LUND
UNIVERSITY

Department of
AUTOMATIC CONTROL

Exam in Optimization for Learning

2020-10-26

Points and grading

All answers must include a clear motivation. Answers should be given in English. The total number of points is 25. The maximum number of points is specified for each subproblem. Preliminary grading scales:

Grade 3: 12 points
4: 17 points
5: 22 points

Accepted aid

All material form the course.

Results

Solutions will be posted on the course webpage, and results will be registered in LADOK. Date and location for display of corrected exams will be posted on the course webpage.

1. Which of the following functions f are convex? Prove or disprove. (Use may use convexity perserving operations.)

a. $f(x) = \frac{1}{g(x)}$ where $g : \mathbb{R} \to \mathbb{R}$ is concave and satisfies $g(x) > 0$ for all $x \in \mathbb{R}$ (1 p)

b. $f(x) = \sqrt{x^T L^T L x}$ where $x \in \mathbb{R}^n$ and $L \in \mathbb{R}^{m \times n}$ (1 p)

c. $f(x) = \begin{cases} \sqrt{x_1 x_2} & \text{if } x_1 > 0,\ x_2 > 0 \\ \infty & \text{otherwise} \end{cases}$, where $x = (x_1, x_2) \in \mathbb{R}^2$ (1 p)

d. $f(x) = \sum_{i=1}^{r} x_{(i)} = x_{(1)} + \ldots + x_{(r)}$, where r is an integer such that $1 \leq r \leq n$ and for any vector $x \in \mathbb{R}^n$, we let $x_{(i)}$ denote the ith largest component of x, i.e.

$$x_{(1)} \geq \ldots \geq x_{(n)}$$

(1 p)

Solution

a. Convex. Let

$$h(u) = \frac{1}{u} + \iota_{\mathbb{R}_{++}}(u),$$

where $\mathbb{R}_{++} = \{u \in \mathbb{R} : u > 0\}$. The function f can be written as $f(x) = h(g(x))$. Since h is convex and nonincreasing, and g is concave, the composition rule gives that f is convex.

b. Convex. Note that

$$f(x) = \sqrt{(Lx)^T Lx} = \|Lx\|_2 .$$

Thus, the function f can be written as a composition between the convex function $h(u) = \|u\|_2$ and the affine mapping $g(x) = Lx$, i.e. $f(x) = h(g(x))$. The composition rule gives that f is convex.

c. Not convex. Let $x = (1, 1)$ and $y = (4, 1)$. Then $f(x) = 1$ and $f(y) = 2$. However, considering the convex combination

$$\frac{1}{2}x + \frac{1}{2}y = (2.5, 1),$$

gives

$$f\left(\frac{1}{2}x + \frac{1}{2}y\right) = \sqrt{2.5} > 1.5 = \frac{1}{2}f(x) + \frac{1}{2}f(y),$$

which violates the definition of convexity.

d. Convex. Note that the function f can be written as

$$f(x) = \max\left\{x_{i_1} + \ldots + x_{i_r} : \forall i_1, \ldots, i_r \in \mathbb{N},\, 1 \leq i_1 < \ldots < i_r \leq n\right\}.$$

Also, each function $a_{i_1,\ldots,i_r} : \mathbb{R}^n \to \mathbb{R}$, where

$$a_{i_1,\ldots,i_r}(x) = x_{i_1} + \ldots + x_{i_r},$$

is linear and therefore convex. We then see that f is given by a point-wise supremum of convex functions, and therefore itself convex.

2. Which of the following sets S are convex?

a. $S = \{x \in \mathbb{R}^n : x_1 + \ldots + x_n = 1\}$ (1 p)

b. $S = \{x \in \mathbb{R}^n : \|x - a\|_2 \leq \|x - b\|_2\}$, where $a \neq b$ and $a, b \in \mathbb{R}^n$ (1 p)

c. $S = \left\{x \in \mathbb{R}^3 : 2x_1 \geq \sqrt{x_2^2 + x_3^2}\right\}$ (1 p)

d. $S = \left\{x \in \mathbb{R}^2 : 2 \leq e^{x_1^2 + x_2^2} \leq 4\right\}$ (1 p)

e. $S = \left\{x \in \mathbb{R}^n : x^T y \leq 1, \forall y \in C\right\}$ where $C \subseteq \mathbb{R}^n$ (1 p)

Solution

a. Convex. The set

$$S = \left\{x \in \mathbb{R}^n : \mathbf{1}^T x = 1\right\},$$

defines a hyperplane in $\mathbb{R}^n$, which we know is convex.

b. Convex. Since norms are nonnegative, we have that

$$\begin{gathered}
\|x - a\|_2 \leq \|x - b\|_2 \\
\Leftrightarrow \\
\|x - a\|_2^2 \leq \|x - b\|_2^2 \\
\Leftrightarrow \\
(x - a)^T (x - a) \leq (x - b)^T (x - b) \\
\Leftrightarrow \\
2(b - a)^T x \leq \|b\|_2^2 - \|a\|_2^2
\end{gathered}$$

which defines a halfspace in $x \in \mathbb{R}^n$. Thus, the set S is convex.

c. Convex. The set S can be written as the zero:th sublevel set

$$S = \left\{x \in \mathbb{R}^3 : g(x) \leq 0\right\},$$

of the function

$$g(x) = \|Lx\|_2 - 2x_1,$$

where

$$L = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The first term $\|Lx\|_2$ of g is convex by **1.b.**. The second term $-2x_1$ is affine and therefore convex. Thus, the function g is convex. However, the set S is then a sublevel set of a convex function and therefore a convex set.

d. Not convex. The set S can be written as

$$S = \left\{ x \in \mathbb{R}^2 : \log 2 \leq x_1^2 + x_2^2 \leq \log 4 \right\}.$$

The set S is nonempty, e.g. $(0, \sqrt{\log 4}) \in S$. Consider any $x \in S$. Then $-x \in S$. However, the convex combination

$$\frac{1}{2}x + \frac{1}{2}(-x) = 0,$$

is not in S. Thus, the set S is not convex.

e. Convex. Note that the set S can be written as

$$S = \bigcap_{y \in C} \left\{ x \in \mathbb{R}^n : x^T y \leq 1 \right\},$$

i.e. an intersection of halfspaces, which we know is convex.

3. Consider the closed convex function $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$f(x) = \left(a^T x - b \right)^2, \quad \forall x \in \mathbb{R}^n,$$

where $a \in \mathbb{R}^n \setminus \{0\}$, $b \in \mathbb{R}$ and $n \geq 2$.

a. Prove or disprove that f is a strongly convex function (1 p)

b. Derive the conjugate function f^* of f (2 p)

c. Derive the proximal operator $\text{prox}_{\gamma f}$ of f, where $\gamma > 0$ (1 p)

Solution

a. The gradient of f is

$$\nabla f(x) = 2aa^T x - 2ab, \quad \forall x \in \mathbb{R}^n,$$

and the Hessian of f is

$$\nabla^2 f(x) = 2aa^T, \quad \forall x \in \mathbb{R}^n.$$

Since $a \in \mathbb{R}^n$ and $n \geq 2$, we know that there exists a vector $z \in \mathbb{R}^n \setminus \{0\}$ such that $a^T z = 0$. Note that

$$z^T \nabla^2 f(x) z = 2 \left(a^T z \right)^2 = 0,$$

i.e. the Hessian of f is not positive definite. However, this shows that f is not strongly convex by the second order condition for strong convexity.

b. For each $s \in \mathbb{R}^n$, decompose s such that

$$s = s^{\parallel} + s^{\perp},$$

where $s^{\parallel}$ is parallel to a, i.e. $s^{\parallel} = \alpha_s a$ for some $\alpha_s \in \mathbb{R}$, and $s^{\perp}$ is orthogonal to a, i.e. $a^T s^{\perp} = 0$. Similarly, for each $x \in \mathbb{R}^n$, decompose x such that

$$x = x^{\parallel} + x^{\perp},$$

where $x^{\parallel} = \alpha_x a$ for some $\alpha_x \in \mathbb{R}$, and $a^T x^{\perp} = 0$. Note that

$$\left(s^{\parallel}\right)^T x^{\perp} = \left(s^{\perp}\right)^T x^{\parallel} = 0.$$

The conjugate function of f can then be written as

$$\begin{aligned}
f^*(s) &= \sup_x \left(s^T x - f(x)\right) \\
&= \sup_x \left(s^T x - \left(a^T x - b\right)^2\right) \\
&= \sup_{x^{\parallel}, x^{\perp}} \left(\left(s^{\parallel} + s^{\perp}\right)^T \left(x^{\parallel} + x^{\perp}\right) - \left(a^T \left(x^{\parallel} + x^{\perp}\right) - b\right)^2\right) \\
&= \sup_{x^{\parallel}, x^{\perp}} \left(\left(s^{\parallel}\right)^T x^{\parallel} + \left(s^{\perp}\right)^T x^{\perp} - \left(a^T x^{\parallel} - b\right)^2\right) \\
&= \sup_{\alpha_x, x^{\perp}} \left(\alpha_x \alpha_s \|a\|_2^2 + \left(s^{\perp}\right)^T x^{\perp} - \left(\alpha_x \|a\|_2^2 - b\right)^2\right).
\end{aligned}$$

First, suppose that $s^{\perp} \neq 0$. Select $\alpha_x = 0$ and $x^{\perp} = t s^{\perp}$, and let $t \to \infty$ to conclude that $f^*(s) = \infty$.
Second, soppose that $s^{\perp} = 0$. We have that

$$\begin{aligned}
f^*(s) &= \sup_{\alpha_x} \left(\alpha_x \alpha_s \|a\|_2^2 - \left(\alpha_x \|a\|_2^2 - b\right)^2\right) \\
&= -\inf_{\alpha_x} \left(\underbrace{-\alpha_x \alpha_s \|a\|_2^2 + \left(\alpha_x \|a\|_2^2 - b\right)^2}_{=h(\alpha_x)}\right).
\end{aligned}$$

Note that $h : \mathbb{R} \to \mathbb{R}$ is convex and differentiable in α_x and therefore $\partial h(\alpha_x) = \{\nabla h(\alpha_x)\}$ for each $\alpha_x \in \mathbb{R}$. Fermat's rule gives that $\alpha_x^\star$ is a minimizer of h if and only if

$$0 = -\alpha_s \|a\|_2^2 + 2\|a\|_2^2 (\alpha_x^\star \|a\|_2^2 - b) \quad \Leftrightarrow \quad \alpha_x^\star = \frac{2b + \alpha_s}{2\|a\|_2^2}.$$

Therefore,

$$f^*(s) = -h(\alpha_x^\star) = \frac{(2b + \alpha_s)\alpha_s}{2} - \frac{\alpha_s^2}{4} = b\alpha_s + \frac{\alpha_s^2}{4}.$$

To summarize, we have that

$$f^*(s) = \begin{cases} b\alpha_s + \dfrac{\alpha_s^2}{4} & \text{if } s = \alpha_s a \text{ for some } \alpha_s \in \mathbb{R}, \\ \infty & \text{else.} \end{cases}$$

c. The proximal operator is given by

$$\text{prox}_{\gamma f}(z) = \underset{x\in\mathbb{R}^n}{\text{argmin}} \left(\left(a^T x - b\right)^2 + \frac{1}{2\gamma}\|x - z\|^2 \right), \quad \forall z \in \mathbb{R}^n.$$

Since the objective in the minimization problem above is a sum of differentiable convex functions with full domain, we can use Fermat's rule to find the minimizer x by

$$0 = 2a(a^T x - b) + \frac{1}{\gamma}(x - z)$$
$$\Leftrightarrow$$
$$0 = 2\gamma a a^T x - 2\gamma a b + x - z = \left(I + 2\gamma a a^T\right) x - 2\gamma a b - z.$$

Since $\left(I + 2\gamma a a^T\right)$ is invertible (it is symmetric with largest eigenvalue greater than or equal to 1), we get

$$\text{prox}_{\gamma f}(z) = \left(I + 2\gamma a a^T\right)^{-1} (z + 2\gamma a b), \quad \forall z \in \mathbb{R}^n.$$

4. Let $C \subseteq \mathbb{R}^n$ be a nonempty closed convex set. Its support function $\sigma_C : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is defined as

$$\sigma_C(y) = \sup_{x\in C} y^T x, \quad \forall y \in \mathbb{R}^n.$$

a. Show that the support function σ_C is convex, independent of the convexity of the set C (1 p)

b. Show that $\sigma_C^* = \iota_C$ (1 p)

c. Find an expression for $\text{prox}_{\gamma\sigma_C}$, where $\gamma > 0$, that involves Π_C, i.e. the projection onto the set C (1 p)

Solution

a. We have that

$$\sigma_C(y) = \sup_{x\in\mathbb{R}^n} \left(y^T x - \iota_C(x)\right) = \iota_C^*(y), \quad \forall y \in \mathbb{R}^n,$$

i.e. σ_C is equal to the conjugate function to the indicator function of C. Thus, σ_C is a convex function since conjugate functions are always convex.

Alternatively, note that σ_C is a points-wise supremum of convex functions (affine to be precise) and therefore itself a convex function.

b. We have that $\sigma_C^* = \iota_C^{**} = \iota_C$ since ι_C is a closed convex function.

c. Moreau decomposition gives that

$$\begin{aligned}\operatorname{prox}_{\gamma\sigma_C}(z) &= z - \gamma \operatorname{prox}_{\gamma^{-1}\sigma_C^\star}\left(\frac{z}{\gamma}\right) \\ &= z - \gamma \operatorname{prox}_{\gamma^{-1}\iota_C}\left(\frac{z}{\gamma}\right) \\ &= z - \gamma \Pi_C\left(\frac{z}{\gamma}\right), \quad \forall z \in \mathbb{R}^n.\end{aligned}$$

5. Consider the problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\, f(x) = \frac{1}{2}\|Ax - b\|_2^2,$$

where $A \in \mathbb{R}^{n\times n}$ satisfies

$$A = \mathbf{diag}(a_1, \ldots, a_n), \quad a_i \neq 0, \quad \forall i \in \{1, \ldots, n\},$$

and $b = (b_1, \ldots, b_n) \in \mathbb{R}^n$.

a. Give a closed-form expression for the solution (0.5 p)

b. Show that $\beta = \max_{i\in\{1,\ldots,n\}} a_i^2$ is a smoothness constant for f (1 p)

c. Show that $\beta_i = a_i^2$ is a coordinate-wise smoothness constants for f for each coordinate $i \in \{1, \ldots, n\}$ (1 p)

d. Consider the gradient method with step-size $1/\beta$, where β is the smoothness constant in **b.**. Suppose you are given the iterate $x^k \in \mathbb{R}^n$, where $k \in \mathbb{N}$ is the iteration number. For each coordinate $i \in \{1, \ldots, n\}$, provide the update formula for x_i^k. Utilize that $A = \mathbf{diag}(a_1, \ldots, a_n)$ (1 p)

e. Let $b_i = 0$ for each $i \in \{1, \ldots, n\}$ and provide an exact linear convergence rate for each of the coordinates for the gradient method in **d.**. This means, find the $\rho_i \in [0, 1)$ such that

$$\left\|x_i^{k+1}\right\|_2 = \rho_i \left\|x_i^k\right\|_2,$$

for each coordinate $i \in \{1, \ldots, n\}$ (Each coordinate will converge linearly to $x_i^\star = 0$ in this case) (1 p)

f. Consider the coordinate gradient method (i.e. no proximal operator) with step-sizes $1/\beta_i$, where β_i are the coordinate smoothness constants in **c.**. Provide an update formula for each coordinate $i \in \{1, \ldots, n\}$. Utilize that $A = \mathbf{diag}(a_1, \ldots, a_n)$. Show that $x_i^{k+1} = x_i^\star$ with $x_i^\star$ from **a.**, independent on $x^k \in \mathbb{R}^n$ (1 p)

Solution

a. The objective function f in the problem is convex and differentiable. Fermat's rule gives that $x^\star \in \mathbb{R}^n$ is a minimizer of the problem if and only if

$$\begin{gathered} 0 = \nabla f\left(x^\star\right) = A^T\left(Ax^\star - b\right) \\ \Leftrightarrow \\ x^\star = (A^TA)^{-1}(A^Tb) \\ \Leftrightarrow \\ x_i^\star = \frac{b_i}{a_i}, \quad \forall i \in \{1,\ldots,n\}, \end{gathered}$$

since $A^TA = \mathbf{diag}\left(a_1^2,\ldots,a_n^2\right)$ is invertible as $a_i \neq 0$ for each $i \in \{1,\ldots,n\}$.

b. Note that

$$\begin{aligned} \|\nabla f(x) - \nabla f(y)\|_2^2 &= \left\|A^T(Ax-b) - A^T(Ay-b)\right\|_2^2 \\ &= \left\|A^TA(x-y)\right\|_2^2 \\ &= \left\|\mathbf{diag}\left(a_1^2,\ldots,a_n^2\right)(x-y)\right\|_2^2 \\ &= \sum_{i=1}^n \left(a_i^2(x_i-y_i)\right)^2 \\ &= \sum_{i=1}^n a_i^4(x_i-y_i)^2 \\ &\leq \max_{i\in\{1,\ldots,n\}} a_i^4 \sum_{j=1}^n (x_j-y_j)^2 \\ &= \max_{i\in\{1,\ldots,n\}} a_i^4 \left\|x-y\right\|_2^2. \end{aligned}$$

Taking square root gives the result since

$$\sqrt{\max_{i\in\{1,\ldots,n\}} a_i^4} = \max_{i\in\{1,\ldots,n\}} a_i^2.$$

Alternative proof: Since f is convex and twice continuously differentiable, $\beta \geq 0$ is a smoothness constant if $\nabla^2 f(x) \preceq \beta I$ for each $x \in \mathbb{R}^n$. This is equivalent to that

$$y^T(\beta I - \nabla^2 f(x))y \geq 0, \quad \forall x, y \in \mathbb{R}^n.$$

Note that

$$\nabla^2 f(x) = A^TA = \mathbf{diag}\left(a_1^2,\ldots,a_n^2\right), \quad \forall x \in \mathbb{R}^n.$$

The condition on β reduces to

$$0 \leq y^T\mathbf{diag}\left(\beta - a_1^2,\ldots,\beta - a_n^2\right)y = \sum_{i=1}^n y_i^2(\beta - a_i^2), \quad \forall y \in \mathbb{R}^n.$$

which holds if $\beta = \max_{i\in\{1,\ldots,n\}} a_i^2$.

c. Consider any coordinate $i \in \{1, \ldots, n\}$. The coordinate-wise smoothness with parameter β_i can be written as

$$|\nabla f(x)_i - \nabla f(y)_i| \leq \beta_i |x_i - y_i|.$$

In our setting, we have $\nabla f(x)_i = a_i(a_i x_i - b_i)$ and

$$|a_i(a_i x_i - b_i) - a_i(a_i y_i - b_i)| = |a_i a_i (x_i - y_i)| \leq a_i^2 |x_i - y_i|.$$

So a_i^2 is a coordinate smoothness constant for coordinate i.

d. The gradient method with step-size $\gamma = 1/\beta$ reads

$$x^{k+1} = x^k - \frac{1}{\beta}\nabla f(x^k) = x^k - \frac{1}{\max_{j\in\{1,\ldots,n\}} a_j^2} A^T(Ax^k - b).$$

Since $A = \mathbf{diag}(a_1, \ldots, a_n)$, this reads as

$$x_i^{k+1} = x_i^k - \frac{1}{\max_{j\in\{1,\ldots,n\}} a_j^2} a_i(a_i x_i^k - b_i),$$

for each coordinate $i \in \{1, \ldots, n\}$.

e. For each coordinate $i \in \{1, \ldots, n\}$, we have

$$\begin{aligned} \left\| x_i^{k+1} \right\|_2 &= \left\| x_i^k - \frac{1}{\max_{j\in\{1,\ldots,n\}} a_j^2} a_i^2 x_i^k \right\|_2 \\ &= \underbrace{\left(1 - \frac{a_i^2}{\max_{j\in\{1,\ldots,n\}} a_j^2}\right)}_{=\rho_i} \left\| x_i^k \right\|_2, \end{aligned}$$

where $\rho_i \in [0, 1)$.

f. The coordinate gradient method when updating coordinate i is

$$x_i^{k+1} = x_i^k - \frac{1}{a_i^2} a_i \left(a_i x_i^k - b_i\right) = \frac{b_i}{a_i} = x_i^\star.$$

6. Consider minimizing a function $f : \mathbb{R}^n \to \mathbb{R}$, with minimizer $x^\star \in \mathbb{R}^n$, using a stochastic optimization algorithm, starting at some predetermined (deterministic) point $x_0 \in \mathbb{R}^n$. Analysis of the algorithm resulted in the following inequality

$$\mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2 \mid x_k\right] \leq \|x_k - x^\star\|_2^2 - 2\gamma(f(x_k) - f(x^\star)) + \gamma^2 G, \quad \forall k \in \mathbb{N},$$

where G is a deterministic positive constant and γ is a deterministic fixed positive step-size of the algorithm. In particular, $(x_k)_{k\in\mathbb{N}}$ is a stochastic process.

a. Apply an expectation to the above inequality to derive a Lyapunov inequality for the algorithm (1 p)

b. Use the obtained Lyapunov inequality to show that

$$\sum_{i=0}^{k} \mathbb{E}[f(x_i) - f(x^\star)] \leq \frac{\|x_0 - x^\star\|_2^2 + G(k+1)\gamma^2}{2\gamma}, \quad \forall k \in \mathbb{N} \tag{1}$$

(1.5 p)

c. The upper bound (1) diverges as $k \to \infty$ unless $G = 0$. Consider the step-size $\gamma = \theta/\sqrt{K+1}$, where $K \in \mathbb{N}$ is the total number of iterations we wish to run in the algorithm and $\theta > 0$. Show that we get a $\mathcal{O}(1/\sqrt{K+1})$ convergence bound

$$\min_{i \in \{0,\ldots,K\}} \mathbb{E}[f(x_i) - f(x^\star)] \leq \frac{\|x_0 - x^\star\|_2^2 + G\theta^2}{2\theta\sqrt{K+1}}$$

that is valid until iteration K (1 p)

Solution

a. We start from the inequality

$$\mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2 \mid x_k\right] \leq \|x_k - x^\star\|_2^2 - 2\gamma(f(x_k) - f(x^\star)) + \gamma^2 G, \quad \forall k \in \mathbb{N}.$$

By monotonicity and linearity of expectation, we get that

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2 \mid x_k\right]\right] &\leq \mathbb{E}\left[\|x_k - x^\star\|_2^2 - 2\gamma(f(x_k) - f(x^\star)) + \gamma^2 G\right] \\ &= \mathbb{E}\left[\|x_k - x^\star\|_2^2\right] - 2\gamma\mathbb{E}\left[f(x_k) - f(x^\star)\right] + \gamma^2 G,\end{aligned}$$

holds for each $k \in \mathbb{N}$. The law of total expectation yields

$$\mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2\right] \leq \mathbb{E}\left[\|x_k - x^\star\|_2^2\right] - 2\gamma\mathbb{E}\left[f(x_k) - f(x^\star)\right] + \gamma^2 G, \quad \forall k \in \mathbb{N}.$$

This is the Lyapunov inequality we pick.

b. Recursively applying the Lyapunov inequality above gives

$$\begin{aligned}\mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2\right] &\leq \mathbb{E}\left[\|x_0 - x^\star\|_2^2\right] - 2\gamma\sum_{i=0}^{k}\mathbb{E}\left[f(x_i) - f(x^\star)\right] + G\gamma^2(k+1) \\ &= \|x_0 - x^\star\|_2^2 - 2\gamma\sum_{i=0}^{k}\mathbb{E}\left[f(x_i) - f(x^\star)\right] + G\gamma^2(k+1), \quad \forall k \in \mathbb{N},\end{aligned}$$

since $\|x_0 - x^\star\|_2^2$ is deterministic. Again, by monotonicity of expectation, we know that

$$0 \leq \mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2\right], \quad \forall k \in \mathbb{N},$$

since

$$0 \leq \|x_{k+1} - x^\star\|_2^2, \quad \forall k \in \mathbb{N}.$$

We conclude that

$$0 \leq \|x_0 - x^\star\|_2^2 - 2\gamma \sum_{i=0}^{k} \mathbb{E}\left[f(x_i) - f(x^\star)\right] + G\gamma^2(k+1), \quad \forall k \in \mathbb{N},$$

or by rearranging

$$\sum_{i=0}^{k} \mathbb{E}[f(x_i) - f(x^\star)] \leq \frac{\|x_0 - x^\star\|_2^2 + G\gamma^2(k+1)}{2\gamma}, \quad \forall k \in \mathbb{N}, \tag{2}$$

as desired.

c. We first note that

$$(K+1) \min_{i \in \{0,\ldots,K\}} \mathbb{E}[f(x_i) - f(x^\star)] \leq \sum_{i=0}^{K} \mathbb{E}[f(x_i) - f(x^\star)].$$

Using this in the bound in **b**. with $\gamma = \theta/\sqrt{K+1}$ gives

$$\begin{aligned} \min_{i \in \{0,\ldots,K\}} \mathbb{E}[f(x_i) - f(x^\star)] &\leq \frac{\|x_0 - x^\star\|_2^2 + G(K+1)\gamma^2}{2\gamma(K+1)} \\ &= \frac{\|x_0 - x^\star\|_2^2 + G\theta^2}{2\theta\sqrt{K+1}}, \end{aligned}$$

as desired.