Introductory Lectures on Optimization

Acceleration Methods (2)

Hui Qian
qianhui@zju.edu.cn

College of Computer Science, Zhejiang University

November 6, 2024

Outline

Part I
Accelerated Gradient Descent

Unconstrained Minimization Problem

Consider the following unconstrained minimization problem:

$$\min_{\boldsymbol{x} \in \mathbb{R}^N} f(\boldsymbol{x}),$$

where f is strongly convex: $f \in \mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n),\ \mu \geq 0.$

This family of classes also contains the class of convex functions with Lipschitz continuous gradient, since

$$\mathcal{S}^{1,1}_{0,L}(\mathbb{R}^n) \equiv \mathcal{F}^{1,1}_{L}(\mathbb{R}^n).$$

Remark. The gradient is Lipschitz continuous.

Effciency Estimation

For the Gradient Descent method, we proved the following convergence rates:

$$\mathcal{F}_L^{1,1}(\mathbb{R}^n) : f(\boldsymbol{x}_k) - f^* \leq \frac{2L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|}{k+4},$$

$$\mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n) : f(\boldsymbol{x}_k) - f^* \leq \frac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2k}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2.$$

These estimates differ from our lower complexity bounds by an order of magnitude.

- Generally speaking, this does not mean that the Gradient Method is not optimal (it may be that the lower bounds are too optimistic) .
- However, we will see that in our case the lower bounds are exact up to a constant factor.

We prove this by constructing a method with rate of convergence proportional to these bounds.

Effciency Estimation

The upper bounds vs. the lowver bounds

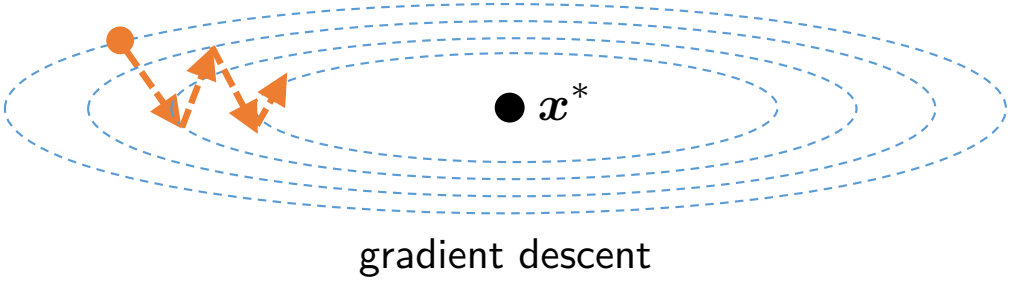
$$\mathcal{F}_L^{1,1}(\mathbb{R}^n): f(x_k) - f^* \leq \frac{2L\|x_0 - x^*\|}{k+4}, \qquad \text{(upper bound of GD)}$$

$$\mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n): f(x_k) - f^* \leq \frac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2k}\|x_0 - x^*\|^2, \qquad \text{(upper bound of GD)}$$

$$\mathcal{F}_L^{1,1}(\mathbb{R}^n): f(x_k) - f^* \geq \frac{3L\|x_0 - x^*\|^2}{32(k+1)^2}, \qquad \text{(lower bound of class)}$$

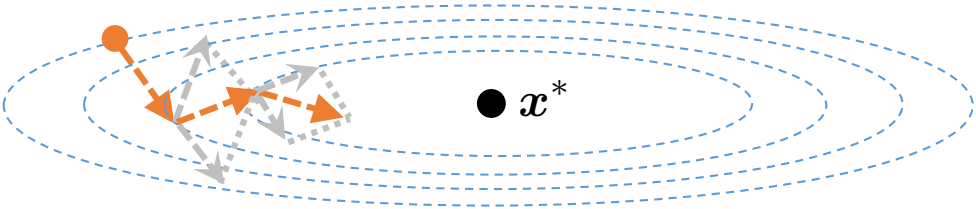
$$\mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n): f(x_k) - f^* \geq \frac{\mu}{2}\left(\frac{\sqrt{Q_f}-1}{\sqrt{Q_f}+1}\right)^{2k}\|x_0 - x^*\|^2. \qquad \text{(lower bound of class)}$$

Effciency Estimation



This picture comes from ELE 522: Large-Scale Optimization for Data Science of Princeton University.

Effciency Estimation



This picture comes from ELE 522: Large-Scale Optimization for Data Science of Princeton University.

Historical Notes on Accelerated GD

Nesterov first discovered an accelerated gradient descent method for smooth functions in 1983[Nesterov, 1983], and proposed a more generalized method in 1988 [Nesterov and Nemirovsky, 1988]. Nesterov's technique boils down to a mechanism called Estimate Sequences.

- In 2004 and 2007, Nesterov generalized the acceleration to non-smooth objective functions [Nesterov, 2013] and combinatorial functions [Nesterov, 2007], respectively.
- In 2008, FISTA(Fast Iterative Shrinkage-Thresholding Algorithm) [Beck and Teboulle, 2009].
- In 2009, Unified Analysis of Smooth and Nonsmooth Function Classes[Tseng and Yun, 2009].

Since then, various first-order algorithms have had their own accelerated versions.

Historical Notes on Accelerated GD

Since then, new approaches have been proposed, with the interpretation and study of Nesterov's acceleration. The understanding has also deepened.

- ODE perspective: Differential Equation[Su et al., 2014], Integral Quadratic Constraints [Lessard et al., 2016], Variational Perspective[Wibisono et al., 2016].
- Geometry perspective: Linear Coupling[Allen-Zhu and Orecchia, 2014], Geometric interpretation loosely inspired by the ellipsoid method[Bubeck et al., 2015]
- Game theory perspective: view this method as a natural iterative buyer-supplier game[Lan and Zhou, 2018]

Estimate Sequences

The Nesterov acceleration schemes and efficiency bounds of optimal methods are based on the notion of estimating sequences.

Definition 6 (Definition 2.2.1 of Nesterov [2013])

A pair of sequences $\{\phi_k(\boldsymbol{x})\}_{k=0}^{\infty}$ and $\{\lambda_k\}_{k=0}^{\infty}$, $\lambda_k \geq 0$, are called the estimating sequences of function of $f(\boldsymbol{x})$ if

$$\lambda_k \to 0,$$

and for any $x \in \mathbb{R}^n$ and all $k \geq 0$ we have

$$\phi_k(\boldsymbol{x}) \leq (1 - \lambda_k) f(\boldsymbol{x}) + \lambda_k \phi_0(\boldsymbol{x}). \tag{5}$$

Estimate Sequences

Lemma 7 (Lemma 2.2.1)

If for some sequence of points $\{\boldsymbol{x}_k\}$, we have

$$f(\boldsymbol{x}_k) \leq \phi_k^* \equiv \min_{\boldsymbol{x} \in \mathbb{R}^n} \phi_k(\boldsymbol{x}), \tag{6}$$

Then $f(\boldsymbol{x}_k) - f^* \leq \lambda_k [\phi_0(\boldsymbol{x}^*) - f^*] \to 0$.

Proof. Indeed,

$$\begin{aligned} f(\boldsymbol{x}_k) \underset{(6)}{\leq} \phi_k^* &= \min_{\boldsymbol{x} \in \mathbb{R}^n} \phi_k(\boldsymbol{x}) \underset{(5)}{\leq} \min_{\boldsymbol{x} \in \mathbb{R}^n} [(1 - \lambda_k) f(\boldsymbol{x}) + \lambda_k \phi_0(\boldsymbol{x})] \\ &\leq (1 - \lambda_k) f(\boldsymbol{x}^*) + \lambda_k \phi_0(\boldsymbol{x}^*). \end{aligned}$$

□

Estimate Sequences

$$f(\boldsymbol{x}_k) \leq \phi_k^*, \quad \phi_\infty(\boldsymbol{x}) \leq f(\boldsymbol{x}),$$

Thus, for any sequence $\{\boldsymbol{x}_k\}$, satisfying (6), we can derive its rate of convergence directly from the convergence rate of the sequence $\{\lambda_k\}$.

However, at this moment we have two serious questions.

- Firstly, we do not know how to form the estimating sequences.
- Secondly, we do not know how to satisfy inequalities (6).

Estimate Sequences

The first question is simpler.

Lemma 8 (Lemma 2.2.2 of Nesterov [2013])

Assume that

1. $f \in \mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$,
2. $\phi_0(\boldsymbol{x})$ is an arbitrary convex function on $\mathbb{R}^n$,
3. $\{\boldsymbol{y}_k\}_{k=0}^{\infty}$ is an arbitrary sequence of points in $\mathbb{R}^n$,
4. the coefficients $\{\alpha_k\}_{k=0}^{\infty}$ satisfy conditions $\alpha_k \in (0,1)$, $\sum_{k=0}^{\infty} \alpha_k = \infty$, ,
5. we choose $\lambda_0 = 1$.

To be continued...

Estimate Sequences

Lemma 8

(Continued.) Then, the pair of sequences $\{\phi_k(\boldsymbol{x})\}_{k=0}^{\infty}$ and $\{\lambda_k\}_{k=0}^{\infty}$ defined recursively by the relations:

$$\begin{aligned} \lambda_{k+1} &= (1-\alpha_k)\lambda_k, \\ \phi_{k+1}(\boldsymbol{x}) &= (1-\alpha_k)\phi_k(\boldsymbol{x}) \\ &\quad + \alpha_k[f(\boldsymbol{y}_k) + \langle \nabla f(\boldsymbol{y}_k),\ \boldsymbol{x}-\boldsymbol{y}_k\rangle + \frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{y}_k\|^2], \end{aligned} \tag{7}$$

are estimating sequences.

注：ϕ_k 是 ϕ_0 和一系列下界的平均。

Estimate Sequences

Proof. Indeed, $\phi_0(\boldsymbol{x}) \leq (1-\lambda_0)f(\boldsymbol{x}) + \lambda_0\phi_0(\boldsymbol{x}) \equiv \phi_0(\boldsymbol{x})$. Further, let (5) hold for some $k \geq 0$. Then

$$\begin{aligned}
\phi_{k+1}(\boldsymbol{x}) &\leq (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k f(\boldsymbol{x}) \text{ (strongly convex)} \\
&= (1-(1-\alpha_k)\lambda_k)f(\boldsymbol{x}) + (1-\alpha_k)(\underbrace{\phi_k(\boldsymbol{x}) - (1-\lambda_k)f(\boldsymbol{x})}_{\leq \lambda_k\phi_0(\boldsymbol{x})}) \\
&\leq (1-(1-\alpha_k)\lambda_k)f(\boldsymbol{x}) + (1-\alpha_k)\lambda_k\phi_0(\boldsymbol{x}) \text{ (definition)} \\
&= (1-\lambda_{k+1})f(\boldsymbol{x}) + \lambda_{k+1}\phi_0(\boldsymbol{x}).
\end{aligned}$$

Thus, we arrive at

$$\phi_k(\boldsymbol{x}) \leq (1-\lambda_k)f(\boldsymbol{x}) + \lambda_k\phi_0(\boldsymbol{x}).$$

Estimate Sequences

Proof. (Continued.) It remains to note that condition 4 ensures $\lambda_k \to 0$.
Since $\lambda_k > 0$ holds for all $k = 0, 1, \ldots, k$. We have $\ln \lambda_{k+1} = \ln\{(1-\alpha_k)\lambda_k\}$. Therefore,

$$\ln \lambda_{k+1} - \ln \lambda_k = \ln(1-\alpha_k) \leq -\alpha_k. \quad (\text{ using } \ln(1+x) \leq x \text{ here.})$$

We telescope the above inequality and obtain

$$\ln \lambda_{k+1} - \ln \lambda_0 \leq -\sum_{i=0}^{k} \alpha_i.$$

With conditoin 4, it implies that $\ln \lambda_{k+1} \to -\infty$, when $k \to \infty$. That is $k \to \infty$, $\lambda_k \to 0$. $\square$

Estimate Sequence

Thus, the above statement provides us with some rules for updating the estimating sequences. Now we have two control sequences which can help us to maintain recursively the relation (6).

At this moment, we are also free in our choice of initial function $\phi_0(x)$.

Let us choose it as a simple quadratic function. Then, we can obtain a closed form recurrence for values ϕ_k^*.

$$f(\boldsymbol{x}_k) \leq \phi_k^* \equiv \min_{\boldsymbol{x} \in \mathbb{R}^n} \phi_k(\boldsymbol{x}), \tag{6}$$

Estimate Sequence

Lemma 9 (Lemma 2.2.3)

Let

$$\phi_0(\boldsymbol{x}) = \phi_0^* + \frac{\gamma_0}{2} \|\boldsymbol{x} - \boldsymbol{v}_0\|^2 .$$

Then the process (7) preserves the canonical form of functions $\{\phi_k(x)\}$:

$$\phi_k(\boldsymbol{x}) \equiv \phi_k^* + \frac{\gamma_k}{2} \|\boldsymbol{x} - \boldsymbol{v}_k\|^2 , \tag{8}$$

Where the sequences $\{\gamma_k\}$, $\{\boldsymbol{v}_k\}$ and $\{\phi_k^*\}$ are defined as follows:

To be continued...

Estimate Sequence

Lemma 9
(Continued.)

$$\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\mu,$$

$$\boldsymbol{v}_{k+1} = \frac{1}{\gamma_{k+1}}\left[(1-\alpha_k)\gamma_k\boldsymbol{v}_k + \alpha_k\mu\boldsymbol{y}_k - \alpha_k\nabla f(\boldsymbol{y}_k)\right],$$

$$\phi^*_{k+1} = (1-\alpha_k)\phi^*_k + \alpha_k f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\left\|\nabla f(\boldsymbol{y}_k)\right\|^2 + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\left\|\boldsymbol{y}_k - \boldsymbol{v}_k\right\|^2 + \langle\nabla f(\boldsymbol{y}_k),\ \boldsymbol{v}_k - \boldsymbol{y}_k\rangle\right).$$

Remark

1. First of all, we have to confirm whether the canonical form can be preserved. If possible, $\phi_k(\boldsymbol{x})$ is quadratic, so it can be formulated as (8);
2. Then, what the lemma does is deduce the definitions of $\{\gamma_k\}$, $\{\boldsymbol{v}_k\}$ and $\{\phi_k^*\}$.
3. Set γ_k first, and deduce the rest directly. If the deduced result conforms to the definitions in the lemma, then the lemma is proved.

$$\phi_k(\boldsymbol{x}) \equiv \phi_k^* + \frac{\gamma_k}{2}\|\boldsymbol{x} - \boldsymbol{v}_k\|^2, \tag{8}$$
$$\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\mu,$$

Estimate Sequence

$$\phi_k(\boldsymbol{x}) \equiv \phi_k^* + \frac{\gamma_k}{2} \|\boldsymbol{x} - \boldsymbol{v}_k\|^2 . \tag{8}$$

Proof. For cannonical form: (Induction) Note that $\nabla^2 \phi_0(\boldsymbol{x}) = \gamma_0 I_n$. Let us show that $k \geq 0$, $\nabla^2 \phi_k(\boldsymbol{x}) = \gamma_k I_n$ for all $k \geq 0$. Indeed, if it is true for some k, then

$$\underbrace{\nabla^2 \phi_{k+1}(\boldsymbol{x}) = (1 - \alpha_k)\nabla^2 \phi_k(\boldsymbol{x}) + \alpha_k \mu I_n}_{\text{by (7)}} = ((1 - \alpha_k)\gamma_k + \alpha_k \mu) I_n \equiv \gamma_{k+1} I_n.$$

This justifies the cannonical form (7) of the function $\phi_k(\boldsymbol{x})$.

Estimate Sequence

Proof. (Continued.) For $\boldsymbol{v}_{k+1}$ (which is the optimal value of ϕ_{k+1}): Further,

$$\begin{aligned}\phi_{k+1}(\boldsymbol{x}) = (1-\alpha_k)\left(\phi_k^* + \frac{\gamma_k}{2}\|\boldsymbol{x}-\boldsymbol{v}_k\|^2\right) \\ + \alpha_k\left[f(\boldsymbol{y}_k) + \langle\nabla f(\boldsymbol{y}_k),\ \boldsymbol{x}-\boldsymbol{y}_k\rangle + \frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{y}_k\|^2\right].\end{aligned} \tag{9}$$

Therefore the equation $\nabla\phi_{k+1}(\boldsymbol{x}) = 0$, which is the first order optimality conditon for the function $\phi_{k+1}(\boldsymbol{x})$, is as follows:

$$(1-\alpha_k)\gamma_k(\boldsymbol{x}-\boldsymbol{v}_k) + \alpha_k\nabla f(\boldsymbol{y}_k) + \alpha_k\mu(\boldsymbol{x}-\boldsymbol{y}_k) = 0.$$

From this equation, we get a closed form expression for the point $\boldsymbol{v}_{k+1}$, the minimum of the function $\phi_{k+1}(\boldsymbol{x})$. This arrives at $\boldsymbol{v}_{k+1} = \frac{1}{\gamma_{k+1}}\left[(1-\alpha_k)\gamma_k\boldsymbol{v}_k + \alpha_k\mu\boldsymbol{y}_k - \alpha_k\nabla f(\boldsymbol{y}_k)\right]$.

Estimate Sequence

Proof. (Continued.) For ϕ_{k+1}^* : In view of the recurrence (7) for the sequence $\{\phi_k(\boldsymbol{x})\}$, we have

$$\begin{aligned}\phi_{k+1}^* + \frac{\gamma_{k+1}}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_{k+1}\|^2 &= \phi_{k+1}(\boldsymbol{y}_k) \quad \text{(Substitude } \boldsymbol{y}_k\text{)} \\ &= (1-\alpha_k)\underbrace{(\phi_k^* + \frac{\gamma_k}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|^2)}_{\phi_k(\boldsymbol{y}_k)} + \alpha_k f(\boldsymbol{y}_k) \\ &= \boxed{(1-\alpha_k)\frac{\gamma_k}{2}}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|^2 + (1-\alpha_k)\phi_k^* + \alpha_k f(\boldsymbol{y}_k).\end{aligned}$$

Remark.

$$\begin{aligned}\phi_{k+1}(x) &= (1-\alpha_k)\phi_k(x) \\ &\quad + \alpha_k[f(y_k) + \langle \nabla f(y_k),\, x - y_k\rangle + \frac{\mu}{2}\|x - y_k\|^2],\end{aligned}$$

Estimate Sequence

Proof. (Continued.) By the recursive relation for $\boldsymbol{v}_{k+1}$, we have (两边都减去 $\boldsymbol{y}_k$),

$$\boldsymbol{v}_{k+1} - \boldsymbol{y}_k = \frac{1}{\gamma_{k+1}} \left[(1-\alpha_k)\gamma_k(\boldsymbol{v}_k - \boldsymbol{y}_k) - \alpha_k \nabla f(\boldsymbol{y}_k) \right].$$

Remark.

$$\begin{aligned} \boldsymbol{v}_{k+1} - \boldsymbol{y}_k &= \frac{1}{\gamma_{k+1}} \left[(1-\alpha_k)\gamma_k \boldsymbol{v}_k + \alpha_k \mu \boldsymbol{y}_k - \alpha_k \nabla f(\boldsymbol{y}_k) - \gamma_{k+1} \boldsymbol{y}_k \right] \\ &= \frac{1}{\gamma_{k+1}} \left[(1-\alpha_k)\gamma_k(\boldsymbol{v}_k - \boldsymbol{y}_k) - \alpha_k \nabla f(\boldsymbol{y}_k) \right]. \end{aligned}$$

The last step is obtained by expanding $\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k \mu$.

Estimate Sequence

Proof. (Continued.) Therefore,

$$\begin{aligned}\frac{\gamma_{k+1}}{2}\|\boldsymbol{v}_{k+1}-\boldsymbol{y}_k\|^2 = &\boxed{\frac{1}{2\gamma_{k+1}}[(1-\alpha_k)^2\gamma_k^2}\|\boldsymbol{v}_k-\boldsymbol{y}_k\|^2\\ &-2\alpha_k(1-\alpha_k)\gamma_k\langle\nabla f(\boldsymbol{y}_k),\,\boldsymbol{v}_k-\boldsymbol{y}_k\rangle\\ &+\alpha_k^2\|\nabla f(\boldsymbol{y}_k)\|^2].\end{aligned}$$

$$\boldsymbol{v}_{k+1}-\boldsymbol{y}_k=\frac{1}{\gamma_{k+1}}\left[(1-\alpha_k)\gamma_k(\boldsymbol{v}_k-\boldsymbol{y}_k)-\alpha_k\nabla f(\boldsymbol{y}_k)\right].$$

Estimate Sequence

Proof. (Continued.) It remains to substitute this relation into (9), taking into account that the multiplicative factor for the term $\|\boldsymbol{y}_k - \boldsymbol{v}_{k+1}\|^2$ in the resulting expression is as follows:

$$\boxed{(1-\alpha_k)\frac{\gamma_k}{2}} - \boxed{\frac{1}{2\gamma_{k+1}}(1-\alpha_k)^2\gamma_k^2} = (1-\alpha_k)\frac{\gamma_k}{2}\left(1 - \frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\right)$$
$$= (1-\alpha_k)\frac{\gamma_k}{2}\cdot\frac{\alpha_k\mu}{\gamma_{k+1}}.$$

□

$$\boxed{\phi_{k+1}^* + \frac{\gamma_{k+1}}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_{k+1}\|^2 = (1-\alpha_k)(\phi_k^* + \frac{\gamma_k}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|^2) + \alpha_k f(\boldsymbol{y}_k).}$$

Constructing x_k

(Induction) Indeed, assume that we already have x_k: $\phi_k^* \geq f(\boldsymbol{x}_k)$. Let's check how to make $\phi_{k+1}^* \geq f(\boldsymbol{x}_{k+1})$. Then, in the view of Lemma 9,

$$\phi_{k+1}^* \geq (1-\alpha_k) f(\boldsymbol{x}_k) + \alpha_k f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}} \|\nabla f(\boldsymbol{y}_k)\|^2 + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}} \langle \nabla f(\boldsymbol{y}_k),\, \boldsymbol{v}_k - \boldsymbol{y}_k \rangle.$$

Lemma 9: $$\phi_{k+1}^* = (1-\alpha_k) \underbrace{\phi_k^*}_{\phi_k^* \geq f(\boldsymbol{x}_k)} + \alpha_k f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}} \|\nabla f(\boldsymbol{y}_k)\|^2 + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}} \left(\frac{\mu}{2} \|\boldsymbol{y}_k - \boldsymbol{v}_k\|^2 + \langle \nabla f(\boldsymbol{y}_k),\, \boldsymbol{v}_k - \boldsymbol{y}_k \rangle \right).$$

Constructing x_k

Since $f(\boldsymbol{x}_k) \geq f(\boldsymbol{y}_k) + \langle \nabla f(\boldsymbol{y}_k),\ \boldsymbol{x}_k - \boldsymbol{y}_k \rangle$, we get the following estimate:

$$\begin{aligned}\phi^*_{k+1} &\geq f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}} \|\nabla f(\boldsymbol{y}_k)\|^2 \\ &\quad + (1-\alpha_k)\langle \nabla f(\boldsymbol{y}_k),\ \frac{\alpha_k \gamma_k}{\gamma_{k+1}}(\boldsymbol{v}_k - \boldsymbol{y}_k) + \boldsymbol{x}_k - \boldsymbol{y}_k \rangle.\end{aligned}$$

$$\begin{aligned}\phi^*_{k+1} &\geq (1-\alpha_k) f(\boldsymbol{x}_k) + \alpha_k f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}} \|\nabla f(\boldsymbol{y}_k)\|^2 \\ &\quad + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}} \langle \nabla f(\boldsymbol{y}_k),\ \boldsymbol{v}_k - \boldsymbol{y}_k \rangle.\end{aligned}$$

Constructing x_k

Let us look at this inequality. We want to have $\phi_{k+1}^* \geq f(\boldsymbol{x}_{k+1})$. Recall that we can ensure the inequality

$$f(\boldsymbol{y}_k) - \frac{1}{2L} \|\nabla f(\boldsymbol{y}_k)\|^2 \geq f(\boldsymbol{x}_{k+1}),$$

in many different ways. The simplest one is just to take the gradient step

$$\boxed{\boldsymbol{x}_{k+1} = \boldsymbol{y}_k - h_k \nabla f(\boldsymbol{y}_k)},$$

with $h_k = \frac{1}{L}$. (此时，我们还有 y_k 和 α_k 两个自由度可以调)

$$0 \leq f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}),\, \boldsymbol{y} - \boldsymbol{x} \rangle \leq \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2 .$$

Constructing x_k

adjust $\boldsymbol{y}_k$ and α_k such that

$$\phi_{k+1}^* \geq f(\boldsymbol{y}_k) - \underbrace{\frac{\alpha_k^2}{2\gamma_{k+1}}}_{=1/2L} \|\nabla f(\boldsymbol{y}_k)\|^2$$
$$+ (1-\alpha_k)\langle \nabla f(\boldsymbol{y}_k), \underbrace{\frac{\alpha_k \gamma_k}{\gamma_{k+1}}(\boldsymbol{v}_k - \boldsymbol{y}_k) + \boldsymbol{x}_k - \boldsymbol{y}_k}_{=0} \rangle.$$

Constructing x_k

Let us define α_k as a positive root of the quadratic equation

$$\boxed{L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k\mu \quad (= \gamma_{k+1})}.$$

Then $\frac{\alpha_k^2}{2\gamma_{k+1}} = \frac{1}{2L}$, and we can replace the previous inequality by the following one:

$$\phi_{k+1}^* \geq f(\boldsymbol{x}_{k+1}) + (1-\alpha_k)\langle \nabla f(\boldsymbol{y}_k), \underbrace{\frac{\alpha_k\gamma_k}{\gamma_{k+1}}(\boldsymbol{v}_k - \boldsymbol{y}_k) + \boldsymbol{x}_k - \boldsymbol{y}_k}_{=0}\rangle.$$

Constructing x_k

Let us now use our freedom in the choice of y_k. It can be found from the equation:

$$\frac{\alpha_k \gamma_k}{\gamma_{k+1}}(\boldsymbol{v}_k - \boldsymbol{y}_k) + \boldsymbol{x}_k - \boldsymbol{y}_k = 0$$

This is

$$\boxed{\boldsymbol{y}_k = \frac{\alpha_k \gamma_k \boldsymbol{v}_k + \gamma_{k+1} \boldsymbol{x}_k}{\gamma_k + \alpha_k \mu}},$$

and we come to the following methods, which are often addressed as accelerated gradient method.

Optimal Scheme

General Scheme of Optimal Method

0. Choose the point $\boldsymbol{x}_0 \in \mathbb{R}^n$, some $\gamma_0 > 0$, and set $\boldsymbol{v}_0 = \boldsymbol{x}_0$.
1. k-th iteration ($k \geq 0$).
   (a) Compute $\alpha_k \in (0, 1)$ from the equation $L\alpha_k^2 = (1 - \alpha_k)\gamma_k + \alpha_k\mu$. Set $\gamma_{k+1} = (1 - \alpha_k)\gamma_k + \alpha_k\mu$.
   (b) Choose

$$\boldsymbol{y}_k = \frac{\alpha_k\gamma_k\boldsymbol{v}_k + \gamma_{k+1}\boldsymbol{x}_k}{\gamma_k + \alpha_k\mu}. \tag{10}$$

   Compute $f(\boldsymbol{y}_k)$ and $\nabla f(\boldsymbol{y}_k)$.
   (c) Find $\boldsymbol{x}_{k+1}$ such that $f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{y}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{y}_k)\|^2$
   (d) Set $\boldsymbol{v}_{k+1} = \frac{(1-\alpha_k)\gamma_k\boldsymbol{v}_k + \alpha_k\mu\boldsymbol{y}_k - \alpha_k\nabla f(\boldsymbol{y}_k)}{\gamma_{k+1}}$.

Optimal Scheme

Note that in Step (c) of this scheme we can choose an arbitrary $\boldsymbol{x}_{k+1}$ satisfying the inequality

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{y}_k) - \frac{\omega}{2} \|\nabla f(\boldsymbol{y}_k)\|^2,$$

with some $\omega > 0$. Then the constant $\frac{1}{\omega}$ replaces L in the equation of step (a).

Part II
Theoretically Analysis for AGD

Analysis of Optimal Scheme

Theorem 10 (Theorem 2.2.1)

Scheme (10) generates a sequence of points $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ such that

$$f(\boldsymbol{x}_k) - f^* \leq \lambda_k [f(\boldsymbol{x}_0) - f^* + \frac{\gamma_0}{2} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2],$$

where $\lambda_0 = 1$ and $\lambda_k = \prod_{i=0}^{k-1}(1 - \alpha_i)$.

Proof. Indeed, let us choose $\phi_0(x) = f(\boldsymbol{x}_0) + \frac{\gamma_0}{2} \|\boldsymbol{x} - \boldsymbol{x}_0\|^2$. Then $f(\boldsymbol{x}_0) = \phi_0^*$ and we get $f(\boldsymbol{x}_k) \leq \phi_k^*$ by the rules of the scheme. It remains to use Lemma 7.
(From Lemma 7, we have $f(\boldsymbol{x}_k) - f^* \leq \lambda_k[\phi_0(\boldsymbol{x}^*) - f^*] \to 0$. 代入 $\phi_0(\boldsymbol{x}^*)$ 即可) 。 □

Analysis of Optimal Scheme

Lemma 11 (Lemma 2.2.4)

If in the method (10) we choose $\gamma_0 \geq \mu$, then

$$\lambda_k \leq \min\left\{\left(1 - \sqrt{\frac{\mu}{L}}\right)^k, \frac{4L}{(2\sqrt{L} + k\sqrt{\gamma_0})^2}\right\}. \tag{11}$$

Proof. Indeed, if $\gamma_k \geq \mu$, then $\gamma_{k+1} = L\alpha_k^2 = (1 - \alpha_k)\gamma_k + \alpha_k\mu \geq \mu$. Since $\gamma_0 \geq \mu$, we conclude that this inequality is valid for all γ_k. Hence $\alpha_k \geq \sqrt{\frac{\mu}{L}}$ and we have proved the first inequality in (11).

Remark. $\gamma_{k+1} = (1 - \alpha_k)\gamma_k + \alpha_k\mu > (1 - \alpha_k)\mu + \alpha_k\mu = \mu$。

Analysis of Optimal Scheme

Since

$$\lambda_k = \prod_{i=0}^{k-1}(1-\alpha_i),$$

and for every α_i we have

$$\alpha_i \geq \sqrt{\frac{\mu}{L}}.$$

Thus, we obtain

$$\lambda_k \leq \left(1-\sqrt{\frac{\mu}{L}}\right)^k.$$

Analysis of Optimal Scheme

Proof. (Continued.) For the second inequality, let us first prove that $\gamma_k \geq \gamma_0 \lambda_k$. Indeed, since $\gamma_0 = \gamma_0 \lambda_0$, we can use induction:

$$\gamma_{k+1} \geq (1 - \alpha_k)\gamma_k \geq (1 - \alpha_k)\gamma_0 \lambda_k = \gamma_0 \lambda_{k+1}.$$

Therefore $L\alpha_k^2 = \gamma_{k+1} \geq \gamma_0 \lambda_{k+1}$. That is

$$\sqrt{\lambda_{k+1}} \leq \alpha_k \sqrt{\frac{L}{\gamma_0}}. \tag{12}$$

$$\gamma_{k+1} = (1 - \alpha_k)\gamma_k + \alpha_k \mu.$$

Analysis of Optimal Scheme

Proof. (Continued.) Denote $b_k = \frac{1}{\sqrt{\lambda_k}}$. Since $\{\lambda_k\}$ is a decreasing sequence, we have

$$\begin{aligned} b_{k+1} - b_k &= \frac{\sqrt{\lambda_k} - \sqrt{\lambda_{k+1}}}{\sqrt{\lambda_k \lambda_{k+1}}} = \frac{\lambda_k - \lambda_{k+1}}{\sqrt{\lambda_k \lambda_{k+1}}(\sqrt{\lambda_k} + \sqrt{\lambda_{k+1}})} \\ &\geq \frac{\lambda_k - \lambda_{k+1}}{2\lambda_k \sqrt{\lambda_{k+1}}} = \frac{\lambda_k - (1-\alpha_k)\lambda_k}{2\lambda_k \sqrt{\lambda_{k+1}}} = \frac{\alpha_k}{2\sqrt{\lambda_{k+1}}} \\ &\geq \frac{1}{2}\sqrt{\frac{\gamma_0}{L}}. \text{(Since (12))} \end{aligned}$$

Thus, $b_k \geq 1 + \frac{k}{2}\sqrt{\frac{\gamma_0}{L}}$ and the lemma is proved. □

Remark. In Theorem 10, set $\lambda_0 = 1$, $\lambda_k = \prod_{i=0}^{k-1}(1-\alpha_i)$.

Analysis of Optimal Scheme

Theorem 12 (Theorem 2.2.2)

Let us take in (10) $\gamma_0 = L$. Then this scheme generates a sequence $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ such that

$$f(\boldsymbol{x}_k) - f^* \leq L \min\left\{\left(1 - \sqrt{\frac{\mu}{L}}\right)^k, \frac{4}{(k+2)^2}\right\} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 .$$

This means that (10) is optimal for unconstrained minimization of the functions from $\mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$, $\mu \geq 0$.

Proof. We get the above inequality using $f(\boldsymbol{x}_0) - f^* \leq \frac{L}{2} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2$ and Theorem 10 with lemma 11.

Analysis of Optimal Scheme

Proof. (Continued.) Let $\mu > 0$. From the lower complexity bounds for the class, we have

$$f(x_k) - f^* \geq \frac{\mu}{2}\left(\frac{\sqrt{Q_f}-1}{\sqrt{Q_f}+1}\right)^{2k} R^2 \geq \frac{\mu}{2}\exp\left(-\frac{4k}{\sqrt{Q_f}-1}\right) R^2,$$

where $Q_f = L/\mu$ and $R = \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|$. Therefore, the worst case bound for finding x_k satisfying $f(\boldsymbol{x}_k) - f^* \leq \epsilon$ cannot be better than

$$k \geq \frac{\sqrt{Q_f}-1}{4}\left[\ln\frac{1}{\epsilon} + \ln\frac{\mu}{2} + 2\ln R\right].$$

Remark. $\ln(1+\boldsymbol{x}) \leq \boldsymbol{x}$

Analysis of Optimal Scheme

Proof. (Continued.) For our scheme we have

$$f(\boldsymbol{x}_k) - f^* \leq LR^2 \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \leq LR^2 \exp\left(-\frac{k}{\sqrt{Q_f}}\right).$$

Therefore, we guarantee that $k \leq \sqrt{Q_f}\left[\ln \frac{1}{\epsilon} + \ln L + 2 \ln R\right]$. Thus, the main term in this estimate, $\sqrt{Q_f} \ln \frac{1}{\epsilon}$, is proportional to the lower bound. The same reasoning can be used for the class $\mathcal{S}^{1,1}_{0,L}(\mathbb{R}^n)$. □

Variant of Optimal Scheme

Let us analyze a variant of the scheme (10).

Constant Step Scheme, I

0. Choose $\boldsymbol{x}_0 \in \mathbb{R}^n$ and $\gamma_0 > 0$. Set $\boldsymbol{v}_0 = \boldsymbol{x}_0$.
1. k-th iteration ($k \geq 0$).
   (a) Compute $\alpha_k \in (0,1)$ from the equation $L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k\mu$. Set $\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\mu$.
   (b) Choose $\boldsymbol{y}_k = \frac{\alpha_k\gamma_k\boldsymbol{v}_k + \gamma_{k+1}\boldsymbol{x}_k}{\gamma_k + \alpha_k\mu}$. Compute $f(\boldsymbol{y}_k)$ and $\nabla f(\boldsymbol{y}_k)$.
   (c) Set $\boldsymbol{x}_{k+1} = \boldsymbol{y}_k - \frac{1}{L}\nabla f(\boldsymbol{y}_k)$ and

$$\boldsymbol{v}_{k+1} = \frac{1}{\gamma_{k+1}}[(1-\alpha_k)\gamma_k\boldsymbol{v}_k + \alpha_k\mu\boldsymbol{y}_k - \alpha_k\nabla f(\boldsymbol{y}_k)]. \tag{13}$$

Variant of Optimal Scheme

Let us demonstrate that this scheme can be rewritten in a simpler form. Note that

$$\boldsymbol{y}_k = \frac{1}{\gamma_k + \alpha_k \mu}(\alpha_k \boxed{\gamma_k \boldsymbol{v}_k} + \gamma_{k+1}\boldsymbol{x}_k), \tag{14}$$

$$\boldsymbol{x}_{k+1} = \boldsymbol{y}_k - \frac{1}{L}\nabla f(\boldsymbol{y}_k), \tag{15}$$

$$\boldsymbol{v}_{k+1} = \frac{1}{\gamma_{k+1}}[(1-\alpha_k)\boxed{\gamma_k \boldsymbol{v}_k} + \alpha_k \mu \boldsymbol{y}_k - \alpha_k \nabla f(\boldsymbol{y}_k)]. \tag{16}$$

Remark. Eliminate $\boldsymbol{v}_k$.

Variant of Optimal Scheme

Therefore,

$$\begin{aligned}
\boldsymbol{v}_{k+1} &= \frac{1}{\gamma_{k+1}}\left\{\frac{1-\alpha_k}{\alpha_k}\left[(\gamma_k+\alpha_k\mu)\boldsymbol{y}_k-\gamma_{k+1}\boldsymbol{x}_k\right]+\alpha_k\mu\boldsymbol{y}_k-\alpha_k\nabla f(\boldsymbol{y}_k)\right\}\\
&= \frac{1}{\gamma_{k+1}}\left\{\frac{(1-\alpha_k)\gamma_k}{\alpha_k}\boldsymbol{y}_k+\mu\boldsymbol{y}_k\right\}-\frac{1-\alpha_k}{\alpha_k}\boldsymbol{x}_k-\frac{\alpha_k}{\gamma_{k+1}}\nabla f(\boldsymbol{y}_k)\\
&= \boldsymbol{x}_k+\frac{1}{\alpha_k}(\boldsymbol{y}_k-\boldsymbol{x}_k)-\frac{1}{\alpha_k L}\nabla f(\boldsymbol{y}_k)\\
&= \boldsymbol{x}_k+\frac{1}{\alpha_k}\left[(\boldsymbol{y}_k-\boldsymbol{x}_k)-\frac{1}{L}\nabla f(\boldsymbol{y}_k)\right]=\boldsymbol{x}_k+\frac{1}{\alpha_k}(\boldsymbol{x}_{k+1}-\boldsymbol{x}_k).
\end{aligned}$$

$$\gamma_{k+1}=L\alpha_k^2=(1-\alpha_k)\gamma_k+\alpha_k\mu.\quad \boldsymbol{x}_{k+1}=\boldsymbol{y}_k-\frac{1}{L}\nabla f(\boldsymbol{y}_k).$$

Variant of Optimal Scheme

Hence,

$$\begin{aligned}\boldsymbol{y}_{k+1} &= \frac{1}{\gamma_{k+1}+\alpha_{k+1}\mu}(\alpha_{k+1}\gamma_{k+1}\boldsymbol{v}_{k+1}+\gamma_{k+2}\boldsymbol{x}_{k+1}) \\ &= \boldsymbol{x}_{k+1}+\frac{\alpha_{k+1}\gamma_{k+1}(\boldsymbol{v}_{k+1}-\boldsymbol{x}_{k+1})}{\gamma_{k+1}+\alpha_{k+1}\mu} = \boldsymbol{x}_{k+1}+\beta_k(\boldsymbol{x}_{k+1}-\boldsymbol{x}_k),\end{aligned}$$

where

$$\beta_k = \frac{\alpha_{k+1}\gamma_{k+1}(1-\alpha_k)}{\alpha_k(\gamma_{k+1}+\alpha_{k+1}\mu)}.$$

Thus, $\boldsymbol{y}_k$ is not involved with $\boldsymbol{v}_k$.

Variant of Optimal Scheme

Further eliminate γ_k: Let us do the same with γ_k. We have

$$\alpha_k^2 L = (1-\alpha_k)\gamma_k + \mu\alpha_k \equiv \gamma_{k+1}.$$

Therefore,

$$\begin{aligned}\beta_k &= \frac{\alpha_{k+1}\gamma_{k+1}(1-\alpha_k)}{\alpha_k(\gamma_{k+1}+\alpha_{k+1}\mu)} = \frac{\alpha_{k+1}\gamma_{k+1}(1-\alpha_k)}{\alpha_k(\gamma_{k+1}+\alpha_{k+1}^2 L-(1-\alpha_{k+1})\gamma_{k+1})} \\ &= \frac{\gamma_{k+1}(1-\alpha_k)}{\alpha_k(\gamma_{k+1}+\alpha_{k+1}L)} = \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2+\alpha_{k+1}}.\end{aligned}$$

Variant of Optimal Scheme

The iteration of α_{k+1} is processed below.

- Note that $\alpha_{k+1}^2 = (1-\alpha_{k+1})\alpha_k^2 + q\alpha_{k+1}$ (since $L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k\mu$ and $\gamma_{k+1} = L\alpha_k^2$), where $q = \mu/L$,
- And we also have
$$\alpha_0^2 L = (1-\alpha_0)\gamma_0 + \mu\alpha_0.$$
The latter relation means that γ_0 can be seen as a function of α_0. (Therefore, setting α_0 is equivalent to setting γ_0.)

Thus, we can completely eliminate the sequence $\{\gamma_k\}$.

Variant of Optimal Scheme

Constant Step Scheme, II

0. Choose $\boldsymbol{x}_0 \in \mathbb{R}^n$ and $\alpha_0 \in (0,1)$. Set $\boldsymbol{y}_0 = \boldsymbol{x}_0$ and $q = \frac{\mu}{L}$.
1. k-th iteration ($k \geq 0$).
   (a) Compute $f(\boldsymbol{y}_k)$ and $\nabla f(\boldsymbol{y}_k)$. Set $\boldsymbol{x}_{k+1} = \boldsymbol{y}_k - \frac{1}{L}\nabla f(\boldsymbol{y}_k)$.
   (b) Compute $\alpha_{k+1} \in (0,1)$ from equation

$$\alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2 + q\alpha_{k+1}, \tag{17}$$

   and set $\beta_k = \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2 + \alpha_{k+1}}$,

$$\boldsymbol{y}_{k+1} = \boldsymbol{x}_{k+1} + \beta_k(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k).$$

Analysis of the Variant

Theorem 13 (Theorem 2.2.3)

If in scheme (17)

$$\alpha_0 \geq \sqrt{\frac{\mu}{L}}, \tag{18}$$

then

$$f(\boldsymbol{x}_k) - f^* \leq \min\left\{\left(1 - \sqrt{\frac{\mu}{L}}\right)^k, \frac{4L}{(2\sqrt{L} + k\sqrt{\gamma_0})^2}\right\} \times \left[f(\boldsymbol{x}_0) - f^* + \frac{\gamma_0}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2\right],$$

where $\gamma_0 = \frac{\alpha_0(\alpha_0 L - \mu)}{1 - \alpha_0}$.

Analysis of the Variant

We do not need to prove this theorem since the initial scheme is not changed. We change only notation. In Theorem13 condition (18) is equivalent to $\gamma_0 \geq \mu$.

If we choose $\alpha_0 = \sqrt{\frac{\mu}{L}}$ (this corresponds to $\gamma_0 = \mu$), Scheme (17) becomes very simple. Then

$$\alpha_k = \sqrt{\frac{\mu}{L}}, \ \beta_k = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}.$$

for all $k \geq 0$.

Analysis of the Variant

Thus, we come to the following process.

Constant step scheme, III

0. Choose $\boldsymbol{y}_0 = \boldsymbol{x}_0 \in \mathbb{R}^n$.
1. k-th iteration ($k \geq 0$).

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{y}_k - \frac{1}{L}\nabla f(\boldsymbol{y}_k), \\ \boldsymbol{y}_{k+1} &= \boldsymbol{x}_{k+1} + \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k). \end{aligned} \tag{19}$$

However, note that this process does not work for $\mu = 0$. The choice $\gamma_0 = L$ (which changes corresponding value of α_0) is safer.

Analysis of the Variant

1. [Nesterov, 1983]

$$\boldsymbol{x}_{t+1} = \boldsymbol{y}_t - \frac{1}{L}\nabla f(\boldsymbol{y}_t).$$

$$\boldsymbol{y}_{t+1} = \boldsymbol{x}_{t+1} + \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t).$$

Analysis of the Variant

1. Heavy-ball method [Polyak, 1964]:

$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \alpha \boxed{\nabla f(\boldsymbol{x}_t)} + \beta(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})$. Set $\alpha = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$, $\beta = \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2$

$$\boldsymbol{x}_{t+1} = \boldsymbol{y}_t - \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}\boxed{\nabla f(\boldsymbol{y}_t)}.$$

$$\boldsymbol{y}_{t+1} = \boldsymbol{x}_{t+1} + \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2 (\boldsymbol{x}_{t+1} - \boldsymbol{x}_t).$$

For $\mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$,

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|_2 \leq \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^t \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2 .$$

Analysis of the Variant

1. [Nesterov and Nemirovsky, 1988]

$$\boldsymbol{x}_{t+1} = \boldsymbol{y}_t - \frac{1}{L}\nabla f(\boldsymbol{y}_t).$$
$$\boldsymbol{y}_{t+1} = \boldsymbol{x}_{t+1} + \frac{t}{t+3}(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t).$$

For $\mathcal{S}^{1,1}_{0,L}(\mathbb{R}^n)$,

$$f(\boldsymbol{x}_t) - f^* \leq 2L\frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{(t+1)^2}.$$

Analysis of the Variant

1. FISTA [Beck and Teboulle, 2009]:

$$\boldsymbol{x}_{t+1} = \boldsymbol{y}_t - \frac{1}{L}\nabla f(\boldsymbol{y}_t).$$
$$\boldsymbol{y}_{t+1} = \boldsymbol{x}_{t+1} + \frac{\lambda_t - 1}{\lambda_{t+1}}(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t),$$

where $\lambda_0 = 0$, and $\lambda_{t+1} = \frac{1+\sqrt{1+4\lambda_t^2}}{2}$ for all $t > 0$.

For $\mathcal{S}^{1,1}_{0,L}(\mathbb{R}^n)$,

$$f(\boldsymbol{x}_t) - f^* \leq 2L\frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{t^2}.$$

Analysis of the Variant

Lemma 14
By definition of $\boldsymbol{x}_{t+1} = \boldsymbol{y}_t - \frac{1}{L}\nabla f(\boldsymbol{y}_t)$, we have

$$f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}) \leq -\frac{L}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{y}_t\|_2^2 - L\langle \boldsymbol{x}_{t+1} - \boldsymbol{y}_t,\ \boldsymbol{y}_t - \boldsymbol{x}\rangle.$$

Proof.
We have the following fact:

$$f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}) = f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{y}_t) + f(\boldsymbol{y}_t) - f(\boldsymbol{x}).$$

Remark. Please refer to the course website of Niao He:
https://github.com/niaohe/Big-Data-Optimization-Course/blob/main/lecture_scribe/IE598-lecture9-gradient-descent-and-acceleration.pdf

Analysis of the Variant

$$f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}) \leq -\frac{L}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{y}_t\|_2^2 - L\langle \boldsymbol{x}_{t+1} - \boldsymbol{y}_t,\ \boldsymbol{y}_t - \boldsymbol{x}\rangle.$$

Proof. (Continued.) Using the facts that $f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{y}_t) \leq -\frac{1}{2L}\|\nabla f(\boldsymbol{y}_t)\|_2^2$ (pp. 26 of nesterov book) and $f(\boldsymbol{y}_t) - f(\boldsymbol{x}) \leq \langle \nabla f(\boldsymbol{y}_t),\ \boldsymbol{y}_t - \boldsymbol{x}\rangle$, we have

$$f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}) \leq -\frac{1}{2L}\|\nabla f(\boldsymbol{y}_t)\|_2^2 + \langle \nabla f(\boldsymbol{y}_t),\ \boldsymbol{y}_t - \boldsymbol{x}\rangle.$$

Hence, by using definition of $\boldsymbol{x}_{t+1} = \boldsymbol{y}_t - \frac{1}{L}\nabla f(\boldsymbol{y}_t)$, we have the result. □

Analysis of the Variant

Theorem 15 (FISTA)
For $\mathcal{S}^{1,1}_{0,L}(\mathbb{R}^n)$,

$$f(\boldsymbol{x}_t) - f^* \leq 2L\frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{t^2}.$$

Proof. According to lemma 14, and let $\boldsymbol{x} = \boldsymbol{x}_t$ and $\boldsymbol{x} = \boldsymbol{x}^*$, respectively. We have

$$[f(\boldsymbol{x}_{t+1}) - f(\boxed{\boldsymbol{x}_t})](\lambda_t - 1) \leq \left[-\frac{L}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{y}_t\|_2^2 + \langle \boldsymbol{x}_{t+1} - \boldsymbol{y}_t,\ \boldsymbol{y}_t - \boxed{\boldsymbol{x}_t}\rangle\right](\lambda_t - 1),$$

$$[f(\boldsymbol{x}_{t+1}) - f(\boxed{\boldsymbol{x}^*})] \leq \left[-\frac{L}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{y}_t\|_2^2 + \langle \boldsymbol{x}_{t+1} - \boldsymbol{y}_t,\ \boldsymbol{y}_t - \boxed{\boldsymbol{x}^*}\rangle\right].$$

Analysis of the Variant

Proof. (Continued.) Adding above two equations, and let $\epsilon_t = f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)$, we arrive at

$$\begin{aligned}\lambda_t f(\boldsymbol{x}_{t+1}) - \lambda_t f(\boldsymbol{x}_t) + f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) &= \lambda_t \epsilon_{t+1} - (\lambda_t - 1)\epsilon_t \\ &\leq -\frac{\lambda_t L}{2} \|\boldsymbol{x}_{t+1} - \boldsymbol{y}_t\|_2^2 \\ &\quad - L\langle(\boldsymbol{x}_{t+1} - \boldsymbol{y}_t),\ \lambda_t \boldsymbol{y}_t - (\lambda_t - 1)\boldsymbol{x}_t - \boldsymbol{x}^*\rangle.\end{aligned}$$

Multiplying both sides by λ_t:

$$\lambda_t^2 \epsilon_{t+1} - \lambda_t(\lambda_t - 1)\epsilon_t \leq -\frac{L}{2}\left[\lambda_t^2 \|\boldsymbol{x}_{t+1} - \boldsymbol{y}_t\|_2^2 + 2\lambda_t\langle(\boldsymbol{x}_{t+1} - \boldsymbol{y}_t),\ \lambda_t \boldsymbol{y}_t - (\lambda_t - 1)\boldsymbol{x}_t - \boldsymbol{x}^*\rangle\right].$$

Analysis of the Variant

Proof. (Continued.) Since $\lambda_0 = 0$, and $\lambda_{t+1} = \frac{1+\sqrt{1+4\lambda_t^2}}{2}$ for all $t \geq 0$, we know the follwing is true,

$$\lambda_{t+1}^2 - \lambda_{t+1} = \lambda_t^2.$$

Thus, rearragning terms, we have

$$\begin{aligned}\lambda_t^2 \epsilon_{t+1} - \lambda_{t-1}^2 \epsilon_t \leq -\frac{L}{2}\big[&\|\lambda_t \boldsymbol{x}_{t+1} - (\lambda_t - 1)\boldsymbol{x}_t - \boldsymbol{x}^*\|_2^2 \\ &- \|\lambda_t \boldsymbol{y}_t - (\lambda_t - 1)\boldsymbol{x}_t - \boldsymbol{x}^*\|_2^2 \big].\end{aligned}$$

Invoking the definition of $\boldsymbol{y}_{t+1}$, one can show that

$$\lambda_t \boldsymbol{x}_{t+1} - (\lambda_t - 1)\boldsymbol{x}_t - \boldsymbol{x}^* = \lambda_{t+1}\boldsymbol{y}_{t+1} - (\lambda_{t+1} - 1)\boldsymbol{x}_{t+1} - \boldsymbol{x}^*.$$

Analysis of the Variant

Proof. (Continued.) Hence, define $\boldsymbol{u}_0 = \boldsymbol{x}_0 - \boldsymbol{x}^*$, $u_t = \lambda_t \boldsymbol{y}_t - (\lambda_t - 1)\boldsymbol{x}_t - \boldsymbol{x}^*$, $\forall t > 1$, the abve equation simplifies to

$$\lambda_t^2 \epsilon_{t+1} - \lambda_{t-1}^2 \epsilon_t \leq -\frac{L}{2}\left[\|\boldsymbol{u}_{t+1}\|_2^2 - \|\boldsymbol{u}_t\|_2^2\right].$$

By induction, we have

$$\lambda_{t-1}^2 \epsilon_t \leq \frac{L}{2}(\|\boldsymbol{u}_0\|_2^2 - \|\boldsymbol{u}_t\|_2^2) \leq \frac{L}{2}\|\boldsymbol{u}_0\|_2^2.$$

Since $\lambda_{t-1} > t/2$, $\forall t \geq 1$, we have

$$\epsilon_t \leq \frac{2L\|\boldsymbol{u}_0\|_2^2}{t^2}.$$

References I

Yu Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. In *Sov. Math. Dokl*, volume 27, 1983.

Yurii Nesterov and A Nemirovsky. A general approach to polynomial-time algorithms design for convex programming. *Report, Central Economical and Mathematical Institute, USSR Academy of Sciences, Moscow*, 1988.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Yu Nesterov. Gradient methods for minimizing composite objective function. Technical report, Université catholique de Louvain, Center for Operations Research, 2007.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

References II

Paul Tseng and Sangwoon Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1-2):387–423, 2009.

Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1): 57–95, 2016.

Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *proceedings of the National Academy of Sciences*, 113 (47):E7351–E7358, 2016.

References III

Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 2015.

Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Mathematical programming*, 171(1-2):167–215, 2018.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Thank You!

Email:qianhui@zju.edu.cn