Large Stepsizes Accelerate Gradient Descent for Regularized Logistic Regression

Jingfeng Wu[*]
UC Berkeley
uuujf@berkeley.edu

Pierre Marion[*†]
Inria, DI ENS, PSL University
pierre.marion@inria.fr

Peter L. Bartlett
UC Berkeley & Google DeepMind
peter@berkeley.edu

Abstract

We study *gradient descent* (GD) with a constant stepsize for ℓ_2-regularized logistic regression with linearly separable data. Classical theory suggests small stepsizes to ensure monotonic reduction of the optimization objective, achieving exponential convergence in $\widetilde{\mathcal{O}}(\kappa)$ steps with κ being the condition number. Surprisingly, we show that this can be *accelerated* to $\widetilde{\mathcal{O}}(\sqrt{\kappa})$ by simply using a large stepsize—for which the objective evolves *nonmonotonically*. The acceleration brought by large stepsizes extends to minimizing the population risk for separable distributions, improving on the best-known upper bounds on the number of steps to reach a near-optimum. Finally, we characterize the largest stepsize for the local convergence of GD, which also determines the global convergence in special scenarios. Our results extend the analysis of Wu et al. (2024) from convex settings with minimizers at infinity to strongly convex cases with finite minimizers.

1 Introduction

Machine learning often involves minimizing regularized empirical risk (see, e.g., Shalev-Shwartz and Ben-David, 2014). An iconic case is *logistic regression with ℓ_2-regularization*, given by

$$\widetilde{\mathcal{L}}(\mathbf{w}) := \mathcal{L}(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2, \quad \text{where } \mathcal{L}(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} \ln\left(1+\exp(-y_i \mathbf{x}_i^\top \mathbf{w})\right). \tag{1}$$

Here, $\lambda > 0$ is the regularization hyperparameter, $\mathbf{w} \in \mathbb{H}$ is the trainable parameter, and $(\mathbf{x}_i, y_i) \in \mathbb{H} \times \{\pm 1\}$ for $i = 1, \dots, n$ are the training data, where $\mathbb{H}$ is a Hilbert space. We consider a generic optimization algorithm, *gradient descent* (GD), defined as

$$\mathbf{w}_{t+1} := \mathbf{w}_t - \eta \nabla \widetilde{\mathcal{L}}(\mathbf{w}_t), \quad t \geq 0, \quad \mathbf{w}_0 \in \mathbb{H}, \tag{GD}$$

where $\eta > 0$ is a constant stepsize and $\mathbf{w}_0$ is an initialization, e.g., $\mathbf{w}_0 = 0$.

This problem is smooth and strongly convex. Classical optimization theory suggests a small stepsize, for which GD decreases the objective $\widetilde{\mathcal{L}}(\mathbf{w}_t)$ *monotonically* (Nesterov, 2018, Section 1.2.3), which we refer to as the *stable* regime. In this regime, GD achieves an ε error in $\mathcal{O}(\kappa \ln(1/\varepsilon))$ steps, where $\kappa > 1$ is the condition number of the Hessian of $\widetilde{\mathcal{L}}$ (the smoothness parameter divided by the strong convexity parameter). This step complexity is known to be suboptimal and can be improved to $\mathcal{O}(\sqrt{\kappa}\ln(1/\varepsilon))$ when GD is modified by Nesterov's momentum (Nesterov, 2018, Section 2.2).

A recent line of work shows that GD converges even with large stepsizes that lead to *oscillation* (Wu et al., 2024, other related works will be discussed later in Section 1.1). This is known as the *edge of*

[*]Equal contribution.
[†]Work done while P.M. was a postdoc at EPFL, visiting the Simons Institute at UC Berkeley.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

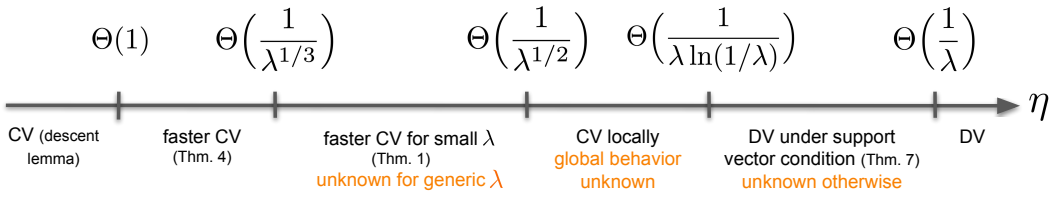

Figure 1: The effect of the stepsize (η) for GD in logistic regression with ℓ_2-regularization (λ). Here, "CV" stands for convergence and "DV" stands for divergence.

stability (EoS) (Cohen et al., 2020) regime. Specifically, Wu et al. (2024) considered (unregularized) logistic regression ((1) with $\lambda = 0$) with linearly separable data. Their problem is smooth and convex, but *not* strongly convex. They showed that GD achieves an $\widetilde{\mathcal{O}}(1/\sqrt{\varepsilon})$ step complexity when operating in the EoS regime, which improves the classical $\widetilde{\mathcal{O}}(1/\varepsilon)$ step complexity when operating in the stable regime. However, it is unclear whether large stepsizes would benefit GD in strongly convex problems such as ℓ_2-regularized logistic regression, for two reasons. First, with linearly separable data, the minimizer of logistic regression is at infinity. Exploiting this property, Wu et al. (2024) showed that GD converges with an arbitrarily large stepsize. However, this is impossible for regularized logistic regression, which is strongly convex and admits a unique, finite minimizer. In this case, GD is *unstable* around the minimizer when the stepsize exceeds a certain threshold (e.g., Hirsch et al., 2013, Section 8), which prevents convergence. Second, Wu et al. (2024) only obtained the accelerated $\widetilde{\mathcal{O}}(1/\sqrt{\varepsilon})$ step complexity for $\varepsilon < 1/n$, where n is the sample size (see their Corollary 2). However, the statistical error (or generalization error) is often larger than $1/n$. In these situations, targeting an optimization error of $\varepsilon < 1/n$ seems less practical, as the statistical error already caps the final population error. It remains unclear whether large stepsizes save computation to minimize population error in the presence of statistical uncertainty.

Contributions. We show that large stepsizes accelerate GD for ℓ_2-regularized logistic regression with linearly separable data, with the following contributions (summarized in Figure 1).

1. For a small regularization hyperparameter ($\lambda = \mathcal{O}(1/n^2)$), we show that GD can achieve an ε error within $\mathcal{O}(\ln(1/\varepsilon)/\sqrt{\lambda})$ steps. This uses an appropriately large stepsize for which GD operates in the EoS regime. Since the condition number of this problem is $\kappa = \Theta(1/\lambda)$, GD matches the accelerated step complexity of Nesterov's momentum by simply using large stepsizes. We further provide a hard dataset showing that this does not always happen if GD operates in the stable regime.
2. For a general λ (independent of n), GD still benefits from large stepsizes, achieving an improved step complexity of $\mathcal{O}(\ln(1/\varepsilon)/\lambda^{2/3})$. Assuming a separable data distribution, GD minimizes the (best-known upper bound on) population risk to the statistical bottleneck in $\widetilde{\mathcal{O}}(n^{2/3})$ steps using large stepsizes and regularization. Without one of these, GD takes $\widetilde{\mathcal{O}}(n)$ steps to achieve the same. This improvement provides evidence that large stepsizes accelerate GD under statistical uncertainty.
3. Finally, under additional data assumptions, we derive a critical threshold $\Theta(1/(\lambda \ln(1/\lambda)))$ on the convergent stepsizes for GD in the following sense. With stepsizes that are smaller by a constant factor, GD converges locally (and globally in 1-dimensional cases); with stepsizes that are larger by a constant factor, GD diverges with almost every initialization $\mathbf{w}_0$.

Terminology. Formally, we say that GD is in the *stable phase* at step t when $\widetilde{\mathcal{L}}(\mathbf{w}_t)$ decreases monotonically from t onwards, and in the *EoS phase* when it does not. Moreover, we say that a GD run is in the *stable regime* if GD is in the stable phase at the initial step, and in the *EoS regime* if it is in the EoS phase in the beginning but transitions to the stable phase afterward. To give intuition, note that, for a strongly convex and sufficiently differentiable objective, if GD converges, it must enter the stable phase in finite time for a generic initialization. This means that a typical convergent GD run is either in the stable regime or the EoS regime.

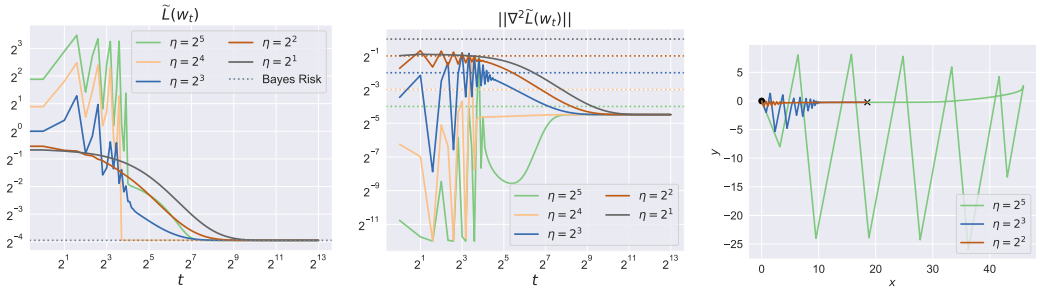

Figure 2: Illustration of large stepsizes accelerating GD. We run constant stepsize GD for an ℓ_2-regularized logistic regression on a two-dimensional separable dataset. The dataset is given by $\mathbf{x}_1 = (\gamma, 1)$, $\mathbf{x}_2 = (\gamma, -2)$, $y_1 = y_2 = 1$, where $\gamma = 0.2$. The regularization is $\lambda = 2^{-12}$. GD is initialized at $\mathbf{w}_0 = 0$. **Left:** Objective value as a function of training steps. **Middle:** Sharpness (the largest eigenvalue of the Hessian of the objective) as a function of training steps. **Right:** GD trajectory in the parameter space, where the black dot is the GD initialization and the black cross is the minimizer. Additional details and plots are given in Appendix E.

Simulations. Our results are illustrated in Figure 2 by running GD for ℓ_2-regularized logistic regression on a toy two-dimensional separable dataset. Figure 2 suggests that GD converges faster with a larger stepsize by entering the EoS regime, in which the sharpness oscillates around $2/\eta$ in the initial phase.

Notation. For two positive-valued functions f and g, we write $f \lesssim g$ or $f \gtrsim g$ if there exists $c > 0$ such that for every x, $f(x) \le cg(x)$ or $f(x) \ge cg(x)$, respectively. We write $f \eqsim g$ if $f \lesssim g \lesssim f$. We use the standard big-O notation, with $\widetilde{\mathcal{O}}$ and $\widetilde{\Omega}$ to hide polylogarithmic factors within the $\mathcal{O}$ and Ω notation, respectively. For two vectors $\mathbf{u}$ and $\mathbf{v}$ in a Hilbert space, we denote their inner product by $\langle \mathbf{u}, \mathbf{v} \rangle$ or, equivalently, $\mathbf{u}^\top \mathbf{v}$. We write $\|\mathbf{u}\| := \sqrt{\mathbf{u}^\top \mathbf{u}}$.

1.1 Related work

Edge of stability. In practice, GD often induces an oscillation yet still converges in the long run (see Wu et al., 2018; Cohen et al., 2020, and references therein). This is referred to by Cohen et al. (2020) as the *edge of stability* (EoS). Since gradient flow would never increase the objective, EoS is essentially a consequence of large stepsizes. Cohen et al. (2020) further pointed out that GD apparently needs to operate in the EoS regime to obtain reasonable optimization and generalization performance in practical deep learning settings. Besides the empirical results, the theoretical mechanism of EoS has been investigated in several papers (see, e.g., Damian et al., 2022; Zhu et al., 2022; Arora et al., 2022, and references therein). In particular, Cohen et al. (2025) proposed a modified ODE called *centrol flow* to approximate the time-averaged GD trajectory in the EoS regime. Instead of focusing on explaining EoS itself, we study the optimization benefits of GD operating in the EoS regime.

Another line of research has focused on the statistical benefits of large stepsizes for neural networks (see, e.g., Mulayoff et al., 2021; Qiao et al., 2024; Wu et al., 2025b), exploiting the observation that GD with a larger stepsize is constrained to converge to flatter minima (Wu et al., 2018). However, those works assumed the convergence of GD with large stepsizes, which itself is a challenging question. In this regard, our work makes partial progress by showing the global convergence of GD with large stepsizes for ℓ_2-regularized logistic regression.

Aggresive stepsize schedulers. A recent line of research discovered that a variant of GD with certain aggressive stepsize schedulers yields improved convergence for smooth and (strongly) convex optimization (see Altschuler and Parrilo, 2025; Grimmer, 2024; Zhang et al., 2024, and references therein). As a representative example, Altschuler and Parrilo (2025) showed that their GD variant with the *silver stepsize scheduler* attains an improved $\widetilde{\mathcal{O}}(\kappa^{0.7864})$ step complexity for smooth and strongly convex problems with condition number κ. Similar to our work, they obtained acceleration by using large stepsizes that operate outside the classical stable regime. However, there are several notable differences. First, our problem class, ℓ_2-regularized logistic regression, is smaller than theirs.

However, we obtain a better $\widetilde{\mathcal{O}}(\kappa^{0.5})$ step complexity. Moreover, we do so with the simpler approach of constant stepsize GD. Finally, from a technical perspective, our analysis is *anytime* as our stepsize choice does not rely on the target error ε (see Theorem 1), while the algorithm of Altschuler and Parrilo (2025) needs to know the target error ε in advance.

Logistic regression. Logistic regression with linearly separable data is a standard class of problems in optimization and statistical learning theory. For GD with small stepsizes in the stable regime, Soudry et al. (2018) and Ji and Telgarsky (2018) showed that GD diverges to infinity while converging in direction to the maximum ℓ_2-margin direction. This result was later extended to GD with an arbitrarily large stepsize in the EoS regime (Wu et al., 2023). More recently, Wu et al. (2024) showed that GD with a large stepsize attains an accelerated $\widetilde{\mathcal{O}}(1/\sqrt{\varepsilon})$ step complexity for logistic regression with linearly separable data, demonstrating the benefits of EoS. For the same problem, Zhang et al. (2025) improved the step complexity to $1/\gamma^2$ by considering an *adaptive* large-stepsize variant of GD, where γ is the margin of the dataset, and they further showed that this is minimax optimal. As discussed earlier, their results rely strongly on the minimizer being at infinity. In comparison, we focus on logistic regression with ℓ_2-regularization, where the minimizer is finite.

For ℓ_p-regularized logistic regression with linearly separable data, Rosset et al. (2004) showed that the regularized empirical risk minimizer converges in direction to the maximum ℓ_p-margin direction as the regularization tends to zero. Our work complements theirs by considering the step complexity of finding the ℓ_2-regularized empirical risk minimizer.

For logistic regression with *strictly nonseparable* data, Meng et al. (2024) constructed examples where GD with large stepsizes does not converge globally (even if the stepsize allows local convergence). Similar to our problem, theirs is also smooth, strictly convex, and admits a unique finite minimizer. However, Meng et al. (2024) focused on negative results, while we provide positive results with separable data for the global convergence of GD with large stepsizes. Proving positive results in the nonseparable case is an interesting direction for future work.

2 Large stepsizes accelerate GD

We make the following standard assumptions (Novikoff, 1962) throughout the paper.

Assumption 1 (Bounded and separable data)**.** *Assume the training data* $(\mathbf{x}_i, y_i)_{i=1}^n$ *satisfies*

A. for every $i = 1, \dots, n$, $\|\mathbf{x}_i\| \le 1$ *and* $y_i \in \{\pm 1\}$*;*

B. there is a margin $\gamma \in (0, 1]$ *and a unit vector* $\mathbf{w}^*$ *such that* $y_i \mathbf{x}_i^\top \mathbf{w}^* \ge \gamma$ *for every* $i = 1, \dots, n$.

Under Assumption 1, the objective function $\widetilde{\mathcal{L}}(\cdot)$ defined in (1) is $(1+\lambda)$-smooth and λ-strongly convex. The condition number of this problem is $\kappa = \Theta(1/\lambda)$, as the regularization hyperparameter λ is typically small. For a small stepsize $\eta = 1/(1+\lambda) = \Theta(1)$, GD operates in the stable regime, achieving a well-known $\mathcal{O}(\ln(1/\varepsilon)/\lambda)$ step complexity (Nesterov, 2018). Quite surprisingly, we will show that this can be improved to $\mathcal{O}(\ln(1/\varepsilon)/\sqrt{\lambda})$ when the regularization hyperparameter λ is small (compared to the reciprocal of the sample size; see Section 2.1), and to $\mathcal{O}(\ln(1/\varepsilon)/\lambda^{2/3})$ for general λ (Section 2.2). We obtain this acceleration by using large stepsizes, where GD operates in the EoS regime.

The minimizer, $\mathbf{w}_\lambda := \arg\min \widetilde{\mathcal{L}}(\cdot)$, is unique and finite when $\lambda > 0$.

2.1 Matching Nesterov's acceleration under small regularization

Our first theorem characterizes the convergence of GD in the EoS regime when the regularization is small. The proof is deferred to Appendix A.2.

Theorem 1 (Convergence under small regularization)**.** *Consider* (GD) *for* ℓ_2*-regularized logistic regression* (1) *under Assumption 1. Assume without loss of generality that* $\mathbf{w}_0 = 0$*. There exist constants* $C_1, C_2, C_3 > 1$ *such that the following holds. For every* $n \ge 2$*,*

$$\lambda \le \frac{\gamma^2}{C_1 n \ln n} \quad \text{and} \quad \eta \le \min\left\{\frac{\gamma}{\sqrt{C_1 \lambda}}, \frac{\gamma^2}{C_1 n \lambda}\right\},$$

we have the following:

- ***Phase transition.*** *GD must be in the stable phase at step τ for*

$$\tau := \frac{C_2}{\gamma^2} \max\left\{\eta,\, n,\, \frac{n \ln n}{\eta}\right\},$$

that is, $\widetilde{\mathcal{L}}(\mathbf{w}_t)$ decreases monotonically for $t \geq \tau$.

- ***The stable phase.*** *Moreover, for every $t \geq \tau$, we have*

$$\widetilde{\mathcal{L}}(\mathbf{w}_t) - \min \widetilde{\mathcal{L}} \leq C_3 e^{-\lambda\eta(t-\tau)}, \quad \|\mathbf{w}_t - \mathbf{w}_\lambda\| \leq C_3 \frac{\eta + \ln(\gamma^2/\lambda)}{\gamma} e^{-\lambda\eta(t-\tau)/2}.$$

Theorem 1 provides a convergence guarantee for GD with a stepsize as large as $\eta = \mathcal{O}(1/\sqrt{\lambda})$ (treating other problem-dependent parameters, γ and n, as constants). With this stepsize, GD might not decrease the objective monotonically—that is, GD might be in the EoS phase at the beginning. Nonetheless, Theorem 1 shows that GD must undergo a phase transition to the stable phase in $\tau = \mathcal{O}(\eta)$ steps. In the stable phase, GD benefits from the large stepsize, achieving an ε error in $\mathcal{O}\big(\ln(1/\epsilon)/(\eta\lambda)\big)$ subsequent steps.

Theorem 1 recovers the classical $\mathcal{O}(\ln(1/\varepsilon)/\lambda)$ step complexity when GD operates in the stable regime with $\eta = \Theta(1)$. Additionally, Theorem 1 suggests that GD achieves faster convergence in the EoS regime when the stepsize is large, but not larger than $\Theta(1/\sqrt{\lambda})$. Choosing the largest allowed stepsize, GD matches the accelerated step complexity of Nesterov's momentum. This is detailed in the following corollary, with proof deferred to Appendix A.3.

Corollary 2 (Step complexity under small regularization)**.** *Under the setting of Theorem 1, by using the largest allowed stepsize,*

$$\eta := \min\left\{\frac{\gamma}{\sqrt{C_1\lambda}},\, \frac{\gamma^2}{C_1 n \lambda}\right\},$$

we have $\widetilde{\mathcal{L}}(\mathbf{w}_t) - \min \widetilde{\mathcal{L}} \leq \varepsilon$ for

$$t \leq C_4 \max\left\{\frac{1}{\gamma\sqrt{\lambda}},\, \frac{n}{\gamma^2}\right\} \ln(1/\varepsilon),$$

where $C_4 > 1$ is a constant. Thus for $\lambda \lesssim \gamma^2/n^2$, $\eta \eqsim 1/\sqrt{\lambda}$ ensures that $t \eqsim \ln(1/\varepsilon)/\sqrt{\lambda}$ suffices.

Matching Nesterov's acceleration. Treat γ as a constant. For a small regularization of $\lambda \lesssim 1/n^2$, Corollary 2 shows that GD achieves a step complexity of $\mathcal{O}(\ln(1/\varepsilon)/\sqrt{\lambda})$ using a large stepsize. Since the condition number is $\kappa = \Theta(1/\lambda)$, this matches the accelerated step complexity of Nesterov's momentum, improving the classical $\mathcal{O}(\ln(1/\varepsilon)/\lambda)$ step complexity for GD in the stable regime.

For a moderately small regularization, $1/n^2 \lesssim \lambda \lesssim 1/(n \ln n)$, Corollary 2 implies a step complexity of $\mathcal{O}(n \ln(1/\varepsilon))$ for GD with a large stepsize, which still improves the classical $\mathcal{O}(\ln(1/\varepsilon)/\lambda)$ step complexity for GD in the stable regime (by at least a logarithmic factor). However, it no longer matches Nesterov's momentum. It is an open question whether large stepsize GD can match Nesterov's momentum for a moderate (or large) regularization.

A lower bound for stable convergence. We have shown that large stepsizes accelerate the convergence of GD. This acceleration effect is closely tied to operating in the EoS regime. To clarify, our next theorem shows that GD in the stable regime suffers from an $\widetilde{\Omega}(1/\lambda)$ step complexity in the worst case. Its proof is deferred to Appendix B.

Theorem 3 (A lower bound)**.** *Consider (GD) for ℓ_2-regularized logistic regression (1) with $\mathbf{w}_0 = 0$ and the following dataset (satisfying Assumption 1):*

$$\mathbf{x}_1 = (\gamma,\, 0.9), \quad \mathbf{x}_2 = (\gamma,\, -0.5), \quad y_1 = y_2 = 1, \quad 0 < \gamma < 0.1.$$

There exist $C_1, C_2, C_3 > 1$ that only depend on γ such that the following holds. For every $\lambda < 1/C_1$ and $\varepsilon < C_2 \lambda \ln^2(1/\lambda)$, if η is such that $(\widetilde{\mathcal{L}}(\mathbf{w}_t))_{t\geq 0}$ is nonincreasing, then

$$\widetilde{\mathcal{L}}(\mathbf{w}_t) - \min \widetilde{\mathcal{L}} \leq \varepsilon \quad \Rightarrow \quad t \geq \frac{\ln\big(1/\varepsilon\big)}{C_3 \lambda \ln^2(1/\lambda)}.$$

It is worth noting that Theorem 3 focuses on the common asymptotic case of a small ε; for $\varepsilon \gtrsim \lambda \ln^2(1/\lambda)$, the step complexity is $\Omega(1/\varepsilon)$, which is reflected by its proof in Appendix B.

A limitation. We conclude this part by discussing a limitation of our Theorem 1. Note that Theorem 1 only allows a small regularization such that $\lambda \lesssim 1/(n \ln n)$. A regularization of this order might be suboptimal in the presence of statistical noise. Moreover, the proof of Theorem 1 implies that, in the stable phase, all training data are classified correctly (see Lemma 13 in Appendix A.2). Therefore, the allowed regularization is too small to prevent the minimizer from perfectly classifying the training data. A similar limitation is encountered in the prior work of Wu et al. (2024), who showed acceleration with large stepsizes only when targeting an optimization error small enough to imply perfect classification of the training data (see their Corollary 2).

Depending on the statistical model, a perfect fit to the training data does not necessarily lead to overfitting (a phenomenon known as *benign overfitting*, see Bartlett et al., 2020). Even so, a larger regularization such as $\lambda \gtrsim 1/n$ often leads to better performance in many statistical models (one such case will be discussed in Section 3). Below, we show that GD can still benefit from large stepsizes even with regularization larger than $1/(n \ln n)$. This allows large regularization that leads to misclassification of the training data.

2.2 Improved convergence under general regularization

Our next theorem characterizes the convergence of GD for ℓ_2-regularized logistic regression in the EoS regime with a general regularization hyperparameter λ. The proof is deferred to Appendix A.4.

Theorem 4 (Convergence under general regularization)**.** *Consider* (GD) *for ℓ_2-regularized logistic regression* (1) *under Assumption 1. Assume without loss of generality that $\mathbf{w}_0 = 0$. There exist constants $C_1, C_2, C_3 > 1$ such that the following holds. For every*

$$\lambda \le \frac{\gamma^2}{C_1}, \quad \eta \le \left(\frac{\gamma^2}{C_1 \lambda}\right)^{1/3},$$

we have the following:

- ***Phase transition time.*** *GD must be in the stable phase at step $\tau := C_2 \max\{1, \eta^2\}/\gamma^2$.*
- ***The stable phase.*** *Moreover, for $t \ge \tau$, we have*

$$\widetilde{\mathcal{L}}(\mathbf{w}_t) - \min \widetilde{\mathcal{L}} \le \frac{C_3}{\eta} e^{-\lambda \eta (t-\tau)}, \quad \|\mathbf{w}_t - \mathbf{w}_\lambda\| \le C_3 \frac{\eta + \ln(\gamma^2/\lambda)}{\gamma} e^{-\lambda \eta (t-\tau)/2}.$$

Similarly to Theorem 1, Theorem 4 allows a large stepsize, in which GD might be in the EoS phase at the beginning, then it must transition to the stable phase in finite steps, achieving an exponential convergence subsequently.

Unlike Theorem 1, where the allowed regularization and phase transition time are functions of the sample size n, Theorem 4 is completely independent of the sample size n. In particular, it allows for a large regularization of order $1/(n \ln n) \lesssim \lambda \lesssim 1$, with which the minimizer of the regularized logistic regression (1) might not correctly classify the training data.

The relaxation of the allowed regularization is obtained at the price of a tighter constraint on the allowed stepsize, $\eta = \mathcal{O}(1/\lambda^{1/3})$, and a slower phase transition time, $\tau = \Theta(\eta^2)$. Nonetheless, Theorem 4 still implies that large stepsizes lead to acceleration, as explained in the following corollary. Its proof is included in Appendix A.5.

Corollary 5 (Step complexity under general regularization)**.** *Under the setting of Theorem 4, by using the largest allowed stepsize, $\eta := (\gamma^2/(C_1 \lambda))^{1/3}$, we have $\widetilde{\mathcal{L}}(\mathbf{w}_t) - \min \widetilde{\mathcal{L}} \le \varepsilon$ for*

$$t \le C_3 \frac{\ln(1/\varepsilon)}{(\gamma \lambda)^{2/3}},$$

where $C_3 > 1$ is a constant.

Ignoring the dependence on γ, Corollary 5 shows that GD achieves an ε error in $\mathcal{O}(\ln(1/\varepsilon)/\lambda^{2/3})$ steps using a large stepsize. This improves the classical $\mathcal{O}(\ln(1/\varepsilon)/\lambda)$ step complexity for GD in the stable regime, although it does not match Nesterov's momentum.

We remark that the predictions of Corollaries 2 and 5 are incomparable even in the regime where both are applicable, that is, $\lambda \lesssim (n \ln n)^{-1}$. Specifically, in this regime, Corollary 5 predicts a step

complexity of $\widetilde{\mathcal{O}}(\lambda^{-2/3})$ while Corollary 2 predicts a step complexity of $\widetilde{\mathcal{O}}(\max\{\lambda^{-1/2}, n\})$. The prediction of Corollary 5 is worse for $\lambda \lesssim n^{-3/2}$ but is better for $n^{-3/2} \lesssim \lambda \lesssim (n \ln n)^{-1}$. Thus, Corollaries 2 and 5 are incomparable even in the joint applicable regime, suggesting our analysis is improvable. Technically, this mismatch stems from two distinct approaches for analyzing phase transition (see Section 2.3). We leave it as future work to improve our analysis.

In statistical learning contexts, the (optimal) regularization hyperparameter λ is often a function of the sample size n, for example, $\lambda = \Theta(1/n^\alpha)$ for some $\alpha > 0$. In contrast to Corollary 2, which only applies to small regularization (with $\alpha > 1$), the acceleration implied by Corollary 5 applies to any such λ (in particular, with $0 < \alpha \le 1$). Specifically, Corollary 5 shows an $\mathcal{O}(n^{2\alpha/3} \ln(1/\varepsilon))$ step complexity for GD when $\lambda = \Theta(1/n^\alpha)$ for any $\alpha > 0$. We will revisit this later in Section 3 and show the acceleration of large stepsizes in a statistical learning setting.

2.3 Technical overview

In this part, we discuss key ideas in our analysis and elaborate on our technical innovations compared to the prior work of Wu et al. (2024).

Bounds in the EoS phase. The following lemma provides bounds on the logistic empirical risk and parameter norm for any time; in particular, it applies to the EoS phase.

Lemma 1 (EoS bounds)**.** *Assume that $\eta\lambda \le 1/2$ and $\mathbf{w}_0 = 0$. Then for every t, and in particular in the EoS phase, we have*

$$\frac{1}{t}\sum_{k=0}^{t-1} \mathcal{L}(\mathbf{w}_k) \le 10\frac{\eta^2 + \ln^2(e + \gamma^2 \min\{\eta t, 1/\lambda\})}{\gamma^2 \min\{\eta t, 1/\lambda\}}, \quad \|\mathbf{w}_t\| \le 4\frac{\eta + \ln(e + \gamma^2 \min\{\eta t, 1/\lambda\})}{\gamma}.$$

Lemma 1 recovers the EoS bounds in (Wu et al., 2024) (see their Lemma 8 in Appendix B) for the special case of $\lambda = 0$. The intuition is that, for $\eta t < 1/\lambda$, the regularization term is negligible compared to the logistic term. However, these bounds are too crude for a large t.

New challenges. The analysis by Wu et al. (2024) relies on the self-boundedness of the logistic loss, $\|\nabla^2 \mathcal{L}(\mathbf{w})\| \le \mathcal{L}(\mathbf{w})$, and that the minimizers of $\mathcal{L}(\cdot)$ appear at infinity. Although GD with a large stepsize oscillates initially, it keeps moving towards infinity along the maximum ℓ_2-margin direction, which reduces the objective $\mathcal{L}(\mathbf{w})$ in the long run. Once GD hits a small objective value, $\mathcal{L}(\mathbf{w}) \lesssim 1/\eta$, it enters the stable phase as the local landscape becomes flat due to the self-boundedness. In the stable phase, GD continues to move towards infinity along the maximum ℓ_2-margin direction.

However, the situation is significantly different in the presence of an ℓ_2-regularization. In this case, the minimizer has a small norm, $\|\mathbf{w}_\lambda\| = \mathcal{O}(\ln(1/\lambda))$ (see Lemma 3 in Appendix A.1). But large stepsizes GD can go as far as $\Theta(\eta) = \mathrm{poly}(1/\lambda)$ in the EoS phase (Lemma 1), which is even more distant from the minimizer than the initialization $\mathbf{w}_0 = 0$. Instead of moving towards infinity, in our case, GD must move backwards (if it converges).

Consider a flat region defined as

$$\{\mathbf{w} : \|\nabla^2 \widetilde{\mathcal{L}}(\mathbf{w})\| = \|\nabla^2 \mathcal{L}(\mathbf{w})\| + \lambda \lesssim 1/\eta\} \approx \{\mathbf{w} : \mathcal{L}(\mathbf{w}) \lesssim 1/\eta\}.$$

When GD enters this region, we expect $\widetilde{\mathcal{L}}(\mathbf{w})$ to decrease in the next step (see Lemma 12 in Appendix A.1). However, different from Wu et al. (2024), this does not guarantee that GD stays in this region. In fact, the regularization term leads to contraction towards zero, so a decrease of $\widetilde{\mathcal{L}}(\mathbf{w})$ may cause an increase of $\mathcal{L}(\mathbf{w})$, and then GD might leave this region. In Theorems 1 and 4, we identify two situations where GD stays in the flat region, respectively, as explained below.

Intuition of Theorem 4. When $\eta \lesssim 1/\lambda^{1/3}$, Lemma 1 implies a small regularization term throughout the training, $\lambda\|\mathbf{w}_t\|^2 = \widetilde{\mathcal{O}}(\lambda\eta^2) = \widetilde{\mathcal{O}}(1/\eta)$. Thus, the logistic term dominates the whole objective in the EoS phase, and $\mathcal{L}(\mathbf{w}) \lesssim 1/\eta$ is nearly the same as $\widetilde{\mathcal{L}}(\mathbf{w}) \lesssim 1/\eta$. By the decrease of $\widetilde{\mathcal{L}}(\mathbf{w})$ within the flat region, we can show GD stays in this region by induction (see Lemma 15 in Appendix A.4).

Table 1: Step complexities for variants of GD to reach a population risk of $\widetilde{\mathcal{O}}(1/n)$.

algorithm	# steps	λ	η	population risk
GD	$\mathcal{O}(n)$	0	$\Theta(1)$	$\widetilde{\mathcal{O}}(1/(\gamma^2 n))$
	$\mathcal{O}(n \ln n)$	$1/n$	1	$\widetilde{\mathcal{O}}(1/(\gamma^2 n))$
	$\mathcal{O}((n/\gamma)^{2/3} \ln n)$	$1/n$	$\Theta((\gamma^2 n)^{1/3})$	$\widetilde{\mathcal{O}}(1/(\gamma^2 n))$
Nesterov's momentum	$\mathcal{O}(n^{1/2} \ln n)$	$1/n$	1	$\widetilde{\mathcal{O}}(1/(\gamma^2 n))$
adaptive GD	$\mathcal{O}(1/\gamma^2)$	0	$\Theta(\ln n)$	$\widetilde{\mathcal{O}}(1/(\gamma^4 n))$

Intuition of Theorem 1. For $\eta \lesssim 1/\lambda^{1/2}$, the above arguments no longer work, since the regularization term could be as large as $\widetilde{\mathcal{O}}(\lambda\eta^2) = \widetilde{\mathcal{O}}(1)$ in the EoS phase. Alternatively, we compare the size of the gradients from the logistic term $\|\nabla\mathcal{L}(\mathbf{w})\|$ and the regularization term $\|\lambda\mathbf{w}\|$. If the former is larger, then the logistic term decreases; if the latter is larger, then by the exponential tail of the logistic loss, we conclude that $\mathcal{L}(\mathbf{w}) \approx \|\nabla\mathcal{L}(\mathbf{w})\| \leq \|\lambda\mathbf{w}\| = \widetilde{\mathcal{O}}(1/\eta)$, where the last equality is by Lemma 1. In both cases, GD stays in the flat region (see Lemma 13 in Appendix A.2).

3 Benefits of large stepsizes under statistical uncertainty

In this section, we apply Theorem 4 in a statistical learning setting, showing that the acceleration of large stepsizes continues to hold even under statistical uncertainty. We make the following natural assumption on the population data distribution.

Assumption 2 (Bounded and separable distribution)**.** *Assume that $(\mathbf{x}_i, y_i)_{i=1}^n$ are independent copies of $(\mathbf{x}, y)$ that follows a distribution such that*

A. the label is binary, $y \in \{\pm 1\}$, and $\|\mathbf{x}\| \leq 1$, almost surely;

B. there exist a margin $\gamma > 0$ and a unit vector $\mathbf{w}^*$ such that $y\mathbf{x}^\top\mathbf{w}^* \geq \gamma$, almost surely.

The population risk of an estimator $\hat{\mathbf{w}}$ is defined as

$$\mathcal{L}_{\text{test}}(\hat{\mathbf{w}}) := \mathbb{E}\ln\left(1 + \exp(-y\mathbf{x}^\top\hat{\mathbf{w}})\right),$$

where the expectation is over the distribution of $(\mathbf{x}, y)$ satisfying Assumption 2.

The following Proposition 6 gives the best-known population risk upper bound (without assuming enormous burn-in samples) in the setting of Assumption 2. This is a direct consequence of the fast rate established by Srebro et al. (2010, Theorem 1) using *local Rademacher complexity* (Bartlett et al., 2005). A variant of Proposition 6 also appears in Schliserman and Koren (2024, Proposition 1). We include its proof in Appendix C.1 for completeness.

Proposition 6 (A population risk bound)**.** *Suppose that $(\mathbf{x}_i, y_i)_{i=1}^n$ satisfies Assumption 2. Then for every $\hat{\mathbf{w}}$, with probability at least $1-\delta$ over the randomness of sampling $(\mathbf{x}_i, y_i)_{i=1}^n$, we have*

$$\mathcal{L}_{\text{test}}(\hat{\mathbf{w}}) \leq C\left(\mathcal{L}(\hat{\mathbf{w}}) + \frac{\max\{1, \|\hat{\mathbf{w}}\|^2\}\left(\ln^3(n) + \ln(1/\delta)\right)}{n}\right),$$

where $C > 1$ is a constant.

Recall that the minimizer of $\mathcal{L}(\cdot)$ is at infinity under Assumption 2. However, Proposition 6 suggests that a good estimator should balance its fit to the training data (measured by $\mathcal{L}(\hat{\mathbf{w}})$) and its complexity (measured by $\|\hat{\mathbf{w}}\|$). It is also worth noting that the upper bound in Proposition 6 is at least $\widetilde{\Omega}(1/n)$—a bottleneck that stems from the statistical uncertainty. With this in mind, we are ready to discuss the number of steps needed by GD (and its variants) to minimize the population risk to the statistical bottleneck. Table 1 summarizes the results, which we explain in detail below.

Logistic regression with ℓ_2-regularization. Let us first consider the minimizer of the ℓ_2-regularized logistic regression, $\mathbf{w}_\lambda := \arg\min \widetilde{\mathcal{L}}(\cdot)$. With direct calculation, setting $\lambda = \Theta(1/n)$ minimizes the upper bound in Proposition 6, resulting in an $\widetilde{\mathcal{O}}(1/(\gamma^2 n))$ population risk (see Appendix C.2 for details). This is nearly optimal ignoring the logarithmic factors and dependence on γ. Clearly, the same bound applies to any approximate minimizer $\hat{\mathbf{w}}$ such that $\|\hat{\mathbf{w}} - \mathbf{w}_\lambda\| \le \varepsilon := 1/\mathrm{poly}(n)$. To obtain such an approximate minimizer,

- GD with a small stepsize $\eta = 1$ needs $\mathcal{O}(n \ln n)$ steps by the classical optimization theory;
- GD with a large stepsize $\eta = \Theta((\gamma^2 n)^{1/3})$ needs $\mathcal{O}((n/\gamma)^{2/3} \ln n)$ steps by Theorem 4;
- Nesterov's momentum needs $\mathcal{O}(n^{1/2} \ln n)$ steps by the classical optimization theory.

These suggest that large stepsizes accelerate GD in the presence of statistical uncertainty, although not as fast as Nesterov's momentum.

Logistic regression without regularization. Instead of solving regularized logistic regression, one can also apply GD to the unregularized logistic regression with early stopping to obtain a small population risk. For instance, Shamir (2021) showed that GD with a small stepsize $\eta = 1$ achieves a population risk of $\widetilde{\mathcal{O}}(1/(\gamma^2 n))$ in $\mathcal{O}(n)$ steps. A similar result is obtained by Schliserman and Koren (2024) using a different proof technique.

For GD with a larger stepsize, Wu et al. (2024) obtained an empirical risk bound of $\mathcal{L}(\mathbf{w}_t) = \mathcal{O}\big((\eta^2 + \ln^2(\eta t))/(\gamma^2 \eta t)\big)$ and a parameter norm bound of $\|\mathbf{w}_t\| = \mathcal{O}\big((\eta + \ln(\eta t))/\gamma\big)$. Note that we do not consider their accelerated empirical risk bound here, as it only applies after $\Theta(n)$ steps. Plugging these bounds into Proposition 6, however, one cannot resolve for a stepsize η better than the choice of $\eta = 1$ (ignoring logarithmic factors). That is, without regularization, solely using a large stepsize does not accelerate GD in the presence of statistical uncertainty. This sets an interesting gap between our acceleration results and those by Wu et al. (2024).

One can also solve logistic regression via *adaptive* GD (Ji and Telgarsky, 2021; Zhang et al., 2025), defined as $\mathbf{w}_{t+1} := \mathbf{w}_t - \eta \nabla \mathcal{L}(\mathbf{w}_t)/\mathcal{L}(\mathbf{w}_t)$. This is faster for optimization than GD as it adapts to the curvature. Specifically, Zhang et al. (2025) obtained an empirical risk bound of $\mathcal{L}(\bar{\mathbf{w}}_t) \le \exp(-\Theta(\gamma^2 \eta t))$ for $t > 1/\gamma^2$ (see their Theorem 2.1) and a parameter norm bound of $\|\bar{\mathbf{w}}_t\| \le \eta t$ (see the proof of their Theorem 2.2), where $\bar{\mathbf{w}}_t$ is the average of the iterates up to step t. Plugging these bounds into Proposition 6, we minimize the upper bound by setting $\eta = \Theta(\ln n)$ and $t = \Theta(1/\gamma^2)$, with which the population risk is $\widetilde{\mathcal{O}}(1/(\gamma^4 n))$. Although the step complexity is much improved, the population risk seems to have a suboptimal dependence on γ.

A limitation. We note that the above discussion is based on the best-known population risk upper bound in a statistical setting specified by Assumption 2. Depending on the actual data distribution, the population risk might be smaller than that (although we suspect the upper bound is nearly sharp in the worst case). We leave it for future work to investigate the effect of large stepsizes in broader statistical learning settings.

4 A critical threshold on the convergent stepsizes

We have shown the global convergence of GD with stepsizes as large as $\mathcal{O}(1/\sqrt{\lambda})$ in Section 2. Clearly, if $\eta > 2/\lambda$, GD diverges with almost every initialization. But the largest convergent stepsize is unclear yet—this is the focus of this section. We will show the largest convergent stepsize is $\Theta(1/(\lambda \ln(1/\lambda))$ under the following technical condition:

Assumption 3 (Support vectors condition)**.** *Let $\mathcal{S}_+$ be the index set of the support vectors associated with nonzero dual variables (formally defined in Appendix D.1). Assume that* $\mathrm{rank}\{\mathbf{x}_i : i \in \mathcal{S}_+\} = \mathrm{rank}\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$.

Assumption 3 is widely used in the literature of logistic regression (Soudry et al., 2018; Ji and Telgarsky, 2021; Wu et al., 2023), requiring the support vectors to be generic. Under this condition, the features $(\mathbf{x}_i)$ can be decomposed into a separable component and a strictly nonseparable component (Wu et al., 2023). Under Assumption 3, our next theorem sharply characterizes the largest convergent stepsizes, with proof deferred to Appendix D.1.

Theorem 7 (The critical stepsize)**.** *Suppose that $\mathbb{H}$ is finite-dimensional and that Assumptions 1 and 3 hold. Consider (GD) for ℓ_2-regularized logistic regression (1). Let $\eta_{\rm crit} := 1/(\lambda \ln(1/\lambda))$. Then there exist $C_1, C_2 > 1$ that only depend on the dataset (but not on λ) such that the following holds. For every $\lambda \le 1/C_1$, we have*

- *If $\eta \le \eta_{\rm crit}/C_2$, then GD converges locally. That is, there exists $r > 0$ such that, for every $\mathbf{w}_0$ satisfying $\|\mathbf{w}_0 - \mathbf{w}_\lambda\| < r$ and every such η, we have $\widetilde{\mathcal{L}}(\mathbf{w}_t) \to \min \widetilde{\mathcal{L}}$.*
- *If $\eta \ge C_2 \eta_{\rm crit}$, then GD diverges for almost every $\mathbf{w}_0$. That is, there exists $\varepsilon > 0$ such that, excluding a measure zero set of $\mathbf{w}_0$, we have $\widetilde{\mathcal{L}}(\mathbf{w}_t) - \min \widetilde{\mathcal{L}} > \varepsilon$ for infinitely many t.*

Theorem 7 suggests that if the stepsize exceeds the critical threshold $\eta_{\rm crit}$ by a constant factor, GD must diverge except with a "lucky" initialization. The critical threshold $\eta_{\rm crit}$ improves the trivial divergent threshold of $2/\lambda$ by a logarithmic factor, and is tight in the sense that GD converges locally for any stepsize smaller than that by a constant factor. This is in sharp contrast to unregularized logistic regression, where GD converges globally for any stepsize (Wu et al., 2023, 2024).

It remains open whether GD converges *globally* with stepsizes of order $1/\sqrt{\lambda} \lesssim \eta \lesssim \eta_{\rm crit}$. In the special case where $\mathbb{H}$ is 1-dimensional, we provide an affirmative answer in Theorem 8 in Appendix D.2 along with a step complexity of $\mathcal{O}\big(\ln(1/(\varepsilon\lambda\ln(1/\lambda)))/(\eta\lambda)\big)$. We also refer the reader to (Meng et al., 2024, 2025) for a fine-grained convergence analysis in this case. In the general finite-dimensional case, we conjecture that the answer is affirmative in the following sense:

Conjecture 1. *Under the setting of Theorem 7, if $\eta \le \eta_{\rm crit}/C_2$ and $\mathbf{w}_0$ is sampled uniformly at random from a unit ball, then GD converges with high probability over the randomness of initialization.*

5 Concluding remarks

We consider gradient descent (GD) with a constant stepsize applied to ℓ_2-regularized logistic regression with linearly separable data. We show that, for a small enough regularization, GD can match the acceleration of Nesterov's momentum by simply using an appropriately large stepsize—with which the objective evolves nonmonotonically. Furthermore, we show that this acceleration brought by large stepsizes holds even under statistical uncertainty. Finally, we calculate the largest possible stepsize with which GD can converge (locally).

This work focuses on the cleanest setup with logistic loss and linear predictors. However, the results presented are ready to be extended to other loss functions (Wu et al., 2024), neural networks in the lazy regime (Wu et al., 2024), and two-layer networks with linearly separable data and bi-Lipschitz activation (Cai et al., 2024). We do not foresee significant new technical challenges here.

There are three future directions worth noting. First, in the context of our paper, does GD converge globally with large stepsizes below the proposed critical threshold? Second, is there a natural statistical learning setting such that GD with large stepsizes generalizes better than GD with small ones? Finally, is there a generic optimization theory for the convergence of GD with large stepsizes? Specifically, is there a general framework to prove the convergence of constant stepsize GD without relying on the descent lemma?

Acknowledgments

We gratefully acknowledge the NSF's support of FODSI through grant DMS-2023505 and of the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and #814639 and of the ONR through MURI award N000142112431. P.M. is supported by a Google PhD Fellowship. The authors are grateful to the Simons Institute for the Theory of Computing for hosting them during parts of this work.

References

Jason M Altschuler and Pablo A Parrilo. Acceleration by stepsize hedging: Multi-step descent and the silver stepsize schedule. *Journal of the ACM*, 72(2):1–38, 2025.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pages 948–1024. PMLR, 2022.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, pages 1497–1537, 2005.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Yuhang Cai, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter Bartlett. Large stepsize gradient descent for non-homogeneous two-layer networks: Margin improvement and fast optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2020.

Jeremy Cohen, Alex Damian, Ameet Talwalkar, J Zico Kolter, and Jason D. Lee. Understanding optimization in deep learning with central flows. In *The Thirteenth International Conference on Learning Representations*, 2025.

Alex Damian, Eshaan Nichani, and Jason D Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *The Eleventh International Conference on Learning Representations*, 2022.

Benjamin Grimmer. Provably faster gradient descent via long steps. *SIAM Journal on Optimization*, 34(3):2588–2608, 2024.

Morris W Hirsch, Stephen Smale, and Robert L Devaney. *Differential equations, dynamical systems, and an introduction to chaos*. Academic press, 2013.

A. Hoorfar and M. Hassani. Inequalities on the Lambert w function and hyperpower function. *JIPAM*, 9(2), 2008.

Daniel Hsu, Vidya Muthukumar, and Ji Xu. On the proliferation of support vectors in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 91–99. PMLR, 2021.

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pages 772–804. PMLR, 2021.

Steven G Krantz and Harold R Parks. *A primer of real analytic functions*. Springer Science & Business Media, 2002.

Si Yi Meng, Antonio Orvieto, Daniel Yiming Cao, and Christopher De Sa. Gradient descent on logistic regression with non-separable data and large step sizes. *arXiv preprint arXiv:2406.05033*, 2024.

Si Yi Meng, Baptiste Goujaud, Antonio Orvieto, and Christopher De Sa. Gradient descent on logistic regression: Do large step-sizes work with data on the sphere? *arXiv preprint arXiv:2507.11228*, 2025.

Rotem Mulayoff, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability: A view from function space. *Advances in Neural Information Processing Systems*, 34:17749–17761, 2021.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer Publishing Company, Incorporated, 2nd edition, 2018. ISBN 3319915770.

Albert BJ Novikoff. On convergence proofs for perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*. New York, NY, 1962.

S. P. Ponomarev. Submersions and preimages of sets of measure zero. *Siberian Mathematical Journal*, 28(1):153–163, 1987. ISSN 1573-9260.

Dan Qiao, Kaiqi Zhang, Esha Singh, Daniel Soudry, and Yu-Xiang Wang. Stable minima cannot overfit in univariate ReLU networks: Generalization by large step sizes. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a regularized path to a maximum margin classifier. *The Journal of Machine Learning Research*, 5:941–973, 2004.

Matan Schliserman and Tomer Koren. Tight risk bounds for gradient descent on separable data. *Advances in Neural Information Processing Systems*, 36, 2024.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Ohad Shamir. Gradient methods never overfit on separable data. *Journal of Machine Learning Research*, 22(85):1–20, 2021.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1): 2822–2878, 2018.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. *Advances in neural information processing systems*, 23, 2010.

Jingfeng Wu, Vladimir Braverman, and Jason D Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. *Advances in Neural Information Processing Systems*, 36: 74229–74256, 2023.

Jingfeng Wu, Peter L Bartlett, Matus Telgarsky, and Bin Yu. Large stepsize gradient descent for logistic loss: Non-monotonicity of the loss improves optimization efficiency. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 5019–5073. PMLR, 2024.

Jingfeng Wu, Peter Bartlett, Matus Telgarsky, and Bin Yu. Benefits of early stopping in gradient descent for overparameterized logistic regression. *arXiv preprint arXiv:2502.13283*, 2025a.

Lei Wu, Chao Ma, and Weinan E. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31, 2018.

Yu-Han Wu, Pierre Marion, Gérard Biau, and Claire Boyer. Taking a big step: Large learning rates in denoising score matching prevent memorization. *arXiv preprint arXiv:2502.03435*, 2025b.

Ruiqi Zhang, Jingfeng Wu, Licong Lin, and Peter L Bartlett. Minimax optimal convergence of gradient descent in logistic regression via large and adaptive stepsizes. *arXiv preprint arXiv:2504.04105*, 2025.

Zihan Zhang, Jason D Lee, Simon S Du, and Yuxin Chen. Anytime acceleration of gradient descent. *arXiv preprint arXiv:2411.17668*, 2024.

Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. In *The Eleventh International Conference on Learning Representations*, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's contributions regarding the acceleration of gradient descent for regularized logistic regression using large stepsizes. The scope, focusing on linearly separable data and the theoretical analysis of convergence rates and stepsize thresholds, is also well-defined.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly discusses limitations of each main result.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper clearly states assumptions in each theoretical result, with complete and rigorous proof referenced and presented in the appendices.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental results are purely illustrative. All necessary information for reproducibility is given in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide an open source implementation of our experiment.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental results are illustrative; we do not claim any statistical significance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The experimental results are illustrative; we do not claim any statistical significance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiments run on a consumer laptop in a few seconds.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The paper is theoretical, with no indication of ethical concerns that would violate the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper is theoretical, with no indication of having potential positive or negative societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper is theoretical and does not contain experimental results.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper is theoretical and does not use external code, data, or models that would require specific license mentions beyond academic citation practices.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper is theoretical and does not introduce new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The paper does not involve LLMs as a component of the core methodology.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Upper bounds on convergence in the EoS regime

Throughout this section, we assume Assumption 1 holds; we set $\mathbf{w}_0 = 0$ without loss of generality. Let $\ell(z) := \ln(1 + \exp(-z))$ be the logistic loss. Define the gradient potential as

$$\mathcal{G}(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} |\ell'(y_i \mathbf{x}_i^\top \mathbf{w})| = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 + \exp(y_i \mathbf{x}_i^\top \mathbf{w})}.$$

Recall that $\mathbf{w}_\lambda := \arg\min \widetilde{\mathcal{L}}(\cdot)$.

We first establish useful lemmas in Appendix A.1. We then prove our first set of upper bounds, Theorem 1 and Corollary 2, in Appendices A.2 and A.3, respectively. Finally, we prove our second set of upper bounds, Theorem 4 and Corollary 5, in Appendices A.4 and A.5, respectively.

A.1 Basic lemmas

We begin with the self-boundedness property.

Lemma 2 (Self-boundedness of the logistic function)**.** *For all $z \in \mathbb{R}$, we have*

$$\ell''(z) < |\ell'(z)| < \ell(z).$$

Proof of Lemma 2. First notice that, for $\alpha > 0$,

$$(1 + \alpha)\ln(1 + \alpha) > \alpha. \tag{2}$$

Indeed, the function $J(\alpha) = (1+\alpha)\ln(1+\alpha) - \alpha$ satisfies $J(0) = 0$ and $J'(\alpha) = \ln(1+\alpha)$, which is positive for $\alpha > 0$. Now, since $\ell'(z) = -1/(1 + \exp(z))$, we have

$$\begin{aligned}
\ell''(z) &= \frac{\exp(z)}{(1 + \exp(z))^2} \\
&< \frac{1}{1 + \exp(z)} = |\ell'(z)| \\
&< \ln(1 + \exp(-z)) = \ell(z),
\end{aligned}$$

where the last inequality uses (2) with $\alpha = \exp(-z)$. □

The following lemma provides bounds on the norm and objective value of $\mathbf{w}_\lambda$.

Lemma 3 (Bounds on the minimizer)**.** *For $\lambda < \gamma^2$, we have*

$$\|\mathbf{w}_\lambda\| \le \frac{\sqrt{2} + \ln(\gamma^2/\lambda)}{\gamma}, \quad \widetilde{\mathcal{L}}(\mathbf{w}_\lambda) \le \frac{\lambda(2 + \ln^2(\gamma^2/\lambda))}{2\gamma^2}.$$

Proof of Lemma 3. For

$$\mathbf{u} := \frac{\ln(\gamma^2/\lambda)}{\gamma}\mathbf{w}^*,$$

we have, by Assumption 1,

$$\mathcal{L}(\mathbf{u}) \le \exp(-\gamma\|\mathbf{u}\|) = \frac{\lambda}{\gamma^2}, \quad \|\mathbf{u}\|^2 = \frac{\ln^2(\gamma^2/\lambda)}{\gamma^2}.$$

Then by definition, we have

$$\widetilde{\mathcal{L}}(\mathbf{w}_\lambda) = \mathcal{L}(\mathbf{w}_\lambda) + \frac{\lambda}{2}\|\mathbf{w}_\lambda\|^2 \le \mathcal{L}(\mathbf{u}) + \frac{\lambda}{2}\|\mathbf{u}\|^2 \le \frac{\lambda\big(2 + \ln^2(\gamma^2/\lambda)\big)}{2\gamma^2}.$$

This completes the proof. □

The following basic facts are due to Assumption 1.

Lemma 4 (Basic facts)**.** *For all $\mathbf{w}$, we have*

1. $\gamma \mathcal{G}(\mathbf{w}) \leq \langle -\nabla \mathcal{L}(\mathbf{w}), \mathbf{w}^* \rangle \leq \mathcal{G}(\mathbf{w})$.
2. $\gamma \mathcal{G}(\mathbf{w}) \leq \|\nabla \mathcal{L}(\mathbf{w})\| \leq \mathcal{G}(\mathbf{w})$.
3. $\|\nabla^2 \mathcal{L}(\mathbf{w})\| \leq \mathcal{G}(\mathbf{w}) \leq \mathcal{L}(\mathbf{w})$.
4. *If* $\mathcal{L}(\mathbf{w}) \leq \ln(2)/n$ *or* $\mathcal{G}(\mathbf{w}) \leq 1/(2n)$, *then* $\mathcal{L}(\mathbf{w}) \leq 2\mathcal{G}(\mathbf{w})$.

Proof of Lemma 4. Since

$$-\nabla \mathcal{L}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i \mathbf{x}_i}{1 + \exp(y_i \mathbf{x}_i^\top \mathbf{w})},$$

the first claim is due to $\gamma \leq y_i \mathbf{x}_i^\top \mathbf{w}^* \leq 1$ by Assumption 1. For the second claim, the lower bound is by the first claim, and the upper bound is by the assumption that $\|\mathbf{x}_i\| \leq 1$. The third claim is due to the self-boundedness of the logistic function (Lemma 2) and the assumption that $\|\mathbf{x}_i\| \leq 1$. In the last claim, both conditions imply that all data are correctly classified, then the claim follows from the fact that $\ln(1 + e^{-t}) \leq e^{-t} \leq 2/(1 + e^t)$ for $t \geq 0$. □

The following lemma suggests that GD aligns well with $\mathbf{w}^*$ throughout the training.

Lemma 5 (Parameter angle)**.** *For* $\lambda \eta < 1$, *we have*

$$\langle \mathbf{w}_t, \mathbf{w}^* \rangle > 0.$$

Proof of Lemma 5. Unrolling (GD) from $\mathbf{w}_0 = 0$, we get

$$\mathbf{w}_t = \sum_{k=0}^{t-1} (1 - \eta\lambda)^{t-1-k} (-\eta \nabla \mathcal{L}(\mathbf{w}_k)).$$

So we have

$$\langle \mathbf{w}_t, \mathbf{w}^* \rangle = \sum_{k=0}^{t-1} (1 - \eta\lambda)^{t-1-k} \langle -\eta \nabla \mathcal{L}(\mathbf{w}_k), \mathbf{w}^* \rangle \geq \gamma\eta \sum_{k=0}^{t-1} (1 - \eta\lambda)^{t-1-k} \mathcal{G}(\mathbf{w}_k) > 0,$$

where the first inequality is by Lemma 4. This completes the proof. □

The following two lemmas are variants of the split optimization lemma introduced by Wu et al. (2024).

Lemma 6 (Split optimization, version 1)**.** *Let* $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2 + \mathbf{u}_3$ *for*

$$\mathbf{u}_2 = \frac{\eta}{\gamma} \mathbf{w}^*, \quad \mathbf{u}_1 = \|\mathbf{u}_1\| \mathbf{w}^*, \quad \mathbf{u}_3 = \|\mathbf{u}_3\| \mathbf{w}^*.$$

For $\lambda \eta < 1$, *we have*

$$\frac{\|\mathbf{w}_t - \mathbf{u}\|^2}{2\eta t} + \frac{\gamma \|\mathbf{u}_3\|}{t} \sum_{k=0}^{t-1} \mathcal{G}(\mathbf{w}_k) + \frac{1}{t} \sum_{k=0}^{t-1} \mathcal{L}(\mathbf{w}_k) \leq \mathcal{L}(\mathbf{u}_1) + \frac{\|\mathbf{u}\|^2}{2\eta t} + \frac{\lambda}{t} \sum_{k=0}^{t-1} \langle \mathbf{w}_k, \mathbf{u} \rangle.$$

Proof of Lemma 6. We use an extended version of the split optimization technique by Wu et al. (2024), which involves three comparators.

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{u}\|^2 &= \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta \langle \nabla \widetilde{\mathcal{L}}(\mathbf{w}_t), \mathbf{u} - \mathbf{w}_t \rangle + \eta^2 \|\nabla \widetilde{\mathcal{L}}(\mathbf{w}_t)\|^2 \\
&= \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta \langle \nabla \mathcal{L}(\mathbf{w}_t) + \lambda \mathbf{w}_t, \mathbf{u} - \mathbf{w}_t \rangle + \eta^2 \|\nabla \mathcal{L}(\mathbf{w}_t) + \lambda \mathbf{w}_t\|^2 \\
&\leq \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta \langle \nabla \mathcal{L}(\mathbf{w}_t) + \lambda \mathbf{w}_t, \mathbf{u} - \mathbf{w}_t \rangle + 2\eta^2 \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2 + 2\eta^2 \lambda^2 \|\mathbf{w}_t\|^2 \\
&\leq \|\mathbf{w}_t - \mathbf{u}\|^2 + 2\eta \langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{u} - \mathbf{w}_t \rangle + 2\eta\lambda \langle \mathbf{w}_t, \mathbf{u} \rangle + 2\eta^2 \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2,
\end{aligned}$$

where the last inequality is because $\lambda \eta < 1$. The choice of $\mathbf{u}_2$ and Lemma 4, parts 1 and 2 imply

$$2\eta \langle -\nabla \mathcal{L}(\mathbf{w}_t), \mathbf{u}_2 \rangle \geq 2\eta^2 \mathcal{G}(\mathbf{w}_t) \geq 2\eta^2 \|\nabla \mathcal{L}(\mathbf{w}_t)\| \geq 2\eta^2 \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2.$$

(See also the proof of Lemma 7 in (Wu et al., 2024).) Then we have

$$\|\mathbf{w}_{t+1}-\mathbf{u}\|^2 \le \|\mathbf{w}_t-\mathbf{u}\|^2 + 2\eta\langle\nabla\mathcal{L}(\mathbf{w}_t),\mathbf{u}_1-\mathbf{w}_t\rangle + 2\eta\langle\nabla\mathcal{L}(\mathbf{w}_t),\mathbf{u}_3\rangle + 2\eta\lambda\langle\mathbf{w}_t,\mathbf{u}\rangle.$$

By convexity and Lemma 4 part 1, we have

$$\|\mathbf{w}_{t+1}-\mathbf{u}\|^2 \le \|\mathbf{w}_t-\mathbf{u}\|^2 + 2\eta\big(\mathcal{L}(\mathbf{u}_1)-\mathcal{L}(\mathbf{w}_t)\big) - 2\eta\gamma\|\mathbf{u}_3\|\mathcal{G}(\mathbf{w}_t) + 2\eta\lambda\langle\mathbf{w}_t,\mathbf{u}\rangle.$$

Telescoping the sum, using $\mathbf{w}_0=0$, and rearranging, we get

$$\frac{\|\mathbf{w}_t-\mathbf{u}\|^2}{2\eta t} + \frac{\gamma\|\mathbf{u}_3\|}{t}\sum_{k=0}^{t-1}\mathcal{G}(\mathbf{w}_k) + \frac{1}{t}\sum_{k=0}^{t-1}\mathcal{L}(\mathbf{w}_k) \le \mathcal{L}(\mathbf{u}_1) + \frac{\|\mathbf{u}\|^2}{2\eta t} + \lambda\left\langle\frac{1}{t}\sum_{k=0}^{t-1}\mathbf{w}_k,\mathbf{u}\right\rangle,$$

which completes the proof. □

Lemma 7 (Split optimization, version 2). *Let* $\mathbf{u}=\mathbf{u}_1+\mathbf{u}_2$ *for*

$$\mathbf{u}_2 = \frac{\eta}{2\gamma(1-\eta\lambda)}\mathbf{w}^*, \quad \mathbf{u}_1 = \|\mathbf{u}_1\|\mathbf{w}^*.$$

For $\lambda\eta<1$*, we have*

$$\|\mathbf{w}_t-\mathbf{u}\|^2 \le (1-\eta\lambda)^t\|\mathbf{u}\|^2 + 2\eta\sum_{k=0}^{t-1}(1-\eta\lambda)^{t-1-k}\Big((1-\eta\lambda)\big(\mathcal{L}(\mathbf{u}_1)-\mathcal{L}(\mathbf{w}_k)\big)+\lambda\|\mathbf{u}\|^2\Big).$$

Proof of Lemma 7. Recall that

$$\mathbf{w}_{t+1}-\mathbf{u} = (1-\eta\lambda)(\mathbf{w}_t-\mathbf{u}) - \eta\lambda\mathbf{u} - \eta\nabla\mathcal{L}(\mathbf{w}_t).$$

Taking the squared norm and expanding, we have

$$\begin{aligned}\|\mathbf{w}_{t+1}-\mathbf{u}\|^2 &= (1-\eta\lambda)^2\|\mathbf{w}_t-\mathbf{u}\|^2 + 2\eta(1-\eta\lambda)\langle\nabla\mathcal{L}(\mathbf{w}_t),\mathbf{u}-\mathbf{w}_t\rangle + \eta^2\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2\\ &\quad + \eta^2\lambda^2\|\mathbf{u}\|^2 + 2\eta\lambda(1-\eta\lambda)\langle\mathbf{u},\mathbf{u}-\mathbf{w}_t\rangle + 2\eta^2\lambda\langle\mathbf{u},\nabla\mathcal{L}(\mathbf{w}_t)\rangle\\ &= (1-\eta\lambda)^2\|\mathbf{w}_t-\mathbf{u}\|^2 + 2\eta(1-\eta\lambda)\langle\nabla\mathcal{L}(\mathbf{w}_t),\mathbf{u}-\mathbf{w}_t\rangle + \eta^2\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2\\ &\quad + \eta\lambda(2-\eta\lambda)\|\mathbf{u}\|^2 - 2\eta\lambda(1-\eta\lambda)\langle\mathbf{u},\mathbf{w}_t\rangle + 2\eta^2\lambda\langle\mathbf{u},\nabla\mathcal{L}(\mathbf{w}_t)\rangle\,.\end{aligned}$$

The sum of the last two terms of the previous identity is negative,

$$-2\eta\lambda(1-\eta\lambda)\langle\mathbf{u},\mathbf{w}_t\rangle + 2\eta^2\lambda\langle\mathbf{u},\nabla\mathcal{L}(\mathbf{w}_t)\rangle = -2\eta\lambda\langle\mathbf{u},\mathbf{w}_{t+1}\rangle < 0,$$

where the last inequality is by Lemma 5. Moreover, the choice of $\mathbf{u}_2$ and Lemma 4, parts 1 and 2 imply that (see also the proof of Lemma 7 in (Wu et al., 2024))

$$2\eta(1-\eta\lambda)\langle\nabla\mathcal{L}(\mathbf{w}_t),\mathbf{u}_2\rangle + \eta^2\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2 \le 0.$$

So we have

$$\begin{aligned}\|\mathbf{w}_{t+1}-\mathbf{u}\|^2 &\le (1-\eta\lambda)^2\|\mathbf{w}_t-\mathbf{u}\|^2 + 2\eta(1-\eta\lambda)\langle\nabla\mathcal{L}(\mathbf{w}_t),\mathbf{u}_1-\mathbf{w}_t\rangle + \eta\lambda(2-\eta\lambda)\|\mathbf{u}\|^2\\ &\le (1-\eta\lambda)^2\|\mathbf{w}_t-\mathbf{u}\|^2 + 2\eta(1-\eta\lambda)\big(\mathcal{L}(\mathbf{u}_1)-\mathcal{L}(\mathbf{w}_t)\big) + \eta\lambda(2-\eta\lambda)\|\mathbf{u}\|^2\\ &\le (1-\eta\lambda)\|\mathbf{w}_t-\mathbf{u}\|^2 + 2\eta(1-\eta\lambda)\big(\mathcal{L}(\mathbf{u}_1)-\mathcal{L}(\mathbf{w}_t)\big) + 2\eta\lambda\|\mathbf{u}\|^2.\end{aligned}$$

Unrolling the recursion, we get

$$\|\mathbf{w}_t-\mathbf{u}\|^2 \le (1-\eta\lambda)^t\|\mathbf{u}\|^2 + 2\eta\sum_{k=0}^{t-1}(1-\eta\lambda)^{t-1-k}\Big((1-\eta\lambda)\big(\mathcal{L}(\mathbf{u}_1)-\mathcal{L}(\mathbf{w}_k)\big)+\lambda\|\mathbf{u}\|^2\Big).$$

This completes the proof. □

Based on these split optimization bounds, the following three lemmas establish bounds on parameter norm, gradient potential, and the logistic empirical risk, respectively.

Lemma 8 (A parameter bound). *For* $\eta\lambda\le 1/2$*, we have*

$$\|\mathbf{w}_t\| \le 4\frac{\eta+\ln(e+\gamma^2\min\{\eta t,1/\lambda\})}{\gamma}.$$

Proof of Lemma 8. By Lemma 7, we have

$$\begin{aligned}\|\mathbf{w}_t - \mathbf{u}\|^2 &\le (1-\eta\lambda)^t\|\mathbf{u}\|^2 + \left(2\eta\sum_{k=0}^{t-1}(1-\eta\lambda)^k\right)\left((1-\eta\lambda)\mathcal{L}(\mathbf{u}_1) + \lambda\|\mathbf{u}\|^2\right)\\ &= (1-\eta\lambda)^t\|\mathbf{u}\|^2 + 2\frac{1-(1-\eta\lambda)^t}{\lambda}\left((1-\eta\lambda)\mathcal{L}(\mathbf{u}_1) + \lambda\|\mathbf{u}\|^2\right)\\ &\le 2\frac{1-(1-\eta\lambda)^t}{\lambda}\mathcal{L}(\mathbf{u}_1) + 2\|\mathbf{u}\|^2\\ &\le 2\min\{\eta t, 1/\lambda\}\mathcal{L}(\mathbf{u}_1) + 2\|\mathbf{u}\|^2.\end{aligned}$$

In the final inequality, the proof that $1-(1-\eta\lambda)^t \le \eta\lambda t$ is by induction. For $\eta\lambda \le 1/2$ and

$$\mathbf{u}_1 = \frac{\ln(e+\gamma^2\min\{\eta t, 1/\lambda\})}{\gamma}\mathbf{w}^*, \quad \mathbf{u}_2 = \frac{\eta}{2\gamma(1-\eta\lambda)}\mathbf{w}^*,$$

we have

$$\|\mathbf{u}_2\| \le \frac{\eta}{\gamma}, \quad \mathcal{L}(\mathbf{u}_1) \le \exp(-\gamma\|\mathbf{u}_1\|) \le \frac{1}{\gamma^2\min\{\eta t, 1/\lambda\}}.$$

Combining, we have

$$\begin{aligned}\|\mathbf{w}_t\| &\le \|\mathbf{w}_t - \mathbf{u}\| + \|\mathbf{u}\|\\ &\le \sqrt{2\min\{\eta t, 1/\lambda\}\mathcal{L}(\mathbf{u}_1)} + \left(\sqrt{2}+1\right)\|\mathbf{u}\|\\ &\le \frac{\sqrt{2}}{\gamma} + \left(\sqrt{2}+1\right)\frac{\eta + \ln(e+\gamma^2\min\{\eta t, 1/\lambda\})}{\gamma}\\ &\le 4\frac{\eta + \ln(e+\gamma^2\min\{\eta t, 1/\lambda\})}{\gamma}.\end{aligned}$$

This completes the proof. □

Lemma 9 (A gradient potential bound)**.** *For $\eta\lambda \le 1/2$, we have*

$$\frac{1}{t}\sum_{k=0}^{t-1}\mathcal{G}(\mathbf{w}_k) \le 11\frac{\eta + \ln(e+\gamma^2\min\{\eta t, 1/\lambda\})}{\gamma^2\min\{\eta t, 1/\lambda\}}.$$

Proof of Lemma 9. Let $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2 + \mathbf{u}_3$ and

$$\mathbf{u}_1 = \frac{\ln(e+\gamma^2\min\{\eta t, 1/\lambda\})}{\gamma}\mathbf{w}^*, \quad \mathbf{u}_2 = \frac{\eta}{\gamma}\mathbf{w}^*, \quad \mathbf{u}_3 = \frac{\eta + \ln(e+\gamma^2\min\{\eta t, 1/\lambda\})}{\gamma}\mathbf{w}^*.$$

Then we have

$$\|\mathbf{u}_3\| \ge \frac{1}{\gamma}, \quad \|\mathbf{u}\| = 2\|\mathbf{u}_3\|, \quad \mathcal{L}(\mathbf{u}_1) \le \frac{1}{\gamma^2\min\{\eta t, 1/\lambda\}}.$$

Moreover, Lemma 8 yields

$$\max_{k\le t}\|\mathbf{w}_k\| \le 4\|\mathbf{u}_3\|.$$

Using Lemma 6, we have

$$\begin{aligned}\frac{1}{t}\sum_{k=0}^{t-1}\mathcal{G}(\mathbf{w}_k) &\le \frac{1}{\gamma\|\mathbf{u}_3\|}\left(\mathcal{L}(\mathbf{u}_1) + \frac{\|\mathbf{u}\|^2}{2\eta t} + \lambda\|\mathbf{u}\|\max_{k\le t}\|\mathbf{w}_k\|\right)\\ &\le \frac{1}{\gamma}\left(\frac{1}{\|\mathbf{u}_3\|\gamma^2\min\{\eta t, 1/\lambda\}} + \frac{2\|\mathbf{u}_3\|}{\eta t} + 8\lambda\|\mathbf{u}_3\|\right)\\ &\le \frac{1}{\gamma}\left(\frac{1}{\gamma\min\{\eta t, 1/\lambda\}} + \left(\frac{2}{\eta t} + 8\lambda\right)\frac{\eta + \ln(e+\gamma^2\min\{\eta t, 1/\lambda\})}{\gamma}\right)\\ &\le 11\frac{\eta + \ln(e+\gamma^2\min\{\eta t, 1/\lambda\})}{\gamma^2\min\{\eta t, 1/\lambda\}}.\end{aligned}$$

This completes the proof. □

Lemma 10 (A logistic empirical risk bound)**.** *For $\eta\lambda \le 1/2$, we have*

$$\frac{1}{t}\sum_{k=0}^{t-1}\mathcal{L}(\mathbf{w}_k) \le 10\frac{\eta^2 + \ln^2(e+\gamma^2\min\{\eta t, 1/\lambda\})}{\gamma^2\min\{\eta t, 1/\lambda\}}.$$

Proof of Lemma 10. Let $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2 + \mathbf{u}_3$ with

$$\mathbf{u}_1 = \frac{\ln(e+\gamma^2\min\{\eta t, 1/\lambda\})}{\gamma}\mathbf{w}^*, \quad \mathbf{u}_2 = \frac{\eta}{\gamma}\mathbf{w}^*, \quad \mathbf{u}_3 = 0.$$

Then we have

$$\mathcal{L}(\mathbf{u}_1) \le \frac{1}{\gamma^2\min\{\eta t, 1/\lambda\}}, \quad \|\mathbf{u}\| = \frac{\eta + \ln(e+\gamma^2\min\{\eta t, 1/\lambda\})}{\gamma}.$$

By Lemma 8, we have

$$\max_{k\le t}\|\mathbf{w}_k\| \le 4\|\mathbf{u}\|.$$

Using Lemma 6, we have

$$\begin{aligned}
\frac{1}{t}\sum_{k=0}^{t-1}\mathcal{L}(\mathbf{w}_k) &\le \mathcal{L}(\mathbf{u}_1) + \frac{\|\mathbf{u}\|^2}{2\eta t} + \lambda\|\mathbf{u}\|\max_{k\le t}\|\mathbf{w}_k\| \\
&\le \mathcal{L}(\mathbf{u}_1) + \left(\frac{1}{2\eta t} + 4\lambda\right)\|\mathbf{u}\|^2 \\
&\le \frac{1}{\gamma^2\min\{\eta t, 1/\lambda\}} + 2\left(\frac{1}{2\eta t} + 4\lambda\right)\frac{\eta^2 + \ln^2(e+\gamma^2\min\{\eta t, 1/\lambda\})}{\gamma^2} \\
&\le 10\frac{\eta^2 + \ln^2(e+\gamma^2\min\{\eta t, 1/\lambda\})}{\gamma^2\min\{\eta t, 1/\lambda\}}.
\end{aligned}$$

This completes the proof. □

The next lemma shows that when the gradient potential is small, it remains small under one step of GD (even when the stepsize is large).

Lemma 11 (Small gradient potential)**.** *Assume that*

$$\lambda \le \frac{1}{3\eta\ln(e+\eta)}.$$

If in the t-th step we have

$$\mathcal{G}(\mathbf{w}_t) \le \frac{1}{2e^2\eta},$$

then for every $\mathbf{v}$ in the line segment between $\mathbf{w}_t$ and $\mathbf{w}_{t+1}$, we have

$$\mathcal{G}(\mathbf{v}) \le \frac{1}{2\eta}.$$

Proof of Lemma 11. There exists an $\alpha \in [0,1]$ such that $\mathbf{v} = \alpha\mathbf{w}_{t+1} + (1-\alpha)\mathbf{w}_t$. Then for every $1 \le i \le n$, we have

$$\begin{aligned}
y_i\mathbf{x}_i^\top\mathbf{v} &= y_i\mathbf{x}_i^\top(\alpha((1-\eta\lambda)\mathbf{w}_t - \eta\nabla\mathcal{L}(\mathbf{w}_t)) + (1-\alpha)\mathbf{w}_t) \\
&= (1-\alpha\lambda\eta)y_i\mathbf{x}_i^\top\mathbf{w}_t - \alpha\eta y_i\mathbf{x}_i^\top\nabla\mathcal{L}(\mathbf{w}_t) \\
&\ge (1-\alpha\lambda\eta)y_i\mathbf{x}_i^\top\mathbf{w}_t - \eta\|\nabla\mathcal{L}(\mathbf{w}_t)\| \\
&\ge (1-\alpha\lambda\eta)y_i\mathbf{x}_i^\top\mathbf{w}_t - \eta\mathcal{G}(\mathbf{w}_t) \\
&\ge (1-\alpha\lambda\eta)y_i\mathbf{x}_i^\top\mathbf{w}_t - 1,
\end{aligned}$$

where the second inequality is by Lemma 4 and the last inequality is because $\mathcal{G}(\mathbf{w}_t) \le 1/\eta$. So we have

$$\begin{aligned}\mathcal{G}(\mathbf{v}) &\le \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+\exp((1-\alpha\lambda\eta)y_i\mathbf{x}_i^\top\mathbf{w}_t - 1)} \\ &\le \frac{e}{n}\sum_{i=1}^{n}\frac{1}{1+\exp((1-\alpha\lambda\eta)y_i\mathbf{x}_i^\top\mathbf{w}_t)} \\ &= \frac{e}{n}\sum_{i=1}^{n}\frac{1}{1+\exp(y_i\mathbf{x}_i^\top\mathbf{w}_t)^{1-\alpha\lambda\eta}}\,.\end{aligned}$$

Recall the following inequality:

$$1 + x^\beta \ge (1+x)^\beta \;\text{ for }\; x \ge 0 \text{ and } 0 < \beta < 1.$$

We see this by verifying that the function $x \mapsto 1 + x^\beta - (1+x)^\beta$ is increasing for $x > 0$ and maps 0 to 0. Applying this inequality and the concavity of the function $x \mapsto x^\beta$ for $0 < \beta < 1$ and $x > 0$, we obtain

$$\begin{aligned}\mathcal{G}(\mathbf{v}) &\le \frac{e}{n}\sum_{i=1}^{n}\left(\frac{1}{1+\exp(y_i\mathbf{x}_i^\top\mathbf{w}_t)}\right)^{1-\alpha\lambda\eta} \\ &\le e\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+\exp(y_i\mathbf{x}_i^\top\mathbf{w}_t)}\right)^{1-\alpha\lambda\eta} \\ &= e\mathcal{G}(\mathbf{w}_t)^{1-\alpha\lambda\eta}.\end{aligned}$$

Using the assumption on $\mathcal{G}(\mathbf{w}_t)$, we get

$$\mathcal{G}(\mathbf{v}) \le e\left(\frac{1}{2e^2\eta}\right)^{(1-\alpha\lambda\eta)} = \frac{1}{2e\eta}\big(2e^2\eta\big)^{\alpha\lambda\eta} = \frac{1}{2e\eta}\exp\big(\alpha\lambda\eta\ln(2e^2\eta)\big) \le \frac{1}{2\eta},$$

where the last inequality is because $\alpha\lambda\eta\ln(2e^2\eta) \le 1$, which is verified by discussing two cases. If $2e^2\eta \le 1$, this is trivial; If $2e^2\eta > 1$, this follows from our assumption on λ and $\alpha \le 1$:

$$\alpha\lambda\eta\ln(2e^2\eta) \le \lambda\eta\ln(2e^2\eta) \le \frac{\ln(2e^2\eta)}{3\ln(e+\eta)} \le 1.$$

This completes the proof. □

The following lemma shows that if the gradient potential is small, then the objective value decreases after one step of GD.

Lemma 12 (One contraction step)**.** *Assume that*

$$\lambda \le \frac{1}{3\eta\ln(e+\eta)}.$$

If in the t-th step we have

$$\mathcal{G}(\mathbf{w}_t) \le \frac{1}{2e^2\eta},$$

then we have

$$\widetilde{\mathcal{L}}(\mathbf{w}_{t+1}) \le \widetilde{\mathcal{L}}(\mathbf{w}_t) - \frac{\eta}{2}\|\nabla\widetilde{\mathcal{L}}(\mathbf{w}_t)\|^2.$$

Furthermore, we have

$$\widetilde{\mathcal{L}}(\mathbf{w}_{t+1}) - \min\widetilde{\mathcal{L}} \le (1-\eta\lambda)\big(\widetilde{\mathcal{L}}(\mathbf{w}_t) - \min\widetilde{\mathcal{L}}\big) \;\text{ and }\; \|\mathbf{w}_{t+1} - \mathbf{w}_\lambda\|^2 \le (1-\eta\lambda)\|\mathbf{w}_t - \mathbf{w}_\lambda\|^2.$$

Proof of Lemma 12. There exists $\mathbf{v}$ in the line segment between $\mathbf{w}_t$ and $\mathbf{w}_{t+1}$ such that

$$\widetilde{\mathcal{L}}(\mathbf{w}_{t+1}) = \widetilde{\mathcal{L}}(\mathbf{w}_t) + \langle\nabla\widetilde{\mathcal{L}}(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + \frac{1}{2}\big\langle\nabla^2\widetilde{\mathcal{L}}(\mathbf{v}), (\mathbf{w}_{t+1} - \mathbf{w}_t)^{\otimes 2}\big\rangle$$

$$\leq \widetilde{\mathcal{L}}(\mathbf{w}_t) - \eta\|\nabla\widetilde{\mathcal{L}}(\mathbf{w}_t)\|^2\left(1 - \frac{\eta\|\nabla^2\widetilde{\mathcal{L}}(\mathbf{v})\|}{2}\right)$$
$$= \widetilde{\mathcal{L}}(\mathbf{w}_t) - \eta\|\nabla\widetilde{\mathcal{L}}(\mathbf{w}_t)\|^2\left(1 - \frac{\eta(\lambda + \|\nabla^2\mathcal{L}(\mathbf{v})\|)}{2}\right).$$

Our assumption on λ implies that $\lambda\eta \leq 1/2$. Then by Lemmas 4 and 11, we have

$$\frac{\eta(\lambda + \|\nabla^2\mathcal{L}(\mathbf{v})\|)}{2} \leq \frac{\eta(\lambda + \mathcal{G}(\mathbf{v}))}{2} \leq \frac{0.5 + 0.5}{2} = \frac{1}{2},$$

which leads to

$$\widetilde{\mathcal{L}}(\mathbf{w}_{t+1}) \leq \widetilde{\mathcal{L}}(\mathbf{w}_t) - \eta\|\nabla\widetilde{\mathcal{L}}(\mathbf{w}_t)\|^2\left(1 - \frac{\eta(\lambda + \|\nabla^2\mathcal{L}(\mathbf{v})\|)}{2}\right) \leq \widetilde{\mathcal{L}}(\mathbf{w}_t) - \frac{\eta}{2}\|\nabla\widetilde{\mathcal{L}}(\mathbf{w}_t)\|^2.$$

The risk contraction follows from the above and the well-known Polyak-Lojasiewicz inequality from the λ-strong convexity:

$$\widetilde{\mathcal{L}}(\mathbf{w}) - \min\widetilde{\mathcal{L}} \leq \frac{1}{2\lambda}\|\nabla\widetilde{\mathcal{L}}(\mathbf{w})\|^2.$$

The norm contraction is because

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}_\lambda\|^2 \\
&= \|\mathbf{w}_t - \mathbf{w}_\lambda\|^2 + 2\eta\langle\nabla\widetilde{\mathcal{L}}(\mathbf{w}_t), \mathbf{w}_\lambda - \mathbf{w}_t\rangle + \eta^2\|\nabla\widetilde{\mathcal{L}}(\mathbf{w}_t)\|^2 \\
&\leq \|\mathbf{w}_t - \mathbf{w}_\lambda\|^2 + 2\eta\left(\widetilde{\mathcal{L}}(\mathbf{w}_\lambda) - \widetilde{\mathcal{L}}(\mathbf{w}_t) - \frac{\lambda}{2}\|\mathbf{w}_t - \mathbf{w}_\lambda\|^2\right) + \eta^2\left(\frac{2}{\eta}\big(\widetilde{\mathcal{L}}(\mathbf{w}_t) - \widetilde{\mathcal{L}}(\mathbf{w}_{t+1})\big)\right) \\
&= (1 - \eta\lambda)\|\mathbf{w}_t - \mathbf{w}_\lambda\|^2 + 2\eta\big(\widetilde{\mathcal{L}}(\mathbf{w}_\lambda) - \widetilde{\mathcal{L}}(\mathbf{w}_{t+1})\big) \\
&\leq (1 - \eta\lambda)\|\mathbf{w}_t - \mathbf{w}_\lambda\|^2,
\end{aligned}$$

where the first inequality is by λ-strong convexity and the first claim, and the second inequality is because $\mathbf{w}_\lambda := \arg\min\widetilde{\mathcal{L}}(\cdot)$. This completes our proof. □

A.2 Proof of Theorem 1

The following lemma is crucial for showing that GD remains in the stable phase.

Lemma 13 (Stable phase)**.** *Assume that $n \geq 2$ and*

$$\lambda \leq \frac{\gamma^2}{C_1}\min\left\{\frac{1}{n\ln n}, \frac{1}{n\eta}, \frac{1}{\eta^2}\right\}$$

for a large constant $C_1 > 1$. If in the s-th step we have

$$\mathcal{L}(\mathbf{w}_s) \leq \min\left\{\frac{1}{2e^2\eta}, \frac{\ln 2}{n}\right\},$$

then for all $t \geq s$ we have

$$\mathcal{L}(\mathbf{w}_t) \leq \min\left\{\frac{1}{2e^2\eta}, \frac{\ln 2}{n}\right\}.$$

Proof of Lemma 13. The condition on λ with $C_1 \geq 6$ guarantees that

$$\lambda \leq \frac{1}{6}\min\left\{\frac{1}{n\ln n}, \frac{1}{\eta^2}\right\} \leq \frac{1}{6}\min\left\{1, \frac{1}{\eta^2}\right\} \leq \frac{1}{3(1+\eta^2)} \leq \frac{1}{3\eta\ln(e+\eta)} \leq \frac{1}{2\eta}.$$

which satisfies the condition on λ required by Lemmas 8, 11 and 12.

We prove the claim by induction. The claim holds for s. Assume the claim holds for t. We then verify the claim for $t+1$. By Taylor's theorem, there exists $\mathbf{v}$ within the line segment between $\mathbf{w}_t$ and $\mathbf{w}_{t+1}$ such that

$$\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_t) = \langle\nabla\mathcal{L}(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + \frac{1}{2}\big\langle\nabla^2\mathcal{L}(\mathbf{v}), (\mathbf{w}_{t+1} - \mathbf{w}_t)^{\otimes 2}\big\rangle$$

$$= -\eta\langle \nabla\mathcal{L}(\mathbf{w}_t), \nabla\mathcal{L}(\mathbf{w}_t) + \lambda\mathbf{w}_t\rangle + \frac{\eta^2}{2}\langle \nabla^2\mathcal{L}(\mathbf{v}), (\nabla\mathcal{L}(\mathbf{w}_t) + \lambda\mathbf{w}_t)^{\otimes 2}\rangle$$
$$\le -\eta\langle \nabla\mathcal{L}(\mathbf{w}_t), \nabla\mathcal{L}(\mathbf{w}_t) + \lambda\mathbf{w}_t\rangle + \frac{\eta^2}{2}\mathcal{G}(\mathbf{v})\|\nabla\mathcal{L}(\mathbf{w}_t) + \lambda\mathbf{w}_t\|^2.$$

where the last inequality is because $\|\nabla^2\mathcal{L}(\mathbf{v})\| \le \mathcal{G}(\mathbf{v})$ by Lemma 4. The induction hypothesis and Lemma 4 imply that

$$\mathcal{G}(\mathbf{w}_t) \le \mathcal{L}(\mathbf{w}_t) \le \min\left\{\frac{1}{2e^2\eta}, \frac{\ln 2}{n}\right\}, \tag{3}$$

which implies $\mathcal{G}(\mathbf{v}) \le 2/\eta \le 1/\eta$ by Lemma 11. Then we have

$$\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_t) \le -\eta\langle \nabla\mathcal{L}(\mathbf{w}_t), \nabla\mathcal{L}(\mathbf{w}_t) + \lambda\mathbf{w}_t\rangle + \frac{\eta}{2}\|\nabla\mathcal{L}(\mathbf{w}_t) + \lambda\mathbf{w}_t\|^2$$
$$= -\frac{\eta}{2}\Big(\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2 - \lambda\|\mathbf{w}_t\|^2\Big).$$

We discuss two cases. If $\|\nabla\mathcal{L}(\mathbf{w}_t)\| \ge \lambda\|\mathbf{w}_t\|$, then $\mathcal{L}(\mathbf{w}_{t+1}) \le \mathcal{L}(\mathbf{w}_t)$, which together with (3) verifies the claim for $t+1$. If $\|\nabla\mathcal{L}(\mathbf{w}_t)\| < \lambda\|\mathbf{w}_t\|$, then

$$\begin{aligned}\mathcal{L}(\mathbf{w}_{t+1}) &\le \mathcal{L}(\mathbf{w}_t) + \frac{\eta\lambda^2\|\mathbf{w}_t\|^2}{2}\\ &\le 2\mathcal{G}(\mathbf{w}_t) + \frac{\eta\lambda^2\|\mathbf{w}_t\|^2}{2}\\ &\le \frac{2}{\gamma}\|\nabla\mathcal{L}(\mathbf{w}_t)\| + \frac{\eta\lambda^2\|\mathbf{w}_t\|^2}{2}\\ &\le \frac{\lambda\|\mathbf{w}_t\|}{\gamma}\left(2 + \frac{\gamma\eta\lambda\|\mathbf{w}_t\|}{2}\right),\end{aligned}$$

where the second inequality is by Lemma 4 and (3), the third inequality is by Lemma 4, and the fourth inequality is because $\|\nabla\mathcal{L}(\mathbf{w}_t)\| < \lambda\|\mathbf{w}_t\|$. By Lemma 8, we have

$$\|\mathbf{w}_t\| \le 4\frac{\eta + \ln(e + \gamma^2\min\{\eta t, 1/\lambda\})}{\gamma} \le 4\frac{\eta + \ln(e + \gamma^2/\lambda)}{\gamma} =: M.$$

Then we have

$$\mathcal{L}(\mathbf{w}_{t+1}) \le \frac{\lambda M}{\gamma}\left(2 + \frac{\gamma\eta\lambda M}{2}\right) \le 3\frac{\lambda M}{\gamma} \le \min\left\{\frac{1}{2e^2\eta}, \frac{\ln 2}{n}\right\},$$

which verifies the claim for $t+1$. Here, the last inequality is by (4) (proved below), and the second inequality is because $\eta\lambda M \le 1$ by (4).

$$\frac{3\lambda M}{\gamma} = 12\lambda\frac{\eta + \ln(e + \gamma^2/\lambda)}{\gamma^2} \le \min\left\{\frac{1}{2e^2\eta}, \frac{\ln 2}{n}\right\} \tag{4}$$

$$\Leftarrow \quad \frac{\eta + \ln(e + \gamma^2/\lambda)}{\gamma^2/\lambda} \le \frac{1}{K_1}\min\left\{\frac{1}{\eta}, \frac{1}{n}\right\} \quad \text{for a sufficiently large constant } K_1 \tag{5}$$

$$\Leftrightarrow \quad \frac{\gamma^2}{\lambda} \ge K_1\big(\eta^2 + \eta\ln(e + \gamma^2/\lambda)\big) \text{ and } \frac{\gamma^2}{\lambda} \ge K_1\big(n\eta + n\ln(e + \gamma^2/\lambda)\big)$$

$$\Leftarrow \quad \frac{\gamma^2}{\lambda} \ge C_1\max\{1, \eta^2\} \text{ and } \frac{\gamma^2}{\lambda} \ge C_1\max\{n\eta, n\ln n\} \text{ for a sufficiently large constant } C_1$$

$$\Leftrightarrow \quad \frac{\gamma^2}{\lambda} \ge C_1\max\{n\ln n, n\eta, \eta^2\}.$$

This completes the proof. □

The next lemma provides a bound on the phase transition time.

Lemma 14 (Phase transition)**.** *Assume that $n \ge 2$ and*

$$\lambda \le \frac{\gamma^2}{C_1}\min\left\{\frac{1}{n\ln n}, \frac{1}{n\eta}, \frac{1}{\eta^2}\right\}$$

for a large constant $C_1 > 1$. Let

$$\tau := \frac{C_2}{\gamma^2} \max\left\{\eta,\, n,\, \frac{n \ln n}{\eta}\right\}$$

for a large constant $C_2 > 1$. Then for all $t \geq \tau$,

$$\mathcal{G}(\mathbf{w}_t) \leq \mathcal{L}(\mathbf{w}_t) \leq \min\left\{\frac{1}{2e^2\eta},\, \frac{\ln 2}{n}\right\}.$$

Proof of Lemma 14. The assumption on λ with $C_1 \geq 2$ implies $\eta\lambda \leq 1/2$. Then by Lemma 9 we have

$$\frac{1}{\tau}\sum_{k=0}^{\tau-1} \mathcal{G}(\mathbf{w}_k) \leq 11 \frac{\eta + \ln(e + \gamma^2 \min\{\eta\tau, 1/\lambda\})}{\gamma^2 \min\{\eta\tau, 1/\lambda\}}.$$

If the right-hand side is smaller than $\min\{1/(4e^2\eta), \ln(2)/(2n)\}$, then there exists $s \leq \tau$ such that

$$\mathcal{G}(\mathbf{w}_s) \leq \min\left\{\frac{1}{4e^2\eta},\, \frac{\ln 2}{2n}\right\} \leq \frac{1}{2n}.$$

This further implies $\mathcal{L}(\mathbf{w}_s) \leq 2\mathcal{G}(\mathbf{w}_s) \leq \min\{1/(2e^2\eta), \ln(2)/n\}$ by Lemma 4, and then Lemmas 4 and 13 imply the result. So it suffices to check that

$$\begin{aligned}
& 11 \frac{\eta + \ln(e + \gamma^2 \min\{\eta\tau, 1/\lambda\})}{\gamma^2 \min\{\eta\tau, 1/\lambda\}} \leq \min\left\{\frac{1}{4e^2\eta},\, \frac{\ln 2}{2n}\right\} \\
\Leftarrow \quad & \begin{cases} \dfrac{\eta + \ln(e + \gamma^2/\lambda)}{\gamma^2/\lambda} \leq \dfrac{1}{K_1} \min\left\{\dfrac{1}{\eta}, \dfrac{1}{n}\right\}, \\ \dfrac{\eta + \ln(e + \gamma^2\eta\tau)}{\gamma^2\eta\tau} \leq \dfrac{1}{K_1} \min\left\{\dfrac{1}{\eta}, \dfrac{1}{n}\right\}, \end{cases}
\end{aligned}$$

for a sufficiently large constant K_1. As shown in (5) in the proof of Lemma 13, the first condition follows from our assumption on λ. The second condition is equivalent to

$$\begin{aligned}
& \gamma^2\tau \geq K_1\big(\eta + \ln(e + \gamma^2\eta\tau)\big) \text{ and } \gamma^2\tau \geq K_1\left(n + \frac{n}{\eta}\ln(e + \gamma^2\eta\tau)\right) \\
\Leftarrow \quad & \gamma^2\tau \geq C_2 \max\{1, \eta\} \text{ and } \gamma^2\tau \geq C_2 \max\left\{n, \frac{n\ln n}{\eta}\right\} \text{ for a sufficiently large constant } C_2 \\
\Leftrightarrow \quad & \gamma^2\tau \geq C_2 \max\left\{\eta, n, \frac{n\ln n}{\eta}\right\}.
\end{aligned}$$

This completes the proof. □

With the above lemmas, we are ready to prove Theorem 1.

Proof of Theorem 1. Our assumption on λ and η satisfies the condition on λ and η required by Lemma 14. That condition with $C_1 \geq 6$ implies that $\lambda \leq 1/(3(1+\eta^2)) \leq 1/(3\eta\ln(e+\eta)) \leq 1/(2\eta)$, satisfying the condition on λ and η required by Lemmas 8 and 12.

The phase transition time bound is by Lemma 14, which further enables Lemma 12 for all $t \geq \tau$. Thus we have

$$\begin{aligned}
\widetilde{\mathcal{L}}(\mathbf{w}_{\tau+t}) - \min\widetilde{\mathcal{L}} &\leq (1-\lambda\eta)^t\big(\widetilde{\mathcal{L}}(\mathbf{w}_\tau) - \min\widetilde{\mathcal{L}}\big) \leq \exp(-\lambda\eta t)\big(\widetilde{\mathcal{L}}(\mathbf{w}_\tau) - \min\widetilde{\mathcal{L}}\big), \\
\|\mathbf{w}_{\tau+t} - \mathbf{w}_\lambda\|^2 &\leq (1-\lambda\eta)^t\|\mathbf{w}_\tau - \mathbf{w}_\lambda\|^2 \leq \exp(-\lambda\eta t)\|\mathbf{w}_\tau - \mathbf{w}_\lambda\|^2.
\end{aligned}$$

It remains to bound $\widetilde{\mathcal{L}}(\mathbf{w}_\tau) - \min\widetilde{\mathcal{L}}$ and $\|\mathbf{w}_\tau - \mathbf{w}_\lambda\|$.

Our assumption on λ implies $\gamma^2/\lambda \geq C_1 \geq e$. Then by Lemma 8 we have

$$\|\mathbf{w}_\tau\| \leq 4\frac{\eta + \ln(e + \gamma^2\min\{\eta s, 1/\lambda\})}{\gamma} \leq 4\frac{\eta + \ln 2 + \ln(\gamma^2/\lambda)}{\gamma} \leq 4\frac{\eta + 2\ln(\gamma^2/\lambda)}{\gamma}.$$

We then use Lemma 3 to bound $\|\mathbf{w}_\tau - \mathbf{w}_\lambda\|$ by

$$\|\mathbf{w}_\tau - \mathbf{w}_\lambda\| \le \|\mathbf{w}_\tau\| + \|\mathbf{w}_\lambda\| \le 4\frac{\eta + 2\ln(\gamma^2/\lambda)}{\gamma} + \frac{\sqrt{2} + \ln(\gamma^2/\lambda)}{\gamma} \le 10\frac{\eta + \ln(\gamma^2/\lambda)}{\gamma}.$$

This completes our proof for the parameter convergence.

Furthermore, the assumption on λ guarantees that

$$\frac{\lambda}{\gamma^2} \lesssim 1, \quad \frac{\lambda}{\gamma^2}\eta^2 \lesssim 1, \quad \frac{\lambda}{\gamma^2}\ln^2(\gamma^2/\lambda) \lesssim 1.$$

Therefore, we have

$$\frac{\lambda}{2}\|\mathbf{w}_\tau\|^2 \le 8\frac{\lambda}{\gamma^2}\big(\eta + 2\ln(\gamma^2/\lambda\})\big)^2 \le C_3 - 1$$

for a constant $C_3 > 1$. Also note that $\mathcal{L}(\mathbf{w}_\tau) \le \min\{1/(2e^2\eta), \ln(2)/n\} \le 1$ by Lemma 14. These two bounds together imply that

$$\widetilde{\mathcal{L}}(\mathbf{w}_\tau) - \min\widetilde{\mathcal{L}} \le \widetilde{\mathcal{L}}(\mathbf{w}_\tau) = \mathcal{L}(\mathbf{w}_\tau) + \frac{\lambda}{2}\|\mathbf{w}_\tau\|^2 \le C_3.$$

This completes our proof for the risk convergence. □

A.3 Proof of Corollary 2

Proof of Corollary 2. By Theorem 1, GD enters the stable phase in τ steps, and then attains an ε error within an additional

$$t - \tau \lesssim \frac{\ln(1/\varepsilon)}{\eta\lambda} \eqsim \max\left\{\frac{1}{\gamma\sqrt{\lambda}}, \frac{n}{\gamma^2}\right\}\ln(1/\varepsilon)$$

steps. We can further upper bound the phase transition time by

$$\begin{aligned}
\tau &\eqsim \frac{1}{\gamma^2}\max\left\{\eta, n, \frac{n}{\eta}\ln\frac{n}{\eta}\right\} \\
&\eqsim \frac{1}{\gamma^2}\max\left\{\eta, n\right\} \\
&\eqsim \frac{1}{\gamma^2}\max\left\{\min\left\{\frac{\gamma}{\sqrt{\lambda}}, \frac{\gamma^2}{n\lambda}\right\}, n\right\} \\
&\eqsim \frac{1}{\gamma^2}\max\left\{\frac{\gamma}{\sqrt{\lambda}}, n\right\},
\end{aligned}$$

where the first equality is by the definition of τ, the second equality is because

$$\eta \eqsim \min\left\{\sqrt{\frac{\gamma^2}{\lambda}}, \frac{\gamma^2}{\lambda n}\right\} \gtrsim \min\left\{\sqrt{n\ln n}, \ln n\right\} \gtrsim \ln n,$$

the third equality is by the choice of η, and the fourth equality is because $\max\{\min\{a, a^2/b\}, b\} = \max\{a, b\}$ for $a, b > 0$. So the total number of steps is $t \lesssim \max\{1/(\gamma\sqrt{\lambda}, n/\gamma^2)\}\ln(1/\varepsilon)$. □

A.4 Proof of Theorem 4

The following lemma shows that GD stays in the stable phase.

Lemma 15 (Stable phase, version 2)**.** *Assume that*

$$\lambda \le \frac{\gamma^2}{C_1}\min\left\{1, \frac{1}{\eta^3}\right\}$$

for a large constant C_1. If in the s-th step we have

$$\mathcal{L}(\mathbf{w}_s) \le \frac{1}{4e^2\eta},$$

then for all $t \ge s$ we have

$$\mathcal{L}(\mathbf{w}_t) \le \frac{1}{2e^2\eta}.$$

Proof of Lemma 15. The condition on λ with $C_1 \geq 2$ implies $\lambda\eta \leq \min\{\eta, 1/\eta^2\}/2 \leq 1/2$. Then by Lemma 8 we have

$$\text{for all } t, \quad \|\mathbf{w}_t\| \leq 4\frac{\eta + \ln(e + \gamma^2 \min\{\eta s, 1/\lambda\})}{\gamma} \leq 4\frac{\eta + \ln(e + \gamma^2/\lambda)}{\gamma} =: M.$$

Our assumption on λ implies that

$$\frac{\lambda}{2}M^2 = \frac{8\lambda\big(\eta + \ln(e + \gamma^2/\lambda)\big)^2}{\gamma^2} \leq \frac{1}{4e^2\eta}.$$

To see this, it is sufficient to check that

$$\begin{aligned}
&\frac{\lambda}{\gamma^2}(\eta^2 + \ln^2(e + \gamma^2/\lambda)) \leq \frac{1}{K_1\eta} \quad \text{for a sufficiently large constant } K_1 \qquad (6)\\
\Leftarrow\quad &\frac{\lambda}{\gamma^2} \leq \frac{1}{K_2\eta^3} \text{ and } \frac{\lambda}{\gamma^2}\ln^2(e + \gamma^2/\lambda)) \leq \frac{1}{K_2\eta} \quad \text{for a sufficiently large constant } K_2\\
\Leftarrow\quad &\frac{\lambda}{\gamma^2} \leq \frac{1}{C_1}\min\left\{1, \frac{1}{\eta^3}\right\} \quad \text{for a sufficiently large constant } C_1.
\end{aligned}$$

With this, we prove the following stronger claim by induction:

$$\text{for all } t \geq s, \quad \mathcal{L}(\mathbf{w}_t) \leq \widetilde{\mathcal{L}}(\mathbf{w}_t) \leq \frac{1}{2e^2\eta}.$$

In the s-th step, we have

$$\mathcal{L}(\mathbf{w}_s) \leq \widetilde{\mathcal{L}}(\mathbf{w}_s) \leq \mathcal{L}(\mathbf{w}_s) + \frac{\lambda}{2}\|\mathbf{w}_s\|^2 \leq \mathcal{L}(\mathbf{w}_s) + \frac{\lambda}{2}M^2 \leq \frac{1}{2e^2\eta},$$

which satisfies the hypothesis. Next, assume the hypothesis holds for t. Then $\mathcal{G}(\mathbf{w}_t) \leq \mathcal{L}(\mathbf{w}_t) \leq 1/(2e^2\eta)$. Additionally, our assumption on λ with $C_1 \geq 6$ implies $\lambda \leq 1/(3(1+\eta^3)) \leq 1/(3\eta\ln(e+\eta))$. Thus we can apply Lemma 12 for t, obtaining that $\widetilde{\mathcal{L}}(\mathbf{w}_{t+1}) \leq \widetilde{\mathcal{L}}(\mathbf{w}_t) \leq 1/(2e^2\eta)$. This verifies the hypothesis for $t+1$, and completes our induction. □

The next lemma provides a bound on the phase transition time.

Lemma 16 (Phase transition, version 2)**.** *Assume that*

$$\lambda \leq \frac{\gamma^2}{C_1}\min\left\{1, \frac{1}{\eta^3}\right\}$$

for a constant $C_1 > 1$. Let

$$\tau := \frac{C_2\max\{1, \eta^2\}}{\gamma^2}$$

for a constant $C_2 > 1$. Then for all $t \geq \tau$, we have

$$\mathcal{G}(\mathbf{w}_t) \leq \mathcal{L}(\mathbf{w}_t) \leq \frac{1}{2e^2\eta}.$$

Proof of Lemma 16. The condition on λ with $C_1 \geq 2$ implies $\lambda\eta \leq \min\{\eta, 1/\eta^2\}/2 \leq 1/2$. Then by Lemma 10 we have

$$\frac{1}{\tau}\sum_{k=0}^{\tau-1}\mathcal{L}(\mathbf{w}_k) \leq 10\frac{\eta^2 + \ln^2(e + \gamma^2\min\{\eta\tau, 1/\lambda\})}{\gamma^2\min\{\eta\tau, 1/\lambda\}}.$$

If the right-hand side is smaller than $1/(4e^2\eta)$, then there exists $s \leq \tau$ such that $\mathcal{L}(\mathbf{w}_s) \leq 1/(4e^2\eta)$. By Lemma 15, we have $\mathcal{L}(\mathbf{w}_t) \leq 1/(2e^2\eta)$ for all $t \geq s$. We then complete the proof by using $\mathcal{G}(\mathbf{w}) \leq \mathcal{L}(\mathbf{w})$ from Lemma 4.

To see the right-hand side is smaller than $1/(4e^2\eta)$, it suffices to show that

$$\frac{\eta^2 + \ln^2(e + \gamma^2/\lambda))}{\gamma^2/\lambda} \leq \frac{1}{K_1\eta} \text{ and } \frac{\eta^2 + \ln^2(e + \gamma^2\eta\tau)}{\gamma^2\eta\tau} \leq \frac{1}{K_1\eta}$$

for a sufficiently large constant K_1. This first condition is implied by our assumption on λ as shown by (6) in the proof of Lemma 15. For the second condition to hold, it suffices to have

$$\gamma^2\tau \geq K_2\eta^2 \text{ and } \gamma^2\tau \geq K_2\ln^2(e+\eta\gamma^2\tau) \quad \text{for a sufficiently large constant } K_2$$
$$\Leftarrow \quad \gamma^2\tau \geq C_2\max\{1,\eta^2\} \quad \text{for a sufficiently large constant } C_2.$$

This completes the proof. □

The proof of Theorem 4 follows from the above lemmas.

Proof of Theorem 4. Our assumption on λ and η satisfies the condition on λ and η required by Lemma 16. That condition with $C_1 \geq 6$ implies that $\lambda \leq 1/(3(1+\eta^3)) \leq 1/(3\eta\ln(e+\eta))$, satisfying the condition on λ and η required by Lemma 12.

The phase transition time bound is by Lemma 16, which further enables Lemma 12 for all $t \geq \tau$. Thus we have

$$\widetilde{\mathcal{L}}(\mathbf{w}_{\tau+t}) - \min\widetilde{\mathcal{L}} \leq (1-\lambda\eta)^t\big(\widetilde{\mathcal{L}}(\mathbf{w}_\tau) - \min\widetilde{\mathcal{L}}\big) \leq \exp(-\lambda\eta t)\big(\widetilde{\mathcal{L}}(\mathbf{w}_\tau) - \min\widetilde{\mathcal{L}}\big),$$
$$\|\mathbf{w}_{\tau+t} - \mathbf{w}_\lambda\|^2 \leq (1-\lambda\eta)^t\|\mathbf{w}_\tau - \mathbf{w}_\lambda\|^2 \leq \exp(-\lambda\eta t)\|\mathbf{w}_\tau - \mathbf{w}_\lambda\|^2.$$

Moreover, from the proof of Lemma 15, we know

$$\widetilde{\mathcal{L}}(\mathbf{w}_s) - \min\widetilde{\mathcal{L}} \leq \widetilde{\mathcal{L}}(\mathbf{w}_s) \leq \frac{1}{2e^2\eta}.$$

This completes our proof for the risk convergence. We need to bound $\|\mathbf{w}_s - \mathbf{w}_\lambda\|$ to complete our proof for the parameter convergence, which follows from the same argument as in the proof of Theorem 1 in Appendix A.2. □

A.5 Proof of Corollary 5

Proof of Corollary 5. Recall that $\eta \eqsim \gamma^{2/3}/\lambda^{1/3}$. By Theorem 4, GD enters the stable phase in τ steps, and then attains an ε-suboptimal error within an additional

$$t - \tau \lesssim \frac{\ln(1/(\eta\varepsilon))}{\eta\lambda} \lesssim \frac{\ln(1/\varepsilon)}{(\gamma\lambda)^{2/3}}$$

steps. We can further upper bound the phase transition time by

$$\tau \eqsim \frac{\max\{1,\eta^2\}}{\gamma^2} \eqsim \frac{1}{(\gamma\lambda)^{2/3}}.$$

So the total number of steps is $t \lesssim \ln(1/\varepsilon)/(\gamma\lambda)^{2/3}$. □

B A lower bound

The next lemma provides a hard dataset for which GD cannot use a large stepsize if it operates in the stable regime. The hard dataset construction is motivated by the lower bound of Wu et al. (2024).

Lemma 17 (A stepsize bound)**.** *Consider the dataset*

$$\mathbf{x}_1 = (\gamma, 0.9), \quad \mathbf{x}_2 = (\gamma, -0.5), \quad y_1 = y_2 = 1, \quad 0 < \gamma < 0.1.$$

Then with $\mathbf{w}_0 = 0$*,* $\widetilde{\mathcal{L}}(\mathbf{w}_1) \leq \widetilde{\mathcal{L}}(\mathbf{w}_0)$ *implies that* $\eta \leq 20$*.*

Proof of Lemma 17. We have $\widetilde{\mathcal{L}}(\mathbf{w}_0) = \ln 2$ and

$$\nabla\widetilde{\mathcal{L}}(\mathbf{w}_0) = -\frac{1}{2}(\gamma, 0.4) \quad \Rightarrow \quad \mathbf{w}_1 = \mathbf{w}_0 - \eta\nabla\widetilde{\mathcal{L}}(\mathbf{w}_0) = \frac{\eta}{2}(\gamma, 0.4).$$

So we have

$$\widetilde{\mathcal{L}}(\mathbf{w}_1) \geq \mathcal{L}(\mathbf{w}_1) \geq \frac{1}{2}\ln(1+\exp(-\mathbf{x}_2^\top\mathbf{w}_1)) = \frac{1}{2}\ln(1+\exp((\gamma^2+0.2)\eta/2)) \geq \frac{(\gamma^2+0.2)\eta}{4}.$$

Thus $\widetilde{\mathcal{L}}(\mathbf{w}_1) \leq \widetilde{\mathcal{L}}(\mathbf{w}_0)$ implies that $\eta \leq 4\ln(2)/(\gamma^2+0.2) \leq 20$, which completes the proof. □

The following lemma establishes upper and lower bounds for the logistic empirical risk. For simplicity, this lemma is stated for the special dataset in Lemma 17. However, this lemma can be extended to general datasets satisfying Assumptions 1 and 3 using techniques from Wu et al. (2023).

Lemma 18 (Upper and lower bounds on the logistic empirical risk)**.** *Assume that $\lambda\eta < 1$. For the dataset in Lemma 17, we have*

$$\frac{1}{Ct} \le \mathcal{L}(\mathbf{w}_t) \le \frac{C}{t} \quad \textit{and} \quad \|\mathbf{w}_t\| \le C\ln(t), \quad \textit{for} \ \ 1 \le t \le \frac{1}{C\lambda\ln(1/\lambda)},$$

where $C > 1$ depends on γ and η but is independent of t and λ.

Proof of Lemma 18. Denote the trainable parameter as $\mathbf{w} = (w, \bar{w})$. Then we have $w_0 = \bar{w}_0 = 0$ and

$$w_{t+1} = (1-\eta\lambda)w_t + \frac{\eta\gamma}{2}\left(\frac{1}{1+e^{\gamma w_t + 0.9\bar{w}_t}} + \frac{1}{1+e^{\gamma w_t - 0.5\bar{w}_t}}\right) \tag{7}$$

$$\bar{w}_{t+1} = (1-\eta\lambda)\bar{w}_t + \frac{\eta}{2}\left(\frac{0.9}{1+e^{\gamma w_t + 0.9\bar{w}_t}} - \frac{0.5}{1+e^{\gamma w_t - 0.5\bar{w}_t}}\right). \tag{8}$$

Bounds on $\bar{w}_t$. Recall that $\eta\lambda < 1$. From (7), we see that $(w_t)_{t\ge 0}$ are all nonnegative. Then by direct computation, we can verify that the factor within the big bracket in (8) is positive when $\bar{w}_t \le 0$ and is negative when $\bar{w}_t \ge 2$. Then (8) implies the following:

$$\begin{aligned}
&\text{if } \bar{w}_t \le 0, && -(1-\eta\lambda)|\bar{w}_t| \le (1-\eta\lambda)\bar{w}_t \le \bar{w}_{t+1} \le (1-\eta\lambda)\bar{w}_t + \eta \le \eta;\\
&\text{if } 0 < \bar{w}_t \le 2, && -\eta \le (1-\eta\lambda)\bar{w}_t - \eta \le \bar{w}_{t+1} \le (1-\eta\lambda)\bar{w}_t + \eta \le 2 + \eta;\\
&\text{if } \bar{w}_t > 2, && -\eta \le (1-\eta\lambda)\bar{w}_t - \eta \le \bar{w}_{t+1} \le (1-\eta\lambda)\bar{w}_t \le (1-\eta\lambda)|\bar{w}_t|.
\end{aligned}$$

In all cases, we have

$$|\bar{w}_{t+1}| \le \max\{(1-\eta\lambda)|\bar{w}_t|,\ \eta + 2\},$$

which implies that $|\bar{w}_t| \le \eta + 2$ for every $t \ge 0$ by induction.

Let

$$\mathcal{H}(\bar{w}) := \frac{1}{2}\big(\exp(-0.9\bar{w}) + \exp(0.5\bar{w})\big).$$

Then for every $t \ge 0$, we have $\mathcal{H}(\bar{w}_t) \le \exp(\eta + 2) := H_{\max}$.

An upper bound on w_t. Using the upper bound on $\mathcal{H}(\bar{w}_t)$ and (7), we have

$$w_{t+1} \le w_t + \frac{\eta\gamma}{2}\left(\frac{1}{e^{\gamma w_t + 0.9\bar{w}_t}} + \frac{1}{e^{\gamma w_t - 0.5\bar{w}_t}}\right) \le w_t + \frac{\eta\gamma H_{\max}}{2} e^{-\gamma w_t}.$$

Let $t_0 := \min\{t : \gamma^2\eta H_{\max}\exp(-\gamma w_t)/2 \le 1\}$. Since w_t is increasing, t_0 exists. For every $t \le t_0$, we have $w_t \le \gamma^{-1}\ln(\gamma^2\eta H_{\max}/2)$. For $t \ge t_0$, we have

$$e^{\gamma w_{t+1}} \le e^{\gamma w_t} e^{\gamma^2\eta H_{\max}/2\exp(-\gamma w_t)} \le e^{\gamma w_t}\left(1 + e\frac{\gamma^2\eta H_{\max}\exp(-\gamma w_t)}{2}\right) \le e^{\gamma w_t} + \frac{e\gamma^2\eta H_{\max}}{2}$$

$$\Rightarrow \quad \text{for } t \ge t_0, \quad w_t \le \frac{1}{\gamma}\ln\left(\frac{e\gamma^2\eta H_{\max}}{2}(t - t_0) + e^{\gamma w_{t_0}}\right).$$

Putting these two bounds together, we have for every $t \ge 0$,

$$w_t \le \frac{1}{\gamma}\ln\big(e\gamma^2\eta H_{\max}(t+1)\big).$$

A lower bound on w_t. From (7), we have

$$\begin{aligned}
w_{t+1} &\ge (1-\eta\lambda)w_t + \frac{\eta\gamma}{2}\big(\min\{1, e^{-\gamma w_t - 0.9\bar{w}_t}\} + \min\{1, e^{-\gamma w_t + 0.5\bar{w}_t}\}\big)\\
&= (1-\eta\lambda)w_t + \frac{\eta\gamma}{2}e^{-\gamma w_t}\big(\min\{e^{\gamma w_t}, e^{-0.9\bar{w}_t}\} + \min\{e^{\gamma w_t}, e^{0.5\bar{w}_t}\}\big)
\end{aligned}$$

$$\geq (1-\eta\lambda)w_t + \frac{\eta\gamma}{2}e^{-\gamma w_t}\big(\min\{1,\, e^{-0.9\bar{w}_t}\} + \min\{1,\, e^{0.5\bar{w}_t}\}\big)$$
$$\geq (1-\eta\lambda)w_t + \frac{\eta\gamma}{2}e^{-\gamma w_t},$$

For t such that

$$\lambda < \frac{1}{4\eta H_{\max}(t+1)\ln\big(e\gamma^2\eta H_{\max}(t+1)\big)},$$

from our upper bound on w_t, we have

$$\frac{2\lambda}{\gamma}w_t e^{\gamma w_t} \leq \frac{1}{2}.$$

Then for such t, we have

$$e^{\gamma w_{t+1}} \geq e^{(1-\eta\lambda)\gamma w_t + \frac{\eta\gamma}{2}e^{-\gamma w_t}} = e^{\gamma w_t}\exp\left(\frac{\eta\gamma}{2}e^{-\gamma w_t}\left(1 - \frac{2\lambda}{\gamma}w_t e^{\gamma w_t}\right)\right)$$
$$\geq e^{\gamma w_t}\exp\left(\frac{\eta\gamma}{4}e^{-\gamma w_t}\right) \geq e^{\gamma w_t}\left(1 + \frac{\eta\gamma}{4}e^{-\gamma w_t}\right) \geq e^{\gamma w_t} + \frac{\eta\gamma}{4}.$$

So we have

$$w_t \geq \frac{1}{\gamma}\ln\left(\frac{\eta\gamma}{4}t + 1\right), \quad \text{for } t \lesssim \frac{1}{\gamma^2\eta H_{\max}\lambda\ln(1/\lambda)}.$$

Bounds on logistic empirical risk. Notice that

$$e^{-\gamma w_t + |\bar{w}_t|} \geq \mathcal{L}(\mathbf{w}_t) \geq \frac{1}{2}\ln(1 + e^{-\gamma w_t}).$$

This, together with our upper and lower bounds on w_t and $\bar{w}_t$, leads to the promised bounds. □

With the above lemmas, we are ready to prove Theorem 3.

Proof of Theorem 3. From Lemma 17 we know $\eta \leq 20$. Then from Lemma 18, we have

$$\frac{1}{C_0 t} \leq \mathcal{L}(\mathbf{w}_t) \leq \frac{C_0}{t} \text{ and } \|\mathbf{w}_t\| \leq C_0\ln(t), \quad \text{for } t \leq \frac{1}{C_0\lambda\ln(1/\lambda)},$$

where $C_0 > 1$ is a large factor that only depends on γ but is independent of t and λ. Here, we can make C_0 independent of η as $\eta \leq 20$. For every sufficiently small λ, we can pick

$$\tau := \frac{1}{C_0^2\lambda\ln^2(1/\lambda)} < \frac{1}{C_0\lambda\ln(1/\lambda)}.$$

For this τ, we have

$$C_0\lambda\ln^2(1/\lambda) \leq \mathcal{L}(\mathbf{w}_\tau) \leq C_0^3\lambda\ln^2(1/\lambda) \text{ and } \|\mathbf{w}_\tau\| \leq C_0\ln(1/\lambda).$$

By Lemma 3 and setting C_0 large enough, we get

$$\min\widetilde{\mathcal{L}} \leq \frac{\lambda(2 + \ln^2(\gamma^2/\lambda))}{2\gamma^2} \leq \frac{1}{2}C_0\lambda\ln^2(1/\lambda).$$

That is, we have

$$\widetilde{\mathcal{L}}(\mathbf{w}_\tau) - \min\widetilde{\mathcal{L}} \geq \mathcal{L}(\mathbf{w}_\tau) - \min\widetilde{\mathcal{L}} \geq \frac{1}{2}C_0\lambda\ln^2(1/\lambda).$$

Step complexity for a large ε. For $\varepsilon \geq 0.5C_0\lambda\ln^2(1/\lambda)$, we have

$$\widetilde{\mathcal{L}}(\mathbf{w}_t) - \min\widetilde{\mathcal{L}} \leq \varepsilon \quad \Rightarrow \quad \mathcal{L}(\mathbf{w}_t) \leq \min\widetilde{\mathcal{L}} + \varepsilon \leq 2\varepsilon$$
$$\Rightarrow \quad t \geq \frac{1}{2C_0\varepsilon},$$

where the second line is because of the lower bound on $\mathcal{L}(\mathbf{w}_t)$ for $t \leq 1/(C_0\lambda\ln(1/\lambda))$ and our choice of λ.

Step complexity for a small ε. The case of $\varepsilon < 0.5C_0\lambda\ln^2(1/\lambda)$ needs some more effort. Since GD operates in the stable regime, we have

$$\text{for all } t \geq \tau, \quad \widetilde{\mathcal{L}}(\mathbf{w}_t) \leq \widetilde{\mathcal{L}}(\mathbf{w}_\tau) \leq \mathcal{L}(\mathbf{w}_\tau) + \frac{\lambda}{2}\|\mathbf{w}_\tau\|^2 \leq 2C_0^3\lambda\ln^2(1/\lambda).$$

That is, $(\mathbf{w}_t)_{t\geq\tau}$ are all within the level set

$$\mathcal{W} := \big\{\mathbf{w} : \widetilde{\mathcal{L}}(\mathbf{w}) \leq 2C_0^3\lambda\ln^2(1/\lambda)\big\}.$$

For parameters in this level set, we have

$$\sup_{\mathbf{w}\in\mathcal{W}} \|\nabla^2\widetilde{\mathcal{L}}(\mathbf{w})\| = \sup_{\mathbf{w}\in\mathcal{W}} \|\nabla^2\mathcal{L}(\mathbf{w})\| + \lambda \begin{cases} \leq \sup_{\mathbf{w}\in\mathcal{W}} \mathcal{L}(\mathbf{w}) + \lambda \leq \sup_{\mathbf{w}\in\mathcal{W}} \widetilde{\mathcal{L}}(\mathbf{w}_t) + \lambda \leq 3C_0^3\lambda\ln^2(1/\lambda), \\ \geq \lambda, \end{cases}$$

where the upper bound is given by Lemma 4 and the definition of $\mathcal{W}$. That is, $\widetilde{\mathcal{L}}$ is β-smooth for $\beta = 3C_0^3\lambda\ln^2(1/\lambda)$ and λ-strongly convex for $\mathbf{w} \in \mathcal{W}$. By standard convex optimization theory, we have

$$\text{for all } \mathbf{w} \in \mathcal{W}, \quad \widetilde{\mathcal{L}}(\mathbf{w}) - \min\widetilde{\mathcal{L}} \geq \frac{1}{2\beta}\|\nabla\widetilde{\mathcal{L}}(\mathbf{w})\|^2;$$

$$\widetilde{\mathcal{L}}(\mathbf{w}_{t+1}) \geq \widetilde{\mathcal{L}}(\mathbf{w}_t) + \langle\nabla\widetilde{\mathcal{L}}(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle = \widetilde{\mathcal{L}}(\mathbf{w}_t) - \eta\|\nabla\widetilde{\mathcal{L}}(\mathbf{w}_t)\|^2.$$

Moreover, our choice of λ implies $\eta \leq 20 < 1/(4\beta)$. Then the above two inequalities imply that

$$\text{for all } t \geq \tau, \quad \widetilde{\mathcal{L}}(\mathbf{w}_{t+1}) - \min\widetilde{\mathcal{L}} \geq (1 - 2\eta\beta)\big(\widetilde{\mathcal{L}}(\mathbf{w}_t) - \min\widetilde{\mathcal{L}}\big),$$

which further implies that

$$\begin{aligned} \widetilde{\mathcal{L}}(\mathbf{w}_t) - \min\widetilde{\mathcal{L}} &\geq (1 - 2\eta\beta)^{t-\tau}\big(\widetilde{\mathcal{L}}(\mathbf{w}_\tau) - \min\widetilde{\mathcal{L}}\big) \\ &\geq (1 - 2\eta\beta)^{t-\tau}\frac{1}{2}C_0\lambda\ln^2(1/\lambda) \\ &\geq \exp\big(-4\eta\beta(t-\tau)\big)\frac{1}{2}C_0\lambda\ln^2(1/\lambda). \end{aligned}$$

For the right-hand side to be smaller than $\varepsilon < 0.5C_0\lambda\ln^2(1/\lambda)$, we need

$$\begin{aligned} t &\geq \tau + \frac{\ln\big(0.5C_0\lambda\ln^2(1/\lambda)/\varepsilon\big)}{4\eta\beta} \\ &\geq \frac{1}{C_0\lambda\ln^2(1/\lambda)} + \frac{\ln\big(0.5C_0\lambda\ln^2(1/\lambda)/\varepsilon\big)}{240C_0^3\lambda\ln^2(1/\lambda)}. \end{aligned}$$

This completes our proof. □

C Population risk analysis

We provide a proof for Proposition 6 in Appendix C.1, then calculate the optimal regularization hyperparameter in Appendix C.2.

C.1 Proof of Proposition 6

Proof of Proposition 6. It is clear that under Assumption 2, $\|\mathbf{w}\| \leq B$ implies $\ell(y\mathbf{x}^\top\mathbf{w}) \leq \ell(0) + B$. Applying Srebro et al. (2010, Theorem 1) to the functional class induced by $\{\mathbf{w} : \|\mathbf{w}\| \leq B\}$, we have the following: with probability $1 - \delta$,

$$\text{for every } \mathbf{w} \text{ such that } \|\mathbf{w}\| \leq B, \quad \mathcal{L}_{\text{test}}(\mathbf{w}) \lesssim \mathcal{L}(\mathbf{w}) + \ln^3(n)\mathcal{R}_n^2(B) + \frac{(B+1)\ln(1/\delta)}{n},$$

where $\mathcal{R}_n(B)$ is the Rademacher complexity of the functional class induced by $\{\mathbf{w} : \|\mathbf{w}\| \le B\}$,

$$\mathcal{R}_n(B) := \sup_{(\mathbf{x}_i, y_i)_{i=1}^n} \mathbb{E}_\sigma \sup_{\|\mathbf{w}\| \le B} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \ell(-y_i \mathbf{x}_i^\top \mathbf{w}) \right|,$$

where $\sigma = (\sigma_i)_{i=}^n$ are n independent Rademacher random varaibles.

We control the Rademacher complexity by (Shalev-Shwartz and Ben-David, 2014, Lemma 26.10)

$$\begin{aligned}
\mathcal{R}_n(B) &\le \sup_{(\mathbf{x}_i, y_i)_{i=1}^n} \mathbb{E}_\sigma \sup_{\|\mathbf{w}\| \le B} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i y_i \mathbf{x}_i^\top \mathbf{w} \right| \le \frac{B}{n} \sup_{(\mathbf{x}_i, y_i)_{i=1}^n} \mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i y_i \mathbf{x}_i \right\| \\
&\le \frac{B}{n} \sup_{(\mathbf{x}_i, y_i)_{i=1}^n} \sqrt{\mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i y_i \mathbf{x}_i \right\|^2} = \frac{B}{n} \sup_{(\mathbf{x}_i, y_i)_{i=1}^n} \sqrt{\sum_{i=1}^n \|\mathbf{x}_i\|^2} \le \frac{B}{\sqrt{n}},
\end{aligned}$$

where the first inequality is by the 1-Lipschitzness of ℓ, the second inequality is by Cauchy–Schwarz inequality, and the last inequality is by Assumption 2.

Putting these together, we have: with probability $1 - \delta$,

$$\text{for every } \mathbf{w} \text{ such that } \|\mathbf{w}\| \le B, \quad \mathcal{L}_{\text{test}}(\mathbf{w}) \lesssim \mathcal{L}(\mathbf{w}) + \frac{B^2 \ln^3(n)}{n} + \frac{(B+1)\ln(1/\delta)\big)}{n}.$$

Now for a given $\mathbf{w}$, consider a sequence of balls with radius $B_i = e^i$ and a sequence of probabilities $\delta_i = \delta/(i+1)^2$. It is clear that $\sum_i \delta_i \lesssim \delta$ and $\mathbf{w}$ belongs to B_i for $i = \ln(\|\mathbf{w}\| + 1)$. Applying the above inequality to each B_i and δ_i, then applying a union bound (motivated by the proof of Theorem 26.14 in (Shalev-Shwartz and Ben-David, 2014)), we get: for every $\mathbf{w}$, with probability $1 - \delta$,

$$\begin{aligned}
\mathcal{L}_{\text{test}}(\mathbf{w}) &\lesssim \mathcal{L}(\mathbf{w}) + \frac{\big(\|\mathbf{w}\| + 1\big)^2 \ln^3(n)}{n} + \frac{(\|\mathbf{w}\| + 1)\ln(\ln(\|\mathbf{w}\| + 1)/\delta)\big)}{n} \\
&\lesssim \mathcal{L}(\mathbf{w}) + \frac{\max\{1,\, \|\mathbf{w}\|^2\}\big(\ln^3(n) + \ln(1/\delta)\big)}{n}.
\end{aligned}$$

This completes the proof. □

C.2 Optimal regularization

We compute the optimal regularization hyperparameter λ such that $\mathbf{w}_\lambda$ minimizes the upper bound in Proposition 6. We assume that $\lambda < 1/\gamma^2$. From Lemma 3, we have

$$\|\mathbf{w}_\lambda\| \le \frac{\sqrt{2} + \ln(\gamma^2/\lambda)}{\gamma}, \quad \mathcal{L}(\mathbf{w}_\lambda) \le \widetilde{\mathcal{L}}(\mathbf{w}_\lambda) \le \frac{\lambda\big(2 + \ln^2(\gamma^2/\lambda)\big)}{2\gamma^2}.$$

Pugging these into Proposition 6, we have

$$\begin{aligned}
\mathcal{L}_{\text{test}}(\mathbf{w}_\lambda) &\lesssim \mathcal{L}(\mathbf{w}_\lambda) + \frac{\ln^3(n) + \ln(1/\delta)}{n} \|\mathbf{w}_\lambda\|^2 \\
&\lesssim \frac{\lambda\big(1 + \ln^2(\gamma^2/\lambda)\big)}{\gamma^2} + \frac{\ln^3(n) + \ln(1/\delta)}{n} \frac{1 + \ln^2(\gamma^2/\lambda)}{\gamma^2}.
\end{aligned}$$

Choosing $\lambda \eqsim 1/n$ minimizes the right-hand side up to constant factors, where we have

$$\mathcal{L}_{\text{test}}(\mathbf{w}_\lambda) \lesssim \frac{\big(\ln^3(n) + \ln(1/\delta)\big)\ln^2(n)}{\gamma^2 n} = \widetilde{\mathcal{O}}\left(\frac{1}{\gamma^2 n}\right).$$

Note that the upper bound provided in Proposition 6 is at least $\Omega(1/n)$. So the choice of $\lambda \eqsim 1/n$ leads to a nearly unimprovable bound, ignoring logarithmic factors and dependence on γ.

D The critical stepsize threshold

We first prove Theorem 7 in Appendix D.1. We then show that the proposed critical threshold also sharply determines the global convergence of GD in Appendix D.2.

D.1 Proof of Theorem 7

Denote the linearly separable dataset in Assumption 1 as

$$\mathbf{X} := (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top, \quad \mathbf{y} := (y_1, \dots, y_n)^\top.$$

Consider the following margin maximization program

$$\min_{\mathbf{w} \in \mathbb{H}} \ \|\mathbf{w}\| \quad \text{s.t. } y_i \mathbf{x}_i^\top \mathbf{w} \geq 1, \ i = 1, \dots, n.$$

Its Lagrangian dual can be written

$$\max_{\boldsymbol{\beta} \in \mathbb{R}^n} \ -\frac{1}{2} \boldsymbol{\beta}^\top \mathbf{X}\mathbf{X}^\top \boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{y} \quad \text{s.t. } y_i \boldsymbol{\beta}_i \geq 0, \ i = 1, \dots, n,$$

where $y_i \boldsymbol{\beta}_i$ is the dual variable associated with the i-th constraint (see, e.g., Hsu et al., 2021). Let $\hat{\boldsymbol{\beta}}$ be the solution to the above problem. Let

$$\mathcal{S}_+ := \{i \in [n] : y_i \hat{\boldsymbol{\beta}}_i > 0\}$$

be the set of support vectors with nonzero dual variables. Then Assumption 3 says that $\{\mathbf{x}_i : i \in \mathcal{S}_+\}$ spans the same space as $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$.

We introduce some additional notation following Wu et al. (2023). By the rotational invariance of the problem, we can assume without loss of generality that the maximum ℓ_2-margin direction is aligned with the first vector of the canonical basis. Then we can write the dataset and the parameters as

$$\mathbf{x}_i = (x_i, \bar{\mathbf{x}}_i), \quad \mathbf{w} = (w, \bar{\mathbf{w}}), \quad \mathbf{w}_\lambda = (w_\lambda, \bar{\mathbf{w}}_\lambda),$$

where $y_i x_i \geq \gamma$ by Assumption 1.

Let

$$\mathcal{S} := \{i : y_i x_i = \gamma\}$$

be the index set of all support vectors (satisfying the constraint with equality). Then Assumption 3 implies that $(\bar{\mathbf{x}}_i, y_i)_{i \in \mathcal{S}}$ are strictly nonseparable (Wu et al., 2023, Lemma 3.1). Define

$$\mathcal{H}(\bar{\mathbf{w}}) := \frac{1}{n} \sum_{i \in \mathcal{S}} \exp(-y_i \bar{\mathbf{x}}_i^\top \bar{\mathbf{w}}),$$

then $\mathcal{H}(\cdot)$ is convex, bounded from below, and with a compact level set. Thus, it admits a finite minimizer, which is denoted as $\bar{\mathbf{w}}_* := \arg\min \mathcal{H}(\cdot)$.

Wu et al. (2025a, Lemma D.2) provided an asymptotic characterization of $\mathbf{w}_\lambda$, which is restated as the following lemma.

Lemma 19 (Lemma D.2 in (Wu et al., 2025a)). *Under Assumption 3, as $\lambda \to 0$, we have*

$$w_\lambda \to \infty, \quad \bar{\mathbf{w}}_\lambda \to \bar{\mathbf{w}}_*.$$

Without loss of generality, we assume that $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ spans the whole space $\mathbb{H}$. Otherwise, we project every quantity into the span of $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$. Under this convention, we have the following sharp characterization of the Hessian at $\mathbf{w}_\lambda$.

Lemma 20 (Hessian bounds). *Under Assumption 3, as $\lambda \to 0$, we have*

$$\nabla^2 \mathcal{L}(\mathbf{w}_\lambda) = \lambda \ln(1/\lambda) \frac{1 \pm o(1)}{\gamma^2 \mathcal{H}(\bar{\mathbf{w}}_*)} \frac{1}{n} \sum_{i \in \mathcal{S}} \mathbf{x}_i \mathbf{x}_i^\top \exp(-y_i \bar{\mathbf{x}}_i^\top \bar{\mathbf{w}}_*).$$

As a direct consequence, for every $\lambda < 1/C_0$, we have

$$\frac{1}{C_1} \lambda \ln(1/\lambda) \mathbf{I} \preceq \nabla^2 \mathcal{L}(\mathbf{w}_\lambda) \preceq C_1 \lambda \ln(1/\lambda) \mathbf{I},$$

where $C_0, C_1 > 1$ depend on the dataset but are independent of λ.

Proof of Lemma 20. The first-order optimality condition for $\mathbf{w}_\lambda$ implies that

$$\begin{aligned}\lambda w_\lambda &= \frac{1}{n}\sum_{i=1}^{n}\frac{y_i x_i}{1+\exp(y_i \mathbf{x}_i^\top \mathbf{w}_\lambda)} \\ &= \frac{1}{n}\sum_{i\in\mathcal{S}}\frac{\gamma}{1+\exp(\gamma w_\lambda + y_i \bar{\mathbf{x}}_i^\top \bar{\mathbf{w}}_\lambda)} + \frac{1}{n}\sum_{i\notin\mathcal{S}}\frac{y_i x_i}{1+\exp(y_i x_i w_\lambda + y_i \bar{\mathbf{x}}_i^\top \bar{\mathbf{w}}_\lambda)},\end{aligned}$$

where $y_i x_i > \gamma$ for $i \notin \mathcal{S}$. Then Lemma 19 implies that

$$\begin{aligned}\lambda w_\lambda \exp(\gamma w_\lambda) &= \frac{1}{n}\sum_{i\in\mathcal{S}}\frac{\gamma\exp(\gamma w_\lambda)}{1+\exp(\gamma w_\lambda + y_i \bar{\mathbf{x}}_i^\top \bar{\mathbf{w}}_\lambda)} + \frac{1}{n}\sum_{i\notin\mathcal{S}}\frac{y_i x_i \exp(\gamma w_\lambda)}{1+\exp(y_i x_i w_\lambda + y_i \bar{\mathbf{x}}_i^\top \bar{\mathbf{w}}_\lambda)} \\ &= \big(1+o(1)\big)\frac{1}{n}\sum_{i\in\mathcal{S}}\frac{\gamma\exp(\gamma w_\lambda)}{1+\exp(\gamma w_\lambda + y_i \bar{\mathbf{x}}_i^\top \bar{\mathbf{w}}_\lambda)} \\ &= \big(1\pm o(1)\big)\frac{1}{n}\sum_{i\in\mathcal{S}}\frac{\gamma\exp(\gamma w_\lambda)}{\exp(\gamma w_\lambda + y_i \bar{\mathbf{x}}_i^\top \bar{\mathbf{w}}_\lambda)} \\ &= \big(1\pm o(1)\big)\gamma\mathcal{H}(\bar{\mathbf{w}}_\lambda) \\ &= \big(1\pm o(1)\big)\gamma\mathcal{H}(\bar{\mathbf{w}}_*).\end{aligned}$$

That is,

$$\gamma w_\lambda \exp(\gamma w_\lambda) = \big(1\pm o(1)\big)\gamma^2\mathcal{H}(\bar{w}_*)/\lambda.$$

Notice that γw_λ is the Lambert W function applied to the right-hand side. By the property of the Lambert W function(see, e.g., Hoorfar and Hassani, 2008, Theorem 2.7), we have

$$\exp(\gamma w_\lambda) = \big(1\pm o(1)\big)\frac{\big(1\pm o(1)\big)\gamma^2\mathcal{H}(\bar{\mathbf{w}}_*)/\lambda}{\ln\big(\big(1\pm o(1)\big)\gamma^2\mathcal{H}(\bar{\mathbf{w}}_*)/\lambda\big)} = \big(1\pm o(1)\big)\frac{\gamma^2\mathcal{H}(\bar{\mathbf{w}}_*)}{\lambda\ln(1/\lambda)}.$$

For the Hessian at $\mathbf{w}_\lambda$, we have

$$\begin{aligned}\nabla^2\mathcal{L}(\mathbf{w}_\lambda) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{(1+\exp(-y_i\mathbf{x}_i^\top\mathbf{w}_\lambda))(1+\exp(y_i\mathbf{x}_i^\top\mathbf{w}_\lambda))} \\ &= \big(1\pm o(1)\big)\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\exp(y_i x_i w_\lambda + y_i\bar{\mathbf{x}}_i^\top\bar{\mathbf{w}}_\lambda)} \\ &= \big(1\pm o(1)\big)\frac{1}{n}\sum_{i\in\mathcal{S}}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\exp(\gamma w_\lambda + y_i\bar{\mathbf{x}}_i^\top\bar{\mathbf{w}}_\lambda)} \\ &= \big(1\pm o(1)\big)\exp(-\gamma w_\lambda)\frac{1}{n}\sum_{i\in\mathcal{S}}\exp(-y_i\bar{\mathbf{x}}_i^\top\bar{\mathbf{w}}_\lambda)\mathbf{x}_i\mathbf{x}_i^\top \\ &= \big(1\pm o(1)\big)\exp(-\gamma w_\lambda)\frac{1}{n}\sum_{i\in\mathcal{S}}\exp(-y_i\bar{\mathbf{x}}_i^\top\bar{\mathbf{w}}_*)\mathbf{x}_i\mathbf{x}_i^\top.\end{aligned}$$

Plugging in the bounds for $\exp(\gamma w_\lambda)$, we get

$$\nabla^2\mathcal{L}(\mathbf{w}_\lambda) = \lambda\ln(1/\lambda)\frac{1\pm o(1)}{\gamma^2\mathcal{H}(\bar{\mathbf{w}}_*)}\frac{1}{n}\sum_{i\in\mathcal{S}}\mathbf{x}_i\mathbf{x}_i^\top\exp(-y_i\bar{\mathbf{x}}_i^\top\bar{\mathbf{w}}_*),$$

which concludes the proof. □

We are ready to prove Theorem 7.

Proof of Theorem 7. Lemma 20 implies that for every $\lambda \le 1/C_0$, we have

$$\frac{1}{C_1}\lambda\ln(1/\lambda)\mathbf{I} \preceq \nabla^2\widetilde{\mathcal{L}}(\mathbf{w}_\lambda) := \lambda\mathbf{I} + \nabla^2\mathcal{L}(\mathbf{w}_\lambda) \preceq C_1\lambda\ln(1/\lambda)\mathbf{I}.$$

Since $\nabla^2\widetilde{\mathcal{L}}(\cdot)$ is continuously differentiable, there exists a neighborhood of $\mathbf{w}_\lambda$ of radius r such that

$$\text{for all } \mathbf{w} \text{ such that } \|\mathbf{w}-\mathbf{w}_\lambda\| \le r, \quad \frac{1}{2C_1}\lambda\ln(1/\lambda)\mathbf{I} \preceq \nabla^2\widetilde{\mathcal{L}}(\mathbf{w}) \preceq 2C_1\lambda\ln(1/\lambda)\mathbf{I}.$$

The first claim. By classical optimization theory, GD with initialization satisfying $\|\mathbf{w}_0 - \mathbf{w}_\lambda\| \le r$ and stepsize satisfying $\eta < 1/(C_1\lambda\ln(1/\lambda))$ converges to $\mathbf{w}_\lambda$.

The second claim. Recall that $\nabla\widetilde{\mathcal{L}}(\mathbf{w}_\lambda) = 0$. For every $\mathbf{w}$ such that $\|\mathbf{w} - \mathbf{w}_\lambda\| \le r$,

$$\nabla\widetilde{\mathcal{L}}(\mathbf{w}) = \int_0^1 \nabla^2\widetilde{\mathcal{L}}(t\mathbf{w} + (1-t)\mathbf{w}_\lambda)(\mathbf{w} - \mathbf{w}_\lambda)dt.$$

This, together with the above Hessian bound, implies that

$$\frac{1}{2C_1}\lambda\ln(1/\lambda)\|\mathbf{w} - \mathbf{w}_\lambda\| \le \|\nabla\widetilde{\mathcal{L}}(\mathbf{w})\| \le 2C_1\lambda\ln(1/\lambda)\|\mathbf{w} - \mathbf{w}_\lambda\|.$$

Consider GD with a stepsize $\eta > 20C_1^3/(\lambda\ln(1/\lambda))$. If GD is within that ball in the t-th step, then we have

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}_\lambda\|^2 = \|\mathbf{w}_t - \mathbf{w}_\lambda\|^2 - 2\eta\langle\nabla\widetilde{\mathcal{L}}(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_\lambda\rangle + \eta^2\|\nabla\widetilde{\mathcal{L}}(\mathbf{w}_t)\|^2 \\
&\ge \|\mathbf{w}_t - \mathbf{w}_\lambda\|^2 - 4\eta C_1\lambda\ln(1/\lambda)\|\mathbf{w}_t - \mathbf{w}_\lambda\|^2 + \left(\frac{1}{2C_1}\eta\lambda\ln(1/\lambda)\right)^2\|\mathbf{w}_t - \mathbf{w}_\lambda\|^2 \\
&\ge \|\mathbf{w}_t - \mathbf{w}_\lambda\|^2 + \eta C_1\lambda\ln(1/\lambda)\|\mathbf{w}_t - \mathbf{w}_\lambda\|^2 \\
&\ge (1 + 20C_1^4)\|\mathbf{w}_t - \mathbf{w}_\lambda\|^2.
\end{aligned}$$

That is, if GD enters the ball centered at $\mathbf{w}_\lambda$ with radius r but is different from $\mathbf{w}_\lambda$, then it must exit the ball in a finite number of steps. However, we will show next that the set of the initializations such that GD exactly hits $\mathbf{w}_\lambda$ has measure zero.

Let $d < \infty$ be the dimension of $\mathbb{H}$, then we can embed $\mathbb{H}$ into $\mathbb{R}^d$. Let

$$g : \mathbb{R}^d \to \mathbb{R}^d, \quad \mathbf{w} \mapsto \mathbf{w} - \eta\nabla\widetilde{\mathcal{L}}(\mathbf{w})$$

be one step of GD. To conclude, it suffices to show that g satisfies the Luzin N^{-1} property, that is, for all subsets $S \subset \mathbb{R}^d$, if S has measure zero then its preimage $g^{-1}(S)$ also has measure zero. Indeed, this ensures that (countably infinite times) iterated preimages of $\{\mathbf{w}_\lambda\}$ remain of measure zero. Conveniently, showing g satisfies the Luzin N^{-1} property is equivalent to showing that the Jacobian determinant of g is nonzero almost everywhere (Ponomarev, 1987, Theorem 1). We denote the Jacobian determinant of g as

$$\Delta : \mathbb{R}^d \to \mathbb{R}, \quad \mathbf{w} \mapsto \det\left(\mathbf{I} - \eta\nabla^2\widetilde{\mathcal{L}}(\mathbf{w})\right).$$

Observe that Δ is a composition of a degree-d polynomial, of the derivatives of the sigmoid function $x \mapsto 1/(1+e^{-x})$, and of linear maps of $\mathbf{w}$. Recall that the sigmoid function is analytic on $\mathbb{R}$, meaning that it is everywhere equal to its Taylor expansion on a ball of positive radius. We conclude that Δ is also analytic on $\mathbb{R}^d$ as a composition of analytic functions. By the identity theorem for analytic functions (Krantz and Parks, 2002, Corollary 1.2.7), we conclude that Δ is either zero everywhere or that its zeros do not have an accumulation point in $\mathbb{R}^d$. We show that the latter holds by discussing the following two cases.

- If $\eta = 1/\lambda$, then

  $$\Delta(\mathbf{0}) = \det\left((1-\eta\lambda)\mathbf{I} - \eta\frac{1}{4n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\right) = \det\left(-\eta\frac{1}{4n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\right),$$

  which is nonzero since we assume $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ spans the whole space.

- If $\eta \ne 1/\lambda$, then by Assumption 1, as $\rho \to \infty$, we have

  $$\Delta(\rho\mathbf{w}^*) = \det\left((1-\eta\lambda)\mathbf{I} - \eta\frac{1}{n}\sum_{i=1}^n \frac{\mathbf{x}_i\mathbf{x}_i^\top}{\left(1+\exp(\rho\mathbf{x}_i^\top\mathbf{w}^*)\right)\left(1+\exp(-\rho\mathbf{x}_i^\top\mathbf{w}^*)\right)}\right) \to 1 - \eta\lambda.$$

  That is, for every pair of η and λ such that $\eta\lambda \ne 1$, we can pick a sufficiently large ρ such that $\Delta(\rho\mathbf{w}^*) = 1 - \eta\lambda \pm o(1)$ is nonzero.

In sum, for any choices of η and λ, Δ cannot be zero everywhere. So the zeros of Δ do not have an accumulation point in $\mathbb{R}^d$, and thus have measure zero. This concludes the proof. □

D.2 Global convergence in the 1-dimensional case

In the 1-dimensional case, the objective function can be written as

$$\widetilde{\mathcal{L}}(w) := \frac{1}{n}\sum_{i=1}^{n} \ln(1+e^{-z_i w}) + \frac{\lambda}{2}w^2\,,$$

where $\gamma \le z_i \le 1$ and there exists an i such that $z_i = \gamma$.

The next theorem shows that in this 1-dimensional case, GD converges globally with stepsizes below the critical threshold by a constant factor. We note that this is a very special situation, where GD with large stepsizes oscillates *at most once* (see the proof). However, in general finite-dimensional cases, GD with large stepsizes can oscillate many times. Thus, it is unclear if the results in this theorem generalize to general finite-dimensional cases.

Theorem 8 (A 1-dimensional anlaysis)**.** *Suppose that Assumption 1 holds and that $\mathbb{H}$ is 1-dimensional. Then for every $\lambda \le 1/C_0$, $\eta \le 1/(C_1\lambda\ln(1/\lambda))$, and w_0, GD converges, and after*

$$t = \mathcal{O}\left(\frac{1}{\eta\lambda}\ln\left(\frac{|w_0|+1}{\varepsilon\lambda\ln(1/\lambda)}\right)\right)$$

steps, $\widetilde{\mathcal{L}}(w_t) - \min\widetilde{\mathcal{L}} \le \varepsilon$. Here, $C_0, C_1 > 1$ depend on the dataset and on γ, but not on λ or η.

Proof of Theorem 8. Let us first compute the derivatives of the objective,

$$\widetilde{\mathcal{L}}'(w) = -\frac{1}{n}\sum_{i=1}^{n}\frac{z_i}{1+e^{z_i w}} + \lambda w, \quad \widetilde{\mathcal{L}}''(w) = \frac{1}{n}\sum_{i=1}^{n}\frac{z_i^2}{(1+e^{-z_i w})(1+e^{z_i w})} + \lambda.$$

Setting $\widetilde{\mathcal{L}}'(w_\lambda) = 0$ and using the same argument as the proof of Lemma 20 shows that

$$\exp(\gamma w_\lambda) = (1 \pm o(1))\frac{\gamma^2 p}{\lambda\ln(1/\lambda)},$$

where $p = |\{i : z_i = \gamma\}|/n$. Then for any $w \ge w_\lambda - 1$, we have

$$\begin{aligned}
\widetilde{\mathcal{L}}''(w) &= \frac{1}{n}\sum_{i=1}^{n}\frac{z_i^2}{(1+e^{-z_i w})(1+e^{z_i w})} + \lambda \\
&\le \frac{1}{n}\sum_{i=1}^{n}\frac{1}{(1+e^{z_i w})} + \lambda \\
&\le \frac{1}{n}\sum_{i=1}^{n} e^{-z_i w} + \lambda \\
&\le e^{\gamma}\exp(-\gamma w_\lambda) + \lambda \\
&= (1\pm o(1))\frac{e}{\gamma^2 p}\lambda\ln(1/\lambda) + \lambda \\
&\le C\lambda\ln(1/\lambda),
\end{aligned}$$

for sufficiently small λ and C that depends on p and γ. Moreover, observe that $\widetilde{\mathcal{L}}''(w) > \lambda$ for all w.

Observe that $\widetilde{\mathcal{L}}'(\cdot)$ is increasing and that $\widetilde{\mathcal{L}}'(w_\lambda) = 0$, so we have

$$\widetilde{\mathcal{L}}'(w)\begin{cases}\le 0 & w \le w_\lambda \\ \ge 0 & w \ge w_\lambda.\end{cases}$$

For any w, by Taylor's theorem, since $\widetilde{\mathcal{L}}'(w_\lambda) = 0$, there exists a v between w and w_λ such that

$$\widetilde{\mathcal{L}}'(w) = \widetilde{\mathcal{L}}''(v)(w - w_\lambda).$$

If $w \ge w_\lambda$, $v \ge w_\lambda$ and so this implies

$$\lambda(w - w_\lambda) \le \widetilde{\mathcal{L}}'(w) \le C\lambda\ln(1/\lambda)(w - w_\lambda).$$

Alternatively, if $w_\lambda - 1 \le w \le w_\lambda$, it implies $\widetilde{\mathcal{L}}'(w) \ge C\lambda\ln(1/\lambda)(w - w_\lambda)$. Finally, for any $w \le w_\lambda$, it implies

$$-1 < \widetilde{\mathcal{L}}'(w) \le \lambda(w - w_\lambda).$$

Let t_0 be the first time such that $w_t > w_\lambda$; note that t_0 might be infinite.

Consider $t < t_0$. By definition, we have $w_t \le w_\lambda$ for all $t < t_0$. Thus

$$\begin{aligned} 0 \ge w_t - w_\lambda &= w_{t-1} - w_\lambda - \eta \widetilde{\mathcal{L}}'(w_{t-1}) \\ &\ge w_{t-1} - w_\lambda - \eta\lambda(w_{t-1} - w_\lambda) = (1 - \eta\lambda)(w_{t-1} - w_\lambda). \end{aligned}$$

This implies that $|w_t - w_\lambda| \le (1 - \eta\lambda)^t |w_0 - w_\lambda|$ for $t < t_0$. Hence, for

$$t \ge \frac{1}{\eta\lambda} \ln(|w_0| + w_\lambda),$$

either $t \ge t_0$ or $w_\lambda - 1 \le w_{t-1} \le w_\lambda$. In the latter case

$$w_t - w_\lambda = w_{t-1} - w_\lambda - \eta \widetilde{\mathcal{L}}(w_{t-1}) \le (w_{t-1} - w_\lambda)\,(1 - C\eta\lambda \ln(1/\lambda)) \le 0,$$

provided C_1 is chosen sufficiently large. And in this case, $w_t \le w_\lambda$ for all subsequent t. That is, either t_0 is infinite, in which case the step complexity is $O(\ln((|w_0| + w_\lambda)/\varepsilon)/(\eta\lambda))$, or

$$t_0 \le \frac{1}{\eta\lambda} \ln(|w_0| + w_\lambda).$$

Consider $t \ge t_0$. By definition we have $w_{t_0} > w_\lambda$. We show $w_t \ge w_\lambda$ for all $t \ge t_0$ by induction. Recall the stepsize condition that $\eta < 1/(2C\lambda \ln(1/\lambda))$. Assume that $w_t \ge w_\lambda$, then

$$w_{t+1} = w_t - \eta \widetilde{\mathcal{L}}'(w_t) \ge w_t - \eta C\lambda \ln(1/\lambda)(w_t - w_\lambda) \ge w_t - \frac{1}{2}(w_t - w_\lambda) \ge \frac{1}{2}(w_t + w_\lambda) \ge w_\lambda.$$

So by induction, we have $w_t \ge w_\lambda$ for all $t \ge t_0$. Also,

$$0 \le w_{t+1} - w_\lambda = w_t - w_\lambda - \eta \widetilde{\mathcal{L}}'(w_t) \le w_t - w_\lambda - \eta\lambda(w_t - w_\lambda) = (1 - \eta\lambda)(w_t - w_\lambda).$$

That is, $|w_t - w_\lambda| \le (1 - \eta\lambda)^{t - t_0} |w_{t_0} - w_\lambda|$ for $t \ge t_0$. Finally, notice that

$$w_\lambda < w_{t_0} \le w_{t_0 - 1} + \eta \le w_\lambda + \eta,$$

so $|w_{t_0} - w_\lambda| \le \eta = \Theta(1/(\lambda \ln(1/\lambda)))$. So we have $|w_t - w_\lambda| \le (1 - \eta\lambda)^{t - t_0} \Theta(1/(\lambda \ln(1/\lambda)))$ for $t \ge t_0$.

Combining with the bound on t_0 shows that the step complexity is

$$O\left(\frac{1}{\eta\lambda} \ln\left(\frac{|w_0| + \ln(1/(\lambda \ln(1/\lambda)))}{\varepsilon\lambda \ln(1/\lambda)}\right)\right) = O\left(\frac{1}{\eta\lambda} \ln\left(\frac{|w_0| + 1}{\varepsilon\lambda \ln(1/\lambda)}\right)\right).$$

This completes our proof. □

E Experimental details

The dataset is composed on two datapoints $x_1 = (\gamma, 1)$ and $x_2 = (\gamma, -2)$ for $\gamma = 0.2$. We run GD on the regularized logistic regression for $\lambda = 2^{-12}$, a logarithmic range of stepsizes from 2^1 to 2^{13}, and 2^{13} steps. Additional plots are given in Figure 3. Two comments are of interest. First, we observe that the dynamics converge for stepsizes up to 2^5. This is consistent with the local stability threshold given by $2/\|\nabla \widetilde{L}(\mathbf{w}_\lambda)\|_2 \approx 44.8$. Second, we observe in the case of $\eta = 2^5$ that even after $\widetilde{L}$ stops oscillating for $t \approx 2^4$ (start of the stable phase), both the regularization and the logistic components continue to evolve nonmonotonically. This is connected to the discussion in Section 2.3, where we outline that a decrease of $\widetilde{\mathcal{L}}(\mathbf{w})$ may cause an increase of $\mathcal{L}(\mathbf{w})$, and then GD might leave the stable region.

The code was implemented in JAX (Bradbury et al., 2018) and takes a few seconds to run on a consumer laptop. Our code is available at https://github.com/PierreMarion23/large-stepsize-regularized-logistic.

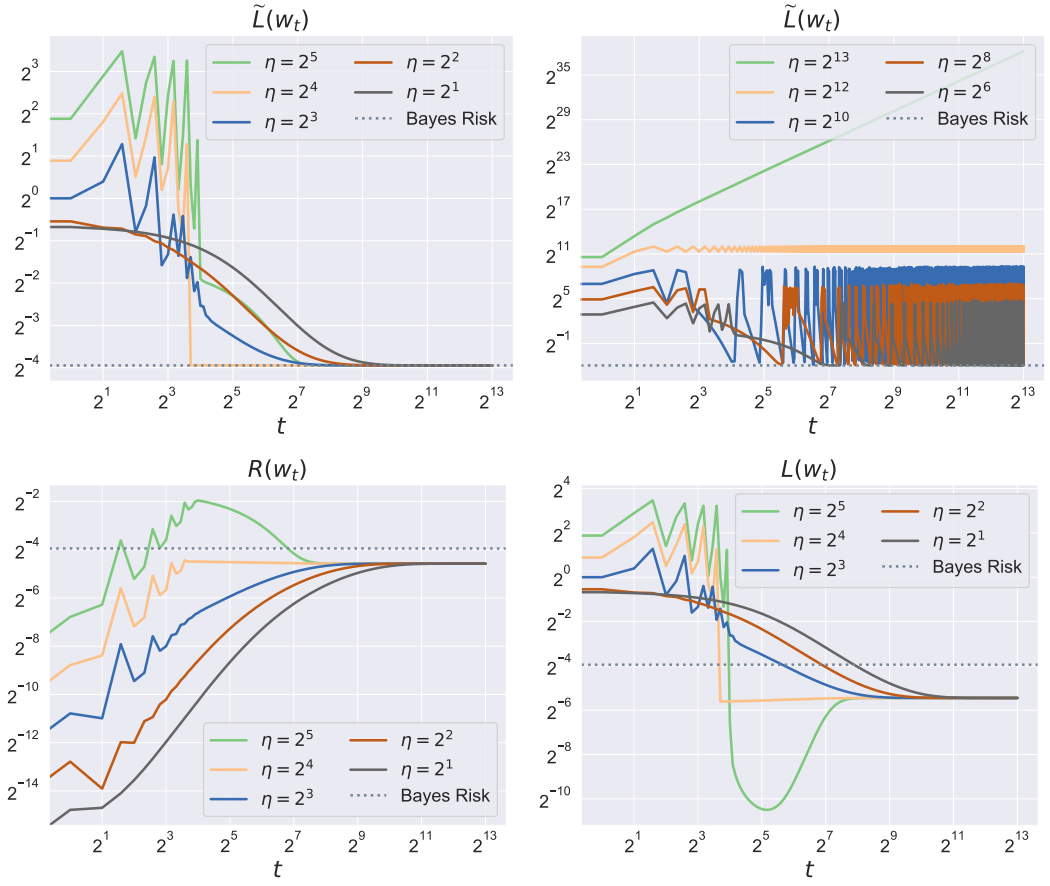

Figure 3: Additional plots for the 2-dimensional experiment. **Top left:** Objective value as a function of training steps. **Top right:** Objective value as a function of training steps for even larger stepsizes. **Bottom left:** Value of the regularization component as a function of training steps. **Bottom right:** Value of the logistic component as a function of training steps.