DSCC/CSC 435 & ECE 412 Optimization for Machine Learning — Lecture 7

Accelerated Gradient Methods

Lecturer: Jiaming Liang — *September 28, 2023*

1 Accelerated composite gradient framework

We are interested in solving

$$\min\{\phi(x) := f(x) + h(x)\}$$

- h is closed and convex;
- f is closed and convex, $\operatorname{dom} h \subseteq \operatorname{dom} f$, and f is L-smooth over some set Ω satisfying $\Omega \supset \operatorname{dom} h$;
- the optimal set X_* is nonempty.

Algorithm 1 Accelerated composite gradient (ACG) framework

Input: Initial point $x_0 \in \operatorname{dom} h$, set $y_0 = x_0$, $A_0 = 0$

for $k \geq 0$ **do**

Step 1. Choose $L_k > 0$ and compute

$$a_k = \frac{1 + \sqrt{1 + 4L_k A_k}}{2L_k}, \quad A_{k+1} = A_k + a_k, \quad \tilde{x}_k = \frac{A_k y_k + a_k x_k}{A_{k+1}} \tag{1}$$

Step 2. Compute x_{k+1} and y_{k+1} using one of the rules listed below.

end for

Definition 1. *Define*

$$y(\tilde{x}_k; L_k) := \operatorname{argmin}\left\{\ell_f(x; \tilde{x}_k) + h(x) + \frac{L_k}{2}\|x - \tilde{x}_k\|^2 : x \in \mathbb{R}^n\right\} \tag{2}$$

and

$$\mathcal{C}(y; \tilde{x}) := \frac{2\left[f(y) - \ell_f(y; \tilde{x})\right]}{\|y - \tilde{x}\|^2}. \tag{3}$$

We say the positive parameter L_k is a good local curvature of f at $\tilde{x}_k$, if

$$\mathcal{C}(y(\tilde{x}_k; L_k); \tilde{x}_k) \leq L_k. \tag{4}$$

We will now describe three possible rules for computing the iterates x_{k+1} and y_{k+1} in step 2 of the above framework.

(i) (FISTA rule) This rule sets $y^f_{k+1} = y(\tilde{x}_k; L_k)$ where $y(\tilde{x}_k; L_k)$ is defined in (2) and $L_k > 0$ is a good upper curvature of f at $\tilde{x}_k$, and computes x_{k+1} as

$$x^f_{k+1} = P_\Omega\left(\frac{A_{k+1}}{a_k} y^f_{k+1} - \frac{A_k}{a_k} y_k\right). \tag{5}$$

FISTA rule was first introduced by Nesterov when h is the indicator function of a nonempty closed convex set and was later extended to general composite closed convex functions by Beck and Teboulle.

(ii) (AT rule) This rule computes

$$x^a_{k+1} = \underset{u\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ a_k \left[\ell_f(u; \tilde{x}_k) + h(u)\right] + \frac{1}{2}\|u - x_k\|^2 \right\} \tag{6}$$

and

$$y^a_{k+1} = \frac{A_k y_k + a_k x_{k+1}}{A_{k+1}}. \tag{7}$$

This rule was introduced by Auslender and Teboulle.

(iii) (LLM rule) This rule sets $y_{k+1} = y^a_{k+1}$ as in the FISTA rule and and x_{k+1} as in the AT rule. LLM rule was introduced by Lu, Lan and Monteiro.

Lemma 1. *For every $k \geq 0$, we define*

$$\tilde{\gamma}_k(u) := \ell_f(u; \tilde{x}_k) + h(u), \tag{8}$$

$$\gamma_k(u) := \tilde{\gamma}_k(y^f_{k+1}) + L_k\langle \tilde{x}_k - y^f_{k+1}, u - y^f_{k+1}\rangle. \tag{9}$$

Then the following statements hold for every $k \geq 0$:

(a) γ_k minorizes $\tilde{\gamma}_k$, $\tilde{\gamma}_k(y^f_{k+1}) = \gamma_k(y^f_{k+1})$,

$$\min_u \left\{ \tilde{\gamma}_k(u) + \frac{L_k}{2}\|u - \tilde{x}_k\|^2 \right\} = \min_u \left\{ \gamma_k(u) + \frac{L_k}{2}\|u - \tilde{x}_k\|^2 \right\},$$

and these minimization problems have y^f_{k+1} as unique optimal solution;

(b) for every $u \in \operatorname{dom} h$, $\tilde{\gamma}_k(u) \leq \phi(u)$;

(c) $x^f_{k+1} = \operatorname{argmin}\left\{a_k\gamma_k(u) + \|u - x_k\|^2/2 : u \in \Omega\right\}$;

(d) $\{x^a_k\}$, $\{y^f_k\}$, and $\{y^a_k\}$ are contained in $\mathcal{H}$, while $\{x^f_k\}$ and $\{\tilde{x}_k\}$ lie in Ω.

Lemma 2. *The following statements about scalars a_k and A_K hold for every $k \geq 0$:*

(a) $A_{k+1} = L_k a_k^2$*;*

(b)

$$A_k \geq \frac{1}{4}\left(\sum_{i=0}^{k-1} \frac{1}{\sqrt{L_i}}\right)^2.$$

Proof. (a) It is easy to see that a_k as in (1) is the solution of

$$L_k a_k^2 - a_k - A_k = 0.$$

Using the second relation in (1), we have

$$A_{k+1} = A_k + a_k = L_k a_k^2.$$

(b) It follows from the first two relations in (1) that

$$\sqrt{A_{i+1}} = \left(A_i + \frac{1 + \sqrt{1 + 4L_i A_i}}{2L_i}\right)^{1/2} \geq \left(A_i + \frac{1 + 2\sqrt{L_i A_i}}{2L_i}\right)^{1/2} \geq \sqrt{A_i} + \frac{1}{2\sqrt{L_i}}.$$

Hence, we have

$$\sqrt{A_k} \geq \sqrt{A_0} + \sum_{i=0}^{k-1} \frac{1}{2\sqrt{L_i}} = \sum_{i=0}^{k-1} \frac{1}{2\sqrt{L_i}},$$

where the equality is due to $A_0 = 0$. □

Proposition 1. *For every $k \geq 0$, assume that L_k is a good local curvature. Then, for every $k \geq 0$ and $u \in \operatorname{dom} h$, the following inequality holds for every rule*

$$A_{k+1}[\phi(y_{k+1}) - \phi(u)] + \frac{1}{2}\|u - x_{k+1}\|^2 \leq A_k[\phi(y_k) - \phi(u)] + \frac{1}{2}\|u - x_k\|^2. \tag{10}$$

Proof. We first show (10) holds for the **FISTA rule**. For the ease of notation, we omit the superscript "f" in the proof for the FISTA rule. Using Lemma 1(c), the fact that $a_k\gamma_k(u) + \frac{1}{2}\|u - x_k\|^2$ is 1-strongly convex, and Theorem 5 of Lecture 3, we have for every $u \in \operatorname{dom} h$,

$$\begin{aligned} A_k\gamma_k(y_k) + a_k\gamma_k(u) + \frac{1}{2}\|u - x_k\|^2 - \frac{1}{2}\|u - x_{k+1}\|^2 &\geq A_k\gamma_k(y_k) + a_k\gamma_k(x_{k+1}) + \frac{1}{2}\|x_{k+1} - x_k\|^2 \\ &\geq A_{k+1}\gamma_k(\hat{y}_{k+1}) + \frac{1}{2}\frac{A_{k+1}^2}{a_k^2}\|\hat{y}_{k+1} - \tilde{x}_k\|^2 = A_{k+1}\left[\gamma_k(\hat{y}_{k+1}) + \frac{L_k}{2}\|\hat{y}_{k+1} - \tilde{x}_k\|^2\right] \end{aligned} \tag{11}$$

where $\hat{y}_{k+1} = (A_k y_k + a_k x_{k+1})/A_{k+1}$ and the second inequality is due to the convexity of γ_k, and the equality is due to Lemma 2(a). It follows from Lemma 1(a) and the assumption that L_k is a good local curvature that

$$\begin{aligned}\gamma_k(\hat{y}_{k+1}) + \frac{L_k}{2}\|\hat{y}_{k+1} - \tilde{x}_k\|^2 &\geq \gamma_k(y_{k+1}) + \frac{L_k}{2}\|y_{k+1} - \tilde{x}_k\|^2 \\ &= \tilde{\gamma}_k(y_{k+1}) + \frac{L_k}{2}\|y_{k+1} - \tilde{x}_k\|^2 \\ &\geq \phi(y_{k+1}).\end{aligned}$$

Plugging the above inequality into (11) and rearraging the terms, we have

$$\begin{aligned}A_{k+1}\phi(y_{k+1}) + \frac{1}{2}\|u - x_{k+1}\|^2 - \frac{1}{2}\|u - x_k\|^2 &\leq A_k\gamma_k(y_k) + a_k\gamma_k(u) \\ &\leq A_k\tilde{\gamma}_k(y_k) + a_k\tilde{\gamma}_k(u) \\ &\leq A_k\phi(y_k) + a_k\phi(u),\end{aligned}$$

where the last two inequalities are due to the fact that $\gamma_k \leq \tilde{\gamma}_k \leq \phi$ (see Lemma 1(a) and (b)). Finally, we conclude that

$$A_{k+1}[\phi(y_{k+1}) - \phi(u)] + \frac{1}{2}\|u - x_{k+1}\|^2 \leq A_k[\phi(y_k) - \phi(u)] + \frac{1}{2}\|u - x_k\|^2.$$

We next show (10) holds for the **AT rule**. For the ease of notation, we omit the superscript "a" in the proof for the AT rule. Using (6), the fact that $a_k\tilde{\gamma}_k(u) + \frac{1}{2}\|u - x_k\|^2$ is 1-strongly convex, and Theorem 5 of Lecture 3, we have for every $u \in \operatorname{dom} h$,

$$\begin{aligned}&A_k\tilde{\gamma}_k(y_k) + a_k\tilde{\gamma}_k(u) + \frac{1}{2}\|u - x_k\|^2 - \frac{1}{2}\|u - x_{k+1}\|^2 \geq A_k\tilde{\gamma}_k(y_k) + a_k\tilde{\gamma}_k(x_{k+1}) + \frac{1}{2}\|x_{k+1} - x_k\|^2 \\ &\geq A_{k+1}\tilde{\gamma}_k(y_{k+1}) + \frac{1}{2}\frac{A_{k+1}^2}{a_k^2}\|y_{k+1} - \tilde{x}_k\|^2 = A_{k+1}\left[\tilde{\gamma}_k(y_{k+1}) + \frac{L_k}{2}\|y_{k+1} - \tilde{x}_k\|^2\right] \\ &\geq A_{k+1}\phi(y_{k+1})\end{aligned}$$

where the second inequality is due to the convexity of γ_k and the definition of y_{k+1} in (7), the equality is due to Lemma 2(a), and the last inequality follows from the assumption that L_k is a good local curvature. The rest of the proof is the same as that for the FISTA rule.

We finally show (10) holds for the **LLM rule**. Using (6), the fact that $a_k\tilde{\gamma}_k(u) + \frac{1}{2}\|u - x_k\|^2$ is 1-strongly convex, and Theorem 5 of Lecture 3, we have for every $u \in \operatorname{dom} h$,

$$\begin{aligned}&A_k\tilde{\gamma}_k(y_k) + a_k\tilde{\gamma}_k(u) + \frac{1}{2}\|u - x_k\|^2 - \frac{1}{2}\|u - x_{k+1}\|^2 \geq A_k\tilde{\gamma}_k(y_k) + a_k\tilde{\gamma}_k(x_{k+1}) + \frac{1}{2}\|x_{k+1} - x_k\|^2 \\ &\geq A_{k+1}\tilde{\gamma}_k(\hat{y}_{k+1}) + \frac{1}{2}\frac{A_{k+1}^2}{a_k^2}\|\hat{y}_{k+1} - \tilde{x}_k\|^2 = A_{k+1}\left[\tilde{\gamma}_k(\hat{y}_{k+1}) + \frac{L_k}{2}\|\hat{y}_{k+1} - \tilde{x}_k\|^2\right] \\ &\geq A_{k+1}\left[\tilde{\gamma}_k(y_{k+1}) + \frac{L_k}{2}\|y_{k+1} - \tilde{x}_k\|^2\right] \geq A_{k+1}\phi(y_{k+1})\end{aligned}$$

where $\hat{y}_{k+1} = (A_k y_k + a_k x_{k+1})/A_{k+1}$ and the second inequality is due to the convexity of γ_k, the equality is due to Lemma 2(a), the third inequality is due to Lemma 1(a), and the last inequality follows from the assumption that L_k is a good local curvature. The rest of the proof is the same as that for the FISTA rule. □

In Step 1 of the ACG framework, we want to choose $L_k > 0$ to be a good local curvature satisfying (4). There two standard ways to do so.

1. **Constant**. Suppose we know the smoothness parameter (i.e., global curvature) L, then we can choose $L_k = L$ for every $k \geq 0$. It follows from Lemma 2(b) that

$$A_k \geq \frac{1}{4}\left(\sum_{i=0}^{k-1} \frac{1}{\sqrt{L_i}}\right)^2 = \frac{k^2}{4L}.$$

2. **Adaptive**. For every $k \geq 0$, we begin from some small $L_k^0 > 0$, check whether (4) is true with L_k replaced by L_k^0, i.e.,

$$\mathcal{C}(y(\tilde{x}_k; L_k^0); \tilde{x}_k) \leq L_k^0.$$

If it is true, then we set $L_k = L_k^0$. Otherwise, we set $L_k^1 = 2L_k^0$, and check whether (4) is true with L_k replaced by L_k^1 and follow the previous step. The search for L_k stays in the loop until (4) is satisfied with some L_k^i, and we output $L_k = L_k^i$. It is easy to show that the interations of this search is bounded by $\log(L_k/L_k^0)$ and $L_k \leq 2L$. Moreover, it follows from Lemma 2(b) that

$$A_k \geq \frac{1}{4}\left(\sum_{i=0}^{k-1} \frac{1}{\sqrt{L_i}}\right)^2 \geq \frac{k^2}{8L}. \tag{12}$$

Theorem 1. *For every $k \geq 0$, we choose $L_k > 0$ to be a good local curvature satisfying* (4)*, then we have*

$$\phi(y_k) - \phi_* \leq \frac{4L\|x_0 - x_*\|^2}{k^2}.$$

Proof. Using Proposition 1 and the fact that $A_0 = 0$, we have

$$A_k[\phi(y_k) - \phi_*] + \frac{1}{2}\|x_k - x_*\|^2 \leq \frac{1}{2}\|x_0 - x_*\|^2.$$

This inequality and (12) imply that

$$\phi(y_k) - \phi_* \leq \frac{1}{2A_k}\|x_0 - x_*\|^2 \leq \frac{4L\|x_0 - x_*\|^2}{k^2}.$$

□