ECE 490: Introduction to Optimization **Fall 2018**

Lecture 6

Unconstrained Optimization for Smooth Strongly-Convex Functions, Part VI

Lecturer: Bin Hu, Date:09/18/2018

In the last lecture, we talked about the dissipation inequality approach that can be used as a general analysis tool for the first-order optimization methods. In this lecture, we routinize the dissipation inequality approach. We talk about a routine that you can follow when analyzing first-order methods described by the following model:

$$\begin{aligned} \xi_{k+1} &= A\xi_k + Bu_k \\ v_k &= C\xi_k \\ u_k &= \nabla f(v_k) \end{aligned} \tag{6.1}$$

Our routine includes the following steps:

1. Replace the nonlinear equation $u_k = \nabla f(v_k)$ in (6.1) by some quadratic inequality in the following form:

$$\begin{bmatrix} \xi_k - \xi^* \\ u_k \end{bmatrix}^{\mathsf{T}} X \begin{bmatrix} \xi_k - \xi^* \\ u_k \end{bmatrix} \leq a(f(x_{k+1}) - f(x^*)) + b(f(x_k) - f(x^*)) \tag{6.2}$$

   where (a, b) are set to be $(0, 0)$ or $(-1, \rho^2)$. Notice (6.2) is required to hold for any (ξ_k, v_k, u_k) satisfying (6.1), and hence the construction of (6.2) relies on the properties of f. Notice we have

$$\{(\xi_k, u_k, v_k) \text{ satisfying (6.1)}\} \subset \{\xi_{k+1} = A\xi_k + Bu_k, v_k = C\xi_k, \ (\xi_k, u_k) \text{ satisfying (6.2)}\}$$

   If we can show ξ_k converges to ξ^* at some rate ρ for any (ξ_k, u_k) satisfying (6.2) and $\xi_{k+1} = A\xi_k + Bu_k$, then we guarantee the original first-order method (6.1) to converge to $(\xi^*, 0, v^*)$ [1] at the same linear rate ρ.

2. Test if there exists $P \geq 0$ such that

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

   If so, then the following inequality holds

$$(\xi_{k+1} - \xi^*)^{\mathsf{T}} P(\xi_{k+1} - \xi^*) - \rho^2(\xi_k - \xi^*)^{\mathsf{T}} P(\xi_k - \xi^*) \leq \begin{bmatrix} \xi_k - \xi^* \\ u_k \end{bmatrix}^{\mathsf{T}} X \begin{bmatrix} \xi_k - \xi^* \\ u_k \end{bmatrix}$$

[1] Recall that we have $v^* = C\xi^*$.

which is exactly the so-called dissipation inequality $V_{k+1}-\rho^2 V_k \leq S(\xi_k, u_k)$ if we define $V_k = (\xi_k - \xi^*)^\mathsf{T} P(\xi_k - \xi^*)$ and $S(\xi_k, u_k) = \begin{bmatrix} \xi_k - \xi^* \\ u_k \end{bmatrix}^\mathsf{T} X \begin{bmatrix} \xi_k - \xi^* \\ u_k \end{bmatrix}$. Clearly $V_k \geq 0$ due to the fact $P \geq 0$.

3. Now directly apply the supply rate condition (6.2) to conclude the linear convergence rate from the dissipation inequality. Depending on the (a, b), two types of linear convergence rate results may be obtained.

   - If $a = b = 0$, then $S(\xi_k, u_k) \leq 0$. The dissipation inequality $V_{k+1} - \rho^2 V_k \leq S(\xi_k, u_k)$ directly leads to $V_{k+1} \leq \rho^2 V_k$ and $V_k \leq \rho^{2k} V_0$. In previous lectures, the convergence rate of the gradient method was proved using such an argument.
   - If $a = -1$, $b = \rho^2$, then the supply rate condition becomes $S(\xi_k, u_k) \leq -(f(x_{k+1}) - f(x^*)) + (f(x_k) - f(x^*))$. Consequently, the dissipation inequality $V_{k+1} - \rho^2 V_k \leq S(\xi_k, u_k)$ leads to $V_{k+1} - \rho^2 V_k \leq -(f(x_{k+1}) - f(x^*)) + (f(x_k) - f(x^*))$. This is exactly a linear convergence rate result: $V_{k+1} + f(x_{k+1}) - f(x^*) \leq \rho^2 \left(V_k + f(x_k) - f(x^*)\right)$. The convergence rate proof for Nesterov's method can be done using such an argument.

   Typically the second type of results is much more involved than the first type of results since the construction of (6.2) can be more difficult. Once the linear convergence rate result is obtained, it is natural to convert the value of ρ^2 into an iteration complexity result $T = O(\frac{1}{1-\rho^2} \log(\frac{1}{\varepsilon}))$.

Now we flesh out more details of the above three-step routine.

6.1 Proof for the Testing Condition in the Second Step

The formal statement for the testing condition given in Step 2 is given as follows. The proof is based on a standard manipulation in the controls literature.

Theorem 1. *Suppose $\xi_{k+1} = A\xi_k + Bu_k$ and $\xi^* = A\xi^*$. If there exists a positive semidefinite matrix $P \in \mathbb{R}^{n_\xi \times n_\xi}$ s.t.*

$$\begin{bmatrix} A^\mathsf{T} P A - \rho^2 P & A^\mathsf{T} P B \\ B^\mathsf{T} P A & B^\mathsf{T} P B \end{bmatrix} - X \leq 0 \tag{6.3}$$

then the following inequality holds

$$(\xi_{k+1} - \xi^*)^\mathsf{T} P (\xi_{k+1} - \xi^*) - \rho^2 (\xi_k - \xi^*)^\mathsf{T} P (\xi_k - \xi^*) \leq \begin{bmatrix} \xi_k - \xi^* \\ u_k \end{bmatrix}^\mathsf{T} X \begin{bmatrix} \xi_k - \xi^* \\ u_k \end{bmatrix}.$$

Proof: Based on (6.3), we have

$$\begin{bmatrix}\xi_k-\xi^*\\ u_k\end{bmatrix}^{\mathsf{T}}\left(\begin{bmatrix}A^{\mathsf{T}}PA-\rho^2P & A^{\mathsf{T}}PB\\ B^{\mathsf{T}}PA & B^{\mathsf{T}}PB\end{bmatrix}-X\right)\begin{bmatrix}\xi_k-\xi^*\\ u_k\end{bmatrix}\leq 0 \tag{6.4}$$

which is equivalent to

$$\begin{bmatrix}\xi_k-\xi^*\\ u_k\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}A^{\mathsf{T}}PA-\rho^2P & A^{\mathsf{T}}PB\\ B^{\mathsf{T}}PA & B^{\mathsf{T}}PB\end{bmatrix}\begin{bmatrix}\xi_k-\xi^*\\ u_k\end{bmatrix}-\begin{bmatrix}\xi_k-\xi^*\\ u_k\end{bmatrix}^{\mathsf{T}}X\begin{bmatrix}\xi_k-\xi^*\\ u_k\end{bmatrix}\leq 0 \tag{6.5}$$

Now we calculate the first term

$$\begin{aligned}
&\begin{bmatrix}\xi_k-\xi^*\\ u_k\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}A^{\mathsf{T}}PA-\rho^2P & A^{\mathsf{T}}PB\\ B^{\mathsf{T}}PA & B^{\mathsf{T}}PB\end{bmatrix}\begin{bmatrix}\xi_k-\xi^*\\ u_k\end{bmatrix}\\
=&\begin{bmatrix}\xi_k-\xi^*\\ u_k\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}A^{\mathsf{T}}PA & A^{\mathsf{T}}PB\\ B^{\mathsf{T}}PA & B^{\mathsf{T}}PB\end{bmatrix}\begin{bmatrix}\xi_k-\xi^*\\ u_k\end{bmatrix}-\begin{bmatrix}\xi_k-\xi^*\\ u_k\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}\rho^2P & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\xi_k-\xi^*\\ u_k\end{bmatrix}\\
=&\begin{bmatrix}\xi_k-\xi^*\\ u_k\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}A^{\mathsf{T}}\\ B^{\mathsf{T}}\end{bmatrix}P\begin{bmatrix}A & B\end{bmatrix}\begin{bmatrix}\xi_k-\xi^*\\ u_k\end{bmatrix}-\rho^2(\xi_k-\xi^*)^{\mathsf{T}}P(\xi_k-\xi^*)\\
=&(A(\xi_k-\xi^*)+Bu_k)^{\mathsf{T}}P(A(\xi_k-\xi^*)+Bu_k)-\rho^2(\xi_k-\xi^*)^{\mathsf{T}}P(\xi_k-\xi^*)\\
=&(\xi_{k+1}-\xi^*)^{\mathsf{T}}P(\xi_{k+1}-\xi^*)-\rho^2(\xi_k-\xi^*)^{\mathsf{T}}P(\xi_k-\xi^*)
\end{aligned}$$

Therefore, the first term in (6.5) is just $(\xi_{k+1}-\xi^*)^{\mathsf{T}}P(\xi_{k+1}-\xi^*)-\rho^2(\xi_k-\xi^*)^{\mathsf{T}}P(\xi_k-\xi^*)$, and we have the desired inequality. ■

6.2 Example: Gradient Method

In Lecture 2, we have presented a testing condition for the gradient method. If there exists $0<\rho<1$ and $\lambda\geq 0$ such that

$$\begin{bmatrix}1-\rho^2 & -\alpha\\ -\alpha & \alpha^2\end{bmatrix}+\lambda\begin{bmatrix}-2mL & m+L\\ m+L & -2\end{bmatrix}\leq 0, \tag{6.6}$$

then the gradient method satisfies $\|x_k-x^*\|\leq\rho^k\|x_0-x^*\|$. In HW1, we have shown the convergence rate for the gradient method by applying this condition. This exactly follows our proposed analysis routine. We did the analysis in three steps:

1. We replace the nonlinear equation $u_k=\nabla(v_k)$ by (1b) in HW1:

$$\begin{bmatrix}v_k-v^*\\ \nabla f(v_k)\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}-2mLI & (m+L)I\\ (m+L)I & -2I\end{bmatrix}\begin{bmatrix}v_k-v^*\\ \nabla f(v_k)\end{bmatrix}\geq 0. \tag{6.7}$$

Notice $v_k = \xi_k$ and $u_k = \nabla f(v_k)$. The above inequality is just a supply rate condition in the form of (6.2):

$$\begin{bmatrix} \xi_k - \xi^* \\ u_k \end{bmatrix}^\mathsf{T} \begin{bmatrix} 2mLI & -(m+L)I \\ -(m+L)I & 2I \end{bmatrix} \begin{bmatrix} \xi_k - \xi^* \\ u_k \end{bmatrix} \leq 0.$$

Specifically, we have $X = \begin{bmatrix} 2mLI & -(m+L)I \\ -(m+L)I & 2I \end{bmatrix}$, $a = 0$, and $b = 0$.

2. Next, we apply the general testing condition. We have $A = I$, $B = -\alpha I$, and $X = \begin{bmatrix} 2mLI & -(m+L)I \\ -(m+L)I & 2I \end{bmatrix}$. We parameterize P as $P = \frac{1}{\lambda} I$. Then the testing condition $\begin{bmatrix} A^\mathsf{T}PA - \rho^2 P & A^\mathsf{T}PB \\ B^\mathsf{T}PA & B^\mathsf{T}PB \end{bmatrix} - X \leq 0$ becomes $\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda \begin{bmatrix} -2mL & m+L \\ m+L & -2 \end{bmatrix} \leq 0$ (verify this!). In (2b) of HW1, you have solved this condition.

3. Now we have $a = b = 0$. Hence the dissipation inequality directly leads to $\|\xi_k - \xi^*\| \leq \rho^k \|\xi_0 - \xi^*\|$. The key fact we use here is $(\xi_k - \xi^*)^\mathsf{T} P (\xi_k - \xi^*) = \frac{1}{\lambda}\|\xi_k - \xi^*\|^2$.

6.3 Difficulty in the Rate Proof of Nesterov's Method

The main difficulty in the rate proof of Nesterov's method is that we need a more complicated X. A natural extension of the analysis for the gradient method is that we still use (6.7) to replace $u_k = \nabla f(v_k)$. Since $v_k = C\xi_k$, we have

$$\begin{bmatrix} C\xi_k - C\xi^* \\ u_k \end{bmatrix}^\mathsf{T} \begin{bmatrix} 2mLI & -(m+L)I \\ -(m+L)I & 2I \end{bmatrix} \begin{bmatrix} C\xi_k - C\xi^* \\ u_k \end{bmatrix} \leq 0$$

which is equivalent to

$$\begin{bmatrix} \xi_k - \xi^* \\ u_k \end{bmatrix}^\mathsf{T} \begin{bmatrix} C^\mathsf{T} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} 2mLI & -(m+L)I \\ -(m+L)I & 2I \end{bmatrix} \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \xi_k - \xi^* \\ u_k \end{bmatrix} \leq 0 \tag{6.8}$$

It seems that we can choose $X = \begin{bmatrix} C^\mathsf{T} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} 2mLI & -(m+L)I \\ -(m+L)I & 2I \end{bmatrix} \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix}$ and have $S(\xi_k, u_k) \leq 0$. Then we can use this X and (A, B) to formulate a testing condition. However, this testing condition is not feasible for $\rho^2 = 1 - \sqrt{\frac{m}{L}}$. It means that (6.7) is too conservative to give a good estimate of $u_k = \nabla f(v_k)$ in this case. Notice the key idea behind our analysis routine is to estimate $u_k = \nabla f(v_k)$ using some inequality in the form of (6.2). The X matrix working for one method may not work for another method. We will present some general guidelines for choosing X and discuss how to choose X for Nesterov's method and other methods in later lectures.

6.4 When to determine (a, b)?

The values of (a, b) are decided in the first step of the routine. Basically we construct X such that (6.2) holds with either $(a, b) = (0, 0)$ or $(a, b) = (-1, \rho^2)$. In Step 3, we are only using the values of (a, b).