ECE 598: Interplay between Control and Machine Learning **Fall 2020**

Homework 1

Instructor: Bin Hu Due date: September 24, 2020

Each problem is worth 5 points, and the total points in HW1 is 15.

1. In this problem, you are asked to implement a few LMIs.
(a) Consider a dynamical system $x_{k+1} = Ax_k$ where

$$A = \begin{bmatrix} -0.5 & -0.85 & 0.33 \\ -0.23 & 0.71 & 0.79 \\ 0.41 & -0.1 & -0.21 \end{bmatrix}.$$

You want to test the stability of this system by using the following SDP

$$P > 0$$
$$A^\mathsf{T} P A - P < 0$$

Your task is to implement and test this SDP. You can use any software of your choosing, e.g. CVX or LMILab. Notice CVX strongly discourages strict inequalities. In addition, if P is a solution for the above LMI, cP is also a solution for any $c > 0$. Hence the LMI is homogeneous. We need to break the homogeneity so that the numerical solver can give us one solution. One way to do this is to replace the original LMI as

$$P \geq \varepsilon I$$
$$A^\mathsf{T} P A - P \leq -\varepsilon I$$

where ε is some small positive number. Another way of doing things is enforcing the trace of P to be a positive constant while using $P \geq 0$ and $A^\mathsf{T} P A - P \leq 0$. A sample code for this LMI implementation in CVX is provided in the course website.

(b) Still consider the same dynamical system in (a). We know the system converges at a rate ρ if there exists $P > 0$ such that $A^\mathsf{T} P A - \rho^2 P \leq 0$. When ρ^2 is fixed, this is a SDP. Hence you can use the code to find the smallest ρ such that the LMI is feasible. Find that value of ρ and compare it with the spectral radius of A. Are they the same?

(c) Now consider the gradient method $x_{k+1} = x_k - \alpha \nabla f(x_k)$ and assume f is L-smooth and m-strongly convex. In the class, we have shown that the gradient method converges at the rate ρ if there exists non-negative λ such that

$$\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} \leq \lambda \begin{bmatrix} 2mL & -(m+L) \\ -(m+L) & 2 \end{bmatrix}$$

The above LMI is already not homogeneous (λ being a solution does not mean $c\lambda$ being a solution) so you can directly implement it. Given the values of (m, L) (let's say set (m, L) to be $(1, 10)$, $(1, 100)$, $(0.1, 23)$, $(1, 1000)$, and $(1, 50000)$), try a few values of α satisfying $\alpha < \frac{2}{L}$ and find the smallest ρ^2 such that the above LMI is feasible. Is ρ equal to $\max\{|1 - m\alpha|, |1 - L\alpha|\}$? What happens to the simulation when $\frac{L}{m}$ is really large?

2. In this problem, you will be asked to test a few LMIs for stochastic gradient descent (SGD) and SAGA. In the class, we talked about how to analyze SGD under the assumptions that f is m-strongly convex and f_i is L-smooth and convex. Here you are asked to analyze SGD under a different set of assumptions:

- f satisfies the "one-point convexity" condition: f has a unique global minimizer x^* and for all x, one has $(x - x^*)^\mathsf{T} \nabla f(x) \geq m\|x - x^*\|^2$ with $m > 0$.
- f_i is L-smooth for all i.

Under the above assumptions, f and f_i are not convex in general. However, you can still obtain a convergence bound for SGD under these assumptions.

(a) Suppose $v_k = \nabla f_{i_k}(w_k)$. Based on the above assumptions, the following two inequalities hold

$$\mathbb{E}\begin{bmatrix} v_k - x^* \\ w_k \end{bmatrix}^\mathsf{T} \begin{bmatrix} -2L^2 I & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} v_k - x^* \\ w_k \end{bmatrix} \leq \frac{2}{n}\sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2 = M$$

$$\mathbb{E}\begin{bmatrix} v_k - x^* \\ w_k \end{bmatrix}^\mathsf{T} \begin{bmatrix} 2mI & -I \\ -I & 0 \end{bmatrix} \begin{bmatrix} v_k - x^* \\ w_k \end{bmatrix} \leq 0$$

Use the above inequalities to prove that if there exists non-negative λ_1 and λ_2 such that

$$\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} - \lambda_1 \begin{bmatrix} -2L^2 & 0 \\ 0 & 1 \end{bmatrix} - \lambda_2 \begin{bmatrix} 2m & -1 \\ -1 & 0 \end{bmatrix} \leq 0, \tag{1}$$

then SGD satisfies the bound

$$\mathbb{E}\|x_k - x^*\|^2 \leq \rho^{2k}\mathbb{E}\|x_0 - x^*\|^2 + \frac{\lambda_1 M}{1-\rho^2}$$

(b) Use the above LMI condition to show that SGD satisfies the following bound:

$$\mathbb{E}\|x_k - x^*\|^2 \leq (1 - 2m\alpha + 2L^2\alpha^2)^k \mathbb{E}\|x_0 - x^*\|^2 + \frac{\alpha^2 M}{2m\alpha - 2L^2\alpha^2}$$

(c) Finally, analyze SAGA under the assumption that f_i is L-smooth and m-strongly convex for all i. Under this assumption, f is also L-smooth and m-strongly convex. Suppose a uniform sampling is used, i.e. $p_i = \frac{1}{n}$. Implement the LMI in Lecture 9. Set $n = 20$. Set (m, L) to be $(1, 10)$ and $(1, 100)$. Set $\alpha = \frac{1}{3L}$ and $\rho^2 = 1 - \min\{\frac{1}{3n}, \frac{m}{3L}\}$. Is the LMI always feasible? Is there any structure in P and λ_j?

3. In this problem, you will be asked to perform several calculations, and these calculations eventually lead to the convergence rate proof for Nesterov's accelerated method applied to smooth strongly-convex objective functions. Recall Nesterov's method is

$$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha \nabla f((1+\beta)x_k - \beta x_{k-1})$$

which can also be written as

$$\begin{aligned} \xi_{k+1} &= A\xi_k + Bw_k \\ v_k &= C\xi_k \\ w_k &= \nabla f(v_k) \end{aligned}$$

where $A = \begin{bmatrix} (1+\beta)I & -\beta I \\ I & 0 \end{bmatrix}$, $B = \begin{bmatrix} -\alpha I \\ 0 \end{bmatrix}$, $C = \begin{bmatrix} (1+\beta)I & -\beta I \end{bmatrix}$, and $\xi_k = \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix}$.

(a) Assume f is L-smooth and m-strongly convex. By L-smoothness and m-strong convexity, we have

$$\begin{aligned} f(x_k) - f(x_{k+1}) &= f(x_k) - f(v_k) + f(v_k) - f(x_{k+1}) \\ &\geq \nabla f(v_k)^\mathsf{T}(x_k - v_k) + \frac{m}{2}\|x_k - v_k\|^2 + \nabla f(v_k)^\mathsf{T}(v_k - x_{k+1}) - \frac{L}{2}\|v_k - x_{k+1}\|^2 \\ &= \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \\ \nabla f(v_k) \end{bmatrix}^\mathsf{T} X_1 \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \\ \nabla f(v_k) \end{bmatrix} \end{aligned}$$

The last step in the above derivation requires substituting $x_{k+1} = (1+\beta)x_k - \beta x_{k-1} - \alpha \nabla f(v_k)$ and $v_k = C\xi_k$ into the second-to-last line $\nabla f(v_k)^\mathsf{T}(x_k - v_k) + \frac{m}{2}\|x_k - v_k\|^2 + \nabla f(v_k)^\mathsf{T}(v_k - x_{k+1}) - \frac{L}{2}\|v_k - x_{k+1}\|^2$ and rewriting the resultant quadratic function. You task is figuring out this symmetric matrix X_1 (actually we have already done this in the class).

(b) Similarly, by L-smoothness and m-strong convexity, we have

$$\begin{aligned} f(x^*) - f(x_{k+1}) &= f(x^*) - f(v_k) + f(v_k) - f(x_{k+1}) \\ &\geq \nabla f(v_k)^\mathsf{T}(x^* - v_k) + \frac{m}{2}\|x^* - v_k\|^2 + \nabla f(v_k)^\mathsf{T}(v_k - x_{k+1}) - \frac{L}{2}\|v_k - x_{k+1}\|^2 \\ &= \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \\ \nabla f(v_k) \end{bmatrix}^\mathsf{T} X_2 \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \\ \nabla f(v_k) \end{bmatrix} \end{aligned}$$

The last step in the above derivation requires substituting $x_{k+1} = (1+\beta)x_k - \beta x_{k-1} - \alpha \nabla f(v_k)$ and $v_k = C\xi_k$ into the second-to-last line $\nabla f(v_k)^\mathsf{T}(x^* - v_k) + \frac{m}{2}\|x^* - v_k\|^2 + \nabla f(v_k)^\mathsf{T}(v_k - x_{k+1}) - \frac{L}{2}\|v_k - x_{k+1}\|^2$ and rewriting the resultant quadratic function. You task is figuring out this symmetric matrix X_2.

(c) Now based on the inequalities in (a) and (b), you can simply choose $X = \rho^2 X_1 + (1-\rho^2)X_2$ for any $0 < \rho < 1$, and we have

$$\begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \\ \nabla f(v_k) \end{bmatrix}^\mathsf{T} X \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \\ \nabla f(v_k) \end{bmatrix} \le -(f(x_{k+1}) - f(x^*)) + \rho^2(f(x_k) - f(x^*))$$

Based on the testing condition presented in the class, if there exists $P \ge 0$ such that

$$\begin{bmatrix} A^\mathsf{T}PA - \rho^2 P & A^\mathsf{T}PB \\ B^\mathsf{T}PA & B^\mathsf{T}PB \end{bmatrix} - X \le 0 \tag{2}$$

then the following inequality holds

$$\begin{aligned}(\xi_{k+1} - \xi^*)^\mathsf{T}P(\xi_{k+1} - \xi^*) - \rho^2(\xi_k - \xi^*)^\mathsf{T}P(\xi_k - \xi^*) &\le \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \\ \nabla f(v_k) \end{bmatrix}^\mathsf{T} X \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \\ \nabla f(v_k) \end{bmatrix} \\ &\le -(f(x_{k+1}) - f(x^*)) + \rho^2(f(x_k) - f(x^*))\end{aligned}$$

which directly leads to the linear convergence rate for Nesterov's method:

$$(\xi_{k+1} - \xi^*)^\mathsf{T}P(\xi_{k+1} - \xi^*) + f(x_{k+1}) - f(x^*) \le \rho^2\left((\xi_k - \xi^*)^\mathsf{T}P(\xi_k - \xi^*) + f(x_k) - f(x^*)\right).$$

Finding P to satisfy (2) is not trivial. Your task is to implement the above LMI and test its feasibility with $\rho^2 = 1 - \sqrt{\frac{m}{L}}$, $\alpha = \frac{1}{L}$, and $\beta = \frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}+\sqrt{m}}$. Again, try a few different values of (m, L) and see what happens to the numerical solution of P. Also, is there any pattern in the resultant matrix on the left side of (2)?

(d) Your task is to prove that (2) holds for $P = \frac{1}{2}\begin{bmatrix} \sqrt{L}I \\ (\sqrt{m} - \sqrt{L})I \end{bmatrix}\begin{bmatrix} \sqrt{L}I & (\sqrt{m} - \sqrt{L})I \end{bmatrix} \ge 0$, $\rho^2 = 1 - \sqrt{\frac{m}{L}}$, and $X = \rho^2 X_1 + (1-\rho^2)X_2$ (X_1 and X_2 are the answers you get in (a) and (b)) when $\alpha = \frac{1}{L}$ and $\beta = \frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}+\sqrt{m}}$.

(Hint: The calculation here can be lengthy. So you are allowed to use some symbolic toolbox to help as long as you turn in the code.)