HIGH-PROBABILITY GUARANTEES FOR SGD UNDER β-HEAVY-TAILED GRADIENT NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Stochastic gradient descent (SGD) is widely used to train machine learning models, but subsampling the training data introduces noise into its updates. The strength and applicability of high-probability guarantees therefore depend critically on how the tails of gradient noise are modeled. Reports of heavy-tailed gradient noise in deep learning motivate relaxing the bounded-noise and sub-Gaussian assumptions commonly used in high-probability analyses of SGD. We use Young functions from Orlicz space theory to describe noise tails in a common framework. We model SGD gradient noise by adopting a Young function that preserves the finiteness of all polynomial moments while allowing tails heavier than sub-Weibull, including lognormal distributions. The resulting class is called β-heavy-tailed, with β controlling the tail heaviness. We establish concentration inequalities for β-heavy-tailed noise and combine them with a uniform bound on the difference between empirical and population gradients along the SGD trajectory to obtain high-probability bounds on optimization and population-risk stationarity for smooth nonconvex losses under trajectory assumptions. The bounds are not restricted to a particular learning-rate decay rule and make explicit the effects of noise tails and learning-rate schedules. Under the Polyak–Łojasiewicz condition, we bound the risk at the last iterate. We also analyze SGD with gradient clipping under the β-heavy-tailed noise model.

1 INTRODUCTION

Stochastic gradient descent (SGD) is a principal optimization algorithm used to train modern machine learning models (Bottou et al., 2018), and characterizing its generalization behavior is a central problem in learning theory. SGD reduces computational cost by computing a gradient from a subset of the training data at each step. This introduces a random difference between that gradient and the full-batch gradient over the entire training set, referred to as gradient noise. Quantifying how this noise affects convergence and learning through successive iterations requires control of its magnitude and the probability of large deviations. Existing analyses therefore impose assumptions such as bounded variance, with boundedness or sub-Gaussian tail conditions often used to obtain high-probability guarantees. However, heavy-tailed gradient noise that is poorly described by a Gaussian approximation has been reported in deep learning (Simsekli et al., 2019; Zhang et al., 2020; Şimşekli, 2025). For quantities distinct from gradient noise itself, near-lognormal distributions have also been reported for the magnitudes of gradients backpropagated through networks (Chmiel et al., 2021), and asymptotic lognormality has been established for gradient norms in randomly initialized fully connected ReLU networks (Hanin & Nica, 2020). These findings motivate analyzing SGD under weaker tail conditions.

Existing analyses of heavy tails take two approaches. One assumes only a bounded p-th moment for some $p \in (1, 2]$, allowing infinite variance when $p < 2$. In this setting, gradient clipping limits the magnitude of a stochastic gradient when its norm exceeds a threshold, reducing the effect of rare large gradients (Zhang et al., 2020). High-probability convergence guarantees have been obtained for methods combining clipping with momentum and normalization (Cutkosky & Mehta, 2021). Clipping, however, changes the update rule and generally introduces bias into the gradient estimate. The other approach relaxes sub-Gaussian assumptions to sub-Weibull conditions and obtains high-probability guarantees for SGD (Li & Liu, 2022; Madden et al., 2024). Lognormal distributions still

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

fall outside this setting: their tails decay more slowly than $\exp(-ct^p)$ for every $p > 0$ and $c > 0$. Although they satisfy the finite p-th moment assumption, they have all polynomial moments, a structure that analyses based only on low-order moments do not fully use. We study high-probability guarantees for SGD in this intermediate regime, which permits tails heavier than sub-Weibull while retaining the finiteness of all polynomial moments.

To describe this intermediate regime, we use the class of β-heavy-tailed random variables introduced by Chamakh et al. (2021) in their study of Orlicz norms and concentration inequalities. This class is defined through the Orlicz norm associated with the Young function $\Psi_\beta(x) = \exp(\log^\beta(x+1)) - 1$ for $\beta > 1$. Sub-Gaussian and sub-exponential conditions are Orlicz conditions associated with $\exp(x^2) - 1$ and $\exp(x) - 1$, respectively; a more slowly growing Young function allows heavier tails. A random variable satisfying $\|X\|_{\Psi_\beta} < \infty$ (Definition 2.1) has a tail probability bounded by an expression of the form $\exp(-c(\log t)^\beta)$ (Proposition A.18). This bound decays faster than any polynomial rate t^{-p}, ensuring that all polynomial moments are finite, but more slowly than $\exp(-ct^\epsilon)$ for any $\epsilon > 0$, allowing tails beyond the sub-Weibull scale. Smaller values of β give weaker tail conditions, and larger values give stronger ones. Our approach is to adopt this probabilistic framework as a model for SGD gradient noise and develop martingale concentration inequalities that control its accumulation along the trajectory, yielding a high-probability analysis of SGD.

Contributions. Our main contributions are as follows.

- We formulate a β-heavy-tailed gradient noise model (Assumption 2.2) and develop concentration tools for its analysis. We show that the associated generating functions are equivalent to Young functions even when squaring induces nonconvexity, enabling control of sums for every $\beta > 1$ (Lemma A.17). We also prove martingale concentration inequalities, including a Freedman-type bound that accommodates noise scales and variance proxies depending on the past trajectory (Proposition A.23).
- For SGD, we combine empirical optimization and uniform gradient generalization bounds to control the weighted average squared gradient norm of the population risk with high probability (Theorems 3.1–3.4). The bounds accommodate different step-size schedules; we instantiate them for $1/t$, $1/\sqrt{t}$, and cosine decay.
- Under a Polyak–Łojasiewicz (PL) condition on the empirical risk and a suitable $1/t$-type schedule, we obtain a high-probability $1/T$ rate for the last-iterate optimization error, up to logarithmic and tail factors (Theorem 3.7). With a PL condition on the population risk as well, $T \asymp n$ iterations yield a d/n excess population risk rate up to similar factors (Theorem 3.8).
- Under the same noise model, a noise-dependent clipping threshold yields a high-probability guarantee without the gradient cap (Assumption 2.5) or a smoothness-based step-size upper bound (Theorem 3.9). The guarantee is quadratic in the gradient norm below the threshold and linear above it, retaining the same tail factor as the bound for SGD.

Organization. Section 1.1 reviews related work, and Section 2 introduces the notation and assumptions. Section 3 presents the main bounds and instantiates them for specific step-size schedules. Sections 3.3 and 3.4 address the PL condition and gradient clipping.

1.1 Related Work

High-probability guarantees for SGD. Nonconvex SGD has been analyzed in expectation (Ghadimi & Lan, 2013), while Madden et al. (2024) give high-probability convergence guarantees under sub-Weibull noise. The work closest to ours, Li & Liu (2022), analyzes both optimization and generalization under the same tail condition. We relax this condition and express the guarantees through the sum of the step sizes and the sum of their squares, rather than restricting the analysis to the $1/\sqrt{t}$ decay used in that work. Approaches to generalization include algorithmic stability (Bousquet & Elisseeff, 2002; Hardt et al., 2016; Lei & Ying, 2020; Lei & Tang, 2021) and uniform convergence of empirical to population gradients (Mei et al., 2018; Foster et al., 2018). To evaluate stationarity, we use the latter approach through the uniform gradient bound of Lei & Tang (2021) (Lemma A.7). Liu & Zhou (2024) remove the compact-domain and bounded-noise restrictions from last-iterate analyses of convex composite objectives under a stochastic oracle.

Under finite p-th moment assumptions, high-probability convergence rates for methods using gradient clipping have been refined (Sadiev et al., 2023; Nguyen et al., 2023), and Liu & Zhou (2025) obtain an optimal rate without clipping by using batched normalized SGD with momentum. Koloskova et al. (2023) study the bias introduced by clipping and its convergence guarantees, while Armacki et al. (2025) consider more general nonlinear transformations. Our clipping analysis concerns noise with all polynomial moments finite and does not cover the full setting in which only low-order moments are assumed finite.

Heavy-tailed gradient noise in neural networks. The precise tail law of gradient noise remains under debate. Following the α-stable model of Simsekli et al. (2019), Xie et al. (2023) reject power-law tails for minibatch gradient noise, while Battash et al. (2024) revisit the question in support of an α-stable model. Gurbuzbalaban et al. (2021) also show that heavy-tailed stationary distributions of SGD iterates can arise from the dynamics themselves. Studies of low-precision training report gradient magnitudes concentrated near zero and spanning a wide dynamic range, without fitting a specific distribution family (Micikevicius et al., 2018; Zhu et al., 2020; Sun et al., 2020). Motivated by these observations, we control noise tails through an Orlicz condition without assuming a specific distribution family.

Orlicz norms and martingale concentration. Deviation bounds based on Orlicz norms have been used to refine Bernstein-type inequalities (van de Geer & Lederer, 2013) and to study concentration for sub-Weibull variables (Vladimirova et al., 2020; Kuchibhotla & Chakrabortty, 2022). Li & Liu (2024) give a general framework for broader families of heavy-tailed variables. In the β-heavy-tailed framework we use, the concentration inequalities of Chamakh et al. (2021) concern sums of independent variables. SGD noise ϵ_t, however, depends on the past sampling history through $\boldsymbol{w}_t$, so these results do not apply directly. We extend the analysis to the martingale setting, following Freedman (1975) (Proposition A.23).

2 Problem Setup

2.1 Notation

For $n \in \mathbb{N}$ we write $[n] := \{1, \dots, n\}$, and $\|\cdot\|$ for the Euclidean norm on $\mathbb{R}^d$. For $x \in \mathbb{R}$ we write $(x)_+ := \max\{x, 0\}$, and $\mathbb{1}_E$ for the indicator of an event E. For positive quantities a and b we write $a \asymp b$ if $c\, b \le a \le C\, b$ for constants $0 < c \le C$ that do not depend on n, d, T or δ. We abbreviate $\log^\beta(x) := (\log x)^\beta$, and for $\beta > 1$ define

$$\Psi_\beta(x) := \exp\Big(\log^\beta(x+1)\Big) - 1, \qquad \overline{\Psi}_\beta(x) := \exp\Big(-\log^\beta(x+1)\Big) = \frac{1}{\Psi_\beta(x)+1}.$$

Here Ψ_β is the Young function whose Orlicz norm $\|X\|_{\Psi_\beta}$ measures the tail of a random variable X throughout the paper. Appendix A.2 recalls the definition of a Young function and of its Orlicz norm, together with the basic properties we use.

Let $\mathcal{Z}$ be the sample space, a Borel subset of a Euclidean space, P a probability distribution on $\mathcal{Z}$, and $\mathcal{W} := \mathbb{R}^d$ the parameter space. The loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ is differentiable in its first argument, and the learning problem is the minimization over $\mathcal{W}$ of the population risk $\mathcal{L}(\boldsymbol{w}) := \mathbb{E}_{z\sim P}[\ell(\boldsymbol{w}; z)]$. We denote its infimum by $\mathcal{L}^\star := \inf_{\boldsymbol{w}\in\mathcal{W}} \mathcal{L}(\boldsymbol{w})$. Since P is accessible only through data, one minimizes instead the empirical risk $\mathcal{L}_S(\boldsymbol{w}) := \frac{1}{n}\sum_{i\in[n]} \ell(\boldsymbol{w}; z_i)$ over a training set $S = \{z_1, \dots, z_n\}$ of i.i.d. samples from P. We write $\boldsymbol{w}(S)$ for a minimizer of $\mathcal{L}_S$ over $\mathcal{W}$, which we assume to exist. We study the standard stochastic gradient descent iteration of Algorithm 1 applied to $\mathcal{L}_S$, where at each step an index j_t is drawn uniformly from $[n]$ and the iterate is updated along the corresponding per-sample gradient. Let T denote the total number of SGD updates (the time horizon). We fix $\boldsymbol{w}_1 = \mathbf{0}$, a normalization that places the origin at the initialization. For a general $\boldsymbol{w}_1$ independent of S, the results below hold with $\|\boldsymbol{w}_t\|$ replaced by $\|\boldsymbol{w}_t - \boldsymbol{w}_1\|$. The choice of the initialization is not studied here.

We work on the joint probability space of the training sample S and the sampled indices. Define

$$\mathcal{F}_0 := \sigma(S), \qquad \mathcal{F}_t := \sigma(S, j_1, \dots, j_t) \quad (t \ge 1).$$

Algorithm 1 Stochastic Gradient Descent

1: **Input:** initial point $\boldsymbol{w}_1 = \mathbf{0}$, step-size sequence $\{\eta_t\}_t$, dataset $S = \{z_1, \ldots, z_n\}$.
2: **for** $t = 1, \ldots, T$ **do**
3: Draw j_t uniformly from $[n]$.
4: $\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \eta_t \nabla \ell(\boldsymbol{w}_t; z_{j_t})$.
5: **end for**
6: **Output:** iterates $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{T+1}$.

At each step, j_t is drawn uniformly from $[n]$, independently of $\mathcal{F}_{t-1}$. Then $\boldsymbol{w}_t$ is $\mathcal{F}_{t-1}$-measurable, and $\mathbb{E}[\nabla \ell(\boldsymbol{w}_t; z_{j_t}) \mid \mathcal{F}_{t-1}] = \nabla \mathcal{L}_S(\boldsymbol{w}_t)$. In this paper, we focus on the per-step *gradient noise* defined by

$$\epsilon_t := \nabla \ell(\boldsymbol{w}_t; z_{j_t}) - \nabla \mathcal{L}_S(\boldsymbol{w}_t).$$

The sequence $(\epsilon_t)_t$ is a martingale difference sequence (MDS) with respect to $(\mathcal{F}_t)_t$. The next subsection models the tail of this sequence.

2.2 β-Heavy-Tailed Noise Model

We model the tail of a random variable using the Ψ_β ($\beta > 1$) Orlicz norm, following the general framework of Orlicz-norm-based tail bounds (cf. Chamakh et al. (2021)).

Definition 2.1 (β-heavy tail property of a random variable)**.** Let $\beta > 1$. A random variable X is *β-heavy tailed* with $K > 0$ if

$$\mathbb{E}\left[\Psi_\beta\left(\frac{|X|}{K}\right)\right] \le 1 \iff \mathbb{E}\left[\exp\left(\log^\beta\left(\frac{|X|}{K} + 1\right)\right)\right] \le 2.$$

Equivalently, the Orlicz norm is at most K, i.e. $\|X\|_{\Psi_\beta} \le K$ (Lemma A.12(iii) in Appendix A.2).

The prototypical example is the lognormal distribution. Every lognormal variable satisfies Definition 2.1 for some K whenever $1 < \beta < 2$. If $\log X$ is Gaussian with variance σ^2, the tail of X decays like $\exp(-\log^2 t/(2\sigma^2))$, matching the $\beta = 2$ tail rate of Proposition A.18 up to the constant in the exponent. At $\beta = 2$, Definition 2.1 holds for some K if and only if $\sigma^2 \le 1/2$, since the scale K does not change the leading coefficient of $\log^2 t$. Lognormal-type tails are the motivating example for this Orlicz condition in Chamakh et al. (2021). Related empirical motivation comes from the approximately lognormal magnitudes of gradients with respect to intermediate layer outputs reported by Chmiel et al. (2021).

Our noise assumption allows the conditional law of ϵ_t given $\mathcal{F}_{t-1}$ to depend on the past trajectory, while a fixed constant K uniformly bounds the conditional Ψ_β-Orlicz norms of $\|\epsilon_t\|$.

Assumption 2.2 (β-heavy tail assumption on the gradient noise)**.** There exist deterministic constants $\beta > 1$ and $K > 0$ such that, for every t, the gradient noise ϵ_t satisfies

$$\mathbb{E}[\epsilon_t \mid \mathcal{F}_{t-1}] = 0, \qquad \mathbb{E}\left[\Psi_\beta\left(\frac{\|\epsilon_t\|}{K}\right) \,\middle|\, \mathcal{F}_{t-1}\right] \le 1.$$

The Freedman-type inequality of Proposition A.23 requires the conditional form of the Orlicz bound above.

Remark 2.3 (Boundedness and noise scales). Some readers may wonder what heavy tails mean when gradients sampled from a finite dataset are necessarily bounded at a fixed iterate. Conditional on the dataset and past iterates, this noise indeed has finite support. Assumption 2.2 instead specifies a uniform Orlicz scale. Boundedness alone does not compare tail-based guarantees, whose scales and constants also matter; the paragraph immediately following Remark 7 in Madden et al. (2024) makes this point for bounded sub-Weibull noise. Appendix A.8 gives the corresponding calculation for our β-heavy-tailed model. In contrast to finite-data sampling, a population stochastic oracle can produce noise with an unbounded conditional law. In this setting, tail assumptions control the probability of large noise realizations without requiring an almost-sure bound; Zhu et al. (2026, Assumption 1.2 and Theorem 4.4) analyze such sub-Weibull noise. That formulation directly optimizes the population risk, whereas our finite-data formulation retains the empirical-risk and generalization questions.

We also impose two assumptions widely used in the optimization and generalization analysis of nonconvex learning problems.

Assumption 2.4 (Smoothness of the loss function)**.** There exists $b > 0$ such that $\|\nabla \ell(\boldsymbol{w}; z) - \nabla \ell(\boldsymbol{w}'; z)\| \leq b\|\boldsymbol{w} - \boldsymbol{w}'\|$ for any sample $z \in \mathcal{Z}$ and parameters $\boldsymbol{w}, \boldsymbol{w}' \in \mathcal{W}$.

Assumption 2.4 is the standard regularity condition in nonconvex stochastic optimization (Ghadimi & Lan, 2013; Hardt et al., 2016). We impose it on the per-sample loss uniformly in z, so that $\mathcal{L}_S$ and $\mathcal{L}$ are b-smooth as well, and it yields the descent lemma (Lemma A.5), the one-step inequality used to derive the trajectory bound of Theorem 3.1.

Assumption 2.5. There exists $G > 0$ such that $\eta_t \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\| \leq G$ for any dataset $S \in \mathcal{Z}^n$ and $t \in \mathbb{N}$.

Assumption 2.5 bounds the product of the step size and the empirical gradient norm. This is an additional trajectory condition; it does not follow from Assumption 2.2, which controls deviations from the empirical gradient but does not bound the empirical gradient itself.

3 Main Results

Assumptions 2.2 and 2.4 hold throughout the rest of the paper, with tail index $\beta > 1$, smoothness constant $b > 0$ and $b' := \sup_{z \in \mathcal{Z}} \|\nabla \ell(\mathbf{0}; z)\| < \infty$, and the step sizes are positive and nonincreasing. For SGD without clipping, we additionally assume Assumption 2.5 and $\eta_t \leq 1/(2b)$. We also assume throughout that there exists a deterministic constant $\Delta_0 < \infty$, independent of n and d, such that, for every sample size n, almost surely over $S \sim P^n$, $\mathcal{L}_S(\mathbf{0}) - \mathcal{L}_S(\boldsymbol{w}(S)) \leq \Delta_0$.

The implicit constants in $\mathcal{O}(\cdot)$ are deterministic and may depend on β and fixed problem parameters such as b, b', K, G, Δ_0 and, in the PL results, μ_0, μ, but not on S, n, d, T, δ. Proofs of all theorems stated in this section are deferred to the supplemental material.

3.1 High-Probability Bound on the SGD Trajectory

Our first result bounds the sum of squared empirical gradient norms along the SGD trajectory.

Theorem 3.1. *For any $\delta \in (0, 1)$, with probability $1 - \delta$,*

$$\sum_{t \in [T]} \eta_t \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2 = \mathcal{O}\left(\exp\Big(2 \log^{1/\beta}(T/\delta)\Big) \left(\log(e/\delta) + \sum_{t \in [T]} \eta_t^2 \right) \right).$$

The factor $\exp(2 \log^{1/\beta}(\cdot))$ arises from the β-heavy-tailed noise. The bound degrades as β decreases, that is, as the noise model admits heavier tails.

Remark 3.2. A threshold at $\beta \approx 1.519$ affects the constant factors in this analysis. Below this threshold, the generating function associated with the square of a β-heavy-tailed random variable loses convexity, requiring an additional constant factor when bounding sums. This parallels the dependence of sub-Weibull sum bounds on the tail parameter θ, where the generating function also loses convexity as the tail condition weakens. The rate stated above applies on both sides of the threshold; see Remark A.15 in the supplemental material for details.

Next, we establish a high-probability bound, uniform over all SGD iterates, on the squared discrepancy between the population and empirical risk gradients.

Theorem 3.3. *For any $\delta \in (0, 1)$, with probability $1 - \delta$ the generalization error of the gradient is bounded simultaneously for all iterates as*

$$\begin{aligned} &\max_{t \in [T+1]} \|\nabla \mathcal{L}(\boldsymbol{w}_t) - \nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2 \\ &= \mathcal{O}\left(\frac{\left(1 + \sum_{t \in [T]} \eta_t\right)(d + \log(1/\delta))}{n} \exp\Big(2 \log^{1/\beta}(T/\delta)\Big) \left(\log(e/\delta) + \sum_{t \in [T]} \eta_t^2 \right) \right). \end{aligned} \tag{1}$$

Finally, combining Theorems 3.1 and 3.3 bounds the step-size-weighted average squared norm of the population risk gradient along the SGD trajectory.

Theorem 3.4. *For any $\delta \in (0,1)$, with probability $1-\delta$ the step-size-weighted average squared gradient norm of the population risk is bounded as*

$$\frac{\sum_{t\in[T]} \eta_t \|\nabla \mathcal{L}(\boldsymbol{w}_t)\|^2}{\sum_{t\in[T]} \eta_t} = \mathcal{O}\left(\exp\Big(2\log^{1/\beta}(T/\delta)\Big)\left(\log(e/\delta) + \sum_{t\in[T]} \eta_t^2\right)\left(\frac{1}{\sum_{t\in[T]} \eta_t} + \frac{1+\sum_{t\in[T]} \eta_t}{n}(d+\log(1/\delta))\right)\right). \tag{2}$$

The two terms in the last factor of (2) correspond to the two sources of error. The first, $1/\sum_{t\in[T]} \eta_t$, is an optimization term that arises because SGD is run for only a finite number of steps, and it decreases as the cumulative step size $\sum_{t\in[T]} \eta_t$ grows. The second, $(1+\sum_{t\in[T]} \eta_t)(d+\log(1/\delta))/n$, is a generalization term inherited from Theorem 3.3. A longer trajectory travels farther from the initialization, enlarging the radius of the ball over which the uniform bound of Lemma A.7 must hold. This term therefore grows with $\sum_{t\in[T]} \eta_t$. The two terms balance when the cumulative step size satisfies $\sum_{t\in[T]} \eta_t \asymp \sqrt{n/(d+\log(1/\delta))}$, and this choice can be made without fixing a particular schedule. The following corollary bounds the step-size-weighted average squared norm of the population risk gradient under this choice of cumulative step size.

Corollary 3.5. *Under the hypotheses of Theorem 3.4, assume in addition that $n \geq d + \log(1/\delta)$ and that the horizon T is chosen so that*

$$\sum_{t\in[T]} \eta_t \asymp \sqrt{\frac{n}{d+\log(1/\delta)}}. \tag{3}$$

Then, with probability $1-\delta$,

$$\frac{\sum_{t\in[T]} \eta_t \|\nabla \mathcal{L}(\boldsymbol{w}_t)\|^2}{\sum_{t\in[T]} \eta_t} = \mathcal{O}\left(\exp\Big(2\log^{1/\beta}(T/\delta)\Big)\left(\log(e/\delta) + \sum_{t\in[T]} \eta_t^2\right)\sqrt{\frac{d+\log(1/\delta)}{n}}\right).$$

Condition (3) constrains the horizon only through the cumulative step size, so solving it for T once a schedule is fixed yields a deterministic iteration number as a function of n, d, and δ. The balancing cumulative step size is finite for a finite sample, so the corollary motivates early stopping within this bound. After imposing (3), the remaining dependence on T is through $\sum_{t\in[T]} \eta_t^2$ and the tail factor $\exp(2\log^{1/\beta}(T/\delta))$. For fixed $\beta > 1$, this factor grows more slowly than any positive power of T/δ, yet faster than any power of $\log(T/\delta)$.

3.2 Learning-Rate Decay

Our main theorems depend on the step sizes only through $\sum_{t\in[T]} \eta_t$ and $\sum_{t\in[T]} \eta_t^2$, leaving the decay schedule unspecified. We instantiate them with three representative schedules, namely the classical $1/t$ decay of stochastic approximation (Robbins & Monro, 1951), the $1/\sqrt{t}$ decay assumed by Li & Liu (2022), and cosine decay, the single-cycle case of the cosine annealing schedule of Loshchilov & Hutter (2017), which is common in modern deep learning, where the number of training steps is fixed in advance. Each is nonincreasing with $\eta_t \leq \eta_0 \leq 1/(2b)$, so all three meet the step-size condition of Section 3. Table 1 lists their step-size sums, stated as Propositions A.26, A.27 and A.28 in Appendix A.5.

The bound of Theorem 3.1 depends on the schedule only through $\sum_{t\in[T]} \eta_t^2$, and the fourth column of Table 1 shows that the three schedules differ in how this sum, and hence the bound, grows with T. Under $1/t$ decay the step-size sums contribute only the constant factor $\log(e/\delta) + \eta_0^2$, so all dependence on T comes from the tail factor and the bound remains sub-polynomial in T even as $T \to \infty$, whereas under cosine decay the bound grows linearly in T. The contrast stems from square-summability, since $1/t$ decay satisfies the classical conditions $\sum_{t=1}^{\infty} \eta_t = \infty$, $\sum_{t=1}^{\infty} \eta_t^2 < \infty$ of

Table 1: Step-size sums for the three schedules.

Schedule	η_t	$\sum_{t\in[T]} \eta_t$	$\sum_{t\in[T]} \eta_t^2$
$1/t$ (Prop. A.26)	$\frac{\eta_0}{t}$	$\eta_0(\log T + \mathcal{O}(1))$	$\mathcal{O}(\eta_0^2)$
$1/\sqrt{t}$ (Prop. A.27)	$\frac{\eta_0}{\sqrt{t}}$	$\eta_0(2\sqrt{T} + \mathcal{O}(1))$	$\eta_0^2(\log T + \mathcal{O}(1))$
Cosine (Prop. A.28)	$\frac{\eta_0}{2}\left(1 + \cos\frac{\pi t}{T}\right)$	$\frac{\eta_0}{2}(T-1)$	$\frac{\eta_0^2}{8}(3T-4)$

stochastic approximation, whereas cosine decay is designed for a fixed horizon T and its $\sum_{t\in[T]} \eta_t^2$ grows linearly in T. The $1/\sqrt{t}$ decay lies between the two. It is not square-summable either, but its $\sum_{t\in[T]} \eta_t^2$ grows only logarithmically.

The third column of Table 1 determines the horizon at which condition (3) is met. Because $\sum_{t\in[T]} \eta_t$ grows only logarithmically under $1/t$ decay, (3) is met only at the exponentially long horizon $T \asymp \exp(\eta_0^{-1}\sqrt{n/(d+\log(1/\delta))})$. A slower decay removes this restriction, because under $1/\sqrt{t}$ decay the cumulative step size diverges as $\sqrt{T}$, so (3) is solved by the polynomial horizon $T \asymp n/(\eta_0^2(d+\log(1/\delta)))$, and Corollary 3.5 then gives the rate $\sqrt{(d+\log(1/\delta))/n}$ up to the tail factor and the extra $\log T$.

3.3 Global Bound under the PL Condition

Theorems 3.1 and 3.4 bound gradient norms along the trajectory. To obtain guarantees for risk values, we introduce the Polyak–Łojasiewicz (PL) condition and analyze the empirical-risk optimization error at the last iterate, followed by the excess population risk $\mathcal{L}(\boldsymbol{w}_{T+1}) - \mathcal{L}^\star$.

A differentiable function F with $F^\star := \inf_{\boldsymbol{w}\in\mathcal{W}} F(\boldsymbol{w}) > -\infty$ satisfies the PL condition with $\mu > 0$ if, for every $\boldsymbol{w} \in \mathcal{W}$,

$$\|\nabla F(\boldsymbol{w})\|^2 \geq 2\mu\,(F(\boldsymbol{w}) - F^\star)\,. \tag{4}$$

The condition stems from Polyak (1963) and Łojasiewicz (1963). Karimi et al. (2016) describe its relations to other conditions under which gradient methods converge linearly. It is implied by μ-strong convexity but does not imply convexity, and it permits a continuum of minimizers. It excludes stationary points that are not global minimizers. Local PL conditions have been established for wide neural networks in a ball around initialization (Liu et al., 2022). Here we assume the PL condition globally on $\mathcal{W} = \mathbb{R}^d$; the local result alone does not establish this assumption. Charles & Papailiopoulos (2018) use the PL condition for stability and generalization guarantees.

Assumption 3.6. There exist deterministic constants $\mu_0, \mu > 0$, independent of n and d, such that, for every sample size n and every $S \in \mathcal{Z}^n$, the empirical risk $\mathcal{L}_S$ satisfies (4) on $\mathcal{W}$ with constant μ_0, and the population risk $\mathcal{L}$ satisfies (4) on $\mathcal{W}$ with constant μ.

The next theorem bounds the optimization error in the empirical risk at the last SGD iterate under the PL condition.

Theorem 3.7. *Under the empirical-risk condition in Assumption 3.6, fix $t_0 \geq \max\{8b/\mu_0,\, 1\}$ and set $\eta_t := 4/(\mu_0(t+t_0))$ for $t = 1, \ldots, T$. Then, for any $\delta \in (0,1)$ and $T \geq 2$, with probability $1-\delta$,*

$$\mathcal{L}_S(\boldsymbol{w}_{T+1}) - \mathcal{L}_S(\boldsymbol{w}(S)) = \mathcal{O}\left(\frac{\exp\left(2\log^{1/\beta}(T/\delta)\right)\sqrt{\log T\,\log(e/\delta)}}{T}\right).$$

The schedule in Theorem 3.7 is nonincreasing and satisfies $\eta_t \leq 4/(\mu_0 t_0) \leq 1/(2b)$, meeting the step-size condition of Section 3. Its step-size sums have the same orders as those of the $1/t$ schedule in Proposition A.26: $\sum_{t\in[T]} \eta_t = \mathcal{O}(\log T)$ and $\sum_{t\in[T]} \eta_t^2 = \mathcal{O}(1)$. Stating the results for a general nonincreasing schedule appears technically possible, but would substantially complicate the argument, so we do not pursue it here.

To pass from empirical to population risk, we also need to control the discrepancy between their gradients at the last iterate. Applying Theorem 3.3 with the step-size sums above gives, for any $\delta \in (0,1)$ and any $T \geq 2$, with probability $1-\delta$,

$$\max_{t\in[T+1]} \|\nabla\mathcal{L}(\boldsymbol{w}_t) - \nabla\mathcal{L}_S(\boldsymbol{w}_t)\|^2 = \mathcal{O}\left(\frac{\exp\Big(2\log^{1/\beta}(T/\delta)\Big)\,(d+\log(1/\delta))\log(e/\delta)\,\log T}{n}\right), \tag{5}$$

where $\sum_{t\in[T]}\eta_t^2$ is absorbed into $\log(e/\delta) \geq 1$.

This bound follows from the step-size schedule and does not require the PL condition. Combining it with Theorem 3.7 and the population PL condition in Assumption 3.6 bounds the excess population risk at the last SGD iterate.

Theorem 3.8. *Under Assumption 3.6, with the step-size schedule in Theorem 3.7 and a horizon $T \asymp n$ with $T \geq 2$, for any $\delta \in (0,1)$ and $n \geq 2$, with probability $1-\delta$,*

$$\mathcal{L}(\boldsymbol{w}_{T+1}) - \mathcal{L}^\star = \mathcal{O}\left(\frac{\exp\Big(2\log^{1/\beta}(n/\delta)\Big)\,(d+\log(1/\delta))\log(e/\delta)\,\log n}{n}\right).$$

Suppressing fixed problem parameters and the common tail factor, the two terms entering Theorem 3.8 are the optimization error of Theorem 3.7, of order $\sqrt{\log T}/T$, and the gradient generalization error (5), of order $\log T/n$. At $T \asymp n$, the generalization term dominates, and its growth with T limits the improvement that this bound can certify from further iterations.

3.4 SGD WITH GRADIENT CLIPPING

We next extend the analysis to SGD with gradient clipping under the same noise model. The preceding analysis of SGD uses Assumption 2.5 and the step-size restriction $\eta_t \leq 1/(2b)$. By choosing a clipping level based on the noise-tail bound, we obtain a high-probability guarantee without these two conditions. The resulting bound controls a quantity that is quadratic in the empirical gradient norm below the clipping level and linear above it, while retaining the same tail factor as the bound for SGD. Our chosen clipping level depends on K, β, T, and δ, making explicit how the noise-tail parameters affect this guarantee.

Algorithm 2 SGD with Gradient Clipping

1: **Input:** initial point $\boldsymbol{w}_1 = \mathbf{0}$, step-size sequence $\{\eta_t\}_t$, dataset $S = \{z_1, \ldots, z_n\}$, clipping level $\tau > 0$.
2: **for** $t = 1, \ldots, T$ **do**
3: Draw j_t uniformly from $[n]$.
4: $\tilde{\boldsymbol{g}}_t \leftarrow \nabla\ell(\boldsymbol{w}_t; z_{j_t}) \min\{1,\ \tau/\|\nabla\ell(\boldsymbol{w}_t; z_{j_t})\|\}$.
5: $\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \eta_t\tilde{\boldsymbol{g}}_t$.
6: **end for**
7: **Output:** iterates $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{T+1}$.

To control large noise realizations over all T steps, we choose a clipping level based on the tail bound of Proposition A.18, using a per-step exceedance probability of $\delta/(4T)$. Specifically, we set

$$\tau := 8K\exp\Big(\log^{1/\beta}(8T/\delta)\Big). \tag{6}$$

The clipping level τ depends on the Orlicz norm bound K, the tail index β, the horizon T, and the confidence parameter δ.

The following theorem bounds a step-size-weighted sum that is quadratic in the empirical gradient norm below the clipping level and linear above it.

Theorem 3.9. *Let $\beta > 1$, $T \geq 4$ and $\delta \in (0,1)$, let $\{\eta_t\}_{t\in[T]}$ be positive and nonincreasing, and suppose Assumptions 2.2 and 2.4 hold. Then Algorithm 2, with the clipping level τ of* (6)*, satisfies*

the following bound with probability $1-\delta$*:*

$$\sum_{t\in[T]} \eta_t \min\left\{\tau\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|,\ \|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|^2\right\} = \mathcal{O}\left(\exp\left(2\log^{1/\beta}(T/\delta)\right)\left(\log(e/\delta)+\sum_{t\in[T]}\eta_t^2\right)\right).$$

At a step with $\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\| \le \tau$ the minimum is the squared gradient norm and the summand matches that in Theorem 3.1. The other branch weakens the statement only where the empirical gradient exceeds the clipping level. An analogous two-regime stationarity measure, without the factor τ in the linear branch, is used in Li & Liu (2022, Theorem 3.19). Clipping bounds the contribution of the gradient noise by the level τ instead of by Assumption 2.5, and the two powers of τ that enter reproduce the tail factor of Theorem 3.1, so both that assumption and the step-size restriction $\eta_t \le 1/(2b)$ are removed without increasing that factor.

4 Conclusion

We establish high-probability optimization and generalization guarantees for SGD under the β-heavy-tailed condition, allowing tails beyond sub-Weibull while retaining all polynomial moments. Under smoothness and trajectory assumptions, we obtain empirical and population stationarity guarantees; with PL conditions, we bound last-iterate empirical optimization error and excess population risk. For gradient-clipped SGD, a two-regime stationarity bound removes the condition on step-size-scaled empirical gradients and the smoothness-dependent step-size restriction.

Our noise assumption requires a uniform deterministic bound on the conditional Orlicz scale. Estimating or bounding this scale along SGD trajectories and determining when uniformity is justified are concrete next steps. Tracking its dependence on sample size and training time would help assess the resulting guarantees. It would also be interesting to extend the analysis to scales that vary along the trajectory without requiring a uniform deterministic bound.

The fixed-dataset setting leaves open an extension to online SGD driven by a population stochastic oracle. Our restriction to standard and clipped SGD also motivates extending the analysis to a broader range of optimization methods, including SGD with momentum, Adam (Kingma & Ba, 2015), and the more recent Muon (Jordan et al., 2024).

More broadly, understanding how architecture, data distributions, learning dynamics, and accumulated numerical errors shape gradient noise would clarify when the β-heavy-tailed model applies and which other tail regimes arise. The update rule itself may alter subsequent noise distributions through its effect on the trajectory. This understanding could guide both the design of training methods suited to the noise they encounter and a general Orlicz-space framework for optimization and generalization across these regimes.

AI USE STATEMENT

We used generative AI tools to assist with writing and copyediting the manuscript, searching for and identifying related literature, and drafting sections of the paper. We also used these tools to assist with proof construction for technical lemmas and to check completed theorems and lemmas for correctness. We have reviewed all AI-assisted work. Every AI-assisted proof step was independently verified by the author against the underlying mathematics, and AI-assisted drafts and text edits were checked for accuracy before inclusion. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Aleksandar Armacki, Shuhua Yu, Pranay Sharma, Gauri Joshi, Dragana Bajovic, Dusan Jakovetic, and Soummya Kar. High-probability convergence bounds for online nonlinear stochastic gradient descent under heavy-tailed noise. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1774–1782, 2025.

Barak Battash, Lior Wolf, and Ofir Lindenbaum. Revisiting the noise model of stochastic gradient descent. In *International Conference on Artificial Intelligence and Statistics*, 2024.

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

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Linda Chamakh, Emmanuel Gobet, and Wenjun Liu. Orlicz norms and concentration inequalities for β-heavy tailed random variables. HAL preprint hal-03175697, version 3, 2021.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 745–754, 2018.

Brian Chmiel, Liad Ben-Uri, Moran Shkolnik, Elad Hoffer, Ron Banner, and Daniel Soudry. Neural gradients are near-lognormal: improved quantized and sparse training. In *International Conference on Learning Representations*, 2021.

Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. In *Advances in Neural Information Processing Systems*, 2021.

Dylan J. Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. In *Advances in Neural Information Processing Systems*, 2018.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In *International Conference on Machine Learning*, 2021.

Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376(1):287–322, 2020.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Petteri Harjulehto and Peter Hästö. *Orlicz Spaces and Generalized Orlicz Spaces*, volume 2236 of *Lecture Notes in Mathematics*. Springer, Cham, 2019. doi: 10.1007/978-3-030-15100-3.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak–łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pp. 795–811, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U. Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, 2023.

Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *Information and Inference*, 11(4):1389–1456, 2022.

Yunwen Lei and Ke Tang. Learning rates for stochastic gradient descent with nonconvex objectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4505–4511, 2021.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, 2020.

Shaojie Li and Yong Liu. High probability guarantees for nonconvex stochastic gradient descent with heavy tails. In *International Conference on Machine Learning*, pp. 12931–12963. PMLR, 2022.

Shaojie Li and Yong Liu. Concentration inequalities for general functions of heavy-tailed random variables. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28343–28367, 2024.

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

Zijian Liu and Zhengyuan Zhou. Revisiting the last-iterate convergence of stochastic gradient methods. In *International Conference on Learning Representations*, 2024. arXiv:2312.08531.

Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *International Conference on Learning Representations*, 2025. arXiv:2412.19529.

Stanisław Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. In *Les Équations aux Dérivées Partielles*, pp. 87–89. Éditions du Centre National de la Recherche Scientifique, Paris, 1963.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Liam Madden, Emiliano Dall'Anese, and Stephen Becker. High probability convergence bounds for non-convex stochastic gradient descent with sub-weibull noise. *Journal of Machine Learning Research*, 25(241):1–36, 2024.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations*, 2018.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, 2004.

Ta Duy Nguyen, Thien H Nguyen, Alina Ene, and Huy Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In *Advances in Neural Information Processing Systems*, volume 36, pp. 24191–24222, 2023.

Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994.

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

M. M. Rao and Z. D. Ren. *Theory of Orlicz Spaces*, volume 146 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker, New York, 1991.

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

Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: The case of unbounded variance. In *International Conference on Machine Learning*, 2023.

Umut Şimşekli. *Heavy Tails in Stochastic Optimization: Emergence and Implications*. PhD thesis, Ecole Normale Supérieure Paris, December 2025. URL `https://hal.science/tel-05434401`.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, 2019.

Xiao Sun, Naigang Wang, Chia-Yu Chen, Jiamin Ni, Ankur Agrawal, Xiaodong Cui, Swagath Venkataramani, Kaoutar El Maghraoui, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. Ultra-low precision 4-bit training of deep neural networks. In *Advances in Neural Information Processing Systems*, 2020.

Sara van de Geer and Johannes Lederer. The Bernstein–Orlicz norm and deviation inequalities. *Probability Theory and Related Fields*, 157(1–2):225–250, 2013. doi: 10.1007/s00440-012-0455-y.

Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. Sub-Weibull distributions: Generalizing sub-Gaussian and sub-Exponential properties to heavier tailed distributions. *Stat*, 9 (1):e318, 2020.

David Williams. *Probability with Martingales*. Cambridge University Press, 1991.

Zeke Xie, Qian-Yuan Tang, Mingming Sun, and Ping Li. On the overlooked structure of stochastic gradients. In *Advances in Neural Information Processing Systems*, 2023.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J. Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems*, 2020.

Feng Zhu, Ruihao Gong, Fengwei Yu, Xianglong Liu, Yanfei Wang, Zhelong Li, Xiuqi Yang, and Junjie Yan. Towards unified int8 training for convolutional neural network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Tianxi Zhu, Yi Xu, Qi Wang, and Xiangyang Ji. Stochastic gradient methods under heavy-tailed noises in weakly convex optimization. In *Forty-third International Conference on Machine Learning*, 2026.

A SUPPLEMENTAL MATERIAL

In this section we collect the technical lemmas and concentration inequalities used throughout the paper (Appendices A.1–A.3), together with the proofs of all theorems, lemmas, propositions, and corollaries stated in the main text (Appendices A.4–A.5).

A.1 GENERAL LEMMAS

For ease of reference, we collect here the standard lemmas used in the proofs throughout this appendix.

Lemma A.1 (Layer-cake representation)**.** *Let Ψ be a Young function (Definition A.10), and let Ψ' denote its (a.e. defined, nonnegative, nondecreasing) derivative, so that $\Psi(x) = \int_0^x \Psi'(w)\,dw$ for all $x \geq 0$. Then for any random variable W,*

$$\mathbb{E}[\Psi(|W|)] = \int_0^\infty \Psi'(w)\,\mathbb{P}(|W| > w)\,dw. \tag{7}$$

Proof. Writing $\Psi(|W|) = \int_0^{|W|} \Psi'(w)\,dw$ and applying Tonelli's theorem to the nonnegative integrand $\Psi'(w)\,\mathbb{1}_{w<|W|}$, we have

$$\begin{aligned}\mathbb{E}[\Psi(|W|)] = \mathbb{E}\left[\int_0^\infty \Psi'(w)\;\mathbb{1}_{w<|W|}\,dw\right] &= \int_0^\infty \Psi'(w)\,\mathbb{E}[\mathbb{1}_{w<|W|}]\,dw \\ &= \int_0^\infty \Psi'(w)\,\mathbb{P}(|W| > w)\,dw.\end{aligned}$$

□

Lemma A.2 (Doob's L^2 maximal inequality; Williams, 1991)**.** *Let $(\mathcal{F}_n)_{n=0}^N$ be a filtration and let $\{S_n\}_{n=0}^N$ be an $\mathbb{R}^d$-valued martingale with respect to $(\mathcal{F}_n)_{n=0}^N$ with $S_0 = 0$. Then for any $z > 0$,*

$$\mathbb{P}\left(\max_{n\in[N]} \|S_n\| \geq z\right) \leq \frac{\mathbb{E}\left[\|S_N\|^2\right]}{z^2}. \tag{8}$$

Lemma A.3 (Ville's maximal inequality; Howard et al., 2020, Lemma 1 and Section 6.1)**.** *Let $(\mathcal{F}_k)_{k=0}^n$ be a filtration and let $\{Z_k\}_{k=0}^n$ be a nonnegative, real-valued supermartingale with respect to $(\mathcal{F}_k)_{k=0}^n$, with $Z_0 \leq 1$ almost surely. Then, for any $z > 0$,*

$$\mathbb{P}\left(\max_{k\in[n]} Z_k \geq z\right) \leq \frac{1}{z}.$$

Lemma A.4 (Pinelis' inequality; Pinelis, 1994, Theorem 3.5)**.** *Let $\{w_i\}_{i\in[N]}$ be an $\mathbb{R}^d$-valued MDS with respect to the filtration $\mathcal{F}_i$ satisfying $\|w_i\| \leq b_i$ almost surely for deterministic constants $b_1, \ldots, b_N$. Then for any $N \geq 1$ and $z > 0$,*

$$\mathbb{P}\left(\max_{n\in[N]} \left\|\sum_{i\in[n]} w_i\right\| \geq z\right) \leq 2\exp\left(-\frac{z^2}{2\sum_{i\in[N]} b_i^2}\right). \tag{9}$$

We also record the following standard consequence of smoothness, used in the proof of Theorem 3.1.

Lemma A.5 (Nesterov, 2004, Lemma 1.2.3)**.** *Let $F : \mathbb{R}^d \to \mathbb{R}$ be differentiable. Suppose that, for some $b > 0$, $\|\nabla F(\boldsymbol{w}) - \nabla F(\boldsymbol{w}')\| \leq b\|\boldsymbol{w} - \boldsymbol{w}'\|$ for all $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$. Then, for all $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$,*

$$F(\boldsymbol{w}) - F(\boldsymbol{w}') \leq \langle \boldsymbol{w} - \boldsymbol{w}', \nabla F(\boldsymbol{w}')\rangle + \tfrac{b}{2}\|\boldsymbol{w} - \boldsymbol{w}'\|^2.$$

Under Assumption 2.4, this lemma applies to $F = \ell(\cdot; z)$ for each fixed $z \in \mathcal{Z}$, and to their average $F = \mathcal{L}_S$.

The next lemma converts a function-value gap into a bound on the gradient norm, in the direction opposite to (4).

Lemma A.6. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be differentiable and suppose that, for some $b > 0$, $\|\nabla F(\boldsymbol{w}) - \nabla F(\boldsymbol{w}')\| \le b\|\boldsymbol{w} - \boldsymbol{w}'\|$ for all $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$. Suppose also that F attains its minimum over $\mathbb{R}^d$ at $\boldsymbol{w}^\dagger$. Then, for every $\boldsymbol{w} \in \mathbb{R}^d$,*

$$\|\nabla F(\boldsymbol{w})\|^2 \le 2b\left(F(\boldsymbol{w}) - F(\boldsymbol{w}^\dagger)\right).$$

Proof. Fix $\boldsymbol{w} \in \mathbb{R}^d$ and set $\boldsymbol{v} := \boldsymbol{w} - \frac{1}{b}\nabla F(\boldsymbol{w}) \in \mathbb{R}^d$. Lemma A.5 gives

$$F(\boldsymbol{v}) - F(\boldsymbol{w}) \le \langle \boldsymbol{v} - \boldsymbol{w}, \nabla F(\boldsymbol{w})\rangle + \tfrac{1}{2}b\|\boldsymbol{v} - \boldsymbol{w}\|^2 = -\frac{1}{2b}\|\nabla F(\boldsymbol{w})\|^2,$$

and $F(\boldsymbol{v}) \ge F(\boldsymbol{w}^\dagger)$ gives the claim. □

The following external result, used in the proof of Theorem 3.3, gives a uniform generalization bound for the gradient of a nonconvex objective.

Lemma A.7 (Lei & Tang, 2021, Corollary 2)**.** *Let $\delta \in (0,1)$ and $R > 0$, write $B_R := \{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\| \le R\}$ for the centered ball of radius R, and let $S = \{z_1, \dots, z_n\}$ be a collection of i.i.d. samples. Suppose that Assumption 2.4 holds and that $b' := \sup_{z\in\mathcal{Z}} \|\nabla \ell(\mathbf{0}; z)\| < \infty$. Then, with probability at least $1 - \delta$,*

$$\sup_{\boldsymbol{w}\in B_R} \|\nabla \mathcal{L}(\boldsymbol{w}) - \nabla \mathcal{L}_S(\boldsymbol{w})\| \le \frac{bR + b'}{\sqrt{n}}\left(2 + 2\sqrt{48e\sqrt{2}(\log 2 + d\log(3e))} + \sqrt{2\log(1/\delta)}\right).$$

The following lemma justifies the rescalings of the confidence level δ by a constant factor performed at the end of the proofs of our main theorems.

Lemma A.8. *Let $\beta > 1$, $a \ge 1$ and $c > 0$. Then for every $x \ge 1$ we have*

$$\exp\Big(c\log^{1/\beta}(ax)\Big) \le \exp\Big(c\log^{1/\beta} a\Big)\exp\Big(c\log^{1/\beta} x\Big),$$

and for every $\delta \in (0,1)$ we have $\log(ea/\delta) \le (1 + \log a)\log(e/\delta)$.

Proof. The first claim follows from the subadditivity of $u \mapsto u^{1/\beta}$ on $[0,\infty)$ and $\log(ax) = \log a + \log x$. The second follows from $\log(ea/\delta) = \log(e/\delta) + \log a$ and $\log(e/\delta) \ge 1$. □

The next technical comparison is used in the truncated moment-generating function bound below.

Lemma A.9. *$\psi(s) := \log^\beta(1+s)/s$ ($s > 0$) attains its unique maximum at some $s^* > 0$, with $\log^\beta(1+s^*) < \beta^\beta$. Hence for any $m > 0$ and $u \in [0, m]$,*

$$\frac{u\log^\beta(1+m)}{m} \le \log^\beta(1+u) + \beta^\beta. \tag{10}$$

Proof. The sign of $\psi'(s)$ is that of $g(s) := \beta s/(1+s) - \log(1+s)$. Since $g(0) = 0$, $g'(s) = (\beta - 1 - s)/(1+s)^2$, and $g(s) \to -\infty$, g has a unique positive zero s^*, at which ψ attains its unique maximum and $\log(1+s^*) = \beta s^*/(1+s^*) < \beta$. For $s^* \le u \le m$, monotonicity gives $u\psi(m) \le u\psi(u) = \log^\beta(1+u)$; for $0 \le u < s^*$, $u\psi(m) \le s^*\psi(s^*) < \beta^\beta$. These bounds imply (10). □

A.2 Orlicz Spaces

We briefly collect the minimal background on Orlicz spaces needed to follow the constructions in this paper. We first recall Young functions and the Luxemburg functional. Squaring a β-heavy-tailed variable then leads us to the modified function $\Psi_{\beta,2^{-\beta}}$, which need not be convex. We handle this loss of convexity by showing that $\Psi_{\beta,c}$ is a weak Φ-function and hence is equivalent to a Young function. This ultimately yields a tail bound for sums with a comparison constant depending only on β and c. See Rao & Ren (1991) for a comprehensive treatment (including proofs of the facts below) and Chamakh et al. (2021) for the specialization to β-heavy tails used throughout this paper.

Definition A.10 (Young function)**.** A function $\Psi : [0,\infty) \to [0,\infty)$ is a *Young function* if it is nondecreasing, convex, and satisfies $\Psi(0) = 0$.

The function Ψ_β is a Young function for every $\beta > 1$. We next recall its standard Luxemburg norm.

Definition A.11 (Luxemburg norm and Orlicz space)**.** Let Ψ be a Young function, and let X be a real-valued random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The *Luxemburg norm* of X with respect to Ψ is

$$\|X\|_\Psi := \inf\left\{\lambda > 0 : \mathbb{E}\left[\Psi\left(\frac{|X|}{\lambda}\right)\right] \le 1\right\}$$

and the associated *Orlicz space* $L^\Psi(\Omega, \mathbb{P})$ is the set of all X with $\|X\|_\Psi < \infty$. (More generally this construction applies to measurable functions on any measure space, but we only need the probability-space case here.)

For a function Ψ that is merely nondecreasing and satisfies $\Psi(0) = 0$, we use the same formula and call the resulting quantity the *Luxemburg functional*. We need this extension for the nonconvex functions that arise after squaring. Although the triangle inequality can then fail, homogeneity and monotonicity remain valid; below we recover a triangle inequality up to a constant by passing to an equivalent Young function.

We use the following standard facts about Luxemburg functionals.

Lemma A.12 (Basic properties of the Luxemburg functional)**.** *Let $\Psi : [0,\infty) \to [0,\infty)$ be nondecreasing with $\Psi(0) = 0$, and let $K > 0$. Properties (i), (iii) and (iv) below need no further assumption, while (ii) assumes in addition that Ψ is convex, that is, that Ψ is a Young function.*

- *(i) (Homogeneity) $\|cX\|_\Psi = |c|\,\|X\|_\Psi$ for any random variable X and scalar c.*
- *(ii) (Triangle inequality) $\|X + Y\|_\Psi \le \|X\|_\Psi + \|Y\|_\Psi$ for any random variables X, Y.*
- *(iii) (Equivalent formulation) If Ψ is continuous, then $\|X\|_\Psi \le K$ if and only if $\mathbb{E}[\Psi(|X|/K)] \le 1$.*
- *(iv) (Monotonicity) If $|X| \le |Y|$ almost surely, then $\|X\|_\Psi \le \|Y\|_\Psi$.*

The proofs of Properties (i)–(iii) can be found in Rao & Ren (1991). Property (iv) is immediate from the definition of the Orlicz norm, since Ψ is nondecreasing on $[0,\infty)$, so that any $c > 0$ with $\mathbb{E}[\Psi(|Y|/c)] \le 1$ also satisfies $\mathbb{E}[\Psi(|X|/c)] \le 1$.

Property (iii), applied to $\Psi = \Psi_\beta$, is exactly the equivalence between the two formulations of β-heavy tailedness recorded in Definition 2.1.

To control squared noise, we introduce a coefficient $c > 0$ in the exponent of Ψ_β.

Definition A.13. We further parametrize the constant inside the exponent of Ψ_β by $c > 0$, defining

$$\Psi_{\beta,c}(x) := \exp(c \cdot \log^\beta(x+1)) - 1. \tag{11}$$

When $c = 1$ we abbreviate $\Psi_\beta = \Psi_{\beta,1}$. We similarly write

$$\overline{\Psi}_{\beta,c}(x) := \exp(-c\log^\beta(x+1)) = 1/(\Psi_{\beta,c}(x) + 1),$$

so that $\overline{\Psi}_\beta = \overline{\Psi}_{\beta,1}$ recovers the notation of Section 2.1.

Every $\Psi_{\beta,c}$ is continuous and increasing, vanishes at the origin, and tends to infinity. The following proposition explains why the parameter c is needed: the square of a β-heavy-tailed random variable remains β-heavy tailed after the coefficient in the exponent is reduced to at most $2^{-\beta}$.

Proposition A.14. *Let $\beta > 1$ and let X be a β-heavy-tailed random variable. Then*

$$\|X^2\|_{\Psi_{\beta,\lambda}} \le \|X\|_{\Psi_\beta}^2 \qquad \text{for every } \lambda \in (0, 2^{-\beta}].$$

Proof. For $u \ge 0$ and $0 < \lambda \le 2^{-\beta}$, the inequality $1 + u^2 \le (1+u)^2$ gives $\Psi_{\beta,\lambda}(u^2) \le \Psi_\beta(u)$. Thus, for every $K > \|X\|_{\Psi_\beta}$,

$$\mathbb{E}\big[\Psi_{\beta,\lambda}(X^2/K^2)\big] \le \mathbb{E}[\Psi_\beta(|X|/K)] \le 1.$$

Lemma A.12(iii) yields $\|X^2\|_{\Psi_{\beta,\lambda}} \le K^2$; letting $K \downarrow \|X\|_{\Psi_\beta}$ proves the claim. □

Remark A.15 (Convexity after squaring). For $\beta > 1$, writing $s = \log(1 + x)$, the sign of $\Psi''_{\beta,c}(x)$ is the sign of $c\beta s^{\beta-1} + (\beta - 1)/s - 1$, whose minimum is $\beta(c\beta)^{1/\beta} - 1$. Thus $\Psi_{\beta,c}$ is convex if and only if $c \geq \beta^{-(\beta+1)}$. The function Ψ_β is therefore convex, while $\Psi_{\beta,2^{-\beta}}$ is convex only for $\beta \geq \beta^\star$, where $\beta^\star = 1.51887\ldots$ solves $(\beta + 1)\log\beta = \beta\log 2$. For $1 < \beta < \beta^\star$, it is nonconvex in an intermediate region but convex near zero and for sufficiently large x. By comparison, the sub-Weibull function $\psi_\theta(x) = \exp(x^{1/\theta}) - 1$ is convex for $0 < \theta \leq 1$ and concave on $(0, (\theta - 1)^\theta)$ for $\theta > 1$. Squaring replaces θ by 2θ, so in both families squaring can require a nonconvex generating function even when the original one is convex.

This loss of convexity matters because the usual triangle inequality for the Luxemburg norm requires a Young function. Instead of restricting β, we use the fact that a nonconvex function can still be equivalent to a Young function if it is a weak Φ-function.

Definition A.16 (Weak Φ-function)**.** Following Harjulehto & Hästö (2019, Definitions 2.1.2 and 2.1.3), an increasing function $\Psi : [0, \infty) \to [0, \infty)$ with $\Psi(0) = 0$, $\lim_{x\downarrow 0} \Psi(x) = 0$, and $\lim_{x\to\infty} \Psi(x) = \infty$ is a *weak Φ-function* if $x \mapsto \Psi(x)/x$ is almost increasing on $(0, \infty)$. That is, there exists $A \geq 1$ such that

$$\frac{\Psi(u)}{u} \leq A\frac{\Psi(v)}{v} \qquad \text{whenever } 0 < u \leq v.$$

In the terminology of the cited reference, this last condition is $(\mathrm{aInc})_1$; ordinary monotonicity is the special case $A = 1$.

We now verify this weaker condition for $\Psi_{\beta,c}$ and invoke the equivalence theorem to restore the triangle inequality up to a comparison constant depending only on β and c.

Lemma A.17. *Let $\beta > 1$ and $c > 0$. There exist a Young function $\Phi_{\beta,c}$ and a finite constant $g_{\beta,c} \geq 1$, depending only on β, c, such that $\Phi_{\beta,c}(x) \leq \Psi_{\beta,c}(x) \leq \Phi_{\beta,c}(g_{\beta,c}x)$ for all $x \geq 0$. Consequently, for every random variable X, $\|X\|_{\Phi_{\beta,c}} \leq \|X\|_{\Psi_{\beta,c}} \leq g_{\beta,c}\|X\|_{\Phi_{\beta,c}}$, and for every $n \geq 1$ and random variables $X_1, \ldots, X_n$,*

$$\left\|\sum_{i\in[n]} X_i\right\|_{\Psi_{\beta,c}} \leq g_{\beta,c} \sum_{i\in[n]} \|X_i\|_{\Psi_{\beta,c}}. \tag{12}$$

If $\Psi_{\beta,c}$ is convex, one may take $\Phi_{\beta,c} = \Psi_{\beta,c}$ and $g_{\beta,c} = 1$.

Proof. Write $\Psi = \Psi_{\beta,c}$. By Harjulehto & Hästö (2019, Theorem 2.2.3), every weak Φ-function is equivalent to a strong Φ-function. Thus it suffices first to verify that Ψ is a weak Φ-function in the sense of Definition A.16. As observed above, Ψ is increasing and continuous, $\Psi(0) = 0$, $\Psi(x) \to 0$ as $x \downarrow 0$, and $\Psi(x) \to \infty$ as $x \to \infty$. It remains to check that $m(x) := \Psi(x)/x$ is almost increasing. Direct calculation gives

$$\frac{x\Psi'(x)}{\Psi(x)} = \frac{x}{1+x}\,\frac{c\beta\log^{\beta-1}(1+x)}{1 - \exp(-c\log^\beta(1+x))} \longrightarrow \begin{cases}\beta & \text{as } x \downarrow 0,\\ \infty & \text{as } x \to \infty.\end{cases}$$

Indeed,

$$m'(x) = \frac{\Psi(x)}{x^2}\left(\frac{x\Psi'(x)}{\Psi(x)} - 1\right).$$

Since $\beta > 1$, the displayed limits show that $m'(x) > 0$ near zero and for all sufficiently large x. Thus m is increasing on both tails. On the compact interval between these tails, m is positive and continuous, so its maximum-to-minimum ratio is finite. Combining the three regions gives a constant $A \geq 1$ such that $m(u) \leq Am(v)$ whenever $0 < u \leq v$. Hence m is almost increasing, and Ψ is a weak Φ-function.

The cited theorem therefore provides a strong Φ-function Φ_0 (in particular, a Young function in the sense of Definition A.10) and $L \geq 1$ such that $\Phi_0(x/L) \leq \Psi(x) \leq \Phi_0(Lx)$. Taking $\Phi_{\beta,c}(x) := \Phi_0(x/L)$ and $g_{\beta,c} := L^2$ gives the claimed comparison. The definition of the Luxemburg functional then gives the functional bounds, and

$$\left\|\sum_{i\in[n]} X_i\right\|_{\Psi_{\beta,c}} \leq g_{\beta,c}\left\|\sum_{i\in[n]} X_i\right\|_{\Phi_{\beta,c}} \leq g_{\beta,c}\sum_{i\in[n]}\|X_i\|_{\Phi_{\beta,c}} \leq g_{\beta,c}\sum_{i\in[n]}\|X_i\|_{\Psi_{\beta,c}}.$$

□

Having recovered control of sums, we record the elementary one-variable tail bound. This estimate uses only the monotonicity of $\Psi_{\beta,c}$ and therefore remains valid even when $\Psi_{\beta,c}$ is nonconvex.

Proposition A.18. *Let $\beta > 1$ and $c > 0$. For a random variable X, we have*

$$\mathbb{P}(|X| > t) \le 2\overline{\Psi}_{\beta,c}\left(\frac{t}{\|X\|_{\Psi_{\beta,c}}}\right) = 2\exp\left(-c\log^{\beta}\left(\frac{t}{\|X\|_{\Psi_{\beta,c}}} + 1\right)\right).$$

Proof. Write $\kappa := \|X\|_{\Psi_{\beta,c}}$. $\mathbb{E}[\Psi_{\beta,c}(|X|/\kappa) + 1] \le 2$, and $\Psi_{\beta,c}$ is increasing on $[0,\infty)$, so Markov's inequality gives

$$\begin{aligned}\mathbb{P}(|X| > t) &= \mathbb{P}\left(\Psi_{\beta,c}\left(\frac{|X|}{\kappa}\right) + 1 > \Psi_{\beta,c}\left(\frac{t}{\kappa}\right) + 1\right)\\ &\le \frac{2}{\Psi_{\beta,c}(t/\kappa) + 1} = 2\overline{\Psi}_{\beta,c}\left(\frac{t}{\kappa}\right).\end{aligned}$$

□

Combining this tail estimate with the quasi-triangle inequality from Lemma A.17 gives the desired bound for sums. Crucially, the comparison constant below depends only on β and c, not on the number or distributions of the summands.

Corollary A.19. *Let $\beta > 1$, $c > 0$, and let $X_1, \dots, X_n$ be random variables with $\|X_i\|_{\Psi_{\beta,c}} \le K_i$ for constants $K_i > 0$. Then for all $x \ge 0$,*

$$\mathbb{P}\left(\left|\sum_{i\in[n]} X_i\right| \ge x\right) \le 2\overline{\Psi}_{\beta,c}\left(\frac{x}{g_{\beta,c}\sum_{i\in[n]} K_i}\right). \tag{13}$$

Proof. By (12), $\|\sum_{i\in[n]} X_i\|_{\Psi_{\beta,c}} \le g_{\beta,c}\sum_{i\in[n]} K_i$. Markov's inequality applied to $1 + \Psi_{\beta,c}(|\sum_{i\in[n]} X_i|/(g_{\beta,c}\sum_{i\in[n]} K_i))$ gives the claim. □

We fix $\Phi_{\beta,c}$ and $g_{\beta,c}$ as in Lemma A.17, and write $\Phi_\beta := \Phi_{\beta,2^{-\beta}}$ and $g_\beta := g_{\beta,2^{-\beta}}$. In particular, g_β is finite for every $\beta > 1$ and depends only on β.

We record the form in which Corollary A.19 is used in the proof of Theorem 3.1. Solving $2\overline{\Psi}_{\beta,c}(x/(g_{\beta,c}\sum_{i\in[n]} K_i)) = \delta$ for x gives, with probability at least $1-\delta$,

$$\left|\sum_{i\in[n]} X_i\right| \le g_{\beta,c}\left(\exp\left(\left(\tfrac{1}{c}\log(2/\delta)\right)^{1/\beta}\right) - 1\right)\sum_{i\in[n]} K_i. \tag{14}$$

In particular, for $c = 2^{-\beta}$ the exponent becomes $(2^\beta\log(2/\delta))^{1/\beta} = 2\log^{1/\beta}(2/\delta)$, so the prefactor is $g_\beta(\exp(2\log^{1/\beta}(2/\delta)) - 1)$. This factor 2, inherited from the weakening of the Young function from Ψ_β to $\Psi_{\beta,2^{-\beta}}$ when squaring in Proposition A.14, is the source of the constant 2 in the exponent of Theorem 3.1. The constant g_β depends only on β.

A.3 Inequalities for Sequences of β-Heavy-Tailed Random Variables

We next give a concentration inequality for β-heavy-tailed martingale difference sequences in $\mathbb{R}^d$, which is the form in which the gradient noise enters. It shows that $\mathbb{P}(\max_{n\in[N]} \|\sum_{i\in[n]} X_i\| \ge z)$ can be bounded by a sub-Gaussian term and a β-heavy tail term with an arbitrary threshold M.

Proposition A.20 (β-heavy tail MDS concentration inequality)**.** *Let $\beta > 1$ and let $\{X_i\}_{i\in[N]}$ be an $\mathbb{R}^d$-valued $(\mathcal{F}_i)$-MDS. Suppose $\sigma_i := \|\,\|X_i\|\,\|_{\Psi_\beta} < \infty$ for all $i \in [N]$, and suppose $M > 0$ satisfies*

$$\beta\log^{\beta-1}(M+1) - 2 > 0. \tag{15}$$

Then for any $z > 0$,

$$\mathbb{P}\left(\max_{n\in[N]}\left\|\sum_{i\in[n]} X_i\right\| \geq z\right) \leq 2\exp\left(-\frac{z^2}{32M^2\sum_{i\in[N]}\sigma_i^2}\right) + \left(2M^2 + \frac{4(M+1)^2}{\beta\log^{\beta-1}(M+1)-2}\right)\frac{4\sum_{i\in[N]}\sigma_i^2}{z^2}\overline{\Psi}_\beta(M). \quad (16)$$

The following corollary, obtained from Proposition A.20, gives a high-probability bound for the β-heavy-tailed MDS only dominated by $\overline{\Psi}_\beta$, in a form more convenient for use.

Corollary A.21. *Let $\beta > 1$, let $\{X_i\}_{i\in[N]}$ be an $\mathbb{R}^d$-valued MDS with $\sigma_i = \|\,\|X_i\|\,\|_{\Psi_\beta} < \infty$, and set*

$$m_\beta := \exp\left(\left(\tfrac{3}{\beta}\right)^{\frac{1}{\beta-1}}\right) - 1, \quad D_\beta := 2 + 4\left(1 + \tfrac{1}{m_\beta}\right)^2, \quad C_\beta := \max\{1, m_\beta\}\max\left\{8,\ 2\sqrt{D_\beta/\log 2}\right\}. \quad (17)$$

Then for any $\delta \in (0, 1]$, with probability at least $1 - \delta$,

$$\max_{n\in[N]}\left\|\sum_{i\in[n]} X_i\right\| \leq C_\beta\sqrt{\log(2/\delta)\sum_{i\in[N]}\sigma_i^2}\cdot\Psi_\beta^{-1}(2/\delta). \quad (18)$$

The threshold M of Proposition A.20 is placed at the quantile of the β-heavy tail at level δ, and is raised to m_β whenever that quantile is too small for condition (15); the level m_β is the one at which $\beta\log^{\beta-1}(m_\beta + 1) = 3$. The constant C_β is finite for every $\beta > 1$ and depends only on β, through m_β.

We first establish the following technical integral estimate.

Lemma A.22. *Let $\beta > 1$ and suppose $M > 0$ satisfies* (15)*. Then*

$$\int_M^\infty t\,\overline{\Psi}_\beta(t)\,dt \leq \frac{(M+1)^2\,\overline{\Psi}_\beta(M)}{\beta\log^{\beta-1}(M+1)-2}. \quad (19)$$

Proof. Substituting $t = \exp(s) - 1$, we have $t < \exp(s)$ and $\overline{\Psi}_\beta(t) = \exp(-s^\beta)$, so

$$\int_M^\infty t\,\overline{\Psi}_\beta(t)\,dt \leq \int_{\log(M+1)}^\infty \exp(2s - s^\beta)\,ds.$$

We bound the right-hand side using Laplace's method. Let $\phi(s) := s^\beta - 2s$. Since $\beta > 1$, $\phi'(s) = \beta s^{\beta-1} - 2$ is monotonically increasing in s, and by (15), $\phi'(\log(M+1)) > 0$. Hence for $s \geq \log(M+1)$, $\phi'(s) \geq \phi'(\log(M+1)) > 0$ and $\lim_{s\to\infty}\phi(s) = \infty$, so

$$\int_{\log(M+1)}^\infty \exp(-\phi(s))\,ds = \int_{\log(M+1)}^\infty \frac{\phi'(s)\exp(-\phi(s))}{\phi'(s)}\,ds \leq \frac{1}{\phi'(\log(M+1))}\int_{\log(M+1)}^\infty \phi'(s)\exp(-\phi(s))\,ds = \frac{\exp(-\phi(\log(M+1)))}{\phi'(\log(M+1))}.$$

Finally, since $\exp(-\phi(\log(M+1))) = \exp(2\log(M+1))\exp(-\log^\beta(M+1)) = (M+1)^2\overline{\Psi}_\beta(M)$ and $\phi'(\log(M+1)) = \beta\log^{\beta-1}(M+1) - 2$, the claim follows. □

We now prove Proposition A.20.

Proof of Proposition A.20. Fix $M > 0$ satisfying (15), and for each $i \in [N]$ decompose

$$X_i' := X_i\mathbb{1}_{\|X_i\|\leq M\sigma_i} - \mathbb{E}[X_i\mathbb{1}_{\|X_i\|\leq M\sigma_i} \mid \mathcal{F}_{i-1}],$$
$$X_i'' := X_i\mathbb{1}_{\|X_i\|> M\sigma_i} - \mathbb{E}[X_i\mathbb{1}_{\|X_i\|> M\sigma_i} \mid \mathcal{F}_{i-1}]. \quad (20)$$

Both are $\mathcal{F}_i$-measurable with conditional mean zero, so $\{X_i'\}$ and $\{X_i''\}$ are both $\mathbb{R}^d$-valued $(\mathcal{F}_i)$-MDS. Since $X_i = X_i' + X_i''$, the triangle inequality gives

$$\mathbb{P}\left(\max_{n\in[N]}\left\|\sum_{i\in[n]} X_i\right\| \geq 2z\right) \leq \mathbb{P}\left(\max_{n\in[N]}\left\|\sum_{i\in[n]} X_i'\right\| \geq z\right) + \mathbb{P}\left(\max_{n\in[N]}\left\|\sum_{i\in[n]} X_i''\right\| \geq z\right), \quad (21)$$

so it suffices to bound each term on the right-hand side.

For the first term, (20) gives $\|X_i'\| \leq 2M\sigma_i$ almost surely, so Lemma A.4 with $b_i = 2M\sigma_i$ yields

$$\mathbb{P}\left(\max_{n\in[N]}\left\|\sum_{i\in[n]} X_i'\right\| \geq z\right) \leq 2\exp\left(-\frac{z^2}{8M^2\sum_{i\in[N]}\sigma_i^2}\right). \quad (22)$$

For the second term, $\{X_i''\}$ is an MDS, so $S_n'' := \sum_{i\in[n]} X_i''$ is a martingale with $S_0'' = 0$, and Lemma A.2 together with the orthogonality of an MDS gives

$$\mathbb{P}\left(\max_{n\in[N]}\|S_n''\| \geq z\right) \leq \frac{1}{z^2}\mathbb{E}\left[\|S_N''\|^2\right] = \frac{1}{z^2}\sum_{i\in[N]}\mathbb{E}\left[\|X_i''\|^2\right]. \quad (23)$$

We now bound $\mathbb{E}[\|X_i''\|^2]$. In general, for an $\mathbb{R}^d$-valued random variable W and a sub-σ-algebra $\mathcal{G}$ we have $\mathbb{E}[\|W - \mathbb{E}[W \mid \mathcal{G}]\|^2] = \mathbb{E}[\|W\|^2] - \mathbb{E}[\|\mathbb{E}[W \mid \mathcal{G}]\|^2] \leq \mathbb{E}[\|W\|^2]$, so by Lemma A.1 applied to the nonnegative random variable $\|X_i\|\mathbb{1}_{\|X_i\|>M\sigma_i}$,

$$\begin{aligned}
\mathbb{E}\left[\|X_i''\|^2\right] &\leq \mathbb{E}[\|X_i\|^2\mathbb{1}_{\|X_i\|>M\sigma_i}] = \int_0^\infty 2y\,\mathbb{P}(\|X_i\|\mathbb{1}_{\|X_i\|>M\sigma_i} > y)\,dy \\
&= M^2\sigma_i^2\mathbb{P}(\|X_i\| > M\sigma_i) + \int_{M\sigma_i}^\infty 2y\,\mathbb{P}(\|X_i\| > y)\,dy \\
&= M^2\sigma_i^2\mathbb{P}(\|X_i\| > M\sigma_i) + \sigma_i^2\int_M^\infty 2t\,\mathbb{P}(\|X_i\| > t\sigma_i)\,dt \quad (24)
\end{aligned}$$

(the last equality using the substitution $y = t\sigma_i$). Applying Proposition A.18 to both terms,

$$\mathbb{E}\left[\|X_i''\|^2\right] \leq \sigma_i^2\left(2M^2\overline{\Psi}_\beta(M) + 4\int_M^\infty t\overline{\Psi}_\beta(t)\,dt\right),$$

and by Lemma A.22,

$$\mathbb{E}\left[\|X_i''\|^2\right] \leq \left(2M^2 + \frac{4(M+1)^2}{\beta\log^{\beta-1}(M+1) - 2}\right)\sigma_i^2\,\overline{\Psi}_\beta(M). \quad (25)$$

Substituting (22), (23), (25) into (21) yields the claim. □

Proof of Corollary A.21. Write $V := \sum_{i\in[N]}\sigma_i^2$, fix $\delta \in (0,1]$, and set

$$M_\delta := \exp\Big(\log^{1/\beta}(2/\delta)\Big) - 1, \qquad \widetilde{M}_\delta := \max\{M_\delta,\, m_\beta\},$$

so that $\overline{\Psi}_\beta(M_\delta) = \delta/2$. We check the hypotheses of Proposition A.20 at $M = \widetilde{M}_\delta$. By (17) we have $\log(m_\beta + 1) = (3/\beta)^{1/(\beta-1)}$ and hence $\beta\log^{\beta-1}(m_\beta + 1) = 3$, so the map $M \mapsto \beta\log^{\beta-1}(M+1) - 2$, which is increasing, is at least 1 on $[m_\beta, \infty)$, and $\widetilde{M}_\delta$ satisfies (15). The same monotonicity, together with the fact that $(1 + 1/M)^2$ is decreasing, shows that

$$\frac{1}{M^2}\left(2M^2 + \frac{4(M+1)^2}{\beta\log^{\beta-1}(M+1) - 2}\right) = 2 + \frac{4\,(1+1/M)^2}{\beta\log^{\beta-1}(M+1) - 2}$$

is nonincreasing on $[m_\beta, \infty)$, where it takes the value D_β at $M = m_\beta$. The constant in the second term of (16) is therefore at most $D_\beta\widetilde{M}_\delta^2$.

We now apply (16) with $M = \widetilde{M}_\delta$ and $z := c\,\widetilde{M}_\delta\sqrt{V\log(2/\delta)}$, where $c := \max\{8, 2\sqrt{D_\beta/\log 2}\}$. Since $c^2 \geq 64$ and $\delta/2 \leq 1/2$, the first term is

$$2\exp\left(-\frac{z^2}{32\widetilde{M}_\delta^2 V}\right) = 2\left(\frac{\delta}{2}\right)^{c^2/32} \leq 2\left(\frac{\delta}{2}\right)^2 \leq \frac{\delta}{2},$$

while the second term, using $\widetilde{M}_\delta \geq M_\delta$ and the monotonicity of $\overline{\Psi}_\beta$, is at most

$$D_\beta \widetilde{M}_\delta^2 \cdot \frac{4V}{z^2}\,\overline{\Psi}_\beta(M_\delta) = \frac{4D_\beta}{c^2\log(2/\delta)} \cdot \frac{\delta}{2} \leq \frac{2D_\beta\,\delta}{c^2\log 2} \leq \frac{\delta}{2},$$

so that $\mathbb{P}(\max_{n\in[N]} \|\sum_{i\in[n]} X_i\| \geq z) \leq \delta$.

It remains to rewrite z. Since $\beta > 1$ and $\delta \leq 1$ we have $\log^{1/\beta}(2/\delta) \geq (\log 2)^{1/\beta} > \log 2$, whence $M_\delta > 1$ and $\widetilde{M}_\delta \leq \max\{1, m_\beta\}M_\delta$. Since $\overline{\Psi}_\beta = 1/(\Psi_\beta + 1)$, we have $\Psi_\beta(M_\delta) = 2/\delta - 1$, so monotonicity of Ψ_β gives $M_\delta \leq \Psi_\beta^{-1}(2/\delta)$, and altogether

$$z \leq C_\beta\sqrt{V\log(2/\delta)}\;\Psi_\beta^{-1}(2/\delta),$$

which is the bound (18). □

Proposition A.20 and Corollary A.21 require deterministic Orlicz norm bounds σ_i, which is too restrictive for SGD, because the increment $\xi_t = -2\eta_t\langle\epsilon_t, \nabla\mathcal{L}_S(\boldsymbol{w}_t)\rangle$ has conditional Orlicz norm bounded by the trajectory-dependent quantity $2\eta_t\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|K$, even though the bound K in Assumption 2.2 is deterministic. We therefore need a Freedman-type inequality (Freedman, 1975) that allows predictable bounds K_{i-1} on the conditional Orlicz norms, subject to deterministic upper bounds, and controls the tail of the martingale sum through the accumulated variance proxy $V_k = \sum_{i\in[k]} a_\beta K_{i-1}^2$. Part (i) of Proposition A.23 moreover permits the self-bounding condition $V_k \leq \alpha S_k + \gamma$, the form used to prove Theorem 3.1. Throughout, the Orlicz condition (A3) below is read with the conventions $u/0 := +\infty$ for $u > 0$ and $0/0 := 0$, so that on the event $\{K_{i-1} = 0\}$ it forces $\xi_i = 0$ almost surely, since any other value makes the left-hand side infinite. The bound K_{i-1} is thus allowed to vanish, in which case the corresponding increment vanishes as well.

Proposition A.23 (β-heavy tail Freedman inequality)**.** *Let $(\Omega, \mathcal{F}, (\mathcal{F}_i)_{i\geq 0}, \mathbb{P})$ be a filtered probability space and $N \in \mathbb{N}$, and suppose the $(\mathcal{F}_i)$-adapted real random variable sequences $(\xi_i)_{i\in[N]}$, $(K_i)_{i\geq 0}$ satisfy the following almost surely for each $i \in [N]$:*

(A1) $0 \leq K_{i-1} \leq m_i$ (m_i *a deterministic constant*), $\overline{m} := \max_{i\in[N]} m_i$;

(A2) $\mathbb{E}[\xi_i \mid \mathcal{F}_{i-1}] = 0$;

(A3) $\mathbb{E}\left[\Psi_\beta\left(\frac{|\xi_i|}{K_{i-1}}\right) \mid \mathcal{F}_{i-1}\right] \leq 1.$

Fix $\delta \in (0,1)$ and set $\Lambda := \log(N/\delta)$. For notational simplicity we define

$$\begin{aligned} J_\beta &:= \int_0^\infty \exp(2s - s^\beta/2)\,ds, \qquad a_\beta := 8\exp(\beta^\beta/2)J_\beta, \\ M &:= \overline{m}(\exp(\Lambda^{1/\beta}) - 1), \qquad b_F := \frac{2(\exp(\Lambda^{1/\beta}) - 1)}{\Lambda}. \end{aligned} \tag{26}$$

Let $S_k := \sum_{i\in[k]} \xi_i$ and define the accumulated variance proxy $V_k := \sum_{i\in[k]} a_\beta K_{i-1}^2$. Then the following hold.

(i) For any $x, \gamma \geq 0$, $\alpha \geq b_F\overline{m}$, $\lambda \in [0, 1/(2\alpha)]$,

$$\mathbb{P}\left(\bigcup_{k\in[N]} \{S_k \geq x \text{ and } V_k \leq \alpha S_k + \gamma\}\right) \leq \exp(-\lambda x + 2\lambda^2\gamma) + 2\delta. \tag{27}$$

(ii) For any $x, \gamma \geq 0$, $\lambda \in [0, 1/(b_F\overline{m})]$,

$$\mathbb{P}\left(\bigcup_{k\in[N]} \{S_k \geq x \text{ and } V_k \leq \gamma\}\right) \leq \exp(-\lambda x + (\lambda^2/2)\gamma) + 2\delta. \tag{28}$$

The event $\{S_k \geq x \text{ and } V_k \leq \alpha S_k + \gamma\}$ can be interpreted as the event that the MDS sum S_k exceeds x while its accumulated variance proxy V_k remains controlled.

We first prove several lemmas needed for the proof of Proposition A.23.

Lemma A.24 (Conditional tail estimate of the variance)**.** *For each $i \in [N]$ and any $t \geq 0$,*

$$\mathbb{P}(|\xi_i| > t \mid \mathcal{F}_{i-1}) \leq 2\overline{\Psi}_\beta\left(\frac{t}{K_{i-1}}\right), \tag{29}$$

$$\mathbb{E}[\xi_i^2 \mid \mathcal{F}_{i-1}] \leq 4J_\beta K_{i-1}^2. \tag{30}$$

Proof. If $K_{i-1} = 0$, then $\xi_i = 0$ almost surely by (A3), and both claims are immediate, because the right-hand side of (29) vanishes for $t > 0$ and equals 2 for $t = 0$, while both sides of (30) vanish. We may therefore assume $K_{i-1} > 0$ in what follows.

The bound (29) follows in the same way as Proposition A.18. For (30), note first that $J_\beta < \infty$: since $\beta > 1$, we have $s^\beta/2 - 2s \to \infty$ as $s \to \infty$, so the integrand of J_β is continuous on $[0, \infty)$ and eventually dominated by e^{-s}. Now (30) follows from Lemma A.1 and (29), substituting $t = K_{i-1}u$ and then $1 + u = e^s$:

$$\begin{aligned}\mathbb{E}[\xi_i^2 \mid \mathcal{F}_{i-1}] = \int_0^\infty 2t\,\mathbb{P}(|\xi_i| > t \mid \mathcal{F}_{i-1})\,dt &\leq 4K_{i-1}^2 \int_0^\infty u\,\overline{\Psi}_\beta(u)\,du \\ &= 4K_{i-1}^2 \int_0^\infty (e^s - 1)e^s \exp(-s^\beta)\,ds \leq 4J_\beta K_{i-1}^2,\end{aligned}$$

where the last inequality uses $(e^s - 1)e^s \leq e^{2s}$ and $s^\beta/2 \leq s^\beta$ for $s \geq 0$. □

Since a truncation technique is used later in the proof, we give a moment generating function (MGF) estimate for a random variable that is bounded above.

Lemma A.25 (MGF estimate for an upper-bounded random variable)**.** *Fix $i \in [N]$. For $M > 0$ and*

$$0 \leq \lambda \leq \frac{1}{2M}\log^\beta\left(1 + \frac{M}{K_{i-1}}\right), \tag{31}$$

we have

$$\mathbb{E}[\exp(\lambda\xi_i)\mathbb{1}_{\xi_i \leq M} \mid \mathcal{F}_{i-1}] \leq \exp\left(\frac{\lambda^2 a_\beta K_{i-1}^2}{2}\right). \tag{32}$$

Proof. If $K_{i-1} = 0$, then $\xi_i = 0$ almost surely by (A3), the constraint (31) is vacuous since its right-hand side is infinite, and (32) reduces to the identity $\mathbb{E}[\mathbb{1}_{0 \leq M} \mid \mathcal{F}_{i-1}] = 1 = \exp(0)$. We may therefore assume $K_{i-1} > 0$, which in particular justifies the substitution $t = K_{i-1}u$ used below.

For brevity write $\mathbb{E}_{i-1}[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{i-1}]$. Let $\phi(u) := e^u - 1 - u$, so that $e^{\lambda z} = 1 + \lambda z + \phi(\lambda z)$. Since $\mathbb{E}_{i-1}[\xi_i] = 0$ and $M > 0$,

$$\mathbb{E}_{i-1}[\xi_i \mathbb{1}_{\xi_i \leq M}] = -\mathbb{E}_{i-1}[\xi_i \mathbb{1}_{\xi_i > M}] \leq 0,$$

and hence

$$\begin{aligned}\mathbb{E}_{i-1}[\exp(\lambda\xi_i)\mathbb{1}_{\xi_i \leq M}] &\leq \mathbb{E}_{i-1}[(1 + \lambda\xi_i + \phi(\lambda\xi_i))\mathbb{1}_{\xi_i \leq M}] \\ &\leq 1 + \underbrace{\mathbb{E}_{i-1}[\phi(\lambda\xi_i)\mathbb{1}_{\xi_i \leq 0}]}_{\text{(A)}} + \underbrace{\mathbb{E}_{i-1}[\phi(\lambda\xi_i)\mathbb{1}_{0 < \xi_i \leq M}]}_{\text{(B)}}\end{aligned} \tag{33}$$

where in the last step we split ξ_i at 0 as well as at M. We bound each of (A) and (B).

(A): Since $\phi(u) \leq u^2/2$ for $u \leq 0$, together with (30),

$$\mathbb{E}_{i-1}[\phi(\lambda\xi_i)\mathbb{1}_{\xi_i \leq 0}] \leq \frac{\lambda^2}{2}\mathbb{E}_{i-1}[\xi_i^2] \leq 2\lambda^2 J_\beta K_{i-1}^2. \tag{34}$$

(B): Since $\phi(\lambda z) = \int_0^z \lambda(e^{\lambda t} - 1)\,dt$, $\phi(\lambda\xi_i)\mathbb{1}_{0<\xi_i\le M} = \int_0^M \lambda(e^{\lambda t} - 1)\mathbb{1}_{t<\xi_i\le M}\,dt$, and $\mathbb{1}_{t<\xi_i\le M} \le \mathbb{1}_{|\xi_i|>t}$. Using Fubini's theorem, (29), and $e^v - 1 \le ve^v$ ($v \ge 0$) in turn, and substituting $t = K_{i-1}u$,

$$\mathbb{E}_{i-1}[\phi(\lambda\xi_i)\mathbb{1}_{0<\xi_i\le M}] \le 2\lambda^2 K_{i-1}^2 \int_0^{M/K_{i-1}} u\, e^{\lambda K_{i-1}u}\, \overline{\Psi}_\beta(u)\,du.$$

Since λ satisfies (31), $\lambda K_{i-1} \le \log^\beta(1 + M/K_{i-1})/(2M/K_{i-1})$, so by Lemma A.9, for $u \in [0, M/K_{i-1}]$, $\lambda K_{i-1}u \le (\log^\beta(1+u) + \beta^\beta)/2$, and hence

$$\begin{aligned}\int_0^{M/K_{i-1}} u\, e^{\lambda K_{i-1}u}\, \overline{\Psi}_\beta(u)\,du &\le \exp(\beta^\beta/2)\int_0^\infty u\,\exp\Big(-\tfrac{1}{2}\log^\beta(1+u)\Big)\,du \\ &= \exp(\beta^\beta/2)\int_0^\infty (e^s - 1)e^s\exp(-s^\beta/2)\,ds \le \exp(\beta^\beta/2)J_\beta,\end{aligned}$$

i.e. $\mathbb{E}_{i-1}[\phi(\lambda\xi_i)\mathbb{1}_{0<\xi_i\le M}] \le 2\lambda^2\exp(\beta^\beta/2)J_\beta K_{i-1}^2$. Substituting these bounds, together with $1 + \exp(\beta^\beta/2) \le 2\exp(\beta^\beta/2)$ and $1 + v \le e^v$, into (33),

$$\mathbb{E}_{i-1}[e^{\lambda\xi_i}\mathbb{1}_{\xi_i\le M}] \le 1 + 4\exp(\beta^\beta/2)J_\beta\lambda^2K_{i-1}^2 \le \exp\left(\frac{\lambda^2 a_\beta K_{i-1}^2}{2}\right).$$

□

Having established the necessary lemmas, we now prove Proposition A.23.

Proof of Proposition A.23. First, noting (29), $K_{i-1} \le \overline{m}$, monotonicity, and the definition $\log^\beta(1 + M/\overline{m}) = \Lambda$,

$$\forall i \in [N], \quad \mathbb{P}(\xi_i > M) \le 2e^{-\Lambda} = \frac{2\delta}{N} \implies \mathbb{P}\left(\max_{i\in[N]}\xi_i > M\right) \le 2\delta. \tag{35}$$

Next fix $\tilde{\lambda} \in [0, 1/(b_F\overline{m})]$ and define

$$Z_0 := 1, \qquad Z_k := \exp\left(\tilde{\lambda}S_k - \frac{\tilde{\lambda}^2 V_k}{2}\right)\prod_{i\in[k]}\mathbb{1}_{\xi_i\le M} \quad \text{for each} \quad k \in [N].$$

We show Z_k is a nonnegative supermartingale. First, from $K_{k-1} \le \overline{m}$ and

$$\frac{1}{b_F\overline{m}} = \frac{\Lambda}{2M} = \frac{1}{2M}\log^\beta(1 + M/\overline{m}),$$

$$\tilde{\lambda} \le \frac{1}{b_F\overline{m}} = \frac{1}{2M}\log^\beta\left(1 + \frac{M}{\overline{m}}\right) \le \frac{1}{2M}\log^\beta\left(1 + \frac{M}{K_{k-1}}\right),$$

where the last inequality also covers $K_{k-1} = 0$, its right-hand side being infinite under the convention stated before Proposition A.23.

Next consider the conditional expectation of Z_k given $\mathcal{F}_{k-1}$. Applying Lemma A.25 with $i = k$ and $\lambda = \tilde{\lambda}$,

$$\mathbb{E}[Z_k \mid \mathcal{F}_{k-1}] = Z_{k-1}\exp\left(-\frac{\tilde{\lambda}^2 a_\beta K_{k-1}^2}{2}\right)\mathbb{E}[e^{\tilde{\lambda}\xi_k}\mathbb{1}_{\xi_k\le M} \mid \mathcal{F}_{k-1}] \le Z_{k-1},$$

so $\{Z_k\}$ is a nonnegative supermartingale with $Z_0 = 1$. For (28) we take $\tilde{\lambda} = \lambda$ with $\lambda \in [0, 1/(b_F\overline{m})]$ as in the statement. Define the event $T_k := \{\max_{i\le k}\xi_i \le M\}$. Setting $A_k := \{S_k \ge x \text{ and } V_k \le \gamma\}$,

$$A_k \cap T_k \implies Z_k = \exp\left(\lambda S_k - \frac{\lambda^2}{2}V_k\right) \ge \exp\left(\lambda x - \frac{\lambda^2}{2}\gamma\right).$$

Hence

$$\bigcup_{k\in[N]} A_k \subseteq \left\{\max_{k\in[N]} Z_k \geq \exp\left(\lambda x - \frac{\lambda^2}{2}\gamma\right)\right\} \cup \left\{\max_{i\in[N]} \xi_i > M\right\} \tag{36}$$

and, by Lemma A.3 and (35),

$$\mathbb{P}\left(\bigcup_{k\in[N]} A_k\right) \leq \exp\left(-\lambda x + \frac{\lambda^2}{2}\gamma\right) + 2\delta,$$

which establishes (28).

Finally we show (27). Let $\alpha \geq b_F\overline{m}$, $\lambda \in [0, 1/(2\alpha)]$, and set $\tilde{\lambda} := 2\lambda$, so that $\tilde{\lambda} \leq 1/\alpha \leq 1/(b_F\overline{m})$, so the argument above applies. Set $A_k := \{S_k \geq x \text{ and } V_k \leq \alpha S_k + \gamma\}$. On $A_k \cap T_k$, we have

$$\tilde{\lambda} S_k - \frac{\tilde{\lambda}^2}{2} V_k \geq 2\lambda(1-\lambda\alpha) S_k - 2\lambda^2\gamma \geq \lambda x - 2\lambda^2\gamma,$$

using $\lambda\alpha \leq 1/2$, hence $2\lambda(1-\lambda\alpha) \geq \lambda \geq 0$, together with $S_k \geq x \geq 0$. Hence on $A_k \cap T_k$, $Z_k \geq \exp(\lambda x - 2\lambda^2\gamma)$, and arguing as in (36),

$$\mathbb{P}\left(\bigcup_{k\in[N]} A_k\right) \leq \exp(-\lambda x + 2\lambda^2\gamma) + 2\delta,$$

which establishes (27). □

A.4 Proofs of the Main Results

Proof of Theorem 3.1. Throughout the proof we write $\epsilon_t := \nabla\ell(\boldsymbol{w}_t; z_{j_t}) - \nabla\mathcal{L}_S(\boldsymbol{w}_t)$ for the gradient noise at step t, as in Assumption 2.2. Since $\mathcal{L}_S$ is an average of the functions $\ell(\cdot\,; z_i)$, each of which is b-smooth by Assumption 2.4, $\mathcal{L}_S$ is itself b-smooth, so Lemma A.5 applies to $\mathcal{L}_S$.

We use Lemma A.5 to bound $\mathcal{L}_S(\boldsymbol{w}_{t+1}) - \mathcal{L}_S(\boldsymbol{w}_t)$:

$$\begin{aligned}
\mathcal{L}_S(\boldsymbol{w}_{t+1}) - \mathcal{L}_S(\boldsymbol{w}_t) &\leq \langle \boldsymbol{w}_{t+1} - \boldsymbol{w}_t, \nabla\mathcal{L}_S(\boldsymbol{w}_t)\rangle + \tfrac{1}{2}b\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2 \\
&= -\eta_t\langle \epsilon_t, \nabla\mathcal{L}_S(\boldsymbol{w}_t)\rangle - \eta_t\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|^2 + \tfrac{1}{2}b\eta_t^2\|\nabla\ell(\boldsymbol{w}_t; z_{j_t})\|^2 \\
&\leq -\eta_t\langle \epsilon_t, \nabla\mathcal{L}_S(\boldsymbol{w}_t)\rangle - (\eta_t - b\eta_t^2)\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|^2 + b\eta_t^2\|\epsilon_t\|^2 \\
&\leq -\eta_t\langle \epsilon_t, \nabla\mathcal{L}_S(\boldsymbol{w}_t)\rangle - \tfrac{1}{2}\eta_t\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|^2 + b\eta_t^2\|\epsilon_t\|^2,
\end{aligned} \tag{37}$$

where the third line uses $\|\nabla\ell(\boldsymbol{w}_t; z_{j_t})\|^2 = \|\epsilon_t + \nabla\mathcal{L}_S(\boldsymbol{w}_t)\|^2 \leq 2\|\epsilon_t\|^2 + 2\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|^2$, and the last line uses $b\eta_t^2 \leq \frac{1}{2}\eta_t$, which follows from the assumption $\eta_t \leq 1/(2b)$.

Summing (37) from $t = 1$ to T,

$$\mathcal{L}_S(\boldsymbol{w}_{T+1}) - \mathcal{L}_S(\boldsymbol{w}_1) \leq -\sum_{t\in[T]} \eta_t\langle \epsilon_t, \nabla\mathcal{L}_S(\boldsymbol{w}_t)\rangle - \tfrac{1}{2}\sum_{t\in[T]} \eta_t\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|^2 + \sum_{t\in[T]} b\eta_t^2\|\epsilon_t\|^2,$$

and since $\boldsymbol{w}(S)$ is a minimizer of $\mathcal{L}_S$ over $\mathcal{W}$ we have $\mathcal{L}_S(\boldsymbol{w}_{T+1}) \geq \mathcal{L}_S(\boldsymbol{w}(S))$, whence we obtain

$$\begin{aligned}
\sum_{t\in[T]} \eta_t\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|^2 \leq\, & 2\left(\mathcal{L}_S(\boldsymbol{w}_1) - \mathcal{L}_S(\boldsymbol{w}(S))\right) \\
& \underbrace{-\sum_{t\in[T]} 2\eta_t\langle \epsilon_t, \nabla\mathcal{L}_S(\boldsymbol{w}_t)\rangle}_{\text{(A)}} + \underbrace{\sum_{t\in[T]} 2b\eta_t^2\|\epsilon_t\|^2}_{\text{(B)}}.
\end{aligned} \tag{38}$$

We now bound the terms (A) and (B) in (38).

(A) With K denoting the uniform bound on the conditional Orlicz norms in Assumption 2.2, set

$$\xi_t := -2\eta_t\langle \epsilon_t, \nabla\mathcal{L}_S(\boldsymbol{w}_t)\rangle, \qquad K_{t-1} := 2\eta_t\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|K.$$

We verify the hypotheses of Proposition A.23 with $N = T$, so that it applies to $S_k = \sum_{t\in[k]} \xi_t$.

Adaptedness and (A1). The iterate $\boldsymbol{w}_t$ is $\mathcal{F}_{t-1}$-measurable, hence so is K_{t-1}, and Assumption 2.5 gives the deterministic bound

$$0 \le K_{t-1} = 2\eta_t \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\| K \le 2GK =: \overline{m}.$$

(A2). $\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = -2\eta_t \langle \mathbb{E}[\epsilon_t \mid \mathcal{F}_{t-1}], \nabla \mathcal{L}_S(\boldsymbol{w}_t)\rangle = 0$ by the mean-zero condition in Assumption 2.2, since $\eta_t \nabla \mathcal{L}_S(\boldsymbol{w}_t)$ is $\mathcal{F}_{t-1}$-measurable.

(A3). By the Cauchy–Schwarz inequality,

$$|\xi_t| = 2\eta_t \, |\langle \epsilon_t, \nabla \mathcal{L}_S(\boldsymbol{w}_t)\rangle| \le 2\eta_t \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\| \cdot \|\epsilon_t\|.$$

On the $\mathcal{F}_{t-1}$-measurable event $\{K_{t-1} > 0\}$, this gives $|\xi_t|/K_{t-1} \le \|\epsilon_t\|/K$. Thus monotonicity of Ψ_β and Assumption 2.2 yield, on this event,

$$\mathbb{E}\left[\Psi_\beta\left(\frac{|\xi_t|}{K_{t-1}}\right) \,\middle|\, \mathcal{F}_{t-1}\right] \le \mathbb{E}\left[\Psi_\beta\left(\frac{\|\epsilon_t\|}{K}\right) \,\middle|\, \mathcal{F}_{t-1}\right] \le 1,$$

which is precisely (A3). In the degenerate case $\nabla \mathcal{L}_S(\boldsymbol{w}_t) = 0$, for which $K_{t-1} = 0$, the displayed Cauchy–Schwarz bound already gives $\xi_t = 0$ almost surely, so (A3) again holds under the convention stated before Proposition A.23.

We may therefore apply Proposition A.23(i) with $\gamma = 0$, $\alpha \ge b_F \overline{m}$, $\lambda = 1/(2\alpha)$ and $x = 2\alpha \log(1/\delta)$, which yields

$$\mathbb{P}\left(\bigcup_{k\in[T]} \{S_k \ge x \text{ and } V_k \le \alpha S_k\}\right) \le \exp(-\lambda x) + 2\delta = \delta + 2\delta = 3\delta.$$

On the complement, every $k \in [T]$ satisfies $S_k < x$ or $S_k < V_k/\alpha$, hence $S_k \le \max(x, V_k/\alpha) \le x + V_k/\alpha$, where by definition

$$V_T = \sum_{t\in[T]} a_\beta K_{t-1}^2 = 4a_\beta K^2 \sum_{t\in[T]} \eta_t^2 \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2.$$

We conclude that with probability at least $1 - 3\delta$,

$$-\sum_{t\in[T]} 2\eta_t \langle \epsilon_t, \nabla \mathcal{L}_S(\boldsymbol{w}_t)\rangle \le 2\alpha \log(1/\delta) + \frac{4a_\beta K^2}{\alpha} \sum_{t\in[T]} \eta_t^2 \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2. \tag{39}$$

(B) Assumption 2.2, Proposition A.14 and the homogeneity of the Orlicz norm (Lemma A.12(i)) give

$$\left\| 2b\eta_t^2 \|\epsilon_t\|^2 \right\|_{\Psi_{\beta,2^{-\beta}}} \le 2b\eta_t^2 \, \| \, \|\epsilon_t\| \, \|_{\Psi_\beta}^2 \le 2b\eta_t^2 K^2.$$

Hence, by Corollary A.19 in the form (14) with $c = 2^{-\beta}$, with probability at least $1 - \delta$,

$$\sum_{t\in[T]} 2b\eta_t^2 \|\epsilon_t\|^2 \le 2g_\beta b K^2 \exp\Big(2\log^{1/\beta}(2/\delta)\Big) \sum_{t\in[T]} \eta_t^2. \tag{40}$$

Substituting (39) and (40) into (38), a union bound gives that with probability at least $1 - 4\delta$,

$$\begin{aligned}\sum_{t\in[T]} \eta_t \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2 \le{}& 2\left(\mathcal{L}_S(\boldsymbol{w}_1) - \mathcal{L}_S(\boldsymbol{w}(S))\right) \\ &+ 2\alpha \log(1/\delta) + \frac{4a_\beta K^2}{\alpha} \sum_{t\in[T]} \eta_t^2 \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2 \\ &+ 2g_\beta b K^2 \exp\Big(2\log^{1/\beta}(2/\delta)\Big) \sum_{t\in[T]} \eta_t^2.\end{aligned}$$

Rearranging in terms of $\sum_{t\in[T]} \eta_t \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2$,

$$\sum_{t\in[T]} \left(1 - \frac{4a_\beta K^2 \eta_t}{\alpha}\right) \eta_t \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2 \le 2\left(\mathcal{L}_S(\boldsymbol{w}_1) - \mathcal{L}_S(\boldsymbol{w}(S))\right) + 2\alpha \log(1/\delta) + 2g_\beta b K^2 \exp\Big(2\log^{1/\beta}(2/\delta)\Big) \sum_{t\in[T]} \eta_t^2.$$

Since η_t is nonincreasing in t, imposing $\alpha \ge 8a_\beta K^2 \eta_1$ makes $4a_\beta K^2 \eta_t/\alpha \le 1/2$ for every t, so the left-hand side is at least $\frac{1}{2}\sum_{t\in[T]} \eta_t \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2$. Multiplying through by 2,

$$\sum_{t\in[T]} \eta_t \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2 \le 4\left(\mathcal{L}_S(\boldsymbol{w}_1) - \mathcal{L}_S(\boldsymbol{w}(S))\right) + 4\alpha \log(1/\delta) + 4g_\beta b K^2 \exp\Big(2\log^{1/\beta}(2/\delta)\Big) \sum_{t\in[T]} \eta_t^2. \tag{41}$$

It remains to choose α and to convert the confidence level. Taking $\alpha := \max(8a_\beta K^2 \eta_1, b_F \overline{m})$, both constraints imposed above are met, and (41) holds with probability at least $1 - 4\delta$. Recalling (26) with $N = T$, so that $\Lambda = \log(T/\delta)$, and $\overline{m} = 2GK$ from Assumption 2.5,

$$b_F \overline{m} = \frac{2\left(\exp\Big(\log^{1/\beta}(T/\delta)\Big) - 1\right)}{\log(T/\delta)} 2GK = \mathcal{O}\Big(\exp\Big(\log^{1/\beta}(T/\delta)\Big)\Big),$$

so that, bounding $\log(1/\delta) \le \log(e/\delta)$, $\alpha \log(1/\delta) = \mathcal{O}(\exp(\log^{1/\beta}(T/\delta)) \log(e/\delta))$. Both this term and the coefficient $g_\beta \exp(2\log^{1/\beta}(2/\delta))$ of $\sum_{t\in[T]} \eta_t^2$ are dominated by $\exp(2\log^{1/\beta}(T/\delta))$ up to a factor depending only on β for $T \ge 2$. Since $\boldsymbol{w}_1 = \mathbf{0}$, the initial-gap assumption in Section 3 bounds the initial-gap term in (41) by $4\Delta_0$ almost surely. This deterministic bound is absorbed into $\mathcal{O}(\cdot)$ because $\log(e/\delta) \ge 1$; the implicit constant may depend on Δ_0, but not on S, n, d, T, δ. Finally, replacing δ by $\delta/4$, we obtain that with probability at least $1 - \delta$,

$$\sum_{t\in[T]} \eta_t \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2 = \mathcal{O}\left(\exp\Big(2\log^{1/\beta}(T/\delta)\Big)\left(\log \tfrac{e}{\delta} + \sum_{t\in[T]} \eta_t^2\right)\right).$$

□

Proof of Theorem 3.3. Rewriting the SGD update rule as $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t(\nabla \ell(\boldsymbol{w}_t; z_{j_t}) - \nabla \mathcal{L}_S(\boldsymbol{w}_t) + \nabla \mathcal{L}_S(\boldsymbol{w}_t))$, we get

$$\boldsymbol{w}_{t+1} = \sum_{i\in[t]} \left(-\eta_i(\nabla \ell(\boldsymbol{w}_i; z_{j_i}) - \nabla \mathcal{L}_S(\boldsymbol{w}_i))\right) - \sum_{i\in[t]} \eta_i \nabla \mathcal{L}_S(\boldsymbol{w}_i)$$

$$\implies \quad \|\boldsymbol{w}_{t+1}\| \le \left\| \sum_{i\in[t]} -\eta_i(\nabla \ell(\boldsymbol{w}_i; z_{j_i}) - \nabla \mathcal{L}_S(\boldsymbol{w}_i)) \right\| + \left\| \sum_{i\in[t]} \eta_i \nabla \mathcal{L}_S(\boldsymbol{w}_i) \right\|.$$

For the first term, the vectors $\eta_i(\nabla \ell(\boldsymbol{w}_i; z_{j_i}) - \nabla \mathcal{L}_S(\boldsymbol{w}_i))$ form an $\mathbb{R}^d$-valued MDS by the mean-zero condition in Assumption 2.2. Taking expectations in the Orlicz bound in Assumption 2.2 and using the homogeneity of the Orlicz norm (Lemma A.12(i)) bounds their Orlicz norms by $\sigma_i = \eta_i K$. Corollary A.21 therefore gives, for any $\delta \in (0, 1)$, with probability $1 - \delta$,

$$\max_{t\in[T]} \left\| \sum_{i\in[t]} \eta_i(\nabla \ell(\boldsymbol{w}_i; z_{j_i}) - \nabla \mathcal{L}_S(\boldsymbol{w}_i)) \right\| = \mathcal{O}\left(\sqrt{\log \tfrac{2}{\delta} \sum_{i\in[T]} \eta_i^2}\, \exp\Big(\log^{1/\beta}(4/\delta)\Big)\right),$$

where $\Psi_\beta^{-1}(2/\delta) = \exp(\log^{1/\beta}(1 + 2/\delta)) - 1 \le \exp(\log^{1/\beta}(4/\delta))$ because $\delta \le 2$.

For the second term, by the Cauchy–Schwarz inequality and (41), for every $t \in [T]$,

$$\left\| \sum_{i\in[t]} \eta_i \nabla \mathcal{L}_S(\boldsymbol{w}_i) \right\|^2 \le \left(\sum_{i\in[t]} \eta_i \right) \left(\sum_{i\in[t]} \eta_i \|\nabla \mathcal{L}_S(\boldsymbol{w}_i)\|^2 \right) \le \left(\sum_{i\in[T]} \eta_i \right) \mathcal{O}\left(\exp\Big(2\log^{1/\beta}(T/\delta)\Big) \left(\log \tfrac{e}{\delta} + \sum_{t\in[T]} \eta_t^2 \right) \right),$$

where both factors were enlarged from t to T using the nonnegativity of the summands. Note that the second term is bounded by the square root of this quantity. Abbreviating

$$\Sigma_1 := \sum_{t\in[T]} \eta_t, \qquad \Sigma_2 := \sum_{t\in[T]} \eta_t^2, \qquad E_\beta := \exp\Big(2\log^{1/\beta}(T/\delta)\Big), \qquad A := \log \tfrac{e}{\delta} + \Sigma_2, \tag{42}$$

and combining the two displays, we obtain

$$\max_{t\in[T]} \|\boldsymbol{w}_{t+1}\| \le \mathcal{O}\Big(\sqrt{E_\beta \Sigma_2 \log \tfrac{e}{\delta}}\Big) + \mathcal{O}\Big(\sqrt{\Sigma_1 E_\beta A}\Big), \tag{43}$$

where the first term uses $\exp(\log^{1/\beta}(4/\delta)) = \mathcal{O}(\exp(\log^{1/\beta}(T/\delta))) = \mathcal{O}(\sqrt{E_\beta})$, by Lemma A.8 with $a = 4$ and $1/\delta \le T/\delta$, and $\log(2/\delta) \le \log(e/\delta)$. Squaring (43) and using $\Sigma_2 \le \eta_1 \Sigma_1$, which holds because η_t is nonincreasing, together with $\log(e/\delta) \le A$,

$$\max_{t\in[T+1]} \|\boldsymbol{w}_t\|^2 = \mathcal{O}\big(E_\beta \Sigma_2 \log \tfrac{e}{\delta} + \Sigma_1 E_\beta A\big) = \mathcal{O}((1+\Sigma_1) E_\beta A) \tag{44}$$

(the case $t = 1$ being trivial, as $\boldsymbol{w}_1 = \mathbf{0}$).

By the initial-gap assumption in Section 3, the implicit constant in (44) can be chosen deterministically, independently of S, n, d, T, δ. Choose such a constant $C_R > 0$ and set $R := \sqrt{C_R(1+\Sigma_1)E_\beta A}$. Thus R is deterministic, and every iterate $\boldsymbol{w}_1, \dots, \boldsymbol{w}_{T+1}$ lies in the ball B_R on the event above. Since Lemma A.7 bounds the generalization error of the gradient uniformly over B_R, it applies simultaneously to all iterates, giving

$$\max_{t\in[T+1]} \|\nabla \mathcal{L}(\boldsymbol{w}_t) - \nabla \mathcal{L}_S(\boldsymbol{w}_t)\| \le \frac{bR + b'}{\sqrt{n}} \left(2 + 2\sqrt{48e\sqrt{2}(\log 2 + d\log(3e))} + \sqrt{2\log(1/\delta)} \right).$$

Squaring both sides, bounding $(bR + b')^2 = \mathcal{O}(1 + R^2)$ and substituting (44), we obtain

$$\max_{t\in[T+1]} \|\nabla \mathcal{L}(\boldsymbol{w}_t) - \nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2 \le \mathcal{O}\left(\frac{(1+\Sigma_1) E_\beta A \, (d + \log(1/\delta))}{n} \right),$$

which is (1) after a union bound over the two events used above and a rescaling of δ by a constant factor, both absorbed into the $\mathcal{O}(\cdot)$ by Lemma A.8. □

Finally, we prove Theorem 3.4.

Proof of Theorem 3.4. We retain the abbreviations $\Sigma_1, \Sigma_2, E_\beta, A$ introduced in the proof of Theorem 3.3. Splitting the population gradient into the empirical gradient and the generalization error and using $\|\boldsymbol{u} + \boldsymbol{v}\|^2 \le 2\|\boldsymbol{u}\|^2 + 2\|\boldsymbol{v}\|^2$, we obtain

$$\sum_{t\in[T]} \eta_t \|\nabla \mathcal{L}(\boldsymbol{w}_t)\|^2 \le 2 \sum_{t\in[T]} \eta_t \|\nabla \mathcal{L}(\boldsymbol{w}_t) - \nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2 + 2 \sum_{t\in[T]} \eta_t \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2 \le 2\Sigma_1 \max_{t\in[T]} \|\nabla \mathcal{L}(\boldsymbol{w}_t) - \nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2 + 2 \sum_{t\in[T]} \eta_t \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2. \tag{45}$$

Because the first term involves the maximum over the whole trajectory, Theorem 3.3 was stated in the uniform form (1).

Applying Theorem 3.3 to the first term of (45) and Theorem 3.1 to the second, and taking a union bound over the two events, we obtain with probability at least $1 - 2\delta$,

$$\sum_{t\in[T]} \eta_t \|\nabla \mathcal{L}(\boldsymbol{w}_t)\|^2 = \mathcal{O}\left(\frac{\Sigma_1(1+\Sigma_1)E_\beta A\,(d+\log(1/\delta))}{n}\right) + \mathcal{O}(E_\beta A)$$
$$= \mathcal{O}\left(E_\beta A\left(1 + \frac{\Sigma_1(1+\Sigma_1)\,(d+\log(1/\delta))}{n}\right)\right). \tag{46}$$

Dividing (46) by Σ_1 gives

$$\frac{\sum_{t\in[T]} \eta_t \|\nabla \mathcal{L}(\boldsymbol{w}_t)\|^2}{\Sigma_1} = \mathcal{O}\left(E_\beta A\left(\frac{1}{\Sigma_1} + \frac{(1+\Sigma_1)\,(d+\log(1/\delta))}{n}\right)\right),$$

which is (2) once E_β and A are expanded and δ is replaced by $\delta/2$, the resulting change of $\log(e/\delta)$ into $\log(2e/\delta)$ and of T/δ into $2T/\delta$ being absorbed by the $\mathcal{O}(\cdot)$ by Lemma A.8. □

Proof of Corollary 3.5. We write $D := d + \log(1/\delta)$, and let $c_2 \geq c_1 > 0$ be constants for which (3) reads $c_1\sqrt{n/D} \leq \Sigma_1 \leq c_2\sqrt{n/D}$. Since $n \geq D$, the lower bound gives $\Sigma_1 \geq c_1$, so $1 + \Sigma_1 \leq (1 + 1/c_1)\Sigma_1$. Substituting the two bounds into the last factor of (2), we obtain

$$\frac{1}{\Sigma_1} + \frac{(1+\Sigma_1)D}{n} \leq \frac{1}{c_1}\sqrt{\frac{D}{n}} + \left(1 + \frac{1}{c_1}\right) c_2\sqrt{\frac{n}{D}} \cdot \frac{D}{n} = \left(\frac{1}{c_1} + \left(1 + \frac{1}{c_1}\right) c_2\right)\sqrt{\frac{D}{n}},$$

and the claim follows from Theorem 3.4. □

A.5 Step-Size Instantiations

Proposition A.26 ($1/t$ decay)**.** *Fix $\eta_0 \in (0, 1/(2b)]$ and set $\eta_t := \frac{\eta_0}{t}$ for $t = 1, \ldots, T$. Then the step-size sums satisfy*

$$\sum_{t\in[T]} \eta_t = \eta_0(\log T + \mathcal{O}(1)), \qquad \sum_{t\in[T]} \eta_t^2 = \mathcal{O}(\eta_0^2). \tag{47}$$

Proof of Proposition A.26. Integral comparison gives $\log(T+1) \leq \sum_{t\in[T]} t^{-1} \leq 1 + \log T$ and $\sum_{t\in[T]} t^{-2} \leq 2$. Multiplying by η_0 and η_0^2, respectively, proves (47). □

Proposition A.27 ($1/\sqrt{t}$ decay)**.** *Fix $\eta_0 \in (0, 1/(2b)]$ and set $\eta_t := \frac{\eta_0}{\sqrt{t}}$ for $t = 1, \ldots, T$. Then the step-size sums satisfy*

$$\sum_{t\in[T]} \eta_t = \eta_0(2\sqrt{T} + \mathcal{O}(1)), \qquad \sum_{t\in[T]} \eta_t^2 = \eta_0^2(\log T + \mathcal{O}(1)). \tag{48}$$

Proof of Proposition A.27. Integral comparison gives $2\sqrt{T+1} - 2 \leq \sum_{t\in[T]} t^{-1/2} \leq 2\sqrt{T} - 1$. This and the harmonic-sum estimate in the preceding proof yield (48) after scaling by η_0 and η_0^2, respectively. □

Proposition A.28 (Cosine decay)**.** *Fix $\eta_0 \in (0, 1/(2b)]$ and the number of training steps $T \in \mathbb{N}$, $T \geq 2$, and set $\eta_t := \frac{\eta_0}{2}\left(1 + \cos\left(\frac{\pi t}{T}\right)\right)$ for $t = 1, \ldots, T$. Then the step-size sums satisfy*

$$\sum_{t\in[T]} \eta_t = \frac{\eta_0}{2}(T-1), \qquad \sum_{t\in[T]} \eta_t^2 = \frac{\eta_0^2}{8}(3T-4). \tag{49}$$

Proof of Proposition A.28. For $T \geq 2$, the standard trigonometric sums give

$$\sum_{t\in[T]} \cos\left(\frac{\pi t}{T}\right) = -1, \qquad \sum_{t\in[T]} \cos^2\left(\frac{\pi t}{T}\right) = \frac{1}{2}\sum_{t\in[T]}\left(1 + \cos\left(\frac{2\pi t}{T}\right)\right) = \frac{T}{2}.$$

Expanding η_t and η_t^2 and summing proves (49). □

A.6 Proof of the Bound under the PL Condition

Proof of Theorem 3.7. We keep the notation $\epsilon_t = \nabla \ell(\boldsymbol{w}_t; z_{j_t}) - \nabla \mathcal{L}_S(\boldsymbol{w}_t)$ and $\eta_t = \frac{4}{\mu_0(t+t_0)}$ for the step size of Theorem 3.7. From (37) we already have

$$\mathcal{L}_S(\boldsymbol{w}_{t+1}) - \mathcal{L}_S(\boldsymbol{w}_t) \le -\eta_t \langle \epsilon_t, \nabla \mathcal{L}_S(\boldsymbol{w}_t) \rangle - \tfrac{1}{2}\eta_t \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2 + b\eta_t^2 \|\epsilon_t\|^2.$$

The PL condition gives $-\frac{1}{4}\eta_t \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2 \le -\frac{1}{2}\mu_0 \eta_t \left(\mathcal{L}_S(\boldsymbol{w}_t) - \mathcal{L}_S(\boldsymbol{w}(S))\right)$, where $\frac{1}{2}\mu_0\eta_t = 2/(t+t_0)$. Substituting it into the above, we obtain

$$\begin{aligned}
&\tfrac{1}{4}\eta_t \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2 + \mathcal{L}_S(\boldsymbol{w}_{t+1}) - \mathcal{L}_S(\boldsymbol{w}(S)) \\
&\qquad \le \left(1 - \frac{2}{t+t_0}\right) \left(\mathcal{L}_S(\boldsymbol{w}_t) - \mathcal{L}_S(\boldsymbol{w}(S))\right) - \eta_t \langle \epsilon_t, \nabla \mathcal{L}_S(\boldsymbol{w}_t) \rangle + b\eta_t^2 \|\epsilon_t\|^2. \quad (50)
\end{aligned}$$

Multiplying (50) by $(t+t_0)(t+t_0-1) > 0$ and using

$$(t+t_0)(t+t_0-1)\eta_t = \frac{4(t+t_0-1)}{\mu_0}, \qquad (t+t_0)(t+t_0-1)\eta_t^2 = \frac{16(t+t_0-1)}{\mu_0^2(t+t_0)} \le \frac{16}{\mu_0^2},$$

we obtain

$$\begin{aligned}
&\frac{t+t_0-1}{\mu_0} \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2 + (t+t_0)(t+t_0-1) \left(\mathcal{L}_S(\boldsymbol{w}_{t+1}) - \mathcal{L}_S(\boldsymbol{w}(S))\right) \\
&- (t+t_0-1)(t+t_0-2) \left(\mathcal{L}_S(\boldsymbol{w}_t) - \mathcal{L}_S(\boldsymbol{w}(S))\right) \le -\frac{4(t+t_0-1)}{\mu_0} \langle \epsilon_t, \nabla \mathcal{L}_S(\boldsymbol{w}_t) \rangle + \frac{16b}{\mu_0^2} \|\epsilon_t\|^2.
\end{aligned}$$

Summing this inequality from $t = 1$ to T telescopes the second and third terms into $(T+t_0)(T+t_0-1)\left(\mathcal{L}_S(\boldsymbol{w}_{T+1}) - \mathcal{L}_S(\boldsymbol{w}(S))\right)$ and $-t_0(t_0-1)\left(\mathcal{L}_S(\boldsymbol{w}_1) - \mathcal{L}_S(\boldsymbol{w}(S))\right)$, whence

$$\begin{aligned}
&\sum_{t\in[T]} \frac{t+t_0-1}{\mu_0} \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2 + (T+t_0)(T+t_0-1) \left(\mathcal{L}_S(\boldsymbol{w}_{T+1}) - \mathcal{L}_S(\boldsymbol{w}(S))\right) \\
&\qquad \le t_0(t_0-1) \left(\mathcal{L}_S(\boldsymbol{w}_1) - \mathcal{L}_S(\boldsymbol{w}(S))\right) + \underbrace{\left(-\sum_{t\in[T]} \frac{4(t+t_0-1)}{\mu_0} \langle \epsilon_t, \nabla \mathcal{L}_S(\boldsymbol{w}_t) \rangle\right)}_{\text{(A)}} \\
&\qquad\qquad + \underbrace{\sum_{t\in[T]} \frac{16b}{\mu_0^2} \|\epsilon_t\|^2}_{\text{(B)}}. \quad (51)
\end{aligned}$$

We now bound the terms (A) and (B) in (51).

(A) With K denoting the uniform bound on the conditional Orlicz norms in Assumption 2.2, set

$$\xi_t := -\frac{4(t+t_0-1)}{\mu_0} \langle \epsilon_t, \nabla \mathcal{L}_S(\boldsymbol{w}_t) \rangle, \qquad K_{t-1} := \frac{4(t+t_0-1)}{\mu_0} \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\| K.$$

Conditions (A2) and (A3) of Proposition A.23 are verified as in the proof of Theorem 3.1, so it remains to bound K_{t-1} by a deterministic m_t, as (A1) requires. Since $\mathcal{L}_S$ averages the per-sample gradients, $\|\nabla \mathcal{L}_S(\mathbf{0})\| \le b'$, and Assumption 2.4 then gives, for every $\boldsymbol{w} \in \mathcal{W}$,

$$\|\nabla \mathcal{L}_S(\boldsymbol{w})\| \le \|\nabla \mathcal{L}_S(\boldsymbol{w}) - \nabla \mathcal{L}_S(\mathbf{0})\| + \|\nabla \mathcal{L}_S(\mathbf{0})\| \le b\|\boldsymbol{w}\| + b', \quad (52)$$

so a deterministic radius confining the trajectory produces such an m_t. The schedule in Theorem 3.7 meets the step-size condition, so (44) in the proof of Theorem 3.3 applies to it. With $\Sigma_1 = \mathcal{O}(\log T)$, $\Sigma_2 = \mathcal{O}(1)$ and hence $A = \mathcal{O}(\log(e/\delta))$ in the notation (42), it gives a deterministic

$$R = \mathcal{O}\Big(\exp\Big(\log^{1/\beta}(T/\delta)\Big) \sqrt{\log T \, \log(e/\delta)}\Big) \quad (53)$$

for which the event $\mathcal{E}_R := \{\max_{t\in[T+1]} \|\boldsymbol{w}_t\| \le R\}$ has probability at least $1-\delta$.

Since (52) bounds K_{t-1} by $m_t := 4(t+t_0-1)(bR+b')K/\mu_0$ on $\mathcal{E}_R$ only, we apply Proposition A.23 to the sequence frozen outside that event, $\tilde{\xi}_t := \chi_t \xi_t$ and $\tilde{K}_{t-1} := \chi_t K_{t-1}$ with the $\mathcal{F}_{t-1}$-measurable $\chi_t := \mathbb{1}_{\max_{s\in[t]} \|\boldsymbol{w}_s\| \le R}$. A predictable indicator preserves (A2) and (A3), so (A1)–(A3) hold with $N = T$ and $\overline{m} := \max_{t\in[T]} m_t = 4(T+t_0-1)(bR+b')K/\mu_0$. Part (i) with $\gamma = 0$, $\alpha \ge b_F \overline{m}$, $\lambda = 1/(2\alpha)$ and $x = 2\alpha \log(1/\delta)$, argued on the complement of the event in (27) as in the proof of Theorem 3.1, together with $\tilde{\xi}_t = \xi_t$ for all $t \in [T]$ on $\mathcal{E}_R$, gives after a union bound with $\mathcal{E}_R$ that with probability at least $1 - 4\delta$,

$$-\sum_{t\in[T]} \frac{4(t+t_0-1)}{\mu_0} \langle \epsilon_t, \nabla \mathcal{L}_S(\boldsymbol{w}_t)\rangle \le 2\alpha \log(1/\delta) + \frac{16 a_\beta K^2}{\mu_0^2 \alpha} \sum_{t\in[T]} (t+t_0-1)^2 \|\nabla \mathcal{L}_S(\boldsymbol{w}_t)\|^2. \tag{54}$$

The sum in (54) is absorbed by the gradient sum in (51) once $16 a_\beta K^2 (t+t_0-1)/(\mu_0 \alpha) \le 1$ for every $t \in [T]$, that is, once $\alpha \ge 16 a_\beta K^2 (T+t_0-1)/\mu_0$.

(B) Assumption 2.2, Proposition A.14 and the homogeneity of the Orlicz norm (Lemma A.12(i)) give $\left\| \frac{16b}{\mu_0^2} \|\epsilon_t\|^2 \right\|_{\Psi_{\beta,2^{-\beta}}} \le \frac{16bK^2}{\mu_0^2}$. Hence, by Corollary A.19 in the form (14) with $c = 2^{-\beta}$, with probability at least $1-\delta$,

$$\sum_{t\in[T]} \frac{16b}{\mu_0^2} \|\epsilon_t\|^2 \le \frac{16 g_\beta b K^2}{\mu_0^2} \exp\Big(2\log^{1/\beta}(2/\delta)\Big)\, T. \tag{55}$$

Substituting (54) and (55) into (51), discarding the remaining nonnegative gradient sum and taking a union bound, we obtain that with probability at least $1 - 5\delta$,

$$\begin{aligned}(T+t_0)(T+t_0-1)\left(\mathcal{L}_S(\boldsymbol{w}_{T+1}) - \mathcal{L}_S(\boldsymbol{w}(S))\right) &\le t_0(t_0-1)\left(\mathcal{L}_S(\boldsymbol{w}_1) - \mathcal{L}_S(\boldsymbol{w}(S))\right) \\ &\quad + 2\alpha\log(1/\delta) + \frac{16 g_\beta b K^2}{\mu_0^2} \exp\Big(2\log^{1/\beta}(2/\delta)\Big)\, T, \end{aligned} \tag{56}$$

for every $\alpha \ge \max\left\{16 a_\beta K^2 (T+t_0-1)/\mu_0,\ b_F \overline{m}\right\}$.

It remains to choose α and to bound the three terms on the right-hand side of (56). We take α equal to this maximum, which is deterministic because R is, and write $E_\beta = \exp(2\log^{1/\beta}(T/\delta))$ as in (42). Recalling (26) with $N = T$, we have $\Lambda = \log(T/\delta)$. Since $T \ge 2$, $t_0 \ge 1$ and $\delta < 1$, we use below that $T + t_0 - 1 \le t_0 T$, $\log T \ge \log 2$ and $\log(1/\delta) \le \Lambda$.

For the first term, the PL condition (4) at $\boldsymbol{w}_1 = \mathbf{0}$ and Assumption 3.6 give $\mathcal{L}_S(\boldsymbol{w}_1) - \mathcal{L}_S(\boldsymbol{w}(S)) \le \|\nabla \mathcal{L}_S(\mathbf{0})\|^2/(2\mu_0) \le (b')^2/(2\mu_0)$, so this term is at most $t_0^2 (b')^2/(2\mu_0)$.

For the second term, (26) gives $b_F \le 2\exp(\Lambda^{1/\beta})/\Lambda$. By (53) and $\exp(\Lambda^{1/\beta})\sqrt{\log T\, \log(e/\delta)} \ge \sqrt{\log 2}$, we have $bR + b' = \mathcal{O}(\exp(\Lambda^{1/\beta})\sqrt{\log T\, \log(e/\delta)})$. Together with $\log(1/\delta) \le \Lambda$ and $\exp(2\Lambda^{1/\beta}) = E_\beta$, this bound gives

$$b_F \overline{m} \log(1/\delta) \le \frac{2\exp(\Lambda^{1/\beta})\log(1/\delta)}{\Lambda} \cdot \frac{4 t_0 T (bR+b')K}{\mu_0} = \mathcal{O}\Big(T E_\beta \sqrt{\log T\, \log(e/\delta)}\Big).$$

The other branch of α gives $16 a_\beta K^2 (T+t_0-1)\log(1/\delta)/\mu_0 \le 16 a_\beta K^2 t_0 T \Lambda/\mu_0$, and $\Lambda \le (\beta/(2e))^\beta E_\beta$ because $u\exp(-2u^{1/\beta}) \le (\beta/(2e))^\beta$ for every $u > 0$. Because $\log(e/\delta) \ge 1$ and $\log T \ge \log 2$, both branches are $\mathcal{O}(T E_\beta \sqrt{\log T\, \log(e/\delta)})$, and hence so is $2\alpha\log(1/\delta)$.

For the third term, $2/\delta \le T/\delta$ gives $\exp(2\log^{1/\beta}(2/\delta)) \le E_\beta$, so this term is $\mathcal{O}(T E_\beta)$.

Dividing (56) by $(T+t_0)(T+t_0-1) \ge T^2$ and bounding the first and third terms by the form of the second through $1/T^2 \le 1/T$, $E_\beta \ge 1$, $\log(e/\delta) \ge 1$ and $\log T \ge \log 2$, we obtain that with probability at least $1 - 5\delta$,

$$\mathcal{L}_S(\boldsymbol{w}_{T+1}) - \mathcal{L}_S(\boldsymbol{w}(S)) = \mathcal{O}\left(\frac{E_\beta \sqrt{\log T\, \log(e/\delta)}}{T}\right).$$

Replacing δ by $\delta/5$ changes E_β into $\exp(2\log^{1/\beta}(5T/\delta))$ and $\log(e/\delta)$ into $\log(5e/\delta)$, both absorbed by the $\mathcal{O}(\cdot)$ by Lemma A.8 with $a = 5$ and $x = T/\delta \ge 1$, which proves the theorem. □

Proof of Theorem 3.8. We retain the abbreviation $E_\beta = \exp(2\log^{1/\beta}(T/\delta))$ of (42). The population PL condition in Assumption 3.6 applied at $\boldsymbol{w}_{T+1}$, followed by $\|\boldsymbol{u}+\boldsymbol{v}\|^2 \le 2\|\boldsymbol{u}\|^2 + 2\|\boldsymbol{v}\|^2$, gives

$$\mathcal{L}(\boldsymbol{w}_{T+1}) - \mathcal{L}^\star \le \frac{1}{2\mu}\|\nabla\mathcal{L}(\boldsymbol{w}_{T+1})\|^2 \le \frac{1}{\mu}\left(\|\nabla\mathcal{L}(\boldsymbol{w}_{T+1}) - \nabla\mathcal{L}_S(\boldsymbol{w}_{T+1})\|^2 + \|\nabla\mathcal{L}_S(\boldsymbol{w}_{T+1})\|^2\right). \tag{57}$$

Assumption 2.4 makes $\mathcal{L}_S$ b-smooth and $\boldsymbol{w}(S)$ minimizes it over $\mathcal{W}$, so Lemma A.6 and Theorem 3.7 bound the second term by, with probability $1-\delta$,

$$\|\nabla\mathcal{L}_S(\boldsymbol{w}_{T+1})\|^2 \le 2b\left(\mathcal{L}_S(\boldsymbol{w}_{T+1}) - \mathcal{L}_S(\boldsymbol{w}(S))\right) = \mathcal{O}\left(\frac{E_\beta\sqrt{\log T\,\log(e/\delta)}}{T}\right). \tag{58}$$

The first term is bounded by (5). Substituting it and (58) into (57) and taking a union bound, we obtain with probability at least $1-2\delta$,

$$\mathcal{L}(\boldsymbol{w}_{T+1}) - \mathcal{L}^\star = \mathcal{O}\left(E_\beta\left(\frac{\sqrt{\log T\,\log(e/\delta)}}{T} + \frac{\left(d+\log\frac{1}{\delta}\right)\log\frac{e}{\delta}\,\log T}{n}\right)\right).$$

Let $cn \le T \le Cn$ with constants $0 < c \le C$ as in Section 2.1, and set $a := \max\{C, 1\}$. For the first term, $d \ge 1$ and $\log(e/\delta) \ge 1$ give $\sqrt{\log(e/\delta)} \le \log\frac{e}{\delta} \le \left(d+\log\frac{1}{\delta}\right)\log\frac{e}{\delta}$, $T \ge 2$ gives $\sqrt{\log T} \le \log T/\sqrt{\log 2}$, and $1/T \le 1/(cn)$, so the first term is dominated by the second. For the second term, $T \le an$ and $n \ge 2$ give $\log T \le \log a + \log n \le (1 + \log a/\log 2)\log n$, and Lemma A.8 with this a and $x = n/\delta \ge 1$ gives $E_\beta \le \exp(2\log^{1/\beta} a)\exp(2\log^{1/\beta}(n/\delta))$. The same lemma absorbs the replacement of δ by $\delta/2$. □

A.7 Proof of the Bound for Clipped SGD

Throughout this subsection τ denotes the clipping level (6), and we abbreviate

$$\boldsymbol{g}_t := \nabla\ell(\boldsymbol{w}_t; z_{j_t}) = \nabla\mathcal{L}_S(\boldsymbol{w}_t) + \epsilon_t, \qquad \bar{\boldsymbol{g}}_t := \mathbb{E}[\tilde{\boldsymbol{g}}_t \mid \mathcal{F}_{t-1}],$$

with $\tilde{\boldsymbol{g}}_t = \boldsymbol{g}_t \min\{1, \tau/\|\boldsymbol{g}_t\|\}$ the clipped gradient of Algorithm 2. We write $\Gamma_t := \min\{\tau\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|, \|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|^2\}$ for the quantity summed in Theorem 3.9 and $A_t := \{\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\| \le \tau/2\}$, and retain the abbreviations Σ_1, Σ_2 of (42). The iterate $\boldsymbol{w}_t$ is $\mathcal{F}_{t-1}$-measurable, hence so are $\bar{\boldsymbol{g}}_t$, Γ_t and $\mathbb{1}_{A_t}$.

Lemma A.29. *Let $s_\beta := g_\beta\left(\exp(2\log^{1/\beta} 2) - 1\right)$. Under Assumption 2.2, for every $t \in [T]$ and every $x \ge 0$,*

$$\mathbb{E}\left[\|\epsilon_t\|^2 \mid \mathcal{F}_{t-1}\right] \le s_\beta K^2, \qquad \mathbb{P}(\|\epsilon_t\| > x \mid \mathcal{F}_{t-1}) \le 2\exp\left(-\log^\beta\left(\frac{x}{K}+1\right)\right),$$

and if τ is given by (6)*, then*

$$\mathbb{P}(\|\epsilon_t\| > \tau/8 \mid \mathcal{F}_{t-1}) \le \delta/(4T). \tag{59}$$

Proof. The pointwise inequality $1 + u^2 \le (1+u)^2$ for $u \ge 0$ gives $\Psi_{\beta,2^{-\beta}}(u^2) \le \Psi_\beta(u)$. Taking conditional expectations and using Assumption 2.2, we obtain $\mathbb{E}[\Psi_{\beta,2^{-\beta}}(\|\epsilon_t\|^2/K^2) \mid \mathcal{F}_{t-1}] \le 1$. Since $\Phi_\beta \le \Psi_{\beta,2^{-\beta}}$, the same bound holds with Φ_β in place of $\Psi_{\beta,2^{-\beta}}$, and Φ_β is convex and increasing, so Jensen's inequality gives $\Phi_\beta(\mathbb{E}[\|\epsilon_t\|^2 \mid \mathcal{F}_{t-1}]/K^2) \le 1$. Writing $r_\beta := \exp(2\log^{1/\beta} 2) - 1$ for the point at which $\Psi_{\beta,2^{-\beta}}(r_\beta) = 1$, we have $\Phi_\beta(g_\beta r_\beta) \ge \Psi_{\beta,2^{-\beta}}(r_\beta) = 1$, and the first bound follows because $s_\beta = g_\beta r_\beta$. For the second statement, conditional Markov's inequality and Assumption 2.2 give

$$\mathbb{P}(\|\epsilon_t\| > x \mid \mathcal{F}_{t-1}) \le \frac{\mathbb{E}[\Psi_\beta(\|\epsilon_t\|/K) + 1 \mid \mathcal{F}_{t-1}]}{\Psi_\beta(x/K) + 1} \le 2\exp\left(-\log^\beta(1 + x/K)\right).$$

By (6), $\log(1+\tau/(8K)) \ge \log^{1/\beta}(8T/\delta)$, so the second bound with $x = \tau/8$ gives

$$\mathbb{P}(\|\epsilon_t\| > \tau/8 \mid \mathcal{F}_{t-1}) \le 2\exp\left(-\log^\beta(1+\tau/(8K))\right) \le 2\exp(-\log(8T/\delta)) = \delta/(4T).$$

□

Proof of Theorem 3.9. We first relate the decrease in empirical risk to $\eta_t\Gamma_t$, keeping track of the bias and fluctuations of the clipped gradient.

Since $\|\tilde{\boldsymbol{g}}_t\| \le \tau$, Lemma A.5 applied to the b-smooth function $\mathcal{L}_S$ gives, for every $t \in [T]$,

$$\mathcal{L}_S(\boldsymbol{w}_{t+1}) - \mathcal{L}_S(\boldsymbol{w}_t) \le -\eta_t \langle \tilde{\boldsymbol{g}}_t, \nabla\mathcal{L}_S(\boldsymbol{w}_t)\rangle + \tfrac{1}{2} b\eta_t^2\tau^2, \tag{60}$$

and we bound $-\eta_t\langle \tilde{\boldsymbol{g}}_t, \nabla\mathcal{L}_S(\boldsymbol{w}_t)\rangle$ separately on A_t and on its complement.

On A_t, the empirical gradient norm is at most $\tau/2$, so clipping can occur only when the noise norm exceeds $\tau/2$. Clipping leaves $\tilde{\boldsymbol{g}}_t = \boldsymbol{g}_t$ when $\|\boldsymbol{g}_t\| \le \tau$ and gives $\|\tilde{\boldsymbol{g}}_t - \boldsymbol{g}_t\| = \|\boldsymbol{g}_t\| - \tau$ otherwise, so in both cases $\|\tilde{\boldsymbol{g}}_t - \boldsymbol{g}_t\| = (\|\boldsymbol{g}_t\| - \tau)_+$. Then we have

$$\|\tilde{\boldsymbol{g}}_t - \boldsymbol{g}_t\| \le (\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\| + \|\epsilon_t\| - \tau)_+ \le (\|\epsilon_t\| - \tau/2)_+ \le \|\epsilon_t\|\mathbb{1}_{\|\epsilon_t\|>\tau/2}.$$

Taking conditional expectations and applying the Cauchy–Schwarz inequality, Lemma A.29 and $\tau/2 \ge \tau/8$, on A_t we obtain

$$\begin{aligned}\|\bar{\boldsymbol{g}}_t - \nabla\mathcal{L}_S(\boldsymbol{w}_t)\| &= \|\mathbb{E}\left[\tilde{\boldsymbol{g}}_t - \boldsymbol{g}_t \mid \mathcal{F}_{t-1}\right]\| \\ &\le \sqrt{\mathbb{E}\left[\|\epsilon_t\|^2 \mid \mathcal{F}_{t-1}\right]\mathbb{P}(\|\epsilon_t\| > \tau/2 \mid \mathcal{F}_{t-1})} \le \tfrac{1}{2}\sqrt{s_\beta}\, K\sqrt{\frac{\delta}{T}} =: \bar{B}.\end{aligned} \tag{61}$$

Decomposing $\langle \tilde{\boldsymbol{g}}_t, \nabla\mathcal{L}_S(\boldsymbol{w}_t)\rangle = \langle \tilde{\boldsymbol{g}}_t - \bar{\boldsymbol{g}}_t, \nabla\mathcal{L}_S(\boldsymbol{w}_t)\rangle + \langle \bar{\boldsymbol{g}}_t - \nabla\mathcal{L}_S(\boldsymbol{w}_t), \nabla\mathcal{L}_S(\boldsymbol{w}_t)\rangle + \|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|^2$ and using (61) and Young's inequality together with $\Gamma_t = \|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|^2$ on A_t, we obtain on A_t

$$-\eta_t\langle \tilde{\boldsymbol{g}}_t, \nabla\mathcal{L}_S(\boldsymbol{w}_t)\rangle \le \xi_t - \tfrac{1}{2}\eta_t\Gamma_t + \tfrac{1}{2}\eta_t\bar{B}^2, \qquad \xi_t := -\eta_t\langle \tilde{\boldsymbol{g}}_t - \bar{\boldsymbol{g}}_t, \nabla\mathcal{L}_S(\boldsymbol{w}_t)\rangle\mathbb{1}_{A_t}. \tag{62}$$

The centering by $\bar{\boldsymbol{g}}_t$ makes ξ_t a martingale difference: the factor $\eta_t\nabla\mathcal{L}_S(\boldsymbol{w}_t)\mathbb{1}_{A_t}$ is $\mathcal{F}_{t-1}$-measurable. Thus (62) separates the descent term from a centered fluctuation and a deterministic bias bound.

Outside A_t we work on the event $\mathcal{E} := \bigcap_{t\in[T]}\{\|\epsilon_t\| \le \tau/8\}$, which has $\mathbb{P}(\mathcal{E}^c) \le \delta/4$ by (59) and a union bound. If $\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\| > \tau/2$, then $\|\epsilon_t\| \le \tau/8 \le \|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|/4$ on $\mathcal{E}$, so $\langle \boldsymbol{g}_t, \nabla\mathcal{L}_S(\boldsymbol{w}_t)\rangle \ge \frac{3}{4}\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|^2$ and $\frac{3}{4}\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\| \le \|\boldsymbol{g}_t\| \le \frac{5}{4}\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|$, whence

$$\begin{aligned}\langle \tilde{\boldsymbol{g}}_t, \nabla\mathcal{L}_S(\boldsymbol{w}_t)\rangle &= \frac{\min\{\tau, \|\boldsymbol{g}_t\|\}}{\|\boldsymbol{g}_t\|}\langle \boldsymbol{g}_t, \nabla\mathcal{L}_S(\boldsymbol{w}_t)\rangle \\ &\ge \frac{\min\{\tau, \frac{3}{4}\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|\}}{\frac{5}{4}\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|}\cdot\tfrac{3}{4}\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|^2 \\ &= \tfrac{3}{5}\min\left\{\tau\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|,\ \tfrac{3}{4}\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|^2\right\} \ge \tfrac{2}{5}\Gamma_t.\end{aligned}$$

Since $\xi_t = 0$ outside A_t and $\frac{2}{5} \le \frac{1}{2}$, this bound and (62) together give, on $\mathcal{E}$ and for every $t \in [T]$,

$$-\eta_t\langle \tilde{\boldsymbol{g}}_t, \nabla\mathcal{L}_S(\boldsymbol{w}_t)\rangle \le \xi_t - \tfrac{2}{5}\eta_t\Gamma_t + \tfrac{1}{2}\eta_t\bar{B}^2. \tag{63}$$

Substituting (63) into (60), summing over $t \in [T]$ and using $\mathcal{L}_S(\boldsymbol{w}_{T+1}) \ge \mathcal{L}_S(\boldsymbol{w}(S))$, we obtain on $\mathcal{E}$

$$\tfrac{2}{5}\sum_{t\in[T]}\eta_t\Gamma_t \le \mathcal{L}_S(\boldsymbol{w}_1) - \mathcal{L}_S(\boldsymbol{w}(S)) + \sum_{t\in[T]}\xi_t + \tfrac{1}{2}\bar{B}^2\Sigma_1 + \tfrac{1}{2}b\tau^2\Sigma_2. \tag{64}$$

We bound $\sum_{t\in[T]}\xi_t$ by Proposition A.23 with confidence parameter $\delta/4$ and $K_{t-1} := c_\beta\eta_t\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|K\mathbb{1}_{A_t}$, where $c_\beta := 2 + \frac{1}{4}\sqrt{s_\beta}$. Condition (A2) holds because $\eta_t, \nabla\mathcal{L}_S(\boldsymbol{w}_t), \mathbb{1}_{A_t}$ are $\mathcal{F}_{t-1}$-measurable. For (A3), the display preceding (61) gives $\|\tilde{\boldsymbol{g}}_t - \nabla\mathcal{L}_S(\boldsymbol{w}_t)\| \le \|\tilde{\boldsymbol{g}}_t - \boldsymbol{g}_t\| + \|\epsilon_t\| \le 2\|\epsilon_t\|$ on A_t, hence $\|\tilde{\boldsymbol{g}}_t - \bar{\boldsymbol{g}}_t\| \le 2\|\epsilon_t\| + \bar{B}$ and $|\xi_t| \le \eta_t\|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|(2\|\epsilon_t\| + \bar{B})\mathbb{1}_{A_t}$ by the Cauchy–Schwarz inequality. Since $T \ge 4$ and $\delta < 1$ give $\sqrt{\delta/T} \le \frac{1}{2}$, the bias satisfies $\bar{B} \le \frac{1}{4}\sqrt{s_\beta}K$ by (61), and the following direct estimate verifies (A3). Since $c_\beta = 2 + \frac{1}{4}\sqrt{s_\beta}$, on the $\mathcal{F}_{t-1}$-measurable event $\{K_{t-1} > 0\}$,

$$\frac{|\xi_t|}{K_{t-1}} \le \frac{2}{c_\beta}\frac{\|\epsilon_t\|}{K} + \frac{c_\beta - 2}{c_\beta}.$$

Since Ψ_β is increasing and convex, and $\Psi_\beta(1) = \exp((\log 2)^\beta) - 1 \le 1$ for $\beta > 1$, we have on this event

$$\mathbb{E}\left[\Psi_\beta\left(\frac{|\xi_t|}{K_{t-1}}\right) \,\middle|\, \mathcal{F}_{t-1}\right] \le \frac{2}{c_\beta}\mathbb{E}\left[\Psi_\beta\left(\frac{\|\epsilon_t\|}{K}\right) \,\middle|\, \mathcal{F}_{t-1}\right] + \frac{c_\beta - 2}{c_\beta}\Psi_\beta(1) \le 1$$

by Assumption 2.2. The degenerate case $K_{t-1} = 0$ forces $\xi_t = 0$ almost surely, as in the proof of Theorem 3.1. For (A1), the definition of A_t and $\eta_t \le \eta_1$ give the deterministic bound $0 \le K_{t-1} \le \frac{1}{2}c_\beta\eta_1\tau K =: \overline{m}$, which is where Assumption 2.5 entered the proof of Theorem 3.1.

To absorb the fluctuation bound into the left-hand side of (64), we now bound its variance proxy V_T by the same sum $\sum_{t\in[T]}\eta_t\Gamma_t$. Since $\Gamma_t = \|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|^2$ on A_t, we have

$$V_T = \sum_{t\in[T]} a_\beta K_{t-1}^2 = c_\beta^2 a_\beta K^2 \sum_{t\in[T]} \eta_t^2 \|\nabla\mathcal{L}_S(\boldsymbol{w}_t)\|^2 \mathbb{1}_{A_t} \le c_\beta^2 a_\beta K^2 \eta_1 \sum_{t\in[T]} \eta_t\Gamma_t. \tag{65}$$

Setting $\alpha := \max\{b_F\overline{m},\, 10c_\beta^2 a_\beta K^2\eta_1\}$ and applying Proposition A.23(i) with $\gamma = 0$, $\lambda = 1/(2\alpha)$ and $x = 2\alpha\log(4/\delta)$ bounds the probability of $\bigcup_{k\in[T]}\{S_k \ge x \text{ and } V_k \le \alpha S_k\}$ by $3\delta/4$. On the complement every $k \in [T]$ satisfies $S_k \le \max(x, V_k/\alpha) \le x + V_k/\alpha$, so (65) and $\alpha \ge 10c_\beta^2 a_\beta K^2\eta_1$ give $\sum_{t\in[T]}\xi_t \le 2\alpha\log(4/\delta) + \frac{1}{10}\sum_{t\in[T]}\eta_t\Gamma_t$ with probability at least $1 - 3\delta/4$. Substituting this into (64) and taking a union bound with $\mathcal{E}$, with probability at least $1 - \delta$,

$$\tfrac{3}{10}\sum_{t\in[T]}\eta_t\Gamma_t \le \mathcal{L}_S(\boldsymbol{w}_1) - \mathcal{L}_S(\boldsymbol{w}(S)) + 2\alpha\log(4/\delta) + \tfrac{1}{2}\bar{B}^2\Sigma_1 + \tfrac{1}{2}b\tau^2\Sigma_2. \tag{66}$$

It remains to bound the terms in (66). By the initial-gap assumption in Section 3 and $\boldsymbol{w}_1 = \mathbf{0}$, the initial gap is at most the deterministic constant Δ_0 almost surely. Since $\Sigma_1 \le \eta_1 T$, the bias contribution $\frac{1}{2}\bar{B}^2\Sigma_1$ is at most $s_\beta K^2\eta_1\delta/8$, while the smoothness remainder equals $32bK^2\exp(2\log^{1/\beta}(8T/\delta))\Sigma_2$ by (6). For the concentration term $2\alpha\log(4/\delta)$, (26) with $N = T$ and confidence parameter $\delta/4$ has $\Lambda = \log(4T/\delta)$ and

$$b_F\overline{m} = \frac{2\left(\exp(\Lambda^{1/\beta}) - 1\right)}{\Lambda}\cdot\tfrac{1}{2}c_\beta\eta_1\tau K \le \frac{8c_\beta\eta_1K^2\exp(\Lambda^{1/\beta})\exp\left(\log^{1/\beta}(8T/\delta)\right)}{\Lambda},$$

so $\log(4/\delta) \le \Lambda$ for the first branch of α, $\log(4/\delta) \le (1 + \log 4)\log(e/\delta)$ for the second, and Lemma A.8 with $a = 4$ and with $a = 8$ give $2\alpha\log(4/\delta) = \mathcal{O}(\exp(2\log^{1/\beta}(T/\delta)) + a_\beta K^2\eta_1\log(e/\delta))$. Every term of (66) is therefore $\mathcal{O}(\exp(2\log^{1/\beta}(T/\delta))(\log(e/\delta) + \Sigma_2))$, the additive constants, including Δ_0, being absorbed because $\log(e/\delta) \ge 1$. □

Implementing the clipping level in (6) requires the noise parameters K and β. Once these parameters, T, δ, and the step-size schedule are fixed, clipping provides a deterministic bound on the trajectory without Assumption 2.5. Indeed, $\|\tilde{\boldsymbol{g}}_t\| \le \tau$ and $\boldsymbol{w}_1 = \mathbf{0}$ imply, for every realization and every $t \in [T+1]$,

$$\|\boldsymbol{w}_t\| \le \sum_{s=1}^{t-1}\eta_s\|\tilde{\boldsymbol{g}}_s\| \le \tau\sum_{s=1}^{t-1}\eta_s \le \tau\sum_{s\in[T]}\eta_s.$$

Thus all clipped iterates lie in a ball centered at the initialization with radius $R = \tau\sum_{s\in[T]}\eta_s$ known before training. This radius can be used when applying the uniform gradient bound of Lemma A.7 to the clipped trajectory, under that lemma's hypotheses. It plays the role of the radius that, for SGD, is only bounded probabilistically in the proof of Theorem 3.3.

A.8 Bounded Noise and Dependence on the Tail Parameter

Even for the same bounded noise, the choice of β affects the constants in our convergence guarantee. The following calculation parallels the discussion immediately after Remark 7 in Madden et al. (2024).

Suppose that $\|\epsilon_t\| \le \rho$ almost surely for every t, with a deterministic $\rho > 0$ uniform over datasets and iterates. This is stronger than finite support at each iterate. Fixing ρ, for every $\beta > 1$ a sufficient Orlicz scale is

$$K(\beta, \rho) = \frac{\rho}{\exp((\log 2)^{1/\beta}) - 1}. \tag{67}$$

Indeed, $\Psi_\beta(\rho/K(\beta,\rho)) = 1$, so boundedness implies the conditional Orlicz bound in Assumption 2.2. As β increases, $K(\beta,\rho)$ decreases from ρ to the positive limit $\rho/(e-1)$.

However, a smaller scale does not by itself give a sharper convergence bound. Under the remaining hypotheses of Theorem 3.1, the choice of α in (41) also involves the coefficient

$$a_\beta K(\beta,\rho)^2 = \frac{8\rho^2 \exp(\beta^\beta/2) J_\beta}{[\exp((\log 2)^{1/\beta}) - 1]^2}, \tag{68}$$

where J_β is defined in (26). Since $J_\beta \ge \int_0^1 e^{2s-1/2}\, ds > 0$, this coefficient diverges as $\beta \to \infty$, despite the decrease in $K(\beta,\rho)$.

Thus, choosing a larger β solely to reduce the displayed tail factors can substantially loosen the guarantee once its constants are included. This concerns the bound supplied by our proof; the noise law and the SGD algorithm remain unchanged. Boundedness therefore does not make the choice of β immaterial: comparing guarantees requires accounting for both the Orlicz scale and the convergence constants.