Q1: MAXQUAD

Preliminary

First we specify the required parameters and define the oracle.

```
%% parameters
clear;clc;
d = 10;
K = 5;
opt(K) = struct();
x1 = ones(d,1);

%% Create optimization oracles
[I,J] = meshgrid(1:d, 1:d);
for k = 1:K
    A = zeros(d);
    b = zeros(1, d);
    [J, I] = meshgrid(1:d, 1:d);
    A = exp(I./J) .* cos(I.*J) .* sin(k);
    A(I>=J) = 0;
    A = (A + A');
    A = A + diag(abs(sin(k)/10.*(1:d)') + sum(abs(A), 2));
    b = exp((1:d)'/k) .* sin((1:d)'*k);
    opt(k).A = A;
    opt(k).b = b;
end
```

Part 1

```
%% evaluating initial point
fx1 = oracle_f(x1, opt, K);
fprintf("Q1 Part 1: \n");
```

```
Q1 Part 1:
```

```
fprintf("The initial objective value is: %5.4f \n\n", fx1);
```

```
The initial objective value is: 5337.0664
```

Part 2

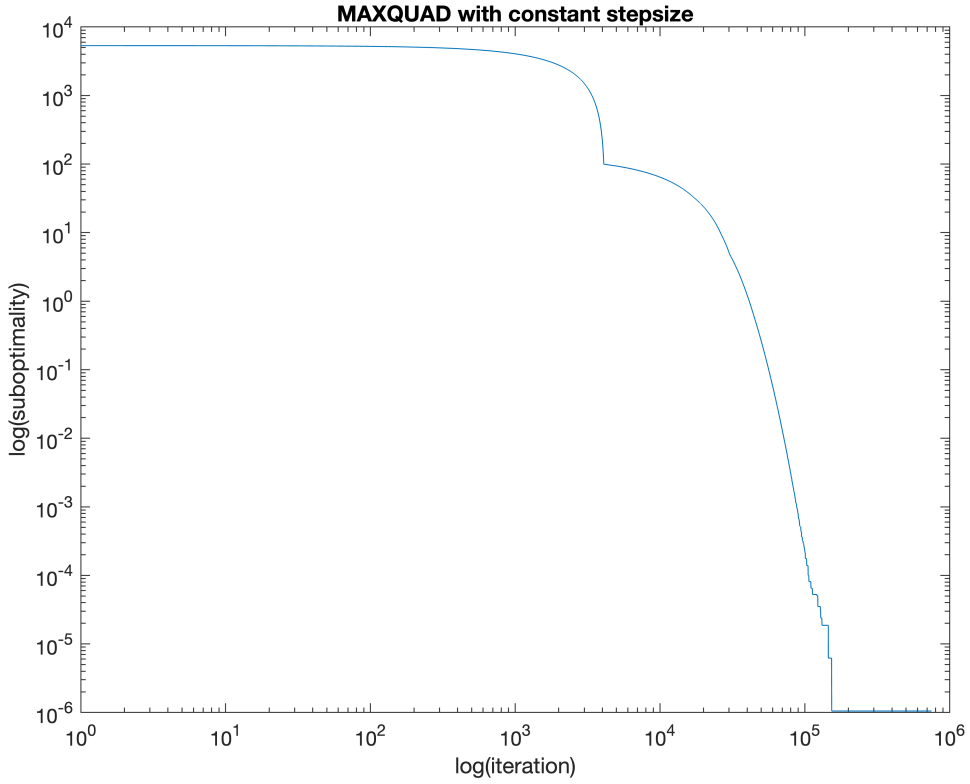
```
%% finding optimal value: full-batch stepsize

T = 1e6; C = 0.1; epsilon = 1e-6;
eta = C/sqrt(T);
xt  = x1;
[fxt, gxt]     =  oracle_f(xt, opt, K);
best_so_far    =  zeros(T,1);
best_so_far(1) =  fxt;

for t = 2:T
```

```
    % if mod(t, 1e4) == 0
    %   disp(["Running iteration: ", num2str(t)]);
    % end
    xt = xt - eta * gxt/norm(gxt);
    [fxt, gxt] = oracle_f(xt, opt, K);
    if (fxt < best_so_far(t-1) - epsilon)
        best_so_far(t) = fxt;
    else
        best_so_far(t) = best_so_far(t-1);
    end
end

best_gap = best_so_far - best_so_far(T);
loglog(1:T, best_gap);
title("MAXQUAD with constant stepsize");
xlabel('log(iteration)');
ylabel('log(suboptimality)');
```



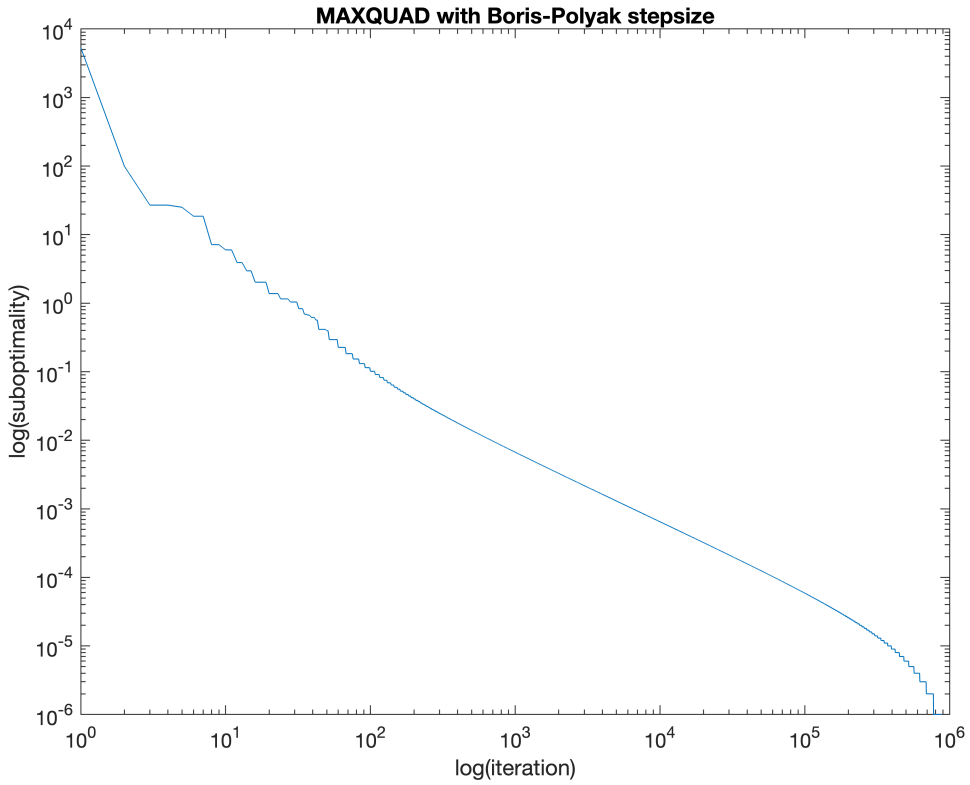

```
fprintf('\n');
```

Part 3

From the results we can see constant stepsize requires a long procedure for burning in, and the suboptimality does not closely follow the theoretical behavior implied by the upper bound(which should be approximately linear on a log-log plot). On the contrary, Boris-Polyak stepsize gives updates that smoothly decreases and aligns better with the theoretical trend.

```
%% finding optimal value: Boris-Polyak stepsize
T = 1e6; C = 1; epsilon = 1e-6;
xt = x1;
[fxt, gxt] =  oracle_f(xt, opt, K);
best_so_far_bp    = zeros(T,1);
best_so_far_bp(1) = fxt;
fmin = best_so_far(T);

for t = 2:T
    % if mod(t, 1e4) == 0
    %  disp(["Running iteration: ", num2str(t)]);
    % end
    eta = (fxt - fmin) / norm(gxt);
    xt  = xt - eta * gxt / norm(gxt);
    [fxt, gxt] =  oracle_f(xt, opt, K);
    if (fxt < best_so_far_bp(t-1) - epsilon)
        best_so_far_bp(t) = fxt;
    else
        best_so_far_bp(t) = best_so_far_bp(t-1);
    end
end
best_gap_bp = best_so_far_bp - best_so_far_bp(T);
loglog(1:T, best_gap_bp);
title("MAXQUAD with Boris-Polyak stepsize");
xlabel('log(iteration)');
ylabel('log(suboptimality)');
```



```
function [fx, gx] = oracle_f(x, opt, K)
   obj = zeros(K,1);
   for k = 1:K
       obj(k) = x' * opt(k).A * x - x' * opt(k).b;
   end
   [fx, ind]  =  max(obj);
   gx = 2 .* opt(ind).A * x - opt(ind).b;
end
```

Q2. Soft-max function.

RHS: $S_\alpha(x) = \frac{1}{\alpha} \log\left(\sum_{i=1}^{n} e^{\alpha x_i}\right)$

$\leq \frac{1}{\alpha} \log\left(n \cdot e^{\alpha \max\{x_i\}}\right)$

$= \frac{\log n}{\alpha} + \max\{x_i\}.$

LHS: $S_\alpha(x) = \frac{1}{\alpha} \log\left(\sum_{i=1}^{n} e^{\alpha x_i}\right)$

$\geq \frac{1}{\alpha} \log\left(e^{\alpha \max\{x_i\}}\right)$

$= \max\{x_i\}.$

$$\frac{\partial S_\alpha(x)}{\partial x_i} = \alpha^{-1} \cdot \frac{\alpha e^{\alpha x_i}}{\sum_{i=1}^{n} e^{\alpha x_i}}$$

$$= \frac{e^{\alpha x_i}}{\sum_{i=1}^{n} e^{\alpha x_i}} \leq 1.$$

Thus $\| \nabla S_\alpha(x) \|_\infty \leq 1.$

Q3: Stochastic GD.

(a) Since K is convex, $x_{t+1} = \frac{t}{t+1} x_t + \frac{1}{t+1} x^t$

By induction it's easy to show $x \in K$.

$$r_{t+1}^2 = \|x_{t+1} - x^*\|_2^2$$

$$= \|\Pi_K(x_t - \eta g(x^t)) - x^*\|_2^2$$

$$\leq \|x_t - \eta g(x_t) - x^*\|_2^2$$

$$= \|x_t - x^*\|^2 - 2\eta \langle g(x_t), x_t - x^* \rangle + \eta^2 \|g(x_t)\|^2$$

$$\leq r_t^2 - 2\eta \cdot (f(x_t) - f^*) + \eta^2 \cdot \|g(x_t)\|^2$$

which gives:

$$2\eta (f(x_t) - f^*) \leq r_t^2 - r_{t+1}^2 + \eta^2 \|g(x_t)\|^2$$

Summing over $t = 0, \cdots, T-1$, we have

$$\frac{2\eta}{T} \sum_{t=0}^{T-1} (f(x_t) - f^*) \leq \frac{r_0^2}{T} + \frac{\eta^2}{T} \sum_{t=0}^{T-1} \|g(x_t)\|^2$$

By convexity of f, we have

$$f\left(\frac{1}{T} \sum_{t=0}^{T-1} x_t\right) \leq \frac{1}{T} \sum_{t=0}^{T-1} f(x_t).$$

Hence

$$f(x_T) - f^* \le \frac{\gamma_0^2}{2\eta T} + \frac{\eta}{2T} \sum_{t=0}^{T-1} \|g(x_t)\|^2$$

Taking expectation, we have

$$\mathbb{E}[f(x_T) - f^*] \le \frac{D^2}{2\eta T} + \frac{\eta G^2}{2}$$

which is minimized at

$$\eta = \frac{D}{G\sqrt{T}} \quad \text{with}$$

$$\mathbb{E}[f(x_T) - f^*] \le \frac{DG}{\sqrt{T}}$$

Hence to achieve ε-optimal, we need

$$T = \left(\frac{DG}{\varepsilon}\right)^2.$$

(b)

(1) $\mathbb{E}[g(x)] = \mathbb{E}\left[2a_i(\langle x, a_i\rangle - b_i)\right]$

$= \frac{1}{m}\sum_{i=1}^{m} 2a_i(\langle x, a_i\rangle - b_i)$

(since each i is taken uniformly from $\{1, \cdots, m\}$)

$= \nabla f(x)$.

$\mathbb{E}[\|g(x)\|_2^2] = \frac{1}{m}\sum_{i=1}^{m} (\langle x, a_i\rangle - b_i)^2 \cdot 4\|a_i\|_2^2$.

Note that

$(\langle x, a_i\rangle - b_i)^2 \le (|\langle x, a_i\rangle| + |b_i|)^2$

$\le (\|x\|_2 \cdot \|a_i\|_2 + |b_i|)^2 \le 4$.

Hence $\mathbb{E}(\|g(x)\|_2^2) \le 16$.

(2) The proved result implies that.

putting stepsize $\eta = \frac{D}{4\sqrt{T}}$,

in $T = \left(\frac{4D}{\varepsilon}\right)^2$ steps

one could achieve ε-optimality by running the stochastic GD.

(3) Efficient: when m is large, no need to compute the true gradient.

Q4. Subgradient descent for nonsmooth functions.

(a) $R(x)$ is convex since it's a norm, which implies:

$$\|\lambda x + (1-\lambda) y\|_1 \le \|\lambda x\|_1 + \|(1-\lambda) y\|_1$$
$$= \lambda \|x\|_1 + (1-\lambda) \|y\|_1 ,$$
$$\forall\, x, y \in \mathbb{R}^n, \quad \lambda \in [0, 1].$$

(b) $R(x)$ is not differentiable at 0.

Take one direction, $x(t) = 0 + t e_1$, where $e_1 = (1, 0, \cdots 0)$.

$$R(x(t)) = \|x(t)\|_1 = |t|.$$

This function has left derivative -1 and right derivative 1. But not differentiable!

(c) $\partial R(x) = \prod_{k=1}^{n} \partial |x_k|$, (The Cartesian product of the subdifferential of the each $|x_k|$)

where $$\partial |x_k| = \begin{cases} \{\operatorname{sign}(x_k)\}, & x_k \neq 0 \\ [-1, 1], & x_k = 0. \end{cases}$$

To see this, for a given x,

$$\|y\|_1 \ge \|x\|_1 + \langle g, y - x \rangle \quad \forall\, y$$
$$\Leftrightarrow \sum_{k=1}^{n} |y_k| \ge \sum_{k=1}^{n} |x_k| + \sum_{k=1}^{n} g_k \cdot (y_k - x_k), \quad (*)$$

Consider coordinate-wisely, take x_1 as example, we take $y_2 = x_2 \cdots y_n = x_n$, then we have

$$|y_1| \geq |x_1| + g_1 \cdot (y_1 - x_1), \quad (**)$$

which suggests g_1 is a sub-gradient of $|x_1|$.

Same for g_k.

Conversely, if (**) holds for all $|x_k|$, we also have (*) holds, simply by adding all parts together.

Now we know from lecture

$$\partial |x_k| = \begin{cases} \{\text{sign}(x_k)\}. \\ [-1, 1]. \end{cases}$$

Q.E.D.

(d) $f(x) = \|Ax - b\|^2$, $F(x) = f(x) + \eta^{-1} \nabla R(x)$

$$\nabla F(x) = 2A^T(Ax - b) + \eta^{-1} \nabla R(x)$$

where $\nabla R(x)$ is some sub-gradient taken from the set $\partial R(x)$.

Let $\lambda = \lambda_{max}(A^TA)$.

-※- Updating rule: $K = \{x : f(x) \leq f(x_0)\}$.

$$D = \max \{\|x - x_0\| : x \in K\}.$$

$$x_{i+1} = \Pi_K\left(x_i - h_i \cdot \nabla F(x_i) / \|\nabla F(x_i)\|\right)$$

$$h_i = D/\sqrt{i}.$$

$\bar{x}_i$ be the best point up to i-th step.

Analysis: $\gamma_i = \|x_i - x^*\|_2 \le \|x_i - x_0\|_2 + \|x^* - x_0\|_2 \le 2D.$

$$\gamma_{i+1}^2 = \|x_{i+1} - x^*\|_2^2$$

$$= \left\| x_i - h_i \cdot \frac{\nabla F(x_i)}{\|\nabla F(x_i)\|} - x^* \right\|_2^2$$

$$= \|x_i - x^*\|_2^2 - 2h_i \cdot \|\nabla F(x_i)\|^{-1} \langle \nabla F(x_i), x_i - x^* \rangle + h_i^2$$

$$\le \gamma_i^2 - 2h_i \frac{F(x_i) - F(x^*)}{\|\nabla F(x_i)\|_2} + h_i^2.$$

$$\|\nabla F(x_i)\|_2 \le \|\nabla f(x_i) - \nabla f(x^*)\|_2 + \eta^{-1} \|\nabla R(x_i) - \nabla R(x^*)\|_2$$

$$\le 2 \cdot \lambda_{max}(A^T A) \cdot \|x_i - x^*\|_2 + \eta^{-1} \cdot \sqrt{n}$$

$$= 4D\lambda_{max}(A^T A) + \eta^{-1}\sqrt{n} := L$$

Let $\epsilon_i = F(x_i) - F(x^*)$.

$$2\sum_{i=M}^{N} h_i \varepsilon_i \le L(\gamma_M^2 - \gamma_{N+1}^2)$$

$$+ L\sum_{i=1}^{N} h_i^2$$

$$\le L\gamma_M^2 + L\sum_{i=1}^{N} h_i^2$$

Consider the best update so far,

$$\varepsilon_N \le \frac{L}{2}\left(\frac{4D^2 + \sum_{i=M}^{N} h_i^2}{\sum_{i=1}^{N} h_i}\right)$$

With $h_i = \frac{D}{N_i}$, $M = L\frac{N}{2}J$,

$$\varepsilon_N \lesssim \frac{LD}{N_N}.$$

More concretely,

$$F(x_N) - F(x^*) \lesssim \Omega\left(\frac{D(\eta^{-1}\sqrt{n} + 4D\lambda)}{N_N}\right).$$

$$T = \Omega\left[\left(\frac{D(\eta^{-1}\sqrt{n} + 4D\lambda)}{\varepsilon}\right)^2\right].$$

Q5. (a) Since $f(x)$ is coordinate-wise twice differentiable, $\left|\frac{\partial^2 f(x)}{\partial x_i^2}\right| \le \beta_i$, we know it's coordinate-wise Lipschitz with β_i. Hence,

$$\mathbb{E} f(x') - f(x) = \sum_{i=1}^{n} \frac{\beta_i}{B} f\left(x - \frac{1}{\beta_i}\frac{\partial f(x)}{\partial x_i} e_i\right) - f(x)$$

$$\le \sum_{i=1}^{n} \frac{\beta_i}{B}\left[-\left\langle \nabla f(x), \frac{1}{\beta_i}\frac{\nabla f(x)}{\partial x_i} e_i\right\rangle + \frac{\beta_i}{2}\cdot\frac{1}{\beta_i^2}\left\|\frac{\partial f(x)}{\partial x_i}\right\|_2^2\right]$$

$$= -\frac{1}{B}\left\langle \nabla f(x), \sum_{i=1}^{n}\frac{\nabla f(x)}{\partial x_i} e_i\right\rangle + \frac{1}{2B}\sum_{i=1}^{n}\left\|\frac{\partial f(x)}{\partial x_i}\right\|_2^2$$

$$= -\frac{1}{2B}\left\|\nabla f(x)\right\|^2$$

(b) Let $K = \{x : f(x) \le f(x_0)\}$

Update rule:

$$x_{k+1} = \Pi_K\left(x_k - \frac{1}{\beta_i}\frac{\partial f(x)}{\partial x_i} e_i\right).$$

where $\Pi_K(\cdot)$ is the projection operator, i is taken by (a).

Analysis:

Let $\Delta_k = f(x_k) - f^*$.

Note that $x_k \in K$ by projection, we have

$$\|x_k - x^*\| \le \|x_k - x_0\| + \|x^* - x_0\| \le 2D.$$

$$\Delta_k \le \langle \nabla f(x_k), x_k - x^* \rangle$$

$$\le \|\nabla f(x_k)\| \cdot \|x_k - x^*\|$$

$$\le 2D \cdot \|\nabla f(x_k)\|$$

$$\mathbb{E}(\Delta_{k+1}) = \mathbb{E}(f(x_{k+1}) - f^*)$$

$$= \mathbb{E}(\Delta_k) + \mathbb{E}\left(f(x_{k+1}) - f(x_k)\right)$$

$$= \mathbb{E}(\Delta_k) + \mathbb{E}\,\mathbb{E}\left(f(x_{k+1}) - f(x_k) \mid x_k\right)$$

$$\le \mathbb{E}(\Delta_k) - \frac{1}{\quad}\mathbb{E}\left(\|\nabla f(x_k)\|^2\right)$$

$$\leq \mathbb{E}(\Delta_k) - \frac{1}{4BD^2}\mathbb{E}(\Delta_k^2)$$

$$\leq \mathbb{E}(\Delta_k) - \frac{1}{4BD^2}[\mathbb{E}(\Delta_k)]^2$$

Dividing each side by $\mathbb{E}\Delta_k \cdot \mathbb{E}\Delta_{k+1}$,

$$\frac{1}{\mathbb{E}(\Delta_{k+1})} \geq \frac{1}{\mathbb{E}(\Delta_k)} + \frac{1}{4BD^2} \cdot \frac{\mathbb{E}[\Delta_k]}{\mathbb{E}[\Delta_{k+1}]}$$

$$\geq \frac{1}{\mathbb{E}(\Delta_k)} + \frac{1}{4BD^2} \cdot 1$$

(Since $\mathbb{E}(\Delta_k)$ is decreasing)

$$\geq \cdots\cdots$$

$$\geq \frac{1}{\Delta_0} + \frac{1}{4BD^2} \cdot (k+1)$$

That is,

$$\mathbb{E}(\Delta_T) \leq \frac{1}{\Delta_0^{-1} + (4BD^2)^{-1} \cdot T} \asymp \Omega\left(\frac{BD^2}{T}\right)$$

Hence

$$T = \Omega\left(\frac{BD^2}{\varepsilon}\right).$$