Convergence of SGD Method with Fixed Step Size

Mason An Convex and Nonsmooth Optimization
May 6th 2024

Formalizing the Problem

$x \in dx$: dimension of input data, $y \in dy$: dimension of output data, $w \in d$: dimension of parameters

$h(.,.) : R^{d_x} \times R^d \rightarrow R^{d_y}$ Prediction function, which is our model

$l(h(x;w),y) : R^{d_y} \times R^{d_y} \rightarrow R$ Loss function for single data output

ε : single sample (x,y) from $R^{d_x} \times R^{d_y}$ or set of samples $(x_i, y_i)_{i \in S}$

$f(w;\varepsilon)$:, combined function of h, l

Expected Risk : $R(w) = E[f(w,\varepsilon)]$

Empirical Risk : $R_n(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w),$

Also denoted as $F(w)$, and is usually the objective function we want to optimize on for a machine learning task

Different Gradient Descent Approaches

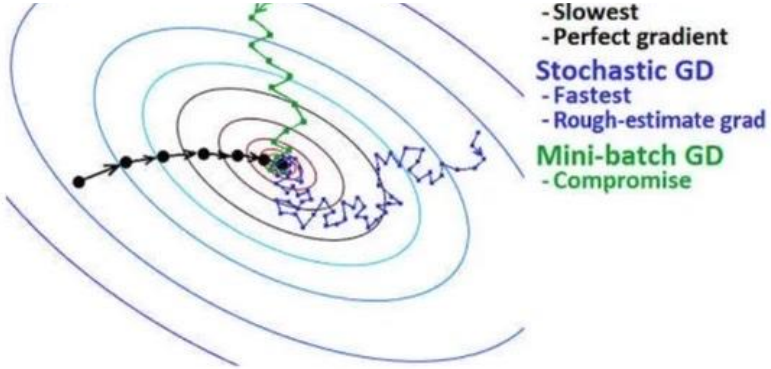
Stochastic Gradient Descent : $w_{k+1} \leftarrow w_k - \alpha_k \nabla f_{ik}(w_k)$. ik is a single data sample

Batch Gradient Descent : $w_{k+1} \leftarrow w_k - \alpha_k \nabla R_n(w_k) = w_k - \frac{a_k}{n} \sum_{i=1}^{n} \nabla f_i(w_i)$

$i \in 1, ..., n$.Whole dataset of sample

Mini Batch Gradient Descent : $w_{k+1} \leftarrow w_k - \alpha_k \nabla R_m(w_k) = w_k - \frac{\alpha_k}{m} \sum_{j=1}^{m} \nabla f_j(w_i)$

m is called batch size. $m < n$, in practice m take values $32, 64, 128...$



Stochastic Gradient Method

Algorithm 4.1 Stochastic Gradient (SG) Method

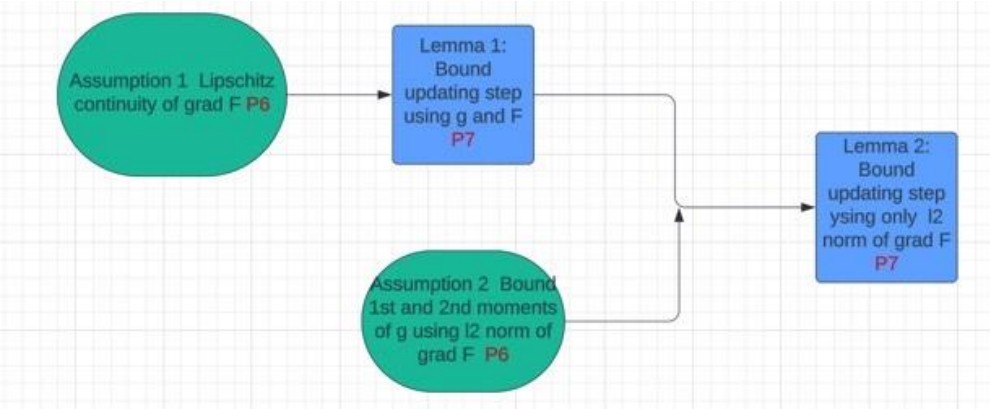
1: Choose an initial iterate w_1.
2: **for** $k = 1, 2, \ldots$ **do**
3: Generate a realization of the random variable ξ_k.
4: Compute a stochastic vector $g(w_k, \xi_k)$.
5: Choose a stepsize $\alpha_k > 0$.
6: Set the new iterate as $w_{k+1} \leftarrow w_k - \alpha_k g(w_k, \xi_k)$.
7: **end for**

$$g(w_k, \xi_k) = \begin{cases} \nabla f(w_k; \xi_k), & \text{Stochastic GD} \\ \frac{1}{n_k} \sum_{i=1}^{n_k} \nabla f(w_k; \xi_{k,i}), & \text{Mini Batch GD} \\ H_k \frac{1}{n_k} \sum_{i=1}^{n_k} \nabla f(w_k; \xi_{k,i}), & \text{Scaled Mini Batch GD} \end{cases}$$

Here g represents stochastic gradient, which is a an unbiased estimator of grad F.

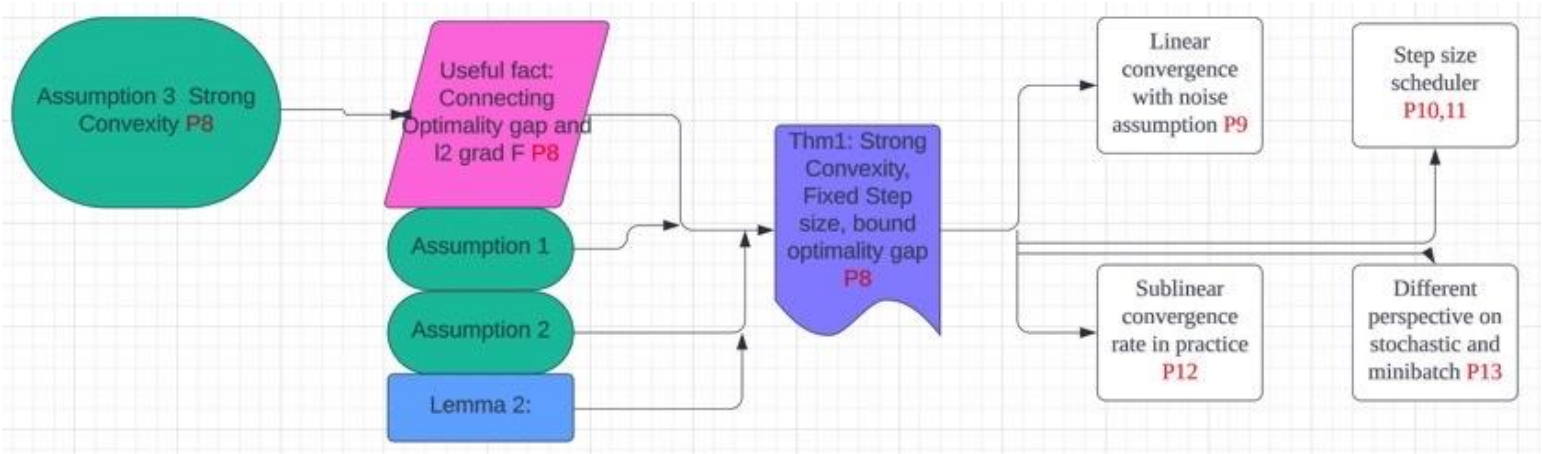
Convergence analysis, mind map

Step 1. Bound update step from above, Using the expression containing F only.



Step 3. What if the Objective is non Convex?

Step 2. For Strong Convexity Objectives, get the bound for optimality gap.



Convergence Analysis

Step 1. Bound update step from above, Using the expression containing F only.

Assumptions:

Assumption 1

$F \in C^1$

∇F Lipschitz continous : $||\nabla F(w) - \nabla F(\bar{w})||_2 \leq L||w - \bar{w}||_2, \forall w, \bar{w} \in R^d$

Assumption 2

$(1).F$ bounded below by F_{inf}

$(2)\exists \mu_G \geq \mu > 0, \forall k \in N, s.t$

$(2a)\nabla F(w_k)^T E_{\varepsilon_k}[g(w_k, \varepsilon_k)] \geq \mu||\nabla F(w_k)||_2^2$

$(2b)||E_{\varepsilon_k}[g(w_k, \varepsilon_k)]||_2 \leq \mu_G||\nabla F(w_k)||_2$

$\exists M \geq 0, M_v \geq 0, s.t, \forall k \in N$

$(3).V_{\varepsilon_k}[g(w_k, \varepsilon_k)] \leq M + M_V||\nabla F(w_k)||_2^2$

$where, V_{\varepsilon_k}[g(w_k, \varepsilon_k)] = E_{\varepsilon_k}[||g(w_k, \varepsilon_k)||_2^2] - ||E_{\varepsilon_k}[g(w_k, \varepsilon_k)]||_2^2$

Direction of g close to direction of grad F. mu = 1 iff g is unbiased estimate of F

L2 norm of g not too big

Variance of g can represent the noise in SGD algorithm.Gives an upper bound for second moment of g. M = 0 if noise decay with ||grad F||. Smaller mini batch, the greater the variance.

Two Useful lemmas

Lemma 1 (Derived From Assumption 1)Bound update step using g and grad F

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\alpha_k \nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] + \tfrac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2].$$

Lemma 2 (Derived From Assumption 2 and Lemma 1)Bound update step using only grad F

$$\begin{aligned}\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) &\leq -\mu\alpha_k \|\nabla F(w_k)\|_2^2 + \tfrac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2] \\ &\leq -(\mu - \tfrac{1}{2}\alpha_k L M_G)\alpha_k \|\nabla F(w_k)\|_2^2 + \tfrac{1}{2}\alpha_k^2 L M.\end{aligned}$$

Little conclusion: Based on Lemma 2, we can have an upper bound for all updating step with a single expression depending only on F. And the optimization process continuous in a Markovian manner.
Why grad F is so important? Because we can establish a connection to optimality gap through it, which is crucial for stating a convergence theorem.

Tips for proof: For Lemma 1, use fundamental theorem for line integrals $F(w) = F(\overline{w}) + \int_0^1 \frac{\partial F(\overline{w} + t(w - \overline{w}))}{\partial t}\, dt$

$$F(w) \leq F(\overline{w}) + \nabla F(\overline{w})^T (w - \overline{w}) + \tfrac{1}{2} L \|w - \overline{w}\|_2^2 \quad \text{for all} \quad \{w, \overline{w}\} \subset \mathbb{R}^d.$$

For Lemma 2, replace g using F and bounds in assumption

Step 2. For Strong Convexity Objectives, get the bound for optimality gap.

Assumption 3, Strong Convexity Objective

$\exists c > 0, s.t$

$$F(\bar{w}) \geq F(w) + \nabla F(w)^T(\bar{w} - w) + \frac{1}{2}c||\bar{w} - w||_2^2, \forall(\bar{w}, w) \in R^d \times R^d$$

An useful fact based on strong convexity assumption

$$2c(F(w) - F_*) \leq ||\nabla F(w)||_2^2$$

Where F_* is the optimal point for F

Total Expectation of $F(w_k)$

$$E[F(w_k)] = E_{\varepsilon_1} E_{\varepsilon_2}, ..., E_{\varepsilon_{k-1}}[F(w_k)]$$

Then we can get a main theorem about optimality gap under strong convexity assumption and fixed step size

Theorem 1: (Proved by assumption 1,2,3, Lemma 2, and the fact above)

Compare to what we get in lecture(by setting noise = 0 and step size = 1/M

Under assumption 1,2,3, and a fixed step size

$$0 \leq \bar{\alpha} \leq \frac{\mu}{LM_G}$$

$$f(x^{(\ell)}) - p^* \leq \left(1 - \frac{m}{M}\right)^{\ell} \left(f(x^{(0)}) - p^*\right).$$

The optimality gap $\forall k \in N$ satisfies

$$E[F(w_k) - F_*] \leq \frac{\bar{\alpha}LM}{2c\mu} + (1 - \bar{\alpha}c\mu)^{k-1}(F(w_1) - F_*) - \frac{\bar{\alpha}LM}{c\mu}) \rightarrow_{k\rightarrow\infty} \frac{\bar{\alpha}LM}{2c\mu}$$

Linear Convergence with Noise Assumption

Observe the expression

$$E[F(w_k) - F_*] \leq \frac{\bar{\alpha}LM}{2c\mu} + (1 - \bar{\alpha}c\mu)^{k-1}(F(w_1) - F_*) - \frac{\bar{\alpha}LM}{c\mu}) \rightarrow_{k\rightarrow\infty} \frac{\bar{\alpha}LM}{2c\mu}$$

Decay Rate $(1 - \bar{\alpha}c\mu) \in (0, 1)$

If there's no noise (we are doing whole batch GD), or noise decay with $||\nabla F(w_k)||_2^2$

$$\Rightarrow M = 0 \Rightarrow E[F(w_k) - F_*] \leq (1 - \bar{\alpha}c\mu)^{k-1}(F(w_1) - F_*)$$

$$\Rightarrow E_{\varepsilon_{k+1}}[F(w_k) - F_*] \leq (1 - \bar{\alpha}c\mu)[F(w_k) - F_*]$$

And this shows the algorithm converges linearly to optimal value under noise assumptions!

However, in practice, when we implement mini batch SGD, the noise is inevitable. And we can only guarantee convergence to a nbhd of optimal point linearly. The size of this nbhd is represented by

$$E_{\varepsilon_{k+1}}[F(w_k) - F_*] \rightarrow_{n\rightarrow\infty} \frac{\bar{\alpha}LM}{2c\mu}$$

And we call it as asymptotic optimality gap

Key factor for convergence speed and accuracy: step size

Theory behind Step Size Scheduler

Based on the theorem, we define the asymptotic optimality gap as: $F_\alpha = \frac{\alpha LM}{2c\mu}$

To balance the convergence accuracy and convergence speed, we propose the following strategy:

Choose α_1, calculate $F_{\alpha_1} = \frac{\alpha_1 LM}{2c\mu}$

Iterate until step k_2, such that $E[F(w_{k_2}) - F_*] \leq 2F_{\alpha 1}$

Update $\alpha_2 = \frac{1}{2}\alpha_1$, continou the above iteration

And by analysing this strategy, we can find that

Stepsize Schedule $\{\alpha_{r+1}\} = \{\alpha_1 2^{-r}\}$

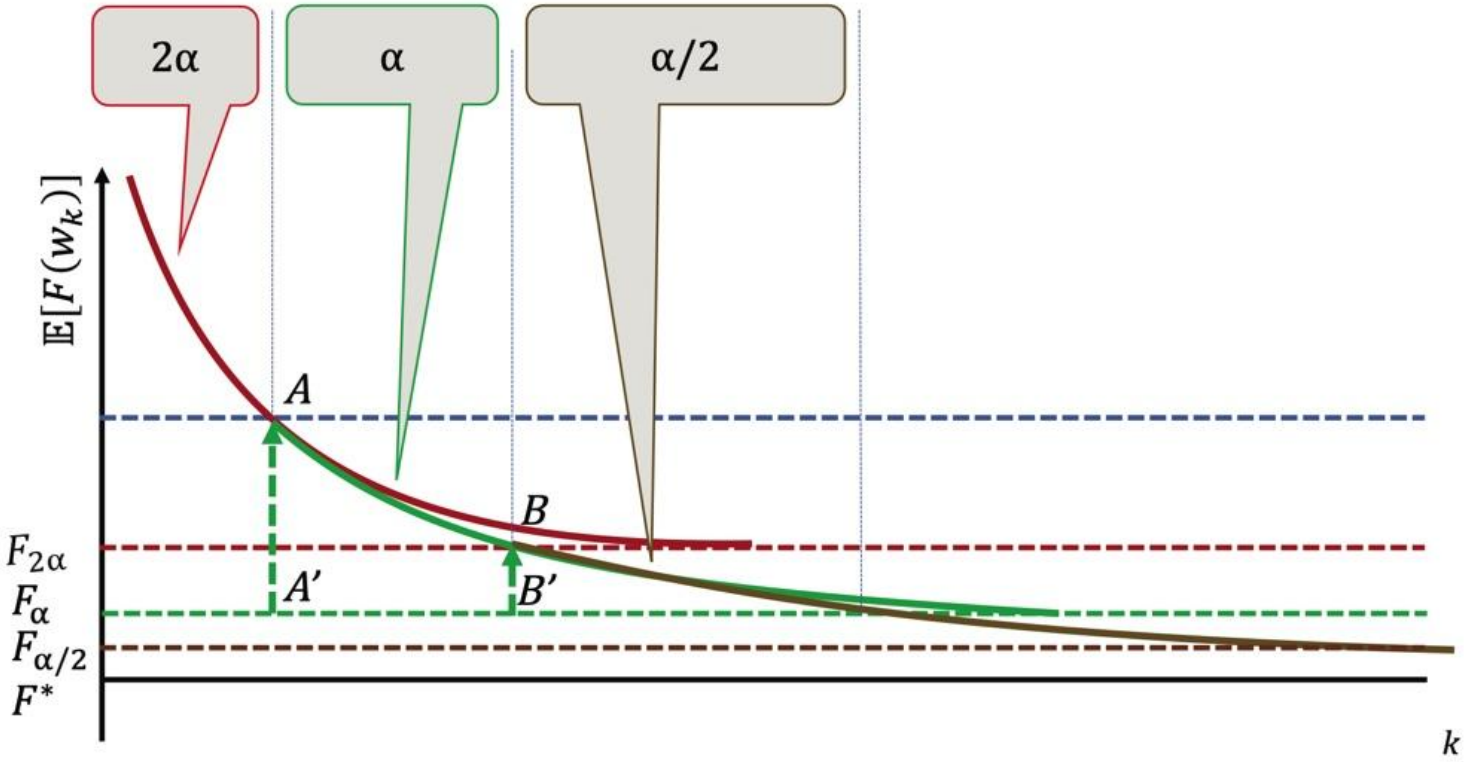
$\{F_{\alpha_r}\} = \{\frac{\alpha_r LM}{2c\mu}\} \rightarrow_{\alpha_r \rightarrow 0} 0$

So by following this strategy, we can finally approach to a neighborhood of optimal point with its size converging to 0. Meanwhile, we don't lose convergence speed in early stage.

LR scheduler in ML:

```
scheduler = StepLR(optimizer, step_size=30, gamma=0.1)
```

A plot illustration



Sublinear Rate of Convergence using the Strategy

Can we analyse how many steps of iteration needed in this strategy?

First we notice:

$$\mathbb{E}[F(w_{k_{r+1}}) - F_*] \leq 2F_{\alpha_r}, \quad \text{where} \quad \mathbb{E}[F(w_{k_r}) - F_*] \approx 2F_{\alpha_{r-1}} = 4F_{\alpha_r}.$$

Also, applying Theorem 1, we have:

$$E[F(w_{k_{r+1}}) - F_*] \leq \frac{\bar{\alpha}LM}{2c\mu} + (1 - \alpha\bar{c}\mu)^{k_{r+1}-k_r}(E[F(w_{k_r}) - F_*] - \frac{\bar{\alpha}LM}{2c\mu}) \leq 2F_{\alpha_r}$$

$$\Rightarrow (1 - \alpha_r c\mu)^{(k_{r+1}-k_r)}(4F_{\alpha_r} - F_{\alpha_r}) \leq F_{\alpha_r}$$

$$\Rightarrow k_{r_1} - k_r \geq \frac{log(3)}{log(1 - \alpha_r c\mu)} \approx \frac{log(3)}{\alpha_r c\mu} = O(2^r)$$

In other words, every time we shrink the stepsize by ½., the iteration number we take doubles!

Also the reason why in practice, SG face sub linear rate of convergence,

Different from what we do in practice, where we usually set a fixed decay step as a hyper parameter. Theoretically speaking, it can give a better guarantee for convergence, compared to what we do in practice.

Exist Theorem 2 about optimality gap under Strong convex objective and diminishing step size (Difficult)

A different perspective for stochastic and mini batch gradient descent

Assume the variance for stochastic gradient on each sample are the same

Let the mini batch size be $n_k \ll n$, and $n_k = |S_k|$

$$V_{\varepsilon_k}[g(w_k, \varepsilon_k)] = V_{\varepsilon_k}[\frac{1}{n_k}\sum_{i \in S_k} \nabla f(w_k, \varepsilon_{k,i})] = \frac{1}{n_k^2}\sum_{i \in S_k} V_{\varepsilon_k}[\nabla f(w_k, \varepsilon_{k,i})] = \frac{1}{n_k^2} n_k V_{\varepsilon_k} \nabla f(w_k, \varepsilon_{k,i}) = \frac{1}{n_k} V_{\varepsilon_k} \nabla f(w_k, \varepsilon_{k,i})$$

For assumption 2,

$$M \to \frac{M}{n_k}, M_V \to \frac{M_V}{n_k}$$

For $0 \leq \bar{\alpha} \leq \dfrac{\mu}{LM_G}$

Cannot always compensate for the higher per iteration loss cost of a mini batch SG method by selecting a larger step size

According to Theorem 1:

Mini batch SG, stepsize $\bar{\alpha}$

$$E[F(w_k) - F_*] \leq \frac{\bar{\alpha}LM}{2c\mu n_k} + [1 - \bar{\alpha}c\mu]^{k-1}(F(w_1) - F_* - \frac{\bar{\alpha}LM}{2c\mu n_k})$$

Same asymptotic optimality gap

Slower contraction scale for Simple SG

Simple SG with stepsize $\bar{\alpha}/n_k$

$$E[F(w_k) - F_*] \leq \frac{\bar{\alpha}LM}{2c\mu n_k} + [1 - \frac{\bar{\alpha}c\mu}{n_k}]^{k-1}(F(w_1) - F_* - \frac{\bar{\alpha}LM}{2c\mu n_k})$$

Less computation per step for SG

Step 3. What if the Objective is non Convex?

Theorem 3 (Using assumption 1,2, Lemma 2)

Under assumption 1,2

$$0 < \bar{\alpha} \leq \frac{\mu}{LM_G}$$

$$E[\frac{1}{K}\sum_{k=1}^{K}||\nabla F(w_k)||_2^2] \leq \frac{\bar{\alpha}LM}{\mu} + \frac{2(F(w_1) - F_{inf})}{K\mu\alpha} \rightarrow_{K\rightarrow+\infty} \frac{\bar{\alpha}LM}{\mu}$$

Based on total Lemma 2, take total expectation and sum up.

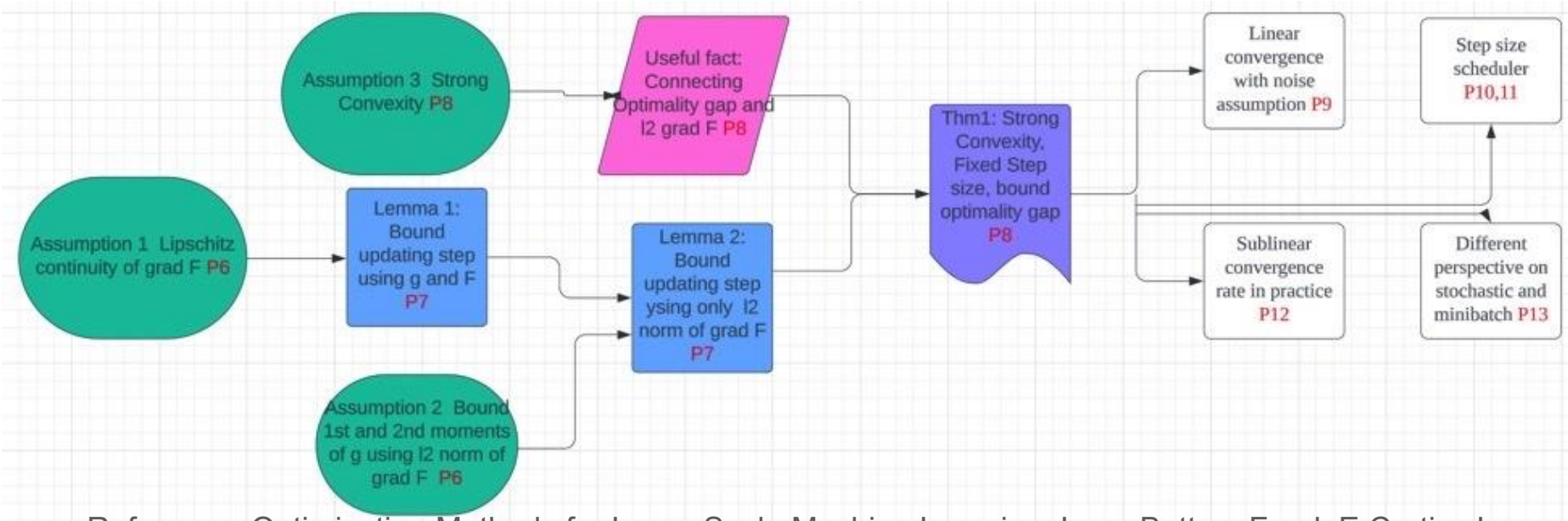
Convex case: bound optimality gap
Non Convex case: bound average sum of normed gradient

Implication of the theorem:

1. Finite sum of gradient norm: $||\nabla F(w_k)||_2^2 \rightarrow_{k\rightarrow\infty} 0$

2. When K increases, the SG method spend more time in regions with small gradient

Conclusion



Reference: Optimization Methods for Large-Scale Machine Learning Leon Bottou, Frank E Curtis, Jorge Nocedal
Note on Gradient Method for Smooth Convex Minimization Michael L Overton
Following the derivation in Boyd and Vandenberghe
Website: https://sweta-nit.medium.com/batch-mini-batch-and-stochastic-gradient-descent-e9bc4cacd461