}
d(x_{n+1}, X^*)^2 \leq \left\|x_{n+1} - x_{n,p}^*\right\|_2^2 =\ & \left\|x_n - x_{n,p}^*\right\|_2^2 - 2\alpha\langle x_n - x_{n,p}^*, \nabla f(x_n)\rangle + \alpha^2\|\nabla f(x_n)\|_2^2 \\
\leq\ & \left\|x_n - x_{n,p}^*\right\|^2 - 2\alpha\left(f(x_n) - f^* + \frac{\mu}{2}\left\|x_n - x_{n,p}^*\right\|_2^2\right) \\
& + 2\alpha^2 L\left(f(x_n) - f^*\right) \\
=\ & (1 - \mu\alpha)\left\|x_n - x_{n,p}^*\right\|_2^2 - 2\alpha(1 - L\alpha)(f(x_n) - f^*) \qquad (90) \\
=\ & \left(1 - \frac{1}{\kappa}\right) d(x_n, X^*)^2 \qquad \text{for } \alpha = \tfrac{1}{L}.
\end{aligned}$$

Note this proof is quite similar to the proof of Theorem 3.1 in Gower et al. (2019) applied directly to the deterministic case.

Next, we show a similar proof that follows the same idea but doesn't require ${}^*\mathrm{SC}^-(\mu)$.

Under ${}^*\mathrm{SC}^-(0)$, $\mathrm{RSI}^-(\mu)$ and $\mathrm{PL}^+(L)$ Assume $f \in {}^*\mathrm{SC}^-(0) \cap \mathrm{RSI}^-(\mu) \cap \mathrm{PL}^+(L)$. From star convexity, we have $\langle \nabla f(x), x - x_p \rangle \geq f(x) - f^*$, and from restricted secant inequality $\langle \nabla f(x), x - x_p \rangle \geq \mu \|x - x_p\|^2$. Combining the two, we obtain

$$\langle \nabla f(x), x - x_p \rangle \geq \frac{1}{2} \left(f(x) - f^* \right) + \frac{\mu}{2} \|x - x_p\|^2.$$

With similar argument as above, denote $x^*_{n,p}$ the projection of x_n on X^*, $\forall\, n \in \mathbb{N}$. Then,

$$\begin{aligned} d(x_{n+1}, X^*)^2 \leq \left\| x_{n+1} - x^*_{n,p} \right\|_2^2 &= \left\| x_n - x^*_{n,p} \right\|_2^2 - 2\alpha \langle x_n - x^*_{n,p}, \nabla f(x_n) \rangle + \alpha^2 \left\| \nabla f(x_n) \right\|_2^2 \\ &\leq \left\| x_n - x^*_{n,p} \right\|^2 - 2\alpha \left(\frac{1}{2} \left(f(x_n) - f^* \right) + \frac{\mu}{2} \left\| x_n - x^*_{n,p} \right\|_2^2 \right) \\ &\qquad + 2\alpha^2 L \left(f(x_n) - f^* \right) \\ &= (1 - \mu\alpha) \left\| x_n - x^*_{n,p} \right\|_2^2 - \alpha(1 - 2L\alpha)(f(x_n) - f^*)] \\ &= \left(1 - \frac{1}{2\kappa} \right) d(x_n, X^*)^2 \end{aligned} \tag{91}$$

for $\alpha = \frac{1}{2L}$.

Under $\mathrm{RSI}^-(\mu)$ and $\mathrm{EB}^+(L)$ Assume $f \in \mathrm{RSI}^-(\mu) \cap \mathrm{EB}^+(L)$, and for some $n \in \mathbb{N}$, x_p^* denotes the projection of x_n on X^*. Then

$$\begin{aligned} d(x_{n+1}, X^*)^2 \leq \left\| x_{n+1} - x_p^* \right\|_2^2 &= \left\| x_n - x_p^* \right\|_2^2 - 2\alpha \langle x_n - x_p^*, \nabla f(x_n) \rangle + \alpha^2 \left\| \nabla f(x_n) \right\|_2^2 \\ &\leq \left\| x_n - x_p^* \right\|^2 - 2\alpha\mu \left\| x_n - x_p^* \right\|_2^2 + \alpha^2 L^2 \left\| x_n - x_p^* \right\|_2^2 \\ &= \left(1 - 2\mu\alpha + L^2\alpha^2 \right) \left\| x_n - x_p^* \right\|_2^2 \\ &= \left(1 - \frac{1}{\kappa^2} \right) d(x_n, X^*)^2 \qquad \text{for } \alpha = \tfrac{\mu}{L^2}. \end{aligned} \tag{92}$$

Under $\mathrm{SC}^-(\mu)$ and $\mathrm{QG}^+(L)$ Assume $f \in \mathrm{SC}^-(\mu) \cap \mathrm{QG}^+(L)$ with some $L \geq \mu > 0$. Note that this implies that f has a unique minimum x^*.

Define $g(x) = \frac{1}{2}\|x - x^*\|_2^2 - \frac{1}{L}(f(x) - f^*)$; then, $g \in C^1(\mathbb{R}^d)$ and $g(x) \geq 0 = g^*$, since $f \in \mathrm{QG}^+(L)$, with $g(x^*) = 0$. Let X^* be the set of all minima of g, including the f minimizer x^*.

$g \in \mathrm{SC}^+\left(1 - \frac{1}{\kappa}\right)$ with $\kappa = \frac{\mu}{L}$: indeed, $\forall\, x, y \in \mathbb{R}^d$

$$\begin{aligned} & g(y) - g(x) - \langle \nabla g(x), y - x \rangle \\ = \frac{1}{2}\|y - x^*\|_2^2 - \frac{1}{L}(f(y) - f^*) - \frac{1}{2}\|x - x^*\|_2^2 + \frac{1}{L}(f(x) - f^*) - \langle (x - x^*) - \frac{1}{L}\nabla f(x), y - x \rangle \\ \leq -\frac{\mu}{2L}\|x - y\|_2^2 + \frac{1}{2}\left(\|y - x^*\|_2^2 - \|x - x^*\|_2^2 - 2\langle x - x^*, y - x \rangle \right) \\ \leq -\frac{\mu}{2L}\|x - y\|_2^2 + \frac{1}{2}\left(\|y - x^*\|_2^2 + \|x - x^*\|_2^2 - 2\langle x - x^*, y - x \rangle \right) \\ = \frac{1}{2}\left(1 - \frac{\mu}{L}\right)\|x - y\|_2^2 \end{aligned} \tag{93}$$

since $f \in \mathrm{SC}^-(\mu)$. This implies $g \in \mathrm{EB}^+\left(1 - \frac{1}{\kappa}\right)$:

$$\|\nabla g(x)\|_2 = \left\| (x - x^*) - \frac{1}{L}\nabla f(x) \right\|_2 \leq \left(1 - \frac{1}{\kappa}\right) d(x, X^*) \leq \left(1 - \frac{1}{\kappa}\right)\|x - x^*\|_2$$

Therefore, in the GD algorithm with step size $\alpha = \frac{1}{L}$, we get

$$\|x_{n+1} - x^*\|_2 = \left\| x_n - x^* - \frac{1}{L}\nabla f(x_n) \right\|_2 \leq \left(1 - \frac{1}{\kappa}\right)\|x_n - x^*\|_2 \tag{94}$$

Hence the linear rate $\left(1 - \frac{1}{\kappa}\right)^2$.

Rates of convergence for any pair of upper and lower condition. We collected all the above results in Table 2. For any pair of upper and lower condition $f \in \mathcal{C}^+(L) \cap \mathcal{C}^-(\mu)$, we define $\kappa = \frac{L}{\mu}$. We will justify here all the entries.

The rates in the first column ($f \in \mathrm{SC}^-(\mu)$) follows from the fact that if $f \in \mathrm{SC}^+(L)$, we recover the classical convergence rate for L-smooth and μ-strongly convex functions, while for any other upper condition $\mathcal{C}^+(L)$, we use the fact that $\mathcal{C}^+(L) \subseteq \mathrm{QG}^+(L)$ and we have convergence rate of $(1 - \frac{1}{\kappa})^2$.

In the first row ($f \in \mathrm{SC}^+(L)$), the rate of convergence $1 - \frac{1}{\kappa}$ holds for $f \in \mathrm{PL}^-(\mu)$ (as proven) and $f \in {}^*\mathrm{SC}^-(\mu)$ (since ${}^*\mathrm{SC}^-(\mu) \subset \mathrm{PL}^-(\mu)$); the rate of convergence $1 - \frac{1}{\kappa^2}$ instead holds for $f \in \mathrm{EB}^-(\mu)$ (since $\mathrm{EB}^-(\mu) \cap \mathrm{SC}^+(L) \subset \mathrm{PL}^-(\frac{\mu^2}{L})$) and consequently also for $f \in \mathrm{RSI}^-(\mu)$ (since $\mathrm{RSI}^-(\mu) \subset \mathrm{EB}^-(\mu)$).

We proved that for $f \in \mathrm{RSI}^-(\mu) \cap \mathrm{EB}^+(L)$ the GD algorithm converges with rate $1 - \frac{1}{\kappa^2}$; the same rate of convergence is also valid for $f \in {}^*\mathrm{SC}^-(\mu) \subset \mathrm{RSI}^-(\mu)$ and/or $f \in \mathrm{PL}^+(L) \subset \mathrm{EB}^+(L)$. This justifies entries $(2,4)$, $(3,2)$ and $(3,4)$ in Table 2.

For entry $(2,2)$, we proved a convergence rate of $1 - \frac{1}{\kappa}$ under assumption $f \in {}^*\mathrm{SC}^-(\mu) \cap \mathrm{PL}^+(L)$. We also completed the entry $(2,4)$ under star convexity. Since $\mathrm{SC}^+(L) \subset \mathrm{PL}^+(L)$, this rate also holds in $(1,4)$.

Similarly, under the additional assumption of star convexity, we have that $f \in \mathrm{QG}^-(\mu) \cap {}^*\mathrm{SC}(0) \subset \mathrm{RSI}^-(\frac{\mu}{2})$, therefore if $f \in \mathrm{QG}^-(\mu) \cap {}^*\mathrm{SC}(0) \cap \mathrm{EB}^+(L)$, GD converges with linear rate $1 - \frac{1}{4\kappa^2}$. Following the same argument, for $f \in \mathrm{QG}^-(\mu) \cap {}^*\mathrm{SC}(0)$ and upper conditions $f \in \mathrm{PL}^+(L)$ or $f \in \mathrm{SC}^+(L)$, GD converges with linear rate $1 - \frac{1}{4\kappa}$. Entries $(2,3)$, $(2,5)$, $(3,3)$ and $(3,5)$ follows from $\mathrm{PL}^-(\mu) \subset \mathrm{QG}^-(\mu)$ and $\mathrm{EB}^-(\mu) \cap \mathrm{QG}^+(L) \subset \mathrm{PL}^-(\frac{\mu^2}{L})$.

If we assume f to be convex, the rates of convergence on the fourth line ($f \in {}^*\mathrm{SC}^+(L)$) follow from the fact that ${}^*\mathrm{SC}^+(L) \cap \mathrm{SC}^-(0) \subset \mathrm{EB}^+(L)$. The rates on the last line ($f \in \mathrm{QG}^+(L)$) follow from $\mathrm{QG}^+(L) \cap \mathrm{SC}^-(0) \subset \mathrm{EB}^+(2L)$; similarly on the fifth line ($\mathrm{RSI}^+(L) \subset \mathrm{QG}^+(L)$).

Table 2: Linear rates for the GD algorithm for each pair of conditions, as function of $\kappa = \frac{L}{\mu}$. Rates marked with * hold under the additional assumption of star-convexity, while rates marked with † hold under the additional assumption of convexity. Rates are colored in green if corresponding to a continuous pair of conditions and red otherwise.

Rates of cv	$\mathrm{SC}^-(\mu)$	${}^*\mathrm{SC}^-(\mu)$	$\mathrm{PL}^-(\mu)$	$\mathrm{RSI}^-(\mu)$	$\mathrm{EB}^-(\mu)$	$\mathrm{QG}^-(\mu)$
$\mathrm{SC}^+(L)$	$\left(\frac{\kappa-1}{\kappa+1}\right)^2$	$1 - \frac{1}{\kappa}$	$1 - \frac{1}{\kappa}$	$1 - \frac{1}{\kappa^2}$ / $1 - \frac{1}{2\kappa}$ *	$1 - \frac{1}{\kappa^2}$	$1 - \frac{1}{4\kappa}$ *
$\mathrm{PL}^+(L)$	$\left(1 - \frac{1}{\kappa}\right)^2$	$1 - \frac{1}{\kappa}$	$1 - \frac{1}{4\kappa}$ *	$1 - \frac{1}{\kappa^2}$ / $1 - \frac{1}{2\kappa}$ *	$1 - \frac{1}{4\kappa^2}$ *	$1 - \frac{1}{4\kappa}$ *
$\mathrm{EB}^+(L)$	$\left(1 - \frac{1}{\kappa}\right)^2$	$1 - \frac{1}{\kappa^2}$	$1 - \frac{1}{4\kappa^2}$ *	$1 - \frac{1}{\kappa^2}$	$1 - \frac{1}{4\kappa^4}$ *	$1 - \frac{1}{4\kappa^2}$ *
${}^*\mathrm{SC}^+(L)$	$\left(1 - \frac{1}{\kappa}\right)^2$	$1 - \frac{1}{\kappa^2}$ †	$1 - \frac{1}{4\kappa^2}$ †	$1 - \frac{1}{\kappa^2}$ †	$1 - \frac{1}{4\kappa^4}$ †	$1 - \frac{1}{4\kappa^2}$ †
$\mathrm{RSI}^+(L)$	$\left(1 - \frac{1}{\kappa}\right)^2$	$1 - \frac{1}{4\kappa^2}$ †	$1 - \frac{1}{16\kappa^2}$ †	$1 - \frac{1}{4\kappa^2}$ †	$1 - \frac{1}{16\kappa^4}$ †	$1 - \frac{1}{16\kappa^2}$ †
$\mathrm{QG}^+(L)$	$\left(1 - \frac{1}{\kappa}\right)^2$	$1 - \frac{1}{4\kappa^2}$ †	$1 - \frac{1}{16\kappa^2}$ †	$1 - \frac{1}{4\kappa^2}$ †	$1 - \frac{1}{16\kappa^4}$ †	$1 - \frac{1}{16\kappa^2}$ †

As a last remark, we show that the additional assumption of f being convex (or star convex) is fundamental in some cases in order to obtain convergence of the GD algorithm. We will show here that the sole pair of conditions $\mathrm{SC}^+(L) \cap \mathrm{QG}^-(\mu)$ doesn't guarantee convergence of gradient descent.

Let $\varepsilon, \eta > 0$. Consider the following function $f \in C^1(\mathbb{R})$:

$$f(x) = \begin{cases} \frac{1}{2}x^2 & x < 1 \\ -\frac{1}{2\varepsilon}x^2 + \frac{1+\varepsilon}{\varepsilon}x - \frac{1+\varepsilon}{2\varepsilon} & 1 \leq x < 1+\varepsilon \\ \frac{1+\varepsilon}{2} & 1+\varepsilon \leq x < 1+\varepsilon+\eta \\ \frac{1}{2}x^2 - (1+\varepsilon+\eta)\,x + \frac{(1+\varepsilon+\eta)^2}{2} + \frac{1+\varepsilon}{2} & 1+\varepsilon+\eta \leq x \end{cases} \tag{95}$$

By inspecting its second derivative (where defined) we can conclude that $f \in \mathrm{SC}^+(1) \cap \mathrm{SC}^-\left(-\frac{1}{\varepsilon}\right)$.

Furthermore, $\frac{2f(x)}{x^2}$ reaches its minimum at $\bar{x} = \frac{(1+\varepsilon+\eta)^2+1+\varepsilon}{1+\varepsilon+\eta}$, with $\frac{2f(\bar{x})}{\bar{x}^2} = \frac{1+\varepsilon}{(1+\varepsilon+\eta)^2+1+\varepsilon} > 0$, therefore

$f \in \mathrm{QG}^- \left(\frac{1+\varepsilon}{(1+\varepsilon+\eta)^2+1+\varepsilon}\right)$.

On the other hand, $f'(x) = 0$ on $[1+\varepsilon, 1+\varepsilon+\eta]$, therefore if one of the iterates x_j of the GD algorithm falls into this interval, then $x_k \in [1+\varepsilon, 1+\varepsilon+\eta]\ \forall\, k \geq j$ and the algorithm fails to converge.

In the following, we will see sublinear convergence analysis under only upper conditions.

Under $^*\mathrm{SC}^-(0)$ and $\mathrm{SC}^+(L)$ This proof is a very classical one (Bansal and Gupta, 2017), and it is based on studying the monotonic properties of the Lyapunov function $V_n = n\left(f(x_n) - f^*\right) + \frac{1}{2\alpha} d(x_n, X^*)^2$. $\forall\, n \in \mathbb{N}$, let $x^*_{n,p}$ be the projection of x_n onto X^*.

$$
\begin{aligned}
V_{n+1} &= (n+1)\left(f(x_{n+1}) - f^*\right) + \frac{1}{2\alpha}\|x_{n+1} - x^*_{n,p}\|^2 \\
&\overset{\mathrm{SC}^+(L)}{\leq} (n+1)\left(f(x_n) - f^* + \left(\frac{L}{2}\alpha^2 - \alpha\right)\|\nabla f(x_n)\|^2\right) \\
&\quad + \frac{1}{2\alpha}\left(\|x_n - x^*_{n,p}\|^2 - 2\alpha\langle \nabla f(x_n), x_n - x^*_{n,p}\rangle + \alpha^2\|\nabla f(x_n)\|^2\right) \\
&= V_n + \left(f(x_n) - f^*\right) + \left((n+1)\left(\frac{L}{2}\alpha^2 - \alpha\right) + \frac{\alpha}{2}\right)\|\nabla f(x_n)\|^2 - \langle \nabla f(x_n), x_n - x^*_{n,p}\rangle \\
&\overset{^*\mathrm{SC}^-(0)}{\leq} V_n + \left((n+1)\left(\frac{L}{2}\alpha^2 - \alpha\right) + \frac{\alpha}{2}\right)\|\nabla f(x_n)\|^2 \\
&\leq V_n \qquad \text{for } \alpha = \tfrac{1}{L}
\end{aligned}
$$

Therefore, V_n is decreasing and in particular

$$
n(f(x_n) - f^*) \leq V_n \leq V_0 \leq \frac{L}{2} d(x_0, X^*)^2 \tag{96}
$$

Leading to the desired rate

$$
f(x_n) - f^* \leq \frac{L}{2n} d(x_0, X^*)^2 \tag{97}
$$

Under $^*\mathrm{SC}^-(0)$ and $\mathrm{PL}^+(L)$ $\forall\, n \in \mathbb{N}$, let $x^*_{n,p}$ be the projection of x_n onto X^*.

$$
\begin{aligned}
d(x_{n+1}, X^*)^2 \leq \|x_{n+1} - x^*_{n,p}\|^2 &= \|x_n - x^*_{n,p}\|^2 - 2\alpha\langle \nabla f(x_n), x_n - x^*_{n,p}\rangle + \alpha^2\|\nabla f(x_n)\|^2 \\
&\leq \|x_n - x^*_{n,p}\|^2 - 2\alpha(f(x_n) - f^*) + \alpha^2 \times 2L(f(x_n) - f^*)
\end{aligned} \tag{98}
$$

therefore,

$$
2\alpha(1 - L\alpha)(f(x_n) - f^*) \leq d(x_n, X^*)^2 - d(x_{n+1}, X^*)^2. \tag{99}
$$

By summing the inequality above for $k = 0, \ldots, n$, we have

$$
2\alpha(1 - L\alpha)\sum_{k=0}^{n}(f(x_k) - f^*) \leq d(x_0, X^*)^2 - d(x_{n+1}, X^*)^2 \leq d(x_0, X^*)^2 \tag{100}
$$

and taking $\alpha = \frac{1}{2L}$,

$$
\frac{1}{n+1}\sum_{k=0}^{n}(f(x_k) - f^*) \leq \frac{2L}{n+1} d(x_0, X^*)^2 \tag{101}
$$

we can conclude

$$\min_{k\in[|0,n|]}(f(x_k)-f^*)\leq\frac{2L}{n+1}d(x_0,X^*)^2 \tag{102}$$

If additionally $f\in\mathrm{SC}^-(0)$ (convex), we have the stronger result

$$f\left(\frac{1}{n+1}\sum_{k=0}^{n}x_k\right)-f^*\leq\frac{2L}{n+1}d(x_0,X^*)^2. \tag{103}$$

F Adaptive step size and application to logistic regression

Let $f\in C^1(\mathbb{R}^d)$ be a function to optimize, and let $g\in C^1\left([f^*,+\infty)\right)$ be an increasing function. It is easy to see that finding the minimum of $g\circ f$ is equivalent to finding the minimum of f, and a GD algorithm with constant step size on $g\circ f$ leads to a GD algorithm on f with adaptive step size:

$$x_{n+1}=x_n-\alpha\nabla\left(g\circ f\right)(x_n)\quad\Leftrightarrow\quad x_{n+1}=x_n-\alpha g'(f(x_n))\nabla f(x_n). \tag{104}$$

We briefly recall here the definition of the Θ notation, because it will be occasionally used in the following proposition and proof in order to preserve their readability.

Definition F.1 (Θ notation). Given two functions $f,g\in C^0(\mathbb{R})$, $g\geq 0$, we say that

$$f(x)\in\Theta\left(g(x)\right)\qquad\text{as } x\to x_0$$

if $\exists\,\delta,m,M>0$ such that $\forall\,x$ with $0<|x-x_0|<\delta$:

$$m\,g(x)\leq|f(x)|\leq M\,g(x). \tag{105}$$

Similarly, we say that

$$f(x)\in\Theta\left(g(x)\right)\qquad\text{as } x\to+\infty$$

if $\exists\,K,m,M>0$ such that $\forall\,x>K$:

$$m\,g(x)\leq|f(x)|\leq M\,g(x). \tag{106}$$

Proposition F.2. *Given $f\in C^1\left(\mathbb{R}^d\right)$, assume that*

$$f(x)-f^*\in\Theta\left(d(x,X^*)^\beta\right)\qquad\text{as } d(x,X^*)\to 0, \tag{107}$$
$$f(x)-f^*\in\Theta\left(d(x,X^*)^\gamma\right)\qquad\text{as } d(x,X^*)\to\infty, \tag{108}$$

for some $\beta,\gamma\in(0,\infty)$. Consider the functions

$$\begin{array}{cc} g:(-c,+\infty)\to\mathbb{R}_+ & h:[f^*,+\infty)\to\mathbb{R}_+\cup\{0\}\\ u\mapsto(u+c)^{\frac{\beta}{\gamma}} & t\mapsto(t-f^*)^{\frac{2}{\beta}}\end{array} \tag{109}$$

where $c>0$ is an arbitrary positive constant. Then, $g\circ h\circ f\in\mathrm{QG}^-(\mu)\cap\mathrm{QG}^+(L)$ for some $\mu,L>0$.

In the case $g\circ f$ is convex, we obtain a linear rate convergence from Table 1. This is easily satisfied when f is convex and $\beta,\gamma\in(0,2]$.

Remark F.3. This property leads to an adaptive step size $\tilde{\alpha}_n=\alpha g'(f(x_n))$ for the adaptive GD algorithm which requires the knowledge of the precise value of f^*. However, in the particular case where $\beta=2$ and $f^*>0$, we can take $c=f^*$ and obtain a step size $\tilde{\alpha}_n=\alpha\frac{2}{\gamma}f(x_n)^{\frac{2}{\gamma}-1}$.

Proof. $g\in C^1((-c,+\infty)))$ and $g(u)>0$ on its domain. It is easy to see that

$$g(u)-c^{\frac{\beta}{\gamma}}\in\Theta\left(u\right)\qquad\text{as } u\to 0 \tag{110}$$
$$g(u)-c^{\frac{\beta}{\gamma}}\in\Theta\left(u^{\frac{\beta}{\gamma}}\right)\qquad\text{as } u\to+\infty \tag{111}$$

Consider the function $h(f(x)) = (f(x) - f^*)^{\frac{2}{\beta}}$: clearly, $h \circ f$ is continuous on $\mathbb{R}^d$ (f is continuous) and $h(f(x)) = 0 \Leftrightarrow x \in X^*$. By continuity of all the functions involved, $\exists\, \delta, m_0, M_0 > 0$ such that

$$m_0 \leq \frac{g(h(f(x))) - g(h(f^*))}{h(f(x))} = \frac{\left((f(x) - f^*)^{\frac{2}{\beta}} + c\right)^{\frac{\beta}{\gamma}} - c^{\frac{\beta}{\gamma}}}{(f(x) - f^*)^{\frac{2}{\beta}}} \leq M_0 \qquad \text{for } 0 < d(x, X^*) < \delta \tag{112}$$

and using the fact that $f(x) - f^* \in \Theta\left(d(x, X^*)^\beta\right)$ as $d(x, X^*) \to 0$

$$\tilde{m}_0 \leq \frac{\left((f(x) - f^*)^{\frac{2}{\beta}} + c\right)^{\frac{\beta}{\gamma}} - c^{\frac{\beta}{\gamma}}}{d(x, X^*)^2} \leq \tilde{M}_0 \qquad \text{for } 0 < d(x, X^*) < \delta \tag{113}$$

i.e. $g(h(f(x))) - g(h(f^*)) \in \Theta\left(d(x, X^*)^2\right)$.

Similarly, $\exists\, K, m_\infty, M_\infty > 0$ such that

$$m_\infty \leq \frac{g(h(f(x))) - g(h(f^*))}{h(f(x))^{\frac{\beta}{\gamma}}} = \frac{\left((f(x) - f^*)^{\frac{2}{\beta}} + c\right)^{\frac{\beta}{\gamma}} - c^{\frac{\beta}{\gamma}}}{(f(x) - f^*)^{\frac{2}{\gamma}}} \leq M_\infty \qquad \text{for } d(x, X^*) > K \tag{114}$$

and using the fact that $f(x) - f^* \in \Theta\left(d(x, X^*)^\gamma\right)$ as $d(x, X^*) \to \infty$

$$\tilde{m}_\infty \leq \frac{\left((f(x) - f^*)^{\frac{2}{\beta}} + c\right)^{\frac{\beta}{\gamma}} - c^{\frac{\beta}{\gamma}}}{d(x, X^*)^2} \leq \tilde{M}_\infty \qquad \text{for } d(x, X^*) > K \tag{115}$$

i.e. $g(h(f(x))) - g(h(f^*)) \in \Theta\left(d(x, X^*)^2\right)$.

In conclusion, $\exists\, R > 0$, $\exists\, \mu_1, \mu_2, L_1, L_2 > 0$ such that

$$\mu_1 \leq \frac{g(h(f(x))) - g(h(f^*))}{d(x, X^*)^2} \leq L_1 \qquad \text{for } 0 < d(x, X^*) \leq R \tag{116}$$

$$\mu_2 \leq \frac{g(h(f(x))) - g(h(f^*))}{d(x, X^*)^2} \leq L_2 \qquad \text{for } d(x, X^*) > R \tag{117}$$

By setting $\mu = \min\{\mu_1, \mu_2\}$ and $L = \max\{L_1, L_2\}$, we have $g \circ h \circ f \in \mathrm{QG}^-(\mu) \cap \mathrm{QG}^+(L)$. □

Logistic regression: settings and notations Logistic regression is a common ML tool that is well studied and documented (see e.g. Bach (2013) and Bach and Moulines (2013)).

Given a distribution of data $X \sim \mathcal{D}$, and their class $Y \in \{-1, 1\}$, logistic regression aims at finding the maximum likelihood of the parametrized set of distributions verifying that $\ln \frac{\mathbb{P}[Y=1|X]}{1-\mathbb{P}[Y=1|X]}$ is linear in X. We call ω the associated coefficient.

$$\ln \frac{\mathbb{P}\left[Y = 1|X\right]}{1 - \mathbb{P}\left[Y = 1|X\right]} = \langle \omega, X \rangle \tag{118}$$

Note the bias can be included in ω by adding an additional dimension to X whose coordinate would always be 1. Eq.(118) is equivalent to

$$\mathbb{P}\left[Y = 1|X\right] = \sigma\left(\langle \omega, X \rangle\right) \tag{119}$$

with $\sigma(x) = \frac{1}{1+e^{-x}}$.

Then the likelihood of $Y|X$ is $\mathbb{P}\left[Y = 1|X\right]^{\mathbf{1}_{Y=1}} \mathbb{P}\left[Y = -1|X\right]^{\mathbf{1}_{Y=-1}}$. We aim at maximizing the log-likelihood (equivalently minimizing its opposite)

$$\begin{aligned} f(\omega) =& -\mathbb{E}\left[\mathbf{1}_{Y=1} \ln \sigma\left(\langle \omega, X \rangle\right) + \mathbf{1}_{Y=-1} \ln \sigma\left(-\langle \omega, X \rangle\right)\right] \\ =& -\mathbb{E}\left[\ln \sigma\left(Y \langle \omega, X \rangle\right)\right] \\ =& -\mathbb{E}_{Z \sim YX}\left[\ln \sigma\left(\langle \omega, Z \rangle\right)\right]. \end{aligned} \tag{120}$$

The function $f(\omega)$ satisfies:

$$f(\omega) = \mathbb{E}\left[-\ln \sigma\left(\langle \omega, Z\rangle\right)\right] \tag{121}$$

$$\nabla f(\omega) = \mathbb{E}\left[-(1-\sigma)\left(\langle \omega, Z\rangle\right) Z\right] \tag{122}$$

$$\nabla^2 f(\omega) = \mathbb{E}\left[\sigma(1-\sigma)\left(\langle \omega, Z\rangle\right) Z Z^\top\right] \tag{123}$$

Proposition F.4. *Under the following assumptions:*

$$\mathbb{P}\left[\langle \omega, Z\rangle > 0\right] > 0, \qquad \forall\, \omega \neq 0 \tag{124}$$

$$\mathbb{E}\left[\|Z\|_2^2\right] < \infty \tag{125}$$

the logistic regression function f is positive, smooth and (strictly) convex on $\mathbb{R}^d$; automatically, as described in Karimi et al. (2016), it is strongly convex on any compact $\mathcal{K} \subset \mathbb{R}^d$. Additionally, f grows linearly at infinity.

Note that the the hypothesis (125) is verified for discrete measure as in practice. The hypothesis (124) ensures there is enough disparity in the data.

Proof. By construction, $f(\omega)$ is the expectation of a positive variable, therefore $f(\omega) > 0\ \forall\, \omega \in \mathbb{R}^d$.

Let $r \in \mathbb{R}^d$ be a unit vector ($\|r\|_2 = 1$), then $\forall\, \omega \in \mathbb{R}^d$

$$\begin{aligned} r^\top \nabla^2 f(\omega) r &= \mathbb{E}\left[\sigma(1-\sigma)\left(\langle \omega, Z\rangle\right)\langle Z, r\rangle^2\right] \leq \mathbb{E}\left[\sigma(1-\sigma)\left(\langle \omega, Z\rangle\right)\|Z\|_2^2\right] \\ &\leq \mathbb{E}\left[\|Z\|_2^2\right] < \infty \end{aligned} \tag{126}$$

thanks to (125). Therefore, $\exists\, M > 0$ such that $M\,\mathrm{I}_d - \nabla^2 f(\omega)$ is positive semi-definite, i.e. f is smooth.

Additionally, $\forall\, r \in \mathbb{R}^d$ unit vector, $\forall\, \omega \in \mathbb{R}^d$

$$r^\top \nabla^2 f(\omega) r = \mathbb{E}\left[\sigma(1-\sigma)\left(\langle \omega, Z\rangle\right)\langle Z, r\rangle^2\right] > 0 \tag{127}$$

thanks to (124), i.e. f is strictly convex. Furthermore, for any $\mathcal{K} \subset \mathbb{R}^d$ compact, f is strongly convex on $\mathcal{K}$.

On the other hand, it is not strongly convex on the full space $\mathbb{R}^d$ and $f \notin \mathrm{QG}^-(\mu)$ for any $\mu \geq 0$, as it grows linearly in infinity: $f(\omega) \in \Theta(\|\omega\|_2)$, as $\|\omega\|_2 \to +\infty$.

Indeed, $\forall\, t \in \mathbb{R}$

$$\ln \sigma(t) = \ln\left(1 + e^{-t}\right) \in \left[\max\{0, -t\}, \ln(2) + \max\{0, -t\}\right],$$

therefore, $\forall\, \omega \in \mathbb{R}^d$, $\mathbb{E}\left[\max\{0, -\langle \omega, Z\rangle\}\right] \leq f(\omega) \leq \ln(2) + \mathbb{E}\left[\max\{0, -\langle \omega, Z\rangle\}\right]$.

On the one hand,

$$\begin{aligned} f(\omega) &\leq \ln(2) + \mathbb{E}\left[\max\{0, -\langle \omega, Z\rangle\}\right] \leq \ln(2) + \mathbb{E}\left[\|\omega\|_2 \|Z\|_2\right] \\ &\leq \ln(2) + \|\omega\|_2 \sqrt{\mathbb{E}\left[\|Z\|_2^2\right]} \leq \ln(2) + K_1 \|\omega\|_2 \end{aligned} \tag{128}$$

for some $K_1 > 0$, thanks to (125). On the other hand,

$$f(\omega) \geq \mathbb{E}\left[\max\{0, -\langle \omega, Z\rangle\}\right] \geq K_2 \|\omega\|_2 \tag{129}$$

where $K_2 = \min_{\|\omega\|_2 = 1} \mathbb{E}\left[\max\{0, -\langle \omega, Z\rangle\}\right]$.

It remains to prove that $K_2 > 0$. Note that the sphere $S^{d-1} \in \mathbb{R}^d$ is a compact set. Hence any continuous function defined on the sphere reaches its minimum and it is clear that $\omega \mapsto \mathbb{E}\left[\max\{0, -\langle \omega, Z\rangle\}\right]$ is Lipschitz continuous hence continuous. Then we only need to show that for any ω with norm 1, we have $\mathbb{E}\left[\max\{0, -\langle \omega, Z\rangle\}\right] > 0$.

We prove the latest by contradiction. Assume $\|\omega\|_2 = 1$ and $\mathbb{E}\left[\max\{0, -\langle \omega, Z\rangle\}\right] = 0$. Since the integrand is non negative, and the integral is 0, the integrand has to be 0 almost surely (i.e. with probability 1). We have $\mathbb{P}\left[-\langle \omega, Z\rangle \leq 0\right] = 1$, or again $\mathbb{P}\left[-\langle \omega, Z\rangle > 0\right] = 0$, which contradicts (124).

□

We conclude that the logistic regression is strongly convex and smooth on every compact set; therefore for any compact set $\mathcal{K} \subset \mathbb{R}^d$ and for any $x_0 \in \mathcal{K}$, one can fine-tune the GD algorithm starting in x_0 such that it converges linearly. However, the logistic regression is not strongly convex on the full space $\mathbb{R}^d$, and global uniform tuning of GD for linear convergence rate is not provided by classical studies of GD algorithm on strongly convex and smooth functions.

On the other hand, $f(\omega)$ verifies all the assumptions of Proposition F.2 with $\beta = 2$ and $\gamma = 1$ and f is convex. Therefore, we can have linear rate of convergence of GD algorithm on the function $g \circ f$ where $g(t) = (t - f^* + c)^2$, for any $c > 0$. In particular, since f is positive, we choose $c = f^*$: then, thanks to Proposition F.2, we have linear convergence rate of GD on the function $f^2(\omega) \in \mathrm{QG}^-(\mu) \cap \mathrm{QG}^-(L)$ for some $\mu, L > 0$ (see Table 1), and the exact knowledge of f^* is not required.

In summary, classical studies of GD with constant step size don't allow to find an optimal global (i.e. independent on the initialization x_0) step size α so that GD algorithm (linearly) converges on f. However, from the study above, we showed that a linear rate convergence can be achieved with an adaptive step size $\tilde{\alpha}_n = \alpha f(x_n)$ for well tuned α (according to the upper and lower properties of f), regardless of the initialization x_0.