On the convergence of gradient descent

Charles Deledalle

November 13, 2018

This document provides two main results of convergence for gradient descent.

1 Convergence

We want to prove the following theorem.

Theorem 1. *Let $F : \mathbb{R}^N \to R$ be differentiable with L Lipschitz gradient:*

$$\|\nabla x - \nabla y\| \leqslant L\|x - y\|, \quad \forall x \in \mathbb{R}^N \quad \textit{and} \quad \forall y \in \mathbb{R}^N \tag{1}$$

and be lower bounded

$$\inf_{x \in \mathbb{R}^N} F(x) = C > -\infty \ . \tag{2}$$

Then, provided $0 < \gamma < \frac{2}{L}$, the gradient descent sequence defined as:

$$x^{t+1} = x^t - \gamma \nabla F(x^t) \tag{3}$$

converges to a stationary point:

$$\lim_{t \to \infty} \nabla F(x^t) = 0 \ . \tag{4}$$

Remark 1. *F does not need to be convex. Nevertheless, to prove the theorem, we will need to prove that $G(x) = \frac{L}{2}\|x\|^2 - F(x)$ is convex. We will need two intermediate lemmas for this.*

1.1 Non-decreasing derivative ⇒ Convexity (1d)

Lemma 1. *Let $g : \mathbb{R} \to \mathbb{R}$ be differentiable with non-decreasing derivative, i.e.:*

$$g'(x) \geqslant g'(y), \quad \forall x \geqslant y \in \mathbb{R} \tag{5}$$

then g is convex

$$g(\lambda x_1 + (1-\lambda)x_5) \leqslant \lambda g(x_1) + (1-\lambda) g(x_5), \quad \forall x_1 \in \mathbb{R}, \forall x_5 \in \mathbb{R} \quad \textit{and} \quad \lambda \in [0,1] \ . \tag{6}$$

Proof. Let $\lambda \in [0,1]$, $x_1 \in \mathbb{R}$ and $x_5 \in \mathbb{R}$.

- If $x_5 = x_1$ or $\lambda = 0$ or $\lambda = 1$ the result is trivial: (6) holds true.
- Consider $x_5 > x_1$ and $0 < \lambda < 1$. Let $x_3 = \lambda x_1 + (1-\lambda)x_5$. We have $x_1 \leqslant x_3 \leqslant x_5$. The mean value theorem clains that there exist x_2, x_4 such that $x_1 \leqslant x_2 \leqslant x_3 \leqslant x_4 \leqslant x_5$ and

$$\frac{g(x_3) - g(x_1)}{x_3 - x_1} = g'(x_2) \quad \text{and} \quad \frac{g(x_5) - g(x_3)}{x_5 - x_3} = g'(x_4) \tag{7}$$

Since $x_2 \leqslant x_4$, $g'(x_2) \leqslant g'(x_4)$ by assumption, and then

$$\frac{g(x_3)-g(x_1)}{x_3-x_1} \leqslant \frac{g(x_5)-g(x_3)}{x_5-x_3} \tag{8}$$
$$\Rightarrow \frac{g(x_3)-g(x_1)}{(1-\lambda)(x_5-x_1)} \leqslant \frac{g(x_5)-g(x_3)}{\lambda(x_5-x_1)} \qquad \text{(by definition of } x_3\text{)} \tag{9}$$
$$\Rightarrow \frac{g(x_3)-g(x_1)}{(1-\lambda)} \leqslant \frac{g(x_5)-g(x_3)}{\lambda} \qquad \text{(since } x_5 > x_1\text{)} \tag{10}$$
$$\Rightarrow \lambda g(x_3) - \lambda g(x_1) \leqslant (1-\lambda)g(x_5) - (1-\lambda)g(x_3) \qquad \text{(since } 0<\lambda<1\text{)} \tag{11}$$
$$\Rightarrow g(x_3) \leqslant \lambda g(x_1) + (1-\lambda)g(x_5) \tag{12}$$

Then (6) holds true.

- If $x_1 < x_5$, the exact same reasoning applies.

Then g is convex. □

Remark 2. *The reciprocal holds true.*

1.2 Monotone gradient ⇒ Convexity (Nd)

Lemma 2. *Let $G : \mathbb{R}^N \to \mathbb{R}$ be differentiable with monotone gradient, i.e.*

$$\langle \nabla G(x) - \nabla G(y),\, x-y\rangle \geqslant 0, \quad \forall x \in \mathbb{R}^N \quad \textit{and} \quad \forall y \in \mathbb{R}^N \tag{13}$$

then G is convex, i.e.:

$$G(\lambda x + (1-\lambda)y) \leqslant \lambda G(x) + (1-\lambda)G(y), \quad \forall x \in \mathbb{R}^N, \forall y \in \mathbb{R}^N \quad \textit{and} \quad \lambda \in [0,1]\ . \tag{14}$$

Proof. Let $x \in \mathbb{R}^N$ and $d \in \mathbb{R}^N$. Let $h : \mathbb{R} \to \mathbb{R}$ be defined as

$$h(t) = G(x+td), \quad \text{for all } t \in \mathbb{R} \tag{15}$$

We have

$$h'(t) = \frac{\partial G(x+td)}{\partial t} = \frac{\partial G(x+td)}{\partial x+td}\frac{\partial x+td}{\partial t} = \left[\nabla G(x+td)\right]^T d = \langle \nabla G(x+td),\, d\rangle \tag{16}$$

Let $t_1 > t_2$. By assumption

$$\langle \nabla G(x+t_1 d) - \nabla G(x+t_2 d),\, (t_1-t_2)d\rangle \geqslant 0 \tag{17}$$
$$\Rightarrow \quad \langle \nabla G(x+t_1 d) - \nabla G(x+t_2 d),\, d\rangle \geqslant 0 \qquad \text{(since } t_1 \geqslant t_2\text{)} \tag{18}$$
$$\Rightarrow \quad h'(t_1) \geqslant h'(t_2) \tag{19}$$

Then, using Lemma 1, h is convex. Then for all $t_1 \in \mathbb{R}$, $t_2 \in \mathbb{R}$ and $\lambda \in [0,1]$

$$G(x + (\lambda t_1 + (1-\lambda)t_2)d) \leqslant \lambda G(x+t_1 d) + (1-\lambda)G(d+t_2 d) \tag{20}$$
$$\Rightarrow \quad G(\lambda(x+t_1 d) + (1-\lambda)(x+t_2 d)) \leqslant \lambda G(x+t_1 d) + (1-\lambda)G(x+t_2 d) \tag{21}$$

In particular it holds for $t_1 = 0$, $t_2 = 1$ and $d = y - x$, which concludes the proof. □

Remark 3. *The reciprocal holds true.*

1.3 Convexity ⇒ Lower bounded by linear functions

Lemma 3. *Let $G : \mathbb{R}^N \to \mathbb{R}$ be differentiable and convex then:*

$$G(y) \geqslant \underbrace{G(x) + \langle \nabla G(x), y - x \rangle}_{\text{1st order Taylor expansion}}, \quad \forall x \in \mathbb{R}^N \quad \text{and} \quad y \in \mathbb{R}^N \ . \tag{22}$$

Proof. Let $x \in \mathbb{R}^N$ and $d \in \mathbb{R}^N$. By definition of convexity, for all $t \in (0, 1]$

$$G(x + td) \leqslant (1 - t)G(x) + tG(x + d) \tag{23}$$

$$\Rightarrow \quad G(x + td) - G(x) \leqslant -tG(x) + tG(x + d) \tag{24}$$

$$\Rightarrow \quad \frac{G(x + td) - G(x)}{t} \leqslant G(x + d) - G(x) \tag{25}$$

Since it is true for all $t \in (0, 1]$, it is also true for $t \to 0$ since G is differentiable and then continuous on $\mathbb{R}^N$. Remark that by definition

$$\lim_{t \to 0} \frac{G(x + td) - G(x)}{t} = \nabla G(x)^T d \tag{26}$$

Consider $d = y - x$ and then

$$\nabla G(x)^T d \leqslant G(x + d) - G(x) \tag{27}$$

$$\Rightarrow \quad \nabla G(x)^T (y - x) \leqslant G(y) - G(x) \tag{28}$$

$$\Rightarrow \quad G(x) + \nabla G(x)^T (y - x) \leqslant G(y) \ . \tag{29}$$

□

1.4 Lipschitz gradient ⇒ Upper-bounded by quadratic functions

Lemma 4. *Let $F : \mathbb{R}^N \to \mathbb{R}$ be differentiable with L Lipschitz gradient:*

$$\|\nabla F(x) - \nabla F(y)\| \leqslant L\|x - y\| \ , \quad \forall x, y \ , \tag{30}$$

then

$$F(y) \leqslant \underbrace{F(x) + \langle \nabla F(x), y - x \rangle}_{\text{1st order Taylor expansion}} + \underbrace{\frac{L}{2}\|y - x\|^2}_{\text{Residual bound}} \ . \tag{31}$$

Proof. Let $G(x) = \frac{L}{2}\|x\|^2 - F(x)$, for all x. Remark that

$$\nabla G(x) = Lx - \nabla F(x) \ . \tag{32}$$

By assumption, we have for any x, y:

$$\|\nabla F(x) - \nabla F(y)\| \leqslant L\|x - y\| \tag{33}$$

$$\Rightarrow \quad \|\nabla F(x) - \nabla F(y)\|\|x - y\| \leqslant L\|x - y\|^2 \qquad \text{(Multiply both sides by } \|x - y\|\text{)} \tag{34}$$

$$\Rightarrow \quad \langle \nabla F(x) - \nabla F(y), x - y \rangle \leqslant L\|x - y\|^2 \qquad \text{(Cauchy-Schwartz inequality)} \tag{35}$$

$$\Rightarrow \quad \langle \nabla F(x) - \nabla F(y), x - y \rangle \leqslant L\langle x - y, x - y \rangle \tag{36}$$

$$\Rightarrow \quad \langle -[Lx - \nabla F(x)] + [Ly - \nabla F(y)], x - y \rangle \leqslant 0 \tag{37}$$

$$\Rightarrow \quad \langle [Lx - \nabla F(x)] - [Ly - \nabla F(y)], x - y \rangle \geqslant 0 \tag{38}$$

$$\Rightarrow \quad \langle \nabla G(x) - \nabla G(y), x - y \rangle \geqslant 0 \tag{39}$$

Since the last inequality hold for any x, y, using Lemma 2, it means that G is convex. Next, based on Lemma 3, we get

$$G(y) \geqslant G(x) + \nabla G(x)^T (y - x) \tag{40}$$

$$\Rightarrow \quad \frac{L}{2}\|y\|^2 - F(y) \geqslant \frac{L}{2}\|x\|^2 - F(x) + (Lx - \nabla F(x))^T(y-x) \tag{41}$$

$$\Rightarrow \quad \frac{L}{2}\|y\|^2 - F(y) \geqslant \frac{L}{2}\|x\|^2 - F(x) + L\langle x,\, y-x\rangle - \langle \nabla F(x),\, y-x\rangle \tag{42}$$

$$\Rightarrow \quad \frac{L}{2}\|y\|^2 + \frac{L}{2}\|x\|^2 - L\langle x,\, y\rangle \geqslant F(y) - F(x) - \langle \nabla F(x),\, y-x\rangle \tag{43}$$

$$\Rightarrow \quad \frac{L}{2}\|x-y\|^2 \geqslant F(y) - F(x) - \langle \nabla F(x),\, y-x\rangle \tag{44}$$

□

Remark 4. *If fact there is an equivalence between* (30) *and* (31)*, see for instance* ***https://xingyuzhou.org/blog/notes/Lipschitz-gradient****.*

1.5 Proof of Theorem 1

Proof. Since F is differentiable with L Lipschitz gradient, based on Lemma 4, we have

$$F(x^{t+1}) \leqslant F(x^t) + \langle \nabla F(x^t),\, x^{t+1} - x^t\rangle + \frac{L}{2}\|x^{t+1} - x^t\|^2 \tag{45}$$

By definition of gradient descent

$$x^{t+1} - x^t = \gamma \nabla F(x^t) \tag{46}$$

Then

$$F(x^{t+1}) \leqslant F(x^t) - \langle \nabla F(x^t),\, \gamma \nabla F(x^t)\rangle + \frac{L}{2}\|\gamma \nabla F(x^t)\|^2 \tag{47}$$

$$\leqslant F(x^t) - \gamma\|\nabla F(x^t)\|^2 + \frac{L\gamma^2}{2}\|\nabla F(x^t)\|^2 \tag{48}$$

$$\leqslant F(x^t) - \left(\gamma - \frac{L\gamma^2}{2}\right)\|\nabla F(x^t)\|^2 \tag{49}$$

If $\|\nabla F(x^t)\| = 0$, we found a solution and GD has converged. Otherwise $\|\nabla F(x^t)\| > 0$, and we have

$$\left(\gamma - \frac{L\gamma^2}{2}\right)\|\nabla F(x^t)\|^2 \leqslant F(x^t) - F(x^{t+1}) \tag{50}$$

We need to characterize when the left hand side is positive

$$\gamma - \frac{L\gamma^2}{2} > 0 \Leftrightarrow \quad 1 - \frac{L\gamma}{2} > 0 \qquad \text{(since } \gamma > 0\text{)} \tag{51}$$

$$\Leftrightarrow \quad \frac{L\gamma}{2} < 1 \quad \Leftrightarrow \quad L\gamma < 2 \quad \Leftrightarrow \quad \gamma < \frac{2}{L} \tag{52}$$

Then, since $0 < \gamma < \frac{2}{L}$, we have

$$0 < \left(\gamma - \frac{L\gamma^2}{2}\right)\|\nabla F(x^t)\|^2 \leqslant F(x^t) - F(x^{t+1}) \tag{53}$$

Then $F(x^t)$ is decresing with t. By summing over $t = 0 \ldots T$, using telescopic cancellation, and using the assumption that $F(x) \geqslant C$, we get

$$0 < \underbrace{\left(\gamma - \frac{L\gamma^2}{2}\right)}_{\text{constant wrt } T} \sum_{t=0}^{T} \|\nabla F(x^t)\|^2 \leqslant F(x^0) - F(x^{T+1}) \leqslant \underbrace{F(x^0) - C}_{\text{constant wrt } T}, \quad \text{for all } T > 0\ . \tag{54}$$

Thus, $0 < \sum_{t=0}^{\infty} \|\nabla F(x^t)\|^2 < \infty$ which yields $\lim_{t\to\infty} \|\nabla F(x^t)\| = 0$. □

2 Speed of convergence

We want to prove the following theorem.

Theorem 2. *Let $F : \mathbb{R}^N \to \mathbb{R}$ be differentiable with L Lipschitz gradient, lower bounded and convex. Then, provided $0 < \gamma < \frac{2}{L}$, the gradient descent sequence defined as:*

$$x^{t+1} = x^t - \gamma \nabla F(x^t) \tag{55}$$

converges to a stationary point $x^\star$

$$\nabla F(x^\star) = 0 \ , \tag{56}$$

with the speed

$$F(x^t) - F(x^\star) \leqslant \frac{\|x^0 - x^\star\|^2}{\left(\gamma - \frac{L\gamma^2}{2}\right) t} \ . \tag{57}$$

Corollary 1. *Under the assumptions of Theorem 2 but with $0 < \gamma < \frac{1}{L}$, the speed becomes*

$$F(x^t) - F(x^\star) \leqslant \frac{2L\|x^0 - x^\star\|^2}{t} \ . \tag{58}$$

2.1 Convexity + Lipschitz gradient ⇒ Co-coercivity of gradient

Lemma 5. *Let $F : \mathbb{R}^N \to \mathbb{R}$ be differentiable with L Lipschitz gradient, and convex. Then we have co-coercivity of the gradient, i.e.:*

$$\frac{1}{L}\|\nabla F(x) - \nabla F(y)\|^2 \leqslant \langle \nabla F(x) - \nabla F(y), \, x - y\rangle \tag{59}$$

Proof. Let $x \in \mathbb{R}^N$, $y \in \mathbb{R}^N$ and $z \in \mathbb{R}^N$. Since F has L Lipschitz gradient, we obtain by Lemma 4:

$$F(z) \leqslant F(x) + \langle \nabla F(x), \, z - x\rangle + \frac{L}{2}\|z - x\|^2 \tag{60}$$

$$\Rightarrow \quad F(z) - F(x) \leqslant \langle \nabla F(x), \, z - x\rangle + \frac{L}{2}\|z - x\|^2 \tag{61}$$

Since F is convex, we obtain by Lemma 3:

$$F(z) \geqslant F(y) + \langle \nabla F(y), \, z - y\rangle \tag{62}$$

$$\Rightarrow \quad F(y) - F(z) \leqslant -\langle \nabla F(y), \, z - x\rangle - \langle \nabla F(y), \, x - y\rangle \tag{63}$$

$$\Rightarrow \quad F(y) - F(z) + \langle \nabla F(y), \, x - y\rangle \leqslant -\langle \nabla F(y), \, z - x\rangle \tag{64}$$

Adding (61) and (64) leads to

$$F(y) - F(x) + \langle \nabla F(y), \, x - y\rangle \leqslant \underbrace{\langle \nabla F(x) - \nabla F(y), \, z - x\rangle + \frac{L}{2}\|z - x\|^2}_{=H(z)} \tag{65}$$

Since the right hand side is true for all z, we want to find z that minimizes this quantity in order to get the tightest upper-bound. We have

$$\nabla H(z) = L(z - x) + \nabla F(x) - \nabla F(y) \quad \text{and} \quad \text{Hessian}[H(z)] = L \cdot \text{Id} \tag{66}$$

Then H is convex and quadratic, and the minimum is reached at:

$$z^\star = x - \frac{1}{L}\left(\nabla F(x) - \nabla F(y)\right) \tag{67}$$

Plugging $z^\star$ in the previous equation leads to

$$F(y) - F(x) + \langle \nabla F(y),\, x - y\rangle \leqslant -\frac{1}{L}\langle \nabla F(x) - \nabla F(y),\, \nabla F(x) - \nabla F(y)\rangle + \frac{1}{2L}\|\nabla F(x) - \nabla F(y)\|^2 \quad (68)$$

$$\leqslant -\frac{1}{2L}\|\nabla F(x) - \nabla F(y)\|^2 \quad (69)$$

We can swap the role of x and y, then

$$F(x) - F(y) + \langle \nabla F(x),\, y - x\rangle \leqslant -\frac{1}{2L}\|\nabla F(x) - \nabla F(y)\|^2 \quad (70)$$

Summing both leads to

$$\langle \nabla F(x) - \nabla F(y),\, x - y\rangle \geqslant \frac{1}{L}\|\nabla F(x) - \nabla F(y)\|^2 \ . \quad (71)$$

□

Remark 5. *For convex functions, the reciprocal holds true.*

Remark 6. *A direct consequence is that:* $\langle \nabla F(x),\, x - x^\star\rangle \geqslant \frac{1}{L}\|\nabla F(x)\|^2$.

2.2 Proof of Theorem 2

Proof. Since F is differentatiable, convex with L Lipschitz gradient, then, by Lemma 5 and using the definition of x^{t+1}, we have

$$\|x^{t+1} - x^\star\|^2 = \|x^t - x^\star - \gamma\nabla F(x^t)\|^2 \quad (72)$$

$$= \|x^t - x^\star\|^2 + \gamma^2\|\nabla F(x^t)\|^2 - 2\gamma\langle \nabla F(x^t),\, x^t - x^\star\rangle \quad (73)$$

$$\leqslant \|x^t - x^\star\|^2 + \left(\gamma^2 - \frac{2\gamma}{L}\right)\|\nabla F(x^t)\|^2 \quad (74)$$

As $\gamma > 0$, we have that

$$\gamma^2 - \frac{2\gamma}{L} < 0 \Rightarrow \gamma - \frac{2}{L} < 0 \Rightarrow \gamma < \frac{2}{L} \quad (75)$$

Then, since $0 < \gamma < 2/L$, we have $\gamma^2 - \frac{2\gamma}{L} < 0$, and then

$$\|x^{t+1} - x^\star\| < \|x^t - x^\star\| \leqslant \ldots \leqslant \|x^0 - x^\star\| \quad (76)$$

By Lemma 3, since F is differentiable and convex we also have

$$F(x^\star) \geqslant F(x^t) + \langle \nabla F(x^t),\, x^\star - x^t\rangle \quad (77)$$

$$\Rightarrow \quad F(x^t) - F(x^\star) \leqslant \langle \nabla F(x^t),\, x^t - x^\star\rangle \quad (78)$$

$$\Rightarrow \quad F(x^t) - F(x^\star) \leqslant \|\nabla F(x^t)\|\|x^t - x^\star\| \qquad \text{(Cauchy-Schwartz inequality)} \quad (79)$$

$$\Rightarrow \quad F(x^t) - F(x^\star) \leqslant \|\nabla F(x^t)\|\|x^0 - x^\star\| \quad (80)$$

$$\Rightarrow \quad \frac{(F(x^t) - F(x^\star))^2}{\|x^0 - x^\star\|^2} \leqslant \|\nabla F(x^t)\|^2 \quad (81)$$

$$\Rightarrow \quad -\|\nabla F(x^t)\|^2 \leqslant -\frac{(F(x^t) - F(x^\star))^2}{\|x^0 - x^\star\|^2} \quad (82)$$

Since F is differentiable with L Lipschitz gradient, based on Lemma 4, we have

$$F(x^{t+1}) \leqslant F(x^t) + \langle \nabla F(x^t),\, x^{t+1} - x^t\rangle + \frac{L}{2}\|x^{t+1} - x^t\|^2 \quad (83)$$

$$\Rightarrow \quad F(x^{t+1}) \leqslant F(x^t) - \langle \nabla F(x^t), \gamma \nabla F(x^t) \rangle + \frac{L}{2} \|\gamma \nabla F(x^t)\|^2 \tag{84}$$

$$\Rightarrow \quad F(x^{t+1}) \leqslant F(x^t) - \gamma \|\nabla F(x^t)\|^2 + \frac{L\gamma^2}{2} \|\nabla F(x^t)\|^2 \tag{85}$$

$$\Rightarrow \quad F(x^{t+1}) \leqslant F(x^t) - \left(\gamma - \frac{L\gamma^2}{2}\right) \|\nabla F(x^t)\|^2 \tag{86}$$

In particular, since $F(x^\star) \leqslant F(x)$, we have

$$0 < \gamma < \frac{2}{L} \Rightarrow \gamma < \frac{2}{L} \Rightarrow 1 - \frac{L\gamma}{2} > 0 \Rightarrow \gamma - \frac{L\gamma^2}{2} > 0 \quad (\text{since } \gamma > 0) \tag{87}$$

$$\Rightarrow \quad F(x^{t+1}) \leqslant F(x^t) \Rightarrow F(x^{t+1}) - F(x^\star) \leqslant F(x^t) - F(x^\star) \tag{88}$$

$$\Rightarrow \quad 1 \leqslant \frac{F(x^t) - F(x^\star)}{F(x^{t+1}) - F(x^\star)} \Rightarrow -\frac{F(x^t) - F(x^\star)}{F(x^{t+1}) - F(x^\star)} \leqslant -1 \tag{89}$$

Injecting (82) into (86) and using the last inequality leads to

$$F(x^{t+1}) \leqslant F(x^t) - \left(\gamma - \frac{L\gamma^2}{2}\right) \frac{(F(x^t) - F(x^\star))^2}{\|x^0 - x^\star\|^2} \tag{90}$$

$$\Rightarrow \quad F(x^{t+1}) - F(x^\star) \leqslant F(x^t) - F(x^\star) - \left(\gamma - \frac{L\gamma^2}{2}\right) \frac{(F(x^t) - F(x^\star))^2}{\|x^0 - x^\star\|^2} \tag{91}$$

$$\Rightarrow \quad \frac{F(x^{t+1}) - F(x^\star)}{F(x^t) - F(x^\star)} \leqslant 1 - \left(\gamma - \frac{L\gamma^2}{2}\right) \frac{F(x^t) - F(x^\star)}{\|x^0 - x^\star\|^2} \tag{92}$$

$$\Rightarrow \quad \frac{1}{F(x^t) - F(x^\star)} \leqslant \frac{1}{F(x^{t+1}) - F(x^\star)} - \frac{\left(\gamma - \frac{L\gamma^2}{2}\right)}{\|x^0 - x^\star\|^2} \frac{F(x^t) - F(x^\star)}{F(x^{t+1}) - F(x^\star)} \tag{93}$$

$$\Rightarrow \quad \frac{1}{F(x^t) - F(x^\star)} \leqslant \frac{1}{F(x^{t+1}) - F(x^\star)} - \frac{\left(\gamma - \frac{L\gamma^2}{2}\right)}{\|x^0 - x^\star\|^2} \tag{94}$$

$$\Rightarrow \quad \frac{\left(\gamma - \frac{L\gamma^2}{2}\right)}{\|x^0 - x^\star\|^2} \leqslant \frac{1}{F(x^{t+1}) - F(x^\star)} - \frac{1}{F(x^t) - F(x^\star)} \tag{95}$$

Summing for $t = 0 \dots T-1$ and using telescopic cancellation leads to

$$T \frac{\left(\gamma - \frac{L\gamma^2}{2}\right)}{\|x^0 - x^\star\|^2} \leqslant \frac{1}{F(x^T) - F(x^\star)} - \frac{1}{F(x^0) - F(x^\star)} \leqslant \frac{1}{F(x^T) - F(x^\star)} \tag{96}$$

$$\Rightarrow \quad F(x^T) - F(x^\star) \leqslant \frac{\|x^0 - x^\star\|^2}{T\left(\gamma - \frac{L\gamma^2}{2}\right)} \tag{97}$$

which concludes the proof □

2.3 Proof of Corollary 1

Proof. By assumption, we have

$$\gamma < 1/L \Rightarrow 1 - L\gamma/2 > 1/2 \Rightarrow \gamma(1 - L\gamma/2) > \gamma/2 \quad (\text{since } \gamma > 0) \tag{98}$$

$$\Rightarrow \quad \frac{1}{\gamma(1 - L\gamma/2)} < \frac{2}{\gamma} \tag{99}$$

and then

$$F(x^T) - F(x^\star) \leqslant \frac{2\|x^0 - x^\star\|^2}{T\gamma} \tag{100}$$

□