Journal of Machine Learning Research 24 (2023) 1-58 Submitted 6/23; Revised 12/23; Published 1/24

A Unified Approach to Controlling Implicit Regularization via Mirror Descent

Haoyuan Sun HAOYUANS@MIT.EDU
Khashayar Gatmiry GATMIRY@MIT.EDU
Kwangjun Ahn KJAHN@MIT.EDU
Navid Azizan AZIZAN@MIT.EDU
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Editor: Mahdi Soltanolkotabi

Abstract

Inspired by the remarkable success of large neural networks, there has been significant interest in understanding the generalization performance of over-parameterized models. Substantial efforts have been invested in characterizing how optimization algorithms impact generalization through their "preferred" solutions, a phenomenon commonly referred to as *implicit regularization*. In particular, it has been argued that gradient descent (GD) induces an implicit ℓ_2-norm regularization in regression and classification problems. However, the implicit regularization of different algorithms are confined to either a specific geometry or a particular class of learning problems, indicating a gap in a general approach for controlling the implicit regularization. To address this, we present a unified approach using mirror descent (MD), a notable generalization of GD, to control implicit regularization in both regression and classification settings. More specifically, we show that MD with the general class of homogeneous potential functions converges in direction to a *generalized maximum-margin* solution for linear classification problems, thereby answering a long-standing question in the classification setting. Further, we show that MD can be implemented efficiently and enjoys fast convergence under suitable conditions. Through comprehensive experiments, we demonstrate that MD is a versatile method to produce learned models with different regularizers, which in turn have different generalization performances.

Keywords: implicit regularization, mirror descent, gradient descent, maximum-margin classification, over-parameterization

1. Introduction

In recent years, deep neural networks have enjoyed a tremendous amount of success in a wide range of applications (Schrittwieser et al., 2020; Ramesh et al., 2021; Brown et al., 2020; Dosovitskiy et al., 2020). Notably, many of these modern machine learning problems operate in the so-called *over-parameterized* regime, where the number of model parameters is sufficiently large to allow for perfectly fitting the training data (Allen-Zhu et al., 2019; Belkin, 2021). However, such highly expressive models have the capacity to have multiple solutions that interpolate training data, and these solutions can often perform very differently on test data. Without knowing which of these interpolating solutions the optimizer finds, it would be difficult to identify whether the learned model *overfits*, where it performs well on

the training data but generalizes poorly on the test data. Therefore, a characterization of the optimizers' solutions is essential to the understanding of the generalization performance of modern over-parameterized 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 many gradient-based optimization algorithms, such as (stochastic) gradient descent, tend to generalize well. Recent research has highlighted that these algorithms converge to solutions with certain properties, i.e., they *implicitly regularize* the learned models. Importantly, it has been argued that such implicit biases play a significant role in determining generalization performance (Neyshabur et al., 2014; Zhang et al., 2021; Wilson et al., 2017; Liang and Rakhlin, 2020; Donhauser et al., 2022).

In the literature, the implicit bias of first-order methods is first studied in linear settings since the analysis is more tractable. Nevertheless, there is significant theoretical and empirical evidence suggesting that certain insights from linear models translate to the case of deep models, e.g., Jacot et al. (2018); Allen-Zhu et al. (2019); Belkin et al. (2019); Bartlett et al. (2017); Nakkiran et al. (2021); Du et al. (2018). In the linear setting, implicit bias has been extensively analyzed in the contexts of different learning problems, of which two have received the most attention. The first case is least-squares regression, where the loss function has a global minimum attainable at a finite value. And the second case is classification with logistic or exponential loss, where the loss function does not have an attainable global minimum. While they are colloquially referred to as regression or classification problems, respectively, we note that there are other examples of the loss function, such as hinge loss, that do not fall into these two categories. In Section 2, we will define these notions more formally.

For the analysis of implicit regularization, the most well-studied optimization algorithm is gradient descent (GD). The implicit bias of gradient descent (GD) for the square loss goes back to Engl et al. (1996), and possibly earlier, where it was shown that GD converges to the global minimum that is closest to the initialization in the Euclidean distance. For the logistic loss, it has been shown that the gradient descent iterates converge to the ℓ_2-maximum-margin SVM solution in direction (Soudry et al., 2018; Ji and Telgarsky, 2019a). Beyond GD, it has been shown that mirror descent (MD), which is an important generalization of GD, converges to the interpolating solution that is closest to the initialization in terms of a Bregman divergence (Gunasekar et al., 2018; Azizan and Hassibi, 2019a). On the classification side, it has been established that the implicit biases of various algorithms such as AdaBoost (Rosset et al., 2004; Telgarsky, 2013) and steepest descent (Gunasekar et al., 2018) maximize the margin with respect to certain norms.

There are also many counter-examples where an optimization algorithm does not exhibit an implicit bias independent of hyper-parameters such as the step size. For instance, it has been shown that the AdaGrad algorithm does not have such an implicit bias in the classification setting (Gunasekar et al., 2018; Wang et al., 2021). Further, it is possible for optimization algorithms to exhibit step-size-invariant implicit bias in regression but not in classification and vice versa (e.g., steepest descent). To our best knowledge, gradient descent is the only first-order algorithm known to induce a step-size-invariant implicit bias in both settings of regression with square loss and classification with logistic loss. Therefore, there is still a significant gap in the understanding of implicit regularization for different classes of losses and some believe that the regression and classification settings are "fundamentally

Table 1: **Conceptual summary of our results.** For both well-specified linear regression and separable linear classification, gradient descent converges to the "smallest" global minimum in ℓ_2-norm. In the case of linear regression, it is known that mirror descent generalizes the implicit bias of gradient descent to any strictly convex potential. However, a similar characterization is missing for mirror descent under the classification setting. In this paper, we prove the implicit regularization of mirror descent with the class of homogeneous potentials and extend the result of gradient descent beyond ℓ_2-norm. In Appendix A, we present a more complete summary where we consider the implicit bias in the regression setting with any initialization.

	Regression (e.g. square loss) with $w_0 = \operatorname{argmin} \psi(w)$	Classification (e.g. logistic loss) with any initialization
Gradient Descent (i.e. $\psi(\cdot) = \frac{1}{2}\|\cdot\|_2^2$)	$\operatorname{argmin}_w \|w\|_2$ 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 (2019a)
Mirror Descent	$\operatorname{argmin}_w \psi(w)$ s.t. w fits all data Gunasekar et al. (2018) Azizan and Hassibi (2019a)	$\operatorname{argmin}_w \psi(w)$ s.t. w classifies all data **This work**

different" (Gunasekar et al., 2018). In this paper, we show that mirror descent can extend the implicit bias of gradient descent to more general geometries in the classification setting. So, we conclude that mirror descent is the first known algorithm exhibiting implicit regularization for both general geometry and different classes of loss functions. Furthermore, we show that under many circumstances, mirror descent can both be efficiently implemented and quickly converge to its implicitly regularized solution. Hence, mirror descent is a versatile way of implicitly enforcing desirable properties on the learned model in a variety of tasks. See Table 1 for a summary.

1.1 Our contributions

In this paper, our theoretical and empirical contributions are as follows:

- We study mirror descent (MD) with the general class of homogeneous potential functions. In Section 3.1, we show that for separable linear classification with logistic loss, MD with homogeneous potential exhibits implicit regularization by converging in direction to a generalized maximum-margin solution with respect to the potential function. More generally, we show that, for any strictly decreasing loss function, MD follows the so-called regularization path. The precise terminologies are defined in Sections 2 and 3.
- We study the rate at which MD with homogeneous potential converges in direction to the generalized maximum-margin solution. In Section 3.2, we show that with fixed step sizes, MD converges in direction to the maximum-margin solution at a poly-logarithmic

rate in the number of iterates T. And in Section 3.3, with additional assumptions on the potential function, we prove the convergence of a variant of normalized MD and show the rate of convergence can be accelerated to be polynomial in the number of iterates T.

- In Section 4, we investigate the implications of our theoretical findings by applying a subclass of MD that is both efficient and scalable. Our experiments involving linear models corroborate our theoretical results in Section 3, 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 mirror descent with respect to different potential functions can lead to different solutions with significantly different generalization performance.

1.2 Related work

Implicit regularization in regression. Regression problems are typically concerned with the case of square loss. For gradient descent (GD) with square loss on a linear model, it is known that GD converges to the minimizer that is closest to the initialization in the Euclidean sense (Engl et al., 1996). We can induce implicit bias with respect to other geometries with a family of algorithms called mirror descent (MD), which is an extension of GD. In particular, it has been shown that mirror descent converges to the interpolating solution that is closest to the initialization in terms of a Bregman divergence with respect to MD's potential function (Gunasekar et al., 2018; Azizan and Hassibi, 2019a).[1] Additionally, Gunasekar et al. (2018) showed that a variant of MD with momentum also converges to the same implicit bias, and Azizan and Hassibi (2019a) analyzed stochastic MD. So, we consider the study of implicit bias in linear regression to be relatively well-understood by now.

Implicit regularization in classification. Classification problems are typically concerned with the case of logistic or exponential loss. Another common loss function for classification problems is the hinge loss, but as we will note in Section 2, this loss is uninteresting in terms of analyzing implicit bias. 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. For the logistic loss and other strictly decreasing losses, it has been shown that gradient descent iterates converge to the ℓ_2-maximum-margin solution in direction (Soudry et al., 2018; Ji and Telgarsky, 2019a; Ji et al., 2020). Further applying various schemes of adaptive step size to gradient descent can accelerate its convergence to the solution induced by its implicit bias (Nacson et al., 2019; Ji and Telgarsky, 2021; Ji et al., 2021). Beyond implicit bias with respect to the ℓ_2-norm, we also know that AdaBoost converges to the ℓ_1-maximum margin direction (Rosset et al., 2004; Telgarsky, 2013). Also, Gunasekar et al. (2018) showed that the steepest descent algorithm converges to the maximum-margin direction with respect to a general norm; however, this algorithm cannot be implemented efficiently in practice. The analysis of mirror descent for this setting had been more limited; the only prior work we are aware of showed a special case where mirror descent's potential is the Mahalanobis distance[2] (Li et al., 2021).

1. MD has also been used for explicit regularization, e.g., Azizan et al. (2021a), which is not the focus of this work.
2. which is equivalent to the Euclidean distance up to a linear transformation in the coordinates.

Connections between regression and classification. For over-parameterized models, empirical evidence shows that using either least-square loss or cross-entropy loss can lead to comparable performance on classification tasks (Hui and Belkin, 2020). This suggests that the distinction between regression and classification problems is blurred when we enter the over-parameterized regime. There has been some progress in theoretically explaining this phenomenon, for example, Muthukumar et al. (2021) showed that for over-parameterized linear models and when the data are drawn from a Gaussian distribution, the minimum-ℓ_2-norm interpolating solution and the ℓ_2-maximum-margin solutions are equivalent with high probability. However, it is unknown whether such connections exist for more general geometries extending beyond the ℓ_2-norm.

Extension to nonlinear models. In addition to linear models, several works have analyzed the implicit bias when we optimize over a nonlinear model such as neural networks. There is now a good understanding of the case of simpler networks such as homogeneous networks without nonlinearity (Lyu and Li, 2019; Vardi et al., 2022; Wang et al., 2021), ReLU networks (Ji and Telgarsky, 2019b; Zou et al., 2020), or networks with dropout (Mianjy et al., 2018; Wei et al., 2020). However, the development of implicit regularization for more general deep neural networks remains an ongoing research direction.

2. Background

2.1 Problem setting

In this paper, we are interested in the standard *empirical risk minimization* problem. Consider a collection of input-output pairs $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ and a model $f_w(x) : \mathbb{R}^d \to \mathbb{R}$ with parameter $w \in \mathcal{W}$. For some convex *loss function* $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, our goal is to minimize the empirical loss:

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

We can categorize the loss functions by the properties of their minimizer. For simplicity, we assume without loss of generality that $\inf \ell(\cdot) = 0$. The first type is concerned with regression problems, whereas the next two types of losses are concerned with classification problems, where the output variable is a binary label $y \in \{\pm 1\}$.

- **Losses with a unique minimizer.** In this case, the minima of the loss function is attained if and only if $f_w(x) = y$. Example: square loss $\ell(f_w(x), y) = (f_w(x) - y)^2$.
- **Losses with non-unique minimizer.** In this case, there are other minimizers in addition to the ones at $f_w(x) = y$. Example: hinge loss $\ell(f_w(x), y) = \max(0, 1 - f_w(x)y)$.
- **Strictly monotone losses.** A finite minimizer is not attainable in this case, but for any fixed value of y, the loss $\ell(f_w(x), y)$ strictly decreases with respect to $f_w(x)y$. Example: exponential loss $\ell(f_w(x), y) = \exp(-f_w(x)y)$, or logistics loss.

We note that gradient descent does not exhibit an implicit bias for losses with a non-unique minimizer because the final solution will depend on the step size (see the example below). Therefore, we are mainly interested in only the first and third types of loss functions. For simplicity, unless otherwise stated, we informally refer to regression as problems with square loss and classification as problems with exponential loss.

Example 1 *We give a simple example where the solution found by gradient descent with the hinge loss is dependent on the step size. Consider a classification dataset with three points in* $\mathbb{R}^3$*:* $((1,0,0),+1),((0,1,0),+1),((0,0,1),+1)$*, a linear model* $f_w(x) = w^\top x$ *and initial weight* $w_0 = (-1,-2,-3)$*. When the step size is* $\eta = 2$*, the iterates are* $w_1 = (1,0,-1), w_2 = (1,2,1)$*; and when the step size is* $\eta = 3$*, the iterates are* $w_1 = (2,1,0), w_2 = (2,1,3)$*. The final solutions are dependent on the step size, so gradient descent does not exhibit implicit regularization with the hinge loss.*

For our theoretical analysis, we focus on a linear model, where the models can be expressed by $f_w(x) = w^\top x$ and $w \in \mathbb{R}^d$, in classification problems with strictly monotone losses. We also make the following assumptions about the data. First, since we are mainly interested in the over-parameterized setting where $d \gg n$, 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, where $\max_i \|x_i\| < C$, for some relevant norm $\|\cdot\|$ which we will specify in Section 3.

2.2 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 ψ defined on the entire domain $\mathbb{R}^n$.[3] 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 strictly convex potential function* 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 $\psi = \frac{\|\cdot\|_2^2}{2}$, where $\|\cdot\|_2$ denotes the Euclidean norm. Then, the Bregman divergence becomes $D_\psi(x,y) = \frac{1}{2}\|x-y\|_2^2$. As an example for more complicated geometry, when the potential is $\psi(x) = \frac{1}{2}x^\top P x$ with positive definite $P \succ 0$, then the Bregman divergence becomes the Mahalanobis distance. For more background on Bregman divergence and its properties, see, e.g., (Bauschke et al., 2017, Section 2.2) and (Azizan and Hassibi, 2019b, Section II.A).

The mirror descent (MD) algorithm (Nemirovskij and Yudin, 1983) is a generalization of gradient descent over geometries beyond the Euclidean distance. In mirror descent with potential ψ, 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):

3. In general, the mirror map is a convex function of Legendre type (see, e.g., (Rockafellar, 1970, Sec. 26)).

Lemma 2 (MD identity) *For any $w \in \mathbb{R}^n$, the following identities hold for (MD)*[4]*:*

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

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

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 B.

Remark 5 *One can relax the condition in Lemma 3 and 4 such that for a sufficiently small step size η, $\psi - \eta L$ only has to be locally convex at the iterates $\{w_t\}_{t=0}^{\infty}$. 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)). In Appendix D, we discuss in detail the existence of such a step size for the exponential loss.*

2.3 Preliminaries on implicit regularization

As we discussed above, the weights vector w_t diverges for mirror descent. Here the main theoretical question is:

What direction does MD converge to? In other words,
under some norm $\|\cdot\|$, can we characterize $w_t / \|w_t\|$ as $t \to \infty$?

To define a notion of "norm" respecting the geometry induced by potential ψ, we let $\|\cdot\|_\psi$ be the Minkowski functional of ψ's unit sub-level set:

$$\|w\|_\psi := \inf\{c > 0 : \psi(w/c) \le 1\} \quad (2)$$

For a wide class of convex ψ, this definition indeed gives us a norm. In Section 3, we shall give a sufficient condition for $\|\cdot\|_\psi$ to be a norm. With the definition of $\|\cdot\|_\psi$ in mind, we introduce two special directions whose importance will be illustrated later.

Definition 6 *The **regularization path** with respect to $\|\cdot\|_\psi$ is defined as*

$$\bar{w}_\psi(B) = \operatorname*{argmin}_{\|w\|_\psi \le B} L(w) \quad (3)$$

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

4. For convenience, for a function f, we write $D_f(x, y) := f(x) - f(y) - \langle \nabla f(y),\ x - y \rangle$. Note that when f is convex, $D_f(\cdot, \cdot) \ge 0$, and when f is strictly convex, $D_f(\cdot, \cdot)$ is the Bregman divergence.

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

$$u_\psi^{\mathsf{m}} := \operatorname*{argmax}_{\psi(w)\le 1} \left\{ \min_{i=1,\dots,n} y_i \langle x_i,\ w\rangle \right\} \tag{4}$$

And let $\hat{\gamma}_\psi$ be the optimal value to the equation above.[5]

For simpler case where we have the ℓ_p-norm (one possible corresponding potential is $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^p$), we overload the notation with u_p^{r} and u_p^{m}. Note that the superscripts in u_ψ^{r} and u_ψ^{m} are not variables and we only use this notation to differentiate the two definitions.

To illustrate the effect of the potential ψ on the maximum-margin solution, we consider a dataset consisting of a single point $((\frac{1}{2}, \frac{3}{2}), +1)$. For u_2^{m}, we get the SVM solution whose decision boundary is orthogonal to the line connecting $(\frac{1}{2}, \frac{3}{2})$ to the origin. And for u_1^{m}, we get a "sparse" max-margin solution that is zero in one coordinate. Lastly, because $\|\cdot\|_{10}$ is very close to $\|\cdot\|_\infty$, the coordinates in the max-margin solution u_{10}^{m} are very close to each other.

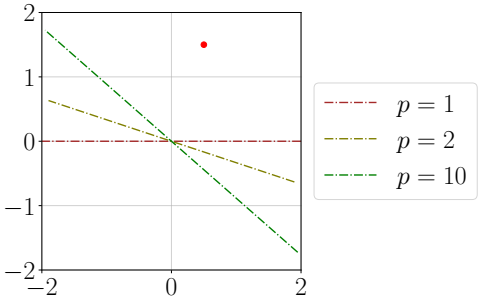


Figure 1: The generalized maximum-margin solution to a single data point (denoted by •) with respect to the ℓ_1, ℓ_2, and ℓ_{10} norms. For each generalized max-margin solution u, we plot the decision boundary $\{x \mid u^\top x = 0\}$.

Prior results had shown that, in linear classification, gradient descent converges in direction to u_2^{m}, which is parallel to the hard-margin SVM solution w.r.t. ℓ_2-norm: $\operatorname{argmin}_w\{\|w\|_2 : \gamma(w) \ge 1\}$.

Theorem 8 (Soudry et al. (2018)) *For separable linear classification problems with logistics/exponential 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_2^{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., $\lim_{t\to\infty} \frac{w_t}{\|w_t\|_2} = u_2^{\mathsf{r}}$.*

As for mirror descent, an earlier version of this paper (Sun et al., 2022) studied when the potential function is $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^p$ and showed that $w_t/\|w_t\|_p$ converges in direction.

Theorem 10 (Sun et al. (2022)) *Given a separable linear classification problem with strictly monotone loss. If the regularized direction u_p^{r} with respect to the ℓ_p-norm exists, then the iterates of mirror descent with $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^p$ and sufficiently small step size converge to the generalized regularized direction u_p^{r}, i.e., $\lim_{t\to\infty} \frac{w_t}{\|w_t\|_p} = u_p^{\mathsf{r}}$.*

In Section 3.1, we will generalize Theorems 8-10 to a more general setting of mirror descent with the class of homogeneous potential function.

5. Note that, when $\|\cdot\|_\psi$ is a norm, the sets $\{\psi(\cdot) \le 1\}$ and $\{\|\cdot\|_\psi \le 1\}$ are equal. In this paper, we will use these two formulations interchangeably.

3. Mirror Descent with Homogeneous Potential

In this section, we investigate the implicit regularization property of mirror descent for classification problems with strictly monotone losses. For the notion of "direction" to be well-defined, we impose the following properties on the potential function.

Assumption 1 *We assume that the potential function ψ has the following properties:*

- *ψ is twice differentiable and strictly convex, i.e. $\nabla^2\psi \succ 0$.*
- *ψ is positive definite in the sense that $\psi(\cdot) \geq 0$ and $\psi(x) = 0$ if and only if $x = 0$.*
- *For some constant $\beta > 1$, ψ is β-absolutely homogeneous in the sense that for any scalar c and vector $x \in \mathbb{R}^d$, we have $\psi(cx) = |c|^\beta \psi(x)$.*

Under this assumption, we can verify that $\|\cdot\|_\psi$ as defined in (2) is indeed a norm. In particular, if $\psi(\cdot) = \|\cdot\|^\beta$ for some norm $\|\cdot\|$, then we have $\|\cdot\|_\psi = \|\cdot\|$. Also, it is not difficult to show that ψ satisfies the following properties, so that the Bregman divergence is also absolutely homogeneous. This would allow us to easily normalize the weight vector w in our computations.

$$\langle \nabla\psi(w),\ w\rangle = \beta \cdot \psi(w) \tag{5a}$$

$$D_\psi\left(cw, cw'\right) = |c|^\beta D_\psi\left(w, w'\right) \quad \forall c \in \mathbb{R}. \tag{5b}$$

For a detailed discussion of potentials under Assumption 1, we refer the readers to Appendix C.

We note that this assumption is quite general and covers several previously studied cases of mirror descent:

- When $\psi = \frac{1}{2}\|\cdot\|_2^2$, we recover gradient descent.
- When $\psi = \frac{1}{2}\|\cdot\|_q^2$, we have the so-called *p-norm algorithm* (where $1/p + 1/q = 1$) (Grove et al., 2001; Gentile, 2003).
- The potential $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^p$ for $p > 1$ 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:

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

  In comparison, the p-norm algorithm is not coordinate-wise separable since it requires computing $\|w_t\|_q$ at each step (see, e.g., (Gentile, 2003, eq. (1))). This potential $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^p$ was first considered by (Azizan et al., 2021b) in the case of regression in which the loss function has unique minimizer. In an earlier version of this work (Sun et al., 2022), we analyzed this potential in the case of classification with strictly monotone losses and named mirror descent with this potential as *p-norm GD*, or p-GD in short, because it naturally generalizes gradient descent to ℓ_p-norms.

Remark 11 *In this paper, we do not consider the case where the potential is the negative entropy function (which recovers the multiplicative weights or the hedge algorithm from the online learning literature) because this potential function requires all the weights to be positive, which would be too restrictive in our problem setting.*

Remark 12 *It is worth noting that the steepest descent algorithm, which follows the update rule*

$$w_{t+1} = \operatorname*{argmin}_{w} \left\{ \frac{1}{2\eta} \|w\|^2 + \langle \nabla L(w_t),\ w \rangle \right\}$$

for a general norm $\|\cdot\|$, is not an instance of mirror descent since $\frac{1}{2}\|\cdot\|^2$ is in general not a Bregman divergence. Further, this update rule does not have a closed-form solution and thus cannot be solved efficiently. In this section, we will see that mirror descent with potential $\psi(\cdot) = \|\cdot\|^\beta$ for any $\beta > 1$ will induce the same implicit bias as steepest descent for strictly monotone losses.

Finally, recall that in Section 2.1, we discussed our assumptions on the dataset. With the definition of $\|\cdot\|_\psi$, we can now state these assumptions more precisely.

Assumption 2 *The dataset $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{\pm 1\}$ is linearly separable so there exists $w^* \in \mathbb{R}^d$ s.t. $\operatorname{sign}(\langle w^*,\ x_i \rangle) = y_i$ for all $i \in [n]$. Let $\|\cdot\|_{\psi,*}$ be the dual norm to $\|\cdot\|_\psi$. The input variables x_i are bounded that there exists constant $C > 0$ where $\max(\|x_i\|_2, \|x_i\|_{\psi,*}) \le C$ for all $i \in [n]$.*

3.1 Main theoretical results

We extend Theorems 8-10 to the setting of mirror descent with potential functions satisfying Assumption 1. Building upon the analysis in our previous work (Sun et al., 2022), we resolve several major obstacles in the analysis of implicit regularization of linear classification where we apply a general class of MD to strictly monotone loss functions:

- We approach the classification setting with strictly monotone loss by considering the limit of a sequence of constrained optimization problems. Then each constrained problem has a unique and finite minimizer, and these solutions trace out the regularization path. Our analysis builds upon the techniques employed by Ji et al. (2020). In addition to generalizing regularized direction to a more general geometry, we derive stronger justification for using regularized direction by connecting the analysis of implicit bias under different types of loss functions.
- Our argument addresses the concern from Gunasekar et al. (2018) that the implicit bias of regression and classification problems are "fundamentally different." In Gunasekar et al. (2018), it was noted that gradient descent's implicit bias is dependent on the initialization for regression problems, but not for classification problems. In this section, we argue that it is sufficient to reframe the classification setting as a sequence of carefully chosen regression problems.[6] Then, we find that the dependence on the initialization vanishes after taking the limit.
- On a more technical note, one challenge in analyzing mirror descent lies in the cross terms of the form $\langle \nabla\psi(w),\ w' \rangle$, which lack direct geometric interpretations. We demonstrate that under Assumption 1, these terms can be simplified nicely and still induce a variety of desired geometric properties on the implicit bias.

6. Recall that, in Section 2.1, we define "regression problem" as the case where the loss function has a unique and attainable minimizer.

We start our discussion with the following main result.

Theorem 13 *For a separable linear classification problem, if the regularized direction u_ψ^{r} exists, then with sufficiently small step size, the iterates of* MD *with a potential function ψ satisfying Assumption 1 converge to u_ψ^{r} in direction:*

$$\lim_{t\to\infty} \frac{w_t}{\|w_t\|_\psi} = u_\psi^{\mathsf{r}}. \tag{6}$$

Next, we shall introduce the key ideas behind this theorem. We motivate our use of regularized direction by first considering the regression setting and then highlighting the additional challenges we must overcome in the classification setting. For over-parameterized regression problems, there exists some weight vector w such that $L(w) = 0$. Then, we can apply Lemma 2 to get

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

By our choice that $L(w) = 0$, 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 $D_\psi(w, \cdot)$ during each step. Using this property, Azizan and Hassibi (2019a) establishes the implicit bias results of mirror descent in the regression setting.

We now return to the classification case and note that for strictly monotone losses, there are no attainable minimizers. Therefore, the choice of w we make in the regression case would not be valid for classification problems. Instead, we relax the "decreasing potential" property to hold for only a single step so that, at each time t, we choose a reference vector $\hat{w}_t$ satisfying $L(\hat{w}_t) \leq L(w_{t+1})$. The following result, which is a generalization of (Ji et al., 2020, Lemma 9), shows that we can let $\hat{w}_t$ be some scalar multiples of the regularized direction.

Lemma 14 *If the regularized direction u_ψ^{r} exists, then $\forall \alpha > 0$, there exists r_α such that for any w with $\|w\|_\psi > r_\alpha$, we have $L((1+\alpha)\|w\|_\psi u_\psi^{\mathsf{r}}) \leq L(w)$.*

We note that, due to Lemma 4, the condition in Lemma 14 is met for any sufficiently large time t. Then, at time t, we can pick Lemma 2's reference vector to be a "moving target" $c_t u_\psi^{\mathsf{r}}$ so that $L(c_t u_\psi^{\mathsf{r}}) \leq L(w_{t+1})$. Recall from the definition of the regularized direction, each choice of $c_t u_\psi^{\mathsf{r}}$ approximately corresponds to the solution of $\operatorname{argmin}_{\|w\|_\psi \leq c_t} L(w)$. Hence, intuitively speaking, our analysis converts the classification problem into a sequence of regression problems by constructing a constrained optimization problem at each update step. Let us formalize this idea. We begin with the following inequality:

$$D_\psi\left(c_t u_\psi^{\mathsf{r}}, w_{t+1}\right) \leq D_\psi\left(c_t u_\psi^{\mathsf{r}}, w_t\right) - \eta L(w_{t+1}) + \eta L(w_t), \tag{7}$$

where c_t is taken to be $\approx \|w_t\|_\psi$.[7]

Now we modify (7) so that it can telescope over different iterations. One way is to add $D_\psi\left(c_{t+1} u_\psi^{\mathsf{r}}, w_{t+1}\right)$ on both sides of (7) and move $D_\psi\left(c_t u_\psi^{\mathsf{r}}, w_{t+1}\right)$ to the right-hand side as

7. 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 15.

follows:

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

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

$$\begin{aligned}
D_\psi\left(c_T u_\psi^{\mathsf{r}}, w_T\right) \le D_\psi\left(c_0 u_\psi^{\mathsf{r}}, w_0\right) - \eta L(w_T) + \eta L(w_0) + \psi(c_T u_\psi^{\mathsf{r}}) - \psi(c_0 u_\psi^{\mathsf{r}})\\
- \sum_{t=0}^{T-1} \left\langle \nabla\psi(w_{t+1}),\ (c_{t+1}-c_t)u_\psi^{\mathsf{r}}\right\rangle
\end{aligned} \tag{8}$$

If we show that right-hand side of (8) is bounded, then exploiting the homogeneity of ψ (Assumption 1) and dividing both sides by c_T^β would yield that $D_\psi(u_\psi^{\mathsf{r}}, w_T/\left\|w_T\right\|_\psi) \to 0$ as $T \to \infty$. Therefore, after normalization, all terms in (8) dependent on the initialization w_0 would vanish. We note that $L(w_T) \in O(1)$ due to Lemma 3 and $\psi(c_T u_\psi^{\mathsf{r}}) = c_T^\beta$ because ψ is homogeneous. So, we turn our attention to the final term $\sum_t \left\langle \nabla\psi(w_{t+1}),\ (c_{t+1}-c_t)u_\psi^{\mathsf{r}}\right\rangle$. We first only consider the product $\left\langle \nabla\psi(w_{t+1}),\ u_\psi^{\mathsf{r}}\right\rangle$. Invoking the MD update rule, we get:

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

The inner product $\left\langle \nabla L(w_t),\ u_\psi^{\mathsf{r}}\right\rangle$ on the right-hand side is difficult to compute directly. So, we exploit the property of u_ψ^{r} to bound this quantity. We recall from the definition of regularized direction that u_ψ^{r} is the direction along which we achieve the smallest loss and hence $\nabla L(w_t)$ must point away from u_ψ^{r}, i.e., it must be that $\left\langle \nabla L(w_t),\ u_\psi^{\mathsf{r}}\right\rangle \lesssim \langle \nabla L(w_t),\ w_t\rangle$ (up to lower-order terms). The following result formalizes this intuition.

Corollary 15 *For w so that $\left\|w\right\|_\psi > r_\alpha$, we have $\langle \nabla L(w),\ w\rangle \ge (1+\alpha)\left\|w\right\|_\psi \left\langle \nabla L(w),\ u_\psi^{\mathsf{r}}\right\rangle$.*

Proof This follows from the following inequality

$$\left\langle \nabla L(w),\ w - (1+\alpha)\left\|w\right\| u_\psi^{\mathsf{r}})\right\rangle \ge L(w) - L((1+\alpha)\left\|w\right\| u_\psi^{\mathsf{r}}) \ge 0\,,$$

where the first inequality is due to the convexity of L and the second inequality is due to Lemma 14. ∎

Therefore, we are left with

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

Under our choice of homogeneous potential as detailed in Assumption 1, one can invoke Lemma 2 and Equation (5) to lower bound the quantity $\langle -\eta\nabla L(w_t),\ w_t\rangle$ in terms of $\psi(w_{t+1})$ and $\psi(w_t)$, and this step is detailed in Lemma 24 in Appendix E.2. It follows that the quantity $\left\langle \nabla\psi(w_{t+1}),\ u_\psi^{\mathsf{r}}\right\rangle$ can be bounded with a telescoping sum, where we can show that

$\left\langle \nabla\psi(w_{t+1}),\ u_\psi^{\mathsf{r}} \right\rangle \in \Omega(\|w_t\|_\psi^{\beta-1})$. Then, the final term in (8) turns into another telescoping sum in c_t's and $\|w_t\|$'s. Unwinding the above process, we show that

$$\sum_{t=0}^{T-1} \left\langle \nabla\psi(w_{t+1}),\ (c_{t+1}-c_t)u_\psi^{\mathsf{r}} \right\rangle \in \Omega(\|w_T\|_\psi^{\beta}),$$

which cancels out the quantity $\psi(c_T u_\psi^{\mathsf{r}})$ in (8). After normalization, it must be the case that $D_\psi\left(u_\psi^{\mathsf{r}}, w_t/\|w_t\|_\psi\right)$ converges to zero in the limit as $t \to \infty$. Putting everything together, we obtain Theorem 13. A formal proof of this result can be found in Appendix E.3.

The final missing piece to Theorem 13 would be the existence of the generalized regularized direction. In general, finding the limit direction u_ψ^{r} would be difficult. Fortunately, we can sometimes appeal to the generalized max-margin direction, which can be computed by solving a convex optimization problem. 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 16 *If we have a loss with exponential tail, e.g.* $\lim_{z\to\infty} \ell(z)e^{az} = b$, *then under a strictly convex potential* ψ, *the generalized regularized direction* u_ψ^{r} *exists and it is equal to the generalized max-margin direction* u_ψ^{m}.

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

3.2 Asymptotic convergence rate

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

Corollary 17 *Under the same setting as Theorem 13, the iterates of* MD *follows the rate of convergence:*

$$D_\psi\left(u_\psi^{\mathsf{r}}, \frac{w_t}{\|w_t\|_\psi}\right) \in O\left(\|w_t\|_\psi^{-(\beta-1)}\right).$$

To fully understand the convergence rate, we need to characterize the asymptotic behavior of $\|w_t\|_\psi$. In the following result, we quantify $\|w_t\|_\psi$ for the exponential loss. We note that similar conclusions can be drawn for other losses with an exponential tail, e.g. logistics loss. For the sake of simplicity, our analysis in this section focuses on the exponential loss.
Recall that from Assumption 2, $\max_i \|x_i\|_{\psi,*} \le C$, and the max-margin direction u_ψ^{m} satisfies $y_i \left\langle x_i,\ u_\psi^{\mathsf{m}} \right\rangle \ge \hat{\gamma}_\psi\ \forall i \in [n]$. Then, we have the following bound on $\|w_t\|_\psi$.

Lemma 18 *For exponential loss, the iterates of* MD *satisfies* $\|w_t\|_\psi \in \Theta(\log t)$. *In particular, we have*

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

The proof of this lemma can be found in Appendix F.

It follows that mirror descent with homogeneous potential has a poly-logarithmic rate of convergence.

Corollary 19 *For exponential loss, the iterates of* MD *have convergence rate*

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

As we previously discussed, the Bregman divergence $D_\psi(\cdot,\cdot)$ generalizes the Euclidean distance to the geometry induced by the potential function ψ. Therefore, the quantity $D_\psi\left(u_\psi^{\mathsf{r}}, w_t/\|w_t\|_\psi\right)$ the most natural measure of the angle between the generalized regularized direction u_ψ^{r} and the MD iterates w_t. For the case of gradient descent, our result recovers the rate derived in (Soudry et al., 2018).

3.3 Accelerated convergence with variable step size

The result in the previous section shows that MD with a fixed step size converges to the maximum-margin solution with a poly-logarithmic rate. In this section, we demonstrate an accelerated convergence rate with adaptive step sizes. In particular, we consider a form of normalized mirror descent:

$$\nabla\psi(w_{t+1}) = \nabla\psi(w_t) - \eta_t \frac{\nabla L(w_t)}{\|\nabla L(w_t)\|_{\psi,*}}.$$

Although the quantity $\|\nabla L(w)\|_{\psi,*}$ may not be simple to compute, we claim that it differs from $L(w)$ by at most a constant. Recall that $\hat{\gamma}_\psi = \max_{\|w\|_\psi=1}\{\min_i y_i w^\top x_i\}$. Hence,

$$\|\nabla L(w)\|_{\psi,*} = \max_{\|v\|_\psi=1} \sum_{i=1}^n \exp(-y_i w^\top x_i) y_i v^\top x_i \geq \hat{\gamma}_\psi \sum_{i=1}^n \exp(-y_i w^\top x_i) = \hat{\gamma}_\psi L(w).$$

For an upper bound on $\|\nabla L(w)\|_{\psi,*}$, we note that, by Assumption 2, x_i's are bounded:

$$\|\nabla L(w)\|_{\psi,*} = \left\| \sum_{i=1}^n \exp(-y_i w^\top x_i) y_i x_i \right\|_{\psi,*} \leq \sum_{i=1}^n \exp(-y_i w^\top x_i) \|x_i\|_{\psi,*} \leq C \cdot L(w).$$

From this observation, we will replace $\|\nabla L(w_t)\|_{\psi,*}$ with the more readily available quantity $L(w_t)$. By letting $\eta_t = \frac{\eta_0}{\sqrt{t+1}}$, we now consider the update rule

$$\nabla\psi(w_{t+1}) = \nabla\psi(w_t) - \frac{\eta_0}{\sqrt{t+1}} \frac{\nabla L(w_t)}{L(w_t)}, \tag{N-MD}$$

which is a direct mirror descent extension of the normalized gradient descent algorithm studied in (Nacson et al., 2019). In Appendix D, we will discuss the choice of step sizes when we use the exponential loss.

To prove the rate of convergence, we follow a similar strategy as we did in the case of fixed step size. First, we show that the magnitude of the iterates w_t can be upper bounded.

Lemma 20 *For exponential loss, the iterates of the normalized mirror descent* (N-MD) *satisfies*

$$\limsup_{t\to\infty} \frac{\|w_t\|_\psi}{\sqrt{t}} \le \hat{\gamma}_\psi^{-1} \frac{\beta}{\beta-1}.$$

And with this lemma, we can proceed to show a lower bound on the magnitude of w_t.

Lemma 21 *For exponential loss, if the potential function satisfies Assumption 1 with $\beta < 3$ and is continuously twice differentiable, and the initial loss is sufficiently small that $L(w_0) \le \frac{1}{2n}$, then the iterates of the normalized mirror descent* (N-MD) *satisfies $\|w_t\|_\psi \in \Omega(t^{(3-\beta)/2-\zeta})$ for any $\zeta > 0$.*

These two lemmas together represent a normalized mirror descent counterpart of Lemma 18. Then, we can apply Lemmas 20 and 21 to the same proof technique as Theorem 13. Compared to the fixed step size case, the main difficulty here lies in that we no longer obtain a telescoping sum like (8) when the step size is not constant. We overcome this difficulty by controlling the terms that do not telescope are of lower order than $D_\psi(c_T u_\psi^{\mathsf{r}}, w_T)$ using the asymptotic growth rate on $\|w_t\|_\psi$ for a certain range of β. In particular, we show that such terms vanish after normalization. Hence, the normalized mirror descent algorithm converges to the same implicit bias as MD, but at a much faster rate.

Theorem 22 *For a separable linear classification problem with exponential loss, if the potential function satisfies Assumption 1 with $1 < \beta < \frac{1}{2}(3+\sqrt{5})$ and is continuously twice differentiable, and the initial loss is sufficiently small that $L(w_0) \le \frac{1}{2n}$, then normalized mirror descent* (N-MD) *converges to the generalized maximum-margin direction at a rate*

$$D_\psi\left(u_\psi^{\mathsf{m}}, \frac{w_t}{\|w_t\|_\psi}\right) \in O\left(t^{-(3\beta-1-\beta^2)/2+\beta\zeta}\right),$$

for any $\zeta > 0$.

In the case of gradient descent, the rate becomes $O(t^{-1/2+2\zeta})$. We note that Nacson et al. (2019) presented their convergence rate for normalized gradient descent in terms of the difference in margin: $\hat{\gamma}_2 - \gamma(w_t/\|w_t\|_2) \in O(\log t/\sqrt{t})$. While their result is not directly comparable to our result here, it intuitively matches the rate we derived through Lemma 21 and Theorem 22.

We also note that our result requires a warm-up period to find a small initial loss $L(w_0) \le 1/(2n)$[8]. One way to achieve this is by warm-starting with MD and switching to N-MD after the condition is met. From what we had shown about the convergence rate of MD with a fixed step size (see the proof of Lemma 18), the warm-up period is at most $O(n)$ steps. In practice, we observe that such a warm-start scheme is not necessary.

The proofs for this section can be found in Appendix G. For experiments illustrating faster convergence with normalized MD, please see Section 4.2.

8. This is not necessary when $\beta = 2$, but our proof requires initial loss $L(w_0)$ be small for other values of β.

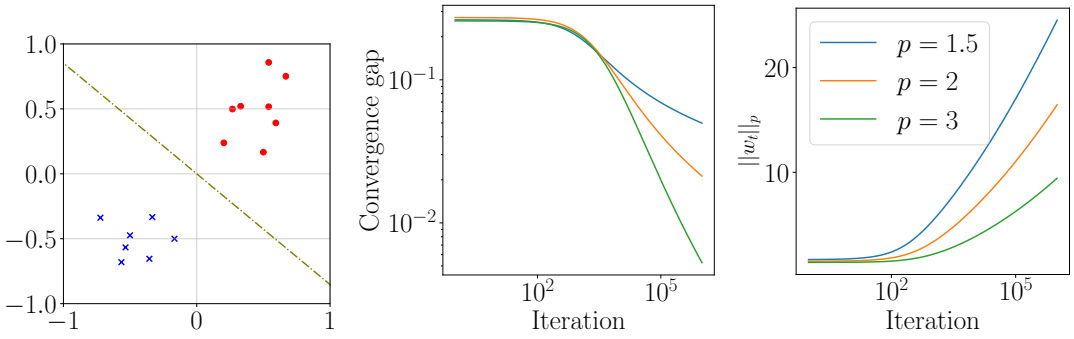


Figure 2: An example of p-GD (MD with potential $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^p$) on randomly generated data with exponential loss and $p = 1.5, 2, 3$. **(1)** The left plot is a scatter plot of the data: $\times$'s and $\bullet$'s denote the two different labels ($y_i = \pm 1$). The dotted line 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 grows. We can observe that the asymptotic behaviors of these plots are consistent with Corollary 19.

4. Experiments

In this section, we perform various experiments to complement our theoretical results in Section 3 and to illustrate the performance of MD in practical settings. We will apply mirror descent (MD) with various homogeneous potentials of the form $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^\beta$, so that, according to our results in Section 3, the iterates converge to the generalized maximum-margin solutions with respect to the ℓ_p-norm. As we previously discussed, when $\beta = p$, the MD update rule can efficiently computed because it is *coordinate-wise separable* (see the update rule (p-GD)). Hence, our experiments also highlight the efficacy of p-GD, which is MD with potential function $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^p$. We naturally pick $p = 2$, which corresponds to gradient descent. Because the mirror map $\nabla\psi$ would not be invertible at $p = 1$ 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. In addition to applying p-GD, we also present several experiments with more general potential functions where $\beta \neq p$. This section will present a summary of our experimental results; the complete experimental setup and full results can be found in Appendices I and J.

4.1 Linear classification

Visualization of the convergence of MD. To visualize the results of Theorem 13 and Corollary 19, we randomly generated a linearly separable set of 15 points in $\mathbb{R}^2$. We then employ MD on this dataset with exponential loss $\ell(z) = \exp(-z)$ and one million iterations at a fixed step size $\eta = 10^{-3}$. For the experiments involving p-GD, we pick $p = 1.5, 2, 3$. And for experiments on more general MD potential, we consider $p = 2, 3$ and $\beta = 1.5, 2, 3$.

In the illustrations of Figure 2, 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 19, the exact rate of convergence for $D_\psi\left(u_p^{\mathsf{m}}, w_t/\|w_t\|_t\right)$ is poly-logarithmic with respect to the number of iterations. And in the third plot of Figure 2,

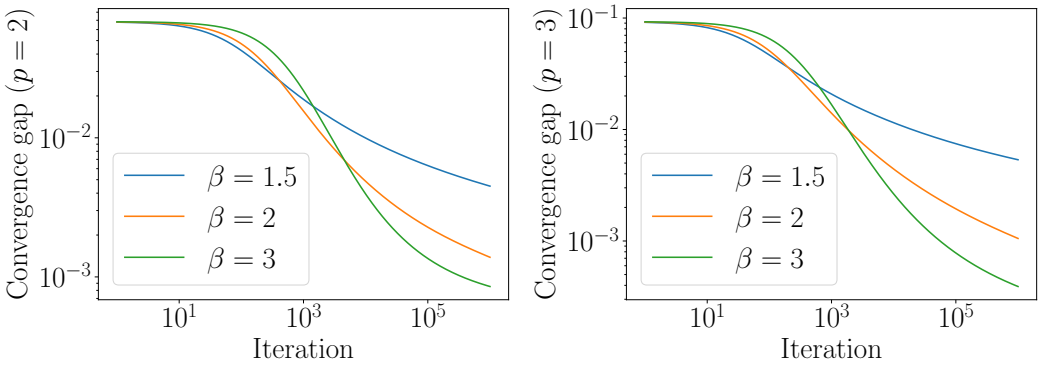


Figure 3: An example of MD with potential $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^\beta$) on the same dataset as in Figure 2. To verify the conclusion of Corollary 19, we plot the quantity $D_\psi\left(u_p^{\mathsf{r}}, w_t/\|w_t\|_t\right)$ for $p = 2$ (**left** figure) and $p = 3$ (**right** figure). We see that the rate of convergence is faster for higher values of the exponent β, which is consistent with Corollary 19.

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}
$p = 1.1$	**5.477**	1.637	1.093	0.696
$p = 2$	6.161	**1.224**	0.684	0.382
$p = 3$	7.299	1.296	**0.667**	0.309
$p = 10$	9.032	1.515	0.740	**0.280**

Table 3: Size of the linear classifiers generated by MD with potential $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^2$ (after rescaling) in $\ell_{1.1}, \ell_2, \ell_3$ and ℓ_{10} norms.

	$\ell_{1.1}$	ℓ_2	ℓ_3	ℓ_{10}
$p = 1.1$	**5.136**	1.864	1.276	0.795
$p = 2$	6.161	**1.224**	0.684	0.382
$p = 3$	7.305	1.275	**0.652**	0.301
$p = 10$	9.132	1.477	0.712	**0.266**

the norm of the iterates w_t grows at a logarithmic rate, which is the same as the prediction by Lemma 18.

In Corollary 19, the convergence rate of MD is dependent on the homogeneity parameter β of the potential function, where larger exponent β leads to faster convergence. We see that this is consistent with our observation in the second plot of Figure 2 and Figure 3. In the second plot of Figure 2, we see that p-GD enjoys faster convergence for larger p, and in Figure 3, we see that for the same value of p, larger exponent β led to faster convergence to the same generalized maximum-margin direction with respect to ℓ_p.

Implicit bias of MD in linear classification. We now verify the conclusions of Theorem 13. To this end, we recall that u_p^{m} is parallel to the generalized 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 and independent of the value of β, the classifier generated by MD with a potential $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^\beta$ 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 selecting 15 points in $\mathbb{R}^{100}$. We used a fixed step size of $\eta = 10^{-4}$ and ran one million iterations for different p's.

Tables 2 and 3 shows the results for $p = 1.1, 2, 3$ and 10; under each norm, we highlight the smallest classifier in **boldface**. In Table 2, we applied p-GD. And in Table 3, we applied MD

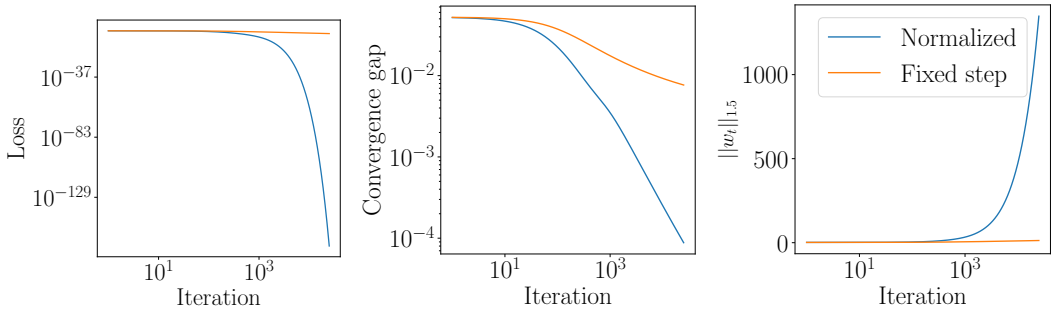


Figure 4: An example of p-GD (MD with potential $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^p$) and normalized p-GD on randomly generated data with exponential loss and $p = 1.5$. **(1)** The left plot is the empirical loss. **(2)** The middle plot shows the rate which the quantity $D_\psi\left(u_p^{\mathsf{r}}, w_t/\|w_t\|_t\right)$ converges to 0. **(3)** The right plot shows how fast the p-norm of w_t grows.

with a fixed exponent $\beta = 2$, which is known as the *p-norm algorithm* in the literature (Grove et al., 2001; Gentile, 2003). 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 MD converges to u_p^{m} at a very slow rate, we can 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. For a more complete result with additional values of p, we refer the readers to Appendix J.1.

4.2 Experiments with normalized MD

We now demonstrate a faster rate of convergence with the normalized mirror descent update (N-MD) from Section 3.3. Because the $\sqrt{t+1}$ term in the update rule (N-MD) would dominate at small t, we rescale the denominator by a factor of $\lambda > 0$ so that

$$\nabla\psi(w_{t+1}) = \nabla\psi(w_t) - \frac{\eta_0}{\sqrt{1+\lambda t}}\frac{\nabla L(w_t)}{L(w_t)}, \tag{9}$$

Here, we present experimental results for p-GD (along with its normalized counterpart) with $p = 1.5$. Additional results for $p = 2$ (which is equivalent to gradient descent) and $p = 2.5$ are deferred to Appendix J.2.

Linear classification. For a clean visualization of the rate of convergence, we again use the 2-dimensional synthetic dataset from Section 4.1. We reuse the same choice of hyper-parameters for mirror descent with a fixed step size. As for normalized mirror descent update (9), we use a base step size $\eta_0 = 10^{-3}$ and scale $\lambda = 10^{-3}$. Since normalized MD converges much more quickly, we only run for 25000 iterations.

As shown in Figure 4, the normalized p-GD algorithm converges to the generalized maximum margin much faster than p-GD with fixed step sizes, which is consistent with the predictions made by Corollary 19 and Theorem 22. Also, we see that the empirical loss decreases very rapidly for normalized p-GD, which means it is able to find classifiers with very large margins in just a few iterations.

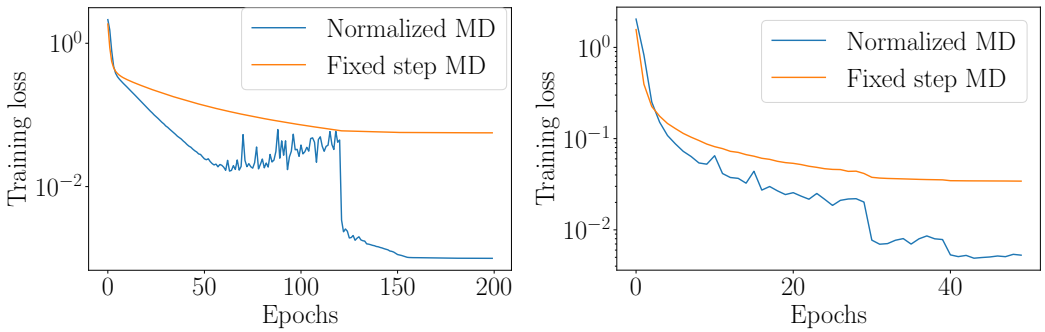


Figure 5: Training loss of p-GD and normalized p-GD on the MNIST dataset and $p = 1.5$. **(1)** The left plot involves a fully connected network. **(2)** The right plot involves a conv-net.

Image classification on MNIST. For a more involved example, we apply p-GD to the MNIST dataset (LeCun et al., 1998). For this task, we use two different architectures: 1) a 2-layer fully connected network with 300 hidden neurons and ReLU activation, and 2) a convolutional network with two convolution layers and batch-norm. We train the fully connected network for 200 epochs and the convolution network for 50 epochs. The detailed specification of this experiment can be found in Appendix I.

As shown in Figure 5, normalized p-GD again achieves faster convergence. Finally, we note that the lower loss achieved by normalized p-GD translates to better generalization performance. For the fully connected network, normalized p-GD has a test accuracy of 98.37%, whereas standard p-GD has a test accuracy of 97.65%. And for the convolutional network, normalized p-GD has a test accuracy of 99.19%, whereas standard p-GD has a test accuracy of 98.68%. These experiments demonstrate that normalized MD enjoys a faster convergence rate and its practical utility in performing better at test time.

4.3 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 the simplicity of implementation, we provide code snippets in Appendix H. Thus, we are able to effectively experiment with the behaviors of 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*[9] 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 6, 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

9. Instead of applying p-GD update with the whole dataset, we use a randomly drawn mini-batch at each iteration.

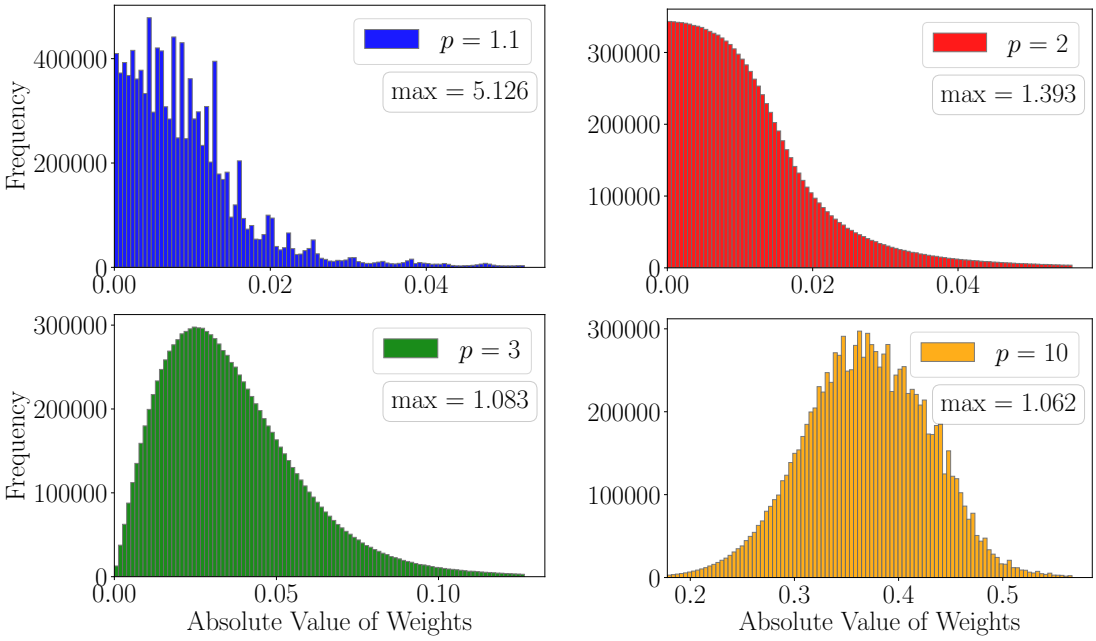


Figure 6: The histogram of weights in ResNet-18 models trained with p-GD (MD with potential $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^p$) for the CIFAR-10 dataset. For clarity, we cropped out the tails and each plot has 100 bins after cropping. The trends of these histograms reflect the implicit biases of 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 4: CIFAR-10 test accuracy (%) of p-GD on various deep neural networks. For each deep network and value of p, the average ± 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 ± .17	92.63 ± .12	91.16 ± .09	91.21 ± .18
$p = 2$ (SGD)	90.15 ± .16	**93.90** ± .14	91.97 ± .10	92.75 ± .13
$p = 3$	**90.85** ± .15	**94.01** ± .13	**93.23** ± .26	**94.07** ± .12
$p = 10$	88.78 ± .37	93.55 ± .21	92.60 ± .22	92.97 ± .16

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 J.3.

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 Table 4 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 understanding them would establish an important step toward understanding the generalization of over-parameterized models. For additional experimental results, we refer the readers to Appendix J.4.

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 J.5.

5. Conclusion and Future Work

In this paper, we provided a unifying view of controlling implicit regularization using mirror descent. More specifically, we analyzed the implicit regularization of the mirror descent algorithm for linear classification problems with strictly monotone losses (e.g. logistics and exponential losses) and showed that for the general class of homogeneous potential functions, mirror descent converges in direction to the generalized regularized/max-margin direction. This result, along with prior literature Gunasekar et al. (2018) and Azizan and Hassibi (2019a), shows that mirror descent can induce implicit regularization with respect to a general geometry for both regression and classification problems. Besides gradient descent, no other algorithm is known to exhibit implicit regularization in both settings. Hence, this work completes the analysis of the first optimization algorithm that can control the implicit regularization for both general geometry and different classes of loss functions. Finally, we ran several experiments to corroborate our theoretical findings and to illustrate the practical applications of mirror descent. 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:

- As discussed in Section 4.3, different choices of p for the p-GD algorithm result in different generalization performance. It would be interesting to develop a theory that explains under what conditions can we show that certain MD potentials lead to better generalization performance.
- Another interesting line of work is to extend this analysis to more sophisticated optimization algorithms such as Adam or RMSprop. While we proved the convergence for a very simple adaptive step size strategy in Section 3.3, it would be interesting to see if such analysis can be strengthened to cover the algorithms most commonly used in practice.

Acknowledgments

The authors thank Christos Thrampoulidis for insightful discussions, his valuable feedback, and his involvement in an earlier version of this work. The authors thank former MIT UROP students Tiffany Huang and Haimoshri Das for contributing to the experiments in Section 4.3. 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, the MIT-IBM Watson AI Lab, and Amazon.

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*, pages 3960–3965. IEEE, 2019b.

Navid Azizan, Sahin Lale, and Babak Hassibi. Explicit regularization via regularizer mirror descent. In *Over-parameterization Workshop at the International Conference on Machine Learning*, 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. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

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.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

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.

Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. *arXiv preprint arXiv:2006.07322*, 2020.

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, 2019a.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *arXiv preprint arXiv:1909.12292*, 2019b.

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.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. 2020.

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.

Poorya Mianjy, Raman Arora, and Rene Vidal. On the implicit bias of dropout. In *International conference on machine learning*, pages 3540–3548. PMLR, 2018.

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *The Journal of Machine Learning Research*, 22(1):10104–10172, 2021.

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.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

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.

Haoyuan Sun, Kwangjun Ahn, Christos Thrampoulidis, and Navid Azizan. Mirror descent maximizes generalized margin and can be implemented efficiently. In *Advances in Neural Information Processing Systems*, volume 35, pages 31089–31101, 2022.

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

Gal Vardi, Ohad Shamir, and Nati Srebro. On margin maximization in linear and relu networks. *Advances in Neural Information Processing Systems*, 35:37024–37036, 2022.

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.

Colin Wei, Sham Kakade, and Tengyu Ma. The implicit and explicit regularization effects of dropout. In *International conference on machine learning*, pages 10181–10192. PMLR, 2020.

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.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109:467–492, 2020.

Appendix A. Extended Version of Table 1

Table 5: **Conceptual summary of our results.** In the case of well-specified linear regression, there is a complete theory of implicit regularization with respect to a general geometry; it is shown that mirror descent converges to the interpolating solution that is closest to the initialization. However, such characterization in the separable linear classification setting is missing in the literature. In this paper, we prove the implicit regularization of mirror descent with a class of homogeneous potentials and extend the result of gradient descent beyond the ℓ_2-norm. Compare to Table 1, here we consider an arbitrary initialization w_0 in the regression setting.

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

Appendix B. Proofs for Section 2

B.1 Proof of Lemma 2

The overall proof follows (Azizan et al., 2021b). We make several modifications to make it better applicable to the classification setting. Note that in the classification setting, there is no $w \in \mathbb{R}^d$ that satisfies $L(w) = 0$.

Proof We start with the definition of Bregman divergence:

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

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

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

We again invoke the definition of Bregman divergence so that:

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

It follows that

$$\begin{aligned} D_\psi(w, w_{t+1}) &= \psi(w) - \psi(w_t) - \langle \nabla\psi(w_t),\ w - w_t\rangle \\ &\quad + \langle \nabla\psi(w_t),\ w_{t+1} - w_t\rangle - \psi(w_{t+1}) + \psi(w_t) \\ &\quad + \eta\langle \nabla L(w_t),\ w - w_{t+1}\rangle \\ &= D_\psi(w, w_t) - D_\psi(w_{t+1}, w_t) + \eta\langle \nabla L(w_t),\ w - w_{t+1}\rangle \end{aligned} \tag{10}$$

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

$$\begin{aligned} \langle \nabla L(w_t),\ w - w_{t-1}\rangle &= \langle \nabla L(w_t),\ w - w_t\rangle - \langle \nabla L(w_t),\ w_{t+1} - w_t\rangle \\ &\quad + L(w_{t+1}) - L(w_t) - L(w_{t+1}) + L(w_t) \\ &= \langle \nabla L(w_t),\ w - w_t\rangle + D_L(w_{t+1}, w_t) - L(w_{t+1}) + L(w_t), \end{aligned} \tag{11}$$

where the last step holds because L is convex.

Combining (10) and (11) yields:

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

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. $D_L(w, w_t) = L(w) - L(w_t) - \langle \nabla L(w_t),\ w - w_t\rangle$:

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

This completes the proof of Lemma 2. ■

B.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. ■

B.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. ■

Appendix C. Properties of Potential Functions under Assumption 1

In this section, we shall establish several useful properties for potential functions satisfying Assumption 1. First, we will show that $\|\cdot\|_\psi$, as defined in (2), is a valid norm induced by the potential ψ.

- Since ψ is positive definite, $\|w\|_\psi = 0$ only when $w = 0$.
- By applying homogeneity and the definition of $\|\cdot\|_\psi$, we have

$$\begin{aligned} \|sw\|_\psi &= \inf\{c > 0 : \psi(sw/c) \le 1\} \\ &= \inf\left\{c > 0 : \left|\frac{s}{|s|}\right|^\beta \psi\left(\frac{w}{c/|s|}\right) \le 1\right\} \\ &= \inf\left\{c > 0 : \psi\left(\frac{w}{c/|s|}\right) \le 1\right\} = |s| \cdot \|w\|_\psi \end{aligned}$$

- To show the triangle inequality, we consider any vectors $w_1, w_2 \in \mathbb{R}^d$ and let $a = \|w_1\|_\psi$ and $b = \|w_2\|_\psi$. Because ψ is convex, we have

$$\psi\left(\frac{w_1 + w_2}{a+b}\right) = \psi\left(\frac{a}{a+b}\cdot\frac{w_1}{a} + \frac{b}{a+b}\cdot\frac{w_2}{b}\right) \le \frac{a}{a+b}\cdot\psi\left(\frac{w_1}{a}\right) + \frac{b}{a+b}\cdot\psi\left(\frac{w_2}{b}\right) = 1.$$

  Therefore, $\|w_1 + w_2\|_\psi \le a + b$, as desired.

Therefore, $\|\cdot\|_\psi$ is a norm. Note that, due to continuity, we have $\psi(w/\|w\|_\psi) = 1$. It follows that, by homogeneity, we can write $\psi(\cdot) = \|\cdot\|_\psi^\beta$.

Next, we show that convexity and β-absolute homogeneity imply strict convexity. So, we can in fact relax the conditions in Assumption 1 where ψ is only required to be convex. Another consequence of this fact is that the potential $\psi(\cdot) = \|\cdot\|^\beta$ for any norm $\|\cdot\|$ satisfies the conditions in Assumption 1. For any $\lambda \in (0, 1)$, we have

$$\|\lambda x + (1-\lambda)y\|_\psi^\beta \le (\lambda\|x\|_\psi + (1-\lambda)\|y\|_\psi)^\beta < \lambda\|x\|_\psi^\beta + (1-\lambda)\|y\|_\psi^\beta$$

The first inequality is due to the triangle inequality and the second inequality holds because the map $z \mapsto |z|^\beta$ is strictly convex whenever $\beta > 1$. Therefore, ψ is strictly convex.

By appealing to the limit definition of gradient, we can show (5), so that the Bregman divergence is also homogeneous.

$$\begin{aligned}
\langle \nabla\psi(w),\ w\rangle &= \lim_{h\to 0}\frac{\psi(w+hw)-\psi(w)}{h}\\
&= \lim_{h\to 0}\frac{(1+h)^\beta-1}{h}\psi(w) = \beta\cdot\psi(w)\\
D_\psi\left(cw, cw'\right) &= \psi(cw)-\psi(cw')-\left\langle\nabla\psi(cw'),\ c(w-w')\right\rangle\\
&= |c|^\beta\psi(w)-|c|^\beta\psi(w')-\lim_{h\to 0}\frac{\psi(cw'+hc(w-w'))-\psi(cw')}{h}\\
&= |c|^\beta\psi(w)-|c|^\beta\psi(w')-\lim_{h\to 0}|c|^\beta\frac{\psi(w'+h(w-w'))-\psi(w')}{h}\\
&= |c|^\beta\psi(w)-|c|^\beta\psi(w')-|c|^\beta\left\langle\nabla\psi(w'),\ w-w'\right\rangle\\
&= |c|^\beta D_\psi\left(w,w'\right)\quad \forall c\in\mathbb{R}.
\end{aligned}$$

Finally, through the triangle inequality, we can show that $\nabla\psi(w)$ and w are "parallel" in the sense that

$$|\langle\nabla\psi(w),\ w\rangle| \geq |\langle\nabla\psi(w),\ v\rangle|, \forall\, v \text{ s.t. } \|v\|_\psi = \|w\|_\psi\,. \tag{12}$$

Proof From the limit definition, we have

$$\langle\nabla\psi(w),\ v\rangle = \lim_{h\to 0}\frac{\psi(w+hv)-\psi(w)}{h}.$$

For any $h>0$, we have

$$\psi(w+hv) = \|w+hv\|_\psi^\beta \leq (\|w\|_\psi + h\|w\|_\psi)^\beta = \psi((1+h)w),$$

and similarly,

$$\psi(w+hv) = \|w+hv\|_\psi^\beta \geq (\|w\|_\psi - h\|w\|_\psi)^\beta = \psi((1-h)w).$$

Therefore, we have that

$$\lim_{h\to 0}\frac{\psi(w-hw)-\psi(w)}{h} \leq \lim_{h\to 0}\frac{\psi(w+hv)-\psi(w)}{h} \leq \lim_{h\to 0}\frac{\psi(w+hw)-\psi(w)}{h},$$

and hence

$$\langle\nabla\psi(w),\ -w\rangle \leq \langle\nabla\psi(w),\ v\rangle \leq \langle\nabla\psi(w),\ w\rangle \implies |\langle\nabla\psi(w),\ v\rangle| \leq |\langle\nabla\psi(w),\ w\rangle|.$$

■

Appendix D. Discussion of Step Sizes

In this section, we provide the details for Remark 5. In particular, we discuss the existence of small step size η such that $\psi-\eta L$ is convex. We break down this discussion into two parts; first, we analyze the case of standard mirror descent update (MD) with fixed step size, and secondly, we consider the case of normalized mirror descent (N-MD) with time-varying step sizes. For concreteness, we focus on the exponential loss $\ell(z)=\exp(-z)$.

D.1 MD with fixed step size

We first use Assumption 2 to directly bound the Hessian $\nabla^2 L$:

$$\left\|\nabla^2 L(w)\right\|_2 = \left\|\frac{1}{n}\sum_{i=1}^n \exp(-y_i w^\top x_i) x_i x_i^\top\right\|_2 \le \sum_{i=1}^n \exp(-y_i w^\top x_i) C^2 = C^2 L(w)$$

It follows that if ψ is μ-strongly convex, then $\eta < \frac{\mu}{C^2 L(w_0)}$ ensures that $\psi - \eta L$ is convex at w_0. By applying Lemma 3 and induction, we can conclude that $\psi - \eta L$ is convex at all iterates $\{w_t\}_{t=0}^\infty$. For instance, gradient descent falls under this case because it has a 1-strongly convex potential $\psi(\cdot) = \frac{1}{2}\|\cdot\|_2^2$.

However, for any $\beta \neq 2$, ψ is not strongly convex, and we shall consider the cases where $\beta \in (1,2)$ and $\beta > 2$ separately. When $\beta \in (1,2)$, we invoke the following fact:

Lemma 23 *Consider a function $\psi : \mathbb{R}^d \to \mathbb{R}$ satisfying Assumption 1,*

- *If there exists $m > 0$ so that $\inf_{\|w\|_2=1} \left\|\nabla^2\psi(w)\right\|_2 \ge m$, then $\left\|\nabla^2\psi(w)\right\|_2 \ge m\|w\|_2^{\beta-2}$.*
- *If there exists $M > 0$ so that $\sup_{\|w\|_2=1} \left\|\nabla^2\psi(w)\right\|_2 \le M$, then $\left\|\nabla^2\psi(w)\right\|_2 \le M\|w\|_2^{\beta-2}$.*

The matrix norm used here is the operator norm induced by the ℓ_2 vector norm.

It follows that $\left\|\nabla^2\psi(w)\right\|_2 \in \Omega(\|w\|_2^{\beta-2})$ as long as it is uniformly positive on the unit circle. Then, since $\left\|\nabla^2 L(w)\right\|_2$ decays exponentially with respect to $\|w\|_2$, there exists $\eta > 0$ so that $\psi - \eta L$ is convex for all iterates $\{w_t\}_{t=0}^\infty$ and this η can be computed from a finite horizon over the iterates. For instance, the p-GD potential $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^p$ satisfies the desired condition on $\nabla^2\psi$ whenever $p \in (1,2)$.

For $\beta > 2$, we alternatively consider the potential $\psi'(\cdot) = \frac{\varepsilon}{2}\|\cdot\|_2^2 + \psi(\cdot)$ for some small $\varepsilon > 0$. Intuitively, this modified potential should induce the same implicit bias as the original potential since their asymptotic tails are the same. Formally, because Bregman divergence is additive, we can plug ψ' into the proof of Theorem 13 and show that mirror descent with potential ψ' converges to u_ψ^r. Since ψ' is strongly convex, the existence of step size η is assured. In practice, we find that directly applying mirror descent with potential ψ already works well.

D.2 Normalized MD with variable step sizes

We claim that there exists sufficiently small $\eta_0 > 0$ so that $\eta_t = \frac{\eta_0}{\sqrt{t+1}} L(w_t)^{-1}$ satisfies $\psi - \eta_t L$ being locally convex at all iterates $\{w_t\}_{t=0}^\infty$. Recall that $\left\|\nabla^2 L(w)\right\|_2 \le C^2 L(w)$. If ψ is μ-strongly convex, then $\eta_0 < \frac{\mu}{C^2}$ guarantees that $\psi - \eta_t L$ is convex at the iterates $\{w_t\}_{t=0}^\infty$. Therefore, our analysis from the previous section still applies for $\beta \ge 2$.

And for $\beta \in (1,2)$, we apply a more careful analysis. We first leverage the upper bound on $\|w\|_\psi$ from Lemma 20. From Assumption 1, we can show that $\inf_{\|w\|_2=1}\psi(w)$ is positive. Therefore,

$$\|w_t\|_2 \le \left(\inf_{\|w\|_2=1}\psi(w)\right)^{-1/\beta}\|w_t\|_\psi \in O\left(\left(\inf_{\|w\|_2=1}\psi(w)\right)^{-1/\beta}\sqrt{t}\right) = O(\sqrt{t}).$$

Combining with Lemma 23, we have that $\left\|\nabla^2\psi(w_t)\right\|_2 \in \Omega(\sqrt{t}^{\beta-2}) \subseteq \omega(1/\sqrt{t})$ for $\beta \in (0,1)$. Since we have

$$\eta_t \left\|\nabla^2 L(w_t)\right\|_2 \leq \frac{\eta_0}{\sqrt{t+1}} C^2 \in O(1/\sqrt{t}),$$

there exists sufficiently small $\eta_0 > 0$ where $\psi - \eta_t L$ is convex at respective iterates $\{w_t\}_{t=0}^{\infty}$.

D.3 Proof of Lemma 23

Proof We directly compute the derivatives through their limit definition. For any vector v,

$$\begin{aligned}
\nabla\psi(w)^\top v &= \lim_{h\to 0} \frac{\psi(w+hv_1)-\psi(w)}{h} \\
&= \|w\|_2^\beta \lim_{h\to 0} \frac{\psi(w/\|w\|_2 + hv/\|w\|_2) - \psi(w/\|w\|_2)}{h} \\
&= \|w\|_2^\beta \nabla\psi(w/\|w\|_2)^\top (v/\|w\|_2) \\
&= \|w\|_2^{\beta-1} \nabla\psi(w/\|w\|_2)^\top v
\end{aligned}$$

Therefore, $\nabla\psi(w) = \|w\|_2^{\beta-1}\nabla\psi(w/\|w\|_2)$.

As for the second derivative, for any vectors v_1, v_2, we have

$$\begin{aligned}
v_2^\top \nabla^2\psi(w) v_1 &= \lim_{h\to 0} \frac{\nabla\psi(w+hv_2)^\top v_1 - \nabla\psi(w)^\top v_1}{h} \\
&= \|w\|_2^{\beta-1} \lim_{h\to 0} \frac{\nabla\psi(w/\|w\|_2 + hv_2/\|w\|_2)^\top v_1 - \nabla\psi(w/\|w\|_2)^\top v_1}{h} \\
&= \|w\|_2^{\beta-1} (v_2/\|w\|_2)^\top \nabla^2\psi(w/\|w\|_2) v_1 \\
&= \|w\|_2^{\beta-2} v_2^\top \nabla^2\psi(w/\|w\|_2) v_1
\end{aligned}$$

Therefore, $\nabla^2\psi(w) = \|w\|_2^{\beta-2}\nabla^2\psi(w/\|w\|_2)$. Now, this Lemma immediately follows from the computation above. ■

Appendix E. Proofs for Section 3.1

E.1 Proof of Lemma 14

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

$$\left\| \frac{\bar{w}(\|w\|_\psi)}{\|w\|_\psi} - u_\psi^{\mathsf{r}} \right\|_\psi \leq \frac{\alpha\bar{\gamma}}{C}$$

whenever $\|w\|_\psi \geq r_\alpha$. Recall that, from Assumption 2, $C \geq \max_{i=1,\dots,n} \|x_i\|_{\psi,*}$. Then, for all $i \in [n]$, we have

$$\begin{aligned} y_i \left\langle \bar{w}(\|w\|_\psi),\ x_i \right\rangle &= y_i \left\langle \bar{w}(\|w\|_\psi) - \|w\|_\psi u^{\mathsf{r}}_\psi,\ x_i \right\rangle + y_i \left\langle \|w\|_\psi u^{\mathsf{r}},\ x_i \right\rangle \\ &\leq \alpha\bar{\gamma} \|w\|_\psi \|x_i\|_{\psi,*} / C + y_i \left\langle \|w\|_\psi u^{\mathsf{r}}_\psi,\ x_i \right\rangle \\ &\leq \alpha\bar{\gamma} \|w\|_\psi + y_i \left\langle \|w\|_\psi u^{\mathsf{r}}_\psi,\ x_i \right\rangle \\ &\leq y_i \left\langle (1+\alpha) \|w\|_\psi u^{\mathsf{r}}_\psi,\ x_i \right\rangle \end{aligned}$$

Since the loss L is decreasing, we have

$$L((1+\alpha) \|w\|_\psi u^{\mathsf{r}}_\psi) \leq L(\bar{w}(\|w\|_\psi)) \leq L(w)\,,$$

as desired. ■

E.2 Lower bounding the mirror descent updates

Lemma 24 *For any potential ψ satisfying Assumption 1, the mirror descent update satisfies the following inequality:*

$$(\beta-1)\psi(w_{t+1}) - (\beta-1)\psi(w_t) + \eta L(w_{t+1}) - \eta L(w_t) \leq \langle -\eta \nabla L(w_t),\ w_t \rangle \tag{13}$$

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

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

Rearranging the terms yields

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

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

$$D_\psi(0, w) = \psi(0) - \psi(w) - \langle \nabla\psi(w),\ -w \rangle = \langle \nabla\psi(w),\ w \rangle - \psi(w) = (\beta-1)\psi(w)\,,$$

where the last equality follows from the homogeneity of Bregman divergence (5). ■

E.3 Proof of Theorem 13

Proof Consider arbitrary $\alpha \in (0,1)$ and define r_α according to Lemma 14. Since $\lim_{t\to\infty} \|w_t\|_\psi = \infty$, we can find t_0 so that $\|w_t\|_\psi > \max(1, r_\alpha)$ for all $t \geq t_0$. Let $c_t = (1+\alpha)\|w_t\|_\psi$.

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

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

By Corollary 15, we have $\left\langle \nabla L(w_t),\ c_t u_\psi^{\mathsf{r}} - w_t\right\rangle \le 0$. Therefore,

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

It follows that

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

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

$$\begin{aligned}
D_\psi\left(c_T u_\psi^{\mathsf{r}}, w_T\right) \le D_\psi\left(c_{t_0} u_\psi^{\mathsf{r}}, w_{t_0}\right) - \eta L(w_T) + \eta L(w_{t_0}) + \psi(c_T u_\psi^{\mathsf{r}}) - \psi(c_{t_0} u_\psi^{\mathsf{r}})\\
- \sum_{t=t_0}^{T-1} \left\langle \nabla\psi(w_{t+1}),\ (c_{t+1} - c_t) u_\psi^{\mathsf{r}}\right\rangle
\end{aligned} \tag{14}$$

Now we want to establish a lower bound on the last term of (14). To do so, we inspect the change in $\nabla\psi(w_t)$ from each successive mirror descent update:

$$\begin{aligned}
&\left\langle \nabla\psi(w_{t+1}) - \nabla\psi(w_t),\ u_\psi^{\mathsf{r}}\right\rangle && \text{(15a)}\\
&= \left\langle -\eta\nabla L(w_t),\ u_\psi^{\mathsf{r}}\right\rangle && \text{(15b)}\\
&\ge \frac{1}{(1+\alpha)\left\|w_t\right\|_\psi} \left\langle -\eta\nabla L(w_t),\ w_t\right\rangle && \text{(15c)}\\
&\ge \frac{1}{(1+\alpha)\left\|w_t\right\|_\psi} \left((\beta-1)\left\|w_{t+1}\right\|_\psi^\beta - (\beta-1)\left\|w_t\right\|_\psi^\beta + \eta L(w_{t+1}) - \eta L(w_t)\right) && \text{(15d)}\\
&\ge \frac{1}{(1+\alpha)\left\|w_t\right\|_\psi} \left((\beta-1)\left\|w_{t+1}\right\|_\psi^\beta - (\beta-1)\left\|w_t\right\|_\psi^\beta\right) + \eta L(w_{t+1}) - \eta L(w_t) && \text{(15e)}
\end{aligned}$$

where we applied Corollary 15 on (15c) and Lemma 24 on (15d).

Now we bound (15e). We claim the following identity and defer its derivation to Section E.4.

$$(\beta-1)(\left\|w_{t+1}\right\|_\psi^\beta - \left\|w_t\right\|_\psi^\beta) \ge \beta(\left\|w_{t+1}\right\|_\psi^{\beta-1} - \left\|w_t\right\|_\psi^{\beta-1})\left\|w_t\right\|_\psi. \tag{16}$$

We are left with

$$\left\langle \nabla\psi(w_{t+1}) - \nabla\psi(w_t),\ u_\psi^{\mathsf{r}}\right\rangle \ge \beta \cdot \frac{\left\|w_{t+1}\right\|_\psi^{\beta-1} - \left\|w_t\right\|_\psi^{\beta-1}}{1+\alpha} + \eta L(w_{t+1}) - \eta L(w_t).$$

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

$$\left\langle \nabla\psi(w_T) - \nabla\psi(w_{t_0}),\ u_\psi^{\mathsf{r}}\right\rangle \ge \beta \cdot \frac{\left\|w_T\right\|_\psi^{\beta-1} - \left\|w_0\right\|_\psi^{\beta-1}}{1+\alpha} + \eta L(w_T) - \eta L(w_{t_0}). \tag{17}$$

With (17), we can bound the last term of (14) as follows:

$$\begin{aligned}\sum_{t=t_0}^{T-1}\left\langle \nabla\psi(w_{t+1}),\ (c_{t+1}-c_t)u_\psi^{\mathsf{r}}\right\rangle &\geq \sum_{t=t_0+1}^{T}\frac{\beta\left\|w_t\right\|_\psi^{\beta-1}+O(1)}{1+\alpha}(c_t-c_{t-1})\\ &=\sum_{t=t_0+1}^{T}\beta(\left\|w_t\right\|_\psi^{\beta-1}+O(1))(\left\|w_t\right\|_\psi-\left\|w_{t-1}\right\|_\psi)\\ &\geq \sum_{t=t_0+1}^{T}(\left\|w_t\right\|_\psi^{\beta}-\left\|w_{t-1}\right\|_\psi^{\beta})+O(1)\cdot(\left\|w_T\right\|_\psi-\left\|w_{t_0}\right\|_\psi)\\ &=\left\|w_T\right\|_\psi^{\beta}+O(\left\|w_T\right\|_\psi) \end{aligned} \tag{18}$$

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

We now apply the inequality in (18) to (14). Note that $\psi(c_T u^{\mathsf{r}}) = (1+\alpha)^\beta \left\|w_T\right\|_\psi^\beta$. We now have the following:

$$D_\psi\left((1+\alpha)\left\|w_T\right\|_\psi u_\psi^{\mathsf{r}}, w_T\right)\leq \left\|w_T\right\|_\psi^\beta\left((1+\alpha)^\beta-1\right)+O(\left\|w_T\right\|_\psi). \tag{19}$$

After applying homogeneity of Bregman divergence, and define $\varepsilon\in(0,1)$ so that $\alpha=\frac{\varepsilon}{1-\varepsilon}$, we have

$$D_\psi\left(u_\psi^{\mathsf{r}},(1-\varepsilon)\frac{w_T}{\left\|w_T\right\|_\psi}\right)\leq\frac{\left\|w_T\right\|_\psi^\beta(1-(1-\varepsilon)^\beta)}{\left\|w_T\right\|_\psi^\beta}+o(1). \tag{20}$$

Let $\tilde{w}_T=\frac{w_T}{\left\|w_T\right\|_\psi}$. We note that Bregman divergence in fact satisfies the Law of Cosines (see, e.g., Azizan and Hassibi (2019b)):

Lemma 25 (Law of Cosines)

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

Therefore,

$$\begin{aligned}D_\psi\left(u_\psi^{\mathsf{r}},\tilde{w}_T\right)&\leq\frac{\left\|w_T\right\|_\psi^\beta(1-(1-\varepsilon)^\beta)}{\left\|w_T\right\|_\psi^\beta}+D_\psi\left((1-\varepsilon)\tilde{w}_T,\tilde{w}_T\right)\\ &\qquad-\left\langle\nabla\psi(\tilde{w}_T)-\nabla\psi((1-\varepsilon)\tilde{w}_T),\ u_\psi^{\mathsf{r}}-(1-\varepsilon)\tilde{w}_T\right\rangle+o(1)\\ &\leq(1-(1-\varepsilon)^\beta)+((1-\varepsilon)^\beta-1)+\beta\varepsilon+2\beta(1-(1-\varepsilon)^{\beta-1})+o(1)\end{aligned} \tag{21}$$

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

$$\limsup_{T\to\infty}D_\psi\left(u_\psi^{\mathsf{r}},\frac{w_T}{\left\|w_T\right\|_\psi}\right)\leq\beta\varepsilon+2\beta(1-(1-\varepsilon)^{\beta-1}) \tag{22}$$

Note that the RHS vanishes in the limit as $\varepsilon\to0$. Since the choice of ε is arbitrary, we have $w_T/\left\|w_T\right\|_\psi\to u_\psi^{\mathsf{r}}$ as $T\to\infty$.

■

Remark 26 *Because Bregman divergence is additive, we can extend this theorem to the case where the potential can be written in the form of $\psi = \psi_1 + \psi_2$, where ψ_1 and ψ_2 satisfy the conditions of Assumption 1 with homogeneity constants β_1 and β_2, respectively, and $\beta_1 < \beta_2$. In the end, all terms associated with ψ_1 would vanish because they are of lower order and we will be left with $D_{\psi_2}(u^{\mathsf{r}}_{\psi_2}, \frac{w_T}{\|w_T\|_{\psi_2}}) \to 0$.*

E.4 Auxiliary Computation for Section E.3

To show (16), we claim that for $\delta \geq -1$ and $\beta > 1$, we have

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

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

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

which is negative when $\delta \in [-1, 0)$ and positive when $\delta > 0$, so this identity holds. Now, (16) follows from setting $\delta = (\|w_{t+1}\|_\psi - \|w_t\|_\psi)/\|w_t\|_\psi$ and then multiplying by $\beta \cdot \|w_t\|_\psi^\beta$ on both sides.

To finish showing (18), we claim that for $\delta \geq -1$ and $\beta > 1$, we have

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

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

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

which is positive when $\delta \in [-1, 0)$ and negative when $\delta > 0$, so this identity holds. Now, the last inequality of (18) follows by setting $\delta = (\|w_t\|_\psi - \|w_{t-1}\|_\psi)/\|w_{t-1}\|_\psi$ and then multiply by $\beta \cdot \|w_t\|_\psi^\beta$ on both sides.

Finally, we simplify the RHS of (21) by taking advantage of the fact that $\tilde{w}_T$ is normalized:

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

And note that for any vector v and $\varepsilon \in (0,1)$, we have

$$\begin{aligned} \langle \nabla\psi((1-\varepsilon)\tilde{w}_T),\ v\rangle &= \frac{1}{1-\varepsilon}\langle \nabla\psi((1-\varepsilon)\tilde{w}_T),\ (1-\varepsilon)v\rangle \\ &= \frac{1}{1-\varepsilon}\lim_{h\to 0}\frac{\psi((1-\varepsilon)\tilde{w}_T + h(1-\varepsilon)v) - \psi((1-\varepsilon)\tilde{w}_T)}{h} \\ &= (1-\varepsilon)^{\beta-1}\lim_{h\to 0}\frac{\psi(\tilde{w}_T + hv) - \psi((1-\varepsilon)\tilde{w}_T)}{h} \\ &= (1-\varepsilon)^{\beta-1}\langle \nabla\psi(\tilde{w}_T),\ v\rangle. \end{aligned}$$

It follows that

$$\begin{aligned}
&\left|\left\langle \nabla\psi(\tilde{w}_T) - \nabla\psi((1-\varepsilon)\tilde{w}_T),\ u_\psi^{\mathsf{r}} - (1-\varepsilon)\tilde{w}_T \right\rangle\right| \\
&= (1-(1-\varepsilon)^{p-1}) \left|\left\langle \nabla\psi(\tilde{w}_T),\ u_\psi^{\mathsf{r}} - (1-\varepsilon)\tilde{w}_T \right\rangle\right| \\
&\le (1-(1-\varepsilon)^{\beta-1}) \left(\left|\left\langle \nabla\psi(\tilde{w}_T),\ u_\psi^{\mathsf{r}} \right\rangle\right| + \left|\left\langle \nabla\psi(\tilde{w}_T),\ (1-\varepsilon)\tilde{w}_T \right\rangle\right|\right) \\
&\le (1-(1-\varepsilon)^{\beta-1}) \left(\left|\left\langle \nabla\psi(\tilde{w}_T),\ \tilde{w}_T \right\rangle\right| + \left|\left\langle \nabla\psi(\tilde{w}_T),\ (1-\varepsilon)\tilde{w}_T \right\rangle\right|\right) \\
&\le (1-(1-\varepsilon)^{\beta-1})\beta(2-\varepsilon) \le 2\beta(1-(1-\varepsilon)^{\beta-1})
\end{aligned}$$

where the second to last line follows from (12).

E.5 Proof of Proposition 16

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

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

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

Define $B' > 0$ so that $\ell(z)e^{az} \in [b/2, 2b]$ for any $z = B\hat{\gamma}_\psi$ where $B > B'$. Note that

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

Suppose the contrary that the regularized direction does not converge to u_ψ^{m}, then there must exist $\varepsilon \in (0, \hat{\gamma}_\psi/2)$ so that there are arbitrarily large values of B satisfying

$$\min_{i=1,\dots,n} y_i \left\langle \frac{\bar{w}(B)}{B},\ x_i \right\rangle \le \hat{\gamma}_\psi - \varepsilon.$$

And this implies

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

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

Appendix F. Proofs for Section 3.2

F.1 Proof of Corollary 17

Proof This is an immediate consequence of (19) and (20). ■

F.2 Proof of Lemma 18

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 $w^\star = \hat{\gamma}_\psi^{-1} u_\psi^{\mathsf{m}}$. By the definition of the max-margin direction, the margin of $w^\star$ is 1 and $\|w^\star\|_\psi = \hat{\gamma}_\psi^{-1}$. 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:

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

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

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

Due to the homogeneity of Bregman divergence with respect to the β-absolutely homogeneous potential ψ, we can divide by a factor of $\log^\beta T$ on both sides:

$$D_\psi\left(w^\star, \frac{w_0}{\log T}\right) \ge D_\psi\left(w^\star, \frac{w_T}{\log T}\right) - o(1). \tag{23}$$

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

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

where the inequality follows from (12).

If $\|w_T/\log T\|_\psi > \hat{\gamma}_\psi^{-1} \cdot \frac{\beta}{\beta-1}$ for arbitrarily large T, then $D_\psi(w^\star, w_T/\log T) > \hat{\gamma}_\psi^{-\beta}$ for those T. This in turn contradicts inequality (23). Therefore, we must have

$$\limsup_{T\to\infty} \frac{\|w_T\|_\psi}{\log T} \le \hat{\gamma}_\psi^{-1} \frac{\beta}{\beta-1}.$$

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,

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

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 $\gamma(\|w_t\|_\psi u_\psi^{\mathsf{m}}) \ge m_t$. Then by linearity of margin, there exists $w^\star$ so that $\gamma(w^\star) = (1 + \frac{\log(2n)}{m_t})m_t$ and $\|w^\star\|_\psi \le (1 + \frac{\log(2n)}{m_t})\|w_t\|_\psi$. It follows that

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

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 standard convergence rate of mirror descent (Lu et al., 2018, Theorem 3.1):

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

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

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

After dropping the lower order terms and recalling the upper bounds on $\|w^\star\|_\psi$ and $\|w_t\|_\psi$, we have

$$\frac{1}{2n}\exp(-m_t) \le O(1)\cdot\frac{1}{\eta t}\cdot\left(1+\frac{\log(2n)}{m_t}\right)^\beta\left(\hat{\gamma}_\psi^{-1}\frac{\beta}{\beta-1}\log t\right)^\beta$$

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

$$m_t \ge \log t - \beta\log\log t + O(1)$$

Finally, we use the definition of margin to conclude that $m_t \le \langle w_t,\ x_i\rangle \le C\cdot\|w_t\|_\psi$. Therefore,

$$\|w_t\|_\psi \ge \frac{1}{C}(\log t - \beta\log\log t) + O(1).$$

■

Appendix G. Proofs for Section 3.3

G.1 Proof of Lemma 20

This proof follows the same technique as Lemma 18. For simplicity, we assume without loss of generality that $y_i = 1$ for all i by replacing every data point of the form $(x_i, -1)$ with $(-x_i, 1)$.

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

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

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

$$\begin{aligned}
&\left\langle \nabla L(w_t),\ w^\star\sqrt{T} - w_t \right\rangle \\
&= \sum_{i=1}^n \left\langle \nabla_w \exp(-\langle w,\ x_i\rangle)\Big|_{w=w_t},\ w^\star\sqrt{T} - w_t \right\rangle \\
&= \sum_{i=1}^n \left\langle \exp(-\langle w_t,\ x_i\rangle) x_i,\ w_t - w^\star\sqrt{T} \right\rangle \\
&= \sum_{i=1}^n \exp\left(-\left\langle w^\star\sqrt{T},\ x_i\right\rangle\right) \exp\left(-\left\langle w_t - w^\star\sqrt{T},\ x_i\right\rangle\right) \left\langle x_i,\ w_t - w^\star\sqrt{T}\right\rangle \\
&\le \sum_{i=1}^n \exp(-\sqrt{T}) \cdot \frac{1}{e} \in o\left(\frac{1}{T}\right),
\end{aligned}$$

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

$$\begin{aligned}
D_\psi(w^\star\sqrt{T}, w_t) &\ge D_\psi(w^\star\sqrt{T}, w_{t+1}) - o(1/T) - \eta_t L(w_t) + \eta_t L(w_{t+1}) \\
&\ge D_\psi(w^\star\sqrt{T}, w_{t+1}) - o(1/T) - \frac{\eta_0}{\sqrt{t+1}}.
\end{aligned}$$

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

$$D_\psi(w^\star\sqrt{T}, w_0) \ge D_\psi(w^\star\sqrt{T}, w_T) - O(\sqrt{T}).$$

Due to the homogeneity of Bregman divergence with respect to the β-absolutely homogeneous potential ψ, we can divide by a factor of $T^{\beta/2}$ on both sides:

$$D_\psi\left(w^\star, \frac{w_0}{\sqrt{T}}\right) \ge D_\psi\left(w^\star, \frac{w_T}{\sqrt{T}}\right) - o(1). \tag{24}$$

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

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

where the inequality follows from (12).

If $\left\|w_T/\sqrt{T}\right\|_\psi > \hat{\gamma}_\psi^{-1}\cdot\frac{\beta}{\beta-1}$ for arbitrarily large T, then $D_\psi\left(w^\star, w_T/\sqrt{T}\right) > \hat{\gamma}_\psi^{-\beta}$ for those T. This in turn contradicts inequality (24). Therefore, we must have

$$\limsup_{T\to\infty}\frac{\|w_T\|_\psi}{\sqrt{T}} \leq \hat{\gamma}_\psi^{-1}\frac{\beta}{\beta-1}\,,$$

as desired. ■

G.2 Proof of Lemma 21

Proof Given sufficiently small η_0, we have that $\psi - \eta_t L$ is convex. From relatively smoothness (Lu et al., 2018, Proposition 1.1), we have

$$L(w_{t+1}) \leq L(w_t) + \langle \nabla L(w_t),\ w_{t+1} - w_t\rangle + \frac{1}{\eta_t}D_\psi(w_{t+1}, w_t).$$

By applying the mirror descent update rule (MD) to the RHS, the following holds for any w:

$$L(w_{t+1}) \leq L(w_t) + \langle \nabla L(w_t),\ w - w_t\rangle + \frac{1}{\eta_t}D_\psi(w, w_t). \tag{25}$$

Let Δw be the vector satisfying

$$-\langle \nabla L(w_t),\ \Delta w\rangle = \|\nabla L(w_t)\|_{\psi,*}\|\Delta w\|_\psi = \|\nabla L(w_t)\|_{\psi,*}^2 = \|\Delta w\|_\psi^2\,.$$

Then, we choose the vector w so that $w - w_t = \eta_t(t+1)^{-c/2}\Delta w$ for some constant $c > 0$ which we will determine later.

Next, we bound the Bregman divergence with Lagrange's remainder:

$$\psi(w) \leq \psi(w_t) + \langle \nabla\psi(w_t),\ w - w_t\rangle + \underbrace{\frac{1}{2}\sup_{\lambda\in(0,1)}(w-w_t)^\top\nabla^2\psi(w_t + \lambda(w-w_t))(w-w_t)}_{R}\,.$$

Note that by construction, $\|\Delta w_t\|_\psi = \|\nabla L(w_t)\|_{\psi,*} \leq C\cdot L(w_t)$. Hence, it holds that

$$\|w - w_t\|_\psi = \eta_t(t+1)^{-c/2}\|\Delta w\|_\psi \leq C\cdot\eta_0\,.$$

Next, when $L(w_0) \le \frac{1}{2n}$, we have

$$\begin{aligned} L(w_t) &= \frac{1}{n}\sum_{i=1}^{n}\exp(-w_t^\top x_i) \ge \frac{1}{n}\exp(-\min_i w_t^\top x_i), \\ \implies & -\log L(w_t) \le \min_i w_t^\top x_i + \log n \le C\cdot\|w_t\|_\psi + \log n, \\ \implies & \|w_t\|_\psi \ge \log 2/C. \end{aligned}$$

Also, from Assumption 1, we can show that for any vector w, the norms $\|w\|_2$ and $\|w\|_\psi$ differ by up to a constant factor:

$$\left(\inf_{\|w\|_2=1}\psi(w)\right)^{-1/\beta}\|w\|_\psi \le \|w\|_2 \le \left(\sup_{\|w\|_2=1}\psi(w)\right)^{1/\beta}\|w\|_\psi . \tag{26}$$

Therefore, for sufficiently small η_0, we have $\|w-w_t\|_2 \le 2\|w_t\|_2$. Applying Lemma 23 yields

$$R \le \frac{1}{2}\|w-w_t\|_2^2 (2\|w_t\|_2)^{\beta-2}\sup_{\|v\|_2=1}\left\|\nabla^2\psi(v)\right\|_2 .$$

Invoking (26) again, there exists a constant $B>0$ so that

$$R \le \frac{B}{2}\|w-w_t\|_\psi^2\|w_t\|_\psi^{\beta-2} = \frac{B}{2}\eta_t^2(t+1)^{-c}\|\Delta w\|_\psi^2\|w_t\|_\psi^{\beta-2} .$$

Because $D_\psi(w,w_t)\le R$, we return to (25) and conclude that

$$\begin{aligned} L(w_{t+1}) &\le L(w_t) - \eta_t(t+1)^{-c/2}\|\nabla L(w_t)\|_{\psi,*}^2 + \frac{B}{2}\eta_t(t+1)^{-c}\|\Delta w\|_\psi^2\|w_t\|_\psi^{\beta-2} \\ &= L(w_t) - \eta_t(t+1)^{-c/2}\|\nabla L(w_t)\|_{\psi,*}^2 + \frac{B}{2}\eta_t(t+1)^{-c}\|\nabla L(w_t)\|_{\psi,*}^2\|w_t\|_\psi^{\beta-2} \\ &= L(w_t)\left(1-\eta_0(t+1)^{-(c+1)/2}\delta_t + \frac{B}{2}\eta_0(t+1)^{-(2c+1)/2}\delta_t\|w_t\|_\psi^{\beta-2}\right) \\ &\le L(w_t)\exp\left(-\eta_0(t+1)^{-(c+1)/2}\delta_t + \frac{B}{2}\eta_0(t+1)^{-(2c+1)/2}\delta_t\|w_t\|_\psi^{\beta-2}\right), \end{aligned}$$

where we let $\delta_t = \left(\frac{\|\nabla L(w_t)\|_{\psi,*}}{L(w_t)}\right)^2$. Since $C\cdot\|w\|_\psi \ge \min_i w^\top x_i \ge -\log L(w)$, we propagate the previous inequality through $t=0,\dots,T-1$ to get

$$\begin{aligned} \|w_T\|_\psi &\ge -\frac{1}{C}\log L(w_T) \\ &\ge \frac{1}{C}\left(\sum_{t=0}^{T-1}\left(\eta_0(t+1)^{-(c+1)/2}\delta_t - \frac{B}{2}\eta_0(t+1)^{-(2c+1)/2}\delta_t\|w_t\|^{\beta-2}\right) - \log L(w_0)\right) \\ &\ge \Theta\left(\sum_{t=0}^{T-1} t^{-(c+1)/2} - t^{-(2c+1)/2}\|w_t\|_\psi^{\beta-2}\right), \end{aligned}$$

where the second inequality holds as $\hat\gamma_\psi L(w) \le \|\nabla L(w)\|_{\psi,*} \le C\cdot L(w) \implies \delta_t$ bounded.

Recall that Lemma 20 gives us $\|w_t\|_\psi \in O(t^{-1/2})$. Picking $c = \max(\beta - 2, 0) + 2\zeta$ for arbitrarily small $\zeta > 0$ gives us

$$\|w_t\|_\psi \in \Omega\left(\sum_{t=0}^{T-1} t^{(1-\beta)/2-\zeta} - t^{(1-\beta)/2-2\zeta}\right) \subseteq \Omega\left(t^{(3-\beta)/2-\zeta}\right),$$

as desired. ■

G.3 Proof of Theorem 22

This proof mostly follows the same steps as that of Theorem 13.

Proof Consider arbitrary $\alpha \in (0,1)$ and define r_α according to Lemma 14. Since $\lim_{t\to\infty} \|w_t\|_\psi = \infty$, we can find t_0 so that $\|w_t\|_\psi > \max(1, r_\alpha)$ for all $t \geq t_0$. Let $c_t = (1+\alpha)\|w_t\|_\psi$.

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

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

By Corollary 15, we have $\left\langle \nabla L(w_t),\ c_t u_\psi^{\mathsf{r}} - w_t \right\rangle \leq 0$. Therefore,

$$\begin{aligned} D_\psi\left(c_t u_\psi^{\mathsf{r}}, w_{t+1}\right) &\leq D_\psi\left(c_t u_\psi^{\mathsf{r}}, w_t\right) - \eta_t L(w_{t+1}) + \eta_t L(w_t) \\ &\leq D_\psi\left(c_t u_\psi^{\mathsf{r}}, w_t\right) + \eta_0 (t+1)^{-1/2}. \end{aligned}$$

It follows that

$$\begin{aligned} &D_\psi\left(c_{t+1} u_\psi^{\mathsf{r}}, w_{t+1}\right) \\ &\leq D_\psi\left(c_t u_\psi^{\mathsf{r}}, w_t\right) + \eta_0 (t+1)^{-1/2} + D_\psi\left(c_{t+1} u_\psi^{\mathsf{r}}, w_{t+1}\right) - D_\psi\left(c_t u_\psi^{\mathsf{r}}, w_{t+1}\right) \\ &= D_\psi\left(c_t u_\psi^{\mathsf{r}}, w_t\right) + \eta_0 (t+1)^{-1/2} + \psi(c_{t+1} u_\psi^{\mathsf{r}}) - \psi(c_t u_\psi^{\mathsf{r}}) - \left\langle \nabla\psi(w_{t+1}),\ (c_{t+1} - c_t) u_\psi^{\mathsf{r}} \right\rangle \end{aligned}$$

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

$$\begin{aligned} D_\psi\left(c_T u_\psi^{\mathsf{r}}, w_T\right) &\leq D_\psi\left(c_{t_0} u_\psi^{\mathsf{r}}, w_{t_0}\right) + \sum_{t=t_0}^{T-1} \eta_0 (t+1)^{-1/2} + \psi(c_T u_\psi^{\mathsf{r}}) - \psi(c_{t_0} u_\psi^{\mathsf{r}}) \\ &\quad - \sum_{t=t_0}^{T-1} \left\langle \nabla\psi(w_{t+1}),\ (c_{t+1} - c_t) u_\psi^{\mathsf{r}} \right\rangle \\ &= D_\psi\left(c_{t_0} u_\psi^{\mathsf{r}}, w_{t_0}\right) + O(\sqrt{T}) + \psi(c_T u_\psi^{\mathsf{r}}) - \psi(c_{t_0} u_\psi^{\mathsf{r}}) \\ &\quad - \sum_{t=t_0}^{T-1} \left\langle \nabla\psi(w_{t+1}),\ (c_{t+1} - c_t) u_\psi^{\mathsf{r}} \right\rangle \end{aligned} \tag{27}$$

Now we want to establish a lower bound on the last term of (27). To do so, we inspect the change in $\nabla\psi(w_t)$ from each successive mirror descent update:

$$\begin{aligned}
&\left\langle \nabla\psi(w_{t+1}) - \nabla\psi(w_t),\ u_\psi^{\mathsf{r}} \right\rangle && \text{(28a)}\\
&= \left\langle -\eta_t \nabla L(w_t),\ u_\psi^{\mathsf{r}} \right\rangle && \text{(28b)}\\
&\geq \frac{1}{(1+\alpha)\left\|w_t\right\|_\psi} \left\langle -\eta_t \nabla L(w_t),\ w_t \right\rangle && \text{(28c)}\\
&\geq \frac{1}{(1+\alpha)\left\|w_t\right\|_\psi} \Big((\beta-1)\left\|w_{t+1}\right\|_\psi^\beta - (\beta-1)\left\|w_t\right\|_\psi^\beta + \eta_t L(w_{t+1}) - \eta_t L(w_t) \Big) && \text{(28d)}\\
&\geq \frac{1}{(1+\alpha)\left\|w_t\right\|_\psi} \Big((\beta-1)\left\|w_{t+1}\right\|_\psi^\beta - (\beta-1)\left\|w_t\right\|_\psi^\beta \Big) - \frac{\eta_t L(w_t)}{\left\|w_t\right\|_\psi} && \text{(28e)}
\end{aligned}$$

where we applied Corollary 15 on (28c) and Lemma 24 on (28d).

Now we bound (28e). We claim the following identity and defer its derivation to Section E.4.

$$(\beta-1)(\left\|w_{t+1}\right\|_\psi^\beta - \left\|w_t\right\|_\psi^\beta) \geq \beta(\left\|w_{t+1}\right\|_\psi^{\beta-1} - \left\|w_t\right\|_\psi^{\beta-1})\left\|w_t\right\|_\psi . \tag{29}$$

From Lemma 21, we have that for any $\zeta > 0$, $\left\|w_t\right\|_\psi \in \Omega(t^{(3-\beta)/2-\zeta})$. Therefore,

$$\frac{\eta_t L(w_t)}{\left\|w_t\right\|_\psi} = \frac{\eta_0 (t+1)^{-1/2}}{\left\|w_t\right\|_\psi} \in O(t^{-2+\beta/2+\zeta}).$$

We are left with

$$\left\langle \nabla\psi(w_{t+1}) - \nabla\psi(w_t),\ u_\psi^{\mathsf{r}} \right\rangle \geq \beta \cdot \frac{\left\|w_{t+1}\right\|_\psi^{\beta-1} - \left\|w_t\right\|_\psi^{\beta-1}}{1+\alpha} - O(t^{-2+\beta/2+\zeta}).$$

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

$$\left\langle \nabla\psi(w_T) - \nabla\psi(w_{t_0}),\ u_\psi^{\mathsf{r}} \right\rangle \geq \beta \cdot \frac{\left\|w_T\right\|_\psi^{\beta-1} - \left\|w_0\right\|_\psi^{\beta-1}}{1+\alpha} - O(T^{-1+\beta/2+\zeta}). \tag{30}$$

With (30), we can bound the last term of (14) as follows:

$$\begin{aligned}
&\sum_{t=t_0}^{T-1} \left\langle \nabla\psi(w_{t+1}),\ (c_{t+1}-c_t) u_\psi^{\mathsf{r}} \right\rangle\\
&\geq \sum_{t=t_0+1}^{T} \frac{\beta\left\|w_t\right\|_\psi^{\beta-1} - O(t^{-1+\beta/2+\zeta})}{1+\alpha} (c_t - c_{t-1})\\
&= \sum_{t=t_0+1}^{T} \beta(\left\|w_t\right\|_\psi^{\beta-1} - O(t^{-1+\beta/2+\zeta}))(\left\|w_t\right\|_\psi - \left\|w_{t-1}\right\|_\psi)\\
&\geq \sum_{t=t_0+1}^{T} (\left\|w_t\right\|_\psi^\beta - \left\|w_{t-1}\right\|_\psi^\beta) - O(T^{-1+\beta/2+\zeta}) \cdot (\left\|w_T\right\|_\psi - \left\|w_{t_0}\right\|_\psi)\\
&= \left\|w_T\right\|_\psi^\beta - O(T^{-1+\beta/2+\zeta}\left\|w_T\right\|_\psi)
\end{aligned} \tag{31}$$

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

We now apply the inequality in (31) to (27). Note that $\psi(c_T u^{\mathsf{r}}) = (1+\alpha)^\beta \|w_T\|_\psi^\beta$. We now have the following:

$$D_\psi\left((1+\alpha)\|w_T\|_\psi u_\psi^{\mathsf{r}}, w_T\right) \le \|w_T\|_\psi^\beta ((1+\alpha)^\beta - 1) + O(\sqrt{T} + T^{-1+\beta/2+\zeta}\|w_T\|_\psi).$$

After applying homogeneity of Bregman divergence, and define $\varepsilon \in (0,1)$ so that $\alpha = \frac{\varepsilon}{1-\varepsilon}$, we have

$$D_\psi\left(u_\psi^{\mathsf{r}}, (1-\varepsilon)\frac{w_T}{\|w_T\|_\psi}\right) \le \frac{\|w_T\|_\psi^\beta (1-(1-\varepsilon)^\beta)}{\|w_T\|_\psi^\beta} + O\left(\frac{\sqrt{T} + T^{-1+\beta/2+\zeta}\|w_T\|_\psi}{\|w_T\|_\psi^\beta}\right).$$

For the remainder term to vanish, we apply Lemma 21 and want that for sufficiently small $\zeta > 0$,

$$\begin{cases} \beta((3-\beta)/2 - \zeta) & > 1/2, \\ (\beta-1)((3-\beta)/2-\zeta) & > -1 + \beta/2 + \zeta. \end{cases}$$

Solving the system gives that we need $1 < \beta < \frac{1}{2}(3+\sqrt{5})$. Therefore, for $1 < \beta < \frac{1}{2}(3+\sqrt{5})$, we have

$$D_\psi\left(u_\psi^{\mathsf{r}}, (1-\varepsilon)\frac{w_T}{\|w_T\|_\psi}\right) \le \frac{\|w_T\|_\psi^\beta (1-(1-\varepsilon)^\beta)}{\|w_T\|_\psi^\beta} + o(1).$$

The lower-order term can be more precisely written as

$$O\left(T^{-(3\beta-1-\beta^2)/2+\beta\zeta}\right),$$

which gives us the rate at which the error term vanishes. Now, to finish showing convergence, the rest of the proof follows exactly as that of Theorem 13. ■

Appendix H. 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 approximately 10% slower than with PyTorch's `optim.SGD` (in the same number of epochs),[10] and we believe that this gap can be closed with optimization.

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):
        # 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, dx = param.data, 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
```

10. This measurement may not be very accurate because we were using shared computing resources.

Appendix I. Experimental Details

All of the following experiments were performed on compute nodes equipped with an Intel Skylake CPU + one Nvidia V100 GPU. The experiments involving linear models were CPU only and experiments with convolutional network models took advantage of the GPU's acceleration.

I.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 fixed step size 10^{-3} 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 have up to 10 nonzero entries. Each nonzero entry is i.i.d. 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.

I.2 Normalized MD experiements

For the first experiment with the synthetic dataset in $\mathbb{R}^2$, we use the same choices of hyper-parameters as above. The only difference is that we only run for 25000 iterations due to faster convergence of normalized MD.

For the MNIST experiments, we used two different models, the first one is a two-layer fully connected neural network with 300 neurons in the hidden layer and ReLU activation and the second one is a convolutional network with two convolution layers (see exact specification next page). For both models, we applied cross-entropy loss and batch size of 512. To avoid numerical issues in the normalized MD update (9), we divide by $\max(L(w_t), 10^{-5})$ instead of the empirical loss directly.

For the fully connected network, we train for 200 epochs in total and use a learning rate schedule that starts with $\eta = 0.1$ and decays by a factor of 5 at the 120th, 150th, and 180th epochs. In the normalized MD update (9), the base step size η_0 follows the same schedule and has scale factor $\lambda = 0.1$. These parameters were chosen to closely match the setup in (Nacson et al., 2019, Section 4.2) when $p = 2$, but our experiments used mini-batch instead to better reflect a practical training scenario.

For the convolutional network, we train for 50 epochs in total and use a learning rate schedule that starts with $\eta = 0.02$ and decays by a factor of 5 at the 30th and 40th epochs. In the normalized MD update (9), the base step size η_0 follows the same schedule and has scale factor $\lambda = 1.0$.

```
==================================================================
Layer                                       Output Shape
==================================================================
|-Conv2d: 3x3 kernel                        [512, 32, 26, 26]
|-BatchNorm2d:                              [512, 32, 26, 26]
|-ReLU:                                     [512, 32, 26, 26]
|-MaxPool2d: 2x2 kernel                     [512, 32, 13, 13]
|-Conv2d: 3x3 kernel                        [512, 32, 11, 11]
|-BatchNorm2d:                              [512, 32, 11, 11]
|-ReLU:                                     [512, 32, 11, 11]
|-MaxPool2d: 2x2 kernel                     [512, 32, 5, 5]
|-Flatten:                                  [512, 800]
|-Linear:                                   [512, 64]
|-ReLU:                                     [512, 64]
|-Linear:                                   [512, 10]
==================================================================
```

I.3 CIFAR-10 experiments

For the experiments with the CIFAR-10 dataset, we adopted the example implementation from the FFCV library.[11] 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 a maximum learning rate of 0.1 and ran for 400 epochs so the training loss is almost equal to 0.[12]

I.4 ImageNet experiments

For the experiments with the ImageNet dataset, we used the example implementation from the FFCV library.[13] 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 a 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.

Appendix J. Additional Experimental Results

J.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 6 and Table 3 is a subset of Table 8, as shown below. Within each trial, we use the same dataset generated from a fixed random seed.

We first observe the results of p-GD in Tables 6 and 7. 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 19 predicts a much slower convergence

11. https://github.com/libffcv/ffcv/tree/main/examples/cifar
12. This differs from the setup from Azizan et al. (2021b), where they used a fixed small learning rate and much larger number of epochs.
13. https://github.com/libffcv/ffcv-imagenet/

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 a 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.

Next, we fix the value of β and look at the results of MD with potential $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^2$ in Tables 8 and 9. We are able to observe the same general trend as we did for p-GD. However, because we fixed the value of the exponent β, the various linear classifiers generated by MD would have similar rates of convergence. We note that in both trials, MD with potential $\psi(\cdot) = \frac{1}{p}\|\cdot\|_{1.1}^2$ led to the smallest classifier in $\ell_{1.1}$-norm, which is consistent with Theorem 13.

Table 6: 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
$p = 1.1$	**7.398**	5.477	2.592	1.637	1.093	0.780	0.696	0.629
$p = 1.5$	7.544	**5.348**	**2.237**	1.296	0.803	0.558	0.514	0.505
$p = 2$	8.985	6.161	2.315	**1.224**	0.684	0.429	0.382	0.366
$p = 3$	10.820	7.299	2.592	1.296	**0.667**	0.369	0.309	0.278
$p = 6$	12.714	8.523	2.957	1.441	0.711	**0.360**	0.281	0.229
$p = 10$	13.484	9.032	3.123	1.515	0.740	0.367	**0.280**	**0.213**

Table 7: 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
$p = 1.1$	10.278	7.731	3.776	2.408	1.610	1.162	1.058	0.973
$p = 1.5$	**8.835**	**6.222**	**2.549**	1.450	0.873	0.577	0.512	0.463
$p = 2$	10.161	6.962	2.609	**1.375**	0.760	0.464	0.406	0.387
$p = 3$	11.681	7.926	2.863	1.449	**0.754**	0.419	0.348	0.316
$p = 6$	13.454	9.083	3.217	1.592	0.797	**0.410**	0.321	0.261
$p = 10$	14.290	9.630	3.392	1.669	0.828	0.417	**0.321**	**0.244**

Table 8: Size of the linear classifiers generated by MD with potential $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^2$ (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 MD with potential $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^2$ 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
$p = 1.1$	**6.526**	**5.136**	2.780	1.864	1.276	0.900	0.795	0.700
$p = 1.5$	7.338	5.231	**2.215**	1.292	0.803	0.552	0.498	0.473
$p = 2$	8.985	6.161	2.315	**1.224**	0.684	0.429	0.382	0.366
$p = 3$	10.871	7.305	2.567	1.275	**0.652**	0.360	0.301	0.276
$p = 6$	12.836	8.553	2.919	1.406	0.687	**0.346**	0.269	0.220
$p = 10$	13.738	9.132	3.091	1.477	0.712	0.349	**0.266**	**0.201**

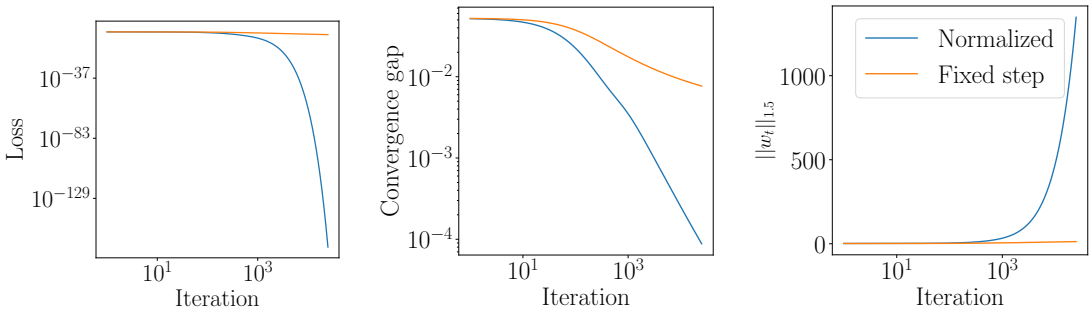
Table 9: Size of the linear classifiers generated by MD with potential $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^2$ (after rescaling) in $\ell_1, \ell_{1.1}, \ell_{1.5}, \ell_2, \ell_3, \ell_6$ and ℓ_{10} norms. For each norm, we highlight the value of p for which MD with potential $\psi(\cdot) = \frac{1}{p}\|\cdot\|_p^2$ 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
$p = 1.1$	**7.277**	**5.646**	2.996	2.032	1.453	1.118	1.039	0.980
$p = 1.5$	8.671	6.133	**2.529**	1.438	0.867	0.582	0.527	0.509
$p = 2$	10.161	6.962	2.609	**1.375**	0.760	0.464	0.406	0.387
$p = 3$	11.783	7.948	2.828	1.420	**0.735**	0.409	0.340	0.306
$p = 6$	13.877	9.272	3.194	1.553	0.767	**0.392**	0.309	0.250
$p = 10$	14.685	9.800	3.360	1.625	0.795	0.397	**0.306**	**0.237**

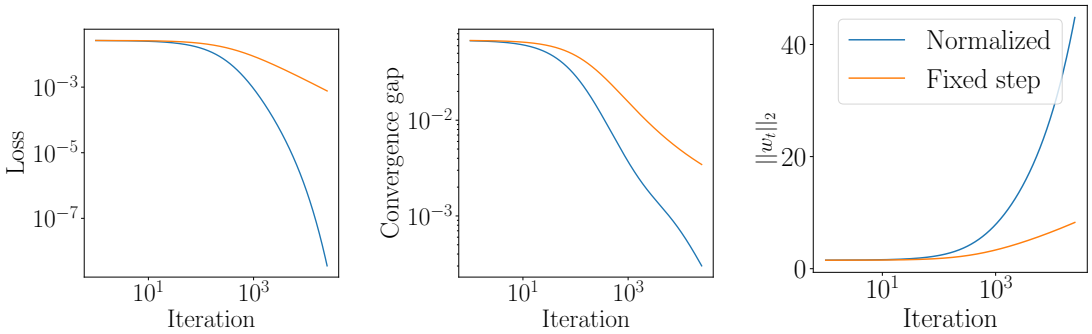
J.2 Experiments with normalized MD

We present additional experiments on normalized MD that expand upon what we presented in Section 4.2. We compare normalized p-GD against the standard p-GD for $p = 1.5, 2$ and 2.5.

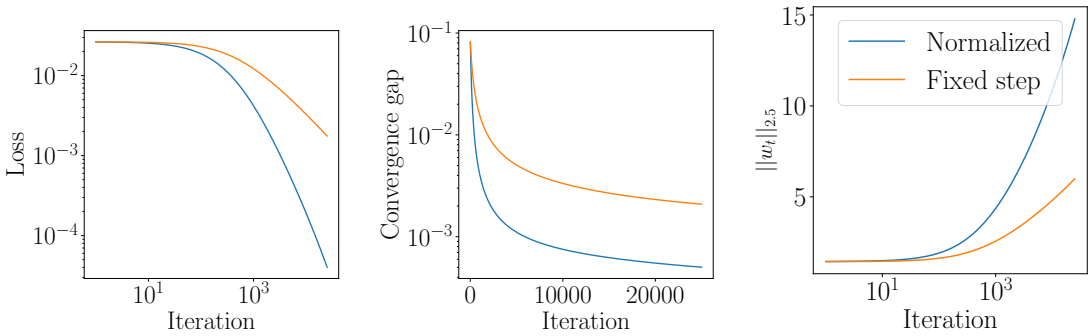
For $p = 1.5$ and 2, normalized p-GD enjoys much faster convergence and we can see that its training loss in both synthetic dataset and MNIST is significantly lower than that of standard p-GD. The picture for $p = 2.5$ is less clear, where normalized p-GD enjoys a similarly sizable advantage in rate of convergence for synthetic dataset while struggling somewhat on the MNIST dataset before learning rate decay kicks in. But in all cases, the final loss achieved by normalized p-GD is much lower and we see that the lower training loss translates to better test performance on the MINST dataset, where normalized p-GD is better than standard p-GD by about 0.2–0.7 percent. These observations are consistent with our analysis on more general normalized mirror descent, as reflected by the statement of Theorem 22.



(a) $p = 1.5$

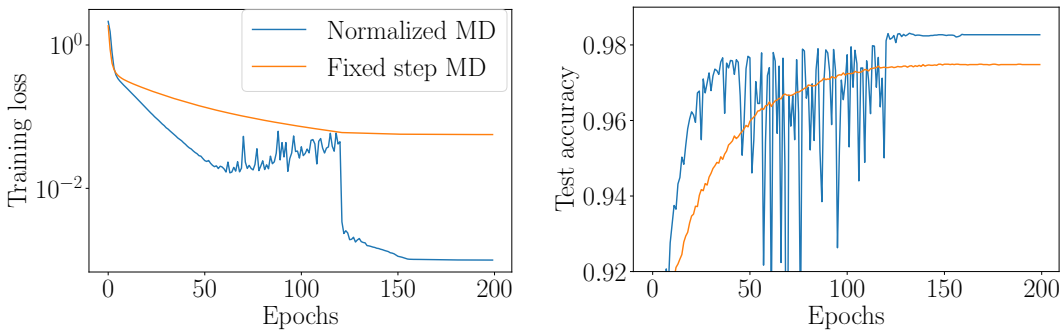


(b) $p = 2$

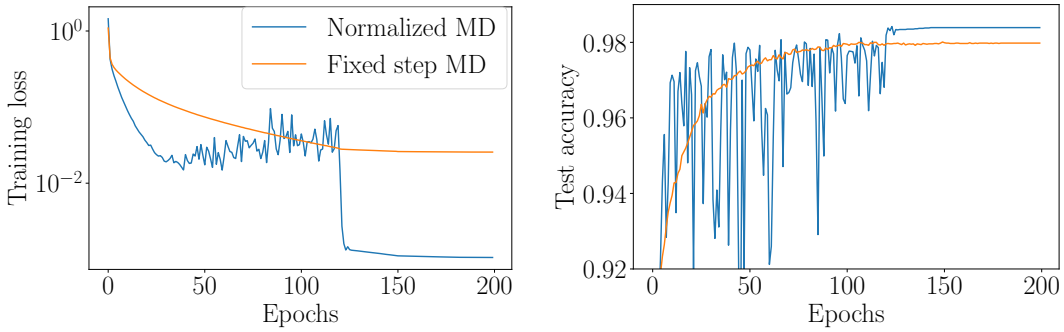


(c) $p = 2.5$

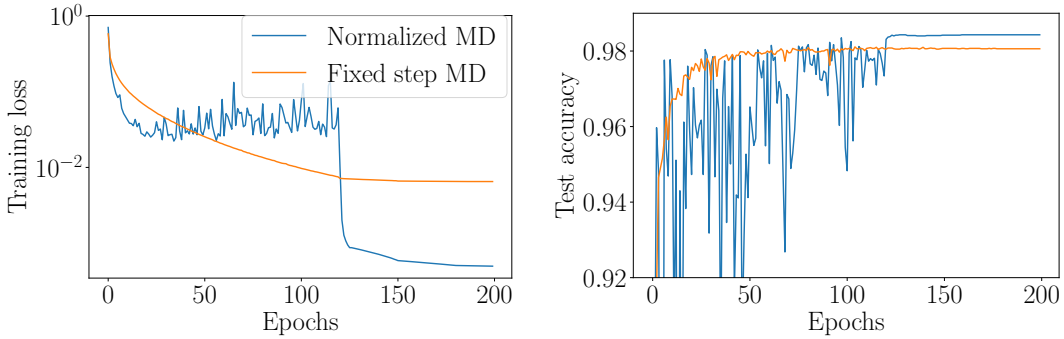
Figure 7: Examples of p-GD and normalized p-GD on randomly generated data with exponential loss and $p = 1.5, 2, 2.5$. **(1)** The left plot is the empirical loss. **(2)** The middle plot shows the rate which the quantity $D_\psi\left(u_p^r, w_t / \|w_t\|_t\right)$ converges to 0. **(3)** The right plot shows how fast the p-norm of w_t growths.



(a) $p = 1.5$



(b) $p = 2$

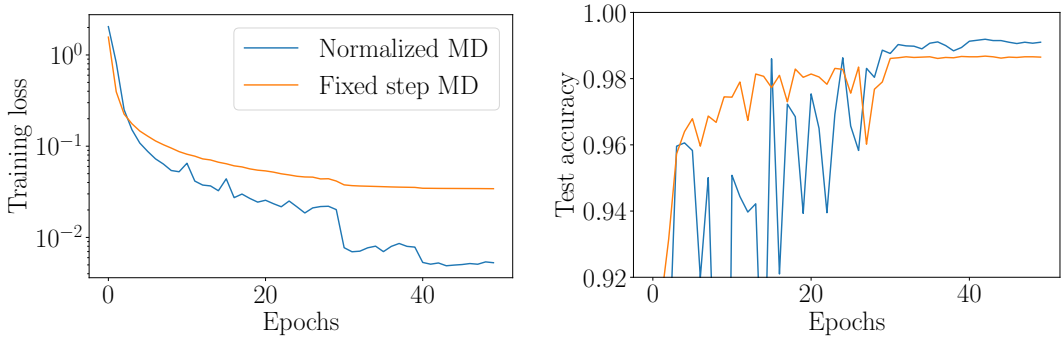


(c) $p = 2.5$

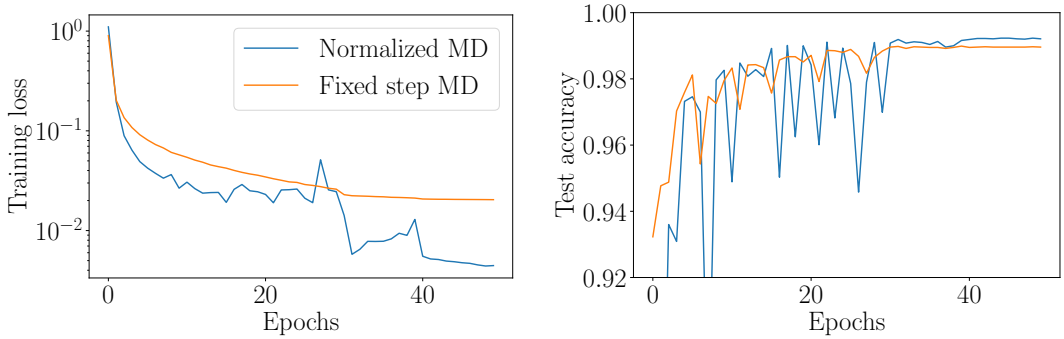
Figure 8: Example of p-GD and normalized p-GD on the MNIST dataset and $p = 1.5, 2, 2.5$. **(1)** The left plot is the empirical loss at training time. **(2)** The right plot is the test accuracy.

Table 10: MNIST accuracy (%) of p-GD versus normalized p-GD for a fully connected network. Note that the normalized version has better generalization for all values of p.

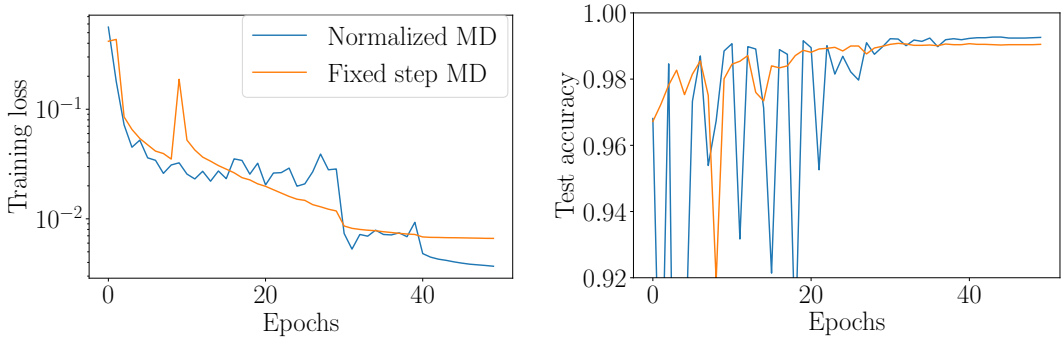
	Mirror descent (p-GD)	Normalized p-GD
$p = 1.5$	97.65	98.37
$p = 2$ (SGD)	97.95	98.34
$p = 2.5$	98.29	98.48



(a) $p = 1.5$



(b) $p = 2$



(c) $p = 2.5$

Figure 9: Example of p-GD and normalized p-GD on the MNIST dataset and $p = 1.5, 2, 2.5$. **(1)** The left plot is the empirical loss at training time. **(2)** The right plot is the test accuracy.

Table 11: MNIST accuracy (%) of p-GD versus normalized p-GD for a convolutional network. Note that the normalized version has better generalization for all values of p.

	Mirror descent (p-GD)	Normalized p-GD
$p = 1.5$	98.68	99.19
$p = 2$ (SGD)	98.99	99.23
$p = 2.5$	99.08	99.27

J.3 CIFAR-10 experiments: implicit bias

We present more complete illustrations of the implicit bias trends of trained models in CIFAR-10. Compared to Figure 6, 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 4.3 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 peaks around zero, and higher p's result in smaller maximum weights.

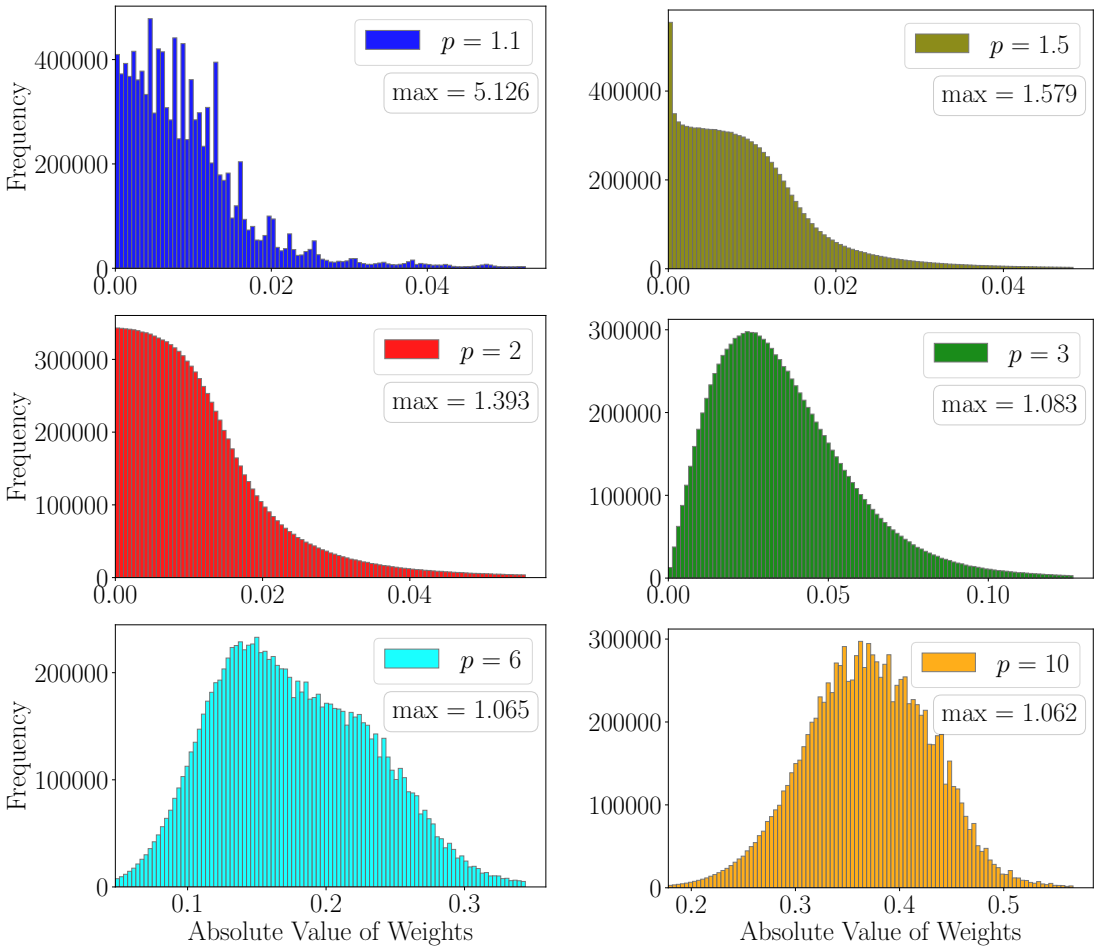


Figure 10: The histogram of weights in ResNet-18 models trained with p-GD for the CIFAR-10 dataset. For clarity, we cropped out the tails and each plot has 100 bins after cropping. Note that the scale on the y-axis differs per graph.

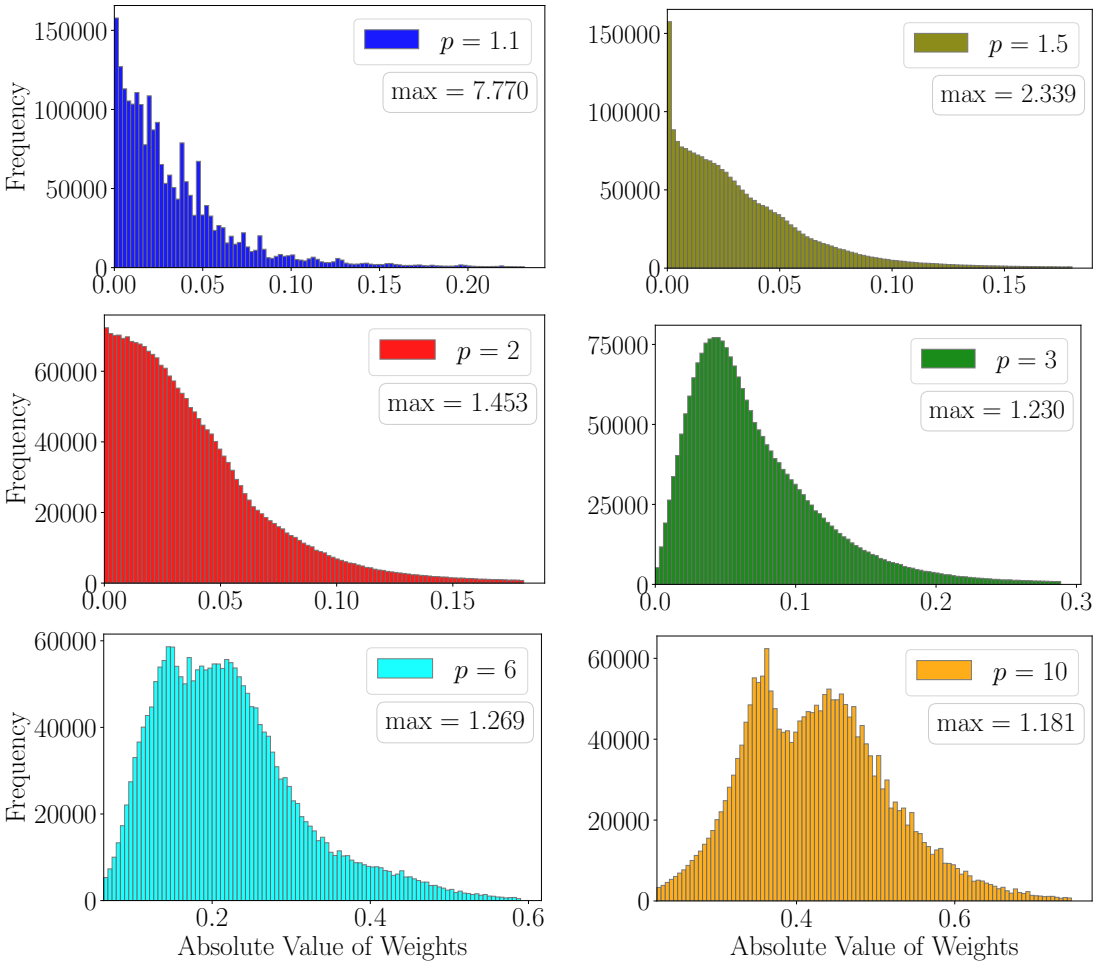


Figure 11: The histogram of weights in MobileNet-v2 models trained with p-GD for the CIFAR-10 dataset. For clarity, we cropped out the tails and each plot has 100 bins after cropping.

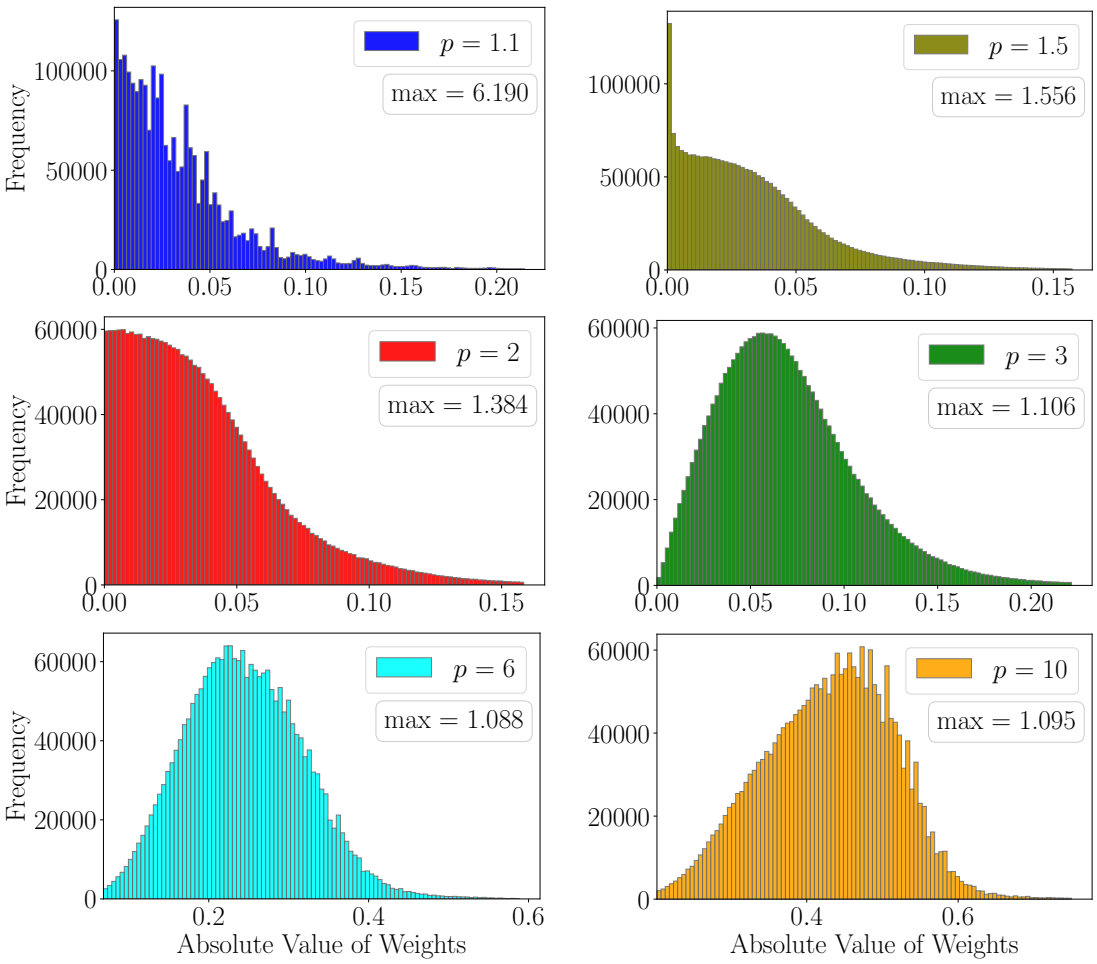


Figure 12: The histogram of weights in RegNetX-200mf models trained with p-GD for the CIFAR-10 dataset. For clarity, we cropped out the tails and each plot has 100 bins after cropping.

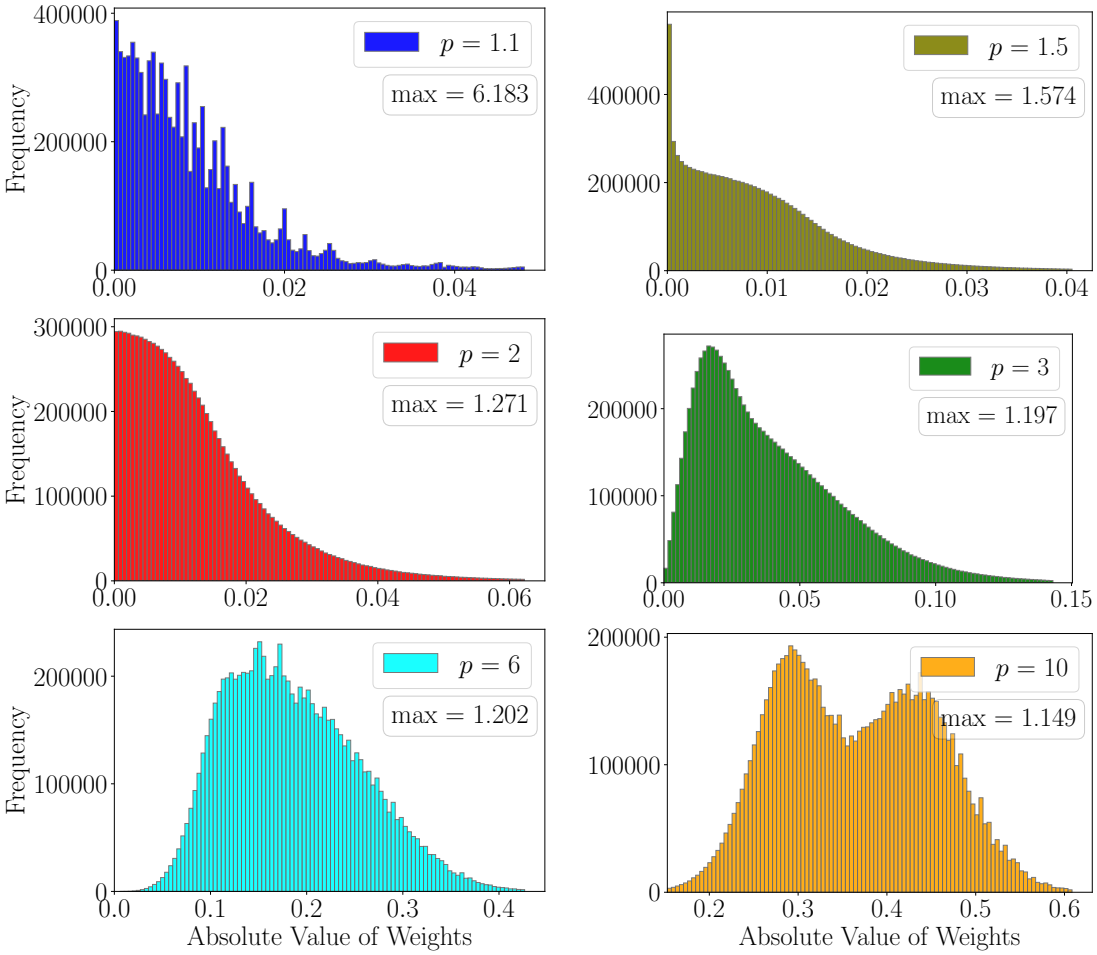


Figure 13: The histogram of weights in VGG-11 models trained with p-GD for the CIFAR-10 dataset. For clarity, we cropped out the tails and each plot has 100 bins after cropping.

J.4 CIFAR-10 experiments: generalization

We present a more complete result for the CIFAR-10 generalization experiment in Section 4.3 with additional values of p.

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

Table 12: CIFAR-10 test accuracy (%) of p-GD on various deep neural networks. For each deep net and value of p, the average ± std. dev. over 5 trials are reported. 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 ± .17	92.63 ± .12	91.16 ± .09	91.21 ± .18
$p = 1.5$	88.45 ± .29	92.73 ± .11	90.81 ± .19	90.91 ± .12
$p = 2$ (SGD)	90.15 ± .16	**93.90** ± .14	91.97 ± .10	92.75 ± .13
$p = 3$	**90.85** ± .15	**94.01** ± .13	**93.23** ± .26	**94.07** ± .12
$p = 6$	89.47 ± .14	**93.87** ± .13	92.84 ± .15	93.03 ± .17
$p = 10$	88.78 ± .37	93.55 ± .21	92.60 ± .22	92.97 ± .16

J.5 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-70s. 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 13: 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