Stochastic Gradient Descent for Large Scale Machine Learning Problems

Ravi Raghavan & Lakshitha Ramanayake

Abstract

This paper provides a comprehensive review of Stochastic Gradient Descent (SGD) and its variants, highlighting their pivotal role in convex optimization for machine learning applications. Beginning with essential theoretical underpinnings such as convexity, differentiability, and smoothness, this paper sets the stage for a deeper exploration of SGD, including its standard form and minibatch variant. These foundational methods are examined for their efficiency in processing large datasets, with a particular focus on their convergence properties and stability under various conditions. Advanced modifications of SGD, such as Momentum and Nesterov Accelerated Gradient (NAG), are discussed, which enhance performance by incorporating mechanisms that expedite convergence and improve algorithmic stability. Additionally, adaptive gradient methods like AdaGrad and RMSprop are explored, which adjust learning rates per parameter based on the accumulation of past gradients, thus optimizing the training process across different scenarios. By integrating theoretical analysis with empirical evidence, this study not only deepens the understanding of these techniques but also provides practical guidance for their application, making it a valuable resource for both new researchers and seasoned practitioners in the field of machine learning optimization.

I. Introduction

A. Rationale

SGD is a pivotal optimization technique that has revolutionized the way machine learning models are trained, particularly in the realm of large-scale data analysis and deep learning. By updating model parameters incrementally using only a subset of the data at each iteration, SGD dramatically reduces computational costs and memory requirements, enabling the training of models on datasets of virtually any size. This efficiency has paved the way for applications in real-time and online learning, where models are dynamically updated as new data becomes available continuously. The efficiency of SGD positions itself as a promising tool for overcoming the computational hurdles we encounter in current applications. This realization has motivated a thorough investigation of SGD, aiming to understand its mechanisms and to explore its applicability within various use cases in convex optimization.

B. Scope

In contemporary Machine Learning Literature, SGD has typically been used to train highly complex and non-convex models, such as Deep Learning models. Interestingly, when trained with SGD, these models exhibit good generalization ability.

This study aims to conduct a comprehensive investigation into the mathematical foundations underpinning SGD, to analyze the algorithmic stability of SGD to understand its generalization ability in convex settings, to study the convergence behavior of SGD, and to experiment with the practical applicability of SGD for convex optimization problems via numerical simulations.

First, the study will discuss the basic concepts underlying Gradient Descent to give a brief introduction. Subsequently, the study will analyze how SGD acts as a stochastic approximation for traditional Gradient Descent. Additionally, the study will also explore different variants of SGD such as momentum, nestrov accelerated gradient, RMSProp optimization and Adagrad optimization, discussing which variants are best suited for various used cases and why.

Another focus of this study will be to understand why SGD is effective for convex optimization, specifically delving into the mathematical principles and properties that underlie its performance. By exploring the convergence guarantees provided by convexity and the properties of stochastic gradients, this study seeks to provide a deeper understanding of the mechanisms through which SGD can efficiently find optimal solutions in convex settings. Additionally, this research will investigate the stability of SGD as an optimization algorithm. Finally, through

numerical experiments, the study will seek to run Stochastic Gradient Descent on various convex problems, hoping to observe the real-time behaviour of Stochastic Gradient Descent.

II. Convex Optimization Basic Concepts

Citation Note: All theory for this section has been taken from [1]. The proofs in this book are rather succinct and have been further developed and analyzed in this section of the paper.

A. Notation

- $[x]^2$ is the vector obtained by raising each element of the vector x to the 2nd power.
- $x \cdot y$ is the vector obtained by doing an element wise multiplication of the vectors x and y
- $x \geq y$, $x \leq y$, where x and y are vectors, means that each component of x is greater than or equal to/less than or equal to the corresponding component of y
- $x = c$, where x is a vector and c is a scalar, means that each component of x is equal to c
- $x \geq c$, $x \leq c$, where x is a vector and c is a scalar, means that each component of x is greater than or equal to/less than or equal to c
- **Expected Value**: For $X \in \mathbb{R}^n$, $\mathbb{E}[X] = [\mathbb{E}[X_1], \mathbb{E}[X_2], \ldots, \mathbb{E}[X_n]]^T$
- **Element-Wise Variance**: For $X \in \mathbb{R}^n$, $\mathbb{V}[X] = \mathbb{E}[\|X - \mathbb{E}[X]\|_2^2]$

Note: Throughout this paper, expected value of the gradient is taken over the samples/batches. $E[\nabla f_i(x)]$ takes the expected value of the gradient over the randomness of sample i. $E[\nabla f_B(x)]$ takes the expected value of the gradient over the randomness of batch B. Furthermore, $E[x^{(t+1)}]$ represents the expected value of the next iterate due to the stochastic gradient(i.e. randomness of sample/batch), given the value of the current iterate(i.e. $x^{(t)}$).

Note: Throughout this paper, the term "Variance" is used to denote "Element-Wise Variance". The term "Variance" is just used as a short-form and as a mere convenience.

B. Background Concepts

Note: The Proofs for all the Lemmas in the Subsection Background Concepts are in the Appendix

Differentiability

Definition 1 (Jacobian). Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ be differentiable, and $x \in$ **int dom**f. $Df(x)$ is the derivative or **Jacobian** of f at x, which is the matrix defined by its first partial derivatives:

$$[Df(x)]_{ij} = \frac{\partial f_i}{\partial x_j}(x), \quad for \;\; i = 1, \ldots, m, j = 1, \ldots, n$$

Definition 2 (Gradient). If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, then $Df(x) \in \mathbb{R}^{1 \times d}$ is a row vector, whose transpose is called the **gradient** of f at x:

$$\nabla f(x) = Df(x)^T \in \mathbb{R}^{d \times 1}$$

Definition 3 (Hessian). Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable, and $x \in$ **int dom**f. $\nabla^2 f(x)$ is the **Hessian** of f at x, which is the matrix defined by its second-order partial derivatives:

$$[\nabla^2 f(x)]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x), \quad for \;\; i = 1, \ldots, n, \;\; j = 1, \ldots, n$$

Definition 4 (Lipschitz). Let $f : \mathbb{R}^n \to \mathbb{R}^m$, and $L > 0$. f is L-**Lipschitz** if

$$\forall x, y \in \textbf{dom} f, \;\; \|f(y) - f(x)\| \leq L\|y - x\|$$

Convexity

Definition 5 (Jensen's Inequality). A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if **dom** f is a convex set and if for all $x, y \in domf$, and θ with $0 \leq \theta \leq 1$

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$$

Definition 6 (First Order Condition of Convexity). A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if and only if **dom** f is a convex set and

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

holds for all $x, y \in domf$

Definition 7 (Second Order Condition of Convexity). A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if and only if **dom** f is a convex set and its Hessian is positive semi-definite:

$$\forall x \in domf, \;\; \nabla^2 f(x) \succeq 0$$

Strong Convexity

Definition 8 (Jensen's Inequality for Strong Convexity). A function $f : \mathbb{R}^n \to \mathbb{R}$ is *p-strongly convex* if **dom** f is a convex set and if for all $x, y \in domf$, and θ with $0 \leq \theta \leq 1$,

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) - \frac{p}{2}\theta(1-\theta)\|x - y\|_2^2$$

Definition 9 (First Order Condition for Strong Convexity). A function $f : \mathbb{R}^n \to \mathbb{R}$ is *p-strongly convex* if and only if **dom** f is a convex set and

$$f(y) \geq f(x) + \nabla(f(x))^T (y - x) + \frac{p}{2}\|y - x\|_2^2$$

holds for all $x, y \in domf$

Definition 10 (Second Order Condition for Strong Convexity). A function $f : \mathbb{R}^n \to \mathbb{R}$ is *p-strongly convex* if and only if **dom** f is a convex set and

$$\forall x \in domf, \;\; \nabla^2 f(x) \succeq pI$$

Smoothness

Definition 11 (L-Smooth Functions). A function $f : \mathbb{R}^n \to \mathbb{R}$ and $L > 0$ is L-Smooth if it is differentiable and if $\nabla f : \mathbb{R}^n \to \mathbb{R}$ is $L-$Lipschitz:

$$\forall x, y \in \mathbb{R}^n, \;\; \|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|$$

Consequently, $L-$Smooth functions have a quadratic upper bound:

$$\forall x, y \in \mathbb{R}^n, f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$$

Lemma 12: If f is $L-$smooth and $\gamma > 0$ then,

$$\forall x, y \in \mathbb{R}^n, \;\; f(x - \gamma \nabla f(x)) - f(x) \leq -\gamma(1 - \frac{\gamma L}{2})\|\nabla f(x)\|^2$$

Smoothness and Convexity

Lemma 13: If $f : \mathbb{R}^d \to \mathbb{R}$ is convex and $L-$smooth, then $\forall x, y \in \mathbb{R}^d$:

$$\frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle$$

Gradient Descent Algorithm

Let $x^{(0)} \in$ **dom**f, and let $\gamma_t > 0$ be a step size. The **Gradient Descent (GD)** algorithm defines a sequence

$(x^{(t)})_{t\in\mathbb{N}}$ satisfying

$$x^{(t+1)} = x^{(t)} - \gamma_t \nabla f(x^{(t)})$$

Function Definitions

Sum of Functions. Let $f : \mathbb{R}^n \to \mathbb{R}$ such that

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

where $f_i : \mathbb{R}^n \to \mathbb{R}$.
Whenever the paper refers to a function f as a "Sum of Functions", it means that f can be expressed as above.
Sum of Convex Functions. Let $f : \mathbb{R}^n \to \mathbb{R}$ such that

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

where $f_i : \mathbb{R}^n \to \mathbb{R}$ and f_i is convex.
Whenever the paper refers to a function f as a "Sum of Convex Functions", it means that f can be expressed as $f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$ where $f_i(x)$ is Convex.

Sum of L-Smooth Functions. Let $f : \mathbb{R}^n \to \mathbb{R}$ such that

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

where $f_i : \mathbb{R}^n \to \mathbb{R}$ and f_i is L_i smooth. Let $L_{max} = \max_{\{}\, 1, \ldots, m\}\{L_i\}$

Whenever the paper refers to a function f as a "Sum of Convex Functions",it means that $f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$ where $f_i(x)$ is L_i-Smooth.

Sum of L-Smooth Functions and Convex Functions. Let $f : \mathbb{R}^n \to \mathbb{R}$ such that

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

where $f_i : \mathbb{R}^n \to \mathbb{R}$ and f_i is L_i smooth. Let $L_{max} = \max_{\{}\, 1, \ldots, m\}\{L_i\}$. Furthermore each f_i is convex.

Whenever the paper refers to a function f as a "Sum of L-Smooth Functions and Convex Functions", it means that $f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$ where $f_i(x)$ is L_i-Smooth and each $f_i(x)$ is Convex

Stochastic Gradient Descent Algorithm

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a sum of functions. It is assumed that $\arg\min f \neq \emptyset$ and that f_i is not unbounded below. Let $x^{(0)} \in \mathbf{dom} f$, and let $\gamma_t > 0$ be a sequence of step sizes. The **Stochastic Gradient Descent (GD)** algorithm defines a sequence $(x^{(t)})_{t\in\mathbb{N}}$ satisfying

$$i_t \in \{1, \ldots, m\}$$

$$x^{(t+1)} = x^{(t)} - \gamma_t \nabla f_{i_t}(x^{(t)})$$

Note: i_t is sampled with probability $\frac{1}{m}$ and $E[\nabla f_{i_t}(x)] = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(x) = \nabla f(x)$.

Minibatch Stochastic Gradient Descent

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a sum of functions. It is assumed that $\arg\min f \neq \emptyset$ and that f_i is not unbounded below. Let $x^{(0)} \in \mathbf{dom} f$, let $b \in [1, m]$ be the batch size, and let $\gamma_t > 0$ be a sequence of step sizes. The **Minibatch**

Stochastic Gradient Descent (Minibatch SGD) algorithm defines a sequence $(x^{(t)})_{t\in\mathbb{N}}$ satisfying

$$B_t \subset \{1,\dots,m\}$$

$$x^{(t+1)} = x^{(t)} - \gamma_t \nabla f_{B_t}(x^{(t)})$$

Note: B_t is sampled uniformly among all sets of size b. This means that given a batch of size b, it has a probability of $\frac{1}{\binom{m}{b}}$ of being selected

$$\nabla f_{B_t}(x^{(t)}) = \frac{1}{b}\sum_{i\in B_t} \nabla f_i(x^{(t)})$$

Observation: For any given value of $i \in \{1,\dots,m\}$, $\nabla f_i(x^{(t)})$ will be sampled in exactly $\binom{m-1}{b-1}$ of the total $\binom{m}{b}$ mini batches
Property: $\binom{m}{b} = \binom{m-1}{b-1}\cdot\frac{m}{b}$

$$\mathbb{E}[\nabla f_{B_t}(x^{(t)})] = \frac{1}{\binom{m}{b}}\sum_{B\subset\{1\dots n\},|B|=b} \nabla f_B(x^{(t)}) = \frac{1}{\binom{m}{b}}(\frac{1}{b})\binom{m-1}{b-1}\sum_{i=1}^{m}\nabla f_i(x^{(t)}) = \frac{1}{m}\sum_{i=1}^{m}\nabla f_i(x^{(t)}) = \nabla f(x^{(t)})$$

Expected Smoothness and Variance [SGD]

Lemma 14: Let $f:\mathbb{R}^n\to\mathbb{R}$ be a sum of Convex functions and a sum of L-Smooth functions

$$\forall x,y\in \mathbf{dom} f,\ \ \frac{1}{2L_{max}}\mathbb{E}[\|\nabla f_i(y)-\nabla f_i(x)\|^2] \le f(y)-f(x)-\langle\nabla f(x), y-x\rangle$$

Lemma 15: Let $f:\mathbb{R}^n\to\mathbb{R}$ be a sum of Convex functions and a sum of L-Smooth functions. When $x=x^*$, where $x^*\in\arg\min f$, and $y=x$

$$\frac{1}{2L_{max}}\mathbb{E}[\|\nabla f_i(x)-\nabla f_i(x^*)\|^2] \le f(x)-\inf f$$

Definition 16 (Interpolation). Let $f:\mathbb{R}^n\to\mathbb{R}$ be a Sum of Functions. **Interpolation** holds if there exists a common $x^*\in\mathbb{R}^n$ such that $f_i(x^*)=\inf f_i, \forall i=1,\dots m$.
Lemma 17. Let $f:\mathbb{R}^n\to\mathbb{R}$ be a Sum of Functions. If interpolation holds at $x^*\in\mathbb{R}^n$, then $x^*\in\arg\min f$
Definition 18(Function Noise). Let $f:\mathbb{R}^n\to\mathbb{R}$ be a Sum of Functions. The **Function Noise**, Δ_f^*, is defined as:

$$\Delta_f^* = \inf f - \frac{1}{m}\sum_{i=1}^{m}\inf f_i$$

Lemma 19.

$$\Delta_f^* \ge 0$$

Interpolation Holds if and only if $\Delta_f^* = 0$

Definition 20.(Gradient Noise) Let $f:\mathbb{R}^n\to\mathbb{R}$ be a Sum of $L-$ Smooth Functions. **Gradient Noise**, σ_f^*, is defined as such:

$$\sigma_f^* = \inf_{x^*\in\arg\min f}\mathbb{V}[\nabla f_i(x^*)]$$

Lemma 21. Let $f:\mathbb{R}^n\to\mathbb{R}$ be a Sum of $L-$ Smooth Functions and a Sum of Convex Functions

$$\sigma_f^* = \mathbb{V}[\nabla f_i(x^*)], \forall x^*\in\arg\min f$$

Interpolation Holds if and only if $\sigma_f^* = 0$

Lemma 22. Let $f:\mathbb{R}^n\to\mathbb{R}$ be a Sum of $L-$ Smooth Functions
1) $\sigma_f^* \le 2L_{max}\Delta_f^*$

2) If each f_i is p strongly convex, then $2p\Delta_f^* \leq \sigma_f^*$

Lemma 23. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions

$$\forall x \in \mathbb{R}^d, \ \ \mathbb{E}[\|\nabla f_i(x)\|^2] \leq 2L_{max}(f(x) - \inf f) + 2L_{max}\Delta_f^*$$

Lemma 24. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions and a Sum of Convex Functions

$$\forall x \in \mathbb{R}^d, \ \ \mathbb{E}[\|\nabla f_i(x)\|^2] \leq 4L_{max}(f(x) - \inf f) + 2\sigma_f^*$$

Expected Smoothness and Variance [Minibatch SGD]

Definition 25: Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions Let $b \in [1, m]$. Then f is L_b smooth in expectation if

$$\forall x, y \in \mathbb{R}^n, \frac{1}{2L_b}\mathbb{E}_B[\|\nabla f_B(y) - \nabla f_B(x)\|^2] \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle$$

If $y = x$ and $x = x^*$, where $x^* \in \arg\min f$, a function being L_b smooth indicates that:

$$\frac{1}{2L_b}\mathbb{E}_B[\|\nabla f_B(x) - \nabla f_B(x^*)\|^2] \leq f(x) - f(x^*) - \langle \nabla f(x^*), x - x^* \rangle$$

Since $\nabla f(x^*) = 0$

$$\frac{1}{2L_b}\mathbb{E}[\|\nabla f_B(x) - \nabla f_B(x^*)\|^2] \leq f(x) - \inf f$$

Definition 26: Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions. **Minibatch Gradient Noise** is defined as such:

$${\sigma_b}^* = \inf_{x^* \in \arg\min f} \mathbb{V}[\nabla f_B(x^*)]$$

Lemma 27. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions and Convex Functions.

$$\sigma_b^* = \mathbb{V}[\nabla f_B(x^*)], \forall x^* \in \arg\min f$$

Lemma 28. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions and a Sum of Convex Functions

$$\mathbb{E}[\|\nabla f_B(x)\|^2] \leq 4L_b(f(x) - \inf f) + 2\sigma_b^*$$

III. Variants on Stochastic Gradient Descent

Cite Note: This section is based on the discussion of variants of stochastic gradient method on [3].

In this section a brief discussion on the variants on the algorithmic variations on which are used in stochastic gradient descents.

A. Momentum

The key idea behind the momentum method is to incorporate a fraction of the previous update vector into the current update. The previous update step is denoted as u_{t-1}. The current update step u_t can then be described according to 1.

$$u^t = \alpha u^{t-1} + \gamma \nabla f(x^{(t)}) \tag{1}$$

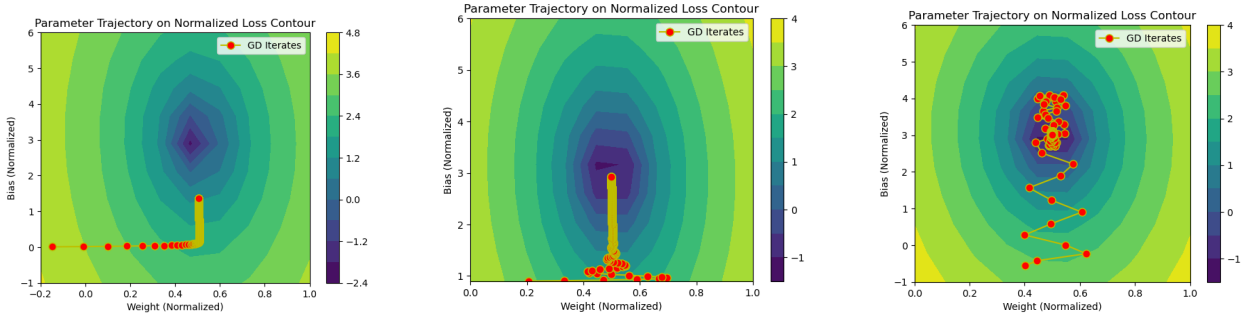
where the $\alpha \in (0, 1)$ is the momentum coefficient, which determines how much of the previous velocity is retained, and γ is the learning rate. This coefficient helps in accumulating a direction of persistent descent, smoothing over the updates. With this step, the current iterate takes the form,

$$x^{t+1} = x^t - u^t \tag{2}$$

The term u_t is also called the velocity term. The momentum term αu^{t-1} in the velocity term serves as a memory of past gradients:

- If gradients continue pointing in the same direction, the velocity grows in magnitude, allowing for faster convergence.
- If gradients change direction, the velocity's magnitude decreases, which helps mitigate oscillations and overshooting in steep regions of the parameter space.

This approach effectively dampens the oscillations and accelerates convergence towards the minimum of the loss function, particularly in landscapes where the surface curves more steeply in one dimension than in another. To visualize the effect of momentum in optimization let's imagine a ball rolling down a slope. If the slope does not have any turns the ball will keep accumulating velocity till it reaches the bottom. However, if there are turns the ball will slow down to navigate more efficiently. Figure 1 illustrates the iteration of stochastic gradient descent on the contour plots of the loss function.



(a) $\gamma = 0.001$, $\alpha = 0$ ***(b)*** $\gamma = 0.001$, $\alpha = 0.8$ ***(c)*** $\gamma = 0.001$, $\alpha = 0.9$

Fig. 1: Comparison between SGD with and without momentum

B. Nestrov Accelerated Gradients (NAG)

NAG is also a momentum-based variant of SGD. The main difference between the momentum method and NAG lies in the gradient calculation stage. It was seen that in the momentum method, the update happens at x^t depending on the previous velocity v^{t-1} and the gradient of the function at x^t (1). In NAG the calculation of the gradient is done at a point ahead given by $\nabla f(x^t - \alpha u^{t-1})$.The intuition behind NAG is looking ahead and anticipating, which leads to better solutions. Again, α is the momentum coefficient and γ is the learning rate. The update rule of NAS can be summarized as mentioned below.

1) Looking Ahead.

$$x^t_{lookahead} = x^t - \alpha v^{t-1} \tag{3}$$

2) Computing the gradient.

$$\nabla f(x^t - \alpha v^{t-1}) \tag{4}$$

3) Taking the gradient step.

$$\begin{aligned} x^{t+1} &= x^t_{lookahead} - \gamma \nabla f(x^t - \alpha v^{t-1}) \\ x^{t+1} &= x^t - \alpha v^{t-1} - \gamma \nabla f(x^t - \alpha v^{t-1}) \end{aligned} \tag{5}$$

This anticipatory step allows NAG to correct its course more responsively than standard Momentum, leading to potentially faster convergence and better handling of the curvature near optimal points. Essentially, NAG adds a level of foresight to updates, which can result in more efficient navigation of complex optimization landscapes.

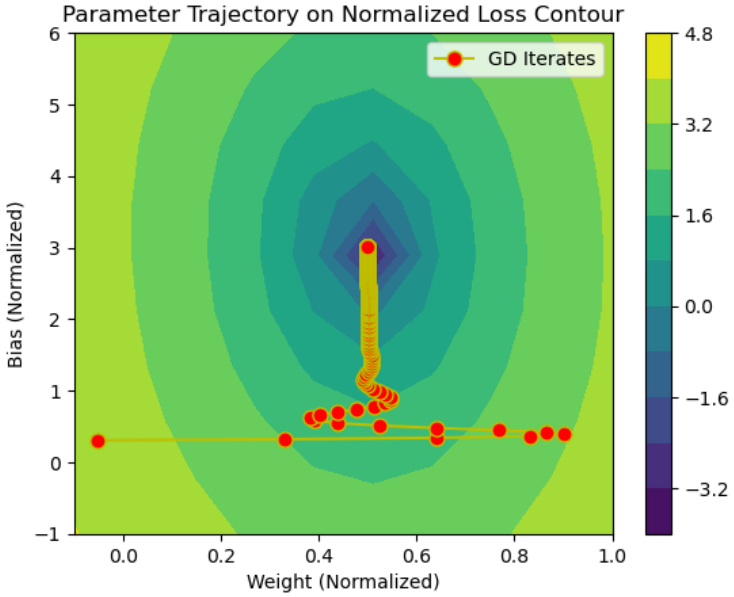


Fig. 2: NAG with $\gamma = 0.001$, $\alpha = 0$

C. Adaptive Gradient Method (AdaGrad)

AdaGrad is an adaptive learning rate method that modifies the general approach of gradient descent by allowing each parameter to have its own learning rate. This method addresses a common challenge in training machine learning models, where choosing an appropriate learning rate can be crucial for effective learning. Traditional gradient descent methods use a single learning rate for all parameters, which might not be optimal. Specifically, in a scenario where the users are presented with a dataset with sparse features the methods discussed above take longer to converge to the extremum. This is because the level sets of the problem are elongated balls.

AdaGrad adjusts the learning rate for each parameter based on the history of gradients that have been computed for that parameter. This means that parameters associated with frequently occurring features will have their learning rates decreased, while parameters associated with infrequent features will have their learning rates increased. Such adjustments are beneficial because they make the model less sensitive to the scale of features and more responsive to each feature's specific behavior and importance. This feature-dependent scaling of the learning rate helps in dealing with data sparsity and enhances the convergence properties of the gradient descent optimization, particularly in complex models dealing with high-dimensional data.

Let's define, $f(x)$ to be the stochastic objective function with parameter x, the function evaluation at step t as $f_t(x)$, the gradient of the function with respect to x at step t to be $g_t(x)$. Further take,

$$\mathbf{G}_t = \sum_{i=1}^{t-1} g_t g_t^T \tag{6}$$

Now the update rule for Adagrad can be written as follows, where γ is the learning rate.

$$x_{t+1} = x_t - \gamma \mathbf{G}_t^{-\frac{1}{2}} g_t \tag{7}$$

A simplified version of the update rule can be written by only considering the diagonal elements of $\mathbf{G}$.

$$x_{t+1} = x_t - \gamma diag(\mathbf{G}_t)^{-\frac{1}{2}} g_t \tag{8}$$

This simplified version of the update step is computationally efficient when we are dealing with high-dimensional data. Additionally, to avoid the problems arise due to the matrix being singular,in practice a small offset is added to the diagonal elements of the matrix $\mathbf{G}$.

$$x_{t+1} = x_t - \gamma diag(\epsilon \mathbf{I} + \mathbf{G}_t)^{-\frac{1}{2}} g_t \tag{9}$$

Finally, let's look at the expanded version of the update rule.

$$\begin{bmatrix} x_{t+1}^{(1)} \\ x_{t+1}^{(2)} \\ \vdots \\ x_{t+1}^{(m)} \end{bmatrix} = \begin{bmatrix} x_t^{(1)} \\ x_t^{(2)} \\ \vdots \\ x_t^{(m)} \end{bmatrix} - \begin{bmatrix} \frac{\gamma}{\sqrt{\varepsilon I + G_t^{(1,1)}}} \\ \frac{\gamma}{\sqrt{\varepsilon I + G_t^{(2,2)}}} \\ \vdots \\ \frac{\gamma}{\sqrt{\varepsilon I + G_t^{(m,m)}}} \end{bmatrix} \odot \begin{bmatrix} g_t^{(1)} \\ g_t^{(2)} \\ \vdots \\ g_t^{(m)} \end{bmatrix}$$

Where $\odot$ is the Hadamard product between two matrices having the same dimensions. This provides a clear idea of how the per-parameter learning rate works. Here, γ is the paramter which describes the global learning rate. It must also be noted that as G accumulates, the learning rate slows down for each parameter and eventually no progress can be made, causing the algorithm to never reach the exact minima. This is once of the major disadvantages of Adagrad.

*D. Root Mean Square Propagation (**RMS Prop**)*

It can be seen that the monotonically decreasing learning rate of Adagrad leads to a scenario where the learning rate becomes too small too quickly while the descent method can still achieve a reduction of the cost function. RMS prop addresses this issue by introducing a moving window of fixed length w over the gradients computed at each step rather than using the full set of gradients. Now the term $G^{(i)}$ for a the coordinate $x^{(i)}$ on the t^{th} iteration can be written as,

$$G_t^{(i)} = \frac{\left(g_{t-w}^{(i)}\right)^2 + \left(g_{t-w+1}^{(i)}\right)^2 + \ldots + \left(g_{t-1}^{(i)}\right)^2}{w} \tag{10}$$

A conceptually equivalent and computationally cheaper way of doing this is to treat 10 as an accumulation of exponentially decaying average of square of gradients. Let, ρ be the decaying factor, then we can write,

$$\mathbb{E}\left[\left(g_t^{(i)}\right)^2\right] = \rho \mathbb{E}\left[\left(g_{t-1}^{(i)}\right)^2\right] + (1-\rho)\left(g_t^{(i)}\right)^2 \tag{11}$$

From 11 we can see that the decay factor causes the older gradient to decay with iteration. This prevents the learning rate from becoming too small too quickly. Now $\left(G_t^i\right)^{-\frac{1}{2}}$ can be seen as,

$$RMS[g_t^{(i)}] = \left(G_{(t)}^i\right)^{-\frac{1}{2}} = \sqrt{\mathbb{E}\left[\left(g_t^{(i)}\right)^2\right]} \tag{12}$$

Finally the update step of the RMSprop algorithm can be written as.

$$\begin{bmatrix} x_{t+1}^{(1)} \\ x_{t+1}^{(2)} \\ \vdots \\ x_{t+1}^{(m)} \end{bmatrix} = \begin{bmatrix} x_{t}^{(1)} \\ x_{t}^{(2)} \\ \vdots \\ x_{t}^{(m)} \end{bmatrix} - \begin{bmatrix} \frac{\gamma}{RMS[g_t^{(1)}]} \\ \frac{\gamma}{RMS[g_t^{(2)}]} \\ \vdots \\ \frac{\gamma}{RMS[g_t^{(m)}]} \end{bmatrix} \odot \begin{bmatrix} g_t^{(1)} \\ g_t^{(2)} \\ \vdots \\ g_t^{(m)} \end{bmatrix}$$

IV. Convergence Behavior of SGD and Minibatch SGD

Citation Note: All theory for this section has been taken from [1]. The proofs in this book are rather succinct and have been further developed and analyzed in this section of the paper.

A. SGD Convergence for Convex and Smooth Functions

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions and Convex Functions. Let the sequence of iterates generated by the SGD Algorithm be $(x^{(t)})_{t\in\mathbb{N}}$ with a sequence of step sizes that satisfy $0 < \gamma_t < \frac{1}{4L_{max}}$. Denote

$$\bar{x}^T = \frac{1}{\sum_{t=0}^{T-1}\gamma_t}\sum_{t=0}^{T-1}\gamma_t x^{(t)}$$

Then for $T \geq 1$,

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=0}^{T-1}\gamma_t} + 2\sigma_f^* \frac{\sum_{t=0}^{T-1}\gamma_t^2}{\sum_{t=0}^{T-1}\gamma_t}$$

Proof:
Let $x^* \in \arg\min f$. According to Lemma 21, $\sigma_f^* = \mathbb{V}[\nabla f_i(x^*)]$.

In SGD, the iterates are as follows: $x^{(t)} = x^{(t-1)} - \gamma_{t-1}\nabla f_{i_{t-1}}(x^{(t-1)})$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - \gamma_{t-1}\nabla f_{i_{t-1}}(x^{(t-1)}) - x^*||^2$$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^* - \gamma_{t-1}\nabla f_{i_{t-1}}(x^{(t-1)})||^2$$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\langle x^{(t-1)} - x^*, \gamma_{t-1}\nabla f_{i_{t-1}}(x^{(t-1)})\rangle + ||\gamma_{t-1}\nabla f_{i_t}(x^{(t-1)})||^2$$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\langle x^{(t-1)} - x^*, \gamma_{t-1}\nabla f_{i_{t-1}}(x^{(t-1)})\rangle + \gamma_{t-1}^2||\nabla f_{i_t}(x^{(t-1)})||^2$$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\gamma_{t-1}\langle x^{(t-1)} - x^*, \nabla f_{i_{t-1}}(x^{(t-1)})\rangle + \gamma_{t-1}^2||\nabla f_{i_t}(x^{(t-1)})||^2$$

Taking the Expectation conditioned on $x^{(t-1)}$. For this proof, $\mathbb{E}[X]$ refers to the Expected value of X given the value of $x^{(t-1)}$:

$$\mathbb{E}[||x^{(t)} - x^*||^2] = ||x^{(t-1)} - x^*||^2 - 2\gamma_{t-1}\langle x^{(t-1)} - x^*, \nabla f(x^{(t-1)})\rangle + \gamma_{t-1}^2\mathbb{E}[||\nabla f_{i_t}(x^{(t-1)})||^2] \tag{13}$$

Due to the definition of convexity, $f(y) \geq f(x) + \nabla(f(x))^T(y - x)$

$$f(x^*) \geq f(x^{(t-1)}) + \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)})$$

$$f(x^*) - f(x^{(t-1)}) \geq \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)})$$

$$\nabla(f(x^{(t-1)}))^T(x^{(t-1)} - x^*) \geq f(x^{(t-1)}) - f(x^*) \tag{14}$$

Substituting 14 into 13

$$\mathbb{E}[||x^{(t)} - x^*||^2] = ||x^{(t-1)} - x^*||^2 - 2\gamma_{t-1}\langle x^{(t-1)} - x^*, \nabla f(x^{(t-1)})\rangle + \gamma_{t-1}^2\mathbb{E}[||\nabla f_{i_t}(x^{(t-1)})||^2]$$

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_{t-1}(f(x^{(t-1)}) - f(x^*)) + \gamma_{t-1}^2 \mathbb{E}[||\nabla f_{i_t}(x^{(t-1)})||^2] \tag{15}$$

Lemma 24 states that $\mathbb{E}[||\nabla f_i(x)||^2] \leq 4L_{max}(f(x) - \inf f) + 2\sigma_f^*$
Substituting this into 15

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_{t-1}(f(x^{(t-1)}) - f(x^*)) + \gamma_{t-1}^2(4L_{max}(f(x^{(t-1)}) - \inf f) + 2\sigma_f^*)$$

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_{t-1}(f(x^{(t-1)}) - f(x^*)) + \gamma_{t-1}^2 4L_{max}(f(x^{(t-1)}) - f(x^*)) + 2\gamma_{t-1}^2\sigma_f^*$$

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq ||x^{(t-1)} - x^*||^2 + (2\gamma_{t-1})(2\gamma_{t-1}L_{max} - 1)(f(x^{(t-1)}) - f(x^*)) + 2\gamma_{t-1}^2\sigma_f^*$$

Since $\gamma_{t-1} < \frac{1}{4L_{max}}$, $2\gamma_{t-1}L_{max} - 1 < \frac{-1}{2}$. Additionally, $(f(x^{(t-1)}) - f(x^*)) > 0$.
Hence,

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq ||x^{(t-1)} - x^*||^2 - \gamma_{t-1}(f(x^{(t-1)}) - f(x^*)) + 2\gamma_{t-1}^2\sigma_f^* \tag{16}$$

Taking Expectation, conditioned on $x^{(t-1)}$, over both sides of 16:

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq \mathbb{E}[||x^{(t-1)} - x^*||^2] - \gamma_{t-1}\mathbb{E}[(f(x^{(t-1)}) - f(x^*))] + 2\gamma_{t-1}^2\sigma_f^*$$

$$\gamma_{t-1}\mathbb{E}[(f(x^{(t-1)}) - f(x^*))] \leq \mathbb{E}[||x^{(t-1)} - x^*||^2] - \mathbb{E}[||x^{(t)} - x^*||^2] + 2\gamma_{t-1}^2\sigma_f^*$$

$$\gamma_{t-1}\mathbb{E}[(f(x^{(t-1)}) - \inf f)] \leq \mathbb{E}[||x^{(t-1)} - x^*||^2] - \mathbb{E}[||x^{(t)} - x^*||^2] + 2\gamma_{t-1}^2\sigma_f^*$$

Build this up recursively

$$\gamma_0\mathbb{E}[(f(x^{(0)}) - \inf f)] \leq \mathbb{E}[||x^{(0)} - x^*||^2] - \mathbb{E}[||x^{(1)} - x^*||^2] + 2\gamma_0^2\sigma_f^*$$

$$\gamma_1\mathbb{E}[(f(x^{(1)}) - \inf f)] \leq \mathbb{E}[||x^{(1)} - x^*||^2] - \mathbb{E}[||x^{(2)} - x^*||^2] + 2\gamma_1^2\sigma_f^*$$

$$\gamma_2\mathbb{E}[(f(x^{(2)}) - \inf f)] \leq \mathbb{E}[||x^{(2)} - x^*||^2] - \mathbb{E}[||x^{(3)} - x^*||^2] + 2\gamma_2^2\sigma_f^*$$

When adding together these inequalities, a general inequality begins to emerge:

$$\sum_{t=0}^{T-1} \gamma_t\mathbb{E}[(f(x^{(t)}) - f(x^*))] = \mathbb{E}[||x^{(0)} - x^*||^2] - \mathbb{E}[||x^{(T)} - x^*||^2] + \sum_{t=0}^{T-1} 2\gamma_t^2\sigma_f^*$$

Known Fact: $\mathbb{E}||x^{(T)} - x^*||^2 > 0$.

$$\sum_{t=0}^{T-1} \gamma_t\mathbb{E}[(f(x^{(t)}) - f(x^*))] = \mathbb{E}[||x^{(0)} - x^*||^2] - \mathbb{E}[||x^{(T)} - x^*||^2] + \sum_{t=1}^{T} 2\gamma_t^2\sigma_f^* \leq \mathbb{E}[||x^{(0)} - x^*||^2] + \sum_{t=1}^{T} 2\gamma_t^2\sigma_f^*$$

$$\sum_{t=0}^{T-1} \gamma_t\mathbb{E}[(f(x^{(t)}) - f(x^*))] \leq \mathbb{E}[||x^{(0)} - x^*||^2] + \sum_{t=0}^{T-1} 2\gamma_t^2\sigma_f^*$$

Note: $\mathbb{E}[||x^{(0)} - x^*||^2] = ||x^{(0)} - x^*||^2$

$$\sum_{t=0}^{T-1} \gamma_t \mathbb{E}[(f(x^{(t)}) - f(x^*))] \leq ||x^{(0)} - x^*||^2 + \sum_{t=0}^{T-1} 2\gamma_t^2 \sigma_f^*$$

Divide both sides of this inequality by $c = \sum_{t=0}^{T-1} \gamma_t$

$$\mathbb{E}[\sum_{t=0}^{T-1} \frac{\gamma_t}{c}(f(x^{(t)}) - f(x^*))] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=0}^{T-1} \gamma_t} + \frac{\sum_{t=0}^{T-1} 2\gamma_t^2 \sigma_f^*}{\sum_{t=0}^{T-1} \gamma_t}$$

$$\mathbb{E}[\sum_{t=0}^{T-1} \frac{1}{c}(\gamma_t f(x^{(t)}) - \gamma_t f(x^*))] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=0}^{T-1} \gamma_t} + \frac{\sum_{t=0}^{T-1} 2\gamma_t^2 \sigma_f^*}{\sum_{t=0}^{T-1} \gamma_t} \tag{17}$$

Since f is convex, the Generalized Jensen's Inequality means that:

$$f(\bar{x}^T) = f(\sum_{t=0}^{T-1} \frac{\gamma_t}{c} x^{(t)}) \leq \sum_{t=0}^{T-1} \frac{\gamma_t}{c} f(x^{(t)}) = \frac{1}{c} \sum_{t=0}^{T-1} \gamma_t f(x^{(t)}) \tag{18}$$

Known Equation:

$$f(x^*) = \frac{1}{c} \sum_{t=0}^{T-1} \gamma_t f(x^*) \tag{19}$$

Since $f(x^*) = \inf f$, 18, 19, and the Properties of Linearity Expectation

$$\mathbb{E}[f(\bar{x}^T) - \inf f] = \mathbb{E}[f(\bar{x}^T)] - \mathbb{E}[\inf f] = \mathbb{E}[f(\bar{x}^T)] - \mathbb{E}[f(x^*)] \leq \mathbb{E}[\frac{1}{c} \sum_{t=0}^{T-1} \gamma_t f(x^{(t)})] - \mathbb{E}[\frac{1}{c} \sum_{t=0}^{T-1} \gamma_t f(x^*)] \tag{20}$$

Using 20 and 17

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \mathbb{E}[\sum_{t=0}^{T-1} \frac{1}{c}(\gamma_t f(x^{(t)}) - \gamma_t f(x^*))] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=0}^{T-1} \gamma_t} + \frac{\sum_{t=0}^{T-1} 2\gamma_t^2 \sigma_f^*}{\sum_{t=0}^{T-1} \gamma_t}$$

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=0}^{T-1} \gamma_t} + \frac{\sum_{t=0}^{T-1} 2\gamma_t^2 \sigma_f^*}{\sum_{t=0}^{T-1} \gamma_t}$$

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions and Convex Functions. The sequence of iterates generated by the SGD Algorithm is $(x^{(t)})_{t \in \mathbb{N}}$ with a constant step size $\gamma_t = \gamma \leq \frac{1}{4L_{max}}$.
Denote

$$\bar{x}^T = \frac{1}{\sum_{t=0}^{T-1} \gamma_t} \sum_{t=0}^{T-1} \gamma_t x^t = \frac{1}{\gamma T} \gamma \sum_{t=0}^{T-1} x^t = \frac{1}{T} \sum_{t=0}^{T-1} x^t$$

Then for every $T \geq 1$,

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\gamma T} + 2\gamma \sigma_f^*$$

Proof. Known Facts: $\sum_{t=0}^{T-1} \gamma_t = \gamma T$ and $\sum_{t=0}^{T-1} \gamma_t^2 = T\gamma^2$.
Known Theorem:

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=0}^{T-1} \gamma_t} + \frac{\sum_{t=0}^{T-1} 2\gamma_t^2 \sigma_f^*}{\sum_{t=0}^{T-1} \gamma_t}$$

Substitute Known Facts

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\gamma T} + \frac{2\sigma_f^* T\gamma^2}{\gamma T}$$

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\gamma T} + 2\sigma_f^* \gamma$$

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions and Convex Functions. The sequence of iterates generated by the SGD Algorithm is $(x^{(t)})_{t\in\mathbb{N}}$ with a vanishing step size $\gamma_t = \frac{\gamma_0}{\sqrt{t+1}}$ where $\gamma_0 \leq \frac{1}{4L_{max}}$
Denote

$$\bar{x}^T = \frac{1}{\sum_{t=0}^{T-1} \gamma_t} \sum_{t=0}^{T-1} \gamma_t x^t$$

Then for every $T \geq 1$,

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{5||x^{(0)} - x^*||^2}{4\gamma_0\sqrt{T}} + \sigma_f^* \frac{5\gamma_0 \log(T+1)}{\sqrt{T}} = \mathbb{O}(\frac{\log(T+1)}{\sqrt{T}})$$

Proof. The stepsize is decreasing. Hence, $\gamma_t \leq \gamma_0 \leq \frac{1}{4L_{max}}$ for $t \geq 0$. Apply the earlier result that was derived:

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=0}^{T-1} \gamma_t} + 2\sigma_f^* \frac{\sum_{t=0}^{T-1} \gamma_t^2}{\sum_{t=0}^{T-1} \gamma_t}$$

Based on the Sum-Integral Bounds[Refer to Appendix], $\sum_{t=0}^{T-1} \gamma_t = \gamma_0 \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \geq \frac{4\gamma_0}{5}\sqrt{T}$ and $\sum_{t=0}^{T-1} \gamma_t^2 = \gamma_0^2 \sum_{t=1}^{T} \frac{1}{t} \leq 2\gamma_0^2 \log(T+1)$

Substituting it into the earlier inequality:

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{5||x^{(0)} - x^*||^2}{4\gamma_0\sqrt{T}} + \sigma_f^* \frac{5\gamma_0 \log(T+1)}{\sqrt{T}} = \mathbb{O}(\frac{\log(T+1)}{\sqrt{T}})$$

B. SGD Convergence for Strongly Convex and Smooth Functions

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions and Convex Functions. f is also p strongly convex. The sequence of iterates generated by the SGD Algorithm is $(x^{(t)})_{t\in\mathbb{N}}$ and there is a constant stepsize satisfying $0 < \gamma < \frac{1}{2L_{max}}$. For each iteration(i.e. $t \geq 0$)

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - \gamma p)^t ||x^{(0)} - x^*||^2 + \frac{2\gamma}{p}\sigma_f^*$$

Proof:
In SGD, the iterates are as follows: $x^{(t)} = x^{(t-1)} - \gamma \nabla f_{i_{t-1}}(x^{(t-1)})$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - \gamma \nabla f_{i_{t-1}}(x^{(t-1)}) - x^*||^2$$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^* - \gamma \nabla f_{i_{t-1}}(x^{(t-1)})||^2$$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\langle x^{(t-1)} - x^*, \gamma \nabla f_{i_{t-1}}(x^{(t-1)})\rangle + ||\gamma \nabla f_{i_{t-1}}(x^{(t-1)})||^2$$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\langle x^{(t-1)} - x^*, \gamma \nabla f_{i_{t-1}}(x^{(t-1)})\rangle + \gamma^2 ||\nabla f_{i_{t-1}}(x^{(t-1)})||^2$$

Take the Expectation conditioned on $x^{(t-1)}$. For this proof, $\mathbb{E}[X]$ refers to the Expected value of X given the value of $x^{(t-1)}$:

$$\mathbb{E}[||x^{(t)} - x^*||^2] = ||x^{(t-1)} - x^*||^2 - 2\gamma\langle x^{(t-1)} - x^*, \nabla f(x^{(t-1)})\rangle + \gamma^2 \mathbb{E}[||\nabla f_{i_{t-1}}(x^{(t-1)})||^2] \tag{21}$$

$f(y) \geq f(x) + \nabla(f(x))^T(y-x) + \frac{p}{2}||y-x||_2^2$ as per Strong Convexity

Hence, $f(x^*) \geq f(x^{(t-1)}) + \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)}) + \frac{p}{2}||x^* - x^{(t-1)}||_2^2$

$$f(x^*) - f(x^{(t-1)}) \geq \nabla f(x^{(t-1)})^T(x^* - x^{(t-1)}) + \frac{p}{2}||x^* - x^{(t-1)}||_2^2$$

$$\nabla f(x^{(t-1)})^T(x^{(t-1)} - x^*) \geq f(x^{(t-1)}) - f(x^*) + \frac{p}{2}||x^{(t-1)} - x^*||_2^2 \tag{22}$$

Substituting 22 into 21:

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq ||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*) + \frac{p}{2}||x^{(t-1)} - x^*||_2^2) + \gamma^2\mathbb{E}[||\nabla f_{i_{t-1}}(x^{(t-1)})||^2] \tag{23}$$

Simplifying RHS on 23:

$$||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*)) - p\gamma||x^{(t-1)} - x^*||_2^2 + \gamma^2\mathbb{E}[||\nabla f_{i_{t-1}}(x^{(t-1)})||^2]$$

$$(1-p\gamma)||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*)) + \gamma^2\mathbb{E}[||\nabla f_{i_{t-1}}(x^{(t-1)})||^2] \tag{24}$$

Putting 24 and 23

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq (1-p\gamma)||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*)) + \gamma^2\mathbb{E}[||\nabla f_{i_{t-1}}(x^{(t-1)})||^2] \tag{25}$$

As per Lemma 24, $\mathbb{E}[||\nabla f_i(x)||^2] \leq 4L_{max}(f(x) - \inf f) + 2\sigma_f^*$

Substituting Lemma 24 into 25

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq (1-p\gamma)||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*)) + \gamma^2(4L_{max}(f(x^{(t-1)}) - \inf f) + 2\sigma_f^*)$$

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq (1-p\gamma)||x^{(t-1)} - x^*||^2 + (2\gamma)(2\gamma L_{max} - 1)(f(x^{(t-1)}) - f(x^*)) + 2\gamma^2\sigma_f^*$$

Again, Taking Expectation Conditioned on $x^{(t-1)}$ on both sides of Inequality:

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq (1-p\gamma)\mathbb{E}[||x^{(t-1)} - x^*||^2] + (2\gamma)(2\gamma L_{max} - 1)\mathbb{E}[(f(x^{(t-1)}) - f(x^*))] + 2\gamma^2\sigma_f^*$$

Since $\gamma < \frac{1}{2L_{max}}$, $2\gamma L_{max} - 1 < 0$. Additionally, $\mathbb{E}[(f(x^{(t-1)}) - f(x^*))] > 0$.

$$\begin{aligned}\mathbb{E}[||x^{(t)} - x^*||^2] &\leq (1-p\gamma)\mathbb{E}[||x^{(t-1)} - x^*||^2] + (2\gamma)(2\gamma L_{max} - 1)\mathbb{E}[(f(x^{(t-1)}) - f(x^*))] + 2\gamma^2\sigma_f^* \\ &\leq (1-p\gamma)\mathbb{E}[||x^{(t-1)} - x^*||^2] + 2\gamma^2\sigma_f^*\end{aligned}$$

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq (1-p\gamma)\mathbb{E}[||x^{(t-1)} - x^*||^2] + 2\gamma^2\sigma_f^*$$

Build this inequality recursively

$$\mathbb{E}[||x^{(1)} - x^*||^2] \leq (1-p\gamma)\mathbb{E}[||x^{(0)} - x^*||^2] + 2\gamma^2\sigma_f^*$$

$$\mathbb{E}[||x^{(2)} - x^*||^2] \leq (1-p\gamma)\mathbb{E}[||x^{(1)} - x^*||^2] + 2\gamma^2\sigma_f^*$$

Note: $\mathbb{E}[||x^{(0)} - x^*||^2] = ||x^{(0)} - x^*||^2$

$$\mathbb{E}[||x^{(2)} - x^*||^2] \leq (1-p\gamma)((1-p\gamma)||x^{(0)} - x^*||^2 + 2\gamma^2\sigma_f^*) + 2\gamma^2\sigma_f^*$$

The following inequality emerges: $\mathbb{E}[||x^{(t)} - x^*||^2] \leq (1-\gamma p)^t||x^{(0)} - x^*||^2 + \sum_{n=0}^{t-1}(1-p\gamma)^n 2\gamma^2\sigma_f^*$
Let's look at the term $\sum_{n=0}^{t-1}(1-p\gamma)^n 2\gamma^2\sigma_f^*$.

$$\sum_{n=0}^{t-1}(1-p\gamma)^n 2\gamma^2\sigma_f^* < \sum_{n=0}^{\infty}(1-p\gamma)^n 2\gamma^2\sigma_f^* = \frac{1}{p\gamma}2\gamma^2\sigma_f^* = \frac{2\gamma\sigma_f^*}{p}$$

Hence,

$$\mathbb{E}[||x^{(t)}-x^*||^2] \leq (1-\gamma p)^t||x^{(0)}-x^*||^2 + \sum_{n=0}^{t-1}(1-p\gamma)^n 2\gamma^2\sigma_f^* \leq (1-\gamma p)^t||x^{(0)}-x^*||^2 + \frac{2\gamma}{p}\sigma_f^*$$

$$\mathbb{E}[||x^{(t)}-x^*||^2] \leq (1-\gamma p)^t||x^{(0)}-x^*||^2 + \frac{2\gamma}{p}\sigma_f^*$$

C. Minibatch SGD Convergence for Convex and Smooth Functions

Theorem. Let $f:\mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions and Convex Functions. The sequence of iterates generated by the Minibatch SGD Algorithm is $(x^{(t)})_{t\in\mathbb{N}}$ with a sequence of step sizes that satisfy $0 < \gamma_t < \frac{1}{4L_b}$. Denote

$$\bar{x}^T = \frac{1}{\sum_{t=0}^{T-1}\gamma_t}\sum_{t=0}^{T-1}\gamma_t x^t$$

Then for $T \geq 1$,

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)}-x^*||^2}{\sum_{t=0}^{T-1}\gamma_t} + 2\sigma_b^*\frac{\sum_{t=0}^{T-1}\gamma_t^2}{\sum_{t=0}^{T-1}\gamma_t}$$

Proof:
Let $x^* \in \arg\min f$.
According to Lemma 21, $\sigma_b^* = \mathbb{V}[\nabla f_B(x^*)]$.
In Minibatch Stochastic Gradient Descent, our iterates are as follows: $x^{(t)} = x^{(t-1)} - \gamma_{t-1}\nabla f_{B_{t-1}}(x^{(t-1)})$

$$||x^{(t)}-x^*||^2 = ||x^{(t-1)} - \gamma_{t-1}\nabla f_{B_{t-1}}(x^{(t-1)}) - x^*||^2$$

$$||x^{(t)}-x^*||^2 = ||x^{(t-1)} - x^* - \gamma_{t-1}\nabla f_{B_{t-1}}(x^{(t-1)})||^2$$

$$||x^{(t)}-x^*||^2 = ||x^{(t-1)}-x^*||^2 - 2\langle x^{(t-1)}-x^*, \gamma_{t-1}\nabla f_{B_{t-1}}(x^{(t-1)})\rangle + ||\gamma_{t-1}\nabla f_{B_{t-1}}(x^{(t-1)})||^2$$

$$||x^{(t)}-x^*||^2 = ||x^{(t-1)}-x^*||^2 - 2\langle x^{(t-1)}-x^*, \gamma_{t-1}\nabla f_{B_{t-1}}(x^{(t-1)})\rangle + \gamma_{t-1}^2||\nabla f_{B_{t-1}}(x^{(t-1)})||^2$$

$$||x^{(t)}-x^*||^2 = ||x^{(t-1)}-x^*||^2 - 2\gamma_{t-1}\langle x^{(t-1)}-x^*, \nabla f_{B_{t-1}}(x^{(t-1)})\rangle + \gamma_{t-1}^2||\nabla f_{B_{t-1}}(x^{(t-1)})||^2$$

Take Expectation conditioned on $x^{(t-1)}$ on both sides of this equation. For this proof, $E[X]$ refers to the Expected value of X given the value of $x^{(t-1)}$:

$$\mathbb{E}[||x^{(t)}-x^*||^2] = ||x^{(t-1)}-x^*||^2 - 2\gamma_{t-1}\langle x^{(t-1)}-x^*, \nabla f(x^{(t-1)})\rangle + \gamma_{t-1}^2\mathbb{E}[||\nabla f_{B_{t-1}}(x^{(t-1)})||^2] \tag{26}$$

$f(y) \geq f(x) + \nabla(f(x))^T(y-x)$ as per Convexity

$$f(x^*) \geq f(x^{(t-1)}) + \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)})$$

$$f(x^*) \geq f(x^{(t-1)}) + \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)})$$

$$\nabla(f(x^{(t-1)}))^T(x^{(t-1)} - x^*) \geq f(x^{(t-1)}) - f(x^*) \tag{27}$$

Substituting 27 into 26:

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_{t-1}(f(x^{(t-1)}) - f(x^*)) + \gamma_{t-1}^2\mathbb{E}[||\nabla f_{B_{t-1}}(x^{(t-1)})||^2] \tag{28}$$

According to Lemma 28, $\mathbb{E}[||\nabla f_B(x)||^2] \leq 4L_b(f(x) - \inf f) + 2\sigma_b^*$

Substitute this into 28:

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_{t-1}(f(x^{(t-1)}) - f(x^*)) + \gamma_{t-1}^2(4L_b(f(x^{(t-1)}) - \inf f) + 2\sigma_b^*)$$

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_{t-1}(f(x^{(t-1)}) - f(x^*)) + \gamma_{t-1}^2 4L_b(f(x^{(t-1)}) - \inf f) + 2\gamma_{t-1}^2\sigma_b^*$$

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq ||x^{(t-1)} - x^*||^2 + (2\gamma_{t-1})(2\gamma_{t-1}L_b - 1)(f(x^{(t-1)}) - f(x^*)) + 2\gamma_{t-1}^2\sigma_b^*$$

Since $\gamma_{t-1} < \frac{1}{4L_b}$, $2\gamma_{t-1}L_b - 1 < \frac{-1}{2}$. Additionally, $(f(x^{(t-1)}) - f(x^*)) > 0$. Hence,

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq ||x^{(t-1)} - x^*||^2 - \gamma_{t-1}(f(x^{(t-1)}) - f(x^*)) + 2\gamma_{t-1}^2\sigma_b^*$$

Take Expectation, conditioned on $x^{(t-1)}$, over both sides of the inequality

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq \mathbb{E}[||x^{(t-1)} - x^*||^2] - \gamma_{t-1}\mathbb{E}[(f(x^{(t-1)}) - f(x^*))] + 2\gamma_t^2\sigma_b^*$$

$$\gamma_{t-1}\mathbb{E}[(f(x^{(t-1)}) - f(x^*))] \leq \mathbb{E}[||x^{(t-1)} - x^*||^2] - \mathbb{E}[||x^{(t)} - x^*||^2] + 2\gamma_{t-1}^2\sigma_b^*$$

$$\gamma_{t-1}\mathbb{E}[(f(x^{(t-1)}) - \inf f)] \leq \mathbb{E}[||x^{(t-1)} - x^*||^2] - \mathbb{E}[||x^{(t)} - x^*||^2] + 2\gamma_{t-1}^2\sigma_b^*$$

Build this up recursively

$$\gamma_0\mathbb{E}[(f(x^{(0)}) - \inf f)] \leq \mathbb{E}[||x^{(0)} - x^*||^2] - \mathbb{E}[||x^{(1)} - x^*||^2] + 2\gamma_0^2\sigma_b^*$$

$$\gamma_1\mathbb{E}[(f(x^{(1)}) - \inf f)] \leq \mathbb{E}[||x^{(1)} - x^*||^2] - \mathbb{E}[||x^{(2)} - x^*||^2] + 2\gamma_1^2\sigma_b^*$$

$$\gamma_2\mathbb{E}[(f(x^{(2)}) - \inf f)] \leq \mathbb{E}[||x^{(2)} - x^*||^2] - \mathbb{E}[||x^{(3)} - x^*||^2] + 2\gamma_2^2\sigma_b^*$$

$$\sum_{t=0}^{T-1}\gamma_t\mathbb{E}[(f(x^{(t)}) - f(x^*))] = \mathbb{E}[||x^{(0)} - x^*||^2] - \mathbb{E}[||x^{(T)} - x^*||^2] + \sum_{t=0}^{T-1}2\gamma_t^2\sigma_b^*$$

Known Fact: $\mathbb{E}[||x^{(T)} - x^*||^2] > 0$.

$$\sum_{t=0}^{T-1}\gamma_t\mathbb{E}[(f(x^{(t)}) - f(x^*))] = \mathbb{E}[||x^{(0)} - x^*||^2] - \mathbb{E}[||x^{(T)} - x^*||^2] + \sum_{t=0}^{T-1}2\gamma_t^2\sigma_b^* \leq \mathbb{E}[||x^{(0)} - x^*||^2] + \sum_{t=0}^{T-1}2\gamma_t^2\sigma_b^*$$

$$\sum_{t=0}^{T-1} \gamma_t \mathbb{E}[(f(x^{(t)}) - f(x^*))] \leq \mathbb{E}[||x^{(0)} - x^*||^2] + \sum_{t=0}^{T-1} 2\gamma_t^2 \sigma_b^*$$

Note: $\mathbb{E}[||x^{(0)} - x^*||^2] = ||x^{(0)} - x^*||^2$

$$\sum_{t=0}^{T-1} \gamma_t \mathbb{E}[(f(x^{(t)}) - f(x^*))] \leq ||x^{(0)} - x^*||^2 + \sum_{t=0}^{T-1} 2\gamma_t^2 \sigma_b^*$$

Divide both sides of this inequality by $c = \sum_{t=0}^{T-1} \gamma_t$

$$\mathbb{E}[\sum_{t=0}^{T-1} \frac{\gamma_t}{c}(f(x^{(t)}) - f(x^*))] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=0}^{T-1} \gamma_t} + \frac{\sum_{t=0}^{T-1} 2\gamma_t^2 \sigma_b^*}{\sum_{t=0}^{T-1} \gamma_t}$$

$$\mathbb{E}[\sum_{t=0}^{T-1} \frac{1}{c}(\gamma_t f(x^{(t)}) - \gamma_t f(x^*))] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=0}^{T-1} \gamma_t} + \frac{\sum_{t=0}^{T-1} 2\gamma_t^2 \sigma_b^*}{\sum_{t=0}^{T-1} \gamma_t} \tag{29}$$

Since f is convex, the Generalized Jensen's Inequality means that:

$$f(\bar{x}^T) = f(\sum_{t=0}^{T-1} \frac{\gamma_t}{c} x^{(t)}) \leq \sum_{t=0}^{T-1} \frac{\gamma_t}{c} f(x^{(t)}) = \frac{1}{c} \sum_{t=0}^{T-1} \gamma_t f(x^{(t)})$$

Known Equation: $f(x^*) = \frac{1}{c} \sum_{t=0}^{T-1} \gamma_t f(x^*)$
Using the above Known Equation and the Inequality that was derived from the Generalized Jensen's Inequality, along with 29

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \mathbb{E}[\sum_{t=0}^{T-1} \frac{1}{c}(\gamma_t f(x^{(t)}) - \gamma_t f(x^*))] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=0}^{T-1} \gamma_t} + \frac{\sum_{t=0}^{T-1} 2\gamma_t^2 \sigma_b^*}{\sum_{t=0}^{T-1} \gamma_t}$$

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=0}^{T-1} \gamma_t} + \frac{\sum_{t=0}^{T-1} 2\gamma_t^2 \sigma_b^*}{\sum_{t=0}^{T-1} \gamma_t}$$

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions and Convex Functions. The sequence of iterates generated by the SGD Algorithm is $(x^{(t)})_{t \in \mathbb{N}}$ with a constant step size $\gamma_t = \gamma \leq \frac{1}{4L_b}$.
Denote

$$\bar{x}^T = \frac{1}{\sum_{t=0}^{T-1} \gamma_t} \sum_{t=0}^{T-1} \gamma_t x^t = \frac{1}{\gamma T} \gamma \sum_{t=0}^{T-1} x^t = \frac{1}{T} \sum_{t=0}^{T-1} x^t$$

Then for every $T \geq 1$,

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\gamma T} + 2\gamma \sigma_b^*$$

Proof.
Known Fact: $\sum_{t=0}^{T-1} \gamma_t = \gamma T$ and $\sum_{t=0}^{T-1} \gamma_t^2 = T\gamma^2$.
Substitute this known fact into the Last Theorem:

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=0}^{T-1} \gamma_t} + \frac{\sum_{t=0}^{T-1} 2\gamma_t^2 \sigma_b^*}{\sum_{t=0}^{T-1} \gamma_t}$$

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\gamma T} + \frac{2\sigma_b^* T\gamma^2}{\gamma T}$$

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\gamma T} + 2\sigma_b^* \gamma$$

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions and Convex Functions. The sequence of iterates generated by the SGD Algorithm is $(x^{(t)})_{t\in\mathbb{N}}$ with a vanishing step size $\gamma_t = \frac{\gamma_0}{\sqrt{t+1}}$ where $\gamma_0 \leq \frac{1}{4L_b}$
Denote

$$\bar{x}^T = \frac{1}{\sum_{t=0}^{T-1} \gamma_t} \sum_{t=0}^{T-1} \gamma_t x^t$$

Then for every $T \geq 1$,

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{5||x^{(0)} - x^*||^2}{4\gamma_0\sqrt{T}} + \sigma_b^* \frac{5\gamma_0 \log(T+1)}{\sqrt{T}} = \mathbb{O}(\frac{\log(T+1)}{\sqrt{T}})$$

Proof. Since the stepsize decreases, $\gamma_t \leq \gamma_0 \leq \frac{1}{4L_b}$ for $t \geq 0$. Applying the earlier result that was derived:

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=0}^{T-1} \gamma_t} + \frac{\sum_{t=0}^{T-1} 2\gamma_t^2 \sigma_b^*}{\sum_{t=0}^{T-1} \gamma_t} = \frac{||x^{(0)} - x^*||^2}{\sum_{t=0}^{T-1} \gamma_t} + 2\sigma_b^* \frac{\sum_{t=0}^{T-1} \gamma_t^2}{\sum_{t=0}^{T-1} \gamma_t}$$

Based on the Sum-Integral Bounds[Refer to Appendix], $\sum_{t=0}^{T-1} \gamma_t = \gamma_0 \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \geq \frac{4\gamma_0}{5}\sqrt{T}$ and $\sum_{t=0}^{T-1} \gamma_t^2 = \gamma_0^2 \sum_{t=1}^{T} \frac{1}{t} \leq 2\gamma_0^2 \log(T+1)$

Substituting it into the expected value

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{5||x^{(0)} - x^*||^2}{4\gamma_0\sqrt{T}} + \sigma_b^* \frac{5\gamma_0 \log(T+1)}{\sqrt{T}} = \mathbb{O}(\frac{\log(T+1)}{\sqrt{T}})$$

D. Minibatch SGD Convergence for Strongly Convex and Smooth Functions

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions and a Sum of Convex Functions. f is also p strongly convex. The sequence of iterates generated by the Minibatch SGD Algorithm is $(x^{(t)})_{t\in\mathbb{N}}$ and there is a constant stepsize satisfying $0 < \gamma < \frac{1}{2L_b}$. Then, for each iteration(i.e. $t \geq 0$)

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - \gamma p)^t ||x^{(0)} - x^*||^2 + \frac{2\gamma}{p}\sigma_b^*$$

Proof:
In Minibatch Stochastic Gradient Descent, the iterates progress as such: $x^{(t)} = x^{(t-1)} - \gamma \nabla f_{B_{t-1}}(x^{(t-1)})$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - \gamma \nabla f_{B_{t-1}}(x^{(t-1)}) - x^*||^2$$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^* - \gamma \nabla f_{B_{t-1}}(x^{(t-1)})||^2$$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\langle x^{(t-1)} - x^*, \gamma \nabla f_{B_{t-1}}(x^{(t-1)})\rangle + ||\gamma \nabla f_{B_{t-1}}(x^{(t-1)})||^2$$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\langle x^{(t-1)} - x^*, \gamma \nabla f_{B_{t-1}}(x^{(t-1)})\rangle + \gamma^2 ||\nabla f_{B_{t-1}}(x^{(t-1)})||^2$$

Take Expectation conditioned on $x^{(t-1)}$ on both sides of this equation. For this proof, $E[X]$ refers to the Expected value of X given the value of $x^{(t-1)}$:

$$\mathbb{E}[||x^{(t)} - x^*||^2] = ||x^{(t-1)} - x^*||^2 - 2\gamma\langle x^{(t-1)} - x^*, \nabla f(x^{(t-1)})\rangle + \gamma^2\mathbb{E}[||\nabla f_{B_{t-1}}(x^{(t-1)})||^2] \tag{30}$$

Due to Strong Convexity, $f(y) \geq f(x) + \nabla(f(x))^T(y-x) + \frac{p}{2}||y-x||_2^2$

$$f(x^*) \geq f(x^{(t-1)}) + \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)}) + \frac{p}{2}||x^* - x^{(t-1)}||_2^2$$

$$f(x^*) - f(x^{(t-1)}) \geq \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)}) + \frac{p}{2}||x^* - x^{(t-1)}||_2^2$$

$$\langle x^{(t-1)} - x^*, \nabla f(x^{(t-1)})\rangle = \nabla(f(x^{(t-1)}))^T(x^{(t-1)} - x^*) \geq f(x^{(t-1)}) - f(x^*) + \frac{p}{2}||x^{(t-1)} - x^*||_2^2 \tag{31}$$

Substitute 31 into 30:

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq ||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*) + \frac{p}{2}||x^{(t-1)} - x^*||_2^2) + \gamma^2\mathbb{E}[||\nabla f_{B_{t-1}}(x^{(t-1)})||^2] \tag{32}$$

Simplify Right Hand Side of 32

$$||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*)) - p\gamma||x^{(t-1)} - x^*||_2^2 + \gamma^2\mathbb{E}[||\nabla f_{B_{t-1}}(x^{(t-1)})||^2]$$

$$(1 - p\gamma)||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*)) + \gamma^2\mathbb{E}[||\nabla f_{B_{t-1}}(x^{(t-1)})||^2]$$

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq (1 - p\gamma)||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*)) + \gamma^2\mathbb{E}[||\nabla f_{B_{t-1}}(x^{(t-1)})||^2] \tag{33}$$

According to Lemma 28, $\mathbb{E}[||\nabla f_B(x)||^2] \leq 4L_b(f(x) - \inf f) + 2\sigma_b^*$
Substitute this into 33

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq (1 - p\gamma)||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*)) + \gamma^2(4L_b(f(x^{(t-1)}) - \inf f) + 2\sigma_b^*)$$

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq (1 - p\gamma)||x^{(t-1)} - x^*||^2 + (2\gamma)(2\gamma L_b - 1)(f(x^{(t-1)}) - f(x^*)) + 2\gamma^2\sigma_b^*$$

Taking Expectation Conditioned on $x^{(t-1)}$ on both sides of Inequality:

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq (1 - p\gamma)\mathbb{E}[||x^{(t-1)} - x^*||^2] + (2\gamma)(2\gamma L_b - 1)\mathbb{E}[(f(x^{(t-1)}) - f(x^*))] + 2\gamma^2\sigma_b^*$$

Since $\gamma < \frac{1}{2L_b}$, $2\gamma L_b - 1 < 0$. Additionally, it is known that $\mathbb{E}(f(x^{(t-1)}) - f(x^*)) > 0$ Hence

$$\begin{aligned}\mathbb{E}[||x^{(t)} - x^*||^2] &\leq (1 - p\gamma)\mathbb{E}[||x^{(t-1)} - x^*||^2] + (2\gamma)(2\gamma L_b - 1)\mathbb{E}[(f(x^{(t-1)}) - f(x^*))] + 2\gamma^2\sigma_b^* \\ &\leq (1 - p\gamma)||x^{(t-1)} - x^*||^2 + 2\gamma^2\sigma_b^*\end{aligned} \tag{34}$$

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq (1 - p\gamma)\mathbb{E}[||x^{(t-1)} - x^*||^2] + 2\gamma^2\sigma_b^*$$

Build this inequality recursively and see how it unfolds:

$$\mathbb{E}[||x^{(1)} - x^*||^2] \leq (1 - p\gamma)\mathbb{E}[||x^{(0)} - x^*||^2] + 2\gamma^2\sigma_b^*$$

$$\mathbb{E}[||x^{(2)} - x^*||^2] \leq (1 - p\gamma)\mathbb{E}[||x^{(1)} - x^*||^2] + 2\gamma^2\sigma_b^*$$

$$\mathbb{E}[||x^{(2)} - x^*||^2] \leq (1 - p\gamma)((1 - p\gamma)\mathbb{E}[||x^{(0)} - x^*||^2] + 2\gamma^2\sigma_b^*) + 2\gamma^2\sigma_b^*$$

Note: $\mathbb{E}[||x^{(0)} - x^*||^2] = ||x^{(0)} - x^*||^2$

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq (1 - \gamma p)^t||x^{(0)} - x^*||^2 + \sum_{n=0}^{t-1}(1 - p\gamma)^n 2\gamma^2\sigma_b^*$$

Analyze the term $\sum_{n=0}^{t-1}(1 - p\gamma)^n 2\gamma^2\sigma_b^*$.

$$\sum_{n=0}^{t-1}(1 - p\gamma)^n 2\gamma^2\sigma_b^* < \sum_{n=0}^{\infty}(1 - p\gamma)^n 2\gamma^2\sigma_b^* = \frac{1}{p\gamma}2\gamma^2\sigma_b^* = \frac{2\gamma\sigma_b^*}{p}$$

Hence, it is evident that

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq (1 - \gamma p)^t||x^{(0)} - x^*||^2 + \sum_{n=0}^{t-1}(1 - p\gamma)^n 2\gamma^2\sigma_b^* \leq (1 - \gamma p)^t||x^{(0)} - x^*||^2 + \frac{2\gamma}{p}\sigma_b^*$$

$$\mathbb{E}[||x^{(t)} - x^*||^2] \leq (1 - \gamma p)^t||x^{(0)} - x^*||^2 + \frac{2\gamma}{p}\sigma_b^*$$

V. Algorithmic Stability of SGD

Cite Note: The insights discussed in this section are derived from the comprehensive analysis presented in [2], which focuses on key stability parameters and their implications.

A. Generalization Error

Generalization error is a measure of how accurately a machine learning model can predict outcome values for previously unseen data. Specifically, it quantifies the difference in performance between training data and new, unseen data. A key goal in areas such as machine learning is to minimize this error, which indicates better model performance on new, unseen data. The generalization error of a learning algorithm can be formally defined as the difference between the expected loss over the distribution of all possible data and the empirical loss calculated on the training set. Following is a build-up for a formal definition of the generalization error adopted from [2], Consider a supervised learning setting where:

- The samples are drawn from a $\mathcal{D}$ unknown distribution.
- The samples $S = (x_1, x_2, \ldots, x_n)$ is samples i.i.d from the aforementioned distribution.
- The main objective is to find the model w with an associated loss function $f(w; x)$
- n denotes the number of samples.

The algorithm tries to find the model w with a minimum population risk defined as the expected loss over $\mathcal{D}$,

$$R[w] \triangleq \mathbb{E}_{x\sim\mathcal{D}}[f(w; x)] \tag{35}$$

However, it is important to note that, not knowing $\mathcal{D}$ makes it impossible to measure $R[w]$ directly. However, having a sufficient number of samples allow us to estimate population risk through the empirical average of the loss function over the samples (empirical risk),

$$R_S[w] \triangleq \frac{1}{n}\sum_{i=1}^{n} f(w; x_i) \tag{36}$$

Now the generalization error of the model w is the difference between the population risk and the empirical risk,

$$\text{Generalization Error} = R_S[w] - R[w] \tag{37}$$

In many applications, the model parameters w are learned using the sample data and response variables. For example, in linear regression, the goal is to establish a linear relationship between the features (or predictors) and the response variable. This relationship is expressed in the form $y = w^T x + b$, where y is the response, x is the feature vector, w is the vector of weights, and b is a bias term. By minimizing the empirical risk, typically represented by the mean squared error between the predicted values and the actual values in the training data, the model learns the parameters that best fit the data.

To learn these parameters effectively, especially in complex or large-scale settings, randomized algorithms such as Stochastic Gradient Descent (SGD) are frequently employed. This makes it computationally more efficient than batch gradient descent, particularly with large datasets. Here the random nature of the algorithm develops through the random selection of data subsets. With this insight it can be seen that the model w can be taken as a function of data S through a randomized algorithm A, where, $w = A(S)$. This allows one to define an expected generalization error over the randomness of the samples and the algorithm.

$$\epsilon_{gen} \triangleq \mathbb{E}_{S,A}[R_S[A(S)] - R[A(S)]] \tag{38}$$

B. Definition of Stability and Preliminaries

Definition 27: A randomized algorithm A is said to be ϵ - uniformly stable if, for a sample S, let,

$$S^{'} = z_1, \ldots, z_{i-1}, z_i^{'}, z_{i+1}, \ldots, z_m$$

be a sample obtained by replacing only the ith observation of S with $z_i^{'}$, we have

$$\sup_z \left(\mathbf{E}_A \left[f\left(A(S); z\right) - f\left(A(S'); z\right)\right]\right) \leq \epsilon. \tag{39}$$

For the rest of this discussion, the following notation on the general update rules is used. Let, $G : \Omega \to \Omega$ be a general update rule which map a point $\omega \in \Omega$ in the parameter space through,

$$G(\omega) = \omega - \alpha \nabla f(\omega)$$

where $\alpha \geq 0$ is the step-size and $f : \Omega \to \mathbb{R}$ is the objective function.
With this general notation, the following definitions provide the foundation for the analysis of how different sequences of an update rule diverge given they were initialized from the same point. These definitions are then used in the analysis of the stability of stochastic gradient descent.

Definition 28 An update rule is η-**expansive** if for all $v, w \in \Omega$,

$$\|G(v) - G(w)\| \leq \eta \|v - w\|.$$

It is σ-**bounded** if

$$\|w - G(w)\| \leq \sigma.$$

These two definitions can be used to develop a lemma on how two sequences of the update rule change when the training set is perturbed.

Lemma 29.[Growth recursion]: Let's consider an arbitrary sequence of updates $G_1, \ldots, G_T$ and another sequence $G'_1, \ldots, G'_T$. Let $w_0 = w'_0$ be a starting point in Ω and define $\delta_t = \|w'_t - w_t\|$ where w_t, w'_t are defined recursively through,

$$w_{t+1} = G_t(w_t) \qquad w'_{t+1} = G'_t(w'_t)\,. \qquad (t > 0) \tag{}$$

Then, the recurrence relation can be derived to be,

$$\delta_{t+1} \leq \begin{cases} \eta\delta_t & G_t = G'_t \text{ is } \eta\text{-expansive} \\ \min(\eta, 1)\delta_t + 2\sigma_t & G_t \text{ and } G'_t \text{ are } \sigma\text{-bounded,} \\ & G_t \text{ is } \eta \text{ expansive} \end{cases}$$

Proof. The first bound on δ_t,

$$\begin{aligned} \delta_{t+1} &= \|\omega_{t+1} - \omega'_{t+1}\| \\ &= \|G(\omega_t) - G(\omega'_t)\| \\ &= \eta\|\omega_t) - \omega'_t\| \\ &= \eta\delta_t \end{aligned} \tag{40}$$

For the second bound, from **Definition 28** given that G_t and G'_t are σ-bounded, then by using the triangle inequality theorem,

$$\begin{aligned} \delta_{t+1} &= \|G(w_t) - G'(w'_t)\| \\ &\leq \|G(w_t) - w_t + w'_t - G'(w'_t)\| + \|w_t - w'_t\| \\ &\leq \delta_t + \|G(w_t) - w_t\| + \|G(w'_t) - w'_t\| \\ &\leq \delta_t + 2\sigma\,, \end{aligned}$$

provides a part of the bound. Furthermore δ_{t+1} can be bounded using the η - expansiveness of G_t,

$$\begin{aligned}\delta_{t+1} &= \|G_t(w_t) - G'_t(w'_t)\| \\ &= \|G_t(w_t) - G_t(w'_t) + G_t(w'_t) - G'_t(w'_t)\| \\ &\leq \|G_t(w_t) - G_t(w'_t)\| + \|G_t(w'_t) - G'_t(w'_t)\| \\ &\leq \|G_t(w_t) - G_t(w'_t)\| + \|w'_t - G_t(w'_t)\| + \|w'_t - G'_t(w'_t)\| \\ &\leq \eta\delta_t + 2\sigma\,.\end{aligned}$$

Considering these two parts, the minimum between $(1, \eta)$ is taken as the bound.

C. Stability of Stochastic Gradient Descent

For the following discussion the the gradient update rule for the stochastic gradient method is defined as follows.

Definition 30 Let, $S = (z_1, \ldots, z_n)$ be a set of labeled examples. Consider a decomposible objective function, $f(\omega) = \frac{1}{n}\sum_{i=1}^{n} f(\omega; z_i)$ denoting loss of ω at example z_i. The stochastic gradient update rule for this kind of problem with a learning rate $\alpha_t \geq 0$ is given by,

$$\omega_{t+1} = \omega_t - \alpha_t \nabla_\omega f(\omega_t; z_{i_t}) \tag{41}$$

where, the sample z_{i_t} is choosen at random as discussed in the the **Introduction**. In parallel with the previous section, the general gradient update rule for a non-negative step size α and a function $f : \omega \to \mathbb{R}$ is stated as,

$$G_{f,\alpha}(\omega) = \omega - \alpha\nabla f(\omega) \tag{42}$$

1) Proof Sketch

This work demonstrates the stability of the stochastic gradient method by examining its outputs on two data sets that only differ at a single point. When the loss function adheres to the L-Lipschitz condition for any given example z, it is expected that the expected difference in the loss between any two weight vectors ω and $omega'$ is bounded by L times the norm of their difference.

$$\mathbb{E}[f(\omega; z) - f(\omega'; z)] \leq L\mathbb{E}[\omega - \omega']$$

The divergence between ω_t and w'_t, the parameters at time t, is of interest here. The idea is to bound the $\delta_t = \|\omega_t - \omega'_t\|$, recursively and in expectation as a function of δ_{t-1}.

In the analysis, two scenarios are considered.

1) Stochastic Gradient Method (SGM) selects an index which is identical in datasets S and S'. Contrary to expectations, δ_t may increase due to the possibility of different gradients at w_t and w'_t. The divergence is managed by leveraging the convexity and smoothness of the gradients.
2) SGM updating based on an example that differs between datasets S and S', which happens with a probability of $\frac{1}{n}$ under random selection. Should divergence be observed, the increase in δ_t is limited by the gradient norms $\nabla f(\omega_{t-1}; z)$ and $\nabla f(\omega'_{t-1}; z')$. The application of a bound of $2\alpha_{t-1}L$ to these norms allows for the conclusion that δ_t is restricted by $\delta_{t-1} + 2\alpha_{t-1}L$.

$$\begin{aligned}\delta_t &= \|\omega_t = \omega'_t\| \\ &= \|\omega_{t-1} - \alpha_t\nabla f(\omega_{t-1}; z) - \omega'_{t-1} + \alpha_{t-1}\nabla f(\omega'_{t-1}; z')\| \\ &= \|\omega_{t-1} - \omega'_{t-1} + \alpha_{t-1}\nabla f(\omega'_{t-1}; z') - \alpha_t\nabla f(\omega_{t-1}; z)\| \\ &\leq \|\omega_{t-1} - \omega'_{t-1}\| + \|\alpha_{t-1}\nabla f(\omega'_{t-1}; z') - \alpha_{t-1}\nabla f(\omega_{t-1}; z)\| \\ &\leq \delta_{t-1} - \alpha_{t-1}\max(\|\nabla f(\omega'_{t-1}; z') - \nabla f(\omega_{t-1}; z)\|) \\ &\leq \delta_{t-1} - \alpha_{t-1}\|\nabla f(\omega'_{t-1}; z') + \nabla f(\omega_{t-1}; z)\| \\ &\leq \delta_{t-1} - 2\alpha_{t-1}L\end{aligned}$$

By considering both scenarios, a bound on δ_T is deduced through solving a recurrence relation that incorporates the impacts observed in each case.

2) Expansion Properties of Stochastic Gradients

Lemma 31. Assume that f is L-Lipschitz. Then, the gradient update $G_{f,\alpha}$ is (αL)-bounded.

Proof

By Lipschitz assumption,

$$\begin{aligned}\|\omega - G_{f,\alpha}(\omega)\| &= \|\alpha \nabla f(\omega)\| \\ \|\omega - G_{f,\alpha}(\omega)\| &= \alpha \|\nabla f(\omega)\|\| \\ \|\omega - G_{f,\alpha}(\omega)\| &\leq \alpha L \end{aligned} \tag{43}$$

Next, the expansiveness of the gradient update rule is explored for the case of convex and strongly convex loss function f. Considering the loss function f to be β-smooth (recall, $\|\nabla f(u) - \nabla f(v)\| \leq \beta \|u - v\|$), it implies a limitation on how much the gradient update step can expand. Generally, smoothness ensures that the gradient updates do not excessively widen the distance between points. Moreover, when the function is convex and the step size is sufficiently small, the gradient update becomes non-expansive. Furthermore, if the function is strongly convex, the gradient update becomes "contractive," meaning that η (the step size) will be less than one, causing points u and v to shrink closer to each other.

Lemma 32. Assume that f is β-smooth. Then,

- **Lemma 32.1.** $G_{f,\alpha}$ is $(1 + \alpha\beta)$-expansive.
- **Lemma 32.2.** If f is in addition convex, then for any $\alpha \leq 2/\beta$, the update $G_{f,\alpha}$ is 1-expansive.
- **Lemma 32.2.** If f is in addition m-strongly convex, then for $\alpha \leq \frac{2}{\beta+\gamma}$, $G_{f,\alpha}$ is $\left(1 - \frac{\alpha\beta\gamma}{\beta+\gamma}\right)$-expansive.

Proof Consider points $v, u \in \Omega$

- **Lemma 32.1.**

$$\begin{aligned}\|G_{f,\alpha}(u) - G_{f,\alpha}(v)\| &= \|u - \alpha\nabla f(u) - v + \alpha\nabla f(v)\| \\ &= \|u - v + \alpha\nabla f(v) - \alpha\nabla f(u)\| \\ &\leq \|u - v\| + \alpha\|\nabla f(v) - \nabla f(u)\| \\ &\leq \|u - v\| + \alpha\beta\|u - v\| \\ &\leq (1 + \alpha\beta)\|u - v\| \end{aligned} \tag{44}$$

  Following Definition 28, It can be concluded that $G_{f,\alpha}$ is $(1 + \alpha\beta)$ expansive.
- **Lemma 32.2.** That convexity and β-smoothness of the function imply that the gradients are co-coercive.

$$\langle \nabla f(u) - \nabla f(v), u - v \rangle \geq \frac{1}{\beta} \|\nabla f(u) - \nabla f(v)\|^2 . \tag{45}$$

  Using the above property,

$$\begin{aligned}\|G_{f,\alpha}(u) - G_{f,\alpha}(v)\|^2 &= \|v - w\|^2 - 2\alpha\langle \nabla f(u) - \nabla f(v), u - v\rangle + \alpha^2\|\nabla f(u) - \nabla f(v)\|^2 \\ &\leq \|u - v\|^2 - \left(\tfrac{2\alpha}{\beta} - \alpha^2\right)\|\nabla f(u) - \nabla f(v)\|^2\end{aligned}$$

  When, $\alpha \leq \frac{2}{\beta}$ ensures that factor $\left(\frac{2\alpha}{\beta} - \alpha^2\right)$ to be non negative making the bound,

$$\begin{aligned}\|G_{f,\alpha}(u) - G_{f,\alpha}(v)\|^2 &\leq \|u - v\|^2 \\ \|G_{f,\alpha}(u) - G_{f,\alpha}(v)\| &\leq \|u - v\| \end{aligned} \tag{46}$$

  From this, it can be concluded that for f, β-smooth and convex, $G_{f,\alpha}$ is 1 expansive.
- **Lemma 32.3.** For this case the function f is taken to be m - strongly convex. Recall for the case of strong convexity the gradients are co-coercive following,

$$\langle \nabla f(u) - \nabla f(v), u - v, \rangle \geq \frac{\beta\gamma}{\beta+\gamma}\|u-v\|^2 + \frac{1}{\beta+\gamma}\|\nabla f(u) - \nabla f(v)\|^2$$

Using the above inequality the square of $\|G_{f,\alpha}(u) - G_{f,\alpha}(v)\|$ can be written,

$$\begin{aligned}\|G_{f,\alpha}(u) - G_{f,\alpha}(v)\|^2 &= \|u-v\|^2 - 2\alpha\langle \nabla f(u) - \nabla f(v), u - v, \rangle + \alpha^2\|\nabla f(u) - \nabla f(v)\|^2 \\ &\leq \left(1 - 2\frac{\alpha\beta\gamma}{\beta+\gamma}\right)\|u-v\|^2 - \alpha\left(\frac{2}{\beta+\gamma} - \alpha\right)\|\nabla f(u) - \nabla f(v)\|^2\end{aligned}$$

Having, $\alpha \leq \frac{2}{\beta+\gamma}$ makes the factor $\left(\frac{2}{\beta+\gamma} - \alpha\right)$ non negative allowing,

$$\|G_{f,\alpha}(u) - G_{f,\alpha}(w)\|^2 \leq \left(1 - 2\frac{\alpha\beta\gamma}{\beta+\gamma}\right)\|v-w\|^2$$

$$\|G_{f,\alpha}(u) - G_{f,\alpha}(w)\| \leq \left(1 - 2\frac{\alpha\beta\gamma}{\beta+\gamma}^{\frac{1}{2}}\right)\|v-w\|$$

By applying the inequality $\sqrt{1-x} \leq 1 - x/2$ which holds for $x \in [0,1]$.

$$\|G_{f,\alpha}(u) - G_{f,\alpha}(w)\| \leq \left(1 - \frac{\alpha\beta\gamma}{\beta+\gamma}\right)\|v-w\| \tag{47}$$

From this, it can be concluded that for f, β-smooth and $m-$ strongly convex, $G_{f,\alpha}$ is $\left(1 - \frac{\alpha\beta\gamma}{\beta+\gamma}\right)$ expansive.

3) Proof for Stability of SGD when f is Convex

Theorem If the loss function $f(\cdot; z)$ is β-smooth, convex, and L-Lipschitz for every z, and if we execute Stochastic Gradient Method (SGM) with step sizes $\alpha_t \leq 2/\beta$ for T steps, then SGM demonstrates uniform stability.

$$\epsilon_{\text{stab}} \leq \frac{2L^2}{n}\sum_{t=1}^{T}\alpha_t \tag{48}$$

Proof

Consider two samples of size n, denoted as S and S'. These samples differ in only one example. Let's examine the gradient updates $G_1, \ldots, G_T$ and $G'_1, \ldots, G'_T$, which are produced by running Stochastic Gradient Method (SGM) on samples S and S', respectively. The outputs of SGM for samples S and S' are represented by w_T and w'_T, respectively.
Initializing by fixing an example $z \in Z$ and applying the Lipschitz condition on the loss function $f(\cdot; z)$.

$$\mathbb{E}\left|f(\omega_T; z) - f(\omega'_T; z)\right| \leq L\mathbb{E}\left[\omega_T - \omega'_T\right]$$
$$\mathbb{E}\left|f(\omega_T; z) - f(\omega'_T; z)\right| \leq L\mathbb{E}\left[\delta_T\right]$$

Next the following two scenarios are considered,

1) In a step t, the likelihood of SGM selecting the same example from both samples S and S' is $1 - \frac{1}{n}$. When this occurs, denoted as $G_t = G'_t$, we can utilize the 1-expansivity property of the update rule G_t. This property is derived from Lemma 31.2, considering the convexity of the objective function and the constraint $\alpha_t \leq 2/\beta$.
2) With a probability of $1/n$, the selected example differs, in which case we leverage the fact that both G_t and G'_t are bounded by $\alpha_t L$. This follows from **Lemma 30**.

Now by applying **Lemma 29** where,

$$\delta_{t+1} \leq \begin{cases}\delta_t & G_t = G'_t \text{ is 1-expansive} \\ \delta_t + 2\alpha_t L & G_t \text{ and } G'_t \text{ are } \alpha_t L\text{-bounded,} \\ & G_t \text{ is 1 -expansive}\end{cases}$$

Computuing the expectation δ_{t+1},

$$\mathbb{E}\left[\delta_{t+1}\right] \leq \left(1 - \frac{1}{n}\right)\mathbb{E}\left[\delta_t\right] + \frac{1}{n}\left(\mathbb{E}\left[\delta_t\right] + 2\alpha_t L\right)$$
$$\mathbb{E}\left[\delta_{t+1}\right] \leq \mathbb{E}[\delta_t] + \frac{2L\alpha_t}{n}$$

Unraveling the recursion gives,

$$\mathbb{E}\left[\delta_T\right] \leq \frac{2L}{n}\sum_{t=1}^{T}\alpha_t \tag{49}$$

From $\mathbb{E}\left|f(\omega_T; z) - f(\omega'_T; z)\right| \leq L\mathbb{E}\left[\delta_T\right]$, the above bound can be restated as:

$$\mathbb{E}\left|f(w_T; z) - f(w'_T; z)\right| \leq \frac{2L^2}{n}\sum_{t=1}^{T}\alpha_t \tag{50}$$

Since this bound holds for all S, S' and z, we obtain the desired bound on the uniform stability. Even though, this section covers the stability of SGD for the case of loss function $f(;z)$ being β -smooth, convex, and L-Lipschitz, the stability guarantees can be extended into the strongly convex and non-convex settings. The reader can gain an insight into these topics by following [2]. Furthermore, it is worth stating that the uniform stability of an algorithm is related to the generalization error. If an algorithm is uniformly stable, then the generalization error of the algorithm is small. Recall the **Definition 27** of uniform stability. The above mention relationship can be explored through the generalization in expectation.

Theorem Let A, the randomize algorithm which we use to find the model parameters be ϵ uniformly stable then,

$$|\mathbb{E}_{S,A}[R_S[A(S] - R[A(S)]]| \leq \epsilon \tag{51}$$

Proof

Let $S = (z_1, \ldots, z_n)$ and $S' = (z'_1, \ldots, z'_n)$ denote two independent random samples. Define,

$$S^{(i)} = (z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_n)$$

as the sample identical to S except for the i'th example, where z_i is replaced by z'_i. With this notation, we can express:

$$\begin{aligned}
\mathbb{E}_S\mathbb{E}_A\left[R_S[A(S)]\right] &= \mathbb{E}_S\mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^{n} f(A(S); z_i)\right] \\
&= \mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^{n} f(A(S^{(i)}); z'_i)\right] \\
&= \mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^{n} f(A(S); z'_i)\right] + \delta \\
&= \mathbb{E}_S\mathbb{E}_A\left[R[A(S)]\right] + \delta,
\end{aligned}$$

where,

$$\delta = \mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_A\left[\frac{1}{n}\sum_{i=1}^{n} f(A(S^{(i)}); z'_i) - \frac{1}{n}\sum_{i=1}^{n} f(A(S); z'_i)\right]$$

Furthermore, by taking the supremum over any two data sets S and S' differing in only one sample, we can bound the difference as follows:

$$|\delta| \leq \sup_{S,S',z} \mathbb{E}_A \left[f(A(S); z) - f(A(S'); z) \right] \leq \epsilon$$

VI. Numerical SGD Experiments

A. Experimentation Setting

The motivation of the set of experiments was to observe the performance of SGD and its variants. Initially, the experiments were carried out to see the impact of varying the mini-batch sizes on the performance of the algorithm. Then, how the different variants of SGD perform under different conditions on the same optimization problem. All the relevant resources related to the experiments can be found on Convex Optimization Final Project repository.

B. Data Generation

The file Code Files/SGD_Experiments_Data_Generation.py was used to generate data for the subsequent numerical experiments. With specified parameters, this code file is capable of creating synthetic data. For example, the 'linear' function generates linear data suitable for numerical experiments. It takes several parameters:

- n_train: The number of training samples. Note: This does not end up being the number of training samples. This parameter is simply named as such.
- n_test: The number of test samples. Note: This does not end up being the number of test samples. This parameter is simply named as such.
- d: The number of features for each data sample.
- noise_std: The standard deviation of Gaussian noise to be added to the response variable (Y).
- intercept: A boolean parameter indicating whether the linear model has an intercept term.

First, the function generates a matrix X of size $(n\_train + n\_test) \times d$ filled with random numbers drawn from a normal distribution. Next, it generates a weight matrix W of size $d \times 1$, representing the weights of the linear model. If the intercept is set to True, it generates a random bias term b drawn from a normal distribution; otherwise, b is set to 0.

The response variable Y is computed as the matrix product of X and W, with the addition of the bias term b. In mathematical terms, $Y = XW + b$ where b is added element-wise to the vector $V = XW$. Once Y is generated, multivariate Gaussian noise with standard deviation noise_std is added to Y to simulate real-world noise.

Finally, the data is then split into actual training and test sets using the train_test_split function from sci-kit-learn, with a test size of 33% and a fixed random state for reproducibility. Although seemingly redundant, this step ensures that the generated train and test splits are randomly selected.

Overall, this data generation process provides synthetic linear datasets suitable for investigation of the behavior of SGD and its variants under various parameter settings.

C. Experiments

For the first set of experiments, a training dataset consisting of around 2 million samples, each with 1000 features and devoid of noise, was generated. Stochastic Gradient Descent (SGD), with varying batch sizes, were applied to the training data. Essentially, given X and Y, the ground truth is that $XW_T + b_T = Y$ where W_T and b_T represent the ground truth Weight and Bias. The goal of the Stochastic Gradient Descent Training Process and its variants is to be able to recover W_T and b_T, given X and Y, as accurately as possible. The experiments conducted are tabulated below.
Mathematically, if W_E and b_E represent the estimated values of W and b respectively, the training process's goal is to minimize

$$L(W_E, b_E) = ||Y - (XW_E + b_E)||_2^2 = ||Y - XW_E - b_E||_2^2 = ||Y - b_E||_2^2 - 2\langle Y - b_E, XW_E\rangle + ||XW_E||_2^2$$

An important fact to keep in mind is that this is a Convex Problem.
The accompanying figure illustrates the SGD training process. The x-axis of the figure represents the number of iterations or epochs, while the y-axis depicts the value of the loss function. For larger batch sizes, convergence appears more gradual, with smoother fluctuations in the loss over iterations. Conversely, smaller batch sizes show faster convergence but with greater fluctuations due to the noisy nature of stochastic updates. This observation

suggests a trade-off between convergence speed and stability, where the choice of batch size influences the learning dynamics and convergence behavior of the algorithm.

Experiment 1 explored the performance of single-sample and mini-batch stochastic gradient descent (SGD) methods, both with and without noise. It specifically examined how varying the mini-batch sizes affects performance under these conditions, revealing insights into the robustness and efficiency of SGD in handling noise and optimizing computational outcomes.

Both Figures 3, illustrating the results on non-noisy data, and 4, illustrating the results on noisy data, show that Minibatch SGD has smoother Loss curves when compared to SGD. Referring to the Appendix, Section A, it is shown that the Element Wise Variance of the Stochastic Gradient Decreases as the Batch Size Increases. Hence, it is expected that when the batch size increases, the stochastic gradient will be less noisy and the loss curve will end up being more smoother.

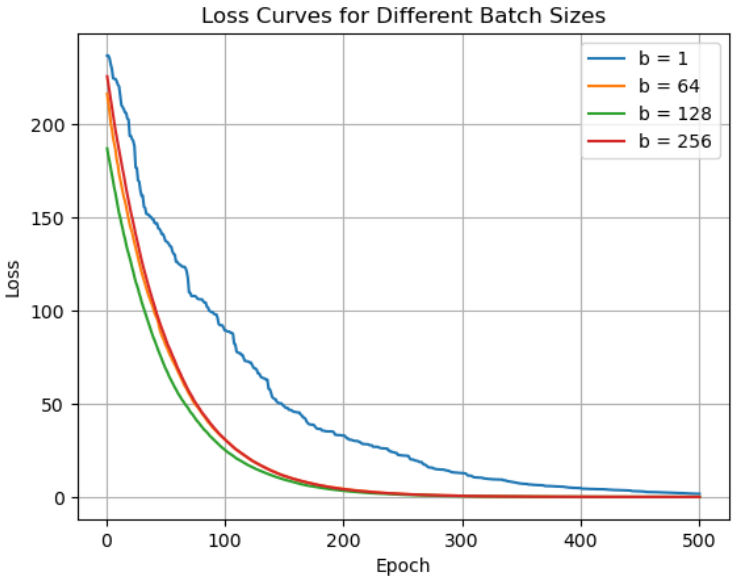


Fig. 3: NAG with $\gamma = 0.001$, $\alpha = 0$

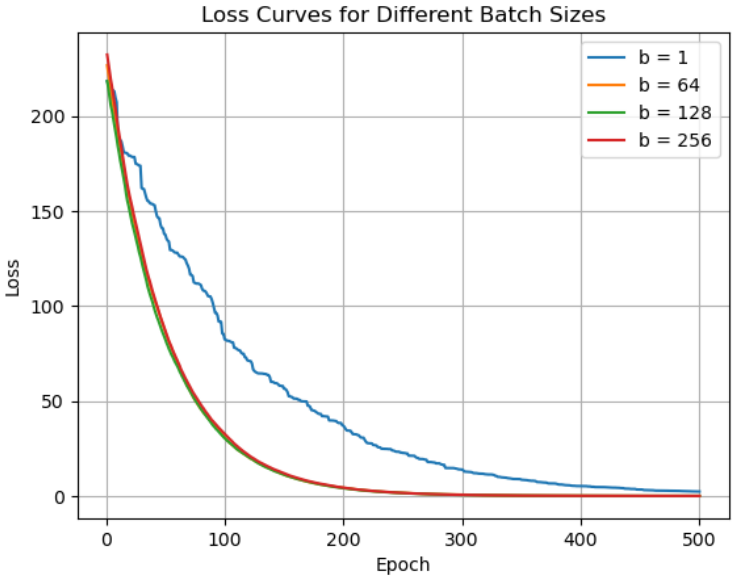


Fig. 4: NAG with $\gamma = 0.001$, $\alpha = 0$

Experiment 2 focused on evaluating different variants of stochastic gradient descent (SGD), such as momentum-based and adaptive learning rate methods. The aim was to understand how these modifications influence learning dynamics and optimization efficiency, on a convex optimization problem. Figure4 illustrate the results on the experiments on the dataset without noise. From the figure, it can be concluded that with the same learning rate, the rate reduction of NAG is faster compared to the momentum variant which aligns well with the discussion carried out in Section 3. It is worth noting that the global learning rates required by Adagrad and the RMSProp were comparatively larger than the rest. Comparing Adagrad and RMSProp algorithms it can be seen that the reduction of loss achieved by the Adagrad variant is small compared to the others. This is due to the monotonically decreasing learning rate which causes the learning rate to be too small too quickly. It is well illustrated in the Figures 5 and 6 that using the decaying factor ρ on the previous gradient terms in the RMSProp method has affected to reduction achieved in the loss.

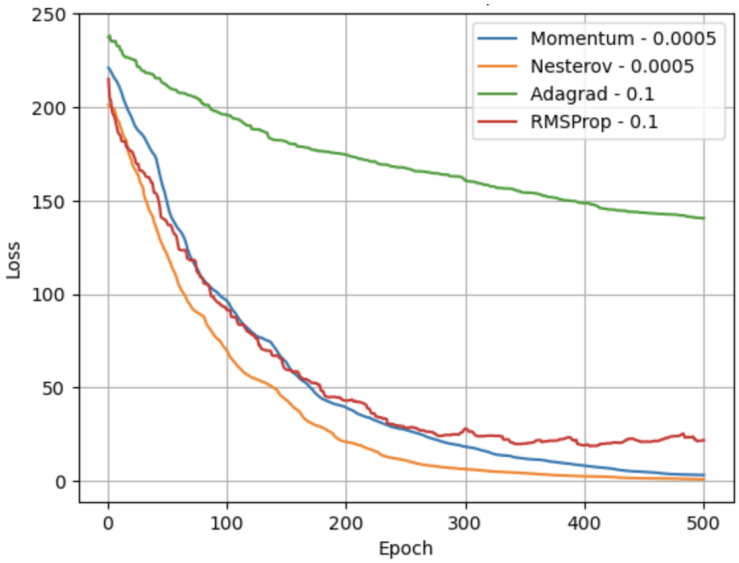


Fig. 5: NAG with $\gamma = 0.001$, $\alpha = 0$

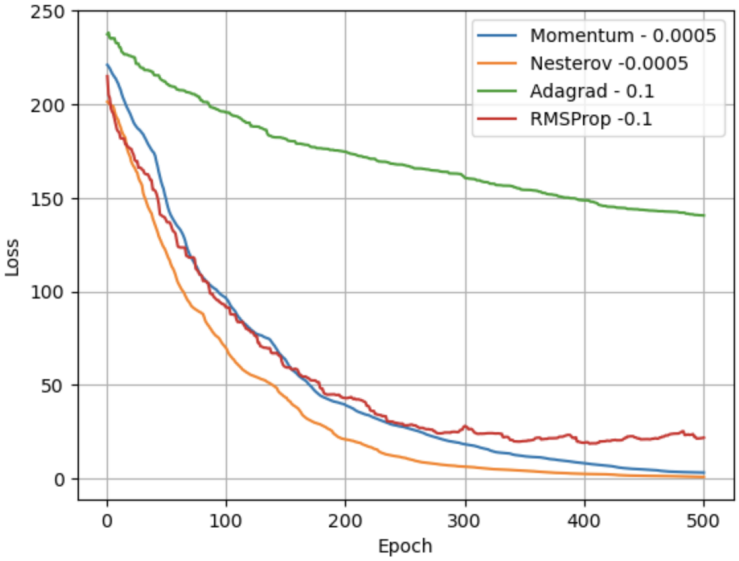


Fig. 6: NAG with $\gamma = 0.001$, $\alpha = 0$

References

[1] G. Garrigos and R. M. Gower. *Handbook of Convergence Theorems for (Stochastic) Gradient Methods*. 2024. arXiv: 2301.11235 [math.OC].

[2] M. Hardt, B. Recht, and Y. Singer. "Train faster, generalize better: Stability of stochastic gradient descent". In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by M. F. Balcan and K. Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1225–1234. URL: https://proceedings.mlr.press/v48/hardt16.html.

[3] N. Ketkar and N. Ketkar. "Stochastic gradient descent". In: *Deep learning with Python: A hands-on introduction* (2017), pp. 113–132.

VII. Appendix

A. Element-Wise Variance Analysis for SGD and Minibatch SGD

1) Notation

$[x]^2$ is the vector obtained by raising each element of the vector x to the 2nd power.

$x \cdot y$ is the vector obtained by doing an element wise multiplication of the vectors x and y

$x \leq y$, where x and y are vectors, means that each component of x is less than or equal to the corresponding component of y

$x \geq y$, where x and y are vectors, means that each component of x is greater than or equal to the corresponding component of y

$x = c$, where x is a vector and c is a scalar, means that each component of x is equal to c

$x \geq c$, where x is a vector and c is a scalar, means that each component of x is greater than or equal to c

$x \leq c$, where x is a vector and c is a scalar, means that each component of x is less than or equal to c

Expected Value: For $X \in \mathbb{R}^n$, $\mathbb{E}[X] = [\mathbb{E}[X_1], \mathbb{E}[X_2], \ldots, \mathbb{E}[X_n]]^T$

Element-Wise Variance: For $X \in \mathbb{R}^n$, $\mathbb{V}[X] = \mathbb{E}[||X - \mathbb{E}[X]||_2^2]$

2) Recap

SGD Expectation: $E[\nabla f_{i_t}(x)] = \nabla f(x)$.

Minibatch SGD Expectation: $\mathbb{E}[\nabla f_{B_t}(x^{(t)})] = \nabla f(x^{(t)})$.

The goal of this section is to analyze the Element-Wise Variance of $\nabla f_{i_t}(x)$ and $\nabla f_{B_t}(x^{(t)})$

3) Element-Wise Variance Analysis [SGD]

Our goal is to compute $\mathbb{V}[\nabla f_{i_t}(x)]$.

Since we know that, for a random variable X, $\mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ is the case, we can solve for the Element-Wise Variance as such

$$\mathbb{V}[\nabla f_{i_t}(x)] = \frac{1}{m}\sum_{i=1}^{m}[\nabla f_i(x)]^2 - [\nabla f(x)]^2$$

4) Element-Wise Variance Analysis [Minibatch SGD]

Our goal is to compute $\mathbb{V}[\nabla f_{B_t}(x)]$. Since we know that, for a random variable X, $\mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ is the case, we can solve for the Element-Wise Variance as such

$$\mathbb{V}[\nabla f_{B_t}(x)] = \frac{1}{\binom{m}{b}}\sum_{B\subset\{1\ldots n\},|B|=b}[\nabla f_B(x)]^2 - [\nabla f(x)]^2$$

$$\mathbb{V}[\nabla f_{B_t}(x)] = \frac{1}{\binom{m}{b}}\sum_{B\subset\{1\ldots n\},|B|=b}[\frac{1}{b}\sum_{j=1}^{b}\nabla f_{B_j}(x)]^2 - [\nabla f(x)]^2$$

$$\mathbb{V}[\nabla f_{B_t}(x)] = \frac{1}{\binom{m}{b}}\frac{1}{b^2}\sum_{B\subset\{1\ldots n\},|B|=b}[\sum_{j=1}^{b}\nabla f_{B_j}(x)]^2 - [\nabla f(x)]^2$$

$$\mathbb{V}[\nabla f_{B_t}(x)] = \frac{1}{\binom{m}{b}}\frac{1}{b^2}\binom{m-1}{b-1}\sum_{i=1}^{m}[\nabla f_i(x)]^2 + \frac{1}{\binom{m}{b}}\frac{1}{b^2}\binom{m-2}{b-2}\sum_i\sum_{j\neq i} 2\nabla f_i(x)\cdot\nabla f_j(x) - [\nabla f(x)]^2$$

$$\mathbb{V}[\nabla f_{B_t}(x)] = \frac{1}{bm}\sum_{i=1}^{m}[\nabla f_i(x)]^2 + \frac{1}{bm}\frac{b-1}{m-1}\sum_i\sum_{j\neq i} 2\nabla f_i(x)\cdot\nabla f_j(x) - [\nabla f(x)]^2$$

Remark: When $b = m$,

$$\mathbb{V}[\nabla f_{B_t}(x)] = \frac{1}{m^2}\sum_{i=1}^{m}[\nabla f_i(x)]^2 + \frac{1}{m^2}\frac{m-1}{m-1}\sum_i\sum_{j\neq i} 2\nabla f_i(x)\cdot\nabla f_j(x) - [\nabla f(x)]^2$$

$$\mathbb{V}[\nabla f_{B_t}(x)] = \frac{1}{m^2}\sum_{i=1}^{m}[\nabla f_i(x)]^2 + \frac{1}{m^2}\sum_i\sum_{j\neq i} 2\nabla f_i(x)\cdot\nabla f_j(x) - [\nabla f(x)]^2$$

Since $\nabla f(x) = \frac{1}{m}\sum_{i=1}^{m}\nabla f_i(x)$, our equation can be further simplified:

$$\mathbb{V}[\nabla f_{B_t}(x)] = [\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(x)]^2 - [\nabla f(x)]^2$$

$$\mathbb{V}[\nabla f_{B_t}(x)] = [\nabla f(x)]^2 - [\nabla f(x)]^2 = 0$$

This corresponds with the notion that, when using all f_i functions to compute the gradient, we will **always** get the true value of $\nabla f(x)$

5) SGD vs Minibatch SGD Analysis

The goal of this section is to show that, as we increase the batch size, the Element-Wise Variance of the Stochastic Gradient decreases(i.e. decreases element-wise).

When $b = k$,

$$\mathbb{V}_k[\nabla f_{B_t}(x)] = \frac{1}{km}\sum_{i=1}^{m}[\nabla f_i(x)]^2 + \frac{1}{km}\frac{k-1}{m-1}\sum_i\sum_{j\neq i} 2\nabla f_i(x)\cdot\nabla f_j(x) - [\nabla f(x)]^2$$

When $b = k+1$,

$$\mathbb{V}_{k+1}[\nabla f_{B_t}(x)] = \frac{1}{(k+1)m}\sum_{i=1}^{m}[\nabla f_i(x)]^2 + \frac{1}{(k+1)m}\frac{k}{m-1}\sum_i\sum_{j\neq i} 2\nabla f_i(x)\cdot\nabla f_j(x) - [\nabla f(x)]^2$$

Goal: Prove that $\mathbb{V}_{k+1}[\nabla f_{B_t}(x)] \leq \mathbb{V}_k[\nabla f_{B_t}(x)]$

Proof:
Let's approach this via a Proof of Contradiction. Let's start by assuming that $\mathbb{V}_{k+1}[\nabla f_{B_t}(x)] > \mathbb{V}_k[\nabla f_{B_t}(x)]$

$$\frac{1}{(k+1)m}\sum_{i=1}^{m}[\nabla f_i(x)]^2 + \frac{1}{(k+1)m}\frac{k}{m-1}\sum_i\sum_{j\neq i} 2\nabla f_i(x)\cdot\nabla f_j(x) - [\nabla f(x)]^2$$
$$> \frac{1}{km}\sum_{i=1}^{m}[\nabla f_i(x)]^2 + \frac{1}{km}\frac{k-1}{m-1}\sum_i\sum_{j\neq i} 2\nabla f_i(x)\cdot\nabla f_j(x) - [\nabla f(x)]^2$$

Getting rid of the $-[\nabla f(x)]^2$ on both sides of the inequality

$$\frac{1}{(k+1)m}\sum_{i=1}^{m}[\nabla f_i(x)]^2 + \frac{1}{(k+1)m}\frac{k}{m-1}\sum_i\sum_{j\neq i}2\nabla f_i(x)\cdot\nabla f_j(x)$$
$$> \frac{1}{km}\sum_{i=1}^{m}[\nabla f_i(x)]^2 + \frac{1}{km}\frac{k-1}{m-1}\sum_i\sum_{j\neq i}2\nabla f_i(x)\cdot\nabla f_j(x)$$

Since $\frac{1}{km} - \frac{1}{(k+1)m} = \frac{1}{m}(\frac{1}{k} - \frac{1}{k+1}) = \frac{1}{m(k+1)k}$ and $\frac{1}{(k+1)m}\frac{k}{m-1} - \frac{1}{km}\frac{k-1}{m-1} = \frac{1}{m(m-1)}(\frac{k}{k+1} - \frac{k-1}{k}) = \frac{1}{m(m-1)k(k+1)}$

Rearranging Terms:

$$\frac{1}{m(m-1)k(k+1)}\sum_i\sum_{j\neq i}2\nabla f_i(x)\cdot\nabla f_j(x) > \frac{1}{m(k+1)k}\sum_{i=1}^{m}[\nabla f_i(x)]^2$$

Divide both sides by $\frac{1}{m(k+1)k}$

$$\frac{1}{m-1}\sum_i\sum_{j\neq i}2\nabla f_i(x)\cdot\nabla f_j(x) > \sum_{i=1}^{m}[\nabla f_i(x)]^2$$

Let's show how this is a contradiction:

$$\sum_i\sum_{j\neq i}2\nabla f_i(x)\cdot\nabla f_j(x) > (m-1)\sum_{i=1}^{m}[\nabla f_i(x)]^2$$

Add $\sum_{i=1}^{m}[\nabla f_i(x)]^2$ to each side of this inequality

$$\sum_{i=1}^{m}[\nabla f_i(x)]^2 + \sum_i\sum_{j\neq i}2\nabla f_i(x)\cdot\nabla f_j(x) > (m-1)\sum_{i=1}^{m}[\nabla f_i(x)]^2 + \sum_{i=1}^{m}[\nabla f_i(x)]^2$$

Since $\nabla f(x) = \frac{1}{m}\sum_{i=1}^{m}\nabla f_i(x)$

$$[m\nabla f(x)]^2 > m\sum_{i=1}^{m}[\nabla f_i(x)]^2$$

$$m^2[\nabla f(x)]^2 > m\sum_{i=1}^{m}[\nabla f_i(x)]^2$$

$$[\nabla f(x)]^2 > \frac{1}{m}\sum_{i=1}^{m}[\nabla f_i(x)]^2$$

However, according to the Cauchy Schwarz Inequality

$$[\nabla f(x)]^2 = [\sum_{i=1}^{m}\frac{1}{m}\nabla f_i(x)]^2 \leq \sum_{i=1}^{m}(\frac{1}{m})^2\sum_{i=1}^{m}[\nabla f_i(x)]^2$$

$$[\nabla f(x)]^2 \leq \frac{1}{m}\sum_{i=1}^{m}[\nabla f_i(x)]^2$$

Hence, by arriving at a contradiction, our original assumption that $\mathbb{V}_{k+1}[\nabla f_{B_t}(x)] > \mathbb{V}_k[\nabla f_{B_t}(x)]$ is False. It must be the case that $\mathbb{V}_{k+1}[\nabla f_{B_t}(x)] \leq \mathbb{V}_k[\nabla f_{B_t}(x)]$. This means that as we increase the batch size, the element-wise variance of the stochastic gradient decreases.

B. Sum-Integral Bounds

Let $f : \mathbb{R}_{++} \to \mathbb{R}_{++}$ which is also decreasing.

$$\int_1^{T+1} f(x)\,dx \leq \sum_{x=1}^{T} f(x) = f(1) + \sum_{x=2}^{T} f(x) \leq f(1) + \int_1^{T} f(x)\,dx$$

$$\int_1^{T+1} f(x)\,dx \leq \sum_{x=1}^{T} f(x) \leq f(1) + \int_1^{T} f(x)\,dx$$

Let's now use the functions $f(x) = \frac{1}{\sqrt{x}}$ and $f(x) = \frac{1}{x}$
Let's start with the function $f(x) = \frac{1}{\sqrt{x}}$

$$\int_1^{T+1} \frac{1}{\sqrt{x}}\,dx \leq \sum_{x=1}^{T} \frac{1}{\sqrt{x}} \leq 1 + \int_1^{T} \frac{1}{\sqrt{x}}\,dx$$

Let's start by working towards the lower bound

$$\int_1^{T+1} \frac{1}{\sqrt{x}}\,dx = [2\sqrt{x}]_1^{T+1} = 2\sqrt{T+1} - 2 = 2(\sqrt{T+1} - 1)$$

We know that $\inf_{T\geq 1} \frac{\sqrt{T+1}-1}{\sqrt{T}} = \sqrt{2} - 1 > \frac{2}{5}$

$$\int_1^{T+1} \frac{1}{\sqrt{x}}\,dx = [2\sqrt{x}]_1^{T+1} = 2\sqrt{T+1} - 2 = 2(\sqrt{T+1} - 1) \geq \frac{4}{5}\sqrt{T}$$

Now let's look at the upper bound

$$1 + \int_1^{T} \frac{1}{\sqrt{x}}\,dx = 1 + 2\sqrt{T+1} - 2 = 2\sqrt{T+1} - 1$$

Combining both the Lower and Upper Bounds gives us:

$$\frac{4}{5}\sqrt{T} \leq \sum_{x=1}^{T} \frac{1}{\sqrt{x}} \leq 2\sqrt{T+1} - 1$$

Now let's analyze the other function $f(x) = \frac{1}{x}$

$$\int_1^{T+1} \frac{1}{x}\,dx \leq \sum_{x=1}^{T} \frac{1}{x} \leq 1 + \int_1^{T} \frac{1}{x}\,dx$$

Let's start by working towards the lower bound

$$\int_1^{T+1} \frac{1}{x}\,dx = [\log(t)]_1^{T+1} = \log(T+1)$$

Now let's look at the upper bound:

$$1 + \int_1^{T} \frac{1}{x}\,dx = 1 + [\log(t)]_1^{T} = 1 + \log(T) \leq 2\log(T+1)$$

Note: We know that $\sup_{T\geq 1} \frac{1+\log(T)}{\log(T+1)} \approx \sqrt{2} < 2$

$$\log(T+1) \leq \sum_{x=1}^{T} \frac{1}{x} \leq 2\log(T+1)$$

C. Proofs of Introductory Lemmas

Proof of Lemma 12:
Let $g(t) = f(x + t(y - x))$. Based on the Fundamental Theorem of Calculus

$$f(y) - f(x) = \int_0^1 g'(t)\, dt$$

$$f(y) - f(x) = \int_0^1 \nabla f(x + t(y - x))^T (y - x)\, dt$$

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle\, dt$$

$$f(y) - f(x) = \langle \nabla f(x), y - x \rangle + \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle\, dt \tag{52}$$

Within the integral of 52, Apply the Cauchy Schwarz Inequality

$$f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| \ \|y - x\|\, dt \tag{53}$$

According to Definition 11.

$$f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + \int_0^1 tL\|y - x\|^2\, dt$$

$$f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \tag{54}$$

Substituting $y = x - \gamma \nabla f(x)$ into 54

$$f(x - \gamma \nabla f(x)) - f(x) \leq \langle \nabla f(x), x - \gamma \nabla f(x) - x \rangle + \frac{L}{2}\|x - \gamma \nabla f(x) - x\|^2$$

$$f(x - \gamma \nabla f(x)) - f(x) \leq \langle \nabla f(x), -\gamma \nabla f(x) \rangle + \frac{L}{2}\| - \gamma \nabla f(x)\|^2$$

$$f(x - \gamma \nabla f(x)) - f(x) \leq -\langle \nabla f(x), \gamma \nabla f(x) \rangle + \frac{L\gamma^2}{2}\|\nabla f(x)\|^2$$

$$f(x - \gamma \nabla f(x)) - f(x) \leq -\gamma \|\nabla f(x)\|^2 + \frac{L\gamma^2}{2}\|\nabla f(x)\|^2$$

$$f(x - \gamma \nabla f(x)) - f(x) \leq (-\gamma + \frac{L\gamma^2}{2})\|\nabla f(x)\|^2 \tag{55}$$

Assuming that $\inf f > -\infty$ and $\gamma = \frac{1}{L}$, 55 can be further simplified

$$\inf f - f(x) \leq f(x - \frac{1}{L}\nabla f(x)) - f(x) \leq (-\frac{1}{2L})\|\nabla f(x)\|^2$$

$$(\frac{1}{2L})\|\nabla f(x)\|^2 \leq f(x) - \inf f$$

Proof of Lemma 13

$$f(x) - f(y) = f(x) - f(z) + f(z) - f(y) \tag{56}$$

Applying the first order condition of convexity

$$f(z) \geq f(x) + \nabla f(x)^T (z - x)$$

$$f(x) - f(z) \leq -\nabla f(x)^T (z - x) = \nabla f(x)^T (x - z) \tag{57}$$

Using the Upper Bound of L-Smooth Functions

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

$$f(z) - f(y) \leq \langle \nabla f(y), z - y \rangle + \frac{L}{2} \|z - y\|^2 \tag{58}$$

Combining 56, 57, and 58

$$f(x) - f(y) = f(x) - f(z) + f(z) - f(y) \leq \nabla f(x)^T (x - z) + \langle \nabla f(y), z - y \rangle + \frac{L}{2} \|z - y\|^2 \tag{59}$$

Minimize the Right Hand Side of 59 with respect to z by first computing the gradient of the Right Hand Side with respect to z

$$-\nabla f(x) + \nabla f(y) + \frac{L}{2}(2z - 2y) \tag{60}$$

Setting 60 to zero means that

$$z = y - \frac{1}{L}(\nabla f(y) - \nabla f(x)) \tag{61}$$

Substitute 61 into the Right Hand Side of 59

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

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

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

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

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

$$\nabla f(x)^T (x - y) - \frac{1}{L} \|\nabla f(y) - \nabla f(x)\|^2 + \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2$$

$$\nabla f(x)^T (x - y) - \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2$$

Since the Right Hand Side of 59 was minimized with respect to z

$$f(x) - f(y) \leq \nabla f(x)^T (x - y) - \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2$$

$$\frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle$$

Proof of Lemma 14: According to Lemma 13 and the fact that $L_i \leq L_{max}$:

$$\frac{1}{2L_{max}}\|\nabla f_i(y) - \nabla f_i(x)\|^2 \leq \frac{1}{2L_i}\|\nabla f_i(y) - \nabla f_i(x)\|^2 \leq f_i(y) - f_i(x) - \langle \nabla f_i(x), y - x\rangle$$

$$\frac{1}{2L_{max}}\|\nabla f_i(y) - \nabla f_i(x)\|^2 \leq f_i(y) - f_i(x) - \langle \nabla f_i(x), y - x\rangle$$

Take Expectation of both sides

$$\frac{1}{2L_{max}}\mathbb{E}[\|\nabla f_i(y) - \nabla f_i(x)\|^2] \leq \mathbb{E}[(f_i(y) - f_i(x) - \langle \nabla f_i(x), y - x\rangle)]$$

Apply Linearity of Expectation on the Right Hand Side of the above Inequality

$$\frac{1}{2L_{max}}\mathbb{E}[\|\nabla f_i(y) - \nabla f_i(x)\|^2] \leq f(y) - f(x) - \langle \nabla f(x), y - x\rangle$$

Proof of Lemma 15:
Using Lemma 14

$$\frac{1}{2L_{max}}\mathbb{E}[\|\nabla f_i(x) - \nabla f_i(x^*)\|^2] \leq f(x) - f(x^*) - \langle \nabla f(x^*), x - x^*\rangle$$

Since $\nabla f(x^*) = 0$

$$\frac{1}{2L_{max}}\mathbb{E}[\|\nabla f_i(x) - \nabla f_i(x^*)\|^2] \leq f(x) - \inf f$$

Proof of Lemma 17. Since interpolation holds at x^*, $f_i(x^*) = \inf f_i, \forall i = 1, \dots m$. Furthermore, $\forall x \in dom f$

$$f(x^*) = \frac{1}{m}\sum_{i=1}^{m} f_i(x^*) = \frac{1}{m}\sum_{i=1}^{m} inf f_i \leq \frac{1}{m}\sum_{i=1}^{m} f_i(x) = f(x)$$

Proof of Lemma 19.
Let $x^* \in \arg\min f$

$$\Delta_f^* = \inf f - \frac{1}{m}\sum_{i=1}^{m} \inf f_i = f(x^*) - \frac{1}{m}\sum_{i=1}^{m} \inf f_i \geq f(x^*) - \frac{1}{m}\sum_{i=1}^{m} f_i(x^*) = f(x^*) - f(x^*) = 0$$

$$\Delta_f^* \geq 0$$

Goal: Prove that Interpolation Holds if and only if $\Delta_f^* = 0$.

Step 1: Prove that if Interpolation holds, then $\Delta_f^* = 0$
Due to interpolation, $\inf f_i = f_i(x^*)$. Hence,

$$\Delta_f^* = f(x^*) - \frac{1}{m}\sum_{i=1}^{m} \inf f_i = f(x^*) - \frac{1}{m}\sum_{i=1}^{m} f_i(x^*) = f(x^*) - f(x^*) = 0$$

$$\Delta_f^* = 0$$

Step 2: Prove that if $\Delta_f^* = 0$, then Interpolation holds

$$\Delta_f^* = f(x^*) - \frac{1}{m}\sum_{i=1}^{m} \inf f_i \geq f(x^*) - \frac{1}{m}\sum_{i=1}^{m} f_i(x^*) = f(x^*) - f(x^*) = 0$$

Subsequently,

$$\sum_{i=1}^{m} \inf f_i = \sum_{i=1}^{m} f_i(x^*)$$

$$\sum_{i=1}^{m}(f_i(x^*) - \inf f_i) = 0$$

Conclusion: $f_i(x^*) = \inf f_i$, meaning Interpolation Holds!

Proof of Lemma 21
Let $x_1, x_2 \in \arg\min f$. Based on Lemma 15,

$$\frac{1}{2L_{max}}\mathbb{E}[\|\nabla f_i(x_1) - \nabla f_i(x_2)\|^2] \leq f(x_1) - \inf f = \inf f - \inf f = 0$$

Since the square of a norm cannot be negative, for the above to be true, $\mathbb{E}[\|\nabla f_i(x_1) - \nabla f_i(x_2)\|^2] = 0$ and $\|\nabla f_i(x_1) - \nabla f_i(x_2)\|^2 = 0$. Subsequently, $f_i(x_1) = f_i(x_2)$ and $\mathbb{V}[\nabla f_i(x_1)] = \mathbb{V}[\nabla f_i(x_2)]$

Since $\mathbb{V}[\nabla f_i(x_1)] = \mathbb{V}[\nabla f_i(x_2)]$, $\sigma_f^* = \mathbb{V}[\nabla f_i(x_1)] = \mathbb{V}[\nabla f_i(x_2)] = \mathbb{V}[\nabla f_i(x^*)]$

Next Step: Prove that Interpolation Holds if and only if $\sigma_f^* = 0$

Step 1: Prove that if Interpolation holds, then $\sigma_f^* = 0$

If Interpolation holds, then there is a $x^* \in \arg\min f$ where $x^* \in \arg\min f_i$. This means that $\nabla f_i(x^*) = 0, \forall i \in [1, m]$. This means that $\mathbb{V}[\nabla f_i(x^*)] = 0$ and that $\sigma_f^* = 0$

Step 2: Prove that if $\sigma_f^* = 0$, then Interpolation holds

Since f is Convex and $\sigma_f^* = 0$,

$$\forall x^* \in \arg\min f, \mathbb{V}[\nabla f_i(x^*)] = 0$$

Since $x^* \in \arg\min f$, $\nabla f(x^*) = 0$. Since $\mathbb{V}[\nabla f_i(x^*)] = 0$, this means that $\nabla f_i(x^*) = 0$. Due to the definition of convexity, since $\nabla f_i(x^*) = 0$, $x^* \in \arg\min f_i$ as well! This means that Interpolation holds.

Proof of Lemma 22:
Let $x^* \in \arg\min f$. Based on Lemma 12

$$\|\nabla f(x)\|^2 \leq 2L(f(x) - \inf f)$$

Since each f_i is $L-$Smooth,

$$\|\nabla f_i(x^*)\|^2 \leq 2L_i(f_i(x^*) - \inf f_i) \leq 2L_{max}(f_i(x^*) - \inf f_i) \tag{62}$$

Taking Expectations on both sides of 62:

$$\mathbb{E}[\|\nabla f_i(x^*)\|^2] \leq \mathbb{E}[2L_{max}(f_i(x^*) - \inf f_i)]$$

$$\mathbb{E}[\|\nabla f_i(x^*)\|^2] \leq 2L_{max}\mathbb{E}[(f_i(x^*) - \inf f_i)] \tag{63}$$

Since $\nabla f(x^*) = 0$

$$\mathbb{E}[\|\nabla f_i(x^*)\|^2] = \mathbb{E}[\|\nabla f_i(x^*) - \nabla f(x^*)\|^2] = \mathbb{E}[\|\nabla f_i(x^*) - \mathbb{E}[\nabla f_i(x^*)]\|^2] = \mathbb{V}[\nabla f_i(x^*)] \geq \sigma_f^* \tag{64}$$

Applying Linearity of Expectation on the Right Hand Side of 63,

$$2L_{max}\mathbb{E}[(f_i(x^*) - \inf f_i)] = 2L_{max}(\mathbb{E}[(f_i(x^*))] - \mathbb{E}[(\inf f_i)]) = 2L_{max}(f(x^*) - \frac{1}{n}\sum_{i=1}^{n}\inf f_i)$$

$$2L_{max}(\inf f - \frac{1}{n}\sum_{i=1}^{n}\inf f_i) = 2L_{max}\Delta_f^* \tag{65}$$

63, 64, and 65 show that:

$$\sigma_f^* \leq 2L_{max}\Delta_f^*$$

Part (2) of Proof: Show that, if each f_i is p strongly convex, then $2p\Delta_f^* \leq \sigma_f^*$

Applying the definition of Strong Convexity

$$f(y) \geq f(x) + \nabla(f(x))^T(y-x) + \frac{p}{2}\|y-x\|_2^2$$

Set $y = x^*$ and $x = x$

$$f(x^*) \geq f(x) + \nabla(f(x))^T(x^*-x) + \frac{p}{2}\|x^*-x\|_2^2$$

$$\begin{aligned} f(x) - f(x^*) &\leq \nabla f(x)^T(x-x^*) - \frac{p}{2}\|x^*-x\|_2^2 \\ &\leq \frac{-1}{2}\|\sqrt{p}(x-x^*) - \frac{1}{\sqrt{p}}\nabla f(x)\|^2 + \frac{1}{2p}\|\nabla f(x)\|^2 \leq \frac{1}{2p}\|\nabla f(x)\|^2 \end{aligned}$$

$$f_i(x) - \inf f_i \leq \frac{1}{2p}\|\nabla f_i(x)\|^2$$

$$f_i(x^*) - \inf f_i \leq \frac{1}{2p}\|\nabla f_i(x^*)\|^2$$

Take Expectation over both sides of the above Inequality

$$\mathbb{E}[f_i(x^*) - \inf f_i] \leq \frac{1}{2p}\mathbb{E}[\|\nabla f_i(x^*)\|^2]$$

Applying Linearity of Expectation to the Left Side of the above inequality:

$$\mathbb{E}[f_i(x^*)] - \mathbb{E}[\inf f_i] \leq \frac{1}{2p}\mathbb{E}[\|\nabla f_i(x^*)\|^2] \tag{66}$$

Since $\nabla f(x^*) = 0$, $\mathbb{E}[\|\nabla f_i(x^*)\|^2] = \mathbb{E}[\|\nabla f_i(x^*) - \nabla f(x^*)\|^2] = \mathbb{E}[\|\nabla f_i(x^*) - \mathbb{E}[\nabla f_i(x^*)]\|^2] = \mathbb{V}[\nabla f_i(x^*)]$

Additionally, $\mathbb{E}[f_i(x^*)] = f(x^*) = \inf f$
Applying the aforementioned facts to 66

$$\inf f - \frac{1}{n}\sum_{i=1}^{n}\inf f_i \leq \frac{1}{2p}\mathbb{V}[\nabla f_i(x^*)]$$

$$\Delta_f^* \leq \frac{1}{2p}\mathbb{V}[\nabla f_i(x^*)]$$

Due to convexity, Lemma 21 states that $\sigma_f^* = \mathbb{V}[\nabla f_i(x^*)]$

$$\Delta_f^* \leq \frac{1}{2p}\sigma_f^*$$

$$2p\Delta_f^* \leq \sigma_f^*$$

Proof of Lemma 23:

As per Lemma 12,

$$\frac{1}{2L_i}\|\nabla f_i(x)\|^2 \leq f_i(x) - \inf f_i$$

$$\begin{aligned}\|\nabla f_i(x)\|^2 \leq 2L_i(f_i(x) - \inf f_i) &\leq 2L_{max}(f_i(x) - \inf f_i) \\ &\leq 2L_{max}(f_i(x) - f_i(x^*)) + 2L_{max}(f_i(x^*) - \inf f_i)\end{aligned}$$

$$\|\nabla f_i(x)\|^2 \leq 2L_{max}(f_i(x) - f_i(x^*)) + 2L_{max}(f_i(x^*) - \inf f_i) \tag{67}$$

Take Expectation over both sides of the above inequality:

$$\mathbb{E}[\|\nabla f_i(x)\|^2] \leq 2L_{max}\mathbb{E}[(f_i(x) - f_i(x^*))] + 2L_{max}\mathbb{E}[(f_i(x^*) - \inf f_i)] \tag{68}$$

Apply Linearity of Expectation on the Right Hand Side of the above inequality

$$\mathbb{E}[\|\nabla f_i(x)\|^2] \leq 2L_{max}(\mathbb{E}[(f_i(x))] - \mathbb{E}[(f_i(x^*))]) + 2L_{max}(\mathbb{E}[(f_i(x^*))] - \mathbb{E}[(\inf f_i)])$$

$$\mathbb{E}[\|\nabla f_i(x)\|^2] \leq 2L_{max}(f(x) - \inf f) + 2L_{max}(\inf f - E[\inf f_i])$$

$$\mathbb{E}[\|\nabla f_i(x)\|^2] \leq 2L_{max}(f(x) - \inf f) + 2L_{max}(\inf f - \frac{1}{n}\sum_{i=1}^{n}\inf f_i)$$

$$\mathbb{E}[\|\nabla f_i(x)\|^2] \leq 2L_{max}(f(x) - \inf f) + 2L_{max}\Delta_f^*$$

Proof of Lemma 24 Let $x^* \in \arg\min f$

$$\|\nabla f_i(x)\|^2 = \|\nabla f_i(x) - \nabla f_i(x^*) + f_i(x^*)\|^2 \tag{69}$$

Apply the Triangle Inequality Theorem to the Right Hand Side of 69

$$\|\nabla f_i(x) - \nabla f_i(x^*) + f_i(x^*)\|^2 \leq (\|\nabla f_i(x) - \nabla f_i(x^*)\| + \|f_i(x^*)\|)^2$$

$$\|\nabla f_i(x) - \nabla f_i(x^*) + f_i(x^*)\|^2 \leq \|\nabla f_i(x) - \nabla f_i(x^*)\|^2 + 2\|\nabla f_i(x) - \nabla f_i(x^*)\|\|\nabla f_i(x^*)\| + \|\nabla f_i(x^*)\|^2 \tag{70}$$

Known Fact:

$$\|\nabla f_i(x) - \nabla f_i(x^*) - f_i(x^*)\|^2 \geq 0$$

$$\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 - 2\|\nabla f_i(x) - \nabla f_i(x^*)\|\|\nabla f_i(x^*)\| + \|\nabla f_i(x^*)\|^2 \geq 0$$

$$\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 + \|\nabla f_i(x^*)\|^2 \geq 2\|\nabla f_i(x) - \nabla f_i(x^*)\|\|\nabla f_i(x^*)\| \tag{71}$$

Putting together 70 and 71

$$\|\nabla f_i(x) - \nabla f_i(x^*) + f_i(x^*)\|^2 \leq 2\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 + 2\|\nabla f_i(x^*)\|^2$$

$$\|\nabla f_i(x)\|^2 \leq 2\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 + 2\|\nabla f_i(x^*)\|^2$$

Take the expectation over the above inequality

$$\mathbb{E}[\|\nabla f_i(x)\|^2] \leq \mathbb{E}[2\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 + 2\|\nabla f_i(x^*)\|^2]$$

$$\mathbb{E}[\|\nabla f_i(x)\|^2] \leq 2\mathbb{E}[\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 + \|\nabla f_i(x^*)\|^2] \tag{72}$$

According to **Lemma 15**,

$$\frac{1}{2L_{max}}\mathbb{E}[\|\nabla f_i(x) - \nabla f_i(x^*)\|^2] \leq f(x) - \inf f$$

$$\mathbb{E}[\|\nabla f_i(x) - \nabla f_i(x^*)\|^2] \leq 2L_{max}(f(x) - \inf f)$$

$$2\mathbb{E}[\|\nabla f_i(x) - \nabla f_i(x^*)\|^2] \leq 4L_{max}(f(x) - \inf f) \tag{73}$$

Since f is a sum of $L-$Smooth and convex functions for $\forall x^* \in \arg\min f.$,

$$\sigma_f^* = \mathbb{V}[\nabla f_i(x^*)]$$

$$\sigma_f^* = E[\|\nabla f_i(x^*) - E[\nabla f_i(x^*)]\|^2]$$

Since f is Convex,

$$E[\nabla f_i(x^*)] = \nabla f(x^*) = 0$$

$$\sigma_f^* = E[\|\nabla f_i(x^*)\|^2]$$

$$2\sigma_f^* = 2E[\|\nabla f_i(x^*)\|^2] \tag{74}$$

Apply Linearity of Expectation to the Right Hand Side of 72

$$2\mathbb{E}[\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 + \|\nabla f_i(x^*)\|^2] = 2\mathbb{E}[\|\nabla f_i(x) - \nabla f_i(x^*)\|^2] + 2\mathbb{E}[\|\nabla f_i(x^*)\|^2]$$

$$\mathbb{E}[\|\nabla f_i(x)\|^2] \leq 2\mathbb{E}[\|\nabla f_i(x) - \nabla f_i(x^*)\|^2] + 2\mathbb{E}[\|\nabla f_i(x^*)\|^2] \tag{75}$$

Substituting 73 and 74 into the Right Hand Side of 75

$$\mathbb{E}[\|\nabla f_i(x)\|^2] \leq 4L_{max}(f(x) - \inf f) + 2\sigma_f^*$$

Proof of Lemma 27. Let $x_1, x_2 \in \arg\min f$.

$$\frac{1}{2L_b}\mathbb{E}[\|\nabla f_B(x_1) - \nabla f_B(x_2)\|^2] \leq f(x_1) - f(x_2) - \langle \nabla f(x_2), x_1 - x_2 \rangle$$

Since $f(x_2) = \inf f$ and $\nabla f(x_2) = 0$

$$\frac{1}{2L_b}\mathbb{E}[\|\nabla f_B(x_1) - \nabla f_B(x_2)\|^2] \leq f(x_1) - \inf f = 0$$

Since the square of a norm must be positive, $\mathbb{E}[\|\nabla f_B(x_1) - \nabla f_B(x_2)\|^2] = 0$, $\nabla f_B(x_1) - \nabla f_B(x_2) = 0$, and

$\nabla f_B(x_1) = \nabla f_B(x_2)$

Since $f_B(x_1) = f_B(x_2)$, $\mathbb{V}[\nabla f_B(x_1)] = \mathbb{V}[\nabla f_B(x_2)]$

Since $\mathbb{V}[\nabla f_B(x_1)] = \mathbb{V}[\nabla f_B(x_2)]$, $\sigma_f^* = \mathbb{V}[\nabla f_B(x_1)] = \mathbb{V}[\nabla f_B(x_2)] = \mathbb{V}[\nabla f_B(x^*)]$

Proof of Lemma 28:
Let $x^* \in \arg\min f$

$$\|\nabla f_B(x)\|^2 = \|\nabla f_B(x) - \nabla f_B(x^*) + f_B(x^*)\|^2 \tag{76}$$

Applying the Triangle Inequality Theorem on 76

$$\|\nabla f_B(x) - \nabla f_B(x^*) + f_B(x^*)\|^2 \leq (\|\nabla f_B(x) - \nabla f_B(x^*)\| + \|f_B(x^*)\|)^2$$

$$\|\nabla f_B(x) - \nabla f_B(x^*) + f_B(x^*)\|^2 \leq \|\nabla f_B(x) - \nabla f_B(x^*)\|^2 + 2\|\nabla f_B(x) - \nabla f_B(x^*)\|\|\nabla f_B(x^*)\| + \|\nabla f_B(x^*)\|^2 \tag{77}$$

Known Facts:

$$\|\nabla f_B(x) - \nabla f_B(x^*) - f_B(x^*)\|^2 \geq 0$$

$$\|\nabla f_B(x) - \nabla f_B(x^*)\|^2 - 2\|\nabla f_B(x) - \nabla f_B(x^*)\|\|\nabla f_B(x^*)\| + \|\nabla f_B(x^*)\|^2 \geq 0$$

$$\|\nabla f_B(x) - \nabla f_B(x^*)\|^2 + \|\nabla f_B(x^*)\|^2 \geq 2\|\nabla f_B(x) - \nabla f_B(x^*)\|\|\nabla f_B(x^*)\| \tag{78}$$

Substituting 78 into 77

$$\|\nabla f_B(x) - \nabla f_B(x^*) + f_B(x^*)\|^2 \leq 2\|\nabla f_B(x) - \nabla f_B(x^*)\|^2 + 2\|\nabla f_B(x^*)\|^2$$

$$\|\nabla f_B(x)\|^2 \leq 2\|\nabla f_B(x) - \nabla f_B(x^*)\|^2 + 2\|\nabla f_B(x^*)\|^2 \tag{79}$$

Taking Expectation over both sides of 79:

$$\mathbb{E}[\|\nabla f_B(x)\|^2] \leq \mathbb{E}[2\|\nabla f_B(x) - \nabla f_B(x^*)\|^2 + 2\|\nabla f_B(x^*)\|^2]$$

$$\mathbb{E}[\|\nabla f_B(x)\|^2] \leq 2\mathbb{E}[\|\nabla f_B(x) - \nabla f_B(x^*)\|^2 + \|\nabla f_B(x^*)\|^2] \tag{80}$$

According to Definition 25

$$\frac{1}{2L_b}\mathbb{E}[\|\nabla f_B(x) - \nabla f_B(x^*)\|^2] \leq f(x) - \inf f$$

$$\mathbb{E}[\|\nabla f_B(x) - \nabla f_B(x^*)\|^2] \leq 2L_b(f(x) - \inf f)$$

$$2\mathbb{E}[\|\nabla f_B(x) - \nabla f_B(x^*)\|^2] \leq 4L_b(f(x) - \inf f) \tag{81}$$

Since f is a Sum of $L-$Smooth and Convex Functions,

$$\sigma_b^* = \mathbb{V}[\nabla f_B(x^*)], \forall x^* \in \arg\min f.$$

$$\sigma_b^* = \mathbb{E}[\|\nabla f_B(x^*) - E[\nabla f_B(x^*)]\|^2], \forall x^* \in \arg\min f.$$

Since f is a Sum of Convex Functions,

$$\mathbb{E}[\nabla f_B(x^*)] = \nabla f(x^*) = 0$$

$$\sigma_b^* = \mathbb{E}[\|\nabla f_B(x^*)\|^2]$$

$$2\sigma_b^* = 2\mathbb{E}[\|\nabla f_B(x^*)\|^2] \tag{82}$$

Applying Linearity of Expectation to the Right Hand Side of 80:

$$2\mathbb{E}[\|\nabla f_B(x) - \nabla f_B(x^*)\|^2 + \|\nabla f_B(x^*)\|^2] = 2\mathbb{E}[\|\nabla f_B(x) - \nabla f_B(x^*)\|^2] + 2\mathbb{E}[\|\nabla f_B(x^*)\|^2]$$

$$\mathbb{E}[\|\nabla f_B(x)\|^2] \leq 2\mathbb{E}[\|\nabla f_B(x) - \nabla f_B(x^*)\|^2] + 2\mathbb{E}[\|\nabla f_B(x^*)\|^2] \tag{83}$$

Substituting 81 and 82 into the Right Hand SIde of 83

$$\mathbb{E}[\|\nabla f_B(x)\|^2] \leq 4L_b(f(x) - \inf f) + 2\sigma_b^*$$