University of Oxford

Unifying (Strong)-Convexity Generalisations for Gradient Descent Methods.

by

Oliver Newcombe

Jesus College

A dissertation submitted in partial fulfilment of the degree of Master of Science in Statistical Science.

Department of Statistics, 24–29 St Giles,
Oxford, OX1 3LB

September 2024

This is my own work (except where otherwise indicated)

Candidate: Oliver Newcombe

September 15, 2024

Date

Signature

Abstract

In statistics, many tasks are fundamentally solving an optimisation problem. Of particular note is the setting of Empirical Risk Minimisation, where a loss function of the form $f(x) = \sum_{i=1}^{n} f_i(x)$ is to be minimised. This problem is often approached with derivatives of gradient descent, wherein properties of the function are leveraged to provide theoretical guarantees on convergence of these gradient methods. Classically the paradigms of (strongly)-convex and non-convex functions have been referred to for these properties, however there is far greater fidelity in these classifications, especially in the non-convex setting where modern machine learning problems tend to lie. This work explores generalisations of base (strong)-convexity, and the important connections between these conditions, summarising previous results and providing implications not found in the literature. These results, along with an analysis of common assumptions in the stochastic gradient descent setting, are used to display how stronger assumptions lead to stronger notions of convergence across three gradient descent schemes.

Acknowledgements

Thanks go to my parents for supporting me through a challenging year and my supervisor Patrick for invaluable advice and the opportunity to research such an interesting area. Special thanks go to my wonderful friends and partner Claire, I could not have done it without you. Of course, a great thanks goes to my feline research assistant Twix for unique insights and reminding me to take breaks (constantly).

Contents

List of Figures

1 Introduction

Phrasing statistical problems in terms of a minimisation of an objective function is fundamental to approaches such as regression, clustering and classification. We focus on problems of the form:

$$f : \mathbb{R}^n \to \mathbb{R}, \ \min_{x \in \mathbb{R}^n} f(x).$$

By studying this in its full generality we can speak to methods across statistics. One setting of interest is in the finite sum setting, [40] [47] where we represent, (definition 17),

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

This can be implicitly assumed throughout to show directly how these results apply to Empirical Risk Minimisation (ERM) style problems, such as training a neural network, an area of intense interest due to their wide ranging capabilities [39].

The solution to these optimisation problems is often approached with iterative methods, usually variants of gradient descent (4). These methods can be shown to guarantee convergence to a solution, with the proofs utilising certain properties of the function being optimised, such as convexity [61] and strong-convexity [3] [29], with functions not satisfying these termed non-convex. The theory behind these convexity conditions is well studied, with concrete proofs showing not only convergence of gradient descent methods, but also the *rate* at which they converge [7]. Optimising this rate of convergence is vital as computational constraints necessitate fast algorithms for practical usage.

However, recent advancements in high dimensional statistics and machine learning have meant this broad non-convex class is of increasing interest [70] [39], prompting further analysis and distinction in the non-convex setting. Analysis of conditions generalising (strong)-convexity has been in the background for some time [56] [48], and now a wide range of generalisations exist in the literature, many with applications to non-trivial problems [73], including modern neural networks [74] [27]. These generalisations can come with their own rates of convergence [32] [12], providing guarantees for many statistical problems. Most results are proven in a vacuum, stating only the conditions needed for that particular result, with no consideration of the relative strength of the conditions. Certain conditions imply others, allowing a range of convergence results to be inherited by stronger conditions via a chain of implications. As such, work linking these conditions has seen increasing interest [32] [59] [21] [42] [53], providing hierarchies of functional classes. A similar vein of work is applied to *stochastic* gradient descent, making assumptions on the function's gradients [34] [20] [14]. This work aims to unify many of these implications, creating a more complete theory, then argues for the area's importance by demonstrating how these implications provide simple proofs for gradient descent convergence rates.

We omit a full formal literature review here in lieu of a continuous introduction of relevant material throughout.

1.1 Key contributions

Below we list our main contributions and indicate where these can be found. We do not introduce terms here, referring the reader to relevant sections.

- We perform a literature review and compilation of a large range of common gradient descent assumptions and convexity-like conditions in sections 3 and 4, along with implications, conversion of constants and example functions. Many works do not clearly report the conversion of constants, we keep track of these to allow for simple comparison of methods and theorems. These results are summarised in figure 5 outlining the broad hierarchy of strengths. We emphasise this *unification* of conditions as the primary contribution made, and discuss why this is in section 6.
- Notation is standardised for many conditions. Of note is the uniform acute angle condition (15), which appeared in multiple different forms across the literature [11] [26].
- We prove implications for certain conditions that could not be found in the literature, of particular note are Theorems 8, with the regularity condition, and 10, with the combination of the weak Polyak-Łojasiewicz (12) and quadratic growth conditions (10).
- We specify equivalences between certain conditions that are trivial, but not clearly listed in any papers found with our naming and format, such as in Theorem 11 with conditions implying variational coherence (16), and Lemma 4, showing the relationship between weak strong-convexity (7) and star strong-convexity (13).
- We show that functions satisfying the weak Polyak-Łojasiewicz condition also satisfy the important property of invexity (11) in Theorem 5. We only found proofs of quasar-convex (13) [27] [2] and Polyak-Łojasiewicz (9) [32] functions implying invexity in the literature, both stronger functions, presented in Theorem 7. Thus, we display a hierarchy with weak Polyak-Łojasiewicz the weakest convexity generalisation to imply invexity in figure 5.
- A compilation of convergence results is provided in section 5, with some simplified proofs and comparison of rates. This framework offers clear directions for future work, discussed in section 6.
- A rate not found in the literature is presented in Theorem 21. This theorem utilises the classic result of Theorem 18 and are used to argue for why this work is important. The quadratic growth with weak Polyak-Łojasiewicz combination uses our results, not found in the literature.
- Along with (strong)-convexity generalisations we present assumptions on the (stochastic) gradients for the stochastic gradient descent (5) case. We unify notation here, in particular with the ABC condition (24) and present a diagrammatic summary in figure 6.
- We also apply the same analysis to another gradient scheme, Nesterov's

accelerated gradient method (6), to briefly display other proposed ways to improve the convergence rate of gradient methods. We present a selection of results, found in Theorems 25, 26 and 27, simplifying with reference to our Lemma 8.

1.2 Outline

The paper begins with statements of key assumptions, results and definitions in section 2. This background includes definitions for smoothness, three gradient descent schemes and (strong)-convexity, and provides discussion of the interpretation and links between these.

Following this, we present our literature review on convexity generalisations, stating definitions and proving theorems. We adopt the convention of stating theorems of implications, then introducing the conditions they concern, to better highlight the differences between them, with the theorem framing the discussion. Throughout we present and prove our own results, displaying how they link into the existing literature. A similar presentation of conditions on the stochastic gradients is found in section 4. We present interpretations of the conditions, along with relevant implications from the literature.

To display the practical use of mapping out these implications, section 5 presents a selection of convergence results for the three gradient methods introduced. These are proven with reference to our own set of implications rather than those of the papers we present from, and as such we acquire slightly different results in some cases. These are summarised, with our contributions outlined, in table 1.

2 Background

Much of this section introducing the basic notation and convex conditions refers closely to [7] for definitions and intuition.

2.1 Setting

Before beginning, we state base definitions, key assumptions and notational conventions used throughout.

- $\mathcal{X}^*$ is defined as the optimal solution set, such that for $x^* \in \mathcal{X}^*$, we have, for a function f, that $f(x^*) = f^* = \min_{x \in \mathbb{R}^n} f(x)$.
- Throughout, for $x, y \in \mathbb{R}^n$, $\langle x, y \rangle = \sum_{i=1}^n x_i y_i$, $\|x\| = \sqrt{\langle x, x \rangle}$ and the Cauchy-Schwarz inequality $\langle x, y \rangle \leq |\langle x, y \rangle| \leq \|x\|\|y\|$ hold.
- The projections x_p are the closest members of the solution set to the value x in euclidean space with the standard norm, such that also $f(x_p) = f^*$.
- We do ***not*** assume a priori a unique minimum, as many other works do.
- We also assume that all functions are lower bounded. This is such that whatever the minimum x^* may be, it is not $-\infty$ and the derivative of the function is 0 when evaluated there ($\nabla f(x^*) = 0$).
- Unless otherwise specified all functions are differentiable.
- For conditions defined like ($\mathbf{SC} : \mu$), when the constant is irrelevant for the given point, we refer to them without, as e.g. ($\mathbf{SC}$).

A foundational assumption is one of smoothness on the function and its gradient. Throughout we refer to it as L-smoothness, although it is equivalently the gradient being Lipschitz-continuous, defined, along with a weaker condition, as follows [15] [20].

Definition 1 (L-Smoothness) *A function f is L-*Smooth*, denoted ($\mathbf{S} : L$) if $\forall x, y$, we have,*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

*We also use the related notion of L-*Weakly Smooth*, denoted ($\mathbf{WS} : L$), if $\forall x, y$, we have,*

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

The form of ($\mathbf{S} : L$) here is not overly useful in proofs, so we use the following implied inequality interchangeably with the definition above, along with the implication of ($\mathbf{WS} : L$):

Theorem 1 (Smoothness Implication [7]) *If a function f satisfies ($\mathbf{S} : L$), then we also have,*

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

Further, we can say that a function f satisfies $(\mathbf{S} : L)$ if and only if *we have,*

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

In addition, $(\mathbf{S} : L) \implies (\mathbf{WS} : L)$.

There is a lower bound that is obtained from this, and it can be shown that the inequality with both upper and lower conditions is exactly equivalent to the original [26]. The $(\mathbf{WS} : L)$ condition will also be used as an intermediary result.

Another key assumption is convexity. It is the core assumption made on the behaviour of functions and their gradients for optimisation problems, so central that key areas of the field are split around it into convex and *non*-convex optimisation. We give its standard definition and the one we will be using throughout below.

Definition 2 (Convexity) *A function f is* Convex*, denoted* $(\mathbf{C})$*, if it satisfies* $\forall x, y$ *and* $\theta \in (0, 1)$*,*

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

Throughout we will use the equivalent definition of,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle.$$

Geometrically this requires any line segment between two points on the function to lie on, or above without intersection, the function at every point [61]. However, this includes functions such as $f(x) = 0$. For a more restrictive, and hence *stronger* notion, we define Strong-Convexity.

Definition 3 (Strong-Convexity) *A function f is* μ Strongly-Convex*, denoted* $(\mathbf{SC} : \mu)$*, if* $\forall x, y$*,*

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

This applies a constraint on the curvature of the function, requiring at least quadratic growth [7]. As such, $f(x) = 0$ for example is not $(\mathbf{SC} : \mu)$ for any μ as it has no curvature. Linear functions similarly are not. Combining $(\mathbf{S} : L)$ and $(\mathbf{SC} : \mu)$ we obtain,

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

giving a positive quadratic lower bound, whereas the existing lower bound from $(\mathbf{S} : L)$ is negative. This is much tighter and imposes a restriction on the parameter values, requiring $\mu \leq L$ [7], illustrated in figure 2 (b). The lower bound from $(\mathbf{S} : L)$ would be represented by $-\frac{2.4}{2}x^2$ in that plot. This is important as the value $\frac{L}{\mu}$, termed the *condition number*, is prominent in convergence proofs, as when it is close to one, it implies tight bounds and as such, more information for faster convergence [23]. From now on, we implicitly assume that $\frac{L}{\mu} \geq 1$.

2.2 Gradient Descent Schemes

A collection of ubiquitous methods for performing optimisation on both convex and non-convex functions are *gradient descent* methods. There are many variations, three of which are covered here, but all hinge on local information informing the direction of a step in an iterative process [66]. We begin with the base case [9].

Definition 4 (Gradient Descent) *Define* gradient descent *for optimising a function f taking inputs $x_k \in \mathbb{R}^n$ with step-size η_k, denoted* (**GD**) *as satisfying the following recursion,*

$$x_{k+1} = x_k - \eta_k \nabla f(x_k).$$

An analogy for this (assuming no saddle points) is to imagine an agent on a hill who wishes to reach the bottom of the valley. The agent may choose to inspect the surrounding terrain, determine the direction of steepest descent, then take a step in this direction. At the bottom of the valley, all ways go uphill, so they stop. This is intuitively what is going on in (**GD**), displayed in blue in figure 1.

Results derived from this definition are used repeatedly, such as the Descent Lemma, stated below. This lemma is important in (**GD**) proofs, having the interpretation of guaranteeing progress at each iteration.

Lemma 1 (Descent Lemma [15]) *Assuming* (**S**: L)*,* (**GD**)*, with step-size η, has the following property:*

$$f(x_{k+1}) - f(x_k) \leq -\|\nabla f(x_k)\|^2 \left(\eta - \frac{\eta^2 L}{2}\right).$$

In particular, for $\eta = \frac{1}{L}$, this reduces to:

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L}\|\nabla f(x_k)\|^2.$$

Proof.

Recall the $(\mathbf{S} : L)$ implication: $f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2$. For the (**GD**) update term of $x_{k+1} = x_k - \eta \nabla f(x_k)$, the smoothness assumption reads

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2.$$

Noting now that $x_{k+1} - x_k = -\eta \nabla f(x_k)$, we obtain,

$$f(x_{k+1}) - f(x_k) \leq \langle \nabla f(x_k), -\eta \nabla f(x_k)\rangle + \frac{L}{2}\|\eta \nabla f(x_k)\|^2 = \left(\frac{L\eta^2}{2} - \eta\right)\|\nabla f(x_k)\|^2.$$

This gives the first result, and for $\eta = \frac{1}{L}$,

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L}\|\nabla f(x_k)\|^2. \ \square$$

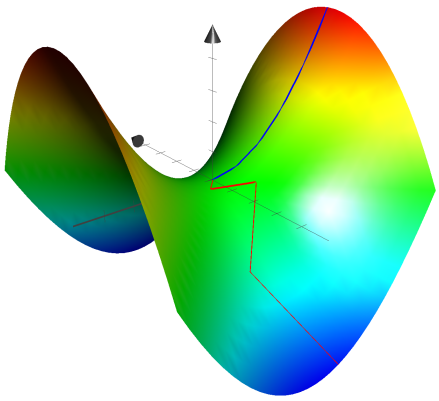


Figure 1: A plot of $f(x, y) = x^2 - y^2$ with a saddle point at $(0, 0)$. The blue path indicates a theoretical (**GD**) trajectory which gets stuck at $(0, 0)$. The red shows a theoretical (accelerated) (**SGD**) trajectory from $(0, 0)$ that escapes the stationary point.

This shows the base (**GD**) setting guarantees progress to a *local* minima or x where $\nabla f(x) = 0$. However, what happens when we get stuck outside of the *global* minima; when the agent is at the bottom of one valley with a deeper one just over the next hill? Another issue with this base setting for statistical problems satisfying (**FS**) is that at each iteration the full gradient, over all terms in the sum, needs to be calculated. This is expensive for complicated data. A solution to both is adding randomness with the following technique.

Definition 5 (Stochastic Gradient Descent [4]) *We define* stochastic gradient descent *for optimising a function f satisfying* (**FS**) *taking inputs $x_k \in \mathbb{R}^n$ with step-size η_k, denoted* (**SGD**) *as satisfying the following recursion,*

$$x_{k+1} = x_k - \eta_k \nabla f_i(x_k),$$

where f_i is selected uniformly at random from $i = 1, ..., n$ and $\mathbb{E}_i \nabla f_i(x_k) = \nabla f(x_k)$.

This definition is easily generalised to the non-finite sum setting [52]. This method, for ERM, can be interpreted as simply using one data point at a time for calculating the direction to move, rather than the whole function. This is far less computationally intensive, and also means that where the full gradient would be stuck in a local minima, by simply considering a single point, the algorithm can escape. As an example, consider figure 1. We can see that there is a saddle point at $(0, 0)$, where $\nabla f(x, y) = 0$, thus if (**GD**) reaches that point, it would stop. (**SGD**), however, would obtain an inaccurate estimate of the gradient based on the one sample, such that it would move in a slightly different direction to dislodge it from the spot with zero gradient, allowing for further convergence [16]. This is an instance of *implicit regularisation*, where the choice of optimisation algorithm enforces certain behaviours and solutions [79].

With (**SGD**), the step-size often matters more for the final convergence result. Due to the randomness, to guarantee convergence an adaptive step-size with decreasing values is often required. The most well known conditions are Robbins-Monro, such that η_k satisfies [60],

$$\sum_{k=0}^{\infty} \eta_k^2 < \infty \text{ and, } \sum_{k=0}^{\infty} \eta_k = \infty.$$

This is because, although we want the added stochasticity to escape undesirable minima, we want it to stay when it finally reaches the global minima. Otherwise, with a constant step-size, it has been observed that often the algorithm oscillates around the solution [30]. This is reflected in the types of convergence we study and in Theorem 24. However, in general, decreasing step-sizes are undesirable as they can slow convergence massively [76] as seen in Theorem 24.

One solution to this speed issue is *momentum*, where memory of previous directions is stored and a component of these directions augments the new trajectory [46]. Much like the physical interpretation, momentum allows local minima to be skipped and surfaces to be traversed faster as the iterates accelerate throughout the optimisation. This idea is ubiquitous in modern machine learning, the most prominent example being the ADAM optimiser [36]. This is somewhat illustrated in figure 1, where the red path continues down the slope at an increasing rate. We will consider a related method, (Nesterov) accelerated gradient.

Definition 6 (Nesterov Accelerated Gradient [67]) *Define the* Nesterov accelerated gradient *method for optimising a function f with sequences (x_k, δ_k, v_k) and parameters $(\eta_k, \alpha_k, \beta_k, s_k)$ with the following recursion,*

$$\delta_k = \alpha_k v_k + (1 - \alpha_k) x_k$$

$$v_{k+1} = \beta_k v_k + (1 - \beta_k)\delta_k - s_k \eta_k \nabla f_i(\delta_k)$$

$$x_{k+1} = \delta_k - \eta_k \nabla f_i(\delta_k).$$

This can be reduced from $\nabla f_i(\delta_k) \to \nabla f(\delta_k)$ in non-stochastic settings.

Many definitions of this method exist [1] [35], we chose this to match our proofs. This method has allows for higher fidelity in parameter choices and aims to *accelerate* the convergence with a momentum-like technique. We will see this can speed up standard methods.

2.3 Types of Convergence

We see convergence proofs later, so introduce the types of convergence used here, for general convergence criteria functions C, rate functions R and constants independent of k, D [15].

- Iterates - $\|x_k - x^*\|^2 \leq R(k)$.
- Function values - $f(x_k) - f^* \leq R(k)$.

- Stationary point - $\min_{t=1,\dots,k} \|\nabla f(x_t)\|^2 \leq R(k)$.
- Average function values - $f(\bar{x}_k) - f^* \leq R(k)$, where $\bar{x}_k := \frac{1}{k}\sum_{i=0}^{k} x_i$.
- Linear - $C(k) \leq c^k D$ where $0 < c < 1$.
- Sub-Linear - $C(k) \leq k^{-r} D$ where $r \in \mathbb{R}^+$.
- Neighbourhood - $C(k) \leq R(k) + \text{const}$.

Both iterates and function value convergence provide the optimal solution in the limit, but do so differently. Function value convergence is more common while iterate convergence is generally only seen in the setting with a unique minimiser. The weakest type is to a stationary point, not necessarily providing the optimal solution if it gets stuck somewhere other than the global optimum. Average function values are also seen, with usage in online learning, where this average provides theoretical guarantees where the function values are harder to calculate [5].

The gold standard for rate of convergence is linear, such that the convergence criteria approaches convergence exponentially fast. Sub-linear is slower, with rate dependent on the power of k, usually $0.5, 1$ or 2. As discussed for (**SGD**), sometimes we can only oscillate around the true value, which leads to neighbourhood convergence. This constant is generally a function of the learning rate and smoothness constant.

With these preliminary ideas, we can explore the main results, beginning with the different classes of functions we consider and how they interact.

3 Generalisations of (Strong)-Convexity

There are many different conditions to measure *how* convex or, indeed, non-convex, classes of functions are. These conditions have a detailed set of implications across the literature. However, several of these definitions are repeated with differing names or subtly different definitions. Many papers exist that create small *chains* of implications between these conditions, and the following sections aim to summarise and combine these conditions to act as a reference for further work.

These generalisations often relate to the nature of local minima in the function, imposing restrictions on variability or growth. For example, a ubiquitous result is that (strong)-convex functions' local minima *are* global minima [7], whilst other weaker function classes may have more stringent requirements to find global minima. We will see this result is not limited to this case.

We present a series of implications individually, joining them in figure 5. The first chain of implications from [32] forms the core of our convex-like tree.

Theorem 2 (Strong-Convexity Generalisations [32]) *For a function satisfying* $(\mathbf{S}: L)$*, we have the following chain of implications:*

$$(\mathbf{SC}: \mu) \implies (\mathbf{ESC}: \mu) \implies (\mathbf{WSC}: \mu) \implies (\mathbf{RSI}: \frac{\mu}{2})$$

$$\implies (\mathbf{EB}: \frac{\mu}{2}) \implies (\mathbf{PL}: \frac{\mu}{2L}) \implies (\mathbf{QG}: \frac{\mu}{2L}).$$

Further, $(\mathbf{PL}: \mu) \implies (\mathbf{EB}: \mu)$ *and if* f *is additionally* $(\mathbf{C})$ *we have,*

$$(\mathbf{QG}: \mu) + (\mathbf{C}) \implies (\mathbf{RSI}: \frac{\mu}{2}).$$

Proof.

We prove a select few illustrative results here. Overly technical or trivial proofs are left to the appendix.

$(\mathbf{WSC}: \mu) \implies (\mathbf{RSI}: \frac{\mu}{2})$:
Re-arranging $(\mathbf{WSC}: \mu)$ gives,
$\langle \nabla f(x), x - x_p \rangle \geq f(x) - f^* + \frac{\mu}{2}\|x_p - x\|^2$. However, f^* is defined as the minimum of f, so $f(x) \geq f^*$ for all x, thus, $f(x) - f^* \geq 0$. This immediately gives

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

converting the constant by a factor of $\frac{1}{2}$.

$(\mathbf{RSI}: \mu) \implies (\mathbf{EB}: \mu)$:
Consider the inner product term from $(\mathbf{RSI}: \mu)$. We can use the Cauchy-Schwartz inequality to yield $\langle \nabla f(x), x - x_p \rangle \leq \|\nabla f(x)\|\|x - x_p\|$. Combining with $(\mathbf{RSI}: \mu)$

gives $\|\nabla f(x)\|\|x - x_p\| \geq \mu\|x - x_p\|^2$. Dividing both sides by $\|x - x_p\|$ gives the result, noting that the result trivially holds if $\|x - x_p\| = 0$.

(**EB**: μ) $\implies$ (**PL**: $\frac{\mu}{L}$) :
Here we leverage (**S** : L), specifically now for x and x_p, giving $f(x) \leq f(x_p) + \langle \nabla f(x_p), x - x_p \rangle + \frac{L}{2}\|x - x_p\|^2$. This simplifies as $\nabla f(x_p) = \nabla f^* = 0$ due to being a minimum, so we have

$$f(x) \leq f^* + \frac{L}{2}\|x - x_p\|^2 \leq \frac{L}{2\mu}\|\nabla f(x)\|^2,$$

where we use (**EB** : μ), to give the result after re-arranging.

The rest of the implications are found in the appendix, with the constants in the theorem found by repeatedly applying the above results. □

This theorem also shows that we can use (**C**) in *conjunction* with other assumptions to make results stronger, and in that regime the lower four conditions are all equivalent, due to transitivity. We will see this technique again later. A similar work [21] classes these conditions as lower bounds, before proving a similar set of implications for the equivalent upper conditions, with the inequality signs reversed. We use a couple of these, e.g. (**SC** : μ) reversed is (**S** : μ) and (**PL** : μ) reversed is (**WS** : μ) and state another in Lemma 3.

Another work [42] provides similar results, but requires the assumption of weak convexity, rather than (**S**). We begin analysis with (**SC**).

(**SC**) is one of the strongest conditions we have and is the gold standard for optimisation problems, encompassing a class of well behaved functions. Thus, we are displaying generalisations of (**SC**) in the downstream implications. (**C**) does not appear immediately in this chain, although is clearly implied for (**SC** : $\mu = 0$). Another very useful quality is that for $\mu > 0$ it implies a unique solution, but there are weaker conditions implied by this, so we save the proof for later in Theorem 8. To gain a simple intuition for how the parameter μ affects the function, we plot $f(x) = \frac{\mu}{2}x^2$ in figure 2 (a). This function is (**SC** : μ). The proof follows from, for $f : \mathbb{R} \to \mathbb{R}$,

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

$$\iff y^2 \geq x^2 + 2xy - 2x^2 + y^2 + x^2 - 2xy \iff y^2 \geq y^2.$$

This is trivially true. This proof holds in exactly the same manner for (**S** : μ), such that $f(x) = \frac{L}{2}x^2$ satisfies (**S** : L).

Figure 2 (a) shows that changing μ corresponds to the rate at which the function grows; a larger value enforces faster growth. This is obvious given the function ($\nabla f(x) = \mu x$ so of course the function grows faster), but the important takeaway is framing this as a function parameter; a larger constant leads to more restrictive behaviour. This is not always the case, but here it is a way to quantify *how convex* the function is. An example with the opposite property is (**S** : L), so we summarise this with the following.

Lemma 2 *For a general function f, we have that,*

$$(\mathbf{SC} : \mu) \implies (\mathbf{SC} : \mu' \leq \mu) \text{ and, } (\mathbf{S} : L) \implies (\mathbf{S} : L' \geq L)$$

Thus, if we are looking for assumptions based upon the constant L, and want a result more likely to be applicable in real-world problems, we can use the notion that a larger L implies generality. Similarly, as we show in section 5, a larger μ often leads to faster convergence. The short proof above shows that $f(x) = x^2$ is both (**SC** : 2) and (**S** : 2). We can then move, as is allowed by Lemma 2, the (**SC**) constant slightly smaller, and (**S**) slightly larger, to show how these conditions bound a given function, in figure 2 (b).

Definition 7 (Essential and Weak Strong-Convexity) *A function f is μ* Essentially Strongly-Convex*, denoted* (**ESC** : μ)*, if $\forall x, y$ such that we have $x_p = y_p$,*

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

Then f is μ Weakly Strongly-Convex*, denoted* (**WSC** : μ)*, if $\forall x$,*

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

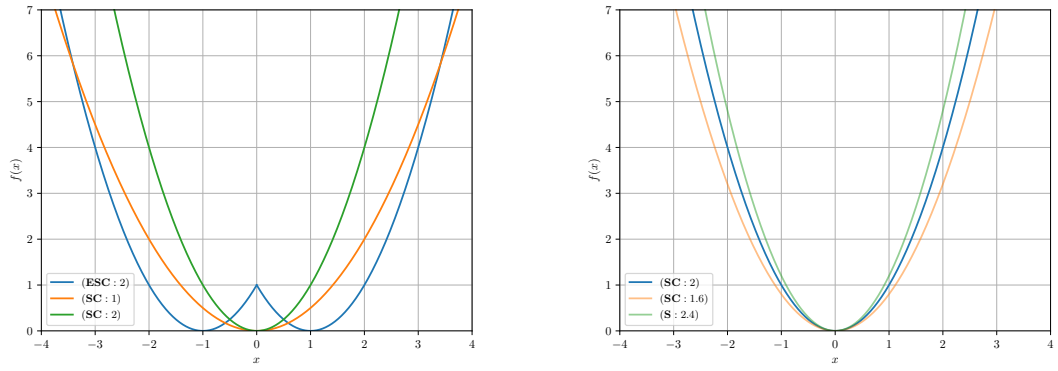
Next is (**ESC**), with the same form as (**SC**) but restricts the function inputs x and y to only those such that $x_p = y_p$, i.e. inputs with the same projection onto the solution set. Crucially, this allows solution sets with multiple feasible values, so no unique solution [45]. To demonstrate this difference, figure 2 (a), shows a piece-wise function, satisfying (**ESC** : 2). Note the similarities to our (**SC** : 2) example, that is the quadratic nature. However, here we are working with a function with two minima, at $x \pm 1$. Thus, $x_p = \pm 1$, dependent on which side of $x = 0$ the x value is; if $x < 0$, $x_p = -1$ and without loss of generality, if $x \geq 0$, $x_p = 1$.

(**WSC**), sometimes referred to as quasi strong-convexity [53], is similar except we restrict to just one input x to be paired with its projection x_p. A key function f that is not (**SC**) but satisfies the previous two is the case of a (**SC**) function g, with $f(x) = g(Ax)$ for some A without full rank. This problem is common in statistical tasks such as regression [53].

Definition 8 (Restricted Secant) *A function f satisfies the μ* Restricted Secant Inequality*, denoted* (**RSI** : μ)*, if $\forall x$,*

$$\langle \nabla f(x), x - x_p \rangle \geq \mu \|x - x_p\|^2.$$

Following these we have (**RSI**), originally appearing in [73] as restricted strong-convexity for (**C**) functions. It has also been referred to as one-point strongly-convex [72]. As for interpretation, it restricts the curvature of our function, specifically, the constant μ acts as a lower bound on the average curvature of the function f, within the range of x and its projection x_p [38]. In many cases, as the goal is to get closer to the solution set from the initialisation, weaker restrictions only applying to the



(a) (b)

Figure 2: (a) A comparison of (**SC** : 2) ($f(x) = x^2$), (**SC** : 1) ($f(x) = \frac{1}{2}x^2$) and (**ESC** : 2) ($f(x) = (x+1)^2\mathbb{I}(x < 0) + (x-1)^2\mathbb{I}(x \geq 0)$). For these and subsequent graphs with specified constants, these are verified empirically. (b) Displaying some upper (**S**) ($f(x) = \frac{2.4}{2}x^2$) and lower (**SC**) ($f(x) = \frac{1.6}{2}x^2$) bounds for the given (**SC** : 2) ($f(x) = x^2$) function.

subspace between x and the closest value to it in the solution set (the projection x_p) can be sufficient for convergence proofs. Some examples of (**RSI**) functions are in [73], however finding a specific *useful* class of functions satisfying (**RSI**) but *not* (**WSC**) is harder.

Definition 9 (Error Bound and Polyak-Łojasiewicz) *A function f satisfies the μ* Error Bound*, denoted* (**EB** : μ)*, if* $\forall x$*,*

$$\|\nabla f(x)\| \geq \mu\|x - x_p\|.$$

Then, f satisfies the μ Polyak-Łojasiewicz Inequality*, denoted* (**PL** : μ)*, if* $\forall x$*,*

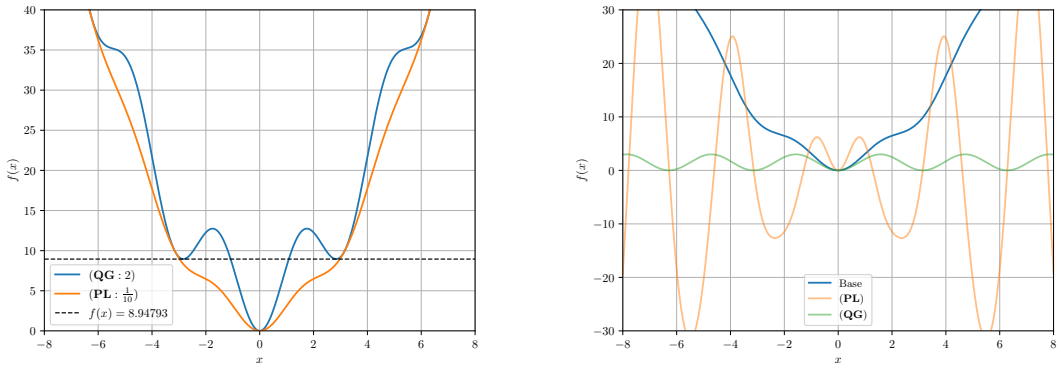
$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*).$$

Next we have (**EB**) and equivalently (under (**S**)) (**PL**). The (**EB**) condition can be traced back to [48], whilst (**PL**) is older still [56]. In practice (**PL**) is more commonly used and has had a recent resurgence due to its relationship to neural networks [13], and even simplified transformers [74]. It also lends itself to simple proofs and is a key part of this dissertation.

Definition 10 (Quadratic Growth) *A function f satisfies the μ* Quadratic Growth *condition, denoted* (**QG** : μ)*, if* $\forall x$*,*

$$f(x) - f^* \geq \frac{\mu}{2}\|x - x_p\|^2.$$

The final and weakest condition here is (**QG**). It is the weakest due to imposing no explicit constraint with the gradient of the function. This class sees less theory but recently some results have emerged combining it with other very strong assumptions



(a) (b)

Figure 3: (a) Two comparative functions for (**QG** : 2) ($f(x) = x^2 + 10\sin^2(x)$) with approximate non-global minima marked and for (**PL** : 0.1) ($f(x) = x^2 + 3\sin^2(x)$). (b) A comparison of conditions for the base function $f(x) = x^2 + 3\sin^2(x)$. The faded (**QG**) plot is $y = f(x) - f^* - \frac{\mu}{2}\|x_p - x\|^2$ and (**PL**) is $y = \frac{1}{2}\|\nabla f(x)\|^2 - \mu(f(x) - f^*)$ for $\mu = 2$.

[43]. We follow this technique in Theorem 10 as, for standard (**GD**), it is shown that it must have additional assumptions to obtain convergence [21]. As such, it is often used as a simplifying step in proofs where it is implied, such as in Theorem 19. Crucially, without those stronger conditions, this condition allows local minima that are not global minima [32]. For example, (**PL** : μ) does not allow this. Take, $f(x) = x^2 + 10\sin^2(x)$, with local non-global minima as seen in figure 3 (a), highlighting how these conditions differ. We can see this further in figure 3 (b) where the condition expressions are plotted for matching parameters ($\mu = 2$). A value below 0 at any point indicates the condition being violated, so we can see the inflection points are where (**PL**) differs with the condition plot dropping below 0. Similarly, areas seeing quadratic-like growth have condition plot values increasing over 0. This visually demonstrates the generality of (**QG**).

For some of these implications, we can obtain different constants by skipping steps. For example, by following the above chain of implications we see (**SC**: μ) $\implies$ (**PL**: $\frac{\mu}{2L}$). However, we can obtain different results directly.

Theorem 3 (Direct Implication [32]) *For a function with* (**S** : L)*, we have the following implication:*

$$(\mathbf{SC} : \mu) \implies (\mathbf{PL} : \mu).$$

This wide range of constant implications will sometimes lead to us obtaining slightly different results than the original papers. In order to ensure validity of some results in Theorem 21, we require the following implication.

Lemma 3 ((QG) upper bound [21]) *Assuming* (**WS** : L)*, f satisfies*

$$\frac{L}{2}\|x - x_p\|^2 \geq f(x) - f^*.$$

This is the equivalent of (**QG**) and is used to phrase convergence rates in terms of the condition number $\frac{L}{\mu}$. The proof is omitted as it is almost identical to (**PL**) $\implies$ (**QG**), a full version is found in ([21] appendix D). We do however prove the alternative constant conversion, also from [21].

Theorem 4 *For a function with* $(\mathbf{S}:L)$*, we have,*

$$(\mathbf{EB}:\mu) \implies (\mathbf{PL}:\frac{\mu^2}{L}).$$

Proof.

$(\mathbf{EB}:\mu)$ squared gives $\|\nabla f(x)\|^2 \geq \mu^2\|x-x_p\|^2$. Combining with the implied Lemma 3 by $(\mathbf{S}:L) \implies (\mathbf{WS}:L)$ gives,

$$\|\nabla f(x)\|^2 \geq \mu^2\|x-x_p\|^2 \geq \frac{2\mu^2}{L}(f(x)-f^*),$$

dividing by 2 for the result. □

There is another weak assumption which is equivalent to local minima implying global minima.

Definition 11 (Invex [24]) *A function* f *is* Invex*, denoted* (**IN**) *if* $\exists\ \phi:\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}^n$ *such that* $\forall x,y$,

$$f(y)\geq f(x)+\langle\nabla f(x),\phi(x,y)\rangle.$$

Note the clear generalisation of (**C**), and that due to the unrestricted form of ϕ, that this must be general. Indeed, it is very weak and most conditions we see imply it. This is because, as will be formally stated, it is equivalent to not having any local minima to get stuck in. Thus, satisfying it eases proofs as, provided it is *some* solution we are looking for, we can use $\nabla f(x)=0$ as a criterion for convergence. We now introduce the weakest condition we found to imply (**IN**), noting we later show most conditions imply this in turn.

Definition 12 (Weak Polyak-Łojasiewicz [11]) *A function* f *satisfies* μ Weak Polyak-Łojasiewicz*, denoted* $(\mathbf{WPL}:\mu)$ *if* $\forall x$,

$$\sqrt{\mu}(f(x)-f^*)\leq\|\nabla f(x)\|\|x-x^*\|.$$

The naming of this inequality is no mistake, it is indeed implied by (**PL**). It is at least strong enough for (**IN**) partially due to specification with the derivative, so even though we found relatively few works with it [11] [22], we believe it is a condition waiting for further work, as discussed in section 6. We can identify one common loss function, the Huber loss, which (when altered as in [11]) has been identified as (**WPL**), but not (**PL**) [11]. We now prove (**WPL**) $\implies$ (**IN**), a result we could not find in the literature.

Theorem 5 ((WPL) implies (IN) [51]) *A function f is* (**IN**) *if and only if every stationary point is a* global *minimum. Further, we have* $(\mathbf{WPL}:\mu) \implies (\mathbf{IN})$.

Proof.

First we prove that a function is (**IN**) if and only if all of its stationary points are global minima, with a proof adapted from [51].

$(\implies)$:

Assume f is (**IN**). That is, there exists ϕ such that $f(y) \geq f(x) + \langle \nabla f(x), \phi(x,y) \rangle$. Consider $x = x^*$, a stationary point such that $\nabla f(x^*) = 0$. This immediately gives, $\forall x,\ f(y) \geq f(x^*)\ \forall y$ thus, x^* is a global minimiser.

$(\impliedby)$:

If we assume that all stationary points are global minimisers, we need only find some ϕ to satisfy (**IN**). If $\nabla f(y) = 0$, that is, y is a stationary point, our assumption implies it is a global minimum. Thus, to verify the (**IN**), we can zero out the inner product (as $\nabla f(y) = 0$) and so only need $f(x) \geq f(y)$ for all $x \in \mathbb{R}$. As y is the global minimum, this is true.

In the case $\nabla f(y) \neq 0$, we claim $\phi(x,y) = \frac{(f(x)-f(y))\nabla f(y)}{\|\nabla f(y)\|^2}$ is a suitable vector valued function. Indeed, this gives:

$$\langle \nabla f(y), \phi(x,y) \rangle = \langle \nabla f(y), \frac{(f(x)-f(y))\nabla f(y)}{\|\nabla f(y)\|^2} \rangle = \frac{f(x)-f(y)}{\|\nabla f(y)\|^2} \langle \nabla f(y), \nabla f(y) \rangle.$$

The definition of an inner product gives $\langle \nabla f(y), \nabla f(y) \rangle = \|\nabla f(y)\|^2$, such that $\langle \nabla f(y), \phi(x,y) \rangle = f(x) - f(y)$. (**IN**) then requires

$$f(x) \geq f(y) + \langle \nabla f(y), \phi(x,y) \rangle \iff f(x) \geq f(y) + (f(x) - f(y)) \iff f(x) \geq f(x),$$

trivially true.

For $(\mathbf{WPL}:\mu) \implies (\mathbf{IN})$, we prove this in a similar way to previous proofs [32], with a squeeze argument. First, recall the definition of $(\mathbf{WPL}:\mu)$, $\sqrt{\mu}(f(x) - f^*) \leq \|\nabla f(x)\| \|x - x^*\|$, for x^* a global minimiser and f^* the minimum. For an arbitrary stationary point $\hat{x}$, $f^* \leq f(\hat{x})$. Next, we substitute $\hat{x}$ into the expression for $(\mathbf{WPL}:\mu)$, noting stationarity gives $\nabla f(\hat{x}) = 0$, to obtain:

$$\sqrt{\mu}(f(\hat{x}) - f^*) \leq \|\nabla f(\hat{x})\| \|\hat{x} - x^*\| \iff \sqrt{\mu}(f(\hat{x}) - f^*) \leq 0 \implies f^* \geq f(\hat{x}).$$

Thus, by a squeeze argument, $f^* = f(\hat{x})$, so is a global minimiser, and as such, as $\hat{x}$ was an arbitrary stationary point, f is (**IN**). $\square$

In generalising (**SC**), this chain focuses on the quadratic difference term, $\|y - x\|^2$, to control convexity. There are other ways to generalise the base definition, one of which that has seen use recently stems from quasar-convexity, another class of non-convexity.

Definition 13 (Quasar-Convex [20]) *A function f is* μ Strongly Star-Convex *for an optimal value x^*, denoted (***SC**, μ), if $\forall x$,*

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

It is simply Star-Convex*, denoted (***C**), if $\mu = 0$. Further, a function is* Strongly Quasar-Convex*, denoted (**SQC** : γ, μ), if $\forall x$ and $\gamma \in (0, 1]$,*

$$f^* \geq f(x) + \frac{1}{\gamma}\langle \nabla f(x), x^* - x \rangle + \frac{\mu}{2}\|x^* - x\|^2.$$

It is simply Quasar-Convex*, denoted (**QC** : γ), if $\mu = 0$.*

These conditions clearly have similarities, so before discussing them we state the simple chain of implications:

Theorem 6 (Quasar-Convex Implications [27]) *For a function satisfying (**S** : L), the following implications hold for any $\gamma \in (0, 1]$ and μ,*

$$(\mathbf{SC} : \mu) \implies (\mathbf{*SC} : \mu),\ (\mathbf{C}) \implies (\mathbf{*C}) \implies (\mathbf{QC} : \gamma),$$

and further,

$$(\mathbf{*SC} : \mu) \implies (\mathbf{SQC} : \gamma, \mu) \implies (\mathbf{QC} : \gamma).$$

The proof in the appendix uses simple implications, often hinging on proving $\langle \nabla f(x), x - x^* \rangle \geq 0$, an important property to be discussed further with (16).

These conditions generalise (**SC**) by a multiplicative constant γ, and move away from using projections onto the optimal solution set (x_p), instead specifying conditions for one globally optimal point (x^*). It is clear to see that if there is only one optimal solution, that is $\forall x$ we have $x_p = x^*$, that many definitions simplify to others already seen, shown later in Theorem 4

(***SC**), has a very similar definition to (**WSC**), only holding for one *fixed* point. Although this is a small difference in terms of notation, the effects are noticeable, as in general $\|x_p - x\| \neq \|x^* - x\|$ and (***SC**) implies a unique minimiser (Theorem 8) There is potential to exploit the fact that $\underset{y \in \mathcal{X}^*}{\operatorname{argmin}} \|y - x\| = x_p$, such that $\|x_p - x\| \leq \|x^* - x\|$ for any other solution x^*, due to the projection being the shortest possible distance to the solution set. We could not find any proofs leveraging this, other than a similar notion in the $(\mathbf{PL} : \mu) \implies (\mathbf{QG} : \mu)$ proof in Theorem 2.

Following this we have the relaxed notion of (***C**) with no quadratic term. This class of functions has seen interest recently for multiple reasons. Firstly, it has been used to augment proofs as an additional assumption to provide more information about the area local to the solution [26]. In many problems, it is often only a small area of the function that we care about specifying behaviour for, allowing for

weaker assumptions to be made globally, whilst still getting sufficient information for convergence proofs. Secondly, a body of work has explored how certain neural networks have close links in terms of behaviour to (***C**). In particular, it can be observed [77] or enforced [75] that the optimisation trajectory follows a (***C**) like structure, with the inequality holding at each iteration of training cumulatively. Further evidence has shown that the neural networks display loss landscapes where the neighbourhood near the optimal solution satisfies star convexity [37].

We note here that this neighbourhood point is general. Many interesting problems will have local minima that are not global. This does not make the exploration of these function classes any less interesting. Aside from the many problems satisfying these conditions in full generality, by specifying conditions for just a small area of the problem, we can still obtain valuable results. Certain conditions, in particular the upcoming regularity condition, are often defined on just a small part of $\mathbb{R}^n$ [10]. In addition, as discussed in section 2, (**SGD**) can escape these minima [37].

(**SQC**) is a similar case, linked to certain neural networks. Of particular note, recurrent neural networks can be seen as a dynamical system wherein, under some assumptions, the loss function is (**QC**) [25]. (**QC**) can be generalised to the regime where gamma can take any value in the positive reals, generally termed quasi-convex [72]. However interest is geared towards more non-convex functions to fit more complicated problems, in particular, highly non-convex neural networks. The smaller the γ value, the more highly non-convex the function is in a sense [20], demonstrated in figure 4 (a). This can be practically seen in how γ values affect convergence rates in section 5. Many of these functions look simple in the one-dimensional case, but these plots are purely for explanatory purposes. These classes of functions include some highly non-convex examples in higher dimensions [27].

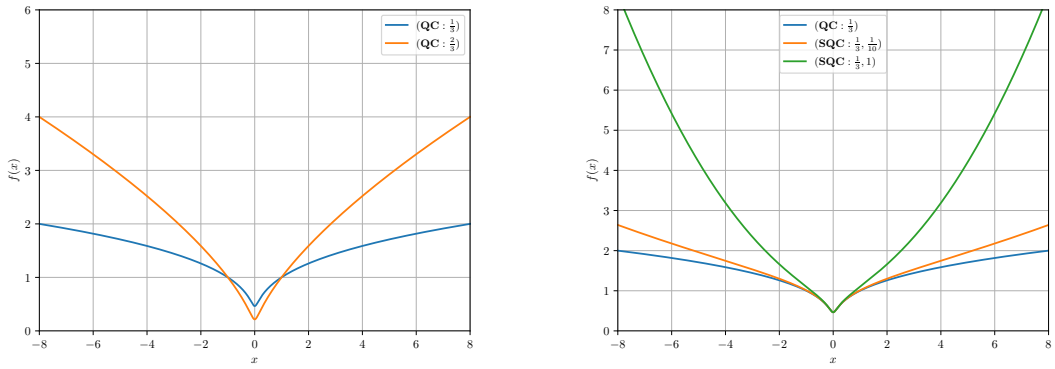
To quantify how much weaker (**QC**) and (**SQC**) are than the standard notions, works show the key difference is local information, due to only specifying the areas near the minimiser x^* [26]. This is similar to previous notions like (**WSC**), but γ allows this local information to be increasingly uncertain. We can explore these ideas with plots of a toy example.

We use the function $f(x) = \left(x^2 + \frac{1}{a}\right)^{\frac{1}{b}} + cx^2$, a generalisation of the form from [27], to show how changing γ and μ affects the sort of functions we obtain. Note, their example is in fact false, claiming for $a = 8$, $b = 6$, $c = 0$ f is $(\mathbf{QC} : \gamma = \frac{1}{2})$, it is however, for example, $(\mathbf{QC} : \gamma = \frac{1}{3})$; they have claimed their function is *more convex* than it actually is. We turn to figure 4 and leverage intuition from the previous example in figure 2 (a), to see how μ affects the function in figure 4 (b), along with how γ affects it from figure 4 (a). The effects of these constants is seen in the figure but formalised in Theorem 10.

Now that we have introduced (**QC**) we state the following:

Theorem 7 ((WPL) implications [11] [22]) *For a function satisfying* $(\mathbf{S} : L)$*, the following implications hold,*

$$(\mathbf{PL} : \mu) \implies (\mathbf{WPL} : \frac{4\mu}{L}), \text{ and } (\mathbf{QC} : \gamma) \implies (\mathbf{WPL} : \gamma^2).$$



(a) (b)

Figure 4: (a) A comparison of two quasar-convex functions with constants (**QC** : $\frac{1}{3}$) ($f(x) = \left(x^2 + \frac{1}{100}\right)^{\frac{1}{6}}$) and (**QC** : $\frac{2}{3}$) ($f(x) = \left(x^2 + \frac{1}{100}\right)^{\frac{1}{3}}$). (b) Comparison of three strongly-quasar convex functions (**SQC** : $\frac{1}{3}, 0$) = (**QC** : $\frac{1}{3}$) as in (a), (**SQC** : $\frac{1}{3}, \frac{1}{10}$) ($f(x) = \left(x^2 + \frac{1}{100}\right)^{\frac{1}{6}} + 0.01x^2$) and (**SQC** : $\frac{1}{3}, 1$) ($f(x) = \left(x^2 + \frac{1}{100}\right)^{\frac{1}{6}} + 0.1x^2$). Note these are not the limiting cases, interpretability of constants was preferred.

Theorem 2 shows there is an increased denominator dependence on L with weaker conditions, which is continued here. We do have some links to the existing tree of implications via (**IN**) and (**WPL**), however there are more connections to be drawn. We present more that are either not explicitly proven; could not be found in the literature; or improve upon literature results, with another condition, termed the regularity condition.

Definition 14 (Regularity Condition [10]) *A function f satisfies the* Regularity Condition *for an optimal value x^*, denoted* (**RC** : α, β)*, if* $\forall x$,

$$\langle \nabla f(x), x - x^* \rangle \geq \alpha \|\nabla f(x)\|^2 + \beta \|x - x^*\|^2.$$

The regularity condition was first identified in work related to matrix factorisation [31] and phase retrieval [8] [28]. It can be interpreted as negatively correlating the direction of (**GD**) iterations ($-\nabla f(x)$) with the iterate error ($x - x^*$). That is, it enforces progression to the optimum by reducing the iterate error [10]. It also implies the useful property of a unique minimiser (seemingly the weakest condition to imply this), and has been shown that it implies (**PL**) [71] and is implied by (***SC**) [10], however we improve these results in Theorem 8. Note, we define (**UAAC**) later in definition 15 but place the proof here for intuition.

Theorem 8 (Regularity Condition Implications) *Assuming* (**S** : L)*, a function satisfying* (**RC** : α, β) *has a unique minimiser. Thus, the following implications hold:*

$$(\mathbf{RC} : \alpha, \beta) \implies (\mathbf{RSI} : \beta) \text{ and } (\mathbf{RC} : \alpha, \beta) \implies (\mathbf{UAAC} : 2\sqrt{\alpha\beta}).$$

Further, we have that,

$$(\mathbf{SQC} : \gamma, \mu) \implies (\mathbf{RC} : \frac{\gamma}{2L}, \frac{\gamma\mu}{2}),$$

and conversely, if we assume $2\alpha\beta \leq L$*, we have,*

$$(\mathbf{RC} : \alpha, \beta) \implies (\mathbf{SQC} : \frac{2\alpha\beta}{L}, \frac{L}{2\alpha}).$$

Proof.

First we show $(\mathbf{RC} : \alpha, \beta)$ implies a unique solution with the standard technique of assuming another minimiser $\hat{x}$ such that $\nabla f(\hat{x}) = 0$. Substituting this in gives:

$$\langle \nabla f(\hat{x}), \hat{x} - x^* \rangle \geq \alpha \|\nabla f(\hat{x})\|^2 + \beta \|\hat{x} - x^*\|^2 \iff 0 \geq \|\hat{x} - x^*\|^2,$$

thus, x^* is unique. For $(\mathbf{RC} : \alpha, \beta) \implies (\mathbf{RSI} : \beta)$, as x^* is unique, $x_p = x^*$, making the implication of $(\mathbf{RSI} : \beta)$ trivial:

$$\langle \nabla f(x), x - x^* \rangle \geq \alpha \|\nabla f(x)\|^2 + \beta \|x - x^*\|^2 \geq \beta \|x - x^*\|^2.$$

For $(\mathbf{SQC} : \gamma, \mu) \implies (\mathbf{RC} : \frac{\gamma}{2L}, \frac{\mu\gamma}{2})$, we alter the proof for (**SC) from [10]. Using the Descent Lemma 1 proof, but for arbitrary $x = x_k$ and $y = x_{k+1} = x - \frac{1}{L}\nabla f(x)$. This gives

$$f^* - f(x) \leq f(y) - f(x) \leq -\frac{1}{2L}\|\nabla f(x)\|^2 \implies f(x) - f^* \geq \frac{1}{2L}\|\nabla f(x)\|^2.$$

Rearranging $(\mathbf{SQC} : \gamma, \mu)$ and using the above result gives,

$$\frac{1}{\gamma}\langle \nabla f(x), x - x^* \rangle \geq (f(x) - f^*) + \frac{\mu}{2}\|x - x^*\|^2 \geq \frac{1}{2L}\|\nabla f(x)\|^2 + \frac{\mu}{2}\|x - x^*\|^2.$$

Multiplying through by $\gamma \in (0, 1]$ finishes the proof. For $(\mathbf{RC} : \alpha, \beta) \implies (\mathbf{SQC} : \frac{2\alpha\beta}{L}, \frac{L}{2\alpha})$, we begin by noting from Theorem 2 that $(\mathbf{RC} : \alpha, \beta) \implies (\mathbf{RSI} : \beta) \implies (\mathbf{PL} : \frac{\beta}{L})$, such that $\|\nabla f(x)\|^2 \geq \frac{2\beta}{L}(f(x) - f^*)$. Using RC then leads to,

$$\langle \nabla f(x), x - x^* \rangle \geq \alpha \|\nabla f(x)\|^2 + \beta \|x - x^*\|^2 \geq \frac{2\alpha\beta}{L}(f(x) - f^*) + \beta \|x - x^*\|^2.$$

Multiplying through by $\frac{L}{2\alpha\beta}$ gives

$$\frac{L}{2\alpha\beta}\langle \nabla f(x), x - x^* \rangle \geq f(x) - f^* + \frac{L}{2\alpha}\|x - x^*\|^2.$$

Thus as long as $2\alpha\beta \leq L$ holds, we have $\gamma := \frac{2\alpha\beta}{L} \in (0, 1]$ as required. Rearranging then gives the result.

For $(\mathbf{RC} : \alpha, \beta) \implies (\mathbf{UAAC} : 2\sqrt{\alpha\beta})$, we have,

$$\begin{aligned}
\langle \nabla f(x), x - x^* \rangle &\geq \alpha \|\nabla f(x)\|^2 + \beta \|x - x^*\|^2 \\
&= \left(\sqrt{\alpha}\|\nabla f(x)\| - \sqrt{\beta}\|x - x^*\|\right)^2 + 2\sqrt{\alpha\beta}\|\nabla f(x)\|\|x - x^*\| \\
&\geq 2\sqrt{\alpha\beta}\|\nabla f(x)\|\|x - x^*\|.
\end{aligned}$$

Thus, rearranging gives,

$$\frac{\langle \nabla f(x), x - x^* \rangle}{\|\nabla f(x)\| \|x - x^*\|} \geq 2\sqrt{\alpha\beta}.$$

Technically we require $2\sqrt{\alpha\beta} \leq 1$ as per the definition, however the same proof in reverse starting from Cauchy-Schwarz implying $\langle \nabla f(x), x - x^* \rangle \leq \|\nabla f(x)\| \|x - x^*\|$ shows this inequality must hold for $(\mathbf{RC} : \alpha, \beta)$ functions [10]. □

It is stated, but not proven, in [27] that $(\mathbf{SQC} : \gamma, \mu) \implies (\mathbf{RSI} : \frac{\gamma\mu}{2})$ directly, but we obtain the same constant going via (**RC**) so do not prove this. This result shows any upstream condition implying (**RC**) implies a unique minimiser. We use this uniqueness to prove another link to the previous tree of implications.

Lemma 4 *We have that in general* $(\mathbf{*SC} : \mu) \implies (\mathbf{WSC} : \mu)$ *but if there is a unique solution* x^* *we have,*

$$(\mathbf{*SC} : \mu) \iff (\mathbf{WSC} : \mu).$$

Proof.

For the first implication, $(\mathbf{*SC} : \mu) \implies (\mathbf{SQC} : \gamma, \mu) \implies (\mathbf{RC} : \frac{\gamma}{2L}, \frac{\gamma\mu}{2})$ so there is a unique minimiser by Theorem 8, such that $x_p = x^*$ and $(\mathbf{WSC} : \mu)$ follows. Similarly for the equivalence statement, it is clearly true as if there is only one minimiser from the start, $\forall x$ we have $x_p = x^*$. □

The condition introduced in Theorem 8 was that of the acute angle condition, defined as follows.

Definition 15 (Uniform Acute Angle Condition [26]) *A function* f *satisfies the* Uniform Acute Angle Condition *for an optimal value* x^**, denoted* $(\mathbf{UAAC} : a)$*, if* $\forall x$

$$1 \geq \frac{\langle \nabla f(x), x - x^* \rangle}{\|\nabla f(x)\| \|x - x^*\|} \geq a > 0.$$

The condition of $1 \geq$ is a consequence of Cauchy-Schwarz. The condition itself, similarly to (**RC**), imposes constraints on the direction our gradient methods can act in. If it were not to hold, it would mean our gradient steps could move in directions orthogonal to the minimum, undoing progress, the strength of the effect controlled by a [11]. However, it does not imply unique minimisers. We will see that (**UAAC**) is often used in conjunction with others to prove upstream implications, by giving some function classes the missing link to a stronger condition. It is hard to prove anything on its own due to it not explicitly imposing constraints on the actual function values. Here are some key examples.

Theorem 9 ((UAAC) Implications [26], [11]) *For a function satisfying* $(\mathbf{S} : L)$*, the following implications hold,*

$$(\mathbf{UAAC} : a) + (\mathbf{WPL} : \mu) \implies (\mathbf{QC} : \frac{\sqrt{\mu}}{a}),$$

which becomes (***C**) *if* $a \leq \frac{1}{\sqrt{\mu}}$. *Further, if there is a unique minimiser for* f,

$$(\mathbf{UAAC} : a) + (\mathbf{EB} : \mu) \implies (\mathbf{RSI} : a\mu).$$

This interpretation of the missing link can also be applied to (**QG**), as shown below.

Theorem 10 ((QG) Implications [69]) *For a function satisfying* $(\mathbf{S} : L)$ *with a unique minimiser* x^*, *the following implication holds for any* $\theta \in (0, 1)$,

$$(\mathbf{QC} : \gamma) + (\mathbf{QG} : \mu) \implies (\mathbf{SQC} : \theta\gamma, \frac{1-\theta}{\theta}\mu),$$

and similarly,

$$(\mathbf{WPL} : \mu_1) + (\mathbf{QG} : \mu_2) \implies (\mathbf{PL} : \frac{\mu_1\mu_2}{4}).$$

Proof. We begin by decomposing the (**QC**) definition, following [69],

$$\langle \nabla f(x), x - x^* \rangle \geq \gamma(f(x) - f^*) = \gamma(\theta(f(x) - f^*) + (1-\theta)(f(x) - f^*)),$$

by splitting the function difference term. Then, for the second term we leverage $(\mathbf{QG} : \mu)$. As there is a unique solution, $x_p = x^*$, this gives,

$$\langle \nabla f(x), x - x^* \rangle \geq \gamma\theta(f(x) - f^*) + (1-\theta)\frac{\mu\gamma}{2}\|x - x^*\|^2.$$

Dividing through by $\gamma\theta$ gives the result.

For the next, we begin by noting that the proof is trivial x is a minimiser. Otherwise, squaring $(\mathbf{WPL} : \mu_1)$ gives $\mu_1(f(x) - f^*)^2 \leq \|\nabla f(x)\|^2\|x - x^*\|^2$. Following this, we can apply $(\mathbf{QG} : \mu_2)$ and rearrange as $(f(x) - f^*) \neq 0$ for,

$$\mu_1(f(x) - f^*)^2 \leq \|\nabla f(x)\|^2\|x - x^*\|^2 \leq \|\nabla f(x)\|^2\frac{2}{\mu_2}(f(x) - f^*)$$
$$\implies \frac{\mu_1\mu_2}{2}(f(x) - f^*) \leq \|\nabla f(x)\|^2 \implies \frac{\mu_1\mu_2}{4}(f(x) - f^*) \leq \frac{1}{2}\|\nabla f(x)\|^2. \square$$

The first is a particularly instructive result in showing that there is a trade-off between the two non-convexity proxies we have. A smaller θ implies a smaller γ term - less convex in the (**QC**) sense, but $\frac{1-\theta}{\theta} = \frac{1}{\theta} - 1$ gets larger - more convex in the (**QG**) and (**SC**) sense. To simplify massively, it seems γ makes the non-convex properties of functions more/less non-convex, whilst μ makes the convex properties more/less convex, with some bounding behaviour between them. This trade-off can be seen somewhat in figure 4. The second is our own and to the best of our knowledge is not in the literature.

Next, we move on to a new function class, similar to (**UAAC**).

Definition 16 (Variational Coherence [78]) *A function* f *satisfies* Variational Coherence, *denoted* (**VC**), *if* $\forall x, x^*$

$$\langle \nabla f(x), x - x^* \rangle \geq 0,$$

with equality if x *is an optimal solution.*

This would clearly be implied by (**UAAC**) if it held for all optimal values, or if there is a unique minimiser. As is, this condition is a little different, using the stronger notion that *all* optimal solutions must satisfy this. It has notably appeared in works concerning another descent algorithm, mirror descent [78] and as such is a valuable condition to satisfy. Many proofs, especially with (**QC**) related conditions, prove this manually each time for the referenced point, thus, stating it as a condition makes sense, and is worth, even if obvious, explicitly analysing implications.

Theorem 11 (Variational Coherence Implications [27]) *For a function with* $(\mathbf{S} : L)$*, the following implications hold,*

$$(\mathbf{SQC} : \gamma, \mu), (\mathbf{RC}, \alpha, \beta) \implies (\mathbf{VC}),$$

and if it is the case that there is a unique minimiser x^**, we have,*

$$(\mathbf{QC} : \gamma), (\mathbf{UAAC} : a), (\mathbf{RSI} : \mu) \implies (\mathbf{VC}).$$

As trivial as the proofs found in the appendix are, the $(\mathbf{RSI} : \mu)$. $(\mathbf{RC}, \alpha, \beta)$ and $(\mathbf{UAAC} : a)$ implications were not found explicitly stated in the literature. This shows that most of our conditions imply (**VC**), but notably not the commonly used (**PL**). This could be the reason it has been found to hold in more general problems [74].

This sections' implications are summarised in 5. The red lines indicate implications not found by our literature review, even if trivial, such as some of these (**VC**) connections and those between (**WSC**) and (***SC**).

This concludes the review of generalisations, used throughout the remainder of the paper. For standard (**GD**), these are the only conditions we need. However, when moving to (**SGD**), due to the stochasticity involved, we often need extra assumptions to guarantee convergence at a given rate. This is not to say (**SGD**) is *worse* than (**GD**) in the sense that it requires stronger assumptions. Rather, as we mentioned in section 1, it provides other benefits.

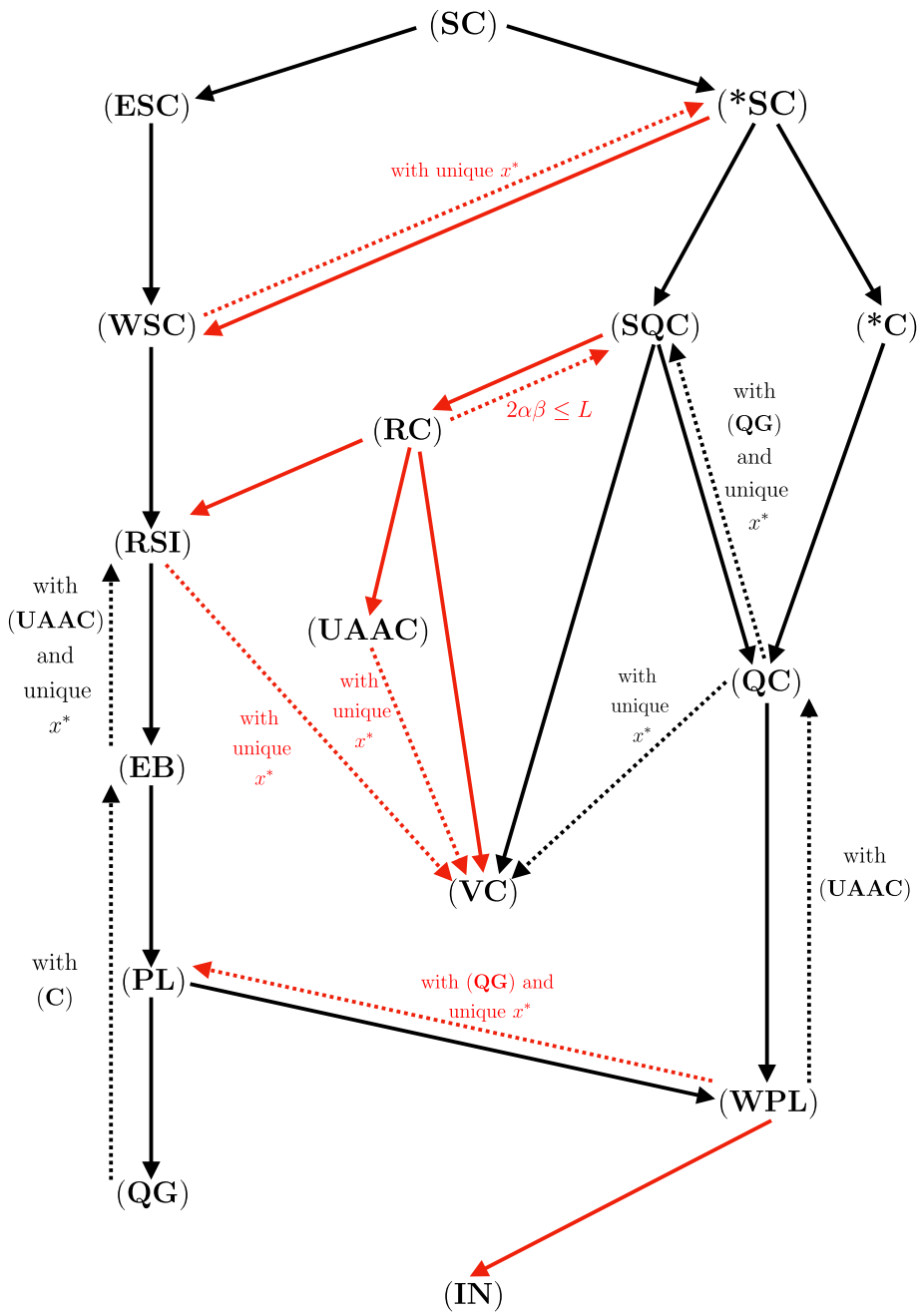


Figure 5: Diagram of implications with (**C**) omitted to avoid clutter. Dotted lines indicate implications with additional assumptions, placed to the side. Red lines indicate implications that, to the best of our knowledge, are not explicitly found in the literature. Refer to theorems for constant conversions.

4 Conditions on Stochastic Gradients

For the gradient conditions, whether or not these are satisfied depends heavily on the dataset used. As such, we provide fewer practical examples, as they would be less illustrative. Many of these conditions bound the stochastic gradients of our function in expectation, with the strength of this bound dependent on how flexible the term bounding it is. More rigid terms, like constants, enforce stronger conditions on the function. Further, many stochastic conditions are not necessarily met in practice, or, indeed, cannot be met. We begin with key definitions, followed by trees of implications, as in section 3.

Definition 17 (Finite Sum [20]) *A function f has a Finite Sum Structure, denoted* (**FS**)*, if we can write it as,*

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

This allows us to apply results to the typical ERM setting, interpreting f as a loss function to minimise over. Further, (**SGD**) selects indices i uniformly at random each iteration, allowing us to take expectations with respect to i with $\mathbb{E}_i\left[f_i(x)\right] = f(x)$. By phrasing this as a sum of functions, it is natural to extend certain definitions for, and make assumptions on, the individual f_i.

Definition 18 (Max Smoothness [15]) *For a function f with* (**FS**)*, suppose each f_i satisfies* $(\mathbf{S} : L_i)$*. Then we define the following notions of the average and max smoothness constants,*

$$L_{max} = \max_{i=1,\dots,n} L_i, \quad L_{avg} = \frac{1}{n}\sum_{i=1}^{n} L_i(x).$$

Further, we say this f satisfies the Sum of Convex *property, denoted if each f_i is convex.*

Assuming each f_i satisfies (**C**) and $(\mathbf{S} : L_i)$, individually is common [18]. For completeness in linking these to our standard $(\mathbf{S} : L)$, we have the following result.

Lemma 5 (Average Smoothness [15]) *For a function f that satisfies* (**FS**)*, if each f_i satisfies* $(\mathbf{S} : L_i)$*, we have that f satisfies* $(\mathbf{S} : L_{avg})$*.*

This can convert between proofs in different settings, but not necessarily for results in and of themselves. Bounding using $L_i \leq L_{max}$ is more common [15]. With these notions established, we now introduce a property that has had a surge in interest [44] [49].

Definition 19 (Interpolation [49]) *For a function f satisfying* (**FS**)*, we say it satisfies* Interpolation*, denoted* (**IP**)*, if, for a minimiser of f, x^*, we have that x^* also minimises each f_i. Formally,*

$$f_i(x^*) = \min_x f_i(x).$$

Definitions of this property differ, but we state the most general noting all are equivalent under (**IN**) [50]. Most works assume (**IN**) implicitly [67] [63], thus using, what we consider here to be an *implied* property from (**IP**) of $\nabla f(x^*) = 0 \implies \nabla f_i(x^*) = 0,\ \forall i$, as the definition for (**IP**). Indeed, under (**IN**) the two definitions are equivalent. (**IP**) has seen interest due to relevance for neural networks. These vastly overparameterised models often have the ability to perfectly interpolate their training data, that is, zero training error in ERM [67]. (**IP**) can be satisfied by models simply overfitting the data, but these larger models also show the capability to *generalise* well too, along with achieving perfect accuracy on the training set, which is fundamentally this (**IP**) property [49]. For convenience, if (**IP**) is assumed, we use the convention that (**FS**) is also implicitly assumed. The final preliminary definition is that of gradient noise.

Definition 20 (Gradient Noise [15]) *For a function f satisfying* (**FS**)*, define the* Gradient Noise *as,*

$$\sigma^2 = \max_{x^* \in \mathcal{X}^*} \mathbb{E}_i \left[\|\nabla f_i(x^*)\|^2 \right].$$

We will assume that $\sigma^2 < \infty$ throughout.

This is a weak assumption [19], primarily due to only needing to be satisfied by points in the solution set, so acts more as a constraint on the variability in the dataset as opposed to the functions we study. For this to hold for general x, in contrast, is exceptionally strong, shown in Lemma 6.

With these defined, we introduce the main tree of assumptions. We begin with the most commonly used classic assumption, then analyse it to motivate why we need other, more general results.

Definition 21 (Bounded Gradient [32]) *A function f with* (**FS**) *satisfies the* Bounded Gradients *condition, denoted* $(\mathbf{BG} : C)$ *if, $\forall x$, we have,*

$$\mathbb{E}_i \left[\|\nabla f_i(x)\|^2 \right] \leq C < \infty.$$

At a glance this seems reasonable, and in many proofs it is used to cancel out troublesome stochastic gradient terms, providing a bound dependent on the constant C [32]. This bounding by a constant is one of the three main ways we can control the stochastic gradients and is stronger than it seems, due to the randomness of the gradients relative to the rigidity of the constant bound. To illustrate this, we have the following result.

Lemma 6 *A function f cannot satisfy both* $(\mathbf{SC} : \mu)$ *and* $(\mathbf{BG} : C)$.

Proof.

This proof, replicated from [55], hinges on assuming both conditions, then proving an upper and lower bound on our function values, showing the feasible region of values is empty. For the upper bound, we utilise Theorem 3 for $(\mathbf{SC} : \mu) \implies (\mathbf{PL} : \mu)$, so we have $f(x) - f^* \leq \frac{1}{2\mu} \|\nabla f(x)\|^2$. We can then use this and $(\mathbf{BG} : C)$ then to obtain,

$$2\mu(f(x) - f^*) \leq \|\nabla f(x)\|^2 = \|\mathbb{E}_i \nabla f_i(x)\|^2 \leq \mathbb{E}_i \|\nabla f_i(x)\|^2 \leq C,$$

where the penultimate inequality is simply Jensen's inequality. Rearranging gives $f(x) \leq f^* + \frac{C}{2\mu}$. On the other hand, applying $(\mathbf{SC}:\mu)$ with $y = x^*$ and $\nabla f^* = 0$, $f(x) \geq f^* + \langle \nabla f^*, x - x^* \rangle + \frac{\mu}{2}\|x - x^*\|^2 \implies f(x) \geq f^* + \frac{\mu}{2}\|x - x^*\|^2$. Thus, if $\frac{C}{2\mu} \leq \frac{\mu}{2}\|x - x^*\|^2$, for any x, we obtain a contradiction. $\square$

This is the first anti-implication we have seen, indicating quite how strong (**BG**) is. Thus, despite a large body of work utilising this assumption [32] [64] [54] [58], we must investigate more general conditions. We begin with the following theorem, using more general notation with ρ here and introducing a unifying norm for symbols soon.

Theorem 12 (Stochastic Gradient Implications [20]) *For a function f satisfying* (**S: L**)*, we have the following chain of implications:*

$$(\mathbf{MSGC}:\rho) \implies (\mathbf{SGC}:\rho) \implies (\mathbf{WGC}:2L\rho) \implies (\mathbf{ES}:2L\rho) \implies (\mathbf{ER}:2L\rho),$$

and for the gradient noise σ^2,

$$(\mathbf{ER}:2L\rho) \implies (\mathbf{ABC}:4L\rho, 1, 2\sigma^2).$$

Proof.

$(\mathbf{MSGC}:\rho) \implies (\mathbf{SGC}:\rho)$, holds trivially by taking the expectation over $(\mathbf{MSGC}:\rho)$.

For $(\mathbf{SGC}:\rho) \implies (\mathbf{WGC}:2L\rho)$, we utilise $(\mathbf{S}:L) \implies (\mathbf{WS}:L)$, proven in the previous section. That, combined with $(\mathbf{SGC}:\rho)$ gives,

$$\mathbb{E}_i\left[\|\nabla f_i(x)\|^2\right] \leq \rho\|\nabla f(x)\|^2 \leq 2L\rho(f(x) - f^*).$$

This is the result.

For $(\mathbf{WGC}:2L\rho) \implies (\mathbf{ES}:2L\rho)$, as $(\mathbf{WGC}:\rho)$ holds for all x, choose $x = x^*$, such that $\mathbb{E}_i\left[\|\nabla f_i(x^*)\|^2\right] \leq 2L\rho(f^* - f^*) = 0$, so $f_i(x^*) = 0$, so the two expressions are equivalent.

For $(\mathbf{ES}:2L\rho) \implies (\mathbf{ER}:2L\rho)$, we begin with restating the LHS of $(\mathbf{ER}:2L\rho)$ and expanding the square,

$$\begin{aligned}\mathbb{E}_i\|(f_i(x) - f_i(x^*)) - (f(x) - f^*)\|^2 &= \mathbb{E}_i\|f_i(x) - f_i(x^*)\|^2 + \|f(x) - f^*\|^2 \\ &- 2\langle \mathbb{E}_i\left[f_i(x) - f_i(x^*)\right], f(x) - f^* \rangle.\end{aligned}$$

Unbiasedness of the gradients gives $\langle \mathbb{E}_i\left[f_i(x) - f_i(x^*)\right], f(x) - f^* \rangle = \|f(x) - f^*\|^2$, such that the RHS gives,

$$= \mathbb{E}_i\|f_i(x) - f_i(x^*)\|^2 - \|f(x) - f^*\|^2 \leq \mathbb{E}_i\|f_i(x) - f_i(x^*)\|^2 \leq 2L\rho(f(x) - f^*),$$

using $(\mathbf{ES}:2L\rho)$ which is the result.

For $(\mathbf{ER} : 2L\rho) \implies (\mathbf{ABC} : 4L\rho, 1, 2\sigma^2)$, we make use of the standard square expansion, $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2$ twice. Starting by adding and subtracting $\nabla f_i(x^*)$,

$$\begin{aligned}
\|\nabla f_i(x) - \nabla f(x)\|^2 &= \|\nabla f_i(x) - \nabla f_i(x^*) + \nabla f_i(x^*) - \nabla f(x)\|^2 \\
&= \|\nabla f_i(x) - \nabla f_i(x^*) - \nabla f(x)\|^2 + \|\nabla f_i(x^*)\|^2 \\
&+ 2\langle \nabla f_i(x) - \nabla f_i(x^*) - \nabla f(x), \nabla f_i(x^*)\rangle \\
&= \|\nabla f_i(x) - \nabla f_i(x^*) - \nabla f(x)\|^2 + \|\nabla f_i(x^*)\|^2 \\
&+ \|\nabla f_i(x) - \nabla f_i(x^*) - \nabla f(x)\|^2 + \|\nabla f_i(x^*)\|^2 \\
&- \|\nabla f_i(x) - 2\nabla f_i(x^*) - \nabla f(x)\|^2 \\
&\leq 2\|\nabla f_i(x) - \nabla f_i(x^*) - \nabla f(x)\|^2 + 2\|\nabla f_i(x^*)\|^2.
\end{aligned}$$

Applying expectations, then $(\mathbf{ER} : 2L\rho)$ noting that $\nabla f^* = 0$ gives,

$$\begin{aligned}
\mathbb{E}_i\|\nabla f_i(x) - \nabla f(x)\|^2 &\leq 2\mathbb{E}_i\|(\nabla f_i(x) - \nabla f_i(x^*)) - (\nabla f(x) - f^*)\|^2 + 2\mathbb{E}_i\|\nabla f_i(x^*)\|^2 \\
&\leq 4L\rho(f(x) - f^*) + 2\sigma^2,
\end{aligned}$$

using that σ^2 is the supremum, as per definition 20. To finish then, we apply a form of the variance formula, $\mathbb{E}_i\|\nabla f_i(x) - \nabla f(x)\|^2 = \mathbb{E}_i\|\nabla f_i(x)\|^2 - \|\mathbb{E}_i\nabla f_i(x)\|^2$, as $\mathbb{E}_i\nabla f_i(x) = \nabla f(x)$. Thus,

$$\begin{aligned}
\mathbb{E}_i\|\nabla f_i(x)\|^2 - \|\mathbb{E}_i\nabla f_i(x)\|^2 &\leq 4L\rho(f(x) - f^*) + 2\sigma^2 \\
\iff \mathbb{E}_i\left[\|\nabla f_i(x)\|^2\right] &\leq 4L\rho(f(x) - f^*) + \|\nabla f(x)\|^2 + 2\sigma^2.
\end{aligned}$$

This is the result. □

We start analysis with the strong growth condition.

Definition 22 (Strong Growth Condition [63]) *A function f satisfies the* Strong Growth Condition*, denoted* $(\mathbf{SGC} : B)$*, if* $\forall x$*,*

$$\mathbb{E}_i\left[\|\nabla f_i(x)\|^2\right] \leq B\|\nabla f(x)\|^2.$$

If this holds $\forall i$*, without the expectation, then it satisfies the* Maximal Strong Growth Condition*, denoted* $(\mathbf{MSGC} : B)$*.*

This gives a second method to bound the stochastic gradients, by bounding it with the full gradient. This enforces regularity and uniformity, such that stochastic gradients cannot deviate too much and must preserve a somewhat similar structure, to ∇f for each x. A smaller constant is more restrictive. It is generally used in proofs to convert stochastic terms into a more familiar deterministic term, allowing similar techniques that are used in the standard gradient descent setting that we will see later [63] [65]. It has also been shown to hold for non-trivial problems, for example, under certain conditions it can be shown that the squared-hinge loss, a commonly used loss function for maximum margin classification tasks, can be shown to satisfy this condition for certain datasets [67]. However, this is not general, so there is a need for weaker conditions, displaying more ways to control the gradients.

Definition 23 (Weak Growth Condition [67]) *A function f satisfies the* Weak Growth Condition*, denoted* (**WGC** : A)*, if* $\forall x$*,*

$$\mathbb{E}_i \left[\|\nabla f_i(x)\|^2 \right] \leq A(f(x) - f^*).$$

This third method is to bound with the difference of a function and its optimal value. This intuitively makes sense; if a function has a smaller variation of values it is, in a sense, better behaved. If there is a smaller range of values a function can be, there is some limit on how quickly the function can change at each point, that is, bounding the (stochastic) gradients. This condition is weaker and as such is satisfied, in general, by the squared-hinge and very common squared loss functions [67]. These two conditions have seen more interest recently due to their links to (**IP**).

Theorem 13 (Interpolation Implications [50] [67]) *If f is such that each f_i satisfies* (**S** : L_i)*, then*

$$(\mathbf{IP}) \implies (\mathbf{WGC} : 2L_{max}).$$

Conversely, we have,

$$(\mathbf{WGC} : A) + (\mathbf{IN}) \implies (\mathbf{IP}).$$

Proof.

The proof for (**IP**) $\implies$ (**WGC** : $2L_{max}$) adapts [50] with results proven within this paper. First, note (**S** : L_i) $\implies$ (**S** : L_{max}) $\forall i$. This is a consequence of $L_i \leq L_{max}$:

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

Similar to the Descent Lemma, choosing $y = x - \frac{1}{L_{max}}\nabla f_i(x)$ leads to $f_i(y) \leq f_i(x) - \frac{1}{2L_{max}}\nabla f_i(x)$. This is when we use that (**IP**) $\implies$ $f_i(x^*) \leq f_i(y)$, to give $f_i(x^*) \leq f_i(x) - \frac{1}{2L_{max}}\|\nabla f_i(x)\|^2$. We then apply expectations with respect to i, giving,

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

Rearranging gives the result.

The converse of (**WGC** : A) + (**IN**) $\implies$ (**IP**) follows from a simple manipulation from [67]. The (**WGC** : A) gives $\mathbb{E}_i \left[\|\nabla f_i(x)\|^2 \right] \leq A(f(x) - f^*)$, simply set $x = x^*$, then we know $\mathbb{E}_i \left[\|\nabla f_i(x^*)\|^2 \right] = 0$, giving $\nabla f_i(x^*) = 0$ $\forall i$. Thus, (**IN**) gives the result, as local minimisers imply global minimisers. □

This result shows the link between this weaker condition and (**IP**), and as such any downstream conditions in figure 6, are of particular interest in the deep learning setting. However, this indicates it is still a rather strong assumption as interpolation is, for many problems without large parameter counts, not satisfied. In fact, we can show it is not far off the stronger condition before it.

Theorem 14 (Weak Growth Implication [67]) *If f satisfies* (**S**: L)*, then we have,*

$$(\mathbf{WGC} : A) + (\mathbf{PL} : \mu) \implies (\mathbf{SGC} : \frac{A}{2\mu}).$$

Proof.

Simply combine (**WGC** : A) then (**PL** : μ) for,

$$\mathbb{E}_i \left[\|\nabla f_i(x)\|^2\right] \leq A(f(x) - f^*) \leq A\frac{1}{2\mu}\|\nabla f(x)\|^2.$$

This is the result. □

Theorem 2 shows the relative generality of (**PL**), making this a powerful result.

The following conditions simplify to a combination of these previous three effects, so we provide the definitions and less discussion.

Definition 24 (Stochastic Condition Definitions) *A function f satisfies the* Expected Smoothness Condition*, denoted*
(**ES**: A)*, if* $\forall x$*,*

$$\mathbb{E}_i \left[\|\nabla f_i(x) - \nabla f_i(x^*)\|^2\right] \leq A(f(x) - f^*).$$

We also have the Expected Residual condition*, denoted* (**ER** : A)*, if* $\forall x$*,*

$$\mathbb{E}_i \left[\|(\nabla f_i(x) - \nabla f_i(x^*)) - (\nabla f(x) - \nabla f^*)\|^2\right] \leq A(f(x) - f^*),$$

noting that $\nabla f^* = 0$*. Then we have the* ABC condition*, denoted*
(**ABC** : A, B, C)*, or just* (**ABC**) *in general, if* $\forall x$*,*

$$\mathbb{E}_i \left[\|\nabla f_i(x)\|^2\right] \leq A(f(x) - f^*) + B\|\nabla f(x)\|^2 + C.$$

These conditions inherit previous intuition, and it has been shown (**ES**) $\not\iff$ (**ER**), the two of them having applications to non-linear least squares problems [20]. (**ES**) also has the following trivial result.

Lemma 7 *For a function f satisfying* (**S**: L)*, we have,*

$$(\mathbf{ES}\text{: A}) + (\mathbf{IP}) \implies (\mathbf{WGC}\text{: A}).$$

Proof.

Simply note (**IP**) $\implies \nabla f_i(x^*) = 0 \, \forall i$. The result follows. □

The emphasis in terms of interpretation should be on the (**ABC**) condition, which encapsulates the three main ways we can control the stochastic gradients. This also explains the choice of A, B, C notation for other conditions and is the most general condition we will consider. To continue, we consider a new chain of implications with links to the previous.

Theorem 15 (ABC Implications [33] [19] [20]) *For a function f satisfying* (**S**: L)*, we have the following chain of implications:*

$$(\mathbf{SGC} : B) \implies (\mathbf{RG} : B, C) \implies (\mathbf{ABC} : A, B, C).$$

Similarly,

$$(\mathbf{BG} : C) \implies (\mathbf{RG} : B, C) \implies (\mathbf{ABC} : A, B, C).$$

Further, we have,

$$(\mathbf{GC} : \rho) \implies (\mathbf{RG} : n, \rho(n-1)), \text{ and } (\mathbf{SS} : L) \implies (\mathbf{ABC} : 2L, 0, 2Lf^*).$$

This shows links of new conditions to our existing tree, with some familiar looking terms.

Definition 25 (Relaxed Growth and Sure Smoothness [20]) *A function f satisfies the* Relaxed Growth Condition*, denoted* $(\mathbf{RG} : B, C)$*, if* $\forall x$*,*

$$\mathbb{E}_i \left[\|\nabla f_i(x)\|^2 \right] \leq B\|\nabla f(x)\|^2 + C.$$

We also have the Sure Smoothness Condition*, denoted* $(\mathbf{SS} : L)$*, if* $\forall x, y$*, we have* $f(x) \geq 0$ *and,*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|.$$

The (**RG**) is a clear generalisation of (**BG**) and (**SGC**), or we can interpret it as simply strengthening (**ABC**) by removing one of the terms. Some other subsets of (**ABC**) terms have been stated in the literature ([55] Lemma 1), but as implied properties, not distinct conditions, so we do not present them here. Due to the zero term in $(\mathbf{ABC} : 2L, 0, C)$, this could be classified as a stronger condition, but, again, as it is not used in any convergence proofs here we omit a full statement. Another common, well-referenced condition is called Bounded Variance, equivalent to (**RG**: $1, C$) under the unbiased stochastic gradient assumption [33].

The (**SS**) condition is also familiar, being the equivalent of $(\mathbf{S} : L)$, with the *stochastic* gradients and the additional assumption of positivity. This assumption is reasonable here as we are mostly considering applications wherein (**FS**) represents a loss function. This condition has also been generalised to settings other than just (**FS**) [41].

The final condition we introduce is a little different however, utilising (**FS**).

Definition 26 (Gradient Confusion [62]) *A function f satisfies the* Gradient Confusion Condition*, denoted* $(\mathbf{GC} : \rho)$*, if* $\forall x$ *and* $\forall i \neq j$ *and some* $\rho > 0$*,*

$$\langle \nabla f_i(x), \nabla f_j(x) \rangle \geq -\rho.$$

This condition is exclusive to the (**FS**) setting, bounding how much stochastic gradients can disagree, or vary. This is a similar idea to previous bounds in expectation, but is approached in a new way. When the confusion is low, stochastic gradients rarely undo the work of previous iterations, leading to faster

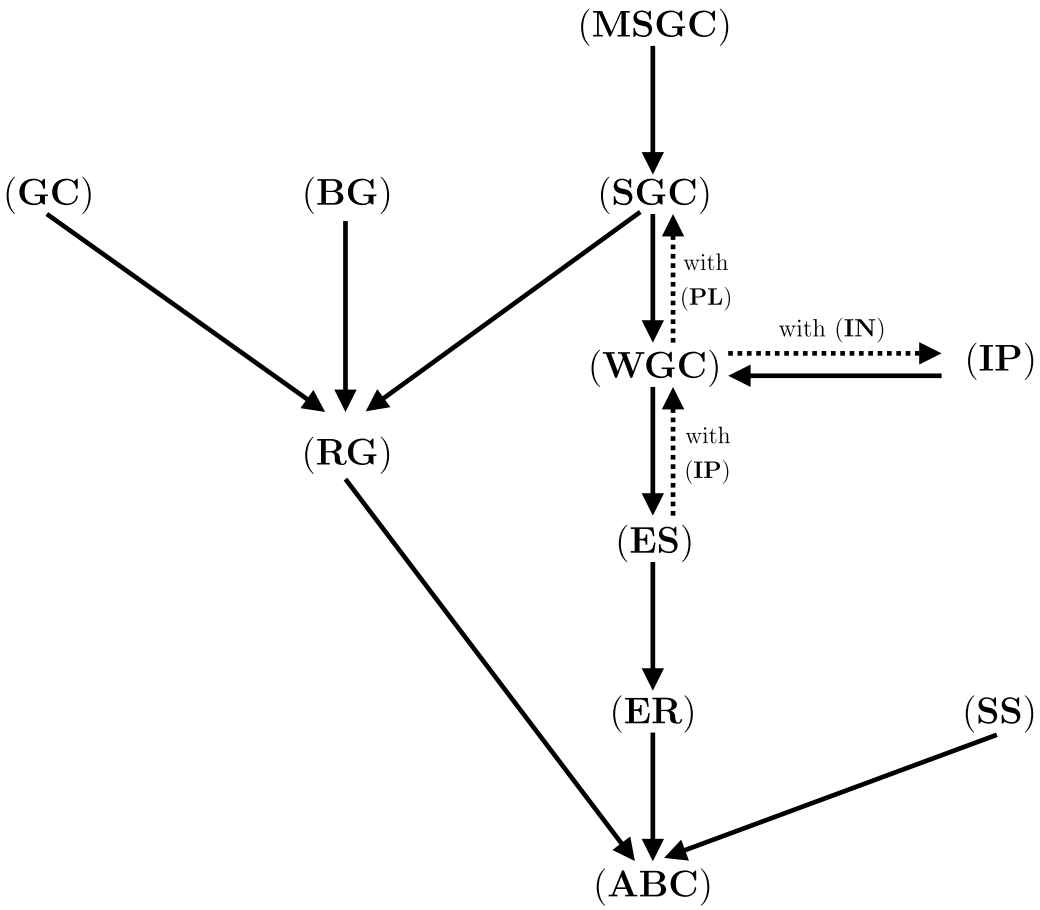


Figure 6: Diagram of discussed implications. Dotted lines indicate implications with additional assumptions, placed to the side. Refer to the theorems for constant conversions.

convergence. It has been proposed that the condition could provide insights into the effectiveness of gradient methods on large neural networks, such as increasing layer width leading to lower confusion and, as such, faster convergence [62].

This concludes the analysis of these stochastic conditions, with the discussion summarised in figure 6.

5 Results

In this section, we highlight results displaying trade-offs and advantages of different methods and assumptions. We then present a table of important results to summarise the argument. The results for (**GD**) seem to have a more complete theory, and some broader results under (**S** : L) for completeness are shown below. For (**SGD**), as the results are more scattered, we present core results, then explore specific examples to display how trade-offs between conditions affect the rate and mode of convergence.

5.1 Gradient Descent Convergence Proofs

To begin, in the standard (**GD**) setting without any major assumptions, the Descent Lemma 1 is strong enough to guarantee some form of convergence on its own. This will be the weakest result we obtain for (**GD**).

Theorem 16 ((GD) under (S) [7]) *Assuming* (**S** : L), (**GD**) *with step-size* $\eta_k = \eta = \frac{1}{L}$ *achieves a convergence rate of,*

$$\min_{t=1,..k} \|\nabla f(x_t)\|^2 \leq \frac{2L}{k}(f(x_0) - f^*).$$

Proof.

We start by rearranging the Descent Lemma 1, an implied result via (**S** : L). Summing over all iterations, we get, $\forall k$,

$$\|\nabla f(x_k)\|^2 \leq 2L(f(x_k) - f(x_{k+1})) \implies \sum_{t=0}^{k-1} \|\nabla f(x_t)\|^2 \leq \sum_{t=0}^{k-1} 2L(f(x_t) - f(x_{t+1})).$$

From this we can simplify. Firstly, the RHS represents a telescoping sum, secondly, note $f(x_k) \geq f^*$, giving:

$$2L \sum_{t=0}^{k-1} (f(x_t) - f(x_{t+1})) = 2L(f(x_0) - f(x_k)) \leq 2L(f(x_0) - f^*).$$

Thirdly, we can recognise that we are aiming to find a stationary point for f, that is, a point where $\nabla f(x) = 0$. Thus, we can aim to minimise $\|\nabla f(x)\|^2$, and take the smallest value seen up to this point and simplify, noting, trivially, that the minimum will be at least as small as every other value:

$$\sum_{t=0}^{k-1} \min_{j=0,...,k-1} \|\nabla f(x_j)\|^2 = k \times \min_{j=0,...,k-1} \|\nabla f(x_j)\|^2 \leq \sum_{t=0}^{k-1} \|\nabla f(x_t)\|^2$$

Combining gives:

$$\min_{j=0,...,k-1} k\|\nabla f(x_j)\|^2 \leq 2L(f(x_0) - f^*) \iff \min_{j=0,...,k-1} \|\nabla f(x_j)\|^2 \leq \frac{2L(f(x_0) - f^*)}{k}.$$

This completes the proof, but some results take square roots on both sides for convergence of $\min_{j=0,\dots,k-1} \|\nabla f(x_j)\|$ at a rate of order $O\left(\frac{1}{\sqrt{k}}\right)$. $\square$

Recalling the discussion of this mode of convergence from section 2, we are not guaranteed to find the global optimum, only *some* stationary point, sometime in the optimisation process. Although Lemma 1 guarantees progress, it says nothing about where we are progressing to. Now that we have shown the most basic form of convergence with the most basic setup, we can start making the additional assumptions we have examined. To begin, we state the classical results in cases of (**C**) and (**SC**).

Theorem 17 ((GD) under (SC)) [15] [7]) *Assuming* $(\mathbf{S}:L)$*, and* $(\mathbf{SC}:\mu)$ *with step-size* $\eta_k = \eta = \frac{1}{L}$ (**GD**) *achieves a convergence rate of,*

$$\|x_k - x^*\|^2 \leq \left(1 - \frac{\mu}{L}\right)^k \|x_0 - x^*\|^2.$$

Similarly, if f *is* (**C**) *we have a convergence rate of,*

$$f(x_k) - f^* \leq \frac{L}{2k}\|x_0 - x^*\|^2.$$

We wish to focus primarily on the non-convex cases for (**GD**),thus, these results will not be proven, but are classic. Proofs can be found in [15] and [7], leveraging similar techniques to upcoming proofs. The results show the strongest notion of convergence we cover for (**SC**) functions. There is linear convergence and, leveraging the uniqueness of the solution, convergence of iterates, not function values. The weaker condition of convexity has the hugely slower factor of $\frac{1}{k}$ convergence rate, emphasising how much stronger the quadratic difference term makes (**SC**).

Turning to non-convexity, we first show a pivotal result, which is a reason (**PL**) has seen much interest.

Theorem 18 ((GD) under (PL)) [56]) *Assuming* $(\mathbf{S}:L)$ *and* $(\mathbf{PL}:\mu)$, (**GD**) *with step-size* $\eta_k = \eta = \frac{1}{L}$ *achieves a convergence rate of,*

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*).$$

Proof.

The implied Descent Lemma gives $f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L}\|\nabla f(x_k)\|^2$. We can then combine this with $(\mathbf{PL}:\mu)$ giving,

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*) \implies -\frac{1}{2L}\|\nabla f(x_k)\|^2 \leq -\frac{\mu}{L}(f(x_k) - f^*).$$

To finish then we combine these:

$$f(x_{k+1}) - f(x_k) \leq -\frac{\mu}{L}(f(x_k) - f^*) \iff f(x_{k+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right)(f(x_k) - f^*).$$

Recursion gives the result. □

This result is incredibly important; not only does it obtain the same linear rate of convergence as (**SC**) with a weaker condition, but also the simplicity of the proof allows for easy modifications. Further, due to it being a relatively weak condition, many others imply it. This means we can find rates simply by finding the (**PL**) constant implied by a stronger condition, then apply this result. This approach is used extensively in Theorem 21. We can then compare this result to one under a different, stronger condition.

Theorem 19 ((GD) under (SQC)) [26]) *Assuming* (**S** : L) *and* (**SQC** : γ, μ), (**GD**) *with step-size* $\eta_k = \eta = \frac{1}{L}$ *achieves a convergence rate of,*

$$f(x_k) - f^* \leq \left(1 + \frac{2L}{\gamma}\right)\left(1 - \frac{\mu\gamma}{L}\right)^k (f(x_0) - f^*).$$

The proof can be found in the appendix and uses the idea of Lyapunov energy to find a decreasing series that provides the result. We note that it is impossible to tell which result due to constants is *better*. We can only find situations in which one is. Comparing the previous two results, the only difference in setup is strength of assumption, where (**PL**) is the weaker. This means from a theoretical standpoint, there could be extra information implied by (**SQC** : γ, μ) that is not leveraged in the previous proof that could lead to a stronger rate. We can directly apply the (**PL**) result with the implied property:

$$(\mathbf{SQC} : \gamma, \mu) \implies (\mathbf{RC} : \frac{\gamma}{2L}, \frac{\mu\gamma}{2}) \implies (\mathbf{RSI} : \frac{\mu\gamma}{2}) \implies (\mathbf{EB} : \frac{\mu\gamma}{2}) \implies (\mathbf{PL} : \frac{\mu\gamma}{2L}).$$

Ignoring the constant factor in front and focusing just on the rate, the (**PL**) implied result for this new condition reads $\left(1 - \frac{\mu\gamma}{2L^2}\right)^k$ compared to $\left(1 - \frac{\mu\gamma}{L}\right)^k$. This means that the rate for the above case decreases faster than the (**PL**) implied result if $L \geq \frac{1}{2}$. However, in cases like this we must be careful that the rate term is greater than 0. We know $\mu \leq L \implies \frac{\mu}{L} \leq 1 \implies \frac{\mu\gamma}{L} \leq 1 \implies 1 - \frac{\mu\gamma}{L} \geq 0$. For the (**PL**) implied result, however, noting $\frac{\mu\gamma}{2L^2} = \frac{\mu}{L}\frac{\gamma}{2L}$, we require $\frac{\gamma}{2L} \leq 1 \implies \gamma \leq 2L$ for a valid result. As above, this result would only be preferable if $L \leq \frac{1}{2}$, which combined with $\gamma \leq 2L$ implies there are certain values where it would seem the implied result is preferable, but it is in fact not valid, such as if $L = \frac{\gamma}{4}$; $L \leq \frac{1}{2}$ but $\gamma \geq 2L = \frac{\gamma}{2}$. From now on, to avoid having to make such distinctions, we present more conservative results, as in Theorem 21, but note that in certain situations additional constraints on constants can improve general results.

In contrast, the previous result in [26] has a multiplicative factor of $\frac{2}{\gamma}$, whilst we obtain $1 + \frac{2L}{\gamma}$. The difference comes from a simplification in the proof using (**QG**), and we obtain a different constant (our (**QG** : $\frac{\mu\gamma}{2L}$) versus their (**QG** : $\frac{\mu\gamma}{2-\gamma}$)). This result has an additional dependence on L, but is worth it if,

$$1 + \frac{2L}{\gamma} \leq \frac{2}{\gamma} \iff \gamma + 2L \leq 2.$$

This will be heavily dependent on the problem in question, but as $\gamma \leq 1$, we can safely assume our constant is better if $L \leq \frac{1}{2}$. This is the opposite of the condition from the previous paragraph!

These two points emphasise that the problem's context matters. Theory would suggest extra information through the stronger condition of Strong Quasar-Convexity provides faster convergence. This is backed up by the first result if $L \geq \frac{1}{2}$, which Lemma 2 implies is the more general case.

It is also worth noting that simplifying with (**QG**), as done in Theorem 19, does not improve the result due to the direction of inequality. The more general form is seen in an upcoming (**NAG**) proof. However, it depends more on the individual initialisation terms x_0 for each case; the advantage of simplifying with quadratic growth is a more uniform bound.

We now move on to a relatively weaker condition.

Theorem 20 ((GD) under (*C)) [21]) *Assuming* (**S** : L) *and* (***C**) *with a unique minimiser* x^*, (**GD**) *with step-size* $\eta_k = \eta = \frac{1}{L}$ *achieves a convergence rate of,*

$$f(x_k) - f^* \leq \frac{L}{2k}\|x_0 - x^*\|^2.$$

This means we obtain the same result as for full (**C**) under the additional assumption of a unique minimiser, showing that it is only certain areas of the function related to the solution that we care about in this optimisation.

We finish our discussion of (**GD**) results by utilising the previous trees of implications to obtain new results.

Theorem 21 ((GD) Implied Results [21]) *Assuming* (**S** : L)*, we have, for* $\eta_k = \eta$ *that under the following conditions,* (**GD**) *achieves convergence rates of,*

$$\begin{aligned}
(\textbf{ESC}\text{: } \mu) &\implies f(x_k) - f^* \leq \left(1 - \frac{\mu^2}{4L^2}\right)^k (f(x_0) - f^*) \\
(\textbf{WSC}\text{: } \mu) &\implies f(x_k) - f^* \leq \left(1 - \frac{\mu^2}{4L^2}\right)^k (f(x_0) - f^*) \\
(\textbf{RSI}\text{: } \mu) &\implies f(x_k) - f^* \leq \left(1 - \frac{\mu^2}{L^2}\right)^k (f(x_0) - f^*) \\
(\textbf{EB}\text{: } \mu) &\implies f(x_k) - f^* \leq \left(1 - \frac{\mu^2}{L^2}\right)^k (f(x_0) - f^*) \\
(\textbf{WPL}\text{: } \mu_1) + (\textbf{QG}\text{: } \mu_2) &\implies f(x_k) - f^* \leq \left(1 - \frac{\mu_1\mu_2}{4L}\right)^k (f(x_0) - f^*) \\
(\textbf{QC}\text{: } \gamma) + (\textbf{QG}\text{: } \mu) &\implies f(x_k) - f^* \leq \left(1 + \frac{4L}{\gamma}\right)\left(1 - \frac{\gamma\mu}{2L}\right)^k (f(x_0) - f^*).
\end{aligned}$$

Proof.

These follow directly from the implications in section 3, in particular Theorems 2, 4 and 10. The first four directly imply (**PL**) from Theorem 2, such that we can then apply Theorem 18, but to ensure validity we perform the final (**EB**) $\implies$ (**PL**) step via Theorem 4 rather than 2. For the fifth, we use the implication
(**WPL** : μ_1) + (**QG** : μ_2) $\implies$ (**PL** : $\frac{\mu_1\mu_2}{4}$) instead. The final implication uses Theorem 10 for (**QC** : γ) + (**QG** : μ) $\implies$ (**SQC** : $\theta\gamma, \frac{1-\theta}{\theta}\mu$). Then, applying 19 would give,

$$f(x_k) - f^* \leq \left(1 + \frac{2L}{\theta\gamma}\right)\left(1 - \frac{\gamma\mu(1-\theta)}{L}\right)^k (f(x_0) - f^*).$$

Recall that $\theta \in (0, 1)$ is a specified quantity. To make the rate as fast as possible, we would want $\theta \to 0$, such that $\left(1 - \frac{\gamma\mu(1-\theta)}{L}\right)^k$ is minimised. However, this takes $\left(1 + \frac{2L}{\theta\gamma}\right) \to \infty$. Thus, we simply present the result for $\theta = \frac{1}{2}$, leaving identification of the optimal θ to future work. □

Note results could be found straight from Theorem 2, but these involve $\frac{\mu}{L^2}$ terms that we can't guarantee are ≤ 1 without changes to the step-size or additional assumptions on L, as discussed after Theorem 19. Hence, we use Theorem 4 instead to leverage $\mu \leq L$. The results here for stronger conditions are worse, in the sense of slower convergence, as $\frac{\mu}{L} \leq 1 \implies \frac{\mu}{L} \geq \frac{\mu^2}{L^2}$. A demonstration of this is found after Theorem 26. The primary point we make here however, is that these proofs are trivial, and in the case of (**WPL**) + (**QG**) and (**QC**) + (**QG**), useful rates are obtained. Being able to rapidly find results, even if sub-optimal for certain conditions, is valuable in practical settings, so we argue for the importance of mapping these conditions out.

5.2 Stochastic Gradient Descent Convergence Proofs

Moving onto (**SGD**), we proceed in the same way as before, establishing basic results. This time we begin with the convex base results.

Theorem 22 ((SGD) under (SC)) [15]) *Assuming, for f satisfying* (**FS**) *with each f_i satisfying* (**S** : L_i)*, that each f_i satisfies* (**C**)*, with step-size $\eta_k = \eta \leq \frac{1}{4L_{max}}$* (**SGD**) *achieves a convergence rate of,*

$$\mathbb{E}_i[f(\bar{x}_k)] - f^* \leq \frac{1}{\eta k}\|x_0 - x^*\|^2 + 2\eta\sigma^2,$$

for the gradient noise σ^2. If additionally (**SC** : μ) *is assumed, with step-size $\eta_k = \eta \leq \frac{1}{2L_{max}}$* (**SGD**) *achieves a convergence rate of,*

$$\mathbb{E}_i\|x_k - x^*\|^2 \leq (1 - \mu\eta)^k \|x_0 - x^*\|^2 + \frac{2\eta}{\mu}\sigma^2.$$

Again, this will not be proven, due to similarity of notation and areas of focus. Proofs are found in [15]. There is a key difference compared to the (**GD**) case, namely

the stochasticity means we can only guarantee convergence to a neighbourhood around the solution. These constant terms are dependent on the step-size; if a large step-size is chosen, the rate is fast to *some* neighbourhood, but the additional constant, and thus inaccuracy, is large. This reinforces the theory discussed in section 2, suggesting the need for additional assumptions to achieve full convergence. This intuitively makes sense; the gradient conditions control how well-behaved the function is in certain neighbourhoods, restricting how much the functions can change and how quickly. By assuming this, local neighbourhoods of the function behave more smoothly, allowing (**SGD**) to converge to more specific points without skipping over them [15]. The same rates as before are found, which raises the question of whether there are any methods than can improve this. The answer is yes, as Theorems 25 and 26 show. Another thing to note is the average iterate convergence in the (**C**) case, showing that it is not a direct port from (**GD**).

The previous result indicated a need for the additional assumptions. We continue with (**WGC** : A).

Theorem 23 ((SGD) under (SC) and (WGC)) [67] [26]) *Assuming, for f satisfying* (**WGC** : $2AL$)*, that f satisfies* (**SC** : μ)*, with step-size $\eta_k = \eta = \frac{1}{AL}$* (**SGD**) *achieves a convergence rate of,*

$$\mathbb{E}_i\|x_k - x^*\|^2 \leq \left(1 - \frac{\mu}{AL}\right)^k \|x_0 - x^*\|^2.$$

If instead f satisfies (**C**)*, with step-size $\eta_k = \eta = \frac{1}{4AL}$* (**SGD**) *achieves a convergence rate of,*

$$\mathbb{E}_i[f(\bar{x}_k)] - f^* \leq \frac{4L(1+A)}{k}\|x_0 - x^*\|^2.$$

By additionally assuming (**WGC**), we obtain the same rates as before but, due to the additional regularity, converge to a true optimal solution, not just the neighbourhood. Again, proofs can be found in [67].

Using these conditions can also improve the existing rates, not just the mode of convergence. We now present a selection of results with varying assumptions on the step-size and stochastic gradients for functions satisfying (**PL**).

Theorem 24 ((SGD) under Various Conditions [32] [67] [62] [33]) *Assuming* (**S** : L) *and* (**PL** : μ)*, we have, for η_k specified in the appendix proofs and stochastic gradient conditions specified below, that* (**SGD**) *achieves*

convergence rates of,

$$
\begin{aligned}
(\textbf{BG}\text{: } C) + (\text{adaptive } \eta_k) \quad &\implies \quad \mathbb{E}_i\left[f(x_k)\right] - f^* \leq \frac{LC}{2k\mu^2} \\
(\textbf{BG}\text{: } C) + (\text{constant } \eta) \quad &\implies \quad \mathbb{E}_i\left[f(x_k)\right] - f^* \leq (1-2\mu\eta)^k \left(f(x_0) - f^*\right) + \frac{LC\eta}{4\mu} \\
(\textbf{SGC}\text{: } B) \quad &\implies \quad \mathbb{E}_i\left[f(x_k)\right] - f^* \leq \left(1 - \frac{\mu}{BL}\right)^k \left(f(x_0) - f^*\right) \\
(\textbf{GC}\text{: } \rho) + (\textbf{PL}\text{: } \mu)\ \forall f_i \quad &\implies \quad \mathbb{E}_i\left[f(x_k)\right] - f^* \leq r_1^k (f(x_0) - f^*) + \frac{\eta\rho}{1-r_1} \\
(\textbf{ABC}\text{: } 2A,\ B,\ C) \quad &\implies \quad \mathbb{E}_i\left[f(x_k)\right] - f^* \leq \frac{9LC}{2k\mu^2} + r_2^k (f(x_0) - f^*)
\end{aligned}
$$

where $r_1 = 1 - \frac{2\mu}{n}\left(\eta - \frac{nL\eta^2}{2}\right)$ *and* $r_2 = \exp\left(-\frac{\mu}{2L\max\{\frac{A}{\mu}, B\}}\right)$.

Proof.

We only prove the second and third here as illustrative examples, with the rest in the appendix. They follow a similar technique, manipulating the $(\mathbf{S}: L)$ expression and (**SGD**) terms to leverage the assumed condition.

The second is from [32]. First we prove an intermediary result required for both the first two theorems above, in the most general notation required for the first (**BG**) proof in the appendix. Using $(\mathbf{S}: L)$, we have

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2.$$

As opposed to previous proofs with (**GD**), we now use the (**SGD**) term with the subtly different stochastic gradient term, leading to,

$$f(x_{k+1}) \leq f(x_k) - \eta_k \langle \nabla f(x_k), \nabla f_i(x_k) \rangle + \frac{L\eta_k^2}{2}\|\nabla f_i(x_k)\|^2.$$

To be able to use a Descent Lemma style technique in this setting, we need to take expectations of both sides. Doing this, then simplifying with $(\mathbf{BG}: C)$ gives,

$$
\begin{aligned}
\mathbb{E}_i[f(x_{k+1})] &\leq f(x_k) - \eta_k \langle \nabla f(x_k), \mathbb{E}_i[\nabla f_i(x_k)] \rangle + \frac{L\eta_k^2}{2}\mathbb{E}_i\left[\|\nabla f_i(x_k)\|^2\right]. \\
&\leq f(x_k) - \eta_k \|\nabla f(x_k)\|^2 + \frac{LC\eta_k^2}{2}.
\end{aligned}
$$

Following this, we can leverage $(\mathbf{PL}: \mu)$ to obtain,

$$\mathbb{E}_i[f(x_{k+1})] \leq f(x_k) - 2\mu\eta_k (f(x_k) - f^*) + \frac{LC\eta_k^2}{2}.$$

To finish this preliminary result, we subtract the optimal f^* from both sides for,

$$
\begin{aligned}
\mathbb{E}_i[f(x_{k+1})] - f^* &\leq f(x_k) - 2\mu\eta_k (f(x_k) - f^*) + \frac{LC\eta_k^2}{2} - f^* \\
&= (1 - 2\mu\eta_k)(f(x_k) - f^*) + \frac{LC\eta_k^2}{2}.
\end{aligned}
$$

For convergence, we switch to constant step-size notation and use recursion of our previous result, taking expectations when necessary for,

$$\begin{aligned}
\mathbb{E}_i[f(x_{k+1})] - f^* &\leq (1-2\mu\eta)(f(x_k) - f^*) + \frac{LC\eta^2}{2} \\
&\leq (1-2\mu\eta)^{k+1}(f(x_0) - f^*) + \frac{LC\eta^2}{2}\sum_{i=0}^{k+1}(1-2\mu\eta)^i \\
&\leq (1-2\mu\eta)^{k+1}(f(x_0) - f^*) + \frac{LC\eta^2}{2}\sum_{i=0}^{\infty}(1-2\mu\eta)^i \\
&\leq (1-2\mu\eta)^{k+1}(f(x_0) - f^*) + \frac{LC\eta}{4\mu},
\end{aligned}$$

where we have finished with the standard result of the limit of a geometric series. This is the result if $k+1 \to k$.

For the third result with (**SGC**), we begin with the same trick using (**S** : L) and taking expectations as above, leading to the Descent Lemma styled,

$$\begin{aligned}
\mathbb{E}_i[f(x_{k+1})] - f^* &\leq \langle \nabla f(x_k), -\eta_k \mathbb{E}_i[\nabla f_i(x_k)]\rangle + \frac{L}{2}\mathbb{E}_i\| - \eta_k \nabla f_i(x_k)\|^2 \\
&= -\eta_k \|\nabla f(x_k)\|^2 + \frac{L\eta_k^2}{2}.
\end{aligned}$$

Then we leverage (**SGC** : B) and substitute $\eta_k = \eta = \frac{1}{BL}$ to obtain,

$$\begin{aligned}
\mathbb{E}_i[f(x_{k+1})] - f(x_k) &\leq -\eta_k \|\nabla f(x_k)\|^2 + \frac{L\eta_k^2}{2} \\
&\leq \left(B\frac{L\eta_k^2}{2} - \eta_k\right)\|\nabla f(x_k)\|^2 \\
&= -\frac{1}{2BL}\|\nabla f(x_k)\|^2.
\end{aligned}$$

Next we apply (**PL** : μ) for,

$$\mathbb{E}_i[f(x_{k+1})] - f(x_k) \leq -\frac{1}{2BL}\|\nabla f(x_k)\|^2 \leq -\frac{\mu}{BL}(f(x_k) - f^*).$$

To finish, subtract f^* from both sides and rearrange for,

$$\begin{aligned}
\mathbb{E}_i[f(x_{k+1})] - f(x_k) &\leq \frac{\mu}{BL}(f^* - f(x_k)) \\
\implies \mathbb{E}_i[f(x_{k+1})] - f^* &\leq \left(1 - \frac{\mu}{BL}\right)(f(x_k) - f^*).
\end{aligned}$$

Recursion gives the result. □

Beginning with the top two, we see the issue with using different step-sizes. In the first case, we use an adaptive step-size $\eta_k = \frac{2k+1}{2\mu(k+1)^2}$, which satisfies Robbins-Monro,

but has a slower rate, of order $\frac{1}{k}$. If one is willing to sacrifice this guarantee, the second result shows (**SGD**) can converge linearly to a neighbourhood around an optimum, up to a constant dependent on the step-size.

If, however, we assume (**SGC**), we can obtain complete linear convergence. This is powerful, especially as this convergence is achieved with the constant step-size of $\eta_k = \eta = \frac{1}{BL}$. The (**GC**) result also has a constant step-size and attains the same level of convergence as the constant step-size (**BG**) result, but requires an additional assumption on the f_i of satisfying (**PL**) individually. One unusual result is that of (**ABC**) achieving a rate of order $\frac{1}{k}$ with a complicated adaptive step-size. This is surprising as it is a strictly weaker condition than the equivalent (**BG**) result at the top, only worse by a constant factor of 9. There is an additional term, however it decays at a linear rate. It appears that this rate is the best possible for adaptive step-sizes [54], so it seems only further improvements to the multiplicative constants are possible.

5.3 Nesterov Accelerated Gradient Convergence Proofs

Now we move onto the accelerated gradient case, stating two base theorems.

Theorem 25 ((NAG) under (SC) and (SGC)) [67]) *Assuming* (**S** : L), (**SC** : μ) *and* (**SGC** : B), (**NAG**) *with parameters*

$$\eta_k = \eta = \frac{1}{BL},\ s_k = s = \frac{1}{\sqrt{\eta B \mu}},\ \beta_k = \beta = 1 - \sqrt{\frac{\mu\eta}{B}},$$

$$b_{k+1} = \sqrt{\mu}\beta_k^{-\frac{k+1}{2}},\ a_{k+1} = \beta_k^{-\frac{k+1}{2}},\ \alpha_k = \frac{s_k\beta_k b_{k+1}^2\eta}{s_k\beta_k b_{k+1}^2\eta + a_k^2},$$

achieves a convergence rate of,

$$\mathbb{E}[f(x_{k+1})] - f^* \leq \left(1 - \sqrt{\frac{\mu}{B^2L}}\right)^k [f(x_0) - f^* + \frac{\mu}{2}\|x_0 - x^*\|^2].$$

Theorem 26 ((NAG) under (C) and (SGC)) [67]) *Assuming* (**S** : L), (**C**) *and* (**SGC** : B), (**NAG**) *with parameters*

$$\eta_k = \eta = \frac{1}{BL},\ s_k = \frac{\frac{1}{B} + \sqrt{\frac{1}{B^2} + 4s_{k-1}^2}}{2},\ \beta_k = \beta = 1,$$

$$b_{k+1} = 1,\ a_{k+1} = s_k\sqrt{B\eta},\ \alpha_k = \frac{s_k\eta}{s_k\eta + a_k^2},$$

achieves a convergence rate of,

$$\mathbb{E}[f(x_{k+1})] - f^* \leq \frac{2B^2L}{k^2}\|x_0 - x^*\|^2.$$

We prove this differently to the original paper by identifying useful properties these assumptions imply, using those to complete the proof.

Lemma 8 *Both initialisations from Theorems 25 and 26 have the following properties:*

$$\beta_k = 1 - s_k\mu\eta \ (\mathbf{N1}),\ s_k = \frac{1}{B}\left(1 + \frac{\beta_k(1-\alpha_k)}{\alpha_k}\right) \ (\mathbf{N2}),\ b_k^2 = \beta_k b_{k+1}^2 \ (\mathbf{N3}),$$

$$b_{k+1}^2 s_k^2 \eta B = a_{k+1}^2 \ (\mathbf{N4}),\ \frac{s_k\eta\beta_k(1-\alpha_k)}{\alpha_k} = \frac{a_k^2}{b_{k+1}^2} \ (\mathbf{N5}),\ b_{k+1}^2 s_k\eta - a_{k+1}^2 + a_k^2 = 0 \ (\mathbf{N6}).$$

Having these properties allows us to prove the main convergence results. We omit the full proof and refer the reader to [67]. With the above Lemma 8, their Theorems 1 and 2 follow the same proof.

The results show (**SC**) reaches the linear rate of convergence that it has enjoyed throughout. Comparing to the (**SGD**) case, we can use Theorems 3 and 24 to give the exact same rate for $(\mathbf{SC} : \mu)$ as $(\mathbf{PL} : \mu)$ under $(\mathbf{SGC} : B)$, of

$$\mathbb{E}[f(x_{k+1})] - f^* \leq \left(1 - \frac{\mu}{BL}\right)^k (f(x_0) - f^*).$$

Note this is a factor of B off the original result from [63]. Comparing, (**NAG**) improves the rate if,

$$\begin{aligned}
\left(1 - \sqrt{\frac{\mu}{B^2L}}\right)^k &\leq \left(1 - \frac{\mu}{BL}\right)^k \\
\iff \frac{\mu}{BL} &\leq \sqrt{\frac{\mu}{B^2L}} \\
\iff \frac{\mu}{L} &\leq \sqrt{\frac{\mu}{L}} \\
\iff \mu &\leq L,
\end{aligned}$$

which we know to be true via assumptions in section 2. Thus, (**NAG**) improves the rate for (**SC**) functions. For (**C**), the $\frac{1}{k}$ rate seen in Theorem 23 before is improved significantly to $\frac{1}{k^2}$, if the stronger assumption of (**SGC**) rather than (**WGC**), can be made.

However, acceleration is not always achieved. The next result has a different set of initialisations and the condition of (**SQC**) but is *not* under stochasticity.

Theorem 27 ((NAG) under (SQC)) [26]) *Assuming* $(\mathbf{S} : L)$ *and* $(\mathbf{SQC} : \gamma, \mu)$, (**NAG**) *with parameters*

$$\eta_k = \eta = \frac{\gamma^2\mu}{L^2},\ s_k = s = \frac{L}{\gamma\mu},\ \beta_k = \beta = 1 - \gamma\sqrt{\mu\eta}, \alpha_k = \alpha = \frac{\sqrt{\mu\eta}}{1 + \sqrt{\mu\eta}},$$

setting $v_0 = x_0$ *achieves a convergence rate of,*

$$f(x_k) - f^* \leq \left(1 + \frac{2L}{\gamma}\right)\left(1 - \gamma^2\frac{\mu}{L}\right)^k (f(x_0) - f^*).$$

Note we still find a linear rate, but it quickly worsens with increasing non-convexity, as seen by the γ^2 term in the rate. This is a case where acceleration has not occurred. As there is no stochasticity, we can refer to Theorem 19 where under the same assumptions, a rate with only γ dependency is obtained for (**GD**), which as $\gamma \leq 1$, is strictly faster.

5.4 Table

Here we summarise the results we have presented. If not otherwise stated, $(\mathbf{S} : L)$ is assumed for all.

Key	Method	$C(k)$	Step	Additional	$R(k)$
$(\mathbf{SC}:\mu)$	**(GD)**	$\|x_k - x^*\|^2$	$\frac{1}{L}$	N/A	$\left(1-\frac{\mu}{L}\right)^k$
(C)	**(GD)**	$f(x_k)-f^*$	$\frac{1}{L}$	N/A	$\frac{1}{k}$
$(\mathbf{SQC}:\gamma,\mu)$	**(GD)**	$f(x_k)-f^*$	$\frac{1}{L}$	N/A	$\left(1-\frac{\gamma\mu}{L}\right)^k$
$(\mathbf{*C}:\mu)$	**(GD)**	$f(x_k)-f^*$	$\frac{1}{L}$	unique x^*	$\frac{1}{k}$
$(\mathbf{ESC}:\mu)$	**(GD)**	$f(x_k)-f^*$	$\frac{1}{L}$	N/A	$\left(1-\frac{\mu^2}{4L^2}\right)^k$
$(\mathbf{WSC}:\mu)$	**(GD)**	$f(x_k)-f^*$	$\frac{1}{L}$	N/A	$\left(1-\frac{\mu^2}{4L^2}\right)^k$
$(\mathbf{RSI}:\mu)$	**(GD)**	$f(x_k)-f^*$	$\frac{1}{L}$	N/A	$\left(1-\frac{\mu^2}{L^2}\right)^k$
$(\mathbf{EB}:\mu)$	**(GD)**	$f(x_k)-f^*$	$\frac{1}{L}$	N/A	$\left(1-\frac{\mu^2}{L^2}\right)^k$
$(\mathbf{PL}:\mu)$	**(GD)**	$f(x_k)-f^*$	$\frac{1}{L}$	N/A	$\left(1-\frac{\mu}{L}\right)^k$
$(\mathbf{QG}:\mu)$	**(GD)**	$f(x_k)-f^*$	$\frac{1}{L}$	$(\mathbf{QC}:\mu)$	$\left(1-\frac{\gamma\mu}{2L}\right)^k$
$(\mathbf{QG}:\mu_2)$	**(GD)**	$f(x_k)-f^*$	$\frac{1}{L}$	$(\mathbf{WPL}:\mu_1)$	$\left(1-\frac{\mu_1\mu_2}{4L}\right)^k$
$(\mathbf{SC}:\mu)$	**(SGD)**	$\mathbb{E}\|x_k - x^*\|^2$	η	N/A	$(1-\eta\mu)^k + \mathrm{c}$
$(\mathbf{SC}:\mu)$	**(SGD)**	$\mathbb{E}\|x_k - x^*\|^2$	$\frac{1}{AL}$	$(\mathbf{WGC}:2AL)$	$\left(1-\frac{\mu}{AL}\right)^k$
$(\mathbf{SC}:\mu)$	**(SGD)**	$\mathbb{E}f(x_k)-f^*$	$\frac{1}{BL}$	$(\mathbf{SGC}:B)$	$\left(1-\frac{\mu}{BL}\right)^k$
(C)	**(SGD)**	$\mathbb{E}f(\bar{x}_k)-f^*$	η	N/A	$\frac{1}{k}+\mathrm{c}$
(C)	**(SGD)**	$\mathbb{E}f(\bar{x}_k)-f^*$	$\frac{1}{4AL}$	$(\mathbf{WGC}:2AL)$	$\frac{1}{k}$
$(\mathbf{PL}:\mu)$	**(SGD)**	$\mathbb{E}f(x_k)-f^*$	$\frac{2k+1}{2\mu(k+1)^2}$	$(\mathbf{BG}:C)$	$\frac{1}{k}$
$(\mathbf{PL}:\mu)$	**(SGD)**	$\mathbb{E}f(x_k)-f^*$	η	$(\mathbf{BG}:C)$	$(1-2\mu\eta)^k+\mathrm{c}$
$(\mathbf{PL}:\mu)$	**(SGD)**	$\mathbb{E}f(x_k)-f^*$	$\frac{1}{BL}$	$(\mathbf{SGC}:B)$	$\left(1-\frac{\mu}{BL}\right)^k$
$(\mathbf{PL}:\mu)$	**(SGD)**	$\mathbb{E}f(x_k)-f^*$	η	$(\mathbf{GC}:\rho)$	$r_1^k+\mathrm{c}$
$(\mathbf{PL}:\mu)$	**(SGD)**	$\mathbb{E}f(x_k)-f^*$	η_k	$(\mathbf{ABC}:2A,B,C)$	$\frac{1}{k}$
$(\mathbf{SC}:\mu)$	**(NAG)**	$\mathbb{E}f(x_k)-f^*$	$\frac{1}{BL}$	$(\mathbf{SGC}:B)$	$\left(1-\sqrt{\frac{\mu}{B^2L}}\right)^k$
$(\mathbf{C}:\mu)$	**(NAG)**	$\mathbb{E}f(x_k)-f^*$	$\frac{1}{BL}$	$(\mathbf{SGC}:B)$	$\frac{1}{k^2}$
$(\mathbf{SQC}:\gamma,\mu)$	**(NAG)**	$f(x_k)-f^*$	$\frac{\gamma^2\mu}{L^2}$	not stochastic	$\left(1-\gamma^2\frac{\mu}{L}\right)^k$

Table 1: Summary of results presented. Terms of step-size η represent a general step-size, sometimes specified in the proof. Similarly, η_k represents adaptive step-sizes, found in the proofs. Red indicates results implied by the work of this paper. The +c represents convergence to a neighbourhood, up to a constant independent of k.

6 Conclusion

This concluding section will provide discussion of the results presented, alongside highlighting our contributions with respect to their limitations, and suggest key areas where future research could be directed.

Overall, we presented a range of results from across the fields of convex and non-convex optimisation. Many implications were identified in a literature review and unified in figures 5 and 6, including some of our own additions. One obvious limitation is the breadth of the field, meaning that we cannot be certain all possible connections have been found or that, despite our best efforts in reviewing the literature, our results are presented for the first time here. Certainly, many of these proofs, especially Theorem 11, are trivial and not theoretical breakthroughs. As such, we wish to emphasise that the true contribution of our work is the unification of many streams of results and statement in this implication tree format. It allows implied proofs along the lines of Theorem 21, and acts as a starting or reference point for future work. Rather than isolated implication results to supplement the main convergence theorems as in [11] [32] and many others, we mainly focused on the conditions, stating convergence results primarily to argue for the importance of mapping them. This is demonstrated especially via the (**WPL**) + (**QG**) rate of Theorem 8 - because we have mapped out the conditions well, generating new rate results can be simple.

This focus on conditions is also a weakness, since we have not performed a full literature review of the best-in-class convergence rates for each condition, method, and step-size. There are many results not reported and many conditions without results. Of note are convergence proofs for (**RC**) [10] (linear) and (**WPL**) [11] ($O(\frac{1}{k})$). A follow-up piece based on this work that compiles the optimal rates for each combination of conditions and methods would allow identification of any weaker areas when cross-referenced with ours. That is, identifying conditions where performance is matched or improved upon by a weaker condition. Throughout we argued for stronger assumptions implying stronger convergence, summarised in table 1, however we only covered a select few results. A full review would allow for a more convincing argument, or alternatively, show areas for future work.

Our work also only briefly mentioned other variable aspects of convergence proofs, notably methods for selecting the step-size. While we simply state the step-size in each theorem we prove, in practice these constants or rates of decrease may not be known, such as the proofs utilising $\eta = \frac{1}{L}$. Tools for selecting step-size represent another level of complexity in our analysis, and future work may wish to consider schemes such as Armijo line search [68]. Similarly, only three optimisation schemes were considered, with countless more specialised methods existing in the literature [36] [55]. Despite not analysing in great detail, (**NAG**) was included to offer insight into how altering the optimisation method used can affect the rate convergence. In future works, other momentum based methods such as Polyak's heavy ball momentum based method [57], mini-batched (**SGD**) [20], proximal gradient methods [32] and mirror descent [78] could be investigated.

We can also summarise and point to results displaying our argument of stronger assumptions implying stronger convergence. For (**GD**), Theorems 17 and 19 (setting $\gamma = 1$ to give a rate for (***SC**)) with 20 give direct implications where the stronger condition provides notably faster convergence. Similarly, Theorem 21 shows how the weaker (**QG**) requires extra assumptions to find convergence. All of these results, compared to Theorem 16, show that any level of assumption has large positive effects on the convergence mode and rate. For (**SGD**), Theorems 22 and 23 directly show how adding regularity assumptions on the stochastic gradients allows more precise convergence to the exact solution, rather than a neighbourhood. Theorem 24 displays a range of results, notably showing how step-sizes affect convergence. Theorem 26 displays that changing the optimisation scheme and making the stronger assumption of (**SGC**), the rate for convex functions can be improved from $O(\frac{1}{k})$ in Theorem 23 to $O(\frac{1}{k^2})$. One area where this argument was not displayed was Theorem 21, as emphasis was on the ease of these proofs.

Another generalisation of our work would be to remove the uniform (**S**) assumption. Many results did not require it, yet it was assumed. Leaving it out, or swapping to a weaker bound, could lead to a very different equivalent of figure 5. A few works have begun investigating this, with (**WS**) [42] or more general conditions [21].

We believe many proofs involving (***SC**) and could be replicated using (**SQC**) instead, to give a weaker but more general result. The same analysis holds for (***C**) and (**QC**). The parameter γ can, in some cases, simply be absorbed into converted constants, as we demonstrated in improving the (***SC**) $\implies$ (**RC**) result of [10] in Theorem 8 with (**SQC**). However, not all techniques allow this. Theorem 20 using (***C**), for example, cannot be altered by assuming the weaker (**QC**) as we cannot show that $f(x_k) - f^* - \langle \nabla f(x_k), x_k - x^* \rangle \leq 0$, a required condition for the proof, and coincidentally the definition of the strictly stronger (***C**).

Similarly, the (**GC**) proof from Theorem 24 required the additional assumption that all f_i satisfy (**PL**). Many proofs require this for (**C**), however we believe alternative rates could be obtained by assuming other conditions for individual f_i.

We showed in Theorem 11 that many conditions imply the important (**VC**) property, however the bound of non-negativity is directly improved in the case of a unique minimiser by (**UAAC**). As such, for conditions that already imply (**VC**), future work making the stronger assumption of (**UAAC**) could improve results. We suggest applying this to (**QC**) functions, via figure 5. Further possible proof techniques could leverage our discussion in section 3 of $\|x_p - x\|^2 \leq \|x^* - x\|^2$.

Throughout, we made an effort to keep track of constants when converting between conditions, and discussed how these affect rates of convergence throughout. Future work could incorporate the collection of known constant conversions into our implication trees 5 and 6. Empirical work to determine the standard scale of these constants in applied work would also allow for a clearer hierarchy of condition strengths, and allow us to verify if conditions for other implications are met, as discussed after Theorems 19 and 21.

Similarly, identifying non-trivial practical examples for each of these classes seems

to be a very open problem. We have identified toy problems to illustrate how certain conditions and constants behave, but these are somewhat contrived and a large weakness of this work is that we only show these for a select few conditions in section 3. Specific use cases have been found in the literature for many and discussed throughout sections 3 and 4. What is not clear is whether they are the *strongest* condition to be applicable in that scenario. The selection of stochastic conditions also sees far fewer practical examples, due to the dependence on the dataset. For example, [67] proves that the squared hinge loss satisfies the (**SGC**) under assumptions on the dataset, primarily, linear separability. For these stochastic conditions, not only does the functional form need to be identified, but the data too, meaning each condition needs a combination of factors to be narrowed down in order to find examples. This makes the task far harder. Further work needs be done here, however this work hopes to convince readers that having this framework of implications set up allows such discoveries in this area to lead to theoretical results very swiftly.

In terms of finding real problems satisfying certain conditions, it may make sense to investigate more general conditions to provide theory for more problems. In this work, out of the strong-convexity generalisations we consider only (**QG**) fails to imply (**IN**). We made the point in section 3 that this does not invalidate the work here, as often it is only small portions of the function to optimise that need to satisfy these constraints. However, more work should be done finding structured conditions not satisfying (**IN**) to speak to the general setting. These could come under the umbrella term often used in the literature of error bounds, not to be confused with our definition [6]. However, considering that (**QG**) required additional assumptions to guarantee convergence, results for these general conditions may be very weak, require moving to different forms of convergence, or, as has been shown in some papers, use a different style of gradient descent method [17]. Alternatively, combined combination proofs similar to 9 and 10 could be utilised.

As for a general area specifically oriented towards statistics, we suggest (**WPL**) as a condition to look into. Both (**PL**) and (**QC**) imply it directly (one step above in figure 5), and both have recently found applications in modern deep learning [74], [13], [75], [77]. A valid criticism of these works, particularly [74], is that simplifications to the neural networks in question (*linear* transformer) are not practical. As a weaker condition then, (**WPL**) may hold for even more general neural network architectures and loss functions, generalising work done in the field to practical real-world examples, in statistics and beyond.

References

[1] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *Information Technology Convergence and Services*, 2014.

[2] Adarsh Barik, Suvrit Sra, and Jean Honorio. Invex programs: First order algorithms and their convergence. *arXiv preprint arXiv:2307.04456*, 2023.

[3] D. Bertsekas. *Convex Optimization Theory*. Athena Scientific optimization and computation series. Athena Scientific, 2009.

[4] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.

[5] Avrim Blum, John Hopcroft, and Ravi Kannan. *Foundations of Data Science*. Cambridge University Press, 2017.

[6] Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

[7] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3–4):231–357, 2015.

[8] Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[9] A. Cauchy. General method for solving systems of simultaneous equations. *Proceedings of the Academy of Sciences*, 25:536–538, 1847.

[10] Yuejie Chi, Yue M. Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

[11] Dominik Csiba and Peter Richtárik. Global convergence of arbitrary-block gradient methods for generalized polyak-łojasiewicz functions. *arXiv preprint arXiv:1709.03014*, 2017.

[12] Marina Danilova, Pavel E. Dvurechensky, Alexander V. Gasnikov, Eduard Gorbunov, Sergey Guminov, Dmitry Kamzolov, and Innokentiy Shibaev. Recent theoretical advances in non-convex optimization. *CoRR*, 2020.

[13] Spencer Frei and Quanquan Gu. Proxy convexity: A unified framework for the analysis of neural networks trained by gradient descent. In *Neural Information Processing Systems*, 2021.

[14] Qiang Fu, Dongchu Xu, and Ashia C. Wilson. Accelerated stochastic optimization methods under quasar-convexity. In *International Conference on Machine Learning*, 2023.

[15] Guillaume Garrigos and Robert M. Gower. Handbook of Convergence Theorems for (Stochastic) Gradient Methods. *arXiv preprint arXiv:2301.11235*, 2023.

[16] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points - online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 797–842. JMLR.org, 2015.

[17] Pinghua Gong and Jieping Ye. Linear convergence of variance-reduced stochastic gradient without strong convexity. *arXiv preprint, arXiv:1406.1102*, 2015.

[18] Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 680–690. PMLR, 2020.

[19] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5200–5209. PMLR, 2019.

[20] Robert Mansel Gower, Othmane Sebbouh, and Nicolas Loizou. Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[21] Charles Guille-Escuret, Baptiste Goujaud, Manuela Girotti, and Ioannis Mitliagkas. A study of condition numbers for first-order optimization. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[22] Sergey Guminov, Alexander Gasnikov, and Ilya Kuruzov. Accelerated methods for weakly-quasi-convex optimization problems. *Computational Management Science*, 20(1):1–19, 2023.

[23] David H. Gutman and Javier F. Peña. The condition number of a function relative to a set. *Mathematical Programming*, 188(1):255–294, 2021.

[24] Morgan A Hanson. On sufficiency of the kuhn-tucker conditions. *Journal of Mathematical Analysis and Applications*, 80(2):545–550, 1981.

[25] Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018.

[26] J Hermant, J. F Aujol, C Dossal, and A Rondepierre. Study of the behaviour of nesterov accelerated gradient in a non convex setting: the strongly quasar convex case. *arXiv preprint arXiv:2405.19809*, 2024.

[27] Oliver Hinder, Aaron Sidford, and Nimit Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1894–1938. PMLR, 2020.

[28] Kishore Jaganathan, Yonina C. Eldar, and Babak Hassibi. *Phase Retrieval:*

An Overview of Recent Developments, pages 264–292. CRC Press, 1st edition, 2016.

[29] Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *Foundations and Trends in Machine Learning*, 10(3–4):142–336, 2017.

[30] Arnulf Jentzen and Philippe von Wurstemberger. Lower error bounds for the stochastic gradient descent optimization algorithm: Sharp convergence rates for slowly and fast decaying learning rates. *J. Complex.*, 57:101438, 2018.

[31] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1724–1732. PMLR, 2017.

[32] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, pages 795–811, Cham, 2016. Springer International Publishing.

[33] Ahmed Khaled and Peter Richtárik. Better theory for sgd in the nonconvex world. *arXiv preprint, arXiv:2002.03329*, 2020.

[34] Ahmed Khaled, Othmane Sebbouh, Nicolas Loizou, Robert M. Gower, and Peter Richtárik. Unified analysis of stochastic gradient methods for composite convex and smooth optimization. *Journal of Optimization Theory and Applications*, 199(2):499–540, 2023.

[35] Jungbin Kim and Insoon Yang. Unifying Nesterov's accelerated gradient methods for convex and strongly convex objective functions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 16897–16954. PMLR, 2023.

[36] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*. ICLR, 2015.

[37] Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does SGD escape local minima? In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2698–2707. PMLR, 2018.

[38] Ming-Jun Lai and Wotao Yin. Augmented ℓ_1 and nuclear-norm models with a globally linearly convergent algorithm. *SIAM Journal on Imaging Sciences*, 6(2):1059–1091, 2013.

[39] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[40] Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via scsg methods. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[41] Yunwen Lei, Ting Hu, Guiying Li, and Ke Tang. Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE Transactions on Neural Networks and Learning Systems*, 31(10):4394–4400, 2020.

[42] Feng-Yi Liao, Lijun Ding, and Yang Zheng. Error bounds, PL condition, and quadratic growth for weakly convex functions, and linear convergences of proximal point methods. In *Proceedings of the 6th Annual Learning for Dynamics and Control Conference*, volume 242 of *Proceedings of Machine Learning Research*, pages 993–1005. PMLR, 2024.

[43] Chaoyue Liu, Dmitriy Drusvyatskiy, Yian Ma, Damek Davis, and Mikhail Belkin. Aiming towards the minimizers: fast convergence of sgd for overparameterized problems. *arXiv preprint arXiv:2306.02601*, 2023.

[44] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022. Special Issue on Harmonic Analysis and Machine Learning.

[45] Ji Liu, Steve Wright, Christopher Re, Victor Bittorf, and Srikrishna Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 469–477, Bejing, China, 2014. PMLR.

[46] Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. In *Advances in Neural Information Processing Systems*, volume 33, pages 18261–18271. Curran Associates, Inc., 2020.

[47] Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. In *AISTATS*, volume 130 of *Proceedings of Machine Learning Research*, pages 1306–1314. PMLR, 2021.

[48] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

[49] Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3325–3334. PMLR, 2018.

[50] Aaron Mishkin. *Interpolation, growth conditions, and stochastic gradient descent*. PhD thesis, University of British Columbia, 2020.

[51] Shashi Kant Mishra and G. Giorgi. *Invexity and Optimization*. Number v. 88 in Nonconvex optimization and its applications. Springer, Berlin, 2008. OCLC: ocn213114192.

[52] Kevin P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.], 2013.

[53] I. Necoara, Yu. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1):69–107, 2019.

[54] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[55] Lam Nguyen, Phuong Ha Nguyen, Marten van Dijk, Peter Richtárik, Katya Scheinberg, and Martin Takac. Sgd and hogwild! convergence without the bounded gradients assumption. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3750–3758. PMLR, 2018.

[56] B.T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

[57] B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[58] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1571–1578, Madison, WI, USA, 2012. Omnipress.

[59] Quentin Rebjock and Nicolas Boumal. Fast convergence to non-isolated minima: four equivalent conditions for C^2 functions. *arXiv preprint arXiv:2303.00096*, 2024.

[60] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951.

[61] R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970.

[62] Karthik A. Sankararaman, Soham De, Zheng Xu, W. Ronny Huang, and Tom Goldstein. The impact of neural network overparameterization on gradient confusion and stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[63] Mark Schmidt and Nicolas Le Roux. Fast Convergence of Stochastic Gradient Descent under a Strong Growth Condition. *arXiv preprint arXiv:1308.6370*, 2013.

[64] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: primal estimated sub-gradient solver for svm. *Mathematical Programming*, 127(1):3–30, 2011.

[65] M. V. Solodov. Incremental gradient algorithms with stepsizes bounded away from zero. *Computational Optimization and Applications*, 11(1):23–35, 1998.

[66] Quoc Tran-Dinh and Marten van Dijk. Gradient descent-type methods: Background and simple unified convergence analysis. *arXiv preprint arXiv:2212.09413*, 2022.

[67] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1195–1204. PMLR, 2019.

[68] Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[69] Jun-Kun Wang and Andre Wibisono. Continuized acceleration for quasar convex functions in non-convex optimization. *arXiv preprint arXiv:2302.07851*, 2023.

[70] John Wright and Yi Ma. *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. Cambridge University Press, 2022.

[71] Kasra Yazdani and Matthew Hale. Asynchronous parallel nonconvex optimization under the polyak-Łojasiewicz condition. *IEEE Control Systems Letters*, 6:524–529, 2022.

[72] Zhuoning Yuan, Yan Yan, Rong Jin, and Tianbao Yang. Stagewise training accelerates convergence of testing error over sgd. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[73] Hui Zhang and Wotao Yin. Gradient methods for convex minimization: better rates under weaker conditions. *arXiv preprint arXiv:1303.4645*, 2013.

[74] Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

[75] Y. Zhang, X. Huang, and Z. Zhang. Prise: Demystifying deep lucas-kanade with strongly star-convex constraints for multimodel image alignment. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13187–13197, Los Alamitos, CA, USA, 2023. IEEE Computer Society.

[76] Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[77] Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh.

SGD converges to global minimum in deep learning via star-convex path. In *7th International Conference on Learning Representations, ICLR*, 2019.

[78] Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Stephen Boyd, and Peter W Glynn. Stochastic mirror descent in variationally coherent optimization problems. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[79] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, Dean P Foster, and Sham Kakade. The benefits of implicit regularization from sgd in least squares problems. In *Advances in Neural Information Processing Systems*, volume 34, pages 5456–5468. Curran Associates, Inc., 2021.

Appendix

A Reference List

(**IP**) Interpolation $f_i(x^*) = \min_x f_i(x)$ for (**FS**) $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$.

(**C**) Convexity - $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle$.

(***C**) Star-Convexity - $f^* = f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x\rangle$.

(***SC** : μ) Star Strong-Convexity - $f^* = f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x\rangle + +\frac{\mu}{2}\|x - x^*\|$.

(**IN**) Invexity - $f(y) \geq f(x) + \langle \nabla f(x), \phi(x, y)\rangle$.

(**SC** : μ) Strong-Convexity - $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2$.

(**ESC** : μ) Essential Strong-Convexity - $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2$, where $y_p = x_p$ for solution set $\mathcal{X}^*$.

(**WSC** : μ) Weak Strong-Convexity - $f^* = f(x_p) \geq f(x) + \langle \nabla f(x), x_p - x\rangle + \frac{\mu}{2}\|x_p - x\|^2$.

(**QC** : γ) Quasar-Convexity $f^* = f(x^*) \geq f(x) + \frac{1}{\gamma}\langle \nabla f(x), x^* - x\rangle$.

(**SQC** : γ, μ) Strong Quasar-Convexity - $f^* = f(x^*) \geq f(x) + \frac{1}{\gamma}\langle \nabla f(x), x^* - x\rangle + \frac{\mu}{2}\|x - x^*\|$.

(**S** : L) L-Smoothness - $f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2$.

(**BG** : C) Bounded Gradients - $\mathbb{E}_i\left[\|\nabla f_i(x)\|^2\right] \leq C$.

(**MSGC** : B) Maximal Strong Growth Condition - $\|\nabla f_i(x)\|^2 \leq B\|\nabla f(x)\|^2$.

(**SGC** : B) Strong Growth Condition - $\mathbb{E}_i\|\nabla f_i(x)\|^2 \leq B\|\nabla f(x)\|^2$.

(**WGC** : A) Weak Growth Condition - If (**S** : L), $\mathbb{E}_i\|\nabla f_i(x)\|^2 \leq A(f(x) - f^*)$.

(**RSI** : μ) Restricted Secant Inequality - $\langle \nabla f(x), x - x_p\rangle \geq \mu\|x - x_p\|^2$.

(**EB** : μ) Error Bound - $\|\nabla f(x)\| \geq \mu\|x_p - x\|$.

(**PL** : μ) Polyak-Łojasiewicz - $\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*)$.

(**WPL** : μ) Weak Polyak-Łojasiewicz - $\sqrt{\mu}(f(x) - f^*) \leq \|\nabla f(x)\|\|x - x^*\|$.

(**QG** : μ) Quadratic Growth - $f(x) - f^* \geq \frac{\mu}{2}\|x_p - x\|^2$.

(**RG** : B, C) Relaxed Growth - $\mathbb{E}_i\|\nabla f_i(x)\|^2 \leq B\|\nabla f(x)\|^2 + C$.

(**ABC** : A, B, C) ABC - $\mathbb{E}_i\|\nabla f_i(x)\|^2 \leq A(f(x) - f^*) + B\|\nabla f(x)\|^2 + C$.

(**GC** : ρ) Gradient Confusion - If (**FS**), $\langle \nabla f_i(x), \nabla f_j(x)\rangle \geq -\rho$ for $i \neq j$.

(**ES** : A) Expected smoothness - $\mathbb{E}_i\left[\|\nabla f_i(x) - \nabla f_i(x^*)\|^2\right] \leq A(f(x) - f(x^*))$.

(**ER** : A) Expected residual-$\mathbb{E}_i\left[\|(\nabla f_i(x) - \nabla f_i(x^*)) - (\nabla f(x) - \nabla f^*)\|^2\right] \leq A(f(x) - f(x^*))$.

(**SS**) Sure smoothness - For $f_i(x) \geq 0$, we have $\|\nabla f_i(x) - \nabla f({}_i y)\| \leq L\|x - y\|$.

(**UAAC**: a) Uniform Acute Angle Condition - $1 \geq \frac{\langle \nabla f(x), x - x^*\rangle}{\|\nabla f(x)\|\|x - x^*\|} \geq a > 0$.

(**RC**: α, β) Regularity condition - $\langle \nabla f(x), x - x^* \rangle \geq \alpha \|\nabla f(x)\| + \beta \|x - x^*\|$.

(**VC**) Variational Coherence - $\langle \nabla f(x), x - x^* \rangle \geq 0$.

B Missing Proofs

B.1 Miscellaneous Results

Lemma 2 - *Proof.*
Simply using the definitions for (**SC**),

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2 \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu'}{2}\|y - x\|^2.$$

The proof for (**S**) is identical. □

Lemma 5 - *Proof.*
The proof is simple using the original definition for (**S**: L) and the triangle inequality. Expanding with the finite sum structure gives,

$$\|\nabla f(x) - \nabla f(y)\| = \|\frac{1}{n}\sum_{i=1}^{n} (\nabla f_i(x) - \nabla f_i(y)) \| \leq \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x) - \nabla f_i(y)\|.$$

Here, we can use the assumption of each f_i satisfying (**S**: L_i), giving,

$$\|\nabla f(x) - \nabla f(y)\| \leq \frac{1}{n}\sum_{i=1}^{n} L_i \|x - y\| = L_{avg}\|x - y\|.$$

This is the result. □

B.2 Implication Proofs

Theorem 1 - *Proof.*
We will only prove the upper bound in the forward implication that we use throughout, but the lower bound and *if and only if* proof can be found in [26]. We follow [15] precisely as this is a well known proof. To begin we consider an auxiliary function in t, $f(x + t(y - x))$, to then invoke the Fundamental Theorem of Calculus on it, which gives

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

Adding and subtracting $\langle \nabla f(x), y - x \rangle$ on the RHS gives,

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

and Cauchy-Schwartz leads to,

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

This allows us to utilise (**S**: L), giving:

$$f(y) - f(x) \leq \int_0^1 L\|x + t(y-x) - x\|\|y - x\| \, dt + \langle \nabla f(x), y - x \rangle.$$

Simplifying then,

$$f(y) - f(x) \leq \int_0^1 Lt\|y - x\|^2 \, dt + \langle \nabla f(x), y - x \rangle = \frac{L}{2}\|y - x\|^2 + \langle \nabla f(x), y - x \rangle.$$

Rearranging gives the result.

For (**S**: L) $\implies$ (**WS**: L), we use the same argument as [20] A.2. The result hinges on a clever choice of $y = x - \frac{1}{L}\nabla f(x)$, just like the Descent Lemma, and we will apply the same simplification. Using the above result for the upper bound from (**S**: L), this then gives,

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

Further, by definition, $f(y) \geq f^*$, so rearranging gives,

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

Multiplying by $-2L$ gives the result. □

Theorem 2 - *Proof.*
These proofs are from [32] and are replicated fairly closely.
(**SC**: μ) $\implies$ (**ESC**: μ) :
Trivial, (**SC**: μ) holds for all inputs so holds in particular for points with the same projection.

(**ESC**: μ) $\implies$ (**WSC**: μ) :
Again trivial as we simply substitute in $y = x_p$. Clearly $y_p = (x_p)_p = x_p$, so by assuming (**ESC**: μ) we immediately get (**WSC**: μ).

(**PL**: μ) $\implies$ (**QG**: μ) :
We skip forward before proving the next step, as this is a required result. Unfortunately it is a little long winded. There are simplifications, however they all require additional assumptions. As this is a key result, we replicate the proof in its fullest generality.

First, define an auxiliary function $g(x) := \sqrt{f(x) - f^*}$, noting it is well defined as $f(x) \geq f^*$. This means showing (**QG**: μ) reduces to

$$f(x) - f^* \geq \frac{\mu}{2}\|x_p - x\|^2 \iff g(x) \geq \sqrt{\frac{\mu}{2}\|x_p - x\|^2} = \sqrt{\frac{\mu}{2}}\|x_p - x^*\|.$$

We also have the following useful fact: If (**PL**: μ) holds, then noting $\nabla g(x) = \nabla(f(x) - f^*)^{\frac{1}{2}} = \frac{1}{2}(f(x) - f^*)^{-\frac{1}{2}}\nabla f(x)$, with standard results, we obtain:

$$\|\nabla g(x)\|^2 = \|\frac{1}{2}(f(x) - f^*)^{-\frac{1}{2}}\nabla f(x)\|^2 = \frac{\|\nabla f(x)\|^2}{4(f(x) - f^*)} \geq \frac{\mu}{2},$$

where the last inequality follows from (**PL**: μ). We further have that g inherits (**IN**) from f implied by (**PL**: μ), is positive, and has gradient bounded from below. Thus, any minimisers of f are minimisers of g, and for any optimal x^*, $g(x^*) = 0$.

Now comes a little jump in the nature of the maths. We must consider a generalisation of the discrete gradient descent method, with something called gradient flow. This involves solving the differential equation:

$$\frac{dx(t)}{dt} = -\nabla g(x(t)), \;\; x(0) = x_0.$$

Interpreting this, it is clear to see that we are moving along a path towards the minimum of our function g. As g is (**IN**) and is bounded below, we will eventually find a minimum, which by (**IN**) is guaranteed to be a global minimum. So, we must identify the time T at which this occurs. We can do that by working with line integrals:

$$g(x_0) - g(x_T) = \int_{x_T}^{x_0} \langle \nabla g(x), dx\rangle = -\int_{x_0}^{x_T} \langle \nabla g(x), dx\rangle = -\int_0^T \langle \nabla g(x(t)), \frac{dx(t)}{dt}\rangle dt.$$

The last line sees a reparameterisation into t, with the limits now the beginning ($x(0) = x_0$) and end time ($x(T) = x_T$) that we are looking for. Substituting in our differential equation where $\frac{dx(t)}{dt} = -\nabla g(x(t))$, we obtain:

$$g(x_0) - g(x_T) = \int_0^T \|\nabla g(x(t))\|^2 dt \geq \int_0^T \frac{\mu}{2} dt = \frac{\mu T}{2},$$

where we have used the previously proved (**PL**) result. This shows there is an expression for T that is finite. This means we can define the length of the trajectory of $x(t)$, starting from x_0:

$$\mathcal{L}(x_0) = \int_0^T \|\frac{dx(t)}{dt}\|^2 dt = \int_0^T \|\nabla g(x(t))\|^2 dt \geq \|x_0 - x_p\|.$$

The last inequality follows from a geometric argument. The fastest path into the solution set from the initialisation of x_0 is a straight line to its projection x_p, thus, the length of the trajectory must be larger than the length of this line, that is, $\|x_0 - x_p\|$. Combining that last inequality with the previous series of expressions and the (**PL**) implied $\|\nabla g(x)\|^2 \geq \sqrt{\frac{\mu}{2}}$ provides:

$$g(x_0) - g(x_T) = \int_0^T \|\nabla g(x(t))\| \|\nabla g(x(t))\| dt \geq \sqrt{\frac{\mu}{2}} \int_0^T \|\nabla g(x(t))\| dt \geq \sqrt{\frac{\mu}{2}} \|x_0 - x_p\|.$$

Using that $g(x_T) = 0$, and squaring gives

$$g(x_0)^2 \geq \frac{\mu}{2}\|x_0 - x_p\|^2 \iff f(x_0) - f^* \geq \frac{\mu}{2}\|x_0 - x_p\|^2.$$

As x_0 was arbitrary, we have the result. □

(**PL**: μ) $\implies$ (**EB**: μ) :
Here we use the fact (**PL**: μ) $\implies$ (**QG**: μ). Writing the standard (**PL**: μ) result and applying (**QG**: μ) immediately gives:

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*) \geq \mu\frac{\mu}{2}\|x - x_p\|^2.$$

Taking square roots gives the complete result.

Finally for (**QG** : μ) + (**C**) $\implies$ (**RSI** : $\frac{\mu}{2}$), we have convexity giving

$$f(x_p) \geq f(x) + \langle \nabla f(x), x_p - x\rangle \iff \langle \nabla f(x), x - x_p\rangle \geq f(x) - f^*.$$

Then simply using (**QG** : μ) gives

$$\langle \nabla f(x), x - x_p\rangle \geq \frac{\mu}{2}\|x - x_p\|^2,$$

completing the proof. Transitivity gives the equivalence between, (**RSI**), (**EB**), (**PL**) and (**QG**) in the case of convexity. □

Theorem 3 - *Proof.*
First consider the statement of (**SC**: μ): $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2$. Investigating (**PL**) then, we need to convert this result into one in terms of function values, not norms of the vectors. Thus, instead of substituting in certain optimal values as we previously would, we leverage the differentiability of f to find the minimum of the LHS and RHS: if the inequality holds for all x and y, it will hold if we minimise over y too.

For the LHS, this is trivial, $\min_y f(y) = f^*$. For the RHS we can differentiate using the standard results $\frac{\partial}{\partial y}\langle c, y\rangle = c$ and $\frac{\partial}{\partial y}\|y - x\|^2 = 2(y - x)$, then set to 0 giving the following:

$$\frac{\partial}{\partial y}\left[f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2\right] = 0 \iff \nabla f(x) + \mu(y - x) = 0.$$

This gives the minimum at $y = x - \frac{1}{\mu}\nabla f(x)$. Substituting this and the LHS result back into Strong-Convexity, we obtain

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

$$= f(x) + \langle \nabla f(x), -\frac{1}{\mu}\nabla f(x)\rangle + \frac{\mu}{2}\|-\frac{1}{\mu}\nabla f(x)\|^2 = f(x) - \frac{1}{\mu}\|\nabla f(x)\|^2 + \frac{\mu}{2\mu^2}\|\nabla f(x)\|^2$$

$$= f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2 \iff \frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*).$$

This gives the desired result with the exact same constant. □

Theorem 6 - *Proof.*
For the first line, the results are trivial, simply specifying $y = x^*$ for (***SC**: μ) and $\mu = 0$ for (**C**). Similarly, (***C**) follows from (**C**) with $y = x^*$ and (***SC**: μ) by noting $\|x - x^*\|^2 \geq 0$. The final requires a little more subtlety as it isn't immediately obvious that $\langle \nabla f(x), x^* - x\rangle \leq 0$, if it is, then as $\frac{1}{\gamma} \geq 1$, the proof follows. This condition would be (**VC**) if it held for all x^*, or if there were a unique minimiser. Note that if (***C**) holds, we have $f^* \geq f(x) + \langle \nabla f(x), x^* - x\rangle$. Moving $f(x)$ over, we then note $f^* \leq f(x)$, thus, $f^* - f(x) \leq 0$. Combining gives:

$$0 \geq f^* - f(x) \geq \langle \nabla f(x), x^* - x\rangle,$$

so $\langle \nabla f(x), x^* - x\rangle$ is negative. The result then follows.

For the second set then, from the previous line we have that

$$(\textbf{*SC} : \mu) \implies (\textbf{*C}) \implies \langle \nabla f(x), x^* - x\rangle \leq 0.$$

Thus, by the exact same reasoning above, $(\textbf{*SC} : \mu) \implies (\textbf{SQC} : \gamma, \mu)$. The final implication is also trivial from $\|x - x^*\|^2 \geq 0$. □

Theorem 7 - *Proof.*
These proofs are replicated from [11] [22]. For $(\textbf{PL} : \mu) \implies (\textbf{WPL} : \frac{4\mu}{L})$, we use (**S**: L) along with (**PL**: μ). We begin with the useful term,

$$\frac{\|\nabla f(x)\|^2\|x - x^*\|^2}{(f(x) - f^*)^2} \geq \frac{2\mu(f(x) - f^*)\|x - x^*\|^2}{(f(x) - f^*)^2} \geq \frac{2\mu\|x - x^*\|^2}{(f(x) - f^*)},$$

where we have used (**PL**) first then simplified. Next, using (**S**: L), substituting in (x, x^*) for,

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

$$\implies \|x - x^*\|^2 \geq \frac{2}{L}(f(x) - f^*).$$

Here we have used that $\nabla f(x^*) = 0$ as this is a minimum, then rearranged. To finish, we combine for:

$$\frac{\|\nabla f(x)\|^2\|x - x^*\|^2}{(f(x) - f^*)^2} \geq \frac{2\mu\|x - x^*\|^2}{(f(x) - f^*)} \geq \frac{4\mu}{L}.$$

Taking sqaure roots of both sides and rearranging gives the result of ($\mathbf{WPL}: \frac{4\mu}{L}$),

$$\sqrt{\frac{\|\nabla f(x)\|^2\|x-x^*\|^2}{(f(x)-f^*)^2}} \geq \sqrt{\frac{4\mu}{L}} \implies \sqrt{\frac{4\mu}{L}}(f(x)-f^*) \leq \|\nabla f(x)\|\|x-x^*\|.$$

For ($\mathbf{QC}: \gamma$) $\implies$ ($\mathbf{WPL}: \gamma^2$), the proof is immediate from Cauchy-Schwarz. Recalling the definition of Quasar-Convex for $\gamma \in (0,1]$:

$$f(x^*) \geq f(x)+\frac{1}{\gamma}\langle \nabla f(x), x^*-x\rangle \implies f(x)-f^* \leq \frac{1}{\gamma}\langle \nabla f(x), x-x^*\rangle \leq \frac{1}{\gamma}\|\nabla f(x)\|\|x-x^*\|.$$

Then simply move γ over for ($\mathbf{WPL}: \gamma^2$). □

Theorem 9 - *Proof.*
For ($\mathbf{UAAC}: a$) + ($\mathbf{WPL}: \mu$) $\implies$ ($\mathbf{QC}: \frac{\sqrt{\mu}}{a}$), we slightly modify the proof from [11] for our more general setting. Manipulating ($\mathbf{UAAC}$: a) leads to $\|\nabla f(x)\|\|x-x^*\| \leq a\langle \nabla f(x), x-x^*\rangle$. So then, from the definition of ($\mathbf{WPL}$: μ) and dividing by $\sqrt{\mu}$ we have the sequence,

$$(f(x)-f^*) \leq \frac{1}{\sqrt{\mu}}\|\nabla f(x)\|\|x-x^*\| \leq \frac{a}{\sqrt{\mu}}\langle \nabla f(x), x-x^*\rangle.$$

Re-arranging gives the result.

For ($\mathbf{UAAC}: a$) + ($\mathbf{EB}: \mu$) $\implies$ ($\mathbf{RSI}: a\mu$), we take the simple proof from [26]. As we have a unique minimiser, $x^* = x_p$, so ($\mathbf{EB}$: μ) reads $\mu\|x-x^*\| \leq \|\nabla f(x)\|$. Rearranging ($\mathbf{UAAC}$: a) gives $a\|x-x^*\|\|\nabla f(x)\| \leq \langle \nabla f(x), x-x^*\rangle$. Combining gives

$$a\mu\|x-x^*\|^2 = a\mu\|x-x^*\|\|x-x^*\| \leq a\|x-x^*\|\|\nabla f(x)\| \leq \langle \nabla f(x), x-x^*\rangle.$$

This gives the result. □

Theorem 11 - *Proof.*
The first step is noting that ($\mathbf{SQC}: \gamma, \mu$) ($\mathbf{RC}, \alpha, \beta$) imply there is a unique minimiser, from Lemma 4. Thus, we can proceed by proving all the given implications under the assumption of the existence of a unique minimiser.

For ($\mathbf{SQC}: \gamma, \mu$) $\implies$ ($\mathbf{VC}$), we note that as the proof of Theorem 6 included the implication ($\mathbf{QC}: \gamma$) $\implies$ ($\mathbf{VC}$), with ($\mathbf{SQC}: \gamma, \mu$) the stronger condition, these two are done. For ($\mathbf{RC}, \alpha, \beta$), simply writing the definition,

$$\langle \nabla f(x), x-x^*\rangle \geq \alpha\|\nabla f(x)\|^2 + \beta\|x-x^*\|^2 \geq 0.$$

For ($\mathbf{UAAC}: a$), the definition gives

$$\frac{\langle \nabla f(x), x-x^*\rangle}{\|\nabla f(x)\|\|x-x^*\|} \geq a \implies \langle \nabla f(x), x-x^*\rangle \geq a\|\nabla f(x)\|\|x-x^*\| \geq 0,$$

as norms are positive. This is (**VC**).

Finally, for (**RSI** : μ) $\implies$ (**VC**), we simply write out the definition,

$$\langle \nabla f(x), x - x^* \rangle \geq \mu \|x - x^*\|^2 \geq 0.$$

This completes the proof. □

Theorem 15 - *Proof.*
These proofs are either trivial and have been phrased in our notation for convenience, or from [33]. The first line of implications simply follows from the fact that, firstly, $C \geq 0$, and similarly, $A(f(x) - f^*) \geq 0$. The second line is almost identical, leveraging that $B\|\nabla f(x)\|^2 \geq 0$.

For (**GC** : ρ) $\implies$ (**RG** : $n, \rho(n-1)$), we use the implied (**FS**) and expand $\|\nabla f(x)\|^2$:

$$\begin{aligned}
\|\nabla f(x)\|^2 &= \langle \nabla f(x), \nabla f(x) \rangle \\
&= \langle \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x), \frac{1}{n} \sum_{j=1}^{n} \nabla f_j(x) \rangle \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \langle \nabla f_i(x), \nabla f_j(x) \rangle \right) \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \left(\langle \nabla f_i(x), \nabla f_i(x) \rangle + \sum_{j \neq i} \langle \nabla f_i(x), \nabla f_j(x) \rangle \right) \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \|\nabla f_i(x)\|^2 + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \langle \nabla f_i(x), \nabla f_j(x) \rangle.
\end{aligned}$$

For the first term, we leverage the fact that the index i is chosen uniformly at random from the n choices, thus, $\mathbb{E}_i \left[\|\nabla f_i(x)\|^2 \right] = \sum_{i=1}^{n} \frac{1}{n} \|\nabla f_i(x)\|^2$. The second term can be simplified with (**GC** : ρ) giving $\langle \nabla f_i(x), \nabla f_j(x) \rangle \geq -\rho$. Substituting these in leads to,

$$\begin{aligned}
\|\nabla f(x)\|^2 &= \frac{1}{n^2} \sum_{i=1}^{n} \|\nabla f_i(x)\|^2 + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \langle \nabla f_i(x), \nabla f_j(x) \rangle \\
&\geq \frac{1}{n} \mathbb{E}_i \left[\|\nabla f_i(x)\|^2 \right] - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \rho. \\
&= \frac{1}{n} \mathbb{E}_i \left[\|\nabla f_i(x)\|^2 \right] - \frac{\rho}{n^2} \sum_{i=1}^{n} |\{j : j \neq i\}| \\
&= \frac{1}{n} \mathbb{E}_i \left[\|\nabla f_i(x)\|^2 \right] - \frac{\rho n(n-1)}{n^2}.
\end{aligned}$$

Rearranging gives,

$$\mathbb{E}_i \left[\|\nabla f_i(x)\|^2 \right] \leq n \|\nabla f(x)\|^2 + \rho(n-1),$$

which is the result.

For (**SS** : L) $\implies$ (**ABC** : $2L, 0, 2Lf^*$), we first notice that (**SS** : L) implies that each f_i satisfies (**S** : L). As such, by Theorem 1, they also satisfy (**WS** : L), giving $\|\nabla f_i(x)\|^2 \leq 2L(f_i(x) - f_i(x^*)) \leq 2Lf_i(x)$, where we have used the positivity condition of (**SS** : L) to give $-f_i(x^*) \leq 0$. A simple decomposition and taking expectations with unbiasedness then gives the result,

$$\mathbb{E}_i\left[\|\nabla f_i(x)\|^2\right] \leq 2L\mathbb{E}_i\left[f_i(x)\right] = 2L(\mathbb{E}_i\left[f_i(x)\right] - f^*) + 2Lf^* = 2L(f(x) - f^*) + 2Lf^*.$$

This completes the proof. □

B.3 Convergence Proofs

Theorem 19 - *Proof.*
A key quantity, often referred to as a Lyapunov energy in many proofs, defined here as $E_k = f(x_k) - f^* + \frac{\mu}{2}\|x_k - x^*\|^2$, for x^* the minimiser of f, and the point we refer to in the definition of (**QC**). By performing induction on this sequence, we will see the result emerge.

We begin by simply taking differences of terms:

$$E_{k+1} - E_k = (f(x_{k+1}) - f(x_k)) + \frac{\mu}{2}\|x_{k+1} - x^*\|^2 - \frac{\mu}{2}\|x_k - x^*\|^2.$$

To begin cancelling some terms, we can use the GD update of $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$, then expand just $\|x_{k+1} - x^*\|^2$:

$$\|x_{k+1} - x^*\|^2 = \|[x_k - x^*] - \frac{1}{L}\nabla f(x_k)\|^2 = \|x_k - x^*\|^2 + \frac{1}{L^2}\|\nabla f(x_k)\|^2 - \frac{2}{L}\langle x_k - x^*, \nabla f(x_k)\rangle.$$

Back into the difference, some terms cancel to give:

$$E_{k+1} - E_k = (f(x_{k+1}) - f(x_k)) - \frac{\mu}{L}\langle x_k - x^*, \nabla f(x_k)\rangle + \frac{\mu}{2L^2}\|\nabla f(x_k)\|^2.$$

To simplify further, we use (**S**: L), as we did in the previous proof to obtain $f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L}\|\nabla f(x_k)\|^2$. Using this then, we can remove the difference in function values at the beginning of the energy difference term:

$$E_{k+1} - E_k \leq -\frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{\mu}{L}\langle x^* - x_k, \nabla f(x_k)\rangle + \frac{\mu}{2L^2}\|\nabla f(x_k)\|^2.$$

The problematic term is now the inner product, however, we can leverage the (**SQC**: γ, μ) assumption. By symmetry of the inner product in argument, and multiplying through by gamma with some re-arranging the definition reads

$$\langle x^* - x_k, \nabla f(x_k)\rangle \leq \gamma(f^* - f(x_k)) - \gamma\frac{\mu}{2}\|x^* - x_k\|^2.$$

This can then be substituted in to achieve:

$$E_{k+1} - E_k \leq -\frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{\mu}{L}(\gamma(f^* - f(x_k)) - \gamma\frac{\mu}{2}\|x^* - x_k\|^2) + \frac{\mu}{2L^2}\|\nabla f(x_k)\|^2.$$

$$= \frac{1}{2L}(\frac{\mu}{L} - 1)\|\nabla f(x_k)\|^2 - \frac{\gamma\mu}{L}(f(x_k) - f^* + \frac{\mu}{2}\|x_k - x^*\|^2).$$

Now note as $\mu \leq L$, from the assumption in section 2 we have,

$$\frac{\mu}{L} \leq \frac{L}{L} = 1 \implies \frac{\mu}{L} - 1 \leq 0.$$

Thus, the first term is negative, giving

$$E_{k+1} - E_k \leq -\frac{\gamma\mu}{L}(f(x_k) - f^* + \frac{\mu}{2}\|x_k - x^*\|^2) = -\frac{\gamma\mu}{L}E_k,$$

where we have substituted in the definition of E_k. This in turn then implies $E_{k+1} \leq \left(1 - \frac{\gamma\mu}{L}\right) E_k$. Starting from E_k and iterating, induction gives

$$E_k \leq \left(1 - \frac{\gamma\mu}{L}\right)^k E_0 \iff f(x_k) - f^* \leq (1 - \frac{\gamma\mu}{L})^k (f(x_0) - f^* + \frac{\mu}{2}\|x_0 - x^*\|^2) - \frac{\mu}{2}\|x_k - x^*\|^2.$$

Trivially cancelling the final term with another inequality (as it is clearly negative), we can then use, in order to simplify and show the dependence on γ more, that (**QG**: $\frac{\mu\gamma}{2L}$) is implied here. This is due to 2 and 8:

$$\begin{aligned}
(\mathbf{SQC} : \gamma, \mu) &\implies (\mathbf{RC} : \frac{\gamma}{2L}, \frac{\mu\gamma}{2}) \\
&\implies (\mathbf{RSI} : \frac{\mu\gamma}{2}) \\
&\implies (\mathbf{EB} : \frac{\mu\gamma}{2}) \\
&\implies (\mathbf{PL} : \frac{\mu\gamma}{2L}) \\
&\implies (\mathbf{QG} : \frac{\mu\gamma}{2L}).
\end{aligned}$$

Note that the referenced paper where this proof is found obtains different constants and hence a different result. However, for consistency with the rest of the paper, we present the result with constants implied wholly by the work found in this review paper.

We can then simplify, using (**QG**: $\frac{\mu\gamma}{2L}$), that is, $\frac{\mu}{2}\|x_0 - x^*\|^2 \leq \frac{2L}{\gamma}(f(x_0) - f^*)$:

$$(f(x_0) - f^* + \frac{\mu}{2}\|x_0 - x^*\|^2) \leq \left(1 + \frac{2L}{\gamma}\right)(f(x_0) - f^*).$$

This then gives the final result. □

Theorem 20 - *Proof.*

For this result we follow [23]. We define a slightly different Lyapunov energy term, $E_k = k(f(x_k) - f^*) + \frac{L}{2}\|x_k - x^*\|^2$ for the unique minimiser x^* and proceed in a similar manner to the last proof, expanding out with $f(x_k)$ and using the definition of (**GD**),

$$\begin{aligned} E_{k+1} &= (k+1)\left(f(x_{k+1}) - f^*\right) + \frac{L}{2}\|x_{k+1} - x^*\|^2 \\ &= (k+1)\left(\left(f(x_{k+1}) - f(x_k)\right) - \left(f^* - f(x_k)\right)\right) \\ &+ \frac{L}{2}\|x_k - \frac{1}{L}\nabla f(x_k) - x^*\|^2. \end{aligned}$$

We can then simplify. Firstly, we have the Descent Lemma 1, and can also expand the norm, giving,

$$\begin{aligned} E_{k+1} &= (k+1)\left(-\frac{1}{2L}\|\nabla f(x_k)\|^2 - \left(f^* - f(x_k)\right)\right) \\ &+ \frac{L}{2}\left(\|x_k - x^*\|^2 + \frac{1}{L^2}\|\nabla f(x_k)\|^2 - \frac{2}{L}\langle \nabla f(x_k), x_k - x^*\rangle\right). \\ &= k(f(x_k) - f^*) + \frac{L}{2}\|x_k - x^*\|^2 \\ &- \left(\frac{(k+1)}{2L} - \frac{1}{2L}\right)\|\nabla f(x_k)\|^2 \\ &+ \left(f(x_k) - f^* - \langle \nabla f(x_k), x_k - x^*\rangle\right) \\ &\leq E_k \\ &+ \left(f(x_k) - f^* - \langle \nabla f(x_k), x_k - x^*\rangle\right). \end{aligned}$$

For the final line we can rearrange the definition of (***C**) for x_k, x^*, which gives us $f^* \geq f(x_k) + \langle \nabla f(x_k), x^* - x_k\rangle \implies f(x_k) - f^* - \langle \nabla f(x_k), x_k - x^*\rangle \leq 0$. Combined with above leaves us with $E_{k+1} \leq E_k$. Using this, we can iterate for $E_k \leq E_0$, then use the definition of $E_0 = \frac{L}{2}\|x_k - x^*\|^2$. Combining gives,

$$k(f(x_k) - f^*) \leq E_k \leq E_0 = \frac{L}{2}\|x_k - x^*\|^2 \implies f(x_k) - f^* \leq \frac{L}{2k}\|x_k - x^*\|^2.$$

This is the result. □

Theorem 24 - *Proof.*
For the second (**BG**) result, we specify the step-size as $\eta_k = \frac{2k+1}{2\mu(k+1)^2}$. Define an auxiliary function $g(k) = k^2(\mathbb{E}_i[f(x_k)] - f^*)$. By multiplying the result proven in the main text, that is,

$$\mathbb{E}_i[f(x_{k+1})] - f^* \leq (1 - 2\mu\eta_k)(f(x_k) - f^*) + \frac{LC\eta_k^2}{2},$$

by $(k+1)^2$ and noting that $\mathbb{E}_i f(x_k) = f(x_k)$ due to conditioning on x_k for this step

to x_{k+1}, we obtain,

$$\begin{aligned}
\mathbb{E}_i[f(x_{k+1})] - f^* &\leq (1 - 2\mu\eta_k)(f(x_k) - f^*) + \frac{LC^2}{2}. \\
\mathbb{E}_i[f(x_{k+1})] - f^* &\leq (1 - 2\mu\frac{2k+1}{2\mu(k+1)^2})(f(x_k) - f^*) + \frac{LC}{2}\left(\frac{2k+1}{2\mu(k+1)^2}\right)^2. \\
&= \frac{k^2}{(k+1)^2}(f(x_k) - f^*) + \frac{LC(2k+1)^2}{8\mu^2(k+1)^4} \iff \\
(k+1)^2(\mathbb{E}_i[f(x_{k+1})] - f^*) &\leq k^2(f(x_k) - f^*) + \frac{LC(2k+1)^2}{8\mu^2(k+1)^2} \iff \\
g(k+1) &\leq g(k) + \frac{LC(2k+1)^2}{8\mu^2(k+1)^2} \leq g(k) + \frac{LC}{2\mu^2},
\end{aligned}$$

simplified as $\left(\frac{2k+1}{k+1}\right)^2 < 2^2 = 4$. Recursion over k gives,

$$g(k+1) \leq g(k) + \frac{LC}{2\mu^2} \leq g(k-1) + \frac{LC}{2\mu^2} + \frac{LC}{2\mu^2} \leq ... \leq g(0) + (k+1)\frac{LC}{2\mu^2}.$$

To finish then, note $g(0) = 0$, then use the definition of g for,

$$g(k+1) = (k+1)^2(\mathbb{E}_i[f(x_{k+1})] - f^*) \leq (k+1)\frac{LC}{2\mu^2} \implies \mathbb{E}_i[f(x_k)] - f^* \leq \frac{LC}{2k\mu^2}.$$

This is the result. □

For the (**GC**) result from [62], again start with the (**S**: L) implied

$$f(x_{k+1}) - f(x_k) \leq -\eta\langle\nabla f(x_k), \nabla f_i(x_k)\rangle + \frac{L\eta^2}{2}\|\nabla f_i(x_k)\|^2,$$

but here rather than taking expectations, we use the(**FS**) implied by (**GC**). As i is chosen uniformly at random from the n possible f_i, we have,

$$\begin{aligned}
f(x_{k+1}) - f(x_k) &\leq -\eta\langle\nabla f(x_k), \nabla f_i(x_k)\rangle + \frac{L\eta^2}{2}\|\nabla f_i(x_k)\|^2 \\
&= -\eta\langle\nabla\left(\frac{1}{n}\sum_{j=1}^{n} f_j(x_k)\right), \nabla f_i(x_k)\rangle + \frac{L\eta^2}{2}\|\nabla f_i(x_k)\|^2 \\
&= -\frac{\eta}{n}\sum_{j\neq i}\langle\nabla f_j(x_k), \nabla f_i(x_k)\rangle + \left(\frac{L\eta^2}{2} - \frac{\eta}{n}\right)\|\nabla f_i(x_k)\|^2,
\end{aligned}$$

where we have separated out the sum and used properties of norms. To deal with the inner product, we then utilise (**GC**: ρ) for,

$$\begin{aligned}
f(x_{k+1}) - f(x_k) &\leq -\frac{\eta}{n}\sum_{j\neq i}(-\rho) + \left(\frac{L\eta^2}{2} - \frac{\eta}{n}\right)\|\nabla f_i(x_k)\|^2 \\
&= \frac{\eta}{n}(n-1)\rho + \left(\frac{L\eta^2}{2} - \frac{\eta}{n}\right)\|\nabla f_i(x_k)\|^2 \\
&\leq \eta\rho + \left(\frac{L\eta^2}{2} - \frac{\eta}{n}\right)\|\nabla f_i(x_k)\|^2.
\end{aligned}$$

The gradient can then be simplified with (**PL**: μ) $\forall i$, for,

$$f(x_{k+1}) - f(x_k) \leq \eta\rho + \left(\frac{L\eta^2}{2} - \frac{\eta}{n}\right)(f_i(x_k) - \min_x f_i(x)).$$

Now we take expectations. For the result to make sense we require that we have $\left(\frac{L\eta^2}{2} - \frac{\eta}{n}\right) < 0 \iff \eta < \frac{2}{nL}$. Further, we note $\min_x f_i(x) = f_i(x_i^*)$ for some optimal x_i^*. Thus, for the full f optimal value x^*, we have $f_i(x_i^*) \leq f_i(x^*)$. Taking expectations then, $\mathbb{E}_i \min_x f_i(x) \leq \mathbb{E}_i f_i(x^*) = f^*$. Applying this, we have,

$$\begin{aligned}\mathbb{E}_i[f(x_{k+1})] - f(x_k) &\leq \eta\rho + \left(\frac{L\eta^2}{2} - \frac{\eta}{n}\right)(f(x_k) - f^*) \\ &= \eta\rho - \left(\frac{\eta}{n} - \frac{L\eta^2}{2}\right)(f(x_k) - f^*).\end{aligned}$$

This allows us to subtract f^* from each side to give,

$$\mathbb{E}_i[f(x_{k+1})] - f^* \leq \eta\rho + r_1(f(x_k) - f^*),$$

for $r_1 = 1 - \frac{2\mu}{n}\left(\eta - \frac{nL\eta^2}{2}\right)$. Iterating gives the same result using the geometric series as in the previous result.

Finally, we have the case for (**ABC**), following [33]. We start in the exact same way, using (**S**: L), taking expectations and subtracting f^* from both sides. This leaves us with,

$$\mathbb{E}_i[f(x_{k+1})] - f^* \leq f(x_k) - f^* - \eta_k\|\nabla f(x_k)\|^2 + \frac{L\eta_k^2}{2}\mathbb{E}_i\|\nabla f_i(x_k)\|^2.$$

Then we leverage (**ABC**: $2A, B, C$) for,

$$\begin{aligned}&\mathbb{E}_i[f(x_{k+1})] - f^* \leq f(x_k) - f^* - \eta_k\|\nabla f(x_k)\|^2 + \frac{L\eta_k^2}{2}\mathbb{E}_i\|\nabla f_i(x_k)\|^2 \\ &\leq f(x_k) - f^* - \eta_k\|\nabla f(x_k)\|^2 + \frac{L\eta_k^2}{2}\left(2A(f(x_k) - f^*) + B\|\nabla f(x_k)\|^2 + C\right) \\ &= (1 + AL\eta_k^2)(f(x_k) - f^*) + \left(\frac{BL\eta_k^2}{2} - \eta_k\right)\|\nabla f(x_k)\|^2 + \frac{L\eta_k^2}{2}C.\end{aligned}$$

Following this, there are some technical details that can be found in [33], that are not overly instructive so we will simply assume the required implications, namely that $1 - \frac{LB\eta_k}{2} \geq \frac{3}{4}$ and $AL\eta_k \leq \frac{\mu}{2}$. The first of these gives, $\left(\frac{BL\eta_k^2}{2} - \eta_k\right) \leq 0$, so that we can apply (**PL**: μ) with the correct inequality direction. The second gives $AL\eta_k^2 \leq \frac{\eta_k\mu}{2}$ which implies $(1 + AL\eta_k^2 - \frac{3\eta_k\mu}{2}) \leq (1 - \eta_k\mu)$. This is then,

$$\begin{aligned}\mathbb{E}_i[f(x_{k+1})] - f^* &\leq (1 + AL\eta_k^2)(f(x_k) - f^*) + \left(\frac{BL\eta_k^2}{2} - \eta_k\right)2\mu(f(x_k) - f^*) + \frac{L\eta_k^2}{2}C. \\ &\leq (1 + AL\eta_k^2 - \frac{3\eta_k\mu}{2})(f(x_k) - f^*) + \frac{L\eta_k^2}{2}C \\ &\leq (1 - \eta_k\mu)(f(x_k) - f^*) + \frac{L\eta_k^2}{2}C.\end{aligned}$$

Referring to [33], if we define $s_k = \mathbb{E}_i[f(x_k)] - f^*$, we can apply their Lemma 3, with $s_{k+1} \leq s_k + \frac{L\eta_k^2}{2}C$ from above. This gives the result and completes the theorem. □

Lemma 8 - *Proof.*
We first restate the conditions for convenience:

$$\beta_k = 1 - s_k\mu\eta \text{ (N1)}, \; s_k = \frac{1}{B}\left(1 + \frac{\beta_k(1-\alpha_k)}{\alpha_k}\right) \text{ (N2)}, \; b_k^2 = \beta_k b_{k+1}^2 \text{ (N3)},$$

$$b_{k+1}^2 s_k^2 \eta B = a_{k+1}^2 \text{ (N4)}, \; \frac{s_k\eta\beta_k(1-\alpha_k)}{\alpha_k} = \frac{a_k^2}{b_{k+1}^2} \text{ (N5)}, \; b_{k+1}^2 s_k\eta - a_{k+1}^2 + a_k^2 = 0 \text{ (N6)}.$$

Beginning for Theorem 25, we see $s_k\mu\eta = \mu\eta\frac{1}{\sqrt{\eta B\mu}} = \sqrt{\frac{\mu\eta}{B}}$, giving (**N1**). Then, note:

$$\beta_k\frac{1-\alpha_k}{\alpha_k} = \beta_k\frac{\frac{a_k^2}{s_k\beta_k b_{k+1}^2\eta + a_k^2}}{\frac{s_k\beta_k b_{k+1}^2\eta}{s_k\beta_k b_{k+1}^2\eta + a_k^2}} = \beta_k\frac{a_k^2}{s_k\beta_k b_{k+1}^2\eta} = \frac{a_k^2}{s_k b_{k+1}^2\eta}.$$

This immediately gives (**N5**) multiplying both sides by $s_k\eta$. To obtain (**N2**) note: $\frac{a_k^2}{b_{k+1}^2} = \frac{\beta_k^{-k}}{\mu\beta_k^{-(k+1)}} = \frac{\beta_k}{\mu}$. Then, substituting in the definitions for β_k and s_k:

$$\frac{1}{B}\left(1 + \frac{\beta_k(1-\alpha_k)}{\alpha_k}\right) = \frac{1}{B}\left(1 + \frac{\beta_k}{\mu}\frac{1}{s_k\eta}\right) = \frac{1}{B}\left(1 + \left(1 - \sqrt{\frac{\mu\eta}{B}}\right)\frac{\sqrt{\eta B\mu}}{\mu\eta}\right)$$

$$= \frac{1}{B}\left(1 + \left(1 - \sqrt{\frac{\mu\eta}{B}}\right)\sqrt{\frac{B}{\mu\eta}}\right) = \frac{1}{B}\left(1 + \sqrt{\frac{B}{\mu\eta}} - 1\right) = \frac{1}{\sqrt{\eta B\mu}} =: s_k.$$

(**N3**) is then clear from the definitions: $\beta_k b_{k+1}^2 = \beta_k\mu\beta_k^{-(k+1)} = \mu\beta_k^{-k} = b_k^2$. For (**N4**), again straight from the definitions:

$$b_{k+1}^2 s_k^2\eta B = b_{k+1}^2\frac{1}{\eta B\mu}\eta B = \frac{b_{k+1}^2}{\mu} = a_{k+1}^2.$$

Finally we prove (**N6**). To begin, first define $C := \frac{\beta_k(1-\alpha_k)}{\alpha_k}$ such that $s_k = \frac{1}{B}(1+C)$, using (**N2**). (**N4**) gives an expression for $a_{k+1}^2 = b_{k+1}^2 s_k^2\eta B$ and (**N5**) gives $a_k^2 = b_{k+1}^2\frac{s_k\eta\beta_k(1-\alpha_k)}{\alpha_k}a_k^2 = b_{k+1}^2 s_k\eta C$. Combining:

$$b_{k+1}^2 s_k\eta - a_{k+1}^2 + a_k^2 = b_{k+1}^2\eta s_k(1 - s_k B + C).$$

Taking just the bracketed term we obtain:

$$1 - s_k B + C = 1 + C - B\frac{1}{B}(1+C) = 0.$$

For Theorem 26:
Note that (**SC**: $\mu = 0$) $\Leftrightarrow$ (**C**), so (**N1**) is trivial. Similarly (**N3**) is trivial as all terms are equal to 1, with (**N4**) following via squaring the definition of a_{k+1},

$a_{k+1}^2 = s_k^2 B\eta$, noting $b_{k+1}^2 = 1$. (**N5**) is equivalent to $a_k^2 = \frac{s_k\eta(1-\alpha_k)}{\alpha_k}$ here. Starting with the definition of α_k,

$$\frac{1-\alpha_k}{\alpha_k} = \frac{\frac{a_k^2}{s_k\eta+a_k^2}}{\frac{s_k\eta}{s_k\eta+a_k^2}} = \frac{a_k^2}{s_k\eta}.$$

Multiplying through with $s_k\eta$ gives the result. For (**N2**) and (**N6**) we need to perform a larger expansion, starting from the definition of s_k:

$$\eta_k = \eta = \frac{1}{BL},\ s_k = \frac{\frac{1}{B} + \sqrt{\frac{1}{B^2} + 4s_{k-1}^2}}{2} \Leftrightarrow 2s_k - \frac{1}{B} = \sqrt{\frac{1}{B^2} + 4s_{k-1}^2}.$$

Squaring both sides and rearranging leads to,

$$4s_k^2 - \frac{4}{B}s_k + \frac{1}{B^2} = \frac{1}{B^2} + 4s_{k-1}^2 \Leftrightarrow Bs_k^2 - s_k = Bs_{k-1}^2 \tag{1}$$

To finish (**N2**) then, recall $\frac{1-\alpha_k}{\alpha_k} = \frac{a_k^2}{s_k\eta}$, such that with (**N4**) giving $a_k^2 = s_{k-1}^2\eta B$:

$$\frac{1}{B}\left(1 + \frac{\beta_k(1-\alpha_k)}{\alpha_k}\right) = \frac{1}{B}\left(1 + \frac{a_k^2}{s_k\eta}\right) = \frac{1}{B}\left(1 + \frac{s_{k-1}^2\eta B}{s_k\eta}\right) = \frac{1}{B}\left(1 + \frac{s_{k-1}^2 B}{s_k}\right).$$

Using (1), we obtain

$$\frac{1}{B}\left(1 + \frac{s_{k-1}^2 B}{s_k}\right) = \frac{1}{B}\left(1 + \frac{Bs_k^2 - s_k}{s_k}\right) = \frac{1}{B}\left(1 + Bs_k - 1\right) = s_k.$$

Finally, (**N6**) follows by using (**N4**) twice to obtain:

$$b_{k+1}^2 s_k\eta - a_{k+1}^2 + a_k^2 = 0 \Leftrightarrow s_k\eta - s_k^2\eta B + s_{k-1}^2\eta B = 0 \Leftrightarrow Bs_k^2 - s_k = Bs_{k-1}^2,$$

which we know to be true by (1). □

Theorem 27 - *Proof.*

We will again be using the same Lyapunov energy idea as in 19 but this proof is notably more precise. As the proof is very finicky, the parameters have been chosen such that we can write the (**NAG**) update terms in a similar form to that of the original paper for clarity. That is, we wish to write the update in terms of just one parameter to avoid confusion with the original, so from $v_{k+1} = \beta_k v_k + (1-\beta_k)\delta_k - s_k\eta_k\nabla f_i(\delta_k)$ to $v_{k+1} = \beta_k v_k + (1-\beta_k)\delta_k - \nu_k\nabla f_i(\delta_k)$. So, define $\nu_k := s_k\eta_k$. Note that $s_k\eta_k$ leads to the same constant as in the original paper, but for simplicity we reduce it to just $\nu_k = s_k\eta_k = \frac{\gamma}{L} = \sqrt{\frac{\eta_k}{\mu}}$. This can also, as it is constant, just be referred to as $\nu = \nu_k$. One change remaining is that $\alpha \to 1-\alpha$ here due to the setup, but the results remain the same.

Taking differences, $E_{k+1} - E_k = (f(x_{k+1}) - f(x_n)) + \frac{\mu}{2}(\|v_{k+1} - x^*\|^2 - \|v_k - x^*\|^2)$. We can then simplify substituting in (**NAG**) values:

$$\|v_{k+1} - x^*\|^2 - \|v_k - x^*\|^2 = \|\beta v_k + (1-\beta)\delta_k - \nu\nabla f(\delta_k) - x^*\|^2 - \|v_k - x^*\|^2$$

$$= \|\beta(v_k - x^*) + \beta x^* + (1-\beta)(\delta_k - x^*) + (1-\beta)x^* - \nu\nabla f(\delta_k) - x^*\|^2 - \|v_k - x^*\|^2$$

$$= \|\beta(v_k - x^*) + (1-\beta)(\delta_k - x^*) - \nu\nabla f(\delta_k) + (\beta + (1-\beta) - 1)x^*\|^2 - \|v_k - x^*\|^2$$

$$= \|[\beta(v_k - x^*)] + [(1-\beta)(\delta_k - x^*) - \nu\nabla f(\delta_k)]\|^2 - \|v_k - x^*\|^2$$

$$= (\beta^2-1)\|v_k-x^*\|^2+\|(1-\beta)(\delta_k-x^*)-\nu\nabla f(\delta_k)\|^2+2\langle\beta(v_k-x^*),(1-\beta)(\delta_k-x^*)-\nu\nabla f(\delta_k)\rangle$$

$$= (\beta^2-1)\|v_k - x^*\|^2 + (1-\beta)^2\|\delta_k - x^*\|^2 + \nu^2\|\nabla f(\delta_k)\|^2$$

$$-2(1-\beta)\nu\langle\delta_k - x^*, \nabla f(\delta_k)\rangle + 2\beta\langle v_k - x^*, (1-\beta)(\delta_k - x^*) - \nu\nabla f(\delta_k)\rangle,$$

using standard properties of inner products and norms, such as $\|a+b\|^2 = \|a\|^2 + \|b\|^2 + 2\langle a, b\rangle$. Focusing now on the second inner product, we substitute in for v_k. Note, from the definition of (**NAG**) we have the relation $\delta_k = \alpha v_k + (1-\alpha)x_k$, then dividing through by α gives $v_k = \delta_k + \frac{1-\alpha}{\alpha}(\delta_k - x_k)$. This can then be substituted in:

$$\langle v_k-x^*,(1-\beta)(\delta_k-x^*)-\nu\nabla f(\delta_k)\rangle = \langle\delta_k+\frac{1-\alpha}{\alpha}(\delta_k-x_k)-x^*,(1-\beta)(\delta_k-x^*)-\nu\nabla f(\delta_k)\rangle$$

$$= \langle\delta_k - x^*, (1-\beta)(\delta_k - x^*) - \nu\nabla f(\delta_k)\rangle + \frac{1-\alpha}{\alpha}\langle\delta_k - x_k, (1-\beta)(\delta_k - x^*) - \nu\nabla f(\delta_k)\rangle$$

$$= (1-\beta)\|\delta_k-x^*\|^2-\nu\langle\delta_k-x^*,\nabla f(\delta_k)\rangle+\frac{1-\alpha}{\alpha}((1-\beta)\langle\delta_k-x_k,\delta_k-x^*\rangle-\nu\langle\delta_k-x_k,\nabla f(\delta_k)\rangle).$$

Now, again just consider the 2nd inner product here. By using the expansion of a norm stated earlier for $a = \frac{1-\alpha}{\alpha}(\delta_k - x_k)$ and $b = \delta_k - x^*$, along with the expression for v_k above:

$$(1-\beta)\langle\frac{1-\alpha}{\alpha}(\delta_k - x_k), \delta_k - x^*\rangle = \frac{1}{2}(1-\beta)[\ \|\delta_k - x^* + \frac{1-\alpha}{\alpha}(\delta_k - x_k)\|^2$$

$$-\|\delta_k - x^*\|^2 - \|\frac{1-\alpha}{\alpha}(\delta_k - x_k)\|^2\ ]$$

$$= \frac{1}{2}(1-\beta)[\ \|v_k - x^*\|^2 - \|\delta_k - x^*\|^2 - \left(\frac{1-\alpha}{\alpha}\right)^2\|\delta_k - x_k\|^2\ ].$$

Finally we can substitute all the previous results into the original $E_{k+1} - E_k$ term:

$$E_{k+1} - E_k = (f(x_{k+1}) - f(x_n)) + \frac{\mu}{2}(\|v_{k+1} - x^*\|^2 - \|v_k - x^*\|^2)$$

$$= (f(x_{k+1}) - f(x_n)) + \frac{\mu}{2}[(\beta^2-1)\|v_k - x^*\|^2 + (1-\beta)^2\|\delta_k - x^*\|^2 + \nu^2\|\nabla f(\delta_k)\|^2$$

$$-2(1-\beta)\nu\langle\delta_k - x^*, \nabla f(\delta_k)\rangle + 2\beta[(1-\beta)\|\delta_k - x^*\|^2 - \nu\langle\delta_k - x^*, \nabla f(\delta_k)\rangle$$

$$+\frac{1-\alpha}{\alpha}((1-\beta)\langle\delta_k - x_k, \delta_k - x^*\rangle - \nu\langle\delta_k - x_k, \nabla f(\delta_k)\rangle)]$$

$$= (f(x_{k+1}) - f(x_n)) + \frac{\mu}{2}[(\beta^2-1)\|v_k - x^*\|^2 + (1-\beta)^2\|\delta_k - x^*\|^2 + \nu^2\|\nabla f(\delta_k)\|^2$$

$$-2(1-\beta)\nu\langle\delta_k - x^*, \nabla f(\delta_k)\rangle + 2\beta[(1-\beta)\|\delta_k - x^*\|^2 - \nu\langle\delta_k - x^*, \nabla f(\delta_k)\rangle$$

$$+(\frac{1-\beta}{2}[\ \|v_k - x^*\|^2 - \|\delta_k - x^*\|^2 - \left(\frac{1-\alpha}{\alpha}\right)^2\|\delta_k - x_k\|^2\ ] - \frac{1-\alpha}{\alpha}\nu\langle\delta_k - x_k, \nabla f(\delta_k)\rangle)].$$

Simplifying leads to the still cumbersome:

$$E_{k+1} - E_k = -(1-\beta)E_k + f(x_{k+1}) - f^* - \beta(f(x_k) - f^*)$$

$$+\frac{\mu}{2}[(1-\beta)\|\delta_k - x^*\|^2 + \nu^2\|\nabla f(\delta_k)\|^2 - \beta(1-\beta)\left(\frac{1-\alpha}{\alpha}\right)^2\|\delta_k - x_k\|^2]$$

$$-\frac{(1-\alpha)\beta\nu\mu}{\alpha}\langle\nabla f(\delta_k), \delta_k - x_k\rangle - \mu\nu\langle\nabla f(\delta_k), \delta_k - x^*\rangle.$$

Now we make use of our assumptions. Firstly we have $(\mathbf{SQC} : \gamma, \mu)$ leading to, by multiplying by γ:

$$\langle\nabla f(\delta_k), \delta_k - x^*\rangle \geq \gamma(f(\delta_k) - f^*) + \frac{\gamma\mu}{2}\|\delta_k - x^*\|^2.$$

Secondly we make use of $(\mathbf{S} : L)$, with a form of the descent lemma. $(\mathbf{S} : L)$ gives $f(x_{k+1}) \leq f(\delta_k) + \langle\nabla f(\delta_k), x_{k+1} - \delta_k\rangle + \frac{L}{2}\|\delta_k - x_{k+1}\|^2$. Using the $(\mathbf{NAG})$ definitions, that is $x_{k+1} = \delta_k - \eta\nabla f(\delta_k)$, we obtain:

$$f(x_{k+1}) \leq f(\delta_k) + \langle\nabla f(\delta_k), \delta_k - \eta\nabla f(\delta_k) - \delta_k\rangle + \frac{L}{2}\|\delta_k - \delta_k - \eta\nabla f(\delta_k)\|^2$$

$$\iff f(\delta_k) - f(x_{k+1}) \geq \|\nabla f(\delta_k)\|^2\left(\eta - \frac{\eta^2 L}{2}\right).$$

If we can choose η such that $\eta \leq \frac{1}{L}$, this then reduces to:

$$\frac{\eta}{2}\|\nabla f(\delta_k)\|^2 \leq f(\delta_k) - f(x_{k+1}).$$

This means we can draw together some terms by using the inequalities. We use the following replacements, from $(\mathbf{S} : L)$,

$$\nu^2\|\nabla f(\delta_k)\|^2 = \nu^2\frac{2}{\eta}\frac{\eta}{2}\|\nabla f(\delta_k)\|^2 \leq \nu^2\frac{2}{\eta}(f(\delta_k) - f(x_{k+1})),$$

and from $(\mathbf{SQC} : \gamma, \mu)$,

$$-\mu\nu\langle\nabla f(\delta_k), \delta_k - x^*\rangle \leq -\gamma\mu\nu(f(\delta_k) - f^*) - \frac{\mu}{2}\gamma\mu\nu\|\delta_k - x^*\|^2.$$

This gives:

$$E_{k+1} - E_k = -(1-\beta)E_k + (f(x_{k+1}) - f^*) - \beta(f(x_k) - f^*)$$

$$+\frac{\mu}{2}[(1-\beta)\|\delta_k - x^*\|^2 + \nu^2\frac{2}{\eta}(f(\delta_k) - f(x_{k+1})) - \beta(1-\beta)\left(\frac{1-\alpha}{\alpha}\right)^2\|\delta_k - x_k\|^2]$$

$$-\frac{(1-\alpha)\beta\nu\mu}{\alpha}\langle\nabla f(\delta_k), \delta_k - x_k\rangle - \gamma\mu\nu(f(\delta_k) - f^*) - \frac{\mu}{2}\gamma\mu\nu\|\delta_k - x^*\|^2.$$

What we notice here is there are a lot of differences in function values, so we aim to simplify here. The problematic one is $f(\delta_k) - f(x_{k+1})$. In order to absorb this with the other terms, we can perform a simple decomposition, that is,

$$\frac{\mu}{2}\nu^2\frac{2}{\eta}(f(\delta_k)-f(x_{k+1})) = \frac{\mu}{\eta}\nu^2(f(\delta_k)-f(x_{k+1})) = \frac{\mu}{\eta}\nu^2(f(\delta_k)-f^*)-\frac{\mu}{\eta}\nu^2(f(x_{k+1})-f^*).$$

We then group terms to obtain,

$$E_{k+1} - E_k = -(1-\beta)E_k + (1 - \frac{\mu}{\eta}\nu^2)(f(x_{k+1}) - f^*) - \beta(f(x_k) - f^*)$$

$$+ \left(\frac{\mu}{\eta}\nu^2 - \gamma\mu\nu\right)(f(\delta_k) - f^*) - \frac{(1-\alpha)\beta\nu\mu}{\alpha}\langle\nabla f(\delta_k), \delta_k - x_k\rangle$$

$$- \frac{\mu\beta(1-\beta)}{2}\left(\frac{1-\alpha}{\alpha}\right)^2\|\delta_k - x_k\|^2 - \frac{\mu}{2}(1-\beta-\gamma\mu\nu)\|\delta_k - x^*\|^2.$$

Now comes some welcome simplification. We can now substitute in the (**NAG**) initialisation values specified at the beginning, in order to cancel out many terms. Our choice of η and s have lead to $\nu = s\eta = \frac{\gamma}{L} = \sqrt{\frac{\eta}{\mu}}$, as show earlier. This means $\left(1 - \frac{\mu}{\eta}\nu^2\right) = 0$ so the $(f(x_{k+1}) - f^*)$ term dissapears. Similarly,

$$1 - \beta = \gamma\sqrt{\mu\eta} = \gamma\sqrt{\frac{\mu\gamma^2\mu}{L^2}} = \frac{\gamma^2\mu}{L} = \gamma\mu\left(\frac{\gamma}{L}\right) = \gamma\mu\nu,$$

So the $\|\delta_k - x^*\|^2$ also disappears. Further, we can perform some nice cancellations with some of the functional difference terms:

$$\left(\frac{\mu}{\eta}\nu^2 - \gamma\mu\nu\right)(f(\delta_k) - f^*) - \beta(f(x_k) - f^*) = (1-\gamma\mu\nu)[(f(\delta_k) - f^*) - (f(x_k) - f^*)]$$

$$= (1-\gamma\mu\nu)(f(\delta_k) - f(x_k)) = (1-\gamma\sqrt{\mu\eta})(f(\delta_k) - f(x_k)).$$

Combining all of these simplifications and re-arranging leads to:

$$E_{k+1} - E_k \leq -(1-\beta)E_k + (1-\gamma\sqrt{\mu\eta})(f(\delta_k) - f(x_k)) - \frac{(1-\alpha)\beta\nu\mu}{\alpha}\langle\nabla f(\delta_k), \delta_k - x_k\rangle$$

$$- \frac{\mu\beta(1-\beta)}{2}\left(\frac{1-\alpha}{\alpha}\right)^2\|\delta_k - x_k\|^2$$

$$= -\gamma\sqrt{\mu\eta}E_k + (1-\gamma\sqrt{\mu\eta})(f(\delta_k) - f(x_k))$$

$$+\frac{1-\alpha}{\alpha}\,(1-\gamma\sqrt{\mu\eta})\sqrt{\mu\eta}\left(\langle\nabla f(\delta_k),x_k-\delta_k\rangle-\frac{1-\alpha}{\alpha}\frac{\gamma\mu}{2}\|\delta_k-x_k\|^2\right).$$

The secondary consequence of (**S**: L) is the following rearranged *lower* bound:

$$\langle\nabla f(\delta_k),x_k-\delta_k\rangle\leq f(x_k)-f(\delta_k)+\frac{L}{2}\|\delta_k-x_k\|^2.$$

Substituting this in, we obtain:

$$E_{k+1}-E_k\leq-\gamma\sqrt{\mu\eta}E_k+(1-\gamma\sqrt{\mu\eta})(f(\delta_k)-f(x_k))$$
$$+\frac{1-\alpha}{\alpha}\,(1-\gamma\sqrt{\mu\eta})\sqrt{\mu\eta}\left(f(x_k)-f(\delta_k)+\frac{L}{2}\|\delta_k-x_k\|^2-\frac{1-\alpha}{\alpha}\frac{\gamma\mu}{2}\|\delta_k-x_k\|^2\right)$$

$$=-\gamma\sqrt{\mu\eta}E_k+(1-\gamma\sqrt{\mu\eta})\left(1-\frac{1-\alpha}{\alpha}\sqrt{\mu\eta}\right)(f(\delta_k)-f(x_k))$$
$$-\frac{1-\alpha}{2\alpha}\,(1-\gamma\sqrt{\mu\eta})\sqrt{\mu\eta}\left(\frac{1-\alpha}{\alpha}\gamma\mu-L\right)\|\delta_k-x_k\|^2.$$

We can cancel further, recall $\alpha=\frac{\sqrt{\mu\eta}}{1+\sqrt{\mu\eta}}$, so:

$$1-\frac{1-\alpha}{\alpha}\sqrt{\mu\eta}=1-\frac{1}{1+\sqrt{\mu\eta}}\frac{1+\sqrt{\mu\eta}}{\sqrt{\mu\eta}}\sqrt{\mu\eta}=0.$$

Thus, the $f(\delta_k)-f(x_k)$ term is cancelled too. We can also remove the $\|\delta_k-x_k\|^2$ term if one of the terms were to equal zero. We see then that,

$$\frac{1-\alpha}{\alpha}\gamma\mu-L=\gamma\mu\frac{1}{\sqrt{\mu\eta}}-L=\gamma\mu\sqrt{\frac{L^2}{\gamma^2\mu^2}}-L=\frac{L\gamma\mu}{\gamma\mu}-L=L-L=0.$$

Finally then we obtain $E_{k+1}-E_k\leq-\gamma\sqrt{\mu\eta}E_k$. Rearrange for $E_{k+1}\leq(1-\gamma\sqrt{\mu\eta})E_k$. This gives us our recursion of the Lyapunov energy as seen before, such that: $E_k=(1-\gamma\sqrt{\mu\eta})^kE_0=(1-\gamma\sqrt{\mu\eta})^k(f(x_0)-f^*+\frac{\mu}{2}\|v_0-x^*\|^2)$. Recall $v_0=x_0$ from the statement of the theorem, and that $\gamma\sqrt{\mu\eta}=\gamma^2\frac{\mu}{L}$. We can simplify as in Theorem 19 with (**QG**: $\frac{\mu\gamma}{2L}$) for the result. □

C Plotting Details

Plots were generated using Python with Matplotlib, and numerical experiments performed using software Desmos. Conditions were verified in the form of figure 3 (b), where the function was determined to be above, or below, 0.