A survey on Large Scale Optimization

Raghav Somani

December 21, 2017

This article contains a summary and survey of the theoretical understandings of Large Scale Optimization by referring some talks, papers and lectures like [8, 1, 6, 5], that I have come across in the recent.

A very important aspect of Machine Learning is Optimization, therefore to have the best results one requires fast and scalable methods before one can appreciate a learning model. Such algorithms involve minimization of a functions $f(\mathbf{x})$. The set of its minimizers usually do not have a closed form solution, or even if they have, computing them is expensive in memory and computation time. Here is where iterative methods turn up to be easy and handy. Analyzing such algorithms involve mathematical analysis of both the function to optimize and the algorithm. Before jumping into some commonly used and theoretically promising algorithms, we first understand some basic concepts.

The below set of definitions might not appear well connected in the beginning, but are important and most of them will blend in when we will start analyzing different algorithms.

1 Definitions

1.1 Convex sets

In a Euclidean space, a convex set contains all the convex combinations of its points. That is, $\mathcal{C}$ is a convex set if for all $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathcal{C}$, $\sum_{i=1}^n \alpha_i \mathbf{x}_i \in \mathcal{C}\ \forall\ \alpha_i > 0$ and $\sum_{i=1}^n \alpha_i = 1$.

1.1.1 Properties

1. If $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_n$ are convex sets then $\bigcap\limits_{i=1}^{n} \mathcal{C}_i$ is also a convex set.

1.1.2 Examples

1. Convex hull and Convex Cone of a set $\mathcal{S}$ is convex.
2. Hyperplanes $\{\mathbf{x} \mid \mathbf{a}^T\mathbf{x} = b\}$ $\mathbf{a} \neq \mathbf{0}$, and Half-spaces $\{\mathbf{x} \mid \mathbf{a}^T\mathbf{x} \leq b\}$ $\mathbf{a} \neq \mathbf{0}$
3. Euclidean ball : $B(\mathbf{x}_c, r) = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{x}_c\| \leq r\} = \{\mathbf{x} + r\mathbf{u} \mid \|\mathbf{u}\| \leq 1\}$
4. Polyhedra : $\{\mathbf{x} \mid \mathbf{A}\mathbf{x} \preceq \mathbf{b}, \mathbf{C}\mathbf{x} = \mathbf{d}\}$

1.2 Convex functions

Let $\mathcal{C}$ be a convex set and $f : \mathcal{C} \to \mathbb{R}$. f is convex if

$$\forall\ \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C},\ \forall\ t \in [0,1] : \qquad f(t\mathbf{x}_1 + (1-t)\mathbf{x}_2) \leq tf(\mathbf{x}_1) + (1-t)f(\mathbf{x}_2)$$

1.2.1 Properties

1. Tangent at all points are under-estimators of the function. That is

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \qquad \forall\ \mathbf{x}, \mathbf{y} \in \mathcal{C}$$

2. If f is twice differentiable then $\nabla^2 f(\mathbf{x})\ \forall\ \mathbf{x} \in \mathcal{C}^{\mathrm{o}}$ is positive semidefinite.
3. All sub-level sets of f, $\{\mathbf{x} \mid f(\mathbf{x}) < a\}$ and $\{\mathbf{x} \mid f(\mathbf{x}) \leq a\}\ \forall\ a \in \mathbb{R}$, are convex sets. Whereas, the functions whose sub-level set are convex, are called Quasi-convex functions.

4. If f_i's are convex functions for $i \in [n]$, then $\max\limits_{1 \le i \le n} f_i$ is also a convex function.

5. Point-wise maximum : If $g(\mathbf{x}, \mathbf{y})$ is a convex function of $\mathbf{x} \ \forall \ \mathbf{y} \in \mathcal{Y}$, then $f(\mathbf{x}) = \max\limits_{\mathbf{y} \in \mathcal{Y}} g(\mathbf{x}, \mathbf{y})$ is convex for any arbitrary set $\mathcal{Y}$.

6. Non-negative weighted sum of convex functions is convex, i.e., if f_i's $\forall \ i \in [n]$ are n convex functions, then $\sum\limits_{i=1}^{n} \alpha_i f_i(\mathbf{x})$ is also a convex function for $\alpha_i \in \mathbb{R}_+ \ \forall \ i \in [n]$.

7. If f is a convex function, then so is $f(\mathbf{A}\mathbf{x} + \mathbf{b})$, where $\mathbf{A} \in \mathbb{R}^{d \times d}$ and $\mathbf{b} \in \mathbb{R}^d$.

1.2.2 Examples

1. Affine functions $f(\mathbf{x}) = \mathbf{a}^T \mathbf{x} + \mathbf{b}$, $\mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}$

2. Affine functions on matrices $f(\mathbf{X}) = tr(\mathbf{A}^T \mathbf{X}) + b, \mathbf{A} \in \mathbb{R}^{n \times d}, b \in \mathbb{R}$

3. $f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle$, $\mathbf{A} \in \mathbb{S}^n_+$.

4. $f(\mathbf{X}) = -\log \det(\mathbf{X})$, $\mathbf{X} \in \mathbb{S}^n_{++}$

5. All p-norms are convex.

6. Spectral norm : $f(\mathbf{X}) = \|\mathbf{X}\|_2 = (\lambda_{\max}(\mathbf{X}^T \mathbf{X}))^{\frac{1}{2}}$

7. Distance to a convex set $\mathcal{X}$, $\text{dist}(x, \mathcal{X}) = \inf\limits_{\mathbf{y} \in \mathcal{X}} \|\mathbf{x} - \mathbf{y}\|$

8. Indicator function of a convex set $\mathcal{X}$, $\mathbb{1}_{\mathcal{X}}(\mathbf{x})$ is a convex function.

$$\mathbb{1}_{\mathcal{X}}(\mathbf{x}) = \begin{cases} 0, & \text{if } x \in \mathcal{X} \\ \infty, & \text{otherwise} \end{cases}$$

1.3 Strongly Convex functions

If a function can be underestimated by some second order Taylor series expansion for all points in its domain, then the function is a strongly convex function. Formally,

$$f(\mathbf{y}) \ge f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 \qquad \mathbf{x}, \mathbf{y} \in \text{ dom } f$$

1.3.1 Properties

1. $f(\mathbf{x})$ is μ-strongly convex iff $g(\mathbf{x}) = f(\mathbf{x}) - \frac{m}{2} \|x\|_2^2$ is convex.

2. If f is twice differentiable, then $\nabla^2 f(x) \succeq \mu \mathbf{I}$.

3. If $\mathbf{x}^*$ is the minimizer of the function, minimizing the right hand side of the definition with respect to $\mathbf{y}$, we get

$$\begin{aligned} f(\mathbf{y}) &\ge f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 \\ &\ge \min_{\mathbf{y} \in \mathbb{R}^d} \left\{ f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 \right\} \\ &= f(\mathbf{x}) - \frac{1}{2\mu} \|\nabla f(\mathbf{x})\|_2^2 \\ \implies f(\mathbf{x}^*) &\ge f(\mathbf{x}) - \frac{1}{2\mu} \|\nabla f(\mathbf{x})\|_2^2 \quad \forall \ \mathbf{x} \in \mathbb{R}^d \end{aligned}$$

4. A function f is μ-strongly convex if and only if it is continuously differentiable and $\forall \ \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\alpha \in [0, 1]$ we have

$$f(\alpha \mathbf{x} + (1 - \alpha)\mathbf{y}) \le \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y}) - \alpha(1 - \alpha) \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|_2^2$$

Lemma 1.1. *If a convex function $f \in C_L^1$, then*

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \mu \left\| \mathbf{x} - \mathbf{y} \right\|_2^2$$

Proof. From strong convexity we have

$$\begin{aligned} f(\mathbf{x}) &\geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu}{2} \left\| \mathbf{x} - \mathbf{y} \right\|_2^2 \\ f(\mathbf{y}) &\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \left\| \mathbf{x} - \mathbf{y} \right\|_2^2 \end{aligned}$$

Adding the above two inequalities we get the result. □

1.4 Smooth functions

If a function can be overestimated by some second order Taylor series expansion for all points in its domain, then the function is a smooth function. Formally,

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \left\| \mathbf{y} - \mathbf{x} \right\|_2^2 \qquad \mathbf{x}, \mathbf{y} \in \text{ dom } f$$

1.4.1 Properties

1. If f is twice differentiable, then $\nabla^2 f(x) \preceq L\mathbf{I}$.
2. Lipschitz gradient implies smoothness.

$$\begin{aligned} g(t) &:= f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) \\ \therefore g(1) &= f(x), g(0) = f(y) \\ \text{and } \nabla g(t) &= \langle f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})), \mathbf{x} - \mathbf{y} \rangle \\ \because \int_0^1 \nabla g(t) dt &= g(1) - g(0) \end{aligned}$$

$$\begin{aligned} \text{Now, } |f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(y), \mathbf{x} - \mathbf{y} \rangle| &= |\int_0^1 \langle f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})), \mathbf{x} - \mathbf{y} \rangle \, dt - \langle \nabla f(y), \mathbf{x} - \mathbf{y} \rangle| \\ &\leq \int_0^1 |\langle f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})), \mathbf{x} - \mathbf{y} \rangle - \langle \nabla f(y), \mathbf{x} - \mathbf{y} \rangle| dt \\ &\leq \int_0^1 \left\| f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) - \nabla f(y) \right\|_2 \left\| \mathbf{x} - \mathbf{y} \right\|_2 dt \\ &\leq \left\| \mathbf{x} - \mathbf{y} \right\|_2 \int_0^1 Lt \left\| \mathbf{x} - \mathbf{y} \right\|_2 dt \\ &= \frac{L}{2} \left\| \mathbf{x} - \mathbf{y} \right\|_2 \end{aligned}$$

Lemma 1.2. *If a convex function $f \in C_L^1$, then*

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L} \left\| \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) \right\|_2$$

Proof. Using smoothness at both $\mathbf{x}$ and $\mathbf{y}$,

$$\begin{aligned} -\langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle &\geq f(\mathbf{y}) - f(\mathbf{x}) + \frac{L}{2} \left\| \mathbf{x} - \mathbf{y} \right\|_2^2 \\ -\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle &\geq f(\mathbf{x}) - f(\mathbf{y}) + \frac{L}{2} \left\| \mathbf{x} - \mathbf{y} \right\|_2^2 \\ \implies \langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle &\geq L \left\| \mathbf{x} - \mathbf{y} \right\|_2^2 \geq \frac{1}{L} \left\| \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) \right\|_2^2 \end{aligned}$$

□

1.5 Condition number

The ratio of smoothness and strong convexity is called the condition number (κ) of the function.

$$\kappa := \frac{L}{\mu}$$

If the function is twice differentiable, and if

$$\mu \mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{L} \ \forall \ \mathbf{x} \in \text{dom } f$$

then $\kappa = \frac{L}{\mu}$.

Lemma 1.3. *If $f \in S^1_{L,\mu}$, then for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we have,*

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{\mu L}{\mu + L} \|\mathbf{x} - \mathbf{y}\|_2^2 + \frac{1}{\mu + L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2$$

Proof. Consider a convex function, $\phi(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2} \|\mathbf{x}\|_2^2 \implies \nabla \phi(\mathbf{x}) = \nabla f(\mathbf{x}) - \mu \mathbf{x}$.
If $\mu = L$, the statement is easily true using strong convexity and Lemma 1.2.
If $\mu < L$, then $\phi \in C^1_{L-\mu}$ and we can invoke Lemma 1.2 to get the result. □

1.6 Fenchel Conjugate

Fenchel transformation is used to transform an optimization problem into its corresponding dual problem, which can often be simpler to solve. The Fenchel conjugate of f is

$$f^*(\mathbf{z}) = \sup_{\mathbf{x} \in \text{dom } f} \{\langle \mathbf{x}, \mathbf{z} \rangle - f(\mathbf{x})\}$$

For intuition in 1-D, for a function $f(x)$, given a slope y, we search for a point x that maximizes the separation between $g(x) := yx$ and $f(x)$. Once we have found the optimal x^*, we define a function with slope y, passing through $(x^*, f(x^*))$. The intercept of the function with the y-axis is $-f^*(y)$.

1.6.1 Properties

1. f^* is always convex.
2. Fenchel-Young inequality : $f(\mathbf{x}) + f^*(\mathbf{y}) \geq \langle \mathbf{x}, \mathbf{y} \rangle$.
3. In general, $f^{**}(\mathbf{x}) \leq f(\mathbf{x})$.
4. f is convex and lower semi-continuous $\iff f^{**}(\mathbf{x}) = f(\mathbf{x})$.
5. $f(\mathbf{x})$ is L-smooth $\iff f^*(\mathbf{x})$ is $\frac{1}{L}$-strongly convex.
6. $f(\mathbf{x})$ is μ-strongly convex $\iff f^*(\mathbf{x})$ is $\frac{1}{\sigma}$-smooth.

1.6.2 Examples

1. $f(\mathbf{x}) = \|x\| \implies f^*(\mathbf{z}) = \mathbb{1}_{\|\cdot\|_* \leq 1}(\mathbf{z})$. Where $\|\mathbf{z}\|_*$ is the operator norm of $\mathbf{z}^T$.
$\because \mathbf{z}^T\mathbf{x} \leq \|\mathbf{x}\| \|\mathbf{z}\|_*$.
2. $f(\mathbf{x}) = \mathbb{1}_{\mathcal{X}}(\mathbf{x}) \implies f^*(\mathbf{z}) = \sup_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{x}, \mathbf{z} \rangle$
3. $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} \implies f^*(\mathbf{y}) = \frac{1}{2}\mathbf{y}^T\mathbf{Q}^{-1}\mathbf{y} \ \forall \ \mathbf{Q} \in \mathbb{S}^n_{++}$

1.7 Sub-gradients

Sub-gradient is a generalization of gradients. $\mathbf{g}$ is a sub-gradient of f at $\mathbf{y}$ if

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \mathbf{g}, \mathbf{x} - \mathbf{y} \rangle$$

When a function is non-differentiable, we can have multiple such vector satisfying the above inequality, therefore the set of all sub-gradients is called the sub-gradient or sub-differential set. That is,

$$\partial f(\mathbf{y}) = \{\mathbf{g} \mid f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \mathbf{g}, \mathbf{x} - \mathbf{y} \rangle \ \forall \ \mathbf{x} \in \text{dom } f\}$$

1.7.1 Properties

1. $\partial f(\mathbf{x})$ is a closed convex set.
2. $\partial f(\mathbf{x})$ is non-empty when f is convex.
3. $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$ if f is differentiable at $\mathbf{x}$.
4. $\partial(\alpha f) = \alpha \partial f \ \forall \ \alpha > 0$.
5. $\partial(f_1 + f_2) \subset \partial f_1 + \partial f_2$.

1.7.2 Examples

1. $f(x) = |x|$, then $\partial f(0) = [-1, 1]$.
2. If $f(\mathbf{x}) = \max\limits_{1 \le i \le m} f_i(\mathbf{x}) \implies \partial f = \text{Conv} \cup \{\partial f_i \mid f_i(\mathbf{x}) = f(\mathbf{x})\}$.

1.8 Projection operator

Projection of a point $\mathbf{y}$ on to a set $\mathcal{X}$ is the closest point on the set.

$$\begin{aligned} P_{\mathcal{X}}(\mathbf{y}) &= \arg\min_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x} - \mathbf{y}\|_2^2 \\ &= \arg\min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{x} - \mathbf{y}\|_2^2 + \mathbb{1}_{\mathcal{X}}(\mathbf{x}) \end{aligned}$$

1.8.1 Properties

1. If $\mathcal{X}$ is closed and convex, projection is unique.
2. $\mathbf{x}^* = P(\mathbf{x})$ iff $\langle \mathbf{x}^* - \mathbf{x}, \mathbf{z} - \mathbf{x}^* \rangle \ge 0 \ \forall \ \mathbf{z} \in \mathcal{X}$
3. If $\mathcal{X}$ is closed and convex, projection is non-expansive, that is

$$\|P_{\mathcal{X}}(\mathbf{x}) - P_{\mathcal{X}}(\mathbf{y})\|^2 \le \|\mathbf{x} - \mathbf{y}\|^2 \ \ \forall \ \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$

1.9 Fermat's rule, Normal Cone and first order condition

$f : \mathbb{R}^n \to (-\infty, \infty]$,
Then, $\arg\min f = \text{zer}(\partial f) := \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{0} \in \partial f(\mathbf{x})\}$

$\therefore \min f(\mathbf{x})$ s.t. $\mathbf{x} \in \mathcal{X}$ becomes $\min f(\mathbf{x}) + \mathbb{1}_{\mathcal{X}}(\mathbf{x})$.
$\therefore$ from Fermat's rule, $\mathbf{0} \in \partial(f + \mathbb{1}_{\mathcal{X}})(\mathbf{x})$.
Or, $\mathbf{0} \in \partial f(\mathbf{x}) + \partial \mathbb{1}_{\mathcal{X}}(\mathbf{x})$. From the definition of sub-gradients, if $\mathbf{x} \in \mathcal{X}$, then $\mathbf{g} \in \partial \mathbb{1}_{\mathcal{X}}(\mathbf{x})$ iff $\mathbb{1}_{\mathcal{X}}(\mathbf{y}) \ge \mathbb{1}_{\mathcal{X}}(\mathbf{y}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \ \ \forall \ \mathbf{y} \in \mathbb{R}^n$.

That is, if $\mathbf{x} \in \mathcal{X}$ and $0 \ge \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle$, then $\mathbf{g} \in \partial \mathbb{1}_{\mathcal{X}}(\mathbf{x})$.
The normal cone of $\mathcal{X}$ at $\mathbf{x}$ is defined as

$$\mathcal{N}_{\mathcal{X}}(\mathbf{x}) := \{\mathbf{g} \in \mathbb{R}^n \mid 0 \ge \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \ \ \forall \ \mathbf{y} \in \mathcal{X}\}$$

$\therefore \min\limits_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$ reduces to finding $\mathbf{x}^* \ni \mathbf{0} \in \nabla f(\mathbf{x}^*) + \mathcal{N}_{\mathcal{X}}(\mathbf{x}^*)$.
Or, $-\nabla f(\mathbf{x}^*) \in \mathcal{N}_{\mathcal{X}}(\mathbf{x}^*) \implies \langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle \ge 0 \ \forall \ \mathbf{y} \in \mathcal{X}$

1.10 Lagrangian Dual problem

Let convex functions $f_i : \mathbb{R}^n \to \mathbb{R}(0 \le i \le m)$

$$\begin{gathered} \min f_0(\mathbf{x}) \\ \text{s.t. } \ f_i(\mathbf{x}) \le 0 \qquad 1 \le i \le m \\ \mathbf{x} \in \bigcup_{i=0}^{m} \text{dom } f_i \end{gathered}$$

For a primal problem, there is a Lagrangian associated with it where the constraints are brought up into the objective function. The Lagrangian is defined as

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) := f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x})$$

where λ_i's are non-negative are called Lagrange multipliers.
If $\mathbf{x}$ is feasible, then clearly $f_0(\mathbf{x}) \geq \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$. $\therefore \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ is a trivial lower-bound to the objective function.
Lagrange dual g as a function of the Lagrange multipliers is the worst such lower bound for the objective function.

$$g(\boldsymbol{\lambda}) := \inf_{\mathbf{x} \in \mathcal{X}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$$

Since the Lagrangian is a linear function in $\boldsymbol{\lambda}$, therefore the point-wise minimization over a family of such functions is concave. Therefore g is concave in $\boldsymbol{\lambda}$.
$\therefore f_0(\mathbf{x}) \geq g(\boldsymbol{\lambda}) \; \forall \; \mathbf{x}$ feasible and $\boldsymbol{\lambda} \in \mathbb{R}^m_+$.
$\therefore p^* := \min_{\mathbf{x}} f_0(\mathbf{x}) \geq g(\boldsymbol{\lambda}) \; \forall \; \boldsymbol{\lambda} \in \mathbb{R}^m_+$.

Dual problem is therefore defined as

$$\sup_{\boldsymbol{\lambda} \in \mathbb{R}^m_+} g(\boldsymbol{\lambda})$$

$\therefore p^* \geq d^* := \sup_{\boldsymbol{\lambda} \in \mathbb{R}^m_+} g(\boldsymbol{\lambda})$.

1.11 KKT Conditions

$$\begin{aligned} &\min f_0(\mathbf{x}) \\ \text{s.t. } &f_i(\mathbf{x}) \leq 0 \qquad 1 \leq i \leq m \\ &\mathbf{x} \in \bigcup_{i=0}^{m} \operatorname{dom} f_i \end{aligned}$$

If strong duality is attained, $\exists \; (\mathbf{x}^*, \boldsymbol{\lambda}^*) \ni$
$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*) \mid_{\mathbf{x}=\mathbf{x}^*} = \nabla f_0(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* f_i(\mathbf{x}^*) = \mathbf{0}$.
$\therefore \sum_{i=1}^{m} \lambda_i^* f_i(\mathbf{x}^*) = 0 \;\therefore\; \lambda_i^* f_i(\mathbf{x}^*) = 0$. If strong duality holds and $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ exists, then KKT conditions are necessary for $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ to be optimal.

If the problem is convex, KKT conditions are sufficient.

1. $f_i(\mathbf{x}^*) \leq 0 \qquad 1 \leq i \leq m$
2. $\lambda_i^* \geq 0 \qquad 1 \leq i \leq m$
3. $\lambda_i^* f_i(\mathbf{x}^*) = 0 \qquad 1 \leq i \leq m \qquad$ (Complementary slackness)
4. $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*) \mid_{\mathbf{x}=\mathbf{x}^*} = \mathbf{0} \qquad$ (Lagrangian stationary)

1.12 Stationary points

A stationary point is a point in the parameter space where the norm of the gradient vanishes. In optimization we define an ϵ-first order stationary point for which the norm of the gradient is at maximum ϵ. That is, $\mathbf{x}$ is an ϵ-first order stationary point of the function $f \in C^1$ if $\|\nabla f(\mathbf{x})\|_2 \leq \epsilon$.

2 Lower bounds on gradient based methods

Below are some famous lower bounds for optimization in the literature.

1. **Lipschitz-continuous :** If f is any L-Lipschitz continuous function, after t iterations, the error of any algorithm is $\Omega(t^{-\frac{1}{d}})$, where d is the dimension of the parameter space.

2. **Non-smooth :** Let A be a first order method starting from $\mathbf{x}_0 \in \mathbb{R}^d$ that has the access to first order non-stochastic oracle. Assume that the solution $\mathbf{x}^*$ to the minimization problem $\min_{\mathbf{x}} f(\mathbf{x})$ exists and $\|\mathbf{x}_0 - \mathbf{x}^*\|_2 \leq R$ and the function is L-Lipschitz on $\{\mathbf{x} \mid \|\mathbf{x}_0 - \mathbf{x}^*\|_2 \leq R\}$. Then for any $t, 0 \leq t \leq d-1$, there exists function f such that
$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \geq \frac{LR}{1+\sqrt{t+1}}$$
Note that sequence $\{\mathbf{x}_t\}$ satisfies $\mathbf{x}_{t+1} = \mathbf{x}_0 + \operatorname{span}(\mathbf{g}(\mathbf{x}_0), \mathbf{g}(\mathbf{x}_1), \ldots, \mathbf{g}(\mathbf{x}_t))$.

3. **Smooth :** Similar setup. For $0 \leq t \leq (d-1)/2$ and any $\mathbf{x}_0$, there exists a function f in the class of functions which is infinitely differentiable with a L-Lipschitz gradient such that any first order method satisfies
$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \geq \frac{3L\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{32(t+1)^2}$$

4. **Strongly convex and Smooth :** Let f be μ-strongly convex and have L-Lipschitz gradient, and be infinitely differentiable, then for any first order method, we have
$$\|\mathbf{x}_t - \mathbf{x}^*\|_2 \geq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2t} \|\mathbf{x}_0 - \mathbf{x}^*\|_2$$
$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \geq \frac{\mu}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2t} \|\mathbf{x}_0 - \mathbf{x}^*\|_2$$

3 Sub-gradient method

We can solve convex optimization problem in polynomial time by interior point methods. But these solvers require $O\left(d^2\right)$ or worse cost per iteration which is practically infeasible when d is large. A greedy, cheap and a locally optimal way to decrease a convex function's value is to iteratively move in a negative sub-gradient direction. Algorithmically,
$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \mathbf{g}_t$$
$\mathbf{g}_t \in \partial f(\mathbf{x}_t)$, and $\eta_t > 0$ is the step length. Here each per iteration cost is just $O(d)$ which makes these methods feasible in practice.
Assumptions : f is L-Lipschitz, therefore $\|g_t\|_2 \leq G := L$, and domain is bounded, i.e., $\|\mathbf{x}_0 - x^*\|_2 \leq R$.

3.1 General convex functions

In general, when a convex function is non-smooth and non-strongly convex, we can guarantee some convergence rates associated with sub-gradient descent.

3.1.1 Convergence

We consider our Lyapunov function to be squared Euclidean distance from $\mathbf{x}^*$ and not the difference in function value to the optimal,
$$\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 &= \|\mathbf{x}_t - \eta_t\mathbf{g}_t - \mathbf{x}^*\|_2^2 \\
&= \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + \eta_t^2\|\mathbf{g}_t\|_2^2 - 2\langle \eta_t\mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^*\rangle \\
&\leq \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + \eta_t^2\|\mathbf{g}_t\|_2^2 - 2\eta_t(f(\mathbf{x}_t) - f(\mathbf{x}^*)) \qquad (\because f(\mathbf{x}^*) \geq f(\mathbf{x}_t) + \langle \mathbf{g}_t, \mathbf{x}^* - \mathbf{x}_t\rangle)
\end{aligned} \tag{3.1.1}$$
Telescoping (3.1.1) from $t=1$ to T, we get
$$\|\mathbf{x}_{T+1} - \mathbf{x}^*\|_2^2 \leq \|\mathbf{x}_1 - \mathbf{x}^*\|_2^2 + \sum_{t=1}^T \eta_t^2\|\mathbf{g}_t\|_2^2 - 2\sum_{t=1}^T \eta_t(f(\mathbf{x}_t) - f(\mathbf{x}^*))$$
$$\implies 2\sum_{t=1}^T \eta_t(f(\mathbf{x}_t) - f(\mathbf{x}^*)) \leq \|\mathbf{x}_1 - \mathbf{x}^*\|_2^2 + \sum_{t=1}^T \eta_t^2\|\mathbf{g}_t\|_2^2 - \|\mathbf{x}_{T+1} - \mathbf{x}^*\|_2^2$$

$$
\begin{aligned}
&\leq \|\mathbf{x}_1 - \mathbf{x}^*\|_2^2 + \sum_{t=1}^{T} \eta_t^2 \|\mathbf{g}_t\|_2^2 \\
&\leq R^2 + G^2 \sum_{t=1}^{T} \eta_t^2 \\
\therefore 2 \min_{1\leq t\leq T}(f(\mathbf{x}_t) - f(\mathbf{x}^*)) \sum_{t=1}^{T} \eta_t &\leq 2 \sum_{t=1}^{T} \eta_t (f(\mathbf{x}_t) - f(\mathbf{x}^*)) \\
&\leq R^2 + G^2 \sum_{t=1}^{T} \eta_t^2 \\
\implies \epsilon_t := \min_{1\leq t\leq T}(f(\mathbf{x}_t) - f(\mathbf{x}^*)) &\leq \frac{R^2 + G^2 \sum\limits_{t=1}^{T} \eta_t^2}{2 \sum\limits_{t=1}^{T} \eta_t}
\end{aligned}
$$

Now we can choose different step-sizes to see how the convergence is affected

1. **Constant :** If $\eta_t = \eta$ $\quad \epsilon_t \leq \frac{R^2 + G^2 T \eta^2}{2T\eta} \to \frac{G^2\eta}{2}$ as $T \to \infty$.
2. **Square summable but not summable :** $\sum\limits_{t=1}^{\infty} \eta_t^2 < \infty$ and $\sum\limits_{t=1}^{\infty} \eta_t = \infty$.
   For fixed t, the best possible step-size is a constant $\eta_t = \frac{R}{G\sqrt{t}}$, then after T steps, $\epsilon_t \leq \frac{RG}{\sqrt{T}}$. Therefore for ϵ accuracy in function value, we need at least $(\frac{RG}{\epsilon})^2 = O(\frac{1}{\epsilon^2})$ steps.

3.2 Strongly convex functions

If a function is μ-strongly convex, we can use this information to modify the convergence analysis to get a better convergence rate.

3.2.1 Convergence

We consider our Lyapunov function to be squared Euclidean distance from $\mathbf{x}^*$ and not the difference in function value to the optimal,

$$
\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 &= \|\mathbf{x}_t - \eta_t \mathbf{g}_t - \mathbf{x}^*\|_2^2 \\
&= \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + \eta_t^2 \|\mathbf{g}_t\|_2^2 - 2\langle \eta_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^* \rangle \\
&\leq \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + \eta_t^2 \|\mathbf{g}_t\|_2^2 - 2\eta_t (f(\mathbf{x}_t) - f(\mathbf{x}^*)) - 2\eta_t \frac{\mu}{2} \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 \quad \text{(using strong convexity at } \mathbf{x}_t\text{)} \\
&= (1 - \eta_t \mu) \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + \eta_t^2 \|\mathbf{g}_t\|_2^2 - 2\eta_t (f(\mathbf{x}_t) - f(\mathbf{x}^*)) \qquad (3.2.1)
\end{aligned}
$$

Setting $\eta_t = \frac{1}{\mu t}$, equation (3.2.1) becomes

$$
\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 &\leq \frac{t-1}{t} \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + \frac{G^2}{\mu^2 t^2} - \frac{2}{\mu t}(f(\mathbf{x}_t) - f(\mathbf{x}^*)) \\
\implies t \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 &\leq (t-1) \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + \frac{G^2}{\mu^2 t} - \frac{2}{\mu}(f(\mathbf{x}_t) - f(\mathbf{x}^*)) \qquad (3.2.2)
\end{aligned}
$$

Telescoping (3.2.2) from $t = 1$ to T, we get

$$
\begin{aligned}
T \|\mathbf{x}_{T+1} - \mathbf{x}^*\|_2^2 &\leq -\frac{2}{\mu} \sum_{t=1}^{T} (f(\mathbf{x}_t) - f(\mathbf{x}^*)) + \frac{G^2}{\mu^2} \sum_{t=1}^{T} \frac{1}{t} \\
\implies \frac{2}{\mu} \sum_{t=1}^{T} (f(\mathbf{x}_t) - f(\mathbf{x}^*)) &\leq -T \|\mathbf{x}_{T+1} - \mathbf{x}^*\|_2^2 + \frac{G^2}{\mu^2} \sum_{t=1}^{T} \frac{1}{t} \\
&\leq \frac{G^2}{\mu^2} \sum_{t=1}^{T} \frac{1}{t}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{G^2}{\mu^2} O(\log T) \\
\because \frac{2T}{\mu} \min_{1\le t\le T} (f(\mathbf{x}_t) - f(\mathbf{x}^*)) &\le \frac{2}{\mu} \sum_{t=1}^{T} (f(\mathbf{x}_t) - f(\mathbf{x}^*)) \\
&\le \frac{G^2}{\mu^2} O(\log T) \\
\implies \epsilon_t := \min_{1\le t\le T} (f(\mathbf{x}_t) - f(\mathbf{x}^*)) &\le \frac{G^2}{2\mu} O\left(\frac{\log T}{T}\right)
\end{aligned}
$$

3.3 Smooth functions

If a convex function has L-Lipschitz gradient, we can use this information to modify the convergence analysis to get a better convergence rate than that for general convex functions.

3.3.1 Convergence

$$
\begin{aligned}
\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 &\le L \|\mathbf{x} - \mathbf{y}\|_2 \\
\implies f(\mathbf{x}) \le f(y) + \langle \nabla f(y), \mathbf{x} - \mathbf{y} \rangle &+ \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|_2^2
\end{aligned}
$$

Considering the squared Euclidean distance from the optimal,

$$
\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 &= \|\mathbf{x}_t - \eta_t \mathbf{g}_t - \mathbf{x}^*\|_2^2 \\
&= \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + \eta_t^2 \|\mathbf{g}_t\|_2^2 - 2\eta_t \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^* \rangle \\
&= \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + \eta_t^2 \|\mathbf{g}_t\|_2^2 - 2\eta_t \langle \mathbf{g}_t - \mathbf{g}^*, \mathbf{x}_t - \mathbf{x}^* \rangle \quad (\text{where } \mathbf{g}^* \in \partial f(\mathbf{x}^*)) \\
&\le \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + \eta_t^2 \|\mathbf{g}_t\|_2^2 - \frac{2\eta_t}{L} \|\mathbf{g}_t - \mathbf{g}^*\|_2^2 \quad (\text{Using Lemma 1.2}) \\
&\le \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + \eta_t^2 \|\mathbf{g}_t\|_2^2 - \frac{2\eta_t}{L} \|\mathbf{g}_t\|_2^2 \\
&= \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \eta_t \left(\frac{2}{L} - \eta_t\right) \|\mathbf{g}_t\|_2^2
\end{aligned}
$$

Therefore for $\eta_t < \frac{2}{L}$, the distance from the optimal decreases monotonically.
Using smoothness for two consecutive iterates,

$$
\begin{aligned}
f(\mathbf{x}_{t+1}) &\le f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2 \\
&= f(\mathbf{x}_t) - \eta_t \|\nabla f(\mathbf{x}_t)\|_2^2 + \frac{\eta_t^2 L}{2} \|\nabla f(\mathbf{x}_t)\|_2^2 \\
&= f(\mathbf{x}_t) - \eta_t \left(1 - \frac{\eta_t L}{2}\right) \|\nabla f(\mathbf{x}_t)\|_2^2 \qquad (3.3.1)
\end{aligned}
$$

Therefore again, for $\eta_t < \frac{2}{L}$ we have descent but we will choose $\eta_t = \frac{1}{L}$ as it is the minimizer.
We define $\Delta_t := f(\mathbf{x}_t) - f(\mathbf{x}^*)$

$$
\begin{aligned}
f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) &\ge \frac{1}{2L} \|\nabla f(\mathbf{x}_t)\|_2^2 \\
\implies \Delta_{t+1} &\le \Delta_t - \frac{1}{2L} \|\nabla f(\mathbf{x}_t)\|_2^2 \qquad (3.3.2)
\end{aligned}
$$

From convexity,

$$
\begin{aligned}
\Delta_t = f(\mathbf{x}_t) - f(\mathbf{x}^*) \le \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle &\le \|\nabla f(\mathbf{x}_t)\|_2 \|\mathbf{x}_t - \mathbf{x}^*\|_2 \\
\implies -\|\nabla f(\mathbf{x}_t)\|_2^2 &\le \frac{-\Delta_t^2}{\|\mathbf{x}_t - \mathbf{x}^*\|_2^2} \qquad (3.3.3)
\end{aligned}
$$

Plugging in equation (3.3.3) in equation (3.3.2), we get

$$
\Delta_{t+1} \le \Delta_t \left(1 - \frac{\Delta_t}{2L \|\mathbf{x}_t - \mathbf{x}^*\|_2^2}\right) \le \Delta_t \left(1 - \frac{\Delta_t}{2L \|\mathbf{x}_1 - \mathbf{x}^*\|_2^2}\right) \qquad (3.3.4)
$$

Since $\frac{\Delta_t}{2L\|\mathbf{x}_t-\mathbf{x}^*\|_2} < 1$,

$$\frac{1}{\Delta_{t+1}} \geq \frac{1}{\Delta_t} \frac{1}{\left(1 - \frac{\Delta_t}{2L\|\mathbf{x}_1-\mathbf{x}^*\|_2^2}\right)} \geq \frac{1}{\Delta_t}\left(1 + \frac{\Delta_t}{2L\left\|\mathbf{x}_1 - \mathbf{x}^*\right\|_2^2}\right)$$

$$\implies \frac{1}{\Delta_{t+1}} \geq \frac{1}{\Delta_t} + \frac{1}{2L\left\|\mathbf{x}_1 - \mathbf{x}^*\right\|_2^2} \tag{3.3.5}$$

Telescoping equation (3.3.5) from $t = 1$ to T, we get

$$\frac{1}{\Delta_T} \geq \frac{1}{\Delta_1} + \frac{T}{2L\left\|\mathbf{x}_1 - \mathbf{x}^*\right\|_2^2} \tag{3.3.6}$$

Re-arranging (3.3.6), we get

$$\Delta_T \leq \frac{2\Delta_1 L\left\|\mathbf{x}_1 - \mathbf{x}^*\right\|_2^2}{T\Delta_1 + 2L\left\|\mathbf{x}_1 - \mathbf{x}^*\right\|_2^2} \tag{3.3.7}$$

Using smoothness at $\mathbf{x}^*$ we have

$$\Delta_1 \leq \frac{L}{2}\left\|\mathbf{x}_1 - \mathbf{x}^*\right\|_2^2 \implies 4\Delta_1 \leq 2L\left\|\mathbf{x}_1 - \mathbf{x}^*\right\|_2^2 \tag{3.3.8}$$

Plugging equation (3.3.8) in equation (3.3.7) we get

$$\Delta_T \leq \frac{2\Delta_1 L\left\|\mathbf{x}_1 - \mathbf{x}^*\right\|_2^2}{(T+4)\Delta_1} = \frac{2L\left\|\mathbf{x}_1 - \mathbf{x}^*\right\|_2^2}{T+4} = O\left(\frac{1}{T}\right)$$

3.4 Smooth and Strongly convex functions

If a convex function is μ-strongly convex as well as if its gradient is L-Lipschitz, we have geometric rates of convergence.

3.4.1 Convergence

Considering the squared Euclidean distance from the optimal,

$$\begin{aligned}
\left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|_2^2 &= \left\|\mathbf{x}_t - \eta_t \mathbf{g}_t - \mathbf{x}^*\right\|_2^2 \\
&= \left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2 + \eta_t^2\left\|\mathbf{g}_t\right\|_2^2 - 2\eta_t\left\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^*\right\rangle \\
&= \left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2 + \eta_t^2\left\|\mathbf{g}_t\right\|_2^2 - 2\eta_t\left\langle \mathbf{g}_t - \mathbf{g}^*, \mathbf{x}_t - \mathbf{x}^*\right\rangle \\
&\leq \left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2 + \eta_t^2\left\|\mathbf{g}_t\right\|_2^2 - 2\eta_t\left(\frac{\mu L}{\mu + L}\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2 + \frac{1}{\mu + L}\left\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}^*)\right\|_2^2\right) \\
&\qquad\qquad \text{(Using Lemma 1.3)} \\
&= \left(1 - \frac{2\eta_t \mu L}{\mu + L}\right)\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2 + +\eta_t\left(\eta_t - \frac{2}{\mu + L}\right)\left\|\mathbf{g}_t\right\|_2^2
\end{aligned}$$

Therefore when $\eta_t < \frac{2}{\mu+L}$, we have a decrease. Therefore for $0 < \eta_t \leq \frac{2}{\mu+L}$, we get

$$\left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|_2^2 \leq \left(1 - \frac{2\eta_t \mu L}{\mu + L}\right)\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2$$

To get the maximum decrease, we set $\eta_t = \frac{2}{\mu+L}$. Therefore,

$$\left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|_2^2 \leq \left(\frac{L-\mu}{L+\mu}\right)^2\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2 \implies \left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|_2^2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^2\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2 \tag{3.4.1}$$

Recursing (3.4.1) we get a geometric convergence rate in parameter space.

$$\left\|\mathbf{x}_{T+1} - \mathbf{x}^*\right\|_2^2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^{2T}\left\|\mathbf{x}_1 - \mathbf{x}^*\right\|_2^2 \tag{3.4.2}$$

Using smoothness at $\mathbf{x}^*$, i.e., $f(\mathbf{x}_{T+1}) - f(\mathbf{x}^*) \leq \frac{L}{2}\left\|\mathbf{x}_{T+1} - \mathbf{x}^*\right\|_2$, we get

$$f(\mathbf{x}_{T+1}) - f(\mathbf{x}^*) \leq \frac{L}{2}\left(\frac{\kappa - 1}{\kappa + 1}\right)^{2T}\left\|\mathbf{x}_1 - \mathbf{x}^*\right\|_2^2 \tag{3.4.3}$$

4 Projected Sub-gradient method

When the feasible set of parameters is constrained, we need to project the iterate to the feasible set.

$$\begin{aligned}\mathbf{z}_{t+1} &= \mathbf{x}_t - \eta_t \mathbf{g}_t \\ \mathbf{x}_{t+1} &= P_{\mathcal{X}}(\mathbf{z}_{t+1})\end{aligned}$$

As we have seen in section 1.8, projection on to closed and convex set is non-expansive. That is

$$\|P_{\mathcal{X}}(\mathbf{x}) - P_{\mathcal{X}}(\mathbf{y})\|_2^2 \leq \|\mathbf{x} - \mathbf{y}\|_2^2 \ \forall \ \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$

where $\mathcal{X}$ is a closed and convex set. Therefore to analyze its convergence, we just need to modify (3.1.1) as

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 &= \|P_{\mathcal{X}}(\mathbf{x}_t - \eta_t \mathbf{g}_t) - \mathbf{x}^*\|_2^2 \\ &\leq \|\mathbf{x}_t - \eta_t \mathbf{g}_t - \mathbf{x}^*\|_2^2 \\ &= \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 + \eta_t^2 \|\mathbf{g}_t\|_2^2 - 2\langle \eta_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^* \rangle\end{aligned}$$

And the analysis proceeds in a similar way with similar convergence rates.
Some commonly used projection operations and their closed form solution are listed below

1. **Non-negativity :** $\mathcal{X} = \{\mathbf{x} \mid x_i \geq 0 \ \forall \ i \in [d]\} \implies P_{\mathcal{X}}(\mathbf{z}) = [\mathbf{z}]_+$
2. l_∞ **ball :** $\mathcal{X} = \{\mathbf{x} \mid \|\mathbf{x}\|_\infty \leq 1\}, \because P_{\mathcal{X}}(\mathbf{z}) = \arg\min_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x} - \mathbf{z}\|_2^2$, this minimization is coordinate separable and $P_{\mathcal{X}}(\mathbf{z}) = \mathbf{y}$ where $y_i = \text{sign}(z_i)\min\{|z_i|, 1\}$.
3. **Linear Equality constraint :** $\mathcal{X} = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}\}$, $\mathbf{A} \in \mathbb{R}^{n \times d}$ has rank n.

$$\begin{aligned}\implies P_{\mathcal{X}}(\mathbf{z}) &= \mathbf{z} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{z} - \mathbf{b}) \\ &= (\mathbf{I} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A})\mathbf{z} + \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{b}\end{aligned}$$

For the update step, using $\mathbf{A}\mathbf{x}_t = \mathbf{b}$,

$$\begin{aligned}\mathbf{x}_{t+1} &= P_{\mathcal{X}}(\mathbf{x}_t - \eta_t \mathbf{g}_t) \\ &= \mathbf{x}_t - \eta_t(\mathbf{I} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A})\mathbf{g}_t\end{aligned}$$

5 Proximal Gradient Descent

Suppose we have a composite objective function of the form

$$f(\mathbf{x}) = l(\mathbf{x}) + r(\mathbf{x})$$

Consider examples like $l(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ and $r(\mathbf{x}) = \lambda \|\mathbf{x}\|_1$. $r(\mathbf{x})$ in this problem is non-smooth which makes $f(\mathbf{x})$ also a non-smooth objective. Therefore according to the lower bounds we saw in Section 2, we cannot achieve a better rate than $O(\frac{1}{\sqrt{T}})$. Therefore any algorithm which takes just the sub-gradient information cannot lead to anything better. What we know about such objective is that it is a sum of a smooth and a non-smooth function. This fact can be exploited by the Proximal Gradient method.

For projected gradient descent we have

$$\begin{aligned}&\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \\ \mathbf{x}_{t+1} &= P_{\mathcal{X}}(\mathbf{x}_t - \eta_t \mathbf{g}_t)\end{aligned}$$

For proximal gradient descent we have

$$\begin{aligned}&\min f(\mathbf{x}) + h(\mathbf{x}) \\ \mathbf{x}_{t+1} &= \text{prox}_{\eta h}\left(\mathbf{x}_t - \eta \nabla f(\mathbf{x})\right)\end{aligned}$$

Here $\text{prox}_{\eta h}(.)$ denotes the Euclidean proximity operator for h.

5.1 Proximity operator

The projection operator as we had defined in Section 1.8 is

$$P_{\mathcal{X}}(\mathbf{y}) = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{x}-\mathbf{y}\|_2^2 + \mathbb{1}_{\mathcal{X}}(\mathbf{x})$$

For defining the proximal gradient we just replace the indicator function with the non-smooth component of the objective, i.e.,

$$\operatorname{prox}_h(\mathbf{y}) = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{x}-\mathbf{y}\|_2^2 + h(\mathbf{x})$$

5.1.1 Examples :

For Lasso linear regression,

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

We can split the problem into each coordinate and have d sub-problems of the form

$$\min_{\mathbb{R}} \frac{1}{2}(x-y)^2 + \lambda x$$

The operator that maps x to the minimizer is called the *soft-thresholding* operator which is

$$\operatorname{soft}(x,\lambda) := \operatorname{sign}(x)(|x|-\lambda)_+$$

5.1.2 Understanding the operator

From Fermat's rule in Section 1.9, we have

$$\begin{aligned}
\mathbf{0} &\in \nabla f(\mathbf{x}^*) + \partial h(\mathbf{x}^*)\\
\mathbf{0} &\in \eta\nabla f(\mathbf{x}^*) + \eta\partial h(\mathbf{x}^*)\\
\mathbf{x}^* &\in \eta\nabla f(\mathbf{x}^*) + (\mathbf{I}+\eta\partial h)(\mathbf{x}^*)\\
\mathbf{x}^* - \eta\nabla f(\mathbf{x}^*) &\in (\mathbf{I}+\eta\partial h)(\mathbf{x}^*)\\
\mathbf{x}^* &= (\mathbf{I}+\eta\partial h)^{-1}(\mathbf{x}^* - \eta\nabla f(\mathbf{x}^*))
\end{aligned}$$

Defining the operator $(\mathbf{I}+\eta\partial h)^{-1}$ as $\operatorname{prox}_{\eta h}(.)$ we obtain the fixed point iteration as

$$\mathbf{x}_{t+1} = \operatorname{prox}_{\eta_t h}(\mathbf{x}_t - \eta_t\nabla f(\mathbf{x}_t))$$

Therefore if $G_\eta(\mathbf{x}) := \frac{1}{\alpha}(\mathbf{x} - \operatorname{prox}_{\eta h}(\mathbf{x}-\eta\nabla f(\mathbf{x})))$, we get an equality similar to the gradient descent step,

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t G_{\eta_t}(\mathbf{x}_t)$$

This gradient descent like mapping is called *gradient mapping*.

5.2 Convergence

For $f \in C_L^1$, let $\mathbf{y} = \mathbf{x} - \eta G_\eta \mathbf{x}$, then $\forall\ \mathbf{z}$ we have

$$\begin{aligned}
f(\mathbf{y}) &\le f(\mathbf{x}) + \langle\nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|_2^2\\
&= f(\mathbf{x}) - \eta\langle\nabla f(\mathbf{x}), G_\eta(\mathbf{x})\rangle + \frac{\eta^2 L}{2}\|G_\eta(\mathbf{x})\|_2\\
&\le f(\mathbf{x}) - \eta\langle\nabla f(\mathbf{x}), G_\eta(\mathbf{x})\rangle + \frac{\eta}{2}\|G_\eta(\mathbf{x})\|_2^2 \qquad (\text{if } 0\le\eta\le\frac{1}{L})
\end{aligned} \tag{5.2.1}$$

From convexity of f, we have

$$f(\mathbf{x}) \le f(\mathbf{z}) - \langle\nabla f(\mathbf{x}), \mathbf{z}-\mathbf{x}\rangle \qquad \forall\ \mathbf{z}\in\mathbb{R}^d \tag{5.2.2}$$

Adding Equation (5.2.1) and (5.2.2) we get

$$f(\mathbf{y}) \le f(\mathbf{z}) - \langle\nabla f(\mathbf{x}), \mathbf{z}-\mathbf{x}\rangle - \langle\nabla f(\mathbf{x}), \eta G_\eta(\mathbf{x})\rangle + \frac{\eta}{2}\|G_\eta(\mathbf{x})\|_2^2 \tag{5.2.3}$$

From convexity of h, we have

$$h(\mathbf{y}) \leq h(\mathbf{z}) - \langle G_\eta(\mathbf{x}) - \nabla f(\mathbf{x}), \mathbf{z} - \mathbf{y} \rangle \tag{5.2.4}$$

Adding equation (5.2.3) and (5.2.4) we get

$$\begin{aligned} f(\mathbf{y}) + h(\mathbf{y}) &\leq f(\mathbf{z}) + h(\mathbf{z}) + \langle G_\eta(\mathbf{x}), \mathbf{x} - \mathbf{z} \rangle + \langle G_\eta(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\eta}{2} \left\| G_\eta(\mathbf{x}) \right\|_2^2 \\ &= f(\mathbf{z}) + h(\mathbf{z}) + \langle G_\eta(\mathbf{x}), \mathbf{x} - \mathbf{z} \rangle - \frac{\eta}{2} \left\| G_\eta(\mathbf{x}) \right\|_2^2 \end{aligned} \tag{5.2.5}$$

The above inequality with $\phi = f + h$, $\mathbf{y} = \mathbf{x}_{t+1}$ and $\mathbf{x} = \mathbf{x}_t$ shows that it is a descent method.

$$\phi(\mathbf{x}_{t+1}) \leq \phi(\mathbf{x}_t) - \frac{\eta}{2} \left\| G_\eta(\mathbf{x}_t) \right\|_2^2$$

With $\mathbf{z} = \mathbf{x}^*$ in Equation (5.2.5) we can start analyzing the convergence of Proximal Gradient method for smooth functions.

$$\begin{aligned} \phi(\mathbf{x}_{t+1}) - \phi(\mathbf{x}^*) &\leq \langle G_\eta(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle - \frac{\eta}{2} \left\| G_\eta(\mathbf{x}_t) \right\|_2^2 \\ &= \frac{1}{2\eta} \left[\langle 2\eta G_\eta(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle - \left\| \eta G_\eta(\mathbf{x}_t) \right\|_2^2 \right] \\ &= \frac{1}{2\eta} \left[\left\| \mathbf{x}_t - \mathbf{x}^* \right\|_2^2 - \left\| \mathbf{x}_{t+1} - \mathbf{x}^* \right\|_2^2 \right] \end{aligned} \tag{5.2.6}$$

Summing Equation (5.2.6) from $t = 1$ to T, and setting $\eta = \frac{1}{L}$ we get

$$\begin{aligned} T(\phi(\mathbf{x}_T) - \phi(\mathbf{x}^*)) \leq \sum_{t=1}^{T} (\phi(\mathbf{x}_t) - \phi(\mathbf{x}^*)) &\leq \frac{L}{2} \sum_{t=1}^{T} (\left\| \mathbf{x}_t - \mathbf{x}^* \right\|_2^2 - \left\| \mathbf{x}_{t+1} - \mathbf{x}^* \right\|_2^2) \\ &= \frac{L}{2} \left[\left\| \mathbf{x}_1 - \mathbf{x}^* \right\|_2^2 - \left\| \mathbf{x}_{T+1} - \mathbf{x}^* \right\|_2^2 \right] \\ &\leq \frac{L}{2} \left\| \mathbf{x}_1 - \mathbf{x}^* \right\|_2^2 \\ \implies \phi(\mathbf{x}_T) - \phi(\mathbf{x}^*) &\leq \frac{L}{2T} \left\| \mathbf{x}_1 - \mathbf{x}^* \right\|_2^2 \end{aligned}$$

Therefore when ϕ is not a completely smooth function, but a sum of smooth and non-smooth function, we can still achieve the known $O(\frac{1}{T})$ sub-optimality rate.

6 Stochastic Gradient Descent

When our data is large scale, computing the exact gradient turns out to be very expensive. Stochastic optimization methods make it possible to reduce the computational complexity of each iterative step and still provide good optimization and generalization rates. Instead of computing the exact gradient, these methods have access to the noisy stochastic first order oracle.
If the function f can be decomposed into an empirical mean of n functions $\{f_i, i = 1, 2, \dots, n\}$, i.e.,

$$f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x})$$

Therefore we have a trivial unbiased estimator for sub-gradients, i.e., $\partial f_i(\mathbf{x})$ for i sampled from $U([n])$
Therefore,

$$g(\mathbf{x}) := \mathbb{E}\left[\partial f_i(\mathbf{x})\right] \qquad \text{s.t. } g(\mathbf{x}) \in \partial f(\mathbf{x})$$

Considering the objective function as

$$\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$$

We get an iterative update rule as

$$\mathbf{x}_{t+1} = P_{\mathcal{X}}(\mathbf{x}_t - \eta_t \mathbf{g}_t)$$

where $\mathbf{g}_t := \partial f_i(\mathbf{x}_t)$ is a random variable for $i \sim U([n])$.
To note, $\mathbf{x}_t$ depends on random variables, $i_1, i_2, \dots, i_{t-1}$ all sampled independently from $U([n])$.

6.1 General convex functions

f is a general convex function with bounded gradient, i.e., $\|\mathbf{g}_t\|_2 \leq G$ and with finite domain, i.e., $\|\mathbf{x}-\mathbf{x}^*\|_2 \leq R \; \forall \; \mathbf{x} \in \mathcal{X}$.

6.1.1 Convergence

Define $R_t := \|\mathbf{x}_t - \mathbf{x}^*\|_2^2$ and $r_t := \mathbb{E}\left[R_t\right] = \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2\right]$.

$$\begin{aligned}
R_{t+1} &= \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 \\
&= \|P_{\mathcal{X}}(\mathbf{x}_t - \eta_t \mathbf{g}_t) - P_{\mathcal{X}}(\mathbf{x}^*)\|_2^2 \\
&\leq \|\mathbf{x}_t - \mathbf{x}^* - \eta_t \mathbf{g}_t\|_2^2 \\
&= R_t + \eta_t^2 \|\mathbf{g}_t\|_2^2 - 2\eta_t \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^* \rangle
\end{aligned}$$

Taking expectations and using the bound on $\|\mathbf{g}_t\|_2^2$, we get

$$r_{t+1} \leq r_t + \eta_t^2 G^2 - 2\eta_t \mathbb{E}\left[\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^* \rangle\right] \tag{6.1.1}$$

Using the fact that $\mathbf{x}_t$ is dependent only on $i_1, i_2, \ldots, i_{t-1}$

$$\begin{aligned}
\mathbb{E}\left[\langle \mathbf{x}_t - \mathbf{x}^*, \mathbf{g}_t \rangle\right] &= \mathbb{E}\left[\mathbb{E}\left[\langle \mathbf{x}_t - \mathbf{x}^*, \mathbf{g}_t \rangle \mid i_1, i_2, \ldots, i_{t-1}\right]\right] \\
&= \mathbb{E}\left[\langle \mathbf{x}_t - \mathbf{x}^*, \mathbb{E}\left[\mathbf{g}_t \mid i_1, i_2, \ldots, i_{t-1}\right] \rangle\right] \\
&= \mathbb{E}\left[\langle \mathbf{x}_t - \mathbf{x}^*, g(\mathbf{x}_t) \rangle\right] \qquad g(\mathbf{x}_t) \in \partial f(\mathbf{x}_t)
\end{aligned} \tag{6.1.2}$$

Plugging Equation (6.1.2) in Equation (6.1.1), we get

$$r_{t+1} \leq r_t + \eta_t^2 G^2 - 2\eta_t \mathbb{E}\left[\langle \mathbf{x}_t - \mathbf{x}^*, g(\mathbf{x}_t) \rangle\right]$$

Because f is convex, we have

$$\begin{aligned}
f(\mathbf{x}^*) &\geq f(\mathbf{x}_t) + \langle g(\mathbf{x}_t), \mathbf{x}^* - \mathbf{x}_t \rangle \\
\langle g(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle &\geq f(\mathbf{x}_t) - f(\mathbf{x}^*) \\
\implies -2\eta_t \mathbb{E}\left[\langle \mathbf{x}_t - \mathbf{x}^*, g(\mathbf{x}_t) \rangle\right] &\leq -2\eta_t (f(\mathbf{x}_t) - f(\mathbf{x}^*))
\end{aligned} \tag{6.1.3}$$

Plugging Equation (6.1.3) in Equation (6.1.1), we get

$$\begin{aligned}
r_{t+1} &\leq r_t + \eta_t^2 G^2 - 2\eta_t (f(\mathbf{x}_t) - f(\mathbf{x}^*)) \\
\implies 2\eta_t (f(\mathbf{x}_t) - f(\mathbf{x}^*)) &\leq r_t - r_{t+1} + \eta_t^2 G^2
\end{aligned} \tag{6.1.4}$$

Telescoping Equation (6.1.4) from $t = 1$ to T we obtain

$$\begin{aligned}
\sum_{t=1}^{T} 2\eta_t (f(\mathbf{x}_t) - f(\mathbf{x}^*)) &\leq r_1 - r_T + G^2 \sum_{t=1}^{T} \eta_t^2 \\
&\leq r_1 + G^2 \sum_{t=1}^{T} \eta_t^2
\end{aligned}$$

Defining $\gamma_t = \frac{\eta_t}{\sum_{i=1}^{T} \eta_i}$ and $\sum_{t=1}^{T} \gamma_t = 1$,

$$\mathbb{E}\left[\sum_{t=1}^{T} \gamma_t (f(\mathbf{x}_t) - f(\mathbf{x}^*))\right] \leq \frac{r_1 + G^2 \sum_{t=1}^{T} \eta_t^2}{2 \sum_{t=1}^{T} \eta_t} \tag{6.1.5}$$

Now we define $\bar{\mathbf{x}}_T := \sum_{t=1}^{T} \gamma_t \mathbf{x}_t$, therefore from the convexity of f, $f(\bar{\mathbf{x}}_T) \leq \sum_{t=1}^{T} \gamma_t f(\mathbf{x}_t)$.
Therefor Equation (6.1.5) can now be written as

$$\begin{aligned} \mathbb{E}\left[f(\bar{\mathbf{x}}_t) - f(\mathbf{x}^*)\right] &\leq \frac{r_1 + G^2 \sum_{t=1}^{T} \eta_t^2}{2 \sum_{t=1}^{T} \eta_t} \\ &\leq \frac{R^2 + G^2 \sum_{t=1}^{T} \eta_t^2}{2 \sum_{t=1}^{T} \eta_t} \end{aligned}$$

If for T fixed and $\eta_t = \eta$ for $1 \leq t \leq T$, we have

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_t) - f(\mathbf{x}^*)\right] \leq \frac{R^2 + G^2 T \eta^2}{2T\eta} \to \frac{G^2 \eta}{2} \text{ as } T \to \infty \tag{6.1.6}$$

Equation (6.1.6) minimizes for $\eta = \frac{R}{G\sqrt{T}}$.

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_t) - f(\mathbf{x}^*)\right] \leq \frac{RG}{\sqrt{T}}$$

Therefore when T is not fixed, we choose $\eta_t = \frac{R}{G\sqrt{t}}$ ($\therefore \gamma_t = O(\frac{1}{t})$) to get a convergence rate of $O(\frac{1}{\sqrt{T}})$.

6.2 Strongly convex functions

If f is μ-strongly convex, we can use this additional information to show a better convergence rate.

6.2.1 Convergence with uniform averaging

Considering the squared Euclidean distance from the optimal,

$$\begin{aligned} \left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|_2^2 &= \left\|P_{\mathcal{X}}(\mathbf{x}_t - \eta_t \mathbf{g}_t) - P_{\mathcal{X}}(\mathbf{x}^*)\right\|_2^2 \\ &\leq \left\|\mathbf{x}_t - \mathbf{x}^* - \eta_t \mathbf{g}_t\right\|_2^2 \\ &= \left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2 + \eta_t^2 \left\|\mathbf{g}_t\right\|_2^2 - 2\eta_t \left\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}^* \right\rangle \end{aligned}$$

Taking expectation with respect to $i_1, i_2, \ldots, i_{t-1}$, and using the fact that $\mathbf{x}_t$ is dependent only on $i_1, i_2, \ldots, i_{t-1}$,

$$\begin{aligned} \mathbb{E}\left[\left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|_2^2\right] &\leq \mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2\right] + \eta_t^2 G^2 - 2\eta_t \left\langle \mathbb{E}\left[\mathbf{g}_t\right], \mathbf{x}_t - \mathbf{x}^* \right\rangle \\ &= \mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2\right] + \eta_t^2 G^2 - 2\eta_t \left\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \right\rangle \end{aligned} \tag{6.2.1}$$

From the strong convexity of f, we have

$$\begin{aligned} & f(\mathbf{x}^*) \geq f(\mathbf{x}_t) + \left\langle \nabla f(\mathbf{x}_t), \mathbf{x}^* - \mathbf{x}_t \right\rangle + \frac{\mu}{2} \left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2 \\ \implies & -2\eta_t \left\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \right\rangle \leq -2\eta_t (f(\mathbf{x}_t) - f(\mathbf{x}^*) + \frac{\mu}{2} \left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2) \end{aligned} \tag{6.2.2}$$

Plugging Equation (6.2.2) in Equation (6.2.1), we get

$$\begin{aligned} & \mathbb{E}\left[\left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|_2^2\right] \leq \mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2\right] + \eta_t^2 G^2 - 2\eta_t (f(\mathbf{x}_t) - f(\mathbf{x}^*) + \frac{\mu}{2} \left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2) \\ \implies & 2\eta_t \mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \eta_t^2 G^2 + (1 - \mu\eta_t) \mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2\right] - \mathbb{E}\left[\left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|_2^2\right] \\ \implies & \mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \frac{\eta_t G^2}{2} + \frac{1 - \mu\eta_t}{2\eta_t} \mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2\right] - \frac{1}{2\eta_t} \mathbb{E}\left[\left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|_2^2\right] \end{aligned} \tag{6.2.3}$$

Therefore if we set $\eta_t = \frac{1}{\mu t}$, we get

$$\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \frac{G^2}{2\mu t} + \frac{\mu(t-1)}{2}\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2\right] - \frac{\mu t}{2}\mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2\right] \tag{6.2.4}$$

Telescoping Equation (6.2.4) from $t = 1$ to T, we obtain

$$\sum_{t=1}^{T}\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \frac{G^2}{2\mu}\sum_{t=1}^{T}\frac{1}{t} - \frac{\mu T}{2}\mathbb{E}\left[\|\mathbf{x}_{T+1} - \mathbf{x}^*\|_2^2\right] \tag{6.2.5}$$

Dividing by T and from the convexity of f, we further get

$$\begin{aligned}\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\right) - f(\mathbf{x}^*)\right] &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \\ &\leq \frac{G^2}{2\mu T}\sum_{t=1}^{T}\frac{1}{t} \\ &\leq \frac{G^2}{2\mu T}(1 + \log(T))\end{aligned}$$

And from Equation (6.2.5) we also get

$$\mathbb{E}\left[\|\mathbf{x}_{T+1} - \mathbf{x}^*\|_2^2\right] \leq \frac{G^2}{\mu^2 T}(1 + \log(T))$$

This rate of the convergence of the last iterate, can be tightened as we will show in Section 6.2.3.

6.2.2 Convergence with weighted averaging

Instead of uniform averaging, we can have a weighted averaging scheme to get a better convergence rate [4]. Following the previous analysis, if we put $\eta_t = \frac{2}{\mu(t+1)}$ in Equation (6.2.3), we get

$$\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \frac{G^2}{\mu(t+1)} + \frac{\mu(t-1)}{4}\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2\right] - \frac{\mu(t+1)}{4}\mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2\right] \tag{6.2.6}$$

Multiplying Equation (6.2.6) by t

$$\begin{aligned}t\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] &\leq \frac{tG^2}{\mu(t+1)} + \frac{\mu}{4}\left[t(t-1)\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2\right] - t(t+1)\mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2\right]\right] \\ &\leq \frac{G^2}{\mu} + \frac{\mu}{4}\left[t(t-1)\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2\right] - t(t+1)\mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2\right]\right]\end{aligned} \tag{6.2.7}$$

Telescoping Equation (6.2.7) from $t = 1$ to T we get

$$\sum_{t=1}^{T} t\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \frac{TG^2}{\mu} + \frac{\mu}{4}\left[-T(T+1)\mathbb{E}\left[\|\mathbf{x}_{T+1} - \mathbf{x}^*\|_2^2\right]\right] \tag{6.2.8}$$

Dividing Equation (6.2.8) by $\frac{T(T+1)}{2}$, using the convexity of f, and defining $\bar{\mathbf{x}}_T = \frac{2}{T(T+1)}\sum_{t=1}^{T}\mathbf{x}_t$, we obtain

$$\begin{aligned}\mathbb{E}\left[f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*)\right] &\leq \frac{2G^2}{\mu(T+1)} \\ \text{and, } \mathbb{E}\left[\|\bar{\mathbf{x}}_T - \mathbf{x}^*\|_2^2\right] &\leq \frac{4G^2}{\mu^2(T+1)}\end{aligned}$$

Therefore by a better averaging method, the convergence of SGD for μ-strongly convex functions can be improved to $O\left(\frac{1}{T}\right)$ from $O\left(\frac{\log T}{T}\right)$.

6.2.3 Convergence of last iterate

Considering the squared Euclidean distance from the optimum, we can again show a convergence rate of $O\left(\frac{1}{T}\right)$ after T iterations. The analysis follows via induction. From strong convexity at $\mathbf{x}_1$, we have

$$\begin{aligned} &\frac{\mu}{2}\left\|\mathbf{x}_1-\mathbf{x}^*\right\|_2^2 \leq f(\mathbf{x}^*)-f(\mathbf{x}_1)+\langle\nabla f(\mathbf{x}_1),\mathbf{x}_1-\mathbf{x}^*\rangle \leq \langle\nabla f(\mathbf{x}_1),\mathbf{x}_1-\mathbf{x}^*\rangle \leq \left\|\nabla f(\mathbf{x}_1)\right\|_2\left\|\mathbf{x}_1-\mathbf{x}^*\right\|_2 \\ \implies &\left\|\nabla f(\mathbf{x}_1)\right\|_2^2 \geq \frac{\mu^2}{4}\left\|\mathbf{x}_1-\mathbf{x}^*\right\|_2^2 \end{aligned}$$

Now,

$$\begin{aligned} \mathbb{E}\left[\left\|\mathbf{g}_1\right\|_2^2\right] = \mathbb{E}\left[\left\|\nabla f(\mathbf{x}_1)+(\mathbf{g}_1-\nabla f(\mathbf{x}_1))\right\|_2^2\right] &= \mathbb{E}\left[\left\|\nabla f(\mathbf{x}_1)\right\|_2^2\right]+\mathbb{E}\left[\left\|\mathbf{g}_1-\nabla f(\mathbf{x}_1)\right\|_2^2\right]+\mathbb{E}\left[\langle\nabla f(\mathbf{x}_1),\mathbf{g}_1-\nabla f(\mathbf{x}_1)\rangle\right] \\ &= \mathbb{E}\left[\left\|\nabla f(\mathbf{x}_1)\right\|_2^2\right]+\mathbb{E}\left[\left\|\mathbf{g}_1-\nabla f(\mathbf{x}_1)\right\|_2^2\right] \\ &\leq \mathbb{E}\left[\left\|\nabla f(\mathbf{x}_1)\right\|_2^2\right] \end{aligned}$$

Therefore we have

$$\mathbb{E}\left[\left\|\mathbf{x}_1-\mathbf{x}^*\right\|_2^2\right] \leq \frac{4}{\mu^2}\mathbb{E}\left[\left\|\nabla f(\mathbf{x}_1)\right\|_2^2\right] \leq \frac{4}{\mu^2}\mathbb{E}\left[\left\|\mathbf{g}_1\right\|_2^2\right] \leq \frac{4G^2}{\mu^2}$$

Therefore for $t=1$, $\mathbb{E}\left[\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2\right] \leq \frac{4G^2}{\mu^2 t}$ holds. For a general t,

$$\begin{aligned} \mathbb{E}\left[\left\|\mathbf{x}_{t+1}-\mathbf{x}^*\right\|_2^2\right] &= \mathbb{E}\left[\left\|P_{\mathcal{X}}(\mathbf{x}_t-\eta_t\mathbf{g}_t)-P_{\mathcal{X}}(\mathbf{x}^*)\right\|_2^2\right] \\ &\leq \mathbb{E}\left[\left\|\mathbf{x}_t-\mathbf{x}^*-\eta_t\mathbf{g}_t\right\|_2^2\right] \\ &= \mathbb{E}\left[\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2\right]+\eta_t^2\mathbb{E}\left[\left\|\mathbf{g}_t\right\|_2^2\right]-2\eta_t\mathbb{E}\left[\langle\mathbf{g}_t,\mathbf{x}_t-\mathbf{x}^*\rangle\right] \\ &= \mathbb{E}\left[\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2\right]+\eta_t^2\mathbb{E}\left[\left\|\mathbf{g}_t\right\|_2^2\right]-2\eta_t\mathbb{E}\left[\langle\nabla f(\mathbf{x}_t),\mathbf{x}_t-\mathbf{x}^*\rangle\right] \end{aligned}$$

Using strong convexity of f at $\mathbf{x}_t$ and $\mathbf{x}^*$, we have

$$\langle\nabla f(\mathbf{x}_t),\mathbf{x}_t-\mathbf{x}^*\rangle \geq f(\mathbf{x}_t)-f(\mathbf{x}^*)+\frac{\mu}{2}\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2 \geq \mu\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2$$

Therefore

$$\begin{aligned} \mathbb{E}\left[\left\|\mathbf{x}_{t+1}-\mathbf{x}^*\right\|_2^2\right] &\leq \mathbb{E}\left[\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2\right]+\eta_t^2\mathbb{E}\left[\left\|\mathbf{g}_t\right\|_2^2\right]-2\eta_t\mu\mathbb{E}\left[\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2\right] \\ &\leq (1-2\eta_t\mu)\mathbb{E}\left[\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2\right]+\eta_t^2G^2 \end{aligned}$$

Plugging in $\eta_t=\frac{1}{\mu t}$, we get

$$\mathbb{E}\left[\left\|\mathbf{x}_{t+1}-\mathbf{x}^*\right\|_2^2\right] \leq \left(1-\frac{2}{t}\right)\mathbb{E}\left[\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2\right]+\frac{G^2}{\mu^2t^2}$$

Therefore for $t=2$, $\mathbb{E}\left[\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2\right] \leq \frac{4G^2}{\mu^2 t}$ holds again. Assuming the hypothesis to be true for $t-1$, we have

$$\mathbb{E}\left[\left\|\mathbf{x}_t-\mathbf{x}^*\right\|_2^2\right] \leq \frac{4G^2}{\mu^2 t}$$

And checking for t,

$$\begin{aligned} \mathbb{E}\left[\left\|\mathbf{x}_{t+1}-\mathbf{x}^*\right\|_2^2\right] &\leq \left(1-\frac{2}{t}\right)\frac{4G^2}{\mu^2t}+\frac{G^2}{\mu^2t^2} \\ &\leq \frac{G}{\mu^2t^2}(4t-7) \\ &\leq \frac{4G^2}{\mu^2(t+1)} \end{aligned}$$

Therefore we have $\mathbb{E}\left[\left\|\mathbf{x}_T-\mathbf{x}^*\right\|_2^2\right] \leq \frac{4G^2}{\mu^2T} = O\left(\frac{1}{T}\right)$.

6.2.4 Convergence using Tail Averaging

Instead of using a weighted average as what analyzed in Section 6.2.2, we can also show similarly good convergence bounds for uniform averaging on the tail of the iterations called α-suffix averaging, as shown in [7]. We define the last α fraction of the iterates as the α-suffix, therefore the result talks about the convergence of $\bar{\mathbf{x}}_t^\alpha = \sum_{t=(1-\alpha)T+1}^{T} \mathbf{x}_t$. Upper bounding $\mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2\right]$,

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2\right] &= \mathbb{E}\left[\|P_{\mathcal{X}}(\mathbf{x}_t - \eta_t \mathbf{g}_t) - P_{\mathcal{X}}(\mathbf{x}^*)\|_2^2\right] \\
&\leq \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^* - \eta_t \mathbf{g}_t\|_2^2\right] \\
&\leq \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2\right] + \eta_t^2 G^2 - 2\eta_t \mathbb{E}\left[\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle\right] \\
\implies \mathbb{E}\left[\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle\right] &\leq \frac{\eta_t G^2}{2} + \frac{1}{2}\left[\frac{\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2\right]}{\eta_t} - \frac{\mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2\right]}{\eta_t}\right]
\end{aligned}$$

By convexity,

$$\sum_{t=(1-\alpha)T+1}^{T} \mathbb{E}\left[\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle\right] \geq \sum_{t=(1-\alpha)T+1}^{T} \mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \geq \alpha T \mathbb{E}\left[f(\bar{\mathbf{x}}_t^\alpha) - f(\mathbf{x}^*)\right]$$

Therefore

$$\begin{aligned}
\mathbb{E}\left[f(\bar{\mathbf{x}}_t^\alpha) - f(\mathbf{x}^*)\right] &\leq \frac{1}{2\alpha T}\left[\sum_{t=(1-\alpha)T+1}^{T} \eta_t G^2 + \sum_{t=(1-\alpha)T+1}^{T}\left[\frac{\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2\right]}{\eta_t} - \frac{\mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2\right]}{\eta_t}\right]\right] \\
&\leq \frac{1}{2\alpha T}\left[\frac{1}{\eta_{(1-\alpha)T}}\mathbb{E}\left[\|\mathbf{x}_{(1-\alpha)T+1} - \mathbf{x}^*\|_2^2\right] + \sum_{t=(1-\alpha)T+1}^{T} \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2\right]\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + G^2 \sum_{t=(1-\alpha)T+1}^{T} \eta_t\right]
\end{aligned}$$

Using the result from section 6.2.3, we have $\mathbb{E}\left[\|\mathbf{x}_T - \mathbf{x}^*\|_2^2\right] \leq \frac{4G^2}{\mu^2 T}$. Using this we get

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_t^\alpha) - f(\mathbf{x}^*)\right] \leq \frac{1}{2\alpha T}\left[\frac{1}{\eta_{(1-\alpha)T}} \frac{4G^2}{\mu^2((1-\alpha)T+1)} + \sum_{t=(1-\alpha)T+1}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\frac{4G^2}{\mu^2 t}\right] + \frac{G^2}{2\alpha T}\sum_{t=(1-\alpha)T+1}^{T} \eta_t$$

Plugging in $\eta_t = \frac{1}{\mu t}$, we now have

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_t^\alpha) - f(\mathbf{x}^*)\right] \leq \frac{2G^2}{\mu \alpha T}\left[1 + \sum_{t=(1-\alpha)T+1}^{T} \frac{1}{t}\right] + \frac{G^2}{2\mu\alpha T}\sum_{t=(1-\alpha)T+1}^{T} \frac{1}{t}$$

Using the fact that $\sum_{t=(1-\alpha)T+1}^{T} \frac{1}{t} \leq \log \frac{1}{1-\alpha}$, we simplify and get

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_T^\alpha) - f(\mathbf{x}^*)\right] \leq \frac{2 + \frac{5}{2}\log\frac{1}{1-\alpha}}{\alpha}\frac{G^2}{\mu T} = O\left(\frac{1}{T}\right)$$

The constant as a function of alpha is minimum at $\alpha \approx 0.7675$.

6.3 Smooth functions

If each of the f_i's are L-smooth, we can guarantee good convergence rates.

6.3.1 Convergence

Let $\mathbf{e}_t := \nabla f(\mathbf{x}_t) - \mathbf{g}_t$, therefore $\mathbb{E}[\mathbf{e}_t] = \mathbf{0}$ and assume $\mathbb{E}\left[\|\mathbf{e}_t\|_2^2\right] \leq \sigma^2$. Then for $\eta_t = \frac{1}{L+\alpha_t}$ where $\alpha_t = O\left(\frac{1}{\sqrt{t}}\right)$, it can be shown that

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*)\right] \leq O\left(\frac{LR^2}{T}\right) + O\left(\frac{\sigma R}{\sqrt{T}}\right)$$

Where $\bar{\mathbf{x}}_t = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t$.

6.4 Smooth and Strongly convex functions

If f is L-smooth and μ-strongly convex, we can use the result in Section 6.2.2 to obtain the convergence rate for this class of functions.

6.4.1 Convergence

From Section 6.2.2, with $\bar{\mathbf{x}}_T = \frac{2}{T(T+1)} \sum_{t=1}^{T} \mathbf{x}_t$ and $\eta_t = \frac{2}{\mu(T+1)}$, we have

$$\mathbb{E}\left[\|\bar{\mathbf{x}}_T - \mathbf{x}^*\|_2^2\right] \leq \frac{4G^2}{\mu^2(T+1)}$$

$$\implies \mathbb{E}\left[f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*)\right] \leq \frac{L}{2}\mathbb{E}\left[\|\bar{\mathbf{x}}_T - \mathbf{x}^*\|_2^2\right] \leq \frac{2G^2L}{\mu^2(T+1)}$$

7 Some faster stochastic algorithms

SGD is popular for large scale optimization but it has slow convergence asymptotically due to the inherent variance. In order to ensure convergence, the learning rate η_t has to decay to zero which leads to slow convergence. The need of small learning rate is due to the variance of SGD.

A popular way that does explicit variance reduction is SVRG [3] and its variants [2] as we discuss below.

7.1 Stochastic Variance Reduced Gradient (SVRG)

We have the same setting where

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$$

We assume f to be μ-strongly convex and each f_i to be L-smooth. The algorithm keeps a snapshot $\tilde{\mathbf{x}}$ after every m iterations. Moreover, the average gradient is maintained, i.e.,

$$\tilde{\mathbf{g}} := \nabla f(\tilde{\mathbf{x}}) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{\mathbf{x}})$$

Note that $\mathbb{E}\left[\nabla f_i(\tilde{\mathbf{x}}) - \tilde{\mathbf{g}}\right] = \mathbf{0}$. Therefore we can have the stochastic update defined as

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t(\nabla f_i(\mathbf{x}_t) - \nabla f_i(\tilde{\mathbf{x}}) + \tilde{\mathbf{g}})$$

where $i \sim U([n])$.
The above update is the normal SGD update of the auxiliary function with $\tilde{f}_i(\mathbf{x}) := f_i(\mathbf{x}) - \langle \nabla f_i(\tilde{\mathbf{x}}) - \tilde{\mathbf{g}}, \mathbf{x}\rangle$. And since $\sum_{i=1}^{n}(\nabla f_i(\tilde{\mathbf{x}}) - \tilde{\mathbf{g}}) = \mathbf{0}$,

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \tilde{f}_i(\mathbf{x})$$

For every stage of m such updates, we have $m+n$ gradient computations. Therefore it is natural to choose m to be of order n or slightly higher (for example $m = 2n$ for convex problems and $m = 5n$ for non-convex problems). The algorithm therefore is described as

Algorithm 1: SVRG

Initialize : $\tilde{\mathbf{x}}_0$

for $s = 1, 2, \dots$ **do**

  $\tilde{\mathbf{x}} = \tilde{\mathbf{x}}_{s-1}$

  $\tilde{\mathbf{g}} = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{\mathbf{x}})$

  $\mathbf{x}_1 = \tilde{\mathbf{x}}$

  for $t = 1, 2, \dots, m$ **do**

    Sample $i \sim U([n])$

    $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t(\nabla f_i(\mathbf{x}_t) - \nabla f_i(\tilde{\mathbf{x}}) + \tilde{\mathbf{g}})$

  end

  option I : set $\tilde{\mathbf{x}}_s = \mathbf{x}_m$

  option II : set $\tilde{\mathbf{x}}_s = \mathbf{x}_t$ for randomly chosen $t \in [m]$

end

7.1.1 Convergence

For all i, define

$$g_i(\mathbf{x}) = f_i(\mathbf{x}) - f_i(\mathbf{x}^*) - \langle \nabla f_i(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \tag{7.1.1}$$

Therefore we have $g_i(\mathbf{x}^*) = \min_{\mathbf{x}} g_i(\mathbf{x}) \quad (\because \nabla g_i(\mathbf{x}^*) = \mathbf{0})$, therefore

$$\begin{aligned} 0 = g_i(\mathbf{x}^*) &\leq \min_{\eta}[g_i(\mathbf{x} - \eta\nabla g_i(\mathbf{x}))] \\ &\leq \min_{\eta}[g_i(\mathbf{x}) - \eta \left\|\nabla g_i(\mathbf{x})\right\|_2^2 + \frac{L}{2}\eta^2 \left\|\nabla g_i(\mathbf{x})\right\|_2^2] \qquad (g \text{ is also } L-\text{smooth}) \\ &= g_i(\mathbf{x}) - \frac{1}{2L}\left\|\nabla g_i(\mathbf{x})\right\|_2^2 \end{aligned}$$

Plugging in $g(\mathbf{x})$ and $\nabla g(\mathbf{x})$ from the definition of g in Equation (7.1.1), we get

$$\left\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)\right\|_2^2 \leq 2L[f_i(\mathbf{x}) - f_i(\mathbf{x}^*) - \langle \nabla f_i(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle] \tag{7.1.2}$$

Summing Equation (7.1.2) over $i = 1, \dots, n$, and using the fact that $\nabla f(\mathbf{x}^*) = 0$, we get

$$\frac{1}{n}\sum_{i=1}^{n} \left\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)\right\|_2^2 \leq 2L[f(\mathbf{x}) - f(\mathbf{x}^*)]$$

Now let $\mathbf{v} := \nabla f_i(\mathbf{x}_t) - \nabla f_i(\tilde{\mathbf{x}}) + \tilde{\mathbf{g}}$. Taking the conditional expectation of $\left\|\mathbf{v}_t\right\|_2^2$ given $\mathbf{x}_t$, we have

$$\begin{aligned} \mathbb{E}\left[\left\|\mathbf{v}_t\right\|_2^2\right] &\leq 2\mathbb{E}\left[\left\|\nabla f_i(\mathbf{x}_t) - \nabla f_i(\mathbf{x}^*)\right\|_2^2\right] + 2\mathbb{E}\left[\left\|[\nabla f_i(\tilde{\mathbf{x}}_t) - \nabla f_i(\mathbf{x}^*)] - \nabla f(\tilde{\mathbf{x}})\right\|_2^2\right] \\ &= 2\mathbb{E}\left[\left\|\nabla f_i(\mathbf{x}_t) - \nabla f_i(\mathbf{x}^*)\right\|_2^2\right] + 2\mathbb{E}\left[\left\|[\nabla f_i(\tilde{\mathbf{x}}_t) - \nabla f_i(\mathbf{x}^*)] - \mathbb{E}\left[\nabla f_i(\tilde{\mathbf{x}}_t) - \nabla f_i(\mathbf{x}^*)\right]\right\|_2^2\right] \\ &\leq 2\mathbb{E}\left[\left\|\nabla f_i(\mathbf{x}_t) - \nabla f_i(\mathbf{x}^*)\right\|_2^2\right] + 2\mathbb{E}\left[\left\|\nabla f_i(\tilde{\mathbf{x}}_t) - \nabla f_i(\mathbf{x}^*)\right\|_2^2\right] \\ &\leq 4L[f(\mathbf{x}_t) - f(\mathbf{x}^*) + f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)] \end{aligned}$$

Considering the squared Euclidean distance from the optimum, and taking its conditional expectation given $\mathbf{x}_t$ we get

$$\begin{aligned} \mathbb{E}\left[\left\|\mathbf{x}_{t+1} - \mathbf{x}^*\right\|_2^2\right] &= \mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2\right] - 2\eta_t \left\langle \mathbf{x}_t - \mathbf{x}^*, \mathbb{E}\left[\mathbf{v}_t\right]\right\rangle + \eta_t^2\mathbb{E}\left[\left\|\mathbf{v}_t\right\|_2^2\right] \\ &\leq \mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2\right] - 2\eta_t \left\langle \mathbf{x}_t - \mathbf{x}^*, \nabla f(\mathbf{x}_t)\right\rangle + 4L\eta_t^2[f(\mathbf{x}_t) - f(\mathbf{x}^*) + f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)] \\ &\leq \mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2\right] - 2\eta_t[f(\mathbf{x}_t) - f(\mathbf{x}^*)] + 4L\eta_t^2[f(\mathbf{x}_t) - f(\mathbf{x}^*) + f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)] \\ &= \mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_2^2\right] - 2\eta_t(1 - 2L\eta_t)[f(\mathbf{x}_t) - f(\mathbf{x}^*)] + 4L\eta_t^2[f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)] \end{aligned}$$

Therefore

$$\mathbb{E}\left[\|\mathbf{x}_{t+1}-\mathbf{x}^*\|_2^2\right]+2\eta_t(1-2L\eta_t)[f(\mathbf{x}_t)-f(\mathbf{x}^*)]\leq\mathbb{E}\left[\|\mathbf{x}_t-\mathbf{x}^*\|_2^2\right]+4L\eta_t^2[f(\tilde{\mathbf{x}})-f(\mathbf{x}^*)] \tag{7.1.3}$$

Telescoping Equation (7.1.3) from $t=1$ to $m-1$, setting $\eta_t=\eta$, and taking expectation with the history, we get

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}_m-\mathbf{x}^*\|_2^2\right]+2\eta(1-2L\eta)m\mathbb{E}\left[f(\tilde{\mathbf{x}}_s)-f(\mathbf{x}^*)\right]&\leq\mathbb{E}\left[\|\mathbf{x}_1-\mathbf{x}^*\|_2^2\right]+4Lm\eta^2\mathbb{E}\left[f(\tilde{\mathbf{x}}_{s-1})-f(\mathbf{x}^*)\right]\\&=\mathbb{E}\left[\|\tilde{\mathbf{x}}_{s-1}-\mathbf{x}^*\|_2^2\right]+4Lm\eta^2\mathbb{E}\left[f(\tilde{\mathbf{x}}_{s-1})-f(\mathbf{x}^*)\right]\\&\leq\frac{2}{\mu}\mathbb{E}\left[f(\tilde{\mathbf{x}}_{s-1})-f(\mathbf{x}^*))\right]+4Lm\eta^2\mathbb{E}\left[f(\tilde{\mathbf{x}}_{s-1})-f(\mathbf{x}^*)\right]\\&=2(\frac{1}{\mu}+2Lm\eta^2)\mathbb{E}\left[f(\tilde{\mathbf{x}}_{s-1})-f(\mathbf{x}^*))\right]\end{aligned}$$

Therefore rearranging the terms we get

$$\mathbb{E}\left[f(\tilde{\mathbf{x}}_s)-f(\mathbf{x}^*)\right]\leq\left[\frac{1}{\mu\eta(1-2L\eta)m}+\frac{2L\eta}{1-2L\eta}\right]\mathbb{E}\left[f(\tilde{\mathbf{x}}_{s-1})-f(\mathbf{x}^*))\right]$$

Defining $\alpha=\frac{1}{\mu\eta(1-2L\eta)m}+\frac{2L\eta}{1-2L\eta}$, we have

$$\mathbb{E}\left[f(\tilde{\mathbf{x}}_s)-f(\mathbf{x}^*)\right]\leq\alpha^s\mathbb{E}\left[f(\tilde{\mathbf{x}}_0)-f(\mathbf{x}^*))\right]$$

Usually, m is chosen to be $O(\kappa)$ and $\eta=O\left(\frac{1}{L}\right)$ to give a convergence rate of $O\left((n+\kappa)\log\frac{1}{\epsilon}\right)$.

7.2 SVRG++

The original SVRG method as described in [3] was for strongly convex objectives, whereas objectives like that of Lasso or Logistic regression etc., are non-strongly convex. A variant of SVRG known as SVRG++ algorithm [2] which gives faster convergence by modifying it in a novel manner.

Consider the composite convex minimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d}\left\{F(\mathbf{x}):=f(\mathbf{x})+\Psi(\mathbf{x}):=\frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})+\Psi(\mathbf{x})\right\}$$

Here f_i's are L-smooth functions and Ψ is a relatively simple (possibly non-differentiable) function. Example - For lasso, $f_i(\mathbf{x}):=\frac{1}{2}(\langle\mathbf{a}_i,\mathbf{x}\rangle-y_i)^2$ and $\Psi:=\sigma\|\mathbf{x}\|_1$.

In the presence of the the proximal function Ψ, the SVRG update becomes

$$\mathbf{x}_{t+1}=\arg\min_{\mathbf{y}\in\mathbb{R}^d}\left\{\frac{1}{2\eta}\|\mathbf{y}-\mathbf{x}_t\|_2^2+\langle\boldsymbol{\xi}_t,\mathbf{y}\rangle+\Psi(\mathbf{y})\right\}$$

where $\boldsymbol{\xi}_t=\nabla f_i(\mathbf{x}_t)$ for SGD and $\boldsymbol{\xi}_t=\nabla f_i(\mathbf{x}_t)-\nabla f_i(\tilde{\mathbf{x}})+f(\tilde{\mathbf{x}})$ for the snapshots $\tilde{\mathbf{x}}$ after every m stochastic updates in SVRG. Each of such definitions of $\boldsymbol{\xi}_t$ satisfy $\mathbb{E}[\boldsymbol{\xi}_t]=\nabla f(\mathbf{x}_t)$.

For SVRG++, the s-th epoch consists of m_s stochastic updates and m_s doubles after epoch, i.e., $m_s=2^s m_0$. Also, unlike SVRG where $\tilde{\mathbf{x}}^s$ is the average point of the previous epoch, for SVRG++ we have $\tilde{\mathbf{x}}_0^s = \mathbf{x}_{m_{s-1}}^{s-1}$.

Algorithm 2: SVRG++($\mathbf{x}^\phi, m_0, S, \eta$)

Initialize : $\tilde{\mathbf{x}}_0 = \mathbf{x}^\phi$, $\mathbf{x}_0^1 = \mathbf{x}^\phi$
for $s = 1, 2, \ldots, S$ **do**
  $\tilde{\mathbf{g}}_{s-1} = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{\mathbf{x}}^{s-1})$
  $m_s = 2^s m_0$
  for $t = 0, 1, \ldots, m_s - 1$ **do**
    Sample $i \sim U([n])$
    $\boldsymbol{\xi}_t^s = \nabla f_i(\mathbf{x}_t^s) - \nabla f_i(\tilde{\mathbf{x}}^{s-1}) + \tilde{\mathbf{g}}_{s-1}$
    $\mathbf{x}_{t+1}^s = \arg\min_{\mathbf{y}\in\mathbb{R}^d} \left\{ \frac{1}{2\eta}\|\mathbf{x}_t^s - \mathbf{y}\|_2^2 + \Psi(\mathbf{y}) + \langle \boldsymbol{\xi}_t^s, \mathbf{y}\rangle \right\}$
  end
  $\tilde{\mathbf{x}}_s = \frac{1}{m_s}\sum_{t=1}^{m_s} \mathbf{x}_t^s$
  $\mathbf{x}_0^{s+1} = \mathbf{x}_{m_s}^s$
end
return $\tilde{\mathbf{x}}^S$

7.2.1 Convergence

Let $i_t^s \sim U([n])$ be the random index for the s-th epoch and t-th inner iteration, and similarly $\boldsymbol{\xi}_t^s = \nabla f_i(\mathbf{x}_t^s) - \nabla f_i(\tilde{\mathbf{x}}^{s-1}) + \tilde{\mathbf{g}}_{s-1}$ be the stochastic gradient.

For all $\mathbf{u} \in \mathbb{R}^d$,

$$\begin{aligned}
\mathbb{E}_{i_t^s}\left[F(\mathbf{x}_{t+1}^s) - F(\mathbf{u})\right] &= \mathbb{E}_{i_t^s}\left[f(\mathbf{x}_{t+1}^s) - f(\mathbf{u}) + \Psi(\mathbf{x}_{t+1}^s) - \Psi(\mathbf{u})\right] \\
&\le \mathbb{E}_{i_t^s}\left[f(\mathbf{x}_t^s) - f(\mathbf{u}) + \left\langle \nabla f(\mathbf{x}_t^s), \mathbf{x}_{t+1}^s - \mathbf{x}_t^s \right\rangle + \frac{L}{2}\left\|\mathbf{x}_t^s - \mathbf{x}_{t+1}^s\right\|_2^2 + \Psi(\mathbf{x}_{t+1}^s) - \Psi(\mathbf{u})\right] \\
&\hspace{20em} \text{(Using smoothness of } f\text{)} \\
&\le \mathbb{E}_{i_t^s}\left[\langle \nabla f(\mathbf{x}_t^s), \mathbf{x}_t^s - \mathbf{u}\rangle + \left\langle \nabla f(\mathbf{x}_t^s), \mathbf{x}_{t+1}^s - \mathbf{x}_t^s \right\rangle + \frac{L}{2}\left\|\mathbf{x}_t^s - \mathbf{x}_{t+1}^s\right\|_2^2 + \Psi(\mathbf{x}_{t+1}^s) - \Psi(\mathbf{u})\right] \\
&\hspace{20em} \text{(Using convexity of } f\text{)} \\
&= \mathbb{E}_{i_t^s}\left[\langle \boldsymbol{\xi}_t^s, \mathbf{x}_t^s - \mathbf{u}\rangle + \left\langle \nabla f(\mathbf{x}_t^s), \mathbf{x}_{t+1}^s - \mathbf{x}_t^s \right\rangle + \frac{L}{2}\left\|\mathbf{x}_t^s - \mathbf{x}_{t+1}^s\right\|_2^2 + \Psi(\mathbf{x}_{t+1}^s) - \Psi(\mathbf{u})\right] \quad (7.2.1) \\
&\hspace{20em} \text{(Since } \mathbb{E}_{i_t^s}\left[\boldsymbol{\xi}_t^s\right] = \nabla f(\mathbf{x}_t^s)\text{)}
\end{aligned}$$

Analyzing the first and the last two terms in Equation (7.2.1),

$$\langle \boldsymbol{\xi}_t^s, \mathbf{x}_t^s - \mathbf{u}\rangle + \Psi(\mathbf{x}_{t+1}^s) - \Psi(\mathbf{u}) = \left\langle \boldsymbol{\xi}_t^s, \mathbf{x}_t^s - \mathbf{x}_{t+1}^s\right\rangle + \left\langle \boldsymbol{\xi}_t^s, \mathbf{x}_{t+1}^s - \mathbf{u}\right\rangle + \Psi(\mathbf{x}_{t+1}^s) - \Psi(\mathbf{u}) \quad (7.2.2)$$

Since $\mathbf{x}_{t+1}^s = \arg\min_{\mathbf{y}\in\mathbb{R}^d} \left\{ \frac{1}{2\eta}\|\mathbf{x}_t^s - \mathbf{y}\|_2^2 + \Psi(\mathbf{y}) + \langle \boldsymbol{\xi}_t^s, \mathbf{y}\rangle \right\} \implies \exists\ \mathbf{g} \in \partial\Psi(\mathbf{x}_{t+1}^s) \ni \frac{1}{\eta}(\mathbf{x}_{t+1}^s - \mathbf{x}_t^s) + \boldsymbol{\xi}_t^s + \mathbf{g} = \mathbf{0}$. From the convexity of Ψ, we have

$$\begin{aligned}
\Psi(\mathbf{u}) - \Psi(\mathbf{x}_{t+1}^s) &\ge \left\langle \mathbf{g}, \mathbf{u} - \mathbf{x}_{t+1}^s\right\rangle \\
\implies \Psi(\mathbf{u}) - \Psi(\mathbf{x}_{t+1}^s) + \left\langle \frac{1}{\eta}(\mathbf{x}_{t+1}^s - \mathbf{x}_t^s) + \boldsymbol{\xi}_t^s, \mathbf{u} - \mathbf{x}_{t+1}^s\right\rangle &\ge \left\langle \frac{1}{\eta}(\mathbf{x}_{t+1}^s - \mathbf{x}_t^s) + \boldsymbol{\xi}_t^s + \mathbf{g}, \mathbf{u} - \mathbf{x}_{t+1}^s\right\rangle = 0 \\
\implies \Psi(\mathbf{u}) - \Psi(\mathbf{x}_{t+1}^s) + \left\langle \boldsymbol{\xi}_t^s, \mathbf{u} - \mathbf{x}_{t+1}^s\right\rangle &\ge \frac{1}{\eta}\left\langle (\mathbf{x}_{t+1}^s - \mathbf{x}_t^s), \mathbf{x}_{t+1}^s - \mathbf{u}\right\rangle \\
\implies \Psi(\mathbf{x}_{t+1}^s) - \Psi(\mathbf{u}) + \left\langle \boldsymbol{\xi}_t^s, \mathbf{x}_{t+1}^s - \mathbf{u}\right\rangle &\le -\frac{1}{\eta}\left\langle (\mathbf{x}_{t+1}^s - \mathbf{x}_t^s), \mathbf{x}_{t+1}^s - \mathbf{u}\right\rangle \quad (7.2.3)
\end{aligned}$$

Pluggin in Equation (7.2.3) in Equation (7.2.2), we get

$$\langle \boldsymbol{\xi}_t^s, \mathbf{x}_t^s - \mathbf{u}\rangle + \Psi(\mathbf{x}_{t+1}^s) - \Psi(\mathbf{u}) \le \left\langle \boldsymbol{\xi}_t^s, \mathbf{x}_t^s - \mathbf{x}_{t+1}^s\right\rangle - \frac{1}{\eta}\left\langle (\mathbf{x}_{t+1}^s - \mathbf{x}_t^s), \mathbf{x}_{t+1}^s - \mathbf{u}\right\rangle$$

$$= \left\langle \boldsymbol{\xi}_t^s, \mathbf{x}_t^s - \mathbf{x}_{t+1}^s \right\rangle + \frac{\|\mathbf{x}_t^s - \mathbf{u}\|_2^2}{2\eta} - \frac{\left\|\mathbf{x}_{t+1}^s - \mathbf{u}\right\|_2^2}{2\eta} - \frac{\left\|\mathbf{x}_{t+1}^s - \mathbf{x}_t^s\right\|_2^2}{2\eta} \tag{7.2.4}$$

(Using the identity, $2\langle \mathbf{x}-\mathbf{y}, \mathbf{x}-\mathbf{z}\rangle = \|\mathbf{x}-\mathbf{y}\|_2^2 + \|\mathbf{x}-\mathbf{z}\|_2^2 - \|\mathbf{y}-\mathbf{z}\|_2^2$)

Combining Equation (7.2.1) and Equation (7.2.4), we get

$$\begin{aligned}
\mathbb{E}_{i_t^s}\left[F(\mathbf{x}_{t+1}^s) - F(\mathbf{u})\right] &\le \mathbb{E}_{i_t^s}\left[\left\langle \boldsymbol{\xi}_t^s - \nabla f(\mathbf{x}_t^s), \mathbf{x}_t^s - \mathbf{x}_{t+1}^s \right\rangle - \frac{1-\eta L}{2\eta}\left\|\mathbf{x}_t^s - \mathbf{x}_{t+1}^s\right\|_2^2 + \frac{\|\mathbf{x}_t^s - \mathbf{u}\|_2^2 - \left\|\mathbf{x}_{t+1}^s - \mathbf{u}\right\|_2^2}{2\eta}\right] \\
&\le \mathbb{E}_{i_t^s}\left[\frac{\eta}{2(1-\eta L)}\|\boldsymbol{\xi}_t^s - \nabla f(\mathbf{x}_t^s)\|_2^2 + \frac{\|\mathbf{x}_t^s - \mathbf{u}\|_2^2 - \left\|\mathbf{x}_{t+1}^s - \mathbf{u}\right\|_2^2}{2\eta}\right] \qquad (7.2.5)
\end{aligned}$$

(Using Young's Inequality)

Now upper bounding $\mathbb{E}_{i_t^s}\left[\|\boldsymbol{\xi}_t^s - \nabla f(\mathbf{x}_t^s)\|_2^2\right]$,

$$\begin{aligned}
\mathbb{E}_{i_t^s}\left[\|\boldsymbol{\xi}_t^s - \nabla f(\mathbf{x}_t^s)\|_2^2\right] &= \mathbb{E}_{i_t^s}\left[\left\|(\nabla f_{i_t^s}(\mathbf{x}_t^s) - \nabla f_{i_t^s}(\tilde{\mathbf{x}}^{s-1})) - (\nabla f(\mathbf{x}_t^s) - \nabla f(\tilde{\mathbf{x}}^{s-1}))\right\|_2^2\right] \\
&\le \mathbb{E}_{i_t^s}\left[\left\|\nabla f_{i_t^s}(\mathbf{x}_t^s) - \nabla f_{i_t^s}(\tilde{\mathbf{x}}^{s-1})\right\|_2^2\right] \qquad \left(\text{Using } \mathbb{E}\left[\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\|_2^2\right] = \mathbb{E}\left[\|\mathbf{X}\|_2^2\right] - \|\mathbb{E}[\mathbf{X}]\|_2^2\right) \\
&= \mathbb{E}_{i_t^s}\left[\left\|(\nabla f_{i_t^s}(\mathbf{x}_t^s) - \nabla f_{i_t^s}(\mathbf{x}^*)) - (\nabla f_{i_t^s}(\tilde{\mathbf{x}}^{s-1}) - \nabla f_{i_t^s}(\mathbf{x}^*))\right\|_2^2\right] \\
&\le 2\mathbb{E}_{i_t^s}\left[\left\|\nabla f_{i_t^s}(\mathbf{x}_t^s) - \nabla f_{i_t^s}(\mathbf{x}^*)\right\|_2^2\right] + 2\mathbb{E}_{i_t^s}\left[\left\|\nabla f_{i_t^s}(\tilde{\mathbf{x}}^{s-1}) - \nabla f_{i_t^s}(\mathbf{x}^*)\right\|_2^2\right] \qquad (7.2.6)
\end{aligned}$$

Lemma 7.1. *If f_i is convex and L-smooth, we have*

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)\|_2^2 \le 2L\left[f_i(\mathbf{x}) - f_i(\mathbf{x}^*) - \langle \nabla f_i(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*\rangle\right]$$

Proof. Define $\phi(\mathbf{x}) := f_i(\mathbf{x}) - f_i(\mathbf{x}^*) - \langle \nabla f_i(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*\rangle$. Therefore $\nabla\phi(\mathbf{x}) = \nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)$.

$$\begin{aligned}
0 = \phi(\mathbf{x}^*) \le \phi(\mathbf{x} - \frac{1}{L}\nabla\phi(\mathbf{x})) &\le \phi(\mathbf{x}) + \left\langle \nabla\phi(\mathbf{x}), -\frac{1}{L}\nabla\phi(\mathbf{x})\right\rangle + \frac{L}{2}\left\|\frac{1}{L}\nabla\phi(\mathbf{x})\right\|_2^2 \\
&= \phi(\mathbf{x}) - \frac{1}{2L}\|\nabla\phi(\mathbf{x})\|_2^2 \\
\implies \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)\|_2^2 &\le 2L\left[f_i(\mathbf{x}) - f_i(\mathbf{x}^*) - \langle \nabla f_i(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*\rangle\right]
\end{aligned}$$

□

Using Lemma 7.1 in Equation (7.2.6), we get

$$\begin{aligned}
\mathbb{E}_{i_t^s}\left[\|\boldsymbol{\xi}_t^s - \nabla f(\mathbf{x}_t^s)\|_2^2\right] &\le 4L\mathbb{E}_{i_t^s}\left[f_{i_t^s}(\mathbf{x}_t^s) - f_{i_t^s}(\mathbf{x}^*) - \left\langle \nabla f_{i_t^s}(\mathbf{x}^*), \mathbf{x}_t^s - \mathbf{x}^*\right\rangle + f_{i_t^s}(\tilde{\mathbf{x}}^{s-1}) - f_{i_t^s}(\mathbf{x}^*) - \left\langle \nabla f_{i_t^s}(\mathbf{x}^*), \tilde{\mathbf{x}}^{s-1} - \mathbf{x}^*\right\rangle\right] \\
&= 4L\left[f(\mathbf{x}_t^s) - f(\mathbf{x}^*) - \langle \nabla f(\mathbf{x}^*), \mathbf{x}_t^s - \mathbf{x}^*\rangle + f(\tilde{\mathbf{x}}^{s-1}) - f(\mathbf{x}^*) - \left\langle \nabla f(\mathbf{x}^*), \tilde{\mathbf{x}}^{s-1} - \mathbf{x}^*\right\rangle\right] \qquad (7.2.7)
\end{aligned}$$

Since $\partial F(\mathbf{x}^*) = \nabla f(\mathbf{x}^*) + \mathbf{g}^* = \mathbf{0}$ for some $\mathbf{g}^* \in \partial\Psi(\mathbf{x}^*)$. Using this we get

$$\begin{aligned}
\mathbb{E}_{i_t^s}\left[\|\boldsymbol{\xi}_t^s - \nabla f(\mathbf{x}_t^s)\|_2^2\right] &\le 4L\left[f(\mathbf{x}_t^s) - f(\mathbf{x}^*) + \langle \mathbf{g}^*, \mathbf{x}_t^s - \mathbf{x}^*\rangle + f(\tilde{\mathbf{x}}^{s-1}) - f(\mathbf{x}^*) + \left\langle \mathbf{g}^*, \tilde{\mathbf{x}}^{s-1} - \mathbf{x}^*\right\rangle\right] \\
&\le 4L\left[f(\mathbf{x}_t^s) - f(\mathbf{x}^*) + \Psi(\mathbf{x}_t^s) - \Psi(\mathbf{x}^*) + f(\tilde{\mathbf{x}}^{s-1}) - f(\mathbf{x}^*) + \Psi(\tilde{\mathbf{x}}^{s-1}) - \Psi(\mathbf{x}^*)\right] \\
&\qquad\qquad \text{(Using the convexity of } \Psi) \\
&= 4L\left[F(\mathbf{x}_t^s) - F(\mathbf{x}^*) + F(\tilde{\mathbf{x}}^{s-1}) - F(\mathbf{x}^*)\right] \qquad (7.2.8)
\end{aligned}$$

Plugging in Equation (7.2.8) in Equation (7.2.7), we get

$$\mathbb{E}_{i_t^s}\left[F(\mathbf{x}_{t+1}^s) - F(\mathbf{u})\right] \le \frac{2\eta L}{1-\eta L}(F(\mathbf{x}_t^s) - F(\mathbf{x}^*) + F(\tilde{\mathbf{x}}^{s-1}) - F(\mathbf{x}^*)) + \frac{\|\mathbf{x}_t^s - \mathbf{u}\|_2^2 - \mathbb{E}_{i_t^s}\left[\left\|\mathbf{x}_{t+1}^s - \mathbf{u}\right\|_2^2\right]}{2\eta} \tag{7.2.9}$$

Choosing $\eta = \frac{1}{7L}$, we get

$$\mathbb{E}_{i_t^s}\left[F(\mathbf{x}_{t+1}^s) - F(\mathbf{u})\right] \le \frac{1}{3}(F(\mathbf{x}_t^s) - F(\mathbf{x}^*) + F(\tilde{\mathbf{x}}^{s-1}) - F(\mathbf{x}^*)) + \frac{\|\mathbf{x}_t^s - \mathbf{u}\|_2^2 - \mathbb{E}_{i_t^s}\left[\left\|\mathbf{x}_{t+1}^s - \mathbf{u}\right\|_2^2\right]}{2\eta} \tag{7.2.10}$$

Telescoping equation (7.2.10) for $\mathbf{u} = \mathbf{x}^*$, from $t = 0$ to $m_s - 1$, dividing by m_s, and taking the full expectation, we get

$$3\mathbb{E}\left[\sum_{t=0}^{m_s-1} \frac{F(\mathbf{x}_{t+1}^s)}{m_s} - F(\mathbf{x}^*)\right] \leq \mathbb{E}\left[\left(\sum_{t=0}^{m_s-1} \frac{F(\mathbf{x}_t^s)}{m_s} - F(\mathbf{x}^*) + F(\tilde{\mathbf{x}}^{s-1}) - F(\mathbf{x}^*)\right) + \frac{\|\mathbf{x}_0^s - \mathbf{x}^*\|_2^2 - \left\|\mathbf{x}_{m_s}^s - \mathbf{x}^*\right\|_2^2}{2\eta/3.m_s}\right] \tag{7.2.11}$$

Rearranging Equation (7.2.11) we have

$$2\mathbb{E}\left[\sum_{t=0}^{m_s-1} \frac{F(\mathbf{x}_{t+1}^s)}{m_s} - F(\mathbf{x}^*)\right] \leq \mathbb{E}\left[\frac{(F(\mathbf{x}_0^s) - F(\mathbf{x}^*)) - (F(\mathbf{x}_{m_s}^s) - F(\mathbf{x}^*))}{m_s} + F(\tilde{\mathbf{x}}^{s-1}) - F(\mathbf{x}^*) + \frac{\|\mathbf{x}_0^s - \mathbf{x}^*\|_2^2 - \left\|\mathbf{x}_{m_s}^s - \mathbf{x}^*\right\|_2^2}{2\eta/3.m_s}\right] \tag{7.2.12}$$

Because F is convex, therefore $F(\tilde{\mathbf{x}}^s) \leq \frac{1}{m_s}\sum_{t=0}^{m_s-1} F(\mathbf{x}_{t+1}^s)$ from the definition of $\tilde{\mathbf{x}}^s = \frac{1}{m_s}\sum_{t=0}^{m_s-1} \mathbf{x}_{t+1}^s$. Also $\mathbf{x}_{m_s}^s = \mathbf{x}_0^{s+1}$. Therefore Equation (7.2.12) becomes

$$2\mathbb{E}\left[F(\tilde{\mathbf{x}}^s) - F(\mathbf{x}^*)\right] \leq \mathbb{E}\left[\frac{(F(\mathbf{x}_0^s) - F(\mathbf{x}^*)) - (F(\mathbf{x}_0^{s+1}) - F(\mathbf{x}^*))}{m_s} + F(\tilde{\mathbf{x}}^{s-1}) - F(\mathbf{x}^*) + \frac{\|\mathbf{x}_0^s - \mathbf{x}^*\|_2^2 - \left\|\mathbf{x}_0^{s+1} - \mathbf{x}^*\right\|_2^2}{2\eta/3.m_s}\right] \tag{7.2.13}$$

Using $m_s = 2m_{s-1}$, and rearranging the terms in Equation (7.2.13), we get

$$\begin{aligned}\mathbb{E}\left[F(\tilde{\mathbf{x}}^s) - F(\mathbf{x}^*) + \frac{\left\|\mathbf{x}_0^{s+1} - \mathbf{x}^*\right\|_2^2}{4\eta/3.m_s} + \frac{F(\mathbf{x}_0^{s+1}) - F(\mathbf{x}^*)}{2m_s}\right] &\leq 2^{-1}\mathbb{E}\left[F(\tilde{\mathbf{x}}^{s-1}) - F(\mathbf{x}^*) + \frac{\|\mathbf{x}_0^s - \mathbf{x}^*\|_2^2}{4\eta/3.m_{s-1}} + \frac{F(\mathbf{x}_0^s) - F(\mathbf{x}^*)}{2m_{s-1}}\right] \\ &\leq 2^{-S}\mathbb{E}\left[F(\tilde{\mathbf{x}}^0) - F(\mathbf{x}^*) + \frac{\left\|\mathbf{x}_0^1 - \mathbf{x}^*\right\|_2^2}{4\eta/3.m_0} + \frac{F(\mathbf{x}_0^1) - F(\mathbf{x}^*)}{2m_0}\right]\end{aligned} \tag{7.2.14}$$

Relaxing the inequality in Equation (7.2.14), we get

$$\mathbb{E}\left[F(\tilde{\mathbf{x}}^s) - F(\mathbf{x}^*)\right] \leq 2^{-S}\left[F(\tilde{\mathbf{x}}^0) - F(\mathbf{x}^*) + \frac{\left\|\mathbf{x}_0^1 - \mathbf{x}^*\right\|_2^2}{4\eta/3.m_0} + \frac{F(\mathbf{x}_0^1) - F(\mathbf{x}^*)}{2m_0}\right]$$

Since $\tilde{\mathbf{x}}^0 = \mathbf{x}_0^1 = \mathbf{x}^\phi$ and $m_0 \geq 1$, we have

$$\mathbb{E}\left[F(\tilde{\mathbf{x}}^s) - F(\mathbf{x}^*)\right] \leq \frac{F(\tilde{\mathbf{x}}^\phi) - F(\mathbf{x}^*)}{2^{S-1}} + \frac{\left\|\mathbf{x}_0^1 - \mathbf{x}^*\right\|_2^2}{2^S\frac{4\eta m_0}{3}}$$

Now let $F(\tilde{\mathbf{x}}^\phi) - F(\mathbf{x}^*) \leq \Delta$ and $\left\|\mathbf{x}^\phi - \mathbf{x}^*\right\|_2^2 \leq \Theta$. By setting $S = \log_2(\frac{\Delta}{\epsilon})$ and $m_0 = \frac{L\Theta}{\Delta}$ and with $\eta = \frac{1}{7L}$, we have

$$\mathbb{E}\left[F(\tilde{\mathbf{x}}^s) - F(\mathbf{x}^*)\right] \leq O(\epsilon)$$

Therefore SVRG++ has a gradient complexity of $O\left(nS + 2^S m_0\right) = O(n\log\left(\frac{\Delta}{\epsilon}\right) + \frac{L\Theta}{\epsilon})$, clearly an improvement over SGD for the same kind of objective.

8 Non convex Gradient method

For non-convex functions, we cannot talk about the convergence to the global minimizer in general, but instead we talk about how close we can reach a p^{th} order stationary point. Here we talk about convergence to an ϵ-first order stationary point as defined in Section 1.12.

The results shown for sub-gradient descent in Section 3 have the assumption that f is convex. But interestingly we can remove this assumption and still talk about convergence to approximate first order stationary points.

In this section f is a general smooth function (possibly non-convex) and the gradient descent algorithm remains the same, i.e.,

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)$$

Using the fact that f is L-smooth, we have

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 \\ &= f(\mathbf{x}_t) - \eta_t \|\nabla f(\mathbf{x}_t)\|_2^2 + \frac{L\eta_t^2}{2} \|\nabla f(\mathbf{x}_t)\|_2^2 \\ &= f(\mathbf{x}_t) - \eta_t \left(1 - \frac{L\eta_t}{2}\right) \|\nabla f(\mathbf{x}_t)\|_2^2 \end{aligned}$$

Choosing $\eta_t = \frac{1}{L}$ we get

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L} \|\nabla f(\mathbf{x}_t)\|_2^2 \tag{8.0.1}$$

Telescoping equation (8.0.1) from $t = 1$ to T,

$$f(\mathbf{x}_{T+1}) \leq f(\mathbf{x}_1) - \frac{1}{2L} \sum_{t=1}^{T} \|\nabla f(\mathbf{x}_t)\|_2^2 \tag{8.0.2}$$

$$\implies f(\mathbf{x}^*) \leq f(\mathbf{x}_1) - \frac{1}{2L} T \min_{1 \leq t \leq T} \|\nabla f(\mathbf{x}_t)\|_2^2 \tag{8.0.3}$$

$$\implies \min_{1 \leq t \leq T} \|\nabla f(\mathbf{x}_t)\|_2^2 \leq \frac{2L(f(\mathbf{x}_1) - f(\mathbf{x}^*))}{T} \tag{8.0.4}$$

Therefore to have $\min_{1 \leq t \leq T} \|\nabla f(\mathbf{x}_t)\|_2 \leq \epsilon$ we would require $T \geq \frac{2L(f(\mathbf{x}_1) - f(\mathbf{x}^*))}{\epsilon^2} = O\left(\frac{1}{\epsilon^2}\right)$ iterations.

References

[1] Zeyuan Allen-Zhu. Icml 2017 tutorial: Recent advances in stochastic convex and non-convex optimization.

[2] Zeyuan Allen-Zhu and Yang Yuan. Improved svrg for non-strongly-convex or sum-of-non-convex objectives. In *International conference on machine learning*, pages 1080–1089, 2016.

[3] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *In NIPS*. Citeseer, 2013.

[4] Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

[5] Mark Schmidt. Non smooth, non finite, and non convex optimization.

[6] Mark Schmidt. Smooth, finite, and convex optimization deep learning summer school.

[7] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79, 2013.

[8] Suvrit Sra. Introduction to large-scale optimization, June 2015.