Gradient descent follows the regularization path for general losses

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

Miroslav Dudík MDUDIK@MICROSOFT.COM
Microsoft Research, New York, NY

Robert E. Schapire SCHAPIRE@MICROSOFT.COM
Microsoft Research, New York, NY

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

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

Recent work across many machine learning disciplines has highlighted that standard descent methods, even without explicit regularization, do not merely minimize the training error, but also exhibit an *implicit bias*. This bias is typically towards a certain regularized solution, and relies upon the details of the learning process, for instance the use of the cross-entropy loss.

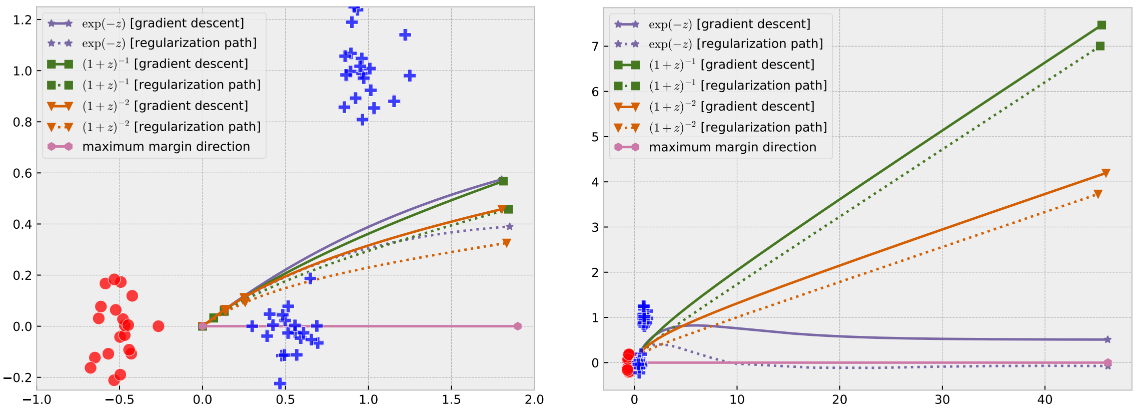
In this work, we show that for empirical risk minimization over linear predictors with *arbitrary* convex, strictly decreasing losses, if the risk does not attain its infimum, then the gradient-descent path and the *algorithm-independent* regularization path converge to the same direction (whenever either converges to a direction). Using this result, we provide a justification for the widely-used exponentially-tailed losses (such as the exponential loss or the logistic loss): while this convergence to a direction for exponentially-tailed losses is necessarily to the maximum-margin direction, other losses such as polynomially-tailed losses may induce convergence to a direction with a poor margin.

Keywords: implicit regularization, gradient descent, exponentially-tailed losses.

1. Introduction

A central problem in machine learning is *overfitting*, where a predictor performs well on training data, but poorly on testing data. A direct way to mitigate overfitting is to add an *explicit* regularizer, such as an ℓ_1 or ℓ_2 penalty on the model parameters. Another approach, achieving strong empirical results in modern models with many parameters (Zhang et al., 2016), is to exploit the *implicit* regularization exhibited by common descent methods, such as coordinate descent (Schapire et al., 1997) and gradient descent (Soudry et al., 2018), simply by running them a long time with no explicit regularization.

In fact, as will be explored in this work, there is a strong relationship between implicit and explicit regularization. For example, coordinate-descent iterates under exponential loss minimization (or, equivalently, *AdaBoost iterates*, see Freund and Schapire, 1997) and ℓ_1-regularized solutions are both biased towards ℓ_1-maximum-margin solutions (Zhang and Yu, 2005; Telgarsky, 2013; Rosset et al., 2004; Zhao and Yu, 2007). Similarly, gradient-descent iterates under exponential or logistic loss minimization and the corresponding ℓ_2-regularized solutions are both biased towards ℓ_2-maximum-margin solutions (Soudry et al., 2018; Ji and Telgarsky, 2019b).



(a) Zoomed in.

(b) Zoomed out.

Figure 1: Behavior of gradient descent and regularization path for three losses: the exponential loss $\exp(-z)$, and two polynomially-tailed losses $(1+z)^{-1}$ and $(1+z)^{-2}$ (with a quadratic extension along $z < 0$ for smoothness). The data has one negative (red) point cloud, and two positive (blue) point clouds; the upper positive cloud pulls the predictors trained with polynomially-tailed losses away from the maximum-margin direction, which points straight to the right.

The preceding methods, which rose to prominence for their empirical performance, all shared a curious property: an insistence upon a loss with *exponential tails*, such as the exponential loss or the logistic loss. This is an odd coincidence, as the classical theory of classification performance of convex losses indicates a wide variety should work well, in both theory and practice (Bartlett et al., 2006; Zhang, 2004). This leads to the central question of this work:

> *For general convex decreasing losses, what is the relationship between gradient descent iterates and the regularized solutions?*

This work focuses on gradient descent and ℓ_2 regularization. Before describing the formal results, we demonstrate on a concrete example the trends we would like to capture. Figure 1 shows the path followed by gradient descent and the *regularization path*, obtained by taking the regularization weight down to 0, for three separate losses on the same data set, consisting of the three depicted point clouds. Zooming in on the data as in Figure 1(a), the behavior is unclear. Zooming out in Figure 1(b), however, a trend emerges: for each loss, its gradient-descent path and regularization path asymptotically follow the same direction. Moreover, the choice of loss function may lead to a different convergent direction, and only the exponential loss converges to the maximum-margin direction.

1.1. Contributions

The goal of this work is to pin down the relationship between gradient-descent paths and regularization paths for linear predictors, but only assuming the losses are convex and strictly decreasing.

Definitions will be mostly deferred, but to summarize the main results, a bit of notation is needed. Throughout, $\mathcal{R}$ will denote the empirical risk, and $(\boldsymbol{w}_t)_{t\geq 0}$ will denote gradient descent iterates given by

$$\boldsymbol{w}_{t+1} := \boldsymbol{w}_t - \eta \nabla \mathcal{R}(\boldsymbol{w}_t), \tag{1}$$

where $\eta > 0$ is a sufficiently small but constant step size. Meanwhile, $\bar{\boldsymbol{w}}(B)$ will denote the regularized solution with ℓ_2 norm B; concretely,

$$\bar{\boldsymbol{w}}(B) := \operatorname*{arg\,min}_{\|\boldsymbol{w}\| \le B} \mathcal{R}(\boldsymbol{w}), \tag{2}$$

and the *regularization path* denotes the curve followed by $\bar{\boldsymbol{w}}$ as B varies, meaning $(\bar{\boldsymbol{w}}(B))_{B \ge 0}$. Choosing regularized rather than constrained solutions does not change our results regarding the regularization path; moreover, in either case, the paths are algorithm-independent.

As in Figure 1, this work is in the setting where the empirical risk $\mathcal{R}$ does not attain its infimum, and consequently (as verified in Section 2), both $\|\boldsymbol{w}_t\| \to \infty$ and $\|\bar{\boldsymbol{w}}(B)\| \to \infty$. As will be shown in Sections 3 and 4, with strictly decreasing losses, $\mathcal{R}$ does not attain its infimum if the training set has a nonempty "separable" part; it is also true in the cases of AdaBoost and deep networks, where perfect classification is possible (cf. Section 1.2). Since the norms grow unboundedly, to compare $\boldsymbol{w}_t$ and $\bar{\boldsymbol{w}}(B)$, this work compares the directions to which they converge: namely, $\lim_{t\to\infty} \frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|}$ and $\lim_{B\to\infty} \frac{\bar{\boldsymbol{w}}(B)}{B}$, when the limits exist. Since we use linear classifiers here, this normalization does not affect their (binary) predictions.

Our core contribution can be summarized as follows.

Theorem 1 (Coarsening of Theorems 4, 5 and 15) *Suppose the loss function is convex, strictly decreasing to* 0*, the empirical risk* $\mathcal{R}$ *does not attain its infimum, and the step size* $\eta > 0$ *is sufficiently small (as discussed in Section 2). Then* $\lim_{t\to\infty} \frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|} = \lim_{B\to\infty} \frac{\bar{\boldsymbol{w}}(B)}{B}$ *whenever either limit exists.*

In words, Theorem 1 states that if either the gradient-descent path or the algorithm-independent regularization path converge to a direction, then both of them converge to the same direction. In more detail, our full contributions and the paper organization are as follows.

Section 2 shows that if the gradient-descent path converges to a direction, then the regularization path converges to the same direction. Interestingly, this proof holds for general convex functions not attaining their infimum, and does not require any properties of the risk.

Section 3 focuses on the case of *linearly separable data*. The primary effort is in showing the converse to Section 2 in this setting, namely that if the regularization path converges to a direction, then the gradient-descent path converges to the same direction. This section also establishes that exponentially-tailed losses (cf. eq. 12) all converge to the same maximum-margin direction, that polynomially-tailed losses (cf. eq. 13) converge to a direction but may only achieve a poor margin, and lastly that for general losses the iterates may fail to converge to a direction.

Section 4 completes the picture in the case of general data which is potentially not linearly separable: that is, if the empirical risk does not attain its infimum, and if the regularization path converges to a direction, then the gradient-descent path converges to the same direction. This setting introduces significant technicalities, but also comes with interesting refinements: while gradient descent and the regularization path do not converge to a point (only to a direction, as in Figure 1) in this nonseparable setting, it is possible to show convergence to a point over a certain subspace.

We provide concluding remarks and open problems in Section 5.

1.2. Related work

Arguably, the earliest relevant literature is the introduction of the support vector machine (SVM), which utilizes explicit regularization to select maximum margin classifiers (Vapnik, 1982)—the property that was eventually tied to generalization performance (Shawe-Taylor et al., 1998; Bartlett, 1996). This use of explicit regularization is significantly different from the setup here: there, the loss is hinge loss (which attains 0) and the regularization level is constant, whereas here, the loss necessarily asymptotes to 0, and the regularization level is also taken to 0. In a concrete sense, exponential losses with this decaying regularization behave asymptotically like the SVM, and this analogy was used explicitly in the aforementioned gradient descent proof of Soudry et al. (2018). Turning back to descent methods, the original use of margins was in the analysis of perceptron (Novikoff, 1962), however there is no implicit bias: the method terminates with 0 classification error, but no reasonable lower bound can be placed on the achieved margin.

The first concrete studies showing an implicit bias of descent methods were for the ℓ_1-regularized case. Coordinate descent, when paired with the exponential loss, is implicitly biased towards ℓ_1-regularized solutions. This observation is the result of separate lines of work on descent methods and on regularization methods. On one hand, AdaBoost was shown to exhibit *positive margins*, meaning its predictions are not only correct, but in a certain sense robust (Schapire et al., 1997); indeed, with some further care on the descent step sizes, AdaBoost finds maximum-margin solutions (Zhang and Yu, 2005; Telgarsky, 2013). On the other hand, the ℓ_1-regularized solutions also converge to maximum-margin solutions as regularization strength is taken to 0 (Rosset et al., 2004; Zhao and Yu, 2007).

Another line of research has shown that gradient descent, when paired with the exponential or logistic loss, converges to ℓ_2-regularized solutions. This was first established for linear methods when the data is linearly separable (Soudry et al., 2018), meaning there exists a linear predictor which perfectly labels all data, but has since been extended to linear predictors on nonseparable data (Ji and Telgarsky, 2019b). Soudry et al. (2018) and Ji and Telgarsky (2019b) only handled exponentially-tailed losses, while in this paper we prove results for general losses and do not require separability.

The implicit bias of gradient descent has also been studied for linear convolutional networks (Gunasekar et al., 2018), deep linear networks (Ji and Telgarsky, 2019a), and homogeneous networks (Lyu and Li, 2020), where empirical results seem to suggest such a bias exists (Neyshabur et al., 2014; Bartlett et al., 2017). Similarly to the situation with AdaBoost, there is a variety of results focusing purely on explicitly-regularized methods Wei et al. (2019).

As a final brief remark, implicit bias and margins have been extended beyond standard classification settings, for instance to adversarial training (Charles et al., 2019; Li et al., 2020).

2. Convergence of gradient descent implies convergence of regularization path

In this section we show one direction of the equivalence, which holds in a more general setting.

Given a differentiable convex function $f : \mathbb{R}^d \to \mathbb{R}$ (not necessarily the empirical risk) and an ℓ_2-norm bound B, the regularized solution is defined as

$$\bar{\boldsymbol{w}}(B) := \arg\min_{\|\boldsymbol{w}\| \le B} f(\boldsymbol{w}). \tag{3}$$

Note that $\bar{\boldsymbol{w}}(B)$ is not unique in general, but we still have $\lim_{B\to\infty} f\left(\bar{\boldsymbol{w}}(B)\right) = \inf_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w})$, as is often the case when working with unregularized losses. In this paper we are particularly interested in the case where the infimum of f is not attained. In that case $\bar{\boldsymbol{w}}(B)$ is uniquely defined, because the set of minimizers is convex and contained in the surface of the ℓ_2 ball, and thus consists of exactly one point due to the curvature of ℓ_2 balls. An example of a function f that does not attain the infimum is e^{-z}: its infimum is 0, which is not attained by any $z \in \mathbb{R}$. A more interesting example is an empirical risk with a nonempty separable part, which will be introduced in Sections 3 and 4.

We minimize f using gradient descent, meaning

$$\boldsymbol{w}_{t+1} := \boldsymbol{w}_t - \eta \nabla f(\boldsymbol{w}_t). \tag{4}$$

Its basic properties are summarized in Lemma 2. If there exists a small step size which ensures decreasing function values, then gradient descent on f can minimize the function value to its infimum; moreover, if the infimum of f is not attained, then gradient descent iterates go to infinity.

Lemma 2 *Given a convex differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, suppose the step size η satisfies*

$$f(\boldsymbol{w}_{t+1}) - f(\boldsymbol{w}_t) \leq -\frac{\eta}{2} \|\nabla f(\boldsymbol{w}_t)\|^2 \tag{5}$$

for all $t \geq 0$. Then for any $\boldsymbol{w} \in \mathbb{R}^d$,

$$\|\boldsymbol{w}_{t+1} - \boldsymbol{w}\|^2 \leq \|\boldsymbol{w}_t - \boldsymbol{w}\|^2 + 2\eta \left(f(\boldsymbol{w}) - f(\boldsymbol{w}_{t+1}) \right), \tag{6}$$

and thus $\|\boldsymbol{w}_{t+1} - \boldsymbol{w}\| \leq \|\boldsymbol{w}_t - \boldsymbol{w}\|$ as long as $f(\boldsymbol{w}) \leq f(\boldsymbol{w}_{t+1})$. Consequently,

$$\lim_{t\to\infty} f(\boldsymbol{w}_t) = \inf_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w}),$$

which implies $\lim_{t\to\infty} \|\boldsymbol{w}_t\| = \infty$ if the infimum of f is not attained.

Remark 3 *The step size condition in eq. (5) holds if f is (globally) β-smooth and $\eta \leq 1/\beta$. There are also standard situations where f merely obeys local smoothness over its sublevel sets; see for example eq. (8), which considers empirical risk minimization with the exponential loss.*

Below is our main result of this section.

Theorem 4 *Consider the gradient descent iterates $(\boldsymbol{w}_t)_{t\geq 0}$ given by eq. (4), and the regularized solutions $(\bar{\boldsymbol{w}}(B))_{B\geq 0}$ given by eq. (3). Suppose f is convex, differentiable, bounded below by 0, and has an unattained infimum, and the step size η satisfies eq. (5) and $\eta \leq 1/(2f(\boldsymbol{w}_0))$. If $\lim_{t\to\infty} \boldsymbol{w}_t/\|\boldsymbol{w}_t\| = \bar{\boldsymbol{u}}$ for some unit vector $\bar{\boldsymbol{u}}$, then also $\lim_{B\to\infty} \bar{\boldsymbol{w}}(B)/B = \bar{\boldsymbol{u}}$.*

The full proof of Theorem 4 is given in Appendix A. Here we sketch the main arguments. The key property used in the proof is eq. (6). Note that given any $B > 0$, by the definition of $\bar{\boldsymbol{w}}(B)$, as long as $\|\boldsymbol{w}_t\|, \|\boldsymbol{w}_{t+1}\| \leq B$, it holds that $\|\boldsymbol{w}_{t+1} - \bar{\boldsymbol{w}}(B)\| \leq \|\boldsymbol{w}_t - \bar{\boldsymbol{w}}(B)\|$. In other words, the distance from the gradient-descent path to $\bar{\boldsymbol{w}}(B)$ is nonincreasing within the ball $\{\boldsymbol{w} : \|\boldsymbol{w}\| \leq B\}$.

Suppose for some $\epsilon > 0$, there exists arbitrarily large B with $\left\|\frac{\bar{\boldsymbol{w}}(B)}{B} - \bar{\boldsymbol{u}}\right\| > \epsilon$. By Euclidean geometry, we can show that

$$\|B\bar{\boldsymbol{u}} - \bar{\boldsymbol{w}}(B)\| - \|\langle \bar{\boldsymbol{w}}(B), \bar{\boldsymbol{u}} \rangle \bar{\boldsymbol{u}} - \bar{\boldsymbol{w}}(B)\| > \frac{B\epsilon^3}{8}.$$

By the assumption, if $\|\boldsymbol{w}_t\|$ is large enough, then $\boldsymbol{w}_t/\|\boldsymbol{w}_t\|$ and $\bar{\boldsymbol{u}}$ can be arbitrarily close. The idea is then to find two gradient descent iterates $\boldsymbol{w}_{t_1}$ and $\boldsymbol{w}_{t_2}$, where $t_1 < t_2$, and $\boldsymbol{w}_{t_1}$ is close to $\langle \bar{\boldsymbol{w}}(B), \bar{\boldsymbol{u}} \rangle \bar{\boldsymbol{u}}$, and $\boldsymbol{w}_{t_2}$ is close to $B\bar{\boldsymbol{u}}$. It then follows that $\|\boldsymbol{w}_{t_2} - \bar{\boldsymbol{w}}(B)\| > \|\boldsymbol{w}_{t_1} - \bar{\boldsymbol{w}}(B)\|$, which violates eq. (6).

3. Convergence to a direction for the linearly separable case

In the remainder of the paper, we consider binary classification with a training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$, and we assume $\|\boldsymbol{x}_i\| \le 1$ without loss of generality. We use a linear classifier $\boldsymbol{w} \in \mathbb{R}^d$, which is learned by minimizing the empirical risk

$$\mathcal{R}(\boldsymbol{w}) := \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle\right),$$

where the loss function ℓ is assumed to be convex, differentiable, and strictly decreasing to 0, such as the logistic loss $\ln(1 + e^{-z})$.

In this section, we assume that the training data is linearly separable: there exists a unit vector $\boldsymbol{u}$ and some $\gamma > 0$ such that $y_i \langle \boldsymbol{u}, \boldsymbol{x}_i \rangle \ge \gamma$ for all $1 \le i \le n$. Results in this section can be extended to the general case with no assumption on the training data, as we will do in Section 4.

Linear separability and a strictly decreasing loss imply that the infimum of $\mathcal{R}$ is not attained, and thus Theorem 4 can be applied. However, we can show a stronger result: the gradient-descent path converges to a direction if and only if the regularization path converges to (the same) direction.

Theorem 5 *Consider the gradient descent iterates $(\boldsymbol{w}_t)_{t \ge 0}$ given by eq. (1), and the regularized solutions $(\bar{\boldsymbol{w}}(B))_{B \ge 0}$ given by eq. (2). Suppose the data is linearly separable, and the step size satisfies $\eta \le 1/\left(2\mathcal{R}(\boldsymbol{w}_0)\right)$ and*

$$\mathcal{R}(\boldsymbol{w}_{t+1}) - \mathcal{R}(\boldsymbol{w}_t) \le -\frac{\eta}{2} \left\|\mathcal{R}(\boldsymbol{w}_t)\right\|^2 \tag{7}$$

for all $t \ge t_0$. Then $\lim_{t \to \infty} \boldsymbol{w}_t / \|\boldsymbol{w}_t\|$ exists if and only if $\lim_{B \to \infty} \bar{\boldsymbol{w}}(B)/B$ exists, and when they exist they are the same.

Remark 6 *It can be verified that if the loss function ℓ is β-smooth, then so is the empirical risk function $\mathcal{R}$, and eq. (7) holds if $\eta \le 1/\beta$. However, it may still hold for a loss function which is not globally smooth. For example, for the exponential loss e^{-z}, Lemma 3.4 of Ji and Telgarsky (2019b) ensures that*

$$\mathcal{R}(\boldsymbol{w}_{t+1}) - \mathcal{R}(\boldsymbol{w}_t) \le -\eta \left(1 - \frac{\eta \mathcal{R}(\boldsymbol{w}_t)}{2}\right) \left\|\nabla \mathcal{R}(\boldsymbol{w}_t)\right\|^2 \tag{8}$$

as long as $\eta \mathcal{R}(\boldsymbol{w}_t) \le 1$. Therefore, eq. (7) holds as long as $\eta \le 1/\mathcal{R}(\boldsymbol{w}_0)$.

The “if” part of Theorem 5 follows directly from Theorem 4. Next we give a proof sketch of the “only if” part of Theorem 5; the full proof is given in Appendix B.

In the remainder of this section, we assume that $\lim_{B \to \infty} \bar{\boldsymbol{w}}(B)/B = \bar{\boldsymbol{u}}$ for some unit vector $\bar{\boldsymbol{u}}$, and define its margin as

$$\bar{\gamma} := \min_{1 \le i \le n} y_i \langle \bar{\boldsymbol{u}}, \boldsymbol{x}_i \rangle.$$

Moreover, the maximum margin $\hat{\gamma}$ and the maximum-margin solution $\hat{\boldsymbol{u}}$ are defined as

$$\hat{\gamma} := \max_{\|\boldsymbol{u}\|=1} \min_{1 \le i \le n} y_i \langle \boldsymbol{u}, \boldsymbol{x}_i \rangle, \quad \text{and} \quad \hat{\boldsymbol{u}} := \arg\max_{\|\boldsymbol{u}\|=1} \min_{1 \le i \le n} y_i \langle \boldsymbol{u}, \boldsymbol{x}_i \rangle.$$

We first show that $\bar{\gamma}$ is always positive.

Lemma 7 *It holds that $\bar{\gamma} \geq \hat{\gamma}^2/(2n) > 0$, where $\hat{\gamma}$ is the maximum margin.*

Remark 8 *Lemma 7 gives a worst-case lower bound on margin, which holds for an arbitrary decreasing loss. The proof technique can be adapted to a specific loss function. For example, if the loss function has a polynomial tail az^{-b}, then $\lim_{B\to\infty} \bar{\boldsymbol{w}}(B)/B$ exists (cf. Proposition 11), and we can prove an $\Omega(n^{-1/(b+1)})$ lower bound on margin. Moreover, there exists a dataset on which this lower bound is tight (cf. Proposition 12).*

Here is a proof sketch of Lemma 7. The starting point is the property that $\bar{\boldsymbol{w}}(B)$ and $\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)$ are collinear, meaning

$$-\left\langle \frac{\bar{\boldsymbol{w}}(B)}{B}, \nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right) \right\rangle = \left\| \nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)\right\|, \tag{9}$$

which is a consequence of the first-order optimality conditions. Next, by the chain rule, the left hand side of eq. (9) is naturally related to the margin of $\bar{\boldsymbol{w}}(B)/B$:

$$-\left\langle \frac{\bar{\boldsymbol{w}}(B)}{B}, \nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right) \right\rangle = \frac{1}{n}\sum_{i=1}^{n} -\ell'\left(\langle \bar{\boldsymbol{w}}(B), y_i\boldsymbol{x}_i\rangle\right) \left\langle \frac{\bar{\boldsymbol{w}}(B)}{B}, y_i\boldsymbol{x}_i \right\rangle, \tag{10}$$

while the right hand side of eq. (9) can be bounded using the Cauchy-Schwarz inequality and the maximum-margin solution $\hat{\boldsymbol{u}}$:

$$\left\| \nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)\right\| \geq \left\langle -\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right), \hat{\boldsymbol{u}} \right\rangle \geq \frac{1}{n}\sum_{i=1}^{n} -\ell'\left(\langle \bar{\boldsymbol{w}}(B), y_i\boldsymbol{x}_i\rangle\right)\hat{\gamma}. \tag{11}$$

If $\bar{\gamma} < \hat{\gamma}^2/(2n)$, then since the regularization path converges to $\bar{\boldsymbol{u}}$, the margin of $\bar{\boldsymbol{w}}(B)/B$ is no larger than $\hat{\gamma}^2/(2n)$ for all large B. To ensure that eq. (10) is upper bounded by eq. (11) would require that

$$-\ell'(B\hat{\gamma}) \geq -\ell'\left(\frac{B\hat{\gamma}^2}{2n}\right)\frac{\hat{\gamma}}{2n}.$$

This would in turn imply $\int_0^\infty -\ell'(z)\,\mathrm{d}z = \infty$, a contradiction.

Next we can show that to minimize the risk, it is almost optimal to move along the direction of $\bar{\boldsymbol{u}}$, thanks to its positive margin.

Lemma 9 *Given any $\alpha > 0$, there exists $\rho(\alpha) > 0$, such that for any $\boldsymbol{w}$ with $\|\boldsymbol{w}\| > \rho(\alpha)$, it holds that*

$$\mathcal{R}\big((1+\alpha)\|\boldsymbol{w}\|\,\bar{\boldsymbol{u}}\big) \leq \mathcal{R}(\boldsymbol{w}).$$

To prove Lemma 9, first note that by definition $\mathcal{R}\big(\bar{\boldsymbol{w}}(\|\boldsymbol{w}\|)\big) \leq \mathcal{R}(\boldsymbol{w})$, and thus it is enough to show that $\mathcal{R}\big((1+\alpha)\|\boldsymbol{w}\|\,\bar{\boldsymbol{u}}\big) \leq \mathcal{R}\big(\bar{\boldsymbol{w}}(\|\boldsymbol{w}\|)\big)$. This is true if for all $1 \leq i \leq n$,

$$y_i\left\langle (1+\alpha)\|\boldsymbol{w}\|\,\bar{\boldsymbol{u}}, \boldsymbol{x}_i\right\rangle \geq y_i\left\langle \bar{\boldsymbol{w}}(\|\boldsymbol{w}\|), \boldsymbol{x}_i\right\rangle, \quad \text{i.e.,} \quad (1+\alpha)y_i\langle \bar{\boldsymbol{u}}, \boldsymbol{x}_i\rangle \geq y_i\left\langle \frac{\bar{\boldsymbol{w}}\left(\|\boldsymbol{w}\|\right)}{\|\boldsymbol{w}\|}, \boldsymbol{x}_i\right\rangle.$$

Since $y_i\langle \alpha\bar{\boldsymbol{u}}, \boldsymbol{x}_i\rangle \geq \alpha\bar{\gamma}$ and $\|\boldsymbol{x}_i\| \leq 1$, we only need to choose $\|\boldsymbol{w}\|$ large enough such that

$$\left\|\bar{\boldsymbol{u}} - \frac{\bar{\boldsymbol{w}}(\|\boldsymbol{w}\|)}{\|\boldsymbol{w}\|}\right\| \leq \alpha\bar{\gamma}.$$

Now we are ready to prove the "only if" part of Theorem 5. The full proof appears in Appendix B, but is a bit cumbersome in our discrete-time setting; here we will illustrate the idea with the *gradient flow*, meaning $\eta \to 0$ and $\dot{\boldsymbol{w}}_t := \mathrm{d}\boldsymbol{w}_t/\,\mathrm{d}t = -\nabla\mathcal{R}(\boldsymbol{w}_t)$. For any $\alpha > 0$, due to $\|\boldsymbol{w}_t\| \to \infty$ and Lemma 9, we can choose t_0 large enough so that $\mathcal{R}\left((1+\alpha)\|\boldsymbol{w}_t\|\bar{\boldsymbol{u}}\right) \leq \mathcal{R}(\boldsymbol{w}_t)$ for all $t \geq t_0$. By convexity,

$$0 \geq \mathcal{R}\big((1+\alpha)\|\boldsymbol{w}_t\|\bar{\boldsymbol{u}}\big) - \mathcal{R}(\boldsymbol{w}_t) \geq \langle \dot{\boldsymbol{w}}_t, \boldsymbol{w}_t - (1+\alpha)\|\boldsymbol{w}_t\|\bar{\boldsymbol{u}}\rangle\,,$$

which rearranges to

$$\langle \dot{\boldsymbol{w}}_t, \bar{\boldsymbol{u}}\rangle \geq \left(\frac{1}{1+\alpha}\right)\left\langle \dot{\boldsymbol{w}}_t, \frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|}\right\rangle = \left(\frac{1}{1+\alpha}\right)\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{w}_t\|.$$

For any $t_1 \geq t_0$, integrating both sides along $[t_0, t_1]$ gives

$$\langle \boldsymbol{w}_{t_1} - \boldsymbol{w}_{t_0}, \bar{\boldsymbol{u}}\rangle = \left\langle \int_{t_0}^{t_1} \dot{\boldsymbol{w}}_t\,\mathrm{d}t, \bar{\boldsymbol{u}}\right\rangle \geq \left(\frac{1}{1+\alpha}\right)\int_{t_0}^{t_1}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{w}_t\|\,\mathrm{d}t = \frac{\|\boldsymbol{w}_{t_1}\| - \|\boldsymbol{w}_{t_0}\|}{1+\alpha}.$$

Dividing both sides by $\|\boldsymbol{w}_{t_1}\|$ and applying $\liminf_{t_1\to\infty}$, since $\liminf_{t_1\to\infty}\boldsymbol{w}_{t_0}/\|\boldsymbol{w}_{t_1}\| = 0$,

$$\liminf_{t_1\to\infty}\left\langle \frac{\boldsymbol{w}_{t_1}}{\|\boldsymbol{w}_{t_1}\|}, \bar{\boldsymbol{u}}\right\rangle = \liminf_{t_1\to\infty}\left\langle \frac{\boldsymbol{w}_{t_1} - \boldsymbol{w}_{t_0}}{\|\boldsymbol{w}_{t_1}\|}, \bar{\boldsymbol{u}}\right\rangle \geq \liminf_{t_1\to\infty}\frac{\|\boldsymbol{w}_{t_1}\| - \|\boldsymbol{w}_{t_0}\|}{(1+\alpha)\|\boldsymbol{w}_{t_1}\|} = \frac{1}{1+\alpha}.$$

Since $\alpha > 0$ was arbitrary, the "only if" part of Theorem 5 is complete.

3.1. What does the regularization path converge to?

Theorem 5 says that the gradient-descent path and regularization path converge to the same direction if either of them converges to a direction. Moreover, the regularization path is independent of the optimization algorithm, and thus easier to study. Here are some examples where $\bar{\boldsymbol{w}}(B)/B$ converges.

A classical example is that if the loss has an exponential tail, then the regularization path converges to the maximum-margin direction (see Rosset et al., 2004, for the case of ℓ_1 regularization).

Proposition 10 *If for some $a, b > 0$,*

$$\lim_{z\to\infty}\frac{\ell(z)}{a\exp(-bz)} = 1, \tag{12}$$

then $\lim_{B\to\infty}\bar{\boldsymbol{w}}(B)/B = \hat{\boldsymbol{u}}$, where $\hat{\boldsymbol{u}}$ is the unique maximum margin solution.

We also prove that if the loss has a polynomial tail, then the regularization path converges to a direction.

Proposition 11 *If for some $a, b > 0$,*

$$\lim_{z\to\infty}\frac{-\ell'(z)}{az^{-b}} = 1, \tag{13}$$

then $\lim_{B\to\infty}\bar{\boldsymbol{w}}(B)/B$ exists.

However, while an exponentially-tailed loss (cf. eq. 12) always induces the maximum-margin direction, a polynomially-tailed loss (cf. eq. 13) may induce a different direction:

Proposition 12 *For any $b > 0$, consider a loss function ℓ which equals z^{-b} for $z \geq 1$. There exists a dataset on which the maximum margin is a universal constant, while the regularization path with ℓ converges to a direction which has margin $\Theta(n^{-1/(b+1)})$.*

Lastly, note that directional convergence should not be taken for granted: we can construct a loss function which satisfies all the conditions in Theorem 5 (i.e., convexity, monotonicity and eq. 7) for which $\bar{\boldsymbol{w}}(B)/B$ does not converge. The constructed loss switches between $\exp(-z)$ and $1/z$ countably infinitely often, with the switching locations chosen carefully so that $\bar{\boldsymbol{w}}(B)/B$ continually oscillates.

Proposition 13 *There exists a loss function ℓ which is convex, strictly decreasing to 0 and 2-smooth for which $\bar{\boldsymbol{w}}(B)/B$ does not converge.*

The proofs of all results in this subsection are given in Appendix B.1.

4. Convergence to a direction for the general case

In this section, we extend the preceding results to the general case of an arbitrary training set, that might or might not be linearly separable. The main idea is to first partition the dataset into a separable part and a nonseparable part using the decomposition studied by Ji and Telgarsky (2019b) (cf. Lemma 14 below). Then we prove (subject to the conditions below) that the gradient-descent path and regularization path are strongly coupled in a highly-refined sense: (1) On the space spanned by the nonseparable part of the dataset, convergence of both gradient descent and the regularization path is to the same unique finite point. (2) On the space perpendicular to the nonseparable part, as in the fully separable case, the gradient-descent path and regularization path converge to the same direction (if either converges to a direction).

Here we define the decomposition formally. Given a dataset $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, we decompose it into $D_s \cup D_c$ in the following way. For each data example $(\boldsymbol{x}_i, y_i)$, if there exists a unit vector $\boldsymbol{u}$ such that $y_i\langle \boldsymbol{u}, \boldsymbol{x}_i\rangle > 0$ and $y_j\langle \boldsymbol{u}, \boldsymbol{x}_j\rangle \geq 0$ for all $1 \leq j \leq n$, then we include $(\boldsymbol{x}_i, y_i)$ into D_c, otherwise we include it into D_s. (The mnemonic is *"s"* for strongly-convex (as justified below) and *"c"* for its complement.) Define

$$\mathcal{R}_s(\boldsymbol{w}) := \frac{1}{n} \sum_{(\boldsymbol{x}_i, y_i) \in D_s} \ell\left(y_i\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle\right), \quad \text{and} \quad \mathcal{R}_c(\boldsymbol{w}) := \frac{1}{n} \sum_{(\boldsymbol{x}_i, y_i) \in D_c} \ell\left(y_i\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle\right),$$

and note that $\mathcal{R} = \mathcal{R}_s + \mathcal{R}_c$. Further define $S := \operatorname{span}\big(\{\boldsymbol{x}_i : (\boldsymbol{x}_i, y_i) \in D_s\}\big)$, and let Π_S denote the projection onto S, and $\Pi_\perp$ denote the projection onto $S^\perp$. Given $\boldsymbol{w} \in \mathbb{R}^d$, let $\boldsymbol{w}_S := \Pi_S \boldsymbol{w}$ and $\boldsymbol{w}_\perp := \Pi_\perp \boldsymbol{w}$.

Lemma 14 *(Ji and Telgarsky, 2019b, Theorem 2.1) The above decomposition satisfies the following properties.*

(1) If ℓ is twice continuously differentiable with $\ell'' > 0$, then $\mathcal{R}_s$ has compact sublevel sets over S, is strongly convex over compact subsets of S, and therefore has a unique minimizer $\bar{\boldsymbol{v}}$ over S.

(2) D_c can be linearly separated in $S^\perp$, meaning that there exists a unit vector $\boldsymbol{u} \in S^\perp$ and some $\gamma > 0$, such that $y_i\langle \boldsymbol{u}, \boldsymbol{x}_i\rangle \geq \gamma$ for all $(\boldsymbol{x}_i, y_i) \in D_c$.

Note that for any $\boldsymbol{v} \in S$, and any $\boldsymbol{u} \in S^\perp$ which can separate D_c, it holds that $\mathcal{R}_s(\boldsymbol{v}) = \lim_{r\to\infty} \mathcal{R}(\boldsymbol{v} + r\boldsymbol{u})$, and thus $\inf_{\boldsymbol{w}\in\mathbb{R}^d} \mathcal{R}(\boldsymbol{w}) = \inf_{\boldsymbol{v}\in S} \mathcal{R}_s(\boldsymbol{v}) = \mathcal{R}_s(\bar{\boldsymbol{v}})$. Moreover, if $D_c \neq \emptyset$, then the infimum of $\mathcal{R}$ is not attained.

With the decomposition and Lemma 14, we can state our equivalence result for general dataset.

Theorem 15 *Consider the gradient descent iterates $(\boldsymbol{w}_t)_{t\geq 0}$ given by eq. (1), and the regularized solutions $(\bar{\boldsymbol{w}}(B))_{B\geq 0}$ given by eq. (2). Suppose ℓ is twice continuously differentiable with $\ell'' > 0$, and the step size $\eta \leq 1/\left(2\mathcal{R}(\boldsymbol{w}_0)\right)$ satisfies eq. (7).*

(1) On S it holds that $\lim_{t\to\infty} \Pi_S \boldsymbol{w}_t = \bar{\boldsymbol{v}}$ and $\lim_{B\to\infty} \Pi_S \bar{\boldsymbol{w}}(B) = \bar{\boldsymbol{v}}$.

(2) If $D_c \neq \emptyset$, then $\lim_{t\to\infty} \|\boldsymbol{w}_t\| = \lim_{B\to\infty} \|\bar{\boldsymbol{w}}(B)\| = \infty$, and $\lim_{t\to\infty} \boldsymbol{w}_t/\|\boldsymbol{w}_t\|$ exists if and only if $\lim_{B\to\infty} \bar{\boldsymbol{w}}(B)/B$ exists, and when they exist they are the same and lie in $S^\perp$.

The convergence result on S is straightforward: it follows from Lemma 2 that $\lim_{t\to\infty} \mathcal{R}(\boldsymbol{w}_t) = \inf_{\boldsymbol{w}\in\mathbb{R}^d} \mathcal{R}(\boldsymbol{w}) = \mathcal{R}_s(\bar{\boldsymbol{v}})$. Since $\mathcal{R}_s(\boldsymbol{w}_t) \leq \mathcal{R}(\boldsymbol{w}_t)$, we also have $\mathcal{R}_s(\boldsymbol{w}_t) \to \mathcal{R}_s(\bar{\boldsymbol{v}})$. Lemma 2 also ensures that $\mathcal{R}_s(\boldsymbol{w}_t) \leq \mathcal{R}(\boldsymbol{w}_t) \leq \mathcal{R}(\boldsymbol{w}_0)$, and since $\mathcal{R}_s$ is strongly convex over sublevel sets, we have $\lim_{t\to\infty} \Pi_S \boldsymbol{w}_t = \bar{\boldsymbol{v}}$. The proof for regularized solutions is similar.

The "if" part of Theorem 15(2) also follows directly from Theorem 4. The limiting direction must lie in $S^\perp$ since $\Pi_S \boldsymbol{w}_t$ is bounded due to Theorem 15(1). Below we give a proof sketch of the "only if" part of Theorem 15(2), and the complete proof is given in Appendix C. The proof is similar to the purely separable case discussed in Section 3, but we must also deal with the interaction between D_s and D_c.

Assume $D_c \neq \emptyset$, and $\lim_{B\to\infty} \bar{\boldsymbol{w}}(B)/B = \bar{\boldsymbol{u}} \in S^\perp$. Define

$$\bar{\gamma} := \min_{(\boldsymbol{x}_i, y_i)\in D_c} y_i\langle \bar{\boldsymbol{u}}, \boldsymbol{x}_i\rangle.$$

Similar to the separable case, it holds that $\bar{\gamma} > 0$.

Lemma 16 *Under the conditions of Theorem 15, it holds that $\bar{\gamma} \geq \hat{\gamma}^2/\left(8|D_c|\right) > 0$.*

The proof of Lemma 16 is similar to the proof of Lemma 7, but uses the fact that $\Pi_\perp \bar{\boldsymbol{w}}(B)$ is collinear with $\Pi_\perp \nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)$.

The following result extends Lemma 9 to the general setting.

Lemma 17 *Under the conditions of Theorem 15, given any $\alpha \in (0,1)$, there exists $\xi(\alpha) > 0$, such that for any $\boldsymbol{w}$ with $\mathcal{R}(\boldsymbol{w}) - \inf_{\boldsymbol{w}\in\mathbb{R}^d} \mathcal{R}(\boldsymbol{w}) \leq \xi(\alpha)$, it holds that*

$$\mathcal{R}\left(\boldsymbol{w}_S + (1+\alpha)\left\|\boldsymbol{w}_\perp\right\|\bar{\boldsymbol{u}}\right) \leq \mathcal{R}(\boldsymbol{w}).$$

The proof of the "only if" part of Theorem 15(2) is similarly based on Lemma 17 and a perceptron-style analysis. Unlike the purely separable case, the tricky part here is that D_c may have a nonzero projection onto S, and thus we need to deal with $\boldsymbol{w}_{t,S}$ carefully. Note that convexity and Lemma 17 ensure that for large enough t,

$$\begin{aligned}
\langle \nabla\mathcal{R}(\boldsymbol{w}_t), \boldsymbol{w}_{t,\perp} - (1+\alpha)\left\|\boldsymbol{w}_{t,\perp}\right\|\bar{\boldsymbol{u}}\rangle &= \langle \nabla\mathcal{R}(\boldsymbol{w}_t), \boldsymbol{w}_{t,S} + \boldsymbol{w}_{t,\perp} - \boldsymbol{w}_{t,S} - (1+\alpha)\left\|\boldsymbol{w}_{t,\perp}\right\|\bar{\boldsymbol{u}}\rangle \\
&= \langle \nabla\mathcal{R}(\boldsymbol{w}_t), \boldsymbol{w}_t - \boldsymbol{w}_{t,S} - (1+\alpha)\left\|\boldsymbol{w}_{t,\perp}\right\|\bar{\boldsymbol{u}}\rangle \\
&\geq \mathcal{R}(\boldsymbol{w}_t) - \mathcal{R}\left(\boldsymbol{w}_{t,S} + (1+\alpha)\left\|\boldsymbol{w}_{t,\perp}\right\|\bar{\boldsymbol{u}}\right) \geq 0,
\end{aligned}$$

which implies

$$\langle -\eta\nabla\mathcal{R}(\boldsymbol{w}_t), \bar{\boldsymbol{u}}\rangle \geq \frac{1}{1+\alpha}\left\langle -\eta\nabla\mathcal{R}(\boldsymbol{w}_t), \frac{\boldsymbol{w}_{t,\perp}}{\|\boldsymbol{w}_{t,\perp}\|}\right\rangle.$$

The remainder of the proof is similar to the proof of Theorem 5.

5. Concluding remarks and open problems

We have established that for a wide variety of losses, gradient descent and the regularization path converge to the same direction if either of them converges to a direction, and while many losses guarantee such convergence, the limit direction may differ across losses.

One avenue for refinement is to go back to the general studies of classification losses (e.g., Bartlett et al., 2006; Zhang, 2004). We have pointed out that polynomially-tailed losses can exhibit worse margin behavior than exponentially-tailed losses, but this does not fully explain why the former are avoided in practice (and in theory). What are some further consequences on time and sample complexity of these two loss classes?

Another question is the role of early stopping. We have established that one can stop a gradient method after a long-enough training and obtain a predictor with roughly the same direction as a minimally-regularized predictor. This, however, requires fairly *late* stopping; what happens for general losses with aggressively early stopping? Moreover, could these observations justify the low levels of regularization encountered in practice?

Lastly, our analysis here does not distinguish the logistic and exponential losses; meanwhile, the logistic loss (and cross-entropy loss) are dominant in the practice of classification. What is a more refined picture for these two losses? Does it boil down to the Lipschitz properties of the logistic loss, or is there more?

Acknowledgments

ZJ and MT are grateful to the NSF for support under grant IIS-1750051, and to Microsoft Research for hosting them during various parts of this work.

References

Peter L. Bartlett. For valid generalization the size of the weights is more important than the size of the network. In *NIPS*, 1996.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *NIPS*, 2017.

Zachary Charles, Shashank Rajput, Stephen Wright, and Dimitris Papailiopoulos. Convergence and margin of adversarial training on separable data. *arXiv preprint arXiv:1905.09209*, 2019.

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.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *NeurIPS*, pages 9461–9471, 2018.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *ICLR*, 2019a.

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. In *COLT*, 2019b.

Yan Li, Ethan X Fang, Huan Xu, and Tuo Zhao. Implicit bias of gradient descent based adversarial training on separable data. In *ICLR*, 2020.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *ICLR*, 2020.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Albert B.J. Novikoff. On convergence proofs on perceptrons. *In Proceedings of the Symposium on the Mathematical Theory of Automata*, 12:615–622, 1962.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a regularized path to a maximum margin classifier. *JMLR*, 5:941–973, 2004.

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.

J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Trans. Inf. Theor.*, 44(5):1926–1940, September 1998.

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

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

Vladimir Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, Berlin, Heidelberg, 1982.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Neurips*, pages 9709–9721, 2019.

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

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32:56–85, 2004.

Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33:1538–1579, 2005.

Peng Zhao and Bin Yu. Stagewise lasso. *JMLR*, 8(Dec):2701–2726, 2007.

Appendix A. Omitted proofs from Section 2

Proof (of Lemma 2) For any $\bar{\boldsymbol{w}} \in \mathbb{R}^d$, it holds that

$$\begin{aligned}
\|\boldsymbol{w}_{t+1} - \boldsymbol{w}\|^2 &= \|\boldsymbol{w}_t - \boldsymbol{w}\|^2 - 2\eta \langle \nabla f(\boldsymbol{w}_t), \boldsymbol{w}_t - \boldsymbol{w}\rangle + \eta^2 \|\nabla f(\boldsymbol{w}_t)\|^2 \\
&= \|\boldsymbol{w}_t - \boldsymbol{w}\|^2 + 2\eta \langle \nabla f(\boldsymbol{w}_t), \boldsymbol{w} - \boldsymbol{w}_t\rangle + 2\eta \cdot \frac{\eta}{2} \|\nabla f(\boldsymbol{w}_t)\|^2 \\
&\le \|\boldsymbol{w}_t - \boldsymbol{w}\|^2 + 2\eta (f(\boldsymbol{w}) - f(\boldsymbol{w}_t)) + 2\eta (f(\boldsymbol{w}_t) - f(\boldsymbol{w}_{t+1})) \\
&= \|\boldsymbol{w}_t - \boldsymbol{w}\|^2 + 2\eta (f(\boldsymbol{w}) - f(\boldsymbol{w}_{t+1})) .
\end{aligned} \tag{14}$$

On the third line we use the convexity of f and eq. (5).

Since $f(\boldsymbol{w}_t)$ is nondecreasing, $\lim_{t\to\infty} f(\boldsymbol{w}_t)$ exists. Suppose $\lim_{t\to\infty} f(\boldsymbol{w}_t) > \inf_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w})$. Let $\bar{\boldsymbol{w}} \in \mathbb{R}^d$ satisfy $f(\bar{\boldsymbol{w}}) < \lim_{t\to\infty} f(\boldsymbol{w}_t) - \epsilon$ for some $\epsilon > 0$. It follows from eq. (14) that $\|\boldsymbol{w}_{t+1} - \bar{\boldsymbol{w}}\|^2 \le \|\boldsymbol{w}_t - \bar{\boldsymbol{w}}\|^2 - 2\eta\epsilon$ for any t, which implies $\|\boldsymbol{w}_{t+1} - \bar{\boldsymbol{w}}\|^2 \to -\infty$, which is a contradiction. ■

Proof (of Theorem 4) First we show that for any $\epsilon > 0$, there exists $B_1(\epsilon) > 0$, such that for any gradient descent iterate $\boldsymbol{w}_t$ with $\|\boldsymbol{w}_t\| > B_1(\epsilon)$, it holds that $\|{\boldsymbol{w}_t}/{\|\boldsymbol{w}_t\|} - \bar{\boldsymbol{u}}\| < \epsilon$. Given any ϵ, by our assumption, there exists t_1 such that $\|{\boldsymbol{w}_t}/{\|\boldsymbol{w}_t\|} - \bar{\boldsymbol{u}}\| < \epsilon$ for any $t > t_1$. It is enough to let $B_1(\epsilon) = \max_{0\le t\le t_1} \|\boldsymbol{w}_t\| + 1$.

Then we show that $\lim_{B\to\infty} \langle \bar{\boldsymbol{w}}(B), \bar{\boldsymbol{u}}\rangle \to \infty$. If this is not true, then there exists a constant $C > 0$ such that there exists arbitrarily large B with $\langle \bar{\boldsymbol{w}}(B), \bar{\boldsymbol{u}}\rangle < C$. Choose B_2 such that

$$B_2 > \max\left\{5\left(\|\boldsymbol{w}_0\| + C + 1\right), B_1\left(\frac{1}{4}\right) + 1\right\}, \quad \text{and} \quad \langle \bar{\boldsymbol{w}}(B_2), \bar{\boldsymbol{u}}\rangle < C.$$

Let t_2 denote the first step such that $\|\boldsymbol{w}_{t_2}\| > B_2 - 1$. Since $B_2 - 1 > \|\boldsymbol{w}_0\|$, we have $t_2 > 0$. Moreover, the conditions of Theorem 4 (i.e., eq. (5) and $\eta \le 1/\left(2f(\boldsymbol{w}_0)\right)$) implies

$$\begin{aligned}
\|\boldsymbol{w}_{t_2} - \boldsymbol{w}_{t_2-1}\| = \eta \|\nabla f(\boldsymbol{w}_{t_2-1})\| =& \sqrt{\eta^2 \|\nabla f(\boldsymbol{w}_{t_2-1})\|^2} \\
&\le \sqrt{2\eta\left(f(\boldsymbol{w}_{t_2-1}) - f(\boldsymbol{w}_{t_2})\right)} \\
&\le \sqrt{2\eta f(\boldsymbol{w}_0)} \le 1.
\end{aligned} \tag{15}$$

Therefore from the definition of t_2,

$$\|\boldsymbol{w}_{t_2}\| \le \|\boldsymbol{w}_{t_2} - 1\| + \|\boldsymbol{w}_{t_2} - \boldsymbol{w}_{t_2-1}\| \le B_2 - 1 + 1 = B_2.$$

By the definition of t_2 and $\bar{\boldsymbol{w}}(B_2)$, we have $f\left(\bar{\boldsymbol{w}}(B_2)\right) \le f(\boldsymbol{w}_t)$ for any $t \le t_2$. Using eq. (6), we can show that

$$\|\boldsymbol{w}_{t_2} - \bar{\boldsymbol{w}}(B_2)\| \le \|\boldsymbol{w}_0 - \bar{\boldsymbol{w}}(B_2)\| . \tag{16}$$

On one hand,

$$\|\boldsymbol{w}_0 - \bar{\boldsymbol{w}}(B_2)\| \le \|\boldsymbol{w}_0\| + \|\bar{\boldsymbol{w}}(B_2)\| = \|\boldsymbol{w}_0\| + B_2. \tag{17}$$

On the other hand,

$$\begin{aligned}\|\boldsymbol{w}_{t_2}-\bar{\boldsymbol{w}}(B_2)\|^2 &= \|\boldsymbol{w}_{t_2}\|^2+B_2^2-2\left\langle \boldsymbol{w}_{t_2},\bar{\boldsymbol{w}}(B_2)\right\rangle \\ &= \|\boldsymbol{w}_{t_2}\|^2+B_2^2-2\|\boldsymbol{w}_{t_2}\|\left\langle \frac{\boldsymbol{w}_{t_2}}{\|\boldsymbol{w}_{t_2}\|},\bar{\boldsymbol{w}}(B_2)\right\rangle \\ &> (B_2-1)^2+B_2^2-2\|\boldsymbol{w}_{t_2}\|\left\langle \frac{\boldsymbol{w}_{t_2}}{\|\boldsymbol{w}_{t_2}\|},\bar{\boldsymbol{w}}(B_2)\right\rangle.\end{aligned}$$

By the definition of t_2 and B_2, we have

$$\|\boldsymbol{w}_{t_2}\|>B_2-1>B_1\left(\frac{1}{4}\right),$$

and thus $\|{}^{\boldsymbol{w}_{t_2}}/_{\|\boldsymbol{w}_{t_2}\|}-\bar{\boldsymbol{u}}\|<1/4$. As a result,

$$\left\langle \frac{\boldsymbol{w}_{t_2}}{\|\boldsymbol{w}_{t_2}\|},\bar{\boldsymbol{w}}(B_2)\right\rangle < \left\langle \bar{\boldsymbol{u}},\bar{\boldsymbol{w}}(B_2)\right\rangle+\frac{1}{4}B_2<C+\frac{1}{4}B_2,$$

and

$$\begin{aligned}\|\boldsymbol{w}_{t_2}-\bar{\boldsymbol{w}}(B_2)\|^2 &> (B_2-1)^2+B_2^2-2\|\boldsymbol{w}_{t_2}\|C-\frac{1}{2}\|\boldsymbol{w}_{t_2}\|B_2 \\ &\ge (B_2-1)^2+B_2^2-2CB_2-\frac{1}{2}B_2^2>\frac{3}{2}B_2^2-2CB_2-2B_2. \end{aligned}\tag{18}$$

Combining eqs. (16) to (18) gives

$$\frac{3}{2}B_2^2-2CB_2-2B_2<\|\boldsymbol{w}_0\|^2+2\|\boldsymbol{w}_0\|B_2+B_2^2,$$

which implies

$$B_2<4\left(\|\boldsymbol{w}_0\|+C+1\right)+\frac{2\|\boldsymbol{w}_0\|^2}{B_2}<4\left(\|\boldsymbol{w}_0\|+C+1\right)+\|\boldsymbol{w}_0\|<5\left(\|\boldsymbol{w}_0\|+C+1\right),$$

a contradiction.

Next we prove the claim that $\lim_{B\to\infty}\bar{\boldsymbol{w}}(B)/B=\bar{\boldsymbol{u}}$. If this is not true, then there exists $\delta>0$, such that there exists arbitrarily large B with $\|{}^{\bar{\boldsymbol{w}}(B)}/_B-\bar{\boldsymbol{u}}\|>\delta$. Choose B_4 such that

$$\left\|\frac{\bar{\boldsymbol{w}}(B_4)}{B_4}-\bar{\boldsymbol{u}}\right\|>\delta,\quad\text{and}\quad\left\langle\bar{\boldsymbol{w}}(B_4),\bar{\boldsymbol{u}}\right\rangle>B_1\left(\frac{\delta^3}{32}\right)+\|\boldsymbol{w}_0\|+1,\quad\text{and}\quad B_4>\frac{32}{\delta^3}.$$

Let $B_3:=\left\langle\bar{\boldsymbol{w}}(B_4),\bar{\boldsymbol{u}}\right\rangle$. By geometric arguments, we have

$$\|\bar{\boldsymbol{w}}(B_4)-B_4\bar{\boldsymbol{u}}\|-\|\bar{\boldsymbol{w}}(B_4)-B_3\bar{\boldsymbol{u}}\|>\frac{B_4\delta^3}{8}.\tag{19}$$

Let t_3 denote the first step such that $\|\boldsymbol{w}_{t_3}\|>B_3-1$. Since $B_3-1>\|\boldsymbol{w}_0\|$, we have $t_3>0$, and similar to eq. (15) we can show that $\|\boldsymbol{w}_{t_3}\|\le B_3$. Since $B_3-1>B_1(\delta^3/32)$, we have $\|{}^{\boldsymbol{w}_{t_3}}/_{\|\boldsymbol{w}_{t_3}\|}-\bar{\boldsymbol{u}}\|<\delta^3/32$. As a result,

$$\|\boldsymbol{w}_{t_3}-B_3\bar{\boldsymbol{u}}\|\le\|\boldsymbol{w}_{t_3}-\|\boldsymbol{w}_{t_3}\|\bar{\boldsymbol{u}}\|+\|\|\boldsymbol{w}_{t_3}\|\bar{\boldsymbol{u}}-B_3\bar{\boldsymbol{u}}\|\le\|\boldsymbol{w}_{t_3}\|\frac{\delta^3}{32}+1\le\frac{B_3\delta^3}{32}+1\le\frac{B_4\delta^3}{32}+1.\tag{20}$$

Similarly, let t_4 denote the first step such that $\|\boldsymbol{w}_{t_4}\| > B_4 - 1$, we can show that $\|\boldsymbol{w}_{t_4}\| \le B_4$, and

$$\|\boldsymbol{w}_{t_4} - B_4\bar{\boldsymbol{u}}\| \le \frac{B_4\delta^3}{32} + 1. \tag{21}$$

Combining eqs. (19) to (21) gives

$$\begin{aligned}
&\|\bar{\boldsymbol{w}}(B_4) - \boldsymbol{w}_{t_4}\| - \|\bar{\boldsymbol{w}}(B_4) - \boldsymbol{w}_{t_3}\| \\
\ge{}& \|\bar{\boldsymbol{w}}(B_4) - B_4\bar{\boldsymbol{u}}\| - \|B_4\bar{\boldsymbol{u}} - \boldsymbol{w}_{t_4}\| - \|\bar{\boldsymbol{w}}(B_4) - B_3\bar{\boldsymbol{u}}\| - \|B_3\bar{\boldsymbol{u}} - \boldsymbol{w}_{t_3}\| \\
\ge{}& \frac{B_4\delta^3}{8} - \frac{B_4\delta^3}{32} - 1 - \frac{B_4\delta^3}{32} - 1 \\
={}& \frac{B_4\delta^3}{16} - 2 > 0.
\end{aligned} \tag{22}$$

On the other hand, using eq. (19) and the triangle inequality,

$$B_4 - B_3 = \|B_4\bar{\boldsymbol{u}} - B_3\bar{\boldsymbol{u}}\| \ge \|\bar{\boldsymbol{w}}(B_4) - B_4\bar{\boldsymbol{u}}\| - \|\bar{\boldsymbol{w}}(B_4) - B_3\bar{\boldsymbol{u}}\| > \frac{B_4\delta^3}{8} > 4,$$

and thus $t_4 > t_3$. Since $\|\boldsymbol{w}_{t_4}\| \le B_4$, by the definition of t_4 and $\bar{\boldsymbol{w}}(B_4)$, we have $f\left(\bar{\boldsymbol{w}}(B_4)\right) \le f(\boldsymbol{w}_t)$ for any $t \le t_4$. Since $t_3 < t_4$, eq. (6) implies $\|\bar{\boldsymbol{w}}(B_4) - \boldsymbol{w}_{t_4}\| \le \|\bar{\boldsymbol{w}}(B_4) - \boldsymbol{w}_{t_3}\|$, which contradicts eq. (22). ■

Appendix B. Omitted proofs from Section 3

We first verify that if ℓ is β-smooth, then $\mathcal{R}$ is also β-smooth. Given $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$, we have

$$\begin{aligned}
\left\|\nabla\mathcal{R}(\boldsymbol{w}) - \nabla\mathcal{R}(\boldsymbol{w}')\right\| &= \left\|\frac{1}{n}\sum_{i=1}^n \ell'\left(y_i\langle\boldsymbol{w},\boldsymbol{x}_i\rangle\right)y_i\boldsymbol{x}_i - \frac{1}{n}\sum_{i=1}^n \ell'\left(y_i\langle\boldsymbol{w}',\boldsymbol{x}_i\rangle\right)y_i\boldsymbol{x}_i\right\| \\
&\le \frac{1}{n}\sum_{i=1}^n \left|\ell'\left(y_i\langle\boldsymbol{w},\boldsymbol{x}_i\rangle\right) - \ell'\left(y_i\langle\boldsymbol{w}',\boldsymbol{x}_i\rangle\right)\right| \|y_i\boldsymbol{x}_i\| \\
&\le \frac{1}{n}\sum_{i=1}^n \beta\left|y_i\langle\boldsymbol{w},\boldsymbol{x}_i\rangle - y_i\langle\boldsymbol{w}',\boldsymbol{x}_i\rangle\right| \\
&\le \frac{1}{n}\sum_{i=1}^n \beta\|\boldsymbol{w} - \boldsymbol{w}'\|\|y_i\boldsymbol{x}_i\| \le \beta\|\boldsymbol{w} - \boldsymbol{w}'\|.
\end{aligned}$$

Therefore $\mathcal{R}$ is β-smooth.

To proceed, we first need the following lemma.

Lemma 18 *It holds that*

$$\frac{\bar{\boldsymbol{w}}(B)}{B} = -\frac{\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)}{\left\|\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)\right\|}.$$

Conversely, if $\|\boldsymbol{w}\| = B$ *and* $\boldsymbol{w}/B = -\nabla\mathcal{R}(\boldsymbol{w})/\left\|\nabla\mathcal{R}(\boldsymbol{w})\right\|$*, then* $\boldsymbol{w} = \bar{\boldsymbol{w}}(B)$*.*

Proof By the first order optimality conditions, $\boldsymbol{w} = \bar{\boldsymbol{w}}(B)$ if and only if for any $\boldsymbol{w}'$ with $\|\boldsymbol{w}'\|_2 \leq B$, it holds that

$$\left\langle \nabla\mathcal{R}(\boldsymbol{w}), \boldsymbol{w}' - \boldsymbol{w} \right\rangle \geq 0. \tag{23}$$

Since the infimum of $\mathcal{R}$ is not attained, the gradient $\nabla\mathcal{R}(\boldsymbol{w})$ is always nonzero. The structure of the ℓ_2 ball implies that eq. (23) holds if and only if $\|\boldsymbol{w}\| = B$ and $\boldsymbol{w}/B = -\nabla\mathcal{R}(\boldsymbol{w})/\left\|\nabla\mathcal{R}(\boldsymbol{w})\right\|$. ■

Proof (of Lemma 7) Since $\bar{\boldsymbol{w}}(B)/B \to \bar{\boldsymbol{u}}$, the margin of $\bar{\boldsymbol{w}}(B)/B$ converges to the margin of $\bar{\boldsymbol{u}}$. For large enough B, the risk $\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right) \leq \ell(0)/n$, which implies $\bar{\boldsymbol{w}}(B)/B$ has a nonnegative margin, and thus $\bar{\boldsymbol{u}}$ also has a nonnegative margin.

The proof of Lemma 7 is by contradiction. Given $\epsilon := \hat{\gamma}^2/(2n)$, suppose there exists $B_0 > 0$, such that for any $B \geq B_0$, the margin of $\bar{\boldsymbol{w}}(B)/B$ is no larger than ϵ. We will derive a contradiction, which implies that the margin of $\bar{\boldsymbol{u}}$ is at least $\hat{\gamma}^2/(2n)$.

For any $B > 0$, Lemma 18 ensures that

$$-\left\langle \frac{\bar{\boldsymbol{w}}(B)}{B}, \nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right) \right\rangle = \left\|\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)\right\|. \tag{24}$$

For simplicity, let $\boldsymbol{z}_i := y_i \boldsymbol{x}_i$. The left hand side of eq. (24) can be rewritten as

$$\frac{1}{n}\sum_{i=1}^{n} -\ell'\left(\langle \bar{\boldsymbol{w}}(B), \boldsymbol{z}_i \rangle\right) \left\langle \frac{\bar{\boldsymbol{w}}(B)}{B}, \boldsymbol{z}_i \right\rangle, \tag{25}$$

while the right hand side of eq. (24) can be bounded below as

$$\left\|\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)\right\| \geq \left\langle -\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right), \hat{\boldsymbol{u}} \right\rangle \geq \frac{1}{n}\sum_{i=1}^{n} -\ell'\left(\langle \bar{\boldsymbol{w}}(B), \boldsymbol{z}_i \rangle\right)\hat{\gamma}, \tag{26}$$

where $\hat{\boldsymbol{u}}$ denotes the unit maximum margin predictor. Let H denote the set of data points on which $\bar{\boldsymbol{w}}(B)/B$ has margin larger than $\hat{\gamma}$, and suppose without loss of generality that $\bar{\boldsymbol{w}}(B)/B$ achieves its minimum margin on $\boldsymbol{z}_1$. It follows from eqs. (24) to (26) that

$$\begin{aligned}\sum_{\boldsymbol{z}_i \in H} -\ell'\left(\langle \bar{\boldsymbol{w}}(B), \boldsymbol{z}_i \rangle\right)\left(\left\langle \frac{\bar{\boldsymbol{w}}(B)}{B}, \boldsymbol{z}_i \right\rangle - \hat{\gamma}\right) &\geq \sum_{\boldsymbol{z}_i \notin H} -\ell'\left(\langle \bar{\boldsymbol{w}}(B), \boldsymbol{z}_i \rangle\right)\left(\hat{\gamma} - \left\langle \frac{\bar{\boldsymbol{w}}(B)}{B}, \boldsymbol{z}_i \right\rangle\right) \\ &\geq -\ell'\left(\langle \bar{\boldsymbol{w}}(B), \boldsymbol{z}_1 \rangle\right)\left(\hat{\gamma} - \left\langle \frac{\bar{\boldsymbol{w}}(B)}{B}, \boldsymbol{z}_1 \right\rangle\right). \end{aligned} \tag{27}$$

Now consider $B \geq B_0$, which implies $\langle \bar{\boldsymbol{w}}(B)/B, \boldsymbol{z}_1 \rangle \leq \epsilon$. Since $\epsilon < \hat{\gamma}/2$, and $\|\boldsymbol{z}_i\|_2 \leq 1$, eq. (27) implies

$$-n\ell'(B\hat{\gamma}) \geq -\ell'(B\epsilon)(\hat{\gamma} - \epsilon) \geq -\ell'(B\epsilon)\frac{\hat{\gamma}}{2},$$

and thus

$$\frac{-\ell'(B\epsilon)}{-\ell'(B\hat{\gamma})} \leq \frac{2n}{\hat{\gamma}} \tag{28}$$

for all $B \geq B_0$. Let $\alpha := B_0\epsilon = B_0\hat{\gamma}^2/(2n)$, and $\lambda := 2n/\hat{\gamma}$. For any $k \geq 1$, we have

$$\int_{\alpha\lambda^k}^{\alpha\lambda^{k+1}} -\ell'(z)\,\mathrm{d}z = \int_{\alpha\lambda^{k-1}}^{\alpha\lambda^k} -\ell'(\lambda y)\lambda\,\mathrm{d}y \geq \int_{\alpha\lambda^{k-1}}^{\alpha\lambda^k} -\ell'(y)\,\mathrm{d}y,$$

where eq. (28) is used. By induction, we have

$$\int_{\alpha\lambda^k}^{\alpha\lambda^{k+1}} -\ell'(z)\,\mathrm{d}z \geq \int_{\alpha}^{\alpha\lambda} -\ell'(z)\,\mathrm{d}z > 0.$$

As a result,

$$\int_{\alpha}^{\infty} -\ell'(z)\,\mathrm{d}z = \infty,$$

which is contradiction, since $\int_\alpha^\infty -\ell'(z)\,\mathrm{d}z = \ell(\alpha)$ should be finite. ■

When the loss function has a polynomial tail az^{-b}, then we can use eq. (28) to prove a margin lower bound of $\hat{\gamma}^{(b+2)/(b+1)}n^{-1/(b+1)}$. The dependency on n cannot be improved in general (cf. Proposition 12).

Proof (of Lemma 9) Since $\lim_{B\to\infty} \bar{\boldsymbol{w}}(B)/B = \bar{\boldsymbol{u}}$, we can choose $\rho(\alpha)$ large enough such that for any $\boldsymbol{w}$ with $\|\boldsymbol{w}\| > \rho(\alpha)$, it holds that

$$\left\|\bar{\boldsymbol{w}}\left(\|\boldsymbol{w}\|\right)/\|\boldsymbol{w}\| - \bar{\boldsymbol{u}}\right\| \leq \alpha\bar{\gamma}.$$

In this case, for any $1 \leq i \leq n$,

$$\begin{aligned} y_i\left\langle \bar{\boldsymbol{w}}\left(\|\boldsymbol{w}\|\right), \boldsymbol{x}_i\right\rangle &= y_i\left\langle \bar{\boldsymbol{w}}\left(\|\boldsymbol{w}\|\right) - \|\boldsymbol{w}\|\bar{\boldsymbol{u}}, \boldsymbol{x}_i\right\rangle + y_i\left\langle \|\boldsymbol{w}\|\bar{\boldsymbol{u}}, \boldsymbol{x}_i\right\rangle \\ &\leq \alpha\bar{\gamma}\|\boldsymbol{w}\| + y_i\left\langle \|\boldsymbol{w}\|\bar{\boldsymbol{u}}, \boldsymbol{x}_i\right\rangle \\ &\leq y_i\left\langle (1+\alpha)\|\boldsymbol{w}\|\bar{\boldsymbol{u}}, \boldsymbol{x}_i\right\rangle. \end{aligned}$$

As a result,

$$\mathcal{R}\left((1+\alpha)\left\|\boldsymbol{w}\right\|\bar{\boldsymbol{u}}\right) \leq \mathcal{R}\left(\bar{\boldsymbol{w}}\left(\|\boldsymbol{w}\|\right)\right) \leq \mathcal{R}(\boldsymbol{w}).$$

■

Next we prove the "only if" part of Theorem 5.

Proof (of Theorem 5, the "only if" part) Given any $\epsilon \in (0,1)$, let α satisfy $1/(1+\alpha) = 1-\epsilon$ (i.e., let $\alpha = \epsilon/(1-\epsilon)$). Since $\lim_{t\to\infty}\|\boldsymbol{w}_t\| = \infty$, we can choose a step t_0 such that for any $t \geq t_0$, it holds that $\|\boldsymbol{w}_t\| > \max\{\rho(\alpha), 1\}$, where ρ is given by Lemma 9.

Now for any $t \geq t_0$, using the convexity of $\mathcal{R}$ and Lemma 9, we have

$$\left\langle \nabla\mathcal{R}(\boldsymbol{w}_t), \boldsymbol{w}_t - (1+\alpha)\|\boldsymbol{w}_t\|\bar{\boldsymbol{u}}\right\rangle \geq \mathcal{R}(\boldsymbol{w}_t) - \mathcal{R}\left((1+\alpha)\|\boldsymbol{w}_t\|\bar{\boldsymbol{u}}\right) \geq 0,$$

meaning

$$\left\langle \nabla\mathcal{R}(\boldsymbol{w}_t), \boldsymbol{w}_t\right\rangle \geq (1+\alpha)\|\boldsymbol{w}_t\|\left\langle \nabla\mathcal{R}(\boldsymbol{w}_t), \bar{\boldsymbol{u}}\right\rangle.$$

Consequently,

$$\begin{aligned}\langle \boldsymbol{w}_{t+1} - \boldsymbol{w}_t, \bar{\boldsymbol{u}} \rangle &= \langle -\eta \nabla \mathcal{R}(\boldsymbol{w}_t), \bar{\boldsymbol{u}} \rangle \\ &\geq \langle -\eta \nabla \mathcal{R}(\boldsymbol{w}_t), \boldsymbol{w}_t \rangle \frac{1}{(1+\alpha)\|\boldsymbol{w}_t\|} \\ &= \langle \boldsymbol{w}_{t+1} - \boldsymbol{w}_t, \boldsymbol{w}_t \rangle \frac{1}{(1+\alpha)\|\boldsymbol{w}_t\|} \\ &= \left(\frac{1}{2}\|\boldsymbol{w}_{t+1}\|^2 - \frac{1}{2}\|\boldsymbol{w}_t\|^2 - \frac{1}{2}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2 \right) \frac{1}{(1+\alpha)\|\boldsymbol{w}_t\|}.\end{aligned}$$

On one hand, we have

$$\left(\frac{1}{2}\|\boldsymbol{w}_{t+1}\|^2 - \frac{1}{2}\|\boldsymbol{w}_t\|^2 \right) / \|\boldsymbol{w}_t\| \geq \|\boldsymbol{w}_{t+1}\| - \|\boldsymbol{w}_t\|.$$

On the other hand, using the step size condition in eq. (7), we have

$$\frac{\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2}{2(1+\alpha)\|\boldsymbol{w}_t\|} \leq \frac{\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2}{2} = \frac{\eta^2 \left\|\nabla \mathcal{R}(\boldsymbol{w}_t)\right\|^2}{2} \leq \eta \left(\mathcal{R}(\boldsymbol{w}_t) - \mathcal{R}(\boldsymbol{w}_{t+1}) \right).$$

As a result,

$$\langle \boldsymbol{w}_t - \boldsymbol{w}_{t_0}, \bar{\boldsymbol{u}} \rangle \geq \frac{\|\boldsymbol{w}_t\| - \|\boldsymbol{w}_{t_0}\|}{1+\alpha} - \eta \mathcal{R}(\boldsymbol{w}_{t_0}) = (1-\epsilon) \left(\|\boldsymbol{w}_t\| - \|\boldsymbol{w}_{t_0}\| \right) - \eta \mathcal{R}(\boldsymbol{w}_{t_0}),$$

meaning

$$\left\langle \frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|}, \bar{\boldsymbol{u}} \right\rangle \geq 1 - \epsilon + \frac{\langle \boldsymbol{w}_{t_0}, \bar{\boldsymbol{u}} \rangle - (1-\epsilon)\|\boldsymbol{w}_{t_0}\| - \eta \mathcal{R}(\boldsymbol{w}_{t_0})}{\|\boldsymbol{w}_t\|}.$$

Consequently,

$$\liminf_{t \to \infty} \left\langle \frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|}, \bar{\boldsymbol{u}} \right\rangle \geq 1 - \epsilon.$$

Since ϵ is arbitrary, we get $\boldsymbol{w}_t / \|\boldsymbol{w}_t\| \to \bar{\boldsymbol{u}}$. ■

B.1. Omitted proofs from Section 3.1

Proof (of Proposition 10) First let us verify that the maximum-margin solution $\hat{\boldsymbol{u}}$ is unique. If this is not true, suppose there exist two unit vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ which both attain the maximum margin $\hat{\gamma}$ but $\boldsymbol{u}_1 \neq \boldsymbol{u}_2$. Consider $\boldsymbol{u}_3 = (\boldsymbol{u}_1 + \boldsymbol{u}_2)/2$. Then for any i, it holds that

$$y_i \langle \boldsymbol{u}_3, \boldsymbol{x}_i \rangle = y_i \langle \boldsymbol{u}_1, \boldsymbol{x}_i \rangle / 2 + y_i \langle \boldsymbol{u}_2, \boldsymbol{x}_i \rangle / 2 \geq \hat{\gamma},$$

and thus $\boldsymbol{u}_3$ also maximizes the margin. However, since $\boldsymbol{u}_1 \neq \boldsymbol{u}_2$, it follows that $\|\boldsymbol{u}_3\| \leq 1$. Consequently, the unit vector $\boldsymbol{u}_3 / \|\boldsymbol{u}_3\|$ should achieve a margin larger than $\hat{\gamma}$, which is a contradiction.

Now note that

$$\mathcal{R}(B\hat{\boldsymbol{u}}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i\langle B\hat{\boldsymbol{u}}, \boldsymbol{x}_i\rangle\right) \le \ell(B\hat{\gamma}).$$

When B is large enough, we have

$$\mathcal{R}(B\hat{\boldsymbol{u}}) \le \ell(B\hat{\gamma}) \le 2a\exp(-bB\hat{\gamma}).$$

Now suppose Proposition 10 is not true. Then there exists $\epsilon > 0$, such that there exists arbitrarily large B with $\|{\bar{\boldsymbol{w}}(B)}/{B} - \hat{\boldsymbol{u}}\| > \epsilon$. Since $\hat{\boldsymbol{u}}$ is the unique maximum-margin solution, it follows that there exists $\epsilon' \in (0, \hat{\gamma})$ such that

$$\min_{1\le i\le n} y_i\left\langle \frac{\bar{\boldsymbol{w}}(B)}{B}, \boldsymbol{x}_i\right\rangle \le \hat{\gamma} - \epsilon',$$

and thus

$$\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right) \ge \frac{1}{n}\ell\left(B(\hat{\gamma} - \epsilon')\right).$$

For large enough B, it follows that

$$\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right) \ge \frac{1}{n}\ell\left(B(\hat{\gamma} - \epsilon')\right) \ge \frac{a}{2n}\exp\left(-bB(\hat{\gamma} - \epsilon')\right) = a\exp\left(-bB\hat{\gamma}\right)\frac{\exp(bB\epsilon')}{2n}.$$

Since B can be arbitrarily large, the factor $\exp(bB\epsilon')/(2n)$ can also be arbitrarily large, which would give $\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right) > \mathcal{R}(B\hat{\boldsymbol{u}})$, a contradiction. ■

Proof (of Proposition 11) The fundamental theorem of calculus implies $\ell(z) = \int_z^\infty -\ell'(z)\,\mathrm{d}z$, and thus $b > 1$. We consider the loss function

$$\tilde{\ell}(z) := \begin{cases} \dfrac{a}{b-1}z^{-b+1}, & \text{if } z \ge 1, \\ -az + \dfrac{ab}{b-1}, & \text{if } z < 1. \end{cases}$$

It can be verified that $\tilde{\ell}$ is convex, differentiable, and strictly decreasing to 0. Moreover, we have $-\tilde{\ell}'(z) = az^{-b}$ for $z \ge 1$.

Let $\widetilde{\mathcal{R}}$ denote the empirical risk function using loss $\tilde{\ell}$. Let B_0 be large enough such that

$$\min_{\boldsymbol{w}:\|\boldsymbol{w}\|_2\le B_0} \widetilde{\mathcal{R}}(\boldsymbol{w}) < \frac{1}{n}\tilde{\ell}(1) = \frac{a}{n(b-1)},$$

and let $\bar{\boldsymbol{u}}$ denote the direction of the optimal solution:

$$\operatorname*{arg\,min}_{\boldsymbol{w}:\|\boldsymbol{w}\|_2\le B_0} \widetilde{\mathcal{R}}(\boldsymbol{w}) = B_0\bar{\boldsymbol{u}}.$$

Due to Lemma 18, we have

$$\bar{\boldsymbol{u}} = -\frac{\nabla\widetilde{\mathcal{R}}(B_0\bar{\boldsymbol{u}})}{\left\|\nabla\widetilde{\mathcal{R}}(B_0\bar{\boldsymbol{u}})\right\|} = -\frac{1}{\left\|\nabla\widetilde{\mathcal{R}}(B_0\bar{\boldsymbol{u}})\right\|}\frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}'\left(y_i\langle B_0\bar{\boldsymbol{u}}, \boldsymbol{x}_i\rangle\right)y_i\boldsymbol{x}_i.$$

Since $\widetilde{\mathcal{R}}(B_0\bar{\boldsymbol{u}}) < \tilde{\ell}(1)/n$, it follows that $y_i\langle B_0\bar{\boldsymbol{u}}, \boldsymbol{x}_i\rangle > 1$ for all i, and thus

$$\bar{\boldsymbol{u}} = -\frac{1}{\left\|\nabla\widetilde{\mathcal{R}}(B_0\bar{\boldsymbol{u}})\right\|}\frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}'\left(y_i\langle B_0\bar{\boldsymbol{u}}, \boldsymbol{x}_i\rangle\right)y_i\boldsymbol{x}_i = \frac{1}{\left\|\nabla\widetilde{\mathcal{R}}(B_0\bar{\boldsymbol{u}})\right\|}\frac{1}{n}\sum_{i=1}^{n}a\left(y_i\langle B_0\bar{\boldsymbol{u}}, \boldsymbol{x}_i\rangle\right)^{-b}y_i\boldsymbol{x}_i.$$

The direction of the right hand side does not depend on B_0 due to the polynomial tail, and thus for any $B > B_0$, we have

$$\bar{\boldsymbol{u}} = -\frac{\nabla\widetilde{\mathcal{R}}(B\bar{\boldsymbol{u}})}{\left\|\nabla\widetilde{\mathcal{R}}(B\bar{\boldsymbol{u}})\right\|},$$

and thus Lemma 18 ensures

$$\underset{\boldsymbol{w}:\|\boldsymbol{w}\|_2\le B}{\arg\min}\ \widetilde{\mathcal{R}}(\boldsymbol{w}) = B\bar{\boldsymbol{u}}.$$

Now we consider the original loss ℓ. We claim that $\lim_{B\to\infty}\bar{\boldsymbol{w}}(B)/B \to \bar{\boldsymbol{u}}$. First note that $\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)/\left\|\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)\right\|$ and $\nabla\widetilde{\mathcal{R}}\left(\bar{\boldsymbol{w}}(B)\right)/\left\|\nabla\widetilde{\mathcal{R}}\left(\bar{\boldsymbol{w}}(B)\right)\right\|$ can become arbitrarily close as $B\to\infty$. To see this, define

$$q_i(B) := \frac{\ell'\left(y_i\left\langle\bar{\boldsymbol{w}}(B), \boldsymbol{x}_i\right\rangle\right)}{\sum_{j=1}^{n}\ell'\left(y_j\left\langle\bar{\boldsymbol{w}}(B), \boldsymbol{x}_j\right\rangle\right)}, \quad \text{and} \quad \tilde{q}_i(B) := \frac{\tilde{\ell}'\left(y_i\left\langle\bar{\boldsymbol{w}}(B), \boldsymbol{x}_i\right\rangle\right)}{\sum_{j=1}^{n}\tilde{\ell}'\left(y_j\left\langle\bar{\boldsymbol{w}}(B), \boldsymbol{x}_j\right\rangle\right)}.$$

Note that

$$-\frac{\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)}{\left\|\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)\right\|} = \frac{\sum_{i=1}^{n}q_i(B)y_i\boldsymbol{x}_i}{\left\|\sum_{i=1}^{n}q_i(B)y_i\boldsymbol{x}_i\right\|}, \quad \text{and} \quad -\frac{\nabla\widetilde{\mathcal{R}}\left(\bar{\boldsymbol{w}}(B)\right)}{\left\|\nabla\widetilde{\mathcal{R}}\left(\bar{\boldsymbol{w}}(B)\right)\right\|} = \frac{\sum_{i=1}^{n}\tilde{q}_i(B)y_i\boldsymbol{x}_i}{\left\|\sum_{i=1}^{n}\tilde{q}_i(B)y_i\boldsymbol{x}_i\right\|}.$$

By the conditions of Proposition 11, it holds that $|q_i(B) - \tilde{q}_i(B)| \to 0$ for all $1 \le i \le n$ as $B\to\infty$, and thus

$$\left|\left\|\sum_{i=1}^{n}q_i(B)y_i\boldsymbol{x}_i\right\| - \left\|\sum_{i=1}^{n}\tilde{q}_i(B)y_i\boldsymbol{x}_i\right\|\right| \to 0.$$

Moreover, for any $q\in\Delta_n$ (i.e., $q_i \ge 0$ and $\sum_{i=1}^{n}q_i = 1$), it holds that

$$\left\|\sum_{i=1}^{n}q_iy_i\boldsymbol{x}_i\right\| \ge \left\langle\sum_{i=1}^{n}q_iy_i\boldsymbol{x}_i, \hat{\boldsymbol{u}}\right\rangle \ge \hat{\gamma} > 0,$$

where $\hat{\boldsymbol{u}}$ and $\hat{\gamma}$ denote the maximum-margin solution and the maximum margin. Consequently $\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)/\left\|\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)\right\|$ and $\nabla\widetilde{\mathcal{R}}\left(\bar{\boldsymbol{w}}(B)\right)/\left\|\nabla\widetilde{\mathcal{R}}\left(\bar{\boldsymbol{w}}(B)\right)\right\|$ can become arbitrarily close. By Lemma 18,

$$\frac{\bar{\boldsymbol{w}}(B)}{B} = -\frac{\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)}{\left\|\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)\right\|},$$

and thus $\bar{\boldsymbol{w}}(B)/B$ and $-\nabla\widetilde{\mathcal{R}}\left(\bar{\boldsymbol{w}}(B)\right) / \left\|\nabla\widetilde{\mathcal{R}}\left(\bar{\boldsymbol{w}}(B)\right)\right\|$ can also become arbitrarily close.

Suppose $\bar{\boldsymbol{w}}(B)/B$ does not converge to $\bar{\boldsymbol{u}}$. Then there exists $\epsilon > 0$ such that there exists arbitrarily large B with $\|{}^{\bar{\boldsymbol{w}}(B)}/_{B} - \bar{\boldsymbol{u}}\| > \epsilon$. When B is large enough, $\bar{\boldsymbol{w}}(B)$ and $\nabla\widetilde{\mathcal{R}}\left(\bar{\boldsymbol{w}}(B)\right)$ can be arbitrarily close to collinear, and due to the structure of the ℓ_2 ball, we have

$$\left\langle \nabla\widetilde{\mathcal{R}}\left(\bar{\boldsymbol{w}}(B)\right), B\bar{\boldsymbol{u}} - \bar{\boldsymbol{w}}(B)\right\rangle > 0,$$

which implies that $\widetilde{\mathcal{R}}(B\bar{\boldsymbol{u}}) > \widetilde{\mathcal{R}}\left(\bar{\boldsymbol{w}}(B)\right)$, a contradiction. ■

Proof (of Proposition 12) Consider the training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ where $\boldsymbol{x}_i = (0.1, 0.1)$ for $1 \le i \le n-1$ and $\boldsymbol{x}_n = (0.6, -0.8)$, and $y_i = +1$ for all $1 \le i \le n$. Note that as we increase n, the maximum margin does not change, and thus is a universal constant. Further consider a loss function such that $\ell(z) = z^{-b}$ for $b > 0$ and $z \ge 1$. We will show that the limiting direction $\bar{\boldsymbol{u}}$ induced by ℓ satisfies

$$y_n\langle \bar{\boldsymbol{u}}, \boldsymbol{x}_n\rangle = \Theta\left(\frac{1}{n^{1/(b+1)}}\right).$$

Consequently, for large enough n it holds that $\bar{\boldsymbol{u}} \ne \hat{\boldsymbol{u}}$.

The existence of $\bar{\boldsymbol{u}}$ is ensured by Proposition 11. Let $\bar{\boldsymbol{u}} = (u_1, u_2)$, and

$$p := \frac{1}{(0.1u_1 + 0.1u_2)^{b+1}}, \quad \text{and} \quad q = \frac{1}{(0.6u_1 - 0.8u_2)^{b+1}}.$$

It follows from the proof of Proposition 11 that $p > 0$, $q > 0$, and (u_1, u_2) is collinear with $(0.1(n-1)p + 0.6q, 0.1(n-1)p - 0.8q)$. Note that we always have $u_1 > 0$, and when n is large enough, we also have $u_2 > 0$. Consequently,

$$\frac{u_1}{u_2} = \frac{0.1(n-1)p + 0.6q}{0.1(n-1)p - 0.8q} = \frac{(n-1)p/q + 6}{(n-1)p/q - 8},$$

and thus

$$\frac{p}{q} = \frac{1}{n-1}\frac{8u_1 + 6u_2}{u_1 - u_2}.$$

Since $0.6u_1 - 0.8u_2 > 0$ and $u_1^2 + u_2^2 = 1$, it can be shown that $u_1 - u_2 > 0.2$, and thus $p/q = \Theta(1/n)$. Moreover,

$$\frac{p}{q} = \frac{(0.6u_1 - 0.8u_2)^{b+1}}{(0.1u_1 + 0.1u_2)^{b+1}},$$

and thus

$$y_n\langle \bar{\boldsymbol{u}}, \boldsymbol{x}_n\rangle = 0.6u_1 - 0.8u_2 = \Theta\left(\frac{1}{n^{1/(b+1)}}\right).$$

■

To prove Proposition 13, we first need the following result which allows us to switch between different tails.

Lemma 19 *Consider the loss functions $\ell_{\text{exp}}(z) := e^{-z}$ and $\ell_{\text{recip}}(z) := 1/z$ on $[1,\infty)$. Given any $C_0 > 1$, there exists $C_1 > C_0$ and a convex loss ℓ_1 such that $\ell_1 = \ell_{\text{exp}}$ on $[1, C_0]$, and $\ell_1 = \ell_{\text{recip}}$ on $[C_1, \infty)$, and ℓ_1 is 2-smooth. Similarly, there also exists $C_2 > C_0$ and convex loss ℓ_2 such that $\ell_2 = \ell_{\text{recip}}$ on $[1, C_0]$, and $\ell_2 = \ell_{\text{exp}}$ on $[C_2, \infty)$, and ℓ_2 is 2-smooth.*

Proof (of Lemma 19) Let C_1 be large enough such that

$$\frac{1}{C_1} + \frac{1}{C_1^2}(C_1 - C_0) + \frac{1}{2}e^{-C_0} - \frac{1}{2C_1^2} < e^{-C_0}, \quad \text{and} \quad C_1 > C_0 + \frac{3}{2}. \tag{29}$$

Consider the two lines

$$f_1(z) := e^{-C_0} - e^{-C_0}(z - C_0), \quad \text{and} \quad f_2(z) := \frac{1}{C_1} - \frac{1}{C_1^2}(z - C_1) + \frac{1}{2}e^{-C_0} - \frac{1}{2C_1^2}.$$

Note that due to eq. (29), we have

$$f_1(C_0) = e^{-C_0} > \frac{1}{C_1} + \frac{1}{C_1^2}(C_1 - C_0) + \frac{1}{2}e^{-C_0} - \frac{1}{2C_1^2} = f_2(C_0),$$

and

$$\begin{aligned} f_1(C_1 - 1) &= e^{-C_0} - e^{-C_0}(C_1 - 1 - C_0) \\ &< e^{-C_0} - \frac{1}{2}e^{-C_0} \\ &< \frac{1}{2}e^{-C_0} + \frac{1}{C_1} + \frac{1}{2C_1^2} = f_2(C_1 - 1). \end{aligned}$$

Consequently, the two lines f_1 and f_2 intersect at some point $C \in (C_0, C_1 - 1)$. Now we define

$$\ell_1'(z) = \begin{cases} -e^{-C_0}, & \text{if } z \in [C_0, C], \\ -e^{-C_0} + \left(e^{-C_0} - \frac{1}{C_1^2}\right)(z - C), & \text{if } z \in [C, C+1], \\ -\frac{1}{C_1^2}, & \text{if } z \in [C+1, C_1]. \end{cases}$$

It is easy to verify that ℓ' is nondecreasing 2-Lipschitz on $[C_0, C_1]$. We only need to show that

$$\int_{C_0}^{C_1} \ell_1'(z)\,\mathrm{d}z = \frac{1}{C_1} - e^{-C_0}. \tag{30}$$

Note that

$$\int_{C_0}^{C} \ell_1'(z)\,\mathrm{d}z = -e^{-C_0}(C - C_0), \tag{31}$$

and

$$\int_{C}^{C+1} \ell_1'(z)\,\mathrm{d}z = -\frac{1}{2}e^{-C_0} - \frac{1}{2C_1^2}, \tag{32}$$

and

$$\int_{C+1}^{C_1} \ell_1'(z)\,\mathrm{d}z = -\frac{1}{C_1^2}(C_1 - C - 1). \tag{33}$$

Moreover, since f_1 and f_2 intersect at C, we have

$$e^{-C_0} - e^{-C_0}(C - C_0) = \frac{1}{C_1} - \frac{1}{C_1^2}(C - C_1) + \frac{1}{2}e^{-C_0} - \frac{1}{2C_1^2}. \tag{34}$$

Combining eqs. (31) to (34) proves eq. (30).

The proof of the other claim is similar. Let C_2 be large enough such that

$$\frac{1}{C_0} > e^{-C_2} - e^{-C_2}(C_0 - C_2) + \frac{1}{2C_0^2} - \frac{1}{2}e^{-C_2}, \quad \text{and} \quad C_2 > 2C_0 + 1.$$

Consider the two lines

$$g_1(z) := \frac{1}{C_0} - \frac{1}{C_0^2}(z - C_0), \quad \text{and} \quad g_2(z) := e^{-C_2} - e^{-C_2}(z - C_2) + \frac{1}{2C_0^2} - \frac{1}{2}e^{-C_2}.$$

It can be verified that $g_1(C_0) > g_2(C_0)$ and $g_1(C_2 - 1) < g_2(C_2 - 1)$, and thus g_1 and g_2 intersect at some point $C' \in (C_0, C_2 - 1)$. Let

$$\ell_2'(z) = \begin{cases} -\frac{1}{C_0^2}, & \text{if } z \in [C_0, C'], \\ -\frac{1}{C_0^2} + \left(\frac{1}{C_0^2} - e^{-C_2}\right)(z - C), & \text{if } z \in [C', C' + 1], \\ -e^{-C_2}, & \text{if } z \in [C + 1, C_2]. \end{cases}$$

It can be verified similarly that

$$\int_{C_0}^{C_2} \ell_2'(z)\,\mathrm{d}z = e^{-C_2} - \frac{1}{C_0}.$$

■

Next we prove Proposition 13. We make ℓ keep switching between e^{-z} and $1/z$ so that the regularization path does not converge.

Proof (of Proposition 13) In this proof, the notation $\bar{\boldsymbol{w}}_\ell(B)$ means the regularized solution using loss ℓ.

Consider the dataset given in the proof of Proposition 12. If n is large enough, then for the exponential loss e^{-z} we have

$$\lim_{B\to\infty} \frac{\bar{\boldsymbol{w}}_{\text{exp}}(B)}{B} = \hat{\boldsymbol{u}},$$

while for the reciprocal loss $1/z$ it holds that

$$\lim_{B\to\infty} \frac{\bar{\boldsymbol{w}}_{\text{recip}}(B)}{B} = \bar{\boldsymbol{u}} \neq \hat{\boldsymbol{u}}.$$

Let B_0 be large enough such that for any $B \geq B_0$,

$$\left\| \frac{\bar{\boldsymbol{w}}_{\text{exp}}(B)}{B} - \hat{\boldsymbol{u}} \right\| \leq \frac{\|\bar{\boldsymbol{u}} - \hat{\boldsymbol{u}}\|}{3}, \quad \text{and} \quad \left\| \frac{\bar{\boldsymbol{w}}_{\text{recip}}(B)}{B} - \bar{\boldsymbol{u}} \right\| \leq \frac{\|\bar{\boldsymbol{u}} - \hat{\boldsymbol{u}}\|}{3},$$

and the margin of $\bar{\boldsymbol{w}}_{\text{exp}}(B)/B$ is at least $\hat{\gamma}/2$, and the margin of $\bar{\boldsymbol{w}}_{\text{recip}}(B)/B$ is at least $\bar{\gamma}/2$.

We construct ℓ in the following way. Let $\ell(z) := z^2 - z + 1$ for $z < 0$, and $\ell(z) := e^{-z}$ for $z \in [0, B_0]$. One can verify that ℓ is convex and 1-smooth on $(-\infty, B_0]$. Let $a_0 = 0$, $b_0 = B_0$. Now for any $k \geq 1$, the construction is as follows.

1. Given $\ell = e^{-z}$ on $[a_{k-1}, b_{k-1}]$, Lemma 19 ensures that we can switch ℓ to $1/z$: there exists $c_k > b_{k-1}$ such that we can let $\ell(z) = 1/z$ for any $z \geq c_k$. We let $\ell(z) = 1/z$ on $[c_k, d_k]$ where $d_k := 2nc_k/\bar{\gamma}$. With this construction it holds that $\bar{\boldsymbol{w}}_\ell(d_k) = \bar{\boldsymbol{w}}_{\text{recip}}(d_k)$. To see this, first note that by our condition

$$y_i \left\langle \bar{\boldsymbol{w}}_{\text{recip}}(d_k), \boldsymbol{x}_i \right\rangle \geq \frac{d_k \bar{\gamma}}{2} = nc_k, \quad \text{and} \quad y_i \left\langle \bar{\boldsymbol{w}}_{\text{recip}}(d_k), \boldsymbol{x}_i \right\rangle \leq \left\| \bar{\boldsymbol{w}}_{\text{recip}}(d_k) \right\| \|\boldsymbol{x}_i\| = d_k$$

for all $1 \leq i \leq n$, which implies

$$\mathcal{R}_\ell \left(\bar{\boldsymbol{w}}_{\text{recip}}(d_k) \right) \leq \frac{1}{nc_k}.$$

On the other hand, if $\bar{\boldsymbol{w}}_\ell(d_k) \neq \bar{\boldsymbol{w}}_{\text{recip}}(d_k)$, then we must have

$$y_i \left\langle \bar{\boldsymbol{w}}_\ell(d_k), \boldsymbol{x}_i \right\rangle < c_k$$

for some $(\boldsymbol{x}_i, y_i)$, and it follows that

$$\mathcal{R}_\ell \left(\bar{\boldsymbol{w}}_\ell(d_k) \right) > \frac{1}{n} \ell(c_k) = \frac{1}{nc_k} \geq \mathcal{R}_\ell \left(\bar{\boldsymbol{w}}_{\text{recip}}(d_k) \right),$$

a contradiction.

2. Given $\ell = 1/z$ on $[c_k, d_k]$, Lemma 19 ensures that we can switch ℓ to e^{-z}: there exists $a_k > d_k$ such that we can let $\ell(z) = e^{-z}$ for any $z \geq a_k$. We let $\ell(z) = e^{-z}$ on $[a_k, b_k]$ where $b_k = 2\left(a_k + \ln(n)\right)/\bar{\gamma}$. Similarly we can show that $\bar{\boldsymbol{w}}_\ell(b_k) = \bar{\boldsymbol{w}}_{\text{exp}}(b_k)$.

Since for any $B \geq B_0$, it holds that

$$\left\| \frac{\bar{\boldsymbol{w}}_{\text{exp}}(B)}{B} - \frac{\bar{\boldsymbol{w}}_{\text{recip}}(B)}{B} \right\| \geq \frac{\|\bar{\boldsymbol{u}} - \hat{\boldsymbol{u}}\|}{3},$$

the loss ℓ constructed above satisfies the requirements in Proposition 13. ■

Appendix C. Omitted proofs from Section 4

Proof (of Lemma 16) Lemma 18 ensures that $\bar{\boldsymbol{w}}(B)$ and $\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)$ are collinear, which also implies $\bar{\boldsymbol{w}}_\perp(B) := \Pi_\perp\bar{\boldsymbol{w}}(B)$ and $\Pi_\perp\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)$ are collinear. Formally,

$$-\left\langle \frac{\bar{\boldsymbol{w}}_\perp(B)}{\|\bar{\boldsymbol{w}}_\perp(B)\|}, \Pi_\perp\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)\right\rangle = \left\|\Pi_\perp\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)\right\|, \tag{35}$$

and the left hand side is equal to

$$\frac{1}{n}\sum_{i=1}^{n} -\ell'\left(\langle\bar{\boldsymbol{w}}(B), y_i\boldsymbol{x}_i\rangle\right)\left\langle \frac{\bar{\boldsymbol{w}}_\perp(B)}{\|\bar{\boldsymbol{w}}_\perp(B)\|}, \Pi_\perp y_i\boldsymbol{x}_i\right\rangle = \frac{1}{n}\sum_{(\boldsymbol{x}_i,y_i)\in D_c} -\ell'\left(\langle\bar{\boldsymbol{w}}(B), y_i\boldsymbol{x}_i\rangle\right)\left\langle \frac{\bar{\boldsymbol{w}}_\perp(B)}{\|\bar{\boldsymbol{w}}_\perp(B)\|}, y_i\boldsymbol{x}_i\right\rangle. \tag{36}$$

Let

$$\hat{\boldsymbol{u}} := \mathop{\arg\max}_{\|\boldsymbol{u}\|=1, \boldsymbol{u}\in S^\perp} \min_{(\boldsymbol{x}_i,y_i)\in D_c} y_i\langle\boldsymbol{u},\boldsymbol{x}_i\rangle, \quad \text{and} \quad \hat{\gamma} := \max_{\|\boldsymbol{u}\|=1, \boldsymbol{u}\in S^\perp} \min_{(\boldsymbol{x}_i,y_i)\in D_c} y_i\langle\boldsymbol{u},\boldsymbol{x}_i\rangle,$$

and we can lower bound the right hand side of eq. (35) as follows:

$$\left\|\Pi_\perp\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right)\right\| \geq \left\langle -\Pi_\perp\nabla\mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right), \hat{\boldsymbol{u}}\right\rangle \geq \frac{1}{n}\sum_{(\boldsymbol{x}_i,y_i)\in D_c} -\ell'\left(\langle\bar{\boldsymbol{w}}(B), y_i\boldsymbol{x}_i\rangle\right)\hat{\gamma}. \tag{37}$$

Since $\mathcal{R}_s\left(\bar{\boldsymbol{w}}(B)\right) \leq \mathcal{R}\left(\bar{\boldsymbol{w}}(B)\right) \leq \ell(0)$, and $\mathcal{R}_s$ has compact sublevel sets on S, we know that $\Pi_S\bar{\boldsymbol{w}}(B)$ is bounded. Consequently

$$\lim_{B\to\infty} \frac{\bar{\boldsymbol{w}}_\perp(B)}{\|\bar{\boldsymbol{w}}_\perp(B)\|} = \lim_{B\to\infty} \frac{\bar{\boldsymbol{w}}(B)}{B} = \bar{\boldsymbol{u}}.$$

If the margin of $\bar{\boldsymbol{u}}$ on D_c is less than $\epsilon := \hat{\gamma}^2/\left(8|D_c|\right)$, then there exists B_0 such that for any $B \geq B_0$, the margin of $\bar{\boldsymbol{w}}_\perp(B)/\left\|\bar{\boldsymbol{w}}_\perp(B)\right\|$ on D_c is no larger than ϵ, and the distance between $\bar{\boldsymbol{w}}(B)/B$ and $\bar{\boldsymbol{w}}_\perp(B)/\left\|\bar{\boldsymbol{w}}_\perp(B)\right\|$ is no larger than ϵ. Let H denote the subset of D_c on which the margin of $\bar{\boldsymbol{w}}_\perp(B)/\left\|\bar{\boldsymbol{w}}_\perp(B)\right\|$ is larger than $\hat{\gamma}$, and suppose the minimum margin of $\bar{\boldsymbol{w}}_\perp(B)/\left\|\bar{\boldsymbol{w}}_\perp(B)\right\|$ on D_c is attained at i_1. Then eqs. (35) to (37) give

$$\begin{aligned}
&\sum_{(\boldsymbol{x}_i,y_i)\in H} -\ell'\left(\langle\bar{\boldsymbol{w}}(B), y_i\boldsymbol{x}_i\rangle\right)\left(\left\langle \frac{\bar{\boldsymbol{w}}_\perp(B)}{\|\bar{\boldsymbol{w}}_\perp(B)\|}, y_i\boldsymbol{x}_i\right\rangle - \hat{\gamma}\right) &(38)\\
\geq &\sum_{(\boldsymbol{x}_i,y_i)\in D_c\setminus H} -\ell'\left(\langle\bar{\boldsymbol{w}}(B), y_i\boldsymbol{x}_i\rangle\right)\left(\hat{\gamma} - \left\langle \frac{\bar{\boldsymbol{w}}_\perp(B)}{\|\bar{\boldsymbol{w}}_\perp(B)\|}, y_i\boldsymbol{x}_i\right\rangle\right)\\
\geq &-\ell'\left(\langle\bar{\boldsymbol{w}}(B), y_{i_1}\boldsymbol{x}_{i_1}\rangle\right)\left(\hat{\gamma} - \left\langle \frac{\bar{\boldsymbol{w}}_\perp(B)}{\|\bar{\boldsymbol{w}}_\perp(B)\|}, y_{i_1}\boldsymbol{x}_{i_1}\right\rangle\right). &(39)
\end{aligned}$$

Note that by our conditions, for any i,

$$\left|\left\langle \frac{\bar{\boldsymbol{w}}(B)}{B}, y_i\boldsymbol{x}_i\right\rangle - \left\langle \frac{\bar{\boldsymbol{w}}_\perp(B)}{\|\bar{\boldsymbol{w}}_\perp(B)\|}, y_i\boldsymbol{x}_i\right\rangle\right| \leq \epsilon.$$

Therefore eq. (38) can be upper bounded by

$$-\ell'\left(B\left(\hat{\gamma}-\epsilon\right)\right)(1-\hat{\gamma})|H| \le -\ell'\left(\frac{B\hat{\gamma}}{2}\right)|D_c|,$$

while eq. (39) can be lower bounded by

$$-\ell'\left(B(\epsilon+\epsilon)\right)\left(\hat{\gamma}-\epsilon\right) \ge -\ell'(2B\epsilon)\frac{\hat{\gamma}}{2}.$$

Consequently, for any $z \ge \alpha := 2B_0\epsilon$,

$$-\ell'\left(\frac{\hat{\gamma}z}{4\epsilon}\right) \ge -\ell'(z)\frac{\hat{\gamma}}{2|D_c|} = -\ell'(z)\frac{4\epsilon}{\hat{\gamma}}.$$

Similar to the proof of Lemma 7, we can show that $\int_{\alpha}^{\infty} -\ell(z)\,\mathrm{d}z = \infty$, a contradiction. ■

Proof (of Lemma 17) Let $\bar{\mathcal{R}} := \inf_{\boldsymbol{w}\in\mathbb{R}^d}\mathcal{R}(\boldsymbol{w})$. Also recall that $\bar{\boldsymbol{v}}$ denote the unique minimizer of $\mathcal{R}_s$ over S. Since for any $\boldsymbol{w}$,

$$\mathcal{R}_s\left(\boldsymbol{w}_S + (1+\alpha)\left\|\boldsymbol{w}_\perp\right\|\bar{\boldsymbol{u}}\right) = \mathcal{R}_s(\boldsymbol{w}),$$

we only need to show that

$$\mathcal{R}_c\left(\boldsymbol{w}_S + (1+\alpha)\left\|\boldsymbol{w}_\perp\right\|\bar{\boldsymbol{u}}\right) \le \mathcal{R}_c(\boldsymbol{w}).$$

Let $\xi(\alpha)$ be small enough such that for any $\boldsymbol{w}$ with $\mathcal{R}(\boldsymbol{w}) - \bar{\mathcal{R}} \le \xi(\alpha)$, the following properties hold.

1. $\|\bar{\boldsymbol{v}} - \boldsymbol{w}_S\| \le 1$.
2. $\|\bar{\boldsymbol{v}}\| + 1/\bar{\gamma} \le \alpha\bar{\gamma}\|\boldsymbol{w}_\perp\|/4 \le \alpha\|\boldsymbol{w}_\perp\|/4$.
3. For any $B \ge \|\boldsymbol{w}_\perp\| - 1/\bar{\gamma}$, it holds that $\|\bar{\boldsymbol{w}}(B)/B - \bar{\boldsymbol{u}}\| \le \alpha\bar{\gamma}/4$.

Consider $\boldsymbol{w}$ which satisfies $\mathcal{R}(\boldsymbol{w}) - \bar{\mathcal{R}} \le \xi(\alpha)$, and define

$$\tilde{\boldsymbol{w}} = \boldsymbol{w} + (\bar{\boldsymbol{v}} - \boldsymbol{w}_S) + \frac{\|\bar{\boldsymbol{v}} - \boldsymbol{w}_S\|}{\bar{\gamma}}\bar{\boldsymbol{u}}.$$

By definition $\mathcal{R}_s(\tilde{\boldsymbol{w}}) = \bar{\mathcal{R}}$, and since for any $(\boldsymbol{x}_i, y_i) \in D_c$ it holds that

$$\begin{aligned} y_i\langle\tilde{\boldsymbol{w}}, \boldsymbol{x}_i\rangle &= y_i\langle\boldsymbol{w}, \boldsymbol{x}_i\rangle + y_i\langle\bar{\boldsymbol{v}} - \boldsymbol{w}_S, \boldsymbol{x}_i\rangle + \frac{\|\bar{\boldsymbol{v}} - \boldsymbol{w}_S\|}{\bar{\gamma}} y_i\langle\bar{\boldsymbol{u}}, \boldsymbol{x}_i\rangle \\ &\ge y_i\langle\boldsymbol{w}, \boldsymbol{x}_i\rangle - \|\bar{\boldsymbol{v}} - \boldsymbol{w}_S\| + \|\bar{\boldsymbol{v}} - \boldsymbol{w}_S\| \\ &= y_i\langle\boldsymbol{w}, \boldsymbol{x}_i\rangle, \end{aligned}$$

we have $\mathcal{R}_c(\tilde{\boldsymbol{w}}) \le \mathcal{R}_c(\boldsymbol{w})$. On the other hand, by definition $\mathcal{R}\left(\bar{\boldsymbol{w}}\left(\|\tilde{\boldsymbol{w}}\|\right)\right) \le \mathcal{R}(\tilde{\boldsymbol{w}})$, and since $\mathcal{R}_s\left(\bar{\boldsymbol{w}}\left(\|\tilde{\boldsymbol{w}}\|\right)\right) \ge \bar{\mathcal{R}} = \mathcal{R}_s(\tilde{\boldsymbol{w}})$, we have

$$\mathcal{R}_c\left(\bar{\boldsymbol{w}}\left(\|\tilde{\boldsymbol{w}}\|\right)\right) \le \mathcal{R}_c(\tilde{\boldsymbol{w}}) \le \mathcal{R}_c(\boldsymbol{w}). \tag{40}$$

Note that due to bullet 1 above,

$$\|\tilde{\boldsymbol{w}}\| \geq \|\Pi_{\perp}\tilde{\boldsymbol{w}}\| \geq \|\boldsymbol{w}_{\perp}\| - \frac{\|\bar{\boldsymbol{v}} - \boldsymbol{w}_S\|}{\bar{\gamma}} \geq \|\boldsymbol{w}_{\perp}\| - \frac{1}{\bar{\gamma}}.$$

Therefore due to bullet 3 above $\|{}^{\bar{\boldsymbol{w}}(\|\tilde{\boldsymbol{w}}\|)}/_{\|\tilde{\boldsymbol{w}}\|} - \bar{\boldsymbol{u}}\| \leq \alpha\bar{\gamma}/4$, which implies for any $(\boldsymbol{x}_i, y_i) \in D_c$,

$$\left(1 + \frac{\alpha}{4}\right) y_i \langle \|\tilde{\boldsymbol{w}}\|\bar{\boldsymbol{u}}, \boldsymbol{x}_i\rangle \geq y_i \left\langle \bar{\boldsymbol{w}}\left(\|\tilde{\boldsymbol{w}}\|\right), \boldsymbol{x}_i\right\rangle. \tag{41}$$

On the other hand, by the triangle inequality and bullet 1 and 2,

$$\begin{aligned}\|\tilde{\boldsymbol{w}}\| &= \left\|\bar{\boldsymbol{v}} + \boldsymbol{w}_{\perp} + \frac{\|\bar{\boldsymbol{v}} - \boldsymbol{w}_S\|}{\bar{\gamma}}\bar{\boldsymbol{u}}\right\| \\ &\leq \|\bar{\boldsymbol{v}}\| + \|\boldsymbol{w}_{\perp}\| + \frac{\|\bar{\boldsymbol{v}} - \boldsymbol{w}_S\|}{\bar{\gamma}} \\ &\leq \|\bar{\boldsymbol{v}}\| + \|\boldsymbol{w}_{\perp}\| + \frac{1}{\bar{\gamma}} \leq \left(1 + \frac{\alpha\bar{\gamma}}{4}\right)\|\boldsymbol{w}_{\perp}\| \leq \left(1 + \frac{\alpha}{4}\right)\|\boldsymbol{w}_{\perp}\|,\end{aligned}$$

and thus

$$\left(1 + \frac{\alpha}{4}\right) y_i \langle \|\tilde{\boldsymbol{w}}\|\bar{\boldsymbol{u}}, \boldsymbol{x}_i\rangle \leq \left(1 + \frac{\alpha}{4}\right)^2 y_i \langle \|\boldsymbol{w}_{\perp}\|\bar{\boldsymbol{u}}, \boldsymbol{x}_i\rangle \leq \left(1 + \frac{3\alpha}{4}\right) y_i \langle \|\boldsymbol{w}_{\perp}\|\bar{\boldsymbol{u}}, \boldsymbol{x}_i\rangle. \tag{42}$$

Moreover, due to bullet 1 and 2,

$$y_i\langle \boldsymbol{w}_S, \boldsymbol{x}_i\rangle \geq -\|\boldsymbol{w}_S\| \geq -\|\bar{\boldsymbol{v}}\| - 1 \geq -\|\bar{\boldsymbol{v}}\| - \frac{1}{\bar{\gamma}} \geq -\frac{\alpha\bar{\gamma}\|\boldsymbol{w}_{\perp}\|}{4} \geq -\frac{\alpha}{4} y_i \langle \|\boldsymbol{w}_{\perp}\|\bar{\boldsymbol{u}}, \boldsymbol{x}_i\rangle. \tag{43}$$

Combining eqs. (41) to (43) gives

$$y_i \left\langle \bar{\boldsymbol{w}}\left(\|\tilde{\boldsymbol{w}}\|\right), \boldsymbol{x}_i\right\rangle \leq \left(1 + \frac{3\alpha}{4}\right) y_i \langle \|\boldsymbol{w}_{\perp}\|\bar{\boldsymbol{u}}, \boldsymbol{x}_i\rangle \leq y_i \left\langle \boldsymbol{w}_S + (1+\alpha)\|\boldsymbol{w}_{\perp}\|\bar{\boldsymbol{u}}, \boldsymbol{x}_i\right\rangle,$$

which implies

$$\mathcal{R}_c\left(\boldsymbol{w}_S + (1+\alpha)\|\boldsymbol{w}_{\perp}\|\bar{\boldsymbol{u}}\right) \leq \mathcal{R}_c\left(\bar{\boldsymbol{w}}\left(\|\tilde{\boldsymbol{w}}\|\right)\right). \tag{44}$$

It follows from eqs. (40) and (44) that

$$\mathcal{R}_c\left(\boldsymbol{w}_S + (1+\alpha)\|\boldsymbol{w}_{\perp}\|\bar{\boldsymbol{u}}\right) \leq \mathcal{R}_c(\boldsymbol{w}),$$

which concludes the proof. ■

Proof (of Theorem 15(2), the "only if" part) Given any $\epsilon \in (0,1)$, let α satisfy $1/(1+\alpha) = 1 - \epsilon$ (i.e., let $\alpha = \epsilon/(1-\epsilon)$).

Since $\lim_{t\to\infty}\mathcal{R}(\boldsymbol{w}_t) = \inf_{\boldsymbol{w}\in\mathbb{R}^d}\mathcal{R}(\boldsymbol{w})$, there exists t_0 such that for any $t \geq t_0$ we have $\mathcal{R}(\boldsymbol{w}_t) - \inf_{\boldsymbol{w}\in\mathbb{R}^d}\mathcal{R}(\boldsymbol{w}) \leq \xi(\alpha)$ and $\|\boldsymbol{w}_{t,\perp}\| \geq 1$. By convexity and Lemma 17, for $t \geq t_0$,

$$\begin{aligned}\langle \nabla\mathcal{R}(\boldsymbol{w}_t), \boldsymbol{w}_{t,\perp} - (1+\alpha)\|\boldsymbol{w}_{t,\perp}\|\bar{\boldsymbol{u}}\rangle &= \langle \nabla\mathcal{R}(\boldsymbol{w}_t), \boldsymbol{w}_{t,S} + \boldsymbol{w}_{t,\perp} - \boldsymbol{w}_{t,S} - (1+\alpha)\|\boldsymbol{w}_{t,\perp}\|\bar{\boldsymbol{u}}\rangle \\ &= \langle \nabla\mathcal{R}(\boldsymbol{w}_t), \boldsymbol{w}_t - \boldsymbol{w}_{t,S} - (1+\alpha)\|\boldsymbol{w}_{t,\perp}\|\bar{\boldsymbol{u}}\rangle \\ &\geq \mathcal{R}(\boldsymbol{w}_t) - \mathcal{R}\left(\boldsymbol{w}_{t,S} + (1+\alpha)\|\boldsymbol{w}_{t,\perp}\|\bar{\boldsymbol{u}}\right) \geq 0.\end{aligned}$$

Consequently,

$$\begin{aligned}\langle \boldsymbol{w}_{t+1}-\boldsymbol{w}_t, \bar{\boldsymbol{u}}\rangle &= \langle -\eta\nabla\mathcal{R}(\boldsymbol{w}_t), \bar{\boldsymbol{u}}\rangle \\ &\geq \langle -\eta\nabla\mathcal{R}(\boldsymbol{w}_t), \boldsymbol{w}_{t,\perp}\rangle \frac{1}{(1+\alpha)\|\boldsymbol{w}_{t,\perp}\|} \\ &= \langle \boldsymbol{w}_{t+1}-\boldsymbol{w}_t, \boldsymbol{w}_{t,\perp}\rangle \frac{1}{(1+\alpha)\|\boldsymbol{w}_{t,\perp}\|} \\ &= \langle \boldsymbol{w}_{t+1,\perp}-\boldsymbol{w}_{t,\perp}, \boldsymbol{w}_{t,\perp}\rangle \frac{1}{(1+\alpha)\|\boldsymbol{w}_{t,\perp}\|} \\ &= \left(\frac{1}{2}\|\boldsymbol{w}_{t+1,\perp}\|^2 - \frac{1}{2}\|\boldsymbol{w}_{t,\perp}\|^2 - \frac{1}{2}\|\boldsymbol{w}_{t+1,\perp}-\boldsymbol{w}_{t,\perp}\|^2\right)\frac{1}{(1+\alpha)\|\boldsymbol{w}_{t,\perp}\|}.\end{aligned}$$

On one hand, we have

$$\left(\frac{1}{2}\|\boldsymbol{w}_{t+1,\perp}\|^2 - \frac{1}{2}\|\boldsymbol{w}_{t,\perp}\|^2\right)/\|\boldsymbol{w}_{t,\perp}\| \geq \|\boldsymbol{w}_{t+1,\perp}\| - \|\boldsymbol{w}_{t,\perp}\|.$$

On the other hand, using the step size condition in eq. (7), we have

$$\begin{aligned}\frac{\|\boldsymbol{w}_{t+1,\perp}-\boldsymbol{w}_{t,\perp}\|^2}{2(1+\alpha)\|\boldsymbol{w}_{t,\perp}\|} \leq \frac{\|\boldsymbol{w}_{t+1,\perp}-\boldsymbol{w}_{t,\perp}\|^2}{2} &\leq \frac{\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\|^2}{2} \\ &= \frac{\eta^2\|\nabla\mathcal{R}(\boldsymbol{w}_t)\|^2}{2} \\ &\leq \eta\left(\mathcal{R}(\boldsymbol{w}_t)-\mathcal{R}(\boldsymbol{w}_{t+1})\right).\end{aligned}$$

As a result,

$$\langle \boldsymbol{w}_t-\boldsymbol{w}_{t_0}, \bar{\boldsymbol{u}}\rangle \geq \frac{\|\boldsymbol{w}_{t,\perp}\|-\|\boldsymbol{w}_{t_0,\perp}\|}{1+\alpha} - \eta\mathcal{R}(\boldsymbol{w}_{t_0}) = (1-\epsilon)\left(\|\boldsymbol{w}_{t,\perp}\|-\|\boldsymbol{w}_{t_0,\perp}\|\right) - \eta\mathcal{R}(\boldsymbol{w}_{t_0}),$$

meaning

$$\left\langle \frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|}, \bar{\boldsymbol{u}}\right\rangle \geq (1-\epsilon)\frac{\|\boldsymbol{w}_{t,\perp}\|}{\|\boldsymbol{w}_t\|} + \frac{\langle \boldsymbol{w}_{t_0}, \bar{\boldsymbol{u}}\rangle - (1-\epsilon)\|\boldsymbol{w}_{t_0,\perp}\| - \eta\mathcal{R}(\boldsymbol{w}_{t_0})}{\|\boldsymbol{w}_t\|}.$$

Consequently,

$$\liminf_{t\to\infty}\left\langle \frac{\boldsymbol{w}_t}{\|\boldsymbol{w}_t\|}, \bar{\boldsymbol{u}}\right\rangle \geq 1-\epsilon.$$

Since ϵ is arbitrary, we get $\boldsymbol{w}_t/\|\boldsymbol{w}_t\| \to \bar{\boldsymbol{u}}$. ■