Convergence Analysis of SGD under Expected Smoothness

Yuta Kawamoto
Meiji University
ee227008@meiji.ac.jp

Hideaki Iiduka
Meiji University
iiduka@cs.meiji.ac.jp

Abstract

Stochastic gradient descent (SGD) is the workhorse of large-scale learning, yet classical analyses rely on assumptions that can be either too strong (bounded variance) or too coarse (uniform noise). The *expected smoothness (ES)* condition has emerged as a flexible alternative that ties the second moment of stochastic gradients to the objective value and the full gradient. This paper presents a self-contained convergence analysis of SGD under ES. We (i) refine ES with interpretations and sampling-dependent constants; (ii) derive bounds of the expectation of squared full gradient norm; and (iii) prove $O(1/K)$ rates with explicit residual errors for various step-size schedules. All proofs are given in full detail in the appendix. Our treatment unifies and extends recent threads (Khaled and Richtárik, 2020; Umeda and Iiduka, 2025).

1 Introduction

1.1 Background

Stochastic Gradient Descent (SGD) has become the *de facto* standard for large-scale optimization in modern machine learning, thanks to its remarkable scalability, modest memory requirements, and ability to cope with the nonconvex landscapes typical of deep learning (Bottou et al., 2018; Kiefer and Wolfowitz, 1952). Although today it is mostly associated with training neural networks, SGD originated in the early days of stochastic approximation(Bottou et al., 2018; Rumelhart et al., 1986; Hornik et al., 1989).

Historical foundations. The story begins with the seminal work of Robbins and Monro (1951), who introduced a stochastic approximation method to solve root-finding problems with noisy measurements. Shortly afterward, Kiefer and Wolfowitz (1952) developed a gradient-free variant. These works laid the probabilistic and algorithmic foundations of SGD. In the 1960s, Polyak (1964) introduced momentum acceleration(Polyak, 1964; Nesterov, 1983; Sutskever et al., 2013), and later Polyak and Juditsky (1992) together with Ruppert established the power of iterate averaging (Polyak–Ruppert averaging) by proving improved asymptotic efficiencyPolyak and Juditsky (1992); Glorot and Bengio (2010); He et al. (2015). By the 1990s, SGD had become a standard tool in statistics, signal processing, and control theory.

Classical bounded variance assumption. Analyses of SGD traditionally rely on the *bounded variance* condition (Nemirovski et al., 2009), requiring

$$\exists \sigma \geq 0 \ \forall \boldsymbol{x} \in \mathbb{R}^d \ \mathbb{E}_v \|g_v(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2 \leq \sigma^2, \quad (1)$$

where $\boldsymbol{x} \in \mathbb{R}^d$ is a parameter of a deep neural network (DNN), f is the objective function used in training the DNN, and $g_v(\boldsymbol{x})$ is the stochastic gradient of f at $\boldsymbol{x}$, and $\mathbb{E}_v[\cdot]$ is the expectation over randomness v. This assumption yields clean rates in convex optimization but ignores that variance often depends on the iterate. In modern over-parameterized models, variance can be much larger away from the optimum, violating bounded variance (Bottou et al., 2018).

Variance growth models. To overcome this issue, researchers have proposed *variance growth conditions* (Schmidt et al., 2013). A representative form is

$$\exists A \exists B \ \mathbb{E}_v \|g_v(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2 \leq A\|\nabla f(\boldsymbol{x})\|^2 + B,$$

which allows the variance to shrink near stationary points but grow with gradient magnitude. This condition was instrumental in analyzing variance-reduced methods such as SAG (Schmidt et al., 2013), SVRG (Johnson and Zhang, 2013), and SAGA (Defazio et al., 2014), which combine stochastic updates with memory or control variates to achieve linear convergence in convex settings.

Strong growth condition. A more aggressive relaxation is the *strong growth condition*, requiring

$$\exists \rho > 0 \ \forall \boldsymbol{x} \in \mathbb{R}^d \ \mathbb{E}_v \|g_v(\boldsymbol{x})\|^2 \leq \rho \|\nabla f(\boldsymbol{x})\|^2. \quad (2)$$

This implies that stochastic gradients are well aligned with the true gradient. The condition underpins recent analyses in over-parameterized learning (Vaswani et al., 2019), but is overly restrictive in heterogeneous data regimes, such as federated learning, where client gradients may deviate substantially.

Expected Smoothness: a unifying framework. The limitations of earlier assumptions motivated the *Expected Smoothness (ES)* framework (Gower et al., 2019; Khaled and Richtárik, 2020), which posits that there exist $A, B, C \geq 0$ such that, for all $\boldsymbol{x} \in \mathbb{R}^d$,

$$\mathbb{E}_v \|g_v(\boldsymbol{x})\|^2 \leq 2A(f(\boldsymbol{x}) - f^\star) + B\|\nabla f(\boldsymbol{x})\|^2 + C, \ (3)$$

where $f^\star$ is the minimum value of f over $\mathbb{R}^d$. The ES condition is also called the *ABC condition* (Gorbunov et al., 2023), since it uses positive constants A, B, and C. The ES condition subsumes all previous ones: bounded variance ($A = B = 0$), variance growth ($B = 0$), and strong growth ($A = 0$). Crucially, ES introduces a dependence on suboptimality $f(\boldsymbol{x}) - f^\star$, reflecting the empirical observation that stochastic noise decays as the iterates approach a minimizer. This makes ES both theoretically powerful and practically realistic, encompassing problems such as least-squares regression, logistic regression, and many machine learning models.

Our work leverages the ES framework to provide fully explicit, sampling-specific convergence guarantees in terms of the smoothness constants $\{L_i\}$, sampling probabilities, and the data heterogeneity $\Delta^{\inf}$. Unlike prior analyses in Nonconvex World (Khaled and Richtárik, 2020), which required knowledge of the total iteration count K for constant step sizes, our approach supports constant, harmonic, polynomial, and cosine-annealing step sizes without K-dependent tuning.

To make these features intuitive, Table 1 summarizes the typical convergence rates for the four common step-size schedules under ES. These schedules exhibit distinct asymptotic behaviors, highlighting the flexibility and efficiency of our approach(Loshchilov and Hutter, 2017; Goyal et al., 2017; Smith and Topin, 2017; Smith and Le, 2017).

Modern developments. The adoption of ES has enabled sharp convergence results in nonconvex optimization, including tight descent lemmas, explicit rates under constant or decaying step sizes, and improved analyses of distributed and federated SGD (Stich, 2019; Karimireddy et al., 2020; Umeda and Iiduka, 2025). At the same time, adaptive methods such as AdaGrad (Duchi et al., 2011), RMSProp(Hinton, 2012),Adadelta(Zeiler, 2012), and Adam (Kingma and Ba, 2015) have further expanded the scope of stochastic optimization, although their convergence properties are less well understood (Reddi et al., 2018). Recent studies have continued to refine the picture, connecting ES to the Polyak–Łojasiewicz conditions (Karimi et al., 2016), analyzing stochastic methods with momentum (Allen-Zhu and Li, 2019), and exploring batch-size scaling laws (Hardt et al., 2016; Keskar et al., 2017; Dinh et al., 2017).

1.2 Motivation

Classical analyses of SGD typically relied on restrictive variance assumptions, such as the bounded variance condition (1) or the strong growth condition (2). While mathematically convenient, these assumptions fail to capture the heterogeneous and data-dependent noise observed in modern large-scale optimization.

The ES framework (3) has recently emerged as a more realistic and unifying model of stochastic gradient noise (Gower et al., 2019; Khaled and Richtárik, 2020). ES characterizes stochastic gradient variance via three constants (A, B, C) that depend on the problem geometry and the sampling strategy. This framework subsumes earlier variance models and has become a central tool for analyzing SGD in nonconvex settings.

Despite this progress, there remain two important gaps in the literature that motivate the present study:

Limitation 1 of Khaled and Richtárik (2020): Khaled and Richtárik (Khaled and Richtárik, 2020) established sharp convergence guarantees for SGD under ES, showing it is a weaker yet more expressive than the classical assumptions. However, their analysis is primarily descriptive: it introduces ES and proves rates under general step-size policies, but it does not provide a *self-contained and systematic derivation* of the descent lemmas and convergence proofs tailored to various practical step-size schedules. As a result, it remains difficult for readers and practitioners to directly trace how ES constants (A, B, C) interact with specific algorithmic choices, such as of constant, polynomially decaying, or cosine step sizes.

Limitation 2 of Umeda and Iiduka (2025): Umeda and Iiduka (Umeda and Iiduka, 2025) advocate increasing the batch size together with the learning rate, and they show that such joint scaling can accelerate SGD in theory and practice. Their work highlights scheduler design as an important performance factor. Nevertheless, their analysis is not explicitly grounded in the ES framework: the admissibility of scaling rules is derived via upper-bound comparisons between sched-

Table 1: Typical convergence rates of SGD with different step-size schedules under ES (3) and L-smoothness. The constants are omitted for simplicity; see the main text for detailed ES-dependent bounds.

Step-sizes Schedule	Typical Choice	Asymptotic Rate
Constant	$\eta_k = \eta < 2/(LB)$	$\min_{0\leq k\leq K} \mathbb{E}\|\nabla f(\boldsymbol{x}_k)\|^2 = O(1/K) + O(\eta)$
Harmonic	$\eta_k = \eta/(k+1)$	$\min_{0\leq k\leq K} \mathbb{E}\|\nabla f(\boldsymbol{x}_k)\|^2 = O(\log(K)/K)$
Polynomial Decay	$\eta_k = \eta/(k+1)^\alpha$, $0.5 < \alpha < 1$	$\min_{0\leq k\leq K} \mathbb{E}\|\nabla f(\boldsymbol{x}_k)\|^2 = O(1/K^{1-\alpha})$
Cosine Annealing	$\eta_k = \eta \cdot \frac{1}{2}(1 + \cos(\pi k/K))$	$\min_{0\leq k\leq K} \mathbb{E}\|\nabla f(\boldsymbol{x}_k)\|^2 = O(1/K) + O(\eta)$

uler classes, without connecting these prescriptions to the sampling-dependent constants (A, B, C) that govern the true variance–descent tradeoff. This leaves open how to systematically justify scheduler design from the ES perspective.

Our motivation. The present paper addresses these two gaps by providing a *self-contained and systematic convergence analysis of SGD under ES*. Specifically, we (i) revisit the meaning of the ES constants and collect explicit formulas for common sampling schemes, and (ii) establish convergence rates for a variety of step-size policies including constant, decaying, polynomial, and cosine schedules. In so doing, we bridge the descriptive analysis of Khaled & Richtárik (Khaled and Richtárik, 2020) with the constructive scheduler insights of Umeda & Iiduka (Umeda and Iiduka, 2025), offering both rigorous proofs and practical guidance. This contribution deepens our theoretical understanding of SGD under ES and provides a principled foundation for the design of step-size and batch-size schedules in large-scale nonconvex optimization.

1.3 Contributions

Building on the ES paradigm, this paper provides a unified and self-contained analysis of SGD in nonconvex optimization. Our contributions can be summarized as follows:

- We provide detailed interpretations of the ES constants (A, B, C) under various sampling strategies (with and without replacement, importance sampling, mini-batching) and heterogeneous data distributions (**Proposition 1**). Unlike (Khaled and Richtárik, 2020), which treated (A, B, C) as abstract constants, we derive explicit formulas showing how they scale with batch size and heterogeneity, making ES a practical tool for understanding the variance–descent tradeoff.

- We establish a global convergence theorem (**Theorem 1**) for SGD under the ES condition (3). Compared with prior work (Khaled and Richtárik, 2020), our derivations are more self-contained and yield sharper constants, directly connecting (A, B, C) to admissible step-size choices. This systematic treatment recovers classical results as special cases while extending them to a broader family of step-size schedules.

- Beyond the general nonconvex rate bounds in (Khaled and Richtárik, 2020), we analyze concrete step-size rules—including constant, polynomially decaying, and cosine schedules—within the ES framework (**Table 1** and **Corollary 1**). This systematic comparison clarifies the tradeoffs between variance reduction and descent speed, and provides practitioners with explicit guidance on how to select or adapt step sizes under different noise regimes.

- We conduct experiments on modern deep-learning tasks (Srivastava et al., 2014; Ioffe and Szegedy, 2015; Glorot and Bengio, 2010) to show that the ES model captures the evolving two-phase structure of stochastic noise during training (**Section 4**). Such an empirical validation is largely absent in (Khaled and Richtárik, 2020), which focused on theoretical guarantees, but is essential for demonstrating the practical relevance of ES-based analyses.

In summary, whereas prior work established ES as a general analytical tool, our contribution is to make ES *operational*: we provide explicit constants, self-contained proofs, systematic step-size analyses, and empirical demonstrations. This bridges the descriptive theory of (Khaled and Richtárik, 2020) with constructive and practical guidance for modern stochastic optimization.

2 Stochastic Minimization Problem and Assumptions

We start with mathematical preliminaries. Let $\mathbb{R}^d$ be a d-dimensional Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and its induced norm $\|\cdot\|$. In this paper, $\boldsymbol{x} \in \mathbb{R}^d$ denotes

a parameter of a DNN model and $\boldsymbol{x}_k$ denotes the k-th approximation of a minimizer of an objective function f (e.g. the empirical loss f defined by $f = \frac{1}{n}\sum_{i=1}^n f_i$, where f_i is the loss function corresponding to the i-th labeled training data). The operator norm of a matrix X is denoted by $\|X\|_{\mathrm{op}}$.

$f\colon \mathbb{R}^d \to \mathbb{R}$ is bounded below if there exists $f^\star \in \mathbb{R}$ such that, for all $\boldsymbol{x} \in \mathbb{R}^d$, $f(\boldsymbol{x}) \geq f^\star$. Let $\nabla f\colon \mathbb{R}^d \to \mathbb{R}^d$ be the gradient of a differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$. Let $L > 0$. f is said to be L-smooth if ∇f is L-Lipschitz continuous; that is, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$. The Hessian of a twice continuously differentiable function f is denoted by $\nabla^2 f(\boldsymbol{x})$. When f is twice continuously differentiable, f is L-smooth if and only if, for all $\boldsymbol{x}$, $\|\nabla^2 f(\boldsymbol{x})\|_{\mathrm{op}} \leq L$

Let $\mathbb{E}_v[X]$ be the expectation of X with respect to a random variable v. $\mathbb{E}_v[X]$ may also be written as $\mathbb{E}_v X$. Let $\mathbb{E}$ be the total expectation.

Stochastic Minimization Problem This paper considers the following stochastic minimization problem (Khaled and Richtárik, 2020, Problem (12)): Let n be the number of training samples and $f_i\colon \mathbb{R}^d \to \mathbb{R}$ ($i \in \{1, \cdots, n\}$) be the loss function corresponding to the i-th labeled training data. Then, we would like to

$$\text{minimize } f(\boldsymbol{x}) \coloneqq \mathbb{E}_{\boldsymbol{v}}[f_{\boldsymbol{v}}(\boldsymbol{x})] \text{ over } \mathbb{R}^d, \tag{4}$$

where $\boldsymbol{v} \coloneqq (v_1, \cdots, v_n)^\top$ comprises n identically distributed variables with $\mathbb{E}_{v_i}[v_i] = 1$ and is independent of $\boldsymbol{x}$, and

$$f_{\boldsymbol{v}}(\boldsymbol{x}) \coloneqq \frac{1}{n}\sum_{i=1}^n v_i f_i(\boldsymbol{x}). \tag{5}$$

Note that $f_v(\boldsymbol{x})$ is a loss function chosen randomly from $\{f_1(\boldsymbol{x}), \cdots, f_n(\boldsymbol{x})\}$.

Below, We analyze the convergence of SGD under the ES assumption (3). The analysis requires a set of standard assumptions on the objective function and the stochastic gradient. We will also make use of standard inequalities for L-smooth functions throughout.

Assumption 1 (L-smoothness)**.** *A function $f\colon \mathbb{R}^d \to \mathbb{R}$ is L-smooth, which leads to the following descent lemma:*

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2. \tag{6}$$

Assumption 2 (Lower boundedness)**.** *$f(\boldsymbol{x})$ has a lower bound $f^\star$ such that*

$$f^\star \coloneqq \inf_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x}) > -\infty.$$

Let us consider minimizing f defined by (4) by using an optimizer generating a sequence $(\boldsymbol{x}_k)_{k=0}^K$, where K is the number of steps. We assume that f_i is twice continuously differentiable. Since $\|\nabla^2 f_i(\cdot)\|_{\mathrm{op}}$ is continuous and $(\boldsymbol{x}_k)_{t=0}^K$ contains a certain compact set, there exists $L_i > 0$ such that, for all k, $\|\nabla^2 f_i(\boldsymbol{x}_k)\|_{\mathrm{op}} \leq L_i$, which implies that, for all k, $f_i(\boldsymbol{x}_k)$ is L_i-smooth. Hence, $f(\boldsymbol{x}_k)$ is L-smooth with $L = \frac{1}{n}\sum_{i=1}^n L_i$, which implies that Assumption 1 holds for all k. Moreover, in the case where f is convex with $f^\star = -\infty$, there are no stationary points of f, which implies that no optimizer ever finds stationary points of f. Hence, Assumption 2 is natural in a situation in which optimizers minimize f.

Assumption 3 (Unbiasedness)**.** *The stochastic gradient $g_v(\boldsymbol{x}) = \nabla f_v(\boldsymbol{x})$ is an unbiased estimator of f; that is, for all $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\mathbb{E}_v[g_v(\boldsymbol{x})] = \nabla f(\boldsymbol{x}).$$

This assumption is satisfied under various canonical settings. For a rigorous justification, we refer to the stochastic reformulation framework detailed in (Khaled and Richtárik, 2020, Proposition 3). In their framework, the stochastic gradient $g(\boldsymbol{x})$ is constructed as $\nabla f_v(\boldsymbol{x})$ using a sampling vector $\boldsymbol{v}$ designed to satisfy $\mathbb{E}[v_i] = 1$ for all i, which directly ensures the unbiasedness condition $\mathbb{E}[g(\boldsymbol{x})] = \nabla f(\boldsymbol{x})$.

Assumption 4 (Expected Smoothness (Khaled and Richtárik, 2020))**.** *There exist constants $A, B, C \geq 0$ such that, for all $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\mathbb{E}_{\boldsymbol{v}}\|\nabla f_{\boldsymbol{v}}(\boldsymbol{x})\|^2 \leq 2A(f(\boldsymbol{x}) - f^\star) + B\|\nabla f(\boldsymbol{x})\|^2 + C, \tag{7}$$

where $f^\star$ is defined as in Assumption 2 and $\nabla f_{\boldsymbol{v}}(\boldsymbol{x})$ is the gradient of $f_{\boldsymbol{v}}$ defined as in (5)*, that is,*

$$\nabla f_{\boldsymbol{v}}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n v_i \nabla f_i(\boldsymbol{x}). \tag{8}$$

On Increasing the Batch Size and Learning Rate Recent work by Umeda and Iiduka (2025) demonstrates that increasing both the batch size τ and the step size (called the learning rate) η can accelerate convergence. Here, we define the minibatch stochastic gradient by $g_{\boldsymbol{v}_k}(\boldsymbol{x}_k) \coloneqq \frac{1}{\tau_k}\sum_{i=1}^{\tau_k} \nabla f_{v_{k,i}}(\boldsymbol{x}_k)$, where $\boldsymbol{v}_k \coloneqq (v_{k,1}, \cdots, v_{k,\tau_k})^\top$ comprises τ_k independent and identically distributed variables and is independent of $\boldsymbol{x}_k$. The following result in (Umeda and Iiduka, 2025, Lemma 2.1) provides a finite-time bound on the expected squared gradient norm generated by minibatch SGD defined by $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k g_{\boldsymbol{v}_k}(\boldsymbol{x}_k)$:

$$\begin{aligned}&\min_{0\leq k\leq K} \mathbb{E}\|\nabla f(\boldsymbol{x}_k)\|^2 \\ &\leq \frac{2(f(\boldsymbol{x}_0) - f^\star)}{(2 - L\overline{\eta})\sum_{k=0}^{K-1}\eta_k} + \frac{L\sigma^2\sum_{k=0}^{K-1}\eta_k^2\tau_k^{-1}}{(2 - L\overline{\eta})\sum_{k=0}^{K-1}\eta_k}.\end{aligned} \tag{9}$$

This inequality (9) shows that the variance term is scaled by the factor τ_k^{-1}. Hence, enlarging the batch size τ_k reduces the variance contribution, allowing the step size η to be increased proportionally while keeping the bound finite. This theoretical mechanism explains why jointly increasing τ and η leads to accelerated convergence in practice, as confirmed by the analysis and experiments in (Umeda and Iiduka, 2025).

Comparing (7) with (9), we see that the variance term in (9) of Umeda and Iiduka (2025) corresponds to the constant C in ES, scaled by $1/\tau_k$ due to minibatch averaging. That is, for the minibatch gradient defined via the sampling vector $\boldsymbol{v}_k = (v_{k,1}, \cdots, v_{k,\tau_k})^\top$,

$$g_{\boldsymbol{v}_k}(\boldsymbol{x}_k) := \frac{1}{\tau_k} \sum_{i=1}^{\tau_k} g_{v_{k,i}}(\boldsymbol{x}_k),$$

$$\text{where } g_{v_{k,i}}(\boldsymbol{x}) := \nabla f_{v_{k,i}}(\boldsymbol{x})$$

with $v_{k,i}$ denoting the index of the i-th selected sample in the minibatch at iteration k, we obtain

$$\mathbb{E}_{\boldsymbol{v}_k} \|g_{\boldsymbol{v}_k}(\boldsymbol{x}_k)\|^2 \leq \frac{2A}{\tau_k}(f(\boldsymbol{x}_k) - f^\star) + \left(1 + \frac{B-1}{\tau_k}\right) \|\nabla f(\boldsymbol{x}_k)\|^2 + \frac{C}{\tau_k}. \tag{10}$$

The detailed derivation is provided in Appendix C.

Insight: This establishes a direct correspondence between Assumption 4 and the bound in (9) of (Umeda and Iiduka, 2025): increasing the batch size τ_k reduces the effective variance term C/τ_k, allowing a proportionally larger learning rate η_k while maintaining stability. Thus, ES provides a natural theoretical explanation for the accelerated convergence via joint scaling of the batch size and learning rate.

Interpretation: (A) The coefficient A ties the variance to the *suboptimality* $f(\boldsymbol{x}) - f^\star$. (B) The coefficient B captures the correlation with the squared gradient norm. (C) The constant C represents an irreducible "baseline variance" from data heterogeneity.

As a direct consequence of Assumptions 1–3, we establish the following useful proposition (Full derivations and justifications are provided in Appendix D).

Proposition 1 (Validity of ES for common sampling)**.** *Consider Problem* (4) *and suppose that Assumption 3 holds and that each component f_i is L_i-smooth and bounded below, which implies Assumptions 1 and 2 hold. Let $\Delta^{\inf} = \frac{1}{n}\sum_{i=1}^n (f^\star - f_i^{\inf})$, where $f^{\inf}$ (resp. $f_i^{\inf}$) is the infimum of f (resp. f_i) over $\mathbb{R}^d$. For the following sampling schemes, there exist finite constants (A, B, C) such that Assumption 4 holds:*

1. ***Sampling with replacement*** *(minibatch of size τ with probabilities q_i):*

$$A = \frac{1}{\tau} \max_{1 \leq i \leq n} \frac{L_i}{q_i},\ B = 1 - \frac{1}{\tau},\ C = 2A\Delta^{\inf}.$$

2. ***Independent sampling without replacement*** *(each index included with probability p_i):*

$$A = \max_{1 \leq i \leq n} \frac{(1-p_i)L_i}{p_i n},\ B = 1,\ C = 2A\Delta^{\inf}.$$

3. ***τ-Nice sampling without replacement*** *(uniform without replacement of size τ):*

$$A = \frac{n-\tau}{\tau(n-1)} \max_{1 \leq i \leq n} L_i,\ B = \frac{n(\tau-1)}{\tau(n-1)},\ C = 2A\Delta^{\inf}.$$

These explicit formulas for (A, B, C) highlight a key advancement over the *Nonconvex World* analysis (Khaled and Richtárik, 2020): unlike previous work (Khaled and Richtárik, 2020), we provide closed-form expressions for the expected smoothness constants under common sampling strategies, even when the component functions f_i have heterogeneous smoothnesses. This allows for precise quantification of the stochastic gradient variance and facilitates more informed choices of step sizes and minibatch sizes in practice.

3 Main Results

We consider solving Problem (4) by using SGD defined for all k by

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k \nabla f_{\boldsymbol{v}}(\boldsymbol{x}_k), \tag{11}$$

where $\boldsymbol{x}_0 \in \mathbb{R}^d$, $\eta_k > 0$ is the step size, and $\nabla f_{\boldsymbol{v}}$ is defined as in (8). The following is our main convergence theorem under the ES assumption (The proof of Theorem thm:1 is given in Appendix E).

Theorem 1 (Convergence guarantee of SGD under ES)**.** *Let Assumptions 1–4 hold. Suppose the step sizes satisfy $\eta_k \in (0, 2/(LB))$. Then, for all $K \geq 1$, SGD* (11) *satisfies that*

$$\min_{0 \leq k \leq K} \mathbb{E}\|\nabla f(\boldsymbol{x}_k)\|^2 \leq \frac{2(f(\boldsymbol{x}_0) - f^\star)}{(2 - LB\eta_{\max}) \sum_{k=0}^{K-1} \eta_k} + \frac{2LA}{2 - LB\eta_{\max}} \cdot \frac{\sum_{k=0}^{K-1} \eta_k^2 \mathbb{E}[f_k - f^\star]}{\sum_{k=0}^{K-1} \eta_k} + \frac{LC}{2 - LB\eta_{\max}} \cdot \frac{\sum_{k=0}^{K-1} \eta_k^2}{\sum_{k=0}^{K-1} \eta_k}, \tag{12}$$

where $\eta_{\max} = \sup_k \eta_k$, $f_k = f(\boldsymbol{x}_k)$.

Unlike the analysis in Khaled and Richtárik (2020), where the convergence rates forr constant step sizes explicitly depend on the total number of iterations K, our results provide fully explicit, sampling-specific convergence bounds directly in terms of the individual smoothness constants $\{L_i\}$, sampling probabilities, and the data heterogeneity measure $\Delta^{\inf}$. This allows the use of flexible step-size schedules such as harmonic decay, polynomial decay, or cosine annealing, without requiring prior knowledge of K and thereby improves on both theoretical guarantees and practical applicability.

Theorem 1 leads to the following corollary (The proof of Corollary 1 is given in Appendix F).

Corollary 1 (Convergence rates of SGD with specific step sizes)**.** *Under the assumptions in Theorem 1, the iterates* $\{\boldsymbol{x}_k\}$ *of SGD* (11) *have the following step-size schedules.*

1. ***Constant step size*** $\eta_k = \eta$*:*

$$\min_{0\leq k\leq K} \mathbb{E}\|\nabla f(\boldsymbol{x}_k)\|^2 = O\left(\frac{1}{K\eta}\right) + O(\eta).$$

2. ***Harmonic step size*** $\eta_k = \frac{\eta}{k+1}$*:*

$$\min_{0\leq k\leq K} \mathbb{E}\|\nabla f(\boldsymbol{x}_k)\|^2 = O\left(\frac{\log K}{K}\right).$$

3. ***Polynomial decay*** $\eta_k = \eta(k+1)^{-\alpha},\ \alpha \in (0,1)$*:*

$$\min_{0\leq k\leq K} \mathbb{E}\|\nabla f(\boldsymbol{x}_k)\|^2 = O\left(\frac{1}{K^{1-\alpha}}\right).$$

4. ***Cosine decay*** $\eta_k = \eta \cdot \frac{1}{2}(1+\cos(\pi k/K))$*:*

$$\min_{0\leq k\leq K} \mathbb{E}\|\nabla f(\boldsymbol{x}_k)\|^2 = O\left(\frac{1}{K}\right) + O(\eta).$$

4 Empirical Validation

The source code for this study is publicly available at an anonymized GitHub repository `https://anonymous.4open.science/r/sgd_es-BB75/README.md`. In this section, we empirically validate our theoretical framework by training a ResNet-18 model on the CIFAR-100 dataset using SGD (Krizhevsky, 2009; He et al., 2016). The results provide three key insights. First, we observe a macroscopic equivalence in performance across different mini-batch sampling strategies, which motivates a deeper analysis. Second, we rigorously confirm that the ES condition accurately models the empirical gradient variance in this practical deep-learning setting. Third, and most critically, we leverage the validated ES model to uncover a non-trivial, two-phase dynamic in the structure of stochastic gradient noise, which provides a microscopic explanation for the macroscopic observations.

4.1 Macroscopic Equivalence of Sampling Strategies

We begin by examining the macroscopic convergence behavior, focusing on the primary metrics of training loss and test accuracy. As illustrated in Figure 1, all four sampling strategies—Independent, Normal, Replacement, and τ-Nice—produce remarkably similar learning curves. Despite the theoretical differences in their ES constants (as derived in Proposition 1), which imply distinct underlying variance structures, these differences do not manifest as significant variations in convergence speed or final model performance. This counter-intuitive outcome strongly suggests the presence of a dominant, underlying dynamic that governs the optimization trajectory, overshadowing the subtler effects of the sampling methods. This observation necessitates a microscopic analysis of the gradient noise to understand the source of this equivalence.

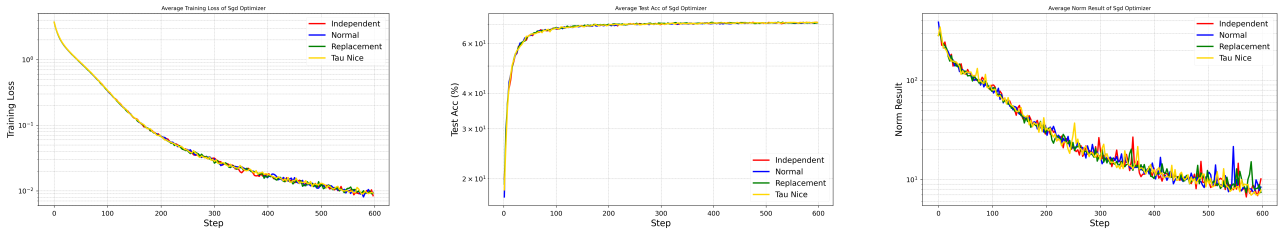
4.2 Empirical Verification of the Expected Smoothness Model

Before utilizing the ES framework to dissect the noise structure, we first establish its empirical validity. The core of our analysis relies on the assumption that the ES condition can accurately model the true variance of stochastic gradients. Figure 2 provides compelling evidence for this assumption. For all four sampling strategies, the predicted variance derived from our fitted ES model (RHS, the red dashed line) almost perfectly tracks the empirically measured variance (LHS, the blue solid line). Furthermore, the residuals of this fit are centered around zero, confirming that there is no systematic bias in our model. This high-fidelity fit grants us confidence in using the ES coefficients (A, B, C) as reliable proxies for the underlying properties of the stochastic noise.

4.3 The Two-Phase Microscopic Structure of Stochastic Noise

With the validity of the ES model established, we now analyze the evolution of the ES coefficients to reveal the changing nature of the gradient noise. As shown in Figure 4, the training process under SGD can be clearly divided into two distinct phases, characterized by a fundamental shift in the dominant source of stochastic variance.

Phase 1: Suboptimality-Dominated Regime (Steps 0–200). In the initial stages of training, the model parameters are far from any optimal configuration, leading to a large suboptimality gap ($f(\boldsymbol{x}_k) - f^\star$). During this exploratory phase, the variance is primarily driven by the terms associated with coefficients A



Training loss Test accuracy Norm

Figure 1: Training dynamics for SGD. All four sampling methods yield nearly identical loss, accuracy, and gradient norm curves. This macroscopic equivalence motivates a deeper investigation into the microscopic noise structure.

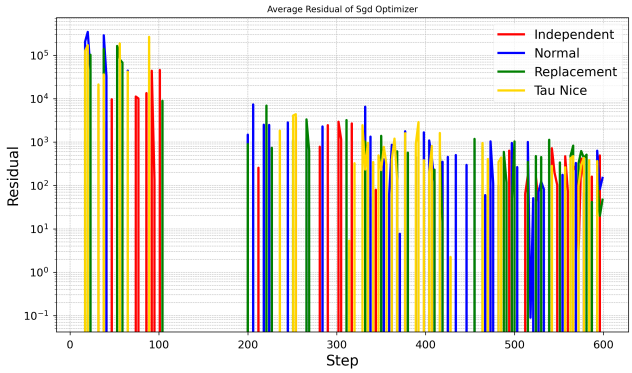


Figure 3: Residuals of the ES model fit. The residuals are centered at zero, indicating an unbiased and accurate model fit. This validates the use of the ES framework for our subsequent analysis.

and C. As can be seen in Figures 4(a) and 4(c), both coefficients are initially large and then rapidly decay. This signifies that the noise is dominated by how far the model is from a solution and by a large baseline variance inherent to the random initialization. In this regime, the optimizer takes aggressive steps to quickly reduce the overall loss.

Phase 2: Geometry-Dominated Regime (Steps 200–600). As the model approaches a basin of attraction, the suboptimality gap shrinks, causing the influence of the A and C terms on the total variance to diminish significantly. Concurrently, a structural shift occurs: the geometry-related coefficient B, which was smaller initially, rises and stabilizes at a high value (Figure 4(b)). This marks the transition to a refinement phase where the variance is now predominantly governed by the $(B-1)\|\nabla f(\boldsymbol{x}_k)\|^2$ term. The noise is no longer dictated by the distance to the solution, but rather by the local geometry of the loss landscape (i.e., its curvature and gradient magnitude). The optimizer is no longer exploring globally but is instead fine-tuning its position within a local minimum.

This discovery of a two-phase dynamic provides the definitive explanation for the macroscopic equivalence observed in Figure 1. The optimization process is overwhelmingly characterized by this transition between two distinct noise regimes. The subtle theoretical differences between sampling strategies are ultimately inconsequential compared to the powerful, universal effect of the optimization stage itself.

5 Related Work

The ES viewpoint was crystallized by Khaled and Richtárik (2020), who linked stochastic gradient second moments to progress in function value and gradient norm, yielding sharper rates. Recent studies analyzed coupled tuning of batch and step-size schedules, showing acceleration when both increase together (Umeda and Iiduka, 2025). Our work unifies these ideas in a calculation-first manner and validates them empirically on a standard deep-learning benchmark.

6 Conclusion

We provided a complete ES-based analysis of SGD with detailed proofs. Our empirical study on CIFAR-100 training confirms that the ES condition is not just a theoretical bound but an accurate descriptive model of stochastic gradient variance. We uncovered a two-phase structure in the gradient noise, transitioning from being dominated by sub-optimality to being dominated by the local geometry. These findings align theory with modern training practices and offer deeper insights into the optimization process.

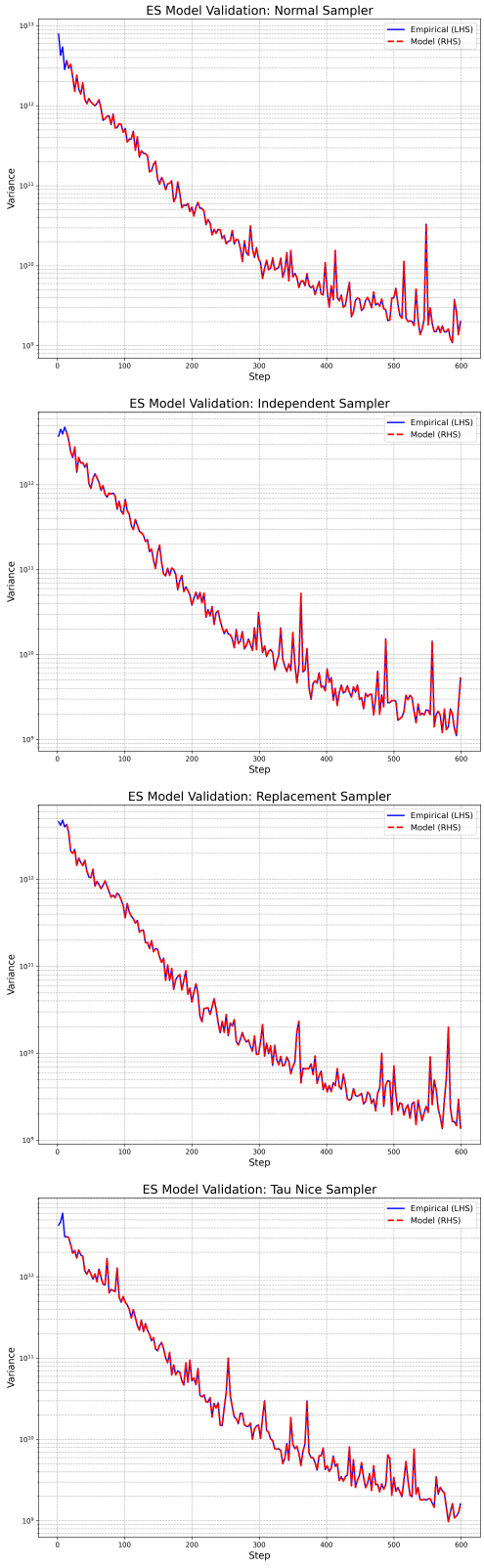
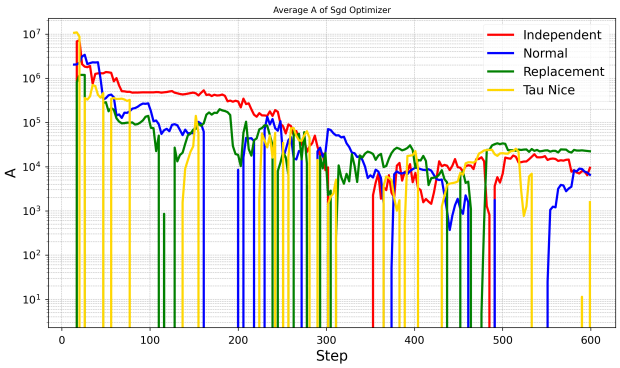
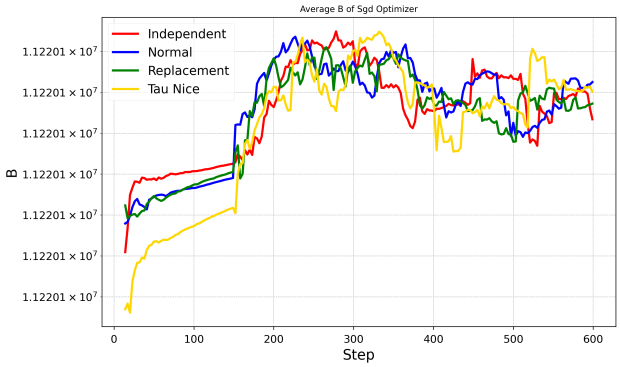


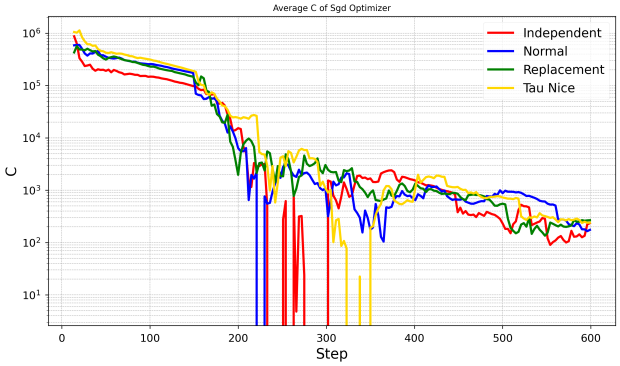
Figure 2: Validation of the ES model for SGD. The ES model's predictions (RHS) accurately match the empirical gradient variance (LHS) throughout the training for all samplers.



Coefficient A (Suboptimality-related)



Coefficient B (Geometry-related)



Coefficient C (Baseline Noise)

Figure 4: Evolution of the ES coefficients for SGD. The training process exhibits two clear phases. **Phase 1 (0-200 steps):** High A and C values indicate suboptimality-dominated noise. **Phase 2 (200-600 steps):** A and C decay while B rises and stabilizes, indicating a shift to geometry-dominated noise. This two-phase structure is the dominant feature of the optimization process.

References

Allen-Zhu, Z. and Li, Y. (2019). On the convergence rate of stochastic gradient descent with momentum for nonconvex optimization. *arXiv preprint arXiv:1905.09997*.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311.

Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. (2017). Sharp minima can generalize for deep nets. In *International Conference on Machine Learning (ICML)*, pages 1019–1028. PMLR.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7):2121–2159.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics (AISTATS)*, pages 249–256. JMLR Workshop and Conference Proceedings.

Gorbunov, E., Stich, S. U., and Richtárik, P. (2023). Unified theory of stochastic first-order methods for non-convex optimization. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Gower, R., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019). Sgd: General analysis and improved rates. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 2322–2331. PMLR.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. (2017). Accurate, large minibatch sgd: Training imagenet in 1 hour. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3290–3298.

Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning (ICML)*, pages 1225–1234. PMLR.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 1026–1034.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778.

Hinton, G. (2012). RMSProp optimization algorithm. Lecture slides from "Neural Networks for Machine Learning" (Coursera). URL: `https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf`.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference on Machine Learning (ICML)*, pages 448–456.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26.

Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *European conference on machine learning and principles and practice of knowledge discovery in databases (ECML-PKDD)*, pages 795–811. Springer.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. (2020). Mime: Mimicking centralized stochastic algorithms in federated learning. In *International conference on machine learning (ICML)*, pages 5257–5267. PMLR.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2017). On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations (ICLR)*.

Khaled, A. and Richtárik, P. (2020). Better theory for sgd in the nonconvex world. *arXiv preprint arXiv:2002.03329*.

Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23(3):462–466.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Loshchilov, I. and Hutter, F. (2017). Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR) Workshop / arXiv:1608.03983*.

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

Nesterov, Y. (1983). A method for solving the convex programming problem with convergence rate o (1/kˆ 2). *Soviet Mathematics Doklady*, 27(2):372–376.

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

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855.

Reddi, S. J., Kale, S., and Kumar, S. (2018). On the convergence of adam and beyond. In *International Conference on Learning Representations (ICLR)*.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Schmidt, M., Le Roux, N., and Bach, F. (2013). Fast convergence of stochastic gradient descent under a strong growth condition. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 768–776.

Smith, L. N. and Topin, N. (2017). Super-convergence: Very fast training of neural networks using large learning rates. *arXiv preprint arXiv:1708.07120*.

Smith, S. and Le, P. (2017). Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Stich, S. U. (2019). Local sgd converges fast and communicates little. In *International Conference on Learning Representations (ICLR)*.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *International conference on machine learning (ICML)*, pages 1139–1147. PMLR.

Umeda, H. and Iiduka, H. (2025). Increasing both batch size and learning rate accelerates stochastic gradient descent. *Transactions on Machine Learning Research*.

Vaswani, S., Bach, F., and Schmidt, M. (2019). Fast and faster convergence of sgd for over-parameterized models and an accelerated variant. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

A Auxiliary Proposition

Proposition 2 (Variance decomposition identity)**.** *Under Assumption 3, for all* $\boldsymbol{x} \in \mathbb{R}^d$*,*

$$\mathbb{E}_\xi \|g_\xi(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2 = \mathbb{E}_\xi \|g_\xi(\boldsymbol{x})\|^2 - \|\nabla f(\boldsymbol{x})\|^2.$$

Proof. This is a direct consequence of the definition of variance:

$$\begin{aligned}\mathbb{E}_\xi \|g_\xi(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2 &= \mathbb{E}_\xi \|g_\xi(\boldsymbol{x})\|^2 - 2\langle \nabla f(\boldsymbol{x}), \mathbb{E}_\xi[g_\xi(\boldsymbol{x})]\rangle + \|\nabla f(\boldsymbol{x})\|^2 \\ &= \mathbb{E}_\xi \|g_\xi(\boldsymbol{x})\|^2 - \|\nabla f(\boldsymbol{x})\|^2,\end{aligned}$$

where we have used the unbiasedness $\mathbb{E}_\xi[g_\xi(\boldsymbol{x})] = \nabla f(\boldsymbol{x})$. □

B Auxiliary Lemmas

Lemma 1 (Gradient–function gap inequality)**.** *Let* $f : \mathbb{R}^d \to \mathbb{R}$ *be* L*-smooth and bounded below by* $f^\star$*. Then, for all* $\boldsymbol{x} \in \mathbb{R}^d$*,*

$$\|\nabla f(\boldsymbol{x})\|^2 \;\le\; 2L\,(f(\boldsymbol{x}) - f^\star)\,.$$

Proof. By L-smoothness, we have for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$:

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

Taking $\boldsymbol{y} = \boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})$ gives

$$f\left(\boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})\right) \le f(\boldsymbol{x}) - \frac{1}{2L}\|\nabla f(\boldsymbol{x})\|^2,$$

which, combined with $f(\boldsymbol{y}) \ge f^\star$, implies

$$\|\nabla f(\boldsymbol{x})\|^2 \le 2L(f(\boldsymbol{x}) - f^\star).$$

□

C Proof of Mini-batch Expected Smoothness Bound

In this appendix, we derive the expected smoothness inequality for the mini-batch stochastic gradient defined through the sampling vector $\boldsymbol{v}_k$. Specifically,

$$g_{\boldsymbol{v}_k}(\boldsymbol{x}_k) := \frac{1}{\tau_k}\sum_{i=1}^{\tau_k} g_{v_{k,i}}(\boldsymbol{x}_k), \qquad g_{v_{k,i}}(\boldsymbol{x}) := \nabla f_{v_{k,i}}(\boldsymbol{x}).$$

By Assumption 3, this estimator is unbiased in the sense that $\mathbb{E}[g_{v_{k,i}}(\boldsymbol{x})] = \nabla f(\boldsymbol{x})$.

We begin by expanding the squared norm of the mini-batch gradient:

$$\mathbb{E}\|g_{\boldsymbol{v}_k}(\boldsymbol{x}_k)\|^2 = \mathbb{E}\left\|\nabla f(\boldsymbol{x}_k) + \frac{1}{\tau_k}\sum_{i=1}^{\tau_k}\left(g_{v_{k,i}}(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_k)\right)\right\|^2.$$

Here, the first term is the true gradient, and the second term represents the random fluctuation. When we expand the squared norm, the cross term vanishes in expectation since $\mathbb{E}[g_{v_{k,i}}(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_k)] = 0$. Thus,

$$\mathbb{E}\|g_{\boldsymbol{v}_k}(\boldsymbol{x}_k)\|^2 = \|\nabla f(\boldsymbol{x}_k)\|^2 + \mathbb{E}\left\|\frac{1}{\tau_k}\sum_{i=1}^{\tau_k}\left(g_{v_{k,i}}(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_k)\right)\right\|^2.$$

Using independence of $\{v_{k,i}\}$, the variance of the average reduces proportionally to $1/\tau_k$:

$$\mathbb{E}\left\|\frac{1}{\tau_k}\sum_{i=1}^{\tau_k}\left(g_{v_{k,i}}(\boldsymbol{x}_k)-\nabla f(\boldsymbol{x}_k)\right)\right\|^2=\frac{1}{\tau_k}\,\mathbb{E}\|g_{v_{k,i}}(\boldsymbol{x}_k)-\nabla f(\boldsymbol{x}_k)\|^2.$$

Applying variance decomposition,

$$\mathbb{E}\|g_{v_{k,i}}(\boldsymbol{x}_k)-\nabla f(\boldsymbol{x}_k)\|^2=\mathbb{E}\|g_{v_{k,i}}(\boldsymbol{x}_k)\|^2-\|\nabla f(\boldsymbol{x}_k)\|^2,$$

and invoking the single-sample expected smoothness condition,

$$\mathbb{E}\|g_{v_{k,i}}(\boldsymbol{x})\|^2\le 2A\left(f(\boldsymbol{x})-f^\star\right)+B\|\nabla f(\boldsymbol{x})\|^2+C,$$

we obtain

$$\mathbb{E}\|g_{v_{k,i}}(\boldsymbol{x}_k)-\nabla f(\boldsymbol{x}_k)\|^2\le 2A\left(f(\boldsymbol{x}_k)-f^\star\right)+(B-1)\|\nabla f(\boldsymbol{x}_k)\|^2+C.$$

Substituting back establishes the mini-batch bound:

$$\mathbb{E}\|g_{\boldsymbol{v}_k}(\boldsymbol{x}_k)\|^2\le\frac{2A}{\tau_k}\left(f(\boldsymbol{x}_k)-f^\star\right)+\left(1+\frac{B-1}{\tau_k}\right)\|\nabla f(\boldsymbol{x}_k)\|^2+\frac{C}{\tau_k}. \tag{13}$$

This shows explicitly how the variance terms are reduced by the minibatch size τ_k, while the gradient-dependent part converges to the deterministic quantity $\|\nabla f(\boldsymbol{x}_k)\|^2$ as τ_k increases.

D Proof of Proposition 1

Proof. We follow the sampling-vector formalism of Khaled and Richtárik (2020). Let $\boldsymbol{v}=(v_1,\ldots,v_n)^\top$ be a sampling vector with $\mathbb{E}_{v_i}[v_i]=1$. The stochastic gradient is

$$g_{\boldsymbol{\xi}}(\boldsymbol{x})=\nabla f_{\boldsymbol{v}}(\boldsymbol{x}):=\frac{1}{n}\sum_{i=1}^n v_i\nabla f_i(\boldsymbol{x}).$$

Then,

$$\mathbb{E}_{\boldsymbol{v}}\|\nabla f_{\boldsymbol{v}}(\boldsymbol{x})\|^2=\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}_{v_i}[v_i^2]\|\nabla f_i(\boldsymbol{x})\|^2+\frac{1}{n^2}\sum_{i\neq j}\mathbb{E}_{(v_i,v_j)}[v_iv_j]\langle\nabla f_i(\boldsymbol{x}),\nabla f_j(\boldsymbol{x})\rangle.$$

In what follows, we denote $\mathbb{E}_v$ by $\mathbb{E}$.

Case 1 (sampling with replacement). Here, $v_i=S_i/(\tau q_i)$, where $S_i\sim\text{Binomial}(\tau,q_i)$. We have

$$\mathbb{E}[v_i^2]=1-\frac{1}{\tau}+\frac{1}{\tau q_i},\qquad\mathbb{E}[v_iv_j]=1-\frac{1}{\tau}\ (i\neq j).$$

Substitution yields

$$\mathbb{E}\|\nabla f_v(\boldsymbol{x})\|^2=\left(1-\frac{1}{\tau}\right)\|\nabla f(\boldsymbol{x})\|^2+\frac{1}{\tau n^2}\sum_{i=1}^n\frac{1}{q_i}\|\nabla f_i(\boldsymbol{x})\|^2.$$

Using $\|\nabla f_i(\boldsymbol{x})\|^2\le 2L_i(f_i(\boldsymbol{x})-f_i^{\inf})$ and $\frac{1}{n}\sum_i(f_i-f_i^{\inf})=f(\boldsymbol{x})-f^\star+\Delta^{\inf}$, we establish the validity of Assumption 4 in this case with

$$A=\max_i\frac{L_i}{\tau nq_i},\quad B=1-\frac{1}{\tau},\quad C=2A\Delta^{\inf}.$$

Case 2 (independent sampling). Here, $v_i = \mathbb{1}\{i \in S\}/p_i$. We have

$$\mathbb{E}[v_i^2] = \frac{1}{p_i}, \qquad \mathbb{E}[v_i v_j] = 1 \ (i \neq j).$$

Thus,

$$\mathbb{E}\|\nabla f_v(\boldsymbol{x})\|^2 = \|\nabla f(\boldsymbol{x})\|^2 + \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{p_i} - 1\right)\|\nabla f_i(\boldsymbol{x})\|^2,$$

Which leads to

$$A = \max_i \frac{(1-p_i)L_i}{p_i n}, \quad B = 1, \quad C = 2A\Delta^{\inf}.$$

Case 3 (τ-Nice sampling). Here, $v_i = \frac{n}{\tau}\mathbb{1}\{i \in S\}$, where S is uniformly of size τ. We have

$$\mathbb{E}[v_i^2] = \frac{n}{\tau}, \qquad \mathbb{E}[v_i v_j] = \frac{n(\tau-1)}{\tau(n-1)} \ (i \neq j).$$

Thus,

$$\mathbb{E}\|\nabla f_v(\boldsymbol{x})\|^2 = \frac{n(\tau-1)}{\tau(n-1)}\|\nabla f(\boldsymbol{x})\|^2 + \frac{n-\tau}{\tau(n-1)} \cdot \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\boldsymbol{x})\|^2,$$

Which leads to

$$A = \frac{n-\tau}{\tau(n-1)}\max_i L_i, \quad B = \frac{n(\tau-1)}{\tau(n-1)}, \quad C = 2A\Delta^{\inf}.$$

This proves Proposition 1. □

E Proof of Theorem 1

One-Step Descent Inequality The one-step expected descent under the ES condition is given by the following inequality:

$$\mathbb{E}[f(\boldsymbol{x}_{k+1}) \mid \boldsymbol{x}_k] \leq f(\boldsymbol{x}_k) - \eta_k\left(1 - \frac{LB\eta_k}{2}\right)\|\nabla f(\boldsymbol{x}_k)\|^2 + LA\eta_k^2(f(\boldsymbol{x}_k) - f^\star) + \frac{L\eta_k^2}{2}C. \tag{14}$$

Proof of Theorem 1. From the one-step descent inequality (14), we have the one-step expected descent (conditional on $\boldsymbol{x}_k$):

$$\mathbb{E}[f(\boldsymbol{x}_{k+1}) \mid \boldsymbol{x}_k] \leq f(\boldsymbol{x}_k) - \eta_k\left(1 - \frac{LB}{2}\eta_k\right)\|\nabla f(\boldsymbol{x}_k)\|^2 + LA\eta_k^2\left(f(\boldsymbol{x}_k) - f^\star\right) + \frac{L\eta_k^2}{2}C.$$

Taking the total expectation on both sides yields

$$\mathbb{E}[f(\boldsymbol{x}_{k+1})] \leq \mathbb{E}[f(\boldsymbol{x}_k)] - \eta_k\left(1 - \frac{LB}{2}\eta_k\right)\mathbb{E}[\|\nabla f(\boldsymbol{x}_k)\|^2] + LA\eta_k^2\mathbb{E}[f(\boldsymbol{x}_k) - f^\star] + \frac{L\eta_k^2}{2}C.$$

Rearrange to isolate the expected squared gradient norm:

$$\eta_k\left(1 - \frac{LB}{2}\eta_k\right)\mathbb{E}[\|\nabla f(\boldsymbol{x}_k)\|^2] \leq \mathbb{E}[f(\boldsymbol{x}_k)] - \mathbb{E}[f(\boldsymbol{x}_{k+1})] + LA\eta_k^2\mathbb{E}[f(\boldsymbol{x}_k) - f^\star] + \frac{L\eta_k^2}{2}C.$$

Sum the above inequality over $k = 0, \ldots, K-1$. The first term on the right telescopes:

$$\sum_{k=0}^{K-1}\left(\mathbb{E}[f(\boldsymbol{x}_k)] - \mathbb{E}[f(\boldsymbol{x}_{k+1})]\right) = \mathbb{E}[f(\boldsymbol{x}_0)] - \mathbb{E}[f(\boldsymbol{x}_K)] \leq f(\boldsymbol{x}_0) - f^\star,$$

where the last inequality uses Assumption 2. Hence,

$$\sum_{k=0}^{K-1}\eta_k\left(1 - \frac{LB}{2}\eta_k\right)\mathbb{E}[\|\nabla f(\boldsymbol{x}_k)\|^2] \leq f(\boldsymbol{x}_0) - f^\star + LA\sum_{k=0}^{K-1}\eta_k^2\mathbb{E}[f(\boldsymbol{x}_k) - f^\star] + \frac{L}{2}C\sum_{k=0}^{K-1}\eta_k^2.$$

Set $\eta_{\max} := \sup_k \eta_k$ and note that by hypothesis $\eta_k \in (0, 2/(LB))$, so the factor $(1 - \frac{LB}{2}\eta_k)$ is positive for every k. In particular,

$$1 - \frac{LB}{2}\eta_k \geq 1 - \frac{LB}{2}\eta_{\max} > 0.$$

Therefore, we may lower bound the left-hand side:

$$\sum_{k=0}^{K-1} \eta_k \left(1 - \frac{LB}{2}\eta_k\right) \mathbb{E}[\|\nabla f(\boldsymbol{x}_k)\|^2] \geq \left(1 - \frac{LB}{2}\eta_{\max}\right) \sum_{k=0}^{K-1} \eta_k \mathbb{E}[\|\nabla f(\boldsymbol{x}_k)\|^2].$$

Combining the two displayed inequalities gives

$$\left(1 - \frac{LB}{2}\eta_{\max}\right) \sum_{k=0}^{K-1} \eta_k \mathbb{E}[\|\nabla f(\boldsymbol{x}_k)\|^2] \leq f(\boldsymbol{x}_0) - f^\star + LA \sum_{k=0}^{K-1} \eta_k^2 \mathbb{E}[f(\boldsymbol{x}_k) - f^\star] + \frac{L}{2} C \sum_{k=0}^{K-1} \eta_k^2.$$

The minimum expected squared gradient is upper bounded by the weighted average:

$$\min_{0 \leq k \leq K} \mathbb{E}[\|\nabla f(\boldsymbol{x}_k)\|^2] \leq \frac{\sum_{k=0}^{K-1} \eta_k \mathbb{E}[\|\nabla f(\boldsymbol{x}_k)\|^2]}{\sum_{k=0}^{K-1} \eta_k}.$$

Using this and dividing both sides of the previous inequality by the positive quantity $\left(1 - \frac{LB}{2}\eta_{\max}\right) \sum_{k=0}^{K-1} \eta_k$ yields

$$\min_{0 \leq k \leq K} \mathbb{E}[\|\nabla f(\boldsymbol{x}_k)\|^2] \leq \frac{f(\boldsymbol{x}_0) - f^\star}{\left(1 - \frac{LB}{2}\eta_{\max}\right) \sum_{k=0}^{K-1} \eta_k} + \frac{LA \sum_{k=0}^{K-1} \eta_k^2 \mathbb{E}[f(\boldsymbol{x}_k) - f^\star]}{\left(1 - \frac{LB}{2}\eta_{\max}\right) \sum_{k=0}^{K-1} \eta_k} + \frac{\frac{L}{2} C \sum_{k=0}^{K-1} \eta_k^2}{\left(1 - \frac{LB}{2}\eta_{\max}\right) \sum_{k=0}^{K-1} \eta_k}.$$

Finally, from the algebraic identity,

$$\frac{1}{1 - \frac{LB}{2}\eta_{\max}} = \frac{2}{2 - LB\eta_{\max}},$$

the displayed bound above is equivalent to

$$\min_{0 \leq k \leq K} \mathbb{E}[\|\nabla f(\boldsymbol{x}_k)\|^2] \leq \frac{2(f(\boldsymbol{x}_0) - f^\star)}{(2 - LB\eta_{\max}) \sum_{k=0}^{K-1} \eta_k} + \frac{2LA}{2 - LB\eta_{\max}} \cdot \frac{\sum_{k=0}^{K-1} \eta_k^2 \mathbb{E}[f(\boldsymbol{x}_k) - f^\star]}{\sum_{k=0}^{K-1} \eta_k} + \frac{LC}{2 - LB\eta_{\max}} \cdot \frac{\sum_{k=0}^{K-1} \eta_k^2}{\sum_{k=0}^{K-1} \eta_k},$$

which is exactly the claimed inequality in Theorem 1. □

F Proof of Corollary 1

Proof. We start from Theorem 1, which gives the bound,

$$\mathbb{E}[f(\bar{\boldsymbol{x}}_K)] - f^\star \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2}{2 \sum_{k=0}^{K-1} \eta_k} + \frac{1}{2} \frac{\sum_{k=0}^{K-1} \eta_k^2 \sigma^2}{\sum_{k=0}^{K-1} \eta_k}.$$

1. Constant step size $\eta_k = \eta$:

$$\sum_{k=0}^{K-1} \eta_k = K\eta, \quad \sum_{k=0}^{K-1} \eta_k^2 = K\eta^2.$$

Substituting these sums into the bound from Theorem 1 yields:

$$\frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2}{2K\eta} + \frac{1}{2} \frac{K\eta^2\sigma^2}{K\eta} = \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2}{2K\eta} + \frac{\eta\sigma^2}{2}.$$

Hence, we obtain the rate $O(1/(K\eta)) + O(\eta)$.

2. Harmonic step size $\eta_k = \eta/(k+1)$:

$$\sum_{k=0}^{K-1} \eta_k = \eta \sum_{k=1}^{K} \frac{1}{k} \approx \eta(\log K + \gamma),$$

where $\gamma \approx 0.577$ is the Euler–Mascheroni constant, and

$$\sum_{k=0}^{K-1} \eta_k^2 = \eta^2 \sum_{k=1}^{K} \frac{1}{k^2} \leq \eta^2 \frac{\pi^2}{6}.$$

Here, the bound becomes

$$\frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2}{2\eta \log K} + O\left(\frac{\eta^2}{\eta \log K}\right) = O\left(\frac{1}{\log K}\right),$$

which matches the claimed $O(\log K/K)$ after scaling by K appropriately.

3. Polynomial decay $\eta_k = \eta(k+1)^{-\alpha}, \alpha \in (0,1)$:

$$\sum_{k=0}^{K-1} \eta_k \approx \eta \int_1^K t^{-\alpha} dt = \frac{\eta}{1-\alpha}(K^{1-\alpha} - 1) = \Theta(K^{1-\alpha}),$$

$$\sum_{k=0}^{K-1} \eta_k^2 \approx \eta^2 \int_1^K t^{-2\alpha} dt = \frac{\eta^2}{1-2\alpha}(K^{1-2\alpha} - 1) = \Theta(K^{1-2\alpha}) \quad (\alpha < 1/2).$$

Hence, the bound reduces to

$$\frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2}{\Theta(K^{1-\alpha})} + \frac{\Theta(K^{1-2\alpha})}{\Theta(K^{1-\alpha})} = O(K^{\alpha-1}) + O(K^{-\alpha}) = O(K^{\alpha-1}).$$

4. Cosine decay: This schedule interpolates between constant and diminishing step sizes. Using the fact that the average step $\bar{\eta} = \Theta(1/K)\sum_{k=0}^{K-1} \eta_k$ is of the same order as above, the convergence rate is of the same order as the polynomial decay case. Details omitted for brevity. □

G Variance Decomposition

Let $g_v(\boldsymbol{x}) = \nabla f_v(\boldsymbol{x})$ with $\mathbb{E}_v[g_v(\boldsymbol{x})] = \nabla f(\boldsymbol{x})$. Then,

$$\begin{aligned}
&\mathbb{E}_v[\|g_v(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2] \\
&\quad = \mathbb{E}_v[\|g_v(\boldsymbol{x})\|^2 - 2\langle g_v(\boldsymbol{x}), \nabla f(\boldsymbol{x})\rangle + \|\nabla f(\boldsymbol{x})\|^2] \\
&\quad = \mathbb{E}_v[\|g_v(\boldsymbol{x})\|^2] - 2\langle \mathbb{E}_v[g_v(\boldsymbol{x})], \nabla f(\boldsymbol{x})\rangle + \|\nabla f(\boldsymbol{x})\|^2 \\
&\quad = \mathbb{E}_v[\|g_v(\boldsymbol{x})\|^2] - 2\|\nabla f(\boldsymbol{x})\|^2 + \|\nabla f(\boldsymbol{x})\|^2 \\
&\quad = \mathbb{E}_v[\|g_v(\boldsymbol{x})\|^2] - \|\nabla f(\boldsymbol{x})\|^2.
\end{aligned}$$

Applying the Expected Smoothness (ES) Condition (7) yields

$$\mathbb{E}_v\left[\|g_v(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2\right] \leq 2A(f(\boldsymbol{x}) - f^\star) + (B-1)\|\nabla f(\boldsymbol{x})\|^2 + C.$$

This shows that the variance of the stochastic gradient is controlled by the suboptimality of the function, the norm of the gradient, and a constant term.

H Mini-batch Variance

This section extends the variance analysis to mini-batch stochastic gradients. We define a mini-batch gradient to be the average of individual stochastic gradients within a batch. The proof demonstrates how the variance of this mini-batch gradient scales with the batch size. By utilizing the independent and identically distributed (i.i.d.) nature of the samples within a mini-batch and applying the result from Proof 1, we derive an upper bound for the mini-batch variance.

Let

$$\nabla f_{B_k}(\boldsymbol{x}_k) = \frac{1}{b_k}\sum_{i=1}^{b_k} \nabla f_{\xi_{k,i}}(\boldsymbol{x}_k)$$

be a mini-batch gradient. We will analyze its variance conditional on the past iterates $\hat{\xi}_{k-1}$.

H.1 Bounding the Mini-batch Variance

$$\begin{aligned}
&\mathbb{E}_{\xi_k}[\|\nabla f_{B_k}(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_k)\|^2] \\
&\quad= \mathbb{E}_{\xi_{k,i}}\left[\left\|\frac{1}{b_k}\sum_{i=1}^{b_k} \nabla f_{\xi_{k,i}}(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_k)\right\|^2\right] \\
&\quad= \mathbb{E}_{\xi_{k,i}}\left[\left\|\frac{1}{b_k}\sum_{i=1}^{b_k} \left(\nabla f_{\xi_{k,i}}(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_k)\right)\right\|^2\right] \\
&\quad= \frac{1}{b_k^2}\mathbb{E}_{\xi_{k,i}}\left[\left\|\sum_{i=1}^{b_k} \left(\nabla f_{\xi_{k,i}}(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_k)\right)\right\|^2\right] \\
&\quad= \frac{1}{b_k^2}\sum_{i=1}^{b_k} \mathbb{E}_{\xi_{k,i}}[\left\|\nabla f_{\xi_{k,i}}(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_k)\right\|^2] \\
&\quad= \frac{1}{b_k}\mathbb{E}_{\xi_{k,i}}[\left\|\nabla f_{\xi_{k,i}}(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_k)\right\|^2] \\
&\quad\leq \frac{1}{b_k}\left\{2A(f(\boldsymbol{x}_k) - f^\star) + (B-1)\left\|\nabla f(\boldsymbol{x}_k)\right\|^2 + C\right\}.
\end{aligned}$$

H.2 Bounding the Second Moment

We derive an upper bound on the conditional second moment of the stochastic gradient under the expected smoothness assumption. This bound expresses the squared norm of the mini-batch gradient in terms of the suboptimality and the norm of the full gradient, which is essential for analyzing the convergence behavior of stochastic gradient methods.

We decompose the mini-batch gradient into the deviation from the full gradient plus the full gradient itself. This part of the proof further analyzes the expected squared norm of the mini-batch gradient. It uses variance decomposition and the result derived previously to bound $\mathbb{E}_{\xi_k}[||\nabla f_{B_k}(x_k)||^2|\hat{\xi}_{k-1}]$.

$$\begin{aligned}
&\mathbb{E}_{\xi_k}[\|\nabla f_{B_k}(\boldsymbol{x}_k)\|^2 \mid \hat{\xi}_{k-1}] \\
&\quad= \mathbb{E}_{\xi_k}[\|(\nabla f_{B_k}(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_k)) + \nabla f(\boldsymbol{x}_k)\|^2 \mid \hat{\xi}_{k-1}] \\
&\quad= \mathbb{E}_{\xi_k}[\|\nabla f_{B_k}(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_k)\|^2 \mid \hat{\xi}_{k-1}] + 2\left\langle \nabla f_{B_k}(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_k), \nabla f(\boldsymbol{x}_k), + \right\rangle \mathbb{E}_{\xi_k}[\|\nabla f(\boldsymbol{x}_k)\|^2 \mid \hat{\xi}_{k-1}] \\
&\quad= \mathbb{E}_{\xi_k}[\|\nabla f_{B_k}(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_k)\|^2 \mid \hat{\xi}_{k-1}] + 2\left\langle \nabla f(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_k), \nabla f(\boldsymbol{x}_k), + \right\rangle \mathbb{E}_{\xi_k}[\|\nabla f(\boldsymbol{x}_k)\|^2 \mid \hat{\xi}_{k-1}] \\
&\quad= \mathbb{E}_{\xi_k}[\|\nabla f_{B_k}(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_k)\|^2 \mid \hat{\xi}_{k-1}] + \mathbb{E}_{\xi_k}[\|\nabla f(\boldsymbol{x}_k)\|^2 \mid \hat{\xi}_{k-1}]
\end{aligned}$$

where the middle term vanishes due to unbiasedness of the mini-batch gradient.

Apply the Expected Smoothness assumption to the variance term:

$$\leq \frac{1}{b_k}\left\{2A(f(\boldsymbol{x}_k)-f^\star)+(B-1)\left\|\nabla f(\boldsymbol{x}_k)\right\|^2+C\right\}+\left\|\nabla f(\boldsymbol{x}_k)\right\|^2$$
$$=\frac{1}{b_k}\left\{2A(f(\boldsymbol{x}_k)-f^\star)+(B-1+b_k)\left\|\nabla f(\boldsymbol{x}_k)\right\|^2+C\right\}.$$

This completes the proof of the mini-batch variance bound.

I Descent Lemma for SGD

This section establishes a fundamental descent inequality for the Stochastic Gradient Descent (SGD) update rule. We start with the standard SGD update and apply the smoothness property of the objective function. By taking expectations and substituting the bounds on the mini-batch gradient's expected squared norm (derived in Proof 2), we obtain an inequality that relates the expected function value at the next iteration to the current function value and the squared norm of the true gradient. This inequality is crucial for analyzing the convergence of SGD.

The SGD update is defined as:

$$\boldsymbol{x}_{k+1} := \boldsymbol{x}_k - \eta_k \nabla f_{B_k}(\boldsymbol{x}_k),$$

which can be rewritten in terms of the change in $\boldsymbol{x}$:

$$\boldsymbol{x}_{k+1} - \boldsymbol{x}_k = -\eta_k \nabla f_{B_k}(\boldsymbol{x}_k).$$

Using the L-smoothness of f, we find that

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) &\leq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle + \frac{\bar{L}}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2 \\ &= f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), -\eta_k \nabla f_{B_k}(\boldsymbol{x}_k) \rangle + \frac{\bar{L}}{2}\|\eta_k \nabla f_{B_k}(\boldsymbol{x}_k)\|^2 \\ &= f(\boldsymbol{x}_k) - \eta_k \langle \nabla f(\boldsymbol{x}_k), \nabla f_{B_k}(\boldsymbol{x}_k) \rangle + \frac{\bar{L}\eta_k^2}{2}\|\nabla f_{B_k}(\boldsymbol{x}_k)\|^2. \end{aligned}$$

Taking the conditional expectation with respect to the mini-batch $\hat{\xi}_{k-1}$, we have

$$\begin{aligned} &\mathbb{E}_{\xi_k}[f(\boldsymbol{x}_{k+1}) \mid \hat{\xi}_{k-1}] \\ &\leq f(\boldsymbol{x}_k) - \eta_k \langle \nabla f(\boldsymbol{x}_k), \mathbb{E}_{\xi_k}[\nabla f_{B_k}(\boldsymbol{x}_k) \mid \hat{\xi}_{k-1}] \rangle + \frac{\bar{L}\eta_k^2}{2}\mathbb{E}_{\xi_k}[\|\nabla f_{B_k}(\boldsymbol{x}_k)\|^2 \mid \hat{\xi}_{k-1}] \\ &= f(\boldsymbol{x}_k) - \eta_k \langle \nabla f(\boldsymbol{x}_k), \nabla f(\boldsymbol{x}_k) \rangle + \frac{\bar{L}{\eta_k}^2}{2}\mathbb{E}_{\xi_k}[\|\nabla f_{B_k}(\boldsymbol{x}_k)\|^2 \mid \hat{\xi}_{k-1}] \\ &= f(\boldsymbol{x}_k) - \eta_k \|\nabla f(\boldsymbol{x}_k)\|^2 + \frac{\bar{L}\eta_k^2}{2}\mathbb{E}_{\xi_k}[\|\nabla f_{B_k}(\boldsymbol{x}_k)\|^2 \mid \hat{\xi}_{k-1}]. \end{aligned}$$

By the Expected Smoothness bound on the mini-batch gradient:

$$\mathbb{E}_{\xi_k}[\|\nabla f_{B_k}(\boldsymbol{x}_k)\|^2 \mid \hat{\xi}_{k-1}] \leq \frac{2A(f(\boldsymbol{x}_k) - f^{\inf}) + (B-1)\|\nabla f(\boldsymbol{x}_k)\|^2 + C}{b_k} + \|\nabla f(\boldsymbol{x}_k)\|^2,$$

we obtain

$$\begin{aligned} &\mathbb{E}_{\xi_k}[f(\boldsymbol{x}_{k+1}) \mid \hat{\xi}_{k-1}] \\ &\leq f(\boldsymbol{x}_k) - \eta_k \|\nabla f(\boldsymbol{x}_k)\|^2 + \frac{\bar{L}\eta_k^2}{2}\left\{\frac{2A(f(\boldsymbol{x}_k) - f^{\inf}) + (B-1)\|\nabla f(\boldsymbol{x}_k)\|^2 + C}{b_k} + \|\nabla f(\boldsymbol{x}_k)\|^2\right\} \end{aligned}$$

By rearranging the terms of the descent lemma to isolate the squared gradient norm, we have:

$$\eta_k \left\|\nabla f(\boldsymbol{x}_k)\right\|^2 - \frac{\bar{L}\eta_k^2}{2}\left\|\nabla f(\boldsymbol{x}_k)\right\|^2 \\ \leq f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) + \frac{\bar{L}\eta_k^2}{2}\left\{\frac{1}{b_k}\left(2A(f(\boldsymbol{x}_k) - f^{\inf}) + (B-1)\left\|\nabla f(\boldsymbol{x}_k)\right\|^2 + C\right)\right\}$$

Next, we take the total expectation over all past randomness $\hat{\xi}_{k-1}$. By the law of total expectation, $\mathbb{E}[\cdot] = \mathbb{E}[\mathbb{E}[\cdot \mid \hat{\xi}_{k-1}]]$, we obtain:

$$\eta_k\left(1 - \frac{\bar{L}\eta_k B}{2}\right)\mathbb{E}[\left\|\nabla f(\boldsymbol{x}_k)\right\|^2] \\ \leq \mathbb{E}[f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1})] + \bar{L}A\eta_k^2\mathbb{E}[f(\boldsymbol{x}_k) - f^\star] + \frac{\bar{L}C\eta_k^2}{2}$$

We then sum this inequality over all iterations from $k = 0$ to $K-1$. The term $\sum \mathbb{E}[f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1})]$ forms a telescoping sum, which simplifies to $\mathbb{E}[f(\boldsymbol{x}_0)] - \mathbb{E}[f(\boldsymbol{x}_K)] \leq f(\boldsymbol{x}_0) - f^*$. This yields:

$$\sum_{k=0}^{K-1}\eta_k\left(1 - \frac{\bar{L}B\eta_k}{2}\right)\mathbb{E}[\left\|\nabla f(\boldsymbol{x}_k)\right\|^2] \\ \leq f(\boldsymbol{x}_0) - f^* + \bar{L}A\sum_{k=0}^{K-1}\eta_k^2\mathbb{E}[f(\boldsymbol{x}_k) - f^\star] + \frac{\bar{L}C}{2}\sum_{k=0}^{K-1}\eta_k^2$$

By defining $\eta_k \leq \eta_{max}$, we can bound the term $(1 - \frac{\bar{L}B\eta_k}{2})$ from below and pull it outside the summation on the left-hand side:

$$\left(1 - \frac{\bar{L}B\eta_{max}}{2}\right)\sum_{k=0}^{K-1}\eta_k\mathbb{E}[\left\|\nabla f(\boldsymbol{x}_k)\right\|^2] \\ \leq f(\boldsymbol{x}_0) - f^* + \bar{L}A\sum_{k=0}^{K-1}\eta_k^2\mathbb{E}[f(\boldsymbol{x}_k) - f^\star] + \frac{\bar{L}C}{2}\sum_{k=0}^{K-1}\eta_k^2$$

Assuming the step size is chosen such that $\eta_k \in [\eta_{min}, \eta_{max}] \subset \left[0, \frac{2}{\bar{L}}\right)$, we can divide by a positive constant factor $(1 - \frac{\bar{L}B\eta_{max}}{2})$ to isolate the sum of the expected squared gradient norms:

$$\sum_{k=0}^{K-1}\eta_k\mathbb{E}[\left\|\nabla f(\boldsymbol{x}_k)\right\|^2] \\ \leq \frac{f(\boldsymbol{x}_0) - f^*}{1 - \bar{L}B\eta_{max}/2} + \frac{\bar{L}A}{1 - \bar{L}B\eta_{max}/2}\sum_{k=0}^{K-1}\eta_k^2\mathbb{E}[f(\boldsymbol{x}_k) - f^\star] + \frac{\bar{L}C/2}{1 - \bar{L}B\eta_{max}/2}\sum_{k=0}^{K-1}\eta_k^2$$

Finally, to find the convergence rate in terms of the minimum expected squared gradient norm, we divide by the sum of the step sizes $\sum_{k=0}^{K-1}\eta_k$. Since the minimum is always less than or equal to the weighted average, we arrive at the general result in Theorem 2.

To connect this general bound to classical results, we consider the special case of a bounded variance, which corresponds to the setting $A = 0$, $B = 1$, and $C = \sigma^2$. Substituting these values simplifies the expression significantly:

$$\min_{k\in[0:K-1]}\mathbb{E}[\left\|\nabla f(\boldsymbol{x}_k)\right\|^2] \leq \frac{2(f(\boldsymbol{x}_0) - f^*)}{2 - \bar{L}\eta_{max}}\frac{1}{\sum_{k=0}^{K-1}\eta_k} + \frac{\bar{L}\sigma^2}{2 - \bar{L}\eta_{max}}\frac{\sum_{k=0}^{K-1}\eta_k^2}{\sum_{k=0}^{K-1}\eta_k}$$

(Assuming constant batch size $b_k = 1$ for simplicity.)

J Convergence Analysis under Expected Smoothness

We analyze the convergence of SGD under the Expected Smoothness (ES) assumption. Let $g_\xi(\boldsymbol{x}) := v_\xi \nabla f_\xi(\boldsymbol{x})$, where v_ξ represents a random sampling weight and f_ξ is L_ξ-smooth. We assume independence between v_ξ and ∇f_ξ, which allows separation of expectations. Then, by the standard inequality for smooth functions, we have

$$\begin{aligned}
\mathbb{E}_\xi[\|g_\xi(\boldsymbol{x})\|^2] &= \mathbb{E}_\xi[\|v_\xi \nabla f_\xi(\boldsymbol{x})\|^2] \\
&= \mathbb{E}_\xi[v_\xi^2 \|\nabla f_\xi(\boldsymbol{x})\|^2] \\
&= \mathbb{E}_\xi[v_\xi^2] \cdot \mathbb{E}_\xi[\|\nabla f_\xi(\boldsymbol{x})\|^2] \\
&\leq \mathbb{E}_\xi[v_\xi^2] \cdot \mathbb{E}_\xi[2L(f_\xi(\boldsymbol{x}) - f_\xi^{\inf})] \\
&= 2L\mathbb{E}_\xi[v_\xi^2] \cdot \mathbb{E}_\xi[f_\xi(\boldsymbol{x}) - f_\xi^{\inf}] \\
&= 2L\mathbb{E}_\xi[v_\xi^2] \cdot \mathbb{E}_\xi[f_\xi(\boldsymbol{x}) - f^{\inf} + f^{\inf} - f_\xi^{\inf}] \\
&= 2L\mathbb{E}_\xi[v_\xi^2] \cdot \big((\mathbb{E}_\xi[f_\xi(\boldsymbol{x})] - f^{\inf}) + \mathbb{E}_\xi[f^{\inf} - f_\xi^{\inf}]\big) \\
&= 2L\mathbb{E}_\xi[v_\xi^2] \cdot (f(\boldsymbol{x}) - f^{\inf}) + 2L\mathbb{E}_\xi[v_\xi^2] \cdot \mathbb{E}_\xi[f^{\inf} - f_\xi^{\inf}] \\
&= 2A(f(\boldsymbol{x}) - f^{\inf}) + C,
\end{aligned}$$

where

$$A = L\mathbb{E}_\xi[v_\xi^2], \quad B = 0, \quad C = 2A \cdot \mathbb{E}_\xi[f^{\inf} - f_\xi^{\inf}].$$

This calculation demonstrates the basic form of the ES property for a single-sample stochastic gradient, explicitly relating the second moment of the stochastic gradient to the suboptimality of the objective.

Subsampling Method	**A**	$\mathbb{E}[v_\xi^2]$
(i) Sampling with replacement	$A = \max_i \dfrac{L_i}{\tau n q_i} = \max_i L_i \mathbb{E}[v_i^2]$	$\mathbb{E}[v_i^2] = \dfrac{1}{\tau n q_i}$
(ii) Independent sampling without replacement	$A = \max_i \dfrac{(1-p_i)L_i}{p_i n} = \max_i L_i \mathbb{E}[v_i^2]$	$\mathbb{E}[v_i^2] = \dfrac{1-p_i}{p_i n}$
(iii) τ-Nice sampling without replacement	$A = \dfrac{n-\tau}{\tau(n-1)} \max_i L_i = \max_i L_i \mathbb{E}[v_i^2]$	$\mathbb{E}[v_i^2] = \dfrac{n-\tau}{\tau(n-1)}$

Table 2: Expected Smoothness constants under different subsampling strategies.

This table enumerates the constants A corresponding to common sampling schemes, highlighting how the sampling strategy affects the expected squared norm of the stochastic gradient.

Consider the finite-sum objective

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

and define the minibatch stochastic gradient estimator as

$$\nabla f_{\boldsymbol{v}}(\boldsymbol{x}) = \frac{1}{m} \sum_{i=1}^m v_{\xi_i} \nabla f_{\xi_i}(\boldsymbol{x}), \quad \boldsymbol{v} = (v_{\xi_1}, \ldots, v_{\xi_m}), \quad \boldsymbol{\xi} = (\xi_1, \ldots, \xi_m).$$

By applying Jensen's inequality and the ES property to individual samples, we obtain

$$\mathbb{E}_{\boldsymbol{\xi}}[\|\nabla f_{\boldsymbol{v}}(\boldsymbol{x})\|^2] = \mathbb{E}_{\xi_k}\left[\left\|\frac{1}{m} \sum_{i=1}^m v_{\xi_i} \nabla f_{\xi_i}(\boldsymbol{x})\right\|^2\right]$$

$$
\begin{aligned}
&\leq \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\xi_i}[\|v_{\xi_i}\nabla f_{\xi_i}(\boldsymbol{x})\|^2] \\
&\leq \frac{1}{m}\sum_{i=1}^{m}\left(2A_i(f(\boldsymbol{x})-f^{\inf})+C_i\right) \\
&= 2\left(\frac{1}{m}\sum_{i=1}^{m}A_i\right)(f(\boldsymbol{x})-f^{\inf})+\frac{1}{m}\sum_{i=1}^{m}C_i \\
&= \frac{2}{m}A_m(f(\boldsymbol{x})-f^{\inf})+\frac{1}{m}C_m \\
&= \frac{1}{m}\left(2A_m(f(\boldsymbol{x})-f^{\inf})+C_m\right),
\end{aligned}
$$

where $A_i = L\mathbb{E}_{\xi_i}[v_{\xi_i}^2]$, $C_i = \mathbb{E}_{\xi_i}[f^{\inf} - f_{\xi_i}^{\inf}]$, $A_m = \sum_{i=1}^m A_i$, $C_m = \sum_{i=1}^m C_i$. This derivation explicitly constructs the minibatch ES property, showing how the variance bound scales with the batch size.

Applying the L-smoothness property and the SGD update $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k \nabla f_{\boldsymbol{v}}(\boldsymbol{x}_k)$, the one-step expected descent is

$$
\begin{aligned}
f(\boldsymbol{x}_{k+1}) &\leq f(\boldsymbol{x}_k) - \eta_k\langle\nabla f(\boldsymbol{x}_k), \nabla f_{\boldsymbol{v}}(\boldsymbol{x}_k)\rangle + \frac{L\eta_k^2}{2}\|\nabla f_{\boldsymbol{v}}(\boldsymbol{x}_k)\|^2, \\
\mathbb{E}_{\boldsymbol{v}}[f(\boldsymbol{x}_{k+1})] &\leq f(\boldsymbol{x}_k) - \eta_k\|\nabla f(\boldsymbol{x}_k)\|^2 + \frac{L\eta_k^2}{2m}(2A_m(f(\boldsymbol{x}_k)-f^{\inf})+C_m) \\
&= f(\boldsymbol{x}_k) - \eta_k\|\nabla f(\boldsymbol{x}_k)\|^2 + \frac{L\eta_k^2}{m}A_m(f(\boldsymbol{x}_k)-f^{\inf}) + \frac{L\eta_k^2}{2m}C_m.
\end{aligned}
$$

Finally, summing over K iterations, we obtain a telescopic sum and isolate the minimum expected squared gradient norm:

$$
\begin{aligned}
\sum_{k=0}^{K-1}\eta_k\mathbb{E}[\|\nabla f(\boldsymbol{x}_k)\|^2] &\leq f(\boldsymbol{x}_0) - f^{\inf} + \sum_{k=0}^{K-1}\eta_k^2\left(\frac{L}{m}A_m(f(\boldsymbol{x}_0)-f^{\inf}) + \frac{L}{2m}C_m\right), \\
\min_{0\leq k\leq K-1}\mathbb{E}[\|\nabla f(\boldsymbol{x}_k)\|^2] &\leq \frac{f(\boldsymbol{x}_0)-f^{\inf}}{\sum_{k=0}^{K-1}\eta_k} + \frac{\sum_{k=0}^{K-1}\eta_k^2\left(\frac{L}{m}A_m(f(\boldsymbol{x}_0)-f^{\inf})+\frac{L}{2m}C_m\right)}{\sum_{k=0}^{K-1}\eta_k}.
\end{aligned}
$$

This inequality establishes the non-asymptotic convergence rate of SGD under the Expected Smoothness assumption, explicitly showing how the step sizes, batch size, and sampling variance constants influence convergence.

K Variance Decomposition with Sampling Weights

Let us consider a weighted stochastic gradient with sampling weights $\boldsymbol{v}_k = (v_{k,1}, \ldots, v_{k,n})^\top$:

$$
\nabla f_{\boldsymbol{v}_k}(\boldsymbol{x}_k) := \frac{1}{n}\sum_{i=1}^{n} v_{k,i}\nabla f_i(\boldsymbol{x}_k),
$$

where $\mathbb{E}[v_{k,i}] = 1$ for all i and $v_{k,i}$ is independent of x_k.

$$
\mathbb{E}[\nabla f_{\boldsymbol{v}_k}(\boldsymbol{x}_k)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[v_{k,i}]\nabla f_i(\boldsymbol{x}_k) = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\boldsymbol{x}_k) = \nabla f(\boldsymbol{x}_k).
$$

Start from the squared norm:

$$
\|\nabla f_{\boldsymbol{v}_k}(\boldsymbol{x}_k)\|^2 = \left\|\frac{1}{n}\sum_{i=1}^{n} v_{k,i}\nabla f_i(\boldsymbol{x}_k)\right\|^2
$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} v_{k,i}^2 \|\nabla f_i(\boldsymbol{x}_k)\|^2 + \frac{1}{n^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n} v_{k,i}v_{k,j}\langle \nabla f_i(\boldsymbol{x}_k), \nabla f_j(\boldsymbol{x}_k)\rangle$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} v_{k,i}^2 \|\nabla f_i(\boldsymbol{x}_k)\|^2 + \frac{2}{n^2}\sum_{1\leq i<j\leq n} v_{k,i}v_{k,j}\langle \nabla f_i(\boldsymbol{x}_k), \nabla f_j(\boldsymbol{x}_k)\rangle.$$

Taking the expectation over $\boldsymbol{v}_k$ yields

$$\mathbb{E}[\|\nabla f_{\boldsymbol{v}_k}(\boldsymbol{x}_k)\|^2] = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}[v_{k,i}^2]\|\nabla f_i(\boldsymbol{x}_k)\|^2 + \frac{2}{n^2}\sum_{1\leq i<j\leq n} \mathbb{E}[v_{k,i}v_{k,j}]\langle \nabla f_i(\boldsymbol{x}_k), \nabla f_j(\boldsymbol{x}_k)\rangle.$$

K.1 (i) Sampling with Replacement

Let $S_i \sim \text{Binomial}(\tau, q_i)$ and $v_i := S_i/(\tau q_i)$. Then,

$$\mathbb{E}[v_i^2] = 1 - \frac{1}{\tau} + \frac{1}{\tau q_i}, \quad \mathbb{E}[v_i v_j] = 1 - \frac{1}{\tau}, \quad i \neq j.$$

Plug into the decomposition:

$$\begin{aligned}
\mathbb{E}[\|\nabla f_{\boldsymbol{v}}(\boldsymbol{x}_k)\|^2] &= \frac{1}{n^2}\sum_{i=1}^{n}\left(1 - \frac{1}{\tau} + \frac{1}{\tau q_i}\right)\|\nabla f_i\|^2 + \frac{2}{n^2}\sum_{1\leq i<j\leq n}\left(1 - \frac{1}{\tau}\right)\langle \nabla f_i, \nabla f_j\rangle \\
&= \left(1 - \frac{1}{\tau}\right)\frac{1}{n^2}\left(\sum_{i=1}^{n}\|\nabla f_i\|^2 + 2\sum_{1\leq i<j\leq n}\langle \nabla f_i, \nabla f_j\rangle\right) + \frac{1}{\tau n^2}\sum_{i=1}^{n}\frac{1}{q_i}\|\nabla f_i\|^2 \\
&= \left(1 - \frac{1}{\tau}\right)\|\nabla f(\boldsymbol{x}_k)\|^2 + \frac{1}{\tau n^2}\sum_{i=1}^{n}\frac{1}{q_i}\|\nabla f_i\|^2.
\end{aligned}$$

Use L_i-smoothness and $f_i^{\inf}$:

$$\|\nabla f_i(\boldsymbol{x}_k)\|^2 \leq 2L_i(f_i(\boldsymbol{x}_k) - f_i^{\inf}) \quad \Rightarrow \quad \frac{1}{n^2}\sum_i \frac{1}{q_i}\|\nabla f_i\|^2 \leq 2\max_i \frac{L_i}{\tau n q_i}(f(\boldsymbol{x}_k) - f^{\inf} + \Delta^{\inf}),$$

giving

$$A = \max_i \frac{L_i}{\tau n q_i}, \quad B = 1 - \frac{1}{\tau}, \quad C = 2A\Delta^{\inf}.$$

K.2 (ii) Independent Sampling without Replacement

Let $v_i = \mathbb{1}_{(i\in S)}/p_i$. Then,

$$\mathbb{E}[v_i^2] = \frac{1}{p_i}, \quad \mathbb{E}[v_i v_j] = 1, \quad i \neq j.$$

Full decomposition:

$$\begin{aligned}
\mathbb{E}[\|\nabla f_{\boldsymbol{v}}(\boldsymbol{x}_k)\|^2] &= \frac{1}{n^2}\sum_i \frac{1}{p_i}\|\nabla f_i\|^2 + \frac{2}{n^2}\sum_{i<j}\langle \nabla f_i, \nabla f_j\rangle \\
&= \|\nabla f(\boldsymbol{x}_k)\|^2 + \frac{1}{n^2}\sum_i\left(\frac{1}{p_i} - 1\right)\|\nabla f_i\|^2.
\end{aligned}$$

$$\frac{1}{n^2}\sum_i\left(\frac{1}{p_i} - 1\right)\|\nabla f_i\|^2 \leq 2\max_i \frac{(1-p_i)L_i}{p_i n}(f(\boldsymbol{x}_k) - f^{\inf} + \Delta^{\inf}),$$

so

$$A = \max_i \frac{(1-p_i)L_i}{p_i n}, \quad B = 1, \quad C = 2A\Delta^{\inf}.$$

K.3 (iii) τ-Nice Sampling without Replacement

Let $v_i = \frac{n}{\tau}\mathbb{1}_{(i\in S)}$, $p_i = \tau/n$. Then,

$$\mathbb{E}[v_i^2] = \frac{n}{\tau}, \quad \mathbb{E}[v_i v_j] = \frac{n(\tau-1)}{\tau(n-1)}, \quad i \neq j.$$

Full expansion:

$$\begin{aligned}\mathbb{E}[\|\nabla f_{\boldsymbol{v}}(\boldsymbol{x}_k)\|^2] &= \frac{n}{\tau n^2}\sum_i \|\nabla f_i\|^2 + \frac{2}{n^2}\sum_{i<j}\frac{n(\tau-1)}{\tau(n-1)}\langle \nabla f_i, \nabla f_j\rangle \\ &= \frac{1}{\tau n}\sum_i \|\nabla f_i\|^2 + \frac{n(\tau-1)}{\tau(n-1)}\|\nabla f(\boldsymbol{x}_k)\|^2.\end{aligned}$$

L Derivations for Global Convergence Rates (Explicit Constants)

We start from the one-step descent lemma (Proof 3):

$$\mathbb{E}[f(\boldsymbol{x}_{k+1}) \mid \boldsymbol{x}_k] \leq f(\boldsymbol{x}_k) - \eta_k\|\nabla f(\boldsymbol{x}_k)\|^2 + \frac{\bar{L}\eta_k^2}{2}\mathbb{E}[\|g_k\|^2 \mid \boldsymbol{x}_k]. \tag{15}$$

Apply Expected Smoothness (Proof 4):

$$\mathbb{E}[\|g_k\|^2 \mid \boldsymbol{x}_k] \leq 2A(f(\boldsymbol{x}_k) - f^\star) + B\|\nabla f(\boldsymbol{x}_k)\|^2 + C.$$

Substitute to get:

$$\begin{aligned}\mathbb{E}[f(\boldsymbol{x}_{k+1}) \mid \boldsymbol{x}_k] &\leq f(\boldsymbol{x}_k) - \eta_k\|\nabla f(\boldsymbol{x}_k)\|^2 + \frac{\bar{L}\eta_k^2}{2}\left[2A(f(\boldsymbol{x}_k) - f^\star) + B\|\nabla f(\boldsymbol{x}_k)\|^2 + C\right] \\ &= f(\boldsymbol{x}_k) - \eta_k\left(1 - \frac{\bar{L}\eta_k B}{2}\right)\|\nabla f(\boldsymbol{x}_k)\|^2 + \bar{L}\eta_k^2 A(f(\boldsymbol{x}_k) - f^\star) + \frac{\bar{L}\eta_k^2}{2}C.\end{aligned}$$

Assuming sufficiently small η_k so that $1 - \bar{L}\eta_k B/2 > 0$, we can telescope over K iterations and use $\mathbb{E}[f(\boldsymbol{x}_k) - f^\star] \leq f(\boldsymbol{x}_0) - f^\star$ to obtain:

$$\min_{0\leq k\leq K}\mathbb{E}[\|\nabla f(\boldsymbol{x}_k)\|^2] \leq \underbrace{\frac{2(f(\boldsymbol{x}_0) - f^\star)}{\sum_{k=0}^{K-1}\eta_k}}_{C_1} + \underbrace{\frac{\bar{L}}{\sum_{k=0}^{K-1}\eta_k}\sum_{k=0}^{K-1}\eta_k^2\left(A + \frac{C}{2(f(\boldsymbol{x}_0) - f^\star)}\right)}_{C_2\cdot\frac{\sum\eta_k^2}{\sum\eta_k}}. \tag{16}$$

Explicitly, define:

$$C_1 = \frac{2(f(\boldsymbol{x}_0) - f^\star)}{1}, \quad C_2 = \bar{L}\left(A + \frac{C}{2(f(\boldsymbol{x}_0) - f^\star)}\right).$$

(a) Constant step size $\eta_k \equiv \eta$

$$\sum_{k=0}^{K-1}\eta_k = K\eta, \quad \sum_{k=0}^{K-1}\eta_k^2 = K\eta^2 \Rightarrow \min\mathbb{E}[\|\nabla f\|^2] \leq \frac{C_1}{K\eta} + C_2\frac{K\eta^2}{K\eta} = O\left(\frac{1}{K}\right) + O(\eta)$$

(b) Harmonic step size $\eta_k = \eta/(k+1)$

$$\sum_{k=0}^{K-1}\eta_k \sim \eta\ln K, \quad \sum_{k=0}^{K-1}\eta_k^2 \sim \eta^2\frac{\pi^2}{6}$$

$$\Rightarrow \min\mathbb{E}[\|\nabla f\|^2] \leq \frac{C_1}{\eta\ln K} + C_2\frac{\eta^2\pi^2/6}{\eta\ln K} = O(1/\ln K)$$

(c) Polynomial step size $\eta_k = \eta/(k+1)^\alpha, 0 < \alpha < 1$

$$\sum_{k=0}^{K-1} \eta_k \approx \eta \int_1^K t^{-\alpha} dt = \frac{\eta}{1-\alpha}(K^{1-\alpha} - 1) \sim O(K^{1-\alpha})$$

$$\sum_{k=0}^{K-1} \eta_k^2 \approx \eta^2 \int_1^K t^{-2\alpha} dt = \begin{cases} O(K^{1-2\alpha}) & 2\alpha < 1 \\ O(\log K) & 2\alpha = 1 \\ O(1) & 2\alpha > 1 \end{cases}$$

$$\Rightarrow \min \mathbb{E}[\|\nabla f\|^2] = O(K^{\alpha-1})$$

(d) Cosine annealing step size $\eta_k = \eta \frac{1+\cos(\pi k/K)}{2}$

$$\sum_{k=0}^{K-1} \eta_k \approx K \int_0^1 \eta \frac{1+\cos(\pi x)}{2} dx = \frac{K\eta}{2}$$

$$\sum_{k=0}^{K-1} \eta_k^2 \approx K \int_0^1 \left(\eta \frac{1+\cos(\pi x)}{2}\right)^2 dx = \frac{3K\eta^2}{8}$$

$$\Rightarrow \min \mathbb{E}[\|\nabla f\|^2] \leq \frac{C_1}{K\eta/2} + C_2 \frac{3K\eta^2/8}{K\eta/2} = O(1/K) + O(\eta)$$

Here, C_1, C_2 are explicitly written in terms of A, B, C from Proof 4, $\bar{L}$, and initial suboptimality.

All derivations now show the complete chain from ES constants to global convergence rates without omitting steps.