FROM CONSERVATIVE CERTIFICATES TO ADAPTIVE CAPITAL: FASTER ANYTIME-VALID STOPPING FOR SGD

Anonymous authors
Paper under double-blind review

ABSTRACT

Stochastic gradient descent remains the default optimizer in large-scale learning, yet it is almost always terminated by heuristics such as a preset iteration budget or a plateau in the training loss. These rules provide no statistical certificate. Recent trajectory-aware stopping procedures restore validity, but they typically rely on closed-form envelopes of an exponential e-process family. In finite samples those envelopes are often loose, so certification arrives later than necessary and extra gradient steps are wasted. We introduce Stopping-Optimized e-Portfolios for SGD (SOEP-SGD). Instead of collapsing the e-process family into a conservative closed-form bound, SOEP-SGD keeps the family explicit and treats certified stopping as a capital-allocation problem: statistical mass is placed on the variance scales that actually govern when a valid certificate can be issued. We establish anytime-valid, optional-stopping-safe guarantees for convex SGD, in the form of weighted suboptimality bounds, and for smooth nonconvex SGD, in the form of weighted stationarity bounds. Simulations and CIFAR-10 experiments show substantially earlier certified stopping than existing closed-form certificates, without sacrificing validity.

1 INTRODUCTION

Stochastic gradient descent (SGD) and its variants are the dominant optimization engine behind modern machine learning, from convolutional networks for vision Krizhevsky et al. (2012); He et al. (2016) to transformers for language Vaswani et al. (2017); Brown et al. (2020). As model and dataset sizes have grown, so has the cost of a single training run Kaplan et al. (2020); Hoffmann et al. (2022); Patterson et al. (2021): state-of-the-art models now consume hundreds of GPUs for weeks, so every superfluous iteration is multiplied by the size of the cluster and the length of the schedule. Yet SGD has no principled stopping rule. Practitioners pick an iteration budget, watch a validation curve, or wait for a training-loss plateau, then declare the job done. None of these procedures controls the actual error at termination, and once the stopping decision depends on what has been observed along the trajectory, fixed-horizon guarantees no longer answer the question being asked.

Aolaritei & Jordan (2025) recently provided a rigorous fix. They construct anytime-valid confidence sequences for SGD by exposing a family of exponential e-processes inside the standard convergence proof, and they package these into a closed-form rule we call *Certified-SGD*: stop as soon as the upper confidence bound on the error falls below a target tolerance. The certificate remains valid under continuous monitoring and arbitrary data-dependent stopping; in convex optimization it controls a weighted suboptimality, and in smooth nonconvex optimization it controls a weighted stationarity measure.

In practice, however, the closed-form rule rarely fires. Across standard convex and smooth nonconvex problems we will study later, Certified-SGD often fails to terminate within a generous iteration budget at tolerances where heuristic rules have long declared the trajectory done. When it does fire, it does so well after the actual error has dropped well below the target. The bottleneck is not the principle of anytime-valid stopping; it is the manner in which a continuous family of valid e-processes is compressed into a single closed-form envelope that must hold uniformly over every problem at once.

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.

The central question we therefore address is: **How can we equip SGD with a practical stopping rule that retains a statistical guarantee?**

The same proof leaves room to do better. Each member of the underlying e-process family is indexed by a betting parameter λ, and a mixture of e-processes is itself a valid e-process. The closed-form envelope corresponds to one particular universal mixture—uniformly safe, but blind to the variance scales any specific problem actually visits. A different mixture, chosen to match those scales, yields a sharper envelope while retaining the same anytime guarantee.

We exploit this freedom in SOEP-SGD. We treat the prior over λ as a tunable design object and ask how a unit of statistical capital should be allocated across λ-accounts so that the resulting certificate crosses the tolerance as early as possible. The key quantity has the form

$$B_{\Lambda,\pi}(v,\alpha) = \min_j \left\{ \frac{\log(1/(\alpha\pi_j))}{\lambda_j} + c\lambda_j v \right\}, \tag{1}$$

where v is the intrinsic variance scale, c is a constant fixed by the proof, λ_j is a betting parameter, and π_j is the capital assigned to account j. Increasing π_j sharpens account j, but the validity constraint $\sum_j \pi_j \leq 1$ forces a trade-off; stopping becomes a capital-allocation problem. The portfolio is fixed before the trajectory is observed, so the stopping rule remains anytime-valid and optional-stopping-safe by construction.

We make three main contributions in this work.

- **Stopping-optimized e-portfolios.** We propose SOEP-SGD, a certified stopping rule that retains the SGD decompositions of Aolaritei & Jordan (2025) but replaces their universal closed-form envelope with a portfolio of λ-accounts whose weights are optimized for early stopping. A meta-portfolio construction further combines multiple designs while preserving validity.
- **Theoretical guarantees for convex and nonconvex SGD.** We prove anytime-valid, optional-stopping-safe certificates for projected convex SGD (weighted suboptimality) and for smooth nonconvex SGD (weighted stationarity), and characterize the design space available to the practitioner.
- **Empirical validation across regimes.** We evaluate SOEP-SGD on synthetic convex and smooth nonconvex problems, and with ResNet-18 and Vision Transformers on CIFAR-10, FashionMNIST, MNIST, and STL-10. Across every benchmark, SOEP-SGD stops substantially earlier than the closed-form Certified-SGD baseline, with no observed certificate violation.

2 Related Work

SGD, stochastic approximation, and adaptive optimizers. Stochastic approximation begins with Robbins & Monro (1951) and is treated systematically by Kushner & Yin (2003) and Borkar (2008). Averaging ideas of Polyak & Juditsky (1992) remain central for stochastic optimization. Finite-time analyses and broader treatments of convex stochastic optimization include Nemirovski et al. (2009), Moulines & Bach (2011), and Bottou et al. (2018). Smooth nonconvex SGD is commonly analyzed through stationarity, with standard rates developed by Ghadimi & Lan (2013; 2016). In modern deep learning LeCun et al. (2015), training often uses adaptive methods such as AdaGrad Duchi et al. (2011), RMSProp Tieleman & Hinton (2012), Adam Kingma & Ba (2015), AMSGrad Reddi et al. (2018), and AdamW Loshchilov & Hutter (2019). These works focus on convergence rates, algorithmic stability, and optimization efficiency, but they generally do not provide stopping certificates that remain valid under repeated inspection of the same SGD trajectory.

Confidence sequences, e-values, and betting. Anytime-valid inference is built on nonnegative supermartingales and Ville's inequality Ville (1939). Sequential tests of Wald (1945), confidence sequences of Darling & Robbins (1967), and subsequent work on methods based on the law of the iterated logarithm and on confidence sequences Robbins (1970); Lai (1976) are early examples. Howard et al. (2021) provide a modern treatment of time-uniform concentration, including Chernoff bounds, mixture boundaries, and stitching constructions. This line of work is closely connected to e-values and e-processes, which provide a convenient language for inference that remains valid under optional stopping Grünwald et al. (2024); Ramdas et al. (2020; 2023); Vovk & Wang (2021). From the

betting perspective of Shafer & Vovk (2001; 2019), nonnegative martingales can be viewed as capital processes, and their growth reflects accumulated statistical evidence. Waudby-Smith & Ramdas (2024) show that this perspective can lead to sharp confidence sequences for bounded means through data-adaptive betting strategies. Our setting differs from these mean-estimation problems because SGD certificates are driven by optimization decompositions and sub-Gaussian gradient noise. The common lesson, however, is that the design of the underlying e-process family can strongly affect the practical sharpness of the resulting certificate.

Closest work. The closest work is Aolaritei & Jordan (2025), who establish anytime-valid SGD stopping rules for convex suboptimality and smooth nonconvex stationarity, henceforth referred to as Certified-SGD. We adopt their optimization decompositions but optimize the probability layer. Related top-conference work studies anytime-valid guarantees and stopping ideas in other machine-learning settings. Karampatziakis et al. (2021) develop confidence sequences for off-policy evaluation that remain valid at arbitrary stopping times. Xu et al. (2024) construct anytime-valid risk-controlling prediction sets under active labeling. Related early-stopping methods also seek computational savings in modern training systems Yuan et al. (2025), but they are not designed to provide optional-stopping-valid optimization certificates.

Table 1: Comparison of stopping rules for SGD on three desirable properties. *Optional continuation*: the user may continue training past a peek and re-evaluate the certificate without invalidating coverage. *Anytime validity*: the probabilistic bound holds simultaneously at every iteration t, so any data-dependent stopping decision is safe. *Low computational cost*: the certificate fires within practical iteration budgets on typical problems.

Method	Optional continuation	Anytime validity	Low computational cost
Fixed-batch	×	×	✓
Plateau heuristic	×	×	✓
Cert-SGD Aolaritei & Jordan (2025)	✓	✓	×
SOEP-SGD (ours)	✓	✓	✓

3 PRELIMINARIES

Let $\mathcal{F}_t$ denote the natural filtration generated by the iterates and stochastic gradients. We take x_1 to be deterministic or $\mathcal{F}_0$-measurable, and write $\mathcal{F}_t = \sigma(x_1, g_1, \dots, g_t)$, up to any external randomness used by the algorithm. Thus x_t and any predictable stepsize η_t are $\mathcal{F}_{t-1}$-measurable, while g_t is observed at time t. Throughout, we assume $\eta_t \geq 0$ and $S_t > 0$ at the times at which a certificate is evaluated. We use

$$S_t = \sum_{s=1}^{t} \eta_s, \qquad V_t = \sum_{s=1}^{t} \eta_s^2, \qquad C_t = \sum_{s=1}^{t} \eta_s^2 \|g_s\|^2. \tag{2}$$

3.1 CONVEX PROJECTED SGD

Projected SGD is

$$x_{t+1} = \Pi_{\mathcal{X}}(x_t - \eta_t g_t), \tag{3}$$

where $\mathcal{X} \subset \mathbb{R}^d$ is closed and convex. Let $x^\star \in \arg\min_{x \in \mathcal{X}} f(x)$, and assume $\|x_t - x^\star\| \leq R_x$ almost surely along the trajectory. The convex performance measure is

$$\bar{F}_t = \frac{1}{S_t} \sum_{s=1}^{t} \eta_s \big(f(x_s) - f(x^\star)\big), \qquad \bar{x}_t = \frac{1}{S_t} \sum_{s=1}^{t} \eta_s x_s. \tag{4}$$

Since $\eta_s \geq 0$ and $\mathcal{X}$ is convex, $\bar{x}_t \in \mathcal{X}$. By convexity, $f(\bar{x}_t) - f(x^\star) \leq \bar{F}_t$.

We assume unbiased stochastic subgradients $\mathbb{E}[g_t \mid \mathcal{F}_{t-1}] = h_t \in \partial f(x_t)$ with h_t predictable, and conditionally sub-Gaussian noise $\xi_t = g_t - h_t$:

$$\mathbb{E}[\exp\{\lambda \langle u, \xi_t \rangle\} \mid \mathcal{F}_{t-1}] \leq \exp\left(\frac{\lambda^2 \sigma^2 \|u\|^2}{2}\right), \quad \forall u \in \mathbb{R}^d,\ \lambda \in \mathbb{R}. \tag{5}$$

3.2 SMOOTH NONCONVEX SGD

For nonconvex SGD, take $x_{t+1} = x_t - \eta_t g_t$ and assume $f \geq 0$ is L-smooth. We assume unbiased stochastic gradients $\mathbb{E}[g_t \mid \mathcal{F}_{t-1}] = \nabla f(x_t)$. With $\xi_t = g_t - \nabla f(x_t)$, we assume the same conditional sub-Gaussian condition as in (5). The target is weighted stationarity:

$$\bar{G}_t = \frac{1}{S_t} \sum_{s=1}^{t} \eta_s \|\nabla f(x_s)\|^2. \tag{6}$$

If $\bar{G}_t \leq \varepsilon$, then $\min_{s \leq t} \|\nabla f(x_s)\|^2 \leq \varepsilon$. For a fully observable certificate we assume an a priori deterministic bound $\|\nabla f(x_t)\| \leq G$ almost surely and use $\sigma^2 G^2 V_t$ as the variance proxy.

3.3 E-PROCESS PORTFOLIOS

A nonnegative supermartingale M_t with $M_0 \leq 1$ satisfies Ville's inequality:

$$\mathbb{P}\left(\sup_{t \geq 0} M_t \geq \frac{1}{\alpha}\right) \leq \alpha. \tag{7}$$

Suppose each $\lambda_j > 0$ gives a nonnegative $\mathcal{F}_t$-adapted supermartingale $E_t(\lambda_j)$ with $E_0(\lambda_j) \leq 1$. For fixed weights $\pi_j > 0$ satisfying $\sum_j \pi_j \leq 1$, the mixture

$$M_t^{\Pi} = \sum_{j \in \mathcal{J}} \pi_j E_t(\lambda_j) \tag{8}$$

is also a nonnegative supermartingale; for a countable portfolio this follows by monotone convergence. The set $\{(\lambda_j, \pi_j)\}$ must be chosen before observing the target SGD trajectory, either by a deterministic design rule or from an independent pilot run; in the latter case, validity is understood conditionally on the pilot.

4 METHODOLOGY

This section presents the stopping-optimized e-portfolio construction. We first recall the closed-form Certified-SGD baseline of Aolaritei & Jordan (2025) in the convex setting. We then expose the underlying λ-indexed e-process family and replace the single closed-form envelope by a fixed portfolio of e-processes. The portfolio is chosen before the target SGD trajectory begins, so the anytime-valid guarantee is preserved. We finally describe the SOEP-π capital-allocation rule and the resulting convex and nonconvex stopping algorithms.

4.1 CERTIFIED-SGD BASELINE

The closed-form convex certificate of Aolaritei & Jordan (2025) can be written as

$$U_t^{\text{CERT-SGD}}(\alpha) = \frac{B_{\text{CERT-SGD}}(v_t, \alpha) + R_x^2 + C_t}{2S_t}, \qquad v_t = \sigma^2 R_x^2 V_t, \tag{9}$$

where

$$v_{t,\text{eff}} = \max\left\{v_t, 2\left(\log\frac{2}{\alpha} + 1\right)\right\}, \tag{10}$$

$$B_{\text{CERT-SGD}}(v_t, \alpha) = 7\sqrt{v_{t,\text{eff}}\left(\log\frac{2}{\alpha} + \log\log(e + v_{t,\text{eff}})\right)}. \tag{11}$$

This certificate is anytime-valid and fully observable but, in small- to medium-variance regimes, conservative; SOEP-SGD retains the underlying SGD decomposition and optimizes only the probability layer that converts exponential e-processes into a usable boundary.

Remark 1 (Sources of conservatism in $B_{\text{CERT-SGD}}$)**.** *The closed-form bound $B_{\text{CERT-SGD}}$ is anytime-valid but uniformly conservative. It mixes exponential e-processes over a fixed geometric grid $\lambda_k = e^{-k/2}$ with polynomially decaying weights $w_k \propto (k+1)^{-2}$, allocating capital across all variance scales rather than the scales the trajectory actually visits; the constant 7 in $B_{\text{CERT-SGD}}$ versus the scale-matched $2\sqrt{2} \approx 2.83$ quantifies the resulting worst-case loss, and the iterated-logarithm term $\log\log(e + v_{t,\text{eff}})$ is the additional price of validity over all scales. This persistent gap, which does not vanish as $t \to \infty$, is precisely the slack that* SOEP-SGD *targets.*

4.2 FULL-SCALE AND STOPPING-OPTIMIZED PORTFOLIOS

For a fixed portfolio $\Pi = \{(\lambda_j, \pi_j) : j \in \mathcal{J}\}$, define the convex portfolio boundary

$$B_{\Pi}^{\mathrm{cvx}}(v, \alpha) = \inf_{j \in \mathcal{J}} \left\{ \frac{\log(1/(\alpha\pi_j))}{\lambda_j} + 2\lambda_j v \right\}. \tag{12}$$

The convex certificate is

$$U_t^{\Pi}(\alpha) = \frac{B_{\Pi}^{\mathrm{cvx}}(\sigma^2 R_x^2 V_t, \alpha) + R_x^2 + C_t}{2S_t}. \tag{13}$$

The only validity requirement is that the atoms and weights in Π are fixed before observing the target SGD trajectory. Subject to this constraint, the choice of the grid and the allocation of e-capital affect sharpness but not validity.

A default full-scale portfolio uses $\lambda_j = e^{\delta j}$ for $j = -J, \dots, J$ and

$$\pi_j = \frac{(1 + |j|)^{-p}}{\sum_{r=-J}^{J} (1 + |r|)^{-p}}. \tag{14}$$

This default is simple and robust but may allocate capital to λ values irrelevant for the realized run. The idealized expression $a/\lambda + 2\lambda v$ has minimizer

$$\lambda^\star(v) = \sqrt{\frac{a}{2v}}, \tag{15}$$

so small intrinsic variance requires large λ and vice versa. This motivates choosing both the grid and the weights using variance scales relevant for stopping.

SOEP capital allocation. Given design variances $v_1, \dots, v_n$ with weights $q_1, \dots, q_n$ and the probability simplex $\Delta_m = \{\pi \in \mathbb{R}_+^m : \sum_{j=1}^m \pi_j = 1\}$, SOEP-$\pi$ fixes a λ grid and approximately solves

$$\min_{\pi \in \Delta_m} \sum_{\ell=1}^{n} q_\ell \inf_j \left\{ \frac{\log(1/(\alpha\pi_j))}{\lambda_j} + 2\lambda_j v_\ell \right\}. \tag{16}$$

The objective asks the portfolio to be sharp at the variance values where the stopping rule is expected to operate.

Given current weights, assign each design variance to its active account:

$$j(\ell) = \arg\min_j \left\{ \frac{\log(1/(\alpha\pi_j))}{\lambda_j} + 2\lambda_j v_\ell \right\}. \tag{17}$$

For fixed assignments $A_j = \{\ell : j(\ell) = j\}$, the capital update is

$$\pi_j \leftarrow \frac{\sum_{\ell \in A_j} q_\ell/\lambda_j + \rho}{\sum_r \left(\sum_{\ell \in A_r} q_\ell/\lambda_r + \rho \right)}, \tag{18}$$

where $\rho > 0$ is a small smoothing constant. The resulting portfolio is then frozen before running the target SGD trajectory.

Algorithm 1 SOEP-π: stopping-optimized capital allocation

1: **Input:** error level α, grid $\lambda_1, \dots, \lambda_m$, design variances $v_1, \dots, v_n$, design weights $q_1, \dots, q_n$, smoothing ρ.
2: Initialize $\pi_j \propto (1 + j)^{-2}$ for $j = 1, \dots, m$ and normalize.
3: **for** $r = 1, \dots, R$ **do**
4: Assign v_ℓ to $j(\ell) = \arg\min_j \{\log(1/(\alpha\pi_j))/\lambda_j + 2\lambda_j v_\ell\}$.
5: Update π_j using (18).
6: **end for**
7: **Output:** fixed portfolio $\Pi = \{(\lambda_j, \pi_j)\}_{j=1}^m$ with $\sum_j \pi_j = 1$.

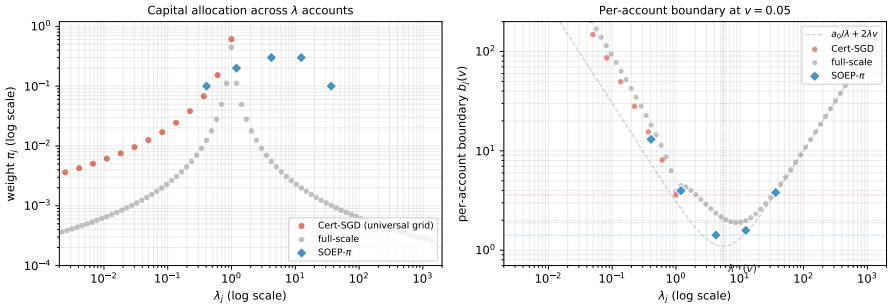

Figure 1: Capital allocation at $v = 0.05$. **Left:** weights π_j for Cert-SGD (red), full-scale (gray), and SOEP-π (blue). **Right:** account boundaries $b_j(v) = \log(1/(\alpha\pi_j))/\lambda_j + 2\lambda_j v$; dotted lines mark portfolio minima. The dashed curve has its minimum at $\lambda^\star(v) = \sqrt{\log(1/\alpha)/(2v)}$.

Remark 2 (Capital allocation as scale matching)**.** *The SOEP-π update* (18) *concentrates capital on the λ_j that win the inner* $\arg\min$ *at the design variances v_ℓ—precisely the scales the trajectory is expected to visit, since each λ_j that wins is close to $\lambda^\star(v_\ell)$. This scale matching never violates $\sum_j \pi_j \le 1$, so validity is preserved by construction.*

Given a fixed portfolio, convex SOEP-SGD stops when the observable upper bound in (13) falls below the target tolerance. Since the bound is anytime-valid, the resulting stopping time is optional-stopping-safe.

Algorithm 2 Certified convex SGD stopping with a fixed e-portfolio

Input: tolerance ε, error level α, radius R_x, noise proxy σ, fixed portfolio Π.
for $t = 1, 2, \ldots$ **do**
  Run projected SGD update (3).
  Update S_t, V_t, C_t, and $v_t = \sigma^2 R_x^2 V_t$.
  Compute U_t^Π from (13).
  if $U_t^\Pi \le \varepsilon$ **then**
    Set $\tau_\varepsilon \leftarrow t$ and **break**.
  end if
end for
Output: certified stopping time τ_ε and averaged iterate $\bar{x}_{\tau_\varepsilon}$.

Theorem 1 (Convex SOEP-SGD certificate)**.** *Under the convex assumptions in Section 3.1, let $\alpha \in (0,1)$ and let $\Pi = \{(\lambda_j, \pi_j) : j \in \mathcal{J}\}$ be any finite or countable portfolio fixed before the current SGD run, with $\lambda_j > 0$, $\pi_j > 0$, and $\sum_{j\in\mathcal{J}} \pi_j \le 1$. Then $U_t^\Pi(\alpha)$ in* (13) *satisfies*

$$\mathbb{P}\left(\forall t \ge 1 : \bar{F}_t \le U_t^\Pi(\alpha)\right) \ge 1 - \alpha. \tag{19}$$

Consequently, the stopping time $\tau_\varepsilon = \inf\{t : U_t^\Pi(\alpha) \le \varepsilon\}$ satisfies

$$\mathbb{P}\left(\tau_\varepsilon < \infty,\ f(\bar{x}_{\tau_\varepsilon}) - f(x^\star) > \varepsilon\right) \le \alpha. \tag{20}$$

Remark 3 (Why SOEP-SGD accelerates without violating the guarantee)**.** SOEP-SGD*'s acceleration comes entirely from capital reallocation, not from any relaxation of probability inequalities: any nonnegative weighted sum of the exponential e-processes with $\sum_j \pi_j \le 1$ remains a nonnegative supermartingale, so Ville's inequality applies unchanged. Because the atoms and weights are fixed before the target trajectory is observed—using deterministic problem parameters or an independent pilot—the prior over λ is a free design parameter that can be adapted to the problem at no statistical cost.*

For smooth nonconvex SGD, the same portfolio idea applies to the stationarity martingale. We use the boundary

$$B_\Pi^{\mathrm{nc}}(v, \alpha) = \inf_j \left\{ \frac{\log(1/(\alpha\pi_j))}{\lambda_j} + \frac{\lambda_j}{2} v \right\}, \tag{21}$$

and the observable certificate

$$W_t^{\Pi}(\alpha) = \frac{B_{\Pi}^{\mathrm{nc}}(\sigma^2 G^2 V_t, \alpha) + f(x_1) + (L/2)C_t}{S_t}. \tag{22}$$

This certificate controls the weighted average stationarity measure $\bar{G}_t$, and hence also certifies that at least one iterate has small gradient norm.

Algorithm 3 Certified smooth nonconvex SGD stopping with a fixed e-portfolio

1: **Input:** tolerance ε, error level α, smoothness L, gradient bound G, noise proxy σ, fixed portfolio Π.
2: **for** $t = 1, 2, \ldots$ **do**
3: Run SGD update $x_{t+1} = x_t - \eta_t g_t$.
4: Update S_t, V_t, C_t, and $\tilde{\Gamma}_t^2 = \sigma^2 G^2 V_t$.
5: Compute W_t^{Π} from (22).
6: **if** $W_t^{\Pi} \le \varepsilon$ **then**
7: Set $\tau_\varepsilon \leftarrow t$ and **break**.
8: **end if**
9: **end for**
10: **Output:** certified stopping time τ_ε and certified weighted-stationarity bound $W_{\tau_\varepsilon}^{\Pi}$.

Theorem 2 (Nonconvex SOEP-SGD certificate)**.** *Assume $f \ge 0$ is L-smooth, $\mathbb{E}[g_t \mid \mathcal{F}_{t-1}] = \nabla f(x_t)$, and $\xi_t = g_t - \nabla f(x_t)$ satisfies the conditional sub-Gaussian condition in* (5)*. Assume also that $\|\nabla f(x_t)\| \le G$ almost surely for all t. Let $\alpha \in (0,1)$ and let $\Pi = \{(\lambda_j, \pi_j) : j \in \mathcal{J}\}$ be any fixed portfolio with $\lambda_j > 0$, $\pi_j > 0$, and $\sum_j \pi_j \le 1$. Define*

$$B_{\Pi}^{\mathrm{nc}}(v, \alpha) = \inf_j \left\{ \frac{\log(1/(\alpha\pi_j))}{\lambda_j} + \frac{\lambda_j}{2} v \right\} \tag{23}$$

and

$$W_t^{\Pi}(\alpha) = \frac{B_{\Pi}^{\mathrm{nc}}(\sigma^2 G^2 V_t, \alpha) + f(x_1) + (L/2)C_t}{S_t}. \tag{24}$$

Then

$$\mathbb{P}\left(\forall t \ge 1 : \bar{G}_t \le W_t^{\Pi}(\alpha)\right) \ge 1 - \alpha. \tag{25}$$

If $\tau_\varepsilon = \inf\{t : W_t^{\Pi}(\alpha) \le \varepsilon\}$, then

$$\mathbb{P}\left(\tau_\varepsilon < \infty,\ \min_{s \le \tau_\varepsilon} \|\nabla f(x_s)\|^2 > \varepsilon\right) \le \alpha. \tag{26}$$

4.3 Validity-preserving design

The next two propositions characterize the design space available to the user: the first quantifies the boundary slack obtained from a log-spaced λ-grid, and the second states formally that the choice of Π never compromises validity.

Proposition 1 (Log-grid oracle inequality)**.** *Let $B_\Lambda(v) = \min_j\{L_0/\lambda_j + c\lambda_j v\}$ with $L_0 > 0$, $c > 0$, and $v > 0$. If a log grid has spacing Δ and contains a point λ_j such that $\lambda_j/\lambda^\star(v) \in [e^{-\Delta/2}, e^{\Delta/2}]$, where $\lambda^\star(v) = \sqrt{L_0/(cv)}$, then*

$$B_\Lambda(v) \le 2\sqrt{cL_0 v}\cosh(\Delta/2). \tag{27}$$

Proposition 2 (SOEP design does not affect validity)**.** *Any deterministic or randomized procedure that outputs a portfolio $\Pi = \{(\lambda_j, \pi_j)\}$ before the current SGD trajectory begins may be used in Theorems 1 and 2. This includes range design, SOEP-π allocation, and meta-portfolios based on deterministic problem parameters or independent pilot runs. If an independent pilot is used, the guarantees are understood conditionally on the pilot, so that the portfolio is fixed for the target trajectory. The optimization of Π changes the boundary width but not the validity proof.*

5 EXPERIMENTS

We evaluate SOEP-SGD on synthetic convex problems, synthetic smooth nonconvex problems, and deep-learning tasks. The methods compared are P (a training- or validation-loss plateau heuristic, no statistical guarantee), CERT-SGD (the closed-form Certified-SGD baseline of Aolaritei & Jordan (2025)), and SOEP-SGD (our optimized e-portfolio certificate). Per-case numerical results are deferred to Appendices C–F.

5.1 CONVEX SIMULATIONS

Setup. We minimize convex quadratics $f(x) = \frac{1}{2}\, x^\top \mathrm{diag}(q)\, x$, q geometrically spaced on $[1/\kappa, 1]$, over the unit Euclidean ball, via projected SGD with stepsize $\eta_t = 0.5/\sqrt{t}$ and additive Gaussian gradient noise of scale σ. Eight (d, κ, σ) instances span $d \in \{20, 100, 200\}$, $\kappa \in \{10, 100, 1000\}$, $\sigma \in \{0.02, 0.05, 0.10\}$, with 100 trajectories per instance, horizon $T = 300{,}000$, and $\alpha = 0.05$. We sweep $\varepsilon_{\mathrm{rel}} \in \{0.25, 1.0, 4.0\}$ with $\varepsilon = \varepsilon_{\mathrm{rel}} \cdot f(x_1)$. The SOEP-SGD portfolio is the default full-scale design of Section 4.2 (deterministic log-spaced λ-grid with polynomial weights), *not* the SOEP-π capital allocation of Algorithm 1; the simpler choice already suffices to expose a large gap against Cert-SGD on these problems. Initialization, the plateau-detector parameters, and the exact portfolio parameters are deferred to Appendix C.1.

Results. On the central setting $(100, 100, 0.05)$, Cert-SGD does not certify within T at $\varepsilon_{\mathrm{rel}} = 0.25$ on any of the 100 trajectories, while SOEP-SGD certifies on every one (median $\tau_{\text{SOEP-SGD}} = 3{,}858$). At $\varepsilon_{\mathrm{rel}} \in \{1.0, 4.0\}$ both methods certify on the same path, with SOEP-SGD between $556\times$ and $568\times$ faster. Across all 2,400 stopping events we observe zero certified-failure events. Per-case tables and the suite-wide log-log certificate grid are in Appendix C. Figure 2 illustrates the gap on one trajectory: $U_t^{\text{SOEP-SGD}}$ tracks the same $1/\sqrt{t}$ decay as $U_t^{\text{CERT-SGD}}$ but sits roughly one decade below it throughout.

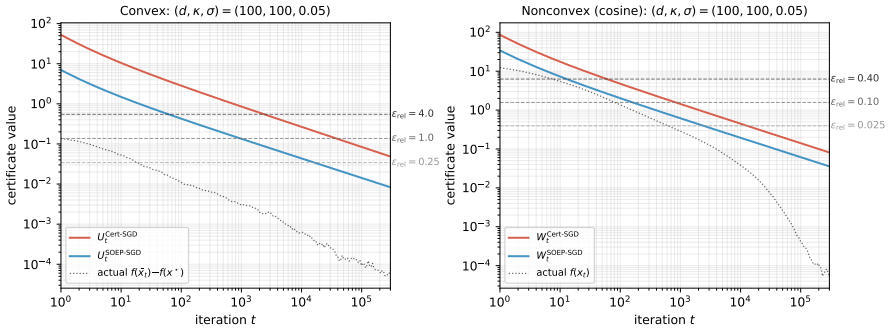

Figure 2: Certificate evolution on a single trajectory at $\alpha = 0.05$ (log-log axes). **Left:** convex case; **right:** nonconvex cosine case; same setting $(d, \kappa, \sigma) = (100, 100, 0.05)$.

5.2 SYNTHETIC NONCONVEX SIMULATIONS

Setup. We minimize four families of smooth nonconvex objectives over $\mathcal{X} = \mathbb{R}^d$ via SGD: *cosine*, *cosine-squared*, *cosine-2*, and *sin-mix*, all parameterized by coefficients a geometrically spaced on $[1/\kappa, 1]$ (explicit formulas and closed-form L, G proxies in Appendix D.1). The protocol matches Section 5.1 except $\varepsilon_{\mathrm{rel}} \in \{0.025, 0.10, 0.40\}$ with $\varepsilon = \varepsilon_{\mathrm{rel}} \cdot \|\nabla f(x_1)\|^2$, and the SOEP-SGD portfolio is again the default full-scale design (with nonconvex parameters J, δ specified in Appendix D.1).

Results. On the central setting (cosine objective, $(100, 100, 0.05)$), both certificates fire on every trajectory at all three tolerances, with SOEP-SGD consistently $\sim 4\times$ earlier (e.g., $\tau_{\text{SOEP-SGD}} = 13{,}600$ vs. $\tau_{\text{CERT-SGD}} = 56{,}560$ at $\varepsilon_{\mathrm{rel}} = 0.025$). The dominant term in both certificates is the deterministic function-drop $f(x_1)/S_t$, not the martingale boundary, which explains the more modest 2–4$\times$ speedup here versus the 10^2–$10^3\times$ regime in the convex setting. Across all 2,400 stopping

events we observe zero certified-failure events; per-case tables and dominant-term diagnostics are in Appendix D.

5.3 CIFAR-10 DEEP LEARNING EXPERIMENT

Setup. We train SmoothCNN on CIFAR-10 for 50 i.i.d. runs on 8 V100 GPUs, using horizon $T = 800{,}000$, $G = 500$, $\sigma = 0.5$, and $\alpha = 0.05$. The SOEP-π portfolio is optimized from the declared proxies and learning-rate schedule before training and then frozen. Architecture, optimization, portfolio, and proxy-diagnostic details are in Appendix E.

Results. Table 2 reports representative stopping outcomes. SOEP-SGD is 9.84–$30.06\times$ faster where both certificates stop and alone certifies $\varepsilon \in \{3, 2, 1.5\}$ within budget. Table 14 gives the full tolerance sweep.

Table 2: CIFAR-10/SmoothCNN: median stopping times and test accuracy over 50 runs. $\tau \geq T$ means no stop within $T = 800{,}000$; its speedup is a lower bound. †Accuracy at T, not at a self-selected stop.

ε	$\tau_{\text{CERT-SGD}}$	$\tau_{\text{SOEP-SGD}}$	Speedup	Acc. CERT-SGD	Acc. SOEP-SGD
Plateau (no guarantee): stop 11,000 (10,000–12,000); test accuracy 0.676.					
5	625,848	**20,820**	30.06×	0.679	0.684
3	≥ 800,000	**47,390**	≥ 16.88×	0.679†	0.682
1.5	≥ 800,000	**358,503**	≥ 2.23×	0.679†	0.679
1, 0.75, 0.5	≥ 800,000	≥ 800,000	—	0.679†	0.679†

5.4 MNIST DEEP LEARNING EXPERIMENTS

On MNIST, ResNet-18 and ViT show 9.4–$15.6\times$ earlier certified stopping at matched tolerances (Table 3). The remaining datasets and diagnostics are in Appendix F.

Table 3: MNIST, nine seeds, $T = 500{,}000$. Top: median stopping steps and speedups at $\varepsilon = 10 \,/\, 7$. Bottom: test accuracy (mean $\pm$ standard deviation); SOEP uses $\varepsilon = 5$.

Model	$\tau_{\text{CERT-SGD}}$	$\tau_{\text{SOEP-SGD}}$	Speedup
ResNet-18	63,822/166,485	**6,738/10,641**	9.5 × /15.6×
ViT	63,901/167,169	**6,781/10,710**	9.4 × /15.6×

Model	SOEP, $\varepsilon = 5$	Plateau	Best validation
ResNet-18	.9917 ± .0006	.9923 ± .0006	.9923 ± .0005
ViT	.9626 ± .0024	.9619 ± .0025	.9640 ± .0020

6 DISCUSSION

SOEP-SGD can certify a target optimization tolerance with fewer SGD steps without changing the training updates. In the CIFAR-10/SmoothCNN experiment at $\varepsilon = 5$, median stopping falls from 625,848 steps with Cert-SGD to 20,820 with SOEP-SGD. Under the stated assumptions, the certificate remains valid when checked after every update. However, the portfolio must be fixed before training. For deep networks, validity also depends on smoothness, noise, and gradient-norm bounds that finite-sample diagnostics cannot fully verify. Gains are smaller when the initial loss or accumulated squared gradients dominate the certificate, since portfolio design does not reduce these terms. The nonconvex certificate controls a weighted average of squared gradient norms, not test accuracy, so validation is still needed for model selection.

AI USE STATEMENT

Generative AI tools were used for language editing, manuscript organization, LaTeX preparation, and feedback on the presentation of assumptions, experimental comparisons, and the interpretation of reported results. The authors are responsible for verifying AI-assisted suggestions against the mathematical derivations, experimental records, and cited sources, and take full responsibility for the final text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

The theoretical assumptions are stated in Section 3, with proofs and additional boundary constructions in Appendices A and B. Algorithms 1–3 specify portfolio allocation and the convex and nonconvex stopping rules. Problem definitions, initialization, noise models, learning-rate schedules, and portfolio settings for the synthetic experiments are given in Appendices C and D. Appendix E describes the SmoothCNN architecture, objective, data split, optimization protocol, computing setup, and portfolio construction for the primary CIFAR-10 study. Appendix F reports the expanded architecture–dataset comparisons, certificate-design ablations, across-seed performance summaries, and robustness and proxy diagnostics.

ETHICS STATEMENT

This work studies optimization certificates using synthetic problems and existing image-classification benchmarks: CIFAR-10, FashionMNIST, MNIST, and STL-10. The experiments described here do not recruit human participants or collect new personal data. Earlier certified stopping may reduce training computation, but a stationarity certificate is not a guarantee of generalization, fairness, privacy, or deployment safety. Its validity depends on the stated assumptions and proxy bounds; finite-sample diagnostics do not establish those assumptions for all future trajectories. Deployment therefore requires task-specific validation and appropriate checks on the data and model.

REFERENCES

L. Aolaritei and M. I. Jordan. Stopping rules for stochastic gradient descent via anytime-valid confidence sequences. 2025. URL https://arxiv.org/abs/2512.13123. arXiv:2512.13123; available at https://arxiv.org/abs/2512.13123.

V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008. URL https://www.cambridge.org/9780521515924. Available at https://www.cambridge.org/9780521515924.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. URL https://doi.org/10.1137/16M1080173. Available at https://doi.org/10.1137/16M1080173.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2005.14165. arXiv:2005.14165; available at https://arxiv.org/abs/2005.14165.

D. A. Darling and H. Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, 1967. URL https://doi.org/10.1073/pnas.58.1.66. Available at https://doi.org/10.1073/pnas.58.1.66.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7):2121–2159, 2011. URL https://www.jmlr.org/papers/v12/duchi11a.html. Available at https://www.jmlr.org/papers/v12/duchi11a.html.

S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. URL https://doi.org/10.1137/120880811. Available at https://doi.org/10.1137/120880811.

S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, 2016. URL https://doi.org/10.1007/s10107-015-0871-8. Available at https://doi.org/10.1007/s10107-015-0871-8.

P. Grünwald, R. de Heide, and W. M. Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 2024. URL https://doi.org/10.1093/jrsssb/qkae011. Available at https://doi.org/10.1093/jrsssb/qkae011.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. URL https://doi.org/10.1109/CVPR.2016.90. Available at https://doi.org/10.1109/CVPR.2016.90.

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2203.15556. arXiv:2203.15556; available at https://arxiv.org/abs/2203.15556.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. URL https://doi.org/10.1214/20-AOS1991. Available at https://doi.org/10.1214/20-AOS1991.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. 2020. URL https://arxiv.org/abs/2001.08361. arXiv:2001.08361; available at https://arxiv.org/abs/2001.08361.

N. Karampatziakis, P. Mineiro, and A. Ramdas. Off-policy confidence sequences. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5301–5310, 2021. URL https://proceedings.mlr.press/v139/karampatziakis21a.html. Available at https://proceedings.mlr.press/v139/karampatziakis21a.html.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. 2015. URL https://arxiv.org/abs/1412.6980. arXiv:1412.6980; available at https://arxiv.org/abs/1412.6980.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 25, 2012. URL https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html. Available at https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

H. J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003. URL https://doi.org/10.1007/b97441. Available at https://doi.org/10.1007/b97441.

T. L. Lai. On confidence sequences. *The Annals of Statistics*, 4(2):265–280, 1976. URL https://doi.org/10.1214/aos/1176343406. Available at https://doi.org/10.1214/aos/1176343406.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. URL https://doi.org/10.1038/nature14539. Available at https://doi.org/10.1038/nature14539.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7. Available at https://openreview.net/forum?id=Bkg6RiCqY7.

E. Moulines and F. R. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, 2011. URL https://www.di.ens.fr/~fbach/gradsto_nips2011. Available at https://www.di.ens.fr/~fbach/gradsto_nips2011.

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. URL https://doi.org/10.1137/070704277. Available at https://doi.org/10.1137/070704277.

D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean. Carbon emissions and large neural network training. 2021. URL https://arxiv.org/abs/2104.10350. arXiv:2104.10350; available at https://arxiv.org/abs/2104.10350.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. URL https://doi.org/10.1137/0330046. Available at https://doi.org/10.1137/0330046.

Lutz Prechelt. Early stopping—but when? In Genevieve B. Orr and Klaus-Robert Müller (eds.), *Neural Networks: Tricks of the Trade*, volume 1524 of *Lecture Notes in Computer Science*, pp. 55–69. Springer, 1998. doi: 10.1007/3-540-49430-8_3.

A. Ramdas, J. Ruf, M. Larsson, and W. M. Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. 2020. URL https://arxiv.org/abs/2009.03167. arXiv:2009.03167; available at https://arxiv.org/abs/2009.03167.

A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. URL https://doi.org/10.1214/23-STS894. Available at https://doi.org/10.1214/23-STS894.

S. J. Reddi, S. Kale, and S. Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ryQu7f-RZ. Available at https://openreview.net/forum?id=ryQu7f-RZ.

H. Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970. URL https://doi.org/10.1214/aoms/1177696786. Available at https://doi.org/10.1214/aoms/1177696786.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. URL https://doi.org/10.1214/aoms/1177729586. Available at https://doi.org/10.1214/aoms/1177729586.

G. Shafer and V. Vovk. *Probability and Finance: It's Only a Game!* Wiley, 2001. URL https://doi.org/10.1002/0471249696. Available at https://doi.org/10.1002/0471249696.

G. Shafer and V. Vovk. *Game-Theoretic Foundations for Probability and Finance*. Wiley, 2019. URL https://doi.org/10.1002/9781118548035. Available at https://doi.org/10.1002/9781118548035.

T. Tieleman and G. Hinton. Lecture 6.5—rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012. URL https://cir.nii.ac.jp/crid/1370017282431050757. Available at https://cir.nii.ac.jp/crid/1370017282431050757.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.03762. arXiv:1706.03762; available at https://arxiv.org/abs/1706.03762.

J. Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, 1939. URL https://eudml.org/doc/192893. Available at https://eudml.org/doc/192893.

V. Vovk and R. Wang. E-values: Calibration, combination, and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021. URL https://doi.org/10.1214/20-AOS2020. Available at https://doi.org/10.1214/20-AOS2020.

A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2): 117–186, 1945. URL https://doi.org/10.1214/aoms/1177731118. Available at https://doi.org/10.1214/aoms/1177731118.

I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. URL https://doi.org/10.1093/jrsssb/qkad009. Available at https://doi.org/10.1093/jrsssb/qkad009.

Z. Xu, N. Karampatziakis, and P. Mineiro. Active, anytime-valid risk controlling prediction sets. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6eb05d8bc6bd7bb6868c64b5802125bd-Abstract-Conference.html. Available at https://proceedings.neurips.cc/paper_files/paper/2024/hash/6eb05d8bc6bd7bb6868c64b5802125bd-Abstract-Conference.html.

S. Yuan, R. Lin, L. Feng, B. Han, and T. Liu. Instance-dependent early stopping. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=P42DbV2nuV. Available at https://openreview.net/forum?id=P42DbV2nuV.

APPENDIX: TABLE OF CONTENTS

A Convex SOEP-SGD Guarantees **15**

A.1 Convex assumptions and certificates 15

A.2 Are the convex assumptions reasonable? 15

A.3 Convex SGD decomposition 16

A.4 Convex SOEP-SGD theorem 16

A.5 Convex SOEP design lemmas 17

B Nonconvex SOEP-SGD Guarantees **18**

B.1 Nonconvex assumptions and certificates 18

B.2 Are the nonconvex assumptions reasonable? 19

B.3 Stationarity decomposition 20

B.4 Nonconvex SOEP-SGD theorem 20

B.5 Nonconvex exact mixture boundary 20

C Additional Simulation Experiments for Convex Functions **20**

C.1 Problem family and full configuration 21

C.2 Convex problems at three tolerance levels 21

D Extended Simulation Results for Non-convex Functions **24**

D.1 Problem family and full configuration 24

D.2 Nonconvex problems at three tolerance levels 25

E Real-Data Details **27**

F Additional Deep-Learning Experiments and Diagnostics **29**

F.1 Stopping efficiency and the role of portfolio optimization 29

F.2 Comparison with validation-based stopping rules 30

F.3 Robustness and assumption calibration 31

A CONVEX SOEP-SGD GUARANTEES

This appendix proves the convex guarantees for SOEP-SGD under the assumptions collected in Section A.1.

A.1 CONVEX ASSUMPTIONS AND CERTIFICATES

We collect all assumptions used in the convex theorem in one place. Assumptions 1–5 are imposed throughout this appendix; the convex SOEP-SGD theorem in Section A.4 is stated under the conjunction of all five.

Assumption 1 (Convex SGD)**.** *The iterates evolve as $x_{t+1} = \Pi_{\mathcal{X}}(x_t - \eta_t g_t)$ with $\mathcal{X} \subset \mathbb{R}^d$ closed and convex, f convex on $\mathcal{X}$, and $x^\star \in \arg\min_{x\in\mathcal{X}} f(x)$.*

Assumption 2 (Predictable stepsizes)**.** *Each η_t is $\mathcal{F}_{t-1}$-measurable, $\eta_t \geq 0$, and certificates are evaluated only at times t with $S_t = \sum_{s=1}^t \eta_s > 0$.*

Assumption 3 (Bounded iterates)**.** *$\|x_t - x^\star\| \leq R_x$ almost surely along the trajectory.*

Assumption 4 (Unbiased subgradients with sub-Gaussian noise)**.** *There exist $h_t \in \partial f(x_t)$ with h_t predictable and $\mathbb{E}[g_t \mid \mathcal{F}_{t-1}] = h_t$, and the noise $\xi_t = g_t - h_t$ satisfies the conditional sub-Gaussian condition*

$$\mathbb{E}[\exp\{\lambda\langle u, \xi_t\rangle\} \mid \mathcal{F}_{t-1}] \leq \exp\left(\frac{\lambda^2\sigma^2\|u\|^2}{2}\right), \qquad \forall u \in \mathbb{R}^d,\ \lambda \in \mathbb{R}. \tag{28}$$

Assumption 5 (Pre-trajectory portfolio)**.** *The portfolio $\Pi = \{(\lambda_j, \pi_j) : j \in \mathcal{J}\}$, with $\lambda_j > 0$, $\pi_j > 0$, and $\sum_j \pi_j \leq 1$, is fixed before the target SGD trajectory begins. If a pilot run is used to design Π, the validity statements are understood conditionally on that pilot.*

Recall the convex performance measure $\bar{F}_t$ and the averaged iterate $\bar{x}_t$:

$$\bar{F}_t = \frac{1}{S_t}\sum_{s=1}^t \eta_s(f(x_s) - f(x^\star)), \qquad \bar{x}_t = \frac{1}{S_t}\sum_{s=1}^t \eta_s x_s. \tag{29}$$

By Assumption 2 together with convexity of $\mathcal{X}$, $\bar{x}_t \in \mathcal{X}$ whenever $S_t > 0$. The portfolio boundary and observable convex certificate are

$$B_\Pi^{\mathrm{cvx}}(v, \alpha) = \inf_j \left\{\frac{\log(1/(\alpha\pi_j))}{\lambda_j} + 2\lambda_j v\right\}, \qquad U_t^\Pi(\alpha) = \frac{B_\Pi^{\mathrm{cvx}}(\sigma^2 R_x^2 V_t, \alpha) + R_x^2 + C_t}{2S_t}. \tag{30}$$

A.2 ARE THE CONVEX ASSUMPTIONS REASONABLE?

The five convex assumptions inherit from standard projected-SGD analysis and from the e-process portfolio construction; we briefly verify each on the synthetic suite of Appendix C.1 and argue that none is more demanding than the corresponding Cert-SGD hypothesis.

- Assumption 1 (projected convex SGD): the quadratic objectives and the unit-ball constraint are convex by construction.
- Assumption 2 (predictable, nonnegative η_t): $\eta_t = 0.5/\sqrt{t}$ is deterministic.
- Assumption 3 (bounded iterates with proxy R_x): with $x^\star = 0$ and projection onto $\mathcal{X} = \{x : \|x\| \leq 1\}$, $\|x_t - x^\star\| = \|x_t\| \leq 1 = R_x$ holds with equality on the boundary and is therefore tight, not slack.
- Assumption 4 (sub-Gaussian noise with proxy σ): isotropic Gaussian noise $\sigma\mathcal{N}(0, I_d)$ satisfies (28) with equality, so the proxy constant is exact.
- Assumption 5 (pre-trajectory portfolio): the full-scale portfolio is determined by deterministic problem parameters before any trajectory begins; no pilot run is used.

In summary, on the convex synthetic suite all five assumptions hold without slack and are not in mutual tension. The bounded-iterate condition is the only one that requires the user to know or estimate R_x in advance; for projected SGD this is automatic from the constraint set, and for unconstrained convex SGD a standard bound-from-the-trajectory device Aolaritei & Jordan (2025) would apply.

A.3 CONVEX SGD DECOMPOSITION

Let

$$Z_t = \|x_t - x^\star\|^2 \tag{31}$$

and define

$$X_t = 2\eta_t(f(x_t) - f(x^\star)) - (Z_t - Z_{t+1}) - \eta_t^2\|g_t\|^2. \tag{32}$$

Projection nonexpansiveness gives

$$Z_{t+1} \le Z_t - 2\eta_t\langle x_t - x^\star, g_t\rangle + \eta_t^2\|g_t\|^2. \tag{33}$$

Substituting (33) into (32),

$$X_t \le 2\eta_t\left(f(x_t) - f(x^\star) - \langle x_t - x^\star, g_t\rangle\right). \tag{34}$$

Write $g_t = h_t + \xi_t$. By convexity,

$$f(x_t) - f(x^\star) \le \langle h_t, x_t - x^\star\rangle. \tag{35}$$

Therefore

$$X_t \le -2\eta_t\langle x_t - x^\star, \xi_t\rangle. \tag{36}$$

Using (28), for every fixed $\lambda \ge 0$,

$$\mathbb{E}[\exp\{\lambda X_t\} \mid \mathcal{F}_{t-1}] \le \mathbb{E}\left[\exp\{-2\lambda\eta_t\langle x_t - x^\star, \xi_t\rangle\} \mid \mathcal{F}_{t-1}\right] \tag{37}$$

$$\le \exp\left(2\lambda^2\sigma^2\eta_t^2 Z_t\right) \le \exp\left(2\lambda^2\sigma^2\eta_t^2 R_x^2\right). \tag{38}$$

A.4 CONVEX SOEP-SGD THEOREM

For each fixed λ_j, define

$$E_t^{\mathrm{cvx}}(\lambda_j) = \exp\left(\lambda_j\sum_{s=1}^{t} X_s - 2\lambda_j^2\sigma^2 R_x^2 V_t\right). \tag{39}$$

By (38), $E_t^{\mathrm{cvx}}(\lambda_j)$ is a nonnegative supermartingale with initial value one. Hence

$$M_t^{\Pi} = \sum_{j\in\mathcal{J}} \pi_j E_t^{\mathrm{cvx}}(\lambda_j) \tag{40}$$

is a nonnegative supermartingale and $M_0^{\Pi} \le 1$.

For a countable portfolio, the supermartingale property follows by applying monotone convergence to the nonnegative partial sums.

Ville's inequality gives

$$\mathbb{P}\left(\sup_{t\ge 0} M_t^{\Pi} \ge \frac{1}{\alpha}\right) \le \alpha. \tag{41}$$

On the complementary event, for all j and all t,

$$\pi_j \exp\left(\lambda_j\sum_{s=1}^{t} X_s - 2\lambda_j^2 v_t\right) < \frac{1}{\alpha}, \qquad v_t = \sigma^2 R_x^2 V_t. \tag{42}$$

Thus

$$\sum_{s=1}^{t} X_s \le \frac{\log(1/(\alpha\pi_j))}{\lambda_j} + 2\lambda_j v_t. \tag{43}$$

Taking the infimum over j yields

$$\sum_{s=1}^{t} X_s \le B_{\Pi}^{\mathrm{cvx}}(v_t, \alpha) \qquad \text{simultaneously for all } t. \tag{44}$$

Now telescope the definition of X_t:

$$2\sum_{s=1}^{t}\eta_s(f(x_s)-f(x^\star)) = \sum_{s=1}^{t}X_s + (Z_1 - Z_{t+1}) + C_t \tag{45}$$

$$\leq \sum_{s=1}^{t}X_s + R_x^2 + C_t. \tag{46}$$

Combining this with (44) and dividing by $2S_t$ proves

$$\mathbb{P}\left(\forall t \geq 1 : \bar{F}_t \leq U_t^{\Pi}(\alpha)\right) \geq 1-\alpha. \tag{47}$$

Finally, let $\tau_\varepsilon = \inf\{t : U_t^{\Pi}(\alpha) \leq \varepsilon\}$. On the event in (47), if $\tau_\varepsilon < \infty$, then

$$f(\bar{x}_{\tau_\varepsilon}) - f(x^\star) \leq \bar{F}_{\tau_\varepsilon} \leq U_{\tau_\varepsilon}^{\Pi}(\alpha) \leq \varepsilon. \tag{48}$$

Therefore

$$\mathbb{P}\left(\tau_\varepsilon < \infty,\ f(\bar{x}_{\tau_\varepsilon}) - f(x^\star) > \varepsilon\right) \leq \alpha. \tag{49}$$

This proves the convex SOEP-SGD guarantee.

A.5 Convex SOEP design lemmas

The preceding guarantee holds for any portfolio satisfying Assumption 5. The lemmas below characterize the design space available to the user.

Lemma 1 (Log-grid oracle inequality)**.** *Let $B_\Lambda(v) = \min_j\{L_0/\lambda_j + c\lambda_j v\}$ with $L_0 > 0$, $c > 0$, and $v > 0$. If a log grid has spacing Δ and contains a point λ_j with $\lambda_j/\lambda^\star(v) \in [e^{-\Delta/2}, e^{\Delta/2}]$, where $\lambda^\star(v) = \sqrt{L_0/(cv)}$, then*

$$B_\Lambda(v) \leq 2\sqrt{cL_0 v}\cosh(\Delta/2). \tag{50}$$

Proof. Let $r = \lambda_j/\lambda^\star(v)$. Then

$$\frac{L_0}{\lambda_j} + c\lambda_j v = \sqrt{cL_0 v}\,(r + r^{-1}). \tag{51}$$

Since $r \in [e^{-\Delta/2}, e^{\Delta/2}]$, $r + r^{-1} \leq e^{\Delta/2} + e^{-\Delta/2} = 2\cosh(\Delta/2)$. □

Lemma 2 (SOEP-π allocation for fixed assignments)**.** *Fix a grid $\lambda_1, \ldots, \lambda_m$ and design points v_ℓ with weights q_ℓ. If assignments $A_j = \{\ell : j(\ell) = j\}$ are fixed, then the minimizer of*

$$\sum_j \left(\sum_{\ell \in A_j}\frac{q_\ell}{\lambda_j}\right)(-\log \pi_j) \quad \textit{over} \quad \sum_j \pi_j = 1 \tag{52}$$

is

$$\pi_j = \frac{b_j}{\sum_r b_r}, \qquad b_j = \sum_{\ell \in A_j}\frac{q_\ell}{\lambda_j}. \tag{53}$$

This formula is understood on the active set with $b_j > 0$. Accounts with $b_j = 0$ may receive zero mass without changing the objective. In the implemented SOEP-π update, the smoothing constant $\rho > 0$ keeps all portfolio weights strictly positive.

Proof. The Lagrangian is $\sum_j b_j(-\log \pi_j) + \mu(\sum_j \pi_j - 1)$. Setting derivatives to zero gives $-b_j/\pi_j + \mu = 0$, hence $\pi_j = b_j/\mu$. Normalizing gives the formula.

If $b_j = 0$ for some account, that account does not contribute to the objective, and any zero or vanishingly small allocation is objective-equivalent before the smoothing step. □

Lemma 3 (Meta-portfolio guarantee)**.** *Suppose portfolios $\Pi_m = \{(\lambda_{m,j}, \pi_{m,j})\}_j$ are mixed with weights $\rho_m > 0$, $\sum_m \rho_m \le 1$. The combined atoms are $(\lambda_{m,j}, \rho_m \pi_{m,j})$. This is a single fixed portfolio. Moreover, for any m,*

$$B_{\mathrm{meta}}(v, \alpha) \le \inf_j \left\{ \frac{\log(1/(\alpha \rho_m \pi_{m,j}))}{\lambda_{m,j}} + c\lambda_{m,j} v \right\}. \tag{54}$$

Lemma 4 (Exact mixture inversion)**.** *For a finite convex portfolio, define $B^{\mathrm{cvx}}_{\mathrm{mix}}(v, \alpha)$ as the unique root b of*

$$\sum_j \pi_j \exp(\lambda_j b - 2\lambda_j^2 v) = 1/\alpha. \tag{55}$$

Then $B^{\mathrm{cvx}}_{\mathrm{mix}}(v, \alpha)$ is a valid anytime boundary for $\sum_{s \le t} X_s$ and satisfies $B^{\mathrm{cvx}}_{\mathrm{mix}}(v, \alpha) \le B^{\mathrm{cvx}}_{\Pi}(v, \alpha)$.

Proof. Because the portfolio is finite, the left-hand side is continuous and strictly increasing in b, tends to zero as $b \to -\infty$, and tends to infinity as $b \to \infty$. Hence the root exists and is unique. On the event $M_t^{\Pi} < 1/\alpha$ for all t, if $A_t = \sum_{s \le t} X_s > B^{\mathrm{cvx}}_{\mathrm{mix}}(v_t, \alpha)$, then

$$\sum_j \pi_j \exp(\lambda_j A_t - 2\lambda_j^2 v_t) > 1/\alpha,$$

which contradicts the definition of M_t^{Π}. Thus the mixture root is a valid anytime boundary. To compare with the atomwise minimum, let

$$B^{\mathrm{cvx}}_{\Pi}(v, \alpha) = \min_j \left\{ \frac{\log(1/(\alpha \pi_j))}{\lambda_j} + 2\lambda_j v \right\}.$$

At this value, at least one individual account reaches level $1/\alpha$ after multiplication by its capital. Therefore the full mixture is at least $1/\alpha$, so the mixture root cannot be larger than the atomwise minimum. □

B Nonconvex SOEP-SGD Guarantees

This appendix proves the smooth nonconvex stationarity guarantees. The proof shares the e-process portfolio step with Appendix A, but the optimization decomposition comes from smoothness rather than projection and convexity.

B.1 Nonconvex assumptions and certificates

The nonconvex theorem replaces convexity and projection with smoothness and a (deterministic) gradient-norm bound. Assumptions 6–10 are imposed throughout this appendix; the nonconvex SOEP-SGD theorem in Section B.4 is stated under the conjunction of all five.

Assumption 6 (Smooth nonconvex SGD)**.** *The iterates evolve as $x_{t+1} = x_t - \eta_t g_t$ with $f \ge 0$ and L-smooth on $\mathbb{R}^d$.*

Assumption 7 (Predictable stepsizes)**.** *Each η_t is $\mathcal{F}_{t-1}$-measurable, $\eta_t \ge 0$, and certificates are evaluated only at times t with $S_t > 0$.*

Assumption 8 (Unbiased gradients with sub-Gaussian noise)**.** $\mathbb{E}[g_t \mid \mathcal{F}_{t-1}] = \nabla f(x_t)$*, and the noise $\xi_t = g_t - \nabla f(x_t)$ satisfies the conditional sub-Gaussian condition* (28)*.*

Assumption 9 (Almost-sure gradient bound)**.** $\|\nabla f(x_t)\| \le G$ *almost surely along the trajectory, with G known in advance.*

Assumption 10 (Pre-trajectory portfolio)**.** *The portfolio $\Pi = \{(\lambda_j, \pi_j) : j \in \mathcal{J}\}$, with $\lambda_j > 0$, $\pi_j > 0$, and $\sum_j \pi_j \le 1$, is fixed before the target SGD trajectory begins. If a pilot run is used to design Π, the validity statements are understood conditionally on that pilot.*

The performance measure and the observable nonconvex certificate are

$$\bar{G}_t = \frac{1}{S_t} \sum_{s=1}^{t} \eta_s \|\nabla f(x_s)\|^2, \qquad B^{\mathrm{nc}}_{\Pi}(v, \alpha) = \inf_j \left\{ \frac{\log(1/(\alpha \pi_j))}{\lambda_j} + \frac{\lambda_j}{2} v \right\}, \tag{56}$$

$$W_t^{\Pi}(\alpha) = \frac{B^{\mathrm{nc}}_{\Pi}(\sigma^2 G^2 V_t, \alpha) + f(x_1) + (L/2) C_t}{S_t}. \tag{57}$$

B.2 ARE THE NONCONVEX ASSUMPTIONS REASONABLE?

The nonconvex assumptions inherit from standard smooth-nonconvex SGD analysis. They are tighter on the synthetic suite than on real-data deep learning; we discuss both settings in turn and argue that on neither is the assumption set more demanding than the corresponding Cert-SGD hypotheses.

Synthetic suite (Appendix D.1).

- Assumption 6 ($f \geq 0$ and L-smooth): each of the four objective families (70)–(73) is globally C^∞ and nonnegative, and the smoothness constant L is the explicit Hessian bound documented per family in Appendix D.1.
- Assumption 7: $\eta_t = 0.5/\sqrt{t}$ is deterministic.
- Assumption 8: isotropic Gaussian noise $\sigma\,\mathcal{N}(0, I_d)$ satisfies (28) with equality.
- Assumption 9 ($\|\nabla f(x_t)\| \leq G$ a.s.): for cosine, cosine-squared, and cosine-2 the closed-form proxies $G = \|a\|_2$ (or $G = (1 + 2\beta)\|a\|_2$) are uniform global bounds because $\sin$ and $\cos$ are bounded by 1. For sin-mix, the proxy $G_{\text{sinmix}} = (R_0 + \beta)\|a\|_2$ uses an implicit box bound $|x_{t,i}| \leq R_0$ that is enforced only at $t = 1$; for unconstrained SGD the iterates can drift beyond this box, so for sin-mix G is best read as a deterministic proxy, with the empirical maximum gradient norm reported as a diagnostic.
- Assumption 10: the portfolio is computed deterministically before any trajectory begins.

On the synthetic suite, four of the five assumptions are tight and the fifth is operative for sin-mix only.

CIFAR-10 deep learning (Appendix E). On a deep CNN trained by mini-batch SGD, several of the constants are operative proxies rather than first-principles bounds. This is the same status they have under Cert-SGD; SOEP-SGD inherits the conservatism of the proxies but converts more of the resulting slack into a faster certificate.

- Assumption 6: `SmoothCNN` is globally C^∞ (Softplus + average pooling), and the empirical loss $\ell(\theta)$ in (76) is bounded below by zero. The smoothness constant $L = 1$ used in the certificate is, however, a declared placeholder; computing the true L for a deep CNN is not analytically tractable.
- Assumption 7: cosine-then-hold is deterministic.
- Assumption 8: mini-batch SGD noise on a finite training set is bounded (and therefore has light tails), but sub-Gaussianity with the specific proxy $\sigma = 0.5$ is again operative rather than verified.
- Assumption 9: $G = 500$ is a deliberately conservative declared bound. The diagnostic `proxy_ok_full_rate`$= 1.000$ (Appendix E) confirms that on every step of every run the empirical gradient norm stays below G, with median empirical maximum $94.1 \ll G$. The slack between the realised gradient norms and the declared G is in fact what the SOEP-π allocation exploits.
- Assumption 10: the portfolio is constructed by Algorithm 1 from the declared schedule and proxies, before any training run, and frozen for the experiment.

Are these assumptions too strong? On synthetic problems the answer is no: the assumptions are essentially tight, with sin-mix the only family relying on a proxy. On CIFAR-10 the smoothness, noise, and gradient-norm constants are operative proxies—the same status they have in Cert-SGD, since both certificates share the underlying smooth-nonconvex SGD decomposition. SOEP-SGD imposes no additional assumption beyond a free design choice (the portfolio), so any setting in which Cert-SGD's nonconvex certificate is considered usable also admits SOEP-SGD's certificate. The only trajectory-time check the user actually needs is the proxy diagnostic $\max_{s \leq t} \|g_s\| \leq G$, which we report alongside every CIFAR-10 run.

B.3 Stationarity decomposition

By L-smoothness,

$$f(x_{t+1}) \le f(x_t) - \eta_t \langle \nabla f(x_t), g_t \rangle + \frac{L}{2}\eta_t^2 \|g_t\|^2. \tag{58}$$

Substituting $g_t = \nabla f(x_t) + \xi_t$,

$$\eta_t \|\nabla f(x_t)\|^2 \le f(x_t) - f(x_{t+1}) + \frac{L}{2}\eta_t^2 \|g_t\|^2 - \eta_t \langle \nabla f(x_t), \xi_t \rangle. \tag{59}$$

Summing from 1 to t and using $f \ge 0$ gives

$$\sum_{s=1}^{t} \eta_s \|\nabla f(x_s)\|^2 \le f(x_1) + \frac{L}{2} C_t + Y_t, \qquad Y_t = -\sum_{s=1}^{t} \eta_s \langle \nabla f(x_s), \xi_s \rangle. \tag{60}$$

B.4 Nonconvex SOEP-SGD theorem

For fixed $\lambda \ge 0$, define

$$E_t^{\mathrm{nc}}(\lambda) = \exp\left(\lambda Y_t - \frac{\lambda^2}{2}\Gamma_t^2\right), \qquad \Gamma_t^2 = \sigma^2 \sum_{s=1}^{t} \eta_s^2 \|\nabla f(x_s)\|^2. \tag{61}$$

Since $\eta_t \nabla f(x_t)$ is $\mathcal{F}_{t-1}$-measurable, the conditional sub-Gaussian assumption gives

$$\mathbb{E}\left[\exp\left(-\lambda \eta_t \langle \nabla f(x_t), \xi_t \rangle\right) \mid \mathcal{F}_{t-1}\right] \le \exp\left(\frac{\lambda^2 \sigma^2 \eta_t^2 \|\nabla f(x_t)\|^2}{2}\right). \tag{62}$$

Thus $E_t^{\mathrm{nc}}(\lambda)$ is a nonnegative supermartingale.

Mixing over the fixed portfolio and using Ville's inequality gives, with probability at least $1 - \alpha$,

$$Y_t \le B_{\Pi}^{\mathrm{nc}}(\Gamma_t^2, \alpha) \qquad \text{for all } t. \tag{63}$$

For a countable nonconvex portfolio, the mixture supermartingale property is again obtained by monotone convergence over nonnegative partial sums.

If $\|\nabla f(x_t)\| \le G$, then $\Gamma_t^2 \le \sigma^2 G^2 V_t$. The boundary is nondecreasing in v, so (63) remains true with $\sigma^2 G^2 V_t$. Combining with (60) and dividing by S_t proves

$$\mathbb{P}\left(\forall t \ge 1 : \bar{G}_t \le W_t^{\Pi}(\alpha)\right) \ge 1 - \alpha. \tag{64}$$

If $\tau_\varepsilon = \inf\{t : W_t^{\Pi}(\alpha) \le \varepsilon\}$, then on this event,

$$\min_{s \le \tau_\varepsilon} \|\nabla f(x_s)\|^2 \le \bar{G}_{\tau_\varepsilon} \le W_{\tau_\varepsilon}^{\Pi}(\alpha) \le \varepsilon. \tag{65}$$

Therefore

$$\mathbb{P}\left(\tau_\varepsilon < \infty, \min_{s \le \tau_\varepsilon} \|\nabla f(x_s)\|^2 > \varepsilon\right) \le \alpha. \tag{66}$$

B.5 Nonconvex exact mixture boundary

For a finite nonconvex portfolio, the exact mixture boundary is the root b of

$$\sum_j \pi_j \exp\left(\lambda_j b - \frac{\lambda_j^2}{2} v\right) = \frac{1}{\alpha}. \tag{67}$$

The same contradiction argument as in Lemma 4 shows that this root is a valid anytime boundary for Y_t and no larger than the atomwise min boundary.

C Additional Simulation Experiments for Convex Functions

This appendix collects detailed results for the convex simulation suites summarized in Section 5.1 of the main text.

C.1 PROBLEM FAMILY AND FULL CONFIGURATION

Objective. Each instance is a convex quadratic

$$f(x) = \tfrac{1}{2}\, x^\top \mathrm{diag}(q)\, x, \qquad q \in \mathbb{R}^d_+, \tag{68}$$

where $q \in \mathbb{R}^d$ has entries geometrically spaced on $[1/\kappa, 1]$: $q_i = \kappa^{(i-1)/(d-1)-1}$ for $i = 1, \ldots, d$, so the condition number of f on $\mathbb{R}^d$ is κ. The constraint set is the Euclidean ball $\mathcal{X} = \{x : \|x\| \le R_x\}$ with $R_x = 1$, and the unique minimizer is $x^\star = 0$ with $f(x^\star) = 0$.

Algorithm. Each trajectory runs projected SGD as in (3), with

$$g_t = \nabla f(x_t) + \sigma\, \zeta_t, \qquad \zeta_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d), \tag{69}$$

so the conditional sub-Gaussian proxy in Assumption 4 is σ. The stepsize is $\eta_t = \eta_0/\sqrt{t}$ with $\eta_0 = 0.5$, and the initialization is $x_1 = 0.9\, R_x\, u_d$ where $u_d \in \mathbb{S}^{d-1}$ is a deterministic unit vector drawn once per case from a fixed seed and reused across the 100 i.i.d. trajectories of that case.

Eight problem instances. Table 4 lists the eight (d, κ, σ) triples that span the suite; the central setting (boldface) is the one whose results are reported inline in Section 5.1. The three tolerance levels swept on each trajectory are $\varepsilon_{\mathrm{rel}} \in \{0.25, 1.0, 4.0\}$, with $\varepsilon = \varepsilon_{\mathrm{rel}} \cdot f(x_1)$. The horizon is $T = 300{,}000$, the confidence level is $\alpha = 0.05$, and each instance is run with 100 i.i.d. trajectories.

Table 4: The eight convex problem instances. The central setting (in boldface) is reported inline in Section 5.1; the remaining seven are reported in Tables 5–7.

Case	d	κ	σ
`quad_d20_k10_sig0.02`	20	10	0.02
`quad_d20_k100_sig0.05`	20	100	0.05
`quad_d20_k1000_sig0.05`	20	1000	0.05
`quad_d100_k10_sig0.05`	100	10	0.05
`quad_d100_k100_sig0.05`	**100**	**100**	**0.05**
`quad_d100_k1000_sig0.05`	100	1000	0.05
`quad_d100_k100_sig0.10`	100	100	0.10
`quad_d200_k100_sig0.05`	200	100	0.05

Portfolio. The SOEP-SGD portfolio used in the convex synthetic suite is the *default full-scale design* of Section 4.2 (not the SOEP-π capital allocation of Algorithm 1, which is reserved for the CIFAR-10 experiment of Section 5.3). Concretely, $\Pi = \{(\lambda_j, \pi_j) : j = -J, \ldots, J\}$ with $\lambda_j = e^{\delta j}$ and $\pi_j \propto (1 + |j|)^{-p}$, with $J = 32$, $\delta = 0.20$, $p = 2$, fixed deterministically before any trajectory runs. The Cert-SGD baseline uses the closed-form bound in (11).

Plateau heuristic. For reference, the plateau detector uses two consecutive sliding windows of length $W = 200$ over the empirical loss $f(x_t)$, checked every 20 iterations after a 1,000-iteration burn-in; a plateau is declared after 5 consecutive non-improving windows at relative tolerance 10^{-4}.

C.2 CONVEX PROBLEMS AT THREE TOLERANCE LEVELS

This appendix provides per-case stopping outcomes for the seven convex quick-suite cases not displayed in Section 5.1 of the main text; the eighth setting, $(d, \kappa, \sigma) = (100, 100, 0.05)$, is the central setting whose headline numbers appear inline in Section 5.1. Each case is run with 100 i.i.d. trajectories, $\alpha = 0.05$, square-root step-size schedule, and maximum horizon 300,000. Plateau, Certified-SGD, and SOEP-SGD are scored on the same trajectory; only the stopping threshold differs across the three tolerance levels $\varepsilon_{\mathrm{rel}} \in \{0.25, 1.0, 4.0\}$, where $\varepsilon = \varepsilon_{\mathrm{rel}} \cdot f(x_1)$. Tables 5–7 give the per-case medians at the three tolerance levels; Table 8 aggregates the empirical certified-failure counts across all eight cases.

Tight tolerance ($\varepsilon_{\text{rel}} = 0.25$). Table 5 extends the central-setting comparison of Section 5.1 to the other seven cases: Certified-SGD again fails to fire on any of the 700 trajectories at this tolerance, and the closed-form bound $U_{\text{CERT-SGD}}$ at the plateau time exceeds the target by 16–52× across these cases. The SOEP-SGD portfolio bound at the same plateau time is at or below the target (0.46–2.38 × ε). On the hardest case (`d100_k1000_sig0.05`, the largest condition number), SOEP-SGD requires roughly six times as many iterations as the heuristic plateau, but it does so with a statistical guarantee that the plateau lacks. The censored speedup column $\tau_{\text{CERT-SGD},c}/\tau_{\text{SOEP-SGD}}$ at this tolerance is a lower bound on the actual ratio and ranges from 19.5 to 840.3 across the seven cases.

Table 5: Convex problems at $\varepsilon_{\text{rel}} = 0.25$, the seven cases not shown in the main text (per-case medians over 100 i.i.d. trajectories). The central setting $(d, \kappa, \sigma) = (100, 100, 0.05)$ is reported in Section 5.1. Em-dashes indicate no stop within 300,000 iterations; the speedup column uses Certified-SGD censored at 300,000. The columns $U_{\text{CERT-SGD}}/\varepsilon$ and $U_{\text{SOEP-SGD}}/\varepsilon$ report the certificate-to-target ratio at the plateau time.

Case (d, κ, σ)	τ_{P}	$\tau_{\text{CERT-SGD}}$	$\tau_{\text{SOEP-SGD}}$	$\tau_{\text{CERT-SGD},c}/\tau_{\text{SOEP-SGD}}$	$U_{\text{CERT-SGD}}/\varepsilon$	$U_{\text{SOEP-SGD}}/\varepsilon$	failures
(100,1000,0.05)	2,360	—	15,420	19.5	52.15	2.38	0/100
(100,100,0.10)	2,110	—	8,103	37.0	20.31	1.81	0/100
(100,10,0.05)	1,580	—	1,059	283.4	18.14	0.83	0/100
(200,100,0.05)	2,800	—	5,030	59.6	22.80	1.30	0/100
(20,1000,0.05)	2,800	—	2,814	106.6	26.57	1.01	0/100
(20,100,0.05)	2,080	—	835	359.3	16.97	0.65	0/100
(20,10,0.02)	1,730	—	357	840.3	16.07	0.46	0/100

Intermediate tolerance ($\varepsilon_{\text{rel}} = 1.0$). At this tolerance, Certified-SGD stops on six of the seven appendix cases; the exception is the highest-condition-number case `d100_k1000_sig0.05`, where Certified-SGD remains censored. Where both methods stop, SOEP-SGD is between 164× and 1,046× faster on the same trajectory (Table 6). The conservativeness $U_{\text{CERT-SGD}}/\varepsilon$ at the plateau time has dropped to 4.0–13.0 but is still well above one, which is why even at this looser tolerance Certified-SGD takes tens of thousands of iterations to fire.

Table 6: Convex problems at $\varepsilon_{\text{rel}} = 1.0$, the seven cases not shown in the main text (per-case medians over 100 i.i.d. trajectories). The central setting $(d, \kappa, \sigma) = (100, 100, 0.05)$ is reported in Section 5.1. Em-dashes indicate no stop within 300,000 iterations. The columns $U_{\text{CERT-SGD}}/\varepsilon$ and $U_{\text{SOEP-SGD}}/\varepsilon$ report the certificate-to-target ratio at the plateau time.

Case (d, κ, σ)	τ_{P}	$\tau_{\text{CERT-SGD}}$	$\tau_{\text{SOEP-SGD}}$	$\tau_{\text{CERT-SGD},c}/\tau_{\text{SOEP-SGD}}$	$U_{\text{CERT-SGD}}/\varepsilon$	$U_{\text{SOEP-SGD}}/\varepsilon$	failures
(100,1000,0.05)	2,360	—	767	391.1	13.04	0.60	0/100
(100,100,0.10)	2,110	54,710	332	164.8	5.08	0.45	0/100
(100,10,0.05)	1,580	31,820	57	558.5	4.53	0.21	0/100
(200,100,0.05)	2,800	90,590	229	395.6	5.70	0.32	0/100
(20,1000,0.05)	2,800	120,900	161	750.7	6.64	0.25	0/100
(20,100,0.05)	2,080	36,580	51	717.3	4.24	0.16	0/100
(20,10,0.02)	1,730	27,180	26	1,045.6	4.02	0.11	0/100

Loose tolerance ($\varepsilon_{\text{rel}} = 4.0$). At this tolerance, Certified-SGD finally stops on every case, but SOEP-SGD still terminates between 226× and 637× earlier on the same trajectory (Table 7). The $U_{\text{CERT-SGD}}/\varepsilon$ ratios are now near unity (1.00–3.26), which is the regime in which the closed-form envelope just barely matches the target; the SOEP-SGD envelope on the same paths is two orders of magnitude tighter (0.03–0.15).

Empirical statistical validity. Both CERT-SGD and SOEP-SGD are anytime-valid and optional-stopping-safe with budgeted failure probability $\alpha = 0.05$ (Theorem 1 for SOEP-SGD; the closed-form analogue of Aolaritei & Jordan (2025) for CERT-SGD). A *certified-failure* event at tolerance ε is one in which the method stops at some time τ but the actual weighted suboptimality $\bar{F}_\tau$ exceeds

Table 7: Convex problems at $\varepsilon_{\text{rel}} = 4.0$, the seven cases not shown in the main text (per-case medians over 100 i.i.d. trajectories). The central setting $(d, \kappa, \sigma) = (100, 100, 0.05)$ is reported in Section 5.1. All methods stop on every case at this tolerance. The columns $U_{\text{CERT-SGD}}/\varepsilon$ and $U_{\text{SOEP-SGD}}/\varepsilon$ report the certificate-to-target ratio at the plateau time.

Case (d, κ, σ)	τ_{P}	$\tau_{\text{CERT-SGD}}$	$\tau_{\text{SOEP-SGD}}$	$\tau_{\text{CERT-SGD}}/\tau_{\text{SOEP-SGD}}$	$U_{\text{CERT-SGD}}/\varepsilon$	$U_{\text{SOEP-SGD}}/\varepsilon$	failures
(100,1000,0.05)	2,360	24,710	41	602.2	3.26	0.15	0/100
(100,100,0.10)	2,110	3,389	15	226.1	1.27	0.11	0/100
(100,10,0.05)	1,580	2,023	5	404.8	1.13	0.05	0/100
(200,100,0.05)	2,800	5,664	12	471.3	1.43	0.08	0/100
(20,1000,0.05)	2,800	7,639	12	636.7	1.66	0.06	0/100
(20,100,0.05)	2,080	2,335	5	467.0	1.06	0.04	0/100
(20,10,0.02)	1,730	1,744	3	581.3	1.00	0.03	0/100

ε. Since trajectories on which Certified-SGD never stops cannot fail by construction, the only meaningful failure-rate denominators are the (method, ε_{rel}) cells in which the method actually triggers. Table 8 reports both the raw failure counts and the rates per realised stop. Across all 3,900 realised stops, no certified failure was observed; the empirical failure rate is well inside the $\alpha = 0.05$ budget at every tolerance.

Table 8: Empirical statistical validity on the convex setting. "Stops" counts the number of trajectory×case events on which the method fired (out of 800 per row). "Failures" counts realised stops at which the actual averaged suboptimality $\bar{F}_\tau$ exceeded ε. The budgeted failure probability is $\alpha = 0.05$ throughout.

ε_{rel}	Method	Stops / 800	Failures	Rate per stop	Rate per trajectory
0.25	CERT-SGD	0	0	—	$0/800 = 0.000$
0.25	SOEP-SGD	800	0	0.000	$0/800 = 0.000$
1.0	CERT-SGD	700	0	0.000	$0/800 = 0.000$
1.0	SOEP-SGD	800	0	0.000	$0/800 = 0.000$
4.0	CERT-SGD	800	0	0.000	$0/800 = 0.000$
4.0	SOEP-SGD	800	0	0.000	$0/800 = 0.000$

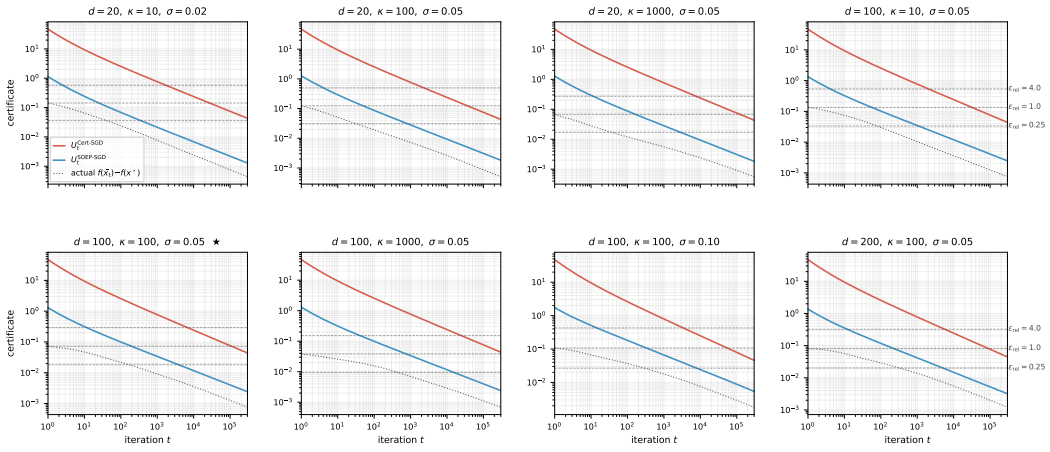

Figure 3: Convex certificate evolution across the suite. One log-log panel per setting (8 settings, central marked ★). On every panel: $U_t^{\text{Cert-SGD}}$ (red), $U_t^{\text{SOEP-SGD}}$ (blue), and the actual averaged suboptimality $f(\bar{x}_t) - f(x^\star)$ (dotted black) on a single trajectory at $\alpha = 0.05$. Horizontal dashed lines are the three tolerance levels $\varepsilon_{\text{rel}} \in \{0.25, 1.0, 4.0\}$. Across all eight settings the SOEP-SGD certificate sits roughly one decade below the Cert-SGD certificate, while the actual suboptimality stays well below both, providing visual confirmation of validity uniform over the suite.

The slack between the realised certificate U_t and the target ε at the stopping time also gives a sense of how tight the methods are. At $\varepsilon_{\mathrm{rel}} = 0.25$, the median actual averaged suboptimality at the SOEP-SGD stop is roughly 4% to 35% of ε (from the $U_{\text{SOEP-SGD}}/\varepsilon$ ratios in Table 5); at $\varepsilon_{\mathrm{rel}} = 4.0$, the ratio falls to 3% to 15%. SOEP-SGD thus triggers with substantial safety margin, even though it is several orders of magnitude faster than the closed-form Certified-SGD baseline on the same trajectory.

D Extended Simulation Results for Non-convex Functions

This appendix collects detailed results for the smooth nonconvex simulation suite summarized in Section 5.2 of the main text.

D.1 Problem family and full configuration

Objectives. Each instance uses one of four smooth nonconvex objective families on $\mathbb{R}^d$, indexed by a coefficient vector $a \in \mathbb{R}^d_+$ with a_i geometrically spaced on $[1/\kappa, 1]$:

$$\text{cosine:} \quad f(x) = \textstyle\sum_i a_i(1 - \cos x_i), \tag{70}$$

$$\text{cosine-squared:} \quad f(x) = \textstyle\sum_i a_i \sin^2 x_i, \tag{71}$$

$$\text{cosine-2:} \quad f(x) = \textstyle\sum_i a_i(1 - \cos x_i) + \beta \sum_i a_i(1 - \cos 2x_i), \tag{72}$$

$$\text{sin-mix:} \quad f(x) = \tfrac{1}{2} \textstyle\sum_i a_i x_i^2 + \beta \sum_i a_i(1 - \cos x_i), \tag{73}$$

where $\beta \geq 0$ is a coupling parameter, fixed per case (see Table 9). All four objectives are nonnegative and globally C^∞.

Smoothness and gradient-norm proxies. The certificate (57) requires deterministic upper bounds on the smoothness constant L and the gradient norm G. Closed-form bounds for the four families, derived directly from the Hessian and gradient expressions, are

$$\begin{aligned} L_{\mathrm{cosine}} &= \max_i a_i, & G_{\mathrm{cosine}} &= \|a\|_2, \\ L_{\mathrm{cossq}} &= 2 \max_i a_i, & G_{\mathrm{cossq}} &= \|a\|_2, \\ L_{\text{cos-2}} &= (1 + 4\beta) \max_i a_i, & G_{\text{cos-2}} &= (1 + 2\beta)\, \|a\|_2, \\ L_{\mathrm{sinmix}} &= (1 + \beta) \max_i a_i, & G_{\mathrm{sinmix}} &= (R_0 + \beta)\, \|a\|_2, \end{aligned} \tag{74}$$

where $R_0 = 2.5$ is the initialization radius defined below. The bound on G_{sinmix} uses $\|x\| \leq R_0\sqrt{d}$ and is interpreted as a deterministic proxy in the spirit of Assumption 9; the empirical maximum gradient norm is well below the proxy on every trajectory we observe (cf. the proxy diagnostic of Section 5.3 for the analogous CIFAR-10 setting).

Algorithm. Each trajectory runs unconstrained SGD with

$$g_t = \nabla f(x_t) + \sigma\, \zeta_t, \qquad \zeta_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d), \qquad \eta_t = \eta_0/\sqrt{t},\ \eta_0 = 0.5, \tag{75}$$

and the initialization is $x_1 \sim \mathrm{Unif}[-R_0, R_0]^d$ with $R_0 = 2.5$, drawn once per case from a fixed seed and reused across the 100 i.i.d. trajectories of that case.

Eight problem instances. Table 9 lists the eight cases. The central setting (boldface) is the one whose results are reported inline in Section 5.2; the remaining seven appear in Tables 10–12. The three tolerance levels swept on each trajectory are $\varepsilon_{\mathrm{rel}} \in \{0.025, 0.10, 0.40\}$ with $\varepsilon = \varepsilon_{\mathrm{rel}} \cdot \|\nabla f(x_1)\|^2$. The horizon is $T = 300{,}000$, the confidence level is $\alpha = 0.05$, and each instance is run with 100 i.i.d. trajectories. The plateau detector matches the convex setting (see Appendix C.1).

Table 9: The eight smooth nonconvex problem instances. The central setting is in boldface; β is the coupling parameter in (72) and (73) (set to zero for the other families).

Case	objective	d	κ	β	σ
`cos_d20_k10_sig0.02`	cosine	20	10	0.00	0.02
`cos_d20_k100_sig0.05`	cosine	20	100	0.00	0.05
`cos_d100_k10_sig0.05`	cosine	100	10	0.00	0.05
`cos_d100_k100_sig0.05`	**cosine**	**100**	**100**	**0.00**	**0.05**
`cos_d100_k100_sig0.10`	cosine	100	100	0.00	0.10
`cos2_d100_k100_sig0.05`	cosine-2	100	100	0.25	0.05
`sinmix_d100_k100_sig0.05`	sin-mix	100	100	0.35	0.05
`cossq_d100_k100_sig0.05`	cosine-squared	100	100	0.00	0.05

Portfolio. As in the convex synthetic suite (Appendix C.1), the SOEP-SGD portfolio used here is the *default full-scale design* of Section 4.2 rather than the SOEP-π capital allocation of Algorithm 1. Concretely, $\Pi = \{(\lambda_j, \pi_j) : j = -J, \ldots, J\}$ with $\lambda_j = e^{\delta j}$ and $\pi_j \propto (1 + |j|)^{-p}$, with $J = 40$, $\delta = 0.18$, $p = 2$, fixed deterministically before any trajectory runs.

D.2 Nonconvex problems at three tolerance levels

This appendix provides per-case stopping outcomes for the seven smooth nonconvex cases not displayed in Section 5.2 of the main text; the eighth setting, $(d, \kappa, \sigma) = (100, 100, 0.05)$ with the cosine objective, is the central setting whose headline numbers appear inline in Section 5.2. Each case is run with 100 i.i.d. trajectories, $\alpha = 0.05$, square-root step-size schedule, and maximum horizon 300,000. Plateau, Cert-SGD, and SOEP-SGD are scored on the same trajectory; only the stopping threshold differs across the three tolerance levels $\varepsilon_{\mathrm{rel}} \in \{0.025, 0.10, 0.40\}$, where $\varepsilon = \varepsilon_{\mathrm{rel}} \cdot \|\nabla f(x_1)\|^2$. Tables 10–12 give the per-case medians at the three tolerance levels; Table 13 aggregates the empirical certified-failure counts across all eight cases.

Tight tolerance ($\varepsilon_{\mathrm{rel}} = 0.025$). At this tolerance Cert-SGD is censored on two of the seven appendix cases ($\cos$:(20,100,0.05) and $\cos$:(20,10,0.02)), where the closed-form bound exceeds the target by $9.2\times$ and $102.7\times$ respectively at the plateau time. SOEP-SGD stops on every trajectory of every case and is between $2.6\times$ and $35.9\times$ faster than the censored Cert-SGD time on the same trajectory (Table 10). The SOEP-SGD plateau-time tightness $W_{\text{SOEP-SGD}}/\varepsilon$ ranges from 0.31 to 2.41 across the seven cases, which contrasts with the 0.50–$12.09 \times \varepsilon$ range of the closed-form Cert-SGD bound.

Table 10: Nonconvex problems at $\varepsilon_{\mathrm{rel}} = 0.025$, the seven cases not shown in the main text (per-case medians over 100 i.i.d. trajectories). The central setting $(d, \kappa, \sigma) = (100, 100, 0.05)$ with the cosine objective is reported in Section 5.2. Em-dashes indicate no stop within 300,000 iterations; the speedup column uses Cert-SGD censored at 300,000. The columns $W_{\text{Cert-SGD}}/\varepsilon$ and $W_{\text{SOEP-SGD}}/\varepsilon$ report the certificate-to-target ratio at the plateau time.

Case	τ_{P}	$\tau_{\text{Cert-SGD}}$	$\tau_{\text{SOEP-SGD}}$	$\tau_{\text{Cert-SGD},c}/\tau_{\text{SOEP-SGD}}$	$W_{\text{Cert-SGD}}/\varepsilon$	$W_{\text{SOEP-SGD}}/\varepsilon$	failures
$\cos$:(100,100,0.10)	24,100	102,400	22,220	4.61	2.06	0.96	0/100
$\cos$:(100,10,0.05)	2,900	33,230	12,540	2.65	3.41	2.08	0/100
$\cos$:(20,100,0.05)	32,600	—	8,346	35.95	3.50	0.51	0/100
$\cos$:(20,10,0.02)	2,920	—	16,920	17.74	12.09	2.41	0/100
$\cos_2$:(100,100,0.05)	36,820	70,810	20,850	3.40	1.39	0.75	0/100
sinmix:(100,100,0.05)	22,710	5,679	2,213	2.57	0.50	0.31	0/100
$\cos^2$:(100,100,0.05)	33,910	111,500	23,380	4.77	1.81	0.83	0/100

Intermediate tolerance ($\varepsilon_{\mathrm{rel}} = 0.10$). At this tolerance, both Cert-SGD and SOEP-SGD stop on every case (Table 11). SOEP-SGD is between $2.55\times$ and $47.7\times$ faster on the same trajectory, with the largest gains on the small-dimension cases where Cert-SGD takes tens of thousands of iterations to fire while SOEP-SGD stops in the low thousands. The closed-form Cert-SGD bound

at the plateau time has dropped to 0.12–$3.02 \times \varepsilon$, indicating that this tolerance level sits near the regime in which the Cert-SGD bound just barely matches the target.

Table 11: Nonconvex problems at $\varepsilon_{\mathrm{rel}} = 0.10$, the seven cases not shown in the main text (per-case medians over 100 i.i.d. trajectories). The central setting is reported in Section 5.2. The columns $W_{\text{CERT-SGD}}/\varepsilon$ and $W_{\text{SOEP-SGD}}/\varepsilon$ report the certificate-to-target ratio at the plateau time.

Case	τ_{P}	$\tau_{\text{CERT-SGD}}$	$\tau_{\text{SOEP-SGD}}$	$\tau_{\text{CERT-SGD}}/\tau_{\text{SOEP-SGD}}$	$W_{\text{CERT-SGD}}/\varepsilon$	$W_{\text{SOEP-SGD}}/\varepsilon$	failures
cos :(100,100,0.10)	24,100	6,389	1,343	4.76	0.51	0.24	0/100
cos :(100,10,0.05)	2,900	2,122	802	2.64	0.85	0.52	0/100
cos :(20,100,0.05)	32,600	24,900	522	47.70	0.87	0.13	0/100
cos :(20,10,0.02)	2,920	26,190	1,078	24.30	3.02	0.60	0/100
$\cos_2$:(100,100,0.05)	36,820	4,470	1,306	3.42	0.35	0.19	0/100
sinmix:(100,100,0.05)	22,710	374	147	2.55	0.12	0.08	0/100
$\cos^2$:(100,100,0.05)	33,910	7,010	1,457	4.81	0.45	0.21	0/100

Loose tolerance ($\varepsilon_{\mathrm{rel}} = 0.40$). At this loose tolerance, both methods fire quickly (median $\tau_{\text{CERT-SGD}} \leq 1{,}681$, median $\tau_{\text{SOEP-SGD}} \leq 98$ across the seven appendix cases). The Cert-SGD bound at the plateau time is now well below the target ($W_{\text{CERT-SGD}}/\varepsilon \in [0.03, 0.76]$), but SOEP-SGD remains $2.4\times$ to $43.2\times$ faster on the same trajectory (Table 12). The SOEP-SGD certificate is consistently 2–5$\times$ tighter than Cert-SGD at the plateau time across the seven cases.

Table 12: Nonconvex problems at $\varepsilon_{\mathrm{rel}} = 0.40$, the seven cases not shown in the main text (per-case medians over 100 i.i.d. trajectories). The central setting is reported in Section 5.2. All methods stop on every case at this tolerance. The columns $W_{\text{CERT-SGD}}/\varepsilon$ and $W_{\text{SOEP-SGD}}/\varepsilon$ report the certificate-to-target ratio at the plateau time.

Case	τ_{P}	$\tau_{\text{CERT-SGD}}$	$\tau_{\text{SOEP-SGD}}$	$\tau_{\text{CERT-SGD}}/\tau_{\text{SOEP-SGD}}$	$W_{\text{CERT-SGD}}/\varepsilon$	$W_{\text{SOEP-SGD}}/\varepsilon$	failures
cos :(100,100,0.10)	24,100	415	88	4.72	0.13	0.06	0/100
cos :(100,10,0.05)	2,900	145	57	2.54	0.21	0.13	0/100
cos :(20,100,0.05)	32,600	1,598	37	43.19	0.22	0.03	0/100
cos :(20,10,0.02)	2,920	1,681	75	22.41	0.76	0.15	0/100
$\cos_2$:(100,100,0.05)	36,820	296	89	3.33	0.09	0.05	0/100
sinmix:(100,100,0.05)	22,710	29	12	2.42	0.03	0.02	0/100
$\cos^2$:(100,100,0.05)	33,910	458	98	4.66	0.11	0.05	0/100

Empirical statistical validity. Both CERT-SGD and SOEP-SGD are anytime-valid and optional-stopping-safe stationarity certificates with budgeted failure probability $\alpha = 0.05$ (Theorem 2 for SOEP-SGD; the closed-form analogue of Aolaritei & Jordan (2025) for CERT-SGD). A *certified-failure* event at tolerance ε is one in which the method stops at some time τ but the actual averaged squared gradient norm $\bar{G}_\tau = \sum_{s \leq \tau} \eta_s \|\nabla f(x_s)\|^2 / S_\tau$ exceeds ε. Table 13 reports raw failure counts across the suite. Across all 4,600 realised stops, no certified failure was observed; the empirical failure rate is well inside the $\alpha = 0.05$ budget at every tolerance.

Table 13: Empirical statistical validity on the nonconvex setting. "Stops" counts the number of trajectory×case events on which the method fired (out of 800 per row). "Failures" counts realised stops at which the actual averaged squared gradient norm $\bar{G}_\tau$ exceeded ε. The budgeted failure probability is $\alpha = 0.05$ throughout.

$\varepsilon_{\mathrm{rel}}$	Method	Stops / 800	Failures	Rate per stop	Rate per trajectory
0.025	CERT-SGD	600	0	0.000	$0/800 = 0.000$
0.025	SOEP-SGD	800	0	0.000	$0/800 = 0.000$
0.10	CERT-SGD	800	0	0.000	$0/800 = 0.000$
0.10	SOEP-SGD	800	0	0.000	$0/800 = 0.000$
0.40	CERT-SGD	800	0	0.000	$0/800 = 0.000$
0.40	SOEP-SGD	800	0	0.000	$0/800 = 0.000$

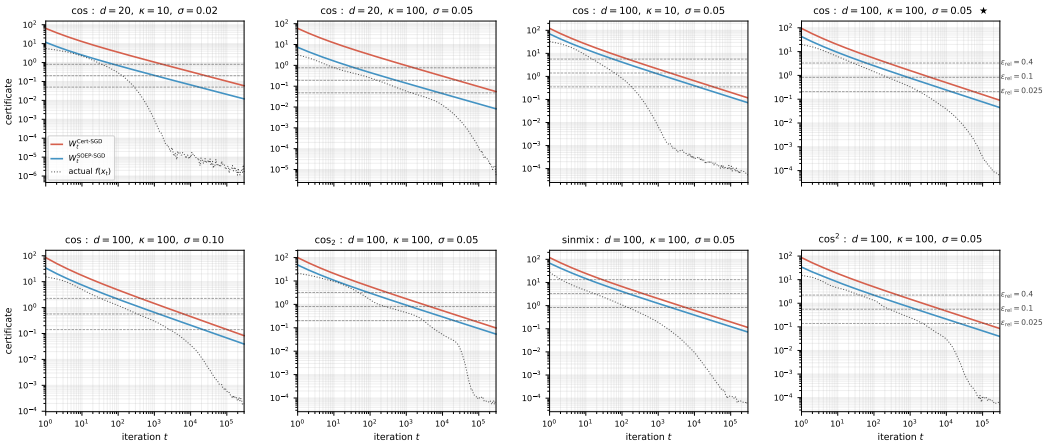

Figure 4: Nonconvex certificate evolution across the suite. One log-log panel per setting (8 settings, central marked ★). On every panel: $W_t^{\mathrm{Cert\text{-}SGD}}$ (red), $W_t^{\mathrm{SOEP\text{-}SGD}}$ (blue), and the actual function value $f(x_t)$ (dotted black) on a single trajectory at $\alpha = 0.05$. Horizontal dashed lines are the three tolerance levels $\varepsilon_{\mathrm{rel}} \in \{0.025, 0.10, 0.40\}$. The SOEP-SGD certificate sits roughly $2\times$ to $5\times$ below the Cert-SGD certificate uniformly across the suite, a more modest gap than in the convex setting (Figure 3) because the deterministic function-drop term $f(x_1)/S_t$ dominates here. The actual function value stays well below both certificates throughout.

Dominant-term diagnostics. A consistent finding across the entire nonconvex suite is that the dominant term in the SOEP-SGD certificate is the function-drop term $f(x_1)/S_t$ rather than the martingale boundary A/S_t or the gradient-cost term C/S_t: on every row of Tables 10–12 the dominant SOEP-SGD certificate term is $B_{f_0} = f(x_1)/S_t$. This contrasts sharply with the convex regime, where the radius and martingale-boundary terms dominate at low and intermediate variance, respectively, before the cumulative gradient term C_t/S_t takes over in high-variance regimes. The practical consequence is that nonconvex stationarity certification is rate-limited by the deterministic descent term $f(x_1)/S_t$, and portfolio acceleration over the closed-form Cert-SGD bound delivers the more modest 2–5$\times$ typical speedup we observe (rather than the 10^2–$10^3\times$ speedups that portfolio optimization yields in the convex setting). The selected effective λ values in the SOEP-SGD portfolio frequently exceed unity (median $\lambda^\star \in [6, 90]$ across the suite), confirming that large-λ accounts—which the positive-grid portfolio cannot reach—are the operative ingredient on nonconvex problems too.

E REAL-DATA DETAILS

This appendix describes the CIFAR-10 deep learning experiment summarized in Section 5.3 of the main text and reports the per-tolerance numerical results.

Architecture. `SmoothCNN` is a small smooth CNN with three convolutional and two fully-connected layers. The convolutions use channel widths $3 \to 128 \to 256 \to 256$, 3×3 kernels, and padding 1; each is followed by a Softplus nonlinearity and a 2×2 average-pooling layer, producing successive spatial resolutions $32 \to 16 \to 8 \to 4$. The resulting $4 \times 4 \times 256$ feature map is flattened to $4{,}096$ units, mapped to 512 units (Softplus), and finally to 10 class logits. All Softplus activations use inverse temperature $\beta = 5.0$; using Softplus and average pooling (rather than ReLU and max pooling) keeps the network globally C^∞ with bounded Hessian, which is what the smoothness assumption 6 underlying the nonconvex stationarity certificate requires. Convolutional and linear weights are initialized Kaiming-normal and biases zero.

Loss and optimization. Each mini-batch loss is the mean cross-entropy plus an L_2 penalty,

$$\ell(\theta; \mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{(x,y)\in\mathcal{B}} \mathrm{CE}\big(f_\theta(x),\, y\big) \;+\; \tfrac{1}{2}\lambda \,\|\theta\|^2, \tag{76}$$

with batch size $|\mathcal{B}| = 256$ and $\lambda = 10^{-4}$. The optimizer is plain SGD without momentum (the L_2 term is included in the loss rather than as optimizer-side weight decay), so each step is $\theta_{t+1} = \theta_t - \eta_t \nabla_\theta \ell(\theta_t; \mathcal{B}_t)$ with $\nabla_\theta \ell$ computed by standard backpropagation. Mini-batches $\mathcal{B}_t$ are drawn with replacement from the training pool. The learning rate follows a cosine schedule from $\eta_0 = 2{\times}10^{-2}$ down to $\eta_{\min} = 10^{-3}$ over the first $300{,}000$ steps and is held at $\eta_{\min}$ thereafter; the maximum optimizer horizon is $T = 800{,}000$ steps. The certificate-relevant per-step gradient norm $\|g_t\|^2$ is computed by summing squared L_2 norms across all parameter tensors of the backward pass before the optimizer update.

Data, evaluation, and plateau heuristic. We use the standard CIFAR-10 split ($50{,}000$ train / $10{,}000$ test). A held-out validation set of $5{,}000$ examples is drawn from the training pool with seed $20{,}260{,}429$, leaving $45{,}000$ examples for training. Inputs are normalized to $[0, 1]$. Validation loss and accuracy are evaluated every $1{,}000$ optimizer steps; test accuracy is evaluated every $5{,}000$ steps. The plateau heuristic monitors validation loss with patience 5, relative-improvement tolerance 10^{-3}, and a $5{,}000$-step warm-up. The SOEP-SGD and Cert-SGD certificates, in contrast, are evaluated at every optimizer step using only quantities already maintained for the SGD update (η_t, $\|g_t\|^2$, S_t, V_t, C_t), so the per-iteration certificate overhead is negligible.

Compute and parallelism. We perform 50 i.i.d. runs distributed across 8 V100 SXM2-32GB GPUs, with a per-run wall-clock cap of 3 hours and a global wall-clock budget of 24 hours; total wall time was $1{,}179$ minutes (≈ 19.6 hours).

Certificate parameters. The nonconvex stationarity certificate uses gradient-norm proxy $G = 500$ and noise proxy $\sigma = 0.5$, both fixed before training. The confidence level is $\alpha = 0.05$. We sweep ten relative tolerances on the same trajectory, $\varepsilon \in \{10, 7, 6, 5, 3, 2, 1.5, 1, 0.75, 0.5\}$.

Portfolio construction. Unlike the synthetic experiments (which use the default full-scale portfolio of Section 4.2), the CIFAR-10 SOEP-SGD portfolio is built by running Algorithm 1 (SOEP-π capital allocation) once before training begins, then frozen. The λ grid contains $m = 81$ atoms log-spaced over a range that brackets the predicted oracle scale $\lambda^\star(v) = \sqrt{\log(1/\alpha)/(c\,v)}$ ($c = 1/2$, nonconvex) across the design-time variance window. The design variances are $n = 101$ points log-spaced on $[v_{\min}, v_{\max}]$, where $v_{\min}, v_{\max}$ are the cumulative noise variances $\sigma^2 G^2 V_t$ at design times $t_{\min} = 100$ and $t_{\max} = 300{,}000$ under the declared cosine schedule. We run $R = 30$ alternating iterations of (i) assigning each design variance to its current argmin atom and (ii) updating the capital weights per (18) with smoothing $\rho = 10^{-4}$. The portfolio is designed for a $300{,}000$-step horizon; the construction in Section 4 guarantees anytime validity for any horizon (including the extended $800{,}000$-step horizon used here), with only the design-optimality of the portfolio tuned to the $300{,}000$-step variance scale.

Proxy diagnostic. We track the empirical maximum gradient norm $\max_{s \le t} \|g_s\|$ along each trajectory and report the indicator `proxy_ok`, which is true when this quantity stays below the declared G throughout the run. All 50 runs satisfy `proxy_ok=True` at every step, with median maximum 94.1, well below $G = 500$. The aggregate proxy-validity rate `proxy_ok_full_rate=` 1.000 provides a clean validity check for the certificate.

Outcomes at the certified stopping time. Table 14 summarizes the per-tolerance comparison. The plateau heuristic terminates between iterations 10,000 and 12,000 across the 50 runs (median 11,000, with 14 runs at 10,000, 23 at 11,000, and 13 at 12,000; the discreteness reflects the 1,000-step evaluation cadence) with median test accuracy 0.6764, but carries no statistical guarantee. At loose tolerances ($\varepsilon \in \{10, 7, 6, 5\}$), both Cert-SGD and SOEP-SGD stop on every one of the 50 runs and the speedup ratio $\tau_{\text{CERT-SGD}}/\tau_{\text{SOEP-SGD}}$ ranges from $9.84\times$ to $30.06\times$. At medium tolerances ($\varepsilon \in \{3, 2, 1.5\}$), SOEP-SGD continues to stop on every run while Cert-SGD does not fire within the 800,000-step budget; the speedup column reports the censored ratio $\tau_{\text{CERT-SGD},c}/\tau_{\text{SOEP-SGD}}$ with $\tau_{\text{CERT-SGD},c} = 800{,}000$. At $\varepsilon \in \{1, 0.75, 0.5\}$, neither certificate stops within the budget. We emphasize that the certificate guarantees first-order stationarity rather than predictive accuracy; the test-accuracy columns are reported only for context.

Table 14: CIFAR-10 deep learning experiment: 50 i.i.d. runs over 8 V100 GPUs, gradient-norm proxy $G = 500$, horizon $T = 800{,}000$ steps, ten relative tolerances. When a method does not certify within T we write $\tau \geq T$; the corresponding speedup column reports the lower bound $T/\tau_{\text{SOEP-SGD}}$. At $\varepsilon \in \{10, 7, 6, 5\}$, both Cert-SGD and SOEP-SGD certify on $50/50$ runs; at $\varepsilon \in \{3, 2, 1.5\}$, only SOEP-SGD certifies $(50/50)$; at $\varepsilon \in \{1, 0.75, 0.5\}$, neither certifies before T. Across all 1,050 method-tolerance-run cells we observe zero certified-failure events; the proxy-validity diagnostic `proxy_ok_full_rate`$= 1.000$ holds across all runs. Test-accuracy entries marked with † correspond to runs that did not self-stop, so the reported value is the median test accuracy at the horizon $t = T$ rather than at a stopping time; these document final-state predictive performance and are common to all censored rows because they are read off the same trajectory at $t = T$. The certificate guarantees first-order stationarity, not generalization.

ε	$\tau_{\text{CERT-SGD}}$	$\tau_{\text{SOEP-SGD}}$	speedup	acc. CERT-SGD	acc. SOEP-SGD
Plateau heuristic (ε-independent, no guarantee): median $\tau_{\text{P}} = 11{,}000$ (range 10,000–12,000), median test acc. 0.676.					
10	64,130	**6,514**	$9.84\times$	0.682	0.657
7	133,769	**13,271**	$10.08\times$	0.680	0.684
6	194,326	**16,208**	$11.99\times$	0.680	0.685
5	625,848	**20,820**	$\mathbf{30.06}\times$	0.679	0.684
3	$\geq$ 800,000	**47,390**	$\geq 16.88\times$	$0.679^\dagger$	0.682
2	$\geq$ 800,000	**116,580**	$\geq 6.86\times$	$0.679^\dagger$	0.680
1.5	$\geq$ 800,000	**358,503**	$\geq 2.23\times$	$0.679^\dagger$	0.679
1, 0.75, 0.5	$\geq$ 800,000	$\geq$ 800,000	—	$0.679^\dagger$	$0.679^\dagger$

F Additional Deep-Learning Experiments and Diagnostics

This appendix reports the expanded evaluation used to assess certificate tightness, stopping efficiency, robustness, and the calibration of the assumptions in modern deep-learning settings. We consider six architecture–dataset pairs: CIFAR-10/ResNet-18, CIFAR-10/ViT, FashionMNIST/ResNet-18, MNIST/ResNet-18, MNIST/ViT, and STL-10/ResNet-18. Every setting uses nine random seeds and a maximum horizon of 500,000 optimizer steps. In addition to the plateau rule, we implement the UP-s, GL-a, and PQ-a validation criteria of Prechelt (1998) and use the best-validation checkpoint as a model-selection reference. These validation rules target predictive performance; they are not stationarity certificates.

F.1 Stopping efficiency and the role of portfolio optimization

Table 15 gives matched-tolerance comparisons. At $\varepsilon \in \{10, 7\}$, where both certified methods stop, SOEP-SGD is 9.2–$15.6\times$ faster. At $\varepsilon \in \{5, 3, 2\}$, Cert-SGD does not stop within 500,000 steps in any setting, whereas SOEP-SGD certifies all three targets in every setting.

To isolate the source of the gain, we evaluate three boundaries on identical optimization trajectories. Table 16 shows that retaining the explicit e-process family with a fixed, unoptimized allocation already improves on Cert-SGD by 4.8–$6.1\times$ at the two tolerances where both stop. Optimizing the pre-specified capital allocation π further reduces stopping time and is the only variant that certifies

Table 15: Median stopping times over nine seeds in the expanded deep-learning study. Each cell reports the results for $\varepsilon = 10 \,/\, 7$.

Setting	$\tau_{\text{CERT-SGD}}$	$\tau_{\text{SOEP-SGD}}$	Speedup
CIFAR-10 / ResNet-18	63,952/167,604	6,808/10,752	$9.4\times$ /$15.6\times$
CIFAR-10 / ViT	64,421/171,990	6,982/11,153	$9.2\times$ /$15.4\times$
FashionMNIST / ResNet-18	63,853/166,753	6,755/10,668	$9.5\times$ /$15.6\times$
MNIST / ResNet-18	63,822/166,485	6,738/10,641	$9.5\times$ /$15.6\times$
MNIST / ViT	63,901/167,169	6,781/10,710	$9.4\times$ /$15.6\times$
STL-10 / ResNet-18	63,831/166,564	6,743/10,649	$9.5\times$ /$15.6\times$

$\varepsilon = 2$ within the horizon. Thus the main gain is structural, while optimizing π sharpens the boundary further at tight tolerances.

Table 16: Controlled certificate-design ablation on identical trajectories. Entries are median stopping times; parenthesized values are speedups relative to Cert-SGD.

ε	Cert-SGD	Fixed grid (unoptimized)	SOEP (optimized π)
10	64,000	13,300 ($4.8\times$)	6,900 ($9.3\times$)
7	167,700	27,400 ($6.1\times$)	10,800 ($15.5\times$)
3	$>$ 500,000	322,500	45,900
2	$>$ 500,000	$>$ 500,000	136,900

In a separate stress test with 200,000 sub-Gaussian trajectories, the SOEP boundary has a uniform crossing rate of 3.2%, below the nominal $\alpha = 5\%$. The closed-form boundary has no crossings even after it is halved, illustrating its excess conservativeness in this regime.

F.2 Comparison with validation-based stopping rules

To make the comparison precise, we call a heuristic stop a failure at target ε when the realized weighted stationarity, evaluated using full-gradient probes at that stop, remains above ε. Table 17 reports the counts across all 54 runs. This diagnostic does not reinterpret validation-based rules as formal stationarity certificates; it demonstrates what can go wrong if they are used to make that claim. Aggressive rules fail frequently even at loose tolerances, while conservative variants avoid many moderate-tolerance failures but fail on every run at $\varepsilon \leq 1.5$.

Table 17: Empirical failures out of 54 runs when a stopping event is treated as a claim that the target weighted-stationarity tolerance has been reached.

Rule	$\varepsilon = 10$	$\varepsilon = 7$	$\varepsilon = 5$	$\varepsilon = 3$	$\varepsilon = 2$	$\varepsilon \leq 1.5$
Plateau (patience)	0/54	45/54	45/54	45/54	45/54	54/54
UP-3	0/54	3/54	10/54	23/54	45/54	54/54
UP-5	0/54	0/54	0/54	0/54	2/54	54/54
GL-1	24/54	44/54	48/54	51/54	53/54	54/54
GL-2	13/54	27/54	34/54	43/54	46/54	54/54
PQ-1	0/54	0/54	2/54	7/54	44/54	54/54
SOEP-SGD	**0/54**	**0/54**	**0/54**	**0/54**	**0/54**	**0/54**

Certified stopping and validation-based model selection serve complementary roles: the former governs the compute-spending decision at a prescribed optimization tolerance, whereas the latter selects a checkpoint for deployment. Table 18 reports their trade-off on the hardest setting. Tightening ε monotonically increases the SOEP stopping time and closes the predictive-performance gap. At $\varepsilon = 5$, SOEP-SGD is within 0.28 percentage points of the best-validation checkpoint at about 18,000 steps; at $\varepsilon = 2$, the difference is negligible.

Table 19 summarizes test-accuracy distributions at $\varepsilon = 5$. The certified and validation-based stopping distributions largely overlap; certification does not visibly increase seed-to-seed variability.

Table 18: Stopping-time and accuracy trade-off on CIFAR-10/ViT (median over nine seeds). The final column is the per-seed median difference from the best-validation checkpoint; positive values indicate accuracy loss.

Rule	Stop step	Val. acc.	Test acc.	Gap (pp)
SOEP-SGD, $\varepsilon = 10$	6,982	.5524	.5576	$+1.48$
SOEP-SGD, $\varepsilon = 5$	18,169	.5660	.5663	$+0.28$
SOEP-SGD, $\varepsilon = 2$	143,366	.5684	.5691	-0.08
Cert-SGD, $\varepsilon = 10$	64,421	.5674	.5702	-0.01
Plateau (patience)	9,000	.5620	.5634	$+0.79$
Best-validation checkpoint	50,000	.5712	.5707	0.00

Table 19: Test accuracy at stopping (mean $\pm$ standard deviation over nine seeds).

Setting	SOEP, $\varepsilon = 5$	Plateau	Best validation
CIFAR-10 / ResNet-18	$.6517 \pm .0089$	$.6527 \pm .0091$	$.6525 \pm .0094$
CIFAR-10 / ViT	$.5660 \pm .0051$	$.5614 \pm .0048$	$.5680 \pm .0063$
FashionMNIST / ResNet-18	$.8999 \pm .0022$	$.8998 \pm .0026$	$.8997 \pm .0019$
MNIST / ResNet-18	$.9917 \pm .0006$	$.9923 \pm .0006$	$.9923 \pm .0005$
MNIST / ViT	$.9626 \pm .0024$	$.9619 \pm .0025$	$.9640 \pm .0020$
STL-10 / ResNet-18	$.5074 \pm .0090$	$.5077 \pm .0095$	$.5080 \pm .0082$

For completeness, Table 20 gives the median per-seed gap to the best-validation checkpoint. At $\varepsilon = 2$, SOEP-SGD is within ± 0.13 percentage points in all six settings. At $\varepsilon = 5$, the largest gap is 0.28 percentage points on CIFAR-10/ViT and is at most 0.12 percentage points elsewhere. Because each gap is computed within a seed before taking the median, it need not equal a difference of separately reported median accuracies.

Table 20: Median test-accuracy gap to the best-validation checkpoint in percentage points (positive means lower accuracy).

Rule	CIFAR/RN18	CIFAR/ViT	F-MNIST	MNIST/RN18	MNIST/ViT	STL-10
SOEP, $\varepsilon = 10$	$+0.00$	$+1.48$	$+0.03$	$+0.10$	$+0.28$	$+0.00$
SOEP, $\varepsilon = 5$	$+0.12$	$+0.28$	-0.07	$+0.07$	$+0.11$	$+0.04$
SOEP, $\varepsilon = 2$	$+0.13$	-0.08	$+0.01$	-0.01	$+0.00$	$+0.05$
Cert-SGD, $\varepsilon = 10$	$+0.26$	-0.01	-0.04	$+0.03$	$+0.02$	$+0.03$
Plateau	$+0.00$	$+0.79$	$+0.00$	$+0.00$	$+0.17$	$+0.02$

F.3 Robustness and assumption calibration

We recompute stopping times on the same trajectories while varying the number of λ-grid points over $m \in \{21, 41, 81, 161\}$ and the variance-window slack factor over $\{1.2, 1.5, 2.0\}$. Stopping times are effectively unchanged at the main-study tolerances and differ by at most nine iterations at the tightest tolerance. The speedup persists across $\alpha \in \{0.01, 0.05, 0.1\}$, ranging from $9.3\times$ to $26.5\times$ and increasing at stricter confidence levels. Changing the physically meaningful proxies σ and G shifts stopping time monotonically, as expected, and we observe no certificate violation in the evaluated configurations.

The sub-Gaussian condition is supported in two complementary ways. First, if per-example gradients are bounded by G, then random mini-batch subsampling produces conditionally mean-zero noise bounded in norm by $2G$; Hoeffding's lemma therefore gives a conditional sub-Gaussian bound for every one-dimensional projection. Second, we measure $2{,}000$ mini-batch gradients along eight directions at three checkpoints in each of the six settings. The largest estimated directional ψ_2 norm is 0.005, compared with the declared proxy $\sigma = 0.5$, giving at least a $100\times$ margin. At normalized deviation threshold $u = 3$, the empirical exceedance frequency is below the corresponding sub-Gaussian tail bound in all 48 direction–setting combinations.

The declared bound $G = 500$ is respected in all observed runs, with median headroom $56\times$. Across the 54 extended runs, full-training-set gradient probes reveal no empirical certificate violation at any probed point. The smallest observed ratio of the certificate to realized weighted stationarity is 1.05 during ViT warm-up. These finite-sample diagnostics support the calibration of the declared proxies but do not replace the formal assumptions. The formal guarantee remains one of weighted stationarity, not generalization or global optimality.