SGD for Structured Nonconvex Functions: Learning Rates, Minibatching and Interpolation

Robert M. Gower
LTCI, Telecom Paris, IPP

Othmane Sebbouh
ENS Paris & CREST-ENSAE

Nicolas Loizou
Mila and DIRO, Université de Montréal

Abstract

Stochastic Gradient Descent (SGD) is being used routinely for optimizing non-convex functions. Yet, the standard convergence theory for SGD in the smooth non-convex setting gives a slow sublinear convergence to a stationary point. In this work, we provide several convergence theorems for SGD showing convergence to a global minimum for non-convex problems satisfying some extra structural assumptions. In particular, we focus on two large classes of structured non-convex functions: (i) Quasar (Strongly) Convex functions (a generalization of convex functions) and (ii) functions satisfying the Polyak-Lojasiewicz condition (a generalization of strongly-convex functions). Our analysis relies on an Expected Residual condition which we show is a strictly weaker assumption than previously used growth conditions, expected smoothness or bounded variance assumptions. We provide theoretical guarantees for the convergence of SGD for different step-size selections including constant, decreasing and the recently proposed stochastic Polyak step-size. In addition, all of our analysis holds for the arbitrary sampling paradigm, and as such, we give insights into the complexity of minibatching and determine an optimal minibatch size. Finally, we show that for models that interpolate the training data, we can dispense of our Expected Residual condition and give state-of-the-art results in this setting.

1. Introduction

We consider the unconstrained finite-sum optimization problem

\[
\min_{x \in \mathbb{R}^d} \left[ f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right].
\]

We use \(\mathcal{X}^* \subset \mathbb{R}^d\) to denote the set of minimizers \(x^*\) of (1) and assume that \(\mathcal{X}^*\) is not empty and that \(f(x)\) is lower bounded. This problem is prevalent in machine learning tasks where \(x\) corresponds to the model parameters, \(f_i(x)\) represents the loss on the training point \(i\) and the aim is to minimize the average loss \(f(x)\) across training points.

When \(n\) is large, stochastic gradient descent (SGD) and its variants are the preferred methods for solving (1) mainly because of their cheap per iteration cost. The standard convergence theory for SGD [2, 13, 37–39, 47, 52] in the smooth nonconvex setting shows slow sub-linear convergence...
SGD for Structured Nonconvex Functions

to a stationary point. Yet in contrast, when applying SGD to many practical nonconvex problems of the form (1) such as matrix completion [49], deep learning [32], and phase retrieval [53] the iterates converge globally, and sometimes, even linearly. This is because these problems often have additional structure and properties, such as all local minima are global minima [18, 49], the model interpolates the data [32] or the function under study is unimodal on all lines through a minimizer [16]. By exploiting these structures and properties one can prove significantly tighter (and more useful) convergence bounds.

Here we present a general analysis of SGD for two large classes of structured nonconvex functions: (i) the Quasar (Strongly) Convex functions and (ii) the functions satisfying the Polyak-Lojasiewicz (PL) condition. In all of our results we provide convergence guarantees for SGD to the global minimum. We also develop several corollaries for functions that interpolate the data.

1.1. Background and Main Contributions

Classes of structured nonconvex functions. The last few years has seen an increased interest in exploiting additional structure prevalent in large classes of nonconvex functions. Such conditions include error bounds property [7], essential strong convexity [25], quasi strong convexity [11, 34], the restricted secant inequality [58], and the quadratic growth (QG) condition [1, 28]. We focus on two of the weakest conditions: the quasar (strongly) convex functions [12, 14, 16] and functions satisfying the PL condition [17, 31, 42]. The class of quasar convex functions include all convex functions as a special case, but it also includes several nonconvex functions. Recently there is also some evidences suggesting that the loss function of neural networks have a quasar-convexity structure [21, 59].

Contributions. We show that SGD converges at a $O(1/\sqrt{k})$ rate on the quasar convex functions and prove linear convergence to a neighborhood for PL functions without any bounded variance assumption or growth assumptions on the stochastic gradients. Instead, we rely on the recently introduced expected residual (ER) condition [10].

Assumptions on the gradient. The standard convergence analysis of SGD in the nonconvex setting relies on the bounded gradients assumption $\mathbb{E}_i \|\nabla f_i(x_k)\|^2 < c$ [15, 44, 45] or a growth condition $\mathbb{E}_i \|\nabla f_i(x_k)\|^2 \leq c_1 + c_2 \mathbb{E} \|\nabla f(x_k)\|^2$ [4, 5, 50]. There is now a line of recent works [11, 19, 22, 24, 41, 54] which aims at relaxing these assumptions.

Contributions. We use the recently introduced Expected Residual (ER) condition [10]. We give the first convergence proofs for SGD under the ER condition and we show that ER is a strictly weaker assumption than the SGC, Weak Growth (WGC) [54] or the Expected Smoothness (ES) [11] assumptions. Furthermore, we show that the ER condition holds for a large class of nonconvex functions including 1) smooth and interpolated functions 2) smooth and $x^*$– convex functions\(^1\). Not only does the ER assumption hold for a larger class of functions, our resulting convergence rates under ER either match or exceed the state-of-the-art for quasar convex and PL functions.

PL condition. The PL condition [31, 42] was introduced as a sufficient condition for the linear convergence of Gradient Descent for nonconvex functions. Assuming bounded gradients, it was shown in [17] that SGD with a decreasing step size converges sublinearly at a rate of $O(1/\sqrt{k})$ for PL functions. In contrast, by using a step size which depends on the last iterate, the same convergence rate can be achieved without the need for the bounded gradient assumption [19]. Assuming in addition the interpolation condition and Strong Growth Condition (SGC) [54] showed that SGD

\(^1\) The $x^*$– convexity includes all convex functions and several nonconvex functions
converges linearly for PL functions, but the specialization of this last result to gradient descent results in a suboptimal dependence on the condition number \(^2\) of the function.

**Contributions.** We propose an analysis of minibatch SGD for PL functions which recovers the best known dependence on the condition number for Gradient Descent [17] while also matching the current state-of-the-art rate derived in [24, 54] for SGD for interpolated functions. All of which relies on the weaker ER condition. Moreover, we propose a switching step size scheme similar to [11] which does not require knowledge of the last iterate of the algorithm. Using this step size, we prove that SGD converges sublinearly at a rate of \(O(1/k)\) for PL functions without any additional bounded gradient of bounded variance assumption or growth assumption.

**Step-size selection for SGD.** The most important parameter that one should select to guarantee the convergence of SGD is the step-size or learning rate. There are several choices that one can use including constant step-size [11, 33, 35, 36, 41], decreasing step-size [8, 11, 17, 39, 47] and adaptive step-size [3, 6, 20, 26, 55, 56].

**Contributions.** We provide convergence theorems for SGD under several step-size rules for minimiz-

**Over-parameterized models and Interpolation.** Recently it was shown that SGD converges considerably faster when the underlying model is sufficiently over-parameterized as to interpolate the data. This includes problems such as deep matrix factorization [43, 48], binary classification using kernels [30], consistent linear systems [9, 27, 29, 46] and multi-class classification using deep networks [54].

**Contributions.** As a corollary of our main theorems we show that for models that interpolate the training data, we can further relax our assumptions, dispense of the ER condition and instead, simply assume that each \(f_i\) is smooth. Our results here match the state-of-the-art convergence results [54] but again under strictly weaker assumptions.

1.2. SGD and Arbitrary Sampling

When the number of terms \(n\) in the summation (1) is large calculating the full gradient becomes prohibitive. So instead, we assume we are given access to unbiased estimates \(g(x) \in \mathbb{R}^d\) of the gradient such that \(\mathbb{E} [g(x)] = \nabla f(x)\). For example, we can use a minibatch to form an estimate of the gradient such as \(g(x) = \frac{1}{b} \sum_{i \in B} \nabla f_i(x)\), where \(B \subset \{1, \ldots, n\}\) will be chosen uniformly at random and \(|B| = b\). To allow for any form of minibatching we use the *arbitrary sampling* notation

\[
g(x) = \nabla f_v(x) := \frac{1}{n} \sum_{i=1}^{n} v_i \nabla f_i(x),
\]

where \(v \in \mathbb{R}_+^n\) is a random sampling vector such that \(\mathbb{E} [v_i] = 1\), for \(i = 1, \ldots, n\) and \(f_v(x) := \frac{1}{n} \sum_{i=1}^{n} v_i f_i(x)\). Note that it follows immediately from this definition of sampling vector that \(\mathbb{E} [g(x)] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} [v_i] \nabla f_i(x) = \nabla f(x)\). In this work we mostly focus on the \(b\)-minibatch sampling, however we highlight that our analysis holds for any form of sampling vectors.

**Definition 1 (Minibatch sampling)** Let \(b \in [n]\). We say that \(v \in \mathbb{R}_+^n\) is a \(b\)-minibatch sampling if for every subset \(S \in [n]\) with \(|S| = b\) we have that \(\mathbb{P} \left[ v = \frac{b}{n} \sum_{i \in S} e_i \right] = 1/(\binom{n}{b}) = b!(n-b)!/n!\)

\(^2\) Theorem 4 in [54] specialized to GD gives a rate of \(\mu^2/L^2\) where \(L\) is the smoothness constant and \(\mu\) the PL constant.
It is easy to verify by using a double counting argument that if $v$ is a $b$–minibatch sampling, it is also a valid sampling vector ($E[v_i] = 1$) [11]. See [11] for other choices of sampling vectors $v$.

With an unbiased estimate of the gradient $g(x)$, we can now use Stochastic gradient descent (SGD) to solve (1) by sampling $g(x^k)$ i.i.d and iterating

$$x^{k+1} = x^k - \gamma^k g(x^k)$$

(3)

We also make the following mild assumption on the gradient noise.

**Assumption 1** The gradient noise $\sigma^2 := \sup_{x^* \in X^*} E[\|g(x^*)\|^2]$ is finite.

### 2. Classes of Structured Nonconvex Functions

We work with two classes of nonconvex problems: the quasar-convex functions and the functions that satisfy the Polyak-Lojasiewicz (PL) condition.

**Definition 2 (Quasar convex)** Let $\zeta > 0$ and $x^* \in X^*$. We that $f$ is $\zeta$-quasar convex with respect to $x^*$ if for all $x \in \mathbb{R}^n$,

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

(4)

For shorthand we write $f \in QC(\zeta)$ to mean (4). The class of quasar convex functions are parameterized by a positive constant $\zeta > 0$. In the case that $\zeta = 1$ then (4) is known as the star convexity [40] (generalization of convexity). One can think of $\zeta$ as the value that controls the non-convexity of the function. In particular, as value of $\zeta$ becomes smaller or larger than 1 the function becomes “more nonconvex” [16]. We highlight that the Quasar convex functions may have multiple solutions.

One of weakest possible assumptions that guarantee a global convergence of gradient descent to the global minimum is the PL condition [17]. This is largely due to the fact that all local minimas of a function satisfying the PL condition are also global minimas.

**Definition 3 (Polyak-Lojasiewicz (PL) Condition)** There exists $\mu > 0$ such that

$$\|\nabla f(x)\|^2 \geq 2\mu [f(x) - f^*]$$

(5)

We write $f \in PL(\mu)$ if function $f$ satisfies inequality (5).

In addition we will also consider in several corollaries the following interpolation condition.

**Assumption 2** We say that the interpolation condition holds if there exists $x^* \in X^*$ such that

$$\min_{x \in \mathbb{R}^n} f_i(x) = f_i(x^*) \quad \text{for} \quad i = 1, \ldots, n.$$  

(6)

This interpolation condition has drawn much attention recently because many overparametrized deep neural networks achieve a zero loss over all training data points [32] and thus satisfy (6).
3. Expected Residual (ER)

In all of our analysis of SGD we rely on the remarkably weak Expected Residual (ER) assumption. In this section we formally define ER, provide new sufficient conditions for it to hold and relate it to the existing assumptions on the gradient.

ER measures how far the gradient estimate $g(x)$ is from the true gradient in the following sense.

**Assumption 3 (Expected residual)** We say that the ER condition holds or $g \in \text{ER} (\rho)$ if

$$
\mathbb{E} \left[ \| g(x) - g(x^*) - (\nabla f(x) - \nabla f(x^*)) \|^2 \right] \leq 2\rho (f(x) - f(x^*)), \quad \forall x \in \mathbb{R}^d.
$$

(ER)

Note that ER depends on both how $g(x)$ is sampled and the properties of the $f(x)$ function.

As a direct consequence of Assumption 3 we have the following bound on the variance of $g(x)$.

**Lemma 4** If $g \in \text{ER} (\rho)$ then

$$
\mathbb{E} \left[ \| g(x) \|^2 \right] \leq 4\rho (f(x) - f^*) + \| \nabla f(x) \|^2 + 2\sigma^2.
$$

(7)

It is this bound on the variance (7) that we use in our proofs and allows us to avoid the stronger bounded gradient or bounded variance assumptions.

**Connections to other Assumptions.** Let us provide some more familiar sufficient conditions which guarantee that the ER condition holds. In doing so, we will also provide simple and informative bounds on the expected residual constant $\rho$, when using minibatching.

We say that $f_i$ is $L_i$–smooth if

$$
f_i(z) - f_i(x) \leq \langle \nabla f_i(x), z - x \rangle + \frac{L_i}{2} \| z - x \|^2, \quad \forall x, z \in \mathbb{R}^d.
$$

(8)

Let $L_{\text{max}} := \max_{i=1,...,n} L_i$. For $x^* \in X^*$, we say that $f_i$ is $x^*$–convex if

$$
f_i(x^*) - f_i(x) \leq \langle \nabla f_i(x^*), x^* - x \rangle, \quad \forall x \in \mathbb{R}^d.
$$

(9)

These two assumptions are sufficient for the $\text{ER}(\rho)$ condition to hold and give a useful bound on $\rho$, as we show in the following proposition.

**Proposition 5** Let $v$ be a sampling vector. If $f_i$ is $L_i$–smooth and there exists $x^* \in X^*$ such that $f_i$ is $x^*$–convex then $g \in \text{ER}(\rho)$. If in addition $v$ is the $b$–minibatch sampling then

$$
\rho(b) = L_{\text{max}} \frac{n-b}{(n-1)^2}, \quad \sigma^2(b) = \frac{1}{b} \frac{n-b}{n-1} \sigma_1^2 \quad \text{where} \quad \sigma_1 := \sup_{x^* \in X^*} \frac{1}{n} \sum_{i=1}^n \| \nabla f_i(x^*) \|.
$$

(10)

The bounds in Proposition 5 have been proven before but under the stronger assumption that each $f_i$ is convex. In this work by dropping the requirement that each $f_i$ is convex we are able to consider interesting classes of nonconvex functions. To this end, the following Theorem is of great importance. It establishes that only smoothness and the interpolation condition are sufficient for the ER to hold. Furthermore, we place the ER within a hierarchy of other popular assumptions used in the analysis of SGD for smooth nonconvex functions. We show that ER is the weakest condition.

---

3. See Proposition 3.10 item (iii) in [11] and Lemma F.3 in [51]
Theorem 6  Let ES, WGC and SGC denote Assumption 2.1 in [11], Eq (7) and Eq (2) in [54], respectively. Let $L_i$ and $x^*$–convex abbreviate (8) and (9), respectively. The hierarchy holds

$$
\text{SGC + L–smooth} \quad \Rightarrow \quad \text{WGC} \quad \Rightarrow \quad \text{ES} \quad \Rightarrow \quad \text{ER}
$$

\[
L_i + \text{Interpolated} \quad \Rightarrow \quad L_i + x^*–\text{convex}
\]

where L–smooth is shorthand for $f(x)$ is L–smooth. Finally, there are problems such that ER holds and ES does not hold. Making ER the strictly weakest assumption among the above.

The important assumptions for analyzing SGD in the nonconvex setting are the ones that are downstream from $L_i + \text{Interpolated}$. This is because these exists a rich class of nonconvex functions that are smooth and satisfy the interpolation condition. In contrast, the WGC is only known to hold for smooth and convex functions that also satisfy the interpolation assumptions (Proposition 2 in [54]). A hierarchy of assumptions that also includes the SGC and a parametrized family of assumptions that include the ER condition as a special was also recently presented in [19].

4. Convergence Analysis

In this section, we present the main convergence results. Proofs of all key results can be found in the Appendix D. In Appendix E, we present additional convergence results on quasar-strongly convex functions (Section E.1) and on convergence under expected smoothness (Section E.2).

4.1. Quasar Convex functions

4.1.1. Constant and Decreasing Stepsizes

Now we present our results for quasar-convex functions for SGD with a constant, finite horizon and decreasing step sizes.

Theorem 7 Assume $f(x)$ is L–smooth, $\zeta$–quasar-convex with respect to $x^*$ and $g \in \text{ER}(\rho)$. Let $0 < \gamma_k < \frac{\zeta}{2\rho + L}$ for all $k \in \mathbb{N}$ and let $r_0 = \|x^0 - x^*\|^2$. SGD (3) converges as

$$
\min_{t=0, \ldots, k-1} \mathbb{E} \left[ f(x^t) - f(x^*) \right] \leq \frac{r_0}{2} + \frac{\gamma_k^2}{2} \sum_{t=0}^{k-1} \gamma_t^2.
$$

Moreover, for $\gamma < \frac{\zeta}{2\rho + L}$ we have that

1. If $\forall k \in \mathbb{N}$, $\gamma_k \equiv \frac{1}{2} \frac{\zeta}{(2\rho + L)}$ then $\forall k \in \mathbb{N}$,

$$
\min_{t=0, \ldots, k-1} \mathbb{E} \left[ f(x^t) - f(x^*) \right] \leq 2r_0 \frac{\rho + L}{\zeta} + \frac{\sigma^2}{2\rho + L}.
$$

2. Suppose SGD (3) is run for $T$ iterations. If $\forall k = 0, \ldots, T - 1$, $\gamma_k = \frac{\gamma}{\sqrt{T}}$ then

$$
\min_{t=0, \ldots, T-1} \mathbb{E} \left[ f(x^t) - f(x^*) \right] \leq \frac{r_0 + 2\gamma^2\sigma^2}{\gamma}. \quad (13)
$$

3. If $\forall k \in \mathbb{N}$, $\gamma_k = \frac{\gamma}{\sqrt{k+1}}$ then $\forall k \in \mathbb{N}$,

$$
\min_{t=0, \ldots, k-1} \mathbb{E} \left[ f(x^t) - f(x^*) \right] \leq \frac{r_0 + 2\gamma^2\sigma^2(\log(k) + 1)}{\zeta^2(\sqrt{k} - 1) - \gamma(\rho + L/2)(\log(k) + 1)} \sim O \left( \frac{\log(k)}{\sqrt{k}} \right).
$$

1
Thus the optimal minibatch size is given by

$$b^* = \begin{cases} 1 & \text{if } (n - 1) \geq 2 \frac{L_{\max}}{L} \\ \frac{1}{n} & \text{if } (n - 1) < 2 \frac{L_{\max}}{L} \end{cases}.$$  

Specializing (15) to the full batch setting ($n = b$), we have that gradient descent (GD) with step size $\gamma = \frac{\zeta}{\mathcal{L}}$ converges as follows$^4$: $f(x^t) - f(x^*) \leq \frac{2L||x^t - x^*||^2}{\zeta^2}$. This is exactly the rate given recently for GD for quasar-convex functions in [12], with the exception that we have a squared dependency on $\zeta$ the quasar convex parameter.

4.1.2. STOCHASTIC POLYAK STEP-SIZE (SPS) - GUARANTEE CONVERGENCE WITHOUT TUNING

The stochastic Polyak step size (SPS) is a recently proposed step size selection for SGD [30]. We generalize the SPS to the arbitrary sampling regime and provide a new convergence analysis of SGD with SPS for the class of smooth, quasar (strongly) convex functions.

Let $\nu$ be a sampling vector and let $f_\nu = \sum_{i=1}^n f_i(x)v_i$. Let $f^{\star}_\nu = \min_x f_\nu(x)$ which we assume exists. Just like the gradient, we have that $f^{\star}_i$ is an unbiased estimate of $f$. Now given a sampling vector $\nu$, we define the Stochastic Polyak Stepsize (SPS) as

$$\text{SPS: } \gamma_k = \frac{f^{\star}_{\nu}(x^k) - f^{\star}_\nu}{c||\nabla f_\nu(x^k)||^2},$$

where $0 < c \in \mathbb{R}$. The SPS rule is particularly useful when the interpolation Assumption 2 holds and each $f_i$ represents a loss function, since then we have that $f^{\star}_i = f^{\star}_\nu = f(x^*) = 0$ for every $i \in [n]$ and realization of $\nu$.

---

$^4$ Here we use that the smoothness of $f$ guarantees that $f(x^1), \ldots, f(x^t)$ for GD is a decreasing sequence.
By assuming that every $f_i$ is $L_i$-smooth, we have that $f_v$ is $L_v$-smooth with $L_v := \frac{1}{n} \sum_{i=1}^{n} v_i L_i$. This smoothness combined with Lemma 16 and Jensen’s inequality gives a lower bound on SPS (18):

$$\frac{1}{2c^2 \mathbb{E}[L_v]} \leq \mathbb{E} \left[ \frac{1}{2c} \right] \leq \mathbb{E} \left[ \gamma_k = \frac{f_i(x^k) - f_v^*}{c \|\nabla f_v(x^k)\|^2} \right],$$

(19)

This lower bound and the following new expected smoothness bound allows us to establish the forthcoming theorem for quasar convex functions.

**Lemma 9** Assume interpolation 2 holds. Let $f_i$ be $L_i$-smooth and let $v$ be a sampling vector. It follows that there exists $L_{max} > 0$ such that

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

(20)

Furthermore, for $B \subset \{1, \ldots, n\}$ let $L_B$ be the smoothness constant of $f_B := \frac{1}{n} \sum_{i \in B} p_i f_i$. If $v$ is the $b$-minibatch sampling then $L_{max} = L_{max}(b) = \max_{i=1, \ldots, n} \sum_{B, i \in B} L_B^{-1}$.

With the above lemma we can now establish our main theorem.

**Theorem 10** Let $v$ be a sampling vector. Assume interpolation 2 holds. Assume that each $f_i$ is $\zeta$-quasar convex with respect to $x^*$ and $L_i$-smooth. Then SGD with SPS (18) and $c > \frac{1}{2\zeta}$ converges as follows:

$$\min_{i=0, \ldots, K-1} \mathbb{E} \left[ f(x^i) - f^* \right] \leq \frac{2c^2}{2c-1} \frac{L_{max}}{K} \|x^0 - x^*\|^2,$$

where $L_{max}$ is the expected smoothness constant defined in Lemma (9).

We now use $L_{max}(b)$ given in Lemma 9 to derive the importance sampling complexity. To the best of our knowledge, this is the first importance sampling result for SGD with SPS in any setting.

**Corollary 11** Consider the setting of Theorem 10 with $c = 1/4\zeta$. Given $\epsilon > 0$ we have that

$$k \geq \frac{L_{max}}{4c^2} \frac{\|x^0 - x^*\|^2}{\epsilon} = O \left( \frac{L_{max}}{c^2 \epsilon} \right) \Rightarrow \min_{i=0, \ldots, K-1} \mathbb{E} \left[ f(x^i) - f^* \right] < \epsilon.$$

(21)

1. (Full batch) If we use full batch sampling we have that $L_{max} = L$ and (21) becomes $O(L/\epsilon \zeta^2)$

2. (Importance sampling). If we use single element sampling with $p_i = L_i / \sum_j L_j$ we have that $L_{max} = \frac{1}{n} \sum_{j=1}^{n} L_j := \mathcal{L}$ and (21) becomes $O(\mathcal{L}/\epsilon \zeta^2)$.

### 4.2. PL condition

Here we present our convergence results for functions satisfying the PL condition.

**Theorem 12** Let $f$ be $L$-smooth. Assume $f \in PL(\mu)$ and $g \in ER(\rho)$. Let $\gamma_k = \gamma \leq \frac{1}{1+2\rho/\mu} \frac{1}{L}$, for all $k$, then SGD given by (3) converges as follows:

$$\mathbb{E}[f(x^k) - f^*] \leq (1 - \gamma \mu)^k [f(x^0) - f^*] + \frac{\rho \sigma^2}{\mu}.$$  

(22)

Hence, given $\epsilon > 0$ and using the step size $\gamma = \frac{1}{L} \min \left\{ \frac{\mu \epsilon}{2\sigma^2}, \frac{1}{1+2\rho/\mu} \right\}$ gives

$$k \geq \frac{L}{\mu} \max \left\{ \frac{2\sigma^2}{\mu \epsilon}, 1 + \frac{2\rho}{\mu} \right\} \log \left( \frac{2(f(x^0) - f^*)}{\epsilon} \right) \implies \mathbb{E} \left[ f(x^k) - f^* \right] \leq \epsilon.$$

(23)
In Appendix D.7 based on Theorem 12, we show how to obtain a $O(1/k)$ convergence for SGD using an insightful stepsize-switching rule. This stepsize-switching rule describes when one should switch from a constant to a decreasing stepsize regime. When the function is interpolated, a constant step size gives a linear rate of convergence, as we show in the next theorem.

**Corollary 13** Consider the setting of Theorem 12 and let assume interpolation 2 holds. Then SGD with $\gamma_k = \gamma \leq \frac{1}{1 + 2\rho/\mu}$, converges linearly at a rate of $1 - \gamma\mu$. Consequently for every $\epsilon > 0$, the iteration complexity of SGD to achieve $\mathbb{E} \left[ f(x^k) - f^* \right] \leq \epsilon$ is

$$k \geq \frac{L}{\mu} \left( 1 + 2 \frac{\sigma}{\mu} \right) \log \left( \frac{f(x^0) - f^*}{\epsilon} \right).$$  \hspace{1cm} (24)

If $b$ is a $b$-minibatch sampling then the total complexity is given by

$$\text{Total Complexity}(b) \geq \frac{L}{\mu} \left( b + 2 \frac{L_{\max}}{\mu} \frac{n-b}{n-1} \right) \log \left( \frac{f(x^0) - f^*}{\epsilon} \right).$$  \hspace{1cm} (25)

Finally, let $\kappa_{\max} := L_{\max}/\mu$. The minibatch size $b^*$ that optimizes the total complexity is given by

$$b^* = \begin{cases} 1 & \text{if } n - 1 \geq 2\kappa_{\max} \\ n & \text{if } n - 1 < 2\kappa_{\max}. \end{cases}$$ \hspace{1cm} (26)

Note that Corollary 13 recovers the linear convergence rate of the gradient descent algorithm under the PL condition [17] as a special case. Indeed for gradient descent we have that $\sigma = 0 = \rho$. Thus by choosing $\gamma = \frac{1}{L}$ the resulting iteration complexity is $\frac{L}{\mu} \log(\epsilon^{-1})$ which was already proven in [17] and is the tightest known convergence result for gradient descent under the PL condition. On the other extreme, we see that for $b = 1$, that is SGD without minibatching, we obtain the convergence rate $1 - \mu^2/3LL_{\max}$ which matches the current state-of-the-art rate [54, Thm. 4], [19, Thm. 2] and [24, Thm 4] known under the exact same assumptions. Thus we recover the best known rate on either end ($b = n$ and $b = 1$), and give the first rates for everything in between $1 < b < n$. To the best of our knowledge our result is the first analysis of SGD for PL functions that recovers the deterministic gradient descent convergence as special case.

5. Conclusions

We establish a hierarchy between the expected residual (ER) condition and a host of other assumptions used in bounding the stochastic gradient in the smooth setting. In particular, we show that ER is the weakest condition and that it holds for nonconvex functions. Using the ER condition, we give the first convergence results for SGD on quasar convex functions without the bounded gradient or bounded variance assumption. We also establish Theorem 12 which recovers the best known convergence results on PL functions for gradient descent and SGD as a special case, and all minibatch sizes in between.

References


SGD FOR STRUCTURED NONCONVEX FUNCTIONS


Supplementary Material

The Supplementary Material is organized as follows:

In Section A we provide some examples of classes of nonconvex functions that satisfy all of the assumptions of our main theorems.
In Section B, we give some lemmas and consequences of smoothness.
In Section C we present the proofs of the proposition, lemma and theorem related to the Expected Residual condition as presented in Section 3 of the main paper.
In Section D we present the proofs of the main theorems.
In Section E we provide additional convergence results under the strongly quasar convex assumption (Section E.1), the Expected Smoothness assumption (Section E.2) and a minibatch analysis that does not rely on the interpolation condition (Section E.3).

Appendix A. Examples

In this section we provide some examples of classes of nonconvex functions that satisfy all of the assumptions of our main theorems.

Separable functions. Our first example that satisfies both the PL and ER condition are separable functions $f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x_i)$. If each $f_i(x_i)$ satisfies the PL condition with constant $\mu_i$ then $f(x)$ satisfies the PL condition with $\mu = \min_{i=1,...,n} \frac{\mu_i}{n}$. If in addition each $f_i$ is a smooth function then according to Theorem 6 we have that the ER condition holds, and thus Theorem 13 holds.

As an example, consider the nonconvex function

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} a_i(x_i^2 + 4b_i \sin^2(x_i)) := f_i(x), \quad (27)$$

where $a_i > 0$ and $1 > b_i > 0$ for $i = 1, \ldots, n$, so that each $f_i$ satisfies the PL condition (see [17]). The function (27) is interpolated since $x^* = 0$ is a global minima for each $f_i$. Furthermore $f_i$ is smooth since $|f_i''(x)| \leq 2a_i + 6b_i$. By the above arguments, so does $f$ satisfy the PL condition. Thus by Theorem 13 we know that SGD converges linearly when applied to (27). To illustrate that such functions (27) are nonconvex, we have a surface plot for $n = 2$ in Figure 1.

Nonlinear least squares. Let $F: \mathbb{R}^d \to \mathbb{R}^n$ be a differentiable function where $DF(x) \in \mathbb{R}^{n \times d}$ is its Jacobian. Now consider the nonlinear least squares problem $\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{2n} \|F(x) - y\|^2 = \frac{1}{2n} \sum_{i=1}^{n} (F_i(x) - y_i)^2$, where $y \in \mathbb{R}^n$. For this class of problems we can show the following:

**Lemma 14** Assume there exists $x^* \in \mathbb{R}^d$ such that $F(x^*) = y$. If the $F_i(x)$ functions are Lipschitz and the $DF(x)$ has full row rank then $F$ satisfies the PL and the ER condition.

5. In [17] the authors claim that $x^2 + 3 \sin^2(x)$ is PL. We then used computer aided analysis to show that $x^2 + 3b \sin^2(x)$ satisfies the PL condition for $0 < b < 4$. 

![Figure 1: Surface plot of $x^2 + 3 \sin^2(x) + 1.5y^2 + 4\sin^2(y)$](image-url)
Star/Quasar convex. Several nonconvex empirical risk problems are quasar convex functions [23]. Let \( f_i : \mathbb{R}^d \rightarrow \mathbb{R} \) be a smooth star-convex (quasar convex with \( \zeta = 1 \)) centered at 0. Each \( f_i \) represent the loss over the \( i \)th data point. Let \( A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m \) such that there exists \( A x^* = b \). Since compositions of affine maps with star convex functions are star convex [23, Section A.4] we have that \( f_i(Ax - b) \) is star convex centered at \( x^* \). Furthermore the average of star convex functions that share the same center are star convex. Thus, \( f(x) = \frac{1}{n} \sum_{i=1}^n f_i(Ax - b) \), is a star-convex function which also satisfies the interpolation condition. Note that this model includes several nonconvex examples, such the case where \( f_i \) is the \( \ell_p \) loss or the square of any monomials [23].

Appendix B. Technical Lemmas on Smoothness

Here we give some lemmas and consequences of smoothness.

For all of our analysis we do not need that the \( f_i \) functions be smooth in all directions. Rather, we just need them to be smooth along the \( x^* \)–direction, as we define next.

**Definition 15** We say that \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) is \( L \)–smooth function along the \( x^* \)–direction if there exists \( x^* \) such that

\[
  f(z) - f(x) \leq \langle \nabla f(x), z - x \rangle + \frac{L}{2} \|z - x\|^2, \quad \forall x \in \mathbb{R}^d,
\]

where

\[
  z = x - \frac{1}{L}(\nabla f(x) - \nabla f(x^*)).
\]

By inserting \( z \) into (28) we can equivalently write (28) as

\[
  f(x - (1/L)(\nabla f(x) - \nabla f(x^*))) \leq f(x) - \frac{1}{2L} \|\nabla f(x)\|^2 + \frac{1}{2L} \|\nabla f(x^*)\|^2. \tag{29}
\]

**Lemma 16** Let \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) be differentiable and suppose \( f \) has a minimizer \( x^* \in \mathbb{R}^d \). Furthermore, let \( f \) be \( L \)–smooth function along the \( x^* \)–direction according to Definition 15. It follows that

\[
  \|\nabla f(x)\|^2 \leq 2L(f(x) - f(x^*)). \tag{30}
\]

**Proof** Since \( x^* \) is a minimizer of \( f \) we have that \( \nabla f(x^*) = 0 \). Furthermore, since \( f \) is \( L \)–smooth function along the \( x^* \)–direction we have by re-arranging (29) that

\[
  f(x^*) - f(x) \leq f(x - (1/L)\nabla f(x)) - f(x) \leq \frac{1}{2L} \|\nabla f(x)\|^2.
\]

Re-arranging the above gives (30). \[\blacksquare\]

Now we provide a lemma that will then be used to establish the simplest and most minimalistic assumptions that imply the expected residual (ER) condition (Assumption 3).

**Lemma 17** Suppose these exists \( x^* \in \mathbb{R}^d \) where

\[
  x^* \in \arg \min \left\{ f(x) := \frac{1}{n} \sum_{i=1}^n f_i(x) \right\},
\]
Suppose the interpolated Assumption 2 holds. Furthermore, suppose that for each \(i\),

\[
\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 \leq 2L_i(f_i(x) - f_i(x^*) - \langle \nabla f_i(x^*), x - x^*\rangle), \quad \forall x \in \mathbb{R}^d. \tag{32}
\]

**Proof** Fix \(i \in \{1, \ldots, n\}\). To prove (32), it follows that

\[
f_i(x^*) - f_i(x) = f_i(x^*) - f_i(z) + f_i(z) - f_i(x) \leq \langle \nabla f_i(x^*) - \nabla f_i(z), x^* - z \rangle + \langle \nabla f_i(z), z - x \rangle + \frac{L_i}{2} \|z - x\|^2, \tag{33}
\]

where

\[
z = x - \frac{1}{L_i}(\nabla f_i(x) - \nabla f_i(x^*)). \tag{34}
\]

Substituting this in \(z\) into (33) gives

\[
f_i(x^*) - f_i(x) = \langle \nabla f_i(x^*), x^* - x + \frac{1}{L_i}(\nabla f_i(x) - \nabla f_i(x^*)) \rangle - \frac{1}{L_i} \langle \nabla f_i(x), \nabla f_i(x) - \nabla f_i(x^*) \rangle + \frac{1}{2L_i} \|\nabla f_i(x) - \nabla f_i(x^*)\|^2
\]

\[
= \langle \nabla f_i(x^*), x^* - x \rangle - \frac{1}{L_i} \|\nabla f_i(x) - \nabla f_i(x^*)\|^2 + \frac{1}{2L_i} \|\nabla f_i(x) - \nabla f_i(x^*)\|^2
\]

\[
= \langle \nabla f_i(x^*), x^* - x \rangle - \frac{1}{2L_i} \|\nabla f_i(x) - \nabla f_i(x^*)\|^2.
\]

Now we present a corollary of the previous lemma for over-parametrized functions. We now develop an immediate consequence of each \(f_i\) being convex around \(x^*\) and smooth along the \(x^*\)-direction.

**Corollary 18** Suppose these exists \(x^* \in \mathbb{R}^d\) where

\[
x^* \in \text{arg min}\left\{ f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\}.
\]

Suppose the interpolated Assumption 2 holds. Furthermore, suppose that for each \(f_i\) there exists \(L_i\) such that

\[
f_i \left( x - \frac{1}{L_i} \nabla f_i(x) \right) \leq f_i(x) - \frac{1}{2L_i} \|\nabla f_i(x)\|^2. \tag{35}
\]

It follows for every \(i \in \{1, \ldots, n\}\) that

\[
\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 \leq 2L_i(f_i(x) - f_i(x^*)). \quad \forall x \in \mathbb{R}^d. \tag{36}
\]

**Proof** Note that for interpolated functions we have that each \(f_i\) is convex around \(x^*\). Furthermore, since each \(\nabla f_i(x^*) = 0\) we have that (29) holds, and thus \(f_i\) is smooth in the \(x^*\)-direction according to Definition 15. Finally all the conditions of Lemma 17 holds, and thus so does (36) holds.
Appendix C. Proofs of results on Expected Residual

C.1. Proof of Lemma 4

Proof Using
\[ \|g(x) - \nabla f(x)\|^2 \leq 2 \|g(x) - g(x^*) - \nabla f(x)\|^2 + 2 \|g(x^*)\|^2, \]
and taking expectation together with (ER) and \( \nabla f(x^*) = 0 \) gives
\[ \mathbb{E} \left[ \|g(x) - \nabla f(x)\|^2 \right] \leq 4 \rho(f(x) - f(x^*)) + 2 \mathbb{E}_D \|g(x^*)\|^2. \]
Taking the supremum over \( x^* \in X^* \) and using Assumption 1 and that \( \mathbb{E} \left[ \|X - \mathbb{E}[X]\|^2 \right] = \mathbb{E} \left[ \|X\|^2 \right] - \|\mathbb{E}[X]\|^2 \) with \( X = g(x) \) gives (7).

C.2. Proof of Proposition 5 and its expansion to all samplings.

In this section we give an expanded version of Proposition 5 that also gives bounds for the Expected Smoothness assumption (ES), a closely related assumption to the Expected Residual condition.

Assumption 4 (Expected smoothness) We say that the stochastic gradient \( g \) satisfy the expected smoothness assumption if for all \( x \in \mathbb{R}^d \), there exists \( L = L(g) > 0 \) such that
\[ \mathbb{E}_D \left[ \|g(x) - g(x^*)\|^2 \right] \leq 2L \left( f(x) - f(x^*) \right). \] (ES)
We use \( g \in ES(L) \) as shorthand for expected smoothness.

Here we show that a sufficient condition for the expected smoothness and the expected residual conditions 4 and 3 to hold if that each \( f_i \) is convex around \( x^* \) and smooth. Furthermore, we give tight bounds on the expected smoothness \( L \) and the expected residual constant \( \rho \) for when \( v \) is an independent sampling and, in particular, a \( b \)-minibatch sampling.

In the main text our minibatch results are stated only for \( b \)-minibatching. But they actually hold for a large family of sampling that we refer to as the independent samplings.

Definition 19 (Independent sampling) Let \( S \subset \{1, \ldots, n\} \) be a random set and let \( v = \sum_{i \in S} \frac{1}{p_i} e_i \) which is a sampling vector. Suppose there exists a constant \( c_2 > 0 \) such that
\[ \frac{\mathbb{P}[i, j \in S]}{\mathbb{P}[i \in S] \mathbb{P}[j \in S]} = c_2, \quad \forall i, j \in \{1, \ldots, n\}, i \neq j. \] (37)
In [11] it was proven that an independent sampling vector is indeed a valid sampling vector. For completeness we also give the proof in In Lemma 23. Furthermore, all the samplings presented in [11] are examples of an independent sampling vector. In particular the minibatch sampling in Definition 1 is also an independent sampling. Finally, note that (37) does not imply that \( i \in S \) and \( j \in S \) are independent events unless \( c_2 = 1 \). Indeed, for \( b \)-minibatch sampling we have that \( \mathbb{P}[i \in S] = \frac{b}{n} = \mathbb{P}[j \in S] \) and \( \mathbb{P}[i, j \in S] = \frac{b}{n} \frac{b-1}{n-1} \) and thus they are not independent events yet satisfy (37) with \( c_2 = \frac{n}{b} \frac{b-1}{n-1} \).

The following Proposition is based on the proof of Proposition 3.8 in [11] with the exception that now we show that only convexity around \( x^* \) is required for the proof to follow, as opposed to assuming convexity everywhere.
Proposition 20  Let $f$ be a finite sum problem $f = \frac{1}{n} \sum_{i=1}^{n} f_i$. Let $f_i$ be $L_i$-smooth and convex around $x^*$ according to (15) and (31), respectively. It follows that

1. If $v$ is a sampling vector then the expected smoothness and expected residual conditions hold $g \in ES(\mathcal{L})$ and $g \in ER(\rho)$ with $\mathcal{L} = \max_v \frac{1}{n} \sum_{i=1}^{n} L_i v_i = \rho$.

2. If $v$ is an independent sampling vector according to Definition 19 then we have that

$$\mathcal{L} = c_2 L + \max_{i=1,\ldots,n} \frac{L_i}{np_i} (1 - p_i c_2).$$

$$\rho = \frac{\lambda_{\max}(\mathbb{E}[(v-1)(v-1)^T])}{n} L_{\max}$$

3. If $v$ is the $b$-minibatch sampling with replacement then

$$\sigma^2 = \frac{1}{b} \frac{n-b}{n-1} \sigma_1^2$$

$$\rho = \frac{1}{b} \frac{n-b}{n-1} L_{\max}$$

$$\mathcal{L} = \frac{n-b}{b} \frac{1}{n-1} L + \frac{1}{b} \frac{n-b}{n-1} L_{\max}.$$  

Proof

1. Assume that $v$ is any sampling vector. Since $f_i$ is $L_i$-smooth and convex around $x^*$ we have that by multiplying each side of

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

$$f_i(x^*) - f_i(x) \leq \langle \nabla f_i(x^*), x^* - x \rangle,$$

by $v_i/n$ and summing up over $i = 1, \ldots, n$ bearing in mind that $v_i \geq 0$ we have that

$$f_v(z) - f_v(x) \leq \langle \nabla f_v(x), z - x \rangle + \frac{1}{n} \sum_{i=1}^{n} v_i L_i \| z - x \|^2$$

$$f_v(x^*) - f_v(x) \leq \langle \nabla f_v(x^*), x^* - x \rangle.$$

Consequently $f_v$ is convex and $x^*$ and is $L_v$-smooth where $L_v := \frac{1}{n} \sum_{i=1}^{n} v_i L_i$. Applying Lemma 17 we thus have that

$$\| \nabla f_v(x) - \nabla f_v(x^*) \|^2 \leq L_v (f_v(x) - f_v(x^*) - \langle \nabla f_v(x^*), x - x^* \rangle), \ \forall x \in \mathbb{R}^d.$$  

(43)

Taking expectation gives

$$\mathbb{E} \left[ \| \nabla f_v(x) - \nabla f_v(x^*) \|^2 \right] \leq \mathbb{E} \left[ L_v (f_v(x) - f_v(x^*) - \langle \nabla f_v(x^*), x - x^* \rangle) \right]$$

$$\leq \max_v L_v \mathbb{E} \left[ (f_v(x) - f_v(x^*) - \langle \nabla f_v(x^*), x - x^* \rangle) \right]$$

$$= \max_v L_v (f(x) - f(x^*)).$$

This proves that the expected smoothness assumption holds with $\mathcal{L} = \max_v L_v$. Consequently by Theorem 6 we have that the expected residual condition holds with $\rho = \mathcal{L}$. 

18
2. Assume that $v_i$ is an independent sampling. First we prove (38).

Since $f_i$ is $L_i$–smooth and convex around $x^*$ we have that $f$ is $L$–smooth and convex around $x^*$ and by Lemma 17

$$
\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 \leq 2L_i(f_i(x) - f_i(x^*) - \langle \nabla f_i(x^*), x - x^* \rangle)
$$

(44)

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

(45)

Noticing that

$$
\|\nabla f_v(x) - \nabla f_v(x^*)\|^2 = \frac{1}{n^2} \left| \sum_{i \in S} \frac{1}{p_i} (\nabla f_i(x) - \nabla f_i(x^*)) \right|^2
$$

$$
= \sum_{i,j \in S} \left( \frac{1}{np_i} (\nabla f_i(x) - \nabla f_i(x^*)), \frac{1}{np_j} (\nabla f_j(x) - \nabla f_j(x^*)) \right),
$$

we have

$$
\mathbb{E}[\|\nabla f_v(x) - \nabla f_v(x^*)\|^2] = \sum_{C} \sum_{i,j \in C} \left( \frac{1}{np_i} (\nabla f_i(x) - \nabla f_i(x^*)), \frac{1}{np_j} (\nabla f_j(x) - \nabla f_j(x^*)) \right)
$$

$$
= \sum_{i,j=1}^{n} \sum_{C : i \in C} p_C \left( \frac{1}{np_i} (\nabla f_i(x) - \nabla f_i(x^*)), \frac{1}{np_j} (\nabla f_j(x) - \nabla f_j(x^*)) \right)
$$

$$
= \sum_{i,j=1}^{n} \frac{\mathbb{P}[i,j \in S]}{p_i p_j} \left( \frac{1}{n} (\nabla f_i(x) - \nabla f_i(x^*)), \frac{1}{n} (\nabla f_j(x) - \nabla f_j(x^*)) \right),
$$

where we used a double counting argument in the 2nd equality. Now since $\mathbb{P}[i,j \in S] / (p_i p_j) = c_2$ for $i \neq j$. Recalling that $\mathbb{P}[i,i \in S] = p_i$ we have from the above that

$$
\mathbb{E}[\|\nabla f_v(x) - \nabla f_v(x^*)\|^2] = \sum_{i \neq j} \frac{c_2}{n} \left( \frac{1}{n} (\nabla f_i(x) - \nabla f_i(x^*)), \frac{1}{n} (\nabla f_j(x) - \nabla f_j(x^*)) \right)
$$

$$
+ \sum_{i=1}^{n} \frac{1}{n^2 p_i} \|\nabla f_i(x) - \nabla f_i(x^*)\|^2
$$

$$
= \sum_{i,j=1}^{n} \frac{c_2}{n} \left( \frac{1}{n} (\nabla f_i(x) - \nabla f_i(x^*)), \frac{1}{n} (\nabla f_j(x) - \nabla f_j(x^*)) \right)
$$

$$
+ \sum_{i=1}^{n} \frac{1}{n^2 p_i} (1 - p_i c_2) \|\nabla f_i(x) - \nabla f_i(x^*)\|^2
$$

$$
\leq \frac{c_2}{n} \|\nabla f(x) - \nabla f(x^*)\|^2
$$

$$
+ 2 \sum_{i=1}^{n} \frac{1}{n^2 p_i} (1 - p_i c_2) \left( f_i(x) - f_i(x^*) - \langle \nabla f_i(x^*), x - x^* \rangle \right)
$$

$$
\leq 2 \left( c_2 L + \max_{i=1, \ldots, n} \frac{L_i}{p_i} \right) \left( f(x) - f(x^*) - \langle \nabla f(x^*), x - x^* \rangle \right).
$$
Comparing the above to the definition of expected smoothness (ES) we have that
\[ \mathcal{L} \leq c_2 L + \max_{i=1,\ldots,n} \frac{L_i}{n p_i} (1 - p_i c_2). \] (46)

Now we will prove that
\[ \mathbb{E} \left[ \left\| \nabla f_v(w) - \nabla f_v(x^*) - (\nabla f(w) - \nabla f(x^*)) \right\|^2 \right] \leq 2 \rho (f(w) - f(x^*)), \] (47)
holds with the constant given in (39). First we expand the squared norm on the left hand side of (47). Define \( DF(w) = [\nabla f_1(w), \ldots, \nabla f_n(w)] \in \mathbb{R}^{d \times n} \) as the Jacobian of \( F(w) \) de\( f \) \([f_1(w), \ldots, f_n(w)] \). We denote \( R := (DF(w) - DF(x^*)) \). It follows that
\[
C := \| \nabla f_v(w) - \nabla f_v(x^*) - (\nabla f(w) - \nabla f(x^*)) \|^2 \\
= \frac{1}{n^2} \| (DF(w) - DF(x^*)) (v - 1) \|^2 \\
= \frac{1}{n^2} \langle (R(v - 1), R(v - 1))_{\mathbb{R}^d} \rangle \\
= \frac{1}{n^2} \text{Trace} \left( (v - 1)^\top R^\top R (v - 1) \right) \\
= \frac{1}{n^2} \text{Trace} \left( R^\top R (v - 1)(v - 1)^\top \right). 
\]

Let \( \text{Var} [v] := \mathbb{E} \left[ (v - 1)(v - 1)^\top \right] \). Taking expectation,
\[
\mathbb{E} [C] = \frac{1}{n^2} \text{Trace} \left( R^\top R \text{Var} [v] \right) \\
\leq \frac{1}{n^2} \text{Trace} \left( R^\top R \right) \lambda_{\max} (\text{Var} [v]). \] (48)

Moreover, since the \( f_i \)'s are convex around \( x^* \) and \( L_i \)-smooth, it follows from (32) that
\[
\text{Trace} \left( R^\top R \right) = \sum_{i=1}^n \| \nabla f_i(w) - \nabla f_i(x^*) \|^2 \\
\leq 2 \sum_{i=1}^n L_i (f_i(w) - f_i(x^*) - (\nabla f_i(x^*), w - x^*)) \\
\leq 2 n L_{\max} (f(w) - f(x^*)). \] (49)

Therefore,
\[
\mathbb{E} [C] \overset{(48)+(49)}{\leq} 2 \frac{\lambda_{\max} (\text{Var} [v])}{n} L_{\max} (f(w) - f(x^*)). \] (50)

Which means
\[ \rho = \frac{\lambda_{\max} (\text{Var} [v])}{n} L_{\max}. \] (51)

3. Finally, if \( v \) is a \( b \)-minibatch sampling, the specialized expressions for \( \mathcal{L} \) in (42) follows by observing that \( \mathbb{P} \{ i \in S \} = p_i = \frac{b}{n}, \mathbb{P} \{ i, j \in S \} = \frac{b \cdot (b - 1)}{n^2} \) and consequently \( c_2 = \frac{n b - 1}{b n - 1} \). The specialized expressions for \( \sigma \) and \( \rho \) in (40) and (41) follow from Proposition 3.8 [11] and Lemma F.3 in [51], respectively.

\[ \blacksquare \]
C.3. Proof of Theorem 6

First we include the formal definition of each of these assumptions named in Theorem 6. Let $g(x) = \nabla f_i(x)$ denote the stochastic gradient. The results in this section carry over verbatim by using $g(x) = \nabla f_i(x)$ and $f_i = f_i$ instead, where $v$ is a sampling vector. But since the sampling only affects the constants in each of the forthcoming assumptions, and here we are only interested in a hierarchy between assumptions, we omit the proof for a general sampling vector.

First we repeat the definitions of ES, WGC and SGC from Assumption 2 [19], Assumption 2.1 in [11], Eq (7) and Eq (2) in [54], respectively.

SGC: Strong Growth Condition. We say that SGC holds with $\rho_{SGC} > 0$ if

$$E\left[\|g(x)\|^2\right] \leq \rho_{SGC} \|\nabla f(x)\|^2.$$  (52)

WGC: Weak Growth Condition. We say that WGC holds with $\rho_{WGC} > 0$ if

$$E\left[\|g(x)\|^2\right] \leq 2\rho_{WGC}(f(x) - f(x^*)).$$  (53)

ES: Expected Smoothness. We say that ES holds with $L > 0$ if

$$E\left[\|g(x) - g(x^*)\|^2\right] \leq 2L(f(x) - f(x^*)).$$  (54)

ER: Expected Residual. We say that ER holds with $\rho > 0$ if

$$E\left[\|g(x) - g(x^*) - (\nabla f(x) - \nabla f(x^*))\|^2\right] \leq 2\rho(f(x) - f(x^*)).$$  (55)

In addition we will use

x*–convex. We say that $x^*$–convex holds if

$$f_i(x^*) - f_i(x) \leq \langle \nabla f_i(x^*), x^* - x \rangle, \text{ for } i = 1, \ldots, n.$$  (56)

$L_i$–smoothness. We say that $L_i$–smoothness holds for $L_i > 0$ if

$$f_i(z) - f_i(x) \leq \langle \nabla f_i(x), z - x \rangle + \frac{L_i}{2}\|z - x\|^2, \forall x, z \in \mathbb{R}^d, i = 1, \ldots, n.$$  (57)

Interpolated. We say that the interpolation condition holds at $x^*$ if

$$f_i(x^*) \leq f_i(x), \text{ for } i = 1, \ldots, n, \text{ and for every } x \in \mathbb{R}^d.$$  (58)

Now we repeat the statement of Theorem 6 for convenience.

**Theorem 21** The following hierarchy holds

\[
\begin{array}{c}
SGC + L\text{-smooth} \quad \Rightarrow \quad WGC \quad \Rightarrow \quad ES \quad \Rightarrow \quad ER
\end{array}
\]

\[
L_i + \text{Interpolated} \quad \Rightarrow \quad L_i + x^*\text{-convex}
\]

**In addition we have** that $ES(L) + PL(\mu) \Rightarrow ER(L - \mu)$ and $ER \not\Rightarrow ES.
SGD for Structured Nonconvex Functions

**Proof** We first prove the top row of implications.

1. **SGC + L-smooth** $\implies$ **WGC**. Using Lemma 16 and (52) we have that

$$\mathbb{E} \left[ \|g(x)\|^2 \right] \leq \rho_{SGC} \|\nabla f(x)\|^2 \overset{(30)}{=} 2L\rho_{SGC}(f(x) - f(x^*)) .$$

Thus (53) holds with $\rho_{WGC} = 2L\rho_{SGC}$.

2. **WGC** $\implies$ **ES**.

Plugging in $x = x^*$ in WGC (53) gives $g(x^*) = 0$ almost surely. Since $g(x^*) = 0$ we have that (53) gives (54).

3. **ES** $\implies$ **ER**. Expanding the squares of the left hand side of (55) gives

$$\|g(x) - g(x^*) - (\nabla f(x) - \nabla f(x^*))\|^2 = \|g(x) - g(x^*)\|^2 + \|\nabla f(x) - \nabla f(x^*)\|^2 - 2 \langle g(x) - g(x^*), \nabla f(x) - \nabla f(x^*) \rangle.$$ 

Now assuming that **ES** (54) holds, taking expectation and using that $\mathbb{E} [g(x)] = \nabla f(x)$ we have that

$$\mathbb{E} \left[ \|g(x) - g(x^*) - (\nabla f(x) - \nabla f(x^*))\|^2 \right] = \mathbb{E} \left[ \|g(x) - g(x^*)\|^2 - \|\nabla f(x) - \nabla f(x^*)\|^2 \right]$$

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

In addition, if the PL condition holds, then we can upper bound $\|\nabla f(x) - \nabla f(x^*)\|^2 \leq -2\mu(f(x) - f^*)$ which combined with the above gives

$$\mathbb{E} \left[ \|g(x) - g(x^*) - (\nabla f(x) - \nabla f(x^*))\|^2 \right] \leq 2(\mathcal{L} - \mu)(f(x) - f^*) .$$

Thus **ER** holds with $\rho = \mathcal{L} - \mu$.

Now we prove the remaining implications.

4. **L$_i$†Interpolated** $\implies$ **L$_i$† + $x^*$-convex**. A direct consequence of the interpolation assumption (2) is that $\nabla f_i(x^*) = 0$ and $f_i(x^*) \leq f_i(x)$. Consequently $f_i(x^*) \leq f_i(x) + \langle \nabla f_i(x^*), x - x^* \rangle$.

5. **L$_i$† + $x^*$-convex** $\implies$ **ES**. Follows from Proposition 20.

Finally

6. **ER** $\not\implies$ **ES**. Since when $v$ encodes the full batch sampling where $g(x) = \nabla f(x)$, the expected residual condition always holds for any $\rho > 0$ since the left hand side of (ER) is zero and $0 \leq \rho(f(x) - f^*)$. On the other hand, in the full batch case the expected smoothness assumption is equivalent to claiming that $f$ is $L$-smooth, and clearly there exist differentiable functions that have gradients that are not Lipschitz. For instance $f(x) = x^4$.

An important assumption created recently [19] is the following $ABC$-assumption.

---

22
ABC. We say that $ABC$ holds with $A, B, C > 0$ if
\[
\mathbb{E} \left[ \|g(x)\|^2 \right] \leq 2A(f(x) - f(x^*)) + B \|\nabla f(x)\|^2 + C.
\] (59)
The $ABC$ condition (59) includes all previous assumptions SGC, WGC, ES and ER as a special case by choosing the three parameters $A, B$ and $C$ appropriately. In this sense, it is rather a family of assumptions. See [19] for more details on this assumption and how it linked to all the other assumptions.

Appendix D. Proofs of Main Convergence Analysis Results

D.1. Proof of Theorem 7

First we need the following lemma.

Lemma 22 Assume $g \in ER(\rho)$. Then for all $x \in \mathbb{R}^d$,
\[
\mathbb{E}_D \left[ \|g(x)\|^2 \right] \leq 2(2\rho + L)(f(x) - f(x^*)) + 2\sigma^2.
\] (60)

Proof Since $f$ is $L-$smooth, we have $\|\nabla f(x)\|^2 \leq 2L(f(x) - f(x^*))$. Using this inequality together with (7) gives (60).

Proof We have:
\[
\left\| x^{k+1} - x^* \right\|^2 = \left\| x^k - x^* \right\|^2 - 2\gamma_k \langle g(x^k), x^k - x^* \rangle + \gamma_k^2 \|g(x^k)\|^2
\]
Hence, taking expectation conditioned on $x_k$, we have:
\[
\mathbb{E}_D \left[ \left\| x^{k+1} - x^* \right\|^2 \right] = \left\| x^k - x^* \right\|^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle + \gamma_k^2 \mathbb{E}_D \left[ \|\nabla f_k(x_k)\|^2 \right]
\]
\[
\tag{4} \leq \left\| x^k - x^* \right\| - 2\gamma_k (\zeta - \gamma_k (2\rho + L))(f(x^k) - f^*)) + 2\gamma_k^2 \sigma^2.
\]
Rearranging and taking expectation, we have
\[
2\gamma_k (\zeta - \gamma_k (2\rho + L)) \mathbb{E} \left[ f(x^k) - f^* \right] \leq \mathbb{E} \left[ \left\| x^k - x^* \right\|^2 \right] - \mathbb{E} \left[ \left\| x^{k+1} - x^* \right\|^2 \right] + 2\gamma_k^2 \sigma^2.
\]
Summing over $k = 0, \ldots, t - 1$ and using telescopic cancellation gives
\[
2 \sum_{k=0}^{t-1} \gamma_k (\zeta - \gamma_k (2\rho + L)) \mathbb{E} \left[ f(x_k) - f^* \right] \leq \left\| x^0 - x^* \right\|^2 - \mathbb{E} \left[ \left\| x^{t+1} - x^* \right\|^2 \right] + 2\sigma^2 \sum_{k=0}^{t-1} \gamma_k^2.
\]
Since $\mathbb{E} \left[ \left\| x^k - x^* \right\|^2 \right] \geq 0$ and $(\zeta - \gamma_k (2\rho + L)) \geq 0$, dividing both sides by $2 \sum_{i=0}^{t-1} \gamma_i (\zeta - \gamma_i (2\rho + L))$ gives:
\[
\sum_{k=0}^{t-1} \mathbb{E} \left[ \frac{\gamma_k (\zeta - \gamma_k (2\rho + L))}{\gamma_i (\zeta - \gamma_i (2\rho + L))} \left( f(x^k) - f^* \right) \right] \leq \frac{\left\| x^0 - x^* \right\|^2}{2 \sum_{i=0}^{t-1} \gamma_i (\zeta - \gamma_i (2\rho + L))} + \frac{\sigma^2 \sum_{i=0}^{t-1} \gamma_i^2}{\sum_{i=0}^{t-1} \gamma_i (\zeta - \gamma_i (2\rho + L))}.
\]
Thus,

\[
\min_{k=0,\ldots,t-1} \mathbb{E} \left[ f(x^k) - f(x^*) \right] \leq \frac{1}{2} \sum_{i=0}^{t-1} \left\| x^i - x^* \right\|^2 + \frac{\sigma^2 \sum_{i=0}^{t-1} \gamma_i^2}{\sum_{i=0}^{t-1} \gamma_i (\zeta - \gamma_i (2\rho + L))}.
\]

For the different choices of step sizes:

1. If \( \forall k \in \mathbb{N}, \gamma_k = \frac{1}{2} \frac{\zeta}{2(2\rho + L)} \), then it suffices to replace \( \gamma_k = \gamma \) in (11).

2. Suppose algorithm (3) is run for \( T \) iterations. Let \( \forall k = 0,\ldots,T - 1, \gamma_k = \frac{\gamma}{\sqrt{k+1}} \) with \( \gamma \leq \frac{\zeta}{2(2\rho + L)} \). Notice that since \( \gamma \leq \frac{\zeta}{2(2\rho + L)} \), we have \( \zeta - \gamma (2\rho + L) \leq \frac{1}{2} \). Then it suffices to replace \( \gamma_k = \frac{\gamma}{\sqrt{k+1}} \) in (11).

3. Let \( \forall k \in \mathbb{N}, \gamma_k = \frac{\gamma}{\sqrt{k+1}} \) with \( \gamma \leq \frac{\zeta}{2(2\rho + L)} \). Note that that since \( \gamma_k = \frac{\gamma}{\sqrt{k+1}} \) and using the integral bound, we have that

\[
\sum_{t=0}^{k-1} \gamma_t^2 = \gamma^2 \sum_{t=0}^{k-1} \frac{1}{t+1} \leq \gamma^2 (\log(k) + 1). \tag{61}
\]

Furthermore using the integral bound again we have that

\[
\sum_{t=0}^{k-1} \gamma_t \geq 2\gamma \left( \sqrt{k} - 1 \right). \tag{62}
\]

Now using (61) and (62) we have that

\[
\sum_{i=0}^{k-1} \gamma_i (\zeta - \gamma_i (2\rho + L)) = \zeta \sum_{i=0}^{k-1} \gamma_i - (2\rho + L) \sum_{i=0}^{k-1} \gamma_i^2 \geq 2\gamma \left( \zeta (\sqrt{k} - 1) - \gamma (\rho + \frac{L}{2}) (\log(k) + 1) \right).
\]

It remains to replace bound the sums in (11) by the values we have computed.

D.2. Proof of Corollary 8

Proof The interpolated assumption 2 implies that \( \nabla f_i(x^*) = g(x^*) = 0 \) and thus \( \sigma = 0 \). Furthermore from (10) we have that the ER condition holds with \( \rho = L_{\max} \frac{n-b}{(n-1)b} \). Combining these two observations with (12) gives (15). The total complexity (16) follows from computing the iteration complexity via (15) and multiplying it by \( b \).

Finally for the optimal minibatch size, since (16) is a linear function in \( b \), the minimum depends on the sign of its slope. Taking the derivative in \( b \) we have the slope is given by \( 2 \frac{L - 2L_{\max}}{n-1} \). If the slope is negative, we want \( b \) to be a large as possible, that is \( b = n \). Otherwise if the slope is positive \( b = 1 \) is optimal.
D.3. Proof of Lemma 9

Before presenting our proof for Lemma 24, we need to present a large family of sampling vectors called the arbitrary samplings.

**Lemma 23 (Lemma 3.3 [11])**  Let $S \subset \{1, \ldots, n\}$ be a random set. Let $\mathbb{P}[i \in S] = p_i$. It follows that $v = \sum_{i \in S} \frac{1}{p_i} e_i$ is a sampling vector. We call $v$ the arbitrary sampling vector.

An arbitrary sampling is sufficiently flexible as to model almost all samplings and minibatching schemes of interest, see Section 3.2 in [11]. For example the $b$–minibatch sampling is a special case where $p_i = \frac{b}{n}$ and $\mathbb{P}[S = B] = 1 / \binom{n}{b}$ for every $B \in \{1, \ldots, n\}$ that has $b$ elements.

Now we prove Lemma 9 and some additional results.

**Lemma 24**  Assume interpolation 2 holds. Let $f_i$ be $L_i$–smooth and let $v$ be a sampling vector as defined in Lemma 23. It follows that there exists $L_{\text{max}} > 0$ such that

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

(63)

For $B \subset \{1, \ldots, n\}$ let $L_B$ be the smoothness constant of $f_B := \frac{1}{n} \sum_{i \in B} p_i f_i$. It follows that

1. If $v$ is an arbitrary sampling vector (Lemma 23) then $L_{\text{max}} = \max_{i=1, \ldots, n} \frac{1}{\sum_{B; i \in B} \frac{p_i}{L_B}}$.

2. If $v$ is the $b$–minibatch sampling then $L_{\text{max}} = L_{\text{max}}(b) = \max_{i=1, \ldots, n} \sum_{B; i \in B} \frac{p_2}{L_B^{-1}}$.

**Proof**  Since $f_i$ is $L_i$–smooth, we have that $f_v$ is $L_v$–smooth with $L_v := \frac{1}{n} \sum_{i=1}^n v_i L_i$. Thus according to Lemma 17 we have that

$$\|\nabla f_v(x)\|^2 \leq 2 L_v (f_v(x) - f_v^*).$$

Consequently we have that

$$\frac{1}{\|\nabla f_v(x)\|^2} \geq \frac{1}{2L_v (f_v(x) - f_v^*)}.$$ (64)

Using this we have the following bound

$$\mathbb{E} \left[ \frac{(f_v(x) - f_v^*)^2}{\|\nabla f_v(x)\|^2} \right] \geq \mathbb{E} \left[ \frac{f_v(x) - f_v^*}{2L_v} \right].$$ (65)

Let $S$ be the random set associated to the arbitrary sampling vector $v$. We use $B \subset \{1, \ldots, n\}$ to denote a realization of $S$ and $p_B := \mathbb{P}[B = S]$. Thus with this notation we have that

$$\mathbb{E} \left[ \frac{(f_v(x) - f_v^*)^2}{\|\nabla f_v(x)\|^2} \right] \geq \sum_{B \subset \{1, \ldots, n\}} p_B \frac{f_B(x) - f^*_B}{2L_B}.$$ (66)

Now let $p_i := \mathbb{P}[i \in S]$. Due to the interpolation condition we have that and the definition of $f_B$ we have that

$$f_B^* = f_B(x^*) = \frac{1}{n} \sum_{i \in B} p_i f_i(x^*) = \frac{1}{n} \sum_{i \in B} p_i f_i^*.$$
Consequently

$$\mathbb{E} \left[ \| f(x) - f^* \|^2 \right] \quad (66)$$

\[
= \sum_{B \subseteq \{1, \ldots, n\}} P_B \sum_{i \in B} \frac{f_i(x) - f_i^*}{2nL_{B,i}} \\
= \frac{1}{2n} \sum_{i=1}^{n} \sum_{B : i \in B} \frac{P_B}{p_i L_B} (f_i(x) - f_i^*) \\
\geq \min_{i=1, \ldots, n} \left\{ \sum_{B : i \in B} \frac{P_B}{p_i L_B} \right\} \frac{1}{2n} \sum_{i=1, \ldots, n} (f_i(x) - f_i^*) \\
= \frac{1}{2} \min_{i=1, \ldots, n} \left\{ \sum_{B : i \in B} \frac{P_B}{p_i L_B} \right\} (f(x) - f^*), \quad (67)
\]

where in the first equality we used a double counting argument to switch the order of the sum over subsets $B$ and elements $i \in B$. The main result $63$ now follows by observing that

$$\frac{1}{\min_{i=1, \ldots, n} \left\{ \sum_{B : i \in B} \frac{P_B}{p_i L_B} \right\}} = \max_{i=1, \ldots, n} \left\{ \frac{p_i}{\sum_{B : i \in B} \frac{P_B}{L_B}} \right\} = L_{\max}.$$

Finally, for a $b$–minibatch sampling we have that

$$p_i = \frac{b}{n}, \quad p_B = 1/ \left( \frac{n}{b} \right) \quad \text{and} \quad L_B \leq \frac{1}{b} \sum_{j \in B} L_j,$$

which in turn gives

$$\frac{1}{L_{\max}} = \min_{i=1, \ldots, n} \sum_{B : i \in B} \frac{1}{\frac{b}{n} L_B} = \min_{i=1, \ldots, n} \sum_{B : i \in B} \frac{1}{b-1} \frac{1}{L_B}.$$

\[\square\]

**D.4. Proof of Theorem 10**

**Proof**

$$\| x^{k+1} - x^* \|^2 = \| x^k - \gamma_k \nabla f_v(x^k) - x^* \|^2 \leq \| x^k - x^* \|^2 - 2 \gamma_k \langle x^k - x^*, \nabla f_v(x^k) \rangle + \gamma_k^2 \| \nabla f_v(x^k) \|^2 \quad (4)$$

$$\leq \| x^k - x^* \|^2 - 2 \gamma_k \left[ f_v(x^k) - f_v(x^*) \right] + \gamma_k^2 \| \nabla f_v(x^k) \|^2 \quad (18)$$

$$= \| x^k - x^* \|^2 - 2 \gamma_k \left[ f_v(x^k) - f_v(x^*) \right] + \frac{2\gamma_k}{c} \left[ f_v(x^k) - f_v^* \right]$$

$$= \| x^k - x^* \|^2 - \gamma_k \left( 2 \zeta - \frac{1}{c} \right) \left[ f_v(x^k) - f_v(x^*) \right]. \quad (68)$$

By rearranging we have that

$$\gamma_k \left( 2 \zeta - \frac{1}{c} \right) \left[ f_v(x^k) - f_v(x^*) \right] \leq \| x^k - x^* \|^2 - \| x^{k+1} - x^* \|^2. \quad (69)$$

26
Taking expectation, and since \(2\zeta - \frac{1}{c} > 0\) we have by Lemma 9 we have that
\[
\frac{2c\zeta - 1}{2c^2} \frac{1}{L_{\max}} \mathbb{E} \left[ f(x^k) - f(x^*) \right] \leq \left( 2\zeta - \frac{1}{c} \right) \mathbb{E} \left[ \frac{(f_{\nu}(x) - f_{\nu}^*)^2}{c\|\nabla f_{\nu}(x)\|^2} \right] \\
= \left( 2\zeta - \frac{1}{c} \right) \mathbb{E} \left[ \gamma_k(f_{\nu}(x) - f_{\nu}^*) \right] \\
\leq \mathbb{E} \left[ \|x^k - x^*\|^2 \right] - \mathbb{E} \left[ \|x^{k+1} - x^*\|^2 \right].
\]

Summing from \(k = 0, \ldots, K - 1\) and using telescopic cancellation gives
\[
\frac{2c\zeta - 1}{2c^2} \frac{1}{L_{\max}} \sum_{k=0}^{K-1} \mathbb{E} \left[ f(x^k) - f(x^*) \right] \leq \|x^0 - x^*\|^2 - \mathbb{E} \left[ \|x^K - x^*\|^2 \right].
\]

Multiplying through by \(L_{\max} \frac{2c^2}{2c^2 - 1} \frac{1}{K}\) gives
\[
\min_{i=0, \ldots, K-1} \mathbb{E} \left[ f(x^k) - f(x^*) \right] \leq \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left[ f(x^k) - f(x^*) \right] \leq \frac{2c^2}{2c^2 - 1} \frac{L_{\max}}{K} \|x^0 - x^*\|^2.
\]

\[\blacksquare\]

**D.5. Proof of Theorem 12**

In the following proof, for ease of reference, we repeat the step-size choice here:
\[
\gamma \leq \frac{1}{1 + \frac{2\mu}{\rho} \frac{1}{L}}.
\]

**Proof** By combining the smoothness of function \(f\) with the update rule of SGD we obtain:
\[
f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2 \\
= f(x^k) - \gamma \langle \nabla f(x^k), \nabla f_{\nu}(x^k) \rangle + \frac{L}{2} \|\nabla f_{\nu}(x^k)\|^2.
\]

By taking expectation conditioned on \(x^k\) we obtain:
\[
\mathbb{E} \left[ f(x^{k+1}) \mid x^k \right] \leq f(x^k) - \gamma \|\nabla f(x^k)\|^2 + \frac{L^2}{2} \mathbb{E} \|\nabla f_{\nu}(x^k)\|^2 \\
\leq f(x^k) - \gamma \|\nabla f(x^k)\|^2 + 2L\gamma^2 \rho(f(x^k) - f^*) \\
+ \frac{L^2}{2} \|\nabla f(x^k)\|^2 + L\gamma^2 \sigma^2 \\
= f(x^k) - \gamma (1 - \frac{L\gamma}{2}) \|\nabla f(x^k)\|^2 \\
+ 2L\gamma^2 \rho(f(x^k) - f^*) + L\gamma^2 \sigma^2 \\
\leq f(x^k) - 2\mu\gamma (1 - \frac{L\gamma}{2}) [(f(x^k) - f^*)] \\
+ 2L\gamma^2 \rho(f(x^k) - f^*) + L\gamma^2 \sigma^2.
\]

where the last inequality holds because \(1 - \frac{L\gamma}{2} > 0\) since \(\gamma \leq \frac{1}{1 + \frac{2\mu}{\rho} \frac{1}{L}} < \frac{1}{L}\).
Taking expectations again and subtracting $f^*$ from both sides yields:
\[
\mathbb{E}[f(x^{k+1}) - f^*] \leq \left( 1 - 2\gamma \left( \mu (1 - \frac{L\gamma}{2}) - L\gamma\rho \right) \right) \mathbb{E}[f(x^k) - f^*] + L\gamma^2\sigma^2.
\]

(71)

Recursively applying the above and summing up the resulting geometric series gives:
\[
\mathbb{E} \left[ f(x^k) - f^* \right] \leq \left( 1 - \mu\gamma \right) [f(x^0) - f^*] + L\gamma^2\sigma^2 \sum_{j=0}^{k-1} (1 - \gamma\mu)^j.
\]

(75)

Using $\sum_{i=0}^{k-1} (1 - \mu\gamma)^i = \frac{1 - (1 - \mu\gamma)^k}{1 - (1 - \mu\gamma)} \leq \frac{1}{\mu\gamma}$, in the above gives (22).

**On Iteration Complexity:** For ease of reference, we we repeat the step-size choice for the iteration complexity result
\[
\gamma = \frac{1}{L} \min \left\{ \frac{\mu\epsilon}{2\sigma^2}, \frac{1}{1 + 2\rho/\mu} \right\}.
\]

(76)

To analyze the iteration complexity, let $\epsilon > 0$ and let us divide the right hand side of (22) into two parts and bound each of them separately by $\frac{\epsilon}{2}$. For the right most part we have that
\[
\frac{L\gamma^2\sigma^2}{\mu} \leq \frac{\epsilon}{2} \Rightarrow \gamma \leq \frac{1}{L} \frac{\mu\epsilon}{2\sigma^2}.
\]

(77)

The derivation in (77) gives us the restriction (76) on the step size.

For the other remaining part we have that
\[
(1 - \mu\gamma)^k [f(x^0) - f^*] < \frac{\epsilon}{2}.
\]

Taking logarithms and re-arranging the above gives
\[
\log \left( \frac{2(f(x^0)-f^*)}{\epsilon} \right) \leq k \log \left( \frac{1}{1-\gamma\mu} \right).
\]

(78)

Now using that $\log \left( \frac{1}{\rho} \right) \geq 1 - \rho$, for $0 < \rho \leq 1$ gives
\[
k \geq \frac{1}{\mu\gamma} \log \left( \frac{1}{2(f(x^0)-f^*)} \right).
\]

Thus restricting the step size according to (76) and inserting $\gamma$ into the above gives the result (23).

**D.6. Proof of Theorem 13**

**Proof** By Theorem 6 we have that the ER condition holds. Thus Theorem 12 holds. Furthermore, since $\tilde{f}$ is interpolated we have that $\sigma = 0$, which when combined with Theorem 12 and (23) gives (24).

The total complexity (25) follows by using Lemma 5 and the expression for $\rho$ in (10) and plugging into (24). Since (25) is a linear function in $b$, the minimum depends on the sign of its slope. Taking the derivative in $b$ we have the sign slope is given by $\left( 1 - \frac{2\kappa_{\text{max}}}{n-1} \right)$. If the slope is negative, we want $b$ to be a large as possible, that is $b = n$. Otherwise if the slope is positive $b = 1$ is optimal.

\[\square\]
D.7. Theorem of PL with switching stepsize

**Theorem 25 (Decreasing step sizes/switching strategy)** Let $f$ be an $L$-smooth. Assume $f \in PL(\mu)$ and $g \in ER(\rho)$. Let $k^* := \frac{2\mu}{\rho} \left( 1 + \frac{2\rho}{\mu} \right)$ and

$$
\gamma_k = \begin{cases} 
\frac{\mu}{L(\mu+2\rho)} & \text{for } k \leq \lceil k^* \rceil \\
\frac{2k+1}{(k+1)^2\mu} & \text{for } k > \lceil k^* \rceil 
\end{cases}
$$

(79)

If $k \geq \lceil k^* \rceil$, then SGD given by (3) satisfies:

$$
\mathbb{E}[f(x^k) - f^*] \leq \frac{4L\sigma^2}{\mu^2} \frac{1}{k} + \frac{(k^*)^2}{k+2\rho} \mathbb{E}[f(x^0) - f^*].
$$

(80)

**Proof** Let $\gamma_k := \frac{2k+1}{(k+1)^2\mu}$ and let $k^*$ be an integer that satisfies

$$
\gamma_{k^*} \leq \frac{\mu}{L(\mu+2\rho)}.
$$

(81)

Note that $\gamma_k$ is decreasing in $k$ and consequently $\gamma_k \leq \frac{\mu}{L(\mu+2\rho)}$ for all $k \geq k^*$. This in turn guarantees that (74) holds for all $k \geq k^*$ with $\gamma_k$ in place of $\gamma$, that is

$$
\mathbb{E}[f(x^{k+1}) - f^*] \leq \frac{k^2}{(k+1)^2} \mathbb{E}[f(x^k) - f^*] + \frac{4L\sigma^2}{\mu^2} \frac{(2k+1)^2}{(k+1)^2}.
$$

(82)

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

$$(k+1)^2 \mathbb{E}[f(x^{k+1}) - f^*] \leq k^2 \mathbb{E}[f(x^k) - f^*] + \frac{4L\sigma^2}{\mu^2} \left( \frac{2k+1}{k+1} \right)^2$$

$$
\leq k^2 \mathbb{E}[f(x^k) - f^*] + 4 \frac{L\sigma^2}{\mu^2},
$$

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

$$
\sum_{t=k^*}^{k} [(t+1)^2 \mathbb{E}[f(x^{t+1}) - f^*] - t^2 \mathbb{E}[f(x^t) - f^*]] \leq \sum_{t=k^*}^{k} 4 \frac{L\sigma^2}{\mu^2}.
$$

(83)

Using telescopic cancellation gives

$$(k+1)^2 \mathbb{E}[f(x^{k+1}) - f^*] \leq (k^*)^2 \mathbb{E}[f(x^{k^*}) - f^*] + 4 \frac{L\sigma^2}{\mu^2} (k - k^*)$$

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

$$
\mathbb{E}[f(x^{k+1}) - f^*] \leq \frac{(k^*)^2}{(k+1)^2} \mathbb{E}[f(x^{k^*}) - f^*] + \frac{4L\sigma^2(k-k^*)}{\mu^2(k+1)^2}.
$$

(84)

For $k \leq k^*$ we have that (22) holds, which combined with (84), gives

$$
\mathbb{E}[f(x^{k+1}) - f^*] \leq \frac{(k^*)^2}{(k+1)^2} \left( 1 - \frac{\mu^2}{(\mu+2\rho)L} \right)^{k^*} [f(x^0) - f^*]
$$

$$
+ \frac{4L\sigma^2}{\mu^2(k+1)^2} \left( 4(k - k^*) \frac{(k^*)^2}{\mu^2} \frac{1}{L} \right),
$$

(85)
It now remains to choose $k^*$. Choosing $k^*$ that minimizes the second line of (85) gives $k^* = 2L\left(1 + \frac{2L}{\mu}\right)$. With this choice of $k^*$ it is easy to show that (81) holds. Furthermore, by using that $\frac{2}{k^*} = \frac{\mu^2}{\mu + 2\rho} \frac{1}{L}$ in (85) gives

$$
\mathbb{E}[f(x^{k+1}) - f^*] \leq \frac{(k^*)^2}{(k+1)^2} \left(1 - \frac{2}{k^*}\right) f(x^0) - f^* + \frac{L_2\sigma^2}{\mu^2(k+1)^2} (4(k - k^*) + 2k^*)
$$

where in the second inequality we have used that $\frac{2k - k^*}{k+1} \leq 2\frac{k^*}{k+1} \leq 2$.

D.8. Proofs of Section A

Separable, smooth and PL. Let $f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x_i)$, which are smooth and interpolated. If in addition each $f_i(x_i)$ satisfies the PL condition with constant $\mu_i$ then there exists $x^* \in \mathcal{X}^*$ such that $f(x)$ satisfies the PL condition with $\mu = \min_{i=1,\ldots,n} \frac{\mu_i}{n}$. Indeed since

$$
\|\nabla f(x)\|^2 = \sum_{i=1}^{n} \frac{1}{n} \|\nabla f_i(x_i)\|^2
$$

$$
\geq \sum_{i=1}^{n} \frac{\mu_i}{n} (f_i(x_i) - f_i(x^*))
$$

$$
\geq \min_{i=1,\ldots,n} \frac{\mu_i}{n} (f(x) - f(x^*)).
$$

D.8.1. PROOF OF LEMMA 14

Consider the problem

$$
\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{2n} \|F(x) - y\|^2 = \frac{1}{2n} \sum_{i=1}^{n} (F_i(x) - y_i)^2
$$

where $y \in \mathbb{R}^n$.

Proof The Jacobian of $F$ is given by $DF(x)^\top = [\nabla F_1(x), \ldots, \nabla F_n(x)] \in \mathbb{R}^{d \times n}$. Note that

$$
\nabla f(x) = \frac{1}{n} DF(x)^\top (F(x) - y),
$$

$$
\nabla f_i(x) = \nabla F_i(x)(F_i(x) - y_i).
$$

Consequently $\nabla f(x^*) = \nabla f_i(x^*) = 0$. Finally, we suppose that the $F_i(x)$ functions are Lipschitz and the Jacobian $DF(x)$ has full row rank, that is,

$$
\|\nabla F_i(x)\| \leq u, \quad \forall i \in \{1, \ldots, n\}, \forall k.
$$

$$
\left\|DF(x)^\top v\right\| \geq \ell \|v\|, \quad \forall v.
$$
Under these assumptions, our objective (87) satisfies the PL condition and the expected smoothness condition. Indeed (ES) holds using
\[
\mathbb{E}_i \left[ \left\| \nabla f_i(x) - \nabla f_i(x^*) \right\|^2 \right] = \mathbb{E}_i \left[ \left\| \nabla f_i(x) \right\|^2 \right] = \frac{1}{n} \sum_{i=1}^{n} \left\| \nabla F_i(x) \right\|^2 (F_i(x) - y_i)^2
\]
\[
\leq \frac{u}{n} \sum_{i=1}^{n} (F_i(x) - y_i)^2 = \frac{u}{n} \left\| F(x) - y \right\|^2
\]
\[
= 2u(f(x) - f(x^*)), \tag{92}
\]
where we used (90) in the inequality. This shows that the expected smoothness condition hold with \(u = \mathcal{L} \).

By using the lower bound (91) we have that
\[
\left\| \nabla f(x) \right\|^2 = \frac{1}{n^2} \left\| DF(x)^\top (F(x) - y) \right\|
\]
\[
\geq \frac{1}{n^2} \left\| F(x) - y \right\|^2 \min_v \frac{\left\| DF(x)^\top v \right\|^2}{\left\| v \right\|^2}
\]
\[
\geq \ell f(x) = \ell (f(x) - f(x^*)),
\]
which shows that the PL condition holds with \(\mu = \ell \).

The condition (91) is hard to verify, and somewhat unlikely to hold for all \(x \in \mathbb{R}^d\). Though if we had consider a closed and bounded constraint \(\mathcal{X} \subset \mathbb{R}^d\), and applied the projected SGD method, then (90) is more likely to hold. For instance, assuming that (91) holds in neighborhood of the solution is the typical assumption used to prove the asymptotic convergence of the Gauss-Newton method (see Theorem 10.1 in [57]).

Appendix E. Additional Convergence Analysis Results

E.1. Convergence of SGD for Quasar Strongly Convex functions

In this section we develop specialized theorems for quasar strongly convex functions,

**Definition 26 (Quasar strongly convex)** Let \(\zeta > 0\) and \(\lambda \geq 0\). Let \(x^* \in \mathcal{X}^*\). We that \(f\) is \((\zeta, \mu)\)-quasar strongly convex with respect to \(x^*\) if for all \(x \in \mathbb{R}^n\),
\[
f(x^*) \geq f(x) + \frac{1}{\zeta} \langle \nabla f(x), x^* - x \rangle + \frac{\lambda}{2} \left\| x^* - x \right\|^2. \tag{93}
\]
For shorthand we write \(f \in QSC(\zeta, \lambda)\).

Note that If (93) holds with \(\lambda = 0\) we say that the function \(f\) is \(\zeta\)-quasar convex and we write \(f \in QC(\zeta)\) (same to (4) from the main paper).

E.1.1. Constant Step-size

When \(\lambda = 1\) we say that \(f\) is star-strong convexity, but it is also known in the literature as quasi-strong convexity [11, 34] or weak strong convexity [17]. Star-strong convexity is also used in [34] for the analysis of gradient and accelerated gradient descent and in [11] for the analysis of SGD. The following theorem is a generalization of Theorem 1 in [11] to quasar strongly convex functions and under the assumption of expected residual.
Theorem 27  Let $\zeta > 0$. Assume $f$ is $(\zeta, \lambda)$-quasar-strongly convex with respect to $x^*$ and $g \in ER(\rho)$. Choose $\gamma^k = \gamma \in (0, \frac{\lambda}{(2\rho + L)})$ for all $k$. Then iterates of SGD given by (3) satisfy:

$$
\mathbb{E}\|x^k - x^*\|^2 \leq (1 - \gamma \zeta \lambda)^k \|x^0 - x^*\|^2 + \frac{2\gamma^2 \sigma^2}{\zeta \lambda}.
$$

(94)

Proof Let $r^k = x^k - x^*$. From (3), we have

$$
\|r^{k+1}\|^2 = \|r^k\|^2 - 2\gamma \langle r^k, \nabla f(x^k) \rangle + \gamma^2 \|\nabla f_{\nu^k}(x^k)\|^2.
$$

Taking expectation conditioned on $x^k$ we obtain:

$$
\mathbb{E}_D\|r^{k+1}\|^2 = \|r^k\|^2 - 2\gamma \mathbb{E}_D \langle r^k, \nabla f(x^k) \rangle + \gamma^2 \mathbb{E}_D \|\nabla f_{\nu^k}(x^k)\|^2 \\
\overset{(93)}{\leq} (1 - \gamma \zeta \lambda)\|r^k\|^2 - 2\zeta \gamma [f(x^k) - f(x^*)] + \gamma^2 \mathbb{E}_D \|\nabla f_{\nu^k}(x^k)\|^2.
$$

Taking expectations again and using (60)

$$
\mathbb{E}\|r^{k+1}\|^2 \leq (1 - \gamma \zeta \lambda)\|r^k\|^2 - 2\zeta \gamma [f(x^k) - f(x^*)] + \gamma^2 (2\rho + L)(f(x) - f(x^*)) + 2\gamma^2 \sigma^2 \\
\leq (1 - \gamma \zeta \lambda)\mathbb{E}\|r^k\|^2 + 2\gamma \mathbb{E}_D \|f(x^k) - f(x^*)\| + 2\gamma^2 \sigma^2 \\
\leq (1 - \gamma \zeta \lambda)\mathbb{E}\|r^k\|^2 + 2\gamma^2 \sigma^2,
$$

where we used in the last inequality that $\gamma (2\rho + L) \leq \zeta$ since $\gamma \leq \frac{\zeta}{(2\rho + L)}$. Recursively applying the above and summing up the resulting geometric series gives

$$
\mathbb{E}\|r^k\|^2 \leq (1 - \gamma \zeta \lambda)^k \|r^0\|^2 + 2 \sum_{j=0}^{k-1} (1 - \gamma \zeta \lambda)^j \gamma^2 \sigma^2 \\
\leq (1 - \gamma \zeta \lambda)^k \|r^0\|^2 + \frac{2\gamma \sigma^2}{\zeta \lambda}.
$$

(95)

E.1.2. Stochastic Polyak Step-size (SPS)

Theorem 28  Assume interpolation 2 holds. Let all $f_i$ be $L_i$-smooth and $(\zeta, \lambda)$-quasar strongly convex functions (4) with respect to $x^* \in \mathcal{X}^*$. SGD with SPS with $c = 1/2\zeta$ converges as:

$$
\mathbb{E}\|x^k - x^*\|^2 \leq \left(1 - \frac{\lambda \zeta}{\mathbb{E}[L_i]}\right)^k \|x^0 - x^*\|^2.
$$

(96)

6. This implies that function $f(x) = \sum_{i=1}^n f_i(x)$ is also $(\zeta, \lambda)$-strongly quasar convex function with respect to $x^* \in \mathcal{X}^*$ (see [16]).
Consequently, by selecting \( c \), we have that \( f_v \) is \( L_v \)-smooth where \( L_v := \frac{1}{n} \sum_{i=1}^{n} v_i L_i \). Consequently, taking expectation condition on \( x^k \)

\[
\mathbb{E}_D \| x^{k+1} - x^* \|^2 \leq (1 - \alpha \mathbb{E}_D [\gamma_k]) \| x^k - x^* \|^2 \leq \left(1 - \frac{\lambda}{2 \mathbb{E}[L_v]} \right) \| x^k - x^* \|^2. \tag{98}
\]

Taking expectations again and using the tower property:

\[
\mathbb{E} \| x^{k+1} - x^* \|^2 \leq \left(1 - \frac{\lambda}{2 \mathbb{E}[L_v]} \right) \mathbb{E} \| x^k - x^* \|^2. \tag{99}
\]

**Corollary 29**  If \( v \) is the \( b \)-minibatch sampling, we have that \( \mathbb{E}[L_v] \leq \frac{L n(b-1)}{(n-1)b} + L_{\max} \frac{n-b}{(n-1)b} \).

Consequently, by selecting \( c = \frac{1}{2\lambda} \) and given \( \epsilon > 0 \), if we take

\[
k \geq \left( \frac{L}{\lambda} \frac{n(b-1)}{(n-1)b} + \frac{L_{\max}}{\lambda} \frac{n-b}{(n-1)b} \right) \log \left( \frac{\| x^0 - x^* \|^2}{\epsilon} \right),
\]

steps of SGD with the SPS step size then \( \mathbb{E} \| x^k - x^* \|^2 \leq \epsilon \).

**Proof** Since the interpolation condition implies that \( f_i \) is convex around \( x^* \), we have by Proposition 20 that the expected smoothness condition ES holds with \( \mathcal{L}(b) \) given in (42). Furthermore, we have that from Lemma E.1 in [11] that \( \mathbb{E}[L_v] \leq \mathcal{L}(b) \). Finally, from (96) we have that the iteration complexity is given by

\[
k \geq \frac{2 \epsilon \mathbb{E}[L_v]}{\lambda} \log \left( \frac{\| x^0 - x^* \|^2}{\epsilon} \right).
\]

Plugging in \( c = \frac{1}{2\lambda} \) and the upperbound (42) for \( \mathcal{L}(b) \) gives the result.
E.2. Convergence Analysis Results under Expected Smoothness

E.2.1. Quasar Convex and Expected Smoothness

For this next theorem we first need the following lemma.

Lemma 30  Suppose $f$ satisfies the expected smoothness Assumption 54. It follows that

$$
\mathbb{E}_D \left[ \|g(x)\|^2 \right] \leq 4\mathcal{L}(f(x) - f^*) + 2\sigma^2,
$$

(100)

Proof Using

$$
\|g(x)\|^2 \leq 2 \|g(x) - g(x^*)\|^2 + 2 \|g(x^*)\|^2,
$$

and taking the supremum over $x^* \in \mathcal{X}$ and expectation together with (ES) gives the result. 

Theorem 31  Assume $f(x)$ is $\zeta$-quasar-convex (4) and $g \in \text{ES}(\mathcal{L})$. Let $0 < \gamma_k < \frac{\zeta}{4\mathcal{L}}$ for all $k \in \mathbb{N}$. Then, for every $x^* \in \mathcal{X}$

$$
\min_{t=0,\ldots,k-1} \mathbb{E} \left[ f(x^t) - f(x^*) \right] \leq \frac{\|x^0 - x^*\|^2}{2\sum_{i=0}^{k-1} \gamma_i (\zeta - 2\gamma\mathcal{L})} + \sigma^2 \frac{\sum_{i=0}^{k-1} \gamma_i^2}{\sum_{i=0}^{k-1} \gamma_i (\zeta - 2\gamma\mathcal{L})},
$$

(101)

Moreover:

1. if $\forall k \in \mathbb{N}, \gamma_k = \gamma \leq \frac{\zeta}{2\mathcal{L}}, \text{then } \forall k \in \mathbb{N}$,

$$
\min_{t=0,\ldots,k-1} \mathbb{E} \left[ f(x^t) - f(x^*) \right] \leq \frac{\|x^0 - x^*\|^2}{2\gamma (\zeta - 2\gamma\mathcal{L})} + \frac{\gamma \sigma^2}{\zeta - 2\gamma\mathcal{L}},
$$

(102)

2. suppose algorithm (3) is run for $T$ iterations. If $\forall k = 0, \ldots, T-1, \gamma_k = \frac{\gamma}{\sqrt{T}}$ with $\gamma \leq \frac{\zeta}{4\mathcal{L}}$,

$$
\min_{t=0,\ldots,T-1} \mathbb{E} \left[ f(x^t) - f(x^*) \right] \leq \frac{\|x^0 - x^*\|^2 + 2\gamma^2 \sigma^2}{\gamma \sqrt{T}},
$$

(103)

3. $\forall k \in \mathbb{N}, \gamma_k = \frac{\gamma}{\sqrt{k+1}} \text{ with } \gamma \leq \frac{\zeta}{4\mathcal{L}}, \text{then } \forall k \in \mathbb{N}$,

$$
\min_{t=0,\ldots,k-1} \mathbb{E} \left[ f(x^t) - f(x^*) \right] \leq \frac{\|x^0 - x^*\|^2 + 2\gamma^2 \sigma^2 (\log(k)+1)}{4\gamma (\zeta (\sqrt{k+1}) - \gamma \mathcal{L} (\log(k)+1))} \sim O \left( \frac{\log(k)}{\sqrt{k}} \right).
$$

(104)

Proof We have:

$$
\|x^{k+1} - x^*\|^2 = \|x^k - x^*\|^2 - 2\gamma_k \langle \nabla g(x^k), x^k - x^* \rangle + \gamma_k^2 \|g(x^k)\|^2.
$$

Hence, taking expectation conditioned on $x_k$, we have:

$$
\mathbb{E}_D \left[ \|x^{k+1} - x^*\|^2 \right] \leq \|x^k - x^*\|^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle + \gamma_k^2 \mathbb{E}_D \left[ \|\nabla f(x_k)\|^2 \right] \overset{(4)+(100)}{\leq} \|x^k - x^*\|^2 - 2\gamma_k (\zeta - 2\gamma\mathcal{L}) (f(x^k) - f^*) + 2\gamma_k^2 \sigma^2.
$$
Rearranging and taking expectation, we have
\[ 2\gamma_k (\zeta - 2\gamma_k \mathcal{L}) \mathbb{E} \left[ f(x^k) - f^* \right] \leq \mathbb{E} \left[ \|x^k - x^*\|^2 \right] - \mathbb{E} \left[ \|x^{k+1} - x^*\|^2 \right] + 2\gamma_k^2 \sigma^2. \]

Summing over \( k = 0, \ldots, t - 1 \) and using telescopic cancellation gives
\[ 2 \sum_{k=0}^{t-1} \gamma_k (\zeta - 2\gamma_k \mathcal{L}) \mathbb{E} \left[ f(x_k) - f^* \right] \leq \|x^0 - x^*\|^2 - \mathbb{E} \left[ \|x^k - x^*\|^2 \right] + 2\sigma^2 \sum_{k=0}^{t-1} \gamma_k^2. \]

Since \( \mathbb{E} \left[ \|x^k - x^*\|^2 \right] \geq 0 \), dividing both sides by \( 2 \sum_{i=1}^{t} \gamma_i (\zeta - 2\gamma_i \mathcal{L}) \) gives:
\[ \sum_{k=0}^{t-1} \mathbb{E} \left[ \frac{\gamma_k (\zeta - 2\gamma_k \mathcal{L})}{\sum_{i=0}^{k} \gamma_i (\zeta - 2\gamma_i \mathcal{L})} (f(x^k) - f^*) \right] \leq \frac{\|x^0 - x^*\|^2}{2 \sum_{i=0}^{t} \gamma_i (\zeta - 2\gamma_i \mathcal{L})} + \frac{\sigma^2 \sum_{i=0}^{t-1} \gamma_i^2}{\sum_{i=0}^{t} \gamma_i (\zeta - 2\gamma_i \mathcal{L})}. \]

Thus,
\[ \min_{k=0, \ldots, t-1} \mathbb{E} \left[ f(x^k) - f(x^*) \right] \leq \frac{\|x^0 - x^*\|^2}{2 \sum_{i=0}^{t} \gamma_i (\zeta - 2\gamma_i \mathcal{L})} + \frac{\sigma^2 \sum_{i=0}^{t-1} \gamma_i^2}{\sum_{i=0}^{t} \gamma_i (\zeta - 2\gamma_i \mathcal{L})}. \]

For the different choices of step sizes:

1. If \( \forall k \in \mathbb{N}, \gamma_k = \gamma \leq \frac{\zeta}{4\mathcal{L}} \), then it suffices to replace \( \gamma_k = \gamma \) in (101).

2. Suppose algorithm (3) is run for \( T \) iterations. Let \( \forall k = 0, \ldots, T - 1, \gamma_k = \frac{\gamma}{\sqrt{T}} \) with \( \gamma \leq \frac{\zeta}{4\mathcal{L}} \).
   Notice that since \( \gamma \leq \frac{\zeta}{4\mathcal{L}} \), we have \( \zeta - 2\gamma \mathcal{L} \leq \frac{1}{2} \). Then it suffices to replace \( \gamma_k = \frac{\gamma}{\sqrt{T}} \) in (101).

3. Let \( \forall k \in \mathbb{N}, \gamma_k = \frac{\gamma}{\sqrt{k+1}} \) with \( \gamma \leq \frac{\zeta}{4\mathcal{L}} \). Note that that since \( \gamma_k = \frac{\gamma}{\sqrt{k+1}} \) and using the integral bound, we have that
   \[ \sum_{t=0}^{k-1} \gamma_t^2 = \gamma^2 \sum_{t=0}^{k-1} \frac{1}{t+1} \leq \gamma^2 (\log(k) + 1). \]
   Furthermore using the integral bound again we have that
   \[ \sum_{t=0}^{k-1} \gamma_t \geq 2\gamma \left( \sqrt{k} - 1 \right). \]

Now using (105) and (106) we have that
\[ \sum_{i=0}^{k-1} \gamma_i (\zeta - 2\gamma_i \mathcal{L}) = \zeta \sum_{i=0}^{k-1} \gamma_i - 2\mathcal{L} \sum_{i=0}^{k-1} \gamma_i^2 \geq 2\gamma \left( \zeta (\sqrt{k} - 1) - \gamma \mathcal{L} (\log(k) + 1) \right). \]

It remains to replace bound the sums in (101) by the values we have computed.
**Specialized results for Interpolated Functions (with expected smoothness)**  Analogously to Corollary 8, when the interpolated Assumption 2 holds, we can drop the expected smoothness assumption in Theorem 31 in lieu for standard smoothness assumptions.

**Theorem 32**  Let $f_i(x)$ be $L_i$-smooth for $i = 1, \ldots, n$, $f(x)$ be $\zeta$-quasar-convex (4) and assume that the interpolated Assumption 2 holds. Consequently $g \in ES(\mathcal{L})$. If we use the step size

$$\gamma_k \equiv \gamma \leq \frac{\zeta}{2\mathcal{L}},$$

for all $k$, then SGD given by (3) converges

$$\min_{i=1,\ldots,k} \mathbb{E}\left[f(x^i) - f^*\right] \leq \frac{1}{k} \frac{\|x^0 - x^*\|^2}{2(\zeta - 2\gamma \mathcal{L})}.$$  \hspace{1cm} (108)

Hence, if $\gamma = \frac{\zeta}{4\mathcal{L}}$ and given $\epsilon > 0$ we have that

$$k \geq \frac{4\zeta^{\epsilon}}{\epsilon^2} \|x^0 - x^*\|^2,$$

implies $\min_{i=1,\ldots,k} \mathbb{E}\left[f(x^i) - f^*\right] \leq \epsilon.$

**Proof** By Theorem 6 we have that the ES condition holds. Thus Theorem 31 holds. Furthermore, since $f$ is interpolated we have that $\sigma = 0$, which when combined with (102) from Theorem 31 gives the result.  \hspace{1cm} \blacksquare

Specializing Theorem 32 to the full batch setting, we have that gradient descent (GD) with step size $\gamma = \frac{\zeta}{4\mathcal{L}}$ converges at a rate of

$$f(x^t) - f(x^*) \leq \frac{4L\|x^0 - x^*\|^2}{\zeta^2 k},$$

where we have used that $\mathcal{L} = L$ in the full batch setting and the smoothness of $f$ guarantees that the sequences $f(x^1), \ldots, f(x^t)$ for GD is a decreasing sequence. Similar to the result of the main paper on GD, we note that this is exactly the rate given recently for GD for quasar-convex functions in [12], with the exception that we have a squared dependency on $\xi$ the quasar convex function.

**Quasar-strongly convex functions and Expected smoothness**  Similar to Theorem 27 we present below the convergence of SGD with constant step-size for $(\zeta, \lambda)$-quasar-strongly convex functions under the expected smoothness.

**Theorem 33**  Let $\zeta > 0$. Assume $f$ is $(\zeta, \lambda)$-quasar-strongly convex and that $(f, D) \sim ES(\mathcal{L})$. Choose $\gamma_k = \gamma \in (0, \frac{\zeta}{2\mathcal{L}})$ for all $k$. Then iterates of SGD given by (3) satisfy:

$$\mathbb{E}\|x^k - x^*\|^2 \leq (1 - \gamma \zeta \lambda)^k \|x^0 - x^*\|^2 + \frac{2\gamma \sigma^2}{\zeta \lambda}.$$  \hspace{1cm} (110)

**Proof** Similar to the proof of Theorem 27 but using ES instead of ER.  \hspace{1cm} \blacksquare
E.2.2. PL AND EXPECTED SMOOTHNESS.

In this section we present four main theorems for the convergence of SGD with constant and decreasing step size. Through our results we highlight the limitations of using expected smoothness in the PL setting and explain why one needs to have the expected residual to prove efficient convergence.

**Theorem 34** Let \( f(x) \) be \( L \)-smooth and PL function and that \( g \in ES(\mathcal{L}) \). Choose \( \gamma^k = \gamma \leq \frac{\mu}{2L} \) for all \( k \). Then SGD given by (3) converges as follows:

\[
\mathbb{E}[f(x^k) - f^*] \leq (1 - \gamma \mu)^k [f(x^0) - f^*] + \frac{\gamma \sigma^2}{\mu} \tag{111}
\]

**Proof** By Proposition 21 we have that the expected smoothness condition holds with \( \rho = L - \mu \). Thus by Theorem 12 we have that with a step size

\[
\gamma_k = \gamma \leq \frac{1}{1 + 2\rho / \mu} \frac{1}{L} = \frac{1}{1 + 2(L - \mu) / \mu} \frac{1}{L} = \frac{\mu}{2L \mathcal{L}}
\]

the iterates converge according to (111).

**Limitation of Theorem 34.** Let us consider the case where \( |S| = n \) with probability one. That is, each iteration (3) uses the full batch gradient. Thus \( \sigma = 0 \) and the expected smoothness parameter becomes \( \mathcal{L} = L \). Consequently, from Theorem 34 we obtain:

\[
\mathbb{E}[f(x^k) - f^*] \leq (1 - \gamma \mu)^k [f(x^0) - f^*]. \tag{112}
\]

For \( \gamma = \frac{\mu}{2L \mathcal{L}} \) (larger possible value) the rate of the gradient descent is \( \rho = 1 - \frac{\mu^2}{2L} \). Thus the resulting iteration complexity (number of iterations to achieve given accuracy) for gradient descent becomes \( k \geq 2L^2 / \mu^2 \). However it is known that for minimizing PL functions, the iteration complexity of gradient descent method is \( k \geq 2L / \mu \). Thus the result of Theorem 34 give as a suboptimal convergence for gradient descent and the gap between the predicted behavior and the known results could potentially be very large.

E.3. Minibatch Corollaries without Interpolation

In this section we state the corollaries for the main theorems when \( v \) is a \( b \) minibatch sampling. Differently than what we did in the main paper, we will not assume interpolation. Instead, we will use the weaker assumptions that each \( f_i \) is \( x^* \)-convex.

E.3.1. QUASAR CONVEX

**Corollary 35** Assume \( f \) is \( \zeta \)-quasar-strongly convex and that each \( f_i \) is \( L_i \)-smooth and \( x^* \)-convex. If \( v \) is a \( b \)-minibatch sampling and \( \gamma_k \equiv \frac{1}{2} \frac{\zeta}{2 \rho + L} \) then,

\[
\min_{t=0, \ldots, k-1} \mathbb{E} [f(x^t) - f(x^*)] \leq 2 \|x^0 - x^*\|^2 \frac{2L_{\max} \frac{n-b}{n-1} k + L}{\zeta^2 k} + \frac{\frac{1}{b} \frac{n-b}{n-1} \sigma^2}{2L_{\max} \frac{n-b}{n-1} b + L}. \tag{113}
\]

**Proof** By Theorem 6 we have that the ER condition holds. Thus, the main Theorem 7 holds. Replacing the constants \( \rho \) and \( \sigma \) by their corresponding minibatch constants in (10) gives the result.

E.3.2. PL Function

**Corollary 36** Let \( b \in \{1, \ldots, n\} \) and let \( v \) be a \( b \)-minibatch sampling with replacement. Furthermore let each \( f_i \) be \( L_i \)-smooth and convex around \( x^* \). If \( f \) satisfies the PL condition (5), then by Theorem 12 if

\[
\gamma = \frac{\mu(n-1)b}{\mu(n-1)b + 2L_{\text{max}}(n-b)\frac{1}{L}},
\]

then

\[
\mathbb{E}\left[f(x^k) - f^*\right] \leq \left(1 - \frac{\mu^2(n-1)b}{\mu(n-1)b + 2L_{\text{max}}(n-b)\frac{1}{L}}\right)^k (f(x^0) - f^*) + \frac{n-b}{\mu(n-1)b + 2L_{\text{max}}(n-b)\sigma^2}.
\]

**Proof** The proof follows by plugging in the values of \( \rho \) and \( \sigma \) given in Proposition 5 into (76) and (22).\]