Problem 1. *Heavy-ball method*

(a) Since $\nabla f(x) = Ax + b$, the gradient method with momentum is then

$$
\begin{aligned}
x_{k+1} &= x_k - t(Ax_k + b) + s(x_k - x_{k-1}) \\
&= ((1+s)I - tA)x_k - sx_{k-1} - tb.
\end{aligned}
$$

Multiply out the linear recursion

$$
z_{k+1} = \begin{bmatrix} x_{k+1} \\ x_k \end{bmatrix}
$$

$$
Mz_k + q = \begin{bmatrix} ((1+s)I - tA)x_k - sx_{k-1} - tb \\ x_k \end{bmatrix}.
$$

We find the iteration is indeed equivalent to the linear recursion. Now suppose this recursion reaches equilibrium at $z^* = Mz^* + q$; rewrite the equilibrium condition with $z^* = \begin{bmatrix} x^* \\ y^* \end{bmatrix}$,

$$
\begin{bmatrix} x^* \\ y^* \end{bmatrix} = \begin{bmatrix} ((1+s)I - tA)x_* - sy^* - tb \\ x^* \end{bmatrix}.
$$

The second half entails $x^* = y^*$; plugging this to the first half, we get $(-tA)x^* - tb = 0$, and $x^* = -A^{-1}b$, indeed.

(b) Suppose $Ax = \lambda x$ for some eigenvalue $\lambda \in$ and eigenvector $x \neq 0$. Make a guess that eigenvectors of M might be of the form $z = \begin{bmatrix} x \\ y \end{bmatrix}$ (here y depends on x and potentially λ). Suppose $M \begin{bmatrix} x \\ y \end{bmatrix} = \nu \begin{bmatrix} x \\ y \end{bmatrix}$ for some $\nu \in$. Multiply out the expression,

$$
M \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} (1 + s - t\lambda)x - sy \\ x \end{bmatrix} = \begin{bmatrix} \nu x \\ \nu y \end{bmatrix}.
$$

This implies $x = \nu y$ and thus $(\nu^2 - (1+s-t\lambda)\nu + s)y = 0$. Solving the algebraic equation,

$$
\nu = \frac{1 + s - t\lambda \pm \sqrt{(1 + s - t\lambda)^2 - 4s}}{2}.
$$

Take the discriminant $D = (1 + s - t\lambda)^2 - 4s$, observe the following several equivalent conditions (note the assumption $t, s > 0$),

$$
\begin{align}
D &\leq 0 \tag{1} \\
(1 + s - t\lambda)^2 &\leq 4s \tag{2} \\
-2\sqrt{s} \leq 1 + s - t\lambda &\leq 2\sqrt{s} \tag{3} \\
-(1 + \sqrt{s})^2 \leq -t\lambda &\leq -(1 - \sqrt{s})^2 \tag{4} \\
(1 - \sqrt{s})^2 \leq t\lambda &\leq (1 + \sqrt{s})^2 \tag{5}
\end{align}
$$

We notice (5) is equivalent to the mentioned condition

$$\frac{(1-\sqrt{s})^2}{m} \le t \le \frac{(1+\sqrt{s})^2}{L}. \tag{6}$$

Under this condition, the eigenvalue ν of M is bound to be a complex number and

$$|\nu|^2 = \frac{1}{4}((1+s-t\lambda)^2 + 4s - (1+s-t\lambda)^2) = s.$$

We conclude that when the condition is satisfied, $\rho(M) = \max_\nu |\nu| = \sqrt{s}$.

(c) In minimizing the spectral radius $\rho(M) = \sqrt{s}$ subject to constraint (6), the two bound $(1-\sqrt{s})^2/m$ and $(1+\sqrt{s})^2/L$ eventually coincide and further yield no feasible t. The critical value is

$$\frac{(1-\sqrt{s})^2}{m} = t = \frac{(1+\sqrt{s})^2}{L}.$$

Taking square root and we get

$$\frac{1-\sqrt{s}}{\sqrt{m}} = \frac{1+\sqrt{s}}{\sqrt{L}}$$

$$\sqrt{s} = \frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}+\sqrt{m}} = \frac{\sqrt{\gamma}-1}{\sqrt{\gamma}+1}$$

The linear convergence rate of the gradient method on page 1.31 of the lecture notes is

$$c = \left(\frac{\gamma-1}{\gamma+1}\right)^2$$

We know that the quotient $\frac{\gamma-1}{\gamma+1}$ gets closer to 1 for large condition number $\gamma \in \mathbb{R}$, and by scaling γ to $\sqrt{\gamma}$ we obtain a smaller convergence rate with $\frac{\sqrt{\gamma}-1}{\sqrt{\gamma}+1}$. However, these two both converge to 1 when $\gamma \to \infty$.

Problem 2.

(a) We aim to find $g \in \mathbb{R}^n$ such that $\forall y \in \mathbb{R}^n$,

$$\sup_{t\in[0,1]} y_1 + y_2 t + \cdots + y_n t^{n-1} \ge \sup_{t\in[0,1]} x_1 + x_2 t + \cdots + x_n t^{n-1} + \sum_{i=1}^n g_i(y_i - x_i).$$

Suppose $s = \arg\max_{t\in[0,1]} x_1 + x_2 t + \cdots + x_n t^{n-1}$; take $g \in \mathbb{R}^n$ with $g_i = s^{i-1}$. Observe that indeed,

$$f(y) = \sup_{t\in[0,1]} y_1 + y_2 t + \cdots + y_n t^{n-1}$$

$$\ge y_1 + y_2 s + \cdots + y_n s^{n-1}$$

$$= x_1 + x_2 s + \cdots + x_n s^{n-1} + \sum_{i=1}^n s^{i-1}(y_i - x_i) = f(x) + g^T(y-x).$$

(b) Denote $S_x^k = \{[1], [2], \cdots, [k]\}$, the index set of the largest k elements of $x \in \mathbb{R}^n$. Take $g \in \{0,1\}^n$ with $g_i = \chi_{S_x^k}(i)$, then

$$f(y) = \text{sum of largest } k \text{ elements of } y$$

$$= \sum_{i \in S_y^k} y_i \geq \sum_{i \in S_x^k} y_i = \sum_{i \in S_x^k} x_i + \sum_{i \in S_x^k} (y_i - x_i)$$

$$= \text{sum of } k \text{ largest elements of } x + \sum_{i=1}^{n} g_i(y_i - x_i)$$

$$= f(x) + g^T(y - x).$$

(c) One known fact is that $\partial \|x\| = \{v \in V^* : \langle v, x \rangle = \|x\|, \|v\|_* \leq 1\}$; in the case of Euclidean norm $\|\cdot\|_2$,

$$\partial \|x\|_2 = \begin{cases} \{x/\|x\|_2\}, x \neq 0 \\ \{v \in \mathbb{R}^n : \|v\|_2 = 1\}, x = 0 \end{cases}$$

We observe that for any function $f : \mathbb{R}^m \to \mathbb{R}$, define $h : \mathbb{R}^n \to \mathbb{R}, h(x) = f(Ax + b)$, then

$$\partial h(x) = A^T \partial f(Ax + b).$$

To verify this, take $g \in \partial f(Ax + b)$; we should have $\forall z \in \mathbb{R}^m, f(z) \geq f(Ax + b) + g^T(z - Ax - b)$. Now for $y \in \mathbb{R}^n$,

$$g(y) = f(Ay + b) \geq f(Ax + b) + g^T(Ay + b - Ax - b)$$
$$= f(Ax + b) + (A^T g)^T(y - x) = g(x) + (A^T g)^T(y - x).$$

This confirms that $A^T g \in \partial g(x)$ indeed. Combine this observation with the additivity of subgradient, we write down the subdifferential of $f(x) = \|Ax + b\|_2 + \|x\|2$: (assuming $b \neq 0$)

$$\partial f(x) = \begin{cases} \left\{ \dfrac{A^T(Ax+b)}{\|Ax+b\|_2} + \dfrac{x}{\|x\|_2} \right\}, Ax + b \neq 0, x \neq 0 \\[2ex] \left\{ \dfrac{A^T b}{\|b\|_2} + v : v \in \mathbb{R}^n, \|v\|_2 = 1 \right\}, x = 0 \\[2ex] \left\{ A^T u + \dfrac{x}{\|x\|_2} : u \in \mathbb{R}^m, \|u\|_2 = 1 \right\}, Ax + b = 0, x \neq 0 \end{cases}$$

I'm skipping to attach the close form for the case that $b = 0$, but it should be very easy to write down from $\partial f(x) = A^T \partial \|Ax + b\|_2 + \partial \|x\|_2$.

(d) Note that for any symmetric $W \in \mathbf{S}^n$,

$$\lambda_{\max}(W) = \max_{\|u\|=1} u^T W u.$$

3

This identity holds true after adding $\mathbf{diag}(x)$ for $x \in \mathbb{R}^n$ as well. Now suppose

$$v = \arg\max_{\|u\|=1} u^T(W + \mathbf{diag}(x))u,$$

take $g \in \mathbb{R}^n$ with $g_i = v_i^2$, then verify that, indeed,

$$\lambda_{\max}(W + \mathbf{diag}(y)) = \max_{\|u\|=1} u^T(W + \mathbf{diag}(y))u \geq v^T(W + \mathbf{diag}(y))v$$

$$= v^T(W + \mathbf{diag}(x))v + \sum_{i=1}^n v_i^2(y_i - x_i)$$

$$= \lambda_{\max}(W + \mathbf{diag}(x)) + g^T(y - x).$$

(e) Suppose

$$u = \arg\max_{Ay \preceq b} z^T y.$$

Take $g = u \in \mathbb{R}^n$; verify that, indeed,

$$f(x) = \sup_{Ay \preceq b} x^T y \geq x^T u = z^T u + u^T(z - x)$$

$$= \sup_{Ay \preceq b} z^T y + g^T(z - x) = f(z) + g^T(z - x).$$