Continuous Optimization: Lecture 2 Exercises

March 28, 2026

Some problems borrowed from courses of Daniel Dadush and Lap Chi Lau.

1. Recall the definitions of α-strong convexity and β-smoothness: for all x, y

$$\frac{\alpha}{2}\|y-x\|_2^2 \leq f(y) - f(x) - \langle \nabla f(x), y-x\rangle \leq \frac{\beta}{2}\|y-x\|_2^2.$$

We will examine equivalent 1-st and 2-nd order definitions of these properties. For this question, assume f is twice continuously differentiable.

1. Show if $f : \mathbb{R}^n \to \mathbb{R}$ is α-strongly convex or β-smooth, then the restriction $g(t) := f(x+tv)$ is $\alpha\|v\|_2^2$-strongly convex and $\beta\|v\|_2^2$-smooth for any $x, v \in \mathbb{R}^n$;
2. Show α-strong convexity and β-smoothness are equivalent to the conditions $\alpha I \preceq \nabla^2 f(x)$ and $\nabla^2 f(x) \preceq \beta I$, respectively. (Here the notation $A \succeq B$ means $A - B \succeq 0$ or equivalently $\forall v : \langle v, Av\rangle \geq \langle v, Bv\rangle$)
3. Show for an α-strongly convex function, we have distance bound on the optimizer

$$\|x - x^*\|_2 \lesssim \frac{\|\nabla f(x)\|_2}{\alpha}.$$

4. Show β-smoothness is equivalent to Lipschitz-ness of the *gradient*:

$$\forall x, y : \quad \|\nabla f(y) - \nabla f(x)\|_2 \leq \beta\|y-x\|_2.$$

(You can use the following fact without proof: $0 \preceq H \preceq \beta I \iff \forall x : \|Hx\|_2 \leq \beta\|x\|_2 \iff \forall x, y : |\langle x, Hy\rangle| \leq \beta\|x\|_2\|y\|_2$)

Solution:

1. Follows by verifying definitions

$$\begin{aligned} g(t) - g(s) - g'(s)(t-s) &= f(x+tv) - f(x+sv) - \langle \nabla f(x+sv), (t-s)v\rangle \\ &\geq \frac{\alpha}{2}\|(t-s)v\|_2^2 = \frac{\alpha\|v\|_2^2}{2}|t-s|^2, \end{aligned}$$

where the first step was by definition of g, and in the second we used α-strong convexity of f. The last expression is exactly the definition of $\alpha\|v\|_2^2$-strong convexity of g. The proof for smoothness is the same with reversed inequalities.

2. We first show that the 2-nd order definition of strong convexity implies the 1-st order definition: by fundamental theorem of calculus, we have

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x \rangle &= \int_{t=0}^{1} \langle \nabla f(x + t(y-x)) - \nabla f(x), y - x \rangle \\ &= \int_{t=0}^{1} \langle \int_{s=0}^{t} \nabla^2 f(x + t(y-x))(y-x), (y-x) \rangle. \\ &\geq \int_{t=0}^{1} \int_{s=0}^{t} \alpha \|y - x\|_2^2 = \frac{\alpha}{2} \|y - x\|_2^2, \end{aligned}$$

From here we see th where we used fundamental theorem of calculus in the first two steps, and in the third we used $\nabla^2 f \succeq \alpha I$ for the lower bound.

Converse: we use Taylor's theorem for f twice-continuously differentiable:

$$\begin{aligned} \langle v, \nabla^2 f(x) v \rangle &= \partial_{t=0}^2 f(x + tv) \\ &= 2 \lim_{t \to 0} \frac{f(x+tv) - f(x) - \langle \nabla f(x+tv), tv \rangle}{t^2} \\ &\geq \lim_{t \to 0} \frac{\alpha \|tv\|_2^2}{t^2} = \alpha \|v\|_2^2, \end{aligned}$$

where the second step is by Taylor's theorem and in the third step we used the first-order definition of strong convexity. Since this is true for every x, v we have $\nabla^2 f \succeq \alpha I$. The proofs for smoothness are the same and we omit them.

3. We claim that for any x, the sub-level set $L := \{y \mid f(y) \leq f(x)\}$ is contained in the ball $B(x, R)$ with $R \leq 2\|\nabla f(x)\|_2 / \alpha$. This gives the result as $f(x^*) \leq f(x)$. We show the contrapositive of our claim, so for $\|y - x\|_2 > R$ we have

$$\begin{aligned} f(y) - f(x) &\geq \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2} \|y - x\|_2^2 \\ &\geq -\|\nabla f(x)\|_2 \|y - x\|_2 + \frac{\alpha}{2} \|y - x\|_2^2 \\ &\geq \inf_{t > R} -\|\nabla f(x)\|_2 t + \frac{\alpha}{2} t^2 > 0 \end{aligned}$$

where the first step was by strong convexity, in the second we applied Cauchy-Schwarz, and in the final step we used that $\|y - x\|_2 > R$. Therefore $y \notin L$, which shows our required distance bound.

4. By the above part, we see that smoothness gives upper bounds on the second derivative, which exactly control the *change* in the first derivative:

$$\begin{aligned}\|\nabla f(y) - \nabla f(x)\|_2 &= \left\| \int_0^1 \nabla^2 f(x + t(y-x))(y-x) \right\|_2 \\ &\leq \int_0^1 \|\nabla^2 f(x + t(y-x))(y-x)\|_2 \\ &\leq \int_0^1 \beta \|y - x\|_2 = \beta \|y - x\|_2,\end{aligned}$$

where the first step was by fundamental theorem of calculus, the second was by triangle inequality for $\|\cdot\|_2$, and in the third step we used the stated fact $\nabla^2 f(z) \preceq \beta I \implies \|\nabla^2 f(z)(y-x)\|_2 \leq \beta \|y - x\|_2$.

For the converse we again take limits:

$$\begin{aligned}\langle v, \nabla^2 f(x) v \rangle = \partial^2_{t=0} f(x + tv) &= \lim_{t \to 0} \frac{\partial_{s=t} f(x + sv) - \partial_{s=0} f(x + sv)}{t} \\ &= \lim_{t \to 0} \frac{\langle v, \nabla f(x + tv) - \nabla f(x) \rangle}{t} \leq \lim_{t \to 0} \frac{\beta \|v\|_2 \|tv\|_2}{t} = \beta \|v\|_2^2,\end{aligned}$$

where the first step was by definition of the Hessian, in the second we used calculus to write the second derivative as a limit of first derivatives, in the third step we used definition of gradient to rewrite the first derivatives, and in the fourth step we used the Lipschitz bound for gradients along with Cauchy-Schwarz. Since this is true for all x, v we have shown the second-derivative condition which is equivalent to smoothness by the previous part.

2. In the analysis of gradient descent for strongly convex and smooth functions we used the following crucial claim, known as the Polyak-Lojaseiwicz inequality (PL):

$$\|\nabla f(x)\|_2^2 \geq 2\alpha (f(x) - f(x^*)),$$

where x^* is the optimizer.

1. Show that α-strong convexity implies this α-PL condition.
2. Show that the analysis goes through if we assume only the PL condition as above, instead of strong convexity.
3. Find a function that satisfies $\alpha = 1$ PL condition but is not $\alpha = 1$-strongly convex. (Hint: consider convex quadratics)

Solution:

1. If we have α-strong convexity, then

$$f(x^*) - f(x) \geq \langle \nabla f(x), x^* - x \rangle + \frac{\alpha}{2}\|x^* - x\|_2^2$$
$$\geq \min_r -\|\nabla f(x)\|_2 r + \frac{\alpha}{2} r^2 = -\frac{\|\nabla f(x)\|_2^2}{2\alpha},$$

   where the first step is by strong convexity, in the second we used Cauchy-Schwarz and change of variable $r := \|x^* - x\|_2$, and the last step was by optimizing the quadratic. Rearranging gives the result.

2. This is the only consequence of strong convexity that is used in the analysis of gradient descent. We omit the details.

3. Let $f(x, y) := \frac{1}{2}x^2$, i.e. f is the norm of the projection onto the first coordinate. The optimizer is clearly $(0, y)$ for any $y \in \mathbb{R}$ with value $f^* = 0$. We can calculate

$$\nabla f(x, y) = (x, 0), \qquad \langle (u, v), \nabla^2 f(x, y)(u, v) \rangle = u^2.$$

   By the previous question we see that α-strong convexity is equivalent to $\nabla^2 f \succeq \alpha I$. But $\langle (0, v), \nabla^2 f(0, v) \rangle = 0$ so f is convex but not α-strongly convex for any $\alpha > 0$. On the other hand,

$$\|\nabla f(x, y)\|_2^2 = x^2 = 2(f(x, y) - f^*),$$

   verifying the 1-PL condition.

3. (Q4 Ex4 from Daniel Dadush Course 2022) Assume $\mathsf{A}\mathsf{A}^\mathsf{T}$ is invertible. Give an explicit formula for $\nu^* := \max g(\nu) := \nu^\mathsf{T}\mathbf{b} - \|\mathsf{A}^\mathsf{T}\nu\|_2^2/2$ in terms of $\mathsf{A}\mathsf{A}^\mathsf{T}$ and $\mathbf{b}$. Prove that $\mathsf{A}^\mathsf{T}\nu^*$ is the optimal solution to the primal least-squares problem $\min \|\mathbf{x}\|_2^2/2, \mathsf{A}\mathbf{x} = \mathbf{b}$.

Solution: g is concave since $\nu^\mathsf{T}\mathbf{b}$ is linear and $\|\mathsf{A}^\mathsf{T}\nu\|_2^2$ is convex. Therefore any critical point is a global optimum, so we can compute

$$\nabla g(\nu) = \mathbf{b} - \mathsf{A}\mathsf{A}^\mathsf{T}\nu$$

$$\mathbf{0} = \nabla g(\nu) \implies \mathbf{b} = \mathsf{A}\mathsf{A}^\mathsf{T}\nu^* \implies \nu^* = (\mathsf{A}\mathsf{A}^\mathsf{T})^{-1}\mathbf{b},$$

where in the last step we used that $\mathsf{A}\mathsf{A}^\mathsf{T}$ is invertible. Note that $g(\nu^*) = ((\mathsf{A}\mathsf{A}^\mathsf{T})^{-1}\mathbf{b})^\mathsf{T}\mathbf{b} - \|\mathsf{A}^\mathsf{T}(\mathsf{A}\mathsf{A}^\mathsf{T})^{-1}\mathbf{b}\|_2^2/2 = \mathbf{b}(\mathsf{A}\mathsf{A}^\mathsf{T})^{-1}\mathbf{b}/2$.

To prove that $\mathsf{A}^\mathsf{T}\nu^*$ is the optimal solution to the primal program, first note that $\mathsf{A}(\mathsf{A}^\mathsf{T}\nu^*) = \mathsf{A}\mathsf{A}^\mathsf{T}(\mathsf{A}\mathsf{A}^\mathsf{T})^1\mathbf{b}$, thus the solution is feasible. Furthermore, $\|\mathsf{A}^\mathsf{T}\nu^*\|_2^2/2 =$

$\mathbf{b}^\mathsf{T}(\mathsf{A}\mathsf{A}^\mathsf{T})^{-1}\mathbf{b}/2 = g(\nu^*)$. Since $\mathsf{A}^\mathsf{T}\nu^*$ is a primal solution whose value is the same as $g(\nu^*)$, the Lagrangian dual function evaluated at ν^*, we get that $\mathsf{A}^\mathsf{T}\nu^*$ is an optimal primal solution by weak duality.

4. Recall our characterization of quadratic convex functions from last homework:

$$\langle x, Hx\rangle + \langle b, x\rangle + c$$

is convex iff $H \succeq 0$, i.e. H is symmetric and $\forall v \in \mathbb{R}^n : \langle v, Hv\rangle \geq 0$ (strictly convex iff $H \succ 0$, i.e. the inequality is strict for $v \neq 0$). In this question we will consider partitioned matrices

$$q(x, y) := \left\langle \begin{pmatrix} x \\ y \end{pmatrix}, \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \right\rangle,$$

where A, C are symmetric.

1. Consider $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be a convex function of variables $x \in \mathbb{R}^n, y \in \mathbb{R}^m$. Show $g(x) := \inf_y f(x, y)$ is also convex.
2. Let $q(x, y)$ be the quadratic defined above and assume $C \succ 0$. Show

$$\inf_y q(x, y) = \langle x, (A - BC^{-1}B^T)x\rangle.$$

   (Hint: use first order optimality conditions for y).
3. Show q is strictly convex iff

$$H := \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \succ 0$$

   iff $C \succ 0$ and $A - BC^{-1}B^T$.
4. (Not a question, just a note): This last expression $A - BC^{-1}B^T$ is known as the *Schur complement* of H onto the x variables. It can be seen as the result of Gaussian elimination of the y variables and is extremely useful throughout math and computer science.

Solution:

1. For now we assume the infimum is always attained. The general case follows by taking a limit. So consider x_0, x_1 and $y_b := \arg\min_y f(x_b, y)$ for $b \in \{0, 1\}$ so $f(x_b, y_b) = g(x_b)$. Then we can verify the 0-th order definition of convexity:

$$\begin{aligned} g((1-\lambda)x_0 + \lambda x_1) &= \inf_y f((1-\lambda)x_0 + \lambda x_1, y) \\ &\leq f((1-\lambda)x_0 + \lambda x_1, (1-\lambda)y_0 + \lambda y_1) \\ &\leq (1-\lambda)f(x_0, y_0) + \lambda f(x_1, y_1) \\ &= (1-\lambda)g(x_0) + \lambda g(x_1), \end{aligned}$$

where the first step was by definition of g, in the third we used convexity of f, and the last step was by definition of y_0, y_1.

2. We can open up the matrix notation first:

$$q(x,y) = \langle x, Ax\rangle + 2\langle x, By\rangle + \langle y, Cy\rangle.$$

Since $C \succ 0$ this is a strictly convex quadratic function in y, so we can find the infimum by optimality conditions

$$\nabla_y q(x,y) = 2B^T x + 2Cy = 0 \iff y = -C^{-1}B^T x;$$

$$\inf_y q(x,y) = q(x, -C^{-1}B^T x) = \langle x, Ax\rangle - 2\langle x, BC^{-1}B^T x\rangle + \langle B^T x, C^{-1}B^T x\rangle.$$

5. We use the following lemma: for convex ℓ and convex differentiable r, let $f := \ell + r$ with $x_* = \arg\min f$. Then

$$f(x) = \ell(x) + r(x) \geq f(x_*) + D_r(x, x_*) = \ell(x_*) + r(x_*) + D_r(x, x_*).$$

1. Let $f(z) := D_r(z, x)$, then $D_f = D_r$, i.e.

$$D_f(z,y) = D_r(z,x) - D_r(y,x) - \langle \nabla r(y) - \nabla r(x), z - y\rangle = D_r(z,y).$$

2. Prove the Bregman Pythagorean result: for strictly convex $r : \mathbb{R}^n \to \mathbb{R}$, let $\Pi_K(x) := \arg\min_{z\in K} D_r(z,x)$, then

$$\forall z \in K : \quad D_r(z,x) \geq D_r(\Pi(x), x) + D_r(z, \Pi(x)).$$

Solution:

1. Note $f(z) = r(z) - (r(x) + \langle \nabla r(x), z - x\rangle)$, so $f - r$ is an affine function, which implies $D_f = D_r$ since the Bregman divergence of an affine function is 0 and Bregman divergence is additive.

This can also be proven by direct calculation.

2. Using the lemma we set $\ell(z) := \delta_K(z)$ and $r'(z) := D_r(z \mid x)$, then

$$\Pi_K(x) = \arg\min_z D_r(z \mid x) \implies D_r(z \mid x) \geq D_r(\Pi(x) \mid x) + D_{r'}(z \mid \Pi(x)),$$

and the result follows from the previous part as $D_{r'} = D_r$.

6. (Lap Chi Lau Course CS798, HW2 Q1) We analyze the multiplicative weights method for solving the following linear program:

$$\min \sum_{i\in[n]} x_i \qquad \text{s.t.} \qquad Ax \geq \vec{1}, \quad x \geq 0,$$

where $A \in \{0,1\}^{m\times n}$, i.e. every entry of A is either 0 or 1. Design an algorithm to solving this linear program using the multiplicative weights update method (including the design of the oracle), and analyze its total running time.

Solution: Our plan is to make each row constraint of A into an expert and apply the multiplicative weights algorithm to maintain a distribution over experts. At each step, we will be required to construct an oracle that outputs a feasible solution to the following simpler problem:

$$\exists x \geq 0 \quad \text{s.t.} \langle p, Ax\rangle \geq 1,$$

where $p \in \Delta_m$ is some arbitrary probability distribution over rows. Since the constraints are $Ax \geq 1$, multiplicative and additive error is equivalent, so we use the following refined form the multiplicative weights guarantee:

$$\sum_t \ell_t(p_t) \leq \min_{j\in[m]} \quad (1+\eta)\sum_t \ell_t(e_j) + \frac{b\log m}{\eta}.$$

Here η is the learning rate and the feedback $\ell_t(p) := \langle g^t, p\rangle$ satisfies $g^t \in [0,b]^m$. Specifically in our case $\ell_t(p) := \langle p, Ax^t\rangle$ where x^t is chosen as follows

$$x^t := \arg\min\{\langle 1_n, x\rangle \mid x \geq 0, \langle p^t, Ax\rangle = 1\}.$$

We claim $\langle 1_n, x^t\rangle \leq n$, otherwise the system is infeasible. Indeed the solution has a simple explicit form: let $i_t := \arg\max_{i\in[n]}\langle p^t, Ae_i\rangle = (A^T p^t)_i$, then it is simply verified that $x^t = \frac{e_{i_t}}{\langle p^t, Ae_{i_t}\rangle}$. If $\langle p^t, Ae_{i_t}\rangle < 1/n$ then

$$\frac{1}{n} > \max_{i\in[n]}\langle p^t, Ae_i\rangle \geq \frac{1}{n}\sum_i \langle p^t, Ae_i\rangle = \frac{\langle p^t, A1_n\rangle}{n} \geq \frac{\min_{j\in[m]}(A1_n)_j}{n},$$

which rearranging gives $\min_{j\in[m]}(A1_n)_j < 1$. But since $A \in \{0,1\}^{m\times n}$, this implies some row of A is all zero, so the system is clearly infeasible.

Given the claim, this implies the following width bound:

$$b := \max_{j\in[m]} \ell_t(e_j) = \max_{j\in[m]}\langle e_j, Ax^t\rangle = \max_{j\in[m]} \frac{A_{i_t j}}{\langle p^t, Ae_{i_t}\rangle} \leq n,$$

where in the last step we used $A_{ij} \in \{0,1\}$ and the denominator is $\geq 1/n$. Plugging this into our MW guarantee gives

$$(1+\eta)\min_{j\in[m]}\sum_{t\in[T]}\langle e_j, Ax^t\rangle \geq \sum_{t\in[T]}\langle p^t, Ax^t\rangle + \frac{n\log m}{\eta} \geq T - \frac{n\log m}{\eta},$$

where in the last step we used that $\langle p^t, Ax^t \rangle = 1$ by definition of our oracle. Finally, choosing $\eta = \varepsilon, T = n \log m/\varepsilon^2$ and letting $\overline{x} := \frac{1}{T} \sum_{t \in [T]} x_t$ gives

$$\min_{j \in [m]} (A\overline{x})_j \geq \frac{(1+\eta)^{-1}}{T} \left(T - \frac{n \log m}{\eta} \right) = \frac{1-\varepsilon}{1+\varepsilon},$$

where in the last step we substituted $\eta = \varepsilon, T = n \log m/\varepsilon^2$. Also, if x^* is the optimum solution to the LP, we can compare

$$\langle 1_n, \overline{x} \rangle = \frac{1}{T} \sum_{t \in [T]} \langle 1_n, x^t \rangle \leq \langle 1_n, x^* \rangle$$

where the last step was by our oracle definition for x^t. Therefore in $T = n \log m/\varepsilon^2$ iterations we have an $O(\varepsilon)$-approximately feasible solution with optimal objective value. We can also instead output $\tilde{x} := \frac{1+\varepsilon}{1-\varepsilon} \overline{x}$ which is feasible and $1 + O(\varepsilon)$-optimal.

7. In our proof of the lower bound for first-order methods for convex Lipschitz functions we made the following technical assumption: the initial point $x_0 = 0$ and the queries $x_t \in \text{span}\{x_0, g_0, ..., g_{t-1}\}$. In this question we show how to design an *adaptive* adversarial gradient oracle to remove this assumption.

   1. Show that we can remove the assumption $x_0 = 0$ by shifting the function based on the initial query.
   2. Show that for each query sequence $x_0, ...x_T$ with $T < n$ we can design a function of the form
   $$f_T(x) = \max_{i \in [n]} \langle g_i, x - x_0 \rangle$$
   such that $f(x_t) \leq 0$ for all $t \in [T]$.
   3. Show that we can choose functions $\{f_0, ..., f_T\}$ as above in a *consistent* way:
   $$\forall s \leq t < n : f_t(x_s) = f_s(x_s).$$
   Conclude that we can remove the technical assumption about queries.

Solution: As an adaptive adversary, we can choose our function to be of the form

$$f(x) := \max_{i \in [n]} \langle g_i, x \rangle$$

where g_i can be chosen as an adversarial response to the query sequence. To define these responses, let

$$X_s := \text{span}\{x_0, ..., x_s\},$$

and let P_s be the orthogonal projector onto X_s. Then in each step we sill define

$$g_t := \frac{(I - P_{t-1})x_t}{\|(I - P_{t-1})x_t\|_2}$$

if the denominator is non-zero, and otherwise we let g_t be an arbitrary unit vector in $X_t^\perp$. By this definition, note that we maintain the invariant

$$\forall t: \quad x_t \in \text{span}\{g_0, ..., g_t\}.$$

Now to define our oracle, let

$$i_t := \arg\max_{i \in [t]} \langle g_i, x_t \rangle.$$

Note $\langle g_{i_t}, x_t \rangle \geq \langle g_t, x_t \rangle = \|(I - P_{t-1})x_t\|_2 \geq 0$. Since every $g_{s>t} \in X_t^\perp$, we have $\langle g_{s>t}, x_t \rangle = 0$. Therefore we can choose function

$$f(x_t) = \max_{i \in [n]} \langle g_i, x_t \rangle = \langle g_{i_t}, x_t \rangle$$

and g_{i_t} is a valid output for our subgradient oracle. Further, for all $t < n$ we have

$$f(x_t) \geq \langle g_t, x_t \rangle = \|(I - P_{t-1})x_t\|_2 \geq 0.$$

Therefore we can follow the remainder of the proof to show the query lower bound.

8. (Lap Chi Course CS798, HW2 Q1) Suppose you are given an 'accelerated' algorithm $\mathcal{A}$ for minimizing strongly convex and smooth functions: given f that is α-strongly convex and β-smooth, the algorithm outputs x_{alg} satisfying $f(x_{\text{alg}}) - f(x^*) \leq \varepsilon$ in

$$T \lesssim \sqrt{\frac{\beta}{\alpha}} \cdot \log\left(\frac{f(x_0) - f(x^*)}{\varepsilon}\right)$$

iteration, where $x^* = \arg\min_x f(x)$ (unknown), and x_0 is the initial point.

You would like to use algorithm $\mathcal{A}$, but apply it to f that is *not strongly convex*. Prove that you can still use $\mathcal{A}$ as a black box to give an 'accelerated' algorithm with the following guarantee: for convex f that is β-smooth and minimizer $x^* = \arg\min_x f(x)$, find output x such that $f(x) - f(x^*) \leq \varepsilon$ in

$$T \lesssim \sqrt{\frac{\beta\|x_0 - x^*\|_2^2}{\varepsilon}} \cdot \log\left(\frac{f(x_0) - f(x^*)}{\varepsilon}\right)$$

iterations, where again x_0 is the initial point. You are allowed to assume that you know the values $\|x_0 - x^*\|_2$ and $f(x_0) - f(x^*)$.

Solution: Let $f_\alpha(x) := f(x) + \frac{\alpha}{2}\|x - x_0\|_2^2$ and note if f is β-smooth then f_α is α-strongly convex and $(\beta+\alpha)$-smooth (can be verified as sum of convex functions). Letting $f_\alpha^* := \inf_x f_\alpha(x)$ for shorthand, we can therefore apply the black box $\mathcal{A}$ with initial point x_0 to compute x_T satisfying $f_\alpha(x_T) - f_\alpha^* \leq \delta$ in

$$T \lesssim \sqrt{\frac{\beta+\alpha}{\alpha}} \log\left(\frac{f_\alpha(x_0) - f_\alpha^*}{\delta}\right)$$

iterations. We will choose δ to give our final solution. Note the following observations:

$$f(x_T) \leq f_\alpha(x_T), \qquad f_\alpha^* \leq f_\alpha(x^*) = f(x^*) + \frac{\alpha}{2}\|x_0 - x_*\|_2^2.$$

Therefore we can bound our algorithms output

$$f(x_T) - f(x^*) \leq f_\alpha(x_T) - f_\alpha(x^*) + \frac{\alpha}{2}\|x_0 - x\|_2^2 \leq \delta + \frac{\alpha}{2}\|x_0 - x\|_2^2.$$

Finally, choosing $\delta = \frac{\varepsilon}{2}$ and $\alpha = \frac{\varepsilon}{\|x_0 - x^*\|_2^2}$ gives

$$f(x_T) - f(x^*) \leq \delta + \frac{\alpha}{2}\|x_0 - x\|_2^2 \leq \varepsilon.$$

The runtime can be bounded

$$T \lesssim \sqrt{\frac{\beta+\alpha}{\alpha}} \log\left(\frac{f_\alpha(x_0) - f_\alpha^*}{\delta}\right) \lesssim \sqrt{1 + \frac{\beta\|x_0 - x^*\|_2^2}{\varepsilon}} \log\left(\frac{f(x_0) - f(x^*)}{\varepsilon}\right),$$

where we substituted δ, α and used $f(x_0) = f_\alpha(x_0), f_\alpha^* \geq f(x^*)$. Finally, if $\frac{\beta}{2}\|x_0 - x^*\|_2^2 \leq \varepsilon$ we can output x_0, otherwise the above bound gives the required runtime.