Yale University
CPSC 516, Spring 2023
Assignment 5

Chang Feng (Felix) Zhou cz397

P.1.

(a)

From our work in class, we know that $\nabla^2 f \succeq mI$ implies that f is m-strongly convex. This yields the inequality

$$f(y) - [f(x) + \langle \nabla f(x), y - x\rangle] \geq \frac{m}{2}\|y - x\|_2^2$$

by the definition of strong convexity.

On the other hand, we know by the second order Taylor expansion about x that

$$\begin{aligned} f(y) &= f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}(y-x)^T \nabla^2 f(\xi)(y-x) && \xi \in [x, y] \\ f(y) - f(x) + \langle \nabla f(x), x - y\rangle &\leq \frac{M}{2}\|y - x\|_2^2. && MI \succeq \nabla^2 f \end{aligned}$$

This shows both inequalities.

(b)

Suppose $f(z^*) = y^*$ and consider the inequality from P.1.(a)

$$y^* \leq f(z) \leq f(x) + \langle \nabla f(x), z - x\rangle + \frac{M}{2}\|z - x\|_2^2.$$

Let us minimize the RHS with respect to z by taking the derivative and setting it to 0. We must have

$$\begin{aligned} \nabla f(x) + M[z - x] &= 0 \\ z &= x - \frac{1}{M}\nabla f(x). \end{aligned}$$

Substituting this particular value of z to the RHS above yields

$$\begin{aligned} &f(x) + \left\langle \nabla f(x), -\frac{1}{M}\nabla f(x)\right\rangle + \frac{1}{2M}\|\nabla f(x)\|_2^2 \\ &= f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2 \end{aligned}$$

which is one of the desired inequalities.

To see the other inequality, We note that we wish to prove

$$\frac{1}{2m}\|\nabla f(x)\|_2^2 \geq [f(x) - f(z^*)]$$

which is known as the *Polyak-Lojasiewicz (PL)* condition.

By the definition of strong convexity,

$$
\begin{aligned}
f(x) - f(z^*) &\le \langle \boldsymbol{\nabla} f(x), x - z^* \rangle - \frac{m}{2}\|x - z^*\|_2^2 \\
&= \langle \boldsymbol{\nabla} f(x), x - z^* \rangle - \frac{m}{2}\|x - z^*\|_2^2 - \frac{1}{2m}\|\boldsymbol{\nabla} f(x)\|_2^2 + \frac{1}{2m}\|\boldsymbol{\nabla} f(x)\|_2^2 \\
&= -\frac{1}{2}\left\|\sqrt{m}(x - z^*) - \frac{1}{\sqrt{m}}\boldsymbol{\nabla} f(x)\right\|_2^2 + \frac{1}{2m}\|\boldsymbol{\nabla} f(x)\|_2^2 \\
&\le \frac{1}{2m}\|\boldsymbol{\nabla} f(x)\|_2^2.
\end{aligned}
$$

Having shown both inequalities, we conclude the proof.

(c)

In P.1.(b), we have shown that

$$f(z) \le f(x) - \frac{1}{2M}\|\boldsymbol{\nabla} f(x)\|_2^2$$

for $z := x - \frac{1}{M}\boldsymbol{\nabla} f(x)$.

But since we chose the step size α to minimize $f(x_{t+1})$, it must be at least as good as $\alpha = \frac{1}{M}$. Thus

$$f(x_{t+1}) \le f(x_t) - \frac{1}{M}\|\boldsymbol{\nabla} f(x_t)\|_2^2$$

as desired.

(d)

We argue by induction, the base case

$$f(x_0) - y^* \le 1 \cdot [f(x_0) - y^*]$$

holds trivially. Suppose it holds up to some t and consider $f(x_{t+1}) - y^*$.

We have

$$
\begin{aligned}
f(x_{t+1}) - y^* &= f(x_{t+1}) - f(x_t) + f(x_t) - y^* \\
&\le f(x_t) - y^* - \frac{1}{2M}\|\boldsymbol{\nabla} f(x_t)\|_2^2 && \text{P.1.(c)} \\
&\le f(x_t) - y^* - \frac{m}{M}[f(x_t) - y^*] && \text{P.1.(b) LHS} \\
&= \left(1 - \frac{m}{M}\right)[f(x_t) - y^*] \\
&= \left(1 - \frac{m}{M}\right)^{t+1}[f(x_0) - y^*]. && \text{induction hypothesis}
\end{aligned}
$$

By induction, we conclude the proof.

The number of iterations to reach ε error can be computed as follows

$$\begin{aligned}
\left(1 - \frac{m}{M}\right)^t [f(x_0) - y^*] &\leq \exp(-mt/M)[f(x_0) - y^*] && 1 - x \leq e^{-x} \\
&\leq \varepsilon \\
-\frac{mt}{M} + \log[f(x_0) - y^*] &\leq \log \varepsilon \\
t &\geq \frac{M}{m} \log \frac{f(x_0) - y^*}{\varepsilon}.
\end{aligned}$$

(e)

Suppose we are given A, b as input.

Consider the minimization problem

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2$$

The objective is the composition of an affine function and a convex, separable, and non-decreasing (in each coordinate) function, which is therefore convex.

We explicitly compute its first and second derivatives

$$\begin{aligned}
\frac{d}{dx}[Ax + b]^T [Ax + b] &= \frac{d}{dx} x A^2 x + 2x^T Ab + b^T b \\
&= 2A^2 x + 2Ab \\
\frac{d^2}{dx^2} &= 2A^2.
\end{aligned}$$

The objective is certainly twice differentiable, and since the eigenvalues of A^2 are just the eigenvalues of A squared,

$$\lambda_1(A)^2 \leq \boldsymbol{\nabla} f^2 \leq \lambda_n(A)^2.$$

By our work above, if we start with an initial solution $x_0 := 0$ and run gradient descent with step size $\alpha = \frac{1}{\lambda_n(A)^2}$, this yields a solution x such that $\|Ax - b\|_2^2 \leq \varepsilon$ after

$$T = O\left(\frac{\lambda_n(A)^2}{\lambda_1(A)^2} \log \frac{\|b\|_2^2}{\varepsilon}\right)$$

iterations. In each iteration, we need to compute the gradient and subtract it from the current iterate. The number of arithmetic operations is dominated by the gradient computation $A(Ax)$, which requires $O(n^2)$ operations if we compute Ax and then $A(Ax)$.

Thus the algorithm terminates after performing

$$O(n^2 T) = O\left(n^2 \kappa^2 \log \frac{\|b\|_2^2}{\varepsilon}\right)$$

arithmetic operations.

P.2.

Lemma 1:
$(BB^T)^\dagger Bg$ is a minimizer of the optimization problem

$$\min_{y \in \mathbb{R}^n} \|B^T y - g\|_2^2.$$

Thus $B^T(BB^T)^\dagger Bg$ is the Euclidean projection of g onto the row space of B.

Proof : Lemma 1

The objective function is a composition of an affine (convex) function with a convex, separable, and coordinate-wise non-decreasing function, which is therefore convex. We can thus solve this problem by taking the gradient and setting it to zero.

The objective is equivalent to

$$(B^T y - g)^T (B^T y - g) = y^T BB^T y - 2g^T B^T y - g^T g.$$

Taking the derivative and setting it to 0 yields

$$\begin{aligned} 2y^T BB^T - 2g^T B^T &= 0 \\ y &= (BB^T)^\dagger Bg. \end{aligned}$$

□

Recall from elmentary linear algebra that the kernel is the orthogonal complement of the rowspace so that we can write

$$\mathbb{R}^m = \ker B \oplus \operatorname{row}(B).$$

In particular, if $r := \operatorname{rank} B$, we can find an orthonormal basis of $\mathbb{R}^m$, say $v_1, \ldots, v_r, w_1, \ldots, w_{m-r}$, where $v_i \in \operatorname{row}(B), w_j \in \ker B$. Thus we can write

$$g = \sum_{i=1}^{r} \langle g, v_i \rangle v_i + \sum_{j=1}^{m-r} \langle g, w_j \rangle w_j.$$

By elementary linear algebra,

$$\begin{aligned} x_g &= \sum_{i=1}^{r} \langle g, v_i \rangle v_i \\ \Pi g &= \sum_{j=1}^{m-r} \langle g, w_j \rangle w_j. \end{aligned}$$

Thus

$$x_g + \Pi g = g$$

as desired.