Characterizing the implicit bias via a primal-dual analysis

Ziwei Ji ZIWEIJI2@ILLINOIS.EDU
University of Illinois, Urbana-Champaign
Matus Telgarsky MJT@ILLINOIS.EDU
University of Illinois, Urbana-Champaign

Editors: Vitaly Feldman, Katrina Ligett and Sivan Sabato

Abstract

This paper shows that the implicit bias of gradient descent on linearly separable data is exactly characterized by the optimal solution of a dual optimization problem given by a smoothed margin, even for general losses. This is in contrast to prior results, which are often tailored to exponentially-tailed losses. For the exponential loss specifically, with n training examples and t gradient descent steps, our dual analysis further allows us to prove an $O\left(\ln(n)/\ln(t)\right)$ convergence rate to the ℓ_2 maximum margin direction, when a constant step size is used. This rate is tight in both n and t, which has not been presented by prior work. On the other hand, with a properly chosen but aggressive step size schedule, we prove $O(1/t)$ rates for both ℓ_2 margin maximization and implicit bias, whereas prior work (including all first-order methods for the general hard-margin linear SVM problem) proved $\widetilde{O}(1/\sqrt{t})$ margin rates, or $O(1/t)$ margin rates to a suboptimal margin, with an implied (slower) bias rate. Our key observations include that gradient descent on the primal variable naturally induces a mirror descent update on the dual variable, and that the dual objective in this setting is smooth enough to give a faster rate.

1. Introduction

Recent work has shown that in deep learning, the solution found by gradient descent not only gives low training error, but also has low complexity and thus generalizes well (Zhang et al., 2016; Bartlett et al., 2017). This motivates the study of the implicit bias of gradient descent: amongst all choices with low training error, which is preferred by gradient descent?

This topic has been extensively studied recently: specifically on linear classifiers, Soudry et al. (2017) show that with linearly separable data and exponentially-tailed losses (such as the exponential loss and the logistic loss), gradient descent converges to the ℓ_2 maximum margin direction. Ji and Telgarsky (2018b) further characterize the implicit bias in the nonseparable setting, while Gunasekar et al. (2018a) consider generic optimization algorithms such as steepest descent and adaptive gradient descent.

However, as detailed below, most prior results rely on exponentially-tailed losses, and do not prove tight rates for a range of step size schedules. In this work, we focus on linear classifiers and linearly separable data, and make contributions along all these directions:

- We prove that for a broad class of losses that asymptote to 0, including exponentially-tailed losses, polynomially-tailed losses and others, the gradient descent iterates grow unboundedly, but their directions (i.e., the normalized gradient descent iterates) converge to some point given by the dual optimization problem corresponding to a specific smoothed margin function with an $O(1/\sqrt{t})$ rate (cf. Theorem 5). Previously, Ji et al. (2020) also handle general losses, and they prove that the gradient descent iterates converge to the same direction as regularized solutions. However, they do not further give a closed-form characterization of the implicit bias, and their convergence result is asymptotic.

- For the exponential/logistic loss, we can use a much more aggressive step size schedule, with which we prove an $O(\ln(n)/t)$ rate for ℓ_2 margin maximization (cf. Theorem 7). For the exponential loss, we can also prove an $O(\ln(n)/t)$ rate for convergence to the implicit bias (cf. Theorem 8). Such a step size schedule is also used in AdaBoost (Freund and Schapire, 1997); however, it does not always maximize the corresponding ℓ_1 margin, with counterexamples given in (Rudin et al., 2004). To maximize the ℓ_1 margin, we need to shrink the step sizes, and prior work has shown either an $O(1/t)$ convergence rate to a suboptimal margin (Telgarsky, 2013), or an $\widetilde{O}(1/\sqrt{t})$ convergence rate to the maximum margin (Nacson et al., 2018). Their proof ideas can be applied to generic steepest descent, but cannot prove our $O(\ln(n)/t)$ margin maximization rate.

- On the other hand, with a constant step size that is more widely used, for the exponential loss we prove a tight $O\left(\ln(n)/\ln(t)\right)$ rate for the directional convergence of gradient descent to the ℓ_2 maximum margin direction (cf. Theorem 8). Previously, Soudry et al. (2017) prove an $O\left(1/\ln(t)\right)$ rate, but the dependency on n is not specified; it should be carefully handled since the denominator only grows at a rate of $\ln(t)$. Ji and Telgarsky (2018b) consider general nonseparable data, but their convergence rate for the implicit bias in the separable setting is $O\left(\sqrt{\ln(n)/\ln(t)}\right)$, which is quadratically slower than our rate.

All of our results are based on a primal-dual analysis of gradient descent. One key observation is that gradient descent on the primal variable induces exactly a mirror descent update on the dual variable. This perspective has been studied in (Freund et al., 2013) for boosting / coordinate descent. However, they only prove an $\widetilde{O}(1/\sqrt{t})$ dual convergence rate, while we can further prove an $O(1/t)$ dual rate by exploiting the smoothness of the dual objective (cf. Theorem 1). More surprisingly, our dual analysis further gives rise to a faster primal convergence guarantee (cf. Theorem 1), which allows us to prove the $O(1/t)$ margin maximization and implicit bias rates.

1.1. Related work

Margin maximization and implicit bias are heavily studied in the context of boosting methods (Schapire et al., 1997; Schapire and Freund, 2012; Shalev-Shwartz and Singer, 2008). Boosting methods are themselves a form of coordinate descent, one whose convergence is difficult to analyze (Schapire, 2010); interestingly, the original proof of AdaBoost's empirical risk convergence also uses an analysis in the dual (Collins et al., 2002), though without any rate. This same dual analysis, and also work by Kivinen and Warmuth (1999), point out that AdaBoost, in the dual, performs iterative Bregman projection.

As mentioned above, prior work did not prove margin maximization rates better than $O(1/\sqrt{t})$, possibly owing to the nonsmoothness of the problem. This topic is discussed in (Nacson et al., 2018, Section 3), where it is stated that the current best rate is $O(1/\sqrt{t})$ for the general hard-margin linear SVM problem via first-order methods, that is, not merely restricting to the framework in this present work, which applies gradient descent to smooth losses which asymptote to zero.

Regarding lower bounds, Clarkson et al. (2012) prove that, under a few conditions including $\epsilon^{-2} = O\left(\min\{n, d\}\right)$ where n is the number of data examples and d is the input dimension, to maximize the margin up to an additive error of ϵ, the optimization algorithm has to read $\Omega\left(\epsilon^{-2}(n+d)\right)$ entries of the data matrix. Due to the required condition, this lower bound is basically $\Omega\left((n+d)\min\{n,d\}\right) = \Omega(nd)$. On the other hand, in this paper we analyze full-batch

gradient descent, which reads nd entries of the data matrix at each step, and therefore does not violate this lower bound. More generally, the lower bound for general nonsmooth optimization is $1/\sqrt{t}$ (Nesterov, 2004, Theorem 3.2.1), which is arguably one source of difficulty, as the limiting hard-margin problem is nonsmooth.

The implicit bias of gradient descent has also been studied in more complicated models, such as deep linear networks and homogeneous networks (Gunasekar et al., 2018b; Ji and Telgarsky, 2018a; Lyu and Li, 2019; Chizat and Bach, 2020; Woodworth et al., 2020; Ji and Telgarsky, 2020).

1.2. Notation

In this paper, $\|\cdot\|$ denotes the ℓ_2-norm. The dataset is denoted by $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ satisfies $\|x_i\| \leq 1$, and $y_i \in \{-1, +1\}$. We consider linear classfiers, and the corresponding unbounded, unregularized empirical risk minimization problem:

$$\min_{w\in\mathbb{R}^d} \mathcal{R}(w) := \frac{1}{n}\sum_{i=1}^n \ell\left(-y_i\langle w, x_i\rangle\right) = \frac{1}{n}\sum_{i=1}^n \ell\left(\langle w, z_i\rangle\right),$$

where $z_i := -y_i x_i$, and we collect them into a matrix $Z \in \mathbb{R}^{n\times d}$ whose i-th row is $z_i^\top$. We assume the loss function $\ell : \mathbb{R} \to \mathbb{R}$ satisfies the following conditions.

Assumption 1 *The loss function ℓ satisfies:*

1. *$\ell, \ell', \ell'' > 0$, and $\lim_{z\to-\infty} \ell(z) = 0$.*
2. *$z\ell'(z)/\ell(z)$ is increasing on $(-\infty, 0)$, and $\lim_{z\to-\infty} z\ell'(z) = 0$.*
3. *For all $b \geq 1$, there exists $c > 0$ (which can depend on b), such that for all $a > 0$, we have $\ell'\left(\ell^{-1}(a)\right)/\ell'\left(\ell^{-1}(ab)\right) \geq c$.*
4. *Given $\xi \in \mathbb{R}^n$, define*

$$\mathcal{L}(\xi) := \sum_{i=1}^n \ell(\xi_i), \quad \text{and} \quad \psi(\xi) := \ell^{-1}\left(\mathcal{L}(\xi)\right).$$

Then ψ is convex and β-smooth with respect to the ℓ_∞ norm.

Note that by definition, $\mathcal{R}(w) = \mathcal{L}(Zw)/n$. The function ψ is called a "generalized sum" (Hardy et al., 1934), and if the dataset is linearly separable, it can be interpreted as a "smoothed margin" (Lyu and Li, 2019). In Sections 3 and 4, we will use ψ to characterize the implicit bias and prove faster margin rates.

The most interesting example satisfying Assumption 1 is the exponential loss $\ell_{\exp}(z) := e^z$, in which case $\psi(\xi) = \ln\left(\sum_{i=1}^n \exp(\xi_i)\right)$. However, many of our results hold for any loss function satisfying Assumption 1, such as the logistic loss $\ell_{\log}(z) := \ln(1 + e^z)$, and polynomially-tailed losses (cf. Theorem 11). Note that the smoothness constant β will affect the convergence rate, and although $\beta = 1$ for the exponential loss, it can be as large as n for other losses such as the logistic loss (cf. Lemma 13). In such settings, we can make a finer analysis which uses the smoothness constant on sublevel sets: for example, for the logistic loss, ψ is 2-smooth on $\{\xi|\psi(\xi) \leq 0\}$ (cf. Lemma 14). We still assume global smoothness in Assumption 1 since it can simplify the analysis a lot and highlight the key ideas.

2. A primal-dual convergence analysis for gradient descent

In this section, we start analyzing gradient descent on the (primal) risk $\mathcal{R}(w)$. We show that it naturally induces a mirror descent update on the dual variable, and prove a primal-dual convergence result (cf. Theorem 1), which will be used in subsequent sections to give a characterization of the implicit bias, and prove fast convergence rates.

Gradient descent on $\mathcal{R}(w)$ starts from some initialization w_0, and sets $w_{t+1} := w_t - \eta_t \nabla \mathcal{R}(w_t)$ for $t \geq 0$. For each gradient descent iterate w_t, let $p_t := Z w_t$ and $q_t := \nabla \psi(p_t)$; we call q_t the corresponding dual variable of w_t. Note that

$$q_{t,i} = \frac{\ell'(p_{t,i})}{\ell'\left(\psi(p_t)\right)} = \frac{\ell'\left(\langle w_t, z_i \rangle\right)}{\ell'\left(\psi(Z w_t)\right)},$$

and thus

$$w_{t+1} = w_t - \eta_t \nabla \mathcal{R}(w_t) = w_t - \hat{\eta}_t Z^\top q_t,$$

where $\hat{\eta}_t := \eta_t \ell'\left(\psi(Z w_t)\right)/n$, which will be extensively used in our analysis.

The key observation is that the induced update on the dual variable q_t is actually a mirror descent (more exactly, a dual averaging) update:

$$p_{t+1} = p_t - \hat{\eta}_t Z Z^\top q_t = p_t - \hat{\eta}_t \nabla f(q_t), \quad \text{and} \quad q_{t+1} = \nabla \psi(p_{t+1}), \tag{1}$$

where $f(q) := \left\| Z^\top q \right\|^2 / 2$. Therefore we can use a mirror descent analysis to prove a dual convergence result for f. Additionally, the analysis also allows us to prove a primal convergence rate for ψ, which is tight up to a constant.

Let ψ^* denote the convex conjugate of ψ. Given $q \in \operatorname{dom} \psi^*$, the generalized Bregman distance (Gordon, 1999) between q and q_t is defined as

$$D_{\psi^*}(q, q_t) := \psi^*(q) - \psi^*(q_t) - \langle p_t, q - q_t \rangle.$$

It is a generalization of the Bregman distance to the nondifferentiable setting, since ψ^* may not be differentiable at q_t; instead we just use p_t in the definition. Here is our main convergence result.

Theorem 1 *Under Assumption 1, for all $q \in \operatorname{dom} \psi^*$, if $\hat{\eta}_t \leq 1/\beta$, then the following results hold:*

1. *Dual convergence: for all $t \geq 0$,*

$$f(q_{t+1}) \leq f(q_t), \quad \text{and} \quad \hat{\eta}_t \left(f(q_{t+1}) - f(q) \right) \leq D_{\psi^*}(q, q_t) - D_{\psi^*}(q, q_{t+1}).$$

   As a result, for all $t > 0$,

$$f(q_t) - f(q) \leq \frac{D_{\psi^*}(q, q_0) - D_{\psi^*}(q, q_t)}{\sum_{j<t} \hat{\eta}_j} \leq \frac{D_{\psi^*}(q, q_0)}{\sum_{j<t} \hat{\eta}_j}.$$

2. *Primal convergence: for all $t \geq 0$,*

$$\psi(p_t) - \psi(p_{t+1}) \geq \hat{\eta}_t \left(f(q_t) + f(q_{t+1}) \right) = \frac{\hat{\eta}_t}{2} \left\| Z^\top q_t \right\|^2 + \frac{\hat{\eta}_t}{2} \left\| Z^\top q_{t+1} \right\|^2,$$

and thus if $\hat{\eta}_t$ is nonincreasing, then

$$\psi(p_0) - \psi(p_t) \geq \sum_{j<t} \hat{\eta}_j \left\|Z^\top q_j\right\|^2 - \frac{\hat{\eta}_0}{2}\left\|Z^\top q_0\right\|^2 + \frac{\hat{\eta}_t}{2}\left\|Z^\top q_t\right\|^2.$$

This rate is tight up to a constant, since $\psi(p_0) - \psi(p_t) \leq \sum_{j<t} \hat{\eta}_j \left\|Z^\top q_j\right\|^2$.

Here are some comments on Theorem 1.

- If we let $\hat{\eta}_t = 1/\beta$, then we get an $O(1/t)$ dual convergence rate. By contrast, Freund et al. (2013) consider boosting, and can only handle step size $\hat{\eta}_t \propto 1/\sqrt{t+1}$ and give an $\widetilde{O}(1/\sqrt{t})$ dual rate. This is because the dual objective $f(q) := \left\|Z^\top q\right\|^2/2$ for gradient descent is smooth, while for boosting the dual objective is given by $\left\|Z^\top q\right\|_\infty^2/2$, which is nonsmooth. In some sense, we can handle a constant $\hat{\eta}_t$ and prove a faster rate because *both the primal objective ψ and the dual objective f are smooth.*

- Moreover, the primal and dual smoothness allow us to prove a super tight primal convergence rate for ψ. By contrast, if we use a standard smoothness guarantee and $\hat{\eta}_t = 1/\beta$, then the error term (compared with the upper bound on $\psi(p_0) - \psi(p_t)$) can be as large as $\sum_{j<t} \hat{\eta}_j \left\|Z^\top q_j\right\|^2/2$ (cf. Lemma 3). While a constant factor does not hurt the risk bound too much, it can stop us from proving an $O(1/t)$ margin maximization rate with a constant step size for the exponential loss (cf. Section 4).

- For the exponential loss (and other exponentially-tailed losses), Soudry et al. (2017) prove that w_t converges to the maximum margin direction. This is called an "implicit bias" result since it does not follow from classical results such as risk minimization, and requires a nontrivial proof tailored to the exponential function. By contrast, Theorem 1 explicitly shows that the dual iterates minimize the dual objective f, and the minimum of f is given exactly by the maximum margin (cf. eq. (9)). Moreover, this dual perspective and Theorem 1 can help us characterize the implicit bias of a general loss function (cf. Theorem 5).

2.1. Proof of Theorem 1

Here we sketch the proof of Theorem 1. Omitted details are given in Appendix B.

The most important property we use is the ℓ_1 smoothness of f. As mentioned before, this is the key tool to prove the fast $1/t$ rate; for boosting (coordinate descent), the dual objective is not smooth, which might be the reason why prior results only have $1/\sqrt{t}$ rates.

Lemma 2 *The function $f : \mathbb{R}^n \to \mathbb{R}$ given by $f(\theta) := \left\|Z^\top \theta\right\|^2/2$ is 1-smooth with respect to the ℓ_1 norm.*

Proof For any $\theta, \theta' \in \mathbb{R}^n$, using the Cauchy-Schwarz inequality and $\|z_i\| \leq 1$,

$$\begin{aligned}\left\|\nabla f(\theta) - \nabla f(\theta')\right\|_\infty = \left\|ZZ^\top(\theta - \theta')\right\|_\infty &= \max_{1\leq i\leq n} \left|\left\langle Z^\top(\theta - \theta'), z_i\right\rangle\right| \\ &\leq \max_{1\leq i\leq n} \left\|Z^\top(\theta - \theta')\right\| \|z_i\| \\ &\leq \left\|Z^\top(\theta - \theta')\right\|.\end{aligned}$$

Furthermore, by the triangle inequality and $\|z_i\| \le 1$,

$$\left\|Z^\top(\theta-\theta')\right\| \le \sum_{i=1}^n \left|\theta_i-\theta_i'\right| \|z_i\| \le \sum_{i=1}^n \left|\theta_i-\theta_i'\right| = \left\|\theta-\theta'\right\|_1.$$

Therefore f is 1-smooth with respect to the ℓ_1 norm. ■

Here are some standard results we need, from the smoothness of ψ; a proof is given in Appendix B for completeness. Some refined results are given in Lemma 14, which use the smoothness constants over sublevel sets that could be much better.

Lemma 3 *We have*

$$\psi(p_{t+1})-\psi(p_t) \le -\hat\eta_t \left\|Z^\top q_t\right\|^2 + \frac{\beta\hat\eta_t^2}{2}\left\|Z^\top q_t\right\|^2 \qquad \text{and} \qquad D_{\psi^*}(q_{t+1},q_t) \ge \frac{1}{2\beta}\|q_{t+1}-q_t\|_1^2.$$

Next is a standard result for mirror descent; a proof is also included in Appendix B.

Lemma 4 *For any $t \ge 0$ and $q \in \operatorname{dom}\psi^*$, it holds that*

$$\hat\eta_t\left(f(q_t)-f(q)\right) \le \langle \hat\eta_t \nabla f(q_t), q_t - q_{t+1}\rangle - D_{\psi^*}(q_{t+1},q_t) + D_{\psi^*}(q,q_t) - D_{\psi^*}(q,q_{t+1}).$$

Moreover, q_{t+1} is the unique minimizer of

$$h(q) := f(q_t) + \langle \nabla f(q_t), q-q_t\rangle + \frac{1}{\hat\eta_t} D_{\psi^*}(q,q_t),$$

and specifically $h(q_{t+1}) \le h(q_t) = f(q_t)$.

With Lemmas 2 to 4, we can prove Theorem 1.

Proof (of Theorem 1) Since f is 1-smooth with respect to the ℓ_1 norm,

$$f(q_{t+1}) - f(q_t) \le \langle \nabla f(q_t), q_{t+1}-q_t\rangle + \frac{1}{2}\|q_{t+1}-q_t\|_1^2.$$

Further invoking Lemma 3, and $\hat\eta_t \le 1/\beta$, and the function h defined in Lemma 4, we have

$$\begin{aligned} f(q_{t+1}) &\le f(q_t) + \langle \nabla f(q_t), q_{t+1}-q_t\rangle + \frac{1}{2}\|q_{t+1}-q_t\|_1^2 \\ &\le f(q_t) + \langle \nabla f(q_t), q_{t+1}-q_t\rangle + \beta D_{\psi^*}(q_{t+1},q_t) \\ &\le f(q_t) + \langle \nabla f(q_t), q_{t+1}-q_t\rangle + \frac{1}{\hat\eta_t} D_{\psi^*}(q_{t+1},q_t) \\ &= h(q_{t+1}) \le f(q_t), \end{aligned} \tag{2}$$

which proves that $f(q_t)$ is nonincreasing.

To prove the iteration guarantee for f, note that rearranging the terms of eq. (2) gives

$$\hat\eta_t \langle \nabla f(q_t), q_{t+1}-q_t\rangle + D_{\psi^*}(q_{t+1},q_t) \ge \hat\eta_t\left(f(q_{t+1})-f(q_t)\right).$$

Lemma 4 then implies

$$\hat{\eta}_t \left(f(q_t) - f(q) \right) \leq \hat{\eta}_t \left(f(q_t) - f(q_{t+1}) \right) + D_{\psi^*}(q, q_t) - D_{\psi^*}(q, q_{t+1}).$$

Rearranging terms gives

$$\hat{\eta}_t \left(f(q_{t+1}) - f(q) \right) \leq D_{\psi^*}(q, q_t) - D_{\psi^*}(q, q_{t+1}). \tag{3}$$

Taking the sum of eq. (3) from 0 to $t-1$, and noting that $f(q_{j+1}) \geq f(q_t)$ for all $j < t$ since f is nonincreasing, the proof is done.

To prove the iteration guarantee for ψ, note that

$$\begin{aligned} D_{\psi^*}(q_{t+1}, q_t) &= \psi^*(q_{t+1}) - \psi^*(q_t) - \langle p_t, q_{t+1} - q_t \rangle \\ &= \langle p_{t+1}, q_{t+1} \rangle - \psi(p_{t+1}) - \langle p_t, q_t \rangle + \psi(p_t) - \langle p_t, q_{t+1} - q_t \rangle \\ &= \psi(p_t) - \psi(p_{t+1}) - \langle q_{t+1}, p_t - p_{t+1} \rangle \\ &= \psi(p_t) - \psi(p_{t+1}) - \hat{\eta}_t \left\langle Z^\top q_t, Z^\top q_{t+1} \right\rangle, \end{aligned} \tag{4}$$

where we used eq. (1). Therefore eq. (2) ensures

$$\begin{aligned} \psi(p_t) - \psi(p_{t+1}) &\geq \hat{\eta}_t \left(f(q_{t+1}) - f(q_t) - \langle \nabla f(q_t), q_{t+1} - q_t \rangle + \left\langle Z^\top q_t, Z^\top q_{t+1} \right\rangle \right) \\ &= \frac{\hat{\eta}_t}{2} \left\| Z^\top q_t \right\|^2 + \frac{\hat{\eta}_t}{2} \left\| Z^\top q_{t+1} \right\|^2. \end{aligned}$$

Telescoping gives the lower bound on $\psi(p_0) - \psi(p_t)$. For the upper bound, note that ψ is convex, and thus

$$\psi(p_t) - \psi(p_{t+1}) \leq \langle q_t, p_t - p_{t+1} \rangle = \left\langle q_t, \hat{\eta}_t Z Z^\top q_t \right\rangle = \hat{\eta}_t \left\| Z^\top q_t \right\|^2.$$

■

3. The dual optimal solution characterizes the implicit bias

For our implicit bias and margin maximization results, we assume the data examples are linearly separable.

Assumption 2 *There exists $u \in \mathbb{R}^d$ such that $y_i \langle u, x_i \rangle > 0$ for all i.*

As mentioned before, most prior results on the implicit bias are focused on exponentially-tailed losses. Ji et al. (2020) consider general losses, and show that the gradient descent iterates and regularized solutions converge to the same direction, but give no closed-form characterization of the implicit bias or convergence rate. In the following result, we characterize the implicit bias using the dual optimal solution.

Theorem 5 *Under Assumptions 1 and 2, suppose $\hat{\eta}_t = \eta_t \ell' \left(\psi(Z w_t) \right) / n \leq 1/\beta$ is nonincreasing, and $\sum_{t=0}^{\infty} \hat{\eta}_t = \infty$.*

1. *The set* $\{q|\psi^*(q) \le 0\}$ *is nonempty, compact and convex. Moreover* $\min_{\psi^*(q)\le 0} f(q) > 0$, *and* $Z^\top \bar{q}$ *is the same for all* $\bar{q} \in \arg\min_{\psi^*(q)\le 0} f(q)$.

2. *For* $\bar{q} \in \arg\min_{\psi^*(q)\le 0} f(q)$, *and all* t *with* $\psi(Zw_t) \le 0$ *(which holds for all large enough* t*), we have*

$$\left\|Z^\top q_t - Z^\top \bar{q}\right\|^2 \le \frac{2D_{\psi^*}(\bar{q}, q_0)}{\sum_{j<t}\hat{\eta}_j}, \qquad \text{and} \qquad \left\langle \frac{w_t}{\|w_t\|}, \frac{-Z^\top \bar{q}}{\|Z^\top \bar{q}\|} \right\rangle \ge 1 - \frac{\delta(w_0, \bar{q})}{\sum_{j<t}\hat{\eta}_j},$$

where $\delta(w_0, \bar{q}) := \left(\psi(p_0) + \hat{\eta}_0 f(q_0) + \|w_0\| \left\|Z^\top \bar{q}\right\|\right) / \left(2f(\bar{q})\right)$ *is a constant depending only on* w_0 *and* $\bar{q}$. *In particular, it holds that the implicit bias is*

$$\lim_{t\to\infty} \frac{w_t}{\|w_t\|} = -\frac{Z^\top \bar{q}}{\|Z^\top \bar{q}\|}.$$

Theorem 5 is partly proved by establishing a lower bound on $-\psi(Zw_t)/\|w_t\|$, which will also help us prove the $O(1/t)$ margin maximization rate for the exponential loss (cf. Theorem 7). Note also that while the condition $\psi^*(q) \le 0$ in the definition of $\bar{q}$ looks technical, it appears naturally when deriving $\bar{q}$ as the solution to the convex dual of the smoothed margin, as in Appendix A.

3.1. Proof of Theorem 5

Here is a proof sketch of Theorem 5. Omitted proofs are given in Appendix C.

Note that for $\bar{q}$ defined in Theorem 5, it already follows from Theorem 1 that $\lim_{t\to\infty} f(q_t) \le f(\bar{q})$. We further need the following result to ensure $f(q_t) \ge f(\bar{q})$; its proof is basically identical to the proof of (Ji and Telgarsky, 2020, Lemma 3.5), and is included in Appendix C for completeness. We then have $\lim_{t\to\infty} = f(\bar{q})$, which is crucial in the proof of Theorem 5.

Lemma 6 *For any* $\xi \in \mathbb{R}^n$ *such that* $\psi(\xi) \le 0$, *it holds that* $\psi^*\left(\nabla\psi(\xi)\right) \le 0$.

Part 1 of Theorem 5 is fully proved in Appendix C; here we sketch its proof. The set $S_0 := \{q|\psi^*(q) \le 0\}$ is nonempty because of Lemma 6: note that if $\ell(\xi_i) \le \ell(0)/n$ for all i, then $\psi(\xi) \le 0$. It holds that S_0 is closed convex since it is a sublevel set of the closed convex function ψ^*. To show the boundedness of S_0, note that (Rockafellar, 1970, Theorem 23.5) ensures $\operatorname{range}\nabla\psi = \operatorname{dom}\partial\psi^*$, while (Rockafellar, 1970, Theorem 23.4) ensures $\operatorname{dom}\partial\psi^*$ contains the relative interior of $\operatorname{dom}\psi^*$; therefore we only need to consider $q = \nabla\psi(\xi)$ for some $\xi \in \mathbb{R}^n$. It follows from the definition of ψ that $\|\nabla\psi(\xi)\|_1 \le n$, and thus S_0 is bounded. On the other hand, it can be shown using Assumption 1 that $\|\nabla\psi(\xi)\|_1$ is bounded below by a positive constant, which implies $\left\|Z^\top\nabla(\xi)\right\|$ is also bounded below by a positive constant, using the following argument: Assumption 2 ensures there exists a unit vector $u \in \mathbb{R}^d$ and $\gamma > 0$ such that for all $1 \le i \le n$, it holds that $\langle u, -z_i\rangle = y_i\langle u, x_i\rangle \ge \gamma$. Further note that $\nabla\psi(\xi)_i > 0$, we have

$$\left\|Z^\top\nabla\psi(\xi)\right\| \ge \left\langle -Z^\top\nabla\psi(\xi), u\right\rangle = \sum_{i=1}^n \langle u, -z_i\rangle\nabla\psi(\xi)_i \ge \gamma\left\|\nabla\psi(\xi)\right\|_1.$$

Finally, to prove the uniqueness of $Z^\top\bar{q}$, note that if $\bar{q}_1, \bar{q}_2$ are two minimizers of f on S_0, with $\left\|Z^\top\bar{q}_1\right\| = \left\|Z^\top\bar{q}_2\right\| > 0$ but $Z^\top\bar{q}_1 \ne Z^\top\bar{q}_2$, then

$$\frac{Z^\top\bar{q}_1 + Z^\top\bar{q}_2}{2} \in S_0, \quad \text{but} \quad \left\|\frac{Z^\top\bar{q}_1 + Z^\top\bar{q}_2}{2}\right\| < \left\|Z^\top\bar{q}_1\right\|,$$

a contradiction.

Next we prove Part 2 of Theorem 5.

Proof (of Theorem 5 Part 2) Since $\sum_{t=0}^{\infty} \hat{\eta}_t = \infty$, Theorem 1 implies that $\lim_{t\to\infty} f(q_t) \le f(\bar{q})$. On the other hand, by Theorem 1 and Part 1 of Theorem 5, $\psi(p_t) \le 0$ for all large enough t, and then Lemma 6 implies that $\psi^*(q_t) \le 0$. By the definition of $\bar{q}$, we have $f(q_t) \ge f(\bar{q})$, and thus $\lim_{t\to\infty} f(q_t) = f(\bar{q})$.

We first prove the convergence of $Z^\top q_t$ to $Z^\top \bar{q}$. Let t be large enough such that $\psi(p_t) \le 0$ and thus $\psi^*(q_t) \le 0$. By the definition of $\bar{q}$ and the first-order optimality condition (Borwein and Lewis, 2000, Proposition 2.1.1), we have

$$\langle \nabla f(\bar{q}), q_t - \bar{q} \rangle \ge 0, \quad \text{and thus} \quad \left\| Z^\top \bar{q} \right\|^2 \le \left\langle Z^\top \bar{q}, Z^\top q_t \right\rangle. \tag{5}$$

Theorem 1 and eq. (5) then imply

$$\begin{aligned}
\left\| Z^\top q_t - Z^\top \bar{q} \right\|^2 &= \left\| Z^\top q_t \right\|^2 - 2 \left\langle Z^\top q_t, Z^\top \bar{q} \right\rangle + \left\| Z^\top \bar{q} \right\|^2 \\
&\le \left\| Z^\top q_t \right\|^2 - \left\| Z^\top \bar{q} \right\|^2 \\
&\le \frac{2 D_{\psi^*}(\bar{q}, q_0)}{\sum_{j<t} \hat{\eta}_j}.
\end{aligned}$$

To prove the other claim, we use an idea from (Ji and Telgarsky, 2018b), but also invoke the tighter guarantee on ψ in Theorem 1. By Fenchel-Young inequality, and recall that $\psi^*(\bar{q}) \le 0$,

$$\left\langle w_t, -Z^\top \bar{q} \right\rangle = -\langle Z w_t, \bar{q} \rangle \ge -\psi(Z w_t) - \psi^*(\bar{q}) \ge -\psi(Z w_t) = -\psi(p_t). \tag{6}$$

Moreover, Theorem 1 implies that

$$\begin{aligned}
-\psi(p_t) &\ge -\psi(p_0) + \sum_{j<t} \hat{\eta}_j \left\| Z^\top q_j \right\|^2 - \frac{\hat{\eta}_0}{2} \left\| Z^\top q_0 \right\|^2 \\
&\ge -\psi(p_0) + \sum_{j<t} \hat{\eta}_j \left\| Z^\top q_j \right\| \left\| Z^\top \bar{q} \right\| - \frac{\hat{\eta}_0}{2} \left\| Z^\top q_0 \right\|^2.
\end{aligned} \tag{7}$$

On the other hand, the triangle inequality implies $\|w_t\| \le \|w_0\| + \sum_{j<t} \hat{\eta}_j \left\| Z^\top q_j \right\|$. Recall that $\psi(p_t) \le 0$, then eq. (7) implies

$$\begin{aligned}
\frac{-\psi(p_t)}{\|w_t\| \left\| Z^\top \bar{q} \right\|} \ge \frac{-\psi(p_t)}{\left(\|w_0\| + \sum_{j<t} \hat{\eta}_j \left\| Z^\top q_j \right\| \right) \left\| Z^\top \bar{q} \right\|} &\ge 1 - \frac{\psi(p_0) + \hat{\eta}_0 f(q_0) + \|w_0\| \left\| Z^\top \bar{q} \right\|}{\left(\|w_0\| + \sum_{j<t} \hat{\eta}_j \left\| Z^\top q_j \right\| \right) \left\| Z^\top \bar{q} \right\|} \\
&\ge 1 - \frac{\psi(p_0) + \hat{\eta}_0 f(q_0) + \|w_0\| \left\| Z^\top \bar{q} \right\|}{2 f(\bar{q}) \sum_{j<t} \hat{\eta}_j}.
\end{aligned} \tag{8}$$

Finally, eqs. (6) and (8) imply that

$$\left\langle \frac{w_t}{\|w_t\|}, \frac{-Z^\top \bar{q}}{\left\| Z^\top \bar{q} \right\|} \right\rangle \ge \frac{-\psi(p_t)}{\|w_t\| \left\| Z^\top \bar{q} \right\|} \ge 1 - \frac{\psi(p_0) + \hat{\eta}_0 f(q_0) + \|w_0\| \left\| Z^\top \bar{q} \right\|}{2 f(\bar{q}) \sum_{j<t} \hat{\eta}_j} = 1 - \frac{\delta(w_0, \bar{q})}{\sum_{j<t} \hat{\eta}_j}.$$

■

4. $1/t$ and $1/\ln(t)$ exponential loss rates with fast and slow steps

In this section we focus primarily on the exponential loss e^z, proving refined margin maximization and implicit bias rates, though the margin maximization rate is also proved for the logistic loss $\ln(1+e^z)$. We fix the initialization to $w_0 = 0$, which can make the bounds cleaner; however our analysis can be easily extended to handle nonzero initialization.

Regarding step sizes, our rates depend on the quantity $\sum_{j<t} \hat{\eta}_j$, which at its largest is t by taking constant $\hat{\eta}_j$, giving rise to both of our $1/t$ rates. This step size choice is in fact extremely aggressive; e.g., for the exponential loss, the induced step sizes on $\nabla\mathcal{R}(w_j)$ are $\eta_j = 1/\mathcal{R}(w_j)$, which end up growing exponentially. While this gives our strongest results, for instance improving the known rates for hard-margin linear SVM as discussed before, such step sizes are rarely used in practice, and moreover it is unclear if they could carry over to deep learning and other applications of these ideas, where the step sizes often have constant or decreasing η_j. These smaller step sizes also figure heavily in prior work, and so we give special consideration to the regime where η_j is constant, and prove a tight $\ln(n)/\ln(t)$ implicit bias rate.

Turning back to additional notation for this section, for the exponential loss, recall $\psi(\xi) = \ln\left(\sum_{i=1}^n \exp(\xi_i)\right)$, and $\psi^*(\theta) = \sum_{i=1}^n \theta_i \ln\theta_i$ with domain the standard probability simplex $\Delta_n := \{\theta \in \mathbb{R}^n | \theta \geq 0, \sum_{i=1}^n \theta_i = 1\}$. Moreover, ψ is 1-smooth with respect to the ℓ_∞ norm.

Let $\gamma := \max_{\|u\|=1} \min_{1\leq i\leq n} y_i\langle u, x_i\rangle$ and $\bar{u} := \arg\max_{\|u\|=1} \min_{1\leq i\leq n} y_i\langle u, x_i\rangle$ respectively denote the maximum margin value and direction on the dataset. As in Appendix A in the general case of ψ, but as presented in prior work for the specific case of losses with bias towards the maximum margin solution, the maximum margin has a dual characterization (for the exponential loss) of

$$\gamma = \min_{q\in\Delta_n} \left\|Z^\top q\right\| = \sqrt{2\min_{q\in\Delta_n} f(q)} = \sup_{\substack{\|w\|\leq 1\\ r>0}} -r\psi(Zw/r). \tag{9}$$

4.1. $O(1/t)$ margin maximization rates

For the exponential loss,

$$\psi(Zw) = \ln\left(\sum_{i=1}^n \exp\left(\langle z_i, w\rangle\right)\right) \geq \ln\left(\exp\left(\max_{1\leq i\leq n}\langle z_i, w\rangle\right)\right) = \max_{1\leq i\leq n}\langle z_i, w\rangle,$$

and thus $\min_{1\leq i\leq n}\langle -z_i, w\rangle = \min_{1\leq i\leq n} y_i\langle w, x_i\rangle \geq -\psi(Zw)$. The next result then follows immediately from eq. (7), and gives $O(1/t)$ margin maximization rates with constant $\hat{\eta}_j$.

Theorem 7 *Under Assumption 2, for the exponential loss, if $\hat{\eta}_t = \eta_t\mathcal{R}(w_t) \leq 1$ is nonincreasing and $w_0 = 0$, then*

$$\frac{\min_{1\leq i\leq n} y_i\langle w_t, x_i\rangle}{\|w_t\|} \geq \frac{-\psi(Zw_t)}{\|w_t\|} \geq \gamma - \frac{\ln(n)+1}{\gamma\sum_{j<t}\hat{\eta}_j}.$$

For the logistic loss, letting $t_0 = (256\ln n)^2/\gamma^2$, and $\eta_t\mathcal{R}(w_t) = 1/2$ for $t < t_0$, and $\hat{\eta}_t = \eta_t\ell'\left(\psi(Zw_t)\right)/n = 1/2$ for $t \geq t_0$, then for any $t > t_0$,

$$\frac{\min_{1\leq i\leq n} y_i\langle w_t, x_i\rangle}{\|w_t\|} \geq \frac{-\psi(Zw_t)}{\|w_t\|} \geq \gamma - \frac{1+512\ln n}{\gamma t - (256\ln n)^2/\gamma}.$$

The full proof of Theorem 7 is given in Appendix D. Here we sketch the proof for the exponential loss. Note that eqs. (7) and (9) imply

$$\frac{-\psi(Zw_t)}{\|w_t\|} \geq \frac{-\psi(p_0) + \sum_{j<t} \hat{\eta}_j \left\|Z^\top q_j\right\| \cdot \gamma - \frac{\hat{\eta}_0}{2}\left\|Z^\top q_0\right\|^2}{\|w_t\|}$$
$$= \gamma \cdot \frac{\sum_{j<t} \hat{\eta}_j \left\|Z^\top q_j\right\|}{\|w_t\|} - \frac{\psi(p_0) + \frac{\hat{\eta}_0}{2}\left\|Z^\top q_0\right\|^2}{\|w_t\|}.$$

By the triangle inequality, $\|w_t\| \leq \sum_{j<t} \hat{\eta}_j \left\|Z^\top q_j\right\|$. Moreover, $\psi(p_0) = \ln(n)$, and $\left\|Z^\top q_0\right\| \leq 1$ since $\|z_i\| \leq 1$. Therefore we have

$$\frac{-\psi(Zw_t)}{\|w_t\|} \geq \gamma - \frac{\ln(n)+1}{\|w_t\|}.$$

Lastly, note that $\|w_t\| \geq \langle w_t, \bar{u}\rangle$, and moreover

$$\langle w_{j+1} - w_j, \bar{u}\rangle = \hat{\eta}_j \left\langle -Z^\top q_j, \bar{u}\right\rangle = \hat{\eta}_j \langle -Z\bar{u}, q_j\rangle \geq \hat{\eta}_j \gamma.$$

Margin maximization has been analyzed in many settings: Telgarsky (2013) proves that for any $\epsilon > 0$, the margin can be maximized by coordinate descent to $\gamma - \epsilon$ with an $O(1/t)$ rate, while Nacson et al. (2018) show an $\widetilde{O}(1/\sqrt{t})$ margin maximization rate for gradient descent by letting $\hat{\eta}_t \propto 1/\sqrt{t+1}$ using our notation. Their proofs also analyze $-\psi(Zw_t)/\|w_t\|$, but use Lemma 3. If we let $\hat{\eta}_t$ be a constant in Lemma 3, the error term $\sum_{j<t} \frac{\beta\hat{\eta}_j^2}{2}\left\|Z^\top q_j\right\|^2$ will be too large to prove exact margin maximization, while if we let $\hat{\eta}_t = 1/\sqrt{t+1}$, then the error term is $O\left(\ln(t)\right)$, but only an $O\left(\ln(t)/\sqrt{t}\right)$ rate can be obtained. By contrast, our analysis uses the tighter guarantee given by Theorem 1, which always has a bounded error term.

4.2. Tight $\ln(n)/t$ and $\ln(n)/\ln(t)$ bias rates

Next we turn to a fast implicit bias rate, where we produce both upper and lower bounds. In the case of aggressive step sizes, there does not appear to be prior work, though a $O(1/t^{1/4})$ rate can be easily derived via the Fenchel-Young inequality from the bias rate in prior work (Nacson et al., 2018). Instead, prior work appears to use constant η_j, and the rate is roughly $\mathcal{O}(1/\ln(t))$, however the dependence on n in unspecified, and unclear from the proofs. Here we provide a careful analysis with a rate $\mathcal{O}(\ln n/\ln t)$ for constant η_j, and rate $\mathcal{O}(\ln(n)/t)$ for constant $\hat{\eta}_j$, which we moreover show are tight. In this subsection we only analyze the exponential loss.

Theorem 8 *Consider the exponential loss and nonincreasing steps $\hat{\eta}_t = \eta_t \mathcal{R}(w_t) \leq 1$ with $\sum_j \eta_j = \infty$. For any data $(z_i)_{i=1}^n$ sampled from a density which is continuous w.r.t. the Lebesgue measure and which satisfies Assumption 2, then almost surely, for every iteration t,*

$$\left\|\frac{w_t}{\|w_t\|} - \bar{u}\right\| = \frac{O(\ln n)}{\sum_{j<t}\hat{\eta}_j} = \begin{cases} O(\frac{\ln n}{t}) & \text{when } \hat{\eta}_j = 1, \\ O(\frac{\ln n}{\ln t}) & \text{when } \eta_j = 1. \end{cases}$$

On the other hand, there exists data $Z \in \mathbb{R}^{n\times 2}$ comprised of n examples in $\mathbb{R}^2$ satisfying Assumption 2, so that for all sufficiently large iterations $t \geq 1$,

$$\left\|\frac{w_t}{\|w_t\|} - \bar{u}\right\| \geq \frac{\ln n - \ln 2}{\|w_t\|} = \begin{cases} \frac{\ln n - \ln 2}{t} & \text{when } \hat{\eta}_j = 1, \\ \frac{\ln n - \ln 2}{\Theta(\ln(t))} & \text{when } \eta_j = 1. \end{cases}$$

Before sketching the proof, it's worth mentioning the use of Lebesgue measure in the upper bound. This assumption ensures the support vectors have reasonable structure and simplifies the behavior orthogonal to the maximum margin predictor $\bar{u}$; this assumption originated in prior work on implicit bias (Soudry et al., 2017).

To state the upper and lower bounds more explicitly and to sketch the proof of Theorem 8 (full details are in the appendices), we first introduce some additional notation. Recall that γ denotes the maximum margin, and we let $\bar{u} := \arg\max_{\|u\|=1} \min_{1\le i\le n} y_i\langle u, x_i\rangle$ denote the maximum margin direction. Given any vector $a \in \mathbb{R}^d$, let $\Pi_\perp[a] := a - \langle a, \bar{u}\rangle\bar{u}$ denote its component orthogonal to $\bar{u}$. Given a gradient descent iterate w_t, let $v_t := \Pi_\perp[w_t]$. Given a data point z_i, let $z_{i,\perp} := \Pi_\perp[z_i]$.

Let $S := \{z_i : \langle \bar{u}, -z_i\rangle = \gamma\}$ denote the set of support vectors, and let

$$\mathcal{R}_\gamma(w) := \frac{1}{n}\sum_{z_i\in S} \exp\left(\langle w, z_i\rangle\right)$$

denote the risk induced by support vectors, and

$$\mathcal{R}_{>\gamma}(w) := \frac{1}{n}\sum_{z_i\notin S} \exp\left(\langle w, z_i\rangle\right)$$

denote the risk induced by non-support vectors. In addition, let $S_\perp := \{z_{i,\perp} : z_i \in S\}$, and

$$\mathcal{R}_\perp(w) := \frac{1}{n}\sum_{z_i\in S} \exp\left(\langle w, z_{i,\perp}\rangle\right) = \frac{1}{n}\sum_{z\in S_\perp} \exp\left(\langle w, z\rangle\right)$$

denote the risk induced by components of support vectors orthogonal to $\bar{u}$. By definition, $\mathcal{R}_\perp(w) = \mathcal{R}_\gamma(w)\exp\left(\gamma\langle w, \bar{u}\rangle\right)$. Lastly, let $\gamma' := \min_{z_i\notin S}\langle \bar{u}, -z_i\rangle - \gamma$ denote the margin between support vectors and non-support vectors. If there is no non-support vector, let $\gamma' = \infty$.

Below is our main result.

Theorem 9 *If the data examples are sampled from a density w.r.t. the Lebesgue measure, then almost surely $\mathcal{R}_\perp$ has a unique minimizer $\bar{v}$ over $\mathrm{span}(S_\perp)$. If all $\hat{\eta}_j = \eta\mathcal{R}(w_j) \le 1$ are nonincreasing, then*

$$\|v_t - \bar{v}\| \le \max\left\{\|v_0 - \bar{v}\|, 2\right\} + \frac{2\ln(n)}{\gamma\gamma'} + 2.$$

The key potential used in the proof of Theorem 9 is $\|v_t - \bar{v}\|^2$. The change in this potential comes from three parts: (i) a part due to support vectors, which does not increase this potential; (ii) a part due to non-support vectors, which is controlled by the dual convergence result Theorem 1; (iii) a squared gradient term, which is again controlled via the dual convergence result Theorem 1. The full proof is given in Appendix D.

Theorem 9 implies that $\|v_t\| \le \|\bar{v}\| + \|v_0 - \bar{v}\| + O\left(\ln(n)\right)$. On the other hand, it is proved in the prior work (Soudry et al., 2017; Ji and Telgarsky, 2018b) that $\|w_t\| = \Theta\left(\ln(t)\right)$. Therefore

$$\left\|\frac{w_t}{\|w_t\|} - \bar{u}\right\| = \frac{\|w_t - \|w_t\|\bar{u}\|}{\|w_t\|} \le \frac{2\|v_t\|}{\|w_t\|} \le O\left(\frac{\ln(n)}{\ln(t)}\right).$$

Below we further show that this bound is tight: $\|v_t - \bar{v}\|$ could be $\Omega\left(\ln(n)\right)$ for certain datasets.

Theorem 10 *Consider the dataset in $\mathbb{R}^2$ where $z_1 = (0.1, 0)$ and $z_2, \ldots, z_n$ are all $(0.2, 0.2)$. Then $\gamma = 0.1$, and $\bar{v} = (0, 0)$, and starting from $w_0 = (0, 0)$, for large enough t, we have*

$$\|v_t - \bar{v}\| = \|v_t\| \geq \ln(n) - \ln(2).$$

The proof of Theorem 10 is also given in Appendix D.

5. Examples of losses satisfying Assumption 1

It remains a question what loss functions satisfy Assumption 1. Here are some examples:

- The exponential loss $\ell_{\exp}(z) := e^z$.
- The logistic loss $\ell_{\log}(z) := \ln(1 + e^z)$.
- The polynomial loss $\ell_{\mathrm{poly},k}(z)$: on $(-\infty, 0]$, it is defined as

$$\ell_{\mathrm{poly},k}(z) := \frac{1}{(1-z)^k}, \text{ and thus } \ell'_{\mathrm{poly},k}(z) = \frac{k}{(1-z)^{k+1}}, \text{ and } \ell''_{\mathrm{poly},k}(z) = \frac{k(k+1)}{(1-z)^{k+2}},$$

  for some $k > 0$. On $(0, \infty)$, we let $\ell'_{\mathrm{poly},k}(z) := 2k - k(1+z)^{-k-1}$, and therefore

$$\ell_{\mathrm{poly},k}(z) = 2kz + \frac{1}{(1+z)^k}, \text{ and } \ell''_{\mathrm{poly},k}(z) = \frac{k(k+1)}{(1+z)^{k+2}}.$$

Theorem 11 *Assumption 1 is satisfied by $\ell_{\exp}$, $\ell_{\log}$, and $\ell_{\mathrm{poly},k}$ for all $k > 0$.*

The tricky thing to verify is the convexity and smoothness of ψ. The following result can help us establish the convexity of ψ; it is basically (Hardy et al., 1934, Theorem 3.106), and a proof is included in Appendix E for completeness.

Lemma 12 *If $\ell'^2/(\ell\ell'')$ is increasing on $(-\infty, \infty)$, then ψ is convex.*

On the smoothness of ψ, we have the following global estimate.

Lemma 13 *For $\ell_{\exp}$, the smoothness constant $\beta = 1$. In general, if $\ell'' \leq c\ell'$ for some constant $c > 0$, then $\beta \leq cn$.*

Note that for $\ell_{\log}$ and $\ell_{\mathrm{poly},k}$, the above upper bound on the smoothness constant is cn, which looks bad. However, in these cases ℓ' is bounded above by some universal constant c'; therefore to satisfy the condition $\hat{\eta}_t = \eta_t \ell'\left(\psi(Zw_t)\right)/n \leq 1/\beta$ in Theorems 1 and 5, it is enough if $\eta_t \leq 1/(cc')$. In other words, we can still handle a constant step size for gradient descent on the empirical risk function $\mathcal{R}$. A finer approach is to use the smoothness constant on sublevel sets; we demonstrate this in the next result for the logistic loss.

Lemma 14 *For the logistic loss, on the sublevel set $\{\xi | \psi(\xi) \leq 0\} = \{\xi | \mathcal{L}(\xi) \leq \ell(0)\}$, it holds that $1 \leq \|\nabla\psi(\xi)\|_1 \leq 2$, and ψ is 2-smooth with respect to the ℓ_∞ norm. Moreover, if $\mathcal{L}(Zw_t) \leq \ell(0)/(2e^2)$, and $\hat{\eta}_t = \eta_t \ell'\left(\psi(Zw_t)\right)/n \leq 1/2$, then*

$$\psi(Zw_{t+1}) - \psi(Zw_t) \leq (-\hat{\eta}_t + \hat{\eta}_t^2)\left\|Z^\top q_t\right\|^2, \quad \text{and} \quad D_{\psi^*}(q_{t+1}, q_t) \geq \frac{1}{4}\|q_{t+1} - q_t\|_1^2.$$

To prove Theorem 7, we can proceed as with the exponential loss over those iterations $[t_0, t]$ where Lemma 14 is in effect, achieving the same $O(\ln(n)/t)$ rate over those iterations. To control the magnitude of t_0 and $\|w_{t_0}\|$, we can use a delicate but more standard analysis, giving $t_0 = O((\ln n)^2/\gamma^2)$ and $\|w_{t_0}\| = O((\ln n)/\gamma)$ (cf. Theorem 21).

The full proofs of results in this section are given in Appendix E.

6. Open problems

One open problem is to extend our results to nonlinear models, such as deep linear or homogeneous networks. For example, Chizat and Bach (2020) prove that gradient descent can maximize the margin on a 2-homogeneous network, assuming (a different kind of) dual convergence. It is very interesting to see if our analysis can be applied to this setting.

Another open problem is to see whether our analysis can be extended to other training algorithms, such stochastic gradient descent and accelerated gradient descent.

Acknowledgements

The authors thank Maxim Raginsky for pointing them to the concept of generalized sums (Hardy et al., 1934), and to Daniel Hsu and Nati Srebro for discussion of lower bounds and the best known rates for the general hard-margin linear SVM problem. The authors are grateful for support from the NSF under grant IIS-1750051, and from NVIDIA under a GPU grant.

References

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

Jonathan Borwein and Adrian Lewis. *Convex Analysis and Nonlinear Optimization*. Springer Publishing Company, Incorporated, 2000.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *arXiv preprint arXiv:2002.04486*, 2020.

Kenneth L Clarkson, Elad Hazan, and David P Woodruff. Sublinear optimization for machine learning. *Journal of the ACM (JACM)*, 59(5):1–49, 2012.

Michael Collins, Robert E. Schapire, and Yoram Singer. Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, 48(1-3):253–285, 2002.

Robert M Freund, Paul Grigas, and Rahul Mazumder. Adaboost and forward stagewise regression are first-order convex optimization methods. *arXiv preprint arXiv:1307.1192*, 2013.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

Geoffrey J Gordon. Approximate solutions to markov decision processes. Technical report, Carnegie Mellon University, School of Computer Science, 1999.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018a.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pages 9461–9471, 2018b.

G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge university press, 1934.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Springer Publishing Company, Incorporated, 2001.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018a.

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300v2*, 2018b.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *arXiv preprint arXiv:2006.06657*, 2020.

Ziwei Ji, Miroslav Dudík, Robert E Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pages 2109–2136, 2020.

Jyrki Kivinen and Manfred K. Warmuth. Boosting as entropy projection. In *COLT*, pages 134–144, 1999.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. *arXiv preprint arXiv:1803.01905*, 2018.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

R Tyrrell Rockafellar. *Convex analysis*. Number 28. Princeton university press, 1970.

Cynthia Rudin, Ingrid Daubechies, and Robert E Schapire. The dynamics of adaboost: Cyclic behavior and convergence of margins. *Journal of Machine Learning Research*, 5(Dec):1557–1595, 2004.

Robert E. Schapire. The convergence rate of AdaBoost. In *COLT*, 2010.

Robert E. Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. MIT Press, 2012.

Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *ICML*, pages 322–330, 1997.

Shai Shalev-Shwartz and Yoram Singer. Online learning: Theory, algorithms, and applications. 2007.

Shai Shalev-Shwartz and Yoram Singer. On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms. In *COLT*, pages 311–322, 2008.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *arXiv preprint arXiv:1710.10345*, 2017.

Matus Telgarsky. Margins, shrinkage, and boosting. In *ICML*, 2013.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. *arXiv preprint arXiv:2002.09277*, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Appendix A. Dual objective to the smoothed margin

This appendix justifies calling $\frac{1}{2}\|Z^{\mathsf{T}}q\|^2$ the dual potential via convex duality, which also gives another appearance of the constraint $\psi^*(q) \leq 0$.

To start, it seems that ideally we would build a duality around $\psi(Zw)/\|w\|$, however this is not convex. Instead, consider the *persective function* $\hat{\psi}$ of ψ (cf. (Hiriart-Urruty and Lemaréchal, 2001, Section B.2.2)), a standard notion in convex analysis:

$$\hat{\psi}(v,r) := \begin{cases} r\psi(v/r) & r > 0, \\ \lim_{r\downarrow 0} r\psi(v/r) & r = 0, \\ \infty & r < 0. \end{cases}$$

The nonnegative scalar r takes on the role of $1/\|w\|$.

A standard fact from convex analysis is that the perspective of a convex function is also convex (in joint parameters $(v,r) \in \mathbb{R}^{n+1}$). We will also need the conjugate of $\hat{\psi}$, and the fact that $\hat{\psi} \to \max$ for both $\ell_{\exp}$ and $\ell_{\log}$.

Lemma 15 *If ψ is closed and convex, then $\hat{\psi}$ is convex (as a function over $\mathbb{R}^{n+1}$), and has conjugate*

$$\hat{\psi}^*((q,b)) = \begin{cases} \infty & b > -\psi^*(q), \\ 0 & b \leq -\psi^*(q). \end{cases}$$

Furthermore, for $\ell \in \{\ell_{\log}, \ell_{\exp}\}$ and $v \in \mathbb{R}^n$ with $v < 0$ (coordinate-wise), then $r \mapsto \hat{\psi}(v,r)$ is nondecreasing, and $\lim_{r\downarrow 0} \hat{\psi}(v,r) = \max_i v_i$.

Proof As mentioned above, the perspective function of a closed convex function is also closed and convex (Hiriart-Urruty and Lemaréchal, 2001, Section B.2.2). For the conjugate, since $\hat{\psi}$ is convex

and closed,

$$\hat{\psi}^*((q,b)) = \sup_{v,r} \langle v, q\rangle + br - \hat{\psi}(v,r) = \sup_{r>0} r\left(b + \sup_v \left[\langle v/r, q\rangle - \psi(v/r)\right]\right) = \sup_{r>0} r\left(b + \psi^*(q)\right)$$
$$= \begin{cases} \infty & b > -\psi^*(q), \\ 0 & b \leq -\psi^*(q). \end{cases}$$

For the second part, let $v < 0$ and $\ell \in \{\ell_{\exp}, \ell_{\log}\}$ be given. By (Ji and Telgarsky, 2020, Lemma C.5, after correcting the signs on the losses), $\langle v, \nabla\psi(v)\rangle \leq \psi(v)$, meaning in particular $\langle v/r, \nabla\psi(v/r)\rangle \leq \psi(v/r)$ for any $r > 0$, and

$$\frac{\mathrm{d}}{\mathrm{d}r}\hat{\psi}(v,r) = \psi(v/r) + r\left\langle \psi(v/r), -v/r^2\right\rangle = \psi(v/r) - \langle \psi(v/r), v/r\rangle \geq 0,$$

meaning $r \mapsto \hat{\psi}(v,r)$ is nondecreasing. It only remains to show that $\lim_{r\downarrow 0}\hat{\psi}(v,r) = \min_i v_i$. For $\ell_{\exp}$, this is a consequence of the standard inequalities

$$\begin{aligned}\hat{\psi}(v,r) &= r\ln\sum_i \exp(v_i/r) \geq r\ln\max_i \exp(v_i/r) \\ &= \max_i v_i \\ &= r\ln\max_i v_i/r \geq r\ln\frac{1}{n}\sum_i v_i/r = \hat{\psi}(v,r) - r\ln n,\end{aligned}$$

and thus $\lim_{r\downarrow 0}\hat{\psi}(v,r) = \max_i v_i$. For $\ell_{\log}$, since $\ell_{\log}^{-1}(z) = \ln(\exp(z)-1)$, defining $M := \max_i v_i$ for convenience, it suffices to note that

$$\begin{aligned}\lim_{r\downarrow 0}\hat{\psi}_{\log}(v,r) &= \lim_{r\downarrow 0} r\ln\left(\exp(\sum_i \ln(1+\exp(v_i/r)) - 1)\right) \\ &= \lim_{r\downarrow 0} r\ln\left(\prod_i (1+\exp(v_i/r)) - 1\right) \\ &= \lim_{r\downarrow 0} r\ln\sum_{\substack{S\subseteq\{1,\ldots,n\}\\ |S|\geq 1}} \exp\left(\sum_{i\in S} v_i/r\right) \\ &= M + \ln\lim_{r\downarrow 0}\left[\sum_{\substack{S\subseteq\{1,\ldots,n\}\\ |S|\geq 1}} \exp\left(\sum_{i\in S}(v_i - M)/r\right)\right]^r = M.\end{aligned}$$

■

With $\hat{\psi}$ and $\hat{\psi}^*$ in hand, we can easily form a relevant pair of primal-dual problems.

Theorem 16 *Suppose ψ is closed convex, and $\hat{\psi}(w,r)$ is bounded below over $\|w\| \leq 1$. Then*

$$\max_{\substack{\|w\|\leq 1\\ r\geq 0}} -\hat{\psi}(Zw, r) = \min_{\substack{q\in\mathbb{R}^n\\ \psi^*(q)\leq 0}} \|Z^{\mathsf{T}}q\|_2.$$

Remark 17 *This form makes the primal and dual explicitly the maximum margin γ for exp-tailed losses. Alternatively, we could use an SVM form of the objective, whereby the dual contains $\|Z^\intercal q\|_2^2$, and is thus closer to f.*

Proof Let $v \in \mathbb{R}^{d+1}$ be a single variable for (w, r), and let ι denote the convex indicator of the set $\{v \in \mathbb{R}^{d+1} : \|v_{1:d}\| \leq 1\}$, whereby

$$\iota^*(s) = \sup_{\|v_{1:d}\| \leq 1} \langle v, s\rangle = \begin{cases} \|s\| & s_{d+1} = 0, \\ \infty & s_{d+1} \neq 0. \end{cases}$$

Moreover, let $M \in \mathbb{R}^{(n+1)\times(d+1)}$ denote the matrix which is obtained by adding a row and a column to Z which are 0 except in the common $(n+1, d+1)$-th entry where they are 1, whereby $M(w, r) = (Zw, r)$. By Fenchel-Rockafellar duality (Rockafellar, 1970, Section 31), since $\hat{\psi}$ is closed convex by Theorem 15,

$$\inf_{\substack{\|w\|\leq 1 \\ r\geq 0}} \hat{\psi}(Zw, r) = \inf_{v\in\mathbb{R}^{d+1}} \hat{\psi}(Mv) + \iota(v) = \max_{s\in\mathbb{R}^{n+1}} -\hat{\psi}^*(-s) - \iota^*(M^\intercal s).$$

By the earlier form of ι^* and the construction of M, we have the constraint $s_{n+1} = (M^\intercal s)_{d+1} = 0$. Writing $q \in \mathbb{R}^n$ for the first n coordinates of S and baking in a 0 for an $(n+1)$-st coordinate, and additionally using the form of $\hat{\psi}^*$ from Theorem 15, we have the simpler form

$$\inf_{\substack{\|w\|\leq 1 \\ r\geq 0}} \hat{\psi}(Zw, r) = \max_{q\in\mathbb{R}^n} -\hat{\psi}^*(-(q, 0)) - \|Z^\intercal q\| = \max\left\{-\|Z^\intercal q\| : q \in \mathbb{R}^n, 0 \geq \psi^*(-q)\right\}.$$

To finish, we replace q with $-q$ in the dual. ■

Appendix B. Omitted proofs from Section 2

Proof (of Lemma 3) Since ψ is β-smooth with respect to the ℓ_∞ norm,

$$\begin{aligned} \psi(p_{t+1}) - \psi(p_t) &\leq \langle \nabla\psi(p_t), p_{t+1} - p_t\rangle + \frac{\beta}{2}\|p_{t+1} - p_t\|_\infty^2 \\ &= \left\langle q_t, -\hat{\eta}_t ZZ^\top q_t \right\rangle + \frac{\beta\hat{\eta}_t^2}{2}\left\|ZZ^\top q_t\right\|_\infty^2 \\ &= -\hat{\eta}_t \left\|Z^\top q_t\right\|^2 + \frac{\beta\hat{\eta}_t^2}{2}\left\|ZZ^\top q_t\right\|_\infty^2. \end{aligned}$$

Moreover, since $\|z_i\| \leq 1$,

$$\left\|ZZ^\top q_t\right\|_\infty = \max_{1\leq i\leq n} \left|\left\langle Z^\top q_t, z_i\right\rangle\right| \leq \max_{1\leq i\leq n} \left\|Z^\top q_t\right\| \|z_i\| \leq \left\|Z^\top q_t\right\|.$$

As a result,

$$\psi(p_{t+1}) - \psi(p_t) \leq -\hat{\eta}_t \left\|Z^\top q_t\right\|^2 + \frac{\beta\hat{\eta}_t^2}{2}\left\|Z^\top q_t\right\|^2.$$

On the second claim, note that since ψ is β-smooth with respect to the ℓ_∞ norm, (Shalev-Shwartz et al., 2011, Lemma 2.19) implies that ψ^* is $(1/\beta)$-strongly convex with respect to the ℓ_1 norm, and in particular $D_{\psi^*}(q_{t+1}, q_t) \geq \|q_{t+1} - q_t\|_1^2/(2\beta)$. ■

Proof (of Lemma 4) Since f is convex, we have

$$\hat{\eta}_t \left(f(q_t) - f(q)\right) \leq \langle \hat{\eta}_t \nabla f(q_t), q_t - q\rangle = \langle \hat{\eta}_t \nabla f(q_t), q_t - q_{t+1}\rangle + \langle \hat{\eta}_t \nabla f(q_t), q_{t+1} - q\rangle .$$

Recall that $p_{t+1} = p_t - \hat{\eta}_t Z Z^\top q_t = p_t - \hat{\eta}_t \nabla f(q_t)$, therefore

$$\begin{aligned}\hat{\eta}_t \left(f(q_t) - f(q)\right) &\leq \langle \hat{\eta}_t \nabla f(q_t), q_t - q_{t+1}\rangle + \langle \hat{\eta}_t \nabla f(q_t), q_{t+1} - q\rangle \\ &= \langle \hat{\eta}_t \nabla f(q_t), q_t - q_{t+1}\rangle + \langle p_t - p_{t+1}, q_{t+1} - q\rangle .\end{aligned}$$

It can be verified by direct expansion that

$$\langle p_t - p_{t+1}, q_{t+1} - q\rangle = D_{\psi^*}(q, q_t) - D_{\psi^*}(q, q_{t+1}) - D_{\psi^*}(q_{t+1}, q_t),$$

and thus

$$\hat{\eta}_t \left(f(q_t) - f(q)\right) \leq \langle \hat{\eta}_t \nabla f(q_t), q_t - q_{t+1}\rangle + D_{\psi^*}(q, q_t) - D_{\psi^*}(q, q_{t+1}) - D_{\psi^*}(q_{t+1}, q_t).$$

On the other claim, let ∂ denote subdifferential. We have

$$\partial h(q) = \{\nabla f(q_t)\} + \frac{1}{\hat{\eta}_t}\left(\partial \psi^*(q) - \{p_t\}\right).$$

Note that $q' \in \arg\min h(q)$ if and only if $0 \in \partial h(q')$, which is equivalent to

$$p_t - \hat{\eta}_t \nabla f(q_t) = p_{t+1} \in \partial \psi^*(q).$$

By (Rockafellar, 1970, Theorem 23.5), $p_{t+1} \in \partial \psi^*(q)$ if and only if $q = \nabla \psi(p_{t+1})$; in other words, q_{t+1} is the unique minimizer of h, and specifically $h(q_{t+1}) \leq h(q_t) = f(q_t)$. ■

Appendix C. Omitted proofs from Section 3

We first prove Lemma 6.

Proof (of Lemma 6) Define $\sigma(s) := \ell'\left(\ell^{-1}(s)\right)\ell^{-1}(s)$. Note that Assumption 1 implies

$$\lim_{s\to 0}\sigma(s) = 0 \quad \text{and} \quad \sigma(s)/s \text{ is increasing on } (0, \ell(0)).$$

It then follows that σ is super-additive on $(0, \ell(0))$, meaning for any $a, b > 0$ such that $a+b < \ell(0)$, it holds that $\sigma(a+b) \geq \sigma(a)+\sigma(b)$. In particular, if $\psi(\xi) \leq 0$, or equivalently if $\sum_{i=1}^n \ell(\xi_i) \leq \ell(0)$, then

$$\sum_{i=1}^n \sigma\left(\ell(\xi_i)\right) - \sigma\left(\sum_{i=1}^n \ell(\xi_i)\right) \leq 0.$$

Now note that

$$\psi^*\left(\nabla\psi(\xi)\right) = \langle \nabla\psi(\xi), \xi\rangle - \psi(\xi) = \sum_{i=1}^n \frac{\ell'(\xi_i)\xi_i}{\ell'\left(\psi(\xi)\right)} - \psi(\xi),$$

and thus

$$\ell'\left(\psi(\xi)\right)\psi^*\left(\nabla\psi(\xi)\right) = \sum_{i=1}^n \ell'(\xi_i)\xi_i - \ell'\left(\psi(\xi)\right)\psi(\xi) = \sum_{i=1}^n \sigma\left(\ell(\xi_i)\right) - \sigma\left(\sum_{i=1}^n \ell(\xi_i)\right) \le 0.$$

Since $\ell' > 0$, it follows that $\psi^*\left(\nabla\psi(\xi)\right) \le 0$. ■

Next we prove Part 1 of Theorem 5

Proof (of Theorem 5 Part 1) Recall that $S_0 := \{q | \psi^*(q) \le 0\}$. It is nonempty since for $\xi \in \mathbb{R}^n$, if $\ell(\xi_i) \le \ell(0)/n$ for all i, then $\mathcal{L}(\xi) \le \ell(0)$ and $\psi(\xi) \le 0$, and Lemma 6 implies $\psi^*\left(\nabla\psi(\xi)\right) \le 0$. Moreover, S_0 is closed convex since it is a sublevel set of the closed convex function ψ^*.

To prove the boundedness of S_0, note that since (Rockafellar, 1970, Theorem 23.5) ensures $\operatorname{range}\nabla\psi = \operatorname{dom}\partial\psi^*$, while (Rockafellar, 1970, Theorem 23.4) ensures $\operatorname{dom}\partial\psi^*$ contains the relative interior of $\operatorname{dom}\psi^*$, it is enough to consider $q = \nabla\psi(\xi)$ for some $\xi \in \mathbb{R}^n$. Recall that

$$\nabla\psi(\xi)_i = \frac{\ell'(\xi_i)}{\ell'\left(\psi(\xi)\right)} = \frac{\ell'\left(\ell^{-1}\left(\ell(\xi_i)\right)\right)}{\ell'\left(\ell^{-1}\left(\sum_{i=1}^n \ell(\xi_i)\right)\right)}. \tag{10}$$

Since $\ell > 0$, we have $\ell(\xi_i) < \sum_{i=1}^n \ell(\xi_i)$, and since ℓ' and ℓ^{-1} are increasing, it follows from eq. (10) that $0 < \nabla\psi(\xi)_i \le 1$, and $\|\nabla\psi(\xi)\|_1 \le n$. Consequently, S_0 is bounded.

Next we prove $\left\|Z^\top\nabla\psi(\xi)\right\| \ge c$ for all $\xi \in \mathbb{R}^n$ and some positive constant c. Let $s := \max_{1\le i\le n}\ell(\xi_i)$, then $\sum_{i=1}^n \ell(\xi_i) \le ns$. Since $\ell' > 0$, and ℓ' and ℓ^{-1} are increasing,

$$\|\nabla\psi(\xi)\|_1 \ge \frac{\ell'\left(\ell^{-1}\left(s\right)\right)}{\ell'\left(\ell^{-1}\left(\sum_{i=1}^n \ell(\xi_i)\right)\right)} \ge \frac{\ell'\left(\ell^{-1}(s)\right)}{\ell'\left(\ell^{-1}(ns)\right)}.$$

By Assumption 1, there exists a constant $c > 0$ such that

$$\|\nabla\psi(\xi)\|_1 \ge \frac{\ell'\left(\ell^{-1}(s)\right)}{\ell'\left(\ell^{-1}(ns)\right)} \ge c.$$

On the other hand, Assumption 2 ensures that there exists $u \in \mathbb{R}^d$ and $\gamma > 0$ such that $\langle u, -z_i\rangle \ge \gamma$ for all $1 \le i \le n$. Therefore

$$\left\|Z^\top\nabla\psi(\xi)\right\| \ge \left\langle -Z^\top\nabla\psi(\xi), u\right\rangle = \sum_{i=1}^n \langle u, -z_i\rangle\nabla\psi(\xi)_i \ge \gamma\|\nabla\psi(\xi)\|_1 \ge \gamma c.$$

Next we prove $Z^\top\bar{q}$ is unique. Suppose $\bar{q}_1$ and $\bar{q}_2$ both minimize f over S_0, with $\left\|Z^\top\bar{q}_1\right\| = \left\|Z^\top\bar{q}_2\right\| > 0$, but $Z^\top\bar{q}_1 \ne Z^\top\bar{q}_2$. It follows that $Z^\top\bar{q}_1$ and $Z^\top\bar{q}_2$ point to different directions, and

$$\left\langle Z^\top\bar{q}_1, Z^\top\bar{q}_2\right\rangle < \left\|Z^\top\bar{q}_1\right\|^2.$$

Since ψ^* is convex, S_0 is convex, and therefore $(\bar{q}_1 + \bar{q}_2)/2 \in S_0$. However, we then have

$$\left\|\frac{Z^\top\bar{q}_1 + Z^\top\bar{q}_2}{2}\right\|^2 = \frac{1}{2}\left\|Z^\top\bar{q}_1\right\|^2 + \frac{1}{2}\left\langle Z^\top\bar{q}_1, Z^\top\bar{q}_2\right\rangle < \left\|Z^\top\bar{q}_1\right\|^2,$$

a contradiction. ■

Appendix D. Omitted proofs from Section 4

Before proving Theorem 7, we first prove a few lemmas that will be needed. First, we need to upper and lower bound $\|w_t\|$.

Lemma 18 *Let $w_0 = 0$, then for the exponential loss,*

$$\gamma \sum_{j<t} \hat{\eta}_j \leq \|w_t\| \leq \sum_{j<t} \hat{\eta}_j.$$

For the logistic loss, we have $\|q_j\|_1 \geq 1$ and $\|w_t\| \geq \gamma \sum_{j<t} \hat{\eta}_j$.

Proof For the exponential loss, since $\|z_i\| \leq 1$ and $q_j \in \Delta_n$, the triangle inequality implies $\left\|Z^\top q_j\right\| \leq 1$, and moreover

$$\|w_t\| \leq \sum_{j<t} \hat{\eta}_j \left\|Z^\top q_j\right\| \leq \sum_{j<t} \hat{\eta}_j.$$

On the other hand, by the definition of the maximum margin γ and the unit maximum margin solution $\bar{u}$, we have $\langle -z_i, \bar{u}\rangle = y_i \langle x_i, \bar{u}\rangle \geq \gamma$ for all i. Moreover, for the exponential loss, $q_j \in \Delta_n$. Therefore

$$\langle w_{j+1} - w_j, \bar{u}\rangle = \hat{\eta}_j \left\langle -Z^\top q_j, \bar{u}\right\rangle = \hat{\eta}_j \langle -Z\bar{u}, q_j\rangle \geq \hat{\eta}_j \gamma.$$

Since $w_0 = 0$, the Cauchy-Schwarz inequality implies

$$\|w_t\| \geq \langle w_t, \bar{u}\rangle = \sum_{j<t} \langle w_{j+1} - w_j, \bar{u}\rangle \geq \gamma \sum_{j<t} \hat{\eta}_j.$$

For the logistic loss, the lower bound proof also works, since $\|q_j\|_1 \geq 1$. To see this, note that given $\xi \in \mathbb{R}^n$, we have

$$\|\nabla \psi(\xi)\|_1 = \sum_{i=1}^n \frac{\ell'(\xi_i)}{\ell'\left(\psi(\xi)\right)} = \sum_{i=1}^n \frac{\ell'\left(\ell^{-1}\left(\ell(\xi_i)\right)\right)}{\ell'\left(\ell^{-1}\left(\sum_{i=1}^n \ell(\xi_i)\right)\right)}.$$

Consider the function $\rho(z) := \ell'\left(\ell^{-1}(z)\right) = 1 - e^{-z}$. It holds that $\rho(0) = 0$, and on $[0, \infty)$, we have ρ is subadditive: for all $a, b > 0$, it holds that $\rho(a+b) \leq \rho(a) + \rho(b)$. Therefore

$$\|\nabla \psi(\xi)\|_1 = \sum_{i=1}^n \frac{\ell'\left(\ell^{-1}\left(\ell(\xi_i)\right)\right)}{\ell'\left(\ell^{-1}\left(\sum_{i=1}^n \ell(\xi_i)\right)\right)} = \frac{\sum_{i=1}^n \rho\left(\ell(\xi_i)\right)}{\rho\left(\sum_{i=1}^n \ell(\xi_i)\right)} \geq 1.$$

■

With these tools in hand, we turn to the margin rates.

Proof (of Theorem 7) We first consider the exponential loss. Note that eqs. (7) and (9) imply

$$\begin{aligned}
\frac{-\psi(Zw_t)}{\|w_t\|} &\geq \frac{-\psi(p_0) + \sum_{j<t} \hat{\eta}_j \left\|Z^\top q_j\right\| \cdot \gamma - \frac{\hat{\eta}_0}{2}\left\|Z^\top q_0\right\|^2}{\|w_t\|} \\
&= \gamma \cdot \frac{\sum_{j<t} \hat{\eta}_j \left\|Z^\top q_j\right\|}{\|w_t\|} - \frac{\psi(p_0) + \frac{\hat{\eta}_0}{2}\left\|Z^\top q_0\right\|^2}{\|w_t\|}.
\end{aligned}$$

By the triangle inequality, $\|w_t\| \le \sum_{j<t} \hat{\eta}_j \left\|Z^\top q_j\right\|$. Moreover, $\psi(p_0) = \ln(n)$, and $\hat{\eta}_0 \le 1$, and $\left\|Z^\top q_0\right\| \le 1$ since $\|z_i\| \le 1$. Therefore

$$\frac{-\psi(Zw_t)}{\|w_t\|} \ge \gamma \cdot \frac{\sum_{j<t} \hat{\eta}_j \left\|Z^\top q_j\right\|}{\sum_{j<t} \hat{\eta}_j \left\|Z^\top q_j\right\|} - \frac{\ln(n) + \frac{1}{2}}{\|w_t\|} \ge \gamma - \frac{\ln(n)+1}{\|w_t\|}.$$

Lemma 18 then implies the bound.

Now consider the logistic loss. The analysis is divided into two phases: let t_0 denote the first iteration where the conditions of Lemma 14 hold, after which we may proceed as for the exponential loss. To bound t_0 and $\|w_{t_0}\|$ and in particular to handle the iterations before t_0, we apply Theorem 21, which guarantees $t_0 = O((\ln n)^2/\gamma^2)$ and $\|w_{t_0}\| = O((\ln n)/\gamma)$.

Now we can apply Lemma 14, and start the analysis from w_{t_0} with smoothness constant 2. All the results in Sections 2 and 3 still hold, and in particular eq. (8) ensures for $t > t_0$,

$$\begin{aligned}
\frac{-\psi(Zw_t)}{\|w_t\|} &\ge \frac{-\psi(p_{t_0}) + \sum_{j=t_0}^{t-1} \hat{\eta}_j \left\|Z^\top q_j\right\| \cdot \gamma - \frac{\hat{\eta}_{t_0}}{2}\left\|Z^\top q_{t_0}\right\|^2}{\|w_{t_0}\| + \sum_{j=t_0}^{t-1} \hat{\eta}_j \left\|Z^\top q_j\right\|} \\
&= \gamma - \frac{\psi(p_{t_0}) + \frac{\hat{\eta}_{t_0}}{2}\left\|Z^\top q_{t_0}\right\|^2 + \|w_{t_0}\|\gamma}{\|w_{t_0}\| + \sum_{j=t_0}^{t-1} \hat{\eta}_j \left\|Z^\top q_j\right\|} \\
&\ge \gamma - \frac{\psi(p_{t_0}) + \frac{\hat{\eta}_{t_0}}{2}\left\|Z^\top q_{t_0}\right\|^2 + \|w_{t_0}\|\gamma}{\gamma \sum_{j=t_0}^{t-1} \hat{\eta}_j},
\end{aligned}$$

where we use $\left\|Z^\top q_j\right\| \ge \gamma$, since $\|q_j\|_1 \ge 1$ as given by Lemma 18. By construction, $\psi(p_{t_0}) \le 0$, and Lemma 14 implies $\|q_{t_0}\|_1 \le 2$. Further letting $\hat{\eta}_j = 1/2$, we get

$$\frac{-\psi(Zw_t)}{\|w_t\|} \ge \gamma - \frac{1 + 2\|w_{t_0}\|\gamma}{\gamma(t-t_0)}.$$

Plugging in the earlier bounds on t_0 and $\|w_{t_0}\|$ from Theorem 21 gives the final left hand side. To upper bound the left hand side by the exact margin, Theorem 15 suffices. ■

We then prove the almost-sure existence of $\bar{v}$.

Proof (of first part of Theorem 9) Theorem 2.1 of (Ji and Telgarsky, 2018b) ensures that $S_\perp$ can be decomposed into two subsets B and C, with the following properties:

- The risk induced by B

$$\mathcal{R}_B(w) := \frac{1}{n}\sum_{z\in B} \exp\left(\langle w, z\rangle\right)$$

  is strongly convex over $\operatorname{span}(B)$.

- If C is nonempty, then there exists a vector $\tilde{u}$, such that $\langle z, \tilde{u}\rangle = 0$ for all $z \in B$, and $\langle z, \tilde{u}\rangle \ge \tilde{\gamma} > 0$ for all $z \in C$.

On the other hand, Lemma 12 of (Soudry et al., 2017) proves that, almost surely there are at most d support vectors, and furthermore the i-th support vector z_i has a positive dual variable θ_i, such that $\sum_{z_i \in S} \theta_i z_i = \gamma \bar{u}$. As a result,

$$\sum_{z_i \in S} \theta_i z_{i,\perp} = \sum_{z_{i,\perp} \in S_\perp} \theta_i z_{i,\perp} = 0.$$

Note that

$$0 = \left\langle \sum_{z_{i,\perp} \in S_\perp} \theta_i z_{i,\perp}, \tilde{u} \right\rangle = \sum_{z_{i,\perp} \in C} \theta_i \langle z_{i,\perp}, \tilde{u} \rangle \geq \tilde{\gamma} \sum_{z_{i,\perp} \in C} \theta_i,$$

which implies that C is empty, and thus $\mathcal{R}_\perp$ is strongly convex over $\mathrm{span}(S_\perp)$. The existence and uniqueness of the minimizer $\bar{v}$ follows from strong convexity. ■

To prove the second part of Theorem 9, we need the following iteration guarantee. Note that it holds for the exponential loss and logistic loss, and will later be used to provide a better "warm start" analysis for the logistic loss (cf. Theorem 21)

Lemma 19 ((Ji and Telgarsky, 2018b) Lemma 3.4) *Suppose ℓ is convex, $\ell' \leq \ell$, and $\ell'' \leq \ell$. For any $t \geq 0$, if $\hat{\eta}_t = \eta_t \mathcal{R}(w_t) \leq 1$, then*

$$\mathcal{R}(w_{t+1}) \leq \mathcal{R}(w_t) - \eta_t \left(1 - \frac{\eta_t \mathcal{R}(w_t)}{2}\right) \|\nabla \mathcal{R}(w_t)\|^2 .$$

Note that under the condition of Theorem 9 that $\eta_t \leq \min\{1, 1/\mathcal{R}(w_0)\}$, Theorem 19 implies that $\mathcal{R}(w_t)$ is nonincreasing. Moreover, we have the following bound on $\sum_{j<t} \hat{\eta}_j$ when η_j is a constant.

Lemma 20 *Let $w_0 = 0$ and $\eta_t = \eta \leq 1$ for all t, then*

$$\sum_{j<t} \hat{\eta}_j \geq \ln\left(1 + \frac{\eta\gamma^2}{2} t\right).$$

Proof We first need a risk upper bound. Recall that Theorem 19 ensures that for any $j < t$, if $\hat{\eta}_j = \eta_j \mathcal{R}(w_j) \leq 1$, then

$$\mathcal{R}(w_{j+1}) \leq \mathcal{R}(w_j) - \eta_j \left(1 - \frac{\eta_j \mathcal{R}(w_j)}{2}\right) \|\nabla \mathcal{R}(w_j)\|^2 . \tag{11}$$

As a result, if we let $\eta_j = \eta \leq 1/\mathcal{R}(w_0) = 1$, then $\mathcal{R}(w_j)$ never increases, and the requirement $\hat{\eta}_j = \eta_j \mathcal{R}(w_j) \leq 1$ of eq. (11) always holds.

Dividing both sides of eq. (11) by $\mathcal{R}(w_j)\mathcal{R}(w_{j+1})$ and rearranging terms gives

$$\frac{1}{\mathcal{R}(w_{j+1})} \geq \frac{1}{\mathcal{R}(w_j)} + \eta \left(1 - \frac{\eta \mathcal{R}(w_j)}{2}\right) \frac{\|\nabla \mathcal{R}(w_j)\|^2}{\mathcal{R}(w_j)\mathcal{R}(w_{j+1})}.$$

Notice that

$$\|\nabla \mathcal{R}(w_j)\| \geq |\langle \nabla \mathcal{R}(w_j), \bar{u} \rangle| = \left| \frac{1}{n} \sum_{i=1}^{n} \exp\left(-\langle w_j, z_i \rangle\right) \langle z_i, \bar{u} \rangle \right| \geq \gamma \mathcal{R}(w_j),$$

and thus

$$\frac{1}{\mathcal{R}(w_{j+1})} \geq \frac{1}{\mathcal{R}(w_j)} + \eta\left(1 - \frac{\eta\mathcal{R}(w_j)}{2}\right)\gamma^2\frac{\mathcal{R}(w_j)}{\mathcal{R}(w_{j+1})} \geq \frac{1}{\mathcal{R}(w_j)} + \eta\left(1 - \frac{\eta\mathcal{R}(w_j)}{2}\right)\gamma^2. \quad (12)$$

Since $\eta\mathcal{R}(w_j) \leq 1$, eq. (12) implies

$$\frac{1}{\mathcal{R}(w_{j+1})} \geq \frac{1}{\mathcal{R}(w_j)} + \eta\left(1 - \frac{\eta\mathcal{R}(w_j)}{2}\right)\gamma^2 \geq \frac{1}{\mathcal{R}(w_j)} + \frac{\eta}{2}\gamma^2,$$

and thus

$$\mathcal{R}(w_t) \leq 1/\left(\frac{1}{\mathcal{R}(w_0)} + \frac{\eta\gamma^2}{2}t\right) \leq 1/\left(1 + \frac{\eta\gamma^2}{2}t\right). \quad (13)$$

Now we prove the lower bound. Notice that $\ln\mathcal{R}$ is also convex, since it is the composition of ln-sum-exp and a linear mapping. Therefore the convexity of $\ln\mathcal{R}$ gives

$$\ln\mathcal{R}(w_{j+1}) - \ln\mathcal{R}(w_j) \geq \langle\nabla\ln\mathcal{R}(w_j), w_{j+1} - w_j\rangle = -\hat{\eta}_j\left\|\nabla\ln\mathcal{R}(w_j)\right\|^2 = -\hat{\eta}_j\left\|Z^\top q_j\right\|^2.$$

The triangle inequality ensures $\left\|Z^\top q_j\right\| \leq \sum_{i=1}^n q_{j,i}\|z_i\| \leq 1$, which implies $\ln\mathcal{R}(w_{j+1}) - \ln\mathcal{R}(w_j) \geq -\hat{\eta}_j$, and thus

$$\sum_{j<t}\hat{\eta}_j \geq \ln\mathcal{R}(w_0) - \ln\mathcal{R}(w_t). \quad (14)$$

Combining eqs. (13) and (14) gives

$$\sum_{j<t}\hat{\eta}_j \geq \ln\mathcal{R}(w_0) + \ln\left(1 + \frac{\eta\gamma^2}{2}t\right) = \ln\left(1 + \frac{\eta\gamma^2}{2}t\right).$$

■

Next we prove the refined rate for the implicit bias.

Proof (of second part of Theorem 9) For technical reasons, we consider a range of steps during which $\|v_j - \bar{v}\| \geq 1$. If $\|v_t - \bar{v}\| \leq 1$, then the proof is done. Otherwise let t_{-1} denote the last step before t such that $\left\|v_{t_{-1}} - \bar{v}\right\| \leq 1$; if such a step does not exist, let $t_{-1} = -1$. Furthermore, let $t_0 = t_{-1} + 1$. Since it always holds that

$$\begin{aligned}
\|\eta_j\nabla\mathcal{R}(w_j)\| &= \eta_j\left\|\frac{1}{n}\sum_{i=1}^n \exp\left(\langle w_j, z_i\rangle\right) z_i\right\| \\
&= \eta_j\mathcal{R}(w_j)\left\|\sum_{i=1}^n \frac{\exp\left(\langle w_j, z_i\rangle\right)}{\sum_{i'=1}^n \exp\left(\langle w_j, z_{i'}\rangle\right)} z_i\right\| \\
&\leq \eta_j\mathcal{R}(w_j) = \hat{\eta}_j \leq 1,
\end{aligned}$$

we have $\|v_{t_0} - \bar{v}\| \leq \max\{\|v_0 - \bar{v}\|, 2\}$.

Note that

$$\begin{aligned}\|v_{j+1}-\bar{v}\|^2 &= \|v_j-\bar{v}-\eta_j\Pi_\perp\left[\nabla\mathcal{R}(w_j)\right]\|^2 \\ &= \|v_j-\bar{v}\|^2 - 2\eta_j\left\langle\Pi_\perp\nabla\mathcal{R}(w_j), v_j-\bar{v}\right\rangle + \eta_j^2\left\|\Pi_\perp\nabla\mathcal{R}(w_j)\right\|^2 \\ &= \|v_j-\bar{v}\|^2 - 2\eta_j\left\langle\nabla\mathcal{R}(w_j), v_j-\bar{v}\right\rangle + \eta_j^2\left\|\Pi_\perp\nabla\mathcal{R}(w_j)\right\|^2, \end{aligned} \tag{15}$$

where the middle $\Pi_\perp$ could be dropped since $\Pi_\perp(v_j-\bar{v}) = v_j-\bar{v}$ and $\Pi_\perp = \Pi_\perp^\intercal$ can be moved across the inner product. Continuing, this inner product term in eq. (15) can be decomposed into two parts, for support vectors and non-support vectors respectively:

$$\begin{aligned}-\left\langle\nabla\mathcal{R}(w_j), v_j-\bar{v}\right\rangle = &\left\langle\frac{1}{n}\sum_{z_i\in S}\exp\left(\langle w_j,z_i\rangle\right)z_i, \bar{v}-v_j\right\rangle \\ &+\left\langle\frac{1}{n}\sum_{z_i\notin S}\exp\left(\langle w_j,z_i\rangle\right)z_i, \bar{v}-v_j\right\rangle. \end{aligned} \tag{16}$$

The support vector part in eq. (16) is non-positive, due to convexity of $\mathcal{R}_\perp$:

$$\begin{aligned}\left\langle\frac{1}{n}\sum_{z_i\in S}\exp\left(\langle w_j,z_i\rangle\right)z_i, \bar{v}-v_j\right\rangle &= \left\langle\frac{1}{n}\sum_{z_i\in S}\exp\left(\langle w_j,z_i\rangle\right)z_{i,\perp}, \bar{v}-v_j\right\rangle \\ &= \exp\left(-\gamma\langle w_j,\bar{u}\rangle\right)\left\langle\frac{1}{n}\sum_{z_i\in S}\exp\left(\langle v_j,z_{i,\perp}\rangle\right)z_{i,\perp}, \bar{v}-v_j\right\rangle \\ &= \exp\left(-\gamma\langle w_j,\bar{u}\rangle\right)\left\langle\nabla\mathcal{R}_\perp(v_j), \bar{v}-v_j\right\rangle \\ &\le \exp\left(-\gamma\langle w_j,\bar{u}\rangle\right)\left(\mathcal{R}_\perp(\bar{v})-\mathcal{R}_\perp(v_j)\right) \le 0. \end{aligned} \tag{17}$$

The part for non-support vectors in eq. (16) is bounded using the Cauchy-Schwarz inequality:

$$\begin{aligned}\left\langle\frac{1}{n}\sum_{z_i\notin S}\exp\left(\langle w_j,z_i\rangle\right)z_i, \bar{v}-v_j\right\rangle &\le \frac{1}{n}\sum_{z_i\notin S}\exp\left(\langle w_j,z_i\rangle\right)\|z_i\|\|v_j-\bar{v}\| \\ &\le \mathcal{R}_{>\gamma}(w_j)\|v_j-\bar{v}\|. \end{aligned} \tag{18}$$

For $t_0 \le j < t$, combining eqs. (15) to (18), and invoking $\|v_j-\bar{v}\| \ge 1$,

$$\begin{aligned}\|v_{j+1}-\bar{v}\|^2 &\le \|v_j-\bar{v}\|^2 + 2\eta_j\mathcal{R}_{>\gamma}(w_j)\|v_j-\bar{v}\| + \eta_j^2\left\|\Pi_\perp\nabla\mathcal{R}(w_j)\right\|^2 \\ &\le \|v_j-\bar{v}\|^2 + 2\eta_j\mathcal{R}_{>\gamma}(w_j)\|v_j-\bar{v}\| + \eta_j^2\left\|\Pi_\perp\nabla\mathcal{R}(w_j)\right\|^2\|v_j-\bar{v}\| \\ &\le \left(\|v_j-\bar{v}\| + \eta_j\mathcal{R}_{>\gamma}(w_j) + \frac{\eta_j^2}{2}\left\|\Pi_\perp\nabla\mathcal{R}(w_j)\right\|^2\right)^2, \end{aligned}$$

and thus

$$\|v_{j+1}-\bar{v}\| \le \|v_j-\bar{v}\| + \eta_j\mathcal{R}_{>\gamma}(w_j) + \frac{\eta_j^2}{2}\left\|\Pi_\perp\nabla\mathcal{R}(w_j)\right\|^2. \tag{19}$$

The middle term with $\mathcal{R}_{>\gamma}$ is bounded using Theorem 1. First we have

$$\begin{aligned}\frac{1}{2}\left\|Z^\top q_j\right\|^2 \geq \frac{1}{2}\left\langle -Z^\top q_j, \bar{u}\right\rangle^2 &= \frac{1}{2}\langle -Z\bar{u}, q_j\rangle^2 \\ &\geq \frac{1}{2}\left(\gamma + \gamma'\frac{\mathcal{R}_{>\gamma}(w_j)}{\mathcal{R}(w_j)}\right)^2 \\ &\geq \frac{1}{2}\gamma^2 + \gamma\gamma'\frac{\mathcal{R}_{>\gamma}(w_j)}{\mathcal{R}(w_j)}. \end{aligned} \tag{20}$$

As a result, let $\bar{q}$ denote a minimizer of $f(q) = \left\|Z^\top q\right\|^2/2$, then Theorem 1 and eq. (20) ensure

$$\begin{aligned} D_{\mathrm{KL}}(\bar{q}, q_j) - D_{\mathrm{KL}}(\bar{q}, q_{j+1}) &\geq \hat{\eta}_j\left(f(q_{j+1}) - \frac{1}{2}\gamma^2\right) \\ &\geq \eta_j\mathcal{R}(w_{j+1})\gamma\gamma'\frac{\mathcal{R}_{>\gamma}(w_{j+1})}{\mathcal{R}(w_{j+1})} \\ &= \eta_j\gamma\gamma'\mathcal{R}_{>\gamma}(w_{j+1}). \end{aligned}$$

Later we will need to evaluate $\sum_j \eta_j\mathcal{R}_{>\gamma}(w_j)$, which by applying the above and telescoping gives

$$\sum_{j=0}^{\infty}\eta_j\mathcal{R}_{>\gamma}(w_j) = \eta_j\mathcal{R}_{>\gamma}(w_0) + \sum_{j=1}^{\infty}\eta_j\mathcal{R}_{>\gamma}(w_j) \leq 1 + \frac{D_{\mathrm{KL}}(\bar{q}, q_0)}{\gamma\gamma'} \leq 1 + \frac{\ln(n)}{\gamma\gamma'}. \tag{21}$$

The squared gradient term in eq. (19) can also be bounded using Theorem 1. To start, by the definition of $\Pi_\perp$ and since $\bar{u} = -Z^\mathsf{T}\bar{q}/\|Z^\mathsf{T}\bar{q}\|$ and using the first order condition $\|Z^\mathsf{T}\bar{q}\|^2 \leq \langle Z^\mathsf{T}\bar{q}, Z^\mathsf{T}q_t\rangle$ (which appeared earlier as eq. (5)), and since $\hat{\eta}_j \leq 1$ are nonincreasing,

$$\begin{aligned}\eta_j^2\|\Pi_\perp\nabla\mathcal{R}(w_j)\|^2 &= \hat{\eta}_j^2\|\nabla\psi(w_j) - \nabla\psi(w_j)^\mathsf{T}\bar{u}\bar{u}\|^2 \\ &= \hat{\eta}_j^2\left(\|Z^\mathsf{T}q_j\|^2 - \frac{\langle Z^\mathsf{T}q_j, Z^\mathsf{T}\bar{q}\rangle^2}{\|Z^\mathsf{T}\bar{q}\|^2}\right) \\ &\leq \hat{\eta}_j\hat{\eta}_{j-1}\left(\|Z^\mathsf{T}q_j\|^2 - \|Z^\mathsf{T}\bar{q}\|^2\right) \\ &\leq \hat{\eta}_j\left(D_{\psi^*}(\bar{q}, q_{j-1}) - D_{\psi^*}(\bar{q}, q_j)\right) \leq D_{\psi^*}(\bar{q}, q_{j-1}) - D_{\psi^*}(\bar{q}, q_j).\end{aligned}$$

As mentioned before, we will need to sum across iterations, which telescopes and gives

$$\sum_{j=0}^{\infty}\eta_j^2\|\Pi_\perp\nabla\mathcal{R}(w_j)\|^2 \leq \eta_0^2\|\Pi_\perp\nabla\mathcal{R}(w_0)\|^2 + \sum_{j=1}^{\infty}\eta_j^2\|\Pi_\perp\nabla\mathcal{R}(w_j)\|^2 \leq 1 + D_{\psi^*}(\bar{q}, q_0) \leq 1 + \ln(n). \tag{22}$$

Combining these pieces, applying eq. (19) recursively and then controlling the summations with eqs. (21) and (22) and using $\max\{\gamma, \gamma'\} \leq 1$ gives

$$\|v_t - \bar{v}\| \leq \|v_{t_0} - \bar{v}\| + \sum_{j=0}^{\infty}\eta_j\mathcal{R}_{>\gamma}(w_j) + \sum_{j=0}^{\infty}\frac{\eta_j^2}{2}\|\Pi_\perp\mathcal{R}(w_j)\|^2 \leq \|v_{t_0} - \bar{v}\| + \frac{2\ln(n)}{\gamma\gamma'} + 2,$$

which finishes the proof. ∎

Below is the proof of the lower bound on $\|v_t - \bar{v}\|$.

Proof (of Theorem 10) By construction, the only support vector is $z_1 = (0.1, 0)$, and $z_{1,\perp} = (0, 0)$. Therefore $\operatorname{span}(S_\perp) = \operatorname{span}(\{(0,0)\}) = \{(0,0)\}$, $\gamma = 0.1$, and $\bar{v} = (0,0)$. Moreover,

$$\mathcal{R}_\gamma(w) = \frac{1}{n}\exp(0.1w_1), \quad \text{and} \quad \mathcal{R}_{>\gamma}(w) = \frac{n-1}{n}\exp(0.2(w_1+w_2)),$$

and for any $t \geq 0$,

$$\nabla\mathcal{R}(w_t)_1 = 0.1\mathcal{R}_\gamma(w_t) + 0.2\mathcal{R}_{>\gamma}(w_t), \quad \text{and} \quad \nabla\mathcal{R}(w_t)_2 = 0.2\mathcal{R}_{>\gamma}(w_t). \tag{23}$$

Recall that $w_0 = 0$, and thus eq. (23) implies that $w_{t,1}, w_{t,2} \leq 0$, and $\mathcal{R}_\gamma(w_t) \leq 1/n$ for all t. As a result, as long as $\mathcal{R}(w_t) \geq 2/n$, it holds that $\mathcal{R}_{>\gamma}(w_t) \geq \mathcal{R}_\gamma(w_t)$ and $|\nabla\mathcal{R}(w_t)_2| \geq |\nabla\mathcal{R}(w_t)_1|/2$.

Let τ denote the first step when the risk is less than $2/n$:

$$\tau = \min\{t : \mathcal{R}(w_t) < 2/n\}.$$

Since $|\nabla\mathcal{R}(w_t)_2| \geq |\nabla\mathcal{R}(w_t)_1|/2$ for all $t < \tau$, we have

$$|w_{\tau,2}| \geq |w_{\tau,1}|/2.$$

On the other hand, since $\|z_i\| \leq 1/3$, it holds that $\mathcal{R}(w_\tau) \geq \exp(-\|w_\tau\|/3)$, which implies that

$$\|w_\tau\| \geq 3\ln(n/2).$$

As a result,

$$|w_{\tau,2}| \geq \ln(n/2).$$

■

Lastly, we put together the preceding pieces to get the main simplified implicit bias bound.

Proof (of Theorem 8) For the upper bound, let Z be given as stated, whereby Theorem 9 holds, and thus almost surely

$$\|v_t - \bar{v}\| = \mathcal{O}(\ln n), \qquad \text{whereby } \|v_t\| = \|\bar{v}\| + \|v_t - \bar{v}\| = \mathcal{O}(\ln n).$$

Next,

$$\|w_t - \bar{u}\|w_t\|\|^2 = \|[\|\,1]\Pi_\perp\left(w_t - \bar{u}\|w_t\|\right)^2 + \|[\|\,1](w_t - \bar{u}\|w_t\|)^\intercal\,\bar{u}\bar{u}^2 = \|v_t\|^2 + \left(w_t^\intercal\bar{u} - \|w_t\|\right)^2.$$

Since $\|v_t\| = \mathcal{O}(\ln n)$ whereas $\|w_t\| \to \infty$ via $\sum_j \eta_j = \infty$ and Lemma 18, then for all sufficiently large t, $w_t^\intercal\bar{u} > \|v_t\|$, and thus $\|w_t\| \leq w_t^\intercal\bar{u} + \|v_t\|$, and

$$\|v_t\|^2 + \left(w_t^\intercal\bar{u} - \|w_t\|\right)^2 \leq 2\|v_t\|^2.$$

As such, combining these pieces with the inequality $\|w_t\| \geq \gamma\sum_{j<t}\hat{\eta}_j$ from Lemma 18,

$$\left\|\frac{w_t}{\|w_t\|} - \bar{u}\right\| \leq \frac{\sqrt{2}\|v_t\|}{\|w_t\|} = \mathcal{O}\left(\frac{\ln n}{\sum_{j<t}\hat{\eta}_j}\right).$$

For $\hat{\eta}_j = 1$, we have $\sum_{j<t} \hat{\eta}_j = t$. For $\eta_j = 1$, we have $\sum_{j<t} \hat{\eta}_j = \Omega(\ln(t))$ from Theorem 20.

For the lower bound, let Z be given by the data in Theorem 10, and by the guarantee there,

$$\left\| \frac{w_t}{\|w_t\|} - \bar{u} \right\| = \frac{\|w_t - \bar{u}\|w_t\|\|}{\|w_t\|} \geq \frac{\|[\|\,1]\Pi_\perp\,(w_t - \bar{u}\|w_t\|)}{\|w_t\|} = \frac{\|v_t\|}{\|w_t\|} \geq \frac{\ln n - \ln 2}{\|w_t\|}.$$

The proof is now complete after upper bounding $\|w_t\|$. For $\hat{\eta}_j = 1$, by Lemma 18, we can just take $\|w_t\| \leq t$. For $\eta_j = 1$, Soudry et al. (2017, Theorem 3) show that $\|w_t\| = \Theta(\ln(t))$. ■

Appendix E. Omitted proofs from Section 5

We first prove Lemmas 12 and 13, which can help us check the convexity and smoothness of ψ in general.

Proof (of Lemma 12) Note that $\nabla\psi(\xi)_i = \ell'(\xi_i)/\ell'\left(\psi(\xi)\right)$, and

$$\nabla^2\psi(\xi) = \operatorname{diag}\left(\frac{\ell''(\xi_1)}{\ell'\left(\psi(\xi)\right)}, \ldots, \frac{\ell''(\xi_n)}{\ell'\left(\psi(\xi)\right)}\right) - \frac{\ell''\left(\psi(\xi)\right)}{\ell'\left(\psi(\xi)\right)}\nabla\psi(\xi)\nabla\psi(\xi)^\top. \tag{24}$$

We need to show that for any $v \in \mathbb{R}^n$,

$$\sum_{i=1}^n \frac{\ell''(\xi_i)}{\ell'\left(\psi(\xi)\right)}v_i^2 \geq \frac{\ell''\left(\psi(\xi)\right)}{\ell'\left(\psi(\xi)\right)}\left(\sum_{i=1}^n \frac{\ell'(\xi_i)}{\ell'\left(\psi(\xi)\right)}v_i\right)^2. \tag{25}$$

Note that by the Cauchy-Schwarz inequality,

$$\left(\sum_{i=1}^n \frac{\ell'(\xi_i)}{\ell'\left(\psi(\xi)\right)}v_i\right)^2 \leq \left(\sum_{i=1}^n \frac{\ell''(\xi_i)}{\ell'\left(\psi(\xi)\right)}v_i^2\right)\left(\sum_{i=1}^n \frac{\ell'(\xi_i)^2}{\ell''(\xi_i)\ell'\left(\psi(\xi)\right)}\right),$$

and therefore to show eq. (25), we only need to show that

$$\frac{\ell'\left(\psi(\xi)\right)^2}{\ell''\left(\psi(\xi)\right)} \geq \sum_{i=1}^n \frac{\ell'(\xi_i)^2}{\ell''(\xi_i)},$$

or

$$\frac{\ell'\left(\psi(\xi)\right)^2}{\ell''\left(\psi(\xi)\right)} = \frac{\ell'\left(\ell^{-1}\left(\sum_{i=1}^n \ell(\xi_i)\right)\right)^2}{\ell''\left(\ell^{-1}\left(\sum_{i=1}^n \ell(\xi_i)\right)\right)} \geq \sum_{i=1}^n \frac{\ell'\left(\ell^{-1}\left(\ell(\xi_i)\right)\right)^2}{\ell''\left(\ell^{-1}\left(\ell(\xi_i)\right)\right)}. \tag{26}$$

Consider the function $\phi : (0, \infty) \to \mathbb{R}$ given by

$$\phi(s) := \frac{\ell'\left(\ell^{-1}(s)\right)^2}{\ell''\left(\ell^{-1}(s)\right)}.$$

Note that $\phi(s)/s = \ell'(z)^2/\left(\ell(z)\ell''(z)\right)$ for $z = \ell^{-1}(s)$, and since $\ell'^2/\ell\ell''$ is increasing, it follows that $\phi(s)/s$ is increasing on $(0, \infty)$, and $\lim_{s\to 0}\phi(s) = 0$. In other words, ϕ is super-additive, which then implies eq. (26). ■

Proof (of Lemma 13) Similarly to the proof of (Shalev-Shwartz and Singer, 2007, Lemma 14), to check that ψ is β-smooth with respect to the ℓ_∞ norm, we only need to ensure for any $\xi, v \in \mathbb{R}^n$, it holds that $v^\top \nabla^2 \psi(\xi) v \le \beta \|v\|_\infty^2$. By eq. (24), it is enough if

$$\sum_{i=1}^n \frac{\ell''(\xi_i)}{\ell'\left(\psi(\xi)\right)} v_i^2 \le \beta \max_{1\le i\le n} v_i^2. \tag{27}$$

For $\ell_{\exp}$,

$$\frac{\ell''(\xi_i)}{\ell'\left(\psi(\xi)\right)} = \frac{e^{\xi_i}}{\sum_{i=1}^n e^{\xi_i}}, \quad \text{and thus} \quad \sum_{i=1}^n \frac{\ell''(\xi_i)}{\ell'\left(\psi(\xi)\right)} v_i^2 \le \max_{1\le i\le n} v_i^2.$$

In general, if $\ell''(z) \le c\ell'(z)$, the since $\ell'(\xi_i) \le \ell'\left(\psi(\xi)\right)$, it holds that

$$\sum_{i=1}^n \frac{\ell''(\xi_i)}{\ell'\left(\psi(\xi)\right)} \le \sum_{i=1}^n \frac{c\ell'(\xi_i)}{\ell'\left(\psi(\xi)\right)} \le cn,$$

and thus we can let $\beta = cn$. ■

Next we prove Theorem 11.

Proof (of Theorem 11) The first two conditions of Assumption 1 are easy to verify in most cases; we only check that $\varphi(z) := z\ell'(z)/\ell(z)$ is increasing on $(-\infty, 0)$ for the logistic loss. We have

$$\varphi(z) = \frac{z}{(1+e^{-z})\ln(1+e^z)}, \text{ and } \varphi'(z) = \frac{(1+e^{-z})\ln(1+e^z) + ze^{-z}\ln(1+e^z) - z}{(1+e^{-z})^2 \ln(1+e^z)^2}.$$

Since $(1+e^{-z})\ln(1+e^z) > 0$, and

$$ze^{-z}\ln(1+e^z) - z = ze^{-z}\left(\ln(1+e^z) - e^z\right) > 0,$$

since $z < 0$ and $\ln(1+e^z) < e^z$, it follows that $\varphi'(z) > 0$.

On the third requirement of Assumption 1, for $\ell_{\exp}$ we have $\ell'\left(\ell^{-1}(s)\right) = s$, and thus the condition holds with $c = 1/b$. For $\ell_{\log}$ we have $\ell'\left(\ell^{-1}(s)\right) = 1 - e^{-s}$, and the condition holds with $c = 1/b$. For $\ell_{\mathrm{poly},k}$, if $a \ge \ell(0)/b$, then

$$\frac{\ell'\left(\ell^{-1}(a)\right)}{\ell'\left(\ell^{-1}(ab)\right)} \ge \frac{\ell'\left(\ell^{-1}\left(\ell(0)/b\right)\right)}{2k},$$

while if $a \le \ell(0)/b$, then $\ell^{-1}(a) \le \ell^{-1}(ab) \le 0$. Note that on $(-\infty, 0)$,

$$\ell'\left(\ell^{-1}(s)\right) = ks^{(k+1)/k}, \text{ and thus } \frac{\ell'\left(\ell^{-1}(a)\right)}{\ell'\left(\ell^{-1}(ab)\right)} = b^{-(k+1)/k}.$$

We use Lemma 12 to verify the convexity of ψ. For $\ell_{\exp}$, we have $\ell'^2/(\ell\ell'') = 1$. For $\ell_{\log}$, we have $\ell'^2/(\ell\ell'') = e^z/\ln(1+e^z)$, which is increasing. For $\ell_{\mathrm{poly},k}$, on $(-\infty, 0]$, we have $\ell'^2/(\ell\ell'') = k/(k+1)$. On $(0, \infty)$,

$$\frac{\ell'^2}{\ell\ell''} = \frac{\left(2k - \frac{k}{(1+z)^{k+1}}\right)^2}{\left(2kz + \frac{1}{(1+z)^k}\right)\frac{k(k+1)}{(1+z)^{k+2}}} = \frac{k^2\left(2(1+z)^{k+1} - 1\right)^2}{\left(2kz(1+z)^k + 1\right)k(k+1)},$$

and thus we only need to show

$$\alpha(z) := \frac{\left(2(1+z)^{k+1}-1\right)^2}{2kz(1+z)^k+1}$$

is increasing on $(0,\infty)$. Note that

$$\begin{aligned}\left(2kz(1+z)^k+1\right)^2\alpha'(z) = 2\left(2(1+z)^{k+1}-1\right)\cdot 2(k+1)(1+z)^k\cdot\left(2kz(1+z)^k+1\right)\\ -\left(2(1+z)^{k+1}-1\right)^2\left(2k(1+z)^k+2kz\cdot k(1+z)^{k-1}\right),\end{aligned}$$

and therefore we only need to show that on $(0,\infty)$,

$$\kappa(z) := 2(k+1)(1+z)\cdot\left(2kz(1+z)^k+1\right)-\left(2(1+z)^{k+1}-1\right)\left(k(1+z)+k^2z\right)\geq 0.$$

Rearranging terms gives

$$\kappa(z) = 2k(1+z)^{k+1}(kz+z-1)+(k^2+3k+2)z+(3k+2).$$

Note that $\kappa(0)=k+2>0$, and when $z\geq 0$,

$$\begin{aligned}\kappa'(z) &= 2k(k+1)(1+z)^k(kz+z-1)+2k(1+z)^{k+1}(k+1)+(k^2+3k+2)\\ &= 2k(k+1)(1+z)^k(kz+2z)+(k^2+3k+2)>0.\end{aligned}$$

Therefore $\kappa>0$ on $(0,\infty)$.

The smoothness of ψ is established by Lemma 13. ■

E.1. Warm start tools for the logistic loss

If we try to prove fast margin rates for the logistic loss directly from Theorem 5, we will pay for the bad initial smoothness of the corresponding ψ, which is n, and the rate will be n/t. The smoothness later improves, which is proved as follows.

Proof (of Lemma 14) We first prove that for all $\xi\in\mathbb{R}^n$ with $\psi(\xi)\leq 0$, it holds that $1\leq\|\nabla\psi(\xi)\|_1\leq 2$. Given $\xi\in\mathbb{R}^n$, we have

$$\|\nabla\psi(\xi)\|_1 = \sum_{i=1}^n\frac{\ell'(\xi_i)}{\ell'\left(\psi(\xi)\right)} = \sum_{i=1}^n\frac{\ell'\left(\ell^{-1}\left(\ell(\xi_i)\right)\right)}{\ell'\left(\ell^{-1}\left(\sum_{i=1}^n\ell(\xi_i)\right)\right)}.$$

Consider the function $\rho(z) := \ell'\left(\ell^{-1}(z)\right) = 1-e^{-z}$. It holds that $\rho(0)=0$, and on $z\in[0,\ell(0)]=[0,\ln(2)]$, we have $\rho(z)'\in[1/2,1]$, and ρ is subadditive: for all $a,b>0$ with $a+b\leq\ln(2)$, it holds that $\rho(a+b)\leq\rho(a)+\rho(b)$. Now note that since $\psi(\xi)\leq 0$, we have $\sum_{i=1}^n\ell(\xi_i)\leq\ell(0)$, and thus the subadditivity of ρ implies

$$\|\nabla\psi(\xi)\|_1 = \sum_{i=1}^n\frac{\ell'\left(\ell^{-1}\left(\ell(\xi_i)\right)\right)}{\ell'\left(\ell^{-1}\left(\sum_{i=1}^n\ell(\xi_i)\right)\right)} = \frac{\sum_{i=1}^n\rho\left(\ell(\xi_i)\right)}{\rho\left(\sum_{i=1}^n\ell(\xi_i)\right)}\geq 1.$$

On the other hand, the mean value theorem implies

$$\|\nabla\psi(\xi)\|_1 = \frac{\sum_{i=1}^n \rho\left(\ell(\xi_i)\right)}{\rho\left(\sum_{i=1}^n \ell(\xi_i)\right)} \le \frac{\sum_{i=1}^n \ell(\xi_i)}{\frac{1}{2}\sum_{i=1}^n \ell(\xi_i)} = 2.$$

Then we show that ψ is 2-smooth with respect to the ℓ_∞ norm on the sublevel set $\{\xi|\psi(\xi)\le 0\}$. Recall from eq. (27) that we only need to check

$$\sum_{i=1}^n \frac{\ell''(\xi_i)}{\ell'\left(\psi(\xi)\right)} v_i^2 \le 2 \max_{1\le i\le n} v_i^2.$$

This is true since $\ell'' \le \ell'$, and $\sum_{i=1}^n \ell'(\xi_i)/\ell'\left(\psi(\xi)\right) = \|\nabla\psi(\xi)\|_1 \le 2$.

Next we prove the iteration guarantee on ψ. Let

$$\tilde\eta := \arg\max\left\{0\le\hat\eta\le 1 \middle| \psi\left(Z(w_t - \hat\eta Z^\top q_t)\right)\le 0\right\}, \text{ and } \tilde w := w_t - \tilde\eta Z^\top q_t.$$

Since $\mathcal{L}(Zw_t) < \ell(0)$, we have $\psi(Zw_t) < 0$, and thus $\tilde\eta > 0$. We claim that $\tilde\eta \ge 1/2$. If this is not true, then we must have $\psi(\tilde w) = 0$. Since ψ is convex, and $\psi(Zw_t) < 0$, the line between Zw_t and $Z\tilde w$ are all in the sublevel set $\{\xi|\psi(\xi)\le 0\}$. Using 2-smoothness of ψ, and the same analysis as in Lemma 3, we have

$$\begin{aligned}
\psi(Z\tilde w) - \psi(Zw_t) &\le \langle q_t, Z\tilde w - Zw_t\rangle + \|Z\tilde w - Zw_t\|_\infty^2 \\
&= -\tilde\eta\left\|Z^\top q_t\right\|^2 + \tilde\eta^2\left\|ZZ^\top q_t\right\|_\infty^2 \\
&\le -\tilde\eta\left\|Z^\top q_t\right\|^2 + \tilde\eta^2\left\|Z^\top q_t\right\|^2. \qquad (28)
\end{aligned}$$

Since $\psi(Zw_t) < 0$, and $0 < \tilde\eta \le 1/2$ due to our assumption, and $\left\|Z^\top q_t\right\| > 0$ by Theorem 5, we have $\psi(Z\tilde w) < 0$, a contradiction. As a result, $\tilde\eta \ge 1/2$, and the iteration guarantee follows from eq. (28).

Next we prove the strong-convexity-style property for ψ^*. Let ξ, ξ' satisfy

$$\psi(\xi), \psi(\xi') \le \ell^{-1}\left(\frac{\ell(0)}{2e^2}\right), \quad \text{or} \quad \mathcal{L}(\xi), \mathcal{L}(\xi') \le \frac{\ell(0)}{2e^2}.$$

Since $\mathcal{L}(\xi) = \sum_{i=1}^n \ell(\xi_i)$, it follows that for all $1\le i\le n$, we have $\ell(\xi_i), \ell(\xi_i') \le \ell(0)/(2e^2)$, and thus $\xi_i, \xi_i' \le 0$. Note that for all $z \le 0$, we have $e^z/2 \le \ln(1+e^z) \le e^z$, therefore

$$\frac{\ell(z+2)}{\ell(z)} = \frac{\ln(1+e^{z+2})}{\ln(1+e^z)} \le \frac{e^{z+2}}{e^z/2} = 2e^2.$$

Consequently, for all $\tilde\xi \in \mathbb{R}^n$ such that $\|\xi - \tilde\xi\|_\infty \le 2$, it holds that $\mathcal{L}(\tilde\xi) \le \ell(0)$, and thus $\psi(\tilde\xi) \le 0$. Now let $\theta = \nabla\psi(\xi)$, and $\theta' = \nabla\psi(\xi')$, and

$$\tilde\xi_i := \xi_i - \frac{\|\theta - \theta'\|_1}{2}\cdot \operatorname{sgn}\left(\theta_i - \theta_i'\right).$$

Recall from the proof of Theorem 1 that

$$\begin{aligned} D_{\psi^*}(\theta', \theta) &= \psi(\xi) - \psi(\xi') - \langle \theta', \xi - \xi' \rangle \\ &= \psi(\xi) - \psi(\tilde{\xi}) + \psi(\tilde{\xi}) - \psi(\xi') - \langle \theta', \xi - \xi' \rangle \\ &= \psi(\xi) - \psi(\tilde{\xi}) - \langle \theta', \xi - \tilde{\xi} \rangle + \psi(\tilde{\xi}) - \psi(\xi') - \langle \theta', \tilde{\xi} - \xi' \rangle. \end{aligned} \tag{29}$$

Note that $\|\theta\|_1, \|\theta'\|_1 \leq 2$, therefore $\|\theta - \theta'\|_1 \leq 4$. It then follows that $\|\xi - \tilde{\xi}\|_\infty = \|\theta - \theta'\|_1/2 \leq 2$ and $\psi(\xi), \psi(\tilde{\xi}) \leq 0$, and since ψ is 2-smooth on the 0-sublevel set, we have

$$\psi(\xi) - \psi(\tilde{\xi}) \geq \langle \theta, \xi - \tilde{\xi} \rangle - \|\xi - \tilde{\xi}\|_\infty^2 = \langle \theta, \xi - \tilde{\xi} \rangle - \frac{\|\theta - \theta'\|_1^2}{4}. \tag{30}$$

Then eqs. (29) and (30) and the convexity of ψ imply

$$\begin{aligned} D_{\psi^*}(\theta', \theta) &\geq \langle \theta, \xi - \tilde{\xi} \rangle - \frac{\|\theta - \theta'\|_1^2}{4} - \langle \theta', \xi - \tilde{\xi} \rangle + \psi(\tilde{\xi}) - \psi(\xi') - \langle \theta', \tilde{\xi} - \xi' \rangle \\ &\geq \langle \theta, \xi - \tilde{\xi} \rangle - \frac{\|\theta - \theta'\|_1^2}{4} - \langle \theta', \xi - \tilde{\xi} \rangle \\ &= \langle \theta - \theta', \xi - \tilde{\xi} \rangle - \frac{\|\theta - \theta'\|_1^2}{4}. \end{aligned}$$

By the construction of $\tilde{\xi}$, we have $\langle \theta - \theta', \xi - \tilde{\xi} \rangle = \|\theta - \theta'\|_1^2/2$, therefore

$$D_{\psi^*}(\theta', \theta) \geq \frac{\|\theta - \theta'\|_1^2}{4}.$$

■

The preceding analysis requires $\mathcal{L}(Zw_t) \leq \ell(0)/(2e^2)$. We now produce a second analysis to handle those initial iterations leading to this condition.

Lemma 21 *Consider the logistic loss $\ln(1 + e^z)$, with step size $\eta_j = 1/(2\mathcal{R}(w_j))$. Suppose Assumption 2 holds, and let γ and $\bar{u}$ denote the corresponding maximum margin value and direction. Then the first iteration t with $\mathcal{L}(Zw_t) \leq \ell(0)/(2e^2)$ satisfies $\psi(Zw_t) \leq 0$ and*

$$t \leq \left(\frac{256 \ln n}{\gamma} \right)^2 \qquad \text{and} \qquad \|w_t\| \leq \frac{256 \ln n}{\gamma}.$$

Proof Let t denote the first iteration with $\mathcal{R}(w_t) \leq 1/(32n)$, whereby $\mathcal{L}(Zw_t) \leq \ln(2)/(2e^2)$ and $\psi(Zw_t) \leq 0$. Additionally, define $r := 128 \ln n/\gamma$ and $u := r\bar{u}$; Expanding the square and invoking Theorem 19 and convexity of $\mathcal{R}$, for any $j < t$,

$$\begin{aligned} \|w_{j+1} - u\|^2 &= \|w_j - u\|^2 - 2\eta_j \langle \nabla \mathcal{R}(w_j), w_j - u \rangle + \eta_j^2 \|\nabla \mathcal{R}(w_j)\|^2 \\ &\leq \|w_j - u\|^2 + 2\eta_j \left(\mathcal{R}(u) - \mathcal{R}(w_j) \right) + \eta_j^2 \|\nabla \mathcal{R}(w_j)\|^2. \end{aligned}$$

Applying $\sum_{j<t}$ to both sides and telescoping,

$$2 \sum_{j<t} \eta_j \mathcal{R}(w_j) + \|w_t - u\|^2 - \|w_0 - u\|^2 \leq 2 \sum_{j<t} \eta_j \mathcal{R}(u) + \sum_{j<t} \eta_j^2 \|\nabla \mathcal{R}(w_j)\|^2. \tag{31}$$

The various terms in this expression can be simplified as follows.

- Since $\eta_j = \frac{1}{2\mathcal{R}(w_j)}$, then $2\sum_{j<t}\eta_j\mathcal{R}(w_j) = t$.
- Since $w_0 = 0$,
$$\|w_t - u\|^2 - \|w_0 - u\|^2 = \|w_t\|^2 - 2\langle w_t, u\rangle = \|w_t\|^2 - 2r\langle w_t, \bar{u}\rangle .$$
- Since $\ell_{\log} \le \ell_{\exp}$, by the choice of r,
$$\mathcal{R}(u) \le \mathcal{R}_{\exp}(u) \le \frac{1}{n}\sum_i \exp(-\langle z_i, \bar{u}\rangle r) \le \frac{1}{128n},$$
and using $\mathcal{R}(w_j) \ge \frac{1}{32n}$ and $\eta_j = \frac{1}{2\mathcal{R}(w_j)} \le 16n$ gives
$$2\sum_{j<t}\eta_j\mathcal{R}(u) \le 2\sum_{j<t}\frac{16n}{128n} = \frac{t}{4}.$$
- Since $\ell' \le \ell$,
$$\|\nabla\mathcal{R}(w_j)\| = \left\|\frac{1}{n}\sum_{i=1}^n \ell'(\langle z_i, w_j\rangle) z_i\right\| \le \frac{1}{n}\sum_{i=1}^n \ell'(\langle z_i, w_j\rangle)\|z_i\| \le \frac{1}{n}\sum_{i=1}^n \ell(\langle z_i, w_j\rangle) = \mathcal{R}(w_j),$$
and since $\eta_j = \frac{1}{2\mathcal{R}(w_j)}$,
$$\sum_{j<t}\eta_j^2\|\nabla\mathcal{R}(w_j)\|^2 \le \sum_{j<t}\eta_j^2\mathcal{R}(w_j)^2 = \frac{t}{4}.$$

Combining these inequalities with eq. (31) gives

$$\begin{aligned} t + \|w_t\|^2 - 2r\langle w_t, \bar{u}\rangle &\le 2\sum_{j<t}\eta_j\mathcal{R}(w_j) + \|w_t - u\|^2 - \|w_0 - u\|^2 \\ &\le 2\sum_{j<t}\eta_j\mathcal{R}(u) + \sum_{j<t}\eta_j^2\|\nabla\mathcal{R}(w_j)\|^2 \\ &\le \frac{t}{4} + \frac{t}{4}, \end{aligned}$$

which rearranges to give

$$\frac{t}{2} + \|w_t\|^2 - 2r\langle w_t, \bar{u}\rangle \le 0. \tag{32}$$

Since $t \ge 0$, then by Cauchy-Schwarz, $\|w_t\| \le 2r$. Similarly, Cauchy-Schwarz grants

$$\frac{t}{2} + \|w_t\|^2 - 2r\langle w_t, \bar{u}\rangle \ge \frac{t}{2} + \|w_t\|^2 - 2r\|w_t\|,$$

which is in fact minimized when $\|w_t\| = 2r$, giving

$$\frac{t}{2} + \|w_t\|^2 - 2r\langle w_t, \bar{u}\rangle \ge \frac{t}{2} - 2r^2,$$

which combined with eq. (32) implies $t \le 4r^2$. ■