Proceedings of Machine Learning Research vol 99:1–27, 2019 32nd Annual Conference on Learning Theory

The implicit bias of gradient descent on nonseparable data

Ziwei Ji ZIWEIJI2@ILLINOIS.EDU

Matus Telgarsky MJT@ILLINOIS.EDU

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

Gradient descent, when applied to the task of logistic regression, outputs iterates which are biased to follow a unique ray defined by the data. The direction of this ray is the maximum margin predictor of a maximal linearly separable subset of the data; the gradient descent iterates converge to this ray *in direction* at the rate $\mathcal{O}(\ln\ln t/\ln t)$. The ray does not pass through the origin in general, and its offset is the bounded global optimum of the risk over the remaining data; gradient descent recovers this offset at a rate $\mathcal{O}((\ln t)^2/\sqrt{t})$.

Keywords: Implicit bias, gradient descent, logistic regression, maximum margin.

1. Introduction

Logistic regression is the task of finding a vector $w \in \mathbb{R}^d$ which approximately minimizes the *empirical logistic risk*, namely

$$\mathcal{R}_{\log}(w) := \frac{1}{n}\sum_{i=1}^{n} \ell_{\log}(\langle w, -x_i y_i\rangle) \qquad \text{with} \quad \ell_{\log}(r) := \ln\left(1 + \exp(r)\right),$$

where $\ell_{\log}$ is the *logistic loss*. A traditional way to minimize $\mathcal{R}_{\log}$ is to pick an arbitrary w_0, and from there recursively construct *gradient descent iterates* $(w_j)_{j\geq 0}$ via $w_{j+1} := w_j - \eta_j \nabla\mathcal{R}_{\log}(w_j)$, where $(\eta_j)_{j\geq 0}$ are step size parameters.

Despite the simplicity of this setting, a general characterization of the gradient descent path has escaped the literature for a variety of reasons. It is possible for the data to be configured so that $\mathcal{R}_{\log}$ is strongly convex, in which case standard convex optimization tools grant the existence of a unique bounded optimum, and moreover a rate at which gradient descent iterates converge to it. It is also possible, however, that data is *linearly separable*, in which case $\mathcal{R}_{\log}$ has an infimum of 0 despite being positive everywhere; the optimum is off at infinity, and convergence analyses operate by establishing a *maximum margin* property of the normalized iterates $w_j/|w_j|$ (Soudry et al., 2017), just as in the analysis of AdaBoost (Schapire and Freund, 2012; Telgarsky, 2013).

In general, data can fail to induce a strongly convex risk or a linearly separable problem. Despite this, there is still a unique characterization of the gradient descent path. Specifically, gradient descent is biased to follow an *optimal ray* $\{\bar{v} + r\cdot\bar{u} \ : \ r \geq 0\}$, which is constructed as follows. First, as detailed in Section 2, the data uniquely determines (via a greedy procedure) a linearly separable subset, with a corresponding maximum margin predictor $\bar{u}$; gradient descent converges to $\bar{u}$ *in direction*, meaning $w_j/|w_j| \to \bar{u}$. The remaining data span a space S; the empirical risk of the remaining data is strongly convex over bounded subsets of S, and possesses a unique optimum $\bar{v}$, to which the projected gradient descent iterates converge, meaning $\Pi_S w_j \to \bar{v}$.

Theorem 1 (Simplification of Theorems 2, 3 and 7) *Let examples $((x_i, y_i))_{i=1}^n$ be given satisfying $|x_i y_i| \le 1$, along with a loss $\ell \in \{\ell_{\log}, \exp\}$, with corresponding risk $\mathcal{R}$ as above. Consider gradient descent iterates $(w_j)_{j\ge 0}$ as above, with $w_0 = 0$.*

1. ***(Convergence in risk.)*** *For any step sizes $\eta_j \le 1$ and any $t \ge 1$,*

$$\mathcal{R}(w_t) - \inf_{w\in\mathbb{R}^d} \mathcal{R}(w) = \mathcal{O}\left(\frac{1}{t} + \frac{\ln(t)^2}{\sum_{j<t}\eta_j}\right) = \begin{cases} \mathcal{O}(\ln(t)^2/t) & \eta_j = \Omega(1), \\ \mathcal{O}(\ln(t)^2/\sqrt{t}) & \eta_j = \Omega(1/\sqrt{j+1}), \end{cases}$$

   where $\mathcal{O}(\cdot)$ hides problem-dependent constants.

2. ***(Convergence in parameters; implicit bias and regularization.)*** *The data uniquely determines a subspace S and a vector $\bar{v} \in S$, such that if $\eta_j := 1/\sqrt{j+1}$ and $t^2 = \Omega(n \ln(t))$, letting Π_S denote orthogonal projection onto S and $\bar{w}_t := \arg\min\{\mathcal{R}(w) : |w| \le |w_t|\}$ denote the solution to the constrained optimization problem, then*

$$|\Pi_S w_t| = \Theta(1) \qquad \text{and} \qquad \max\left\{|\Pi_S w_t - \bar{v}|^2,\ |\Pi_S \bar{w}_t - \bar{v}|^2\right\} = \mathcal{O}\left(\frac{\ln(t)^2}{\sqrt{t}}\right).$$

   If there are examples outside S, their projection onto $S^\perp$ is linearly separable with maximum margin predictor $\bar{u} \in S^\perp$, and

$$|\Pi_{S^\perp} w_t| = \Theta(\ln(t)) \qquad \text{and} \qquad \max\left\{\left|\frac{w_t}{|w_t|} - \bar{u}\right|^2,\ \left|\frac{\bar{w}_t}{|\bar{w}_t|} - \bar{u}\right|^2\right\} = \mathcal{O}\left(\frac{\ln\ln t}{\ln t}\right).$$

   In particular, $w_t/|w_t| \to \bar{u}$ and $\Pi_S w_t \to \bar{v}$.

This theorem captures *implicit bias* by showing that gradient descent follows the unique ray $\{\bar{v} + r \cdot \bar{u} : r \ge 0\}$, even though the risk itself may be minimized by any vector which lies in the relative interior of a convex cone defined by the problem. Similarly, the theorem captures *implicit regularization* by showing that the gradient descent iterates also track the sequence of constrained optima $(\bar{w}_j)_{j\ge 1}$.

This paper is organized as follows.

Problem structure (Section 2). This section first builds up the case of general data with a few illustrative examples, including strongly convex and linearly separable cases. Thereafter, a complete construction and characterization of the optimal ray $\{\bar{v}+r\cdot\bar{u} : r \ge 0\}$ and related objects is provided in Theorem 2.

Risk convergence (Section 3). The preceding section on problem structure reveals that the (bounded) point $\bar{v}+\bar{u}\,(\ln(t)/\gamma)$ achieves low risk; plugging this into a modified smoothness argument yields converge in risk with no apparent dependence on the optimum at infinity.

Parameter convergence (Section 4). The preceding problem structure reveals $\mathcal{R}$ is strongly convex over bounded subsets of S, which gives convergence to $\bar{v}$ via standard convex optimization tools. To prove $w_j/|w_j| \to \bar{u}$, the first key to the analysis is to study not $\mathcal{R}$ but instead $\ln \mathcal{R}$, which more conveniently captures local smoothness (extreme flattening) of $\mathcal{R}$. To complete the proof, a number of technical issues must be worked out, including bounds on $|w_t|$, which rely upon an adaptation of the perceptron convergence proof. This proof goes through much more easily for the exponential loss, which is the main reason for its appearance in Theorem 1.

Related work (Section 5). The paper closes with a discussion of related work.

1.1. Notation

The data sample will be denoted by $((x_i, y_i))_{i=1}^n$. Collect these examples into a matrix $A \in \mathbb{R}^{n\times d}$, with i^{th} row $A_i := -y_i x_i^\top$; it is assumed that $\max_i |A_i| = \max_i |x_i y_i| \leq 1$, where $|\cdot|$ denote the ℓ_2-norm in this paper. Given loss function $\ell : \mathbb{R} \to \mathbb{R}_{\geq 0}$, for any k and any $v \in \mathbb{R}^k$, define a coordinate-wise form $L(v) := \sum_{i=1}^k \ell(v_i)$, whereby the empirical risk $\mathcal{R}(w) := \sum_{i=1}^n \ell\left(\langle w, -x_i y_i\rangle\right)/n$ satisfies $\mathcal{R}(w) = L(Aw)/n$, with gradient $\nabla\mathcal{R}(w) = A^\top \nabla L(Aw)/n$. Define $\ell_{\text{exp}}(z) := \exp(z)$ and $\ell_{\text{log}}(z) := \ln(1 + \exp(z))$, and correspondingly L_{exp}, $\mathcal{R}_{\text{exp}}$, L_{log}, and $\mathcal{R}_{\text{log}}$.

As in Theorem 1, and will be elaborated in Section 2, the matrix A defines a unique division of $\mathbb{R}^d$ into a direct sum of subspaces $\mathbb{R}^d = S \oplus S^\perp$. The rows of A are either within S or S^c (i.e., $\mathbb{R}^d \setminus S$), and without loss of generality reorder the examples (and permute the rows of A) so that $A := \begin{bmatrix} A_S \\ A_c \end{bmatrix}$ where the rows of A_S are within S and the rows of A_c are within S^c; tying this to the earlier discussion, A_c is the linearly separable part of the data, and A_S is the strongly convex part. Furthermore, let Π_S and $\Pi_\perp$ respectively denote orthogonal projection onto S and $S^\perp$, and define $A_\perp = \Pi_\perp A_c$, where each row of A_c is orthogonally projected onto $S^\perp$. By this notation,

$$\Pi_\perp \nabla(L \circ A)(w) = \Pi_\perp \begin{bmatrix} A_c \\ A_S \end{bmatrix}^\top \nabla L\left(\begin{bmatrix} A_c \\ A_S \end{bmatrix} w\right) = \Pi_\perp \begin{bmatrix} A_c \\ A_S \end{bmatrix}^\top \begin{bmatrix} \nabla L(A_c w) \\ \nabla L(A_S w) \end{bmatrix} = \begin{bmatrix} A_\perp \\ 0 \end{bmatrix}^\top \begin{bmatrix} \nabla L(A_c w) \\ \nabla L(A_S w) \end{bmatrix}$$
$$= A_\perp^\top \nabla L(A_c w),$$

which has made use of L at varying input dimensions.

Gradient descent here will always start with $w_0 := 0$, and thereafter set $w_{j+1} := w_j - \eta_j \nabla\mathcal{R}(w_j)$. It is convenient to define $\gamma_j := |\nabla(\ln\mathcal{R})(w_j)| = |\nabla\mathcal{R}(w_j)|/\mathcal{R}(w_j)$ and $\hat{\eta}_j := \eta_j \mathcal{R}(w_j)$, whereby

$$|w_t| \leq \sum_{j<t} \eta_j |\nabla\mathcal{R}(w_j)| = \sum_{j<t} \hat{\eta}_j \gamma_j.$$

Moreover, let $\bar{w}_t := \arg\min\left\{\mathcal{R}(w) : |w| \leq |w_t|\right\}$ denote the solution to the corresponding constrained optimization problem.

2. Problem structure

This section culminates in Theorem 2, which characterizes the unique ray $\{\bar{v} + r \cdot \bar{u} : r \geq 0\}$. To build towards this, first consider the following examples.

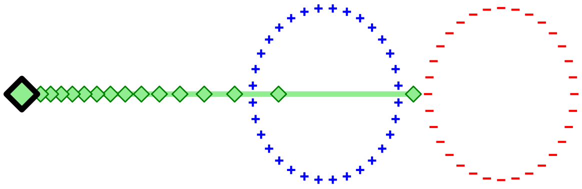

Figure 1: Separable.

Linearly separable. Consider the data at right in Figure 1: a blue circle of positive points, and a red circle of negative points. The data is *linearly separable:* there exist vectors $u \in \mathbb{R}^d$ with positive margin, meaning $\min_i \langle u, x_i y_i\rangle > 0$. Taking any such u and extending it to infinity will achieve 0 risk, but this is not what gradient descent chooses. Constraining u to have unit norm, a unique maximum margin point $\bar{u} = -\mathbf{e}_1$ is obtained. The green gradient descent iterates follow $\bar{u}$ exactly.

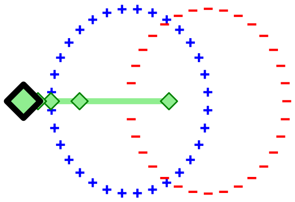

Figure 2: Strongly convex.

Strong convexity. Now consider moving the circles of data in Figure 1 until they overlap, obtaining Figure 2. This data is *not* linearly separable; indeed, given any nonzero vector $u \in \mathbb{R}^d$, there exist data points incorrectly classified by u, and therefore extending u indefinitely will cause the risk to also increase to infinity. It follows that the risk itself is 0-coercive (Hiriart-Urruty and Lemaréchal, 2001), and moreover strongly convex over bounded subsets, with a unique optimum $\bar{v}$. Gradient descent converges towards $\bar{v}$.

An intermediate setting. The preceding two settings either had the circles overlapping, or far apart. What if they are pressed together so that they touch at the origin? Excluding the point at the origin, the circles may still be separated with the maximum margin separator $\bar{u} = -\mathbf{e}_1$ from the linearly separable instance Figure 1. This example is our first taste of the general ray $\{\bar{v} + r \cdot \bar{u} : r \geq 0\}$, albeit still with some triviality: $\bar{v} = 0$. Specifically, $\bar{u}$ is the maximum margin separator of all data excluding the point at the origin; the risk in this instance is bounded below by $\ell(0)/n$, which is the necessary error on the point at the origin; the global optimum for that single point is $\bar{v} = 0$.

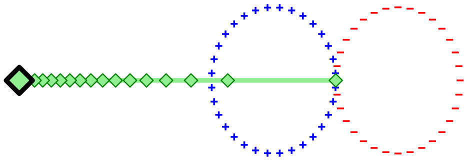

Figure 3: Mixed data.

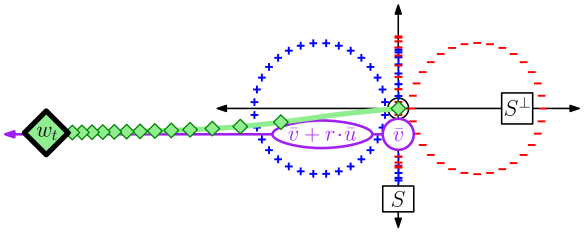

Figure 4: The general case.

The general case. Combining elements from the preceding examples, the general case may be characterized as follows; it appears in Figure 4, with all relevant objects labeled. In the general case, the dataset consists of a *maximal linearly separable subset* Z, with the remaining data falling into a subset over which the empirical risk is strongly convex. Specifically, Z is constructed with the following greedy procedure: for each example (x_i, y_i), include it in Z if there exists u_i with $\langle u_i, x_i y_i \rangle > 0$ and $\min_j \langle u_j, x_j y_j \rangle \geq 0$. The aggregate $u := \sum_{i \in Z} u_i$ satisfies $\langle u, x_i y_i \rangle > 0$ for $i \in Z$ and $\langle u, x_i y_i \rangle = 0$ otherwise. Therefore Z can be strictly separated by some vector u orthogonal to Z^c; let $\bar{u}$ denote the maximum margin separator of Z which is orthogonal to Z^c.

Turning now to Z^c, any vector v which is correct on some $(x_i, y_i) \in Z^c$ (i.e., $\langle u, x_i y_i \rangle > 0$) must also be incorrect on some other example (x_j, y_j) in Z^c (i.e., $\langle v, x_j y_j \rangle < 0$); otherwise, (x_i, y_i) would have been included in Z! Consequently, as in Figure 2 above, the empirical risk restricted to Z^c is strongly convex, with a unique optimum $\bar{v}$. The gradient descent iterates follow the ray $\{\bar{v} + r \cdot \bar{u} : r \geq 0\}$, which means they are globally optimal along Z^c, and achieve zero risk and follow the maximum margin direction $\bar{u}$.

Turning back to the construction in Figure 4, the linearly separable data Z is the two red and blue circles, while Z^c consists of data points on the vertical axis. The points in Z^c do not affect $\bar{u}$, and have been adjusted to move $\bar{v}$ away from 0, where it rested in Figure 3.

Now using the notation $A_i = -y_i x_i^\top$, for $i \in Z$, the vector $-y_i x_i$ is collected into A_c, while for $i \in Z^c$, the vector $-y_i x_i^\top$ is put into A_S. The above constructions are made rigorous in the following theorem. The proof of Theorem 2, presented in the appendix, follows the intuition above.

Theorem 2 *The rows of A can be uniquely partitioned into matrices (A_S, A_c), with a corresponding pair of orthogonal subspaces $(S, S^\perp)$ where $S = \text{span}(A_S^\top)$ satisfying the following properties.*

1. ***(Strongly convex part.)*** *If ℓ is twice continuously differentiable with $\ell'' > 0$, and $\ell \geq 0$, and $\lim_{z\to-\infty} \ell(z) = 0$, then $L \circ A$ is strongly convex over compact subsets of S, and $L \circ A_S$ admits a unique minimizer $\bar{v}$ with $\inf_{w\in\mathbb{R}^d} L(Aw) = \inf_{v\in S} L(A_S v) = L(A_S\bar{v})$.*

2. ***(Separable part.)*** *If A_c is nonempty (and thus so is $A_\perp$), then $A_\perp$ is linearly separable. The maximum margin is given by*

$$\gamma := -\min\left\{ \max_i (A_\perp u)_i : |u| = 1 \right\} = \min\left\{ |A_\perp^\top q| : q \geq 0, \sum_i q_i = 1 \right\} > 0,$$

and the maximum margin solution $\bar{u}$ is the unique optimum to the primal problem, satisfying $\bar{u} = -A_\perp^\top \bar{q}/\gamma$ for every dual optimum $\bar{q}$. If $\ell \geq 0$ and $\lim_{z\to-\infty} \ell(z) = 0$, then

$$\inf_{w\in\mathbb{R}^d} L(Aw) = L(A_S\bar{v}) + \lim_{r\to\infty} L\left(A_c(\bar{v} + r\bar{u})\right) = L(A_S\bar{v}).$$

3. Risk convergence

Gradient descent decreases the risk as follows.

Theorem 3 *For $\ell \in \{\ell_{\log}, \ell_{\exp}\}$, given step sizes $\eta_j \leq 1$ and $w_0 = 0$, then for any $t \geq 1$,*

$$\mathcal{R}(w_t) - \inf_w \mathcal{R}(w) \leq \frac{\exp(|\bar{v}|)}{t} + \frac{|\bar{v}|^2 + \ln(t)^2/\gamma^2}{2\sum_{j=0}^{t-1}\eta_j}.$$

This proof relies upon three essential steps.

1. A slight generalization of standard smoothness-based gradient descent bounds (cf. Lemma 4).
2. A useful comparison point to feed into the preceding gradient descent bound (cf. Lemma 5): the choice $\bar{v} + \bar{u}\,(\ln(t)/\gamma)$, made possible by Theorem 2.
3. Smoothness estimates for L when $\ell \in \{\ell_{\log}, \ell_{\exp}\}$ (cf. Lemma 6).

In more detail, the first step, a refined smoothness-based gradient descent guarantee, is as follows. While similar bounds are standard in the literature (Bubeck, 2015; Nesterov, 2004), this version has a short proof and no issue with unbounded domains.

Lemma 4 *Suppose f is convex, and there exists $\beta \geq 0$ so that $1 - \eta_j\beta/2 \geq 0$ and gradient iterates $(w_0, \ldots, w_t)$ with $w_{j+1} := w_j - \eta_j \nabla f(w_j)$ satisfy*

$$f(w_{j+1}) \leq f(w_j) - \eta_j\left(1 - \frac{\eta_j\beta}{2}\right)|\nabla f(w_j)|^2.$$

Then for any $z \in \mathbb{R}^d$,

$$2\sum_{j=0}^{t-1}\eta_j\left(f(w_j) - f(z)\right) - \sum_{j=0}^{t-1}\frac{\eta_j}{1 - \beta\eta_j/2}\left(f(w_j) - f(w_{j+1})\right) \leq |w_0 - z|^2 - |w_t - z|^2.$$

The proof is similar to the standard ones, and appears in the appendix.

The second step, as above, is to produce a reference point z to plug into Lemma 4.

Lemma 5 *Let $\ell \in \{\ell_{\exp}, \ell_{\log}\}$ be given. Then $z := \bar{v} + \bar{u}\left({}^{\ln(t)}/_{\gamma}\right)$ satisfies $|z|^2 = |\bar{v}|^2 + {}^{\ln(t)^2}/_{\gamma^2}$ and*

$$\mathcal{R}(z) \leq \inf_w \mathcal{R}(w) + \frac{\exp(|\bar{v}|)}{t}.$$

Proof By Theorem 2, and since $\ell_{\log} \leq \ell_{\exp}$ and $|A_i| \leq 1$,

$$L(Az) = L(A_S\bar{v}) + L(A_cz) \leq \inf_w L(Aw) + n\exp\left(|\bar{v}| - \ln(t)\right) = \inf_w L(Aw) + \frac{n\exp\left(|\bar{v}|\right)}{t}.$$

■

Lastly, the smoothness guarantee on $\mathcal{R}$. Even though the logistic loss is smooth, this proof gives a refined smoothness inequality where the j^{th} step is $\mathcal{R}(w_j)$-smooth; this refinement will be essential when proving parameter convergence. This proof is based on the convergence guarantee for AdaBoost (Schapire and Freund, 2012). Recall the definitions $\gamma_j := |\nabla(\ln\mathcal{R})(w_j)| = |\nabla\mathcal{R}(w_j)|/\mathcal{R}(w_j)$ and $\hat{\eta}_j := \eta_j\mathcal{R}(w_j)$.

Lemma 6 *Suppose ℓ is convex, $\ell' \leq \ell$, $\ell'' \leq \ell$, and $\hat{\eta}_j = \eta_j\mathcal{R}(w_j) \leq 1$. Then*

$$\mathcal{R}(w_{j+1}) \leq \mathcal{R}(w_j) - \eta_j\left(1 - \frac{\eta_j\mathcal{R}(w_j)}{2}\right)|\nabla\mathcal{R}(w_j)|^2 = \mathcal{R}(w_j)\left(1 - \hat{\eta}_j(1-\hat{\eta}_j/2)\gamma_j^2\right)$$

and thus

$$\mathcal{R}(w_t) \leq \mathcal{R}(w_0)\prod_{j<t}\left(1 - \hat{\eta}_j(1-\hat{\eta}_j/2)\gamma_j^2\right) \leq \mathcal{R}(w_0)\exp\left(-\sum_{j<t}\hat{\eta}_j(1-\hat{\eta}_j/2)\gamma_j^2\right).$$

Additionally, $|w_t| \leq \sum_{j<t}\hat{\eta}_j\gamma_j$.

This proof mostly proceeds in a usual way via recursive application of a Taylor expansion

$$\mathcal{R}\left(w - \eta\nabla\mathcal{R}(w)\right) \leq \mathcal{R}(w) - \eta|\nabla\mathcal{R}(w)|^2 + \frac{1}{2}\max_{v\in[w,w']}\sum_i (A_i(w-w'))^2\ell''(A_iv)/n.$$

The interesting part is the inequality $\ell'' \leq \ell$: this allows the final term in the preceding expression to replace ℓ'' with ℓ, and some massaging with the maximum allows this term to be replaced with $\mathcal{R}(w)$ (since a descent direction was followed).

A direct but important consequence is obtained by applying $\ln$ to both sides:

$$\ln\mathcal{R}(w_t) \leq \ln\mathcal{R}(w_0) - \sum_{j<t}\hat{\eta}_j(1-\hat{\eta}_j/2)\gamma_j^2. \tag{1}$$

It follows that $\ln\mathcal{R}$ is smooth, but moreover has *constant* smoothness unlike $\mathcal{R}$ above.

Note that for $\ell \in \{\ell_{\log}, \ell_{\exp}\}$, since $\mathcal{R}(w_0) \leq 1$, choosing $\eta_j \leq 1$ will ensure $\hat{\eta}_j \leq 1$, and then Lemma 6 holds. Also, Lemma 6 shows that Lemma 4 holds for $\{\mathcal{R}_{\exp}, \mathcal{R}_{\log}\}$ with $\beta = 1$.

Combining these pieces now leads to a proof of Theorem 3, given in full in the appendix. As a final remark, note that step size $\eta_j = 1$ led to a $\widetilde{\mathcal{O}}(1/t)$ rate, whereas $\eta_j = 1/\sqrt{j+1}$ leads to a $\widetilde{\mathcal{O}}(1/\sqrt{t})$ rate.

4. Parameter convergence

As in Theorem 1, the parameter convergence guarantee gives convergence to $\bar{v} \in S$ over the strongly convex part S (that is, $\Pi_S w_t \to \bar{v}$ and $\Pi_S \bar{w}_t \to \bar{v}$), and convergence *in direction* (convergence of the normalized iterates) to $\bar{u} \in S^\perp$ over the separable part $S^\perp$ (${w_t}/{|w_t|} \to \bar{u}$ and ${\bar{w}_t}/{|\bar{w}_t|} \to \bar{u}$). In more detail, the convergence rates are as follows.

Theorem 7 *Let loss $\ell \in \{\ell_{\exp}, \ell_{\log}\}$ be given.*

1. ***(Separable case.)*** *Suppose A is separable (thus $S^\perp = \mathbb{R}^d$, and $A = A_c = A_\perp$, and $A\bar{u} = A_\perp \bar{u} \le -\gamma$). Furthermore, suppose $\eta_j = 1$, and $t \ge 5$, and ${t}/{\ln(t)^3} \ge {n}/{\gamma^4}$. Then*
$$\max\left\{\left|\frac{w_t}{|w_t|} - \bar{u}\right|^2, \left|\frac{\bar{w}_t}{|\bar{w}_t|} - \bar{u}\right|^2\right\} = \mathcal{O}\left(\frac{\ln n + \ln\left(|w_t|\right)}{|w_t|\gamma^2}\right) = \mathcal{O}\left(\frac{\ln n + \ln\ln t}{\gamma^2 \ln t}\right).$$

2. ***(General case.)*** *Suppose $\eta_j = 1/\sqrt{j+1}$, and $t \ge 5$, and ${\sqrt{t}}/{\ln(t)^3} \ge {n(1+R)}/{\gamma^2}$, where $R := \sup_{j<t} |\Pi_S w_j| = \mathcal{O}(1)$. Then*
$$|\Pi_S w_t| = \Theta\left(1\right) \quad \text{and} \quad \max\left\{|\Pi_S \bar{w}_t - \bar{v}|^2, |\Pi_S \bar{w}_t - \bar{v}|^2\right\} = \mathcal{O}\left(\frac{\ln(t)^2}{\sqrt{t}}\right),$$
and if A_c is nonempty, then $|\Pi_{S^\perp} w_t| = \Theta(\ln(t))$, and
$$\max\left\{\left|\frac{w_t}{|w_t|} - \bar{u}\right|^2, \left|\frac{\bar{w}_t}{|\bar{w}_t|} - \bar{u}\right|^2\right\} = \mathcal{O}\left(\frac{\ln n + \ln|w_t|}{|w_t|\gamma^2}\right) = \mathcal{O}\left(\frac{\ln n + \ln\ln|t|}{\gamma^2 \ln(t)}\right).$$

Establishing rates of parameter convergence not only relies upon all previous sections, but also is much more involved, and will be split into multiple subsections.

The easiest place to start the analysis is to dispense with the behavior over S. For convenience, define
$$\mathcal{R}_S(w) := \frac{L(A_S w)}{n}, \quad \mathcal{R}_c(w) := \frac{L(A_c w)}{n}, \quad \bar{\mathcal{R}} := \inf_{w\in\mathbb{R}^d} \mathcal{R}(w),$$
and note $\mathcal{R}(w) = \mathcal{R}_S(w) + \mathcal{R}_c(w)$.

Convergence over S is a consequence of strong convexity and risk convergence (cf. Theorem 3).

Lemma 8 *Let $\ell \in \{\ell_{\exp}, \ell_{\log}\}$, and λ denote the modulus of strong convexity of $\mathcal{R}_S$ over the 1-sublevel set (guaranteed positive by Theorem 2). Then for any $t \ge 1$,*
$$\max\left\{|\Pi_S w_t - \bar{v}|^2,\ |\Pi_S \bar{w}_t - \bar{v}|^2\right\} \le \frac{2}{\lambda}\min\left\{1,\ \frac{\exp\left(|\bar{v}|\right)}{t} + \frac{|\bar{v}|^2 + \ln(t)^2/\gamma^2}{2\sum_{j=0}^{t-1}\eta_j}.\right\}$$

Proof By Theorem 2, $\mathcal{R}_S(\bar{v}) = \bar{\mathcal{R}}$. Thus, by strong convexity, for $w \in \{w_t, \bar{w}_t\}$ (whereby $\mathcal{R}(w) \le \mathcal{R}(w_t)$),
$$|\Pi_S w - \bar{v}|^2 \le \frac{2}{\lambda}\left(\mathcal{R}_S(w) - \mathcal{R}_S(\bar{v})\right) \le \frac{2}{\lambda}\left(\mathcal{R}(w_t) - \inf_{w\in\mathbb{R}^d}\mathcal{R}(w)\right).$$
The bound follows by noting $\mathcal{R}(w_t) \le \mathcal{R}(w_0) \le 1$, and alternatively invoking in Theorem 3. ■

If A_c is empty, the proof is complete by plugging $\eta_j = 1/\sqrt{j+1}$ into Lemma 8. The rest of this section establishes convergence to $\bar{u} \in S^\perp$.

4.1. Bounding $|\Pi_\perp w_t|$

Before getting into the guts of Theorem 7, this section will establish bounds on $|w_t|$. These bounds are used in Theorem 7 in two ways. First, as will be clear in the next section, it is natural to prove rates with $|w_t|$ in the denominator, thus lower bounding $|w_t|$ with a function of t gives the desired bound.

On the other hand, it will be necessary to also produce an upper bound since the proofs will need a warm start, which will place a $|w_{t_0}|$ in the numerator. The need for this warm start will be discussed in subsequent sections, but a key purpose is to make ℓ_{exp} and ℓ_{log} appear more similar.

Note that this section focuses on behavior within A_c; suppose that A_c is nonempty, since otherwise $A = A_S$ and Theorem 7 follows from Lemma 8. When A_c is nonempty, the solution is off at infinity, and $|w_t|$ grows without bound. For this reason, these bounds will be on w_t rather than $\Pi_\perp w_t$, since $|w_t - \Pi_\perp w_t| = |\Pi_S w_t| = \mathcal{O}(1)$ by Lemma 8.

Lemma 9 *Suppose A_c is nonempty, $\ell \in \{\ell_{\text{exp}}, \ell_{\text{log}}\}$, and step sizes satisfy $\eta_j \le 1$. Define $R := \sup_{j<t} |\Pi_\perp w_j - w_j|$, where Lemma 8 guarantees $R = \mathcal{O}(1)$. Let $n_c > 0$ denote the number of rows in A_c. For any $t \ge 1$,*

$$|w_t| \le \max\left\{ \frac{4\ln(t)}{\gamma^2},\ \frac{4R}{\gamma^2},\ 2 \right\},$$

$$|w_t| \ge \min\left\{ \ln(t) - \ln(2) - |\bar{v}|,\ \ln\left(\sum_{j=0}^{t-1} \eta_j\right) - \ln\left(|\bar{v}|^2 + \frac{\ln(t)^2}{\gamma^2}\right) \right\} - R + \ln\ln(2) - \ln\left(\frac{n}{n_c}\right).$$

The proof of Lemma 9 is involved, and analyzes a key quantity which is extracted from the perceptron convergence proof (Novikoff, 1962). Specifically, the perceptron convergence proof establishes bounds on the number of mistakes up through time t by upper and lower bounding $\langle w_t, \bar{u}\rangle$. As the perceptron algorithm is stochastic gradient descent on the ReLU loss $z \mapsto \max\{0, z\}$, the number of mistakes up through time t is more generally the quantity $\sum_{j<t} \ell'(\langle w_j, -x_j y_j\rangle)$. Generalizing this, the proof here controls the quantity $\ell'_{<t} := \frac{1}{n}\sum_{j<t} \eta_j |\nabla L(A_c w_j)|_1$. Rather than obtaining bounds on this by studying $\langle w_t, \bar{u}\rangle$, the proof here works with $\langle w_t - \bar{u}\cdot r, \bar{u}\rangle = \langle \Pi_\perp w_t - \bar{u}\cdot r, \bar{u}\rangle$, with the strategic choice $r := \ln(t)/\gamma$.

On one hand, $\langle \Pi_\perp w_t - \bar{u}\cdot r, \bar{u}\rangle$ is upper bounded by $\|\Pi_\perp w_t - \bar{u}\cdot r\|$, whose upper bound is given by the following lemma, which is a modification of Lemma 4 with $\mathcal{R}$ replaced with $\mathcal{R}_c$.

Lemma 10 *Let any $\ell \in \{\ell_{\text{exp}}, \ell_{\text{log}}\}$ be given, and suppose $\eta_j \le 1$. For any $u \in S^\perp$ and $t \ge 1$,*

$$|\Pi_\perp w_t - u|^2 \le |u|^2 + 2 + \sum_{j<t} 2\eta_j \left(\mathcal{R}_c(u) - \mathcal{R}_c(w_j)\right) + \sum_{j<t} 2\eta_j \left\langle \nabla\mathcal{R}_c(w_j), w_j - \Pi_\perp w_j \right\rangle.$$

The proof is similar to that of Lemma 4, but inserts $\Pi_\perp$ in a few key places.

On the other hand, to lower bound $\langle \Pi_\perp w_t - \bar{u}\cdot r, \bar{u}\rangle$, notice that

$$\begin{aligned}\langle \Pi_\perp w_t, \bar{u}\rangle &= \frac{1}{n}\left\langle -\Pi_\perp \sum_{j<t} \eta_j A^\top \nabla L(A w_j), \bar{u}\right\rangle = \frac{1}{n}\left\langle -\sum_{j<t} \eta_j A_\perp^\top \nabla L(A_c w_j), \bar{u}\right\rangle \\ &= \sum_{j<t} \frac{\eta_j |\nabla L(A_c w_j)|_1}{n} \left\langle -A_\perp^\top \frac{\nabla L(A_c w_j)}{|\nabla L(A_c w_j)}|_1, \bar{u}\right\rangle \ge \sum_{j<t} \frac{\gamma\eta_j |\nabla L(A_c w_j)|_1}{n} = \gamma \ell'_{<t},\end{aligned}$$

where the last step uses the definition of $\ell'_{<t}$ from above. After a fairly large amount of careful but uninspired manipulations, Lemma 9 follows. A key issue throughout this and subsequent subsections is the handling of cross terms between A_c and A_S.

4.2. Parameter convergence over $S^\perp$: separable case

First consider the simple case where the data is separable; in other words, $S^\perp = \mathbb{R}^d$ and $A = A_c = A_\perp$. The analysis in this case hinges upon two ideas.

1. To show that w is close to $\bar{u}$ in direction, it is essential to lower bound $\langle \bar{u}, w \rangle / |w|$, since

$$|\bar{u} - {}^{w}/_{|w|}|^2 = 2 - \frac{2\langle \bar{u}, w\rangle}{|w|}.$$

   To control this, recall the representation $\bar{u} = -A^\top \bar{q}/\gamma$, where $\bar{q}$ is a dual optimum (cf. Theorem 2). With this in hand, and an appropriate choice of convex function g, the Fenchel-Young inequality gives for any w_t, $t \geq 1$, and any w such that $g(Aw) \leq g(Aw_t)$,

$$-\frac{\langle \bar{u}, w\rangle}{|w|} = \frac{\langle \bar{q}, Aw\rangle}{\gamma|w|} \leq \frac{g^*(\bar{q}) + g(Aw)}{\gamma|w|} \leq \frac{g^*(\bar{q}) + g(Aw_t)}{\gamma|w|}. \tag{2}$$

   The significance of working with w with $g(Aw) \leq g(Aw_t)$ is to allow the proof to handle both $\bar{w}_t$ and w_t simultaneously.

2. The other key idea is to use $g = \ln(L_{\exp}/n)$. With this choice, both preceding numerator terms can be bounded: $g^*(q) = \ln n + \sum_{i=1}^n q_i \ln q_i \leq \ln n$ for any probability vector q, whereas $g(Aw_t)$ can be upper bounded by applying $\ln$ to both sides of Lemma 6, as eq. (1), yielding an expression which will cancel with the denominator since $|w_t| \leq \sum_{j<t} \hat{\eta}_j \gamma_j$.

For illustration, suppose temporarily that $\ell = \ell_{\exp}$. Still let w_t be any gradient descent iterate with $t \geq 1$, and w satisfy $g(Aw) \leq g(Aw_t)$. Continuing as in eq. (2), but also assuming $\eta_j \leq 1$ and invoking Lemma 6 to control $g(Aw_t) = \ln \mathcal{R}(w_t)$ and noting $g^*(\bar{q}) \leq \ln(n)$, $\mathcal{R}(w_0) = 1$,

$$\begin{aligned}
\frac{1}{2}\left|\frac{w}{|w|} - \bar{u}\right|^2 &\leq 1 + \frac{\ln \mathcal{R}(w_t)}{|w|\gamma} + \frac{\ln(n)}{|w|\gamma} \\
&\leq 1 + \frac{\ln \mathcal{R}(w_0)}{|w|\gamma} - \frac{\sum_{j=0}^{t-1} \hat{\eta}_j (1 - \hat{\eta}_j/2)\gamma_j^2}{|w|\gamma} + \frac{\ln(n)}{|w|\gamma} \\
&= 1 - \frac{\sum_{j=0}^{t-1} \hat{\eta}_j \gamma_j^2}{|w|\gamma} + \frac{\sum_{j=0}^{t-1} \hat{\eta}_j^2 \gamma_j^2}{2|w|\gamma} + \frac{\ln(n)}{|w|\gamma}.
\end{aligned}$$

To simplify this further, note $\eta_j \leq 1$ and Lemma 6 also imply

$$\sum_{j=t_0}^{t-1} \hat{\eta}_j^2 \gamma_j^2 = \sum_{j=t_0}^{t-1} \eta_j^2 |\nabla \mathcal{R}(w_j)|^2 \leq 2 \sum_{j=t_0}^{t-1} (\mathcal{R}(w_j) - \mathcal{R}(w_{j+1})) \leq 2,$$

and moreover the definition $\gamma = \min\left\{|A_\perp^\top q| : q \geq 0, \sum_i q = 1\right\}$ (cf. Theorem 2) implies

$$\gamma_j = |\nabla(\ln \mathcal{R})(w_j)| = \frac{|A^\top \nabla L(w_j)|}{L(Aw_j)} \geq \gamma.$$

Combining these simplifications,

$$\frac{1}{2}\left|\frac{w}{|w|}-\bar{u}\right|^2 \leq 1-\frac{\sum_{j=0}^{t-1}\hat{\eta}_j\gamma_j\gamma}{|w|\gamma}+\frac{2}{2|w|\gamma}+\frac{\ln(n)}{|w|\gamma}\leq\frac{1+\ln n}{|w|\gamma}.$$

To finish, invoke the preceding inequality with $w\in\{\bar{w}_t,w_t\}$, noting that $|w_t|=|\bar{w}_t|$. To produce a rate depending on t and not $|w_t|$, the lower bound on $|w_t|$ in Lemma 9 is applied with $|\bar{v}|=R=0$ and $n=n_c$ thanks to separability. The following proposition summarizes this derivation.

Proposition 11 *Suppose $\ell=\ell_{\exp}$ and $A=A_c=A_\perp$ (the separable case). Then, for all $t\geq 1$,*

$$\max\left\{\left|\frac{w_t}{|w_t|}-\bar{u}\right|^2,\left|\frac{\bar{w}_t}{|\bar{w}_t|}-\bar{u}\right|^2\right\}\leq\frac{2+2\ln n}{|w_t|\gamma},$$

where $|w_t|\geq\min\left\{\ln(t)-\ln 2,\ \ln\left(\sum_{j<t}\eta_j\right)-2\ln\ln t+2\ln\gamma\right\}+\ln\ln 2.$

Adjusting this derivation for $\ell_{\log}$ is clumsy, and will only be sketched here, instead mostly appearing in the appendix. The proof will still follow the same scheme, even defining $g=\ln\mathcal{R}_{\exp}$ (rather than $g=\ln\mathcal{R}$), and will use the following lemma to control ℓ and $\ell_{\exp}$ simultaneously.

Lemma 12 *For any $0<\epsilon\leq 1$, $\ell\in\{\ell_{\exp},\ell_{\log}\}$, if $\ell(z)\leq\epsilon$, then*

$$\frac{\ell'(z)}{\ell(z)}\geq 1-\epsilon\qquad\text{and}\qquad\frac{\ell_{\exp}(z)}{\ell(z)}\leq 2.$$

The general Fenchel-Young scheme from above can now be adjusted. The bound requires $\mathcal{R}(w_t)\leq\epsilon/n$, which implies $\max_i\ell((Aw_t)_i)\leq\epsilon$, meaning each point has some good margin, rendering $\ell_{\log}$ and $\ell_{\exp}$ similar.

Lemma 13 *Let $\ell\in\{\ell_{\exp},\ell_{\log}\}$. For any $0<\epsilon\leq 1$, and any t with $\mathcal{R}(w_t)\leq\epsilon/n$, and any w with $\mathcal{R}(w)\leq\mathcal{R}(w_t)$,*

$$\frac{\langle\bar{u},w\rangle}{|\bar{u}|\cdot|w|}\geq-\frac{\ln\mathcal{R}(w_t)}{|w|\gamma}-\frac{g^*(\bar{q})}{|w|\gamma}-\frac{\ln 2}{|w|\gamma}.$$

Proof By the condition, $\mathcal{R}(w)\leq\mathcal{R}(w_t)\leq\epsilon/n$, and thus for any $1\leq i\leq n$, $\ell(A_iw)\leq\epsilon\leq 1$. By Lemma 12, $\mathcal{R}_{\exp}(w)\leq 2\mathcal{R}(w)$. Then by Theorem 2 and the Fenchel-Young inequality,

$$\begin{aligned}\frac{\langle\bar{u},w\rangle}{|\bar{u}|\cdot|w|}&=-\frac{\langle\bar{q},Aw\rangle}{|w|\gamma}\geq-\frac{g^*(\bar{q})}{|w|\gamma}-\frac{g(Aw)}{|w|\gamma}=-\frac{g^*(\bar{q})}{|w|\gamma}-\frac{\ln\mathcal{R}_{\exp}(w)}{|w|\gamma}\\&\geq-\frac{g^*(\bar{q})}{|w|\gamma}-\frac{\ln 2}{|w|\gamma}-\frac{\ln\mathcal{R}(w)}{|w|\gamma}\geq-\frac{g^*(\bar{q})}{|w|\gamma}-\frac{\ln 2}{|w|\gamma}-\frac{\ln\mathcal{R}(w_t)}{|w|\gamma}.\end{aligned}$$

■

Proceeding as with the earlier $\ell_{\exp}$ derivation, since $g^*(\bar{q})\leq\ln(n)$, and using the upper bound on $\ln\mathcal{R}(w_t)$ from Lemma 6,

$$\frac{\langle\bar{u},w\rangle}{|\bar{u}|\cdot|w|}\geq-\frac{\ln(n)}{|w|\gamma}-\frac{\ln 2}{|w|\gamma}-\frac{\ln\mathcal{R}(w_{t_0})-\sum_{j=t_0}^{t-1}\hat{\eta}_j(1-\hat{\eta}_j/2)\gamma_j^2}{|w|\gamma}.$$

The role of the warm start t_0 here is to make $\ell_{\exp}$ and $\ell_{\log}$ behave similarly, in concrete terms allowing the application of Lemma 12 for $j \geq t_0$. For instance, the earlier $\ell_{\exp}$ proof used $\gamma_j \geq \gamma$, but this now becomes $\gamma_j \geq (1-\epsilon)\gamma$. Completing the proof of the separable case of Theorem 7 requires many such considerations which did not appear with $\ell_{\exp}$, including an upper bound on $|w_{t_0}|$ via Lemma 9.

4.3. Parameter convergence over $S^\perp$: general case

The scheme from the separable case does not directly work: for instance, the proofs relied upon $\gamma_i \geq \gamma$, but now $\gamma_i \to 0$. This term γ_i arose by applying Lemma 6 to control $\ln \mathcal{R}(w_t)$, which in the separable case decreased to $-\infty$, in the general case, however, it can be bounded below.

The fix is to replace $\mathcal{R}(w_t)$ with $\mathcal{R}_c(w_t)$, or rather $\mathcal{R}(w_t) - \bar{\mathcal{R}} = \mathcal{R}(w_t) - \inf_w \mathcal{R}(w)$; this quantity goes to 0, and there is again a hope of exhibiting the fortuitous cancellations which proved parameter convergence. More abstractly, by subtracting $\bar{\mathcal{R}}$, the proof is again trying to work in the separable case, though there will be cross-terms to contend with.

The first step, then, is to replace the appearance of $\mathcal{R}$ in earlier Fenchel-Young approach (cf. Lemma 13) with $\mathcal{R}(w_t) - \bar{\mathcal{R}}$.

Lemma 14 *Let $\ell \in \{\ell_{\exp}, \ell_{\log}\}$. For any $0 < \epsilon \leq 1$, $t \geq 1$, and any w with $\mathcal{R}(w) - \bar{\mathcal{R}} \leq \mathcal{R}(w_t) - \bar{\mathcal{R}} \leq \epsilon/n$,*

$$\frac{\langle \bar{u}, w\rangle}{|\bar{u}|\cdot|w|} \geq \frac{-\ln\left(\mathcal{R}(w_t) - \bar{\mathcal{R}}\right)}{\gamma|w|} - \frac{\ln 2 + g^*(\bar{q}) + |\Pi_S(w)|}{\gamma|w|}.$$

Note the appearance of the additional cross term $|\Pi_S(w)|$; by Lemma 8, this term is bounded.

The next difficulty is to replace the separable case's use of Lemma 6 to control $\ln \mathcal{R}(w_t)$ with something controlling $\ln(\mathcal{R}(w_t) - \bar{\mathcal{R}})$.

Lemma 15 *Suppose $\ell \in \{\ell_{\log}, \ell_{\exp}\}$ and $\hat{\eta}_j \leq 1$ (meaning $\eta_j \leq 1/\mathcal{R}(w_j)$). Also suppose that j is large enough such that $\mathcal{R}(w_j) - \bar{\mathcal{R}} \leq \min\{\epsilon/n, \lambda(1-r)/2\}$ for some $\epsilon, r \in (0,1)$, where λ is the strong convexity modulus of $\mathcal{R}_S$ over the 1-sublevel set. Then*

$$\mathcal{R}(w_{j+1}) - \bar{\mathcal{R}} \leq \left(\mathcal{R}(w_j) - \bar{\mathcal{R}}\right)\exp\left(-r(1-\epsilon)\gamma\gamma_j\hat{\eta}_j\left(1-\hat{\eta}_j/2\right)\right).$$

Moreover, if there exists a sequence $(w_j)_{j=t_0}^{t-1}$ such that the above condition holds, then

$$\mathcal{R}(w_t) - \bar{\mathcal{R}} \leq \left(\mathcal{R}(w_{t_0}) - \bar{\mathcal{R}}\right)\exp\left(-r(1-\epsilon)\gamma\sum_{j=t_0}^{t-1}\hat{\eta}_j\left(1-\hat{\eta}_j/2\right)\gamma_j\right).$$

The proof of Lemma 15 is quite involved, but boils down to the following case analysis. The first case is that there is more error over S, whereby a strong convexity argument gives a lower bound on the gradient. Otherwise, the error is larger over S^c, which leads to a big step in the direction of $\bar{u}$.

A key property of the upper bound in Lemma 15 is that it has replaced γ_j^2 in Lemma 6 with $\gamma_j\gamma$. Plugging this bound into the Fenchel-Young scheme in Lemma 14 will now fortuitously cancel γ, which leads to the following promising bound.

Lemma 16 *Let $\ell \in \{\ell_{\log}, \ell_{\exp}\}$ and $0 < \epsilon \le 1$ be given, and select t_0 so that $\mathcal{R}(w_{t_0}) - \bar{\mathcal{R}} \le \epsilon/n$. Then for any $t \ge t_0$ and any w such that $\mathcal{R}(w) \le \mathcal{R}(w_t)$,*

$$\frac{\langle \bar{u}, w \rangle}{|\bar{u}| \cdot |w|} \ge \frac{r(1-\epsilon)\sum_{j=t_0}^{t-1} \hat{\eta}_j(1-\hat{\eta}_j/2)\gamma_j}{|w|} - \frac{\ln 2}{\gamma|w|} - \frac{|\Pi_S(w)|}{\gamma|w|}.$$

As in the separable case, an extended amount of careful massaging, together with Lemma 6 and Lemma 9 to control the warm start, is sufficient to establish the parameter convergence of Theorem 7 in the general case.

5. Related work

The technical basis for this work is drawn from the literature on AdaBoost, which was originally stated for separable data (Freund and Schapire, 1997), but later adapted to general instances (Mukherjee et al., 2011; Telgarsky, 2012). This analysis revealed not only a problem structure which can be refined into the $(S, S^\perp)$ used here, but also the convergence to maximum margin solutions (Telgarsky, 2013). Since the structural analysis is independent of the optimization method, the key structural result, Theorem 2 in Section 2, can be partially found in prior work; the present version provides not only an elementary proof, but moreover differs by providing S and $S^\perp$ (and not just a partition of the data) and the subsequent construction of a unique $\bar{u}$ and its properties.

The remainder of the analysis has some connections to the AdaBoost literature, for instance when providing smoothness inequalities for $\mathcal{R}$ (cf. Lemma 6). There are also some tools borrowed from the convex optimization literature, for instance smoothness-based convergence proofs of gradient descent (Nesterov, 2004; Bubeck, 2015), and also from basic learning theory, namely an adaptation of ideas from the perceptron convergence proof in order to bound $|w_t|$ (Novikoff, 1962).

Another close line of work is an analysis of gradient descent for logistic regression when the data is separable (Soudry et al., 2017; Gunasekar et al., 2018; Nacson et al., 2018). The analysis is conceptually different (tracking $(w_j)_{j\ge 0}$ in all directions, rather than the Fenchel-Young and smoothness approach here), and (assuming linear separability) achieves a better rate than the one here, although it is not clear if this is possible in the nonseparable case. Other work shows that not just gradient descent but also steepest descent with other norms can lead to margin maximization (Gunasekar et al., 2018; Telgarsky, 2013), and that constructing loss functions with an explicit goal of margin maximization can lead to better rates (Nacson et al., 2018). Another line of work uses condition numbers to analyze these problems with a different parameterization (Freund et al., 2017).

There has been a large literature on risk convergence of logistic regression. (Hazan et al., 2007; Mahdavi et al., 2015) assume a bounded domain, exploit the exponential concavity of the logistic loss and apply online Newton step to give a $\mathcal{O}(1/t)$ rate. However, on a domain with norm bound D, the exponential concavity term is $1/\beta = e^D$. In the unbounded setting considered in this paper, this term is a factor of $\mathrm{poly}(t)$ since $|w_t| = \Theta(\ln(t))$ (cf. Lemma 9). (Bach and Moulines, 2013; Bach, 2014) assume optima exist, and the rates depend inversely on the smallest eigenvalue of the Hessian at optima, which similarly introduce a factor of $\mathrm{poly}(t)$ when there is no finite optimum.

There is some work in online learning on optimization over unbounded sets, for instance bounds where the regret scales with the norm of the comparator (Orabona and Pal, 2016; Streeter and McMahan, 2012). By contrast, as the present work is not adversarial and instead has a fixed training set, part of the work (a consequence of the structural result, Theorem 2) is the existence of a good, small comparator.

References

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $\mathcal{O}(1/n)$. In *Advances in neural information processing systems*, pages 773–781, 2013.

Francis R. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627, 2014.

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

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 2015.

Robert M. Freund, Paul Grigas, and Rahul Mazumder. Condition number analysis of logistic regression, and its implications for first-order solution methods. INFORMS, 2017.

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

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

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

Mehrdad Mahdavi, Lijun Zhang, and Rong Jin. Lower and upper bounds on the generalization of stochastic exponentially concave optimization. In *Conference on Learning Theory*, pages 1305–1320, 2015.

Indraneel Mukherjee, Cynthia Rudin, and Robert Schapire. The convergence rate of AdaBoost. In *COLT*, 2011.

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

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

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

Francesco Orabona and David Pal. Coin betting and parameter-free online learning. In *Advances in neural information processing systems*, 2016.

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

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

Matthew Streeter and Brendan McMahan. No-regret algorithms for unconstrained online convex optimization. In *Advances in neural information processing systems*, 2012.

Matus Telgarsky. A primal-dual convergence analysis of boosting. *JMLR*, 13:561–606, 2012.

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

Appendix A. Omitted proofs from Section 2

Before proving Theorem 2, note the following result characterizing margin maximization over $S^\perp$.

Lemma 17 *Suppose $A_\perp$ has $n_c > 0$ rows and there exists u with $A_\perp u < 0$. Then*

$$\gamma := -\min\left\{\max_i (A_\perp u)_i : |u| = 1\right\} = \min\left\{|A_\perp^\top q| : q \geq 0, \sum_i q_i = 1\right\} > 0.$$

Moreover there exists a unique nonzero primal optimum $\bar{u}$, and every dual optimum $\bar{q}$ satisfies $\bar{u} = -A_\perp^\top \bar{q}/\gamma$.

Proof To start, note $\gamma > 0$ since there exists u with $A_\perp u < 0$.

Continuing, for convenience define simplex $\Delta := \{q \in \mathbb{R}^{n_c} : q \geq 0, \sum_i q_i = 1\}$, and convex indicator $\iota_K(z) = \infty \cdot \mathbb{1}[z \in K]$. With this notation, note the Fenchel conjugates

$$\begin{aligned}\iota_\Delta^*(v) &= \sup_{q\in\Delta} \langle v, q\rangle = \max_i v_i,\\ (|\cdot|_2)^*(q) &= \iota_{|\cdot|_2 \leq 1}(q).\end{aligned}$$

Combining this with the Fenchel-Rockafellar duality theorem (Borwein and Lewis, 2000, Theorem 3.3.5, Exercise 3.3.9.f),

$$\begin{aligned}\min |A_\perp^\top q|_2 + \iota_\Delta(q) &= \max -\iota_{|\cdot|_2\leq 1}(u) - \iota_\Delta^*(-A_\perp u)\\ &= \max\left\{-\max_i(-A_\perp u)_i : |u|_2 \leq 1\right\}\\ &= -\min\left\{\max_i (A_\perp u)_i : |u|_2 \leq 1\right\},\end{aligned}$$

and moreover every primal-dual optimal pair $(\bar{u}, \bar{q})$ satisfies $A_\perp^\top \bar{q} \in \partial\left(\iota_{|\cdot|_2\leq 1}\right)(-\bar{u})$, which means $\bar{u} = -A_\perp^\top \bar{q}/\gamma$.

It only remains to show that $\bar{u}$ is unique. Since $\gamma > 0$, necessarily any primal optimum has unit length, since the objective value will only decrease by increasing the length. Consequently, suppose u_1 and u_2 are two primal optimal unit vectors. Then $u_3 := (u_1 + u_2)/2$ would satisfy

$$\max_i (A_\perp u_3)_i = \frac{1}{2}\max_i (A_\perp u_1 + A_\perp u_2)_i \leq \frac{1}{2}\left(\max_i (A_\perp u_1) + \max_j (A_\perp u_2)_j\right) = \max_i (A_\perp u_1)_i,$$

but then the unit vector $u_4 := u_3/|u_3|$ would have $|u_4| > |u_3|$ when $u_1 \neq u_2$, which implies $\max_i(A_\perp u_4) < \max_i(A_\perp u_3) = \max_i(A_\perp u_1)$, a contradiction. ■

The proof of Theorem 2 follows.

Proof (of Theorem 2) Partition the rows of A into A_c and A_S as follows. For each row i, put it in A_c if there exists u_i so that $Au_i \leq 0$ (coordinate-wise) and $(Au_i)_i < 0$; otherwise, when no such u_i exists, add this row to A_S. Define $S := \text{span}(A_S^\top)$, the linear span of the rows of A_S. This has the following consequences.

- To start, $S^\perp = \text{span}(A_S^\top)^\perp = \ker(A_S) \subseteq \ker(A)$.
- For each row i of A_c, the corresponding u_i has $A_S u_i = 0$, since otherwise $Au_i \leq 0$ implies there would be a negative coordinate of $A_S u_i$, and this row should be in A_c not A_S. Combining this with the preceding point, $u_i \in \ker(A_S) = S^\perp$. Define $\tilde{u} := \sum_i u_i \in S^\perp$, whereby $A_c\tilde{u} < 0$ and $A_S\tilde{u} = 0$. Lastly, $\tilde{u} \in \ker(A_S)$ implies moreover that $A_\perp \tilde{u} = A_c\tilde{u} < 0$. As such, when A_c has a positive number of rows, Lemma 17 can be applied, resulting in the desired unique $\bar{u} = -A_\perp^\top/\gamma \in S^\perp$ with $\gamma > 0$.
- S, $S^\perp$, A_S, and A_c, and $\bar{u}$ are unique and constructed from A alone, with no dependence on ℓ.
- If A_c is empty, there is nothing to show, thus suppose A_c is nonempty. Since $\lim_{z\to-\infty} \ell(z) = 0$,
$$0 \leq \inf_{w\in\mathbb{R}^d} L(A_c w) \leq \inf_{w\in S^\perp} L(A_c w) \leq \inf_{u\in S^\perp} L(A_c u) \leq \lim_{r\to\infty} L(r \cdot A_c\bar{u}) = 0.$$
Since these inequalities start and end with 0, they are equalities, and consequently $\inf_{w\in\mathbb{R}^d} = \inf_{u\in S^\perp} L(A_c u) = 0$. Moreover,
$$\begin{aligned}\inf_{w\in\mathbb{R}^d} L(Aw) &\leq \inf_{\substack{v\in S\\ u\in S^\perp}} \left(L(A_S(u+v) + L(A_c(u+v))\right) = \inf_{v\in S}\left(L(A_S v) + \inf_{u\in S^\perp} L(A_c(u+v))\right)\\ &\leq \left(\inf_{v\in S} L(A_S v)\right) + \left(\inf_{u\in S^\perp} L(A_c u)\right) = \left(\inf_{v\in S} L(A_S v)\right) \leq \inf_{w\in\mathbb{R}^d} L(Aw).\end{aligned}$$
which again is in fact a chain of equalities.
- For every $v \in S$ with $|v| > 0$, there exists a row a of A_S such that $\langle a, v\rangle > 0$. To see this, suppose contradictorily that $A_S v \leq 0$. It cannot hold that $A_S v = 0$, since $v \neq 0$ and $\ker(A_S) \subseteq S^\perp$. this means $A_S v \leq 0$ and moreover $(A_S v)_i < 0$ for some i. But since $A\bar{u} \leq 0$ and $A_c\bar{u} < 0$, then for a sufficiently large $r > 0$, $A(v + r\bar{u}) \leq 0$ and $(A_S(v + r\bar{u}))_j < 0$, which means row j of A_S should have been in A_c, a contradiction.
- Consider any $v \in S \setminus \{0\}$. By the preceding point, there exists a row a of A_S such that $\langle a, v\rangle > 0$. Since $\ell(0) > 0$ (because $\ell'' > 0$) and $\lim_{z\to-\infty})$ and $\lim_{z\to-\infty} = 0$, there exists $r > 0$ so that $\ell(-r\langle a, v\rangle) = \ell(0)/2$. By convexity, for any $t > 0$, setting $\alpha := r/(t+r)$ and noting $\alpha\langle a, tv\rangle + (1-\alpha)\langle a, -rv\rangle = 0$,
$$\alpha\ell(t\langle a, v\rangle) \geq \ell(0) - (1-\alpha)\ell(-r\langle a, v\rangle) = \left(\frac{1+\alpha}{2}\right)\ell(0),$$

thus $\ell(t\langle a, v\rangle) \geq \left(\frac{1+\alpha}{2\alpha}\right)\ell(0) = \left(\frac{t+2r}{2r}\right)\ell(0)$, and

$$\lim_{t\to\infty} \frac{L(tAv) - L(0)}{t} \geq \lim_{t\to\infty} \frac{\ell(t\langle a, v\rangle) - n\ell(0)}{t} \geq \lim_{t\to\infty} \frac{\ell(0)}{2r}\left(\frac{(t+2r) - 2nr}{t}\right) \geq \frac{\ell(0)}{2r} > 0.$$

Consequently, $L \circ A$ has compact sublevel sets over S (Hiriart-Urruty and Lemaréchal, 2001, Proposition B.3.2.4).

- Note $\nabla^2 L(v) = \operatorname{diag}(\ell''(v_1), \ldots, \ell''(v_n))$. Moreover, since $\ker(A) \subseteq S^\perp$, then the image $B_0 := \{Av : v \in S, |v| = 1\}$ over the surface of the ball in S through A is a collection of vectors with positive length. Thus for any compact subset $S_0 \subseteq S$,

$$\begin{aligned}\inf_{\substack{v_1\in S_0\\ v_2\in S, |v_2|=1}} v_2^\top \nabla^2(L\circ A)(v_1)v_2 = \inf_{\substack{v_1\in S_0\\ v_2\in S, |v_2|=1}} (Av_2)^\top \nabla^2 L(Av_1)(Av_2) = \inf_{\substack{v_1\in S_0\\ v_3\in B_0}} v_3^\top \nabla^2 L(Av_1)v_3 \\ \geq \inf_{\substack{v_1\in S_0\\ v_3\in B_0}} |v_3|^2 \min_i \ell''((v_1)_i) > 0,\end{aligned}$$

  the final inequality since the minimization is of a continuous function over a compact set, thus attained at some point, and the infimand is positive over the domain. Consequently, $L \circ A$ is strongly convex over compact subsets of S.

- Since $L \circ A$ is strongly convex over S and moreover has bounded sublevel sets over S, it attains a unique optimum over S.

■

Appendix B. Omitted proofs from Section 3

To start, note how the three key lemmas provided in the main text lead to a proof of Theorem 3.

Proof (of Theorem 3) Since $\eta_j \leq 1$, then Lemma 6 guarantees both that the desired smoothness inequality holds and that function values decrease, whereby $\hat\eta_j \leq \eta_j \mathcal{R}(w_j) \leq \eta_j$. Thus, by Lemma 4, for any $z \in \mathbb{R}^d$,

$$\begin{aligned}2\left(\sum_{j<t}\eta_j\right)(\mathcal{R}(w_t) - \mathcal{R}(z)) &\leq 2\sum_{j<t}\eta_j\left(\mathcal{R}(w_j) - \mathcal{R}(z)\right) + 2\sum_{j<t}\eta_j\left(\mathcal{R}(w_{j+1} - \mathcal{R}(w_j)\right) \\ &\leq 2\sum_{j<t}\eta_j\left(\mathcal{R}(w_j) - \mathcal{R}(z)\right) - \sum_{j<t}\frac{\eta_j}{1-\eta_j/2}\left(\mathcal{R}(w_j) - \mathcal{R}(w_{j+1})\right) \\ &\leq |w_0 - z|^2 - |w_t - z|^2 \\ &\leq |z|^2.\end{aligned}$$

Consequently, by the choice $z := \bar v + \bar u(\ln(t)/\gamma$ and Lemma 5,

$$\mathcal{R}(w_t) \leq \mathcal{R}(z) + \frac{|z|^2}{2\sum_{j<t}\eta_j} \leq \inf_w \mathcal{R}(w) + \frac{\exp(|\bar v|)}{t} + \frac{|\bar v|^2 + \ln(t)^2/\gamma^2}{2\sum_{j<t}\eta_j}.$$

■

To fill out the proof, first comes the smoothness-based risk guarantee.

Proof (of Lemma 4) Define $r_j := \eta_j(1 - \beta\eta_j/2)$. For any j,

$$\begin{aligned} |w_{j+1} - z|^2 &= |w_j - z|^2 + 2\eta_j \langle \nabla f(w_j), z - w_j \rangle + \eta_j^2 |\nabla f(w_j)|^2 \\ &\leq |w_j - z|^2 + 2\eta_j (f(z) - f(w_j)) + \frac{\eta_j^2}{r_j} (f(w_j) - f(w_{j+1})). \end{aligned}$$

Summing this inequality over $i \in \{0, \ldots, t-1\}$ and rearranging gives the bound. ■

With that out of the way, the remainder of this subsection establishes smoothness properties of $\mathcal{R}$. For convenience, for the rest of this subsection define $w' := w - \eta\nabla\mathcal{R}(w) = w - \eta A^\top \nabla L(Aw)/n$. Additionally, suppose throughout that ℓ is twice differentiable and $\max_i |A_i| \leq 1$.

Lemma 18 *For any $w \in \mathbb{R}^d$,*

$$\mathcal{R}(w') \leq \mathcal{R}(w) - \eta|\nabla\mathcal{R}(w)|^2 + \frac{\eta^2}{2}|\nabla\mathcal{R}(w)|^2 \max_{v\in[w,w']} \sum_i \ell''(A_i v)/n.$$

Proof By Taylor expansion,

$$\mathcal{R}(w') \leq \mathcal{R}(w) - \eta|\nabla\mathcal{R}(w)|^2 + \frac{1}{2} \max_{v\in[w,w']} \sum_i (A_i(w - w'))^2 \ell''(A_i v)/n.$$

By Hölder's inequality,

$$\max_{v\in[w,w']} \sum_i (A_i(w - w'))^2 \ell''(A_i v) \leq \max_{v\in[w,w']} |A(w - w')|_\infty^2 \sum_i \ell''(A_i v).$$

Since $\max_i |A_i| \leq 1$,

$$|A(w-w')|_\infty^2 = \eta^2 |A\nabla\mathcal{R}(w)|_\infty^2 = \eta^2 \max_i \langle A_i, \nabla\mathcal{R}(w)\rangle^2 \leq \eta^2 \max_i |A_i|^2 |\nabla\mathcal{R}(w)|^2 \leq \eta^2 |\nabla\mathcal{R}(w)|^2.$$

Thus

$$\mathcal{R}(w') \leq \mathcal{R}(w) - \eta|\nabla\mathcal{R}(w)|^2 + \frac{\eta^2}{2}|\nabla\mathcal{R}(w)|^2 \max_{v\in[w,w']} \sum_i \ell''(A_i v)/n.$$

■

Lemma 19 *Suppose $\ell', \ell'' \leq \ell$ and ℓ is convex. Then, for any $w \in \mathbb{R}^d$,*

$$\max_{v\in[w,w']} \sum_i \ell''(A_i v)/n \leq \max\left\{\mathcal{R}(w), \mathcal{R}(w')\right\}.$$

Define $\hat\eta := \eta\mathcal{R}(w)$ and suppose $\hat\eta \leq 1$; then $\mathcal{R}(w') \leq \mathcal{R}(w)$ and

$$\mathcal{R}(w') \leq \mathcal{R}(w) \left(1 - \hat\eta(1 - \hat\eta/2) \frac{|\nabla\mathcal{R}(w)|^2}{\mathcal{R}(w)^2}\right).$$

Proof Since $\ell'' \leq \ell$ and ℓ is convex,

$$\max_{v\in[w,w']} \sum_i \ell''(A_i v)/n \leq \max_{v\in[w,w']} \sum_i \ell(A_i v)/n = \max_{v\in[w,w']} \mathcal{R}(v) = \max\left\{\mathcal{R}(w), \mathcal{R}(w')\right\}.$$

Combining this, the choice of η, and Lemma 18,

$$\begin{aligned}\mathcal{R}(w') &\leq \mathcal{R}(w) - \eta|\nabla\mathcal{R}(w)|^2 + \frac{\eta^2}{2}|\nabla\mathcal{R}(w)|^2 \max\left\{\mathcal{R}(w), \mathcal{R}(w')\right\}\\ &= \mathcal{R}(w) - \frac{\hat\eta|\nabla\mathcal{R}(w)|^2}{\mathcal{R}(w)}\left(1 - \frac{\hat\eta}{2}\frac{\max\left\{\mathcal{R}(w), \mathcal{R}(w')\right\}}{\mathcal{R}(w)}\right).\end{aligned}$$

As a final simplification, suppose $\mathcal{R}(w') > \mathcal{R}(w)$; since $\hat\eta \leq 1$ and $\ell' \leq \ell$ and $\max_i |A_i| \leq 1$,

$$\frac{\mathcal{R}(w')}{\mathcal{R}(w)} - 1 \leq \frac{\hat\eta|\nabla\mathcal{R}(w)|^2}{\mathcal{R}(w)^2}\left(\frac{\hat\eta}{2}\frac{\mathcal{R}(w')}{\mathcal{R}(w)} - 1\right) \leq \hat\eta\left(\frac{\hat\eta}{2}\frac{\mathcal{R}(w')}{\mathcal{R}(w)} - 1\right) \leq \frac{\hat\eta}{2}\frac{\mathcal{R}(w')}{\mathcal{R}(w)} - 1 \leq \frac{1}{2}\frac{\mathcal{R}(w')}{\mathcal{R}(w)} - 1,$$

a contradiction. Therefore $\mathcal{R}(w') \leq \mathcal{R}(w)$, which in turn implies

$$\mathcal{R}(w') \leq \mathcal{R}(w) - \frac{\hat\eta|\nabla\mathcal{R}(w)|^2}{\mathcal{R}(w)}\left(1 - \frac{\hat\eta}{2}\right).$$

■

Together, these pieces prove the desired smoothness inequality.

Proof (of Lemma 6) For any $j < t$, by Lemma 19 and the definition of γ_j,

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

Applying this recursively gives the bound.

Lastly,

$$|w_t| = \left|\sum_{j<t} \hat\eta_j q_j\right| \leq \sum_{j<t} |\hat\eta_j q_j| = \sum_{j<t} \hat\eta_j \gamma_j.$$

■

Appendix C. Omitted proofs from Section 4

This section will be split into subsections paralleling those in Section 4.

C.1. Bounding $|\Pi_\perp w_t|$

To start, the proof of the smoothness-based convergence guarantee, but with sensitivity to A_c.

Proof (of Lemma 10) Fix any $u \in S^\perp$. Expanding the square,

$$|\Pi_\perp w_{j+1} - u|^2 = |\Pi_\perp w_j - u|^2 + 2\eta_j \left\langle \Pi_\perp \nabla\mathcal{R}(w_j), u - \Pi_\perp w_j\right\rangle + \eta_j^2 \left|\Pi_\perp \nabla\mathcal{R}(w_j)\right|^2,$$

whose two key terms can be bounded as

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

the last inequality making use of smoothness, namely Lemma 6. Therefore

$$|\Pi_\perp w_{j+1} - u|^2 \le |\Pi_\perp w_j - u|^2 + 2\eta_j \left(\mathcal{R}_c(u) - \mathcal{R}_c(w_j) + \langle \nabla \mathcal{R}_c(w_j), w_j - \Pi_\perp w_j \rangle \right) + 2 \left(\mathcal{R}(w_j) - \mathcal{R}(w_{j+1}) \right) .$$

Applying $\sum_{j<t}$ to both sides and canceling terms yields

$$|\Pi_\perp w_t - u|^2 \le |u|^2 + 2 + \sum_{j<t} 2\eta_j \left(\mathcal{R}_c(u) - \mathcal{R}_c(w_j) \right) + \sum_{j<t} 2\eta_j \left\langle \nabla \mathcal{R}_c(w_j), w_j - \Pi_\perp w_j \right\rangle$$

as desired. ■

Proving Lemma 9 is now split into upper and lower bounds.

Proof (of upper bound in Lemma 9) For a fixed $t \ge 1$, define

$$u := \frac{\ln(t)}{\gamma} \bar{u}, \qquad \ell'_{<t} := \sum_{j<t} \eta_j \frac{|\nabla L(A_c w_j)|_1}{n}, \qquad R := \sup_{j<t} |\Pi_\perp w_j - w_j| \le |\bar{v}| + \sqrt{\frac{2}{\lambda}},$$

where the last inequality comes from Lemma 8.

The strategy of the proof is to rewrite various quantities in Lemma 10 with $\ell'_{<t}$, which after applying Lemma 10 cancel nicely to obtain an upper bound on $\ell'_{<t}$. This in turn completes the proof, since

$$|\Pi_\perp w_t| \le \sum_{j<t} \eta_j |\Pi_\perp \nabla \mathcal{R}(w_t)| \le \sum_{j<t} \eta_j |\nabla \mathcal{R}_c(w_j)| \le \sum_{j<t} \eta_j |\nabla L(A_c w_j)|/n = \ell'_{<t}.$$

Proceeding with this plan, first note (similarly to the main text)

$$\begin{aligned}
|\Pi_\perp w_t - u| &\ge \langle \Pi_\perp w_t - u, \bar{u} \rangle \\
&= \left\langle - \sum_{j<t} \eta_j \Pi_\perp \nabla \mathcal{R}(w_j)/n, \bar{u} \right\rangle - \langle u, \bar{u} \rangle \\
&= \sum_{j<t} \eta_j \left\langle -A_\perp^\top \nabla L(A_c w_j)/n, \bar{u} \right\rangle - \frac{\ln(t)}{\gamma} \\
&= \sum_{j<t} \eta_j \frac{|\nabla L(A_c w_j)|_1}{n} \left\langle -A_\perp^\top \frac{\nabla L(A_c w_j)}{|\nabla L(A_c w_j)|_1}, \bar{u} \right\rangle - \frac{\ln(t)}{\gamma} \\
&\ge \sum_{j<t} \eta_j \frac{|\nabla L(A_c w_j)|_1}{n} \gamma - \frac{\ln(t)}{\gamma} \\
&= \gamma \ell'_{<t} - \frac{\ln(t)}{\gamma},
\end{aligned}$$

and since $\ell' \leq \ell$

$$\begin{aligned}\sum_{j<t} \eta_j \mathcal{R}_c(w_j) &= \sum_{j<t} \eta_j L(A_c w_j)/n \\ &\geq \sum_{j<t} \eta_j |\nabla L(A_c w_j)|_1/n \\ &= \ell'_{<t},\end{aligned}$$

and

$$\begin{aligned}\sum_{j<t} \eta_j \left\langle \nabla \mathcal{R}_c(w_j), w_j - \Pi_\perp w_j \right\rangle &\leq \sum_{j<t} \eta_j \left|\nabla \mathcal{R}_c(w_j)\right| \left|w_j - \Pi_\perp w_j\right| \\ &\leq \sum_{j<t} \eta_j \frac{|\nabla L(A_c w_j)|_1}{n} \left| A_c^\top \frac{\nabla L(A_c w_j)}{|\nabla L(A_c w_j)|_1} \right| R \\ &\leq R\ell'_{<t}.\end{aligned}$$

Combining these terms with Lemma 10,

$$\begin{aligned}2\ell'_{<t} + \left(\gamma \ell'_{<t} - {\ln(t)}/{\gamma}\right)^2 &\leq \sum_{j<t} 2\eta_j \mathcal{R}_c(w_j) + |\Pi_\perp w_t - u|^2 \\ &\leq |u|^2 + \sum_{j<t} 2\eta_j \left\langle \nabla \mathcal{R}_c(w_j), w_j - \Pi_\perp w_j \right\rangle + \sum_{j<t} 2\eta_j \mathcal{R}_c(u) + 2 \\ &\leq \frac{\ln(t)^2}{\gamma^2} + 2R\ell'_{<t} + \frac{2}{t}\sum_{j<t} \eta_j + 2.\end{aligned}$$

Equivalently,

$$2\ell'_{<t} + \gamma^2(\ell'_{<t})^2 \leq 2\ell'_{<t}\ln(t) + 2R\ell'_{<t} + \frac{2}{t}\sum_{j<t}\eta_j + 2,$$

which implies

$$\ell'_{<t} \leq \max\left\{ \frac{4\ln(t)}{\gamma^2}, \frac{4R}{\gamma^2}, \frac{2}{t}\sum_{j<t}\eta_j, 2 \right\},$$

since otherwise

$$\begin{aligned}2\ell'_{<t}\ln(t) + 2R\ell'_{<t} + \frac{2}{t}\sum_{j<t}\eta_j + 2 &< \frac{\gamma^2(\ell'_{<t})^2}{2} + \frac{\gamma^2(\ell'_{<t})^2}{2} + \ell'_{<t} + \ell'_{<t} \\ &\leq 2\ell'_{<t}\ln(t) + 2R\ell'_{<t} + \frac{2}{t}\sum_{j<t}\eta_j + 2,\end{aligned}$$

a contradiction. ■

Proof (of lower bound in Lemma 9) First note

$$\begin{aligned} n_c \exp(-|\Pi_\perp w_t|) &\le n_c \ell(-|\Pi_\perp w_t|)/\ln 2 \\ &\le L(A_c \Pi_\perp w_t)/\ln 2 \\ &= L\left(A_c w_t - A_c(w_t - \Pi_\perp w_t)\right)/\ln 2 \\ &= L\left(A_c w_t - A_c \Pi_S w_t\right)/\ln 2. \\ &\le L\left(A_c w_t + R\right)/\ln 2. \end{aligned}$$

If $\ell = \ell_{\exp}$, then $L(A_c w_t + R) = e^R L(A_c w_t)$. Otherwise, by Bernoulli's inequality,

$$\sum_i \ell_{\log}((A_c w_t)_i + R) = \sum_i \ln\left(1 + e^R \exp((A_c w_t)_i)\right) \le \sum_i e^R \ln\left(1 + \exp((A_c w_t)_i)\right).$$

Combining these steps, and invoking Theorem 2,

$$\begin{aligned} n_c \ln(2) \exp(-|\Pi_\perp w_t|) &\le \exp(R) L(A_c w_t) \\ &= \exp(R)\left(L(A w_t) - L(A_S w_t)\right) \le \exp(R)\left(L(A w_t) - \inf_w L(Aw)\right). \end{aligned}$$

By Theorem 3,

$$\begin{aligned} \ln\left(\mathcal{R}(w_t) - \inf_w \mathcal{R}(w)\right) &\le \ln\ \max\left\{\frac{2\exp(|\bar v|)}{t},\ \frac{|\bar v|^2 + \ln(t)^2/\gamma^2}{\sum_{j=0}^{t-1}\eta_j}\right\} \\ &= \max\left\{\ln(2) + |\bar v| - \ln(t),\ \ln\left(|\bar v|^2 + \ln(t)^2/\gamma^2\right) - \ln\left(\sum_{j=0}^{t-1}\eta_j\right)\right\}. \end{aligned}$$

Together,

$$\begin{aligned} |\Pi_\perp w_t| &\ge -\ln\left(\mathcal{R}(Aw_t) - \inf_w \mathcal{R}(Aw)\right) - R + \ln\ln(2) - \ln(n/n_c) \\ &\ge \min\left\{-\ln(2) - |\bar v| + \ln(t),\ -\ln\left(|\bar v|^2 + \ln(t)^2/\gamma^2\right) + \ln\left(\sum_{j=0}^{t-1}\eta_j\right)\right\} - R + \ln\ln(2) - \ln(n/n_c). \end{aligned}$$

■

C.2. Parameter convergence when separable

First, the technical inequality on $\ell_{\log}$ and $\ell_{\exp}$.

Proof (of Lemma 12) The claims are immediate for $\ell = \ell_{\exp}$, thus consider $\ell = \ell_{\log}$. First note that $r \mapsto (e^r - 1)/r$ is increasing and not smaller than 1 when $r \ge 0$. Now set $r := \ell_{\log}(z)$, whereby $\ell'_{\log}(z) = e^z/(1+e^z) = (e^r - 1)/e^r$. Suppose $r \le \epsilon$; since $\exp(\cdot)$ lies above its tangents, then $1 - \epsilon \le 1 - r \le e^{-r}$, and

$$\frac{\ell'_{\log}(z)}{\ell_{\log}(z)} = \frac{e^r - 1}{re^r} \ge \frac{1}{e^r} \ge 1 - \epsilon.$$

For $\ell_{\exp}(z) \le 2\ell_{\log}(z)$, note

$$\frac{\ell_{\exp}(z)}{\ell_{\log}(z)} = \frac{e^r - 1}{r}$$

is increasing for $r = \ell_{\log}(z) > 0$, and $e - 1 < 2$. ■

Leveraging this inequality, the following proof handles Theorem 7 in the general separable case (not just $\ell_{\exp}$, as in the main text).

Proof (of Theorem 7 when $A = A_c = A_\perp$ (separable case)) Let $0 < \epsilon \le 1$ be arbitrary, and select t_0 so that $\mathcal{R}(w_{t_0}) \le \epsilon/n$. By Lemma 6, since $\eta_j \le 1$, the loss decreases at each step, and thus for any $t \ge t_0$, $\mathcal{R}(w_t) \le \epsilon/n$. Now let $t \ge t_0$ and $\mathcal{R}(w) \le \mathcal{R}(w_t)$ with arbitrary t and w. By Lemma 13 and Lemma 6,

$$\begin{aligned}
\frac{1}{2}\left|\frac{w}{|w|} - \bar{u}\right|^2 &\le 1 + \frac{\ln \mathcal{R}(w_t)}{|w|\gamma} + \frac{g^*(\bar{q})}{|w|\gamma} + \frac{\ln 2}{|w|\gamma} \\
&\le 1 + \frac{\ln \mathcal{R}(w_{t_0})}{|w|\gamma} - \frac{\sum_{j=t_0}^{t-1} \hat{\eta}_j(1 - \hat{\eta}_j/2)\gamma_j^2}{|w|\gamma} + \frac{g^*(\bar{q})}{|w|\gamma} + \frac{\ln 2}{|w|\gamma} \\
&\le 1 + \frac{\ln(\epsilon/n)}{|w|\gamma} - \frac{\sum_{j=t_0}^{t-1} \hat{\eta}_j\gamma_j^2}{|w|\gamma} + \frac{\sum_{j=t_0}^{t-1} \hat{\eta}_j^2\gamma_j^2/2}{|w|\gamma} + \frac{\ln n}{|w|\gamma} + \frac{\ln 2}{|w|\gamma} \\
&\le 1 - \frac{\sum_{j=t_0}^{t-1} \hat{\eta}_j\gamma_j^2}{|w|\gamma} + \frac{1 + \ln 2}{|w|\gamma},
\end{aligned} \tag{3}$$

where (similarly to the proof for $\ell_{\exp}$ in the main text) the last inequality uses the following smoothness consequence (cf. Lemma 6 with $\eta_j \le 1$)

$$\sum_{j=t_0}^{t-1} \hat{\eta}_j^2\gamma_j^2 = \sum_{j=t_0}^{t-1} \eta_j^2 |\nabla\mathcal{R}(w_j)|^2 \le 2\sum_{j=t_0}^{t-1}(\mathcal{R}(w_j) - \mathcal{R}(w_{j+1})) \le 2.$$

For $j \ge t_0$, by Lemma 12 and $\gamma = \min\left\{|A_\perp^\top q| : q \ge 0, \sum_i q = 1\right\}$ (cf. Theorem 2),

$$\gamma_j = |\nabla(\ln\mathcal{R})(w_j)| = \frac{|A^\top \nabla L(w_j)|}{L(Aw_j)} = \frac{|A^\top \nabla L(Aw_j)|}{|\nabla L(Aw_j)|_1}\frac{|\nabla L(Aw_j)|_1}{L(Aw_j)} \ge (1 - \epsilon)\gamma, \tag{4}$$

Invoking eq. (3) and eq. (4) with $w = w_t$ or $\bar{w}_t$ (notice that in both cases $|w| = |w_t|$) gives

$$\begin{aligned}
\frac{1}{2}\left|\frac{w}{|w_t|} - \bar{u}\right|^2 &\le 1 - \frac{\sum_{j=t_0}^{t-1} \hat{\eta}_j\gamma_j^2}{|w_t|\gamma} + \frac{1 + \ln 2}{|w_t|\gamma} \\
&\le 1 - \frac{\sum_{j=t_0}^{t-1} \hat{\eta}_j\gamma_j}{|w_t|}(1 - \epsilon) + \frac{1 + \ln 2}{|w_t|\gamma} \\
&= 1 - \frac{|w_{t_0}| + \sum_{j=t_0}^{t-1} \hat{\eta}_j\gamma_j}{|w_t|}(1 - \epsilon) + \frac{|w_{t_0}|}{|w_t|}(1 - \epsilon) + \frac{1 + \ln 2}{|w_t|\gamma} \\
&\le \epsilon + \frac{|w_{t_0}|}{|w_t|} + \frac{1 + \ln 2}{|w_t|\gamma}.
\end{aligned} \tag{5}$$

Next, ϵ and t_0 are tuned as follows. First, set $\epsilon := \min\{4/|w_t|, 1\}$; this is possible if the corresponding $t_0 \le t$, or equivalently $\mathcal{R}(w_t) \le 4/n|w_t|$. This in turn is true as long as $t \ge 5$ and

$$\frac{t}{(\ln t)^3} \ge \frac{n}{\gamma^4},$$

because then $(\ln t)^2 \ge 2$, and since $\gamma \le 1$, $|w_t| \le 4 \ln t/\gamma^2$ given by Lemma 9, Theorem 3 gives

$$\mathcal{R}(w_t) \le \frac{1}{t} + \frac{(\ln t)^2}{2\gamma^2 t} \le \frac{(\ln t)^2}{2\gamma^2 t} + \frac{(\ln t)^2}{2\gamma^2 t} = \frac{(\ln t)^2}{\gamma^2 t} \le \frac{\gamma^2}{n \ln t} \le \frac{4}{n|w_t|}.$$

By Theorem 3, $\mathcal{R}(w_{t_0}) \le \epsilon/n$ if

$$\frac{1}{t_0} + \frac{(\ln t_0)^2}{\gamma^2 t_0} \le \frac{\epsilon}{n}.$$

By Lemma 9, $|w_{t_0}| \le \mathcal{O}\left(\ln(n/\epsilon)\right)/\gamma^2$. Together with eq. (5),

$$\max\left\{\left|\frac{w_t}{|w_t|} - \bar{u}\right|^2, \left|\frac{\bar{w}_t}{|\bar{w}_t|} - \bar{u}\right|^2\right\} \le \frac{\mathcal{O}\left(\ln n + \ln\left(|w_t|\right)\right)}{|w_t|\gamma^2},$$

where the constants hidden in the $\mathcal{O}$ do not depend on the problem. Combining this with the lower and upper bounds on $|w_t|$ in Lemma 9,

$$\max\left\{\left|\frac{w_t}{|w_t|} - \bar{u}\right|^2, \left|\frac{\bar{w}_t}{|\bar{w}_t|} - \bar{u}\right|^2\right\} \le \mathcal{O}\left(\frac{\ln n + \ln \ln t}{\gamma^2 \ln t}\right).$$

■

C.3. Parameter convergence in general

For convenience in these proofs, define $v_t := \Pi_S w_t$.

The general application of Fenchel-Young is as follows.

Proof (of Lemma 14) First note that

$$A_\perp w = A_c \Pi_\perp w = A_c w + A_c(\Pi_\perp w - w) = A_c w - A_c \Pi_S w,$$

thus

$$\begin{aligned} \langle \bar{q}, A_\perp w\rangle &= \langle \bar{q}, A_c w\rangle - \langle \bar{q}, A_c \Pi_S(w)\rangle \le \langle \bar{q}, A_c w\rangle + |\bar{q}|_1 |A_c \Pi_S(w)|_\infty \\ &= \langle \bar{q}, A_c w\rangle + \max_i (A_c)_{i:} \Pi_S(w) \le \langle \bar{q}, A_c w_t\rangle + |\Pi_S(w)|. \end{aligned}$$

Thus

$$\begin{aligned}\frac{\langle \bar{u}, w\rangle}{|\bar{u}|\cdot|w|} &= \frac{-\left\langle A_\perp^\top \bar{q}, w\right\rangle}{\gamma|w|} = \frac{-\left\langle \bar{q}, A_\perp w\right\rangle}{\gamma|w|} \\ &= \frac{-\left\langle \bar{q}, A_c w\right\rangle}{\gamma|w|} - \frac{|\Pi_S(w)|}{\gamma|w|} \\ &\geq \frac{-\ln \mathcal{R}_{c,\exp}(w) - g^*(\bar{q})}{\gamma|w|} - \frac{|\Pi_S(w)|}{\gamma|w|} \\ &\geq \frac{-\ln \mathcal{R}_c(w_t) - \ln 2 - g^*(\bar{q})}{\gamma|w|} - \frac{|\Pi_S(w)|}{\gamma|w|} \\ &\geq \frac{-\ln\left(\mathcal{R}(w_t) - \bar{\mathcal{R}}\right) - \ln 2 - g^*(\bar{q})}{\gamma|w|} - \frac{|\Pi_S(w)|}{\gamma|w|}.\end{aligned}$$

■

Next, the adjustment of Lemma 6 to upper bounding $\ln(\mathcal{R}(w_t) - \bar{\mathcal{R}})$, which leads to an upper bound with $\gamma\gamma_i$ rather than γ_i^2, and the necessary cancellation.

Proof (of Lemma 15) The first inequality implies the second via the same direct induction in Lemma 6, so consider the first inequality.

Making use of Lemmas 18 and 19 and proceeding as in Lemma 19,

$$\begin{aligned}\mathcal{R}(w_{j+1}) - \bar{\mathcal{R}} &\leq \mathcal{R}(w_j) - \bar{\mathcal{R}} - \eta_j|\nabla\mathcal{R}(w_j)|^2 + \frac{\eta_j^2\mathcal{R}(w_j)}{2}|\nabla\mathcal{R}(w_j)|^2 \\ &\leq \left(\mathcal{R}(w_j) - \bar{\mathcal{R}}\right)\left(1 - \frac{\eta_j|\nabla\mathcal{R}(w_j)|^2}{\mathcal{R}(w_j) - \bar{\mathcal{R}}}\left(1 - \eta_j\mathcal{R}(w_j)/2\right)\right) \\ &= \left(\mathcal{R}(w_j) - \bar{\mathcal{R}}\right)\left(1 - \frac{|\nabla\mathcal{R}(w_j)|}{\mathcal{R}(w_j) - \bar{\mathcal{R}}}\cdot\frac{\eta_j\mathcal{R}(w_j)|\nabla\mathcal{R}(w_j)|}{\mathcal{R}(w_j)}\left(1 - \eta_j\mathcal{R}(w_j)/2\right)\right) \\ &\leq \left(\mathcal{R}(w_j) - \bar{\mathcal{R}}\right)\left(1 - \frac{|\nabla\mathcal{R}(w_j)|}{\mathcal{R}(w_j) - \bar{\mathcal{R}}}\cdot\hat{\eta}_j\gamma_j\left(1 - \hat{\eta}/2\right)\right). \qquad (6)\end{aligned}$$

Next it will be shown, by analyzing two cases, that

$$\frac{|\nabla\mathcal{R}(w_j)|}{\mathcal{R}(w_j) - \bar{\mathcal{R}}} \geq r\gamma. \qquad (7)$$

In the following, for notational simplicity let w denote w_j.

- Suppose $\mathcal{R}_c(w) < r\left(\mathcal{R}(w) - \bar{\mathcal{R}}\right)$. Consequently,

$$\mathcal{R}_S(w) - \bar{\mathcal{R}} > (1-r)\left(\mathcal{R}(w) - \bar{\mathcal{R}}\right).$$

Then, since $4\left(\mathcal{R}(w) - \bar{\mathcal{R}}\right) \leq 2\lambda(1-r)$,

$$\begin{aligned}
\frac{|\nabla\mathcal{R}(w)|}{\mathcal{R}(w) - \bar{\mathcal{R}}} &\geq \frac{-|\nabla\mathcal{R}_c(w)| + |\nabla\mathcal{R}_S(w)|}{\mathcal{R}(w) - \bar{\mathcal{R}}} \\
&\geq \frac{-\mathcal{R}_c(w) + \sqrt{2\lambda(\mathcal{R}_S(w) - \bar{\mathcal{R}})}}{\mathcal{R}(w) - \bar{\mathcal{R}}} \\
&> \frac{-r(\mathcal{R}(w) - \bar{\mathcal{R}}) + \sqrt{2\lambda(1-r)(\mathcal{R}(w) - \bar{\mathcal{R}})}}{\mathcal{R}(w) - \bar{\mathcal{R}}} \\
&\geq \frac{(2-r)(\mathcal{R}(w) - \bar{\mathcal{R}})}{\mathcal{R}(w) - \bar{\mathcal{R}}} \\
&\geq 1 \geq r\gamma.
\end{aligned}$$

- Otherwise, suppose $\mathcal{R}_c(w) \geq r\left(\mathcal{R}(w) - \bar{\mathcal{R}}\right)$. Using an expression inspired by a general analysis of AdaBoost (Mukherjee et al., 2011, Lemma 16 of journal version), and introducing $(1-\epsilon)$ by invoking Lemma 12 as in the separable case,

  $$|\nabla\mathcal{R}(w)| \geq \langle -\bar{u}, \nabla\mathcal{R}(w)\rangle = \langle -A\bar{u}, \nabla L(Aw)/n\rangle = \langle -A_c\bar{u}, \nabla L(A_c w)/n\rangle \geq \gamma(1-\epsilon)\mathcal{R}_c(w),$$

  Thus

  $$\frac{|\nabla\mathcal{R}(w)|}{\mathcal{R}(w) - \bar{\mathcal{R}}} \geq \frac{\gamma(1-\epsilon)\mathcal{R}_c(w)}{\mathcal{R}(w) - \bar{\mathcal{R}}} \geq \frac{\gamma(1-\epsilon)\left(r\left(\mathcal{R}(w) - \bar{\mathcal{R}}\right)\right)}{\mathcal{R}(w) - \bar{\mathcal{R}}} = r\gamma(1-\epsilon).$$

Combining eq. (6) with eq. (7),

$$\mathcal{R}(w_{j+1}) - \bar{\mathcal{R}} \leq \left(\mathcal{R}(w_j) - \bar{\mathcal{R}}\right)\left(1 - r(1-\epsilon)\gamma\gamma_j\hat{\eta}_j\left(1 - \hat{\eta}_j/2\right)\right).$$

■

Next, the proof of the intermediate inequality by combining Lemma 15 and Lemma 14.

Proof (of Lemma 16) By Lemma 6, since $\eta_j \leq 1$, the loss decreases at each step, and thus for any $t \geq t_0$, $\mathcal{R}(w_t) \leq \epsilon/n$. Combining Lemma 14 and Lemma 15,

$$\begin{aligned}
\frac{\langle\bar{u}, w\rangle}{|\bar{u}|\cdot|w|} &\geq \frac{-\ln\left(\mathcal{R}(w) - \bar{\mathcal{R}}\right)}{\gamma|w|} - \frac{\ln 2 + g^*(\bar{q}) + |\Pi_S(w)|}{\gamma|w|}. \\
&\geq \frac{r(1-\epsilon)\gamma\sum_{j=t_0}^{t-1}\hat{\eta}_j(1-\hat{\eta}_j/2)\gamma_j - \ln\left(\mathcal{R}(w_{t_0}) - \bar{\mathcal{R}}\right)}{\gamma|w|} - \frac{\ln 2 + g^*(\bar{q}) + |\Pi_S(w)|}{\gamma|w|} \\
&\geq \frac{r(1-\epsilon)\sum_{j=t_0}^{t-1}\hat{\eta}_j(1-\hat{\eta}_j/2)\gamma_j}{|w|} - \frac{\ln(\epsilon/n)}{\gamma|w|} - \frac{\ln 2 + \ln n + |\Pi_S(w)|}{\gamma|w|} \\
&\geq \frac{r(1-\epsilon)\sum_{j=t_0}^{t-1}\hat{\eta}_j(1-\hat{\eta}_j/2)\gamma_j}{|w|} - \frac{\ln 2}{\gamma|w|} - \frac{|\Pi_S(w)|}{\gamma|w|}.
\end{aligned}$$

■

The pieces are in place to prove parameter convergence in general.

Proof (of general case in Theorem 7) The guarantee on $\bar{v}_t$ and the $A_c = \emptyset$ case have been discussed in Lemma 8, therefore assume $A_c \neq \emptyset$. The proof will proceed via invocation of the Fenchel-Young scheme in Lemma 16, applied to $w \in \{w_t, \bar{w}_t\}$ since $\mathcal{R}(\bar{w}_t) \le \mathcal{R}(w_t)$.

It is necessary to first control the warm start parameter t_0. Fix an arbitrary $\epsilon \in (0,1)$, set $r := 1 - \epsilon/3$, and let t_0 be large enough such that

$$\mathcal{R}(w_{t_0}) - \bar{\mathcal{R}} \le \frac{\epsilon}{3n}, \qquad \mathcal{R}(w_{t_0}) - \bar{\mathcal{R}} \le \frac{\lambda(1-r)}{2} = \frac{\lambda\epsilon}{6}, \qquad 1 - \frac{\hat{\eta}_{t_0}}{2} \ge 1 - \frac{\eta_{t_0}}{2} \ge 1 - \frac{\epsilon}{3}. \quad (8)$$

By Theorem 3 and the choice of step sizes, it is enough to require

$$\frac{\exp(|\bar{v}|)}{t_0} \le \min\left\{\frac{\epsilon}{6n}, \frac{\lambda\epsilon}{12}\right\}, \qquad \frac{|\bar{v}|^2 + \ln(t_0)^2/\gamma^2}{2\gamma^2\sqrt{t_0}} \le \min\left\{\frac{\epsilon}{6n}, \frac{\lambda\epsilon}{12}\right\}, \qquad \frac{1}{2\sqrt{t_0+1}} \le \frac{\epsilon}{3}.$$

Therefore, choosing $t_0 = \widetilde{\mathcal{O}}\left(\frac{n^2}{\epsilon^2}\right)$ suffices.

Invoking Lemma 16 with the above choice for $w \in \{w_t, \bar{w}_t\}$,

$$\begin{aligned}
\frac{1}{2}\left|\frac{w}{|w_t|} - \bar{u}\right|^2 &= 1 - \frac{\langle \bar{u}, w\rangle}{|\bar{u}| \cdot |w_t|} \\
&\le 1 - \frac{r(1-\epsilon/3)\sum_{j=t_0}^{t-1}\hat{\eta}_j(1-\hat{\eta}_j/2)\gamma_j}{|w_t|} + \frac{\ln 2}{\gamma|w_t|} + \frac{|\Pi_S(w)|}{\gamma|w_t|} \\
&\le 1 - \frac{(1-\epsilon/3)(1-\epsilon/3)\sum_{j=t_0}^{t-1}\hat{\eta}_j(1-\epsilon/3)\gamma_j}{|w_t|} + \frac{\ln 2}{\gamma|w_t|} + \frac{|\Pi_S(w)|}{\gamma|w_T|} \\
&\le 1 - \frac{(1-\epsilon)\sum_{j=t_0}^{t-1}\hat{\eta}_j\gamma_j}{|w_t|} + \frac{\ln 2}{\gamma|w_t|} + \frac{|\Pi_S(w)|}{\gamma|w_t|} \\
&= 1 - \frac{(1-\epsilon)\left(|w_{t_0}| + \sum_{j=t_0}^{t-1}\hat{\eta}_j\gamma_j\right)}{|w_t|} + (1-\epsilon)\frac{|w_{t_0}|}{|w_t|} + \frac{\ln 2}{\gamma|w_t|} + \frac{|\Pi_S(w)|}{\gamma|w_t|} \\
&\le \epsilon + \frac{|w_{t_0}|}{|w_t|} + \frac{\ln 2}{\gamma|w_t|} + \frac{|\Pi_S(w)|}{\gamma|w_t|}.
\end{aligned} \quad (9)$$

Suppose $t \ge 5$ and $\sqrt{t}/\ln^3 t \ge n(1+R)/\gamma^4$, where $R = \sup_{j<t}|\Pi_\perp w_j - w_j| = \mathcal{O}(1)$ is introduced in Lemma 9. As will be shown momentarily, t satisfies (8) with $\epsilon \le C/|w_t|$ for some constant C, and therefore this choice of ϵ can be plugged into (9). To see this, note that Theorem 3 gives

$$\begin{aligned}
\mathcal{R}(w_t) - \bar{\mathcal{R}} &\le \frac{\exp(|\bar{v}|)}{t} + \frac{|\bar{v}|^2 + \ln(t)^2}{2\gamma^2\sqrt{t}} \\
&\le \frac{\exp(|\bar{v}|)}{\sqrt{t}/\ln(t)^2} + \frac{|\bar{v}|^2}{2\gamma^2\sqrt{t}/\ln(t)^2} + \frac{\ln(t)^2}{2\gamma^2\sqrt{t}} \\
&\le \frac{\exp(|\bar{v}|)\gamma^4}{n(1+R)\ln t} + \frac{|\bar{v}|^2\gamma^2}{2n(1+R)\ln t} + \frac{\gamma^2}{2n(1+R)\ln t} \\
&\le C_1\frac{\gamma^2}{n \cdot 4(1+R)\ln t} \\
&\le C_1\frac{1}{n|w_t|},
\end{aligned}$$

where the last line uses Lemma 9. It can be shown similarly that other parts of (8) hold.

Continuing with Equation (9) but using $\epsilon \leq {}^{C}\!/\!_{|w_t|}$ and upper bounding $|w_{t_0}|$ via Lemma 9,

$$\left|\frac{w_t}{|w_t|} - \bar{u}\right|^2 \leq \frac{\mathcal{O}\left(\ln n + \ln\left(1/|w_t|\right)\right)}{|w_t|\gamma^2}.$$

Lastly, controlling the denominator with the lower bound on $|w_t|$ in Lemma 9,

$$\left|\frac{w_t}{|w_t|} - \bar{u}\right|^2 \leq \mathcal{O}\left(\frac{\ln n + \ln\ln t}{\gamma^2 \ln t}\right).$$

■