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

Lecture 5

Lur'e-Postnikov Lyapunov Functions

Lecturer: Bin Hu, Date:09/15/2020

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.

5.1 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 $\beta = \frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}+\sqrt{m}}$ sat-

isfies 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.

5.2 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? Apply the dissipation inequality approach with a new supply rate! Recall that Nesterov's method can be written as

$$\begin{aligned} \xi_{k+1} &= A\xi_k + Bu_k \\ v_k &= C\xi_k \\ u_k &= \nabla f(v_k) \end{aligned} \tag{5.1}$$

where $A = \begin{bmatrix} (1+\beta)I & -\beta I \\ I & 0 \end{bmatrix}$, $B = \begin{bmatrix} -\alpha I \\ 0 \end{bmatrix}$, and $C = \begin{bmatrix} (1+\beta)I & -\beta I \end{bmatrix}$. The convergence rate proof of Nesterov's method can be done by applying the dissipation inequality routine presented in past lectures.

1. Replace the nonlinear equation $u_k = \nabla f(v_k)$ in (5.1) by some quadratic inequality in the following form:

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

   The key issue is how to figure out X.

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

$$\begin{bmatrix} A^\mathsf{T}PA - \rho^2 P & A^\mathsf{T}PB \\ B^\mathsf{T}PA & B^\mathsf{T}PB \end{bmatrix} - X \leq 0. \tag{5.2}$$

If so, then the following inequality holds

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

which is exactly the so-called dissipation inequality $V_{k+1}-\rho^2V_k\le S(\xi_k,u_k)$ 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 $V_k\ge 0$ due to the fact $P\ge 0$. In Homework 1, you will figure out that (5.15) holds with $(\rho^2,\alpha,\beta)=(1-\sqrt{\frac{m}{L}},\frac{1}{L},\frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}+\sqrt{m}})$ if P is chosen properly.

3. Now directly apply the supply rate condition to conclude $V_{k+1}+f(x_{k+1})-f(x^*)\le\rho^2\left(V_k+f(x_k)-f(x^*)\right)$. This rate result can be converted into an ε-optimal iteration complexity result $O(\sqrt{\frac{L}{m}}\log\frac{1}{\varepsilon})$.

How to construct the supply rate condition for Nesterov's method? As mentioned previously, if we can construct a symmetric matrix X such that the following supply rate condition holds

$$\begin{aligned}S(\xi_k,w_k)&=\begin{bmatrix}\xi_k-\xi^*\\w_k\end{bmatrix}^\mathsf{T}X\begin{bmatrix}\xi_k-\xi^*\\w_k\end{bmatrix}\\&\le-(f(x_{k+1})-f(x^*))+\rho^2(f(x_k)-f(x^*))\\&=\rho^2(f(x_k)-f(x_{k+1}))+(1-\rho^2)(f(x^*)-f(x_{k+1})),\end{aligned}\tag{5.3}$$

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

$$\begin{bmatrix}\xi_k-\xi^*\\w_k\end{bmatrix}^\mathsf{T}X_1\begin{bmatrix}\xi_k-\xi^*\\w_k\end{bmatrix}\le f(x_k)-f(x_{k+1})\tag{5.4}$$

$$\begin{bmatrix}\xi_k-\xi^*\\w_k\end{bmatrix}^\mathsf{T}X_2\begin{bmatrix}\xi_k-\xi^*\\w_k\end{bmatrix}\le f(x^*)-f(x_{k+1}),\tag{5.5}$$

then we can set $X=\rho^2X_1+(1-\rho^2)X_2$ to obtain the condition (5.3). 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 (5.4) 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:

$$f(x)\le f(y)+\nabla f(y)^\mathsf{T}(x-y)+\frac{L}{2}\|x-y\|^2\tag{5.6}$$

$$f(x)\ge f(y)+\nabla f(y)^\mathsf{T}(x-y)+\frac{m}{2}\|x-y\|^2\tag{5.7}$$

How can we choose (x, y) in the above inequalities to obtain (5.4). One idea is to set $(x, y) \to (x_{k+1}, x_k)$ in (5.6). 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 (5.4) is w_k which is the gradient evaluated on v_k! Therefore, when applying (5.6) and (5.7) to construct (5.4), 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.

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

$$w_k = \begin{bmatrix} \nabla f_1(v_k) \\ \nabla f_2(v_k) \\ \vdots \\ \nabla f_n(v_k) \end{bmatrix} \tag{5.8}$$

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).

5.3 Sublinear Rate Analysis

So far we have talked about how to obtain a linear convergence rate using the dissipation inequality approach. What if the convergence rate is sublinear, e.g. $O(1/k)$? Recall a differentiable function f is convex if the following inequality holds for all x, y

$$f(x) \geq f(y) + \nabla f(y)^{\mathsf{T}}(x - y)$$

You can think convex functions as "0-strongly convex functions," although the m-strong convexity typically implicitly assume $m > 0$. What performance guarantees can we say about the optimization of such functions? For the gradient method, we can show $f(x_k) - f(x^*) = O(1/k)$. For Nesterov's accelerated method, we can show $f(x_k) - f(x^*) = O(1/k^2)$. We don't have linear convergence anymore. The convergence rate $O(1/k)$ and $O(1/k^2)$ are significantly slower than the linear convergence rate $O(\rho^{-k})$, and categorized as "sublinear convergence rates." Now we discuss how to modify the dissipation inequality approach to show such sublinear convergence rates.

5.3.1 The $O(1/k)$ rate of Gradient Descent Method

The dissipation inequality has the following form

$$V(\xi_{k+1}) - \rho^2 V(\xi_k) \leq S(\xi_k, u_k)$$

where V is non-negative. Depending on the properties of S, the dissipation inequality reaches different conclusions.

1. If $S(\xi_k, u_k) \leq 0$, then we have $V(\xi_{k+1}) - \rho^2 V(\xi_k) \leq 0$. This is a linear convergence in V. We used this argument to show the linear convergence of the gradient method.

2. If $S(\xi_k, u_k) \leq -(f(x_{k+1}) - f(x^*)) + \rho^2(f(x_k) - f(x^*))$, we have $V(\xi_{k+1}) + f(x_{k+1}) - f(x^*) \leq \rho^2(V(\xi_k) + f(x_k) - f(x^*))$. This is also linear convergence. We used this argument to show the linear convergence of Nesterov's method.

3. How about having the condition $S(\xi_k, u_k) \leq f(x^*) - f(x_k)$ and $\rho^2 = 1$? Then the dissipation inequality leads to the inequality $V(\xi_{k+1}) - V(\xi_k) + f(x_k) - f(x^*) \leq 0$. Summing this inequality leads to

$$\sum_{t=0}^{k}(f(x_t) - f(x^*)) \leq V(\xi_0) - V(\xi_{k+1}) \leq V(\xi_0)$$

   The last step relies on the fact $V \geq 0$. If we know $f(x_t) \leq f(x_{t-1})$, then the left side of the above inequality can be lower bounded by $(k+1)(f(x_k) - f(x^*))$. Therefore, we eventually have

$$f(x_k) - f(x^*) \leq \frac{V(\xi_0)}{k+1} \tag{5.9}$$

This is a sublinear rate result. We will use this argument to show that the gradient method is guaranteed to converge at the sublinear rate $O(1/k)$ when the objective function is smooth and convex.

Again, the dissipation inequality is constructed by solving the following condition

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

When f is smooth and convex, we want to show the gradient method converges at the rate $O(1/k)$. We have $A = I$, $B = -\alpha I$, and $\rho^2 = 1$. The key issue is how to choose X such that

$$\begin{bmatrix} x_k - x^* \\ \nabla f(x_k) \end{bmatrix}^\mathsf{T} X \begin{bmatrix} x_k - x^* \\ \nabla f(x_k) \end{bmatrix} \leq f(x^*) - f(x_k) \tag{5.10}$$

If f is L-smooth and convex, the following inequality holds for all x and y

$$f(x) \geq f(y) + \nabla f(y)^\mathsf{T}(x - y) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2$$

We set $x = x^*$ and $y = x_k$ in the above inequality. This leads to

$$f(x^*) \geq f(x_k) + \nabla f(x_k)^\mathsf{T}(x^* - x_k) + \frac{1}{2L}\|\nabla f(x^*) - \nabla f(x_k)\|^2$$

which can be rewritten in the form of (5.10) with $X = \begin{bmatrix} 0 & -\frac{1}{2}I \\ -\frac{1}{2}I & \frac{1}{2L}I \end{bmatrix}$. We can set $P = pI$ and the LMI condition just becomes

$$\begin{bmatrix} 0 & -\alpha p \\ -\alpha p & \alpha^2 p \end{bmatrix} - \begin{bmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2L} \end{bmatrix} \leq 0 \tag{5.11}$$

When $\alpha = \frac{1}{L}$, we can set $p = \frac{L}{2}$. The left side of the above inequality just becomes a zero matrix and the testing condition is met. Now the dissipation inequality holds. In addition, we have $f(x_{k+1}) - f(x_k) \leq -\left(\alpha - \frac{L\alpha^2}{2}\right)\|\nabla f(x_k)\|^2 \leq 0$. So we do have $f(x_{k+1}) \leq f(x_k)$, and (5.9) follows as a consequence. Finally, we have shown the following bound holds for the gradient method with a smooth and convex objective function

$$f(x_k) - f(x^*) \leq \frac{L\|x_0 - x^*\|^2}{2(k+1)}$$

5.3.2 Extension for Nesterov's Method

Nesterov's method can be modified for smooth and convex objective functions. In this case, we will use time-varying parameters, i.e. α_k and β_k.

Specifically, given L-smooth convex f, Nesterov's method iterates as follows:

$$x_{k+1} = x_k - \alpha \nabla f((1+\beta_k)x_k - \beta_k x_{k-1}) + \beta_k(x_k - x_{k-1}) \tag{5.12}$$

where $\alpha = \frac{1}{L}$ and β_k is a prescribed sequence. One typical choice is setting $\beta_k = \frac{k-1}{k+2}$. Another popular choice is defining β_k recursively as follows.

$$\zeta_{-1} = 0, \quad \zeta_{k+1} = \frac{1+\sqrt{1+4\zeta_k^2}}{2}, \quad \beta_k = \frac{\zeta_{k-1}-1}{\zeta_k}.$$

The sequence $\{\zeta_k\}$ satisfies $\zeta_k^2 - \zeta_k = \zeta_{k-1}^2$. Due to the time-varying nature of β_k, we need to use the following time-varying model for (5.12):

$$\begin{aligned} \xi_{k+1} &= A_k \xi_k + B_k u_k \\ v_k &= C_k \xi_k \\ u_k &= \nabla f(v_k) \end{aligned} \tag{5.13}$$

We choose $A_k = \begin{bmatrix} (1+\beta_k)I & -\beta_k I \\ I & 0 \end{bmatrix}$, $B_k = \begin{bmatrix} -\alpha I \\ 0 \end{bmatrix}$, $C_k = \begin{bmatrix} (1+\beta_k)I & -\beta_k I \end{bmatrix}$, and $\xi_k = \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix}$. Then $v_k = C_k\xi_k = \begin{bmatrix} (1+\beta_k)I & -\beta_k I \end{bmatrix} \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix} = (1+\beta_k)x_k - \beta_k x_{k-1}$, and $u_k = \nabla f(v_k) = \nabla f((1+\beta_k)x_k - \beta_k x_{k-1})$. We can see B_k actually does not depend on k. But if we let α depend on k, then we need B to depend on k. So (5.13) is general. We will modify the dissipation inequality to provide a sublinear rate analysis for (5.13).

Now for the general model (5.13), (A, B, C) depend on k. Therefore, we need to modify the above condition as

$$\begin{bmatrix} A_k^\mathsf{T} P_{k+1} A_k - P_k & A_k^\mathsf{T} P_{k+1} B_k \\ B_k^\mathsf{T} P_{k+1} A_k & B_k^\mathsf{T} P_{k+1} B_k \end{bmatrix} - X_k \le 0.$$

The key question is what X_k we should use. We can choose (M_k, N_k) to show

$$\begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \\ \nabla f(y_k) \end{bmatrix}^\mathsf{T} M_k \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \\ \nabla f(y_k) \end{bmatrix} \le f(x_k) - f(x_{k+1})$$

$$\begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \\ \nabla f(y_k) \end{bmatrix}^\mathsf{T} N_k \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \\ \nabla f(y_k) \end{bmatrix} \le f(x^*) - f(x_{k+1})$$

where N_k and M_k directly depend on β_k. Now we define $f^* = f(x^*)$. If we choose $X_k := \mu_k M_k + (\mu_{k+1} - \mu_k) N_k$ for all k. then the supply rate S satisfies the condition

$$S(\xi_k, w_k) = \begin{bmatrix} \xi_k \\ w_k \end{bmatrix}^\mathsf{T} X_k \begin{bmatrix} \xi_k \\ w_k \end{bmatrix} \le \mu_k(f(x_k) - f^*) - \mu_{k+1}(f(x_{k+1}) - f^*).$$

which can be used to show the rate $O(1/k^2)$ if μ_k is chosen properly. Specifically, if we can find positive semidefinite P_k and non-negative increasing sequence $\{\mu_k\}$ such that

$$\begin{bmatrix} A_k^\mathsf{T} P_{k+1} A_k - P_k & A_k^\mathsf{T} P_{k+1} B_k \\ B_k^\mathsf{T} P_{k+1} A_k & B_k^\mathsf{T} P_{k+1} B_k \end{bmatrix} - \mu_k M_k - (\mu_{k+1} - \mu_k) N_k \leq 0.$$

then we will be able to use our standard dissipation inequality arguments to show

$$\begin{aligned}(\xi_{k+1} - \xi^*)^\mathsf{T} P_{k+1} (\xi_{k+1} - \xi^*) - (\xi_k - \xi^*)^\mathsf{T} P_k (\xi_k - \xi^*) &\leq \begin{bmatrix} \xi_k \\ w_k \end{bmatrix}^\mathsf{T} X_k \begin{bmatrix} \xi_k \\ w_k \end{bmatrix} \\ &\leq \mu_k (f(x_k) - f^*) - \mu_{k+1} (f(x_{k+1}) - f^*).\end{aligned}$$

This is equivalent to

$$(\xi_{k+1} - \xi^*)^\mathsf{T} P_{k+1} (\xi_{k+1} - \xi^*) + \mu_{k+1} (f(x_{k+1}) - f^*) \leq (\xi_k - \xi^*)^\mathsf{T} P_k (\xi_k - \xi^*) + \mu_k (f(x_k) - f^*).$$

We can iterate the above inequality to show

$$\mu_k (f(x_k) - f^*). \leq (\xi_k - \xi^*)^\mathsf{T} P_k (\xi_k - \xi^*) + \mu_k (f(x_k) - f^*) \leq (\xi_0 - \xi^*)^\mathsf{T} P_0 (\xi_0 - \xi^*) + \mu_0 (f(x_0) - f^*).$$

If $\frac{1}{\mu_k} = O(1/k^2)$, then we immediately obtain the rate $O(1/k^2)$ for Nesterov's method.

Clearly, obtaining a rate $O(1/k^2)$ is more subtle than obtaining other rates due to the fact that now we need to find a sequence of P_k and μ_k. A specific choice for Neseterov's method is setting $\mu_k := (\zeta_{k-1})^2$ and $P_k := \frac{L}{2} \begin{bmatrix} \zeta_{k-1} \\ 1 - \zeta_{k-1} \end{bmatrix} \begin{bmatrix} \zeta_{k-1} & 1 - \zeta_{k-1} \end{bmatrix}$. We will not talk about $O(1/k^2)$ rate in future lectures.

5.4 Summary

The dissipation inequality appeared in the previous lectures has the following form:

$$V(\xi_{k+1}) - \rho^2 V(\xi_k) \leq S(\xi_k, u_k)$$

We can use the dissipation inequality to prove various results:

1. If $S(\xi_k, u_k) \leq 0$, then the dissipation inequality becomes $V(\xi_{k+1}) - \rho^2 V(\xi_k) \leq 0$. This is a linear convergence in V. We have used this type of arguments to show the linear convergence of the gradient method.

2. If $S(\xi_k, u_k) \leq -(f(x_{k+1}) - f(x^*)) + \rho^2 (f(x_k) - f(x^*))$, we have $V(\xi_{k+1}) + f(x_{k+1}) - f(x^*) \leq \rho^2 (V(\xi_k) + f(x_k) - f(x^*))$. This is a linear convergence in $V(\xi_k) + f(x_k) - f(x^*)$. We have shown the linear convergence of Nesterov's method via this type of arguments.

3. Sometimes the algorithms are stochastic. Then the supply rate condition also has to take the randomness into accounts. If $\mathbb{E}S(\xi_k, u_k) \leq M$, then the dissipation inequality leads to a bound in the form of $\mathbb{E}V(\xi_k) \leq \rho^{2k}\mathbb{E}V(\xi_0) + \frac{M}{1-\rho^2}$. This means the algorithm goes to a small ball around the optimal solution at a linear rate. We have used this argument to show the behaviors of the stochastic gradient method with a constant stepsize.

4. If $S(\xi_k, u_k) \leq f(x^*) - f(x_k)$ and $\rho^2 = 1$, then the dissipation inequality leads to the inequality $V(\xi_{k+1}) - V(\xi_k) + f(x_k) - f(x^*) \leq 0$. Summing this inequality leads to $\sum_{t=0}^{k}(f(x_t) - f(x^*)) \leq V(\xi_0) - V(\xi_{k+1})$. We have used this argument to show that the gradient method is guaranteed to converge at the sublinear rate $O(1/k)$ when the objective function is smooth and convex.

5. For Nesterov's method under the convex assumption, the supply rate condition changes with k. We can obtain an $O(1/k^2)$ rate in this case.

Bibliography

[1] L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.