Implicit bias of gradient-descent: fast convergence rates

Elvis Dohmatob e.dohmatob@criteo.com
Criteo AI Lab

We consider gradient-flow (GF) and gradient-descent (GD) on linear classification problems. For exponential-tailed loss functions, including the usual exponential and logistic loss functions, we establish $\mathcal{O}(\log(n)/t)$ convergence rate for the bias in case of GF, and $\widetilde{\mathcal{O}}(\log(n)/\sqrt{t})$ in case of GD. Upto logarithmic factors, our GD rate already matches the very recent parallel work from Ji and Telgarsky (2020) which uses an aggressive stepsize schedule to establish a $\mathcal{O}(1/\sqrt{t})$ convergence rate for the bias. Finally, using the aggressive stepsize schedule proposed by Ji and Telgarsky (2020), we are able to obtain a convergence rate of $\mathcal{O}(\log(n)/t)$ for the bias. This is presently the fastest known rates for the convergence of the bias of gradient-descent. Our methods of analysis are quite general and radically different from the usual techniques used in the literature: we use nonlinear error analysis for convex functions, in the spirit of Kurdyka-Łojasiewicz theory. One major advantage of our method is that it allows us to convert any convergence rate for the margin, to a convergence rate on the bias, which is at least as good as the former. We believe our work will provide an alternative approach for analyzing the implicit bias of gradient-flow / gradient-descent in very general settings.

1 Introduction

1.1 Problem setup

All through this manuscript, $\mathbb{R}^m$ will be equipped with the euclidean / ℓ_2-norm, which we will simply write, $\|\cdot\|$ (without the subscript 2). We consider binary classification problems with data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ drawn frop an unknown distribution on $\mathbb{R}^m \times \{\pm 1\}$. For each $i \in [n]$, $y_i \in \{\pm 1\}$ is the label and $\mathbf{x}_i \in \mathbb{R}^m$ are the features of the ith example. For simplicity, we will assume $\|\mathbf{x}_i\| \le 1$ for all $i \in [i]$ The integer $n \ge 2$ is the sample size, while m is the dimensionality of the problem. We are interested in "large margin" linear classifiers. Any such model is indexed by a vector of parameters $\mathbf{w} \in \mathbb{R}^m$. The prediction on an input example $\mathbf{x} \in \mathbb{R}^m$ is $h_{\mathbf{w}}(\mathbf{x}) := \operatorname{sign}(\langle \mathbf{x}, \mathbf{w}\rangle) \in \{\pm 1\}$, where $\langle \mathbf{x}, \mathbf{w}\rangle$ is the inner product of $\mathbf{x}$ and $\mathbf{w}$, also usually denoted $\mathbf{x}^\top \mathbf{w}$. Note that $h_{\mathbf{w}}$ does not depend on the norm of the parameters $\mathbf{w}$, as $h_{\lambda \mathbf{w}}(\mathbf{x}) = h_{\mathbf{w}}(\mathbf{x})$ for all $\lambda > 0$ and $\mathbf{x} \in \mathbb{R}^m$. The *hard margin* of $\mathbf{w}$ (i.e of $h_{\mathbf{w}}$), denoted $\gamma_n(\mathbf{w})$, is defined by

$$\gamma_n(\mathbf{w}) := \min_{i \in [n]} \frac{y_i \langle \mathbf{x}_i, \mathbf{w}\rangle}{\|\mathbf{w}\|} \tag{1}$$

is the *hard-margin* of the model $\mathbf{w} \in \mathbb{R}^m$. This measures the minimum (signed) distance of the samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ to the decision boundary of the linear classifier $h_{\mathbf{w}}$. Consider the optimal / maximum margin $\overline{\gamma}_n \in [0, 1]$ for the problem, defined by

$$\overline{\gamma}_n := \max_{\mathbf{w} \in \mathbb{B}_m} \gamma_n(\mathbf{w}) = \max_{\|\mathbf{w}\| \le 1} \min_{i \in [n]} y_i \langle \mathbf{x}_i, \mathbf{w}\rangle. \tag{2}$$

where $\mathbb{B}_m := \{\mathbf{w} \in \mathbb{R}^m \mid \|\mathbf{w}\| \le 1\}$ is the (closed) unit-ball in m-dimensional euclidean space $\mathbb{R}^m$. We will suppose the problem is (linearly) separable, meaning that $\overline{\gamma}_n > 0$.

Finally, let $\mathsf{OPT}_n := \arg\max_{\mathbf{w} \in \mathbb{B}_m} \gamma_n(\mathbf{w}) := \{\mathbf{w} \in \mathbb{B}_m \mid \gamma_n(\mathbf{w}) = \overline{\gamma}_n\}$ be set of all max-margin models. Note that OPT_n is nonempty, compact, and convex.

Remark 1.1. *For our results, we will not require OPT_n to be a singleton (i.e that there is a unique max-margin classifier), an unnecessary restriction usually placed for convenience sake, by most other works in the literature.*

Optimization, smooth margin, empirical risk. Producing a point in OPT_n (i.e maximizing the hard-margin function γ_n) is challenging due to the nonsmooth and combinatorial nature of the problem. For this reason, one usually tries to instead maximize a proxy of γ_n instead, via a loss / risk functional. In this paper we will focus on the so-called exponential risk or smoothed margin function, constructed as thus. For any $\beta \in (0, \infty)$ and $\mathbf{w} \in \mathbb{R}^m$, define a Boltzman distribution $\widehat{\mathbf{q}}_{n,\beta}(\mathbf{w}) \in \Delta_{n-1}$ by setting $q_i \propto e^{-\beta y_i \langle \mathbf{x}_i, \mathbf{w}\rangle}$ for each $i \in [n]$. Define the smooth margin at temperature β, of a model $\mathbf{w}$

$$F_n(\mathbf{w}) := -\ell_{\exp}^{-1}(\mathcal{R}_n) = -\frac{1}{\beta}\log(\frac{1}{n}\sum_{i=1}^{n} e^{-\frac{1}{\beta}y_i\langle\mathbf{x}_i,\mathbf{w}\rangle}), \tag{3}$$

where $\ell_{\exp}(u) := e^{\beta u}$ is the exponential loss, and $\mathcal{R}_n(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n}\ell_{\exp}(-y_i\langle\mathbf{x}_i,\mathbf{w}\rangle))$ is the empirical risk. Note that

$$\gamma_n(\mathbf{w}) \le F_n(\mathbf{w}) \le \gamma_n(\mathbf{w}) + \log n. \tag{4}$$

Implicit bias. One is then interested in the convergence of approximate minimizers of $-F_n$ via, say gradient-descent, to a point in OPT_n. This is the so-called *implicit bias* of gradient-descent. See Soudry et al. (2017) and Ji and Telgarsky (2019), for example.

Summary of main contributions. Our main contributions can be summarized as follows.

- For exponential-tailed losses (as defined in Soudry et al. (2017)), including the exponential loss $\ell_{\exp}$ in the definition of the smoothed margin F_n defined above, we show in Theorem 2.1 that gradient-flow on the empirical risk produces a sequence of iterates whose margin and bias both converge at a rate of $\mathcal{O}(\log n/t)$, where $n \ge 2$ is the number of samples in the training sample, and $t > 0$ is the clock of the gradient-flow (analogous to number of iterations in gradient-descent). Via time discretization, we show in Theorem 2.2 that this leads to a $\widetilde{\mathcal{O}}(\log n/\sqrt{t})$ bias convergence rate in the case of gradient-descent.
- An important by-product of our technique is that it allows us to convert any convergence rate for the margin, to a convergence rate for the bias which is at least as good as the former. In particular, using our methods we derive a convergence rate of $\mathcal{O}(\log n/t)$ for the bias, directly from the $\mathcal{O}(\log n/t)$ margin rate that has been very recently established in Ji and Telgarsky (2019). We believe our methods provide a powerful new way to derive convergence rates for the implicit bias of gradient-flow / gradient-descent.

2 Implicit bias of gradient-flow and gradient-descent: fast rates

In this section, develop and prove our main contributions.

More notations. If A is a nonempty subset of $\mathbb{R}^m$, the distance $\mathsf{dist}(\mathbf{w}, A)$ of a point $\mathbf{w} \in \mathbb{R}^m$ from A is defined by

$$\mathsf{dist}(\mathbf{w}, A) := \inf\{\|\mathbf{w}' - \mathbf{w}\| \mid \mathbf{w}' \in A\}.$$

The indicator function of A is the function $i_A : M \to \mathbb{R} \cup \{\infty\}$ defined by

$$i_A(\mathbf{w}) := \begin{cases} 0, & \text{if } \mathbf{w} \in A, \\ \infty, & \text{else.} \end{cases}$$

The $(n-1)$-dimensional probability simplex, denoted Δ_{n-1}, is defined by $\Delta_{n-1} := \{\mathbf{v} \in \mathbb{R}^n \mid q_1, \ldots, q_n \ge 0,\ \sum_{i=1}^{n} q_i = 1\}$.

2.1 Statement of main results

Consider the following gradient-flow (aka continuous-time limit of gradient-descent) on $-F_n$

$$\frac{d\mathbf{w}(t)}{dt} = \nabla F_n(\mathbf{w}(t)),\ \mathbf{w}(0) = \mathbf{0}. \tag{5}$$

The initial condition $\mathbf{w}(0) = \underline{0}$ is only for convenience and plays no crucial role in our results and proofs. We will analyse the gradient-flow scheme (5) and show that its solution converges wit time, to a limit point (after normalization by norm) of OPT_n, at a rate $\mathcal{O}(1/t)$

The following is one of our main results. It can be extended to any exponential-tail loss function (as defined in Soudry et al. (2017)). Viz

Theorem 2.1 ($\mathcal{O}(1/t)$ convergence rate of the bias of gradient-flow)**.** *Let $\mathbf{w}_n(t)$ be a solution of* (5) *at time $t > 0$. Then, we have the the following bound*

$$\mathsf{dist}\left(\widetilde{\mathbf{w}}_n(t), \mathsf{OPT}_n\right) \le \frac{\log n}{\overline{\gamma}_n^2 t}, \tag{6}$$

where $\widetilde{\mathbf{w}}_n(t) := \mathbf{w}_n(t)/\|\mathbf{w}_n(t)\| \in \mathbb{B}_m$ is the normalized version of $\mathbf{w}_n(t)$.

Remark 2.1. *Note that it is already known via standard gradient-flow arguments (see Bach (2020), for example) that the corresponding margin $\gamma_n(\mathbf{w}_n(t))$ in the above theorem converges to the optimal margin $\overline{\gamma}_n$ at a rate $\mathcal{O}(1/t)$.*

Definition 2.1. *The quantity $\mathsf{dist}(\widetilde{\mathbf{w}}_n(t), \mathsf{OPT}_n)$ will be referred to as the bias of gradient-flow. In general, the implicit bias of an algorithm will be the rate of convergence of the (normalized) iterates produced by that algorithm and the set of max-margin classifiers.*

Paper	*Margin convergence*	*Bias convergence*	*Algorithm*
Soudry et al. (2017)	$\mathcal{O}(\log\log t/\log t)$	$\mathcal{O}(1/\log t)$	GD, constant stepsize
Nacson et al. (2018)	$\widetilde{\mathcal{O}}(1/\sqrt{t})$	$\mathcal{O}(1/\log t)$	GD, adaptive stepsize
Ji and Telgarsky (2019)	$\mathcal{O}(1/\log t)$	$\mathcal{O}(1/\log t)$	GD, constant stepsize
Ji and Telgarsky (2020)	$\mathcal{O}(1/t)$	$\mathcal{O}(1/\sqrt{t})$	GD, aggressive stpesize
*Our paper	$\mathcal{O}(1/t)$	$\mathcal{O}(1/t)$	GF (no stepsize)
*Our paper	$\widetilde{\mathcal{O}}(1/\sqrt{t})$	$\widetilde{\mathcal{O}}(1/\sqrt{t})$	GD (GF discretization)
*Our paper	$\mathcal{O}(1/t)$	$\mathcal{O}(1/t)$	GD, aggressive stepsize

Table 1: Convergence rates established in different papers, for both the margin and the bias of gradient-descent / gradient-flow. Here, only a handful of the most representative papers are shown here for the purpose of comparison. It should be noted that in our work as well as in the works of Ji and Telgarsky above, the dependence on the sample size n is a multiplicative factor $\log n$ in the numerator. The notation $\widetilde{\mathcal{O}}(\ldots)$ means all terms which are logarithmic in t have been ignored.

The next theorem, which is a discrete-time version of Theorem 2.1 is our second main contribution.

Theorem 2.2 ($\widetilde{\mathcal{O}}(1/\sqrt{t})$ convergence rate of the bias of gradient-descent with adaptive stepsizes)**.** *Consider gradient-descent $\mathbf{w}_n(t+1) := \mathbf{w}_n(t) + \eta_t \nabla F_n(\mathbf{w}_n(t))$, with stepsizes $\eta_t = 1/\sqrt{t+1}$ on $-F_n$, where $t \in \mathbb{N}$ is the iteration count. Then, we have the following bound*

$$\mathsf{dist}(\widetilde{\mathbf{w}}_n(t), \mathsf{OPT}_n) = \widetilde{\mathcal{O}}(\frac{\log n}{\overline{\gamma}_n^2 \sqrt{t}}), \tag{7}$$

where $\widetilde{\mathbf{w}}_n(t) := \mathbf{w}_n(t)/\|\mathbf{w}_n(t)\| \in \mathbb{B}_m$ is the normalized version of $\mathbf{w}_n(t)$ as usual.

Finally, we have the following theorem which improves the main result in Ji and Telgarsky (2020).

Theorem 2.3 ($\mathcal{O}(1/t)$ convergence rate of the bias of gradient-descent with aggressive stepsizes)**.** *Consider gradient-descent $\mathbf{w}_n(t+1) := \mathbf{w}_n(t) - \eta_t \nabla \mathcal{R}_n(\mathbf{w}_n(t))$ on the empirical risk $\mathcal{R}_n$, with stepsizes $\eta_t = \Theta(1/\mathcal{R}_n(\mathbf{w}_n(t)))$. We have the bound*

$$\mathsf{dist}(\widetilde{\mathbf{w}}_n(t), \mathsf{OPT}_n) = \mathcal{O}(\frac{\log n}{t}), \tag{8}$$

where $\widetilde{\mathbf{w}}_n(t) := \mathbf{w}_n(t)/\|\mathbf{w}_n(t)\| \in \mathbb{B}_m$.

Remark 2.2. *A few important points to note about results.*

- ***Improved SOTA.*** *Our results in Theorems* (2.2) *and* (2.3) *are a net improvement on best known results from Ji and Telgarsky (2019), where a convergence rate of $\mathcal{O}(1/\log t)$ was obtained for the bias.*
- ***Parallel work Ji and Telgarsky (2020).*** *The very recent work Ji and Telgarsky (2020) has obtained using an adaptive stepsize strategy, a convergence rate of $\mathcal{O}(1/t)$ for the GD margin, and $\mathcal{O}(1/\sqrt{t})$ rate for the convergence of the bias. Upto a logarithmic factor in*

t, this rate matches the bias rate in our Theorem 2.2. Finally, using the same aggressive stepsizes as in Ji and Telgarsky (2020), our Theorem 2.3 establishes a $\mathcal{O}(1/t)$ convergence rate for the bias.

- ***Universality of our approach.*** *The bias rates* (6) *and* (7) *are obtained as follows. Once we have a margin rate, our technique can always ensure a bias rate not worst than the former. This is a desirable property, because it allows us to convert any margin rate to a bias rate which is at least as good, and without any specialized further analysis for the latter. This is unlike the approaches in the literature wherein each paper requires a separate case-by-case analysis for both the margin and bias rate.*

Sketch of proof of Theorems 2.1,2.2, and 2.3. Using a standard gradient-flow argument (see Bach (2020), for example), we know that

$$\gamma_n(\widetilde{\mathbf{w}}_n(t)) \geq \overline{\gamma}_n - \frac{\log n}{\overline{\gamma}_n t}. \tag{9}$$

We will first establish an inequality of the form

$$\mathsf{dist}\left(\mathbf{w}, \mathsf{OPT}_n\right) \leq 1 - \frac{\gamma_n(\mathbf{w})}{\overline{\gamma}_n}, \tag{10}$$

which holds for all $\mathbf{w} \in \mathbb{B}_m$. To this end, we will employ tools from nonsmooth analysis (section 4.1). Combining (9) and (10) with $\mathbf{w} = \widetilde{\mathbf{w}}_n(t) \in \mathbb{B}_m$ then gives the result.

For the proof of Theorem 2.2, we discretize the gradient-flow (5) with stepsizes of order $\eta_t = 1/\sqrt{t+1}$, where $t \in \mathbb{N}$ is the iteration counter. We obtain the following discrete version of (9)

$$\gamma_n(\widetilde{\mathbf{w}}_n(t)) \geq \overline{\gamma}_n - \widetilde{\mathcal{O}}(\frac{\log n}{\overline{\gamma}_n \sqrt{t}}), \tag{11}$$

via arguments similar to the proof of (Nacson et al., 2018, Theorem 5).

The bound (7) then follows from combining (11) and (10).

Finally, for the proof of Theorem 2.3, we combine the $\mathcal{O}(1/t)$ margin rate from (Ji and Telgarsky, 2020, Theorem 4.2) and our (10) above. □

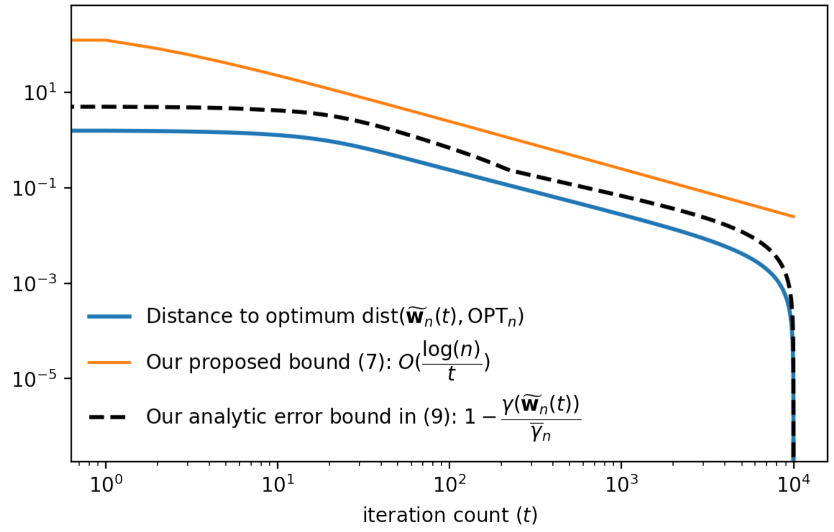

Figure 1: Convergence rate of bias of gradient-descent. This toy example is on a separable binary classification problem on simulated data with $m = 20$ features, $n = 4000$ samples, and optimal margin $\overline{\gamma}_n = 0.2$. Note that $\widetilde{\mathbf{w}}_n(t) := \mathbf{w}_n(t)/\|\mathbf{w}_n(t)\| \in \mathbb{B}_m$, where $\mathbf{w}_n(t) \in \mathbb{R}^m$ are the model parameters after t steps of gradient-descent, as predicted by Theorem 2.3.

3 Related works

There is rich history of research trying to understand the so-called implicit bias of gradient-descent In the case of linear models with so-called exponential-tailed losses (the scenario considered in our paper), let us mention Soudry et al. (2017), Nacson et al. (2018), Gunasekar et al. (2018), Ji and Telgarsky (2019, 2020), etc.

Parallel work. The very recent result Ji and Telgarsky (2020) establishes a convergence rate of $\mathcal{O}(1/\sqrt{t})$ for the bias of gradient-descent. (Theorem 2.1). More precisely, they show that gradient-descent on the smoothed (negative) margin F_n using stepsizes $(\eta_t)_{t\in\mathbb{N}}$ produces iterates $\mathbf{w}_n(t)$ with

$$\gamma_n(\widetilde{\mathbf{w}}_n(t)) \ge \overline{\gamma}_n - \mathcal{O}(\frac{\log n}{\sum_{j=1}^t \widehat{\eta}_j}), \tag{12}$$

where $\widetilde{\mathbf{w}}_n(t) := \mathbf{w}_n(t)/\|\mathbf{w}_n(t)\| \in \mathbb{B}_m$ is the normalized sequence of iterates and $\widehat{\eta}_t := \eta_t F_n(\mathbf{w}_n(t))$. If the stepsizes η_t are chosen aggressively such that $\widehat{\eta}_t$ is constant (or bounded), then the above result matches gives a margin bound of $\overline{\gamma}_n - \gamma_n(\widetilde{\mathbf{w}}_n(t)) = \mathcal{O}(\log n/t)$. The authors further prove (see (Ji and Telgarsky, 2020, Theorem 3.1)) that the bias rate (refer to Definition 2.1) of $\mathcal{O}(1/\sqrt{t})$, which matches the result in our Theorem 2.2. A comparative view of all the existing convergence rates is presented in Table 1.

Neural networks with homogeneous activation functions. In this delicate scenario, let us mention Chizat and Bach (2020) which analyzes gradient-descent on neural networks with one hidden-layer with logistic loss function, and Lyu and Li (2020) which studies deep neural networks with positive-homogeneous activation functions (e.g RELU) and exponential-tail loss functions.

4 Tools and full proof of the main result

4.1 Nonsmooth analysis in Banach spaces: a central tool

We start with a swift digression to nonsmooth analysis, the final purpose being to arrive at (10), and ultimately a proof of our main result (Theorem 2.1). The material developed here is valid for any Banach space $M = (M, \|\cdot\|)$ with topological dual $M^\star = (M^\star, \|\cdot\|_\star)$; recall that this is made of all bounded linear maps $M \to \mathbb{R}$. The self-dual finite-dimensional euclidean case $M = (\mathbb{R}^m, \|\cdot\|_2)$ is only a peculiar instance of this general setup.

Definition 4.1 (Subdifferential, see Rockafellar (1970))**.** *Given an function $f : M \to \mathbb{R} \cup \{\infty\}$ and a point $\mathbf{w} \in M$, the subdifferential of f at $\mathbf{w}$, denoted $\partial f(\mathbf{w})$, is defined by*

$$\partial f(\mathbf{w}) := \{\mathbf{x} \in M^\star \mid f(\mathbf{w}') \ge f(\mathbf{w}) + \langle \mathbf{x}, \mathbf{w}' - \mathbf{w}\rangle,\ \forall \mathbf{w}' \in M\},$$

f is said to be subdifferentiable at $\mathbf{w}$ just in case $\partial f(\mathbf{w}) \ne \emptyset$.

The following fundamental theorem from Corvellec and Motreanu (2007) (see Proposition 5.2 therein) gives a functional link between the function value at a point, the norm of subgradients at the point, and the distance of the point from the minimizers of a function.

Proposition 4.1 (Generalized Hoffman error-bound)**.** *Let $f : M \to \mathbb{R} \cup \{\infty\}$ be a proper convex lower-semicontinuous (l.s.c) function function with* $\arg\min f \ne \emptyset$*. Then for every* $\mathbf{w} \in M$*, we have*

$$\|\partial f(\mathbf{w})\|_\star \,\mathsf{dist}(\mathbf{w}, \arg\min f) \le f(\mathbf{w}) - \min f, \tag{13}$$

where $\|\partial f(\mathbf{w})\|_\star$ is the infimal norm of subgradients of f at $\mathbf{w}$, defined by

$$\|\partial f(\mathbf{w})\|_\star := \mathsf{dist}(\mathbf{0}, \partial f(\mathbf{w})) =: \inf\{\|\mathbf{x}\|_\star \mid \mathbf{x} \in \partial f(\mathbf{w})\}.$$

The quantity $\|\partial f(\mathbf{w})\|_\star$ is nothing but the *strong slope* (see De Giorgi et al. (1980) for definition) of the function f at the point $\mathbf{w}$. The concept of strong slope is definable for general nonconvex functions. In the case of convex functions functions, it is known to coincide with the definition of $\|\partial f(\mathbf{w})\|_\star$. There is a rich history surrounding results in the spirit of Proposition 4.1 above. For example, it has deep connections with the *Kurdika-Łojasiewicz theory* that has been very successful in analyzing the convergence of gradient-descent and proximal schemes in very general spaces (e.g *Asplund* spaces). The interested reader might want to consult Corvellec and Motreanu (2007); Azé and Corvellec (2017); Bolte and Blanchet (2016) and there references therewithin.

Proposition 4.1 has many interesting consequences. In particular, one immediately deduces that if C is a nonempty subset of M such that $\|\partial f(\mathbf{w})\|_\star \geq \alpha$ for every $\mathbf{w} \in C$ and for some $\alpha > 0$, then

$$\mathsf{dist}(\mathbf{w}, \arg\min f) \leq (f(\mathbf{w}) - \min f)/\alpha, \ \forall \mathbf{w} \in C. \tag{14}$$

Idea. Most importantly for our purposes, results of the form (14) can used to convert rates of convergence of function values $f(\mathbf{w}(t)) \to \min f$ produced by any algorithm (e.g gradient-descent / flow), to rates of convergence of iterates $\mathbf{w}(t) \to \mathbf{w}_{\text{opt}}$, i.e $\mathsf{dist}(\mathbf{w}(t), \arg\min f) \to 0$.

We conclude this detour with the following result which is one of our main contributions. Its a nontrivial corollary to Proposition 4.1, which –in combination with (14)– will play an important role in our proof of Theorem 2.1. Viz

Theorem 4.1 (Error analysis for piecewise-linear functions)**.** *Let $\mathbf{a}_1, \ldots, \mathbf{a}_n \in M^\star$ and consider the concave function $g : M \to \mathbb{R} \cup \{-\infty\}$ defined by*

$$g(\mathbf{w}) := \begin{cases} \min_{i \in [n]} \langle \mathbf{a}_i, \mathbf{w} \rangle, & \textit{if } \|\mathbf{w}\| \leq 1, \\ -\infty, & \textit{else.} \end{cases}$$

Suppose $\arg\max g \neq \emptyset$ *and* $\gamma := \max g > 0$*. Then, for every* $\mathbf{w} \in M$ *with* $\|\mathbf{w}\| \leq 1$*, it holds that*

$$\mathsf{dist}(\mathbf{w}, \arg\max g) \leq 1 - \frac{g(\mathbf{w})}{\gamma}. \tag{15}$$

Remark 4.1. *Note that the assumption* $\arg\max g \neq \emptyset$ *is only needed in case of infinite-dimensional spaces (indeed, in finite-dimensions, closed balls are compact, plus, it is an elementary fact that a continuous function on a compact set attains its maximum on the that set).*

The proof of Theorem 4.1 is instructive and is provided in the main manuscript. To this end, we will need the following lemma is a direct consequence of the well-known Danskin-Bertsekas theorem for subdifferentials (Bertsekas, 1971, Proposition A.22). See appendix for a proof.

Lemma 4.1 (Subdifferential of maximum of n functions)**.** *Let $f_1, \ldots, f_n : M \to \mathbb{R} \cup \{\infty\}$ be convex functions and let f be their pointwise maximum. Then for any $\mathbf{w} \in M$, we have the subdifferential identity $\partial f(\mathbf{w}) = \mathsf{conv}(\cup_{i \in I(\mathbf{w})} \partial f_i(\mathbf{w}))$, where $I(\mathbf{w}) := \{i \in [n] \mid f_i(\mathbf{w}) = f(\mathbf{w})\}$.*

Let $\mathbb{B}_M := \{\mathbf{w} \in M \mid \|\mathbf{w}\| \leq 1\}$ be the unit-ball in M. We will also need the elementary result which is proven in the appendix.

Lemma 4.2 (Subdifferential of indicator of closed unit-ball)**.** *Let $i_{\mathbb{B}_M} : M \to \mathbb{R} \cup \{\infty\}$ be the indicator function of the unit-ball of $\mathbb{B}_M$, defined by*

$$i_{\mathbb{B}_M}(\mathbf{w}) := \begin{cases} 0, & \textit{if } \mathbf{w} \in \mathbb{B}_M, \\ +\infty, & \textit{else.} \end{cases}$$

If $\mathbf{w} \in M$ with $\|\mathbf{w}\| < 1$, then $\partial i_{\mathbb{B}_M}(\mathbf{w}) = \{\mathbf{0}\}$.

We are now ready to proof Theorem 4.1.

Proof of Theorem 4.1. Let $f := -g$, and note that $f(\mathbf{w}) := \max_{i \in [n]} -\langle \mathbf{a}_i, \mathbf{w} \rangle$ if $\|\mathbf{w}\| \leq 1$ and $f(\mathbf{w}) = \infty$ otherwise. Now, fix $\epsilon \in (0, 1)$. For $\mathbf{w} \in \mathbb{B}_M$, let $\mathbf{w}_\epsilon := (1 - \epsilon)\mathbf{w}$. The proof is broken into several steps.

Step 1: Bounding the strong slope $\|\partial f(\mathbf{w}_\epsilon)\|_\star$. Invoking the a special case of the "subdifferential sum-rule" (Verona and Verona, 2001, Lemma 2) and the definition of f, we have $\partial f(\mathbf{w}) =$

$\partial(\max_{i\in[n]} -\mathbf{a}_i)(\mathbf{w}_\epsilon) + \partial i_{\mathbb{B}_M}(\mathbf{w}_\epsilon)$. But because $\|\mathbf{w}_\epsilon\| = 1-\epsilon < 1$, we have $\partial i_{\mathbb{B}_M}(\mathbf{w}_\epsilon) = \{\mathbf{0}\}$ by Lemma 4.2. Thus $\partial f(\mathbf{w}) = \partial(\max_{i\in[n]} -\mathbf{a}_i)(\mathbf{w}_\epsilon)$, and invoking Lemma 4.1 then gives

$$\partial f(\mathbf{w}_\epsilon) = \partial(\min_i(-\mathbf{a}_i))(\mathbf{w}_\epsilon) = \mathsf{conv}(\{-\nabla \mathbf{a}_i(\mathbf{w}_\epsilon) \mid i \in I(\mathbf{w}_\epsilon)\}\} = \mathsf{conv}(\{-\mathbf{a}_i \mid i \in I(\mathbf{w}_\epsilon)\},$$

where $I(\mathbf{w}_\epsilon) := \{i \in [n] \mid \langle -\mathbf{a}_i, \mathbf{w}_\epsilon\rangle = f(\mathbf{w}_\epsilon)\}$. Putting things together, one then computes

$$\begin{aligned}\|\partial f(\mathbf{w}_\epsilon)\|_\star &:= \inf\{\|\mathbf{x}\|_\star \mid \mathbf{x} \in \partial f(\mathbf{w}_\epsilon)\} = \inf\{\|\mathbf{x}\|_\star \mid \mathbf{x} \in \mathsf{conv}(\{-\mathbf{a}_i \mid i \in I(\mathbf{w}_\epsilon)\}\} \\ &\geq \inf\{\|\mathbf{x}\|_\star \mid \mathbf{x} \in \mathsf{conv}(\{-\mathbf{a}_i \mid i \in [n]\}\},\ \text{since infimum on larger set is smaller} \\ &= \min_{\mathbf{q}\in\Delta_{n-1}} \|\sum_{i=1}^n q_i \mathbf{a}_i\|_\star = \min_{\mathbf{q}\in\Delta_{n-1}} \max_{\|\mathbf{w}'\|\leq 1} \sum_{i=1}^n q_i \langle \mathbf{a}_i, \mathbf{w}'\rangle,\ \text{by duality} \\ &\geq \max_{\|\mathbf{w}\|\leq 1} \min_{\mathbf{q}\in\Delta_{n-1}} \sum_{i=1}^n q_i \langle \mathbf{a}_i, \mathbf{w}'\rangle,\ \text{because } \min\max \geq \max\min \\ &= \max_{\|\mathbf{w}'\|\leq 1} \min_{i\in[n]} \langle \mathbf{a}_i, \mathbf{w}'\rangle = \max_{\|\mathbf{w}'\|\leq 1} g(\mathbf{w}') =: \max g =: \gamma > 0,\ \text{by definition of } g.\end{aligned}$$

Thus $\|\partial f(\mathbf{w}_\epsilon)\|_\star \geq \gamma := \max g > 0$.

Step 2: Bounding distance to maximizers. Using Proposition 4.1 with $f = -g$ then gives

$$\begin{aligned}\mathsf{dist}(\mathbf{w}, \arg\max g) &\leq \|\mathbf{w} - \mathbf{w}_\epsilon\| + \mathsf{dist}(\mathbf{w}_\epsilon, \arg\max g),\ \text{by triangle inequality} \\ &= \epsilon + \mathsf{dist}(\mathbf{w}_\epsilon, \arg\gamma),\ \text{by definition of } \mathbf{w}_\epsilon \\ &\leq \epsilon + 1 - \frac{g(\mathbf{w}_\epsilon)}{\gamma},\ \text{by Proposition 4.1 and the fact that } \|\mathbf{w}_\epsilon\| = 1-\epsilon < 1 \\ &\leq \epsilon + 1 - \frac{g(\mathbf{w}) - \epsilon \max_{i\in[n]} \|\mathbf{a}_i\|_\star}{\gamma},\ \text{because } g \text{ is } (\max_{i\in[n]} \|\mathbf{a}_i\|_\star)\text{-Lipschitz} \\ &= 1 - \frac{g(\mathbf{w})}{\gamma} + \epsilon(1 + \frac{\max_{i\in[n]} \|\mathbf{a}_i\|_\star}{\gamma}).\end{aligned}$$

Finally, letting $\epsilon \to 0^+$ then gives $\mathsf{dist}(\mathbf{w}, \arg\max g) \leq 1 - g(\mathbf{w})/\gamma$ as claimed. □

4.2 Proof of main results, namrly Theorems 2.1, 2.2, and 2.3

For proving the first part of Theorem 2.1, we will need the following elementary lemma proven in the appendix (see also Ji and Telgarsky (2019) for the finite-dimensional case), which gives a dual formulation of the optimal margin $\overline{\gamma}_n$.

Symbol	*Meaning*	*Definition / Comment*
$\mathbf{w}$	Parameters of a linear model	Element of $\mathbb{R}^m$
γ_n	Margin function	$\min_{i\in[n]} y_i\langle \mathbf{x}_i, \mathbf{w}\rangle$
$\overline{\gamma_n}$	Maximum possible value of margin $\gamma_n(\mathbf{w})$	$\max_{\|\mathbf{w}\|\leq 1} \gamma_n(\mathbf{w})$
OPT_n	Set of all max-margin (optimal) models	$\{\mathbf{w} \in \mathbb{B}_m \mid \gamma_n(\mathbf{w}) = \overline{\gamma}_n\}$
F_n	Empirical risk	defined in (3)
$\mathbf{w}_n(t)$	Model after t iterations of GF / GD on F_n	
$\widetilde{\mathbf{w}}_n(t)$	Normalized version $\mathbf{w}_n(t)$	$\widetilde{\mathbf{w}}_n(t) := \mathbf{w}_n(t)/\|\mathbf{w}_n(t)\| \in \mathbb{B}_m$

Table 2: Summary of key notations for margin / bias analysis, recalled here for convenience.

Lemma 4.3. *For every* $\mathbf{q} = (q_1, \ldots, q_n) \in \Delta_{n-1}$*, it holds that* $\overline{\gamma}_n \leq \|\sum_{i=1}^n q_i y_i \mathbf{x}_i\| \leq 1$.

We are now ready to give a full detailed proof our main results.

Proof of Theorem 2.1. For ease of exposition, the prove is broken down into several steps.

Step 1: Uniform lower- and upper-bounds on gradient of loss. Mindful of the definition (3) of the soft-margin / exponential risk $F_n(\mathbf{w})$ of a linear model $\mathbf{w} \in \mathbb{R}^m$, one computes $\nabla F_n(\mathbf{w}) =$

$\mathbb{E}_{i\sim\widehat{\mathbf{q}}_{n,\beta}(\mathbf{w})}[-y_i\mathbf{x}_i] = -\sum_{i=1}^n \widehat{q}_i y_i \mathbf{x}_i$, where $(\widehat{q}_1,\ldots,\widehat{q}_n) := \widehat{\mathbf{q}}_{n,\beta}$. Invoking Lemma 4.3 with $\mathbf{q} = \widehat{\mathbf{q}}_{n,\beta}$ then gives

$$\overline{\gamma}_n \le \|\nabla F_n(\mathbf{w})\| \le 1,\ \forall \mathbf{w}\in\mathbb{R}^m. \tag{16}$$

Step 2: Lower-bounding the margin. We now prove the first part of the theorem. This step is elementary and well-known (see Bach (2020), for example). So, let $t\mapsto \mathbf{w}_n(t)$ be a solution of the gradient-flow (5), so that $\mathbf{w}_n(t) = \int_0^t \nabla F(\mathbf{w}_n(\tau))d\tau$ for all $t\ge 0$. We note the following facts

- **Fact A:** $\|\mathbf{w}_n(t)\| = \|\int_0^t \nabla F_n(\mathbf{w}_n(\tau))d\tau\| \le \int_0^t \|\nabla F_n(\mathbf{w}_n(\tau))\|d\tau$, by triangle inequality.
- **Fact B:** $\int_0^t \|\nabla F_n(\mathbf{w}_n(\tau))\|d\tau \ge \overline{\gamma}_n t$, which follows from (16).

Now, one computes $\frac{d}{dt}F_n(\mathbf{w}(t)) = \langle \nabla F_n(\mathbf{w}_n(t)), \frac{d}{dt}\mathbf{w}_n(t)\rangle = \|\nabla F_n(\mathbf{w}_n(t))\|^2$, from which

$$\begin{aligned}\frac{\gamma_n(\mathbf{w}_n(t))}{\|\mathbf{w}_n(t)\|} &\ge \frac{F_n(\mathbf{w}_n(t)) - \log n}{\|\mathbf{w}_n(t)\|},\ \text{because } F_n(\mathbf{w}_n(t)) \ge \gamma_n(\mathbf{w}_n(t)) - \log n\\ &= \frac{-\log n + \int_0^t \|\nabla F_n(\mathbf{w}_n(\tau))\|^2 d\tau}{\|\mathbf{w}_n(t)\|},\ \text{by integrating } \frac{d}{dt}F(\mathbf{w}(t)) \text{ w.r.t } t\\ &= \frac{-\log n + \int_0^t \|\nabla F_n(\mathbf{w}_n(\tau))\|^2 d\tau}{\int_0^t \|\nabla F_n(\mathbf{w}_n(\tau))\|d\tau},\ \text{by \textbf{Fact A}}\\ &\ge (\frac{-\log n}{\overline{\gamma}_n t} + \overline{\gamma}_n),\ \text{by (16) and \textbf{Fact B}, provided } t \ge (\log n)/\overline{\gamma}_n^2\end{aligned}$$

Thus for $t \ge (\log n)/\overline{\gamma}_n^2$, we have that

$$\gamma_n(\widetilde{\mathbf{w}}_n(t)) = \gamma_n(\frac{\mathbf{w}_n(t)}{\|\mathbf{w}_n(t)\|}) = \frac{\gamma_n(\mathbf{w}_n(t))}{\|\mathbf{w}_n(t)\|} \ge \overline{\gamma}_n - \frac{\log n}{\overline{\gamma}_n t}, \tag{17}$$

from which the first part of the theorem follows.

Step 2: Bounding distance to set of minimizers. Consider the function $g : \mathbb{R}^m \to \mathbb{R}\cup\{\infty\}$, define by $g(\mathbf{w}) := \min_{i\in[n]}\langle \mathbf{a}_i, \mathbf{w}\rangle$ if $\|\mathbf{w}\|\le 1$ and $f(\mathbf{w}) = \infty$ else, where $\mathbf{a}_i = y_i\mathbf{x}_i$ for all $i\in[n]$. Note that $\arg\max g = \mathsf{OPT}_n$ and $\max g = \overline{\gamma}_n$. Now, invoke Theorem 4.1 with $M = (\mathbb{R}^m, \|\cdot\|_2)$ to get

$$\mathsf{dist}\,(\widetilde{\mathbf{w}}_n(t), \mathsf{OPT}_n) \le 1 - \frac{\gamma_n(\widetilde{\mathbf{w}}_n(t))}{\overline{\gamma}_n}. \tag{18}$$

Combining (12) and (18) then gives

$$\mathsf{dist}\,(\widetilde{\mathbf{w}}_n(t), \mathsf{OPT}_n) \le \frac{\log n}{\overline{\gamma}_n^2 t},$$

which concludes the prove of the theorem. □

Proof of Theorem 2.2. As explained in the proof sketch at the end of section 2.1, the prove of Theorem 2.2 follows from a time-discretization of the proof of Theorem 2.1, with stepsize $\eta_t = \mathcal{O}(1/\sqrt{t+1})$, where $t\in\mathbb{N}$ is the iteration counter. □

Proof of Theorem 2.3. We combine the *(A)* $\mathcal{O}(\log(n)/t)$ margin bound from (Ji and Telgarsky, 2020, Theorem 4.2) which establishes a margin rate $\gamma_n(\widetilde{\mathbf{w}}_n(t)) \ge \overline{\gamma}_n - \log(n)/(\overline{\gamma}_n t)$ via gradient-descent on the empirical risk $\mathcal{R}_n(\mathbf{w}) := \ell_{\exp}^{-1}(-F_n(\mathbf{w})) = (1/n)\sum_{i=1}^n \ell_{\exp}(-y_i\langle \mathbf{x}_i, \mathbf{w}\rangle)$ with aggressive stepsizes $\eta_t = \Theta(1/\mathcal{R}_n(\mathbf{w}_n(t)))$, and *(B)* our inequality (18) above, to get the bound

$$\mathsf{dist}\,(\widetilde{\mathbf{w}}_n(t), \mathsf{OPT}_n) = \mathcal{O}(\frac{\log n}{\overline{\gamma}_n^2 t}),$$

as claimed. □

5 Concluding remarks

Using techniques from gradient-flow and nonsmooth error analysis, we have derived improved bounds on the convergence of the margin and the bias of gradient-descent on on linear classification problems with so-called exponential-tailed loss functions. Our method for obtaining the convergence of bias rate is new and should be applicable to a vast array of scenarios: it allows the transfer of convergence rates on the margin, to convergence rates on the bias, irrespective of the algorithms / constructs (gradient-flow, gradient-descent, what stepsize, etc.) used to establish the former, and does not require any uniqueness assumption on the max-margin solution.

Future work. In a followup paper, we plan to extend our analysis to the more complicated scenario of nonlinear neural networks. We would also like to get an analogue of our analysis in the discrete-time setting (i.e gradient-descent), as in Ji and Telgarsky (2020) for example. *Maybe the aggressive gradient-descent stepsizes proposed in Ji and Telgarsky (2020) are just a super efficient discretization of the gradient-flow approach we used here ?*

Acknowlegement. We are thankful to the authors of Ji and Telgarsky (2019) for a private correspondence in which they explained to us certain details of their contribution (after their manuscript was public).

References

Azé, D. and Corvellec, J.-N. (2017). Nonlinear error bounds via a change of function. *Journal of Optimization Theory and Applications*, 172.

Bach, F. (2020). Generalization and implicit bias. `https://francisbach.com/gradient-descent-for-wide-two-layer-neural-networks-implicit-bias/`.

Bertsekas, D. P. (1971). Control of Uncertain Systems with a Set-Membership Description of Uncertainty.

Bolte, J. and Blanchet, A. (2016). A family of functional inequalities: Lojasiewicz inequalities and displacement convex functions. *arXiv e-prints*.

Chizat, L. and Bach, F. (2020). Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*. PMLR.

Corvellec, J.-N. and Motreanu, V. V. (2007). Nonlinear error bounds for lower semicontinuous functions on metric spaces. *Mathematical Programming*, 114(2):291.

De Giorgi, E., Marino, A., and Tosques, M. (1980). Problemi di evoluzione in spazi metrici e curve di massima pendenza. *Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti*, 68(3):180–187.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. (2018). Characterizing Implicit Bias in Terms of Optimization Geometry. *arXiv e-prints*, page arXiv:1802.08246.

Ji, Z. and Telgarsky, M. (2019). A refined primal-dual analysis of the implicit bias. *arXiv:1906.04540 (version v1 of manuscript)*.

Ji, Z. and Telgarsky, M. (2020). Characterizing the implicit bias via a primal-dual analysis. *arXiv:1906.04540 (version v2 of manuscript)*.

Lyu, K. and Li, J. (2020). Gradient descent maximizes the margin of homogeneous neural networks. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net.

Nacson, M., Lee, J. D., Gunasekar, S., Savarese, P. H. P., Srebro, N., and Soudry, D. (2018). Convergence of Gradient Descent on Separable Data. *arXiv e-prints*, page arXiv:1803.01905.

Rockafellar, R. T. (1970). *Convex analysis*. Princeton Mathematical Series. Princeton University Press.

Soudry, D., Hoffer, E., Shpigel Nacson, M., Gunasekar, S., and Srebro, N. (2017). The Implicit Bias of Gradient Descent on Separable Data. *arXiv e-prints*, page arXiv:1710.10345.

Verona, A. and Verona, M. E. (2001). A simple proof of the sum formula. *Bulletin of the Australian Mathematical Society*, 63.

A Omitted technical proofs

Lemma 4.1 (Subdifferential of maximum of n functions)**.** *Let $f_1, \ldots, f_n : M \to \mathbb{R} \cup \{\infty\}$ be convex functions and let f be their pointwise maximum. Then for any $\mathbf{w} \in M$, we have the subdifferential identity $\partial f(\mathbf{w}) = \mathsf{conv}(\cup_{i \in I(\mathbf{w})} \partial f_i(\mathbf{w}))$, where $I(\mathbf{w}) := \{i \in [n] \mid f_i(\mathbf{w}) = f(\mathbf{w})\}$.*

Proof. Let $f := \max(f_1, \ldots, f_n)$. Define $\Phi : \mathbb{R}^n \times M \to (-\infty, +\infty]$ by $\Phi(\mathbf{q}, \mathbf{w}) := \sum_{i=1}^n q_i f_i(\mathbf{w})$ and note that $f(\mathbf{w}) = \sup_{\mathbf{q} \in \Delta_{n-1}} \Phi(\mathbf{q}, \mathbf{w})$ for every $\mathbf{w} \in M$. Also note that $\partial_{\mathbf{w}} \Phi(\mathbf{q}, \mathbf{w}) = \sum_{i=1}^n q_i \partial f_i(\mathbf{w})$. Now invoke (Bertsekas, 1971, Proposition A.22). □

Lemma 4.2 (Subdifferential of indicator of closed unit-ball)**.** *Let $i_{\mathbb{B}_M} : M \to \mathbb{R} \cup \{\infty\}$ be the indicator function of the unit-ball of $\mathbb{B}_M$, defined by*

$$i_{\mathbb{B}_M}(\mathbf{w}) := \begin{cases} 0, & \textit{if } \mathbf{w} \in \mathbb{B}_M, \\ +\infty, & \textit{else.} \end{cases}$$

If $\mathbf{w} \in M$ with $\|\mathbf{w}\| < 1$, then $\partial i_{\mathbb{B}_M}(\mathbf{w}) = \{\mathbf{0}\}$.

Proof. If $\mathbf{w} \in M$ with $\|\mathbf{w}\| < 1$, then direct computation we have

$$\begin{aligned} \partial i_{\mathbb{B}_M}(\mathbf{w}) &:= \{\mathbf{x} \in M^\star \mid i_{\mathbb{B}_M}(\mathbf{w}') \geq i_{\mathbb{B}_M}(\mathbf{w}) + \langle \mathbf{x}, \mathbf{w}' - \mathbf{w} \rangle \ \forall \mathbf{w}' \in M\} \\ &= \{\mathbf{x} \in M^* \mid \langle \mathbf{x}, \mathbf{w} \rangle \geq \langle \mathbf{x}, \mathbf{w}' \rangle \ \forall \mathbf{w}' \in \mathbb{B}_M\} \\ &= \{\mathbf{x} \in M^* \mid \langle \mathbf{x}, \mathbf{w} \rangle \geq \|\mathbf{x}\|_\star\} \\ &= \{\mathbf{0}\}, \end{aligned}$$

since $\langle \mathbf{x}, \mathbf{w} \rangle \leq \|\mathbf{x}\|_\star \|\mathbf{w}\| \leq \|\mathbf{x}\|_\star$ by the Cauchy-Schwarz inequality. □

Lemma 4.3. *For every $\mathbf{q} = (q_1, \ldots, q_n) \in \Delta_{n-1}$, it holds that $\overline{\gamma}_n \leq \| \sum_{i=1}^n q_i y_i \mathbf{x}_i \| \leq 1$.*

Proof. For any $\mathbf{q} = (q_1, \ldots, q_n) \in \Delta_{n-1}$, one computes

$$\begin{aligned} \| \sum_{=1}^{,} q_i y_i \mathbf{x}_i \| &= \sup_{\mathbf{u} \in \mathbb{B}_M} \langle \sum_{i=1}^n q_i y_i \mathbf{x}_i, \mathbf{u} \rangle = \sup_{\mathbf{u} \in \mathbb{B}_M} \mathbb{E}_{i \sim \mathbf{q}}[y_i \langle \mathbf{x}_i, \mathbf{u} \rangle] \\ &\geq \mathbb{E}_{i \sim \mathbf{q}}[y_i \langle \mathbf{x}_i, \mathbf{w}_n^{\text{opt}} \rangle] = \overline{\gamma}_n, \text{ by taking any } \mathbf{u} = \mathbf{w}_n^{\text{opt}} \in \mathsf{OPT}_n\,. \end{aligned}$$

This proves the lower-bound. On the other hand, using the fact that

$$\sup_{\mathbf{u} \in \mathbb{B}_M} \mathbb{E}_{i \sim \mathbf{q}}[y_i \langle \mathbf{x}_i, \mathbf{u} \rangle] \leq \mathbb{E}_{i \sim \mathbf{q}}[\sup_{\mathbf{u} \in \mathbb{B}_M} y \langle \mathbf{x}_i, \mathbf{u} \rangle] = \mathbb{E}_{i \sim \mathbf{q}}[\|\mathbf{x}\|] \leq 1,$$

since $\|\mathbf{x}_i\| \leq 1$ for all $i \in [n]$ by hypothesis. This proves the upper-bound. □