ECE 586 BH: Interplay between Control and Machine Learning **Fall 2023**

Lecture 12

Control Tools for Stochastic Optimization and Supervised Learning, Part II

Lecturer: Bin Hu, Date:09/28/2023

In this lecture, we will discuss how to tailor the quadratic constraint (QC) approach as a unified analysis tool for stochastic optimization methods.

12.1 Unified Analysis via QCs and Dissipativity

In Lecture 3, we have presented the dissipation inequality approach as a general analysis tool for feedback systems. In today's lecture, we will discuss how to tailor the dissipation inequality approach for stochastic finite-sum methods.

Suppose G is an LTI system satisfying $\xi_{k+1}-\xi^*=A(\xi_k-\xi^*)+B(w_k-w^*)$. Suppose we know $S=\begin{bmatrix}\xi_k-\xi^*\\ w_k-w^*\end{bmatrix}^\mathsf{T} X\begin{bmatrix}\xi_k-\xi^*\\ w_k-w^*\end{bmatrix}\le 0$ for any $w=\Delta(C\xi)$.[1] If there exists a positive definite matrix P s.t.

$$\begin{bmatrix}A^\mathsf{T}PA-\rho^2P & A^\mathsf{T}PB\\ B^\mathsf{T}PA & B^\mathsf{T}PB\end{bmatrix}-X\le 0, \tag{12.1}$$

then we have $V(\xi_{k+1})\le\rho^2V(\xi_k)+S\le\rho^2V(\xi_k)$ where $V(\xi_k):=(\xi_k-\xi^*)^\mathsf{T}P(\xi_k-\xi^*)$. This establishes the linear convergence rate bound $\|\xi_k-\xi^*\|\le\sqrt{\operatorname{cond}(P)}\rho^k\|\xi_0-\xi^*\|$. We have already discussed how to perform such an analysis for the gradient method. To handle stochastic finite-sum methods, we only need to make some minor modification to the dissipation inequality approach. We first present the high-level ideas.

- **Interconnection of an LTI system G and stochastic Δ:** In this case, one typically will be able to construct some expected supply rate condition $\mathbb{E}S\le M$. Then the LMI condition (12.1) can still be used to construct a (almost sure) dissipation inequality $V(\xi_{k+1})\le\rho^2V(\xi_k)+S$. How to obtain a convergence bound from such a dissipation inequality? Since the supply rate condition holds in the average sense, we have to take expectation of the dissipation inequality and obtain $\mathbb{E}V(\xi_{k+1})\le\rho^2\mathbb{E}V(\xi_k)+\mathbb{E}S$. Depending on what M is , this expected dissipation inequality can be used to prove various things. For example, when analyzing the stochastic gradient method for smooth strongly-convex f_i, we will figure out that M is just a constant, and the dissipation inequality can be iterated to show $\mathbb{E}V(\xi_k)\le\rho^2V(\xi_0)+\frac{M}{1-\rho^2}$. This just states that the stochastic gradient method converges linearly to a small ball whose size is controlled by $\frac{M}{1-\rho^2}$. Notice in this case, the supply rate is not decreasing to 0 and the total internal energy is not going to converge to 0.

[1] Here we assume $v_k=C\xi_k$ and hence $w=\Delta(v)=\Delta(C\xi)$.

- **Interconnection of a jump system G and a deterministic nonlinearity Δ:** As discussed in Lecture 3, we can use the property of Δ to construct some quadratic supply rate conditions. Then we can analyze the feedback interconnection using the following LMI (can you figure out why?)

$$\sum_{i=1}^{n}\left(p_i\begin{bmatrix}A_i^\mathsf{T}PA_i-\rho^2P & A_i^\mathsf{T}PB_i\\ B_i^\mathsf{T}PA_i & B_i^\mathsf{T}PB_i\end{bmatrix}-X\right)\leq 0.$$

  Here we assume $P(i_k=i)=p_i$, and X is independent of i_k. This type of supply rate conditions arise naturally when the matrix C in G does not depend on i_k and Δ is deterministic. The above LMI can be directly applied to SAGA-like methods.

12.1.1 Dissipation Inequality for Stochastic Gradient

Now we present a detailed analysis for the SG method under the following two assumptions:

1. f is m-strongly convex.
2. f_i is L-smooth and convex for all i.

Under these two assumptions, we can show that SGD satisfies a bound in the following form:

$$\mathbb{E}\|x_k-x^*\|^2\leq\rho^{2k}\|x_0-x^*\|^2+H \tag{12.2}$$

where $\rho^2=1-2m\alpha+O(\alpha^2)$ and $H=O(\alpha)$. Here ρ^2 quantifies the convergence speed and H quantifies the accuracy. Therefore, for SGD, there is a fundamental trade-off between the convergence speed and the accuracy. If one wants a very accurate solution, one has to decrease α so that H is decreased. However, ρ^2 increases as α decreases and the convergence speed becomes slower.

As mentioned before, the supply rate condition used to prove (12.2) has a form $\mathbb{E}S\leq M$. Recall that the SG method is equivalent to a feedback system $F_u(G,\Delta)$. Here Δ is a stochastic operator mapping v to w as $w_k=\nabla f_{i_k}(v_k)$. In addition, G is governed by an LTI model with $A=I$, $B=-\alpha I$, and $C=I$. We emphasize that for the SG method we have $\xi_{k+1}-\xi^*=A(\xi_k-\xi^*)+Bw_k$ and we do not shift w_k to (w_k-w^*). Again, we perform our analysis in two steps. In Step 1, we construct the supply rates. In Step 2, we solve an LMI to construct the dissipation inequality.

1. Based on $w_k=\nabla f_{i_k}(v_k)$, we can show the following inequalities:

$$\mathbb{E}\begin{bmatrix}v_k-x^*\\ w_k\end{bmatrix}^\mathsf{T}\begin{bmatrix}0 & -LI\\ -LI & I\end{bmatrix}\begin{bmatrix}v_k-x^*\\ w_k\end{bmatrix}\leq\frac{2}{n}\sum_{i=1}^{n}\|\nabla f_i(x^*)\|^2=M \tag{12.3}$$

$$\mathbb{E}\begin{bmatrix}v_k-x^*\\ w_k\end{bmatrix}^\mathsf{T}\begin{bmatrix}2mI & -I\\ -I & 0\end{bmatrix}\begin{bmatrix}v_k-x^*\\ w_k\end{bmatrix}\leq 0 \tag{12.4}$$

We skip the proofs here. Now we just set $X_1 = \begin{bmatrix} 0 & -LI \\ -LI & I \end{bmatrix}$ and $X_2 = \begin{bmatrix} 2mI & -I \\ -I & 0 \end{bmatrix}$. Notice that it is the first supply rate that causes the convergence issue for the SG method. Since this supply rate keeps on delivering energy to the system, the internal energy does not decrease to 0.

2. Now we test if there exists $P > 0$ and non-negative scalers (λ_1, λ_2) such that

$$\begin{bmatrix} A^\mathsf{T} PA - \rho^2 P & A^\mathsf{T} PB \\ B^\mathsf{T} PA & B^\mathsf{T} PB \end{bmatrix} - \lambda_1 X_1 - \lambda_2 X_2 \leq 0. \tag{12.5}$$

If so, we have

$$\begin{aligned} &\mathbb{E}(\xi_{k+1} - \xi^*)^\mathsf{T} P(\xi_{k+1} - \xi^*) - \rho^2 \mathbb{E}(\xi_k - \xi^*)^\mathsf{T} P(\xi_k - \xi^*) \\ \leq &\lambda_1 \mathbb{E} \begin{bmatrix} \xi_k - \xi^* \\ w_k \end{bmatrix}^\mathsf{T} X_1 \begin{bmatrix} \xi_k - \xi^* \\ w_k \end{bmatrix} + \lambda_2 \mathbb{E} \begin{bmatrix} \xi_k - \xi^* \\ w_k \end{bmatrix}^\mathsf{T} X_2 \begin{bmatrix} \xi_k - \xi^* \\ w_k \end{bmatrix} \\ \leq &\lambda_1 M \end{aligned}$$

For simplicity, we can choose $P = I$. Recall for SGD we have $A = I$ and $B = -\alpha I$. Hence (12.5) is equivalent to

$$\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} - \lambda_1 \begin{bmatrix} 0 & -L \\ -L & 1 \end{bmatrix} - \lambda_2 \begin{bmatrix} 2m & -1 \\ -1 & 0 \end{bmatrix} \leq 0 \tag{12.6}$$

Now we set the left side to be a zero matrix. We have $\lambda_1 = \alpha^2$, $\lambda_2 = \alpha - \lambda_1 L$, and $\rho^2 = 1 - 2m\lambda_2 = 1 - 2m\alpha + 2mL\alpha^2$. Now the dissipation inequality leads to

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

Iterating the above bound leads to

$$\begin{aligned} \mathbb{E}\|x_k - x^*\|^2 \leq &\rho^2 \mathbb{E}\|x_{k-1} - x^*\|^2 + \lambda_1 M \\ \leq &\rho^4 \mathbb{E}\|x_{k-1} - x^*\|^2 + (\rho^2 + 1)\lambda_1 M \\ \leq &\rho^{2k} \mathbb{E}\|x_0 - x^*\|^2 + \left(\sum_{t=0}^{\infty} \rho^{2t} \right) \lambda_1 M \\ = &\rho^{2k} \mathbb{E}\|x_0 - x^*\|^2 + \frac{\lambda_1 M}{1 - \rho^2} \end{aligned}$$

From Step 2, we have $\rho^2 = 1 - 2m\alpha + 2mL\alpha^2 = 1 - 2m\alpha + O(\alpha^2)$, and $H = \frac{\lambda_1 M}{1-\rho^2} = O(\alpha)$. This leads to the desired conclusion (12.2).

12.1.2 Dissipation Inequality for SAGA-like Methods

To show convergence behaviors, we are actually looking at the following iteration:

$$\begin{aligned}
\xi_{k+1} - \xi^* &= A_{i_k}(\xi_k - \xi^*) + B_{i_k}(w_k - w^*)\\
v_k - v^* &= C(\xi_k - \xi^*)\\
w_k - w^* &= \begin{bmatrix} \nabla f_1(v_k) - \nabla f_1(v^*)\\ \nabla f_2(v_k) - \nabla f_2(v^*)\\ \vdots \\ \nabla f_n(v_k) - \nabla f_n(v^*) \end{bmatrix}
\end{aligned} \tag{12.7}$$

Again, we can follow the two steps in the dissipation inequality framework.

1. First, we try to construct the following supply rate conditions for $j = 1, \ldots, J$.

$$\begin{bmatrix} \xi_k - \xi^* \\ w_k - w^* \end{bmatrix}^\mathsf{T} X_j \begin{bmatrix} \xi_k - \xi^* \\ w_k - w^* \end{bmatrix} \leq 0. \tag{12.8}$$

The supply rate constructions typically require using the matrix C and some properties of Δ. We will cover this in more details in next section. For now, let's look at one example. Suppose we know f_1 is L-smooth and m-strongly convex. Hence we know

$$\begin{bmatrix} v_k - v^* \\ \nabla f_1(v_k) - \nabla f_1(v^*) \end{bmatrix}^\mathsf{T} \begin{bmatrix} 2mLI & -(L+m)I \\ -(L+m)I & 2I \end{bmatrix} \begin{bmatrix} v_k - v^* \\ \nabla f_1(v_k) - \nabla f_1(v^*) \end{bmatrix} \leq 0. \tag{12.9}$$

Now notice we have $v_k - v^* = C(\xi_k - \xi^*)$ and $\nabla f_1(v_k) - \nabla f_1(w^*) = (e_1^\mathsf{T} \otimes I)(w_k - w^*)$ where e_1 is a vector whose first entry is 1 and all other entries are 0. Therefore, we have

$$\begin{bmatrix} v_k - v^* \\ \nabla f_1(v_k) - \nabla f_1(v^*) \end{bmatrix} = \begin{bmatrix} C & 0_{p\times(np)} \\ 0 & e_1^\mathsf{T} \otimes I \end{bmatrix} \begin{bmatrix} \xi_k - \xi^* \\ w_k - w^* \end{bmatrix}$$

Substituting the above equation into (12.9) leads to

$$\begin{bmatrix} \xi_k - \xi^* \\ w_k - w^* \end{bmatrix}^\mathsf{T} \begin{bmatrix} C & 0_{p\times(np)} \\ 0 & e_1^\mathsf{T} \otimes I \end{bmatrix}^\mathsf{T} \begin{bmatrix} 2mLI & -(L+m)I \\ -(L+m)I & 2I \end{bmatrix} \begin{bmatrix} C & 0_{p\times(np)} \\ 0 & e_1^\mathsf{T} \otimes I \end{bmatrix} \begin{bmatrix} \xi_k - \xi^* \\ w_k - w^* \end{bmatrix} \leq 0$$

Therefore, we can just choose $X_1 = \begin{bmatrix} C & 0_{p\times(np)} \\ 0 & e_1^\mathsf{T} \otimes I \end{bmatrix}^\mathsf{T} \begin{bmatrix} 2mLI & -(L+m)I \\ -(L+m)I & 2I \end{bmatrix} \begin{bmatrix} C & 0_{p\times(np)} \\ 0 & e_1^\mathsf{T} \otimes I \end{bmatrix}$. Clearly X_1 depends on m, L, and C. You can imagine that properties of f_i and f can all be transformed into quadratic inequalities in the form of (12.8) via similar algebraic manipulations.

2. Now we can perform our LMI-based analysis. If there exists a positive definite matrix P and non-negative scalers λ_j s.t.

$$\sum_{i=1}^{n}\left(p_i\begin{bmatrix} A_i^\mathsf{T} P A_i - \rho^2 P & A_i^\mathsf{T} P B_i \\ B_i^\mathsf{T} P A_i & B_i^\mathsf{T} P B_i\end{bmatrix}\right) \le \sum_{j=1}^{J}\lambda_j X_j,$$

then we have the expected dissipation inequality $\mathbb{E}V(\xi_{k+1}) \le \rho^2\mathbb{E}V(\xi_k) + \mathbb{E}S(\xi_k, w_k)$ where the storage function is defined as $V(\xi_k) = (\xi_k - \xi^*)^\mathsf{T} P(\xi_k - \xi^*)$ and the supply rate S is defined as

$$S(\xi_k, w_k) = \sum_{j=1}^{J}\lambda_j\begin{bmatrix}\xi_k - \xi^* \\ w_k - w^*\end{bmatrix}^\mathsf{T} X_j\begin{bmatrix}\xi_k - \xi^* \\ w_k - w^*\end{bmatrix}.$$

The proof for this part is based on standard Lyapunov arguments you have seen many times. We just left and right multiply both sides of the LMI condition with $\begin{bmatrix}(\xi_k - \xi^*)^\mathsf{T} & (w_k - w^*)^\mathsf{T}\end{bmatrix}$ and $\begin{bmatrix}\xi_k - \xi^* \\ w_k - w^*\end{bmatrix}$. This directly leads to the desired dissipation inequality. The non-negativity of λ_j guarantees $S \le 0$ and hence we have the linear convergence bound $\mathbb{E}V(\xi_k) \le \rho^{2k}\mathbb{E}V(\xi_0)$. For given (A_i, B_i, ρ) and X_j, out testing condition is linear in the decision variables P and λ_j, and can be solved as LMIs.

Numerical solutions of the above LMIs can be obtained via using existing SDP solvers. However, solving the LMIs analytically may require case-by-case constructions of P. The good news is that we can use the numerical solutions of LMIs to guide our constructions of analytical proofs.

Once we have the supply rate conditions, the constructions of dissipation inequality can be somehow routinized by solving LMIs. Now we are ready to construct supply rates for stochastic finite-sum methods. In many situations, the analysis of stochastic finite-sum methods only require simple supply rates that can be obtained by manipulating the quadratic constraints covered in the last lecture. First we will focus on SAGA-like methods. Then we will briefly discuss SVRG which is another important finite-sum method.

12.2 Supply Rates for SAGA-Like Methods

Recall that SAGA-like methods can be represented as $F_u(G, \Delta)$ where G is a jump system and the operator Δ maps v to w as

$$w_k = \begin{bmatrix}\nabla f_1(v_k) \\ \nabla f_2(v_k) \\ \vdots \\ \nabla f_n(v_k)\end{bmatrix} \tag{12.10}$$

For this operator Δ, we want to construct pointwise quadratic constraints on the input/output pair (v, w):

$$\begin{bmatrix} v_k - v^* \\ w_k - w^* \end{bmatrix}^\mathsf{T} M \begin{bmatrix} v_k - v^* \\ w_k - w^* \end{bmatrix} \leq 0, \tag{12.11}$$

where M is a symmetric matrix, and (w^*, v^*) are determined by the fixed points of the feedback interconnection $F_u(G, \Delta)$. For SAGA, we know $v^* = x^*$ and $w^* = \begin{bmatrix} \nabla f_1(x^*)^T \cdots \nabla f_n(x^*)^T \end{bmatrix}$ where $\nabla f(x^*) = \frac{1}{n} \sum_{k=1}^n \nabla f_i(x^*) = 0$.

Again, if we know $v_k - v^* = C(\xi_k - \xi^*)$, the above quadratic constraint (12.11) just gives the following supply rate condition

$$\begin{bmatrix} \xi_k - \xi^* \\ w_k - w^* \end{bmatrix}^\mathsf{T} \left(\begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix}^\mathsf{T} M \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} \right) \begin{bmatrix} \xi_k - \xi^* \\ w_k - w^* \end{bmatrix} \leq 0.$$

Hence we just focus on how to obtain the quadratic constraint (12.11). Various assumptions on f_i and f can be converted into inequalities in the form of (12.11). Now let's look at a few concrete examples.

- Assumption 1: f_i is L-smooth and m-strongly convex. In this case, we know

$$\begin{bmatrix} v_k - v^* \\ \nabla f_i(v_k) - \nabla f_i(v^*) \end{bmatrix}^\mathsf{T} \begin{bmatrix} 2mLI & -(L+m)I \\ -(L+m)I & 2I \end{bmatrix} \begin{bmatrix} v_k - v^* \\ \nabla f_i(v_k) - \nabla f_i(v^*) \end{bmatrix} \leq 0. \tag{12.12}$$

We need to make use of the following key relation:

$$w_k - w^* = \begin{bmatrix} \nabla f_1(v_k) - \nabla f_1(v^*) \\ \nabla f_2(v_k) - \nabla f_2(v^*) \\ \vdots \\ \nabla f_n(v_k) - \nabla f_n(v^*) \end{bmatrix} \tag{12.13}$$

which leads to $\nabla f_i(v_k) - \nabla f_i(v^*) = (e_i^\mathsf{T} \otimes I)(w_k - w^*)$ where e_i is a vector whose i-th entry is 1 and all other entries are 0. Therefore, we have

$$\begin{bmatrix} v_k - v^* \\ \nabla f_i(v_k) - \nabla f_i(v^*) \end{bmatrix} = \begin{bmatrix} I & 0_{p \times (np)} \\ 0 & e_i^\mathsf{T} \otimes I \end{bmatrix} \begin{bmatrix} v_k - v^* \\ w_k - w^* \end{bmatrix} \tag{12.14}$$

Substituting the above equation into (12.12) leads to

$$\begin{bmatrix} v_k - v^* \\ w_k - w^* \end{bmatrix}^\mathsf{T} \begin{bmatrix} I & 0_{p \times (np)} \\ 0 & e_i^\mathsf{T} \otimes I \end{bmatrix}^\mathsf{T} \begin{bmatrix} 2mLI & -(L+m)I \\ -(L+m)I & 2I \end{bmatrix} \begin{bmatrix} I & 0_{p \times (np)} \\ 0 & e_i^\mathsf{T} \otimes I \end{bmatrix} \begin{bmatrix} v_k - v^* \\ w_k - w^* \end{bmatrix} \leq 0$$

Therefore, we can just choose $M = \begin{bmatrix} I & 0_{p \times (np)} \\ 0 & e_i^\mathsf{T} \otimes I \end{bmatrix}^\mathsf{T} \begin{bmatrix} 2mLI & -(L+m)I \\ -(L+m)I & 2I \end{bmatrix} \begin{bmatrix} I & 0_{p \times (np)} \\ 0 & e_i^\mathsf{T} \otimes I \end{bmatrix}$.

- Assumption 2: f is L-smooth and m-strongly convex. In this case, we know

$$\begin{bmatrix} v_k - v^* \\ \nabla f(v_k) - \nabla f(v^*) \end{bmatrix}^\mathsf{T} \begin{bmatrix} 2mLI & -(L+m)I \\ -(L+m)I & 2I \end{bmatrix} \begin{bmatrix} v_k - v^* \\ \nabla f(v_k) - \nabla f(v^*) \end{bmatrix} \le 0. \tag{12.15}$$

Based on (12.13), we have $\nabla f(v_k) - \nabla f(v^*) = \frac{1}{n}(e^\mathsf{T} \otimes I)(w_k - w^*)$ where $e := \sum_{i=1}^n e_i$ is a vector whose entries are all 1. Therefore, we have

$$\begin{bmatrix} v_k - v^* \\ \nabla f(v_k) - \nabla f(v^*) \end{bmatrix} = \begin{bmatrix} I & 0_{p\times(np)} \\ 0 & \frac{1}{n}e^\mathsf{T} \otimes I \end{bmatrix} \begin{bmatrix} v_k - v^* \\ w_k - w^* \end{bmatrix}$$

Substituting the above equation into (12.15) leads to

$$\begin{bmatrix} v_k - v^* \\ w_k - w^* \end{bmatrix}^\mathsf{T} \begin{bmatrix} I & 0_{p\times(np)} \\ 0 & \frac{1}{n}e^\mathsf{T} \otimes I \end{bmatrix}^\mathsf{T} \begin{bmatrix} 2mLI & -(L+m)I \\ -(L+m)I & 2I \end{bmatrix} \begin{bmatrix} I & 0_{p\times(np)} \\ 0 & \frac{1}{n}e^\mathsf{T} \otimes I \end{bmatrix} \begin{bmatrix} v_k - v^* \\ w_k - w^* \end{bmatrix} \le 0.$$

Therefore, we can just choose $M = \begin{bmatrix} I & 0_{p\times(np)} \\ 0 & \frac{1}{n}e^\mathsf{T} \otimes I \end{bmatrix}^\mathsf{T} \begin{bmatrix} 2mLI & -(L+m)I \\ -(L+m)I & 2I \end{bmatrix} \begin{bmatrix} I & 0_{p\times(np)} \\ 0 & \frac{1}{n}e^\mathsf{T} \otimes I \end{bmatrix}$.

- Assumption 3: f_i is L-smooth but may not be convex. In this case, we know

$$\begin{bmatrix} v_k - v^* \\ \nabla f_i(v_k) - \nabla f_i(v^*) \end{bmatrix}^\mathsf{T} \begin{bmatrix} -L^2 I & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} v_k - v^* \\ \nabla f_i(v_k) - \nabla f_i(v^*) \end{bmatrix} \le 0. \tag{12.16}$$

Similarly, we can substitute (12.14) into (12.16) and get

$$\begin{bmatrix} v_k - v^* \\ w_k - w^* \end{bmatrix}^\mathsf{T} \begin{bmatrix} I & 0_{p\times(np)} \\ 0 & e_i^\mathsf{T} \otimes I \end{bmatrix}^\mathsf{T} \begin{bmatrix} -L^2 I & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0_{p\times(np)} \\ 0 & e_i^\mathsf{T} \otimes I \end{bmatrix} \begin{bmatrix} v_k - v^* \\ w_k - w^* \end{bmatrix} \le 0$$

Therefore, we can just choose $M = \begin{bmatrix} I & 0_{p\times(np)} \\ 0 & e_i^\mathsf{T} \otimes I \end{bmatrix}^\mathsf{T} \begin{bmatrix} -L^2 I & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0_{p\times(np)} \\ 0 & e_i^\mathsf{T} \otimes I \end{bmatrix}$.

- Assumption 4: f satisfies the "one-point convexity" condition:

$$\begin{bmatrix} v_k - x^* \\ \nabla f(v_k) \end{bmatrix}^\mathsf{T} \begin{bmatrix} 2mLI & -(L+m)I \\ -(L+m)I & 2I \end{bmatrix} \begin{bmatrix} v_k - x^* \\ \nabla f(v_k) \end{bmatrix} \le 0. \tag{12.17}$$

Notice the difference between (12.17) and (12.15) is that v^* is allowed to be any point in (12.15). Due to the facts $v^* = x^*$ and $\frac{1}{n}(e^\mathsf{T} \otimes I)w^* = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x^*) = 0$, we still have $\nabla f(v_k) - \nabla f(v^*) = \frac{1}{n}(e^\mathsf{T} \otimes I)(w_k - w^*)$. Similar to before, we can just choose $M = \begin{bmatrix} I & 0_{p\times(np)} \\ 0 & \frac{1}{n}e^\mathsf{T} \otimes I \end{bmatrix}^\mathsf{T} \begin{bmatrix} 2mLI & -(L+m)I \\ -(L+m)I & 2I \end{bmatrix} \begin{bmatrix} I & 0_{p\times(np)} \\ 0 & \frac{1}{n}e^\mathsf{T} \otimes I \end{bmatrix}$.

How to use the above quadratic constraints? Depending on the assumptions on f_i and f, we can choose multiple M_j ($j = 1, \ldots, J$) accordingly and formulate the following LMI

$$\sum_{i=1}^{n} \left(p_i \begin{bmatrix} A_i^\mathsf{T} P A_i - \rho^2 P & A_i^\mathsf{T} P B_i \\ B_i^\mathsf{T} P A_i & B_i^\mathsf{T} P B_i \end{bmatrix} \right) \leq \sum_{j=1}^{J} \lambda_j \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix}^\mathsf{T} M_j \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix},$$

where the positive definite matrix P and non-negative scalers λ_j are decision variables. When the assumptions on f_i and f change, typically one only needs to modify M_j accordingly. The convergence rates of SAGA and several standard finite-sum methods (SDCA, Finito, etc) can be obtained using the above quadratic constraints and LMI formulations.

Comments on SAG. The convergence rate proof of SAG is more subtle. Quadratic Lyapunov functions and the pointwise quadratic constraints mentioned above are not enough for proving the convergence rate of SAG. The analysis of SAG requires the use of the Lure-Postnikov Lyapunov function (this is similar to the proof of Nesterov's method). For the 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).

12.3 Supply Rates for SVRG

Finally we briefly discuss SVRG that is built upon the idea of variance reduction. Originally we model the SG method as $F_u(G, \Delta)$ where Δ maps v to w as $w_k = \nabla f_{i_k}(v_k)$. We directly developed the supply rate condition for Δ and obtain some condition in the form of $\mathbb{E}S \leq C$ where C is a positive constant. A physical interpretation is that the stochastic gradient $\nabla f_{i_k}(v_k)$ keeps on supplying energy into the system and hence the system is not going to converge to its fixed point. Now we take a closer look. We can actually rewrite the SG method as

$$x_{k+1} = x_k - \alpha(\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x^*)) - \alpha \nabla f_{i_k}(x^*)$$

If we choose $\xi_k = x_k$, $v_k = \xi_k$, $w_k = \begin{bmatrix} \nabla f_{i_k}(v_k) - \nabla f_{i_k}(x^*) \\ \nabla f_{i_k}(x^*) \end{bmatrix}$, $A = I$, $B = \begin{bmatrix} -\alpha I & -\alpha I \end{bmatrix}$, and $C = I$, we obtain a new feedback representation for the SG method. Now the input w_k has two entries. Actually it is trivial to construct a supply rate condition to couple the first entry of w_k with $x_k - x^*$. For example, if f_i is L-smooth and m-strongly convex, the following inequality holds in an almost sure sense

$$\begin{bmatrix} v_k - x^* \\ \nabla f_{i_k}(v_k) - \nabla f_{i_k}(x^*) \end{bmatrix}^\mathsf{T} \begin{bmatrix} 2mLI & -(m+L)I \\ -(m+L)I & 2I \end{bmatrix} \begin{bmatrix} v_k - x^* \\ \nabla f_{i_k}(v_k) - \nabla f_{i_k}(x^*) \end{bmatrix} \leq 0. \qquad (12.18)$$

Hence the first entry of w_k is not delivering energy into the system. The troublesome term is the second entry of w_k. The term $\nabla f_{i_k}(x^*)$ keeps on delivering energy into the system.

SVRG modifies the second entry of w_k as $\nabla f_{i_k}(x^*) - \nabla f_{i_k}(x_0) + \nabla f(x_0)$. Now this input depends on the initial state x_0. One will be able to obtain a supply rate condition in the form of $\mathbb{E}S \leq L\|x_0 - x^*\|^2$. SVRG is an epoch-based algorithm and at the beginning of each epoch it will update x_0 as the last (or average) iterate of the last epoch. Notice for each epoch, one needs to evaluate one full gradient $\nabla f(x_0)$. Hence the selection of the epoch length is going to affect the performance of SVRG. Within one epoch, x_0 is a fixed vector. As more epochs are run, x_0 gets closer to x^*. The supplied energy eventually decreases to 0 as x_0 converges to x^*. This is a rough physical explanation for the convergence mechanism of SVRG. The dissipation inequality approach can be applied to analyze SVRG and its accelerated variant Katyusha. We omit the details here.