Exercise List: Convergence rates and complexity

Francis Bach and Robert M. Gower

February 10, 2019

1 Rate of convergence and complexity

All the algorithm we discuss in the course generate a sequence of random vectors x^t that converge to a desired x^* in some sense. Because the x^t's are random we always prove convergence in expectation. In particular, we focus on two forms of convergence, either showing that the difference of function values converges

$$\mathbb{E}\left[f(x^t) - f(x^*)\right] \longrightarrow 0,$$

or the expected norm difference of the iterates converges

$$\mathbb{E}\left[\|x^t - x^*\|^2\right] \longrightarrow 0.$$

Two important questions: 1) How fast is this convergence and 2) given an ϵ how many iterations t are needed before $\mathbb{E}\left[f(x^t) - f(x^*)\right] < \epsilon$ or $\mathbb{E}\left[\|x^t - x^*\|^2\right] < \epsilon$.

Ex. 1 — Consider a sequence $(\alpha_t)_t \in \mathbb{R}_+$ that converge to zero according to

$$\alpha_t \leq \frac{C}{t},$$

where $C > 0$. Given an $\epsilon > 0$, show that

$$t \geq \frac{C}{\epsilon} \quad \Rightarrow \alpha_t < \epsilon.$$

We refer to this result as a $O(1/\epsilon)$ iteration complexity.

Ex. 2 — Using that

$$\frac{1}{1-\rho}\log\left(\frac{1}{\rho}\right) \geq 1, \tag{1}$$

prove the following lemma.

Lemma 1.1. Consider the sequence $(\alpha_k)_k \in \mathbb{R}_+$ of positive scalars that converges to zero according to

$$\alpha_k \leq \rho^k \, \alpha_0, \tag{2}$$

where $\rho \in [0, 1)$. For a given $1 > \epsilon > 0$ we have that

$$k \geq \frac{1}{1-\rho} \log\left(\frac{1}{\epsilon}\right) \quad \Rightarrow \quad \alpha_k \leq \epsilon \, \alpha_0. \tag{3}$$

We refer to this as a $O(\log(1/\epsilon))$ iteration complexity.

Following the introduction, we can write $\alpha^t \overset{\text{def}}{=} \mathbb{E}\left[f(x^t) - f(x^*)\right]$ or $\alpha^t \overset{\text{def}}{=} \mathbb{E}\left[\|x^t - x^*\|^2\right]$. The type of convergence (2) is known as *linear convergence at a rate of* ρ^k.

Exercise List: Proving convergence of the Stochastic Gradient Descent and Coordinate Descent on the Ridge Regression Problem.

Robert M. Gower & Francis Bach & Nidham Gazagnadou

February 10, 2019

Introduction

Consider the task of learning a rule that maps the *feature vector* $x \in \mathbb{R}^d$ to outputs $y \in \mathbb{R}$. Furthermore you are given a set of labelled observations (x_i, y_i) for $i = 1, \ldots, n$. We restrict ourselves to linear mappings. That is, we need to find $w \in \mathbb{R}^d$ such that

$$x_i^\top w \approx y_i, \quad \text{for } i = 1, \ldots, n. \tag{1}$$

That is the *hypothesis function* is parametrized by w and is given by $h_w : x \mapsto w^\top x$.[1] To choose a w such that each $x_i^\top w$ is close to y_i, we use the squared loss $\ell(y) = y^2/2$ and the squared regularizor. That is, we minimize

$$w^* = \arg\min_w \frac{1}{n}\sum_{i=1}^n \frac{1}{2}(x_i^\top w - y_i)^2 + \frac{\lambda}{2}\|w\|_2^2, \tag{2}$$

where $\lambda > 0$ is the regularization parameter. We now have a complete training problem (2)[2].

Using the matrix notation

$$X \stackrel{\text{def}}{=} [x_1, \ldots, x_n] \in \mathbb{R}^{d\times n}, \quad \text{and} \quad y = [y_1, \ldots, y_n] \in \mathbb{R}^n, \tag{3}$$

we can re-write the objective function in (2) as

$$f(w) \stackrel{\text{def}}{=} \frac{1}{2n}\|X^\top w - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2. \tag{4}$$

First we introduce some necessary notation.

[1] We need only consider a linear mapping as opposed to the more general *affine* mapping $x_i \mapsto w^\top x_i + \beta$, because the zero order term $\beta \in \mathbb{R}$ can be incorporated by defining a new feature vectors $\hat{x}_i = [x_1, 1]$ and new variable $\hat{w} = [w, \beta]$ so that $\hat{x}_i^\top \hat{w} = x_i^\top w + \beta$

[2] Excluding the issue of selection λ using something like crossvalidation `https://en.wikipedia.org/wiki/Cross-validation_(statistics)`

Notation: For every $x, w, \in \mathbb{R}^d$ let $\langle x, w\rangle \overset{\text{def}}{=} x^\top y$ and let $\|x\|_2 = \sqrt{\langle x, x\rangle}$. Let $A \in \mathbb{R}^{d\times d}$ be a matrix and let $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ be the smallest and largest singular values of A defined by

$$\sigma_{\min}(A) \overset{\text{def}}{=} \min_{x\in\mathbb{R}^d,\, x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} \quad \text{and} \quad \sigma_{\max}(A) \overset{\text{def}}{=} \max_{x\in\mathbb{R}^d,\, x\neq 0} \frac{\|Ax\|_2}{\|x\|_2}. \tag{5}$$

Finally, a result you will need, if A is a symmetric positive semi-definite matrix the largest singular value of A can be defined instead as

$$\sigma_{\max}(A) = \max_{x\in\mathbb{R}^d,\, x\neq 0} \frac{\langle Ax, x\rangle_2}{\|x\|_2^2} = \max_{x\in\mathbb{R}^d,\, x\neq 0} \frac{\|Ax\|_2}{\|x\|_2}. \tag{6}$$

Therefore

$$\frac{\langle Ax, x\rangle}{\|x\|_2^2} \leq \sigma_{\max}(A), \quad \forall x \in \mathbb{R}^d \setminus \{0\}. \tag{7}$$

and

$$\frac{\|Ax\|_2}{\|x\|_2} \leq \sigma_{\max}(A), \quad \forall x \in \mathbb{R}^d \setminus \{0\}. \tag{8}$$

We will now solve the following ridge regression problem

$$w^* = \arg\min_{w\in\mathbb{R}^d} \left(\frac{1}{2n}\|X^\top w - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2 \overset{\text{def}}{=} f(w) \right), \tag{9}$$

using stochastic gradient descent and stochastic coordinate descent.

Exercise 1 : Stochastic Gradient Descent (SGD)

Some more notation: Let $\|A\|_F^2 \overset{\text{def}}{=} \mathrm{Tr}\left(A^\top A\right)$ denote the Frobenius norm of A. Let

$$A \overset{\text{def}}{=} \tfrac{1}{n} X X^\top + \lambda I \in \mathbb{R}^{d\times d} \quad \text{and} \quad b \overset{\text{def}}{=} \tfrac{1}{n} X y. \tag{10}$$

We can exploit the separability of the objective function (2) to design a *stochastic* gradient method. For this, first we re-write the problem $Aw = b$ as different linear least squares problem

$$\hat{w}^* = \arg\min_w \tfrac{1}{2}\|Aw - b\|_2^2 \quad = \quad \arg\min_w \sum_{i=1}^d \tfrac{1}{2}(A_{i:}w - b_i)^2 \quad \overset{\text{def}}{=} \quad \arg\min_w \sum_{i=1}^d p_i f_i(w), \tag{11}$$

where $f_i(w) = \frac{1}{2p_i}(A_{i:}w - b_i)^2$, $A_{i:}$ denotes the ith row of A, b_i denotes the ith element of b and $p_i = \frac{\|A_{i:}\|_2^2}{\|A\|_F^2}$ for $i = 1, \ldots, d$. Note that $\sum_{i=1}^d p_i = 1$ thus the p_i's are probabilities.

From a given $w^0 \in \mathbb{R}^d$, consider the iterates

$$w^{t+1} = w^t - \alpha \nabla f_j(w^t), \tag{12}$$

where

$$\alpha = \frac{1}{\|A\|_F^2}, \tag{13}$$

and j is a random index chosen from $\{1, \ldots, d\}$ sampled with probability p_j. In other words, $\mathbb{P}(j = i) = p_i = \frac{\|A_{i:}\|_2^2}{\|A\|_F^2}$ for all $i \in \{1, \ldots, d\}$.

Question 1.1: Show that the solution $\hat{w}^*$ to (11) and the solution to w^* to (9) are equal.

Question 1.2: Show that

$$\nabla f_j(w) = \frac{1}{p_j} A_{j:}^\top A_{j:}(w - w^*) \tag{14}$$

and that

$$\mathbb{E}_{j\sim p}\left[\nabla f_j(w)\right] \overset{\text{def}}{=} \sum_{i=1}^{d} p_i \nabla f_i(w) = A^\top A(w - w^*) \ ,$$

thus $\nabla f_j(w)$ is an unbiased estimator of the full gradient of the objective function in (11). This justifies applying the stochastic gradient method.

Question 1.3: Let $\Pi_j \overset{\text{def}}{=} \frac{A_{j:}^\top A_{j:}}{\|A_{j:}\|_2^2}$, show that

$$\Pi_j \Pi_j = \Pi_j \ , \tag{15}$$

and

$$(I - \Pi_j)(I - \Pi_j) = I - \Pi_j. \tag{16}$$

In other words, Π_j is a projection operator which projects orthogonally onto $\textbf{Range}\left(A_{j:}\right)$. Furthermore, if $j \sim p_j$ verify that

$$\mathbb{E}\left[\Pi_j\right] = \sum_{i=1}^{d} p_i \Pi_i = \frac{A^\top A}{\|A\|_F^2}. \tag{17}$$

Question 1.4: Show the following equality ruling the squared norm of the distance to the solution

$$\|w^{t+1} - w^*\|_2^2 \quad = \quad \|w^t - w^*\|_2^2 - \left\langle \frac{A_{j:}^\top A_{j:}}{\|A_{j:}\|_2^2}(w^t - w^*), w^t - w^* \right\rangle \ . \tag{18}$$

Question 1.5: Using previous answer and analogous techniques from the course, show that the iterates (12) converge according to

$$\mathbb{E}\left[\|w^{t+1} - w^*\|_2^2\right] \quad \leq \quad \left(1 - \frac{\sigma_{\min}(A)^2}{\|A\|_F^2}\right) \mathbb{E}\left[\|w^t - w^*\|_2^2\right] \quad . \tag{19}$$

Remark: This is an amazing and recent result [2], since it shows that SGD converges exponentially fast despite the fact that the iterates (14) only require access to a single row of A at a time! This result can be extended to solving any linear system $Aw = b$, including the case where A rank deficient. Indeed, so long as there exists a solution to $Aw = b$, the iterates (14) converge to the solution of least norm and at rate of $\left(1 - \frac{\sigma_{\min}^+(A)^2}{\|A\|_F^2}\right)$ where $\sigma_{\min}^+(A)$ is the smallest nonzero singular value of A [1]. Thus this method can solve any linear system.

BONUS

Exercise 2: Stochastic Coordinate Descent (CD)

Consider the minimization problem

$$w^* = \arg\min_{x \in \mathbb{R}^d} \left(f(w) \stackrel{\text{def}}{=} \frac{1}{2} w^\top A w - w^\top b \right), \tag{20}$$

where $A \in \mathbb{R}^{d \times d}$ is a symmetric positive definite matrix, and $w, b \in \mathbb{R}^d$.

Question 2.1: First show that, using the notation (10), solving (20) is equivalent to solving (9).

Question 2.2: Show that

$$\frac{\partial f(w)}{\partial w_i} = A_{i:} w - b_i \ \ , \tag{21}$$

where $A_{i:}$ is the ith row of A. Furthermore note that $w^* = A^{-1} b$, thus

$$\frac{\partial f(w)}{\partial w_i} = e_i^\top (Aw - b) = e_i^\top A(w - w^*) \ \ . \tag{22}$$

Question 2.3: Consider a step of the stochastic coordinate descent method

$$w^{k+1} = w^k - \alpha_i \frac{\partial f(w^k)}{\partial x_i} e_i, \tag{23}$$

where $e_i \in \mathbb{R}^d$ is the ith unit coordinate vector, $\alpha_i = \dfrac{1}{A_{ii}}$, and $i \in \{1, \ldots, d\}$ is sampled i.i.d at each step according to $i \sim p_i$ where $p_i = \dfrac{A_{ii}}{\mathrm{Tr}\,(A)}$. Let $\|x\|_A^2 \stackrel{\text{def}}{=} x^\top A x$.

First, prove that

$$\|w^{k+1} - w^*\|_A^2 = \left\langle (I - \Pi_i^\top) A (I - \Pi_i)(w^k - w^*), w^k - w^* \right\rangle \ \ . \tag{24}$$

Question 2.4: Let $r^k \stackrel{\text{def}}{=} A^{1/2}(w^k - w^*)$. Deduce from (24) that

$$\|r^{k+1}\|_2^2 \quad = \quad \|r^k\|_2^2 - \left\langle \frac{A^{1/2} e_i e_i^\top A^{1/2}}{A_{ii}} r^k, r^k \right\rangle \ \ . \tag{25}$$

Question 2.5: Finally, prove the convergence of the iterates of CD (23) converge according to

$$\mathbb{E}\left[\|w^{k+1} - w^*\|_A^2\right] \quad \leq \quad \left(1 - \frac{\lambda_{\min}(A)}{\mathrm{Tr}\,(A)}\right) \mathbb{E}\left[\|w^k - w^*\|_A^2\right] \tag{26}$$

thus (23) converges to the solution.
Hint: Since A is symmetric positive definite you can use that

$$\lambda_{\min}(A) = \inf_{x \in \mathbb{R}^d, x \neq 0} \frac{x^\top A x}{\|x\|_2^2}.$$

You will need to use that $x^\top A x \geq \lambda_{\min}(A) \|x\|_2^2$ at some point.

Question 2.6: When is this stochastic gradient method (14) *faster* than the stochastic coordinate descent method of gradient descent? Note that the each iteration of SGD and CD costs $O(d)$ floating point operations while an iteration of the GD method costs $O(d^2)$ floating point operations (assuming that A has been previously calculated and stored). What happens if d is very big? What if $\|A\|_F^2$ is very large? Discuss this.

References

[1] R. M. Gower and P. Richtárik. "Stochastic Dual Ascent for Solving Linear Systems". In: *arXiv:1512.06890* (2015).

[2] T. Strohmer and R. Vershynin. "A Randomized Kaczmarz Algorithm with Exponential Convergence". In: *Journal of Fourier Analysis and Applications* 15.2 (2009), pp. 262–278.

(BONUS) Exercise List: Proving convergence of the Stochastic Gradient Descent for smooth and convex functions.

Robert M. Gower

February 10, 2019

1 Introduction

Consider the problem

$$w^* \in \arg\min_w \left(\frac{1}{n} \sum_{i=1}^n f_i(w) \overset{\text{def}}{=} f(w) \right), \tag{1}$$

where we assume that $f(w)$ is μ–strongly quasi-convex

$$f(w^*) \geq f(w) + \langle w^* - w, \nabla f(w) \rangle + \frac{\mu}{2} \|w - w^*\|^2, \tag{2}$$

and each f_i is convex and L_i–smooth

$$f_i(w+h) \leq f_i(w) + \langle \nabla f_i(w), h \rangle + \frac{L_i}{2} \|h\|^2, \quad \text{for } i = 1, \ldots, n. \tag{3}$$

Here we will provide a modern proof of the convergence of the SGD algorithm

$$w^{t+1} \quad = \quad w^t - \gamma^t \nabla f_{i_t}(w^t), \quad \text{where } i_t \sim \frac{1}{n}. \tag{4}$$

The result we will prove is given in the following theorem.

Theorem 1.1. Assume f is μ-quasi-strongly convex and the f_i's are convex and L_i–smooth. Let $L_{\max} = \max_{i=1,\ldots,n} L_i$ and let

$$\sigma^2 \quad \overset{\text{def}}{=} \quad \sum_{i=1}^n \frac{1}{n} \|\nabla f_i(w^*)\|^2. \tag{5}$$

Choose $\gamma^t = \gamma \in (0, \frac{1}{2L_{\max}}]$ for all t. Then the iterates of SGD given by (4) satisfy:

$$\mathbb{E}\|w^t - w^*\|^2 \leq (1 - \gamma\mu)^t \|w^0 - w^*\|^2 + \frac{2\gamma\sigma^2}{\mu}. \tag{6}$$

2 Proof of Theorem 1.1

We will now give a modern proof of the convergance of SGD.

Ex. 1 — Let $\mathbb{E}_t\left[\cdot\right] \stackrel{\text{def}}{=} \mathbb{E}\left[\cdot \,|\, w^t\right]$ and consider the tth iteration of the SGD method (4). Show that

$$\mathbb{E}_t\left[\nabla f_{i_t}(w^t)\right] = \nabla f(w^t).$$

Ex. 2 — Let $\mathbb{E}_t\left[\cdot\right] \stackrel{\text{def}}{=} \mathbb{E}\left[\cdot \,|\, w^t\right]$ be the expectation conditioned on w^t. Using a step of SGD (4) show that

$$\mathbb{E}_t\left[\|w^{t+1} - w^*\|^2\right] \quad = \quad \|w^t - w^*\|^2 - 2\gamma\left\langle w^t - w^*, \nabla f(w^t)\right\rangle + \gamma^2 \sum_{i=1}^n \frac{1}{n}\|\nabla f_i(w^t)\|^2. \tag{7}$$

Ex. 3 — Now we need to bound the term $\sum_{i=1}^n \frac{1}{n}\|\nabla f_i(w^t)\|^2$ to continue the proof. We break this into the following steps.

Part I

Using that each f_i is L_i–smooth and convex and using Lemma A.1 in the appendix show that

$$\sum_{i=1}^n \frac{1}{2nL_i}\|\nabla f_i(w) - \nabla f_i(w^*)\|_2^2 \quad \leq \quad f(w) - f(w^*). \tag{9}$$

Hint: Remember that $\nabla f(w^*) = 0$.
Now let $L_{\max} = \max_{i=1,\ldots,n} L_i$ and conlude that

$$\sum_{i=1}^n \frac{1}{n}\|\nabla f_i(w) - \nabla f_i(w^*)\|_2^2 \quad \leq \quad 2L_{\max}(f(w) - f(w^*)). \tag{10}$$

Part II

Using (10) and Definition 5 show that

$$\sum_{i=1}^n \frac{1}{n}\|\nabla f_i(w)\|^2 \quad \leq \quad 4L_{\max}(f(w) - f(w^*)) + 2\sigma^2. \tag{11}$$

Ex. 4 — Using (11) together with (7) and the strong quasi-convexity (2) of $f(w)$ show that

$$\mathbb{E}_t\left[\|w^{t+1} - w^*\|^2\right] \quad \leq \quad (1-\mu\gamma)\|w^t - w^*\|^2 + 2\gamma(2\gamma L_{\max} - 1)(f(w^t) - f(w^*)) + 2\sigma^2\gamma^2. \tag{15}$$

Ex. 5 — Using that $\gamma \in (0, \frac{1}{2L_{\max}}]$ conclude the proof by taking expectation again, and unrolling the recurrence.

Ex. 6 — BONUS importance sampling: Let $i_t \sim p_i$ in the SGD update (4), where $p_i > 0$ are probabilities with $\sum_{i=1}^n p_i = 1$. What should the p_i's be so that SGD has the fastest convergence?

3 Decreasing step-sizes

Based on Theorem 1.1 we can introduce a decreasing stepsize.

Theorem 3.1 (Decreasing stepsizes)**.** Let f be μ–strongly quasi-convex and each f_i be L_i–smooth and convex. Let $\mathcal{K} \overset{\text{def}}{=} L_{\max}/\,\mu$ and

$$\gamma^t = \begin{cases} \frac{1}{2L_{\max}} & \text{for} \quad t \leq 4\lceil \mathcal{K} \rceil \\ \frac{2t+1}{(t+1)^2\mu} & \text{for} \quad t > 4\lceil \mathcal{K} \rceil. \end{cases} \tag{18}$$

If $t \geq 4\lceil \mathcal{K} \rceil$, then SGD iterates given by (4) satisfy:

$$\mathbb{E}\|w^t - w^*\|^2 \leq \frac{\sigma^2}{\mu^2}\frac{8}{t} + \frac{16}{e^2}\frac{\lceil \mathcal{K} \rceil^2}{t^2}\|w^0 - w^*\|^2. \tag{19}$$

Proof. Let $\gamma_t \overset{\text{def}}{=} \frac{2t+1}{(t+1)^2\mu}$ and let t^* be an integer that satisfies $\gamma_{t^*} \leq \frac{1}{2L_{\max}}$. In particular this holds for

$$t^* \geq \lceil 4\mathcal{K} - 1 \rceil.$$

Note that γ_t is decreasing in t and consequently $\gamma_t \leq \frac{1}{2L_{\max}}$ for all $t \geq t^*$. This in turn guarantees that (6) holds for all $t \geq t^*$ with γ_t in place of γ, that is

$$\mathbb{E}\|r^{t+1}\|^2 \leq \frac{t^2}{(t+1)^2}\mathbb{E}\|r^t\|^2 + \frac{2\sigma^2}{\mu^2}\frac{(2t+1)^2}{(t+1)^4}. \tag{20}$$

Multiplying both sides by $(t+1)^2$ we obtain

$$\begin{aligned} (t+1)^2\mathbb{E}\|r^{t+1}\|^2 \quad &\leq \quad t^2\mathbb{E}\|r^t\|^2 + \frac{2\sigma^2}{\mu^2}\left(\frac{2t+1}{t+1}\right)^2 \\ &\leq \quad t^2\mathbb{E}\|r^t\|^2 + \frac{8\sigma^2}{\mu^2}, \end{aligned}$$

where the second inequality holds because $\frac{2t+1}{t+1} < 2$. Rearranging and summing from $j = t^* \ldots t$ we obtain:

$$\sum_{j=t^*}^{t} \left[(j+1)^2\mathbb{E}\|r^{j+1}\|^2 - j^2\mathbb{E}\|r^j\|^2\right] \leq \sum_{j=t^*}^{t} \frac{8\sigma^2}{\mu^2}. \tag{21}$$

Using telescopic cancellation gives

$$(t+1)^2\mathbb{E}\|r^{t+1}\|^2 \leq (t^*)^2\mathbb{E}\|r^{t^*}\|^2 + \frac{8\sigma^2(t-t^*)}{\mu^2}.$$

Dividing the above by $(t+1)^2$ gives

$$\mathbb{E}\|r^{t+1}\|^2 \leq \frac{(t^*)^2}{(t+1)^2}\mathbb{E}\|r^{t^*}\|^2 + \frac{8\sigma^2(t-t^*)}{\mu^2(t+1)^2}. \tag{22}$$

For $t \leq t^*$ we have that (6) holds, which combined with (22), gives

$$\begin{aligned} \mathbb{E}\|r^{t+1}\|^2 \quad &\leq \quad \frac{(t^*)^2}{(t+1)^2}\left(1 - \frac{\mu}{2L_{\max}}\right)^{t^*}\|r^0\|^2 \\ &+ \quad \frac{\sigma^2}{\mu^2(t+1)^2}\left(8(t-t^*) + \frac{(t^*)^2}{\mathcal{K}}\right). \end{aligned} \tag{23}$$

Choosing t^* that minimizes the second line of the above gives $t^* = 4\lceil\mathcal{K}\rceil$, which when inserted into (23) becomes

$$\begin{aligned}\mathbb{E}\|r^{t+1}\|^2 &\leq \frac{16\lceil\mathcal{K}\rceil^2}{(t+1)^2}\left(1-\frac{1}{2\mathcal{K}}\right)^{4\lceil\mathcal{K}\rceil}\|r^0\|^2 \\ &\quad +\frac{\sigma^2}{\mu^2}\frac{8(t-2\lceil\mathcal{K}\rceil)}{(t+1)^2} \\ &\leq \frac{16\lceil\mathcal{K}\rceil^2}{e^2(t+1)^2}\|r^0\|^2+\frac{\sigma^2}{\mu^2}\frac{8}{t+1}, \end{aligned} \tag{24}$$

where we have used that $\left(1-\frac{1}{2x}\right)^{4x} \leq e^{-2}$ for all $x \geq 1$. $\square$

A Appendix: Auxiliary smooth and convex lemma

As a consequence of the f_i's being smooth and convex we have that f is also smooth and convex. In particular f is convex since it is a convex combination of the f_i's. This gives us the following useful lemma.

Lemma A.1. If f is both L–smooth

$$f(z) \leq f(w) + \langle \nabla f(w), z-w\rangle + \frac{L}{2}\|z-w\|_2^2 \tag{25}$$

and convex

$$f(z) \geq f(y) + \langle \nabla f(y), z-y\rangle, \tag{26}$$

then we have that

$$f(y)-f(w) \leq \langle \nabla f(y), y-w\rangle - \frac{1}{2L}\|\nabla f(y)-\nabla f(w)\|_2^2. \tag{27}$$

Proof. To prove (27), it follows that

$$\begin{aligned}f(y)-f(w) &= f(y)-f(z)+f(z)-f(w) \\ &\overset{(26)+(25)}{\leq} \langle \nabla f(y), y-z\rangle + \langle \nabla f(w), z-w\rangle + \frac{L}{2}\|z-w\|_2^2.\end{aligned}$$

To get the tightest upper bound on the right hand side, we can minimize the right hand side in z, which gives

$$z = w - \frac{1}{L}(\nabla f(w) - \nabla f(y)). \tag{28}$$

Substituting this in gives

$$\begin{aligned}f(y)-f(w) &= \left\langle \nabla f(y), y-w+\frac{1}{L}(\nabla f(w)-\nabla f(y))\right\rangle \\ &\quad -\frac{1}{L}\langle \nabla f(w), \nabla f(w)-\nabla f(y)\rangle + \frac{1}{2L}\|\nabla f(w)-\nabla f(y)\|_2^2 \\ &= \langle \nabla f(y), y-w\rangle - \frac{1}{L}\|\nabla f(w)-\nabla f(y)\|_2^2 + \frac{1}{2L}\|\nabla f(w)-\nabla f(y)\|_2^2 \\ &= \langle \nabla f(y), y-w\rangle - \frac{1}{2L}\|\nabla f(w)-\nabla f(y)\|_2^2. \quad \square\end{aligned}$$