ECE 598: Interplay between Control and Machine Learning **Spring 2019**

Lecture 10

Lure-Postnikov Lyapunov Functions

Lecturer: Bin Hu, Date:02/19/2019

The gradient method and SAGA are relatively easy to analyze since they only require quadratic Lyapunov functions and pointwise quadratic constraints. Specifically, we only need to construct a dissipation inequality in the form of $V(\xi_{k+1}) - \rho^2 V(\xi_k) \leq S(\xi_k, w_k)$ (or an expected version) with V being quadratic and S being smaller or equal to 0 for all k. Then we immediately have $V(\xi_{k+1}) \leq \rho^2 V(\xi_k)$. As a matter of fact, one can prove the linear convergence of both methods by only exploiting the one-point convexity of f. The convexity of f is not really needed. That is not the case for Nesterov's method and SAG.

Nesterov's method and SAG are more difficult to analyze. One reason is that more advanced Lyapunov functions and more sophisticated supply rates are required to exploit the properties of f. Just imagine that we still use the sector bound condition to analyze Nesterov's method. Hence we can set $X = \begin{bmatrix} C^\mathsf{T} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} 2mLI & -(m+L)I \\ -(m+L)I & 2I \end{bmatrix} \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix}$ and have $S(\xi_k, w_k) \leq 0$. If we use this X to formulate the LMI, we will see that the testing condition is not feasible for the rate we want to test. Therefore, the sector bound condition is too conservative for Nesterov's method. Notice the key idea behind the dissipation inequality framework is to approximate $w_k = \nabla f(v_k)$ using some supply rate conditions. For Nesterov's method, we will need some supply rate conditions that can exploit the convexity better and give us Lyapunov functions in more general forms. In this lecture, we will talk about the Lure-Postnikov Lyapunov function approach for Nesterov's method and SAG.

10.1 Iteration Complexity of Nesterov's Method

The convergence rate ρ naturally leads to an iteration number T guaranteeing the algorithm to achieve the so-called ε-optimality, i.e. $\|x_T - x^*\|^2 \leq \varepsilon$ or $f(x_T) - f(x^*) \leq \varepsilon$.

Based on the rate bound $c\rho^k$, if we choose $T = \log\left(\frac{c}{\varepsilon}\right)/(-\log\rho) = O\left(\log(\frac{c}{\varepsilon})/(1-\rho)\right)$, we guarantee the ε-optimal solution. The number T just gives the "ε-optimal iteration complexity". It is straightforward to verify that the convergence rate bound that we obtained for the gradient method can be converted to an iteration complexity $T = O\left(\frac{L}{m}\log(\frac{1}{\varepsilon})\right)$.

Nesterov's method improves the iteration complexity from $O\left(\frac{L}{m}\log(\frac{1}{\varepsilon})\right)$ to $O\left(\sqrt{\frac{L}{m}}\log(\frac{1}{\varepsilon})\right)$. This improvement is significant. Just consider $\frac{L}{m} = 10000$. Then $\sqrt{\frac{L}{m}} = 100$. Hence Nesterovs method is roughly 100 times faster than the gradient method in this case. The convergence rate corresponding to this iteration complexity is $\rho^2 = 1 - \sqrt{\frac{m}{L}}$. When f is L-smooth and m-strongly convex, Nesterov's method with $\alpha = \frac{1}{L}$ and $\beta = \frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}+\sqrt{m}}$ satisfies a convergence bound $f(x_k) - f(x^*) \leq c\left(1 - \sqrt{\frac{m}{L}}\right)^k$ where c is a constant. If we only use $X = \begin{bmatrix} C^\mathsf{T} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} 2mLI & -(m+L)I \\ -(m+L)I & 2I \end{bmatrix} \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix}$ to form the supply rate function, the resultant LMI is not feasible with $\rho^2 = 1 - \sqrt{\frac{m}{L}}$. Now we will show how to analyze Nesterov's method by modifying the dissipation inequality and constructing the so-called Lure-Postnikov Lyapunov function.

10.2 Lure-Postnikov Lyapunov Functions

Although quadratic Lyapunov functions are not enough for proving the accelerated rate of Nesterov's method, we can use a Lyapunov function in the form of $(\xi_k - \xi^*)^\mathsf{T} P(\xi_k - \xi^*) + f(x_k) - f(x^*)$ to fix the issue. This type of Lyapunov functions are exactly the so-called "Lure-Postnikov Lyapunov functions" in the controls literature. The quadratic term $(\xi_k - \xi^*)^\mathsf{T} P(\xi_k - \xi^*)$ can be thought as a kinetic energy and the function term $f(x_k) - f(x^*)$ can be interpreted as a potential energy. For Nesterov's method, one can show that the total energy (or Hamiltonian) decreases at every step although the kinetic energy itself may not decrease in that way.

How to construct a Lure-Postnikov Lyapunov function? The answer is using a new supply rate! Suppose we can construct a symmetric matrix X such that the following supply rate condition holds

$$\begin{aligned} S(\xi_k, w_k) &= \begin{bmatrix} \xi_k - \xi^* \\ w_k \end{bmatrix}^\mathsf{T} X \begin{bmatrix} \xi_k - \xi^* \\ w_k \end{bmatrix} \\ &\leq -(f(x_{k+1}) - f(x^*)) + \rho^2(f(x_k) - f(x^*)) \\ &= \rho^2(f(x_k) - f(x_{k+1})) + (1-\rho^2)(f(x^*) - f(x_{k+1})), \end{aligned} \tag{10.1}$$

then the dissipation inequality $V(\xi_{k+1}) - \rho^2 V(\xi_k) \leq S(\xi_k, w_k)$ will directly leads to the desired convergence bound $V(\xi_{k+1}) + f(x_{k+1}) - f(x^*) \leq \rho^2(V(\xi_k) + f(x_k) - f(x^*))$. The key issue is how to figure out X. If we can find X_1 and X_2 such that

$$\begin{bmatrix} \xi_k - \xi^* \\ w_k \end{bmatrix}^\mathsf{T} X_1 \begin{bmatrix} \xi_k - \xi^* \\ w_k \end{bmatrix} \leq f(x_k) - f(x_{k+1}) \tag{10.2}$$

$$\begin{bmatrix} \xi_k - \xi^* \\ w_k \end{bmatrix}^\mathsf{T} X_2 \begin{bmatrix} \xi_k - \xi^* \\ w_k \end{bmatrix} \leq f(x^*) - f(x_{k+1}), \tag{10.3}$$

then we can set $X = \rho^2 X_1 + (1-\rho^2) X_2$ to obtain the condition (10.1). Now let's look at how to obtain X_1 and X_2.

For illustrative purposes, we focus on the construction of X_1. The construction of X_2 will be similar. The condition (10.2) involves $f(x_{k+1})$ and $f(x_k)$. Hence it is reasonable to think that its construction requires some inequalities involving the function value f. Recall that L-smoothness and m-strong convexity give the following two inequalities:

$$f(x) \leq f(y) + \nabla f(y)^\mathsf{T}(x-y) + \frac{L}{2}\|x-y\|^2 \tag{10.4}$$

$$f(x) \geq f(y) + \nabla f(y)^\mathsf{T}(x-y) + \frac{m}{2}\|x-y\|^2 \tag{10.5}$$

How can we choose (x, y) in the above inequalities to obtain (10.2). One idea is to set $(x, y) \to (x_{k+1}, x_k)$ in (10.4). However, $\nabla f(y)$ becomes $\nabla f(x_k)$ and this is not w_k! The only term involving the gradient information on the left side of (10.2) is w_k which is the gradient evaluated on v_k! Therefore, when applying (10.4) and (10.5) to construct (10.2), one has to set y to be v_k! By doing this, we can show

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

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. This gives us the matrix X_1. We can see that the key trick is just subtracting and adding $f(v_k)$.

Similarly, X_2 can be derived as

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

You will be asked to figure out the details of X_2 in the homework.

Now we have the matrix X that ensures the desired supply rate condition (10.1). We are ready to apply our dissipation inequality approach to analyze Nesterov's method. All we need to do is to test if there exists $P \geq 0$ such that

$$\begin{bmatrix} A^\mathsf{T}PA - \rho^2 P & A^\mathsf{T}PB \\ B^\mathsf{T}PA & B^\mathsf{T}PB \end{bmatrix} - X \leq 0. \tag{10.6}$$

If so, then the following inequality holds

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

which is exactly the desired dissipation inequality $V(\xi_{k+1})-\rho^2 V(\xi_k)\le S(\xi_k,w_k)$ if we define $V(\xi_k)=(\xi_k-\xi^*)^\mathsf{T} P(\xi_k-\xi^*)$ and $S(\xi_k,w_k)=\begin{bmatrix}\xi_k-\xi^*\\ w_k\end{bmatrix}^\mathsf{T} X\begin{bmatrix}\xi_k-\xi^*\\ w_k\end{bmatrix}$. Clearly $V_k\ge 0$ due to the fact $P\ge 0$. Then the dissipation inequality and the supply rate condition together will give us the desired Lure-Postnikov Lyapunov function for Nesterov's method. In the homework, you will be asked to test the above LMI. I will also provide the analytical formula of P and you will be asked to analytically verify that (10.6) holds with that P and $(\rho^2,\alpha,\beta)=(1-\sqrt{\frac{m}{L}},\frac{1}{L},\frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}+\sqrt{m}})$.

Comments on SAG. Quadratic Lyapunov functions and the pointwise quadratic constraints developed in the last lecture are also not enough for proving the convergence rate of SAG. However, the analysis of SAG can also be addressed by using the Lure-Postnikov Lyapunov functions. Recall that SAG can be represented as $F_u(G,\Delta)$ where G is a jump system and the operator Δ maps v to w as

$$w_k=\begin{bmatrix}\nabla f_1(v_k)\\ \nabla f_2(v_k)\\ \vdots\\ \nabla f_n(v_k)\end{bmatrix} \tag{10.7}$$

For this operator Δ, we can use similar tricks (adding and subtracting $f(v_k)$) to construct a desired supply rate condition which will eventually gives us the Lure-Postnikov Lyapunov function. Actually the original convergence rate proof of SAG is based on a similar idea (although the Lure-Postnikov Lyapunov function construction there is not based on dissipation inequality).