Identifying stability regions of SGD with constant learning rates

Or Goldreich[1] Soham Bonnerjee[1] Qianqian Lei[1] Jiaqi Li[1]
Wei Biao Wu[1]

[1]Department of Statistics, University of Chicago

January 29, 2026

Abstract

The trade-off inherent in constant learning rate stochastic gradient descent (SGD) has been well-documented empirically: larger learning rates often yield faster convergence, but risk the possibility of exploding. However, the relevant question of an appropriate choice of learning rate has rarely received systematic treatment; one often chooses learning schedules based on domain knowledge and preliminary numerical experiments without theoretical guidance. This question is intimately related to the concept of "edge of stability", which refers to a regime where the chain neither converges nor explodes. Despite rich literature on deterministic gradient descent, the rigorous characterization of "edge of stability" for the more ubiquitous SGD chains, remains an open question. In this paper, we formalize the notion of the stability region, and develop theoretical guarantees for estimating the stability region for SGD for a wide class of strongly convex objectives. We introduce a stochastic version of Lyapunov exponent for SGD, which yields a practical, data-driven threshold for admissible learning rates. Moreover, all of our theoretical results are backed by extensive experiments. Collectively, these findings demonstrate a practically implementable as well as theoretically valid way of choosing learning rate parameters in various problems, while also paving the way to potential generalization to more complicated nonconvex landscapes.

1 Introduction

The dynamics of stochastic gradient descent (SGD) and related optimization methods have been studied extensively from the perspective of stability, generalization, and convergence. Foundational analyses such as [24] established stability guarantees for SGD and connected them to generalization, while subsequent works have investigated SGD as an approximate Bayesian inference procedure [34] and as a stochastic process with heavy-tailed gradient noise [40]. More recently, SGD has also been analyzed as a random dynamical system with almost sure convergence properties [20] and from a nonlinear time series perspective [28]. However, a consistent theme with the majority of these literature is the lack of principled guidelines on how to choose the (small enough) step-size that ensures the stability of the system. On the other hand, choosing a learning rate that is too small leads to excruciatingly slow convergence. Edge of stability analysis reflects the sweet spot between stability and convergence.

However, until recently, the *edge of stability* literature has largely focused on deterministic gradient descent (GD). Conventional theoretical analyses typically focus on the inverted problem of the stability threshold—namely, convergence guarantees at the sharpness threshold (i.e., the maximum eigenvalue of the Hessian) that guarantees stability for a GD algorithm with a given step size. The practically relevant problem of determining a problem and data-dependent threshold of learning rate that ensures stability, is much less explored. Moreover, often stochastic gradient descent is used over vanilla GD in an online setting, and much less is known about the edge-of-stability threshold for the SGD algorithms. In this article, we bridge this gap between theory and practice by proposing a theoretically valid, as well as practically implementable data-driven estimate of edge-of-stability for SGD algorithms in strongly convex setting. Our main contributions are as follows.

1.1 Main Contributions

Maximal expansion parameter. As a stepping stone to the notion of edge-of-stability, we analyze the geometric moment contraction of the SGD dynamics and define the *maximal expansion parameter* $L^\ell(\gamma)$ as the maximal Lipschitz parameter for ℓ-step SGD dynamics given $\ell \in \mathbb{N}_+$ and step size $\gamma > 0$. This parameter can be understood as the value of the weakest possible contraction of the SGD functional with step-size γ. Leveraging tools from time-series theory, we provide asymptotic theory for estimating $L^\ell(\gamma)$ uniformly over γ;

Theorem 1.1 (Theorem 3.5, informal)**.** *Under standard regularity conditions, it follows that* $\sup_{\gamma\in\Gamma} |\widehat{L}^\ell(\gamma) - L^\ell(\gamma)| = O_{\mathbb{P}}(\frac{\log n}{\sqrt{n}})$*, where* Γ *is a compact set.*

Towards the development of this result, we also borrow insights from high-dimensional statistics literature to provide a sharp uniform moment bound on the partial sums of i.i.d. random functions. We expect this result to be of independent interest.

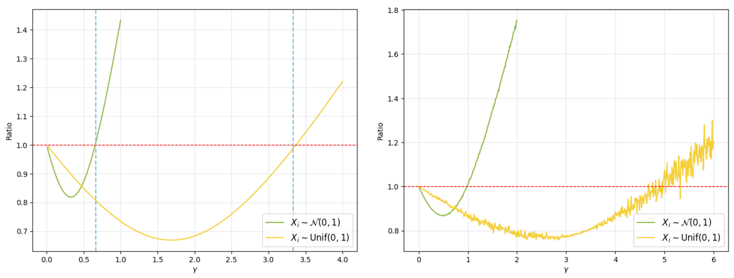
Central limit theorems for the estimators $\widehat{L}^n$. Alongside our novel conception of the maximal expansion parameter, we also establish the asymptotic distributions of the estimators we define for the `MEP`. Along the way, we provide a general central limit theory for supremum of random functions – a result that might also be of independent interest. The twin estimation and inferential results lead naturally to a statistical understanding of edge of stability as in the following.

Conceptual development and estimation of edge-of-stability for SGD. Developing on the concept of maximal expansion parameter, we rigorously characterize the *edge-of-stability*, denoted by γ_ℓ, in terms of the smallest learning rate that pushes the l-step maximal expansion parameter beyond 1, thereby making the chain explode. Our definition leads to a natural estimation strategy for this *edge-of-stability* threshold, denoted by $\widehat{\gamma}_{\ell,n}$.

Theory for the estimator $\widehat{\gamma}_n$. To the best of our knowledge, this work is the first one to provide finite-sample error bounds on the convergence property of $\gamma_{\ell,n}$; in particular, we prove the following theorem.

Theorem 1.2 (Theorem 4.4, informal)**.** *Under standard regularity conditions, it follows that* $|\widehat{\gamma}_{\ell,n} - \gamma_\ell| = O_{\mathbb{P}}(\frac{\log n}{\sqrt{n}})$.

Here we present two examples on linear regression and expectile regression respectively. The detailed settings are deferred to Remark 2.6 and Section C. In particular, the exact forms of the learning-rate boundary can be provided in the linear regression model, which are $\gamma = 2/3$ and $\gamma = 10/3$ for the random samples generated from standard normal distribution and standard uniform distribution, respectively, with $p = 2$ and $d = 1$. As shown in Figure 1, by our proposed methodology, we can very accurately hit the boundary that we derived theoretically (denoted by vertical dashed lines in Figure 1(a)).



(a) Linear regression. (b) Expectile regression.

Figure 1: Examples for edge of stability. Green: $\mathcal{N}(0,1)$; Yellow: Unif$[0,1]$. All the experiments are repeated 30 times. The detailed setting is provided in Section C.

Connections with Lyapunov theory. Our framework admits a natural interpretation in terms of Lyapunov theory once we adopt an asymptotic point of view on edge-of-stability. Indeed, by letting $\ell \to \infty$ in the definition of the maximal expansion parameter, submultiplicativity and Fekete's lemma ensure that the limit:

$$\lambda_p(\gamma) := \lim_{\ell\to\infty} \frac{1}{\ell} \log L_p^\ell(\gamma)$$

exists. This quantity is precisely the maximal *Lyapunov exponent* associated with the stochastic dynamics of SGD at learning rate γ: it measures the exponential rate at which moment distances between trajectories grow (if positive) or decay (if negative). In particular,

$$\begin{aligned} \lambda_p(\gamma) < 0 \quad &\Rightarrow \quad \text{exponential contraction of } p\text{-th moments,} \\ \lambda_p(\gamma) > 0 \quad &\Rightarrow \quad \text{exponential expansion of } p\text{-th moments.} \end{aligned}$$

Accordingly, the oracle edge of stability can be equivalently characterized as the zero-crossing of this exponent:

$$\gamma_\infty(p) := \inf\{\gamma \in \Gamma \mid \lambda_p(\gamma) \geqslant 0\}.$$

This viewpoint places our notion of edge-of-stability squarely within the classical Lyapunov framework: SGD dynamics remain stable as long as the maximal Lyapunov exponent is negative, and instability begins exactly at the point where it reaches zero. The construction aligns with classical work on Lyapunov exponents for products of random matrices and random dynamical systems, beginning with Oseledets' multiplicative ergodic theorem [36] and subsequent developments in the monographs of Bougerol and Lacroix [11] and Arnold [4]. In those settings, the sign of the maximal Lyapunov exponent governs long-run stability of the system. Our moment-based definition $\lambda_p(\gamma)$ can be interpreted as an analogue tailored to stochastic approximation schemes such as SGD, and places the edge-of-stability phenomenon within the same analytical framework.

Notation. In this paper, we denote the set $\{1, \dots, n\}$ by $[n]$. The d-dimensional Euclidean space is $\mathbb{R}^d$. For a vector $a \in \mathbb{R}^d$, $|a|$ denotes its Euclidean norm. For a matrix $M \in \mathbb{R}^{d\times m}$, $|A|$ denotes its Euclidean operator norm. For a random vector $X \in \mathbb{R}^d$, we denote $\|X\| := \sqrt{\mathbb{E}[|X|^2]}$. We also denote in-probability convergence, and stochastic boundedness by $o_{\mathbb{P}}$ and $O_{\mathbb{P}}$ respectively. We write $a_n \lesssim b_n$ if $a_n \leqslant Cb_n$ for some constant $C > 0$, and $a_n \asymp b_n$ if $C_1 b_n \leqslant a_n \leqslant C_2 b_n$ for some constants $C_1, C_2 > 0$. Often we denote $a_n \lesssim b_n$ by $a_n = O(b_n)$. Additionally, if $a_n/b_n \to 0$, we write $a_n = o(b_n)$. For a compact convex set $\Gamma \subset \mathbb{R}^d$, we denote by $\mathrm{int}(\Gamma) := \{x \in \Gamma : \exists \varepsilon > 0 \text{ such that } B_\varepsilon(x) \subset \Gamma\}$, where $B_\varepsilon(x) := \{y : |x - y| < \varepsilon\}$ is the ε-ball around $x \in \mathbb{R}^d$. In particular, we denote the closed unit ball in $\mathbb{R}^d$ by $\mathcal{B} := \overline{B_1(0)}$. Let $\mathcal{L}(\mathbb{R}^d, \mathbb{R}^m)$ denote the set of all smooth, measurable functions $f : \mathbb{R}^d \to \mathbb{R}^m$.

2 Edge of stability: preliminaries

For a function $G : \mathbb{R}^d \mapsto \mathbb{R}$, consider the following optimization problem:

$$\theta^* = \arg\min_{\theta \in \mathcal{D}} G(\theta),\ \mathcal{D} \subset \mathbb{R}^d \text{ is compact and convex,}$$

and let $\xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{P}$ be the innovations. Subsequently, all the probability statements are carried out on the same measure space as $\mathcal{P}$. Define $F \in \mathcal{C}^1$. With an online stream of $\xi_1, \xi_2, \ldots$, the classical SGD algorithm estimates θ^* via the recursion

$$\theta_i = F_{\xi_i}^{\gamma}(\theta_{i-1}),\ \text{with } F_{\xi_i}^{\gamma}(\theta) = \theta - \gamma \nabla g(\theta, \xi_i),\ i = 1, 2, \ldots, \tag{1}$$

where g is a measurable function, and $g(\cdot, x) \in \mathcal{C}^2$ satisfies $\mathbb{E}[\nabla g(\theta, \xi)] = \nabla G(\theta)$. Here $\gamma > 0$ is the constant learning rate. Before proceeding further, we introduce two key assumptions that are ubiquitous in SGD literature, as well as heavily used throughout our article.

Assumption 2.1 (μ-strong convexity)**.** There exists a $\mu > 0$ such that g is μ-strongly convex; in other words, for all $\theta, \theta' \in \mathbb{R}^d$,

$$\left\langle m(\theta) - m\left(\theta'\right), \theta - \theta' \right\rangle \geqslant \mu |\theta - \theta'|^2,$$

where $m(\theta) := \mathbb{E}[\nabla g(\theta, \xi)],\ \xi \sim \mathcal{P}$.

Strong convexity is a textbook assumption in the stochastic approximation literature [39, 37, 9]. It guarantees uniqueness of the minimizer and provides a quadratic lower bound that underlies contraction arguments. This assumption is standard in convex SGD theory, and is satisfied by canonical problems such as linear or regularized logistic regression. While it does not extend to general nonconvex objectives, it is well aligned with our focus on strongly convex settings.

Assumption 2.2 (Stochastic Lipschitz continuity)**.** Let $p \geqslant 1$. There exists some constant $N_p > 0$ such that, for all $\theta, \theta' \in \mathbb{R}^d$,

$$\|\nabla g(\theta, \xi) - \nabla g\left(\theta', \xi\right)\|_p \leqslant N_p |\theta - \theta'|.$$

Strong convexity guarantees uniqueness of the minimizer and provides a quadratic lower bound on the objective. This ensures that the SGD iterates are attracted toward a single point rather than drifting among multiple optima, and it underlies the contraction arguments that follow. On the other hand, stochastic Lipschitz-ness controls the variability of the stochastic gradients across different parameter values. This assumption enables us to bound deviations of the stochastic gradients uniformly, which is essential when passing from local to global statements in a concentration analysis. We remark that Assumptions 2.1 and 2.2 are standard features of statistical analysis of convex stochastic optimization, and have appeared extensively in [39, 37, 10, 13, 53, 43, 27].

2.1 Maximal expansion parameter: introduction

As discussed in §1, the learning rate $\gamma > 0$ plays a fundamental role in the performance of SGD; a larger value of γ may lead to θ_i being divergent. However, one can preclude the possibility of explosion by theoretically analyzing the maximum possible contraction after a given number of iterates from the current instance. We formalize this insight by borrowing the notion of contractive maps in dynamic systems defined by [46];

Definition 2.3 (MEP-ℓ)**.** The p-th *Maximal Expansion Parameter of lag 1* (MEP-1) is defined as

$$L_p(\gamma) := \sup_{\theta \neq \theta'} \frac{\mathbb{E}\left[\left|F^{\gamma}_{\xi_i}(\theta) - F^{\gamma}_{\xi_i}(\theta')\right|^p\right]}{|\theta - \theta'|^p}. \tag{2}$$

Generalizing (2), for $\ell \in \mathbb{N}_+$, the ℓ-lag maximal expansion (MEP-ℓ) can be defined as:

$$L^{\ell}_p(\gamma) := \sup_{\theta \neq \theta' \in \mathcal{D}} \frac{\mathbb{E}\left[|F^{\gamma}_{\xi_{i+\ell-1}:\xi_i}(\theta) - F^{\gamma}_{\xi_{i+\ell-1}:\xi_i}(\theta')|^p\right]}{|\theta - \theta'|^p},$$

where the composite map $F^{\gamma}_{(a+b):a}(\cdot) := F^{\gamma}_{a+b} \circ \ldots \circ F^{\gamma}_{a+1} \circ F^{\gamma}_{a}(\cdot)$.

The quantity $L_p(\gamma)$ can be interpreted as the maximal possible value of the Lipschitz constant in equation (17) of [28]; as we will discuss in §4, this interpretation readily leads to a notion of edge-of-stability through the need to ensure geometric moment contraction. However, before proceeding further, we take a pause here to make a crucial observation regarding the tractability of the maximal expansion parameter.

The maximal expansion parameter, as is defined, concerns computing a supremum over pairs of distinct points θ, θ'. This form may appear cumbersome for both analysis, as well as any direct approach to estimation. However, in Lemma 2.4, we transform the corresponding sample version into a tractable quantity through equivalent characterization through $\nabla_\theta F^{\gamma}_{\xi_i}(\theta)$ for all ξ_i and θ.

Lemma 2.4. *Let $\mathcal{D} \subset \mathbb{R}^d$ be a compact convex set and $\gamma > 0$ be given. Suppose $F^{\gamma}_{\xi_i}(\theta)$ be as in Equation* (1)*. Then, under Assumption 2.2 it follows that:*

$$\sup_{\theta \neq \theta' \in \mathcal{D}} \frac{1}{n} \sum_{i=1}^{n} \frac{\left|F^{\gamma}_{\xi_i}(\theta) - F^{\gamma}_{\xi_i}(\theta')\right|^p}{|\theta - \theta'|^p} = \sup_{\theta \in \mathcal{D}} \sup_{u:|u|=1} \frac{1}{n} \sum_{i=1}^{n} \left|\nabla_\theta F^{\gamma}_{\xi_i}(\theta) u\right|^p. \tag{3}$$

Additionally, it follows that

$$\sup_{\theta \neq \theta' \in \mathcal{D}} \frac{\mathbb{E}\left[\left|F^{\gamma}_{\xi_i}(\theta) - F^{\gamma}_{\xi_i}(\theta')\right|^p\right]}{|\theta - \theta'|^p} = \sup_{\theta \in \mathcal{D}} \sup_{u \in \mathbb{R}^d:|u|=1} \mathbb{E}\left[\left|\nabla_\theta F^{\gamma}_{\xi_i}(\theta)\, u\right|^p\right].$$

Remark 2.5. Virtually the same arguments as Lemma 2.4 allow us to write

$$\sup_{\theta \in \mathcal{D}} \sup_{u \in \mathbb{R}^d:|u|=1} \frac{1}{n} \sum_{i=1}^{n} \left[\left|\nabla_\theta F^{\gamma}_{\xi_i}(\theta)\, u\right|^p\right] = \sup_{\theta \in \mathcal{D}} \lim_{\delta \to 0} \sup_{v:|v|=1} \frac{1}{n} \sum_{i=1}^{n} \frac{|F^{\gamma}_{\xi_i}(\theta) - F^{\gamma}_{\xi_i}(\theta + \delta v)|^p}{|\delta|^p}. \tag{4}$$

Equation (4) is especially useful in situations where the computation of $\nabla_\theta F^{\gamma}_{\xi_i}(\theta)$ is intractable. It allows us perform numerical differentiation by considering a fine-grained mesh around θ in different directions.

Remark 2.6 (Range of γ in linear regression). Consider the linear regression model

$$Y_i = X_i^T \theta + \epsilon_i,$$

where $\theta \in \mathbb{R}^d$ is the population parameter vector of interest and $\epsilon_i \in \mathbb{R}$ are i.i.d. random noise independent of $\{X_i\}_{i \geqslant 1}$. In this setting, the Assumption 2.1 and Assumption 2.2 holds with $\mu = \lambda_{\min}\{\mathbb{E}(X_i X_i^\top)\}$ and $N_2 = \sup_{\delta \in \mathbb{R}^d:|\delta|=1} : \|X_i X_i^\top \delta\|_2^2$, where $\lambda_{\min}\{\cdot\}$ refers to the smallest eigenvalue. Consequently, Theorem 2.2 in [28] ensures $L_2(\gamma) < 1$ as long as

$$0 < \gamma < \frac{2\lambda_{\min}\{\mathbb{E}(X_i X_i^T)\}}{\sup_{\delta \in \mathbb{R}^d:|\delta|=1} : \|X_i X_i^T \delta\|_2^2}.$$

It can be demonstrated that this range reaches the optimum in general. By this expression, for $d = 1$, the boundary of γ is $2/3 \approx 0.67$ when X_i follows the standard normal distribution $\mathcal{N}(0,1)$ and is $10/3 \approx 3.3$ when X_i follows the standard uniform distribution $U[0,1]$.

Lemma 2.4 allows us to consider a supremum over a single parameter, boosting tractability by eliminating dependence on arbitrary pairs. In lieu of Lemma 2.4, one can approach estimating $L_p(\gamma)$ (and in general $L_p^{\ell}(\gamma)$) by way of the corresponding empirical versions:

$$\widehat{L}_p^n(\gamma) := \sup_{\theta \in \mathcal{D}} \sup_{u:|u|=1} \frac{1}{n} \sum_{i=1}^{n} \left| \nabla_\theta F_{\xi_i}^{\gamma}(\theta)\, u \right|^p,$$

and in general for any $\ell \in \mathbb{N}_+$, define $\widehat{L}_p^{\ell,n}(\gamma)$ as:

$$\sup_{\theta \neq \theta' \in \mathcal{D}} \frac{\frac{1}{n}\sum_{i=1}^{n} |F_{\xi_{i+\ell-1}:\xi_i}^{\gamma}(\theta) - F_{\xi_{i+\ell-1}:\xi_i}^{\gamma}(\theta')|^p}{|\theta - \theta'|^p}. \tag{5}$$

Following from Lemma 2.4, we would also like to introduce a similar notion for $L_p^{\ell}(\gamma)$ and its sample version $\widehat{L}_p^{\ell,n}(\gamma)$:

$$L_p^{\ell}(\gamma) = \sup_{\theta \in \mathcal{D}} \sup_{u:|u|=1} \mathbb{E}\left[\left|\nabla_\theta (F_{(i+\ell-1):i(\theta)})u\right|^p\right] = \sup_{\theta \in \mathcal{D}} \sup_{u:|u|=1} \mathbb{E}\left[\left|\left(\prod_{k=1}^{\ell} \nabla_\theta F_{\xi_{i+\ell-k}}^{\gamma}(\theta^{l-k})\right) u\right|^p\right], \text{ and,}$$

$$\widehat{L}_p^{\ell,n}(\gamma) = \sup_{\theta \in \mathcal{D}} \sup_{u:|u|=1} \frac{1}{n}\sum_{i=1}^{n} \left|\nabla_\theta (F_{(i+\ell-1):i(\theta)})u\right|^p = \sup_{\theta \in \mathcal{D}} \sup_{u:|u|=1} \frac{1}{n}\sum_{i=1}^{n} \left|\left(\prod_{k=1}^{\ell} \nabla_\theta F_{\xi_{i+\ell-k}}^{\gamma}(\theta^{l-k})\right) u\right|^p$$

where $\theta^{(0)} = \theta$ and for $k > 0$, $\theta^{(k)} = F_{\xi_{i+k-1}}^{\gamma}(\theta^{(k-1)})$.

Subsequently, we primarily focus on $L_p(\gamma)$ and its estimator $\widehat{L}_p^n(\gamma)$. A naive treatment of the general ℓ-case can be understood to be quite similar; however, we mention another interesting property of the function MEP-ℓ that renders the general case practically trivial after one has considered the $\ell = 1$ scenario. In particular, it follows that the sequence $\{L_p^{\ell}(\gamma)\}_{\ell \in \mathbb{N}_+}$ is submultiplicative.

Proposition 2.7. *Set $p \geqslant 1$ and $\gamma \in \Gamma$, and let $k, \ell \in \mathbb{N}$. Then:*

$$L_p^{\ell+k}(\gamma) \leqslant L_p^{k}(\gamma) \cdot L_p^{\ell}(\gamma).$$

Armed with these additional insights, in the next section we develop an asymptotic theory for $\widehat{L}_p^n(\gamma)$.

3 Theoretical results on maximal expansion parameters

Before stating our main results, we collect a set of regularity assumptions that ensure both well-posedness of the optimization problem and tractability of the analysis. Some of these are standard in the study of SGD, but we briefly comment on their roles.

Assumption 3.1 (Compact and convex domains)**.** The parameters θ and γ are confined to compact convex domains $\mathcal{D} \subset \mathbb{R}^d$ and $\Gamma = [a, b]$, for $b > a > 0$.

Compactness of the parameter and learning-rate domains is not intrinsic to SGD, but serves as a standard technical device in empirical process theory. It guarantees well-posedness when taking suprema over continuous index sets and facilitates the use of covering arguments and δ-nets. Although unconstrained optimization problems such as linear regression are typically posed on $\mathbb{R}^d$, in practice SGD iterates remain bounded due to regularization, explicit projection, or simply because divergence leads to algorithmic instability (see, e.g., projection-based variants of SGD in [35], [26]). Finally, the assumption that $a \wedge b > 0$ excludes the trivial case $L_p^{\ell,n}(\gamma) = 1$, where the SGD chain does not move at all.

Assumption 3.2 (Lipschitz property, B.1, informal)**.** There exists a constant $K_p < \infty$ such that the operator norm of $\nabla_\theta F_\xi^\gamma(\theta)$ fulfills a stochastic Lipschitz property with respect to θ and γ.

In order to control the first derivative of the SGD process, we must bound second order derivative behavior of the function, giving rise Assumption 3.2. This condition is satisfied by many practical smooth models, and rules out only highly irregular loss landscapes.

Assumption 3.3 ($2p$-moment bound)**.** Fix $p \geqslant 1$. Assume

$$A := \mathbb{E}\left[\sup_{\theta\in\mathcal{D},\gamma\in\Gamma}\ \sup_{u:|u|=1}\left|\nabla_\theta F_{\xi_i}^\gamma(\theta)u\right|^{2p}\right] < \infty.$$

Finite $2p$-th moments of the stochastic gradients strengthen Assumption 2.1 and are standard when deriving concentration inequalities for SGD. In our setting, this condition ensures that deviation inequalities for the empirical expansion parameter hold with high probability, which is essential for establishing nonasymptotic confidence statements about the edge of stability.This requirement is reasonable in practice for smooth models where gradients have sub-Gaussian or sub-exponential tails.

Assumption 3.4 (Differentiability)**.** Fix $p \geqslant 1$. We assume that $\frac{\partial}{\partial\gamma}L_p^\ell(\gamma)$ is defined for all $\gamma \in \Gamma$, and that there exists some $K_p > 0$ such that $\sup_{\gamma\in\Gamma}\left|\frac{\partial}{\partial\gamma}L_p^\ell(\gamma)\right| \leqslant K_p$.

Differentiability of $L_\ell^\gamma(p)$ with respect to γ ensures that the stability threshold behaves regularly in a neighborhood of the edge. This smoothness enables a first-order expansion around $\gamma_\ell(p)$, which is the key step in transferring concentration of $\widehat{L}_{\ell,n}^\gamma(p)$ into consistency of $\widehat{\gamma}_{\ell,n}(p)$.

3.1 Asymptotics of `MEP`-ℓ

In this section, we control the estimation error of $\widehat{L}_p^n(\gamma)$ uniformly over $\gamma \in \Gamma$, setting the stage of eventual estimation of the edge-of-stability upon its definition. To that end, we recognize that $\widehat{L}_p^n(\gamma) = \sup_{\theta\in\mathcal{D}}\frac{1}{n}\sum_{i=1}^n\left|\nabla_\theta F_{\xi_i}^\gamma(\theta)\right|^p$ as the $\mathcal{L}_\infty$ norm of mean of random functions. Subsequently, adapting the tools of [16], we provide a general result controlling the partial sums of i.i.d. random functions.

Theorem 3.5. *Let $\Phi \subset \mathbb{R}^d$ be a compact convex set and $X_1(\varphi), \dots, X_n(\varphi)$ be* i. i. d. *random functions with $X_i : \mathbb{R}^d \mapsto \mathbb{R}^m$ for some $d, m \geqslant 1$. For $p \geqslant 1$, denote*

$$K_\Phi := \mathbb{E}\left[\sup_{\varphi\neq\varphi'\in\Phi}\frac{|X_i(\varphi) - X_i(\varphi')|^p}{|\varphi-\varphi'|^p}\right] < \infty, \ \ \textit{and} \tag{6}$$

$$A_{\Phi,p} := \mathbb{E}\left[\sup_{\varphi\in\Phi}|X_i(\varphi)|^{2p}\right] < \infty. \tag{7}$$

Let the n-th partial sum be defined as $S_{n,p}(\varphi) := \sum_{i=1}^{n} |X_i(\varphi)|^p$. Then it holds that:

$$\mathbb{E}\left[\sup_{\varphi\in\Phi} |S_{n,p}(\varphi) - \mathbb{E}\left[S_{n,p}(\varphi)\right]|\right] = O\left(\sqrt{n}\log n\right),$$

where $O(\cdot)$ hides constants solely related to p, d, m, Φ and μ.

Theorem 3.5, to the best of our knowledge, is the *first such result* controlling the moments of sums of random function. As such, it may be of independent interest. Importantly, due to the compactness of Φ, the mean discrepancy between the $\mathcal{L}_\infty$ norm of empirical and oracle average, decays at the near-parametric rate $O(\sqrt{(\log n)/n})$.

This general result serves as the workhorse for bounding the estimation error of $\widehat{L}_p^n(\gamma)$. As an application of Theorem 3.5, we recover the following guarantee on $\widehat{L}_p^n(\gamma)$, and more generally $\widehat{L}_p^{\ell,n}(\gamma)$.

Theorem 3.6. *Fix $\ell \in \mathbb{N}_+$, and recall $L_p^\ell(\gamma)$ and $\widehat{L}_p^{\ell,n}(\gamma)$ from Definition 2.3 and* (5) *respectively. Then, under Assumptions 2.1-2.2 and 3.1-3.3, it holds that:*

$$\mathbb{E}\left[\sup_{\gamma\in\Gamma} \left|\widehat{L}_p^{\ell,n}(\gamma) - L_p^\ell(\gamma)\right|\right] = O\left(\frac{\log n}{\sqrt{n}}\right).$$

Establishing such guarantees is essential: without quantitative control of the estimation error, any attempt to approximate the edge of stability would remain heuristic. Theorem 3.6 provides precisely this control, paving the way for our eventual goal: precisely estimating edge-of-stability.

3.2 Central limit theory of $\hat{L}_p^{\ell,n}(\gamma)$

In this section, we expand on Theorem 3.6 to facilitate statistical inference with $\hat{L}$. Similar to Theorem 3.5, we start off with a general result.

Theorem 3.7. *Consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and for $i \in [n]$, let $X_i : \Omega \to \mathcal{L}(\mathbb{R}^d, \mathbb{R}^m)$ be i.i.d. random elements on $\mathcal{L}(\mathbb{R}^d, \mathbb{R}^m)$ for some $d, m \geqslant 1$. Denote:*

$$S_n(\cdot) := \sum_{i=1}^{n} (X_i(\cdot) - \mathbb{E}[X_i(\cdot)]).$$

On the same probability space, let $Z_1, \ldots, Z_n$ be an i.i.d. mean-zero $\mathbb{R}^m$-valued Gaussian random field on $\mathbb{R}^d$, such that for all $x \in \mathbb{R}^d$, $Z_i(x) \sim N(0, \mathrm{Cov}(X_i(x)))$, and

$$\mathrm{Cov}(Z_i(x), Z_i(y)) = \mathrm{Cov}(X_i(x), X_i(y)).$$

Denote $S_n^Z(\cdot) := \sum_{i=1}^{n} Z_i(\cdot)$. Let $\Phi \subset \mathbb{R}^d$ be a compact convex set. Then, under the conditions of Theorem 3.5, it holds that

$$\sup_{z\in\mathbb{R}} \left|\mathbb{P}(\sup_{x\in\Phi} |S_n(x)| \geqslant z) - \mathbb{P}(\sup_{x\in\Phi} \left|S_n^Z(x)\right| \geqslant z)\right| \lesssim n^{-C},$$

where, $\lesssim$ and C hide constants depending solely on K_Φ and $A_{\Phi,p}$.

As with Theorem 3.5, this statement is more general than the focus of this paper, but can be applied in order to derive central limit theorems for the estimators of $L_p^\ell(\gamma)$ and $\gamma_\ell(p)$.

Theorem 3.8. *Fix $\ell \in \mathbb{N}_+$, and recall $L_p^\ell(\gamma)$ and $\widehat{L}_p^{\ell,n}(\gamma)$ from Definition 2.3 and (5) respectively. Additionally, set $S_n(\gamma) = n(\widehat{L}_p^{\ell,n}(\gamma) - L_p^\ell(\gamma))$ and $S_n^Z(\gamma)$ in accordance with its definition in the theorem statement of Theorem 3.7. Then, under Assumptions 2.1-2.2 and 3.1-3.3, it holds that:*

$$\sup_{z\in\mathbb{R}} \Big|\mathbb{P}\big(\sup_{\gamma\in\Gamma} \Big|\widehat{L}_p^{\ell,n}(\gamma) - L_p^\ell(\gamma)\Big| \geqslant z\big) \\ - \mathbb{P}\big(\sup_{\gamma\in\Gamma} \big|S_n^Z(\gamma)\big| \geqslant nz\big)\Big| \lesssim n^{-C},$$

where $\lesssim$ and C hide constants depending solely on K_Φ and $A_{\Phi,p}$.

The proof of Theorem 3.8 follows directly from Theorem 3.7, and hereby omitted.

In order to practically use the Theorem 3.8 for statistical inference without apriori knowledge of the corresponding asymptotic covariances, one can implement multiplier bootstrap [14], whose theoretical validity can be established likewise as in the above results. We also empirically demonstrate the validity of the Gaussian approximation in the case of linear regression in Figure 2.

4 Edge of stability: definition and estimation

In this section, we endeavor to precisely characterize the edge of stability through the explosions of `MEP`-ℓ. Subsequently, we propose a corresponding version of data-driven edge-of-stability, and provide finite sample error bounds.

Definition 4.1 (`EOS`-ℓ)**.** Fix $\ell \in \mathbb{N}_+$. The oracle *edge-of-stability of lag ℓ* (`EOS`-ℓ) is defined as

$$\gamma_\ell\,(p) \;:= \inf\left\{\gamma > 0 \mid L_p^\ell(\gamma) \geqslant 1\right\}.$$

Clearly, $L_p^\ell(0) = 1$. By recalling Assumption 3.1, $\gamma_\ell(p)$ can be interpreted as smallest $\gamma > 0$ such that the geometric moment contraction no longer holds for the SGD dynamics. As with $L_p^\ell(\gamma)$, we ignore the subscript ℓ whenever $\ell = 1$.

Remark 4.2. It is not yet evident why $\gamma_\ell(p)$ even exists. To ensure its existence, we proceed via the following argument.

1. Recall Theorem 2.2 in [28]; under Assumptions 2.1-2.2, there exists a function $\kappa : \mathbb{R}_+ \mapsto \mathbb{R}_+$, such that for $0 < \gamma < \kappa(p)$, we have $L_p(\gamma) < 1$. Here we remark that [28] dealt with the $p > 1$ case; which however can imply the case with $p = 1$ by Hölder's inequality as shown in [47].

2. Since $g(\cdot, \xi) \in \mathcal{C}^2$, by the Lebesgue Dominate Convergence Theorem (DCT), $\lim_{\gamma\to\infty} L_p(\gamma)/\gamma^p = \sup_\theta \sup_{u\in\mathbb{R}^d:|u|=1} \mathbb{E}[|\nabla_\theta^2 g(\theta, \xi)u|^p]$.

Conditions 1 and 2 above ensure that $\gamma_\ell(p) \in \Gamma$ exists.

Definition of the empirical version of $\gamma_\ell(p)$, denoted by $\widehat{\gamma}_{\ell,n}(p)$, is not straight-forward, since the guarantees in [28] extend only to $L_p^\ell(\gamma)$, and not to its empirical version. However, Theorem 3.6 ensures that for all $\gamma \in \Gamma$ for any compact set Γ, $L_p^\ell(\gamma)$ is closely approximated by its empirical version $\widehat{L}_p^{\ell,n}(\gamma)$. Therefore, it is conceivable to leverage Theorem 3.6 to obtain a precisely-defined compact convex set Γ, such that, with high probability, $\widehat{L}_p^{\ell,n}(\gamma)$ crosses 1 on $\mathrm{int}(\Gamma)$. More formally, by Theorem 2.2 in [28] and continuity of $L_p^\ell(\cdot) < 1$, there exists some $\delta > 0$ and $\gamma_0 > 0$ such that $L_p^\ell(\gamma) < 1$ for all $\gamma \in B_\delta(\gamma_0)$. On the other hand, let $\gamma_\ell^\dagger(p) := \inf\left\{\gamma > 0 \mid L_p^\ell(\gamma) > 2\right\}$. Similar to Remark 4.2, $\gamma_\ell^\dagger(p)$ is well-defined. Then we proceed to define the edge-of-stability at lag ℓ.

Definition 4.3. Denote $\Gamma := [\gamma_0, \gamma_\ell^\dagger(p)]$. The oracle $\texttt{EOS-}\ell$ is defined as

$$\widehat{\gamma}_{\ell,n}(p) := \min\left\{\gamma \in \Gamma \mid \widehat{L}_p^{\ell,n}(\gamma) \geqslant 1\right\}.$$

Note that, by definition of Γ, it also follows that $\gamma_\ell(p) \in \mathrm{int}(\Gamma)$. We establish the following conventions for the edge-cases: if $\sup_{\gamma\in\Gamma} \widehat{L}_p^{\ell,n}(\gamma) < 1$, then $\widehat{\gamma}_{\ell,n}(p) = 0$; on the other hand, if $\inf_{\gamma>0} \widehat{L}_p^{\ell,n}(\gamma) = 1$, then $\widehat{\gamma}_{\ell,n}(p) = \infty$. In fact, in the following we prove that these edge cases have vanishing probability, and consequently, we recover the asymptotic consistency of $\gamma_{\ell,n}(p)$ as an estimator of $\gamma_\ell(p)$.

Theorem 4.4. *Fix $\ell \in \mathbb{N}_+$, and recall $\gamma_\ell(p)$ and $\widehat{\gamma}_{\ell,n}(p)$ from Definitions 4.1 and 4.3 respectively. Then, under Assumptions 2.1-2.2 and 3.1-3.3,*

$$\mathbb{P}(\widehat{\gamma}_{\ell,n}(p) \in \mathrm{int}(\Gamma)) \to 1 \quad \textit{as } n \to \infty.$$

Additionally, it holds that:

$$|\widehat{\gamma}_{\ell,n}(p) - \gamma_\ell(p)| = O_{\mathbb{P}}\left(\frac{\log n}{\sqrt{n}}\right).$$

To the best of our knowledge, Theorem 4.4 provides the *only*, provably consistent estimator of $\texttt{EOS-}\ell$ in the context of SGD. Beyond theoretical interest, the practical relevance of estimator cannot be overstated; $\widehat{\gamma}_{\ell,n}(p)$ indicates a data-driven threshold of the learning rate, beyond which the SGD dynamics explode with high probability. Moreover, employing multiplier bootstrap and the central limit theory (Theorem 3.8), it is possible to produce asymptotically valid confidence intervals for $\hat{\gamma}_{\ell,n}(p)$; We defer the technical details for future work.

5 Simulation

In this section, we empirically characterize the edge-of-stability region, and assess its statistical optimality. In particular, in §5.1, we estimate the contraction ratio $L_p(\gamma)^{1/p}$ as a function of γ and identify the smallest γ at which contraction fails. The resulting empirical boundary closely matches our theoretical prediction, demonstrating that the proposed "edge of stability" is tight. Moving on, in 5.2, we exhibit the asymptotic gaussianity of our estimator $\widehat{L}$ and $\widehat{\gamma}$. Additional numerical experiments are provided in Appendix §C.

5.1 Tightness of "edge of stability"

We consider the data generating mechanism $Y_i = X_i^\top \theta^* + \epsilon_i$, and let $\xi_i = \{X_i, y_i\}_{i\in\mathbb{N}_+}$ denote the observed sequential data and θ^* is the unknown population parameter of interest. For the purpose of this study, we consider linear regression with squared loss. Let the true feature vector $X \in \mathbb{R}^d$ is drawn either from a Gaussian design $\mathcal{N}(0, I_d)$, and the noise ξ is drawn from a standard Gaussian distribution, independent of $\{X_i\}_{i\in\mathbb{N}}$. We vary the ambient dimension $d \in \{1, 2, 3, 5, 10\}$, the composition lag $l \in \{1, 5, 10\}$, and the moment index $p \in \{2, 4\}$. We sweep γ on a grid $\Gamma_{\text{norm}} = \{0.01, 0.02, \ldots, 1.00\}$ under the Gaussian design, and for the Uniform design we use $\Gamma_{\text{unif}} = \{0.01, 0.02, \ldots, 4.00\}$ to account for the different curvature scales observed in practice.

Across all configurations, the curve $\gamma \mapsto L_p(\gamma)^{1/p}$ exhibits a clear elbow: it decreases from 1, reaches a minimum, then increases, crossing 1 at the stability edge $\widehat{\gamma}$, after which it grows rapidly (diverging). Since this transition occurs well within the plotted range, $\widehat{\gamma}$ is visually sharp and can be localized to a narrow interval.

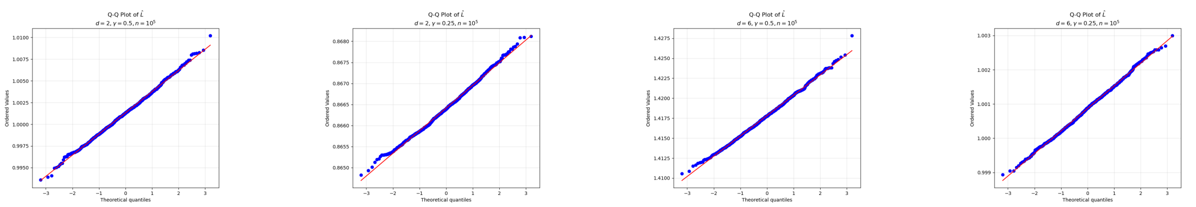
In Figure 3, subplot (a) shows that increasing $p \in \{1, 2, 3, 5, 10\}$ shifts the crossing leftward, i.e., stronger tail sensitivity yields a smaller admissible step-size, consistent with [28]. Subplot (b) shows that increasing lag ℓ enlarges the stable region: by sub-multiplicativity,

$$L_p^\ell(\gamma) \leqslant L_p(\gamma)^\ell,$$

larger ℓ pushes ratios further below 1 on the stable side and further above 1 on the unstable side, allowing larger γ while maintaining contraction. Subplot (c) shows that the stable region contracts as the dimension d increases. Moreover, the empirical edge $\widehat{\gamma}_{\ell,n}(p)$ (yellow curve in (a), red curve in (b)) closely matches the theoretical boundary in Remark 2.6 for $d = 1$ and $\ell = 1$, namely $\widehat{\gamma}_{\ell,n}(2) = 2/3$ for $X_i \sim \mathcal{N}(0, 1)$ and $\widehat{\gamma}_{\ell,n}(2) = 10/3$ for $X_i \sim \text{Unif}[0, 1]$. Overall, Figure 3 validates that $\{\gamma : L_p(\gamma) < 1\}$ is a single interval starting at 0, whose boundary is captured by the unique intersection with level 1, and whose dependence on (p, ℓ, d) matches the theoretical predictions.

5.2 Gaussianity of $\widehat{L}$

We also provide plots demonstrating the validity of Theorem 3.8. As with §5.1, we consider the linear regression setting with the feature vector $X \in \mathbb{R}^d$ drawn from a Gaussian design $\mathcal{N}(0, I_d)$, and the noise ξ drawn from a standard Gaussian distribution, independent of $\{X_i\}_{i\in\mathbb{N}}$. Setting $n = 10^5$, $\ell = 1$ and $p = 2$, we use Q-Q plots with 1000 i.i.d. samples to demonstrate that for $d \in \{2, 6\}$ and $\gamma \in \{0.5, 0.25\}$, $\widehat{L}_p^{\ell,n}(\gamma)$ is asymptotically normal. We have already demonstrated in the simulations in Figure 3 and proved in Theorem 3.6 that $\widehat{L}_p^{\ell,n}(\gamma)$ is consistent, meaning that $\widehat{L}_p^{\ell,n}(\gamma) - L_p^\ell(\gamma)$ is asymptotically mean zero normal.



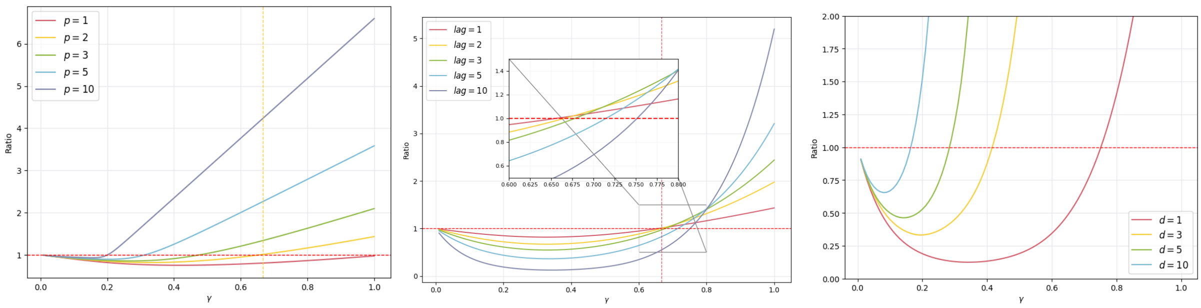
(a) $d = 2$, $\gamma = 0.5$ (b) $d = 2$, $\gamma = 0.25$ (c) $d = 6$, $\gamma = 0.5$ (d) $d = 6$, $\gamma = 0.25$

Figure 2: Q-Q plots demonstrating asymptotic normality of the estimator for the `MEP`.

6 Conclusions and Discussion

In this work, we provided a principled characterization of the stability region of SGD with constant learning rates. By introducing the notion of the maximal expansion parameter and connecting it to Lyapunov exponents, we established a rigorous definition of the edge-of-stability and developed a consistent, data-driven estimator for identifying admissible learning rates. Our theoretical results, complemented by extensive simulations on linear and expectile regression, confirm that the proposed framework accurately captures the transition from stable to unstable regimes. These findings supply both a theoretical foundation and a practical tool for selecting constant step sizes in online learning algorithms.

Looking ahead, the observed dependence of stability thresholds on factors such as dimension, lag, and moment index underscores the importance of adaptive, data-driven tuning strategies, rather than relying on fixed heuristics. Moreover, by situating SGD stability with the Lyapunov exponent in dynamic systems, our work lays the groundwork for unifying deterministic and stochastic stability analyses, potentially leading to sharper guidelines for learning rate selection across a broad range of optimization problems.



(a) $d = 1, \ell = 1, p \in \{1, 2, 3, 5, 10\}$ (b) $d = 1, \ell \in \{1, 2, 3, 5, 10\}, p = 2$ (c) $d = \{1, 3, 5, 10\}, \ell = 10, p = 2$

Figure 3: **Linear regression with $X_i \sim \mathcal{N}(0, I_d)$.** Each panel plots $\widehat{L}_p^\ell(\gamma)^{1/p}$ versus the constant step size γ for linear regression. Experimental factors and grids follow the setup marked in subplot labels.

Software and Data

All the relevant reproducible codes can be found in the anonymous Github repository. All the theoretical results and assumptions are rigorously proved and validated in the Appendix §D-§F.1.

References

[1] A. Agarwala and J. Pennington. High dimensional analysis reveals conservative sharpening and a stochastic edge of stability, 2025.

[2] K. Ahn, S. Bubeck, S. Chewi, Y. T. Lee, F. Suarez, and Y. Zhang. Learning threshold neurons via edge of stability. In *Advances in Neural Information Processing Systems*, volume 36, pages 19540–19569, 2023.

[3] A. Andreyev and P. Beneventano. Edge of Stochastic Stability: Revisiting the Edge of Stability for SGD, 2025.

[4] L. Arnold. *Random Dynamical Systems.* Springer, Berlin, 1998.

[5] S. Arora, Z. Li, and A. Panigrahi. Understanding Gradient Descent on the Edge of Stability in Deep Learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 948–1024, 2022.

[6] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *Advances in Neural Information Processing Systems*, pages 773–781, 2013.

[7] D. Barrett and B. Dherin. Implicit Gradient Regularization. In *International Conference on Learning Representations*, 2021.

[8] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

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

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

[11] P. Bougerol and J. Lacroix. *Products of Random Matrices with Applications to Schrödinger Operators.* Birkhäuser, Boston, 1985.

[12] L. Chen and J. Bruna. Beyond the edge of stability via two-step gradient updates. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[13] X. Chen, J. D. Lee, X. T. Tong, and Y. Zhang. Statistical inference for model parameters in stochastic gradient descent. *Ann. Statist.*, 48(1):251–273, 2020.

[14] V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6), Dec. 2013.

[15] V. Chernozhukov, D. Chetverikov, and K. Kato. Detailed proof of nazarov's inequality. *arXiv preprint arXiv:1711.10696*, 2017.

[16] V. Chernozhukov, D. Chetverikov, and K. Kato. Inference on Causal and Structural Parameters Using Many Moment Inequalities. *arXiv:1312.7614v6*, 2018.

[17] J. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar. Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability. In *International Conference on Learning Representations*, 2021.

[18] J. M. Cohen, B. Ghorbani, S. Krishnan, N. Agarwal, S. Medapati, M. Badura, D. Suo, D. Cardoze, Z. Nado, G. E. Dahl, and J. Gilmer. Adaptive Gradient Methods at the Edge of Stability, 2024.

[19] A. Damian, E. Nichani, and J. D. Lee. Self-Stabilization: The Implicit Bias of Gradient Descent at the Edge of Stability. In *ICLR*, 2023.

[20] H. Daneshmand, N. Le Roux, and A. M. Bronstein. Stochastic Gradient Descent as a Dynamical System: Almost Sure Convergence Analysis. *Journal of Machine Learning Research*, 25(1):1–35, 2024.

[21] S. Dereich, R. Graeber, A. Jentzen, and A. Riekert. Asymptotic stability properties and a priori bounds for Adam and other gradient descent optimization methods, 2025.

[22] M. Even, S. Pesme, S. Gunasekar, and N. Flammarion. (S)GD over Diagonal Linear Networks: Implicit bias, Large Stepsizes and Edge of Stability. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 29406–29448. Curran Associates, Inc., 2023.

[23] A. Ghosh, S. M. Kwon, R. Wang, S. Ravishankar, and Q. Qu. Learning Dynamics of Deep Matrix Factorization Beyond the Edge of Stability. In *The Thirteenth International Conference on Learning Representations*, 2025.

[24] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

[25] P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Parallelizing stochastic gradient descent for least squares regression: Mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 19(1):1–42, 2018.

[26] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

[27] J. Li, Z. Lou, S. Richter, and W.-B. Wu. The stochastic gradient descent from a nonlinear time series perspective. *preprint*, 2024.

[28] J. Li, Z. Lou, S. Richter, and W. B. Wu. The Stochastic Gradient Descent from a Nonlinear Time Series Perspective. *Manuscript*, TBD(TBD):TBD, 2025.

[29] R. Li, J. Chen, and T. Liang. How does Batch Normalization Help Optimization in Deep Learning: A Look from Stability Analysis. *Transactions on Machine Learning Research*, 2023.

[30] C. Liu, L. Zhu, and M. Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Foundations and Trends in Machine Learning*, 15(1):1–159, 2022.

[31] L. Liu, Z. Zhang, S. Du, and T. Zhao. A Minimalist Example of Edge-of-Stability and Progressive Sharpening, 2025.

[32] P. M. Long and P. L. Bartlett. Sharpness-aware minimization and the edge of stability. *J. Mach. Learn. Res.*, 25(1), 2024.

[33] A. Ly and P. Gong. Optimization on multifractal loss landscapes explains a diverse range of geometrical and dynamical properties of deep learning. *Nature Communications*, 16, 04 2025.

[34] S. Mandt, M. D. Hoffman, and D. M. Blei. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 18(1):4873–4907, 2017.

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

[36] V. I. Oseledets. A multiplicative ergodic theorem. characteristic Lyapunov exponents of dynamical systems. *Trans. Moscow Math. Soc.*, 19:197–231, 1968.

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

[38] V. Roulet, A. Agarwala, J.-B. Grill, G. Swirszcz, M. Blondel, and F. Pedregosa. Stepping on the edge: Curvature aware learning rate tuners. In *Advances in Neural Information Processing Systems*, volume 37, pages 47708–47740, 2024.

[39] D. Ruppert. Efficient estimations from a slowly convergent robbins-monro process. *Technical Report*, 1988.

[40] U. Simsekli, L. Sagun, and M. Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5834–5843, 2019.

[41] M. Song and C. Yun. Trajectory Alignment: Understanding the Edge of Stability Phenomenon via Bifurcation Theory. In *Advances in Neural Information Processing Systems*, volume 36, pages 71632–71682, 2023.

[42] Z. Wang, Z. Li, and J. Li. Analyzing sharpness along GD trajectory: Progressive sharpening and edge of stability. In *Advances in Neural Information Processing Systems*, 2022.

[43] Z. Wei, W. Zhu, and W. B. Wu. Weighted averaged stochastic gradient descent: Asymptotic normality and optimality. *arXiv preprint arXiv:2307.06915*, 2023.

[44] J. Wu, V. Braverman, and J. D. Lee. Implicit Bias of Gradient Descent for Logistic Regression at the Edge of Stability. In *Advances in Neural Information Processing Systems*, volume 36, pages 74229–74256, 2023.

[45] L. Wu, C. Ma, and E. Weinan. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, pages 8279–8288, 2018.

[46] W. B. Wu. Nonlinear system theory: another look at dependence. *Proc. Natl. Acad. Sci. USA*, 102(40):14150–14154, 2005.

[47] W. B. Wu and X. Shao. Limit theorems for iterated random functions. *J. Appl. Probab.*, 41(2):425–436, 2004.

[48] Y. Yang, Z. Hao, X. Zhang, and W. Wang. A Statistical Perspective on Adaptive Learning Rates in Deep Learning. In *International Conference on Learning Representations*, 2023.

[49] S. Zeng and Y. Lei. Stability and Generalization Analysis of Decentralized SGD: Sharper Bounds Beyond Lipschitzness and Smoothness. In *Forty-second International Conference on Machine Learning*, 2025.

[50] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 65(6):107–115, 2022.

[51] J. Zhang, S. P. Karimireddy, A. Veit, S. Kim, S. J. Reddi, S. Kumar, and S. Sra. Why ADAM beats SGD for attention models. In *Advances in Neural Information Processing Systems*, volume 33, pages 22863–22876, 2020.

[52] R. Zhang, J. Wu, L. Lin, and P. L. Bartlett. Minimax Optimal Convergence of Gradient Descent in Logistic Regression via Large and Adaptive Stepsizes, 2025.

[53] W. Zhu, X. Chen, and W. B. Wu. Online covariance matrix estimation in stochastic gradient descent. *J. Amer. Statist. Assoc.*, 118(541):393–404, 2023.

[54] X. Zhu, Z. Wang, X. Wang, M. Zhou, and R. Ge. Understanding Edge-of-Stability Training Dynamics with a Minimalist Example. *ArXiv*, abs/2210.03294, 2022.

[55] D. Zou, Y. Cao, D. Zhou, and Q. Gu. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning*, 110:247–284, 2021.

This appendix is devoted to additional discussion, collection of mathematical arguments and additional simulation results. In particular, in §A we discuss some other approaches to edge-of-stability analysis, as well as the existing gaps in the literature. In §D-§F.1, we provide detailed proofs to our theoretical results.

A Related Literature

The edge of stability phenomenon was first systematically identified by [17] in the context of neural networks, showing empirically that GD trajectories typically operate at the stability threshold. Subsequent work such as [2] extended these insights to simple neuron models, establishing that edge of stability behavior arises even in minimal architectures. These contributions built on a long line of optimization analyses [9, 6] that emphasized the importance of step-size selection and convergence guarantees.

Several papers seek to isolate the mechanisms behind the edge of stability using simplified or tractable models. [5] developed a theoretical framework for GD at the edge of stability, while [54] and [31] employed minimalist examples to clarify the core dynamics. Variants such as diagonal linear networks [22] and two-step updates [12] further illuminate how the phenomenon arises across different formulations. Parallel lines of work have also explored how normalization or regularization mechanisms affect optimization stability, e.g. [29] on batch normalization and [7] on implicit gradient regularization.

Another strand of work interprets the edge of stability through the geometry of the loss landscape. Progressive sharpening along training trajectories was analyzed by [42], while [41] provided a bifurcation-theoretic view. More recent work has refined these ideas via high-dimensional analysis [1], sharpness-aware methods [32], and curvature-aware learning-rate tuning [38]. These developments resonate with broader optimization perspectives on adaptive learning rates [48] and comparisons of adaptive methods with SGD [51].

Beyond stability, the edge of stability has been connected to implicit bias and generalization. For instance, [44] and [19] study logistic regression at the edge of stability, highlighting the implicit regularization induced by GD. Related work considers minimax optimal convergence [52] and generalization in decentralized SGD settings [49]. This complements a broader literature on benign overfitting and generalization in over-parameterized models [8, 50, 55, 30], where stability considerations play a central role.

While the majority of results concern deterministic GD, several papers have begun exploring extensions. [3] revisited the notion of stability under stochastic gradient descent, whereas [18] and [21] examined adaptive and Adam-type methods, respectively. Other directions extend edge of stability analysis to deep linear networks [23] and multi-fractal loss landscapes [33]. These developments connect naturally to classical work on stochastic approximation [45, 25] and continue the trend of relating stochastic dynamics to stability properties.

Despite this growing body of work, the focus has remained predominantly on GD. By contrast, our work develops a systematic analysis of edge of stability in the context of *stochastic gradient descent*, providing a sharper understanding of how stochasticity modifies, stabilizes, or destabilizes the classical GD picture. In this way, we broaden the scope of the edge of stability framework to settings of practical relevance.

B Detailed Assumptions

We expand upon the definitions in the main text for which a concise form was given, giving their full details and the explanation for those forms and their inclusions.

Assumption B.1 (Lipschitz property)**.** We assume that there exists a constant $K_p < \infty$ such that the operator norm of $\nabla_\theta F_\xi^\gamma(\theta)$ adheres to the following property with respect to θ and γ:

$$\mathbb{E}\left[\sup_{(\theta,\gamma,u)\neq(\theta',\gamma',u')\in\mathcal{D}\times\Gamma\times\mathcal{B}} \frac{\left|\left|\nabla_\theta F_{\xi_i}^{\gamma}(\theta)u\right|^p - \left|\nabla_\theta F_{\xi_i}^{\gamma'}(\theta')u'\right|^p\right|}{(|\theta-\theta'|+|\gamma-\gamma'|+|u-u'|)^p}\right] \leqslant K_p$$

Assumption B.2 ($2p$-moment bound, 3.3 fully explained)**.** Fix $p \geqslant 1$. Assume

$$A := \mathbb{E}\left[\sup_{\theta\in\mathcal{D},\gamma\in\Gamma}\sup_{u:|u|=1}\left|\nabla_\theta F_{\xi_i}^{\gamma}(\theta)u\right|^{2p}\right] < \infty.$$

Finite $2p$-th moments of the stochastic gradients strengthen Assumption 2.1 and are standard when deriving concentration inequalities for SGD. Higher-moment assumptions of this type are routinely employed in empirical process theory (see, e.g., [16]) to obtain exponential tail bounds, and they also appear in modern analyses of statistical inference for SGD [13]. In our setting, this condition ensures that deviation inequalities for the empirical expansion parameter hold with high probability, which is essential for establishing nonasymptotic confidence statements about the edge of stability. While stronger than bounded variance, this requirement remains reasonable in practice for smooth models where gradients have sub-Gaussian or sub-exponential tails.

Bounding higher-order derivatives of the stochastic update map is not a universal assumption, but is a reasonable strengthening of smoothness. In the SGD chain, its contraction dynamics are characterized by its the first derivative of the iterate function. In order to control this derivative, we must bound second order derivative behavior of the function, giving rise Assumption 3.2. Although quite strong, this condition is satisfied by many smooth models of practical interest (e.g. generalized linear models), and rules out only highly irregular loss landscapes.

C Extended Simulations

In this section, we expand upon Section 5 to empirically characterize the edge-of-stability region, and assess its optimality. Across a suite of synthetic settings (linear and expectile regression with varying dimension, lag, and data distributions), we estimate the contraction ratio $L_p(\gamma)^{1/p}$ as a function of γ and identify the smallest γ at which contraction fails. The resulting empirical boundary closely matches our theoretical prediction, demonstrating that the proposed "edge of stability" is tight. Taken together, these results validate the theory and provide actionable guidance for selecting constant step sizes that guarantee convergence in practice.

We first demonstrate our result focusing on the following data generating mechanism:

$$Y_i = X_i^\top \theta^* + \epsilon_i,$$

and let $\xi_i = \{X_i, y_i\}_{i\in\mathbb{N}_+}$ denote the observed sequential data and θ^* is the unknown population parameter of interest. We study two convex models: (i) linear regression with squared loss

$$G_1(\theta) = \mathbb{E}_{\xi_i=(X_i,y_i)\sim\Pi_2}(X_i^\top\theta - y_i)^2/2,$$

where $F^{\gamma}_{\xi_i}(\theta)$ takes the following form:

$$F^{\gamma}_{\xi_i}(\theta) = \theta - \gamma X_i(X_i^{\top}\theta - y_i),$$

and (ii) expectile regression with the asymmetric least-square loss

$$G_2(\theta) = \mathbb{E}_{\xi_i=(X_i,y_i)\sim\Pi_2}\left|w - 1_{\{X_i^{\top}\theta - y_i>0\}}\right|(X_i^{\top}\theta - y_i)^2/2,$$

with weight $w \in (0,1)$, and corresponding $F^{\gamma}_{\xi_i}(\theta)$ is given by

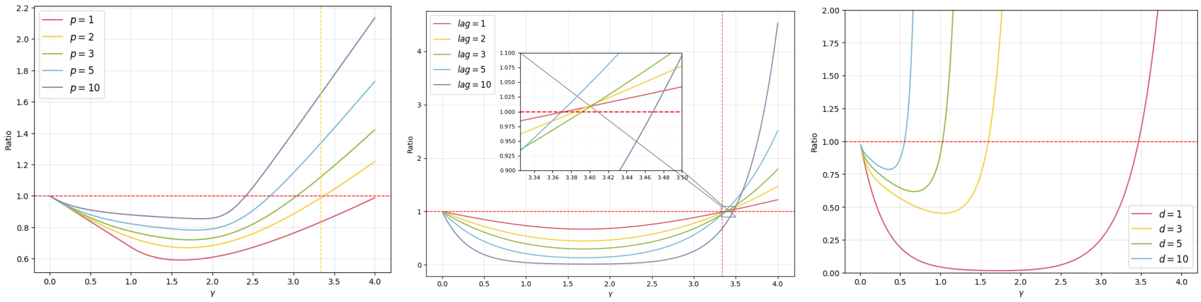
$$F^{\gamma}_{\xi_i}(\theta) = \theta - \left|w - 1_{\{X_i^{\top}\theta - y_i>0\}}\right| X_i(X_i^{\top}\theta - y_i).$$

The feature vector $X \in \mathbb{R}^d$ is drawn either from a Gaussian design $\mathcal{N}(0, I_d)$ or a product Uniform design $\mathrm{Unif}([0,1]^d)$ and the noise ξ is drawn from standard Gaussian distribution, independent of $\{X_i\}_{i\in\mathbb{N}}$. We vary the ambient dimension $d \in \{1,2,3,5,10\}$, the composition lag $l \in \{1,5,10\}$, and the moment index $p \in \{2,4\}$. For linear regression we sweep γ on a grid $\Gamma_{\text{norm}} = \{0.01, 0.02, \dots, 1.00\}$ under the Gaussian design, and for the Uniform design we use $\Gamma_{\text{unif}} = \{0.01, 0.02, \dots, 4.00\}$ to account for the different curvature scales observed in practice.

Across all configurations, the mapping $\gamma \mapsto L_p(\gamma)^{1/p}$ exhibits a pronounced elbow shape, where the estimated ratio initially declines from 1, reaches a minimum, and then reverses, crossing 1 at the stability edge; beyond the crossing it grows rapidly, ultimately diverging. The transition occurs well within the plotted range, so the edge $\hat{\gamma}$ is visually stable and can be localized to a narrow interval.

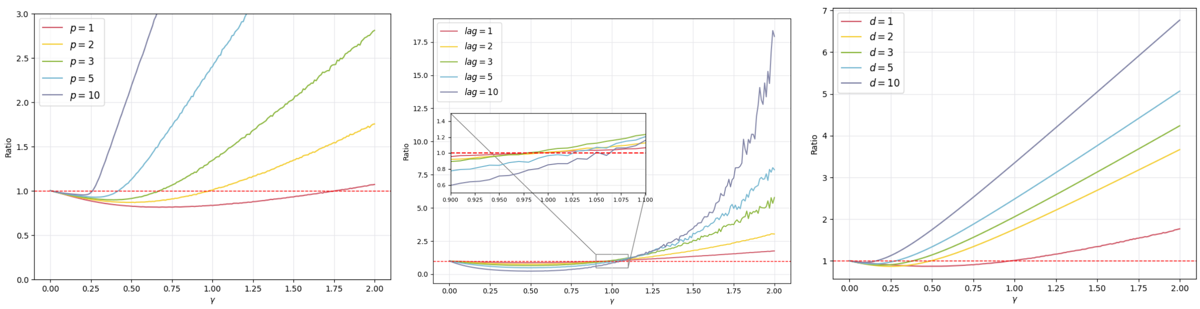
In Figure 3 and Figure 4, subplots (a) demonstrate, that increasing the moment index $p \in \{1,2,3,5,10\}$ shifts the crossing leftward while keeping the minimum shallow. This indicates that heavier emphasis on tail deviations tightens the admissible step-size, which aligns with the result proposed in [28]. Varying the lag ℓ in subplots (b) of Figure 3 and Figure 4 primarily enlarge the edge of stability as lag ℓ increases: for any fixed γ, sub-multiplicative gives $L_p^{\ell}(\gamma) \leqslant L_p(\gamma)^{\ell}$, so increasing ℓ pushes ratios further below on the stable side and further above 1 on the unstable side. Thus the increase of ℓ allows larger γ to ensure the contraction. The dimensional study in subplots (c) of Figure 3 and Figure 4 shows the early contraction of the stable region as d increases. In addition, the empirical edge $\hat{\gamma}_{\ell,n}(p)$ extracted at the yellow curve in subplots (a) and red curve in subplots (b) closely matches the theoretical boundary proposed in Remark 2.6 for $d = 1$ and $\ell = 1$ case, where $\hat{\gamma}_{\ell,n}(2) \approx \frac{2}{3}$ for $X_i \sim \mathcal{N}(0,1)$ and $\hat{\gamma}_{\ell,n}(2) \approx \frac{10}{3}$ for $X_i \sim \mathrm{Unif}[0,1]$. As a conclusion, the results displayed in Figure 3 and Figure 4 validate that the stability set $\gamma : L_p(\gamma) < 1$ is a single interval starting at 0, its boundary is accurately captured by the unique intersection with level 1, and its dependence on p, ℓ, and d follows the theoretical predictions.

Figure 5 shows that expectile regression mirrors the linear case: the edge of stability (the unique crossing of $L_p(\gamma)^{1/p}$ with level 1) decreases as the moment index p increases and increases as the lag ℓ grows. The first trend follows the p-sensitivity of the contraction metric via Hölder's inequality. The second follows from the sub-multiplicativity of the maximal expansion parameter (`MEP`), $L_{p,\ell+k}(\gamma) \leqslant L_{p,\ell}(\gamma)L_{p,k}(\gamma)$, which strengthens contraction on the stable side and steepens growth on the unstable side with right shifting the crossing in γ. The same qualitative dependencies appear for expectile regression for dimension d: the edge moves left as d grows (a smaller stable γ). Taken together, these curves confirm that the qualitative and quantitative dependence of the stability edge on p, ℓ and d persists beyond squared loss.



(a) $d = 1, \ell = 1, p \in \{1, 2, 3, 5, 10\}$ (b) $d = 1, \ell \in \{1, 2, 3, 5, 10\}, p = 2$ (c) $d = \{1, 3, 5, 10\}, \ell = 10, p = 2$

Figure 4: **Linear regression with $X_i \sim \mathrm{Unif}\{[0, 1]^d\}$.** Each panel plots $\widehat{L}_p^{\ell}(\gamma)^{1/p}$ versus the constant step size γ for linear regression. Experimental factors and grids follow the setup marked in subplot labels.



(a) $d = 1,\ \ell = 1,\ p \in \{1, 2, 3, 5, 10\}$ (b) $d = 1, \ell \in \{1, 2, 3, 5, 10\}, p = 2$ (c) $d = \{1, 3, 5, 10\}, \ell = 1, p = 2$

Figure 5: **Expectile regression with $X_i \sim \mathcal{N}\{(0, 1)\}$ and weight $\omega = 0.2$.** Each panel plots $\widehat{L}_p^{\ell}(\gamma)^{1/p}$ versus the constant step size γ for expectile regression and is averaged over 30 experiments. Experimental factors and grids follow the setup marked in subplot labels.

D Proofs of §2

D.1 Proof of Lemma 2.4

Proof. W.l.o.g., we consider the case $\ell = 1$; the case for general $\ell \in \mathbb{N}$ is similar. Note that since $F^{\gamma}_{\xi}(\cdot) \in \mathcal{C}^1$, it follows that

$$\sup_{\theta\in\mathcal{D}} \sup_{u:|u|=1} \frac{1}{n}\sum_{i=1}^{n} \left|\nabla_\theta F^{\gamma}_{\xi_i}(\theta)u\right|^p \leqslant \sup_{\theta\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n} \limsup_{\theta'\to\theta} \frac{\left|F^{\gamma}_{\xi_i}(\theta) - F^{\gamma}_{\xi_i}(\theta')\right|^p}{\left|\theta-\theta'\right|^p} \leqslant \sup_{\theta\neq\theta'\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n} \frac{\left|F^{\gamma}_{\xi_i}(\theta) - F^{\gamma}_{\xi_i}(\theta')\right|^p}{\left|\theta-\theta'\right|^p}. \tag{8}$$

On the other hand, by Jensen's inequality and the convexity of $\mathcal{D}$,

$$\begin{aligned}
\sup_{\theta\neq\theta'\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n} \frac{\left|F^{\gamma}_{\xi_i}(\theta) - F^{\gamma}_{\xi_i}(\theta')\right|^p}{|\theta-\theta'|^p} &= \sup_{\theta\neq\theta'\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n} \frac{\left|\int_0^1 \frac{\partial}{\partial t} F^{\gamma}_{\xi_i}(\theta' + t(\theta-\theta'))dt\right|^p}{|\theta-\theta'|^p} \\
&\leqslant \sup_{\theta\neq\theta'\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n} \frac{\int_0^1 \left|\frac{\partial}{\partial t} F^{\gamma}_{\xi_i}(\theta' + t(\theta-\theta'))\right|^p dt}{|\theta-\theta'|^p} \\
&= \sup_{\theta\neq\theta'\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n} \frac{\int_0^1 \left|\nabla_\theta F^{\gamma}_{\xi_i}(\theta' + t(\theta-\theta'))\,(\theta-\theta')\right|^p dt}{|\theta-\theta'|^p} \\
&\leqslant \sup_{\theta\neq\theta'\in\mathcal{D}, t\in[0,1]} \sup_{u:|u|=1} \frac{1}{n}\sum_{i=1}^{n} \left|\nabla_\theta F^{\gamma}_{\xi_i}(\theta' + t(\theta-\theta'))u\right|^p \\
&\leqslant \sup_{\theta\in\mathcal{D}} \sup_{u:|u|=1} \frac{1}{n}\sum_{i=1}^{n} \left|\nabla_\theta F^{\gamma}_{\xi_i}(\theta)u\right|^p.
\end{aligned} \tag{9}$$

Equations (8) and (9) jointly conclude the proof of (3). In lieu of $\sup_{\theta\neq\theta'\in\mathcal{D}} \mathbb{E}\big[\frac{\left|F^{\gamma}_{\xi_i}(\theta) - F^{\gamma}_{\xi_i}(\theta')\right|^p}{\left|\theta-\theta'\right|^p}\big] < \infty$ from Assumption 3.1, Dominated Convergence Theorem entails Lemma 2.4.

□

D.2 Proof of Proposition 2.7

Proof. We denote $H_\ell(\theta) := F_{i+\ell-1:i}(\theta)$ and $\mathcal{F}_{\ell-1} := \sigma(\xi_i, \ldots, \xi_{i+\ell-1})$. Then:

$$L^{\ell+k}_p(\gamma) = \sup_{\theta\neq\theta'\in\mathcal{D}} \frac{\mathbb{E}\left[|F_{i+\ell+k-1:i}(\theta) - F_{i+\ell+k-1:i}(\theta')|^p\right]}{|\theta-\theta'|^p} = \sup_{\theta\neq\theta'\in\mathcal{D}} \mathbb{E}\left[\frac{|F_{i+\ell+k-1:i}(\theta) - F_{i+\ell+k-1:i}(\theta')|^p}{|\theta-\theta'|^p}\right]$$

$$\begin{aligned}
&= \sup_{\theta\neq\theta'\in\mathcal{D}} \mathbb{E}\left[\frac{|F_{i+\ell+k-1:i+\ell}(H_\ell(\theta)) - F_{i+\ell+k-1:i+\ell}(H_\ell(\theta'))|^p}{|H_\ell(\theta) - H_\ell(\theta')|^p} \cdot \frac{|H_\ell(\theta) - H_\ell(\theta')|^p}{|\theta-\theta'|^p}\right] \\
&= \sup_{\theta\neq\theta'\in\mathcal{D}} \mathbb{E}\left[\mathbb{E}\left[\frac{|F_{i+\ell+k-1:i+\ell}(H_\ell(\theta)) - F_{i+\ell+k-1:i+\ell}(H_\ell(\theta'))|^p}{|H_\ell(\theta) - H_\ell(\theta')|^p} \cdot \frac{|H_\ell(\theta) - H_\ell(\theta')|^p}{|\theta-\theta'|^p} \mid \mathcal{F}_{\ell-1}\right]\right] \\
&= \sup_{\theta\neq\theta'\in\mathcal{D}} \mathbb{E}\left[\mathbb{E}\left[\frac{|F_{i+\ell+k-1:i+\ell}(H_\ell(\theta)) - F_{i+\ell+k-1:i+\ell}(H_\ell(\theta'))|^p}{|H_\ell(\theta) - H_\ell(\theta')|^p} \mid \mathcal{F}_{\ell-1}\right] \cdot \frac{|H_\ell(\theta) - H_\ell(\theta')|^p}{|\theta-\theta'|^p}\right].
\end{aligned}$$

Conditionally on $\mathcal{F}_{\ell-1}$, $F_{i+\ell+k-1:i+\ell}$ is driven by k new i. i. d. innovations which are independent of $\mathcal{F}_{\ell-1}$. Therefore we deduce that:

$$\mathbb{E}\left[\frac{|F_{i+\ell+k-1:i+\ell}(H_\ell(\theta)) - F_{i+\ell+k-1:i+\ell}(H_\ell(\theta'))|^p}{|H_\ell(\theta) - H_\ell(\theta')|^p} \mid \mathcal{F}_{\ell-1}\right] \leqslant L_p^k(\gamma).$$

Therefore:

$$L_p^{\ell+k}(\gamma) \leqslant \sup_{\theta\neq\theta'\in\mathcal{D}} \mathbb{E}\left[L_p^k(\gamma)\cdot\frac{|H_\ell(\theta) - H_\ell(\theta')|^p}{|\theta-\theta'|^p}\right] = L_p^k(\gamma)\cdot \sup_{\theta\neq\theta'\in\mathcal{D}} \frac{|H_\ell(\theta) - H_\ell(\theta')|^p}{|\theta-\theta'|^p} = L_p^k(\gamma)\cdot L_p^\ell(\gamma).$$

□

E Proofs of §3

Before we proceed to the key arguments behind the theoretical results of §3, it is instrumental to introduce a key result that serves as the backbone of our arguments. This result originate from [16], and serves as sharp probabilistic controls on the fluctuations of empirical sums indexed by high-dimensional parameter sets. We restate it here in a form adapted to our setting.

Lemma E.1. *Let $X_1,\dots,X_n \in \mathbb{R}^p$ be independent random vectors with $p \geqslant 2$. Define $M := \max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}|X_{ij}|$ and $\sigma^2 := \max_{1\leqslant j\leqslant p}\sum_{i=1}^n \mathbb{E}[X_{ij}^2]$. Then:*

$$\mathbb{E}\left[\max_{1\leqslant j\leqslant p}\left|\sum_{i=1}^n (X_{ij} - \mathbb{E}[X_{ij}])\right|\right] \leqslant K(\sigma\sqrt{\log p} + \sqrt{\mathbb{E}[M^2]}\log p),$$

where $K > 0$ is a universal constant.

This lemma complements the previous one by providing an expectation bound for the same maximal deviation and quantifies the typical size of the deviation, showing that it scales as $O\big(\sqrt{\log p}\big)$ up to constants depending on variance and maximal moments. In summary, it provides the empirical process tools that underpin our general moment bound in Theorem 3.5. We note that the for the sake of brevity, the results are proved for $\ell = 1$; the general ℓ-cases follow by a simple conditional argument akin to Proposition 2.7.

E.1 Proof of Theorem 3.5

The key idea of Theorem 3.5 is to discretize the set Φ with suitably selected grid, before applying Lemma E.1 to control the deviations of functions evaluated on those grid-points. This grid is carefully chosen to have appropriate packing radius, that allows us to move seamlessly into the compact set Φ while maintaining the rate derived on the grid-points. We formalize this ideas through a novel technique leveraging ε-nets.

Proof. Let $N := n^c$ for some $c > p/2$. For a given $\varphi \in \Phi$, we denote $\lfloor\varphi\rfloor_N := \frac{1}{N}\left(\lfloor N\varphi^1\rfloor,\dots,\lfloor N\varphi^d\rfloor\right)$, with φ^k being the kth coordinate of φ. Then, by compactness and convexity of Φ, $\mathcal{N} :=$

$\{\lfloor\varphi\rfloor_N \mid \varphi \in \Phi\}$ is a δ_n-net for Φ, where $\delta := \delta_n \leqslant L_\Phi n^{-c}$ for some constant $L_\Phi > 0$ that depends only on Φ. Enumerate its elements as $\{\varphi_1, \ldots, \varphi_J\}$ and observe $J \leqslant L_\Phi \cdot N^d$. Recall $A_{\Phi,p}$ defined in Theorem 3.5, and set $X_{ij} := |X_i(\varphi_j)|^p$. Clearly, with $\sigma^2 := \max_{1\leqslant j\leqslant J} \sum_{i=1}^n \mathbb{E}\left[X_{ij}^2\right]$, we obtain, via (7),

$$\sigma^2 \leqslant n\mathbb{E}\left[\sup_{\varphi\in\Phi} |X_i(\varphi)|^{2p}\right] = n \cdot A_{\Phi,p}. \tag{10}$$

On the other hand, letting $M^2 := \max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant J} |X_{ij}|^2$, it follows

$$\mathbb{E}\left[M^2\right] = \mathbb{E}\left[\max_{1\leqslant i\leqslant n} \sup_{\varphi\in\Phi} |X_i(\varphi)|^2\right] \leqslant \sum_{i=1}^n \mathbb{E}\left[\sup_{\varphi\in\Phi} |X_i(\varphi)|^2\right] = n \cdot A_{\Phi,p}. \tag{11}$$

In view of (10) and (11), Lemma E.1 entails

$$\begin{aligned}\mathbb{E}\left[\max_{1\leqslant j\leqslant J}\left|\sum_{i=1}^n (X_{ij} - \mathbb{E}[X_{ij}])\right|\right] &\leqslant K\left(\sigma\sqrt{\log J} + \sqrt{\mathbb{E}[M^2]}\log J\right)\\ &= K\left(\sqrt{n\cdot A_{\Phi,p}}\sqrt{\log L_\Phi + cd\log n} + \sqrt{n\cdot A_{\Phi,p}}(\log L_\Phi + cd\log n)\right)\\ &\leqslant B\cdot\sqrt{n}\log n, \qquad (12)\end{aligned}$$

where $K > 0$ is a universal constant and $B > 0$ depends only on $A_{\Phi,p}$, c and d. With this necessary derivations taken care of, we proceed towards the main arguments. By definition, $|\varphi - \lfloor\varphi\rfloor_N| < \delta$. Recall $S_{n,p}(\cdot)$ from the statement of Theorem 3.5. Note that

$$\begin{aligned}&\mathbb{E}\big[\sup_{\varphi\in\phi}\big|S_{n,p}(\varphi) - \mathbb{E}\big[S_{n,p}(\varphi)\big]\big|\big]\\ &\leqslant \mathbb{E}\big[\max_{1\leqslant j\leqslant J}\big|S_{n,p}(\lfloor\varphi\rfloor_N) - \mathbb{E}\big[S_{n,p}(\lfloor\varphi\rfloor_N)\big]\big|\big] + \mathbb{E}\big[\sup_{\varphi\in\Phi}|S_{n,p}(\varphi) - S_{n,p}(\lfloor\varphi\rfloor_N)|\big] + \sup_{\phi\in\Phi}\big|\mathbb{E}\big[S_{n,p}(\varphi)\big] - \mathbb{E}\big[S_{n,p}(\lfloor\varphi\rfloor_N)\big]\big|\\ &:= T_1 + T_2 + T_3. \qquad (13)\end{aligned}$$

We tackle (13) one-by-one. Equation (12) instructs that $T_1 = O\left(\sqrt{n}\log n\right)$. Next, moving on to T_2, we observe that

$$\begin{aligned}\mathbb{E}\left[\sup_{\varphi\in\Phi}|S_{n,p}(\varphi) - S_{n,p}(\lfloor\varphi\rfloor_N)|\right] &\leqslant n\cdot\mathbb{E}\left[\sup_{\varphi\in\Phi}|X_i^p(\varphi) - X_i^p(\lfloor\varphi\rfloor_N)|\right]\\ &\leqslant np\cdot\mathbb{E}\left[2\sup_{\varphi\in\Phi}|X_i(\varphi)|^{p-1}\cdot\sup_{\varphi\in\Phi}|X_i(\varphi) - X_i(\lfloor\varphi\rfloor_N)|\right] \qquad (14)\\ &\leqslant 2np\left(\mathbb{E}\left[\sup_{\varphi\in\Phi}|X_i(\varphi)|^p\right]\right)^{\frac{p-1}{p}}\left(\mathbb{E}\left[\sup_{\varphi\in\Phi}|X_i(\varphi) - X_i(\lfloor\varphi\rfloor_N)|^p\right]\right)^{\frac{1}{p}} \qquad (15)\\ &\leqslant 2np\sqrt{A_{\Phi,p}}^{\frac{p-1}{p}}(K_\Phi\delta)^{\frac{1}{p}} = O(n\cdot\delta^{-c/p}) = O(n^{1-c/p}), \qquad (16)\end{aligned}$$

where, (14) follows due to the elementary inequality $|\,|a|^p - |b|^p\,| \leqslant p\left(|a|^{p-1} + |b|^{p-1}\right)\cdot|a-b|$, for $p \geqslant 1$, $a, b \in \mathbb{R}$; (15) involves an application of Hölder's inequality, and finally, (16) invokes (6) and (7). Note that, trivially $T_3 \leqslant T_2$. Therefore, (13), along with $\delta = O(n^{-c})$ with $c > p/2$, begets,

$$\mathbb{E}\left[\sup_{\varphi\in\Phi}|S_{n,p}(\varphi) - \mathbb{E}[S_{n,p}(\varphi)]|\right] \lesssim \cdot\sqrt{n}\log n + n^{1-c/p} = O\left(\sqrt{n}\log n\right),$$

where $\lesssim$ hides constants pertaining p, d and φ. This completes the proof. $\square$

E.2 Proof of Theorem 3.6

The key idea behind Theorem 3.6 is to express the data-driven `MEP`'s as supremum of random functions, before invoking Theorem 3.5.

Proof. For $\theta \in \mathcal{D}, \gamma \in \Gamma$ and $u \in \mathcal{B}$, denote

$$M_n(\theta, \gamma, u) := \frac{1}{n}\sum_{i=1}^{n}\left|\nabla_\theta F^{\gamma}_{\xi_i}(\theta)u\right|^p, \text{ and, } M(\theta, \gamma, u) := \mathbb{E}\left[\left|\nabla_\theta F^{\gamma}_{\xi_i}(\theta)u\right|^p\right].$$

We start off by establishing

$$\mathbb{E}\left[\sup_{\theta\in\mathcal{D},\gamma\in\Gamma,u\in\mathcal{B}}\left|M_n\left(\theta,\gamma,u\right) - M\left(\theta,\gamma,u\right)\right|\right] = O\left(\frac{\log n}{\sqrt{n}}\right). \tag{17}$$

Observe that $\Phi := \mathcal{D} \times \Gamma \times \mathcal{B}$ is a compact set, and $F^{\gamma}_{\xi_i}(\theta)$ are i. i. d. random functions taking values in $\varphi \in \Phi$. Moreover, Assumptions 3.2 and 3.3 correspond to (6) and (7) respectively. Therefore, a direct application of Theorem 3.5 entails (17). Finally, in lieu of Lemma 2.4, (17) yields

$$\begin{aligned}
\mathbb{E}\left[\sup_{\gamma\in\Gamma}\left|\widehat{L}^{\ell,n}_p(\gamma) - L^{\ell}_p(\gamma)\right|\right] &= \mathbb{E}\left[\sup_{\gamma\in\Gamma}\left|\sup_{\theta\in\mathcal{D}}\sup_{u:|u|=1}\frac{1}{n}\sum_{i=1}^{n}\left|\nabla_\theta F^{\gamma}_{\xi_i}(\theta)u\right|^p - \sup_{\theta\in\mathcal{D}}\sup_{u:|u|=1}\mathbb{E}\left[\nabla_\theta F^{\gamma}_{\xi_i}(\theta)u\right]\right|\right] \\
&\leqslant \frac{1}{n}\mathbb{E}\left[\sup_{\theta\in\mathcal{D},\gamma\in\Gamma,u\in\mathcal{B}}\left|\sum_{i=1}^{n}\left(\left|\nabla_\theta F^{\gamma}_{\xi_i}(\theta)u\right|^p - \mathbb{E}\left[\left|\nabla_\theta F^{\gamma}_{\xi_i}(\theta)u\right|^p\right]\right)\right|\right] \\
&= O\left(\frac{\log n}{\sqrt{n}}\right),
\end{aligned}$$

which completes the proof. □

E.3 Proof of Theorem 3.7

Proof. It is necessary to establish a bound on $\mathbb{P}(\sup_{x\in\Phi}|S_n(x)| \geqslant z)$. Let $N := n^c$ for some $c > p/2$. To this end, for a given $y \in \Phi$, we denote $\lfloor y\rfloor_N := \frac{1}{N}\left(\lfloor Ny^1\rfloor, \ldots, \lfloor Ny^d\rfloor\right)$, with y^k being the kth coordinate of y. Then, by compactness and convexity of Φ, $\mathcal{N} := \{\lfloor y\rfloor_N \mid y \in \Phi\}$ is a δ_n-net for Φ, where $\delta := \delta_n \leqslant L_\Phi n^{-c}$ for some constant $L_\Phi > 0$ that depends only on Φ. Enumerate its elements as $\{y_1, \ldots, y_J\}$ and observe $J \leqslant L_\Phi \cdot N^d$. This allows for the decomposition:

$$T := \sup_{y\in\Phi}|S_n(y)| \leqslant \sup_{y\in\Phi}|S_n(y) - S_n(\lfloor y\rfloor_n)| + \sup_{y\in\Phi}|S_n(\lfloor y\rfloor_n)| := T_1 + T_2.$$

A similar decomposition holds for the Gaussian process:

$$Z := \sup_{y\in\Phi}\left|S^Z_n(y)\right| \leqslant \sup_{y\in\Phi}\left|S^Z_n(y) - S^Z_n(\lfloor y\rfloor_n)\right| + \sup_{y\in\Phi}\left|S^Z_n(\lfloor y\rfloor_n)\right| := Z_1 + Z_2.$$

It is also useful to observe that $Z_2 \leqslant Z + Z_1$.

These decompositions allow for the following, setting $\varepsilon > 0$:

$$
\begin{aligned}
&\sup_{z\in\mathbb{R}} |\mathbb{P}(T \geqslant z) - \mathbb{P}(Z \geqslant z)| \\
&\leqslant \sup_{z\in\mathbb{R}} |\mathbb{P}(T \geqslant z) - \mathbb{P}(T_2 \geqslant z+\varepsilon)| + \sup_{z\in\mathbb{R}} |\mathbb{P}(T_2 \geqslant z+\varepsilon) - \mathbb{P}(Z_2 \geqslant z+\varepsilon)| + \sup_{z\in\mathbb{R}} |\mathbb{P}(Z \geqslant z) - \mathbb{P}(Z_2 \geqslant z+\varepsilon)| \\
&\leqslant \mathbb{P}(T_1 \geqslant \varepsilon) + \sup_{z\in\mathbb{R}} |\mathbb{P}(T_2 \geqslant z) - \mathbb{P}(Z_2 \geqslant z)| + \sup_{z\in\mathbb{R}} |\mathbb{P}(Z \geqslant z) - \mathbb{P}(Z_2 \geqslant z+\varepsilon)| \,. \qquad (18)
\end{aligned}
$$

For the first summand in (18), bound as follows using Markov's inequality:

$$
\begin{aligned}
\mathbb{P}(T_1 > \varepsilon) &= \mathbb{P}(\sup_{y\in\Phi} |S_n(y) - S_n(\lfloor y\rfloor_n)| > \varepsilon) \leqslant \varepsilon^{-1}\mathbb{E}[\sup_{y\in\Phi} |S_n(y) - S_n(\lfloor y\rfloor_n)|] \\
&\leqslant n\varepsilon^{-1}\mathbb{E}[\sup_{y\in\Phi} |X_i(y) - X_i(\lfloor y\rfloor_n)|] \leqslant n\varepsilon^{-1}\sqrt{K_\Phi}\delta_n \lesssim \varepsilon^{-1}n^{1-c}. \qquad (19)
\end{aligned}
$$

For the second summand, it is prudent to fulfill condition (ii) for Corollary 2.1 in [14]. Denote $x_{ij} := X_i(y^j)$. Then for $p \in \{1, 2\}$ in particular, recall Equation (7). Set $C_1 := A_{\Phi,1}$. Nondegeneracy guarantees existence of some $c_1 > 0$ such that:

$$
c_1 \leqslant \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[x_{ij}^2].
$$

Observe via Jensen's inequality that for all $i \in [n]$:

$$
\sup_{j\in[J]} \mathbb{E}[|x_{ij}|^3] \leqslant \sup_{j\in[J]} (\mathbb{E}[|x_{ij}|^4])^{3/4},
$$

so setting $B_n := A_{\Phi,2}^{\frac{1}{4}}$ yields $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[|x_{ij}|^4] \leqslant 2B_n^4,$, guaranteeing condition (E.2). Additionally, observe:

$$
\frac{B_n^4(\log(Jn))^7}{n} \asymp \frac{(\log(n^{cd+1}))^7}{n} = (cd+1)\frac{\log^7 n}{n},
$$

so to fulfill $\frac{B_n^4(\log(Jn))^7}{n} \leqslant C_2 n^{-c_2}$, it is sufficient to choose $c_2 > 1$. Thus, there exist constants $C_3, c_3 > 0$ depending only on C_1, c_1, C_2, c_2 such that:

$$
\sup_{j\in[J]} \left|\mathbb{P}(S_n(y_j) \geqslant z) - \mathbb{P}(S_n^Z(y_j) \geqslant z)\right| \leqslant C_3 n^{-c_3}. \qquad (20)
$$

The third summand is bounded by applying Nazarov's inequality. Denote for a random vector $V \in \mathbb{R}^J$:

$$
\underline{\sigma}(V) := \min_{j\in[J]} \operatorname{Var}(V^j).
$$

Observe for all $z \in \mathbb{R}$ and $\eta > 0$: $\mathbb{P}(Z \geqslant z) \leqslant \mathbb{P}(Z_1 \geqslant \eta) + \mathbb{P}(Z_2 \geqslant z - \eta)$. It follows

$$
\sup_{z\in\mathbb{R}} (\mathbb{P}(Z \geqslant z) - \mathbb{P}(Z_2 \geqslant z+\varepsilon)) \leqslant \mathbb{P}(Z_1 \geqslant \eta) + \sup_{w\in\mathbb{R}} \mathbb{P}(w \leqslant Z_2 \leqslant w+\eta+\varepsilon). \qquad (21)
$$

For the first term in (21), we apply (6) to compute:

$$
\begin{aligned}
\mathbb{E}[Z_1] &\leqslant \sqrt{\mathbb{E}[Z_1^2]} = \sup_{y\in\Phi}\sqrt{\operatorname{Var}(S_n^Z(y) - S_n^Z(\lfloor y\rfloor_n))} = \sqrt{n}\sup_{y\in\Phi}\sqrt{\operatorname{Var}(Z_i(y) - Z_y(\lfloor y\rfloor_n))} \\
&\leqslant \sqrt{n}\mathbb{E}[\sup_{y\in\Phi} |Z_i(y) - Z_i(\lfloor y\rfloor_n)|] \leqslant \sqrt{n}\mathbb{E}[\sup_{y\in\Phi} |X_i(y) - X_i(\lfloor y\rfloor_n)|] \leqslant nK_\Phi\delta_n \asymp n^{\frac{1}{2}-c}. \qquad (22)
\end{aligned}
$$

Equation (22) yields that $\mathbb{P}(Z_1 \geqslant \eta) \lesssim \eta^{-1} n^{\frac{1}{2}-c}$. For the second term in (21), apply Nazarov's inequality [15] to deduce:

$$\sup_{w \in \mathbb{R}} \mathbb{P}(w \leqslant Z_2 \leqslant w + \eta + \varepsilon) \leqslant (\sqrt{2 \log J} + 2)(\eta + \varepsilon) \sup_{y \in \Phi} \underline{\sigma}^{-1}(S_n^Z(\lfloor y \rfloor_n)) \lesssim (\eta + \varepsilon)\sqrt{\frac{\log n}{n}} \tag{23}$$

Plugging (22) and (23) into (21) and setting $\eta = \varepsilon$ delivers:

$$\sup_{z \in \mathbb{R}} \left(\mathbb{P}(Z \geqslant z) - \mathbb{P}(Z_2 \geqslant z + \varepsilon)\right) \lesssim \varepsilon^{-1} n^{\frac{1}{2}-c} + \varepsilon n^{-\frac{1}{2}} \sqrt{\log n}.$$

An analogous derivation exists for $\mathbb{P}(Z_2 \geqslant z + \varepsilon) - \mathbb{P}(Z \geqslant z)$, so

$$\sup_{z \in \mathbb{R}} |\mathbb{P}(Z \geqslant z) - \mathbb{P}(Z_2 \geqslant z + \varepsilon)| \lesssim \varepsilon^{-1} n^{\frac{1}{2}-c} + \varepsilon n^{-\frac{1}{2}} \sqrt{\log n}. \tag{24}$$

Recalling (19), (20) and (24), conclude:

$$\sup_{z \in \mathbb{R}} |\mathbb{P}(T \geqslant z) - \mathbb{P}(Z \geqslant z)| \lesssim \varepsilon^{-1} n^{1-c} + n^{-c_3} + \varepsilon^{-1} n^{\frac{1}{2}-c} + \varepsilon n^{-\frac{1}{2}} \sqrt{\log n}. \tag{25}$$

Choosing $\varepsilon = n^{\frac{3}{4}-\frac{c}{2}} \log^{-\frac{1}{4}} n$, one concludes from (25) that

$$\sup_{z \in \mathbb{R}} |\mathbb{P}(T \geqslant z) - \mathbb{P}(Z \geqslant z)| \lesssim n^{\frac{1}{4}-\frac{c}{2}} \log^{\frac{1}{4}} n + n^{-c_3},$$

which completes the proof. □

F Proofs of §4

F.1 Proof of Theorem 4.4

Proof. We provide the proof for $\ell = 1$. By definition, $\widehat{\gamma}_n(p) \in \Gamma$. Fix some $M > 0$ such that $L_p(\gamma_0) + M < 1$. Therefore, invoking Theorem 3.6, it follows,

$$\mathbb{P}\left(\widehat{L}_p^n(\gamma_0) < 1\right) \geqslant \mathbb{P}\left(|\widehat{L}_p^n(\gamma_0) - L_p(\gamma_0)| < M\right) \to 1 \quad \text{as } n \to \infty. \tag{26}$$

Additionally, suppose $0 < M' < 1$. By the continuity of $L_p(\cdot)$, $L_p(\gamma^+(p)) = 2$, hence, yet another application of Theorem 3.6 entails that

$$\mathbb{P}\left(\widehat{L}_p^n(\gamma^\dagger(p)) > 1\right) \geqslant \mathbb{P}\left(|\widehat{L}_p^n(\gamma^\dagger(p)) - 2| < M'\right) \to 1 \quad \text{as } n \to \infty. \tag{27}$$

In view of continuity of $\widehat{L}_p^n(\cdot)$, equations (26) and (27) combined, yield that

$$\mathbb{P}(\widehat{\gamma}_n(p) \in \operatorname{int}(\Gamma)) \geqslant \mathbb{P}\left(\widehat{L}_p^n(\gamma_0) < 1,\ \widehat{L}_p^n(\gamma^\dagger(p)) > 1\right) \to 1 \quad \text{as } n \to \infty. \tag{28}$$

This completes the proof of our first assertion. We leverage (28) en route to our second assertion. To that end, observe that following from Assumption 3.4, $L_p(\cdot)$ is differentiable at $\gamma(p)$ with its derivative bounded by K_p. So there exists some $K \leqslant K_p$, such that we can use it to write out first order Taylor expansion of $L(\cdot)$ about $\gamma(p)$:

$$L(\gamma) - L(\gamma(p)) = K(\gamma - \gamma(p)) + o(\gamma - \gamma(p)). \tag{29}$$

From Theorem 3.6, it follows given $\varepsilon > 0$ that there exist some $G_\varepsilon > 0$ and $N_\varepsilon > 0$ such that for all $n > N_\varepsilon$:

$$\mathbb{P}\left(\sup_{\gamma\in\Gamma} |L_n(\gamma) - L(\gamma)| > G_\varepsilon \frac{\log n}{\sqrt{n}}\right) \leqslant \varepsilon.$$

If $\widehat{\gamma}_{\ell,n} \in \Gamma$, then following from the continuity of L_n, we have $L_n(\widehat{\gamma}_{\ell,n}) = 1 = L(\gamma_\ell)$. Therefore,

$$\begin{aligned}\mathbb{P}\left(\sup_{\gamma\in\Gamma} |L_n(\gamma) - L(\gamma)| > G_\varepsilon \frac{\log n}{\sqrt{n}}\right) &\geqslant \mathbb{P}\left(\widehat{\gamma}_{\ell,n} \in \Gamma,\, |L_n(\widehat{\gamma}_{\ell,n}) - L(\widehat{\gamma}_{\ell,n})| > G_\varepsilon \frac{\log n}{\sqrt{n}}\right) \\ &\geqslant \mathbb{P}\left(\widehat{\gamma}_{\ell,n} \in \Gamma,\, |\widehat{\gamma}_{\ell,n} - \gamma_\ell| > K_1 \cdot \frac{\log n}{\sqrt{n}}\right), \end{aligned} \tag{30}$$

where, $K_1 := 2\frac{G_\varepsilon}{K}$, and in (30), we invoke (29). Combined with (29), (30) yields

$$\mathbb{P}\left(\widehat{\gamma}_{\ell,n} \in \Gamma,\, |\widehat{\gamma}_{\ell,n} - \gamma_\ell| > K_1 \cdot \frac{\log n}{\sqrt{n}}\right) \leqslant \varepsilon. \tag{31}$$

Equations (28), (31) jointly conclude the proof. □