Mirror Descent Maximizes Generalized Margin and Can Be Implemented Efficiently

Haoyuan Sun
MIT
haoyuans@mit.edu

Kwangjun Ahn
MIT
kjahn@mit.edu

Christos Thrampoulidis
UBC
cthrampo@ece.ubc.ca

Navid Azizan
MIT
azizan@mit.edu

Abstract

Driven by the empirical success and wide use of deep neural networks, understanding the generalization performance of overparameterized models has become an increasingly popular question. To this end, there has been substantial effort to characterize the implicit bias of the optimization algorithms used, such as gradient descent (GD), and the structural properties of their preferred solutions. This paper answers an open question in this literature: For the classification setting, what solution does mirror descent (MD) converge to? Specifically, motivated by its efficient implementation, we consider the family of mirror descent algorithms with potential function chosen as the p-th power of the ℓ_p-norm, which is an important generalization of GD. We call this algorithm p-GD. For this family, we characterize the solutions it obtains and show that it converges in direction to a *generalized maximum-margin* solution with respect to the ℓ_p-norm for linearly separable classification. While the MD update rule is in general expensive to compute and perhaps not suitable for deep learning, p-GD is fully parallelizable in the same manner as SGD and can be used to train deep neural networks with virtually *no additional computational overhead*. Using comprehensive experiments with both linear and deep neural network models, we demonstrate that p-GD can noticeably affect the structure and the generalization performance of the learned models.

1 Introduction

Overparameterized deep neural networks have enjoyed a tremendous amount of success in a wide range of machine learning applications [Brown et al., 2020, Dosovitskiy et al., 2020, Ramesh et al., 2021, Schrittwieser et al., 2020]. However, as these highly expressive models have the capacity to have multiple solutions that interpolate training data, and not all these solutions perform well on test data, it is important to characterize which of these interpolating solutions the optimization algorithms converge to. Such characterization is important as it helps understand the generalization performance of these models, which is one of the most fundamental questions in machine learning.

Notably, it has been observed that even in the absence of any explicit regularization, the interpolating solutions obtained by the standard gradient-based optimization algorithms, such as (stochastic) gradient descent, tend to generalize well. Recent research has highlighted that such algorithms favor particular types of solutions, i.e., they *implicitly regularize* the learned models. Importantly, such implicit biases are shown to play a crucial role in determining generalization performance, e.g., [Donhauser et al., 2022, Neyshabur et al., 2014, Wilson et al., 2017, Zhang et al., 2021].

In the literature, the implicit bias of first-order methods is first studied in linear settings since the analysis is more tractable, and also, there have been several theoretical and empirical evidence that certain insights from linear models translate to deep learning, e.g. [Allen-Zhu et al., 2019, Bartlett et al., 2017, Belkin et al., 2019, Jacot et al., 2018, Lyu and Li, 2019, Nakkiran et al., 2021]. In the linear setting, it is easier to establish implicit bias for regression tasks, where square loss is typically

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Table 1: **Conceptual summary of our results.** In the case of linear regression, the implicit regularization results are complete; it is shown that mirror descent converges to the interpolating solution that is closest to the initialization. However, such characterization in the classification setting is missing in the literature and this is precisely the goal of this work. In particular, motivated by its practical application, we consider the potential function $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^p$ and extend the result of the gradient descent to such mirror descents.

	Regression	Classification
Gradient Descent ($\psi(\cdot) = \frac{1}{2}\|\cdot\|_2^2$)	$\operatorname{argmin}_w \|w - w_0\|_2$ s.t. w fits all data [Engl et al., 1996, Thm 6.1]	$\operatorname{argmin}_w \|w\|_2$ s.t. w classifies all data Soudry et al. [2018] Ji and Telgarsky [2019]
Mirror Descent (e.g. $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^p$)	$\operatorname{argmin}_w \|w - w_0\|_p$ s.t. w fits all data Gunasekar et al. [2018] Azizan and Hassibi [2019a]	$\operatorname{argmin}_w \|w\|_p$ s.t. w classifies all data **This work**

used and it attains its minimum at a finite value. For example, the implicit bias of gradient descent (GD) for square loss goes back to Engl et al. [1996]. Beyond GD, analysis of other popular algorithms such as the family of mirror descent (MD), which is an important generalization of GD, is more involved and was established only recently by [Azizan and Hassibi, 2019a, Gunasekar et al., 2018]. Specifically, those works showed that mirror descent converges to the interpolating solution that is closest to the initialization in terms of a Bregman divergence. Thus, the implicit bias in linear regression is relatively well-understood by now.

On the other hand, **in the classification setting, the implicit bias analysis becomes significantly more challenging, and several questions remain open** despite significant progress in the past few years. A key differentiating factor in the classification setting is that the loss function does not attain its minimum at a finite value, and the weights have to grow to infinity. It has been shown that for the logistics loss, the gradient descent iterates converge to the ℓ_2-maximum margin SVM solution in direction [Ji and Telgarsky, 2019, Soudry et al., 2018]. However, such characterizations for mirror descent are missing in the literature. Because it is possible for optimization algorithms to exhibit implicit bias in regression but not in classification (and vice versa) [Gunasekar et al., 2018], resolving this gap of knowledge warrants careful analysis. See Table 1 for a summary.

In this paper, we advance the understanding of the implicit regularization of mirror descent in the classification setting. In particular, inspired by their practicality, we focus on mirror descents with potential function $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^p$ for $p > 1$. More specifically, such choice of potential results in an update rule that *can be applied coordinate-wise*, in the sense that updating the value at one coordinate does not depend on the values at other coordinates. Thanks to this property, this subclass of mirror descent can be implemented with *no additional computational overhead*, making it much more practical than other algorithms in the literature; see Remark 10 for more details.

Our contributions. In this paper, we make the following contributions:

- We study mirror descent with potential $\frac{1}{p}\|\cdot\|_p^p$ for $p > 1$, which will call *p-norm GD*, and abbreviated as p-GD, as a practical and efficient generalization of the popular gradient descent.
- We show that for separable linear classification with logistics loss, p-GD exhibits implicit regularization by converging in direction to a "generalized" maximum-margin solution with respect to the ℓ_p norm. More generally, we show that, for monotonically decreasing loss functions, p-GD follows the so-called regularization path, which is defined in Section 2.
- We investigate the implications of our theoretical findings with two sets of experiments: Our experiments involving linear models corroborate our theoretical results, and real-world experiments with deep neural networks and popular datasets suggest that our findings carry over to such

nonlinear settings. Our deep learning experiments further show that p-GD with different p lead to significantly different generalization performance.

Additional related work. We remark that recent works also attempt to accelerate the convergence of gradient descent to its implicit regularization, either by using an aggressive step size schedule [Ji and Telgarsky, 2021, Nacson et al., 2019] or with momentum [Ji et al., 2021]. Further, there have been several results for other optimization methods, including steepest descent, AdaBoost, and various adaptive methods such as RMSProp and Adam [Gunasekar et al., 2018, Min et al., 2022, Rosset et al., 2004, Telgarsky, 2013, Wang et al., 2021]. A mirror-descent-based algorithm for explicit regularization was recently proposed by Azizan et al. [2021a]. Comparatively, there has been very little progress on mirror descent in the classification setting. Li et al. [2021] consider a mirror descent, but their assumptions are not applicable beyond the ℓ_2 geometry.[1] To the best of our knowledge, there is no result for more general mirror descent algorithms in the classification setting.

2 Background and Problem Setting

We are interested in the well-known classification setting. Consider a collection of input-label pairs $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{\pm 1\}$ and a classifier $f_w(x)$, where $w \in \mathcal{W}$. For some *loss function* $\ell : \mathbb{R} \times \{\pm 1\} \to \mathbb{R}$, our goal is to minimize the empirical loss:

$$L(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i \cdot f_w(x_i)).$$

Throughout the paper, we assume that the classification loss function ℓ is decreasing, convex and does not attain its minimum, as in most common loss functions in practice (e.g., logistics loss and exponential loss). Without loss of generality, we assume that $\inf \ell(\cdot) = 0$.

For our theoretical analysis, we consider a linear model, where the models can be expressed by $f_w(x) = w^\top x$ and $w \in \mathbb{R}^d$. We also make the following assumptions about the data. First, since we are mainly interested in the over-parameterized setting where $d > n$, we assume that the data is linearly separable, i.e., there exists $w^* \in \mathbb{R}^d$ s.t. $\operatorname{sign}(\langle w^*,\ x_i \rangle) = y_i$ for all $i \in [n]$. We also assume that the inputs x_i's are bounded. More specifically, for our later purpose, we assume that for $p > 0$, there exists some constant C so that $\max_i \|x_i\|_q < C$, where $1/q + 1/p = 1$.

Preliminaries on mirror descent. The key component of mirror descent is a *potential function*. In this work, we will focus on differentiable and strictly convex potentials defined on the entire domain $\mathbb{R}^n$.[2] We call $\nabla\psi$ the corresponding *mirror map*. Given a potential, the natural notion of "distance" associated with the potential ψ is given by the Bregman divergence.

Definition 1 (Bregman divergence [Bregman, 1967])**.** *For a mirror map ψ, the Bregman divergence $D_\psi(\cdot, \cdot)$ associated to ψ is defined as*

$$D_\psi(x, y) := \psi(x) - \psi(y) - \langle \nabla\psi(y),\ x - y \rangle, \qquad \forall x, y \in \mathbb{R}^n.$$

An important case is the potential $\psi = \frac{\|\cdot\|^2}{2}$, where $\|\cdot\|$ denotes the Euclidean norm. Then, the Bregman divergence becomes $D_\psi(x, y) = \frac{1}{2}\|x - y\|^2$. For more background on Bregman divergence and its properties, see, e.g., [Bauschke et al., 2017, Section 2.2] and [Azizan and Hassibi, 2019b].

Mirror descent (MD) with respect to the mirror map ψ is a generalization of gradient descent where we use Bregman divergence as a measure of distance:

$$w_{t+1} = \operatorname*{argmin}_w \left\{ \frac{1}{\eta} D_\psi(w, w_t) + \langle \nabla L(w_t),\ w \rangle \right\} \tag{MD}$$

Equivalently, MD can be written as $\nabla\psi(w_{t+1}) = \nabla\psi(w_t) - \eta\nabla L(w_t)$. We refer readers to [Bubeck, 2015, Figure 4.1] for a nice illustration of mirror descent. Also, see [Juditsky et al., 2011, Section 5.7] for various examples of potentials depending on applications.

One property we will repeatedly use is the following [Azizan and Hassibi, 2019a]:

[1] To be precise, they assume that the Bregman divergence is lower and upper bounded by a constant factor of the squared Euclidean distance, e.g., as in the case of a squared Mahalanobis distance.

[2] In general, the mirror map is a convex function of Legendre type (see, e.g., [Rockafellar, 1970, Section 26]).

Lemma 2 (MD identity)**.** *For any $w \in \mathbb{R}^n$, the following identities hold for* MD*:*

$$D_\psi(w, w_t) = D_\psi(w, w_{t+1}) + D_{\psi-\eta L}(w_{t+1}, w_t) + \eta D_L(w, w_t) - \eta L(w) + \eta L(w_{t+1}), \quad (1a)$$

$$= D_\psi(w, w_{t+1}) + D_{\psi-\eta L}(w_{t+1}, w_t) - \eta \langle \nabla L(w_t),\ w - w_t \rangle - \eta L(w_t) + \eta L(w_{t+1}). \quad (1b)$$

Using Lemma 2, we make several new observations and prove the following useful statements.

Lemma 3. *For sufficiently small step size η such that $\psi - \eta L$ is convex, the loss is monotonically decreasing after each iteration of* MD*, i.e., $L(w_{t+1}) \le L(w_t)$.*

Lemma 4. *In a separable linear classification problem, if η is chosen sufficiently small s.t. $\psi - \eta L$ is convex, then we have $L(w_t) \to 0$ as $t \to \infty$. Hence, $\lim_{t\to\infty} \|w_t\| = \infty$ for any norm $\|\cdot\|$.*

The formal proofs of these lemmas can be found in Appendix A.

Remark 5. We can relax the condition in Lemma 3 and 4 such that for a sufficiently small step size η, $\psi - \eta L$ is only locally convex at the iterates w_t. The relaxed condition allows us to analyze losses such as the exponential loss (see, e.g. Footnote 2 of Soudry et al. [2018]). This condition can be considered as the mirror descent counterpart to the standard smoothness assumption in the analysis of gradient descent (see Lu et al. [2018]).

Preliminaries on the convergence of linear classifier. As we discussed above, the weights vector w_t diverges for mirror descent. Here the main theoretical question is:

What direction does MD diverge to? In other words, can we characterize $w_t / \|w_t\|$ as $t \to \infty$?

To answer this question, we define two special directions whose importance will be illustrated later.

Definition 6. *The **regularization path** with respect to the ℓ_p-norm is defined as*

$$\bar{w}_p(B) = \operatorname*{argmin}_{\|w\|_p \le B} L(w) \quad (2)$$

*And if the limit $\lim_{B\to\infty} \bar{w}_p(B)/B$ exists, we call it the **regularized direction** and denote it by u_p^{r}.*

Definition 7. *The **margin** γ of the a linear classifier w is defined as $\gamma(w) = \min_{i=1,\dots,n} y_i \langle x_i,\ w \rangle$. The* **max-margin direction** *with respect to the ℓ_p-norm is defined as:*

$$u_p^{\mathsf{m}} := \operatorname*{argmax}_{\|w\|_p \le 1} \left\{ \min_{i=1,\dots,n} y_i \langle x_i,\ w \rangle \right\} \quad (3)$$

And let $\hat{\gamma}_p$ be the optimal value to the equation above.

Note that u_p^{m} is parallel to the hard-margin SVM solution w.r.t. ℓ_p-norm: $\operatorname{argmin}_w \{\|w\|_p : \gamma(w) \ge 1\}$. Also note that the superscripts in u_p^{r} and u_p^{m} are not variables and we only use this notation to differentiate the two definitions. Prior results had shown that, in linear classification, gradient descent converges in direction.

Theorem 8 (Soudry et al. [2018])**.** *For separable linear classification with logistics loss, the gradient descent iterates with sufficiently small step size converge in direction to u_2^{m}, i.e., $\lim_{t\to\infty} \frac{w_t}{\|w_t\|_2} = u_2^{\mathsf{m}}$.*

Theorem 9 (Ji et al. [2020])**.** *If the regularized direction u_p^{r} with respect to the ℓ_2-norm exists, then the gradient descent iterates with sufficiently small step size converge to the regularized direction u_2^{r}, i.e., $\lim_{t\to\infty} \frac{w_t}{\|w_t\|_2} = u_2^{\mathsf{r}}$.*

3 Mirror Descent with the p-th Power of ℓ_p-norm

In this section, we investigate theoretical properties of the main algorithm of interest, namely mirror descent with $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^p$ and for $p > 1$.[3] We shall call this algorithm *p-norm GD* because it naturally generalizes gradient descent to ℓ_p geometry, and for conciseness, we will refer to this algorithm by the shorthand p-GD. As noticed by Azizan et al. [2021b], this choice of mirror potential

[3] Because the potential function must be *strictly* convex for Bregman divergence to be well-defined, the value of p cannot be exactly 1.

is particularly of practical interest because the mirror map $\nabla\psi$ updates becomes *separable* in coordinates and thus can be implemented *coordinate-wise* independent of other coordinates:

$$\forall j \in [d], \quad \begin{cases} w_{t+1}[j] \leftarrow \left|w_t^+[j]\right|^{\frac{1}{p-1}} \cdot \operatorname{sign}\left(w_t^+[j]\right) \\ w_t^+[j] := |w_t[j]|^{p-1}\operatorname{sign}(w_t[j]) - \eta \nabla L(w_t)[j] \end{cases} \tag{p-GD}$$

Furthermore, we can extend upon the observation of Azizan et al. [2021b] and derive these identities that allow us to better manipulate p-GD:

$$\langle \nabla\psi(w),\ w\rangle = \operatorname{sign}(w_1)w_1 \cdot |w_1|^{p-1} + \cdots + \operatorname{sign}(w_d)w_d \cdot |w_d|^{p-1} = \|w\|^p \tag{4a}$$

$$D_\psi\left(cw, cw'\right) = |c|^p D_\psi\left(w, w'\right) \quad \forall c \in \mathbb{R}. \tag{4b}$$

Remark 10. Note that the coordinate-wise separability property is not shared among other related algorithms in the literature. For instance, the choice $\psi = \frac{1}{2}\|\cdot\|_q^2$ for $1/p + 1/q = 1$, which is referred to as the p-norm algorithm [Gentile, 2003, Grove et al., 2001] is not fully coordinate-wise separable since it requires computing $\|w_t\|_p$ at each step (see, e.g., [Gentile, 2003, eq. (1)]). Another related algorithm is steepest descent, where the Bregman divergence in MD is replaced with $\|\cdot\|^2$ for general norm $\|\cdot\|$.[4] However, for similar reasons, the update rule is not fully separable.

3.1 Main theoretical results

We extend Theorems 8 and 9 to the setting of p-GD. We will resolve two major obstacles in the analysis of implicit regularization in linear classification:

- Our analysis approaches the classification setting as a limit of the regression implicit bias. This argument gives stronger theoretical justification for utilizing the regularized direction (as employed by Ji et al. [2020]) and partially addresses the concern from Gunasekar et al. [2018] that the implicit bias of regression and classification problems are "fundamentally different."
- On a more technical note, analyzing the implicit bias requires handling the cross terms of the form $\langle \nabla\psi(w),\ w'\rangle$, which lack direct geometric interpretations. We demonstrate that for our potential functions of interest, these terms can be nicely written and can be handled in the analysis.

We begin with the motivation behind the regularized direction, and consider the regression setting in which there exists some weight vector w such that $L(w) = 0$. Then, we can apply Lemma 2 to get

$$D_\psi(w, w_t) = D_\psi(w, w_{t+1}) + D_{\psi-\eta L}(w_{t+1}, w_t) + \eta D_L(w, w_t) + \eta(L(w_{t+1}) - L(w))$$

Since we assumed $L(w) = 0$, the equation above implies that $D_\psi(w, w_t) \geq D_\psi(w, w_{t+1})$ for sufficiently small step-size η. This can be interpreted as MD having a decreasing "potential" of the from $D_\psi\left(w, \cdot\right)$ during each step. Using this property, Azizan and Hassibi [2019a] establishes the implicit bias results of mirror descent in the regression setting.

However, such weight vector w does not exist in the classification setting. One natural workaround would then be to choose a vector w so that $L(w) \leq L(w_t)$ for all $t \leq T$. The following result, which is a generalization of [Ji et al., 2020, Lemma 9], shows that one can in fact choose the reference vector w as a scalar multiple of the regularized direction.

Lemma 11. *If the regularized direction u_p^r exists, then $\forall \alpha > 0$, there exists r_α such that for any w with $\|w\|_p > r_\alpha$, we have $L((1+\alpha)\|w\|_p u_p^r) \leq L(w)$.*

However, this does not resolve the issue altogether. Recall from Lemma 4 that the loss approaches 0, and therefore one cannot choose a fixed reference vector w in the limit as $T \to \infty$. But due to the homogeneity of Bregman divergence (4b), we can scale u_p^r by a constant factor during each iteration, and, by doing so, we choose the reference vector w to be a "moving target." In other words, the idea behind our analysis is that the classification problem is chasing after a regression one and would behave similar to it in the limit. Let us formalize this idea. We begin with the following inequality:

$$D_\psi\left(c_t u_p^r, w_{t+1}\right) \leq D_\psi\left(c_t u_p^r, w_t\right) - \eta L(w_{t+1}) + \eta L(w_t), \tag{5}$$

[4] It is also worth noting that steepest descent is not an instance of mirror descent since $\|\cdot\|^2$ is not a Bregman divergence for a general norm $\|\cdot\|$.

where c_t is taken to be $\approx \|w_t\|_p$.[5]

Now we modify (5) so that it can telescope over different iterations. One way is to add $D_\psi\left(c_{t+1}u_p^{\mathsf{r}}, w_{t+1}\right)$ on both sides of (5) and move $D_\psi\left(c_t u_p^{\mathsf{r}}, w_{t+1}\right)$ to the right-hand side as follows:

$$\begin{aligned}
&D_\psi\left(c_{t+1}u_p^{\mathsf{r}}, w_{t+1}\right) \\
&\le D_\psi\left(c_t u_p^{\mathsf{r}}, w_t\right) - \eta L(w_{t+1}) + \eta L(w_t) + D_\psi\left(c_{t+1}u_p^{\mathsf{r}}, w_{t+1}\right) - D_\psi\left(c_t u_p^{\mathsf{r}}, w_{t+1}\right) \\
&= D_\psi\left(c_t u_p^{\mathsf{r}}, w_t\right) - \eta L(w_{t+1}) + \eta L(w_t) + \psi(c_{t+1}u_p^{\mathsf{r}}) - \psi(c_t u_p^{\mathsf{r}}) - \left\langle \nabla\psi(w_{t+1}),\ (c_{t+1}-c_t)u_p^{\mathsf{r}}\right\rangle
\end{aligned}$$

Summing over $t = 0, \dots, T-1$ gives us

$$\begin{aligned}
D_\psi\left(c_T u_p^{\mathsf{r}}, w_T\right) \le D_\psi\left(c_0 u_p^{\mathsf{r}}, w_0\right) - \eta L(w_1) + \eta L(w_T) + \psi(c_T u_p^{\mathsf{r}}) - \psi(c_1 u_p^{\mathsf{r}}) \\
- \sum_{t=1}^{T-1} \left\langle \nabla\psi(w_{t+1}),\ (c_{t+1}-c_t)u_p^{\mathsf{r}}\right\rangle
\end{aligned} \tag{6}$$

The rest of the argument deals with simplifying quantities that do not cancel under telescoping sum. For instance, in order to deal with $\left\langle \nabla\psi(w_{t+1}),\ u_p^{\mathsf{r}}\right\rangle$, we invoke the MD update rule as follows

$$\left\langle \nabla\psi(w_{t+1}) - \nabla\psi(w_t),\ u_p^{\mathsf{r}}\right\rangle = \left\langle -\eta\nabla L(w_t),\ u_p^{\mathsf{r}}\right\rangle \gtrsim \left\langle -\eta\nabla L(w_t),\ w_t\right\rangle,$$

where the last inequality follows from the intuition that u_p^{r} is the direction along which the loss achieves the smallest value and hence $\nabla L(w_t)$ must point away from u_p^{r}, i.e., it must be that $\left\langle \nabla L(w_t),\ u_p^{\mathsf{r}}\right\rangle \lesssim \left\langle \nabla L(w_t),\ u\right\rangle$ for any direction u. The following result formalizes this intuition.

Corollary 12. *For w so that $\|w\|_p > r_\alpha$, we have $\left\langle \nabla L(w),\ w\right\rangle \ge (1+\alpha)\|w\|_p \left\langle \nabla L(w),\ u_p^{\mathsf{r}}\right\rangle$.*

Proof. This follows from the convexity of L and Lemma 11: $\left\langle \nabla L(w),\ w - (1+\alpha)\|w\| u_p^{\mathsf{r}})\right\rangle \ge L(w) - L((1+\alpha)\|w\| u_p^{\mathsf{r}}) \ge 0$. □

Now we are left with the terms $\langle -\eta\nabla L(w_t),\ w_t\rangle$. For general potential ψ, the quantity $\langle -\eta\nabla L(w_t),\ w_t\rangle = \langle \nabla\psi(w_{t+1}) - \nabla\psi(w_t),\ w_t\rangle$ cannot be simplified. On the other hand, due to our choice of potential, one can invoke Lemma 2 to lower bound these quantities in terms of $\|w_{t+1}\|_p$ and $\|w_t\|_p$, and this step is detailed in Lemma 18 in Appendix B.2. Once we have established a lower bound on $\left\langle \nabla\psi(w_{t+1}),\ u_p^{\mathsf{r}}\right\rangle$, we can turn (6) entirely into a telescoping sum and unwind the above process to show that $D_\psi\left(u_p^{\mathsf{r}}, w_t/\|w_t\|_p\right)$ must converge to zero in the limit as $t \to \infty$. Putting this all together, we obtain the following result.

Theorem 13. *For a separable linear classification problem, if the regularized direction u_p^{r} exists, then with sufficiently small step size, the iterates of p-GD converge to u_p^{r} in direction:*

$$\lim_{t\to\infty} \frac{w_t}{\|w_t\|_p} = u_p^{\mathsf{r}}. \tag{7}$$

A formal proof of this theorem can be found in Appendix B.3. We note that our proof further simplifies derivations using the separability of the mirror map. The final missing piece would be the existence of the regularized direction. In general, finding the limit direction u_p^{r} would be difficult. Fortunately, we can sometimes appeal to the max-margin direction that is much easier to compute. The following result is a generalization of [Ji et al., 2020, Proposition 10] and shows that for common losses in classification, the regularized direction and the max-margin direction are the same, hence proving the existence of the former.

Proposition 14. *If we have a loss with exponential tail, e.g. $\lim_{z\to\infty} \ell(z)e^{az} = b$, then the regularized direction exists and it is equal to the max-margin direction u_p^{m}.*

The proof of this result can be found in Appendix B.5. Note that many commonly used losses in classification, e.g., logistic loss, have exponential tail.

[5] To be more precise, we want $c_t = (1+\alpha)\|w_t\|_p$; and reason behind this choice is self-evident after we present Corollary 12.

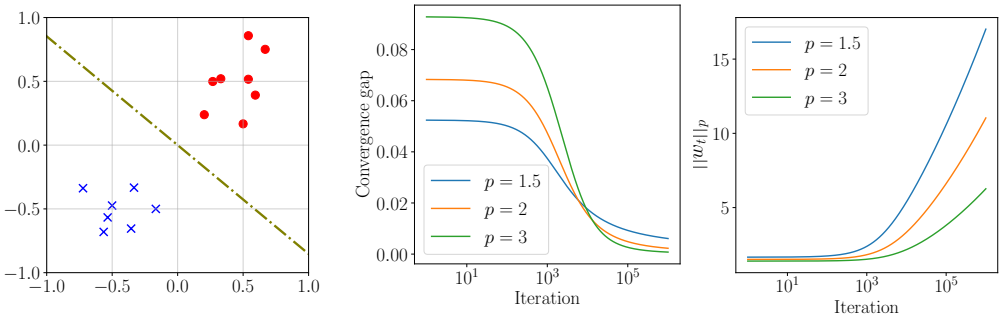

Figure 1: An example of p-GD on randomly generated data with exponential loss and $p = 1.5, 2, 3$. **(1)** The left plot is a scatter plot of the data: $\times$'s and $\bullet$'s denote the two different labels ($y_i = \pm 1$). The dotted line is the ℓ_2 max-margin classifier. For clarity, other ℓ_p max-margin classifiers are omitted from the plot. **(2)** The middle plot shows the rate which the quantity $D_\psi\left(u_p^{\mathsf{r}}, w_t / \|w_t\|_t\right)$ converges to 0. **(3)** The right plot shows how fast the p-norm of w_t growths. We can observe that the asymptotic behaviors of these plots are consistent with Corollary 17.

3.2 Asymptotic convergence rate

In this section, we characterize the rate of convergence in Theorem 13. Following the proof of Theorem 13, one can show the following result in the case of linearly separable data.

Corollary 15. *The following rate of convergence holds:*

$$D_\psi\left(u_p^{\mathsf{r}}, \frac{w_t}{\|w_t\|_p}\right) \in O\left(\|w_t\|_p^{-(p-1)}\right).$$

In order to fully understand the convergence rate, we need to characterize the asymptotic behavior of $\|w_t\|_p$. The next result precisely does that. Recall that we assumed the dataset is bounded so that $\max_i \|x_i\|_q \le C$ for $1/p + 1/q = 1$, and the max-margin direction u_p^{m} satisfies $\langle x_i,\ u_p^{\mathsf{m}} \rangle \ge \hat{\gamma}_p\, \forall i \in [n]$. Then, we have the following bound on $\|w_t\|_p$.

Lemma 16. *For exponential loss $\ell(z) = \exp(-z)$, the asymptotic growth of $\|w_t\|_p$ is contained in $\Theta(\log t)$. In particular, we have*

$$\liminf_{t\to\infty} \|w_t\|_p \ge \frac{1}{C}(\log t - p \log\log t) + O(1) \text{ and } \limsup_{t\to\infty} \|w_t\|_p \le \hat{\gamma}_p^{-1} \frac{p}{p-1} \log t.$$

The proof of this lemma can be found in Appendix C. Similar conclusions can be reached for other losses with exponential tail. Therefore, in such cases, p-GD has poly-logarithmic rate of convergence.

Corollary 17. *For exponential loss, we have convergence rate*

$$D_\psi\left(u_p^{\mathsf{r}}, \frac{w_t}{\|w_t\|_p}\right) \in O\left(\frac{1}{\log^{p-1}(t)}\right).$$

4 Experiments

In this section, we investigate the behavior and performance of p-GD for various values of p. We naturally pick $p = 2$ that corresponds to gradient descent, Because p-GD does not directly support $p = 1$ and ∞, we choose $p = 1.1$ as a surrogate for ℓ_1, and $p = 10$ as a surrogate for ℓ_∞. We also consider $p = 1.5, 3, 6$ to interpolate these points. This section will present a summary of our results; the complete experimental setup and full results can be found in Appendices E and F.

4.1 Linear classification

Visualization of the convergence of p-GD. To visualize the results of Theorem 13 and Corollary 17, we randomly generated a linearly separable set of 15 points in $\mathbb{R}^2$. We then employed p-GD on this

dataset with exponential loss $\ell(z) = \exp(-z)$ and fixed step size $\eta = 10^{-4}$. We ran this experiment for $p = 1.5, 2, 3$ and for 10^6 iterations.

In the illustrations of Figure 1, the mirror descent iterates w_t have unbounded norm and converge in direction to u_p^{m}. These results are consistent with Lemma 4 and with Theorem 13. Moreover, as predicted by Corollary 17, the exact rate of convergence for $D_\psi\left(u_p^{\mathsf{m}}, w_t/\left\|w_t\right\|_t\right)$ is poly-logarithmic with respect to the number of iterations. Corollary 17 also indicates that the convergence rate would be faster for larger p due to the larger exponent, and this is consistent with our observation in the second plot of Figure 1. Finally, in the third plot of Figure 1, the norm of the iterates w_t grows at a logarithmic rate, which is the same as the prediction by Lemma 16.

Implicit bias of p-GD in linear classification. We now verify the conclusions of Theorem 13. To this end, we recall that u_p^{m} is parallel to the SVM solution $\operatorname{argmin}_w\{\|w\|_p : \gamma(w) \geq 1\}$. Hence, we can exploit the linearity and rescale any classifier so that its margin is equal to 1. If the prediction of Theorem 13 holds, then for each fixed value of p, the classifier generated by p-GD should have the smallest ℓ_p-norm after rescaling.

To ensure that u_p^{m} are sufficiently different for different values of p, we simulate an over-parameterized setting by randomly select 15 points in $\mathbb{R}^{100}$. We used fixed step size of 10^{-4} and ran 250 thousand iterations for different p's.

Table 2 shows the results for $p = 1.1, 2, 3$ and 10; under each norm, we highlight the smallest classifier in **boldface**. Among the four classifiers we presented, p-GD with $p = 1.1$ has the smallest $\ell_{1.1}$-norm. And similar conclusions hold for $p = 2, 3, 10$. Although p-GD converges to u_p^{m} at a very slow rate, we are able to observe a very strong implicit bias of p-GD classifiers toward their respective ℓ_p geometry in a highly over-parameterized setting. This suggests we should be able to take advantage of the implicit regularization in practice and at a moderate computational cost. Due to space constraints, we defer a more complete result with additional values of p to Appendix F.1.

Table 2: Size of the linear classifiers generated by p-GD (after rescaling) in $\ell_{1.1}, \ell_2, \ell_3$ and ℓ_{10} norms.

	$\ell_{1.1}$	ℓ_2	ℓ_3	ℓ_{10}
$p = 1.1$	**5.670**	1.659	1.100	0.698
$p = 2$	6.447	**1.273**	0.710	0.393
$p = 3$	7.618	1.345	**0.691**	0.318
$p = 10$	9.086	1.520	0.742	**0.281**

4.2 Deep neural networks

Going beyond linear models, we now investigate p-GD in deep-learning settings in its impact on the structure of the learned model and potential implications on the generalization performance. As we had discussed in Section 3, **the implementation of p-GD is straightforward**; to illustrate simplicity of implementation, we provide code snippets in Appendix D. Thus, we are able to effectively experiment with the behaviors p-GD in neural network training. Specifically, we perform a set of experiments on the CIFAR-10 dataset [Krizhevsky et al., 2009]. We use the *stochastic* version of p-GD with different values of p. We choose a variety of networks: VGG [Simonyan and Zisserman, 2014], ResNet [He et al., 2016], MobileNet [Sandler et al., 2018] and RegNet [Radosavovic et al., 2020].

Implicit bias of p-GD in deep neural networks. Since the notion of margin is not well-defined in this highly nonlinear setting, we instead visualize the impacts of p-GD's implicit regularization on the histogram of weights (in absolute value) in the trained model.

In Figure 2, we report the weight histograms of ResNet-18 models trained under p-GD with $p = 1.1, 2, 3$ and 10. Depending on p, we observe interesting differences between the histograms. Note that the deep network is most sparse when $p = 1.1$ as most weights clustered around 0. Moreover, comparing the maximum weights, one can see that the case of $p = 10$ achieves the smallest value. Another observation is that the network becomes denser as p increases; for instance, there are more weights away from zero for the cases $p = 3, 10$. These overall tendencies are also observed for other deep neural networks; see Appendix F.2.

Generalization performance. We next investigate the generalization performance of networks trained with different p's. To this end, we adopt a fixed selection of hyper-parameters and then train four deep neural network models to 100% training accuracy with p-GD with different p's. As

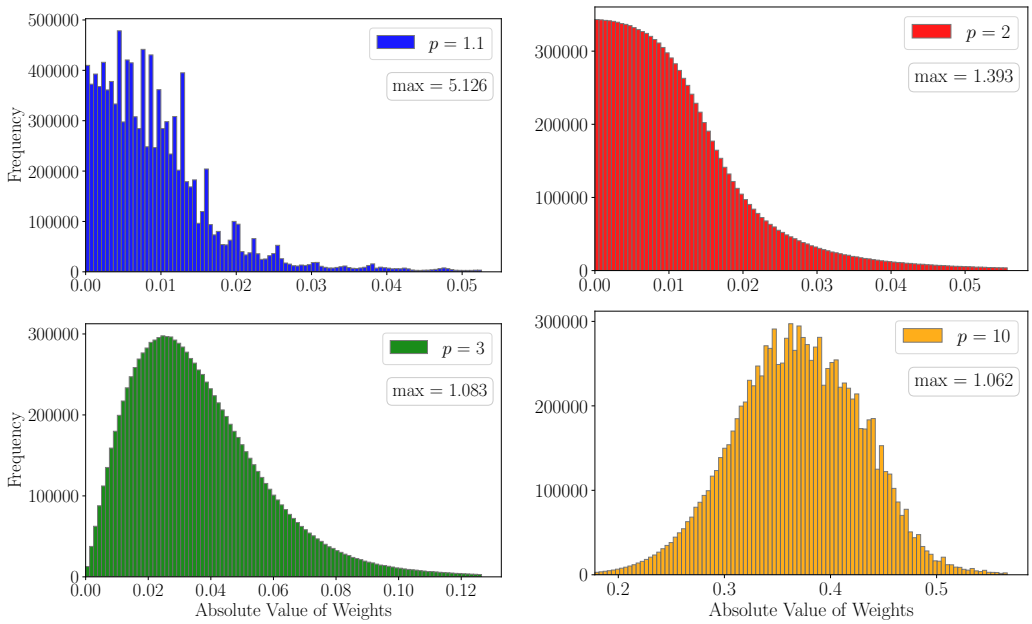

Figure 2: The histogram of weights in RESNET-18 models trained with p-GD for the CIFAR-10 dataset. For clarity, we cropped out the tails and each plot has 100 bins after cropping. The trends of these histograms reflect the implicit biases of p-GD: the distribution of $p = 1.1$ has the most number of weights around zero; and the maximum weight is smallest when $p = 10$.

Table 3: CIFAR-10 test accuracy (%) of p-GD on various deep neural networks. For each deep network and value of p, the average $\pm$ std. dev. over 5 trials are reported. And the best performing value(s) of p for each individual deep network is highlighted in **boldface**.

	VGG-11	RESNET-18	MOBILENET-V2	REGNETX-200MF
$p = 1.1$	88.19 $\pm$ .17	92.63 $\pm$ .12	91.16 $\pm$ .09	91.21 $\pm$ .18
$p = 2$ (SGD)	90.15 $\pm$ .16	**93.90** $\pm$ .14	91.97 $\pm$ .10	92.75 $\pm$ .13
$p = 3$	**90.85** $\pm$ .15	**94.01** $\pm$ .13	**93.23** $\pm$ .26	**94.07** $\pm$ .12
$p = 10$	88.78 $\pm$ .37	93.55 $\pm$ .21	92.60 $\pm$ .22	92.97 $\pm$ .16

Table 3 shows, interestingly the networks trained by p-GD with $p = 3$ consistently outperform other choices of p's; notably, for MOBILENET and REGNET, the case of $p = 3$ outperforms the others by more than 1%. Somewhat counter-intuitively, the sparser network trained by p-GD with $p = 1.1$ does not exhibit better generalization performance, but rather shows worse generalization than other values of p. Although these observations are not directly predicted by our theoretical results, we believe that they nevertheless establish an important step toward understanding generalization of overparameterized models. Due to space limit, we defer other experimental results to Appendix F.3.

IMAGENET experiments. We also perform a similar set of experiments on the IMAGENET dataset [Russakovsky et al., 2015], and these results can be found in Appendix F.4.

5 Conclusion and Future Work

In this paper, we establish an important step towards better understanding implicit bias in the classification setting, by showing that p-GD converges in direction to the generalized regularized/max-margin directions. We also run several experiments to corroborate our main findings along with the practicality of p-GD. The experiments are conducted in various settings: (i) linear models in both low and high dimensions, (ii) real-world data with highly over-parameterized nonlinear models.

We conclude this paper with several important future directions:

- Our analysis holds for $\psi(\cdot) = \|\cdot\|_p^p$, where we argued that this choice is key practical interest due to its efficient algorithmic implementations. It is mathematically interesting to generalize our analysis to other potential functions regardless of practical interest.
- As we discussed in Section 4.2, different choices of p's for our p-GD algorithm result in different generalization performance. It would be interesting to investigate this phenomenon and to develop theory that explains why certain values of p lead to better generalization performance.
- Another interesting question is to further investigate how practical techniques used in training neural networks (such as weight decay and adaptive learning rate) can affect the implicit bias and generalization properties of p-GD.

Acknowledgement

The authors thank MIT UROP students Tiffany Huang and Haimoshri Das for contributing to the experiments in Section 4.2. The authors acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing computing resources that have contributed to the results reported within this paper. This work was supported in part by MathWorks and the MIT-IBM Watson AI Lab. K.A. acknowledges support through graduate assistantship in part from the NSF grant 1846088, the ONR grant N00014-20-1-2394, and the VBFF. N.A. also acknowledges support from the Edgerton Career Development Professorship.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

Navid Azizan and Babak Hassibi. Stochastic gradient/mirror descent: Minimax optimality and implicit regularization. In *International Conference on Learning Representations*, 2019a.

Navid Azizan and Babak Hassibi. A stochastic interpretation of stochastic mirror descent: Risk-sensitive optimality. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3960–3965. IEEE, 2019b.

Navid Azizan, Sahin Lale, and Babak Hassibi. Beyond implicit regularization: Avoiding overfitting via regularizer mirror descent. In *International Conference on Machine Learning, Workshop on Overparameterization: Pitfalls & Opportunities*, 2021a.

Navid Azizan, Sahin Lale, and Babak Hassibi. Stochastic mirror descent on overparameterized nonlinear models. *IEEE Transactions on Neural Networks and Learning Systems*, 2021b.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Heinz H Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Lev M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Sébastien Bubeck. Convex optimization: algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Konstantin Donhauser, Nicolo Ruggeri, Stefan Stojanovic, and Fanny Yang. Fast rates for noisy interpolation require rethinking the effects of inductive bias. *arXiv preprint arXiv:2203.03597*, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

Claudio Gentile. The robustness of the p-norm algorithms. *Machine Learning*, 53(3):265–299, 2003.

Adam J Grove, Nick Littlestone, and Dale Schuurmans. General convergence results for linear discriminant updates. *Machine Learning*, 43(3):173–210, 2001.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798. PMLR, 2019.

Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pages 772–804. PMLR, 2021.

Ziwei Ji, Miroslav Dudík, Robert E Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pages 2109–2136. PMLR, 2020.

Ziwei Ji, Nathan Srebro, and Matus Telgarsky. Fast margin maximization via dual acceleration. In *International Conference on Machine Learning*, pages 4860–4869. PMLR, 2021.

Anatoli Juditsky, Arkadi Nemirovski, et al. First order methods for nonsmooth convex large-scale optimization, i: general purpose methods. *Optimization for Machine Learning*, 30(9):121–148, 2011.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yan Li, Caleb Ju, Ethan X Fang, and Tuo Zhao. Implicit regularization of bregman proximal point algorithm and mirror descent on separable data. *arXiv preprint arXiv:2108.06808*, 2021.

Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2019.

Youngjae Min, Kwangjun Ahn, and Navid Azizan. One-pass learning via bridging orthogonal gradient descent and recursive least-squares. In *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

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

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10428–10436, 2020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

R Tyrrell Rockafellar. *Convex analysis*. Number 28. Princeton University Press, 1970.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a regularized path to a maximum margin classifier. *The Journal of Machine Learning Research*, 5:941–973, 2004.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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

Matus Telgarsky. Margins, shrinkage, and boosting. In *International Conference on Machine Learning*, pages 307–315. PMLR, 2013.

Bohan Wang, Qi Meng, Wei Chen, and Tie-Yan Liu. The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In *International Conference on Machine Learning*, pages 10849–10858. PMLR, 2021.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. *Advances in neural information processing systems*, 30, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] All of our claims accurately reflect the results in Sections 3 and 4.
   (b) Did you describe the limitations of your work? [Yes] In Section 5, we described several directions where we can improve this work.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our paper investigates the foundational properties of mirror descent algorithms in learning; we do not believe there are any direct societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2 for the assumptions we used and the motivation behind them.
   (b) Did you include complete proofs of all theoretical results? [Yes] All proofs are included in the Appendix, and we have referenced them as we present our claims.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] They are included in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We gave an overview of our training setup in Section 4 and the full details are given in Appendix E.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We reported the standard deviation whenever multiple trials were performed.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] They are reported in Appendix E.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A] We only used public datasets.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] We did not curate any new assets.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The datasets we used are well-known; so, we did not feel it was necessary to repeat that they do not contain sensitive information.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] There were no human subjects in our work.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proofs for Section 2

A.1 Proof of Lemma 2

The following proof is adopted from [Azizan et al., 2021b]. We make several small modifications to better fit the classification setting in this paper. In particular, in classification, there is no $w \in \mathbb{R}^d$ that satisfies $L(w) = 0$.

Proof. We start with the definition of Bregman divergence:

$$D_\psi\left(w, w_{t+1}\right) = \psi(w) - \psi(w_{t+1}) - \langle \nabla\psi(w_{t+1}),\ w - w_{t+1}\rangle\,.$$

Now, we plugin the MD update rule $\nabla\psi(w_{t+1}) = \nabla\psi(w_t) - \eta\nabla L(w_t)$:

$$D_\psi\left(w, w_{t+1}\right) = \psi(w) - \psi(w_{t+1}) - \langle \nabla\psi(w_t),\ w - w_{t+1}\rangle + \eta\,\langle \nabla L(w_t),\ w - w_{t+1}\rangle\,.$$

We again invoke the definition of Bregman divergence so that:

$$\begin{aligned} D_\psi\left(w, w_{t+1}\right) &= \psi(w) - \psi(w_{t+1}) - \langle \nabla\psi(w_{t+1}),\ w - w_{t+1}\rangle\,,\\ D_\psi\left(w_{t+1}, w_t\right) &= \psi(w_{t+1}) - \psi(w_t) - \langle \nabla\psi(w_t),\ w_{t+1} - w_t\rangle\,. \end{aligned}$$

It follows that

$$\begin{aligned} D_\psi\left(w, w_{t+1}\right) &= \psi(w) - \psi(w_t) - \langle \nabla\psi(w_t),\ w - w_t\rangle \\ &\qquad + \langle \nabla\psi(w_t),\ w_{t+1} - w_t\rangle - \psi(w_{t+1}) + \psi(w_t) \\ &\qquad + \eta\,\langle \nabla L(w_t),\ w - w_{t+1}\rangle \\ &= D_\psi\left(w, w_t\right) - D_\psi\left(w_{t+1}, w_t\right) + \eta\,\langle \nabla L(w_t),\ w - w_{t+1}\rangle \end{aligned} \tag{8}$$

Next, we consider the term $\langle \nabla L(w_t),\ w - w_{t-1}\rangle$:

$$\begin{aligned} \langle \nabla L(w_t),\ w - w_{t-1}\rangle &= \langle \nabla L(w_t),\ w - w_t\rangle - \langle \nabla L(w_t),\ w_{t+1} - w_t\rangle \\ &\qquad + L(w_{t+1}) - L(w_t) - L(w_{t+1}) + L(w_t) \\ &= \langle \nabla L(w_t),\ w - w_t\rangle + D_L(w_{t+1}, w_t) - L(w_{t+1}) + L(w_t), \end{aligned} \tag{9}$$

where the last step holds because L is convex.

Combining (8) and (9) yields:

$$\begin{aligned} &D_\psi\left(w, w_t\right) \\ &= D_\psi\left(w, w_{t+1}\right) + D_\psi\left(w_{t+1}, w_t\right) - \eta\big(\,\langle \nabla L(w_t),\ w - w_t\rangle + D_L(w_{t+1}, w_t) - L(w_{t+1}) + L(w_t)\big) \\ &= D_\psi\left(w, w_{t+1}\right) + D_{\psi-\eta L}\left(w_{t+1}, w_t\right) - \eta\,\langle \nabla L(w_t),\ w - w_t\rangle + \eta L(w_{t+1}) - \eta L(w_t), \end{aligned}$$

where in the last step, we note that Bregman divergence is additive in its potential. This gives us (1b). And for (1a), we use the definition of Bregman divergence again, i.e. $D_L(w, w_t) = L(w) - L(w_t) - \langle \nabla L(w_t),\ w - w_t\rangle$:

$$\begin{aligned} D_\psi\left(w, w_t\right) &= D_\psi\left(w, w_{t+1}\right) + D_{\psi-\eta L}\left(w_{t+1}, w_t\right) - \eta\,\langle \nabla L(w_t),\ w - w_t\rangle \\ &\qquad + \eta L(w) - \eta L(w_t) + \eta L(w_{t+1}) - \eta L(w) \\ &= D_\psi\left(w, w_{t+1}\right) + D_{\psi-\eta L}\left(w_{t+1}, w_t\right) + \eta D_L(w, w_t) - \eta L(w) + \eta L(w_{t+1}) \end{aligned}$$

□

A.2 Proof of Lemma 3

Proof. This is an application of Lemma 2 with $w = w_t$:

$$\begin{aligned} 0 &= D_\psi(w_t, w_{t+1}) + D_{\psi-\eta L}(w_{t+1}, w_t) - \eta L(w_t) + \eta L(w_{t+1}) \\ \implies \eta L(w_t) &= D_\psi(w_t, w_{t+1}) + D_{\psi-\eta L}(w_{t+1}, w_t) + \eta L(w_{t+1}) \geq \eta L(w_{t+1}) \end{aligned}$$

where we used the fact that Bregman divergence with a convex potential function is non-negative. □

A.3 Proof of Lemma 4

Proof. By Lemma 3, $L(w_t)$ is decreasing with respect to t, therefore the limit exists. Suppose the contrary that $\lim_{t\to\infty} L(w_t) = \varepsilon > 0$. Since the data is separable, we can pick w so that $L(w) \le \varepsilon/2$. Applying Lemma 2, the following holds for all t:

$$\begin{aligned} D_\psi(w, w_{t+1}) &= D_\psi(w, w_t) - D_{\psi-\eta L}(w_{t+1}, w_t) - \eta D_L(w, w_t) + \eta L(w) - \eta L(w_{t+1}) \\ &\le D_\psi(w, w_t) + \eta\varepsilon/2 - \eta\varepsilon = D_\psi(w, w_t) - \eta\varepsilon/2 \end{aligned}$$

Hence, $D_\psi(w, w_t) \le D_\psi(w, w_0) - t\eta\varepsilon/2$. This implies that $\limsup_{t\to\infty} D_\psi(w, w_t) = -\infty$, contradiction. □

B Proofs for Section 3

B.1 Proof of Lemma 11

Proof. Let $\bar\gamma$ be the margin of u_p^{r}. Under separability, we know $\bar\gamma > 0$. Recall the definition of the regularization path. There exists sufficiently large r_α so that

$$\left\| \frac{\bar w_p(\|w\|_p)}{\|w\|_p} - u_p^{\mathsf{r}} \right\|_p \le \frac{\alpha\bar\gamma}{C}$$

whenever $\|w\|_p \ge r_\alpha$. Recall the definition that $C = \max_{i=1,\dots,n} \|x_i\|_q$, $1/p + 1/q = 1$. Then, for all $i \in [n]$, we have

$$\begin{aligned} y_i \left\langle \bar w(\|w\|_p),\ x_i \right\rangle &= y_i \left\langle \bar w(\|w\|_p) - \|w\|_p u_p^{\mathsf{r}},\ x_i \right\rangle + y_i \left\langle \|w\|_p u_p^{\mathsf{r}},\ x_i \right\rangle \\ &\le \alpha\bar\gamma \|w\|_p \|x_i\|_q / C + y_i \left\langle \|w\|_p u_p^{\mathsf{r}},\ x_i \right\rangle \\ &\le \alpha\bar\gamma \|w\|_p + y_i \left\langle \|w\|_p u_p^{\mathsf{r}},\ x_i \right\rangle \\ &\le y_i \left\langle (1+\alpha)\|w\|_p u_p^{\mathsf{r}},\ x_i \right\rangle \end{aligned}$$

Since the loss L is decreasing, we have

$$L((1+\alpha)\|w\|_p u_p^{\mathsf{r}}) \le L(\bar w(\|w\|_p)) \le L(w).$$

□

B.2 Lower bounding the mirror descent updates

Lemma 18. *For $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^p$ with $p > 1$, the mirror descent update satisfies the following inequality:*

$$\frac{p-1}{p}\|w_{t+1}\|_p^p - \frac{p-1}{p}\|w_t\|_p^p + \eta L(w_{t+1}) - \eta L(w_t) \le \langle -\eta\nabla L(w_t),\ w_t\rangle \tag{10}$$

Proof. This result follows from Lemma 2 with $w = 0$:

$$\begin{aligned} D_\psi(0, w_t) &= D_\psi(0, w_{t+1}) + D_{\psi-\eta L}(w_{t+1}, w_t) + \eta D_L(0, w_t) + \eta L(w_{t+1}) - \eta L(0) \\ &\ge D_\psi(0, w_{t+1}) + \eta D_L(0, w_t) + \eta L(w_{t+1}) - \eta L(0) \\ &= D_\psi(0, w_{t+1}) + \eta(L(0) - L(w_t) - \langle \nabla L(w_t),\ -w_t\rangle) + \eta L(w_{t+1}) - \eta L(0) \\ &= D_\psi(0, w_{t+1}) + \eta\langle \nabla L(w_t),\ w_t\rangle + \eta L(w_{t+1}) - \eta L(w_t) \end{aligned}$$

Rearranging the terms yields

$$D_\psi(0, w_{t+1}) - D_\psi(0, w_t) + \eta L(w_{t+1}) - \eta L(w_t) \le \langle -\eta\nabla L(w_t),\ w_t\rangle$$

We conclude the proof by noting that for any $w \in \mathbb{R}^d$,

$$D_\psi(0, w) = \psi(0) - \psi(w) - \langle \nabla\psi(w),\ -w\rangle = \langle \nabla\psi(w),\ w\rangle - \psi(w) = \frac{p-1}{p}\|w\|_p^p$$

□

B.3 Proof of Theorem 13

Proof. Consider arbitrary $\alpha \in (0,1)$ and define r_α according to Lemma 11. Since $\lim_{t\to\infty} \|w_t\|_p = \infty$, we can find t_0 so that $\|w_t\|_p > \max(1, r_\alpha)$ for all $t \geq t_0$. Let $c_t = (1+\alpha)\|w_t\|_p$.

We list some properties about $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^p$ that will be useful in our analysis:

$$\begin{aligned}
\nabla\psi(w) &= (\operatorname{sign}(w_1)|w_1|^{p-1}, \cdots, \operatorname{sign}(w_d)|w_d|^{p-1}) \\
\langle \nabla\psi(w),\ w\rangle &= \operatorname{sign}(w_1)w_1|w_1|^{p-1} + \cdots + \operatorname{sign}(w_d)w_d|w_d|^{p-1} = \|w\|_p^p \\
\|\nabla\psi(w)\|_q &= \|w\|_p^{p-1},\ \text{for } 1/p + 1/q = 1 \\
D_\psi(cw, cw') &= |c|^p D_\psi(w, w') \quad \forall c \in \mathbb{R}.
\end{aligned}$$

Substitute $w = c_t u_p^{\mathsf{r}}$ into Lemma 2, we get

$$D_\psi\left(c_t u_p^{\mathsf{r}}, w_{t+1}\right) \leq D_\psi\left(c_t u_p^{\mathsf{r}}, w_t\right) + \eta\left\langle \nabla L(w_t),\ c_t u_p^{\mathsf{r}} - w_t\right\rangle - \eta L(w_{t+1}) + \eta L(w_t).$$

By Corollary 12, we have $\left\langle \nabla L(w_t),\ c_t u_p^{\mathsf{r}} - w_t\right\rangle \leq 0$. Therefore,

$$D_\psi\left(c_t u_p^{\mathsf{r}}, w_{t+1}\right) \leq D_\psi\left(c_t u_p^{\mathsf{r}}, w_t\right) - \eta L(w_{t+1}) + \eta L(w_t).$$

It follows that

$$\begin{aligned}
&D_\psi\left(c_{t+1} u_p^{\mathsf{r}}, w_{t+1}\right) \\
&\leq D_\psi\left(c_t u_p^{\mathsf{r}}, w_t\right) - \eta L(w_{t+1}) + \eta L(w_t) + D_\psi\left(c_{t+1} u_p^{\mathsf{r}}, w_{t+1}\right) - D_\psi\left(c_t u_p^{\mathsf{r}}, w_{t+1}\right) \\
&= D_\psi\left(c_t u_p^{\mathsf{r}}, w_t\right) - \eta L(w_{t+1}) + \eta L(w_t) + \psi(c_{t+1} u_p^{\mathsf{r}}) - \psi(c_t u_p^{\mathsf{r}}) - \left\langle \nabla\psi(w_{t+1}),\ (c_{t+1} - c_t) u_p^{\mathsf{r}}\right\rangle
\end{aligned}$$

Summing over $t = t_0, \ldots, T-1$ gives us

$$\begin{aligned}
D_\psi\left(c_T u_p^{\mathsf{r}}, w_T\right) \leq D_\psi\left(c_{t_0} u_p^{\mathsf{r}}, w_{t_0}\right) &- \eta L(w_{t_0}) + \eta L(w_T) + \psi(c_T u_p^{\mathsf{r}}) - \psi(c_{t_0} u_p^{\mathsf{r}}) \\
&- \sum_{t=t_0}^{T-1} \left\langle \nabla\psi(w_{t+1}),\ (c_{t+1} - c_t) u_p^{\mathsf{r}}\right\rangle
\end{aligned} \tag{11}$$

Now we want to establish a lower bound on the last term of (11). To do so, we inspect the change in $\nabla\psi(w_t)$ from each successive mirror descent update:

$$\begin{aligned}
&\left\langle \nabla\psi(w_{t+1}) - \nabla\psi(w_t),\ u_p^{\mathsf{r}}\right\rangle && \text{(12a)} \\
&= \left\langle -\eta\nabla L(w_t),\ u_p^{\mathsf{r}}\right\rangle && \text{(12b)} \\
&\geq \frac{1}{(1+\alpha)\|w_t\|_p}\left\langle -\eta\nabla L(w_t),\ w_t\right\rangle && \text{(12c)} \\
&\geq \frac{1}{(1+\alpha)\|w_t\|_p}\left(\frac{p-1}{p}\|w_{t+1}\|_p^p - \frac{p-1}{p}\|w_t\|_p^p + \eta L(w_{t+1}) - \eta L(w_t)\right) && \text{(12d)} \\
&\geq \frac{1}{(1+\alpha)\|w_t\|_p}\left(\frac{p-1}{p}\|w_{t+1}\|_p^p - \frac{p-1}{p}\|w_t\|_p^p\right) + \eta L(w_{t+1}) - \eta L(w_t) && \text{(12e)}
\end{aligned}$$

where we applied Corollary 12 on (12c) and Lemma 18 on (12d).

Now we bound (12e). We claim the following identity and defer its derivation to Section B.4.

$$\frac{p-1}{p}(\|w_{t+1}\|_p^p - \|w_t\|_p^p) \geq (\|w_{t+1}\|_p^{p-1} - \|w_t\|_p^{p-1})\|w_t\|_p. \tag{13}$$

We are left with

$$\left\langle \nabla\psi(w_{t+1}) - \nabla\psi(w_t),\ u_p^{\mathsf{r}}\right\rangle \geq \frac{\|w_{t+1}\|_p^{p-1} - \|w_t\|_p^{p-1}}{1+\alpha} + \eta L(w_{t+1}) - \eta L(w_t).$$

Summing over $t = t_0, \ldots, T-1$ gives us

$$\left\langle \nabla\psi(w_T) - \nabla\psi(w_{t_0}),\ u_p^{\mathsf{r}} \right\rangle \geq \frac{\|w_T\|_p^{p-1} - \|w_{t_0}\|_p^{p-1}}{1+\alpha} + \eta L(w_T) - \eta L(w_{t_0}). \tag{14}$$

With (14), we can bound the last term of (11) as follows:

$$\begin{aligned}
\sum_{t=t_0}^{T-1} \left\langle \nabla\psi(w_{t+1}),\ (c_{t+1} - c_t)u_p^{\mathsf{r}} \right\rangle &\geq \sum_{t=t_0+1}^{T} \frac{\|w_t\|_p^{p-1} + O(1)}{1+\alpha}(c_t - c_{t-1}) \\
&= \sum_{t=t_0+1}^{T} (\|w_t\|_p^{p-1} + O(1))(\|w_t\|_p - \|w_{t-1}\|_p) \\
&\geq \sum_{t=t_0+1}^{T} \frac{1}{p}(\|w_t\|_p^p - \|w_{t-1}\|_p^p) + O(1) \cdot (\|w_T\|_p - \|w_{t_0}\|_p) \\
&= \frac{1}{p}\|w_T\|_p^p + O(\|w_T\|_p)
\end{aligned} \tag{15}$$

where we defer the computation on the last inequality to Section B.4.

We now apply the inequality in (15) to (11). Note that $\psi(c_T u_p^{\mathsf{r}}) = \frac{1}{p}(1+\alpha)^p \|w_T\|_p^p$. We now have the following:

$$D_\psi\left((1+\alpha)\|w_T\|_p u_p^{\mathsf{r}}, w_T\right) \leq \frac{1}{p}\|w_T\|_p^p ((1+\alpha)^p - 1) + O(\|w_T\|_p).$$

After applying homogeneity of Bregman divergence, and recalling that $\alpha = \frac{\varepsilon}{1-\varepsilon}$, we have

$$D_\psi\left(u_p^{\mathsf{r}}, (1-\varepsilon)\frac{w_T}{\|w_T\|_p}\right) \leq \frac{\frac{1}{p}\|w_T\|_p^p (1 - (1-\varepsilon)^p)}{\|w_T\|_p^p} + o(1).$$

Let $\tilde{w}_T = \frac{w_T}{\|w_T\|_p}$. We note that Bregman divergence in fact satisfies the law of cosines:

Lemma 19 (Law of Cosines)**.**

$$D_\psi(w, w') = D_\psi(w, w'') + D_\psi(w'', w') - \langle \nabla\psi(w') - \nabla\psi(w''),\ w - w'' \rangle$$

Therefore,

$$\begin{aligned}
D_\psi\left(u_p^{\mathsf{r}}, \tilde{w}_T\right) &\leq \frac{\frac{1}{p}\|w_T\|_p^p (1 - (1-\varepsilon)^p)}{\|w_T\|_p^p} + D_\psi((1-\varepsilon)\tilde{w}_T, \tilde{w}_T) \\
&\quad - \left\langle \nabla\psi(\tilde{w}_T) - \nabla\psi((1-\varepsilon)\tilde{w}_T),\ u_p^{\mathsf{r}} - (1-\varepsilon)\tilde{w}_T \right\rangle + o(1) \\
&\leq \frac{1}{p}(1 - (1-\varepsilon)^p) + \frac{1}{p}((1-\varepsilon)^p - 1) + \varepsilon + 2d^{1/p}(1 - (1-\varepsilon)^p) + o(1)
\end{aligned} \tag{16}$$

And we defer the computation for the last inequality to Section B.4. Taking the limit as $T \to \infty$ and $\varepsilon \to 0$, we have that

$$\limsup_{T\to\infty} D_\psi\left(u_p^{\mathsf{r}}, \frac{w_T}{\|w_T\|_p}\right) \leq \varepsilon + 2d^{1/p}(1 - (1-\varepsilon)^p) \tag{17}$$

Note that the RHS vanishes in the limit as $\varepsilon \to 0$. Since the choice of ε is arbitrary, we have $w_T / \|w_T\|_p \to u_p^{\mathsf{r}}$ as $T \to \infty$.

□

B.4 Auxiliary Computation for Section B.3

To show (13), we claim that for $\delta \geq -1$ and $p > 1$, we have

$$\frac{p-1}{p}((1+\delta)^p - 1) \geq (1+\delta)^{p-1} - 1.$$

Note that we equality when $\delta = 0$, and now we consider the first derivative:

$$\frac{d}{d\delta}\left\{\frac{p-1}{p}((1+\delta)^p - 1) - (1+\delta)^{p-1} + 1\right\} = (p-1)\delta(1+\delta)^{p-2},$$

which is negative when $\delta \in [-1, 0)$ and positive when $\delta > 0$, so this identity holds. Now, (13) follows from setting $\delta = (\|w_{t+1}\|_p - \|w_t\|_p)/\|w_t\|_p$ and then multiplying by $\|w_t\|_p^p$ on both sides.

To finish showing (15), we claim that for $\delta \geq -1$ and $p > 1$, we have

$$\frac{1}{p}((1+\delta)^p - 1) \leq \delta(1+\delta)^{p-1}.$$

Note that we equality when $\delta = 0$, and now we consider the first derivative:

$$\frac{d}{d\delta}\left\{\frac{1}{p}((1+\delta)^p - 1) - \delta(1+\delta)^{p-1}\right\} = -(p-1)\delta(1+\delta)^{p-2},$$

which is positive when $\delta \in [-1, 0)$ and negative when $\delta > 0$, so this identity holds. Now, the last inequality of (15) follows by setting $\delta = (\|w_t\|_p - \|w_{t-1}\|_p)/\|w_{t-1}\|_p$ and then multiply by $\|w_t\|_p^p$ on both sides.

Finally, we simplify the RHS of (16) by taking advantage of the fact that $\tilde{w}_T$ is normalized:

$$\begin{aligned} D_\psi\left((1-\varepsilon)\tilde{w}_T, \tilde{w}_T\right) &= (1-\varepsilon)^p\psi(\tilde{w}_T) - \psi(\tilde{w}_T) + \langle \nabla\psi(\tilde{w}_T),\ \varepsilon\tilde{w}_T\rangle \\ &= \frac{1}{p}((1-\varepsilon)^p - 1) + \varepsilon \end{aligned}$$

$$\begin{aligned} &\left|\left\langle \nabla\psi(\tilde{w}_T) - \nabla\psi((1-\varepsilon)\tilde{w}_T),\ u_p^{\mathsf{r}} - (1-\varepsilon)\tilde{w}_T\right\rangle\right| \\ =\ &\left|\left\langle (1-(1-\varepsilon)^p)\nabla\psi(\tilde{w}_T),\ u_p^{\mathsf{r}} - (1-\varepsilon)\tilde{w}_T\right\rangle\right| \\ \leq\ &(1-(1-\varepsilon)^p)\left\|\nabla\psi(\tilde{w_T})\right\|_q \cdot \left\|u_p^{\mathsf{r}} - (1-\varepsilon)\tilde{w}_T\right\|_p \\ =\ &(1-(1-\varepsilon)^p)\left\|\tilde{w_T}\right\|_p^{p-1} \cdot \left\|u_p^{\mathsf{r}} - (1-\varepsilon)\tilde{w}_T\right\|_p \\ \leq\ &2d^{1/p}(1-(1-\varepsilon)^p) \end{aligned}$$

B.5 Proof of Theorem 14

Proof. We first show that u_p^{m} is unique. Suppose the contrary that there are two distinct unit p-norm vectors $u_1 \neq u_2$ both achieving the maximum-margin $\hat{\gamma}_p$. Then $u_3 = (u_1 + u_2)/2$ satisfies

$$\forall i, y_i\langle u_3,\ x_i\rangle = \frac{1}{2}y_i\langle u_1,\ x_i\rangle + \frac{1}{2}y_i\langle u_2,\ x_i\rangle \geq \hat{\gamma}_p$$

Therefore, u_3 has margin of at least $\hat{\gamma}_p$. Since $\|\cdot\|_p$ is strictly convex, we must have $\|u_3\|_p < 1$. Therefore, the margin of $u_3/\|u_3\|_p$ is strictly greater than $\hat{\gamma}_p$, contradiction.

Define $\beta > 0$ so that $\ell(z)e^{az} \in [b/2, 2b]$ for $z = B\hat{\gamma}_p/2$ and whenever $B > \beta$. Note that

$$L(Bu_p^{\mathsf{m}}) = \sum_{i=1}^n \ell(y_i\langle Bu_p^{\mathsf{m}},\ x_i\rangle) \leq n \cdot \ell(B\hat{\gamma}_p) \leq 2bn \cdot \exp(-aB\hat{\gamma}_p)$$

Suppose the contrary that the regularized direction does not converge to u_p^{m}, then there must exist $\hat{\gamma}_p/2 > \varepsilon > 0$ so that there are arbitrarily large values of B satisfying

$$\min_{i=1,\ldots,n} y_i\left\langle \frac{\bar{w}(B)}{B},\ x_i\right\rangle \leq \hat{\gamma}_p - \varepsilon.$$

And this implies

$$L(\bar{w}(B)) \geq \ell(B(\hat{\gamma}_p - \varepsilon)) \geq \frac{b}{2}\exp(-aB\hat{\gamma}_p)\exp(aB\varepsilon)$$

Then, for sufficiently large $B > \beta$, we have $\exp(aB\varepsilon) > 4n \Rightarrow L(\bar{w}(B)) > L(Bu_p^{\mathsf{m}})$, contradiction. Therefore, the regularized direction exists and $u_p^{\mathsf{r}} = u_p^{\mathsf{m}}$. □

B.6 Simpler proof of Theorem 13

For potential function $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^p$, we can avoid most calculations involving (11) by directly computing for Bregman divergence. However, we want to note that this approach is less general, and does not highlight the role of u_p^{r} as clearly.

Proof. Consider arbitrary $\alpha \in (0,1)$. Since $\lim_{t\to\infty}\|w_t\|_p = \infty$, we can find t_0 so that $\|w_t\| > \max(1, r_\alpha)$ for all $t \geq t_0$. For $T > t_0$, define $\tilde{w}_T = \frac{w_T}{\|w_T\|_p}$.

We can perform the following manipulation on Bregman divergence:

$$\begin{aligned} D_\psi\left(u_p^{\mathsf{r}}, \tilde{w}_T\right) &= \psi(u_p^{\mathsf{r}}) - \psi\left(\tilde{w}_T\right) - \left\langle \nabla\psi\left(\tilde{w}_T\right),\, u_p^{\mathsf{r}} - \tilde{w}_T\right\rangle \\ &= \psi(u_p^{\mathsf{r}}) - \psi\left(\tilde{w}_T\right) + \langle \nabla\psi(\tilde{w}_T),\, \tilde{w}_T\rangle - \left\langle \nabla\psi(\tilde{w}_T),\, u_p^{\mathsf{r}}\right\rangle \\ &= \frac{1}{p}\left\|u_p^{\mathsf{r}}\right\|_p^p - \frac{1}{p}\|\tilde{w}_T\|_p^p + \|\tilde{w}_T\|_p^p - \left\langle \nabla\psi(\tilde{w}_T),\, u_p^{\mathsf{r}}\right\rangle \\ &= 1 - \left\langle \nabla\psi(\tilde{w}_T),\, u_p^{\mathsf{r}}\right\rangle \end{aligned} \tag{18}$$

We divide both sides of (14) by $\|w_T\|$ and then taking the limit as $T \to \infty$ yields

$$\liminf_{T\to\infty} \frac{1}{\|w_T\|_p^{p-1}}\left\langle \nabla\psi(w_T),\, u_p^{\mathsf{r}}\right\rangle \geq \frac{1}{1+\alpha}. \tag{19}$$

Now, substituting (19) into (18) results in

$$\begin{aligned} \limsup_{T\to\infty} D_\psi\left(u_p^{\mathsf{r}}, \frac{w_T}{\|w_T\|_p}\right) &= 1 - \liminf_{T\to\infty}\left\langle \nabla\psi\left(\frac{w_T}{\|w_T\|_p}\right),\, u_p^{\mathsf{r}}\right\rangle \\ &= 1 - \liminf_{T\to\infty}\frac{1}{\|w_T\|_p^{p-1}}\left\langle \nabla\psi(w_T),\, u_p^{\mathsf{r}}\right\rangle \\ &\leq 1 - \frac{1}{1+\alpha} < \alpha \end{aligned}$$

Since the value of α is arbitrary, we can conclude that

$$\lim_{T\to\infty} D_\psi\left(u_p^{\mathsf{r}}, \frac{w_T}{\|w_T\|_p}\right) = 0.$$

□

C Proofs for Section 3.2

C.1 Proof of Corollary 15

Proof. This is an immediate consequence of (18) and (19). □

C.2 Proof of Lemma 16

For the following proof, we assume without loss of generality that $y_i = 1$ by replacing every instance of $(x_i, -1)$ with $(-x_i, 1)$.

Proof. For the upper bound, we consider a reference vector $w^\star = \hat{\gamma}_p^{-1} u_p^{\mathsf{m}}$. By the definition of the max-margin direction, the margin of $w^\star$ is 1 and $\|w^\star\|_p = \hat{\gamma}_p^{-1}$. From Lemma 2, we have

$$D_\psi(w^\star \log T, w_t) = D_\psi(w^\star \log T, w_{t+1}) + D_{\psi - \eta L}(w_{t+1}, w_t) - \langle \nabla L(w_t),\ w^\star \log T - w_t \rangle \\ - \eta L(w_t) + \eta L(w_{t+1}).$$

We first bound the quantity $\langle \nabla L(w_t),\ w^\star \log T - w_t \rangle$ by expanding the definition of exponential loss:

$$\begin{aligned}
&\langle \nabla L(w_t),\ w^\star \log T - w_t \rangle \\
&= \sum_{i=1}^n \langle \nabla \exp(-\langle w_t,\ x_i \rangle),\ w^\star \log T - w_t \rangle \\
&= \sum_{i=1}^n \langle \exp(-\langle w_t,\ x_i \rangle) x_i,\ w_t - w^\star \log T \rangle \\
&= \sum_{i=1}^n \exp(-\langle w^\star \log T,\ x_i \rangle) \exp(-\langle w_t - w^\star \log T,\ x_i \rangle) \langle x_i,\ w_t - w^\star \log T \rangle \\
&\le \sum_{i=1}^n \frac{1}{T} \cdot \frac{1}{e} = \frac{n}{eT}
\end{aligned}$$

where the last line follows from the definition of $w^\star$ and the fact that for any $x \in \mathbb{R}$, we have $e^{-x} x \le 1/e$. It follows that

$$D_\psi(w^\star \log T, w_t) \ge D_\psi(w^\star \log T, w_{t+1}) - \frac{n}{eT} - \eta L(w_t) + \eta L(w_{t+1}).$$

Summing over $t = 0, \dots, T-1$ gives us

$$D_\psi(w^\star \log T, w_0) \ge D_\psi(w^\star \log T, w_T) - \frac{n}{e} - \eta L(w_0) + \eta L(w_T).$$

Since Bregman divergence with respect to the pth power of ℓ_p-norm is homogeneous, we can divide by a factor of $\log^p T$ on both sides:

$$D_\psi\left(w^\star, \frac{w_0}{\log T}\right) \ge D_\psi\left(w^\star, \frac{w_T}{\log T}\right) - o(1). \tag{20}$$

As $T \to \infty$, the left-hand side converges to $D_\psi(w^\star, 0) = \psi(w^\star) = \frac{1}{p}\hat{\gamma}_p^{-p}$. Let $\tilde{w} = w_T / \log T$, we expand the right-hand side as

$$\begin{aligned}
D_\psi(w^\star, \tilde{w}) &= \psi(w^\star) - \psi(\tilde{w}) - \langle \nabla \psi(\tilde{w}),\ w^\star - \tilde{w} \rangle \\
&= \frac{1}{p}\hat{\gamma}_p^{-p} + \frac{p-1}{p}\|\tilde{w}\|_p^p - \langle \nabla \psi(\tilde{w}),\ w^\star \rangle \\
&\ge \frac{1}{p}\hat{\gamma}_p^{-p} + \frac{p-1}{p}\|\tilde{w}\|_p^p - \hat{\gamma}_p^{-1}\|\nabla \psi(\tilde{w})\|_q
\end{aligned}$$

for $1/p + 1/q = 1$. Recall that $\psi = \frac{1}{p}\|\cdot\|_p^p$ has the following nice properties:

$$\begin{aligned}
\nabla \psi(w) &= (\operatorname{sign}(w_1)|w_1|^{p-1}, \cdots, \operatorname{sign}(w_d)|w_d|^{p-1}) \\
\langle \nabla \psi(w),\ w \rangle &= \operatorname{sign}(w_1) w_1 |w_1|^{p-1} + \cdots + \operatorname{sign}(w_d) w_d |w_d|^{p-1} = \|w\|_p^p
\end{aligned}$$

So, we can further simplify $\|\nabla \psi(\tilde{w})\|_q$:

$$\begin{aligned}
\|\nabla \psi(\tilde{w})\|_q &= \left(\sum_{i=1}^d |\tilde{w}_i|^{(p-1)q}\right)^{1/q} \\
&= \left(\sum_{i=1}^d |\tilde{w}_i|^p\right)^{1/q} \\
&= \|\tilde{w}\|_p^{p/q} = \|\tilde{w}\|_p^{p-1},
\end{aligned}$$

where we note that because $1/p + 1/q = 1$, we also have $pq = p + q$ and $1 + p/q = p$.

Now, we have

$$D_\psi\left(w^\star, \tilde{w}\right) \geq \frac{1}{p}\hat{\gamma}_p^{-p} + \frac{p-1}{p}\left\|\tilde{w}\right\|_p^p - \hat{\gamma}_p^{-1}\left\|\tilde{w}\right\|_p^{p-1}$$

If $\left\|w_T/\log T\right\|_p > \hat{\gamma}_p^{-1} \cdot \frac{p}{p-1}$ for arbitrarily large T, then $D_\psi\left(w^\star, w_T/\log T\right) > \frac{1}{p}\hat{\gamma}_p^{-p}$ for those T. This in turn contradicts inequality (20). Therefore, we must have

$$\limsup_{T\to\infty}\left\|w_T\right\|_p \leq \hat{\gamma}_p^{-1}\frac{p}{p-1}\log T.$$

Now we can turn our attention to the lower bound. Let $m_t = \gamma(w_t)$ be the margin of the mirror descent iterates. Then,

$$L(w_t) = \frac{1}{n}\sum_{i=1}^n \exp(-\left\langle w_t,\ x_i\right\rangle) \geq \frac{1}{n}\exp(-m_t).$$

Due to Lemma 4, we also know that $m_t \xrightarrow{t\to\infty} \infty$.

By the definition of the max-margin direction, we know that $\gamma(\left\|w_t\right\|_p u_p^{\mathsf{m}}) \geq m_t$. Then by linearity of margin, there exists $w^\star$ so that $\gamma(w^\star) \geq (1 + \frac{2n}{m_t})m_t$ and $\left\|w^\star\right\|_p \leq (1 + \frac{2n}{m_t})\left\|w_t\right\|_p$. It follows that

$$L(w^\star) = \frac{1}{n}\sum_{i=1}^n \exp(-\left\langle w_t,\ x_i\right\rangle) \leq \exp(-\gamma(w^\star)) = \frac{1}{2n}\exp(-m_t).$$

Under the assumption that the step size η is sufficiently small so that $\psi - \eta L$ is convex on the iterates, we can apply the convergence rate of mirror descent [Lu et al., 2018, Theorem 3.1]:

$$L(w_t) - L(w^\star) \leq \frac{1}{\eta t}D_\psi\left(w^\star, w_0\right)$$

From our choice of $w^\star$, we have

$$\begin{aligned}\frac{1}{2n}\exp(-m_t) &\leq \frac{1}{\eta t}D_\psi\left(w^\star, w_0\right)\\ &= \frac{1}{\eta t}(\psi(w^\star) - \psi(w_0) - \left\langle\nabla\psi(w_0),\ w^\star - w_0\right\rangle)\end{aligned}$$

After dropping the lower order terms and recall the upper bounds on $\left\|w^\star\right\|_p$ and $\left\|w_t\right\|_p$, we have

$$\frac{1}{2n}\exp(-m_t) \leq O(1)\cdot\frac{1}{\eta t}\cdot\frac{1}{p}\left(1 + \frac{\log(2n)}{m_t}\right)^p\left(\hat{\gamma}_p^{-1}\frac{p}{p-1}\log t\right)^p$$

Since m_t is unbounded, the quantity $1 + \frac{\log(2n)}{m_t}$ is upper bounded by a constant. Taking the logarithm on both sides yields

$$m_t \geq \log t - p\log\log t + O(1)$$

Finally, we use the definition of margin to conclude that $m_t \leq \left\langle w_t,\ x_i\right\rangle \leq C\cdot\left\|w_t\right\|_p$. Therefore,

$$\left\|w_t\right\|_p \geq \frac{1}{C}(\log t - p\log\log t) + O(1).$$

□

D Practicality of p-GD

To illustrate that p-GD can be easily implemented, we show a proof-of-concept implementation in PyTorch. This implementation can directly replace existing optimizers and thus require only minor changes to any existing training code.

We also note that while the p-GD update step requires more arithmetic operations than a standard gradient descent update, this does not significantly impact the total runtime because differentiation is the most computationally intense step. We observed from our experiments that training with p-GD is approximate 10% slower than with PyTorch's `optim.SGD` (in the same number of epochs),[6] and we believe that this gap can be closed with a more optimized code.

Listing 1: Sample PyTorch implementation of p-GD

```
import torch
from torch.optim import Optimizer

class pnormSGD(Optimizer):
    def __init__(self, params, lr=0.01, pnorm=2.0):
        if not 0.0 <= lr:
            raise ValueError("Invalid learning rate: {}".format(lr))
        # p-norm must be strictly greater than 1
        if not 1.01 <= pnorm:
            raise ValueError("Invalid p-norm value: {}".format(pnorm))

        defaults = dict(lr=lr, pnorm=pnorm)
        super(pnormSGD, self).__init__(params, defaults)

    def __setstate__(self, state):
        super(pnormSGD, self).__setstate__(state)

    def step(self, closure=None):
        loss = None
        if closure is not None:
            with torch.enable_grad():
                loss = closure()

        for group in self.param_groups:
            lr = group["lr"]
            pnorm = group["pnorm"]

            for param in group["params"]:
                if param.grad is None:
                    continue

                x = param.data
                dx = param.grad.data

                # \ell_p^p potential function
                update = torch.pow(torch.abs(x), pnorm-1) * \
                                torch.sign(x) -  lr * dx
                param.data =  torch.sign(update) * \
                            torch.pow(torch.abs(update), 1/(pnorm-1))

        return loss
```

[6] This measurement may not be very accurate because we were using shared computing resources.

E Experimental details

E.1 Linear classification

Here, we describe the details behind our experiments from Section 4.1. First, we note that we can absorb the labels y_i by replacing (x_i, y_i) with $(y_i x_i, 1)$. This way, we can choose points with the same $+1$ label.

For the $\mathbb{R}^2$ experiment, we first select three points $(\frac{1}{6}, \frac{1}{2}), (\frac{1}{2}, \frac{1}{6})$ and $(\frac{1}{3}, \frac{1}{3})$ so that the maximum margin direction is approximately $\frac{1}{\sqrt{2}}(1, 1)$. Then we sample 12 additional points from $\mathcal{N}((\frac{1}{2}, \frac{1}{2}), 0.15 I_2)$. The initial weight w_0 is selected from $\mathcal{N}(0, I_2)$. We ran p-GD with step size 10^{-4} for 1 million steps. As for the scatter plot of the data, we randomly re-assign a label and plot out $(x_i, 1)$ or $(-x_i, -1)$ uniformly at random.

For the $\mathbb{R}^{100}$ experiment, we select 15 sparse vectors that each has up to 10 nonzero entries. Each nonzero entry is independently sampled from $\mathcal{U}(-2, 4)$. Because we are in the over-parameterized case, these vectors are linearly separable with high probability. The initial weight w_0 is selected from $\mathcal{N}(0, 0.1 I_{100})$. We ran p-GD with step size 10^{-4} for 1 million steps.

These experiments were performed on an Intel Skylake CPU.

E.2 CIFAR-10 experiments

For the experiments with the CIFAR-10 dataset, we adopted the example implementation from the `FFCV` library.[7] For consistency, we ran p-GD with the same hyper-parameters for all neural networks and values of p. We used a cyclic learning rate schedule with maximum learning rate of 0.1 and ran for 400 epochs so the training loss is approximately 0.[8]

This experiment was performed on a single Nvidia V100 GPU.

E.3 ImageNet experiments

For the experiments with the ImageNet dataset, we used the example implementation from the `FFCV` library.[9] For consistency, we ran p-GD with the same hyper-parameters for all neural networks and values of p. We used a cyclic learning rate schedule with maximum learning rate of 0.5 and ran for 120 epochs. Note that, to more accurately measure the effect of p-GD on generalization, we turned off any parameters that may affect regularization, e.g. with momentum set to 0, weight decay set to 0, and label smoothing set to 0, etc.

This experiment was performed on a single Nvidia V100 GPU.

F Additional experimental results

F.1 Linear classification

We present a more complete result for the setting of Section 4.1 with more values of p. Note that Table 2 is a subset of Table 4, as shown below.

Except for $p = 1.1$, p-GD produces the smallest linear classifier under the corresponding ℓ_p-norm and thus consistent with the prediction of Theorem 13. When $p = 1.1$, Corollary 17 predicts a much slower convergence rate. So, for the number of iterations we have, p-GD with $p = 1.1$ in fact cannot compete against p-GD with $p = 1.5$, which has much faster convergence rate but similar implicit bias. The second trial shows a rare case where p-GD with $p = 1.1$ could not even match p-GD with $p = 2$ under the $\ell_{1.1}$-norm. Therefore, before we come up with techniques to speed up the convergence of p-GD, it is not advisable to pick p that is too close to 1.

[7] https://github.com/libffcv/ffcv/tree/main/examples/cifar

[8] This differs from the setup from Azizan et al. [2021b], where they used a fixed small learning rate and much larger number of epochs.

[9] https://github.com/libffcv/ffcv-imagenet/

Table 4: Size of the linear classifiers generated by p-GD (after rescaling) in $\ell_1, \ell_{1.1}, \ell_{1.5}, \ell_2, \ell_3, \ell_6$ and ℓ_{10} norms. For each norm, we highlight the value of p for which p-GD generates the smallest classifier under that norm. (Trial 1)

	ℓ_1 norm	$\ell_{1.1}$ norm	$\ell_{1.5}$ norm	ℓ_2 norm	ℓ_3 norm	ℓ_6 norm	ℓ_{10} norm	ℓ_∞ norm
$p = 1.1$	**7.692**	5.670	2.650	1.659	1.100	0.782	0.698	0.634
$p = 1.5$	7.924	**5.607**	**2.333**	1.346	0.830	0.573	0.526	0.515
$p = 2$	9.417	6.447	2.413	**1.273**	0.710	0.444	0.393	0.374
$p = 3$	11.307	7.618	2.696	1.345	**0.691**	0.381	0.318	0.285
$p = 6$	13.115	8.787	3.044	1.481	0.729	**0.369**	0.288	0.233
$p = 10$	13.572	9.086	3.137	1.520	0.742	0.367	**0.281**	**0.213**

Table 5: Size of the linear classifiers generated by p-GD (after rescaling) in $\ell_1, \ell_{1.1}, \ell_{1.5}, \ell_2, \ell_3, \ell_6$ and ℓ_{10} norms. For each norm, we highlight the value of p for which p-GD generates the smallest classifier under that norm. (Trial 2)

	ℓ_1 norm	$\ell_{1.1}$ norm	$\ell_{1.5}$ norm	ℓ_2 norm	ℓ_3 norm	ℓ_6 norm	ℓ_{10} norm	ℓ_∞ norm
$p = 1.1$	10.688	8.013	3.883	2.465	1.644	1.187	1.082	1.009
$p = 1.5$	**9.308**	**6.546**	**2.674**	1.518	0.913	0.602	0.535	0.488
$p = 2$	10.735	7.340	2.735	**1.435**	0.790	0.479	0.418	0.397
$p = 3$	12.298	8.327	2.991	1.508	**0.782**	0.432	0.359	0.324
$p = 6$	13.817	9.322	3.297	1.631	0.816	**0.418**	0.328	0.265
$p = 10$	14.545	9.798	3.447	1.695	0.841	0.423	**0.325**	**0.247**

F.2 CIFAR-10 experiments: implicit bias

We present more complete illustrations of the implicit bias trends of trained models in CIFAR-10. Compared to Figure 2, the plots below include data from additional values for additional values of p and more deep neural network architectures.

We see that the trends we observed in Section 4.2 continue to hold under architectures other than RESNET. In particular, for smaller p's, the weight distributions of models trained with p-GD have higher peak around zero, and higher p's result in smaller maximum weights.

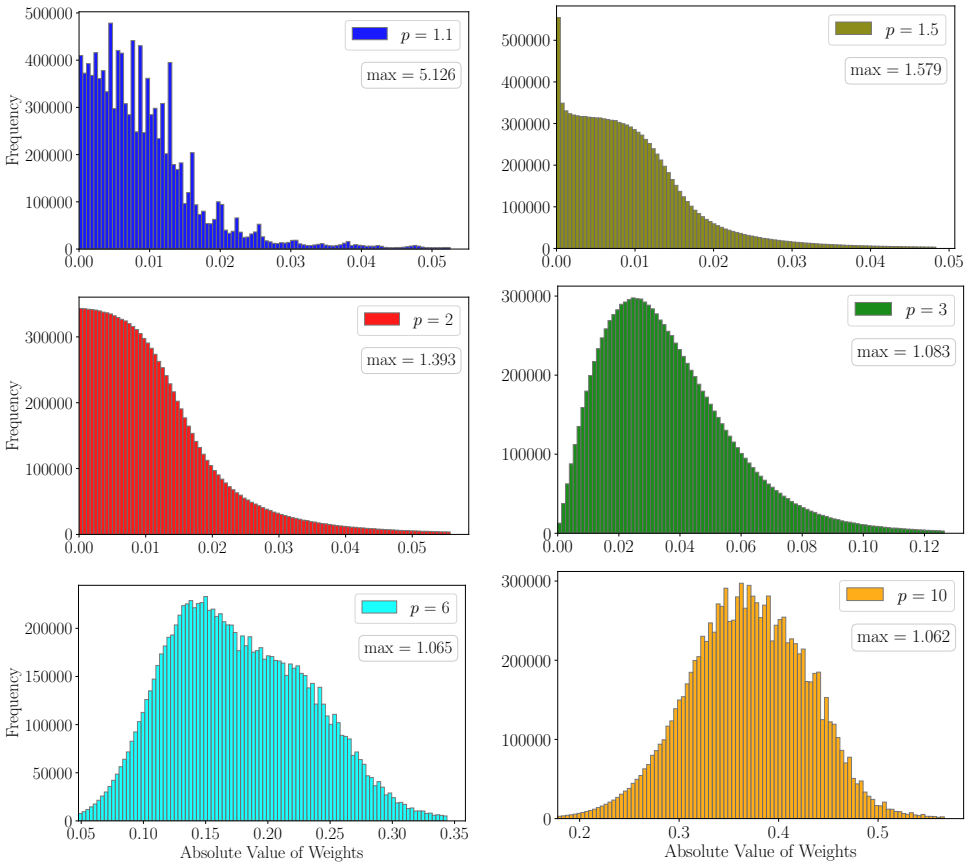

Figure 3: The histogram of weights in RESNET-18 models trained with p-GD for the CIFAR-10 dataset. For clarity, we cropped out the tails and each plot has 100 bins after cropping. Note that the scale on the y-axis differs per graph.

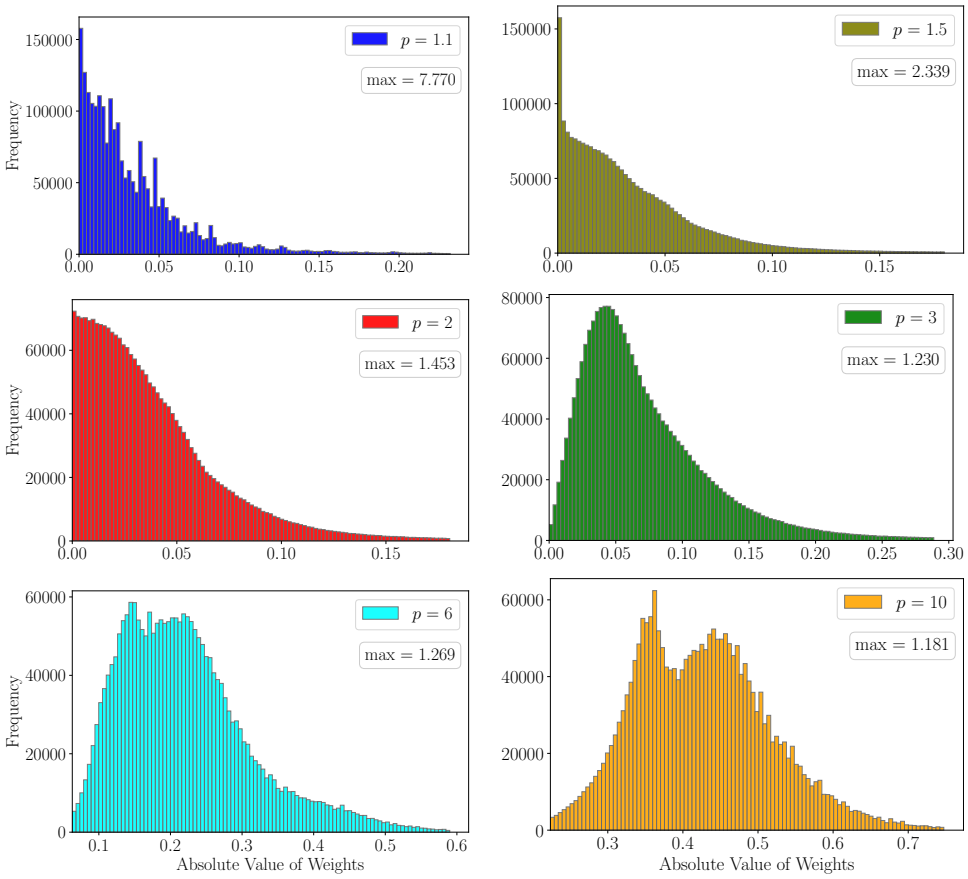

Figure 4: The histogram of weights in MOBILENET-V2 models trained with p-GD for the CIFAR-10 dataset. For clarity, we cropped out the tails and each plot has 100 bins after cropping.

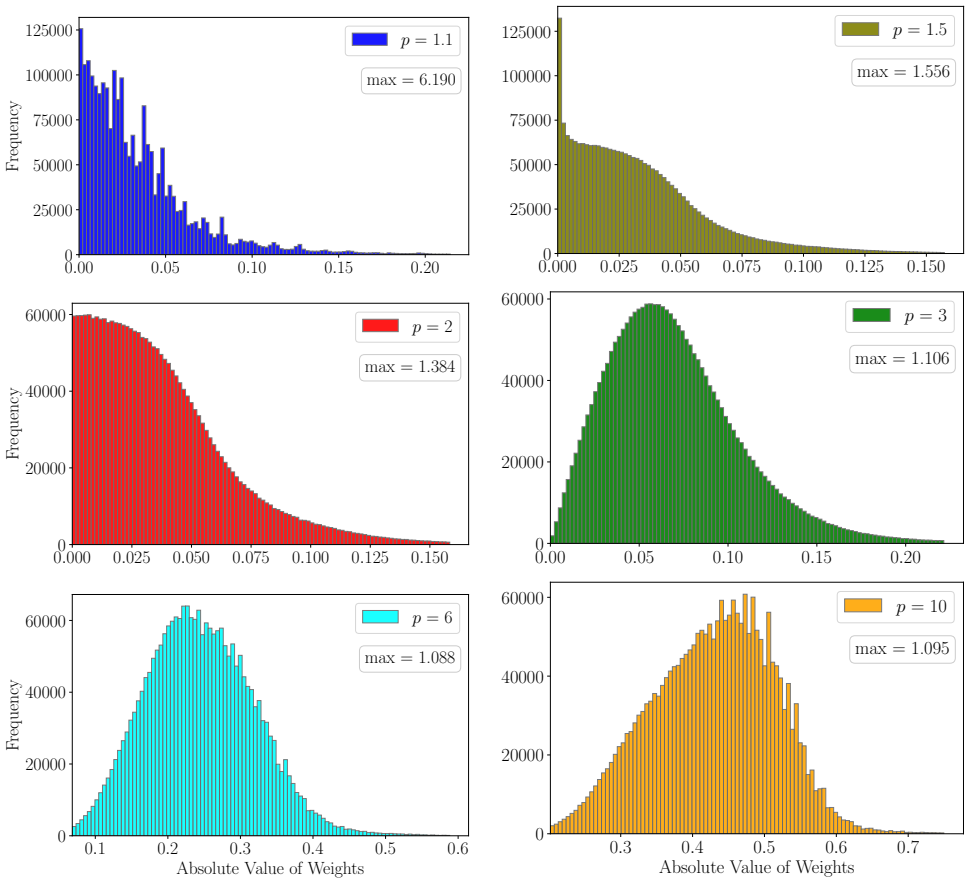

Figure 5: The histogram of weights in REGNETX-200MF models trained with p-GD for the CIFAR-10 dataset. For clarity, we cropped out the tails and each plot has 100 bins after cropping.

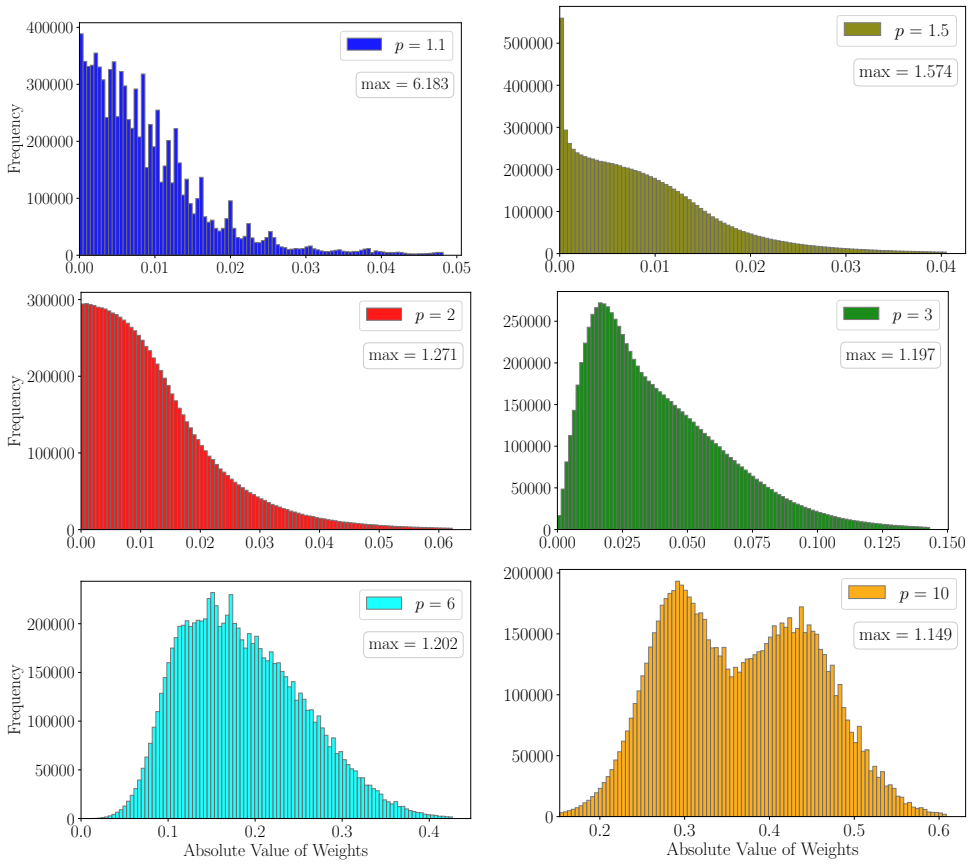

Figure 6: The histogram of weights in VGG-11 models trained with p-GD for the CIFAR-10 dataset. For clarity, we cropped out the tails and each plot has 100 bins after cropping.

F.3 CIFAR-10 experiments: generalization

We present a more complete result for the CIFAR-10 generalization experiment in Section 4.2 with additional values of p.

In the following table, we see that p-GD with $p = 3$ continues have the highest generalization performance for all deep neural networks.

Table 6: CIFAR-10 test accuracy (%) of p-GD on various deep neural networks. For each deep net and value of p, the average $\pm$ std. dev. over 5 trials are reported. And the best performing value(s) of p for each individual deep net is highlighted in **boldface**.

	VGG-11	RESNET-18	MOBILENET-V2	REGNETX-200MF
$p = 1.1$	88.19 $\pm$ .17	92.63 $\pm$ .12	91.16 $\pm$ .09	91.21 $\pm$ .18
$p = 1.5$	88.45 $\pm$ .29	92.73 $\pm$ .11	90.81 $\pm$ .19	90.91 $\pm$ .12
$p = 2$ (SGD)	90.15 $\pm$ .16	**93.90** $\pm$ .14	91.97 $\pm$ .10	92.75 $\pm$ .13
$p = 3$	**90.85** $\pm$ .15	**94.01** $\pm$ .13	**93.23** $\pm$ .26	**94.07** $\pm$ .12
$p = 6$	89.47 $\pm$ .14	**93.87** $\pm$ .13	92.84 $\pm$ .15	93.03 $\pm$ .17
$p = 10$	88.78 $\pm$ .37	93.55 $\pm$ .21	92.60 $\pm$ .22	92.97 $\pm$ .16

F.4 ImageNet experiments

To verify if our observations on the CIFAR-10 generalization performance hold up for other datasets, we also performed similar experiments for the much larger ImageNet dataset. Due to computational constraints, we were only able to experiment with the RESNET-18 and MOBILENET-V2 architectures and only for one trial.

It is worth noting that the neural networks we used cannot reach 100% training accuracy on Imagenet. The models we employed only achieved top-1 training accuracy in the mid-70's. So, we are not in the so-called *interpolation regime*, and there are many other factors that can significantly impact the generalization performance of the trained models. In particular, we find that not having weight decay costs us around 3% in validation accuracy in the $p = 2$ case and this explains why our reported numbers are lower than PyTorch's baseline for each corresponding architecture. Despite this, we find that p-GD with $p = 3$ has the best generalization performance on the ImageNet dataset, matching our observation from the CIFAR-10 dataset.

Table 7: ImageNet top-1 validation accuracy (%) of p-GD on various deep neural networks. The best performing value(s) of p for each individual deep network is highlighted in **boldface**.

	RESNET-18	MOBILENET-V2
$p = 1.1$	64.08	63.41
$p = 1.5$	65.14	65.75
$p = 2$ (SGD)	66.76	67.91
$p = 3$	**67.67**	**69.74**
$p = 6$	66.69	67.05
$p = 10$	65.10	62.32