Implicit bias of any algorithm: bounding bias via margin

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

Abstract

Consider n points $\mathbf{x}_1,\ldots,\mathbf{x}_n$ in finite-dimensional euclidean space, each having one of two colors. Suppose there exists a separating hyperplane (identified with its unit normal vector $\mathbf{w}$) for the points, i.e a hyperplane such that points of same color lie on the same side of the hyperplane. We measure the quality of such a hyperplane by its margin $\gamma(\mathbf{w})$, defined as minimum distance between any of the points $\mathbf{x}_i$ and the hyperplane. In this paper, we prove that the margin function γ satisfies a nonsmooth Kurdyka-Lojasiewicz inequality with exponent $1/2$. This result has far-reaching consequences. For example, let γ^{opt} be the maximum possible margin for the problem and let $\mathbf{w}^{opt}$ be the parameter for the hyperplane which attains this value. Given any other separating hyperplane with parameter $\mathbf{w}$, let $d(\mathbf{w}) := \|\mathbf{w}-\mathbf{w}^{opt}\|$ be the euclidean distance between $\mathbf{w}$ and $\mathbf{w}^{opt}$, also called the bias of $\mathbf{w}$. From the previous KL-inequality, we deduce that $(\gamma^{opt}-\gamma(\mathbf{w}))/R \le d(\mathbf{w}) \le 2\sqrt{(\gamma^{opt}-\gamma(\mathbf{w}))/\gamma^{opt}}$, where $R := \max_i \|\mathbf{x}_i\|$ is the maximum distance of the points $\mathbf{x}_i$ from the origin. Consequently, for any optimization algorithm (gradient-descent or not), the bias of the iterates converges at least as fast as the square-root of the rate of their convergence of the margin. Thus, our work provides a generic tool for analyzing the implicit bias of any algorithm in terms of its margin, in situations where a specialized analysis might not be available: it is sufficient to establish a good rate for converge of the margin, a task which is usually much easier.

1 Introduction

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 from an unknown distribution on $\mathbb{R}^m\times\{\pm 1\}$. For each $i\in[n]$, $y_i\in\{\pm1\}$ 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 $n\ge 1$ is the sample size, while m is the dimensionality of the problem. Let $\mathbb{S}_{m-1} := \{\mathbf{w}\in\mathbb{R}^m \mid \|\mathbf{w}\|_2 = 1\}$ is the $(m-1)$-dimensional unit-sphere.

We are interested in "large margin" linear classifiers. Any such model is indexed by a unit-vector $\mathbf{w}\in\mathbb{S}_{m-1}$. The prediction on an input example $\mathbf{x}\in\mathbb{R}^m$ is $\text{sign}(\mathbf{x}^\top\mathbf{w})\in\{\pm1\}$, where $\mathbf{x}^\top\mathbf{w}$ is the inner product of $\mathbf{x}$ and $\mathbf{w}$. The margin of any $\mathbf{w}\in\mathbb{S}_{m-1}$, denoted $\gamma(\mathbf{w})$, i.e the maximum possible margin attainable by a linear classifier on the problem, defined by

$$\gamma(\mathbf{w}) := \min_{i\in[n]} y_i\mathbf{x}_i^\top\mathbf{w}. \tag{1}$$

This measures the minimum (signed) distance of the samples $\mathbf{x}_1,\ldots,\mathbf{x}_n$ to the induced hyperplane $\mathbf{w}^\perp := \{\mathbf{x}\in\mathbb{R}^m \mid \mathbf{x}^\top\mathbf{w}=0\}$. Consider the optimal / maximum margin $\gamma_{opt}\in[0,1]$ for the problem, defined by

$$\gamma_{opt} := \max_{\mathbf{w}\in\mathbb{S}_{m-1}} \gamma(\mathbf{w}) = \max_{\mathbf{w}\in\mathbb{S}_{m-1}} \min_{i\in[n]} y_i\mathbf{x}_i^\top\mathbf{w}. \tag{2}$$

Finally, let $\gamma_{opt} := \arg\max_{\mathbf{w}\in\mathbb{S}_{m-1}}\gamma(\mathbf{w}) := \{\mathbf{w}\in\mathbb{S}_{m-1} \mid \gamma(\mathbf{w}) = \gamma_{opt}\}$ be the max-margin model (unique). We will assume that the problem is (linearly) separable, meaning that $\gamma_{opt}>0$.

1.1 Summary of main contributions

Let $\gamma:\mathbb{S}_{m-1}\to\mathbb{R}$ be the margin function defined in (1) and let γ_{opt} be the optimal margin, defined in (2). Our main contributions can be summarized as follows

- With a certain nonsmooth replacement of the notion of gradient norm (namely the so-called *strong slope* De Giorgi et al. (1980)), we prove in Theorem 2.1 that the function $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ defined by

$$f(\mathbf{w}) := \begin{cases} -\gamma(\mathbf{w}), & \text{if } \mathbf{w} \in \mathbb{S}_{m-1}, \\ +\infty, & \text{else.} \end{cases}$$

  satisfies a *Kurdyka-Lojasiewicz inequality* with exponent 1/2, on the unit-sphere $\mathbb{S}_{m-1}$. A highlight of this result is that it hints on the possibility of the existence of very fast (perhaps quasi-linear time) algorithms for finding the max-margin model on separable data. These algorithms need not necessarily be gradient-descent in the usual sense.

- For our second contribution, we prove in Theorem 2.2 that that

$$\frac{1}{R} \cdot \left(1 - \frac{\gamma(\mathbf{w})}{\gamma_{opt}}\right) \le \|\mathbf{w} - \mathbf{w}_{opt}\| \le 2 \cdot \left(1 - \frac{\gamma(\mathbf{w})}{\gamma_{opt}}\right)^{1/2}. \tag{3}$$

  where $R = \max_i \|\mathbf{x}_i\|$. This is graphically illustrated in Figure 1. Of course, the LHS side is trivial since gamma is Lipschitz w.r.t w. Consequently, for any optimization algorithm (gradient-descent or not), the *bias* $\|\mathbf{w}(t) - \mathbf{w}_{opt}\|$ of the iterates $\mathbf{w}(t)$ converges at least as fast as the square-root of the rate of their convergence of the margin (deficit of) $\gamma_{opt} - \gamma(\mathbf{w}(t))$. Thus, our work provides a generic tool for analyzing the implicit bias of any algorithm in terms of its margin. This can be especially useful in situations where a specialized analysis might not be available: it is sufficient to establish a good rate for convergence of the margin, a task which is usually much easier

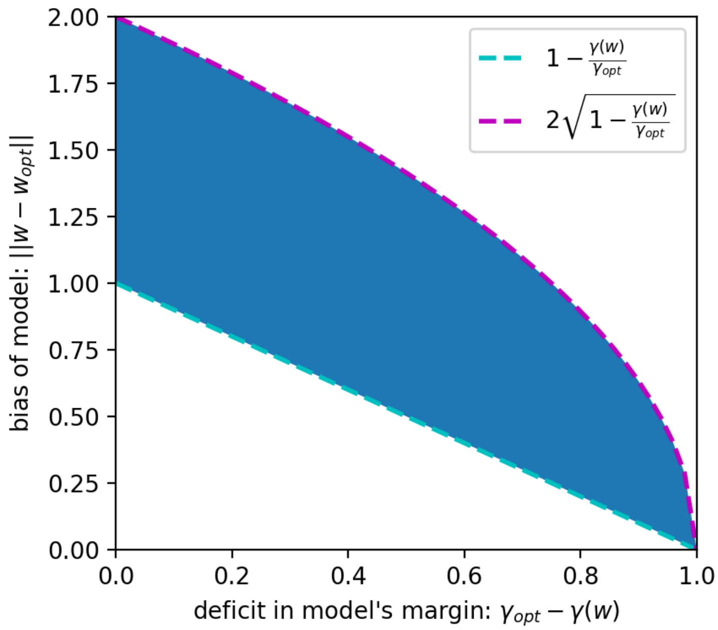

Figure 1: Graphical illustration 3. The result, developed in Theorem 2.1, states that the plots of the margin deficit versus the bias of the iterates generated by any algorithm must lie in within the shaded area. This allows one to transfer convergence rates for the margin to of iterates generated by any optimization algorithm, to convergence rates for the bias and vise versa.

1.2 Related works

There is rich history of research trying to understand the limiting dynamics of the iterates generated by gradient-descent, its so-called *implicit bias*. In the case of linear models with exponential-tailed losses, Soudry et al. (2017), Nacson et al. (2018), Gunasekar et al. (2018), Ji and Telgarsky

(2019, 2020) are part of the standard literature. They all prove that iterates of gradient-descent on linearly separable binary classification problems converge to the max-margin linear classifier $\mathbf{w}_{opt}$ with margin γ_{opt}. These papers also contain explicit rates of convergence. The very recent work Ji and Telgarsky (2020) establishes a convergence rate of $\mathcal{O}(1/t)$ for both the margin the bias of gradient-descent. (Theorem 2.2). 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

$$\max(\|\widetilde{\mathbf{w}}(t) - \mathbf{w}_{opt}\|, \gamma_{opt} - \gamma_n(\widetilde{\mathbf{w}}(t))) = \mathcal{O}(\frac{1}{\sum_{j=1}^{t} \widehat{\eta}_j}),$$

where $\widetilde{\mathbf{w}}(t) := \mathbf{w}(t)/\|\mathbf{w}(t)\| \in \mathbb{S}_{m-1}$ is the normalized sequence of iterates and $\widehat{\eta}_t := \eta_t \mathcal{R}(\mathbf{w}(t))$, with $\mathcal{R}(\mathbf{w}(t))$ being the empirical risk evaluated at $\mathbf{w}(t)$. 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 $\mathcal{O}(1/t)$. These convergence rates are the best known currently in the literature.

Finally, in the case of neural network classifiers, 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.

2 Main results

2.1 Preliminaries on nonsmooth error bounds

Central to our paper will be the notations of *strong slope* De Giorgi et al. (1980) and generalized nonsmooth *Kurdyka-Lojasiewicz inequalities* Corvellec and Motreanu (2007); Azé and Corvellec (2017); Bolte and Blanchet (2016). These concepts are now standard in optimization.

Definition 2.1 (Nonsmooth Kurdyka-Lojasiewicz inequalites via strong slope)**.** *Let $M = (M, d)$ be a complete metric space. An extended-value function $f : M \to \mathbb{R} \cup \{\pm\infty\}$ is said to satisfy a generalized Kurdyka-Lojasiewicz inequality of order $\theta > 0$ around the point $\mathbf{w}_0 \in M$ if there exists $\sigma > 0$, $\nu \in \mathbb{R}$, and some neigborhood $U \subseteq M$ of $\mathbf{w}_0$ such that*

$$|\partial^-|f(\mathbf{w}) \ge \frac{\sigma}{\theta}(f(\mathbf{w}) - f(\mathbf{w}_0))^{1-\theta} \ \forall \mathbf{w} \in U \text{ with } f(\mathbf{w}_0) < f(\mathbf{w}) < f(\mathbf{w}_0) + \nu. \tag{4}$$

Here, $|\partial^-|f(\mathbf{w}) \in [0, +\infty]$ is the strong slope De Giorgi et al. (1980); Corvellec and Motreanu (2007); Azé and Corvellec (2017) of f at $\mathbf{w}_0$, a "synthetic" lower-bound of the the rate of change of f at $\mathbf{w}_0$, in any direction, defined by

$$|\partial^-|f(\mathbf{w}) := \limsup_{\mathbf{w}' \to \mathbf{w}} \frac{(f(\mathbf{w}) - f(\mathbf{w}'))_+}{d(\mathbf{w}, \mathbf{w}_0)}. \tag{5}$$

Strong slopes are difficult to compute in general. Fortunately, they can be bounded in terms of more familiar quantities. For example, if $M = (M, \|\cdot\|)$ is a Banach space with topological dual $M^\star = (M^\star, \|\cdot\|_\star)$ and $\partial f(\mathbf{w}) \subseteq M^\star$ is the is the *Fréchet subdifferential* of f at $\mathbf{w}$, defined by

$$\partial f(\mathbf{w}) := \left\{ \mathbf{w}^\star \in M^\star \mid \liminf_{\mathbf{w}' \to \mathbf{w},\ \mathbf{w}' \ne \mathbf{w}} \frac{f(\mathbf{w}') - f(\mathbf{w}) - \langle \mathbf{w}^\star, \mathbf{w}' - \mathbf{w}\rangle}{\|\mathbf{w}' - \mathbf{w}\|} \ge 0 \right\},$$

then the following bounds hold

$$\liminf_{(\mathbf{w}', f(\mathbf{w}')) \to (\mathbf{w}, f(\mathbf{w}))} \|\partial f(\mathbf{w}')\| \le |\partial^-|f(\mathbf{w}) \le \|\partial f(\mathbf{w})\|, \tag{6}$$

where $\|\partial f(\mathbf{w})\|$ is the minimum norm of subgradients of F at $\mathbf{w}$, i.e $\|\partial f(\mathbf{w})\| := \inf\{\|\mathbf{w}^\star\| \mid \mathbf{w}^\star \in \partial f\}$. See Azé and Corvellec (2017), for example. In particular, if f is convex, then the second inequality in (6) is an equality. In this case, ∂f is given by the familiar formula $\partial f(\mathbf{w}) := \{\mathbf{w}^\star \in \mathbb{R}^m \mid f(\mathbf{w}') \ge f(\mathbf{w}) + \langle \mathbf{w}^\star, \mathbf{w}' - \mathbf{w}\rangle \ \forall \mathbf{w}' \in M\}$. Furthermore, if $\mathbf{w}$ is not not a local minmum point of f, then both inequalities in (6) are equalities.

Our interest in strong slopes is motivated by the following fundamental result, which will be the main workhorse in proving our theorems.

Proposition 2.1 (Nonlinear error-bound via Kurdyka-Lojasiewicz)**.** *Let M be a complete metric space and $f : M \to \mathbb{R} \cup \{+\infty\}$ be a proper l.s.c function which satisfies a KL-inequality around $\mathbf{w}_0$ with exponent $\theta > 0$ and modulus $\alpha > 0$. Then, if $a := f(\mathbf{w})$, we have the error-bound*

$$\mathsf{dist}(\mathbf{w}, \{f \le a\}) \ge \frac{(f(\mathbf{w}) - a)^\theta}{\alpha},\ \forall \mathbf{w} \in M \text{ with } f(\mathbf{w}) > a. \tag{7}$$

2.2 Statement of main results

The following is the first of our main results. Proofs of all our theorems will be provided in section 3.

Theorem 2.1 (Kurdyka-Lojasiewicz inequality for margin function)**.** *The negative marging function defined by $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ defined by*

$$f(\mathbf{w}) = \begin{cases} -\gamma(\mathbf{w}), & \text{if } \mathbf{w} \in \mathbb{S}_{m-1}, \\ +\infty, & \text{else.} \end{cases}$$

satisfies a KL-inequality (4) *with exponent $\theta = 1/2$.*

Note that the (negative) margin function $-\gamma$ which is the subject of the above theorem is neither smooth nor convex. For our second main contribution, we have the following result.

Theorem 2.2 (Bias bounds from margin bounds)**.** *For every unit-vector $\mathbf{w} \in \mathbb{S}_{m-1}$, we have*

$$\frac{1}{R}\left(1 - \frac{\gamma(\mathbf{w})}{\gamma_{opt}}\right) \le \|\mathbf{w} - \mathbf{w}_{opt}\| \le 2\left(1 - \frac{\gamma(\mathbf{w})}{\gamma_{opt}}\right)^{1/2}, \tag{8}$$

where $R := \max_{i \in [n]} \|\mathbf{x}_i\|$.

This theorem, illustrated in Figure can be used to convert rates of convergence of function values $\gamma(\mathbf{w}(t)) \to \gamma_{opt}$ produced by any algorithm (e.g gradient descent), to rates of convergence of iterates, i.e $\|\mathbf{w}(t) - \mathbf{w}_{opt}\| \to 0$.

3 Proof of main results

In this section, we will prove our main results, namely Theorem 2.2 and 2.1. Before that, we need some auxiliary results which might be of independent interest themselves. Let $\Delta_{n-1} := \{(q_1, \ldots, q_n) \in \mathbb{R}^n \mid \sum_{i=1}^n q_i = 1, \min_{i \in [n]} q_i \ge 0\}$ be the unit $(n-1)$-dimensional probability simplex. Given a subset $I \subseteq [n]$ of indices, let $\Delta_{n-1}(I) := \{\mathbf{q} \in \Delta_{n-1} \mid \sum_{i \in I} q_i = 1\}$ be the face of Δ_{n-1} generated by vertices in I.

Theorem 3.1 (Fréchet subdifferential of negative margin function)**.** *The extended-value function f in Theorem 2.1 has Fréchet subdifferential which given for any $\mathbf{w} \in \mathbb{S}_{m-1}$, by*

$$\partial f(\mathbf{w}) = \{b\mathbf{w} - \textstyle\sum_{i=1}^n q_i y_i \mathbf{x}_i \mid \mathbf{q} = (q_1, \ldots, q_n) \in \Delta_{n-1}(I(\mathbf{w})),\ b \in \mathbb{R}\},$$

where $I(\mathbf{w}) := \{i \in [n] \mid y_i \mathbf{x}_i^\top \mathbf{w} = \gamma(\mathbf{w})\}$ is the set of indices of "support vectors" for $\mathbf{w}$.

Proof of Theorem 3.1. Let $\mathbf{A}$ be the $n \times m$ matrix whose i row is $\mathbf{a}_i := -y_i \mathbf{x}_i$, and bserve that we can decompose $f = g + i_{\mathbb{S}_{m-1}}$ where $f : \mathbb{R}^m \to \mathbb{R}$ is defined by $g(\mathbf{w}) := \max_{i \in [n]} g_i(\mathbf{w})$, with $g_i(\mathbf{w}) := \mathbf{a}_i^\top \mathbf{w}$, and $i_{\mathbb{S}_{m-1}} : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is the indicator function of $\mathbb{S}_{m-1}$ defined by $i_{\mathbb{S}_{m-1}}(\mathbf{w}) = 0$ if $\mathbf{w} \in \mathbb{S}_{m-1}$, and $i_{\mathbb{S}_{m-1}} = +\infty$ else. By the sum rule for Fréchet subdifferentials, we have $\partial g(\mathbf{w}) \subseteq \partial f(\mathbf{w}) + \partial i_{\mathbb{S}_{m-1}}(\mathbf{w})$ for all $\mathbf{w} \in \mathbb{S}_{m-1}$.

Now, by a well-known result for the subdifferential of a maximum of convex functions (e.g see (Van Ngai et al., 2002, Corollary 3.6)), one has

$$\begin{aligned}\partial g(\mathbf{w}) &= \{\textstyle\sum_{i=1}^n q_i \mathbf{w}_i^\star \mid \mathbf{q} \in \Delta_{n-1}(I(\mathbf{w})),\ \mathbf{w}_i^\star \in \partial f_i(\mathbf{w})\ \forall i \in [n]\} \\ &= \{-\textstyle\sum_{i=1}^n q_i y_i \mathbf{w}_i \mid \mathbf{q} \in \Delta_{n-1}(I(\mathbf{w}))\} = \{\mathbf{A}^\top \mathbf{q} \mid \mathbf{q} \in \Delta_{n-1}(I(\mathbf{w}))\},\end{aligned}$$

for all $\mathbf{w} \in \mathbb{R}^m$. On the other hand, by Example 2.6 of Bauschke et al. (2013), it holds $\mathbf{w} \in \mathbb{S}_{m-1}$ that $\partial i_{\mathbb{S}_{m-1}}(\mathbf{w}) = \mathbf{w}\mathbb{R} := \{b\mathbf{w} \mid b \in \mathbb{R}\}$. Putting things together then gives the result. $\square$

For the proof of Theorem 2.2, we will also need the following elementary result (also see Ji and Telgarsky (2019))

Lemma 3.1. *For every* $\mathbf{q} = (q_1, \ldots, q_n) \in \Delta_{n-1}$, *it holds that* $\overline{\gamma}_n \le \| \sum_{i=1}^n q_i y_i \mathbf{x}_i \| \le 1$.

We are now ready to proof Theorem 2.2.

Proof of Theorem 2.2. Let $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be the negative margin function appearing in the theorem. Thanks to Theorem 3.1, we know that $\partial f(\mathbf{w}) \subseteq \{\mathbf{A}^\top \mathbf{q} + b\mathbf{w} \mid \mathbf{q} \in \Delta_{n-1}(I(\mathbf{w})),\ b \in \mathbb{R}\}$ for all $\mathbf{w} \in \mathbb{S}_{m-1}$. Thus, we may lower-bound the minimum norm of Fréchet subgradients of g at any point $\mathbf{w} \in \mathbb{S}_{m-1}$ as follows

$$
\begin{aligned}
\|\partial f(\mathbf{w})\|^2 &= \inf_{\mathbf{w}^\star \in \partial f(\mathbf{w})} \|\mathbf{w}^\star\|^2 \ge \inf_{\mathbf{q} \in \Delta_{n-1}(I(\mathbf{w})),\, b \in \mathbb{R}} \|\mathbf{A}^\top \mathbf{q} + b\mathbf{w}\|^2, \text{ by Theorem 3.1} \\
&= \inf_{\mathbf{q} \in \Delta_{n-1}(I(\mathbf{w}))} \| \mathsf{P}_{\mathbf{w}^\perp}(\mathbf{A}^\top \mathbf{q})\|^2, \text{ distance between line and origin} \\
&= \inf_{\mathbf{q} \in \Delta_{n-1}(I(\mathbf{w}))} \|\mathbf{A}^\top \mathbf{q} - (\mathbf{w}^\top \mathbf{A}^\top \mathbf{q})\mathbf{w}\|^2, \text{ orthogonal projection formula} \\
&= \inf_{\mathbf{q} \in \Delta_{n-1}(I(\mathbf{w}))} \|\mathbf{A}^\top \mathbf{q}\|^2 - (\mathbf{w}^\top \mathbf{A}^\top \mathbf{q})^2, \text{ basic linear algebra} \\
&= \inf_{\mathbf{q} \in \Delta_{n-1}(I(\mathbf{w}))} \|\mathbf{A}^\top \mathbf{q}\|^2 - \gamma(\mathbf{w})^2, \text{ because } \mathbf{a}_i^\top \mathbf{w} = \gamma(\mathbf{w})\ \forall \imath \in I(\mathbf{w}) \\
&\ge \inf_{\mathbf{q} \in \Delta_{n-1}} \|\mathbf{A}^\top \mathbf{q}\|^2 - \gamma(\mathbf{w})^2, \text{ since } \Delta_{n-1}(I(\mathbf{w})) \subseteq \Delta_{n-1} \\
&= \gamma_{opt}^2 - \gamma(\mathbf{w})^2, \text{ by Lemma 3.1} \\
&= (\gamma(\mathbf{w}) + \gamma_{opt}) \cdot (-\gamma(\mathbf{w}) + \gamma_{opt}) \\
&= (\gamma(\mathbf{w}) + \gamma_{opt}) \cdot (f(\mathbf{w}) - \min f), \text{ by definition of } f \\
&\ge \gamma_{opt} \cdot (f(\mathbf{w}) - \min f), \text{ since } \gamma(\mathbf{w}) \ge 0 \text{ by assumption.}
\end{aligned}
$$

Combining the above inequality with the RHS of (6) then gives

$$
|\partial^-| f(\mathbf{w}) \ge \liminf_{(\mathbf{w}', f(\mathbf{w}')) \to (\mathbf{w}, f(\mathbf{w}))} \|\partial f(\mathbf{w}')\| \ge \gamma_{opt}^{1/2} \cdot (f(\mathbf{w}) - \min f)^{1/2}.
$$

Thus, the negative margin function g satisfies a KL-inequality with exponent $\theta = 1/2$ and modulus $\alpha = 2/\sqrt{\gamma_{opt}}$ as claimed. $\square$

Proof of Theorem 2.2. The LHS of the inequality is trivial since the margin function γ is R-Lipschitz on $\mathbb{S}_{m-1}$. For the RHS, note from Theorem 2.1 that the negative margin function g (defined in Theorem 2.2) satisfies a KL-inequality with exponent $\theta = 1/2$ and modulus $\alpha = 2/\sqrt{\gamma_{opt}}$. The result then follows as upon invoking Proposition 2.1. $\square$

Acknowledgement. The author is grateful to Ji Ziwei and Matus Telgarsky for spotted an error in our of the theorems announced in a previous version of our paper.

4 Concluding remarks

We have established Kurdyka-Lojasiewicz inequality of exponent $1/2$ for the margin function in linearly separable classification problems. This result gives hopes for the existence of fast (perhaps quasi-linear) optimization schemes for such problems. Also, we have employed our result to establish a generic inequality linking the convergence rates of the bias and of margin. This immediately allows for 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.

Acknowledgement. The author is thankful to Ji Ziwei and Matus Telgarsky for reporting an error in one of the theorems in an earlier version of this preprint.

References

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

Bauschke, H. H., Luke, D. R., Phan, H. M., and Wang, X. (2013). Restricted normal cones and the method of alternating projections: Applications. *Set-Valued and Variational Analysis*, 21(3).

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.

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.

Van Ngai, H., The Luc, D., and Théra, M. (2002). Extensions of fréchet ϵ-subdifferential calculus and applications. *Journal of Mathematical Analysis and Applications*, 268(1).

A Omitted technical proofs

Lemma 3.1. *For every* $\mathbf{q} = (q_1, \ldots, q_n) \in \Delta_{n-1}$*, it holds that* $\overline{\gamma}_n \le \|\sum_{i=1}^n q_i y_i \mathbf{x}_i\| \le 1$.

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

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

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

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

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