Journal of Machine Learning Research 21 (2020) 1-19 Submitted 11/18; Revised 2/20; Published 11/20

Theory of Curriculum Learning, with Convex Loss Functions

Daphna Weinshall DAPHNA@MAIL.HUJI.AC.IL
Dan Amir DAN.AMIR@MAIL.HUJI.AC.IL
School of Computer Science and Engineering
Hebrew University of Jerusalem
Jerusalem 91904, Israel

Editor: Koby Crammer

Abstract

Curriculum Learning is motivated by human cognition, where teaching often involves gradually exposing the learner to examples in a meaningful order, from easy to hard. Although methods based on this concept have been empirically shown to improve performance of several machine learning algorithms, no theoretical analysis has been provided even for simple cases. To address this shortfall, we start by formulating an ideal definition of difficulty score - the loss of the optimal hypothesis at a given datapoint. We analyze the possible contribution of curriculum learning based on this score in two convex problems - linear regression, and binary classification by hinge loss minimization. We show that in both cases, the convergence rate of SGD optimization decreases monotonically with the difficulty score, in accordance with earlier empirical results. We also prove that when the difficulty score is fixed, the convergence rate of SGD optimization is monotonically increasing with respect to the loss of the current hypothesis at each point. We discuss how these results settle some confusion in the literature where two apparently opposing heuristics are reported to improve performance: curriculum learning in which easier points are given priority, vs hard data mining where the more difficult points are sought out.

Keywords: curriculum learning, linear regression, hinge loss minimization

1. Introduction

Many popular machine learning algorithms involve sampling of examples from a large labeled data set and gradually improving the model performance on those examples. In particular, any algorithm which employs Stochastic Gradient Descent (SGD) falls under this category. In the standard and most common form of SGD, examples are drawn uniformly from the training data. This approach is well justified theoretically as it guarantees that the expected value of the gradient in each step is equal to the gradient of the empirical loss.

Although this approach is both simple and theoretically sound, it differs dramatically from our intuition of how living organisms learn from examples. Both humans and animals usually benefit from being exposed to examples in a meaningful order as defined by some curriculum. The efficacy of learning new concepts is usually improved, or even only made possible, when the learner is exposed to gradually more difficult examples or more complex concepts. The learner usually uses the easier examples to acquire capabilities which facilitate the grasping of the more complex examples. This concept is well grounded in cognitive research, where it has been investigated within both a behavioral approach (e.g. Skinner, 1990) and a computational approach (e.g. Elman, 1993).

The idea of incorporating the concept of curriculum learning into the framework of supervised machine learning has been introduced early on (e.g. Sanger, 1994), while being identified as a key challenge for machine learning throughout (Mitchell, 1980, 2006; Wang and Cottrell, 2015). Several formulations have been suggested both in the context of SGD (Bengio et al., 2009) and in the context of other iterative optimization algorithms (Kumar et al., 2010). Most empirical studies, involving non-convex problems for the most part, demonstrated beneficial effects of curriculum learning, including faster convergence rate and better final performance. Even so, this approach has not been widely adopted by practitioners (but see Schroff et al., 2015; Oh et al., 2015). Moreover, this idea has not been theoretically analyzed, and no guarantees have ever been obtained for its success even on simple learning problems.

One inherent limitation of current curriculum learning approaches is the absence of a formal definition of the difficulty score of datapoints. In their empirical research, Bengio et al. (2009) relied on a manually crafted, domain-specific curriculum. This approach fails when the manual definition of easier sub-tasks or subsets of examples is impossible to acquire, especially with large scale and complex data. Moreover, even when it is possible to manually design a curriculum, the scoring of difficulty based on human intuition may not match the difficulty of the example or sub-problem for a learning algorithm.

The framework of Self Paced learning (SPL) (Kumar et al., 2010) overcomes this limitation by focusing on the intrinsic information of the learner, namely, the loss with respect to the learner's current hypothesis, in order to avoid the need to obtain a curriculum from an extrinsic source. In this approach, a new optimization problem is introduced where the training loss is minimized jointly with a regularizing term, which attaches greater significance to points that better fit the current learner's hypothesis (namely, incur lower loss). While SPL obviates the need for a predefined curriculum, new difficulties are introduced as the new optimization problem is more difficult to solve. Moreover, by relying only on the learner's training loss, the optimization is more susceptible to problems such as over-fitting and training instability. Finally, the SPL heuristics seems to contradict other commonly used heuristics, which attach greater significance to points that *do not* fit well with the current learner's hypothesis (namely, incur *higher* loss). Examples include hard data mining (Shrivastava et al., 2016) and boosting (Schapire et al., 1998).

In this paper, we address those challenges from a theoretical point of view. We first define a measure for the convergence rate of SGD optimization, and offer a formal definition of a point's difficulty score - the loss at the optimal hypothesis with respect to the example. We then analyze how these two concepts are related, and specifically, how the convergence rate of SGD optimization changes with the difficulty score. This is done in the context of two convex optimization problems - linear regression and classification with hinge loss minimization. Our analysis shows that under some reasonable assumptions, the convergence rate is expected to decrease monotonically with the difficulty of the sampled examples. This analysis is consistent with empirical results as discussed above.

Another challenge involves the success of apparently contradictory methods, which are based on the idea that the more difficult examples should be given higher weight (Shrivastava et al., 2016; Schapire et al., 1998). We hypothesize that this apparent contradiction can be explained in part by some confusion in the literature with respect to how difficulty is measured. More specifically, we formally differentiate between the *global* difficulty score as defined above, and the *local* difficulty score as defined by the loss at a point with respect to the current hypothesis. In agreement with the intuition underlying both approaches, we claim that ideally a learner should follow a curriculum based on extrinsic (global) difficulty, while not "wasting time" on examples that are easy for the current (local) learning hypothesis. In accordance, we formally show that when examples are drawn conditioned on some fixed global difficulty score, the convergence rate of SGD optimization in linear regression and hinge loss minimization is *monotonically increasing with the local difficulty* of the example.

The rest of the paper is organized as follows: We start by introducing some notations and definitions in Section 2. In Sections 3 and 4 we develop the theory and prove the main results for the two convex problems of linear regression and hinge loss classification, respectively.

Related Work. Jiang et al. (2017) addressed the automatic generation of curriculum by developing a general framework for the joint training of two deep neural networks, where one network (the *MentorNet*) is trained to generate an adaptive curriculum for the other network. In their work, they show both empirically and theoretically that the data-driven generation of curriculum by MentorNet can improve the learner robustness to noisy data.

The apparent contradiction in empirical reports, showing the advantage of both curriculum learning and hard example mining, motivated Chang et al. (2017) to suggest the active bias method. This method circumvents the problem of "easy vs. hard" by focusing on certainty instead of difficulty. In their approach, the training schedule is designed according to the model's prediction variance over the previous training steps, where distribution is biased in favor of examples with high prediction variance.

Our approach differs from these two ideas in that it addresses the question of difficulty definition directly. In contrast, MentorNet and active bias can, in theory, learn to generate biases over the data distribution which do not necessarily reflect a difficulty based curriculum. Future work should examine whether any curriculum generated by these methods complies with the intuition derived from our theoretic results. Namely, a curriculum should rank the examples so that they are negatively correlated with some global difficulty score, and positively correlated with the local difficulty.

We note that in practice, there is no easy way to define a curriculum based on the concept of *global difficulty score*, since the optimal hypothesis is not known to the learner. Nevertheless, many practical scenarios that employ machine learning involve a sequence of iterations of model improvement. In such scenarios, results from earlier iterations can be used to generate a curriculum for subsequent iterations. Another scenario involves transfer learning from a strong learner to a weaker learner. Thus, it has been shown by Hacohen and Weinshall (2019) that curriculum based on the stronger model's difficulty scores can be used to train the weak model faster, and lead it to a better solution.

2. Notations and Definitions

Let $\mathbb{X} = \{[\mathbf{x}_i, y_i]\}_{i=1}^n$ denote the set of the training examples, where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the i-th data point and y_i its corresponding label. Examples are drawn from a distribution $\mathcal{D}$. Let $\mathcal{H}$ denote a set of hypotheses $\{h_\mathbf{w}\}$ defined by some parameter vector $\mathbf{w}$. Let $L(\mathbf{X}_i, h)$ denote the loss of hypothesis h at point $\mathbf{X}_i = [\mathbf{x}_i, y_i]$. In the risk Minimization framework, we seek a hypothesis $\bar{h}$ that minimizes the expected loss $L_{\mathcal{D}}(h)$

$$\begin{aligned} L_{\mathcal{D}}(h) &= \mathbb{E}_{\mathbf{X}_i \sim \mathcal{D}}(L(\mathbf{X}_i, h)) \\ \bar{\mathbf{w}} &= \operatorname*{argmin}_{\mathbf{w}} L_{\mathcal{D}}(h_{\mathbf{w}}), \qquad \bar{h} = h_{\bar{\mathbf{w}}} \end{aligned} \tag{1}$$

2.1 Convergence rate of SGD optimization

When $L_{\mathcal{D}}(h)$ is convex, $\bar{h}$ can be found using Gradient Descent or Stochastic Gradient Descent. In pure SGD optimization, at time $t \in [T]$ a single example $\mathbf{X}_t = [\mathbf{x}_t, y_t]$ is drawn from distribution $\mathcal{D}$ and used to estimate the gradient step. In practice, optimization is often achieved using mini-batch stochastic gradient descent, where at each time t a set of examples is drawn and jointly used to estimate the gradient step.

Our analysis assumes pure SGD optimization as defined above. Given a sequence $\{\mathbf{X}_t\}_{t=1}^{T}$, this optimization method generates a sequence of estimators $\{\mathbf{w}_t\}_{t=1}^{T}$. Although in practice many variations on SGD are used, we analyze here the basic form in which the update rule is defined as follows

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \frac{\partial L(\mathbf{X}_t, \mathbf{w})}{\partial \mathbf{w}}\bigg|_{\mathbf{w}=\mathbf{w}_t} \tag{2}$$

where η is a hyper-parameter that controls the learning rate of the algorithm.

We can now define *the convergence rate of SGD* at time t:

Definition 1 (Convergence Rate) *The improvement achieved by SGD using point $\mathbf{X}_t$ is measured by the change in the distance between the current estimate of the optimal hypothesis and the optimal hypothesis $\|\mathbf{w}_t - \bar{\mathbf{w}}\|$. The* convergence rate *of SGD at time t is measured by the average change in this divergence measure over all points, namely*

$$\Delta = \mathbb{E}_{\mathbf{X}_t \sim \mathcal{D}}[\|\mathbf{w}_t - \bar{\mathbf{w}}\|^2 - \|\mathbf{w}_{t+1} - \bar{\mathbf{w}}\|^2] \tag{3}$$

2.2 Measuring the difficulty score

Definition 2 (Difficulty Score)

- *The* global Difficulty Score *of example* $\mathbf{X}$ *is defined as*

$$\Psi(\mathbf{X}) = g(L(\mathbf{X}, \bar{h})) \tag{4}$$

  where $g(\cdot)$ is some monotonic function.

- *The* local Difficulty Score *of example* $\mathbf{X}$ *is defined as*

$$\Upsilon(\mathbf{X}) = g(L(\mathbf{X}, h_{w_t})) \tag{5}$$

  where $\mathbf{w}_t$ is the current hypothesis, and $g(\cdot)$ is the same function as in (4).

For clarity, in the rest of this paper we will omit the index t when it is clear from context.

2.3 Outline of the main results

In general, SGD is only guaranteed to converge to a local minimum of the loss function. We therefore limit our analysis to simple convex problems, one continuous - linear regression, and one discrete - binary classification with hinge loss minimization.

In these two study cases, we analyze the convergence of SGD as defined above. First we define the conditional convergence rate, Δ from (3) conditioned on example difficulty Ψ or Υ. We then investigate the differential change in the conditional convergence rate as the difficulty score changes. We show two results: (1) This differential change is monotonically decreasing with the global difficulty score Ψ, namely, SGD converges faster when given easier examples. (2) When Ψ is fixed, this differential change is monotonically increasing with the local difficulty score Υ, namely, SGD converges faster when given examples that are more challenging for the current hypothesis $\mathbf{w}_t$.

3. Linear Regression

In linear regression, the learner's goal is to predict a real value $y = h(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^d$, where $h \in \mathcal{H}$ is a linear function of $\mathbf{x}$ and the loss is defined by the sum of least squares. Formally, using the notations above, the loss function can be written as

$$\begin{aligned} L(\mathbf{X}, \mathbf{w}) &= (\mathbf{a} \cdot \mathbf{x} + b - y)^2 \\ &\doteq (\mathbf{x} \cdot \mathbf{w} - y)^2 \end{aligned} \tag{6}$$

where $\mathbf{w} \doteq [\mathbf{a}, b]^t \in \mathbb{R}^{d+1}$ denotes the linear separator concatenated with the bias term. With some abuse of notation, $\mathbf{x}$ henceforth will denote the vector $[\mathbf{x}, 1]^t \in \mathbb{R}^{d+1}$. Let $\mathbf{s}$ denote the gradient vector at time t. We obtain from (2) and (6)

$$\begin{aligned} \mathbf{w}_{t+1} &= \mathbf{w}_t - 2\eta(\mathbf{x} \cdot \mathbf{w} - y)\mathbf{x} = \mathbf{w}_t + \mathbf{s} \\ \mathbf{s} &\doteq -2\eta(\mathbf{x} \cdot \mathbf{w} - y)\mathbf{x} \end{aligned} \tag{7}$$

3.1 Convergence rate decreases with *global difficulty*

The main theorem in this section states that the conditional convergence rate of SGD is monotonically *decreasing* with the *global Difficulty Score* of sample $\mathbf{X}_t$. We prove it below for the gradient vector as defined in (7). We note in passing that if the size of the gradient step is fixed at η, a somewhat stronger theorem can be obtained where the constraint on the step size being small is not required.

Recall that $\mathbf{x}, \mathbf{w} \in \mathbb{R}^{d+1}$. The analysis in carried out in the parameter space $\mathbf{w} \in \mathbb{R}^{d+1}$, where parameter vector $\mathbf{w}$ corresponds to a point, and data vector $\mathbf{x}$ describes a hyperplane. In this space, let $\Omega_{\mathbf{x}}$ denote the hyperplane on which the gradient $\mathbf{s}$ vanishes, i.e. $\mathbf{s} = 0$. It follows from (7) that this hyperplane is defined by $\mathbf{x} \cdot \mathbf{w} = y$, namely, $\mathbf{x}$ defines its normal direction. This implies that the gradient vector at time t is perpendicular to $\Omega_{\mathbf{x}}$ as illustrated in Fig. 1. Let $\bar{\mathbf{z}}$ denote the projection of $\bar{\mathbf{w}}$, the parameters of the optimal hypothesis, on $\Omega_{\mathbf{x}}$.

Due to the nature of the regression loss, which is based on the squared Euclidean distance, we use $g(x) = \sqrt{x}$ in (4), to obtain the difficulty score $\Psi(\mathbf{X}) = \sqrt{L(\mathbf{X}, \bar{\mathbf{w}})}$.

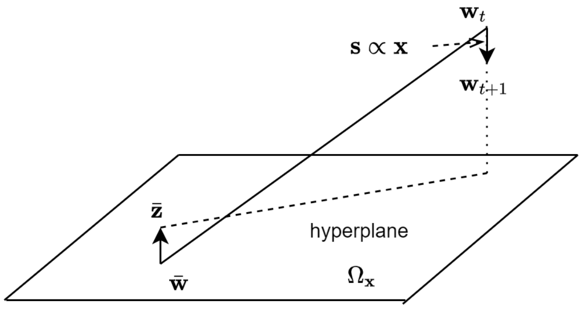

Figure 1: The geometry of the gradient step at time t, where $\mathbf{w}_t \rightarrow \mathbf{w}_{t+1}$.

Lemma 1 *The Difficulty Score (squared) of* $\mathbf{X}$ *is* $\Psi^2 = \|\mathbf{x}\|^2 \|\bar{\mathbf{w}} - \bar{\mathbf{z}}\|^2$.

Proof

$$\begin{aligned} \Psi(\mathbf{X})^2 = L(\mathbf{X}, \bar{\mathbf{w}}) &= L(\mathbf{X}, \bar{\mathbf{z}} + (\bar{\mathbf{w}} - \bar{\mathbf{z}})) \\ &= [\mathbf{x} \cdot \bar{\mathbf{z}} + \mathbf{x} \cdot (\bar{\mathbf{w}} - \bar{\mathbf{z}}) - y]^2 \\ &= [\mathbf{x} \cdot (\bar{\mathbf{w}} - \bar{\mathbf{z}})]^2 \\ &= \|\mathbf{x}\|^2 \|\bar{\mathbf{w}} - \bar{\mathbf{z}}\|^2 \end{aligned} \tag{8}$$

The transition in the third line follows from $\bar{\mathbf{z}} \in \Omega_{\mathbf{x}} \implies \mathbf{x} \cdot \bar{\mathbf{z}} - y = 0$. The last transition follows from the fact that both $\mathbf{x}$ and $(\bar{\mathbf{w}} - \bar{\mathbf{z}})$ are perpendicular to $\Omega_{\mathbf{x}}$, and are therefore parallel to each other. ■

Next, we embed the data points in the parameters space, representing each datapoint $\mathbf{x}$ using a hyperspherical coordinate system $[r, \vartheta, \Phi]$, with pole (origin) fixed at $\bar{\mathbf{w}}$ and polar axis (zenith direction) $\vec{\mathcal{O}} = \bar{\mathbf{w}} - \mathbf{w}_t$ (see Fig. 2). r denotes the vector's length, while $0 \leq \vartheta \leq \pi$ denotes the polar angle with respect to $\vec{\mathcal{O}}$. Let $\Phi = [\varphi_1, \ldots, \varphi_{d-1}]$ denote the remaining polar angles.

To illustrate, Fig. 2 shows a planar section of the parameter space - the $2D$ plane formed by the two intersecting lines $\vec{\mathcal{O}}$ and $\bar{\mathbf{z}} - \bar{\mathbf{w}}$. The gradient vector $\mathbf{s}$ points from $\mathbf{w}_t$ towards $\Omega_{\mathbf{x}}$. $\Omega_{\mathbf{x}}$ is perpendicular to $\mathbf{x}$, which is parallel to $\bar{\mathbf{z}} - \bar{\mathbf{w}}$ and to $\mathbf{s}$, and therefore $\Omega_{\mathbf{x}}$ is projected onto a line in this plane. We introduce the notation $\lambda = \|\bar{\mathbf{w}} - \mathbf{w}_t\|$.

Let $\mathbf{s}_{\mathcal{O}}$ denote the projection of the gradient vector $\mathbf{s}$ on the polar axis $\vec{\mathcal{O}}$, and let $\mathbf{s}_{\perp}$ denote the perpendicular component. From (7) and the definition of Ψ

$$\begin{aligned} \mathbf{s} &= -2\eta\mathbf{x}(\mathbf{x} \cdot \mathbf{w}_t - y) \\ &= -2\eta\mathbf{x}[\mathbf{x} \cdot (\mathbf{w}_t - \bar{\mathbf{w}}) \pm \Psi] \end{aligned} \tag{9}$$

and

$$\begin{aligned} \mathbf{s}_{\mathcal{O}} &= \mathbf{s} \cdot \frac{\bar{\mathbf{w}} - \mathbf{w}_t}{\lambda} \\ &= 2\frac{\eta}{\lambda}[r^2\lambda^2 \cos^2 \vartheta \mp \Psi r \lambda \cos \vartheta] \end{aligned} \tag{10}$$

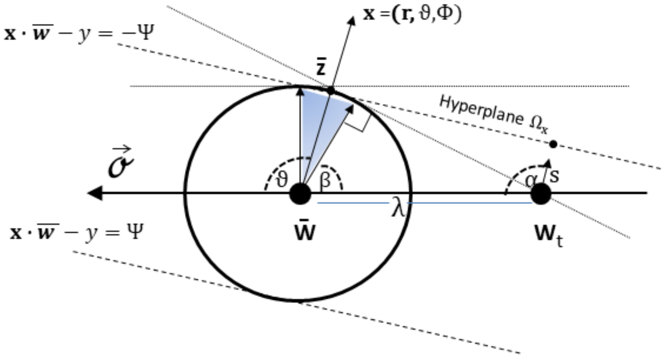
$\mathbf{x}\cdot\overline{\boldsymbol{w}} - y = -\Psi$
$\mathbf{x} = (\mathbf{r}, \vartheta, \Phi)$
$\bar{\mathbf{z}}$
Hyperplane Ω_x
$\vec{\mathcal{O}}$
ϑ
β
α
s
λ
$\mathbf{x}\cdot\overline{\boldsymbol{w}} - y = \Psi$
$\bar{\mathbf{w}}$
$\mathbf{w}_t$

Figure 2: The $2D$ planar section defined by the vectors $\vec{\mathcal{O}} = \bar{\mathbf{w}} - \mathbf{w}_t$ and $\bar{\mathbf{z}} - \bar{\mathbf{w}}$. The circle centered on $\bar{\mathbf{w}}$ has radius $\|\bar{\mathbf{w}} - \bar{\mathbf{z}}\| = \frac{\Psi}{\|\mathbf{x}\|}$ from Lemma 1. It traces the location of $\bar{\mathbf{z}}$ at points $\mathbf{x}$ for which $\frac{\Psi}{\|\mathbf{x}\|}$ is constant.

Assumption 1 (independence assumption) *Assume that the probability of label y depends only on the error $|y - \mathbf{x}^t \cdot \bar{\mathbf{w}}|$, and that the error probability is independent of $\mathbf{x}$.*

With this assumption[1] we can introduce the following notation for the density of $\mathcal{D}$

$$f_{\mathcal{D}}(\mathbf{X}) \doteq f(\mathbf{x}) g(|y - \mathbf{x}^t \cdot \bar{\mathbf{w}}|) \tag{11}$$

where $\int f(\mathbf{x}) d\mathbf{x} = 1$.

The following analysis requires the conditional distribution of the data given difficulty score Ψ. Note that when the difficulty score is fixed, the label y must take one of the following two values: $y_1 = \mathbf{x} \cdot \bar{\mathbf{w}} + \Psi$ or $y_2 = \mathbf{x} \cdot \bar{\mathbf{w}} - \Psi$. By Assumption 1 both labels are equally likely, where from (11) $f_{\mathcal{D}|\Psi}(\mathbf{X}) \propto 2f(\mathbf{x})g(\Psi)$. Let $\mathbf{x} = (r, \vartheta, \Phi)$. It follows that

$$f_{\mathcal{D}|\Psi}(\mathbf{X}) = f(\mathbf{x}) = f(r, \vartheta, \Phi) \tag{12}$$

Based on (3), let $\Delta(\Psi)$ denote the conditional convergence rate at $\mathbf{w}_t$ given Ψ:

$$\Delta(\Psi) = \mathbb{E}_{\mathbf{X}_t \sim \mathcal{D}|\Psi}[\|\mathbf{w}_t - \bar{\mathbf{w}}\|^2 - \|\mathbf{w}_{t+1} - \bar{\mathbf{w}}\|^2] \tag{13}$$

Lemma 2

$$\Delta(\Psi) = 2\lambda \mathbb{E}_{\mathbf{X}_t \sim \mathcal{D}|\Psi}[\mathbf{s}_{\mathcal{O}}] - \mathbb{E}_{\mathbf{X}_t \sim \mathcal{D}|\Psi}[\mathbf{s}^2] \tag{14}$$

Proof From (13), using Cartesian coordinates in the planar section shown in Fig. 2 where $s = [\mathbf{s}_{\mathcal{O}}, \mathbf{s}_{\perp}]$, $\mathbf{w}_t - \bar{\mathbf{w}} = [-\lambda, 0]$ and $\mathbf{w}_{t+1} - \bar{\mathbf{w}} = [-\lambda + \mathbf{s}_{\mathcal{O}}, \mathbf{s}_{\perp}]$, it follows that

$$\begin{aligned}
\Delta(\Psi) &= (-\lambda)^2 - \mathbb{E}_{\mathbf{X}_t \sim \mathcal{D}|\Psi}[(-\lambda + \mathbf{s}_{\mathcal{O}})^2 + \mathbf{s}_{\perp}^2] \\
&= \lambda^2 - (\lambda^2 - 2\lambda \mathbb{E}_{\mathbf{X}_t \sim \mathcal{D}|\Psi}[\mathbf{s}_{\mathcal{O}}] + \mathbb{E}_{\mathbf{X}_t \sim \mathcal{D}|\Psi}[\mathbf{s}_{\mathcal{O}}^2]) - \mathbb{E}_{\mathbf{X}_t \sim \mathcal{D}|\Psi}[\mathbf{s}_{\perp}^2] \\
&= 2\lambda \mathbb{E}_{\mathbf{X}_t \sim \mathcal{D}|\Psi}[\mathbf{s}_{\mathcal{O}}] - \mathbb{E}_{\mathbf{X}_t \sim \mathcal{D}|\Psi}[\mathbf{s}^2]
\end{aligned}$$

■

[1] In Appendix A we show that in a Bayesian framework, this assumption can be replaced by $\mathbb{E}_{f_v}[f_{\mathcal{D}}([\mathbf{x}, \mathbf{x} \cdot \bar{\mathbf{w}} + u])] = \mathbb{E}_{f_v}[f_{\mathcal{D}}([\mathbf{x}, \mathbf{x} \cdot \bar{\mathbf{w}} - u])]$ $\forall u$, where expectation is taken with respect to some prior distribution f_v over $\mathcal{D}$.

To simplify the notations, henceforth $\mathbb{E}$ stands for $\mathbb{E}_{\mathbf{X}_t \sim \mathcal{D}|\Psi}$. In addition, we define a shorthand notation $(\pm\Psi)$ to be used inside the expectation operator $\mathbb{E}[\cdot]$. It conveys that the operand of $\mathbb{E}[]$ should be multiplied by either $+\Psi$ or $-\Psi$, depending on whether the label y equals $\mathbf{x} \cdot \bar{\mathbf{w}} + \Psi$ or $\mathbf{x} \cdot \bar{\mathbf{w}} - \Psi$ respectively. When expectation is computed, each case is assigned the conditional probability of the corresponding label as defined above. Using this notation and Lemma 2, it follows from (9),(10),(14) that

$$\begin{aligned}\frac{1}{4}\Delta(\Psi) = \; & \eta\mathbb{E}[r^2\lambda^2\cos^2\vartheta] - \eta^2\mathbb{E}[r^4\lambda^2\cos^2\vartheta] - \eta^2\Psi^2\mathbb{E}[r^2] - \eta\mathbb{E}[(\pm\Psi)r\lambda\cos\vartheta] \\ & - 2\eta^2\mathbb{E}[(\pm\Psi)r^3\lambda\cos\vartheta] \end{aligned} \tag{15}$$

Invoking Assumption 1 and using (12), it can be readily shown that

$$\mathbb{E}[(\pm\Psi)r\lambda\cos\vartheta] = \mathbb{E}[(\pm\Psi)r^3\lambda\cos\vartheta] = 0$$

from which it follows that

$$\frac{1}{4}\Delta(\Psi) = \eta\mathbb{E}[r^2\lambda^2\cos^2\vartheta] - \eta^2\mathbb{E}[r^4\lambda^2\cos^2\vartheta] - \eta^2\Psi^2\mathbb{E}[r^2] \tag{16}$$

We can now state the main theorem of this section.

Theorem 3 *Given Assumption 1, the conditional convergence rate $\Delta(\Psi)$ is monotonically decreasing with the Difficulty Score Ψ. If the step size coefficient is sufficiently small so that $\eta \leq \frac{\mathbb{E}[r^2\cos^2\vartheta]}{\mathbb{E}[r^4\cos^2\vartheta]}$, it is likewise monotonically increasing with the distance λ between the current estimate of the hypothesis $\mathbf{w}_t$ and the optimal hypothesis $\bar{\mathbf{w}}$.*

Proof From (16)

$$\frac{\partial\Delta(\Psi)}{\partial\Psi} = -8\eta^2\mathbb{E}[r^2]\Psi \leq 0$$

which proves the first statement. In addition,

$$\frac{\partial\Delta(\Psi)}{\partial\lambda} = 8\eta\lambda\left(\mathbb{E}[r^2\cos^2\vartheta] - \eta\mathbb{E}[r^4\cos^2\vartheta]\right)$$

If $\eta \leq \frac{\mathbb{E}[r^2\cos^2\vartheta]}{\mathbb{E}[r^4\cos^2\vartheta]}$ then $\frac{\partial\Delta(\Psi)}{\partial\lambda} \geq 0$, and the second statement follows. ■

Corollary 3.1 *Although $\Delta(\Psi)$ may be negative, $\mathbf{w}_t$ always converges faster to $\bar{\mathbf{w}}$ when the training points are sampled from easier examples with smaller difficulty score Ψ.*

Corollary 3.2 *If the step size coefficient η is small enough so that $\eta \leq \frac{\mathbb{E}[r^2\cos^2\vartheta]}{\mathbb{E}[r^4\cos^2\vartheta]}$, we should expect faster convergence at the beginning of curriculum-based learning.*

We note, outside the scope of the present discussion, that the predictions of these two corollaries have been observed in simulations with deep CNN network, where the loss function is far from being convex, see (Weinshall et al., 2018; Hacohen and Weinshall, 2019).

3.2 Convergence rate increases with *local difficulty*

The main theorem in this section states that for a fixed global difficulty score Ψ, when the gradient step is small enough, convergence is monotonically *increasing* with the *local difficulty*, or the loss of the point with respect to the current hypothesis. *This is not true in general.* The second theorem in this section shows that when the difficulty score is not fixed, there exist hypotheses $\mathbf{w} \in \mathcal{H}$ for which the convergence rate is decreasing with the local difficulty.

Given (5), let $\Upsilon^2 = L(\mathbf{X}, \mathbf{w}_t)$ denote the loss of $\mathbf{X}$ with respect to the current hypothesis $\mathbf{w}_t$. Define the angle $\beta \in [0, \frac{\pi}{2})$ as follows (see Fig. 2):

$$\beta = \beta(r, \Psi, \lambda) = \arccos(\min(\frac{\Psi}{\lambda r}, 1)) \tag{17}$$

Lemma 4 *The relation between $\Upsilon, \Psi, r, \vartheta$ can be written separately in 4 regions as follows (see Fig. 2):*

$$\begin{array}{lll}
A1 & 0 \leq \vartheta \leq \pi - \beta,\ y = \mathbf{x} \cdot \bar{\mathbf{w}} + \Psi & \implies y = \mathbf{x} \cdot \mathbf{w}_t + \Upsilon, \lambda r \cos\vartheta = \mathbf{x} \cdot (\bar{\mathbf{w}} - \mathbf{w}_t) = -\Psi + \Upsilon \\
A2 & \pi - \beta \leq \vartheta \leq \pi,\ y = \mathbf{x} \cdot \bar{\mathbf{w}} + \Psi & \implies y = \mathbf{x} \cdot \mathbf{w}_t - \Upsilon, \lambda r \cos\vartheta = -\Psi - \Upsilon \\
A3 & 0 \leq \vartheta \leq \beta,\ y = \mathbf{x} \cdot \bar{\mathbf{w}} - \Psi & \implies y = \mathbf{x} \cdot \mathbf{w}_t + \Upsilon, \lambda r \cos\vartheta = \Psi + \Upsilon \\
A4 & \beta \leq \vartheta \leq \pi,\ y = \mathbf{x} \cdot \bar{\mathbf{w}} - \Psi & \implies y = \mathbf{x} \cdot \mathbf{w}_t - \Upsilon, \lambda r \cos\vartheta = \Psi - \Upsilon
\end{array}$$

Proof We keep in mind that $\forall \mathbf{x}$ and Ψ, there are 2 possible labels y whose probability is equal from assumption (12). Recall that $\bar{\mathbf{z}}$ denotes the projection of $\bar{\mathbf{w}}$ on $\Omega_{\mathbf{x}}$. Thus, on the planar section shown in Fig. 2:

- $\bar{\mathbf{z}}$ lies in the upper half space $\iff y = \mathbf{x} \cdot \bar{\mathbf{w}} + \Psi$
- $\bar{\mathbf{z}}$ lies in the lower half space $\iff y = \mathbf{x} \cdot \bar{\mathbf{w}} - \Psi$

This follows from 3 observations: (i) $\bar{\mathbf{x}}$ lies in the upper half space by the definition of the polar coordinate system; (ii) $\mathbf{x} \cdot \bar{\mathbf{w}} - y = \pm\Psi$; and (iii) $0 = \mathbf{x} \cdot \bar{\mathbf{z}} - y = \mathbf{x} \cdot (\bar{\mathbf{z}} - \bar{\mathbf{w}}) + \mathbf{x} \cdot \bar{\mathbf{w}} - y$.

Next, let $\mathbf{z}_t$ denote the projection of $\mathbf{w}_t$ on $\Omega_{\mathbf{x}}$. Then

$$0 = \mathbf{x} \cdot \mathbf{z}_t - y = \mathbf{x} \cdot (\mathbf{z}_t - \mathbf{w}_t) + \mathbf{x} \cdot \mathbf{w}_t - y$$

When $\bar{\mathbf{z}}$ lies in the upper half space, the following can be verified geometrically from Fig. 2:

$$\begin{array}{ccccc}
0 \leq \vartheta \leq \pi - \beta & \implies & \mathbf{x} \cdot (\mathbf{z}_t - \mathbf{w}_t) \geq 0 & \implies & y = \mathbf{x} \cdot \mathbf{w}_t + \Upsilon \\
\pi - \beta \leq \vartheta \leq \pi & \implies & \mathbf{x} \cdot (\mathbf{z}_t - \mathbf{w}_t) \leq 0 & \implies & y = \mathbf{x} \cdot \mathbf{w}_t - \Upsilon
\end{array}$$

■

Next we analyze how the conditional convergence rate changes with Υ. Let $\Delta(\Psi, \Upsilon)$ denote the conditional convergence rate at $\mathbf{w}_t$, given fixed global difficulty Ψ and local difficulty Υ. From (16)

$$\Delta(\Psi, \Upsilon) = 4\eta \mathbb{E}_{\mathbf{X}_t \sim \mathcal{D}|\Psi,\Upsilon}[r^2 \lambda^2 \cos^2 \vartheta] + O(\eta^2)$$

It is easier to analyze $\Delta(\Psi,\Upsilon)$ in a Cartesian coordinates system, rather than polar. We focus again on the $2D$ plane defined by the vectors $\vec{\mathcal{O}} = \bar{\mathbf{w}} - \mathbf{w}_t$ and $\bar{\mathbf{z}} - \bar{\mathbf{w}}$ (see Fig. 2), where we define $u = r\cos\vartheta$, $v = r\sin\vartheta$. The 4 cases listed in Lemma 4 can be readily transformed to this coordinate system as follows: $\{0 \le \vartheta \le \beta\} \Leftrightarrow \{\lambda u \ge \Psi\}$, $\{\beta \le \vartheta \le \pi - \beta\} \Leftrightarrow \{-\Psi \le \lambda u \le \Psi\}$, and $\{\pi - \beta \le \vartheta \le \pi\} \Leftrightarrow \{\lambda u \le -\Psi\}$:

$$\begin{array}{cccc}
A1 & \lambda u \ge -\Psi & \Longrightarrow & \lambda u = -\Psi + \Upsilon \\
A2 & \lambda u \le -\Psi & \Longrightarrow & \lambda u = -\Psi - \Upsilon \\
A3 & \lambda u \ge \Psi & \Longrightarrow & \lambda u = \Psi + \Upsilon \\
A4 & \lambda u \le \Psi & \Longrightarrow & \lambda u = \Psi - \Upsilon
\end{array}$$

Define
$$\nabla = \frac{f(\frac{\Psi+\Upsilon}{\lambda}) - f(\frac{\Psi-\Upsilon}{\lambda}) - f(\frac{-\Psi+\Upsilon}{\lambda}) + f(\frac{-\Psi-\Upsilon}{\lambda})}{f(\frac{\Psi+\Upsilon}{\lambda}) + f(\frac{\Psi-\Upsilon}{\lambda}) + f(\frac{-\Psi+\Upsilon}{\lambda}) + f(\frac{-\Psi-\Upsilon}{\lambda})}$$
Clearly $-1 \le \nabla \le 1$.

Theorem 5 *Assume that the gradient step size is small enough so that second order terms $O(\eta^2)$ can be neglected. Assume that $\frac{\partial\nabla}{\partial\Upsilon} \ge \frac{\Psi}{\Upsilon} - \frac{\Upsilon}{\Psi}$ $\forall\Upsilon$, and invoke Assumption 1. Fix the difficulty score at Ψ. Then the conditional convergence rate $\Delta(\Psi,\Upsilon)$ is monotonically increasing with the local difficulty Υ.*

Proof In the coordinate system defined above $\Delta(\Psi,\Upsilon) = 4\eta\mathbb{E}_{\mathbf{X}_t\sim\mathcal{D}|\Psi,\Upsilon}[\lambda^2u^2] + O(\eta^2)$. We compute $\Delta(\Psi,\Upsilon)$ separately in each region, marginalizing out v based on the following
$$\int\int_0^\infty \lambda^2u^2v^{d-1}f_{\mathcal{D}|\Psi,\Upsilon}(u,v)dvdu = \int\lambda^2u^2f(u)du$$
where $f(u)$ denotes the conditional marginal distribution of u.

Let u_i denote the value of u corresponding to score Υ in each region A1-A4, and $\frac{1}{2}f(u_i)$ its density. $\Delta(\Psi,\Upsilon)$ takes on 4 discrete values, one in each region, and its expected value is therefore $\Delta(\Psi,\Upsilon) = 4\eta\sum_{i=1}^4\lambda^2u_i^2\frac{f(u_i)}{\sum_{i=1}^4 f(u_i)}$. It can be readily shown that
$$\frac{1}{4\eta}\Delta(\Psi,\Upsilon) = \Psi^2 + \Upsilon^2 + 2\Psi\Upsilon\nabla \tag{18}$$
and therefore
$$\begin{aligned}\frac{1}{4\eta}\frac{\partial\Delta(\Psi,\Upsilon)}{\partial\Upsilon} &= 2\Upsilon + 2\Psi\Upsilon\,\frac{\partial\nabla}{\partial\Upsilon} + 2\Psi\,\nabla \\ &\ge 2\Upsilon + 2\Psi\Upsilon\,\frac{\partial\nabla}{\partial\Upsilon} - 2\Psi\end{aligned} \tag{19}$$

From the assumption that $\frac{\partial\nabla}{\partial\Upsilon} \ge \frac{\Psi}{\Upsilon} - \frac{\Upsilon}{\Psi}$ $\forall\Upsilon$, it follows that
$$\frac{1}{8\eta}\frac{\partial\Delta(\Psi,\Upsilon)}{\partial\Upsilon} \ge \Upsilon + \Psi\Upsilon\,\frac{\Psi-\Upsilon}{\Psi\Upsilon} - \Psi = 0$$

■

Corollary 5.1 *For any $c \in \mathbb{R}^+$, if ∇ is $(c - \frac{1}{c})$-Lipschitz then $\frac{\partial \Delta(\Psi,\Upsilon)}{\partial \Upsilon} \geq 0$ for any $\Upsilon \geq c\ \Psi$.*

Corollary 5.2 *If the conditional distribution $\mathcal{D}|\Psi = k(\Psi)$ (i.e., constant for a given Ψ) over a compact region and η small enough, then $\nabla = 0$ and $\frac{\partial \nabla}{\partial \Upsilon} = 0$ $\forall \Upsilon$ excluding the boundaries of the compact region. If in addition $\Upsilon \geq \Psi$ $\forall \mathbf{x}, \mathbf{w}_t$, then $\frac{\partial \Delta(\Psi,\Upsilon)}{\partial \Upsilon} \geq 0$ almost surely.*

Theorem 6 *Assume $\mathcal{D}(\mathbb{X})$ is continuous and $\bar{\mathbf{w}}$ is realizable, and invoke Assumption 1. Then there are always hypotheses $\mathbf{w} \in \mathcal{H}$ for which the conditional convergence rate under $f_{\mathcal{D}|\Psi,\Upsilon}$ is monotonically decreasing with the local difficulty Υ.*

Proof We shift to a hyperspherical coordinate system in $\mathbb{R}^{d+1}$ similar as before, but now the pole (origin) is fixed at $\mathbf{w}_t$. For the gradient vector $\mathbf{s}$, it can be shown that:

$$\begin{aligned} \mathbf{s} &= -\operatorname{sgn}(\mathbf{x} \cdot \mathbf{w}_t - y) 2\eta \mathbf{x} \Upsilon \\ \mathbf{s}_{\mathcal{O}} &= \mathbf{s} \cdot \frac{\bar{\mathbf{w}} - \mathbf{w}_t}{\lambda} = \pm \frac{2\eta}{\lambda} r \lambda \cos\vartheta\ \Upsilon \end{aligned} \tag{20}$$

Let $\Delta(\Upsilon)$ denote the conditional convergence rate at $\mathbf{w}_t$ given Υ. From Lemma 2

$$\begin{aligned} \Delta(\Upsilon) &= 2\eta\Upsilon \Big(\mathbb{E}[r\cos\vartheta|_{\mathbf{x}\cdot\mathbf{w}_t - y = -\Upsilon}] - \mathbb{E}[r\cos\vartheta|_{\mathbf{x}\cdot\mathbf{w}_t - y = \Upsilon}] \Big) - \mathbb{E}[(2\eta r \Upsilon)^2] \\ &\doteq 2\eta\Upsilon Q(r, \vartheta, \mathbf{w}_t) - 4\eta^2\Upsilon^2\mathbb{E}[r^2] \end{aligned}$$

If $\mathbf{w} = \bar{\mathbf{w}}$, then $Q(r, \vartheta, \mathbf{w}) = 0$ from the symmetry implied in Assumption 1. From the continuity of $\mathcal{D}(\mathbb{X})$, there exists $\delta > 0$ such that if $\|\mathbf{w} - \bar{\mathbf{w}}\|_2 < \delta$, then $\|Q(r,\vartheta,\mathbf{w}) - Q(r,\vartheta,\bar{\mathbf{w}})\|_2 < \eta\Upsilon\mathbb{E}[r^2]$, which implies that $\Delta(\Upsilon) < -2\eta^2\Upsilon^2\mathbb{E}[r^2] < 0$. ■

4. Classification with the Hinge Loss

In this section we analyze the hinge loss optimization in the context of binary classification. As in (6), we adopt the notation where $\mathbf{x}$ denotes the vector $[\mathbf{x}, 1]^t \in \mathbb{R}^{d+1}$. The hypothesis $\mathbf{w} \in \mathbb{R}^{d+1}$ defines a linear separator which includes a bias term, and the predicted class for example $\mathbf{x}$ is $y = sign(\mathbf{x} \cdot \mathbf{w})$. The hinge loss function is defined as:

$$L(\mathbf{X}, \mathbf{w}) = \max(1 - (\mathbf{x} \cdot \mathbf{w})y, 0) \tag{21}$$

Since in (21) the margin is fixed at 1, it is desirable (and customarily done) to impose a constraint over the length of the parameter vector $\|\mathbf{w}\|$. Without loss of generality we use the constraint $\|\mathbf{w}\|^2 = 1$ (the relaxation of this constraint is discussed in Appendix B). Introducing a Lagrange multiplier λ, the solution $\bar{\mathbf{w}}$ to the ensuing optimization problem is:

$$\bar{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}} \left[\max(1 - (\mathbf{x} \cdot \mathbf{w})y, 0) + \lambda\|\mathbf{w}\|^2 \right] \tag{22}$$

Note that (22) defines the soft-margin SVM classifier.

When using GD, instead of minimizing the argument of (22), one can minimize (21) directly in each step and subsequently project the solution onto the feasible set (aka *projected gradient descent*). This is the procedure we analyze here, with an update rule similar to (7):

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \eta \mathbf{s}, \quad \mathbf{s} = \begin{Bmatrix} \mathbf{x}y & (\mathbf{x}\cdot\mathbf{w})y \le 1 \\ 0 & elsewhere \end{Bmatrix} \tag{23}$$

A projection $\mathbf{w}_{t+1} = \frac{\mathbf{w}_{t+1}}{\|\mathbf{w}_{t+1}\|}$ follows this gradient step.

Given the normalization constraint on the parameter vector $\mathbf{w}$, a suitable metric for comparing two such vectors is the cosine similarity between them (or their normalized inner product), in preference over the Euclidean distance between the vectors. We therefore define the conditional convergence rate for a given *Difficulty Score* Ψ as

$$\Delta(\Psi) = \mathbb{E}_{\mathbf{X}_t \sim \mathcal{D}|\Psi} \left[\frac{\mathbf{w}_{t+1}\cdot\bar{\mathbf{w}}}{\|\mathbf{w}_{t+1}\|\|\bar{\mathbf{w}}\|} - \frac{\mathbf{w}_t\cdot\bar{\mathbf{w}}}{\|\mathbf{w}_t\|\|\bar{\mathbf{w}}\|} \right]$$

Note that by definition $\|\bar{\mathbf{w}}\| = \|\mathbf{w}_t\| = 1$. Because the hinge loss is piece-wise linear, we insert the identity function $g(x) = x$ into (4)-(5), so that $\Psi(\mathbf{X}) = L(\mathbf{X}, \bar{\mathbf{w}})$ and $\Upsilon(\mathbf{X}) = L(\mathbf{X}, \mathbf{w}_t)$.

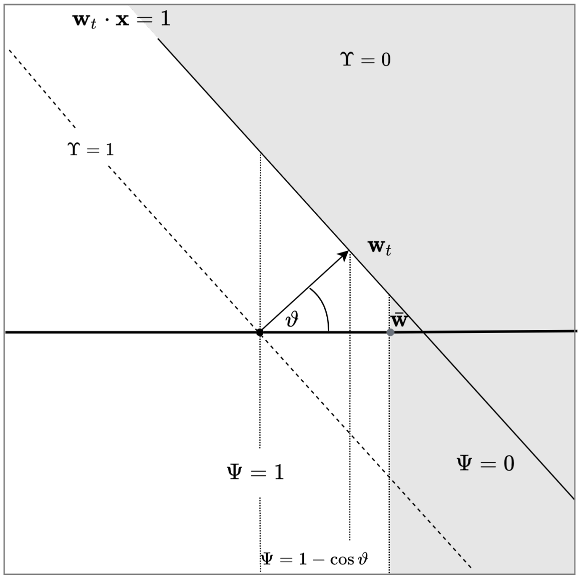

Figure 3: The geometry of the gradient step at time t (see text).

In the following analysis we use a fixed Cartesian coordinate system where the first coordinate axis is defined by $\bar{\mathbf{w}}$, while the plane defined by the first and second axes is the subspace spanned by $\bar{\mathbf{w}}$ and $\mathbf{w}_t$ (see Fig. 3). We assume w.l.o.g that $y_t = 1$ (similar analysis can be repeated in the symmetrical case of $y_t = -1$). By definition, in this coordinate system we have

$$\bar{\mathbf{w}} = [1, 0, \ldots]^t, \quad \mathbf{w}_t = [\cos\vartheta, \sin\vartheta \ldots]^t \tag{24}$$

where $0 \le \vartheta \le \pi$ denotes the angle between $\bar{\mathbf{w}}$ and $\mathbf{w}_t$.

It follows that all the points with *Difficulty Score* $\Psi > 0$ lie on a hyperplane defined by $\mathbf{x} \cdot \bar{\mathbf{w}} = 1 - \Psi$, where from (24)

$$\mathbf{x}|\Psi = [1 - \Psi, x_2, \cdots, x_{d+1}]^t \tag{25}$$

The conditional convergence rate can now be written as follows

$$\Delta(\Psi) = \mathbb{E}_{\mathbf{X}_t \sim \mathcal{D}|\Psi} \left[\frac{\cos\vartheta + \eta(1-\Psi)}{\|\mathbf{w}_{t+1}\|} - \cos\vartheta \right] \tag{26}$$

4.1 Convergence rate decreases with *global difficulty*

The main theorem in this section states that when minimizing the hinge loss, the conditional convergence rate decreases with the global difficulty score Ψ.

Before stating the first lemma, we note that from (23)-(25)

$$\|\mathbf{w}_{t+1}\| = \sqrt{1 + 2\eta[(1-\Psi)\cos\vartheta + x_2 \sin\vartheta] + \eta^2 \|\mathbf{x}\|^2} \tag{27}$$

and

$$\bar{\mathbf{w}} \cdot \mathbf{w}_{t+1} = \cos\vartheta + \eta(1 - \Psi)$$

Lemma 7 *Let $\mathbf{X} = [\mathbf{x}, y]$ denote some example with Difficulty Score $\Psi > 0$, and let*

$$\mathcal{B}(\Psi) \doteq \frac{\Psi - 1}{\tan\vartheta} + \frac{1}{\sin\vartheta} \tag{28}$$

then

$$\mathbf{x} \cdot \mathbf{w}_t < 1 \iff x_2 < \mathcal{B}(\Psi) \tag{29}$$

Proof From (24)-(25) it follows that

$$\mathbf{x} \cdot \mathbf{w}_t = (1 - \Psi)\cos\vartheta + x_2 \sin\vartheta$$

and therefore

$$\mathbf{x} \cdot \mathbf{w}_t < 1 \iff \frac{\cos\vartheta}{\sin\vartheta}(1 - \Psi) + x_2 < \frac{1}{\sin\vartheta}$$

■

Lemma 7 defines the range of x_2 for which $\Upsilon > 0$, namely, the local *Difficulty Score* is positive (see Fig. 3), while the global *Difficulty Score* is fixed at Ψ. This can be used to compute $\Delta(\Psi)$ from (26) and obtain

Lemma 8 *Assume η is small enough, then*

$$\Delta(\Psi) = \int_{-\infty}^{\mathcal{B}(\Psi)} \eta[(1-\Psi)\sin^2\vartheta - x_2 \sin\vartheta\cos\vartheta] \cdot f(x_2) dx_2 + O(\eta^2)$$

where $f(x_2)$ denotes the conditional marginal distribution of $\mathbf{x}$ over the second axis.

Proof Recall that the first coordinate of $\mathbf{x}$ with *Difficulty Score* fixed at $\Psi > 0$ is constant at $x_1 = 1 - \Psi$. We compute $\Delta(\Psi)$ using (26) and Lemma 7:

$$\begin{aligned} \Delta(\Psi) &= \int_{-\infty}^{\mathcal{B}(\Psi)} \int \dots \int \mathcal{I}\ f_{\mathcal{D}|\Psi}(\mathbf{x}) dx_{d+1} \dots dx_3 dx_2 \\ \mathcal{I} &= \frac{\cos\vartheta + (1-\Psi)\cdot\eta}{\|\mathbf{w}_{t+1}\|} - \cos\vartheta \end{aligned} \tag{30}$$

where $\|\mathbf{w}_{t+1}\|$ is defined in (27).

Under the assumption that η is small enough, we approximate the integrand $\mathcal{I}$ in (30) using the first terms of its Taylor expansion at $\eta = 0$, which yields

$$\begin{aligned} \mathcal{I} &\approx \eta\Big[(1-\Psi) - \frac{2\big((1-\Psi)\cos\vartheta + x_2\sin\vartheta\big)\cos\vartheta}{2}\Big] \\ &= \eta[(1-\cos^2\vartheta)(1-\Psi) - x_2\sin\vartheta\cos\vartheta] \\ &= \eta[(1-\Psi)\sin^2\vartheta - x_2\sin\vartheta\cos\vartheta] \end{aligned}$$

Note that $\mathcal{I}$ only depends on x_2, and we can therefore integrate out the remaining integration variables $x_3, \dots, x_{d+1}$. Let $f(x_2)$ denote the marginal distribution of x_2 given $x_1 = 1 - \Psi$. Then

$$\Delta(\Psi) \approx \int_{-\infty}^{\mathcal{B}(\Psi)} \eta[(1-\Psi)\sin^2\vartheta - x_2\sin\vartheta\cos\vartheta] f(x_2) dx_2$$

The derivation above relies on the assumption that the resulting integral is finite, as is the integral in $\mathbb{R}^{d+1}$ of the remaining terms in the Taylor expansion corresponding to $O(\eta^2)$. ■

We can now state the main theorem of this section:

Theorem 9 *Assume that the gradient step size is small enough so that second order terms $O(\eta^2)$ can be neglected. The conditional convergence rate $\Delta(\Psi)$ decreases monotonically as a function of Ψ for every $\Psi > (1-\cos\vartheta)$ when $\cos\vartheta > 0$ (i.e., $\bar{\mathbf{w}}$ and $\mathbf{w}_t$ are positively correlated), and for every $\Psi < (1-\cos\vartheta)$ when $\cos\vartheta < 0$ (i.e., $\bar{\mathbf{w}}$ and $\mathbf{w}_t$ are negatively correlated). Monotonicity holds $\forall\Psi$ when $\cos\vartheta = 0$.*

Proof Using Lemma 8 and the Leibniz Theorem for derivation under the integral sign, we get

$$\frac{\partial\Delta(\Psi)}{\partial\Psi} = \Delta_1 + \Delta_2$$

where

$$\begin{aligned} \Delta_1 &= \eta[(1-\Psi)\sin^2\vartheta - x\sin\vartheta\cos\vartheta] f(\mathcal{B}(\Psi)) \frac{\partial\mathcal{B}(\Psi)}{\partial\Psi} \quad \text{and} \quad \frac{\partial\mathcal{B}(\Psi)}{\partial\Psi} = \frac{\cos\vartheta}{\sin\vartheta} \\ \Delta_2 &= \int_{-\infty}^{\mathcal{B}(\Psi)} \frac{\partial}{\partial\Psi}\eta[(1-\Psi)\sin^2\vartheta - x_2\sin\vartheta\cos\vartheta] f(x_2) dx_2 \\ &= \int_{-\infty}^{\mathcal{B}(\Psi)} -\eta\sin^2\vartheta f(x_2) dx_2 \end{aligned}$$

Clearly $\Delta_2 \leq 0$. It therefore suffices to prove the sufficient condition $\Delta_1 \leq 0$ in order to conclude the proof.

Case 1: $\cos\vartheta = 0$, where $\Delta_1 = 0 \implies \frac{\partial\Delta(\Psi)}{\partial\Psi} < 0 \;\; \forall\Psi$.

Case 2: $\cos\vartheta > 0$. Since $f(x) \geq 0$ (a density function), $\Delta_1 \leq 0$ iff the first multiplicand in the expression describing Δ_1 above is non-negative. Using inequality (29) and substituting $\mathcal{B}(\Psi)$ into this term, we get the following upper bound:

$$\begin{aligned}(1-\Psi)\sin^2\vartheta - \mathcal{B}(\Psi)\sin\vartheta\cos\vartheta &= (1-\Psi)\sin^2\vartheta - [(\Psi-1)\cos\vartheta + 1]\cos\vartheta \\ &= 1 - \cos\vartheta - \Psi\end{aligned}$$

Clearly $\forall\ \Psi > (1-\cos\vartheta)$ this term is negative.

Case 3: $\cos\vartheta < 0$. Using the same line of reasoning, now $\Psi < (1-\cos\vartheta) \implies \Delta_1 < 0$ since $\frac{\partial\mathcal{B}(\Psi)}{\partial\Psi} < 0$. ■

Early in the training procedure we expect **Case 2**, when $\cos\vartheta > 0$ and $\bar{\mathbf{w}}, \mathbf{w}_t$ are positively correlated, to dominate SGD learning. This is because in a high dimensional space, two randomly picked vectors are expected to be almost orthogonal to each other, and therefore only a small step towards the optimal hypothesis is needed in order to satisfy this condition. Now the relevant condition is $\Psi > 1 - \cos\vartheta$, defining a range which includes almost all the training examples with non-zero *Difficulty Score.*

The condition on Ψ in the theorem is necessary. To see this, we next show that when $\cos\vartheta > 0$ and $0 < \Psi < 1 - \cos\vartheta$, there are cases for which the theorem does not hold. Similar construction exists when $\cos\vartheta < 0$ and $\Psi > 1 - \cos\vartheta$.

Theorem 10 *For all w_t and when $\cos\vartheta > 0$, there exists $\mathcal{D}$ for which $\Delta(\Psi)$ is not monotonically decreasing with Ψ in the range $[0, 1-\cos\vartheta]$.*

Proof Let $0 < \Psi_1 < \Psi_2 < 1 - \cos\vartheta$. Assume that $f(x_2) = 0 \;\forall x_2 \leq \mathcal{B}(\Psi_1)$, thus $\Delta(\Psi_1) = 0$. It remains to show that $\Delta(\Psi_2) > 0$. From Lemma 8 and neglecting second order terms in η

$$\Delta(\Psi_2) \approx \eta \int_{\mathcal{B}(\Psi_1)}^{\mathcal{B}(\Psi_2)} \mathcal{J}(x_2) f(x_2) dx_2$$

$$\mathcal{J}(x_2) = (1-\Psi_2)\sin^2\vartheta - x_2\sin\vartheta\cos\vartheta$$

We next observe that $\mathcal{J}(x) > 0 \;\forall x$ where $\mathcal{B}(\Psi_1) \leq x \leq \mathcal{B}(\Psi_2)$. This is because $\mathcal{J}(x)$ is monotonically decreasing with x, and $\mathcal{B}(\Psi_2) > 0$ for $\Psi_2 < 1 - \cos\vartheta$. It thus follows that $\Delta(\Psi_2) > 0$, which concludes the proof. ■

4.2 Convergence rate increases with *local difficulty*

In a similar manner to the case of linear regression and under the same assumptions, we show that when Ψ is fixed, the conditional convergence rate with respect to the local difficulty Υ is increasing, opposite to its trend with Ψ.

As in Section 3.2, we define:

$$\Delta(\Psi, \Upsilon) = \mathbb{E}_{\mathbf{X}_t \sim \mathcal{D}|\Psi,\Upsilon}\left[\frac{\mathbf{w}_{t+1}\cdot\bar{\mathbf{w}}}{\|\mathbf{w}_{t+1}\|\;\|\bar{\mathbf{w}}\|} - \frac{\mathbf{w}_t\cdot\bar{\mathbf{w}}}{\|\mathbf{w}_t\|\;\|\bar{\mathbf{w}}\|}\right]$$

Theorem 11 *Assume that the gradient step size is small enough so that we can neglect second order terms $O(\eta^2)$. Assume further that $\cos\vartheta \geq 0$. Fixing Ψ and $\forall\Psi$, the conditional convergence rate is monotonically increasing with Υ for every $\Upsilon > 0$.*

Proof From Fig. 3 we see that when Ψ, Υ are given, the projection of data point $\mathbf{x}$ onto $X_1 \times X_2$ is a point where $x_1 = 1 - \Psi$, and

$$\begin{aligned} &(\cos\vartheta, \sin\vartheta) \cdot (1 - \Psi, x_2) = 1 - \Upsilon \\ \implies\ & x_2 \doteq \mathcal{X}(\Psi, \Upsilon) = \frac{\Psi - 1}{\tan\vartheta} + \frac{1 - \Upsilon}{\sin\vartheta} \end{aligned}$$

In the same manner used to prove Lemma 8, we can show that

$$\Delta(\Psi, \Upsilon) = \eta[(1 - \Psi)\sin^2\vartheta - \mathcal{X}(\Psi, \Upsilon)\sin\vartheta\cos\vartheta] + O(\eta^2)$$

It follows that

$$\frac{\partial\Delta(\Psi, \Upsilon)}{\partial\Upsilon} = \eta\cos\vartheta \geq 0$$

which concludes the proof. ■

5. Summary and Discussion

This paper offers the first theoretical investigation of curriculum learning, in the context of convex optimization. In its simplest form, curriculum learning can be viewed as a variant of stochastic gradient descent, where easy examples are more frequently sampled at the beginning of training. In order to formalize this intuition, we must first define how to measure difficulty. Here we define the *global Difficulty Score* of a point as its loss with respect to the optimal hypothesis. This definition allows us to analyze the benefits of curriculum learning in two representative convex optimization problems - binary classification with hinge loss minimization, and linear regression. In the context of these two optimization problems we show that curriculum learning, with an initial bias in favor of training points whose loss with respect to the *optimal hypothesis* is **lower**, accelerates learning. We also show that when the *Difficulty Score* is fixed, convergence of SGD optimization is accelerated when preferring training points whose loss with respect to the *current hypothesis* (*local Difficulty Score*) is **higher**.

These theoretical results can direct us towards the development of new practical methods which will incorporate both global and local scores in order to balance between easy and hard examples. One simple approach to achieve this end may control the pace of the curriculum schedule by employing the local score. More sophisticated algorithms can combine biases based on both scores.

Our results suggest that the correlation between local and global difficulty scores can predict whether methods like SPL that favor currently easier examples, or rather methods like hard example mining that favor currently hard examples, should be preferred in specific tasks. For example, when learning from noisy data, we expect to see high correlation between the local and global difficulty scores, and therefore preference towards examples with low local score will also bias towards examples with low global difficulty score. In such cases SPL, which gives preference to examples with lower local score, is predicted by our theoretical analysis to improve convergence. On the other hand, if the local and global difficulty scores are not correlated, hard data mining is likely to perform better based on our theoretical analysis.

Acknowledgements

This work was supported in part by a grant from the Israeli Science Foundation (ISF) and by the Gatsby Charitable Foundations.

Appendix A. Bayesian Formulation

The results in Section 3 depend on the assumption that the two remaining labels when the difficulty score is fixed, $y_1(\mathbf{x}) = \mathbf{x} \cdot \bar{\mathbf{w}} + \Psi$ and $y_2(\mathbf{x}) = \mathbf{x} \cdot \bar{\mathbf{w}} - \Psi$, are equally likely: $f_{\mathcal{D}}([\mathbf{x}, y_i(\mathbf{x})]) = \frac{1}{2} f(r, \vartheta, \Phi)$. We now describe how this assumption can be relaxed in a Bayesian framework.

Let $\mathbf{v}$ denote an additional (vector) hyper-parameter of the distribution $\mathcal{D}(\mathbb{X})$, such that $f_{\mathcal{D}|\Psi} = f_{\mathbf{v}}(r, \vartheta, \Phi, y)$. Let $q(\mathbf{v})$ denote the distribution of $\mathbf{v}$. Assume that the conditional marginal distribution of the data over v does not depend on the label, namely

$$\int f_v(r, \theta, \Phi, y) q(v) dv = f(r, \vartheta, \Phi) \quad \forall \Psi \tag{31}$$

It follow that the marginal conditional distribution of the data satisfies the required condition

$$\int f_{\mathcal{D}|\Psi}(\mathbf{x}, y) dy = \int f_v(r, \theta, \Phi, y) q(v) dv = f(r, \vartheta, \Phi)$$

Thus assumption (31) suffices for Theorems 3 and 5 to hold true in a Bayesian framework, when taking the average over all hyper-parameter values.

Appendix B. Normalization of the Parameter Vector

Throughout the analysis in Section 4 we assumed the constraint $\|\mathbf{w}\| = 1$, but the results also apply to any norm A where $\|\mathbf{w}\| = A$. To see this, let us define $\mathbf{x}' = A\mathbf{x}$. Define the following distribution $\mathcal{D}'$ on $\mathbf{X}'$

$$\forall \mathbf{x}',\ y:\ \mathcal{D}'([\mathbf{x}', y]) = \mathcal{D}([A\mathbf{x}, y])$$

We note that

$$\begin{aligned} \operatorname*{argmin}_{\mathbf{w},\ s.t. \|\mathbf{w}\|=A} \max(1 - (\mathbf{x} \cdot \mathbf{w}) y, 0) &= \operatorname*{argmin}_{\mathbf{w},\ s.t. \|\mathbf{w}\|=1} \max(1 - (A\mathbf{x} \cdot \mathbf{w}) y, 0) \\ &= \operatorname*{argmin}_{\mathbf{w},\ s.t. \|\mathbf{w}\|=1} \max(1 - \mathbf{x}' \cdot \mathbf{w}) y, 0) \end{aligned}$$

The latter is the problem we have analyzed for any distribution on the training examples, including $\mathcal{D}'$. Thus the theorems we have proved hold true for this problem as well.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009.

Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *Advances in Neural Information Processing Systems*, pages 1002–1012, 2017.

Jeffrey L Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2535–2544, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Regularizing very deep neural networks on corrupted labels. *arXiv preprint arXiv:1712.05055*, 2017.

M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, pages 1189–1197, 2010.

Tom M Mitchell. *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ. New Jersey, 1980.

Tom Michael Mitchell. *The discipline of machine learning*, volume 9. Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006.

Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. In *Advances in neural information processing systems*, pages 2863–2871, 2015.

Terence D Sanger. Neural network learning control of robot manipulators using gradually increasing task difficulty. *IEEE transactions on Robotics and Automation*, 10(3):323–333, 1994.

Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26 (5):1651–1686, 1998. ISSN 00905364. URL `http://www.jstor.org/stable/120016`.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016.

Burrhus Frederic Skinner. *The behavior of organisms: An experimental analysis*. BF Skinner Foundation, 1990.

Panqu Wang and Garrison W Cottrell. Basic level categorization facilitates visual object recognition. *arXiv preprint arXiv:1511.04103*, 2015.

Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5238–5246, 2018.