Adaptive and General Acceleration Methods

1 Convex Optimization Problem

In this lecture, we consider the general unconstrained smooth convex optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x).$$

Here we assume that $f(x)$ is an L-smooth convex function, and λ-strongly convex.

In this lecture, we assume that L is known, and we can pick learning rate $\eta \leq 1/L$. However, the strong-convexity parameter λ is either unknown, or $\lambda = 0$. In the case of $\lambda = 0$, the function is convex but not strongly convex. This is the worst case scenario for smooth convex optimization.

In the previous lectures, we have shown that for strongly convex problems, it is possible to improve the gradient descent method by using Nesterov's acceleration method, which is the following algorithm:

$$\begin{aligned} y_t &= x_{t-1} + \beta(x_{t-1} - x_{t-2}) \\ x_t &= y_t - \alpha \nabla f(y_t), \end{aligned}$$

with appropriately chosen α and β.

2 Adaptive Tuning of β with unknown λ

Assume we can pick $\alpha \leq 1/L$, and $\beta = (1-\theta)/(1+\theta)$ with $\theta \leq \sqrt{\eta\lambda}$. Then it follows from Lecture-06 that

$$f(x_t) \leq f(x_*) + (1-\theta)^t \left[f(x_0) - f(x_*) + \frac{\lambda}{2}\|x_* - x_0\|_2^2 \right]. \tag{1}$$

If we know λ, then for fast convergence, we should choose the largest θ possible such that the result holds. This means we should choose a relatively small β so that $\theta = \sqrt{\eta\lambda}$. However, if we do not know λ, then we have to guess it. In such case, we want to choose θ so that θ is as large as possible, and then check whether the convergence result is satisfied. If satisfied, then we have good convergence. If not, then we reduce θ by a factor of 2. We keep doing so until the convergence result (1) is always satisfied.

Unfortunately, it is not possible to check (1) directly because we do not know x_* and $f(x_*)$. In practice, one of the most frequently used method is to check how small the gradient $\|\nabla f(x)\|_2$ is at any point x. Therefore, we have to derive a convergence of gradients from convergence of function values in (1).

Using strong convexity, we have

$$f(x_0) \geq f(x_*) + \frac{\lambda}{2}\|x_* - x_0\|_2^2,$$

therefore

$$f(x_t) \leq f(x_*) + 2(1-\theta)^t \left[f(x_0) - f(x_*)\right].$$

Now from Lecture-04, we obtain

$$\frac{1}{2L}\|\nabla f(x)\|_2^2 \leq [f(x) - f(x_*)] \leq \frac{1}{2\lambda}\|\nabla f(x)\|_2^2.$$

Therefore

$$\frac{1}{2L}\|\nabla f(x_t)\|_2^2 \leq \frac{2}{2\lambda}(1-\theta)^t\|\nabla f(x_0)\|_2^2.$$

Since $\eta \leq 1/L$, we know that if $\theta \leq \sqrt{\eta\lambda}$, then we must have

$$\|\nabla f(x_t)\|_2^2 \leq \frac{2}{\eta\lambda}(1-\theta)^t\|\nabla f(x_0)\|_2^2 \leq \frac{2}{\theta^2}(1-\theta)^t\|\nabla f(x_0)\|_2^2. \tag{2}$$

This is the result we can use to check the convergence of optimization process. This leads to Algorithm 1 to automatically tune β without knowing the strong convexity parameter λ a priori. It tries θ from 0.5, and reduces it if (2). When reduced, it picks a non-increasing gradient and restart the acceleration process from the beginning by setting the initial values to be equal. At each θ, it checks (2) after Δ steps, where Δ is picked so that be bound $\frac{2}{\theta^2}(1-\theta)^t \leq 1$. This means the gradient is expected to reduce if θ satisfies (2).

Algorithm 1: Adaptive Acceleration Method (Theoretically Motivated)

```
Input: $f(x)$, $x_0$, $\alpha = \eta \leq 1/L$
Output: $x_T$
Let $x_{-1} = x_0$
Let $s = 0$
Let $\theta = 0.5$
Let $\beta = (1-\theta)/(1+\theta)$
Let $\Delta = \lceil \log(0.5\theta^2)/\log(1-\theta) \rceil$
for $t = 1, \ldots, T$ do
    Let $y_t = x_{t-1} + \beta(x_{t-1} - x_{t-2})$
    Let $x_t = y_t - \alpha\nabla f(y_t)$
    if $(t-s)\%\Delta == 0$ and $\|\nabla f(x_t)\|_2^2 > 2\theta^{-2}(1-\theta)^{t-s}\|\nabla f(x_s)\|_2^2$ then
        if $\|\nabla f(x_t)\|_2^2 \geq \|\nabla f(x_{t-\Delta})\|_2^2$ then
            Let $x_{t-1} = x_t = x_{t-\Delta}$
        else
            Let $x_{t-1} = x_t$
        Let $s = t$
        Let $\theta = \theta/2$
        Let $\beta = (1-\theta)/(1+\theta)$
        Let $\Delta = \lceil \log(0.5\theta^2)/\log(1-\theta) \rceil$
Return: $x_T$
```

Theorem 1 *Assume that $f(x)$ is λ-strongly convex with $\lambda > 0$. Let $m = \lceil \log_2(0.5/\sqrt{\eta\lambda}) \rceil$, and let*

$$T_0 = \left\lceil \frac{4(m+1)}{\sqrt{\eta\lambda}} \right\rceil,$$

then when $T \geq T_0$, we have

$$\|\nabla f(x_t)\|_2^2 \leq 2\bar{\theta}^{-2}(1-\bar{\theta})^{T-T_0}\|\nabla f(x_0)\|_2^2,$$

where $\bar{\theta} = 2^{-m-1} \geq 0.5\sqrt{\eta\lambda}$.

The theorem implies that to achieve ϵ sub-optimality in terms of gradient, we need

$$O(\sqrt{\kappa}\log(1/\epsilon) + \sqrt{\kappa}\log\kappa)$$

total number of iterations. Here we assume that $\eta \geq c/L$ for some constant c, and $\kappa = L/\lambda$ is the condition number. Therefore even if λ is unknown, we can still guess the right λ by paying an extra penalty of $\sqrt{\kappa}\log\kappa$.

However, this algorithm is too conservative in practice. We may choose a simpler and more aggressive method where we tune β using observed convergence rate, measured by

$$\gamma = \log\frac{\|\nabla f(y_t)\|_2^2}{\|\nabla f(y_{t-1})\|_2^2}.$$

We obtain the following Algorithm 2, which performs better than Algorithm 1. Note that according to our convergence theorem we should set $\theta = 1-\gamma$ and set $\beta = \gamma/(2-\gamma)$. However, in practice, we find that the more aggressive choice of $\beta = \gamma$ is better. We can use this idea for adaptation in the heavy ball method as well, and the corresponding choice is closely related to the FletcherReeves method for nonlinear generalization of Conjugate Gradient.

Algorithm 2: Adaptive Acceleration Method (Practically Simplified)

```
Input: $f(x)$, $x_0$, $\alpha = \eta \leq 1/L$
Output: $x_T$
Let $x_{-1} = x_0$
Let $\gamma = 0$
Let $y_0 = x_0$
for $t = 1, \ldots, T$ do
    Let $\beta = \min(1, \exp(\gamma))$
    Let $y_t = x_{t-1} + \beta(x_{t-1} - x_{t-2})$
    Let $x_t = y_t - \alpha\nabla f(y_t)$
    Let $\gamma = 0.8\gamma + 0.2\ln(\|\nabla f(y_t)\|_2^2/\|\nabla f(y_{t-1})\|_2^2)$
Return: $x_T$
```

3 Non-Strongly-Convex Problems

If the function is not strongly convex, then theory of Algorithm 1 does not apply. However, one can modify the acceleration algorithm so that a theory can still be obtained even when $\lambda = 0$. The general situation can be stated as in Algorithm 3. If we set $\gamma_0 = 1/\eta$ and $\lambda = 0$, we obtain an algorithm for the non-strongly-convex case.

The key idea of this algorithm is to start with a large θ, and gradually decrease it until $\theta \to \lambda$. This is similar to what is done in Algorithm 1. However, the decrease of θ is manually controlled by a deterministic sequence, and not adaptive. Therefore a theory can be proved accordingly even when $\lambda = 0$.

```
Algorithm 3: Nesterov's General Acceleration Method
  Input: $f(x)$, $x_0$, $\eta \le 1/L$
      $\lambda \in [0, 1/\eta]$ (default is $\lambda = 0$)
      $\gamma_0 \in [\lambda, 1/\eta]$ (default is $\gamma_0 = 1/\eta$)
  Output: $x_T$
  Let $x_{-1} = x_0$
  Let $\theta_0 = 1$
  for $t = 1, \dots, T$ do
      Solve for $\theta_t$: $\theta_t^2/\eta = \theta_t\lambda + (1-\theta_t)\gamma_{t-1}$
      Let $\gamma_t = (1-\theta_t)\gamma_{t-1} + \theta_t\lambda$
      Let $\beta_t = (\theta_t^{-1} - 1)(\theta_{t-1}^{-1} - 1)\gamma_{t-1}/(\eta^{-1} - \lambda)$
      Let $y_t = x_{t-1} + \beta_t(x_{t-1} - x_{t-2})$
      Let $x_t = y_t - \eta\nabla f(y_t)$
  Return: $x_T$
```

We will have the following general Theorem.

Theorem 2 *Assume $f(x)$ is L-smooth and λ-strongly convex. Then for all $x_* \in \mathbb{R}^d$, we have*

$$f(x_t) \le f(x_*) + \lambda_t \left[f(x_0) - f(x_*) + \frac{\gamma_0}{2}\|x_* - x_0\|_2^2 \right],$$

where

$$\lambda_t = \prod_{s=1}^{t}(1-\theta_s).$$

If we pick $\gamma_0 = \lambda$, then $\gamma_t = \lambda$, and $\theta_t = \theta$. This leads to the result of last lecture.

If we set $\lambda = 0$ and $\gamma_0 = 1/\eta$, then we have the recursion:

$$\theta_t^2 = (1-\theta_t)r_{t-1}\eta \qquad \gamma_t = (1-\theta_t)\gamma_{t-1}.$$

It can be shown that $\theta_t = O(1/t)$ and $\theta_t = O(1/t^2)$.

Therefore for nonstrongly convex functions, we have a convergence of

$$f(x_t) \le f(x_*) + O(t^{-2}) \left[f(x_0) - f(x_*) + \frac{1}{2\eta}\|x_* - x_0\|_2^2 \right].$$

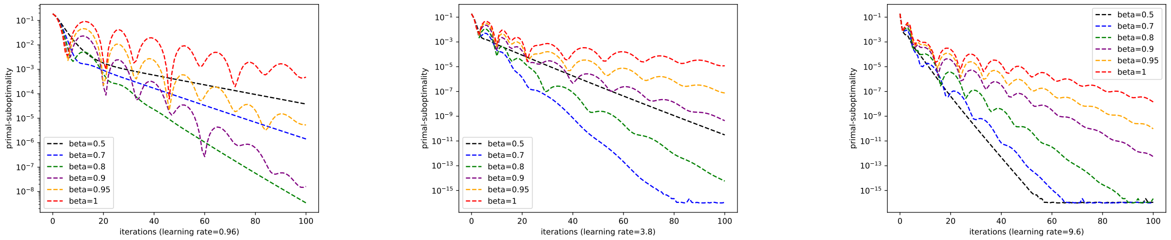
This is faster than that of gradient descent, which is $O(1/t)$.

4 Empirical Studies

We illustrate the sensitivity of β with different learning rate α. using a regularized logistic regression problem with condition number $\kappa \approx 10$. Figure 1 compares the convergence of Nesterov acceleration method, with different learning rate α and different β. In Figure 1a, $\alpha \approx 1$, In Figure 1b, $\alpha \approx 4$, In Figure 1c, $\alpha \approx 10$. We can see that parameter tuning of β is important. Small β leads to slower convergence, and larger β leads to oscillation. Moreover, we can have a larger β without oscillation if α is small. As a rule of thumb, we want to choose as large a β as possible without oscillation. This would lead to optimal trade-off. If we observe oscillation, then β is too large.

Figure 2 shows the performance of Adaptive Algorithms under different learning rates.

Figure 3 shows the performance of Adaptive Algorithms to the general Acceleration Algorithm.

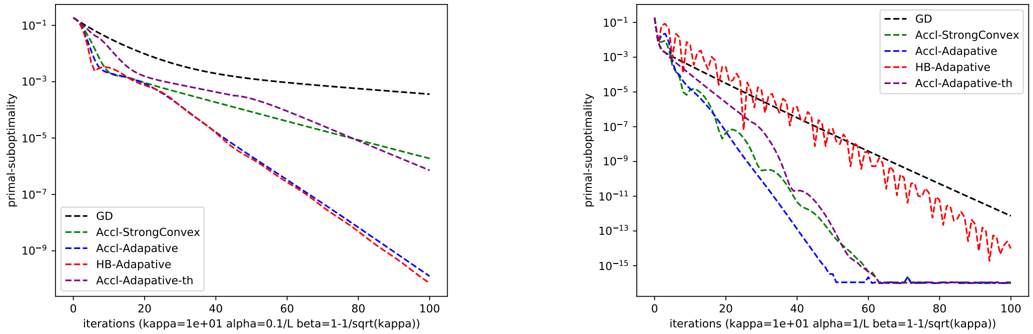


(a) $\alpha \approx 1$

(b) $\alpha \approx 4$

(c) $\alpha \approx 10$

Figure 1: Convergence Comparisons with Fixed α



(a) $\alpha = 0.1/L$

(b) $\alpha \approx 1/L$

Figure 2: Convergence Comparisons with Fixed Learning Rate

5 Proof of Theorem 1

First, when $\theta \leq \sqrt{\eta\lambda}$, The condition

$$\|\nabla f(x_t)\|_2^2 > 2\theta^{-2}(1-\theta)^{t-s}\|\nabla f(x_s)\|_2^2$$

will always be false. If follows that line 9 in the algorithm can be true only for no more m times because after m times, $\theta \leq \sqrt{\eta\lambda}$.

Note that when line 9 in the algorithm holds true, and after executing the if statements, we have

$$\|\nabla f(x_t)\|_2^2 \leq (1-\theta)^{t-s-2\Delta}\|\nabla f(x_s)\|_2^2.$$

If $t - s = \Delta$, the claim holds trivially. Otherwise, $t - s \geq 2\Delta$, and since $2\theta^{-2}(1-\theta)^{\Delta} \leq 1$, we have

$$\|\nabla f(x_t)\|_2^2 \leq \|\nabla f(x_{t-\Delta})\|_2^2 \leq 2\theta^{-2}(1-\theta)^{t-s-\Delta}\|\nabla f(x_s)\|_2^2 \leq (1-\theta)^{t-s-2\Delta}\|\nabla f(x_s)\|_2^2.$$

Let Δ_j be the Δ at j-th time when the if-statement at line 9 of the algorithm holds true ($j = 1, \ldots, m_0 \leq m$), then we have $\Delta_j = \lceil \log(0.5\theta^2)/\log(1-\theta) \rceil$ with $\theta = 2^{-j}$. We have the bound

$$\Delta_j = \lceil \log(0.5\theta^2)/\log(1-\theta) \rceil \leq 2^j(j+1)\log 2 + 1.$$

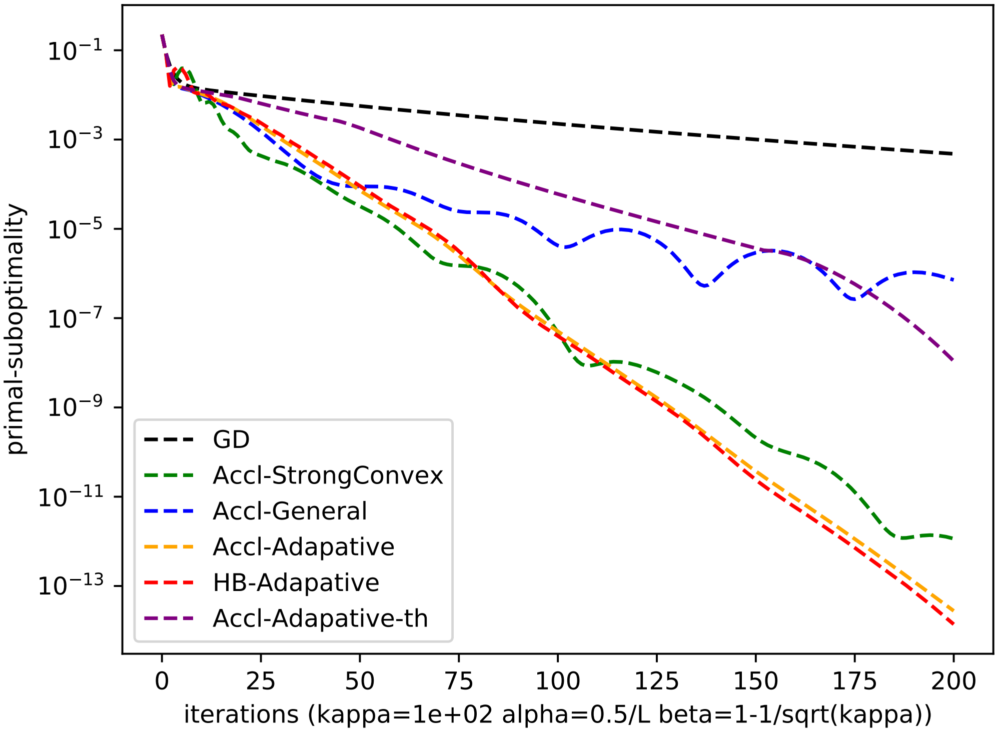


Figure 3: Convergence Comparison with the General Acceleration Algorithm

Therefore

$$T_0 \geq 2 \sum_j \Delta_j$$

Let s_j be the s after the j-th time, then

$$\|\nabla f(x_{s_{m_0}})\|_2^2 \leq \prod_{j=1}^{m_0} (1-\bar{\theta})^{s_j - s_{j-1} - 2\Delta_j} \|\nabla f(x_0)\|_2^2 = (1-\bar{\theta})^{s_{m_0} - 2\sum_j \Delta_j} \|\nabla f(x_0)\|_2^2 \leq (1-\bar{\theta})^{s_{m_0} - T_0} \|\nabla f(x_0)\|_2^2.$$

Therefore when $T \geq T_0$, we have

$$\|\nabla f(x_{s_{m_0}})\|_2^2 \leq 2\bar{\theta}^{-2}(1-\bar{\theta})^{T - s_{m_0}} \|\nabla f(x_{s_{m_0}})\|_2^2 \leq (1-\bar{\theta})^{T-T_0} \|\nabla f(x_0)\|_2^2.$$

This proves the result.

6 Proof of Theorem 2

The following lemma can be used to define estimate-sequence.

Lemma 1 *Let $x^+ = y - \eta \nabla f(y)$. We define*

$$\phi(z; y) = f(x^+) - \frac{1}{2\eta} \|x^+ - y\|_2^2 + \frac{1}{\eta}(y - x^+)^\top (z - x^+) + \frac{\lambda}{2} \|z - y\|_2^2.$$

Then the following inequality holds:

$$\phi(z;y) \leq f(z).$$

Proof We have

$$\begin{aligned}
f(z) \geq & f(y) + \nabla f(y)^\top (z-y) + \frac{\lambda}{2}\|z-y\|_2^2 \\
= & f(y) + \nabla f(y)^\top (x^+ - y) + \nabla f(y)^\top (z - x^+) + \frac{\lambda}{2}\|z-y\|_2^2 \\
\geq & f(x^+) - \frac{1}{2\eta}\|x^+ - y\|_2^2 + \frac{1}{\eta}(y - x^+)^\top (z - x^+) + \frac{\lambda}{2}\|z-y\|_2^2 \\
= & \phi(z;y).
\end{aligned}$$

The first inequality uses the strong convexity. The second inequality uses the smoothness with $1/\eta \geq L$, and replaces $\nabla f(y)$ by $\eta^{-1}(y - x^+)$. ■

Using notations in Lecture 06, we may define an estimate sequence recursively as

$$\phi_t(z) = (1-\theta_t)\phi_{t-1}(z) + \theta_t \phi(z; y_t), \quad \lambda_t = (1-\theta_t)\lambda_{t-1},$$

with

$$\phi_0(z) = f(x_0) + \frac{\gamma_0}{2}\|z - x_0\|_2^2, \quad \lambda_0 = 1.$$

We prove that for this estimate sequence, the following holds. Results of Lecture 06 then implies the theorem.

Lemma 2 *We have*

$$f(x_t) \leq \phi_t(v_t) = \min_z \phi_t(z).$$

Proof We have

$$\phi_t(z) = \phi_t(v_t) + \frac{\gamma_t}{2}\|z - v_t\|_2^2,$$

where $\gamma_t = (1-\theta_t)\gamma_{t-1} + \lambda\theta_t$. Thus

$$\phi_t(z) = (1-\theta_t)\phi_{t-1}(v_{t-1}) + \theta_t \phi(z; y_t) + (1-\theta_t)\frac{\gamma_{t-1}}{2}\|z - v_{t-1}\|_2^2.$$

Since v_t minimizes $\phi_t(z)$, we have the first order condition

$$\theta_t \frac{1}{\eta}(y_t - x_t) + \theta_t \lambda (v_t - y_t) + (1-\theta_t)\gamma_{t-1}(v_t - v_{t-1}) = 0. \tag{3}$$

By induction, it can be shown from this equation that

$$v_t = \theta_t^{-1}(x_t + (\theta_t - 1)x_{t-1}). \tag{4}$$

Moreover, from (3), we obtain

$$x_t - y_t = \eta\lambda(v_t - y_t) + a_t(v_t - v_{t-1})], \quad a_t = (\theta_t^{-1} - 1)\eta\gamma_{t-1}.$$

It also implies that

$$-\|x_t - y_t\|_2^2 = -(\eta\lambda + a_t)[\eta\lambda\|v_t - y_t\|_2^2 - a_t\|v_t - v_{t-1}\|_2^2] + \eta\lambda a_t\|v_{t-1} - y_t\|_2^2. \tag{5}$$

It follows by induction that

$$\begin{aligned}
\phi_t(v_t) =&(1-\theta_t)\left[\phi_{t-1}(v_{t-1}) + \frac{\gamma_{t-1}}{2}\|v_t - v_{t-1}\|_2^2\right] + \theta_t\phi(v_t; y_t) \\
\geq&(1-\theta_t)\left[f(x_{t-1}) + \frac{\gamma_{t-1}}{2}\|v_t - v_{t-1}\|_2^2\right] + \theta_t\phi(v_t; y_t) \\
\geq&(1-\theta_t)\left[\phi(x_{t-1}; y_t) + \frac{\gamma_{t-1}}{2}\|v_t - v_{t-1}\|_2^2\right] + \theta_t\phi(v_t; y_t) \\
=&f(x_t) - \frac{1}{2\eta}\|y_t - x_t\|_2^2 + \frac{1}{\eta}(y_t - x_t)^\top((1-\theta_t)x_{t-1} + \theta_t v_t - x_t) \\
&+ (1-\theta_t)\frac{\lambda}{2}\|x_{t-1} - y_t\|_2^2 + \theta_t\frac{\lambda}{2}\|v_t - y_t\|_2^2 + (1-\theta_t)\frac{\gamma_{t-1}}{2}\|v_t - v_{t-1}\|_2^2 \\
=&f(x_t) - \frac{1}{2\eta}\|y_t - x_t\|_2^2 + (1-\theta_t)\frac{\lambda}{2}\|x_{t-1} - y_t\|_2^2 + \theta_t\frac{\lambda}{2}\|v_t - y_t\|_2^2 + (1-\theta_t)\frac{\gamma_{t-1}}{2}\|v_t - v_{t-1}\|_2^2 \\
=&f(x_t) + (1-\theta_t)\frac{\lambda}{2}\|x_{t-1} - y_t\|_2^2 + (\theta_t^{-1} - 1)\eta\gamma_{t-1}\frac{\lambda}{2}\|v_{t-1} - y_t\|_2^2.
\end{aligned}$$

This finishes the induction. In the above derivation, the first equality is definition. The first inequality uses the induction hypothesis. The second inequality uses Lemma 1. The second equality is by definition of $\phi(x; y)$. The third equality uses (4). The final equality uses (5). ■