Numerical Optimization Final Exam Solutions

Fan Zhang and Xiangyu Yang

December 30, 2020

1 (15 = 5 + 5 + 5 points) Consider a linear system of equations $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ with $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\boldsymbol{x} \in \mathbb{R}^n$ and that $\boldsymbol{A}$ is positive definite. This is equivalent to minimizing a quadratic function $\phi(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}$.

(i) Show that if a set of nonzero vectors $\{\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_m\} \in \mathbb{R}^n$ $(m < n)$ are $\boldsymbol{A}$-conjugate, then they are linearly independent.

(ii) Suppose we have an initial point $\boldsymbol{x}_0$ and initial search direction $\boldsymbol{p}_0 = -\nabla\phi(\boldsymbol{x}_0)$. What is the exact line-search stepsize along $\boldsymbol{p}_0$?

(iii) Suppose we define the new direction as $\boldsymbol{p}_1 = -\boldsymbol{r}_1 + \beta\boldsymbol{p}_0$ (where $\boldsymbol{r}_1$ is the residual at $\boldsymbol{x} = \boldsymbol{x}_1$) and require $\boldsymbol{p}_0, \boldsymbol{p}_1$ are $\boldsymbol{A}$-conjugate. What is the value for β?

Solution:

(i) *Proof.* We prove this by contradiction. Suppose this is not true. Then there exits $\alpha_0, \ldots, \alpha_m$ not all zeros such that

$$\alpha_0\boldsymbol{p}_0 + \alpha_1\boldsymbol{p}_1 + \ldots + \alpha_m\boldsymbol{p}_m = 0. \tag{0.1}$$

Without loss generality, we assume $\alpha_0 \neq 0$. Multiplying $\boldsymbol{p}_0^T A$ on both sides of eq. (0.1), we have

$$\alpha_0\boldsymbol{p}_0^T A\boldsymbol{p}_0 + \alpha_1\boldsymbol{p}_0^T A\boldsymbol{p}_1 + \ldots + \alpha_m\boldsymbol{p}_0^T A\boldsymbol{p}_m = 0, \tag{0.2}$$

where all but the first term vanish because of $\boldsymbol{A}$-conjugate. This implies

$$\alpha_0\boldsymbol{p}_0^T A\boldsymbol{p}_0 = 0.$$

On the other hand, since $\boldsymbol{p}_0 \neq \boldsymbol{0}$ and $\boldsymbol{A}$ is positive definite, we have $\boldsymbol{p}_0^T\boldsymbol{A}\boldsymbol{p}_0 > 0$. It therefore leads to $\alpha_0\boldsymbol{p}_0^T\boldsymbol{A}\boldsymbol{p}_0 \neq 0$. This contradiction completes the proof. $\square$

(ii) This is achieved exactly via solving the following one-dimensional minimization problem

$$\min_{\alpha>0} \quad \phi(\boldsymbol{x}_0 + \alpha \boldsymbol{p}_0), \tag{0.3}$$

where $\alpha \in \mathbb{R}_{++}$ is the stepsize we aim to compute.

It is easy to compute the stepsize α_0 that minimizes $\phi(\boldsymbol{x}_0 - \alpha \nabla \phi(\boldsymbol{x}_0))$. By differentiating the function

$$\phi(\boldsymbol{x}_0 - \alpha \nabla \phi(\boldsymbol{x}_0)) = \frac{1}{2}(\boldsymbol{x}_0 - \alpha \nabla \phi(\boldsymbol{x}_0))^T \boldsymbol{A} (\boldsymbol{x}_0 - \alpha \nabla \phi(\boldsymbol{x}_0)) - \boldsymbol{b}^T (\boldsymbol{x}_0 - \alpha \nabla \phi(\boldsymbol{x}_0))$$

with respect to α, and setting the derivative to zero, we obtain

$$\alpha_0 = \frac{\nabla \phi(\boldsymbol{x}_0)^T \nabla \phi(\boldsymbol{x}_0)}{\nabla \phi(\boldsymbol{x}_0)^T \boldsymbol{A} \nabla \phi(\boldsymbol{x}_0)} = \frac{\boldsymbol{p}_0^T \boldsymbol{p}_0}{\boldsymbol{p}_0^T \boldsymbol{A} \boldsymbol{p}_0}.$$

(iii) Since $\boldsymbol{p}_0, \boldsymbol{p}_1$ are $\boldsymbol{A}$-conjugate, we know

$$0 = \boldsymbol{p}_0^T \boldsymbol{A} \boldsymbol{p}_1 = \boldsymbol{p}_0^T \boldsymbol{A} (-\boldsymbol{r}_1 + \beta \boldsymbol{p}_0).$$

This implies

$$\beta = \frac{\boldsymbol{r}_1^T \boldsymbol{A} \boldsymbol{p}_0}{\boldsymbol{p}_0^T \boldsymbol{A} \boldsymbol{p}_0}.$$

2 (10 points) Find the projection of $\boldsymbol{y} \in \mathbb{R}^n$ onto the half-hyperspace $\{\boldsymbol{x} \mid \boldsymbol{a}^T\boldsymbol{x} \leq \boldsymbol{b}\}$ with $\boldsymbol{a} \in \mathbb{R}^n$ and $\boldsymbol{a} \neq \boldsymbol{0}$. In other words, solve the following Euclidean projection problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \tfrac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \quad \text{s.t. } \boldsymbol{a}^T\boldsymbol{x} \leq b. \tag{0.4}$$

Solution: The Lagrangian is

$$\mathcal{L}(\boldsymbol{x}, \lambda) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda(\boldsymbol{a}^T\boldsymbol{x} - b), \tag{0.5}$$

where $\lambda \in \mathbb{R}_+$ is the introduced Lagrange multiplier. Then the Karush–Kuhn–Tucker (KKT) conditions read

$$\begin{aligned} \boldsymbol{x}^* - \boldsymbol{y} + \lambda\boldsymbol{a} &= \boldsymbol{0}, && \text{(stationay conditions)} \\ \boldsymbol{a}^T\boldsymbol{x}^* &\leq b, && \text{(primal feasibility)} \\ \lambda &\geq 0, && \text{(dual feasibility)} \\ \lambda(\boldsymbol{a}^T\boldsymbol{x}^* - b) &= 0. && \text{(slackness)} \end{aligned}$$

As we want to eliminate λ, we start with eq. (slackness). Then we consider the following two cases

1). $\lambda = 0$. eq. (stationay conditions) immediately gives $\boldsymbol{x}^* = \boldsymbol{y}$ and eq. (primal feasibility) gives $\langle \boldsymbol{a}, \boldsymbol{x}^* \rangle = \langle \boldsymbol{a}, \boldsymbol{y} \rangle \leq b$. This corresponds to the case $\boldsymbol{y}$ inside the closed half-hyperspace.

2). $\lambda > 0$. eq. (slackness) gives $\langle \boldsymbol{a}, \boldsymbol{x}^* \rangle = b$. Multiplying eq. (stationay conditions) by $\boldsymbol{a}^T$ and using eq. (slackness), we have

$$\boldsymbol{a}^T\boldsymbol{x}^* - \boldsymbol{a}^T\boldsymbol{y} + \lambda\|\boldsymbol{a}\|_2^2 = 0 \longrightarrow \lambda = \frac{\langle \boldsymbol{y}, \boldsymbol{a} \rangle - b}{\|\boldsymbol{a}\|_2^2}. \tag{0.7}$$

In addition, this implies $\langle \boldsymbol{y}, \boldsymbol{a} \rangle > b$, which is the case $\boldsymbol{y}$ outside the closed half-hyperspace.

Furthermore, by substituting eq. (0.7) into eq. (stationay conditions), we have

$$\boldsymbol{x}^* = \boldsymbol{y} - \frac{\langle \boldsymbol{y}, \boldsymbol{a} \rangle - b}{\|\boldsymbol{a}\|_2^2}\boldsymbol{a}.$$

Combine the 2 cases gives

$$\boldsymbol{x}^* = \begin{cases} \boldsymbol{y} & \text{if } \langle \boldsymbol{a}, \boldsymbol{y} \rangle \leq b, \\ \boldsymbol{y} - \frac{\langle \boldsymbol{y}, \boldsymbol{a} \rangle - b}{\|\boldsymbol{a}\|_2^2}\boldsymbol{a} & \text{if } \langle \boldsymbol{a}, \boldsymbol{y} \rangle > b. \end{cases}$$

3 (15 points) Consider the inequality constrained strictly convex quadratic programming (QP) problem

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & \tfrac{1}{2}\boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x} + \boldsymbol{g}^T\boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{a}_i^T\boldsymbol{x} + b_i = 0, \quad i = 1,\dots,m \\ & \boldsymbol{a}_i^T\boldsymbol{x} + b_i \le 0, \quad i = m+1,\dots,t. \end{aligned} \tag{0.8}$$

Suppose $\boldsymbol{x}^*$ is the first-order optimal solution and the active-set at $\boldsymbol{x}^*$ is $\mathcal{A}(\boldsymbol{x}^*) := \{i \in \{m+1,\dots,t\} \mid \boldsymbol{a}_i^T\boldsymbol{x}^* + b_i = 0\}$. Show that $\boldsymbol{d} = \boldsymbol{0}$ is optimal for the following problem

$$\begin{aligned} \min_{\boldsymbol{d}\in\mathbb{R}^n} \quad & \tfrac{1}{2}(\boldsymbol{x}^*+\boldsymbol{d})^T\boldsymbol{H}(\boldsymbol{x}^*+\boldsymbol{d}) + \boldsymbol{g}^T(\boldsymbol{x}^*+\boldsymbol{d}) \\ \text{s.t.} \quad & \boldsymbol{a}_i^T(\boldsymbol{x}^*+\boldsymbol{d}) + b_i = 0, \quad i = 1,\dots,m \\ & \boldsymbol{a}_i^T(\boldsymbol{x}^*+\boldsymbol{d}) + b_i \le 0, \quad i \in \mathcal{A}(\boldsymbol{x}^*). \end{aligned} \tag{0.9}$$

Solution:

Proof. For the sake of convenience, denote $\mathcal{A}^* = \mathcal{A}(\boldsymbol{x}^*)$ and $\mathcal{E} = \{1,\cdots,m\}$, and $\boldsymbol{A} = (\boldsymbol{a}_1,\cdots,\boldsymbol{a}_t)^T$. Since $\boldsymbol{x}^*$ is the first-order optimal solution to eq. (0.8), we know

$$\boldsymbol{H}\boldsymbol{x}^* + \boldsymbol{g} + \boldsymbol{A}^T\lambda = \boldsymbol{0} \tag{0.10a}$$
$$\boldsymbol{a}_i^T\boldsymbol{x}^* + b_i = 0, \quad i \in \mathcal{E}\cup\mathcal{A}^* \tag{0.10b}$$
$$\lambda_i \ge 0, \quad i \in \mathcal{A}^* \tag{0.10c}$$

The KKT conditions for the subproblem eq. (0.9) are

$$\boldsymbol{H}\boldsymbol{x}^* + \boldsymbol{g} + \boldsymbol{A}^T\mu = \boldsymbol{0} \tag{0.11a}$$
$$\boldsymbol{a}_i^T(\boldsymbol{x}^*+\boldsymbol{d}) + b_i = 0, \quad i \in \mathcal{E} \tag{0.11b}$$
$$\boldsymbol{a}_i^T(\boldsymbol{x}^*+\boldsymbol{d}) + b_i \le 0, \quad i \in \mathcal{A}^* \tag{0.11c}$$
$$\mu_i(\boldsymbol{a}_i^T(\boldsymbol{x}^*+\boldsymbol{d}) + b_i) = 0, \quad i \in \mathcal{A}^* \tag{0.11d}$$
$$\mu_i \ge 0, \quad i \in \mathcal{A}^*, \tag{0.11e}$$

where $\mu_i,\ i \in \mathcal{E}\cup\mathcal{A}^*$ are dual multipliers. Setting $\boldsymbol{d} = \boldsymbol{0}$, eq. (0.11) can be rewritten as

$$\boldsymbol{H}\boldsymbol{x}^* + \boldsymbol{g} + \boldsymbol{A}^T\mu = \boldsymbol{0} \tag{0.12a}$$
$$\boldsymbol{a}_i^T\boldsymbol{x}^* + b_i = 0, \quad i \in \mathcal{E}\cup\mathcal{A}^* \tag{0.12b}$$
$$\mu_i \ge 0, \quad i \in \mathcal{A}^*, \tag{0.12c}$$

which coincides eq. (0.10) with $\mu = \lambda$. Hence, $\boldsymbol{d} = \mathbf{0}$ is optimal for eq. (0.9). $\square$

4 (15 = 5 + 10 points) In a quasi-Newton method for solving the unconstrained optimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \ f(\boldsymbol{x}).$$

We use local model

$$m_k(\boldsymbol{d}) = f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^T\boldsymbol{d} + \tfrac{1}{2}\boldsymbol{d}^T\boldsymbol{H}_k\boldsymbol{d}$$

to approximate $f(\boldsymbol{x})$ at $\boldsymbol{x}_k$. The secant equation is obtained by requiring the gradient of $m_k(\boldsymbol{d})$ at $\boldsymbol{x}_{k-1}$ is equivalent to $\nabla f(\boldsymbol{x}_{k-1})$.

(i) Derive the secant equation that must satisfy.

(ii) Suppose you were using a multiple of identity matrix $\alpha\boldsymbol{I}$ to approximate the Hessian matrix (i.e., $\boldsymbol{H}_k = \alpha\boldsymbol{I}$), which may not satisfy the secant equation. Find the α as the least-squares solution of the secant equation. (The least squares solution of a linear equation $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ is the minimizer of $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$.)

Solution:

(i) Note that with respect to $\boldsymbol{x}_k$, we get to $\boldsymbol{x}_{k-1}$ with the step $-\alpha_{k-1}\boldsymbol{d}_{k-1}$. Since the gradient values should match at $\boldsymbol{x}_k$ and $\boldsymbol{x}_{k-1}$, we thus have

$$\nabla m_k(-\alpha_{k-1}\boldsymbol{d}_{k-1}) = \nabla f(\boldsymbol{x}_{k-1}).$$

Rearranging

$$\nabla f(\boldsymbol{x}_{k-1}) = \nabla m_k(-\alpha_{k-1}\boldsymbol{d}_{k-1}) = \nabla f(\boldsymbol{x}_k) + \boldsymbol{H}_k(-\alpha_{k-1}\boldsymbol{d}_{k-1}),$$

we obtain the secant equation as

$$\boldsymbol{H}_k\boldsymbol{s}_{k-1} = \boldsymbol{y}_{k-1},$$

where $\boldsymbol{s}_{k-1} = \boldsymbol{x}_k - \boldsymbol{x}_{k-1} = \alpha_{k-1}\boldsymbol{d}_{k-1}$ and $\boldsymbol{y}_{k-1} = \nabla f(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_{k-1})$.

(ii) We aim to find a α that matches the secant equation in the sense that it minimizes the sum of squared errors

$$\min_{\alpha>0} \quad \tfrac{1}{2}\|\alpha\boldsymbol{I}\boldsymbol{s}_{k-1} - \boldsymbol{y}_{k-1}\|_2^2. \tag{0.13}$$

This gives

$$\alpha = \frac{\boldsymbol{s}_{k-1}^T\boldsymbol{y}_{k-1}}{\|\boldsymbol{s}_{k-1}\|_2^2}.$$

Remark 0.1. It should be noticed that this is similar to the Barzilai-Borwein (BB) method, which is motivated by Newton's method but not involves any Hessian. Now, we derive the BB method.

Consider a quadratic function

$$f(\boldsymbol{x}) = \tfrac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x},$$

where $\boldsymbol{A}$ is positive definite.

- Newton step: $\boldsymbol{d}_k^{\text{Newton}} = -\boldsymbol{A}^{-1}\nabla f(\boldsymbol{x}_k)$
- Goal: Choose α_k such that $-\alpha_k\nabla f(\boldsymbol{x}_k) = -(\alpha_k^{-1}\boldsymbol{I})\nabla f(\boldsymbol{x}_k)$ approximates $-\boldsymbol{A}^{-1}\nabla f(\boldsymbol{x}_k)$.
- Define $\boldsymbol{s}_{k-1} = \boldsymbol{x}_k - \boldsymbol{x}_{k-1}$ and $\boldsymbol{y}_{k-1} = \nabla f(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_{k-1})$. Then $\boldsymbol{A}$ satisfies
$$\boldsymbol{A}\boldsymbol{s}_{k-1} = \boldsymbol{y}_{k-1}.$$
- Therefore, give $\boldsymbol{s}_{k-1}$ and $\boldsymbol{y}_{k-1}$, how about choose α_k so that
$$(\alpha_k^{-1}\boldsymbol{I})\boldsymbol{s}_{k-1} \approx \boldsymbol{y}_{k-1}.$$
- BB method
  - Least-squares problem: let $\beta = \alpha^{-1}$
  $$\alpha_k^{-1} = \arg\min_{\beta} \frac{1}{2}\|\boldsymbol{s}_{k-1}\beta - \boldsymbol{y}_{k-1}\|_2^2 \longrightarrow \alpha_k^1 = \frac{\boldsymbol{s}_{k-1}^T\boldsymbol{s}_{k-1}}{\boldsymbol{s}_{k-1}^T\boldsymbol{y}_{k-1}}.$$
  - Alternative Least-squares problem:
  $$\alpha_k = \arg\min_{\beta} \frac{1}{2}\|\boldsymbol{s}_{k-1} - \boldsymbol{y}_{k-1}\alpha\|_2^2 \longrightarrow \alpha_k^2 = \frac{\boldsymbol{s}_{k-1}^T\boldsymbol{y}_{k-1}}{\boldsymbol{y}_{k-1}^T\boldsymbol{y}_{k-1}}.$$
- α_k^1 and α_k^2 are called the BB step sizes.

The material of this remark is adapted from Math 164: Optimization.

5 (30 = 10 × 3 points) Consider the unconstrained optimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \; f(\boldsymbol{x}),$$

where $f \in C^2$. In addition, we have the following assumptions on f:

(1) f is L-smooth.

(2) f is bounded below over $\boldsymbol{x} \in \mathbb{R}^n$.

(i) Suppose $\boldsymbol{d}_k \in \mathbb{R}^n$ is a descent direction at $\boldsymbol{x}_k$. Show that

$$f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k) \le f(\boldsymbol{x}_k)$$

for sufficiently small stepsize $\alpha_k > 0$.

(ii) The Armijo line search condition is

$$f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k) \le f(\boldsymbol{x}_k) + \eta\alpha_k \nabla f(\boldsymbol{x}_k)^T \boldsymbol{d}_k$$

with $\eta \in (0,1)$. Show that for a sufficiently small stepsize, this condition must hold (provide the expression of this stepsize).

(iii) Now let $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$ with $\boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k)$, show that

$$\|\nabla f(\boldsymbol{x}_k)\|_2 \to 0.$$

Solution:

Proof. (i) Applying the first order Taylor expansion of $f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k)$ at the point $\boldsymbol{x}_k$,

$$f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k) = f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \alpha_k \boldsymbol{d}_k\rangle + \frac{\alpha_k^2}{2}(\boldsymbol{d}_k)^T \nabla^2 f(\boldsymbol{x}_k + \theta\alpha_k \boldsymbol{d}_k)\boldsymbol{d}_k,$$

where $\theta \in (0,1)$. For a sufficiently small α_k, the term $\frac{\alpha_k^2}{2}(\boldsymbol{d}_k)^T \nabla^2 f(\boldsymbol{x}_k + \theta\alpha_k \boldsymbol{d}_k)\boldsymbol{d}_k$ can be ignored. Since $\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k\rangle < 0$ and $\alpha_k > 0$, it holds

$$f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k) = f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \alpha_k \boldsymbol{d}_k\rangle < f(\boldsymbol{x}_k),$$

which completes the proof.

(ii) For an arbitrary α, we achieve an upper bound of $f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)$ by the Lipschitz continuity of ∇f

$$f(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k) \le f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \alpha \boldsymbol{d}_k\rangle + \frac{\alpha^2}{2} L\|\boldsymbol{d}_k\|_2^2. \tag{0.14}$$

By making α further satisfying the sufficient decrease condition, it holds

$$f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \alpha \boldsymbol{d}_k \rangle + \frac{\alpha^2}{2} L \|\boldsymbol{d}_k\|_2^2 \leq f(\boldsymbol{x}_k) + \eta \langle \nabla f(\boldsymbol{x}_k), \alpha \boldsymbol{d}_k \rangle .$$

Therefore, for any $\alpha \in [0, \frac{2(\eta-1)\langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k \rangle}{L\|\boldsymbol{d}_k\|_2^2}]$, the Armijo line search condition is satisfied. And then, the backtracking procedure must end up with

$$\alpha_k \geq 2\gamma \frac{(\eta - 1) \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k \rangle}{L\|\boldsymbol{d}_k\|_2^2}, \tag{0.15}$$

where γ is the decay constant of the line search.

(iii) Combining the Armijo line search condition with eq. (0.15) and substituting $\boldsymbol{d}_k$ by $-\nabla f(\boldsymbol{x}_k)$ gives

$$\begin{aligned} f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) =& f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k) \\ \geq& - \eta \alpha_k \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k \rangle \\ \geq& \eta \frac{2\gamma(1-\eta) \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k \rangle}{L\|\boldsymbol{d}_k\|_2^2} \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}_k \rangle \\ =& \frac{2\gamma\eta(1-\eta)\|\nabla f(\boldsymbol{x}_k)\|_2^2}{L}. \end{aligned} \tag{0.16}$$

By rearranging eq. (0.16) and summing up both sides from 1 to t, we have

$$\sum_{k=0}^{t} \|\nabla f(\boldsymbol{x}_k)\|_2^2 < \frac{L}{\eta(1-\eta)} \sum_{k=0}^{t} (f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1})) \leq \frac{L}{2\gamma\eta(1-\eta)} \left(f(\boldsymbol{x}^0) - \underline{f} \right) .$$

Thus, $\|\nabla f(\boldsymbol{x}_t)\|_2^2 \to 0$ as $t \to \infty$.

$\square$

6 (15 points) Consider the unconstrained optimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x}).$$

The trust region subproblem subproblem is given by

$$\min_{\boldsymbol{d}\in\mathbb{R}^n} \nabla f(\boldsymbol{x}_k)^T\boldsymbol{d} + \tfrac{1}{2}\boldsymbol{d}^T\boldsymbol{H}\boldsymbol{d} \qquad \text{s.t.} \|\boldsymbol{d}\|_2 \le \Delta_k. \tag{0.17}$$

Derive the Cauchy-point of this subproblem.

Solution:

First, it is easy to obtain the solution of eq. (0.17), which reads

$$\boldsymbol{d}_k^s = -\frac{\Delta_k}{\|\nabla f(\boldsymbol{x}^k)\|}\nabla f(\boldsymbol{x}^k). \tag{0.18}$$

To obtain τ_k, we consider two cases

1. Suppose $\nabla f(\boldsymbol{x}^k)^T H_k \nabla f(\boldsymbol{x}^k) \le 0$. Then the function $m(\tau \boldsymbol{d}_k^s)$ decreases monotonically with τ whenever $\nabla f(\boldsymbol{x}^k) \neq \mathbf{0}$. Therefore, τ_k is simply the largest value that satisfies the trust-region bound, namely, $\tau_k = 1$.

2. Suppose $\nabla f(\boldsymbol{x}^k)^T H_k \nabla f(\boldsymbol{x}^k) > 0$. Then $m(\tau \boldsymbol{d}_k^s)$ is a convex quadratic in τ, so τ_k is either the unconstrained minimizer of this quadratic, $\|\nabla f(\boldsymbol{x}^k)\|^3/(\Delta_k \nabla f(\boldsymbol{x}^k)^T H_k \nabla f(\boldsymbol{x}^k))$, or the boundary value 1, whichever comes first.

   Overall, we have

$$\boldsymbol{d}_k^c = -\tau_k \frac{\Delta_k}{\|\nabla f(\boldsymbol{x}^k)\|}\nabla f(\boldsymbol{x}^k), \tag{0.19}$$

   where

$$\tau_k = \begin{cases} 1 & \text{if } \nabla f(\boldsymbol{x}^k)^T H_k \nabla f(\boldsymbol{x}^k) \le 0, \\ \min(1, \|\nabla f(\boldsymbol{x}^k)\|^3/(\Delta_k \nabla f(\boldsymbol{x}^k)^T H_k \nabla f(\boldsymbol{x}^k))) & \text{otherwise.} \end{cases}$$

Setting $\boldsymbol{d}_k^c = \tau_k \boldsymbol{d}_k^s$ yields the solution.