STOCHASTIC MIRROR DESCENT ON OVERPARAMETERIZED NONLINEAR MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Most modern learning problems are highly overparameterized, meaning that the model has many more parameters than the number of training data points, and as a result, the training loss may have infinitely many global minima (in fact, a manifold of parameter vectors that perfectly interpolates the training data). Therefore, it is important to understand which interpolating solutions we converge to, how they depend on the initialization point and the learning algorithm, and whether they lead to different generalization performances. In this paper, we study these questions for the family of stochastic mirror descent (SMD) algorithms, of which the popular stochastic gradient descent (SGD) is a special case. Recently it has been shown that, for overparameterized linear models, SMD converges to the global minimum that is “closest” (in terms of the Bregman divergence of the mirror used) to the initialization point, a phenomenon referred to as implicit regularization. Our contributions in this paper are both theoretical and experimental. On the theory side, we show that in the overparameterized nonlinear setting, if the initialization is close enough to the manifold of global optima, SMD with sufficiently small step size converges to a global minimum that is approximately the closest global minimum in Bregman divergence, thus attaining approximate implicit regularization. For highly overparametrized models, this closeness comes for free: the manifold of global optima is so high dimensional that with high probability an arbitrarily chosen initialization will be close to the manifold. On the experimental side, our extensive experiments on the MNIST and CIFAR-10 datasets, using various initializations, various mirror descents, and various Bregman divergences, consistently confirms that this phenomenon indeed happens in deep learning: SMD converges to the closest global optimum to the initialization point in the Bregman divergence of the mirror used. Our experiments further indicate that there is a clear difference in the generalization performance of the solutions obtained from different SMD algorithms. Experimenting on the CIFAR-10 dataset with different regularizers, $\ell_1$ to encourage sparsity, $\ell_2$ (yielding SGD) to encourage small Euclidean norm, and $\ell_{10}$ to discourage large components in the parameter vector, consistently and definitively shows that, for small initialization vectors, $\ell_{10}$-SMD has better generalization performance than SGD, which in turn has better generalization performance than $\ell_1$-SMD. This surprising, and perhaps counter-intuitive, result strongly suggests the importance of a comprehensive study of the role of regularization, and the choice of the best regularizer, to improve the generalization performance of deep networks.

1 INTRODUCTION

Deep learning has demonstrably enjoyed a great deal of success in a wide variety of tasks (Amodei et al., 2016; Graves et al., 2013; Krizhevsky et al., 2012; Mnih et al., 2015; Silver et al., 2016; Wu et al., 2016; LeCun et al., 2015). Despite its tremendous success, the reasons behind the good performance of these methods on unseen data is not fully understood (and, arguably, remains somewhat of a mystery). While the special deep architecture of these models seems to be important to the success of deep learning, the architecture is only part of the story, and it has been now widely recognized that the optimization algorithms used to train these models, typically stochastic gradient descent (SGD) and its variants, play a key role in learning parameters that generalize well.
Since these deep models are highly overparameterized, they have a lot of capacity, and can fit to virtually any (even random) set of data points (Zhang et al. [2016]). In other words, these highly overparameterized models can “interpolate” the training data, so much so that this regime has been called the “interpolating regime” (Ma et al., 2018b). In fact, on a given dataset, the loss function typically has (infinitely) many global minima, which however can have drastically different generalization properties (many of them perform very poorly on the test set). Which minimum among all the possible minima we converge to in practice is determined by the initialization and the optimization algorithm that we use for training the model.

Since the loss functions of deep neural networks are non-convex—sometimes even non-smooth—in theory, one may expect the optimization algorithms to get stuck in local minima or saddle points. In practice, however, such simple stochastic descent algorithms almost always reach zero training error, i.e., a global minimum of the training loss (Zhang et al. [2016]; Lee et al. [2016]). More remarkably, even in the absence of any explicit regularization, dropout, or early stopping (Zhang et al. [2016]), the global minima obtained by these algorithms seem to generalize quite well (contrary to some other global minima). It has been also observed that even among different optimization algorithms, i.e., SGD and its variants, there is a discrepancy in the solutions achieved by different algorithms and how they generalize (Wilson et al., 2017). Therefore, it is important to ask

**Which global minima do these algorithms converge to? And what properties do they have?**

In this paper, we answer this question for the SGD algorithm, and more generally, for the family of stochastic mirror descent (SMD) algorithms, which includes SGD as a special case. For any choice of potential function, there is a corresponding mirror descent algorithm. We show that, for overparameterized nonlinear models, if one initializes close enough to the manifold of parameter vectors that interpolates the data, then the SMD algorithm for any particular potential converges to a global minimum that is approximately the closest one to the initialization, in Bregman divergence corresponding to the potential. In highly overparameterized models, this closeness of the initialization comes for free, something that is occasionally referred to as “the blessing of dimensionality.” For the special case of SGD, this means that it converges to a global minimum which is approximately the closest one to the initialization in the usual Euclidean sense.

We perform extensive systematic experiments with various initial points and various mirror descent algorithms for the MNIST and CIFAR-10 datasets using standard off-the-shelf deep neural network architectures for these datasets with standard random initialization, and we measure all the resulting pairwise Bregman divergences. We found that every single result is exactly consistent with the above theory. Indeed, in all our experiments, the global minimum achieved by any particular mirror descent algorithm is the closest, among all other global minima obtained by other mirrors and other initializations, to its initialization in the corresponding Bregman divergence. In particular, the global minimum obtained by SGD from any particular initialization is closest to the initialization in Euclidean sense, both among the global minima obtained by different mirrors and among the global minima obtained by different initializations.

This result, proven theoretically and backed up by extensive experiments, further implies that, even in the absence of any explicit regularization, these algorithms perform an implicit regularization. In particular, it implies that, when initialized around zero, SGD acts as an $\ell_2$ regularizer. Similarly, by choosing other mirrors, one obtains different forms of implicit regularization (such as $\ell_1$ or $\ell_\infty$), which may have different performances on test data. This raises the question

**How well do different mirrors perform in practice?**

Perhaps, one might expect an $\ell_1$ regularizer to perform better, due to the fact that it promotes sparsity, and “pruning” in neural networks is believed to be helpful for generalization. On the other hand, one may expect SGD ($\ell_2$ regularizer) to work best among different mirrors, because typical architectures have been tuned for and tailored to SGD. We run experiments with four different mirror descents, i.e., $\ell_1$ (sparse), $\ell_2$ (SGD), and $\ell_{\infty}$ (as a surrogate for $\ell_\infty$), on a standard off-the-shelf deep neural network architecture for CIFAR-10, namely ResNet-18. Somewhat counter-intuitively, our results for test errors of different mirrors consistently and definitively show that the $\ell_{10}$ regularizer performs better than the other mirrors including SGD, while $\ell_1$ consistently performs worse. This flies in the face of the conventional wisdom that sparser weights (which are obtained by an $\ell_1$ regularizer) generalize better, and suggests that $\ell_\infty$, which roughly speaking penalizes all the weights uniformly, may be a better regularizer for deep neural networks.
1.1 Related Work

There have been many efforts in the past few years to study deep learning from an optimization perspective, e.g., (Achille & Soatto [2017], Chaudhari & Soatto [2018], Shwartz-Ziv & Tishby [2017], Allen-Zhu et al. [2019], Oymak & Soltanolkotabi [2019], Azizan & Hassibi [2019], Ma et al. [2018b], Du et al. [2018], Li & Liang [2018], Cao & Gu [2019]). While it is not possible to review all the contributions here, we comment on the ones that are most closely related to ours and highlight the distinctions between our results and those.

Many recent papers have studied the convergence of the (S)GD algorithm in the so-called “overparameterized” setting (or “interpolating” regime), which is common in deep learning (Oymak & Soltanolkotabi [2019], Allen-Zhu et al. [2019], Soltanolkotabi et al. [2017], Ma et al. [2018b]). All these works, similar to ours, assume that the initialization is close to the solution space (of global minima), which is a reasonable assumption in highly overparameterized models, as we argue in Section [A.4] of the supplementary material. However, our results are more general because they extend to SMD.

Furthermore, even for the case of SGD, our results are stronger than those in the literature, in the sense that not only do we show convergence to a global minimum, but we also show that the weight vector we converge to, $w_\infty$, say, is close to the interpolating weight vector closest to the initialization, $w^*$, say. Denoting the initialization by $w_0$, Oymak & Soltanolkotabi [2019] showed that for SGD, $\|w_\infty - w_0\|$ is bounded by a constant factor of $\|w^* - w_0\|$. Our Theorem 4 shows the much stronger statement that $\|w_\infty - w_0\| = \|w^* - w_0\| + o(\|w^* - w_0\|)$. We further show that $w_\infty$ and $w^*$ are very close to one another, viz. $\|w_\infty - w^*\|^2 = o(\|w^* - w_0\|)$, something that could not be inferred from the previous work.

There exist a number of results that characterize the implicit regularization properties of different algorithms in different contexts (Neyshabur et al. [2017], Ma et al. [2018a], Gunasekar et al. [2017, 2018a], Soudry et al. [2017], Gunasekar et al. [2018b], Azizan & Hassibi [2019], Mianjy et al. [2018]). The closest ones to our results, since they concern mirror descent, are the works of Gunasekar et al. [2018a], Azizan & Hassibi [2019]. The authors in Gunasekar et al. [2018a] consider linear overparameterized models, and show that if SMD happens to converge to a global minimum, then that global minimum will be the one that is closest in Bregman divergence to the initialization, a result they obtain by examining the KKT conditions. However, they do not provide any conditions for convergence and whether SMD converges with a fixed step size or not. Azizan & Hassibi [2019] also study linear models, but derive conditions on the step size for which SMD converges to the aforementioned global minimum. Our results extend the aforementioned to nonlinear overparameterized models, and show that, for small enough fixed step size, and for initializations close enough to the space of interpolating solutions, SMD converges to a global minimum, something which had not been shown in any of the previous work. Assuming every data point is revisited often enough, the convergence we establish is deterministic. Finally, we show that the solution we converge to exhibits approximate implicit regularization, something that was not not known for nonlinear models.

The remainder of the paper is organized as follows. In Section 2 we review the family of mirror descent algorithms and briefly revisit the linear overparameterized case. Section 3 provides our main theoretical results, which are (1) convergence of SMD to a global minimum, and (2) proximity of the obtained global minimum to the closest one from the initialization in Bregman divergence. Our proofs are remarkably simple and are based on a powerful fundamental identity that holds for all SMD algorithms in a general setting. In Section 4 we provide our experimental results, which consists of two parts, (1) testing the theoretical claims about the distances for different mirrors and different initializations, and (2) assessing the generalization properties of different mirrors. The proofs of the theoretical results and more details on the experiments are relegated to the supplementary material.

2 Background and Warm-Up

Let us denote the training dataset by $\{(x_i, y_i) : i = 1, \ldots, n\}$, where $x_i \in \mathbb{R}^d$ are the inputs, and $y_i \in \mathbb{R}$ are the labels. The model (which can be, e.g., linear, a deep neural network, etc.) is defined by the general function $f(x, w) = f_i(w)$ with some parameter vector $w \in \mathbb{R}^p$. Since typical deep models have a lot of capacity and are highly overparameterized, we are particularly interested in the overparameterized (or so-called interpolating) regime, where $p > n$ (often $p \gg n$). In this case,
there are many parameter vectors \( w \) that are consistent with the training data points. We denote the set of these parameter vectors by
\[
\mathcal{W} = \{ w \in \mathbb{R}^p \mid f(x_i, w) = y_i, i = 1, \ldots, n \}
\] (1)

This a high-dimensional set (e.g. a \((p-n)\)-dimensional manifold) in \( \mathbb{R}^p \) and depends only on the training data \( \{(x_i, y_i) : i = 1, \ldots, n \} \) and the model \( f(\cdot, \cdot) \).

The total loss on the training set (empirical risk) can be expressed as \( L(w) = \sum_{i=1}^{n} L_i(w) \), where \( L_i(\cdot) = \ell(y_i, f(x_i, w)) \) is the loss on the individual data point \( i \), and \( \ell(\cdot, \cdot) \) is a differentiable non-negative function, with the property that \( \ell(y_i, f(x_i, w)) = 0 \) iff \( y_i = f(x_i, w) \). Often \( \ell(y_i, f(x_i, w)) = \ell(\cdot - f(x_i, w)) \), with \( \ell(\cdot) \) convex and having a global minimum at zero (such as square loss, Huber loss, etc.). In this case, \( L(w) = \sum_{i=1}^{n} \ell(y_i - f(x_i, w)) \).

\( \mathcal{W} \) is the set of global minima, and every parameter vector \( w \) in \( \mathcal{W} \) renders the loss on each data point zero, i.e., \( L_i(w) = 0 \forall i \). The loss function is often attempted to be minimized by stochastic gradient descent (SGD):
\[
w_i = w_{i-1} - \eta \nabla L_i(w_{i-1}), \quad i \geq 1
\] (2)
assuming the data is indexed randomly. We use one index \( i \) for both the loss and the parameter vector at step \( i \) (for \( i > n \), one can cycle through the data, or select them at random, etc.).

2.1 Stochastic Mirror Descent

Mirror descent, first introduced by Nemirovski & Yudin [1983], is one of the most widely used families of algorithms for optimization (Beck & Teboulle; 2003; Cesa-Bianchi et al.; 2012; Zhou et al.; 2017), which includes the popular gradient descent as a special case. Consider a strictly convex differentiable function \( \psi(\cdot) \), called the potential function. Then updates for stochastic mirror descent (SMD) are defined as
\[
\nabla \psi(w_i) = \nabla \psi(w_{i-1}) - \eta \nabla L_i(w_{i-1}).
\] (3)

Note that, due to the strict convexity of \( \psi(\cdot) \), the gradient \( \nabla \psi(\cdot) \) defines an invertible map, so the recursion in (3) yields a unique \( w_i \) at each iteration, and thus is a well-defined update, i.e.,
\[
w_i = \nabla \psi^{-1}(\nabla \psi(w_{i-1}) - \eta \nabla L_i(w_{i-1})).
\]

Compared to classical SGD, rather than update the weight vector along the direction of the negative gradient, the update is done in the “mirrored” domain determined by the invertible transformation \( \nabla \psi(\cdot) \). Mirror descent was originally conceived to exploit the geometrical structure of the problem by choosing an appropriate potential. Note that SMD reduces to SGD when \( \psi(w) = \frac{1}{2} \|w\|^2 \), since the gradient \( \nabla \psi(\cdot) \) is simply the identity map.

Alternatively, the update rule (3) can be expressed as
\[
w_i = \arg \min_w \eta w^T \nabla L_i(w_{i-1}) + D_\psi(w, w_{i-1}),
\] (4)

where
\[
D_\psi(w, w_{i-1}) := \psi(w) - \psi(w_{i-1}) - \nabla \psi(w_{i-1})^T (w - w_{i-1})
\] (5)
is the Bregman divergence with respect to the potential function \( \psi(\cdot) \). Note that \( D_\psi(\cdot, \cdot) \) is non-negative, convex in its first argument, and that, due to strict convexity, \( D_\psi(w, w') = 0 \) iff \( w = w' \).

Different choices of the potential function \( \psi(\cdot) \) yield different optimization algorithms, which will potentially have different implicit biases. A few examples follow.

Gradient Descent. For the potential function \( \psi(w) = \frac{1}{2} ||w||^2 \), the Bregman divergence is \( D_\psi(w, w') = \frac{1}{2} ||w - w'||^2 \), and the update rule reduces to that of SGD.

Exponentiated Gradient Descent. For \( \psi(w) = \sum_j w_j \log w_j \), the Bregman divergence becomes the unnormalized relative entropy (Kullback-Leibler divergence) \( D_\psi(w, w') = \sum_j w_j \log \frac{w_j}{w'_j} - \sum_j w_j + \sum_j w'_j \), which corresponds to the exponentiated gradient descent (aka the exponential weights) algorithm (Kivinen & Warmuth; 1997).

\( p \)-norms Algorithm. For any \( q \)-norm squared potential function \( \psi(w) = \frac{1}{2} ||w||^2_q \), with \( \frac{1}{p} + \frac{1}{q} = 1 \), the algorithm will reduce to the so-called \( p \)-norms algorithm (Grove et al.; 2001; Gentile; 2003).

Sparse Mirror Descent. For \( \psi(w) = ||w||_1^{1+\epsilon} \), the algorithm reduces to sparse mirror descent, which is used in compressed sensing (Azizan & Hassibi; 2019).
2.2 OVERPARAMETERIZED LINEAR MODELS

Overparameterized (or underdetermined) linear models have been recently studied in many papers due to their simplicity, and the fact that there are interesting insights than one can obtain from them. In this case, the model is \( f(x_i, w) = x_i^T w \), the set of global minima is \( \mathcal{W} = \{ w \mid y_i = x_i^T w, \ i = 1, \ldots, n \} \), and the loss is \( L_i(w) = \ell(y_i - x_i^T w) \). The following result characterizes the solution that SMD converges to (Azizan & Hassibi, 2019; Gunasekar et al., 2018a).

**Proposition 1.** Consider a linear overparameterized model. For sufficiently small step size, i.e., for any \( \eta > 0 \) for which \( \psi(\cdot) - \eta L_i(\cdot) \) is convex, and for any initialization \( w_0 \), the SMD iterates converge to
\[
w_\infty = \text{arg min}_{w \in \mathcal{W}} D_\psi(w, w_0).
\]

Note that the step size condition, i.e., the convexity of \( \psi(\cdot) - \eta L_i(\cdot) \), depends on both the loss and the potential function. For the case of SGD, \( \psi(w) = \frac{1}{2} \| w \|^2 \), and \( \ell(y_i - x_i^T w) = \frac{1}{2} (y_i - x_i^T w)^2 \), so the condition reduces to the well-known \( \eta \leq \frac{1}{\|x_i\|^2} \). In this case, \( D_\psi(w, w_0) \) is simply \( \frac{1}{2} \| w - w_0 \|^2 \).

**Corollary 2.** In particular, for the initialization \( w_0 = \text{arg min}_{w \in \mathbb{R}^p} \psi(w) \), under the conditions of Proposition 1 the SMD iterates converge to
\[
w_\infty = \text{arg min}_{w \in \mathcal{W}} \psi(w).
\]

This means that running SMD for a linear model with the aforementioned \( w_0 \), without any explicit regularization, results in a solution that has the smallest potential \( \psi(\cdot) \) among all solutions, i.e., SMD implicitly regularizes the solution with \( \psi(\cdot) \). In particular, this means that SGD initialized around zero acts as an \( \ell_2 \)-norm regularizer. In this paper, we show that these results continue to hold for highly overparameterized nonlinear models in an approximate sense.

3 THEORETICAL RESULTS

In this section, we provide our main theoretical results. In particular, we show that for highly overparameterized models (1) SMD converges to a global minimum, (2) the global minimum obtained by SMD is approximately the closest one to the initialization in Bregman divergence corresponding to the potential.

![Figure 1: An illustration of the parameter space. \( \mathcal{W} \) represents the set of global minima, \( w_0 \) is the initialization, \( B \) is the local neighborhood, \( w^* \) and the closest global minimum to \( w_0 \) (in Bregman divergence), and \( w_\infty \) is the minimum that SMD converges to.](image)

3.1 CONVERGENCE OF STOCHASTIC MIRROR DESCENT

Let us define
\[
D_{L_i}(w, w') := L_i(w) - L_i(w') - \nabla L_i(w')^T (w - w'),
\]
which is defined in a similar way to a Bregman divergence for the loss function. The difference though is that, unlike the potential function of the Bregman divergence, the loss function \( L_i(\cdot) = \ell(y_i - f(x_i, \cdot)) \) need not be convex (even when \( \ell(\cdot) \) is) due to the nonlinearity of \( f(\cdot, \cdot) \).
Theorem 3. Consider the set of interpolating parameters on the convex hull of $B_1$.

Assumption 2. Consider the region are bounded.

Our second assumption states that in this local region, the first and second derivatives of the model are bounded.

Assumption 1. Denote the initial point by $w_0$. There exists $w \in \mathcal{W}$ and a region $\mathcal{B} = \{ w' \in \mathbb{R}^p \mid D_\psi(w, w') \leq \epsilon \}$ containing $w_0$, such that $D_{L_i}(w, w') \geq 0$, $i = 1, \ldots, n$, for all $w' \in \mathcal{B}$.

It is important to understand what this assumption means. Since $L_i(\cdot)$ is not necessarily convex, it is certainly not the case that $D_{L_i}(w, w') \geq 0$ for all $w'$. However, since $w$ is a minimizer of $L_i(\cdot)$, there will be a neighborhood around it such that for all $w'$ in this neighborhood $D_{L_i}(w, w') \geq 0$ (see Fig. 2 for an illustration). What we are requiring is that the initialization $w_0$ be inside the intersection of all such neighborhoods for $i = 1, 2, \ldots, n$. In other words, we require a $w_0$ close enough to $\mathcal{W}$.

![Figure 2](image.png)

Figure 2: An illustration of $D_{L_i}(w, w') \geq 0$ in a local region in Assumption 1.

Our second assumption states that in this local region, the first and second derivatives of the model are bounded.

Assumption 2. Consider the region $\mathcal{B}$ in Assumption 1, $f_i(\cdot)$ have bounded gradient and Hessian on the convex hull of $\mathcal{B}$, i.e., $\| \nabla f_i(w') \| \leq \gamma$, and $\alpha \leq \lambda_{\min}(H_{f_i}(w')) \leq \lambda_{\max}(H_{f_i}(w')) \leq \beta$, $i = 1, \ldots, n$, for all $w' \in \text{conv} \mathcal{B}$.

This is again a mild assumption, which is assumed in other related work such as Oymak & Soltanolkotabi [2019] as well. Note that we do not require $\alpha$ to be positive (just its boundedness).

The following theorem states that under Assumption 1, SMD converges to a global minimum.

Theorem 3. Consider the set of interpolating parameters $\mathcal{W} = \{ w \in \mathbb{R}^p \mid f(x_i, w) = y_i, i = 1, \ldots, n \}$, and the SMD iterates given in (5), where every data point is revisited after some steps. Under Assumption 1 for sufficiently small step size, i.e., for any $\eta > 0$ for which $\psi(\cdot) - \eta L_i(\cdot)$ is strictly convex on $\mathcal{B}$ for all $i$, the following holds.

1. All the iterates $\{w_i\}$ remain in $\mathcal{B}$.
2. The iterates converge (to $w_\infty$).
3. $w_\infty \in \mathcal{W}$.

Note that, while convergence (to some point) with decaying step size is almost trivial, this result establishes convergence to the solution set with a fixed step size. Furthermore, the convergence is deterministic, and is not in expectation or with high probability. For example, this result also applies to the case where we cycle through the data deterministically.

We should also remark that the choice of distance in the definition of the “ball” $\mathcal{B}$ was important to be the Bregman divergence with respect to $\psi(\cdot)$ and in that particular order. In fact, one cannot guarantee that the SMD iterates get closer to an interpolating $w$ at every step in the usual Euclidean sense. However, once can establish that it gets closer in $D_\psi(w, \cdot)$. Finally, it is important to note that we need the step size to be small enough to guarantee the strict convexity of $\psi(\cdot) - \eta L_i(\cdot)$ in $\mathcal{B}$, not globally.

Denote the global minimum that is closest to the initialization in Bregman divergence by $w^*$, i.e.,

$$w^* = \arg \min_{w \in \mathcal{W}} D_\psi(w, w_0).$$

Recall that in the linear case, this was what SMD converges to. We show that in the nonlinear case, under Assumptions 1 and 2, SMD converges to a point $w_\infty$ which is “very close” to $w^*$.
Theorem 4. Define $w^* = \arg\min_{w \in W} D_\psi(w, w_0)$. Under the conditions of Theorem 2 and Assumption 2, the following holds:

1. $D_\psi(w_\infty, w_0) = D_\psi(w^*, w_0) + o(\epsilon)$
2. $D_\psi(w^*, w_\infty) = o(\epsilon)$

In other words, if we start with an initialization that is $O(\epsilon)$ away from $W$ (in Bregman divergence), we converge to a point $w_\infty \in W$ that is $o(\epsilon)$ away from $w^*$. The Bregman divergence of this point is $o(\epsilon)$ from the minimum value it can take.

Corollary 5. For the initialization $w_0 = \arg\min_{w \in \mathbb{R}^p} \psi(w)$, under the conditions of Theorem 4, $w^* = \arg\min_{w \in W} \psi(w)$ and the following holds:

1. $\psi(w_\infty) = \psi(w^*) + o(\epsilon)$
2. $D_\psi(w^*, w_\infty) = o(\epsilon)$

3.2 Proof Technique: Fundamental Identity of SMD

The main tool used for the proofs is a fundamental identity that holds for SMD.

Lemma 6. For any model $f(\cdot, \cdot)$, any differentiable loss $\ell(\cdot)$, any parameter $w \in W$, and any step size $\eta > 0$, the following relation holds for the SMD iterates $\{w_i\}$

$$D_\psi(w_i, w_{i-1}) = D_\psi(w, w_i) + D_\psi - \eta L_i(w_i, w_{i-1}) + \eta L_i(w_i) + \eta D_L(w, w_{i-1}), \quad (9)$$

for all $i \geq 1$.

This identity allows one to prove the results in a remarkably simple and direct way. Due to space limitations, the proofs are relegated to the supplementary material.

The ideas behind this identity are related to $H_\infty$ estimation theory [Hassibi et al., 1999; Simon, 2006], which was originally developed in the 1990’s in the context of robust control theory. In fact, it has connections to the minimax optimality of SGD, which was shown by [Hassibi et al., 1994] for linear models, and recently extended to nonlinear models and general mirrors by [Azizan & Hassibi, 2019].

4 Experimental Results

In this section, we provide our experimental results, which consist of two main parts. In the first part, we evaluate the theoretical claims by running systematic experiments for different initializations and different mirrors, and evaluating the distances between the global minima achieved and the initializations, in different Bregman divergences. In the second part, we assess the generalization error of different mirrors, which correspond to different regularizers, in order to understand which regularizer performs better.

Figure 3: An illustration of the experiments in Table 1
Table 1: Fixed Initialization. Distances from final points (global minima) obtained by different algorithms (columns) from the same initialization (Fig. 3), measured in different Bregman divergences (rows). First Row: The closest initialization to $w_0$ in $\ell_1$ Bregman divergence, among the four final points, is exactly the one obtained by SMD with 1-norm potential. Second Row: The closest point to $w_0$ in $\ell_2$ Bregman divergence (Euclidean distance), among the four final points, is exactly the one obtained by SGD. Third Row: The closest point to $w_0$ in $\ell_3$ Bregman divergence, among the four final points, is exactly the one obtained by SMD with 3-norm potential. Fourth Row: The closest point to $w_0$ in $\ell_{10}$ Bregman divergence, among the four final points, is exactly the one obtained by SMD with 10-norm potential.

4.1 Do SMDs Converge to the Closest Point in Bregman Divergence?

While accessing all the points on $\mathcal{W}$ and finding the closest one is impossible, we design systematic experiments to test this claim. We run experiments on some standard deep learning problems, namely, a standard CNN on MNIST (LeCun et al. 1998) and the ResNet-18 (He et al. 2016) on CIFAR-10 (Krizhevsky & Hinton 2009). We train the models from different initializations, and with different mirror descents from each particular initialization, until we reach 100% training accuracy, i.e., a point on $\mathcal{W}$. We randomly initialize the parameters of the networks around zero. We choose 6 independent initializations for the CNN, and 8 for ResNet-18, and for each initialization, we run different SMD algorithms with the following four potential functions: (a) $\ell_1$ norm, (b) $\ell_2$ norm (which is SGD), (c) $\ell_3$ norm, (d) $\ell_{10}$ norm (as a surrogate for $\ell_\infty$). See Supplementary Material B for more details on the experiments.

Table 2: Fixed Mirror: SGD. Pairwise distances between different initial points and the final points obtained from them by SGD (Fig. 4). Row i: The closest final point to the initial point $i$, among all the eight final points, is exactly the one obtained by the algorithm from initialization $i$.

![Figure 4: An illustration of the experiments in Table 2](image-url)
distances between a particular initial point and all the final points obtained from different initializations and different mirrors (the distances are often orders of magnitude different, so we show them in logarithmic scale). The global minimum achieved by any mirror from any initialization is the closest in the correct Bregman divergence, among all mirrors, among all initializations, and among both. This trend is very consistent among all our experiments, which can be found in Supplementary Material B.

Figure 5: Distances between a particular initial point and all the final points obtained by both different initializations and different mirrors. The smallest distance, among all initializations and all mirrors, corresponds exactly to the final point obtained from that initial point by SGD. This trend is observed consistently for all other mirror descents and all initializations (see the results in Tables 8 and 9 in the appendix).

4.2 Distribution of the Weights of the Network

One may be curious to see how the final weights obtained by these different mirrors look like, and whether, for example, mirror descent corresponding to the $\ell_1$-norm potential induces sparsity. Fig. 6 shows the histogram of the absolute value of the weights for different SMDs, when initialized by the same set of close to zero weights. The histograms of the final weights look substantially different and, since they all started from the same initial weights, this difference is fully attributable to the different mirrors used. The histogram of the $\ell_1$-SMD has more weights at and close to zero, which confirms that it induces sparsity. The histogram of the $\ell_2$-SMD (SGD) looks almost perfectly Gaussian, whereas the $\ell_3$ and $\ell_{10}$ histograms are shifted to the right, so much so that almost all weights in the $\ell_{10}$ solution are non-zero. See Appendix B for more details.

4.3 Generalization Errors of Different Mirrors

We compare the performance of the SMD algorithms discussed before on the test set. For MNIST, perhaps not surprisingly, all the four SMD algorithms achieve around 99% or higher accuracy. For CIFAR-10, however, there is a significant difference between the test errors of different mirrors/regularizers on the same ResNet-18 architecture. Fig. 7 shows the test accuracies of different algorithms with eight random initializations around zero, as discussed before. Counter-intuitively, $\ell_{10}$ performs consistently best, while $\ell_3$ performs consistently worse. This result suggests the importance of a comprehensive study of the role of regularization, and the choice of the best regularizer, to improve the generalization performance of deep neural networks.
Figure 6: Histogram of the absolute value of the final weights in the network for different SMD algorithm with different potentials. Note that each of the four histograms corresponds to an $11 \times 10^6$-dimensional weight vector that perfectly interpolates the data. Even though the weights remain quite small, the histograms are drastically different. $\ell_1$-SMD induces sparsity on the weights, as expected. SGD appears to be produce a Gaussian distribution on the weights. $\ell_3$-SMD starts to reduce the sparsity, and $\ell_{10}$ shifts the distribution of the weights significantly, so much so that almost all the weights are non-zero.

Figure 7: Generalization performance of different SMD algorithms on the CIFAR-10 dataset using ResNet-18. $\ell_{10}$ performs consistently better, while $\ell_1$ performs consistently worse.
REFERENCES


Supplementary Material

A PROOFS OF THE THEORETICAL RESULTS

In this section, we prove the main theoretical results. The proofs are based on a fundamental identity about the iterates of SMD, which holds for all mirrors and all overparametrized (even nonlinear) models. We first prove this identity, and then use it to prove the convergence and implicit regularization results.

A.1 FUNDAMENTAL IDENTITY OF SMD

Let us prove the fundamental identity.

Lemma

For any model \( f(\cdot, \cdot) \), any differentiable loss \( \ell(\cdot) \), any parameter \( w \in \mathcal{W} \), and any step size \( \eta > 0 \), the following relation holds for the SMD iterates \( \{ w_i \} \)

\[
D_\psi(w, w_{i-1}) = D_\psi(w, w_i) + D_\psi - \eta L_i(w_i, w_{i-1}) + \eta L_i(w_i) + \eta D_{L_i}(w, w_{i-1}),
\]

for all \( i \geq 1 \).

Proof of Lemma

Let us start by expanding the Bregman divergence \( D_\psi(w, w_i) \) based on its definition

\[
D_\psi(w, w_i) = \psi(w) - \psi(w_i) - \nabla \psi(w_i)^T (w - w_i).
\]

By plugging the SMD update rule \( \nabla \psi(w_i) = \nabla \psi(w_{i-1}) - \eta \nabla L_i(w_{i-1}) \) into this, we can write it as

\[
D_\psi(w, w_i) = \psi(w) - \psi(w_i) - \nabla \psi(w_{i-1})^T (w - w_i) + \eta \nabla L_i(w_{i-1})^T (w - w_i).
\]

Using the definition of Bregman divergence for \( (w, w_{i-1}) \) and \( (w_i, w_{i-1}) \), i.e., \( D_\psi(w, w_{i-1}) = \psi(w) - \psi(w_{i-1}) - \nabla \psi(w_{i-1})^T (w - w_{i-1}) \) and \( D_\psi(w_i, w_{i-1}) = \psi(w_i) - \psi(w_{i-1}) - \nabla \psi(w_{i-1})^T (w_i - w_{i-1}) \), we can express this as

\[
D_\psi(w, w_i) = D_\psi(w, w_{i-1}) + \psi(w_{i-1}) + \nabla \psi(w_{i-1})^T (w - w_{i-1}) - \psi(w_i)
- \nabla \psi(w_{i-1})^T (w - w_i) + \eta \nabla L_i(w_{i-1})^T (w - w_i)
\]

\[
= D_\psi(w, w_{i-1}) + \psi(w_{i-1}) - \psi(w_i) + \nabla \psi(w_{i-1})^T (w_i - w_{i-1})
+ \eta \nabla L_i(w_{i-1})^T (w - w_i)
\]

\[
= D_\psi(w, w_{i-1}) - D_\psi(w_i, w_{i-1}) + \eta \nabla L_i(w_{i-1})^T (w - w_i).
\]

Expanding the last term using \( w - w_i = (w - w_{i-1}) - (w_i - w_{i-1}) \), and following the definition of \( D_{L_i}(\cdot, \cdot) \) from (7) for \( (w, w_{i-1}) \) and \( (w_i, w_{i-1}) \), we have

\[
D_\psi(w, w_i) = D_\psi(w, w_{i-1}) - D_\psi(w_i, w_{i-1}) + \eta \nabla L_i(w_{i-1})^T (w - w_{i-1})
- \eta \nabla L_i(w_{i-1})^T (w_i - w_{i-1})
\]

\[
= D_\psi(w, w_{i-1}) - D_\psi(w_i, w_{i-1}) + \eta (L_i(w) - L_i(w_{i-1}) - D_{L_i}(w, w_{i-1}))
\]

\[
= D_\psi(w, w_{i-1}) - D_\psi(w_i, w_{i-1}) + \eta (L_i(w) - D_{L_i}(w, w_{i-1}))
- \eta (L_i(w_i) - D_{L_i}(w_i, w_{i-1}))
\]

\[
= D_\psi(w, w_{i-1}) - D_\psi(w_i, w_{i-1}) + \eta L_i(w_i) + \eta D_{L_i}(w, w_{i-1})
\]

Note that for all \( w \in \mathcal{W} \), we have \( L_i(w) = 0 \). Therefore, for all \( w \in \mathcal{W} \)

\[
D_\psi(w, w_i) = D_\psi(w, w_{i-1}) - D_\psi(w_i, w_{i-1}) - \eta D_{L_i}(w, w_{i-1}) - \eta L_i(w_i) + \eta D_{L_i}(w, w_{i-1}).
\]

Combining the second and the last terms in the right-hand side leads to

\[
D_\psi(w, w_i) = D_\psi(w, w_{i-1}) - D_\psi - \eta L_i(w_i, w_{i-1}) - \eta D_{L_i}(w, w_{i-1}) - \eta L_i(w_i),
\]

for all \( w \in \mathcal{W} \), which concludes the proof.
A.2 Convergence of SMD to the Interpolating Set

Now that we have proved Lemma 6, we can use it to prove our main results, in a remarkably simple fashion. Let us first prove the convergence of SMD to the set of solutions.

Assumption 1. Denote the initial point by \( w_0 \). There exists \( w \in W \) and a region \( B = \{ w' \in \mathbb{R}^p \mid D_{\psi}(w, w') \leq \epsilon \} \) containing \( w_0 \), such that \( D_{L_i}(w, w') \geq 0 \), \( i = 1, \ldots, n \), for all \( w' \in B \).

Theorem 3. Consider the set of interpolating parameters \( W = \{ w \in \mathbb{R}^p \mid f(x_i, w) = y_i, i = 1, \ldots, n \} \), and the SMD iterates given in (3), where every data point is revisited after some steps. Under Assumption 7 for sufficiently small step size, i.e., for any \( \eta > 0 \) for which \( \psi(\cdot) - \eta L_i(\cdot) \) is strictly convex for all \( i \), the following holds.

1. All the iterates \( \{w_i\} \) remain in \( B \).
2. The iterates converge to \( w_\infty \).
3. \( w_\infty \in W \).

Proof of Theorem 3. First we show that all the iterates will remain in \( B \). Recall the identity of SMD from Lemma 6:

\[
D_{\psi}(w, w_{i-1}) = D_{\psi}(w, w_i) + D_{\psi-L_i}(w_i, w_{i-1}) + \eta L_i(w_i) + \eta D_{L_i}(w, w_{i-1}) \tag{9}
\]

which holds for all \( w \in W \). If \( w_{i-1} \) is in the region \( B \), we know that the last term \( D_{L_i}(w, w_{i-1}) \) is non-negative. Furthermore, if the step size is small enough that \( \psi(\cdot) - \eta L_i(\cdot) \) is strictly convex, the second term \( D_{\psi-L_i}(w_i, w_{i-1}) \) is a Bregman divergence and is non-negative. Since the loss is non-negative, \( \eta L_i(w_i) \) is always non-negative. As a result, we have

\[
D_{\psi}(w, w_{i-1}) \geq D_{\psi}(w, w_i), \tag{19}
\]

This implies that \( D_{\psi}(w, w_i) \leq \epsilon \), which means \( w_i \) is in \( B \) too. Since \( w_0 \) is in \( B \), \( w_1 \) will be in \( B \), and therefore, \( w_2 \) will be in \( B \), and similarly all the iterates will remain in \( B \).

Next, we prove that the iterates converge and \( w_\infty \in W \). If we sum up identity (9) for all \( i = 1, \ldots, T \), the first terms on the right- and left-hand side cancel each other telescopically, and we have

\[
D_{\psi}(w, w_0) = D_{\psi}(w, w_T) + \sum_{i=1}^{T} [D_{\psi-L_i}(w_i, w_{i-1}) + \eta L_i(w_i) + \eta D_{L_i}(w, w_{i-1})]. \tag{20}
\]

Since \( D_{\psi}(w, w_T) \geq 0 \), we have \( \sum_{i=1}^{T} [D_{\psi-L_i}(w_i, w_{i-1}) + \eta L_i(w_i) + \eta D_{L_i}(w, w_{i-1})] \leq D_{\psi}(w, w_0) \). If we take \( T \to \infty \), the sum still has to remain bounded, i.e.,

\[
\sum_{i=1}^{\infty} [D_{\psi-L_i}(w_i, w_{i-1}) + \eta L_i(w_i) + \eta D_{L_i}(w, w_{i-1})] \leq D_{\psi}(w, w_0). \tag{21}
\]

Since the step size is small enough that \( \psi(\cdot) - \eta L_i(\cdot) \) is strictly convex for all \( i \), the first term \( D_{\psi-L_i}(w_i, w_{i-1}) \) is non-negative. The second term \( \eta L_i(w_i) \) is non-negative because of the non-negativity of the loss. Finally, the last term \( D_{L_i}(w, w_{i-1}) \) is non-negative because \( w_{i-1} \in B \) for all \( i \). Hence, all the three terms in the summand are non-negative, and because the sum is bounded, they should go to zero as \( i \to \infty \). In particular,

\[
D_{\psi-L_i}(w_i, w_{i-1}) \to 0 \tag{22}
\]

implies \( w_i \to w_{i-1} \), i.e., convergence \( (w_i \to w_\infty) \) (Note that the functions \( \psi - \eta L_i \) do not go to zero, as there is a fixed number, i.e., \( n \), of them). Further,

\[
\eta L_i(w_i) \to 0. \tag{23}
\]

This implies that all the individual losses are going to zero, and since every data point is being revisited after some steps, all the data points are being fit. Therefore, \( w_\infty \in W \).
A.3 Closeness of the Final Point to the Regularized Solution

In this section, we show that with the additional Assumption 2 (which is equivalent to \( f_i(\cdot) \) having bounded Hessian in \( B \)), not only do the iterates remain in \( B \) and converge to the set \( W \), but also they converge to a point which is very close to \( w^* \) (the closest solution to the initial point, in Bregman divergence). The proof is again based on our fundamental identity for SMD.

**Assumption 2.** Consider the region \( B \) in Assumption 1, \( f_i(\cdot) \) have bounded gradient and Hessian on the convex hull of \( B \), i.e., \( \| \nabla f_i(w') \| \leq \gamma \), and \( \alpha \leq \lambda_{\min}(H_f(w')) \leq \lambda_{\max}(H_f(w')) \leq \beta \), \( i = 1, \ldots, n \), for all \( w' \in \text{conv } B \).

**Theorem 4.** Define \( w^* = \arg \min_{w \in W} D_\psi(w, w_0) \). Under the assumptions of Theorem 3 and Assumption 2, the following holds.

1. \( D_\psi(w_\infty, w_0) = D_\psi(w^*, w_0) + o(\epsilon) \)
2. \( D_\psi(w^*, w_\infty) = o(\epsilon) \)

**Proof of Theorem 4.** Recalling the identity of SMD from Lemma 1,

\[
D_\psi(w, w_{i-1}) = D_\psi(w, w_i) + D_{\psi - \eta L_i}(w_i, w_{i-1}) + \eta L_i(w_i) + \eta D_{L_i}(w, w_{i-1})
\]

which holds for all \( w \in W \). Summing the identity for all \( i \geq 1 \), we have

\[
D_\psi(w, w_0) = D_\psi(w, w_\infty) + \sum_{i=1}^{\infty} [D_{\psi - \eta L_i}(w_i, w_{i-1}) + \eta L_i(w_i) + \eta D_{L_i}(w, w_{i-1})]
\]

for all \( w \in W \). Note that the only terms in the right-hand side which depend on \( w \) are the first one \( D_\psi(w, w_\infty) \) and the last one \( \eta \sum_{i=1}^{\infty} D_{L_i}(w, w_{i-1}) \). In what follows, we will argue that, within \( B \), the dependence on \( w \) in the last term is weak and therefore \( w_\infty \) is close to \( w^* \).

To further spell out the dependence on \( w \) in the last term, let us expand \( D_{L_i}(w, w_{i-1}) \)

\[
D_{L_i}(w, w_{i-1}) = 0 - L_i(w_{i-1}) - \nabla L_i(w_{i-1})^T (w - w_{i-1})
\]

\[
= -L_i(w_{i-1}) + \ell'(y_i - f_i(w_{i-1})) \nabla f_i(w_{i-1})^T (w - w_{i-1})
\]

for all \( w \in W \), where the first equality comes from the definition of \( D_{L_i}(\cdot, \cdot) \) and the fact that \( L_i(w) = 0 \) for \( w \in W \). The second equality is from taking the derivative of \( L_i(\cdot) = \ell(y_i - f_i(\cdot)) \) and evaluating it at \( w_{i-1} \).

By Taylor expansion of \( f_i(w) \) around \( w_{i-1} \) and using Taylor’s theorem (Lagrange’s mean-value form), we have

\[
f_i(w) = f_i(w_{i-1}) + \nabla f_i(w_{i-1})^T (w - w_{i-1}) + \frac{1}{2} (w - w_{i-1})^T H_{f_i}(\hat{w}_i)(w - w_{i-1}),
\]

for some \( \hat{w}_i \) in the convex hull of \( w \) and \( w_{i-1} \). Since \( f_i(w) = y_i \) for all \( w \in W \), it follows that

\[
\nabla f_i(w_{i-1})^T (w - w_{i-1}) = y_i - f_i(w_{i-1}) - \frac{1}{2} (w - w_{i-1})^T H_{f_i}(\hat{w}_i)(w - w_{i-1}),
\]

for all \( w \in W \). Plugging this into (26), we have

\[
D_{L_i}(w, w_{i-1}) = \ell'(y_i - f_i(w_{i-1}))(y_i - f_i(w_{i-1}) - \frac{1}{2} (w - w_{i-1})^T H_{f_i}(\hat{w}_i)(w - w_{i-1})
\]

for all \( w \in W \). Finally, by plugging this back into the identity (24), we have

\[
D_\psi(w, w_0) = D_\psi(w, w_\infty) + \sum_{i=1}^{\infty} [D_{\psi - \eta L_i}(w_i, w_{i-1}) + \eta L_i(w_i) - \eta L_i(w_{i-1})
+
\eta \ell'(y_i - f_i(w_{i-1}))(y_i - f_i(w_{i-1}) - \frac{1}{2} (w - w_{i-1})^T H_{f_i}(\hat{w}_i)(w - w_{i-1}))]
\]

for all \( w \in W \). Note that this can be expressed as

\[
D_\psi(w, w_0) = D_\psi(w, w_\infty) + C - \sum_{i=1}^{\infty} \frac{1}{2} \eta \ell'(y_i - f_i(w_{i-1}))(w - w_{i-1})^T H_{f_i}(\hat{w}_i)(w - w_{i-1}),
\]

where

\[
C = \frac{1}{2} \sum_{i=1}^{\infty} \eta \ell'(y_i - f_i(w_{i-1}))(y_i - f_i(w_{i-1}) - \frac{1}{2} (w - w_{i-1})^T H_{f_i}(\hat{w}_i)(w - w_{i-1})
\].
for all $w \in \mathcal{W}$, where $C$ does not depend on $w$:

$$C = \sum_{i=1}^{\infty} [D_{\psi \eta L_i}(w_i, w_{i-1}) + \eta L_i(w_i) - \eta L_i(w_{i-1}) + \eta \ell'(y_i - f_i(w_{i-1}))(y_i - f_i(w_{i-1}))].$$

From Theorem 3, we know that $w_\infty \in \mathcal{W}$. Therefore, by plugging it into equation (31), and using the fact that $D_{\psi}(w_\infty, w_\infty) = 0$, we have

$$D_{\psi}(w_\infty, w_0) = C - \sum_{i=1}^{\infty} \frac{1}{2} \eta \ell'(y_i - f_i(w_{i-1}))(w_\infty - w_{i-1})^T H_{f_i}(w'_i)(w_\infty - w_{i-1}),$$

(32)

where $w'_i$ is a point in the convex hull of $w_\infty$ and $w_{i-1}$ (and therefore also in conv $\mathcal{B}$), for all $i$. Similarly, by plugging $w^*$, which is also in $\mathcal{W}$, into (31), we have

$$D_{\psi}(w^*, w_0) = D_{\psi}(w^*, w_\infty) + C - \sum_{i=1}^{\infty} \frac{1}{2} \eta \ell'(y_i - f_i(w_{i-1}))(w^* - w_{i-1})^T H_{f_i}(w''_i)(w^* - w_{i-1}),$$

where $w''_i$ is a point in the convex hull of $w^*$ and $w_{i-1}$ (and therefore also in conv $\mathcal{B}$), for all $i$. Subtracting the last two equations from each other yields

$$D_{\psi}(w_\infty, w_0) - D_{\psi}(w^*, w_0) = -D_{\psi}(w^*, w_\infty) + \sum_{i=1}^{\infty} \frac{1}{2} \eta \ell'(y_i - f_i(w_{i-1})).$$

$$\left[(w^* - w_{i-1})^T H_{f_i}(w''_i)(w^* - w_{i-1}) - (w_\infty - w_{i-1})^T H_{f_i}(w'_i)(w_\infty - w_{i-1})\right].$$

(34)

Note that since all $w'_i$ and $w''_i$ are in conv $\mathcal{B}$, by Assumption 2 we have

$$\alpha \|w_\infty - w_{i-1}\|^2 \leq (w_\infty - w_{i-1})^T H_{f_i}(w'_i)(w_\infty - w_{i-1}) \leq \beta \|w_\infty - w_{i-1}\|^2,$$

(35)

and

$$\alpha \|w^* - w_{i-1}\|^2 \leq (w^* - w_{i-1})^T H_{f_i}(w''_i)(w^* - w_{i-1}) \leq \beta \|w^* - w_{i-1}\|^2.$$  

(36)

Further, again since all the iterates $\{w_i\}$ are in $\mathcal{B}$, it follows that $\|w_\infty - w_{i-1}\|^2 = O(\epsilon)$ and $\|w^* - w_{i-1}\|^2 = O(\epsilon)$. As a result the difference of the two terms, i.e., $[(w^* - w_{i-1})^T H_{f_i}(w''_i)(w^* - w_{i-1}) - (w_\infty - w_{i-1})^T H_{f_i}(w'_i)(w_\infty - w_{i-1})]$, is also $O(\epsilon)$, and we have

$$D_{\psi}(w_\infty, w_0) - D_{\psi}(w^*, w_0) = -D_{\psi}(w^*, w_\infty) + \sum_{i=1}^{\infty} \eta \ell'(y_i - f_i(w_{i-1}))O(\epsilon).$$

(37)

Now note that $\ell'(y_i - f_i(w_{i-1})) = \ell'(f_i(w) - f_i(w_{i-1})) = \ell'(\nabla f_i(\bar{w}_i)^T(\bar{w} - w_{i-1}))$ for some $\bar{w}_i \in \text{conv } \mathcal{B}$. Since $\|w - w_{i-1}\|^2 = O(\epsilon)$ for all $i$, and since $\ell(\cdot)$ is differentiable and $f_i(\cdot)$ have bounded derivatives, it follows that $\ell'(y_i - f_i(w_{i-1})) = o(\epsilon)$. Furthermore, the sum is bounded. This implies that $D_{\psi}(w_\infty, w_0) - D_{\psi}(w^*, w_0) = -D_{\psi}(w^*, w_\infty) + o(\epsilon)$, or equivalently

$$(D_{\psi}(w_\infty, w_0) - D_{\psi}(w^*, w_0)) + D_{\psi}(w^*, w_\infty) = o(\epsilon).$$

(38)

The term in parentheses $D_{\psi}(w_\infty, w_0) - D_{\psi}(w^*, w_0)$ is non-negative by definition of $w^*$. The second term $D_{\psi}(w^*, w_\infty)$ is non-negative by convexity of $\psi$. Since both terms are non-negative and their sum is $o(\epsilon)$, each one of them is at most $o(\epsilon)$, i.e.

$$\begin{cases}D_{\psi}(w_\infty, w_0) - D_{\psi}(w^*, w_0) = o(\epsilon) \\ D_{\psi}(w^*, w_\infty) = o(\epsilon)\end{cases}$$

(39)

which concludes the proof. \qed

**Corollary 5.** For the initialization $w_0 \in \arg \min_{w \in \mathbb{R}^p} \psi(w)$, under the conditions of Theorem 4, $w^* \in \arg \min_{w \in \mathcal{W}} \psi(w)$ and the following holds.

1. $\psi(w_\infty) = \psi(w^*) + o(\epsilon)$
2. $D_{\psi}(w^*, w_\infty) = o(\epsilon)$
Proof of Corollary\textsuperscript{5} The proof is a straightforward application of Theorem\textsuperscript{4}. Note that we have

\[ D_\psi(w, w_0) = \psi(w) - \psi(w_0) - \nabla \psi(w_0)^T (w - w_0) \]  

(40)

for all \( w \). When \( w_0 = \arg \min_{w \in \mathbb{R}^p} \psi(w) \), it follows that \( \nabla \psi(w_0) = 0 \), and

\[ D_\psi(w, w_0) = \psi(w) - \psi(w_0). \]  

(41)

In particular, by plugging in \( w_\infty \) and \( w^* \), we have \( D_\psi(w_\infty, w_0) = \psi(w_\infty) - \psi(w_0) \) and \( D_\psi(w^*, w_0) = \psi(w^*) - \psi(w_0) \). Subtracting the two equations from each other yields

\[ D_\psi(w_\infty, w_0) - D_\psi(w^*, w_0) = \psi(w_\infty) - \psi(w^*), \]  

(42)

which along with the application of Theorem\textsuperscript{4} concludes the proof. \( \square \)

A.4 Closeness to the Interpolating Set in Highly Overparameterized Models

As we mentioned earlier, it has been argued in a number of recent papers that for highly overparameterized models, any random initial point is, whp, close to the solution set \( \mathcal{W} \) \cite{AzizanHassibi2019, LiLiang2018, Duetal2018, AllenZhu2019, CaoG2019}. In the highly overparameterized regime, \( p \gg n \), and so the dimension of the manifold \( \mathcal{W} \), which is \( p - n \), is very large. For simplicity, we outline an argument for the case of Euclidean distance, bearing in mind that a similar argument can be used for general Bregman divergence. Note that the distance of an arbitrarily chosen \( w_0 \) to \( \mathcal{W} \) is given by

\[ \min_{w} \| w - w_0 \|^2 \]

\[ \text{s.t. } y = f(x, w) \]

where \( y = \text{vec}(y_i, i = 1, \ldots, n) \) and \( f(x, w) = \text{vec}(f(x_i, w), i = 1, \ldots, n) \). This can be approximated by

\[ \min_{w} \| w - w_0 \|^2 \]

\[ \text{s.t. } y \approx f(x, w_0) + \nabla f(x, w_0)^T (w - w_0) \]

where \( \nabla f(x, w_0)^T = \text{vec}(\nabla f(x_i, w)^T, i = 1, \ldots, n) \) is the \( n \times p \) Jacobian matrix. The latter optimization can be solved to yield

\[ \| w_\ast - w_0 \|^2 \approx (y - f(x, w_0))^T (\nabla f(x, w_0)^T \nabla f(x, w_0))^{-1} (y - f(x, w_0)) \]  

(43)

Note that \( \nabla f(x, w_0)^T \nabla f(x, w_0) \) is an \( n \times n \) matrix consisting of the sum of \( p \) outer products. When the \( x_i \) are sufficiently random, and \( p \gg n \), it is not unreasonable to assume that whp

\[ \lambda_{\text{min}} (\nabla f(x, w_0)^T \nabla f(x, w_0)) = \Omega(p), \]

from which we conclude

\[ \| w_\ast - w_0 \|^2 \approx \| y - f(x, w_0) \|^2 \cdot O(\frac{1}{p}) = O(\frac{n}{p}), \]  

(44)

since \( y - f(x, w_0) \) is \( n \)-dimensional. The above implies that \( w_0 \) is close to \( w_\ast \) and hence \( \mathcal{W} \).
B MORE DETAILS ON THE EXPERIMENTAL RESULTS

In order to evaluate the claim, we run systematic experiments on some standard deep learning problems.

Datasets. We use the standard MNIST [LeCun et al., 1998] and CIFAR-10 [Krizhevsky & Hinton, 2009] datasets.

Architectures. For MNIST, we use a 4-layer convolutional neural network (CNN) with 2 convolution layers and 2 fully connected layers. The convolutional layers and the fully connected layers are picked wide enough to obtain 2 × 10^6 trainable parameters. Since MNIST dataset has 60,000 training samples, the number of parameters is significantly larger than the number of training data points, and the problem is highly overparameterized. For the CIFAR-10 dataset, we use the standard ResNet-18 [He et al., 2016] architecture without any modifications. CIFAR-10 has 50,000 training samples and with the total number of 11 × 10^6 parameters in ResNet-18, the problem is again highly overparameterized.

Loss Function. We use the cross-entropy loss as the loss function in our training. We train the models from different initializations, and with different mirror descents from each particular initialization, until we reach 100% training accuracy, i.e., until we hit W.

Initialization. We randomly initialize the parameters of the networks around zero (N(0, 0.0001)). We choose 6 independent initializations for the CNN, and 8 for ResNet-18, and for each initialization, we run the following 4 different SMD algorithms.

Algorithms. We use the mirror descent algorithms defined by the norm potential \( \psi(w) = \frac{1}{q} \|w\|^q \) for the following four different norms: (a) \( \ell_1 \) norm, i.e., \( q = 1 + \epsilon \), (b) \( \ell_2 \) norm, i.e., \( q = 2 \) (which is SGD), (c) \( \ell_3 \) norm, i.e., \( q = 3 \), (d) \( \ell_{10} \) norm, i.e., \( q = 10 \) (as a surrogate for \( \ell_\infty \) norm). The update rule can be expressed as follows.

\[
\begin{align*}
    w_{i,j} & = \left| w_{i-1,j} \right|^{q-1} \text{sign}(w_{i-1,j}) - \eta \nabla L_i(w_{i-1})_j \left| w_{i-1,j} \right|^{q-1} \cdot \\
    & \quad \text{sign} \left( \left| w_{i-1,j} \right|^{q-1} \text{sign}(w_{i-1,j}) - \eta \nabla L_i(w_{i-1})_j \right) \right), \quad (45)
\end{align*}
\]

where \( w_{i-1,j} \) denotes the j-th element of the \( w_{i-1} \) vector.

We use a fixed step size \( \eta \). The step size is chosen to obtain convergence to global minima.

B.1 MNIST Experiments

B.1.1 Closest Minimum for Different Mirror Descents with Fixed Initialization

We provide the distances from final points (global minima) obtained by different algorithms from the same initialization, measured in different Bregman divergences for MNIST classification task using a standard CNN. Note that in all tables the smallest element in each row is on the diagonal, which means the point achieved by each mirror has the smallest Bregman divergence to the initialization corresponding to that mirror, among all mirrors. Tables 3, 4, 5, 6, 7, 8 depict these results for 6 different initializations. The rows are the distance metrics used as the Bregman Divergences with specified potentials. The columns are the global minima obtained using specified SMD algorithms.

<table>
<thead>
<tr>
<th>SMD 1-norm</th>
<th>SMD 2-norm (SGD)</th>
<th>SMD 3-norm</th>
<th>SMD 10-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-norm BD</td>
<td>2.767</td>
<td>937.8</td>
<td>1.05 × 10^4</td>
</tr>
<tr>
<td>2-norm BD</td>
<td>301.6</td>
<td>58.61</td>
<td>261.3</td>
</tr>
<tr>
<td>3-norm BD</td>
<td>1720</td>
<td>37.45</td>
<td>7.143</td>
</tr>
<tr>
<td>10-norm BD</td>
<td>7.453 × 10^8</td>
<td>773.4</td>
<td>0.2939</td>
</tr>
</tbody>
</table>

Table 3: MNIST Initial Point 1
Table 4: MNIST Initial Point 2

<table>
<thead>
<tr>
<th></th>
<th>SMD 1-norm</th>
<th>SMD 2-norm (SGD)</th>
<th>SMD 3-norm</th>
<th>SMD 10-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-norm BD</td>
<td>2.78</td>
<td>945</td>
<td>1.37 × 10^3</td>
<td>2.01 × 10^7</td>
</tr>
<tr>
<td>2-norm BD</td>
<td>292</td>
<td>59.3</td>
<td>374</td>
<td>2.29 × 10^4</td>
</tr>
<tr>
<td>3-norm BD</td>
<td>1.51 × 10^3</td>
<td>38.6</td>
<td>11.6</td>
<td>2.71 × 10^3</td>
</tr>
<tr>
<td>10-norm BD</td>
<td>1.06 × 10^8</td>
<td>831</td>
<td>0.86</td>
<td>0.00321</td>
</tr>
</tbody>
</table>

Table 5: MNIST Initial Point 3

<table>
<thead>
<tr>
<th></th>
<th>SMD 1-norm</th>
<th>SMD 2-norm (SGD)</th>
<th>SMD 3-norm</th>
<th>SMD 10-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-norm BD</td>
<td>3.02</td>
<td>968</td>
<td>1.06 × 10^4</td>
<td>1.9 × 10^7</td>
</tr>
<tr>
<td>2-norm BD</td>
<td>291</td>
<td>60.9</td>
<td>272</td>
<td>2.12 × 10^4</td>
</tr>
<tr>
<td>3-norm BD</td>
<td>1.49 × 10^3</td>
<td>39.1</td>
<td>7.82</td>
<td>2.49 × 10^3</td>
</tr>
<tr>
<td>10-norm BD</td>
<td>1.1 × 10^8</td>
<td>900</td>
<td>0.411</td>
<td>0.00318</td>
</tr>
</tbody>
</table>

Table 6: MNIST Initial Point 4

<table>
<thead>
<tr>
<th></th>
<th>SMD 1-norm</th>
<th>SMD 2-norm (SGD)</th>
<th>SMD 3-norm</th>
<th>SMD 10-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-norm BD</td>
<td>2.78</td>
<td>1.21 × 10^4</td>
<td>1.08 × 10^4</td>
<td>1.92 × 10^7</td>
</tr>
<tr>
<td>2-norm BD</td>
<td>291</td>
<td>77.3</td>
<td>271</td>
<td>2.15 × 10^4</td>
</tr>
<tr>
<td>3-norm BD</td>
<td>1.48 × 10^3</td>
<td>49.7</td>
<td>7.56</td>
<td>2.52 × 10^3</td>
</tr>
<tr>
<td>10-norm BD</td>
<td>9.9 × 10^7</td>
<td>1.72 × 10^3</td>
<td>0.352</td>
<td>0.00296</td>
</tr>
</tbody>
</table>

Table 7: MNIST Initial Point 5

<table>
<thead>
<tr>
<th></th>
<th>SMD 1-norm</th>
<th>SMD 2-norm (SGD)</th>
<th>SMD 3-norm</th>
<th>SMD 10-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-norm BD</td>
<td>2.79</td>
<td>958</td>
<td>1.08 × 10^4</td>
<td>2 × 10^5</td>
</tr>
<tr>
<td>2-norm BD</td>
<td>292</td>
<td>60.4</td>
<td>271</td>
<td>2.28 × 10^4</td>
</tr>
<tr>
<td>3-norm BD</td>
<td>1.49 × 10^3</td>
<td>39</td>
<td>7.52</td>
<td>2.69 × 10^3</td>
</tr>
<tr>
<td>10-norm BD</td>
<td>9.09 × 10^7</td>
<td>846</td>
<td>0.342</td>
<td>0.00309</td>
</tr>
</tbody>
</table>

Table 8: MNIST Initial Point 6

<table>
<thead>
<tr>
<th></th>
<th>SMD 1-norm</th>
<th>SMD 2-norm (SGD)</th>
<th>SMD 3-norm</th>
<th>SMD 10-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-norm BD</td>
<td>2.96</td>
<td>930</td>
<td>1.08 × 10^4</td>
<td>1.9 × 10^7</td>
</tr>
<tr>
<td>2-norm BD</td>
<td>308</td>
<td>59</td>
<td>271</td>
<td>2.12 × 10^4</td>
</tr>
<tr>
<td>3-norm BD</td>
<td>1.63 × 10^3</td>
<td>38.6</td>
<td>7.46</td>
<td>2.47 × 10^3</td>
</tr>
<tr>
<td>10-norm BD</td>
<td>1.65 × 10^8</td>
<td>864</td>
<td>0.334</td>
<td>0.00295</td>
</tr>
</tbody>
</table>

B.1.2 Closest Minimum for Different Initializations with Fixed Mirror

We provide the pairwise distances between different initial points and the final points (global minima) obtained by using fixed SMD algorithms in MNIST dataset using a standard CNN. Note that the smallest element in each row is on the diagonal, which means the closest final point to each initialization, among all the final points, is the one corresponding to that point. Tables 9, 10, 11 and 12 depict these results for 4 different SMD algorithms. The rows are the initial points and the columns are the final points corresponding to each initialization.

B.1.3 Closest Minimum for Different Initializations and Different Mirrors

Now we assess the pairwise distances between different initial points and final points (global minima) obtained by all different initializations and all different mirrors (Table 9). The smallest element in each row is exactly the final point obtained by that mirror from that initialization, among all the mirrors and all the initial points.
Table 9: MNIST 1-norm Bregman Divergence Between the Initial Points and the Final Points obtained by SMD 1-norm

<table>
<thead>
<tr>
<th>Initial Point</th>
<th>Final 1</th>
<th>Final 2</th>
<th>Final 3</th>
<th>Final 4</th>
<th>Final 5</th>
<th>Final 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Point 1</td>
<td>2.7671</td>
<td>20311</td>
<td>20266</td>
<td>20331</td>
<td>20340</td>
<td>20282</td>
</tr>
<tr>
<td>Initial Point 2</td>
<td>20332</td>
<td>2.7774</td>
<td>20281</td>
<td>20299</td>
<td>20312</td>
<td>20323</td>
</tr>
<tr>
<td>Initial Point 3</td>
<td>20319</td>
<td>20312</td>
<td>3.018</td>
<td>20344</td>
<td>20309</td>
<td>20322</td>
</tr>
<tr>
<td>Initial Point 4</td>
<td>20339</td>
<td>20279</td>
<td>20310</td>
<td>2.781</td>
<td>20321</td>
<td>20297</td>
</tr>
<tr>
<td>Initial Point 5</td>
<td>20347</td>
<td>20317</td>
<td>20273</td>
<td>20316</td>
<td>2.7902</td>
<td>20311</td>
</tr>
<tr>
<td>Initial Point 6</td>
<td>20344</td>
<td>20323</td>
<td>20340</td>
<td>20318</td>
<td>20321</td>
<td>2.964</td>
</tr>
</tbody>
</table>

Table 10: MNIST 2-norm Bregman Divergence Between the Initial Points and the Final Points obtained by SMD 2-norm (SGD)

<table>
<thead>
<tr>
<th>Initial Point</th>
<th>Final 1</th>
<th>Final 2</th>
<th>Final 3</th>
<th>Final 4</th>
<th>Final 5</th>
<th>Final 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Point 1</td>
<td>58.608</td>
<td>670.75</td>
<td>667.03</td>
<td>684.18</td>
<td>671.36</td>
<td>667.84</td>
</tr>
<tr>
<td>Initial Point 2</td>
<td>669.84</td>
<td>59.315</td>
<td>669.16</td>
<td>682.04</td>
<td>669.45</td>
<td>669.98</td>
</tr>
<tr>
<td>Initial Point 3</td>
<td>666.35</td>
<td>670.22</td>
<td>60.858</td>
<td>683.44</td>
<td>667.57</td>
<td>669.99</td>
</tr>
<tr>
<td>Initial Point 4</td>
<td>669.71</td>
<td>668.86</td>
<td>671.19</td>
<td>77.275</td>
<td>670.33</td>
<td>669.7</td>
</tr>
<tr>
<td>Initial Point 5</td>
<td>671.1</td>
<td>669.12</td>
<td>668.45</td>
<td>683.61</td>
<td>60.39</td>
<td>666.04</td>
</tr>
<tr>
<td>Initial Point 6</td>
<td>669.46</td>
<td>670.92</td>
<td>671.59</td>
<td>684.32</td>
<td>667.37</td>
<td>59.043</td>
</tr>
</tbody>
</table>

Table 11: MNIST 3-norm Bregman Divergence Between the Initial Points and the Final Points obtained by SMD 3-norm

<table>
<thead>
<tr>
<th>Initial Point</th>
<th>Final 1</th>
<th>Final 2</th>
<th>Final 3</th>
<th>Final 4</th>
<th>Final 5</th>
<th>Final 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Point 1</td>
<td>7.143</td>
<td>35.302</td>
<td>32.077</td>
<td>32.659</td>
<td>32.648</td>
<td>32.309</td>
</tr>
<tr>
<td>Initial Point 2</td>
<td>32.507</td>
<td>11.578</td>
<td>32.256</td>
<td>32.325</td>
<td>32.225</td>
<td>32.46</td>
</tr>
<tr>
<td>Initial Point 3</td>
<td>31.594</td>
<td>34.643</td>
<td>7.8239</td>
<td>32.521</td>
<td>31.58</td>
<td>32.519</td>
</tr>
<tr>
<td>Initial Point 4</td>
<td>32.303</td>
<td>34.811</td>
<td>32.937</td>
<td>7.5589</td>
<td>32.617</td>
<td>32.284</td>
</tr>
<tr>
<td>Initial Point 5</td>
<td>32.673</td>
<td>34.678</td>
<td>32.071</td>
<td>32.738</td>
<td>7.5188</td>
<td>31.558</td>
</tr>
<tr>
<td>Initial Point 6</td>
<td>32.116</td>
<td>34.731</td>
<td>32.376</td>
<td>32.431</td>
<td>31.699</td>
<td>7.4593</td>
</tr>
</tbody>
</table>

Table 12: MNIST 10-norm Bregman Divergence Between the Initial Points and the Final Points obtained by SMD 10-norm

<table>
<thead>
<tr>
<th>Initial Point</th>
<th>Final 1</th>
<th>Final 2</th>
<th>Final 3</th>
<th>Final 4</th>
<th>Final 5</th>
<th>Final 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Point 1</td>
<td>0.00354</td>
<td>0.37</td>
<td>0.403</td>
<td>0.286</td>
<td>0.421</td>
<td>0.408</td>
</tr>
<tr>
<td>Initial Point 2</td>
<td>0.33</td>
<td>0.00321</td>
<td>0.369</td>
<td>0.383</td>
<td>0.415</td>
<td>0.422</td>
</tr>
<tr>
<td>Initial Point 3</td>
<td>0.347</td>
<td>0.318</td>
<td>0.00318</td>
<td>0.401</td>
<td>0.312</td>
<td>0.406</td>
</tr>
<tr>
<td>Initial Point 4</td>
<td>0.282</td>
<td>0.38</td>
<td>0.458</td>
<td>0.00296</td>
<td>0.491</td>
<td>0.376</td>
</tr>
<tr>
<td>Initial Point 5</td>
<td>0.405</td>
<td>0.418</td>
<td>0.354</td>
<td>0.484</td>
<td>0.00309</td>
<td>0.48</td>
</tr>
<tr>
<td>Initial Point 6</td>
<td>0.403</td>
<td>0.353</td>
<td>0.422</td>
<td>0.331</td>
<td>0.503</td>
<td>0.00295</td>
</tr>
</tbody>
</table>
Figure 8: Different Bregman divergences between all the final points and all the initial points for different mirrors in MNIST dataset using a standard CNN. Note that the smallest element in every single row is on the diagonal, which confirms the theoretical results.
B.2 CIFAR-10 Experiments

B.2.1 Closest Minimum for Different Mirror Descents with Fixed Initialization

We provide the distances from final points (global minima) obtained by different algorithms from the same initialization, measured in different Bregman divergences for CIFAR-10 classification task using ResNet-18. Note that in all tables the smallest element in each row is on the diagonal, which means the point achieved by each mirror has the smallest Bregman divergence to the initialization corresponding to that mirror, among all mirrors. Tables 13, 14, 15, 16, 17, 18, 19, 20 depict these results for 8 different initializations. The rows are the distance metrics used as the Bregman Divergences with specified potentials. The columns are the global minima obtained using specified SMD algorithms.

<table>
<thead>
<tr>
<th>Table 13: CIFAR-10 Initial Point 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMD 1-norm</td>
</tr>
<tr>
<td>1-norm BD</td>
</tr>
<tr>
<td>2-norm BD</td>
</tr>
<tr>
<td>3-norm BD</td>
</tr>
<tr>
<td>10-norm BD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 14: CIFAR-10 Initial Point 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMD 1-norm</td>
</tr>
<tr>
<td>1-norm BD</td>
</tr>
<tr>
<td>2-norm BD</td>
</tr>
<tr>
<td>3-norm BD</td>
</tr>
<tr>
<td>10-norm BD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 15: CIFAR-10 Initial Point 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMD 1-norm</td>
</tr>
<tr>
<td>1-norm BD</td>
</tr>
<tr>
<td>2-norm BD</td>
</tr>
<tr>
<td>3-norm BD</td>
</tr>
<tr>
<td>10-norm BD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 16: CIFAR-10 Initial Point 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMD 1-norm</td>
</tr>
<tr>
<td>1-norm BD</td>
</tr>
<tr>
<td>2-norm BD</td>
</tr>
<tr>
<td>3-norm BD</td>
</tr>
<tr>
<td>10-norm BD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 17: CIFAR-10 Initial Point 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMD 1-norm</td>
</tr>
<tr>
<td>1-norm BD</td>
</tr>
<tr>
<td>2-norm BD</td>
</tr>
<tr>
<td>3-norm BD</td>
</tr>
<tr>
<td>10-norm BD</td>
</tr>
</tbody>
</table>
Table 18: CIFAR-10 Initial Point 6

<table>
<thead>
<tr>
<th></th>
<th>SMD 1-norm</th>
<th>SMD 2-norm (SGD)</th>
<th>SMD 3-norm</th>
<th>SMD 10-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-norm BD</td>
<td>128</td>
<td>9.25 × 10³</td>
<td>4.25 × 10³</td>
<td>2.34 × 10⁷</td>
</tr>
<tr>
<td>2-norm BD</td>
<td>2.71 × 10³</td>
<td>558</td>
<td>1.29 × 10³</td>
<td>6.89 × 10³</td>
</tr>
<tr>
<td>3-norm BD</td>
<td>3.34 × 10⁴</td>
<td>104</td>
<td>55.3</td>
<td>1.85 × 10²</td>
</tr>
<tr>
<td>10-norm BD</td>
<td>2.61 × 10⁵³</td>
<td>612</td>
<td>4.74 × 10⁻⁵</td>
<td>2.62 × 10⁻⁸</td>
</tr>
</tbody>
</table>

Table 19: CIFAR-10 Initial Point 7

<table>
<thead>
<tr>
<th></th>
<th>SMD 1-norm</th>
<th>SMD 2-norm (SGD)</th>
<th>SMD 3-norm</th>
<th>SMD 10-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-norm BD</td>
<td>223</td>
<td>9.76 × 10³</td>
<td>4.38 × 10³</td>
<td>2.27 × 10⁷</td>
</tr>
<tr>
<td>2-norm BD</td>
<td>2.41 × 10³</td>
<td>599</td>
<td>1.37 × 10³</td>
<td>6.65 × 10³</td>
</tr>
<tr>
<td>3-norm BD</td>
<td>2.3 × 10⁴</td>
<td>116</td>
<td>61</td>
<td>1.78 × 10²</td>
</tr>
<tr>
<td>10-norm BD</td>
<td>4.22 × 10¹²</td>
<td>679</td>
<td>6.42 × 10⁻⁵</td>
<td>2.55 × 10⁻⁸</td>
</tr>
</tbody>
</table>

Table 20: CIFAR-10 Initial Point 8

<table>
<thead>
<tr>
<th></th>
<th>SMD 1-norm</th>
<th>SMD 2-norm (SGD)</th>
<th>SMD 3-norm</th>
<th>SMD 10-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-norm BD</td>
<td>145</td>
<td>9.37 × 10³</td>
<td>4.17 × 10³</td>
<td>2.36 × 10⁷</td>
</tr>
<tr>
<td>2-norm BD</td>
<td>2.48 × 10³</td>
<td>576</td>
<td>1.26 × 10³</td>
<td>6.99 × 10³</td>
</tr>
<tr>
<td>3-norm BD</td>
<td>2.85 × 10⁴</td>
<td>108</td>
<td>54.5</td>
<td>1.89 × 10²</td>
</tr>
<tr>
<td>10-norm BD</td>
<td>1.81 × 10⁹³</td>
<td>1.22 × 10³</td>
<td>5.2 × 10⁻⁵</td>
<td>2.64 × 10⁻⁸</td>
</tr>
</tbody>
</table>

B.2.2 Closest Minimum for Different Initializations with Fixed Mirror

We provide the pairwise distances between different initial points and the final points (global minima) obtained by using fixed SMD algorithms in CIFAR-10 dataset using ResNet-18. Note that the smallest element in each row is on the diagonal, which means the closest final point to each initialization, among all the final points, is the one corresponding to that point. Tables 21, 22, 23, 24 depict these results for 4 different SMD algorithms. The rows are the initial points and the columns are the final points corresponding to each initialization.
Table 21: CIFAR-10 1-norm Bregman Divergence Between the Initial Points and the Final Points obtained by SMD 1-norm

<table>
<thead>
<tr>
<th>Initial</th>
<th>Final 1</th>
<th>Final 2</th>
<th>Final 3</th>
<th>Final 4</th>
<th>Final 5</th>
<th>Final 6</th>
<th>Final 7</th>
<th>Final 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial 1</td>
<td>$1.9 \times 10^4$</td>
<td>$8.1 \times 10^4$</td>
<td>$8.1 \times 10^4$</td>
<td>$8.4 \times 10^4$</td>
<td>$8 \times 10^4$</td>
<td>$8.2 \times 10^4$</td>
<td>$7.8 \times 10^4$</td>
<td>$7.8 \times 10^4$</td>
</tr>
<tr>
<td>Initial 2</td>
<td>$8.1 \times 10^4$</td>
<td>$2.7 \times 10^2$</td>
<td>$8.1 \times 10^4$</td>
<td>$8.3 \times 10^4$</td>
<td>$8 \times 10^4$</td>
<td>$8.2 \times 10^4$</td>
<td>$7.8 \times 10^4$</td>
<td>$7.9 \times 10^4$</td>
</tr>
<tr>
<td>Initial 3</td>
<td>$8.1 \times 10^4$</td>
<td>$8.1 \times 10^4$</td>
<td>$1.4 \times 10^2$</td>
<td>$8.4 \times 10^4$</td>
<td>$8 \times 10^4$</td>
<td>$8.1 \times 10^4$</td>
<td>$7.8 \times 10^4$</td>
<td>$7.8 \times 10^4$</td>
</tr>
<tr>
<td>Initial 4</td>
<td>$8.1 \times 10^4$</td>
<td>$8.1 \times 10^4$</td>
<td>$8.1 \times 10^4$</td>
<td>$2.5 \times 10^2$</td>
<td>$8 \times 10^4$</td>
<td>$8.2 \times 10^4$</td>
<td>$7.8 \times 10^4$</td>
<td>$7.9 \times 10^4$</td>
</tr>
<tr>
<td>Initial 5</td>
<td>$8.1 \times 10^4$</td>
<td>$8.1 \times 10^4$</td>
<td>$8.1 \times 10^4$</td>
<td>$8.3 \times 10^4$</td>
<td>$1.1 \times 10^2$</td>
<td>$8.1 \times 10^4$</td>
<td>$7.8 \times 10^4$</td>
<td>$7.8 \times 10^4$</td>
</tr>
<tr>
<td>Initial 6</td>
<td>$8.1 \times 10^4$</td>
<td>$8.1 \times 10^4$</td>
<td>$8.1 \times 10^4$</td>
<td>$8.4 \times 10^4$</td>
<td>$8 \times 10^4$</td>
<td>$1.3 \times 10^2$</td>
<td>$7.8 \times 10^4$</td>
<td>$7.8 \times 10^4$</td>
</tr>
<tr>
<td>Initial 7</td>
<td>$8.1 \times 10^4$</td>
<td>$8.1 \times 10^4$</td>
<td>$8.1 \times 10^4$</td>
<td>$8.4 \times 10^4$</td>
<td>$8 \times 10^4$</td>
<td>$8.1 \times 10^4$</td>
<td>$2.2 \times 10^2$</td>
<td>$7.8 \times 10^4$</td>
</tr>
<tr>
<td>Initial 8</td>
<td>$8.1 \times 10^4$</td>
<td>$8.1 \times 10^4$</td>
<td>$8.1 \times 10^4$</td>
<td>$8.4 \times 10^4$</td>
<td>$7.9 \times 10^4$</td>
<td>$8.1 \times 10^4$</td>
<td>$7.8 \times 10^4$</td>
<td>$1.5 \times 10^2$</td>
</tr>
</tbody>
</table>

Table 22: CIFAR-10 2-norm Bregman Divergence Between the Initial Points and the Final Points obtained by SMD 2-norm (SGD)

<table>
<thead>
<tr>
<th>Initial</th>
<th>Final 1</th>
<th>Final 2</th>
<th>Final 3</th>
<th>Final 4</th>
<th>Final 5</th>
<th>Final 6</th>
<th>Final 7</th>
<th>Final 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial 1</td>
<td>$6 \times 10^2$</td>
<td>$2.9 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
</tr>
<tr>
<td>Initial 2</td>
<td>$2.8 \times 10^3$</td>
<td>$6.1 \times 10^2$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
</tr>
<tr>
<td>Initial 3</td>
<td>$2.8 \times 10^3$</td>
<td>$2.9 \times 10^3$</td>
<td>$5.6 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
</tr>
<tr>
<td>Initial 4</td>
<td>$2.8 \times 10^3$</td>
<td>$2.9 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$5.9 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
</tr>
<tr>
<td>Initial 5</td>
<td>$2.8 \times 10^3$</td>
<td>$2.9 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$5.7 \times 10^2$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
</tr>
<tr>
<td>Initial 6</td>
<td>$2.8 \times 10^3$</td>
<td>$2.9 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$5.6 \times 10^2$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
</tr>
<tr>
<td>Initial 7</td>
<td>$2.8 \times 10^3$</td>
<td>$2.9 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$6 \times 10^2$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
</tr>
<tr>
<td>Initial 8</td>
<td>$2.8 \times 10^3$</td>
<td>$2.9 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$2.8 \times 10^3$</td>
<td>$5.8 \times 10^2$</td>
<td>$2.8 \times 10^3$</td>
</tr>
</tbody>
</table>

Table 23: CIFAR-10 3-norm Bregman Divergence Between the Initial Points and the Final Points obtained by SMD 3-norm

<table>
<thead>
<tr>
<th>Initial</th>
<th>Final 1</th>
<th>Final 2</th>
<th>Final 3</th>
<th>Final 4</th>
<th>Final 5</th>
<th>Final 6</th>
<th>Final 7</th>
<th>Final 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial 1</td>
<td>55.844</td>
<td>103.47</td>
<td>103.61</td>
<td>104.05</td>
<td>106.2</td>
<td>105.32</td>
<td>110.88</td>
<td>104.56</td>
</tr>
<tr>
<td>Initial 2</td>
<td>105.87</td>
<td>53.455</td>
<td>103.68</td>
<td>104.04</td>
<td>106.31</td>
<td>105.34</td>
<td>110.93</td>
<td>104.58</td>
</tr>
<tr>
<td>Initial 3</td>
<td>105.89</td>
<td>103.59</td>
<td>53.527</td>
<td>104.09</td>
<td>106.29</td>
<td>105.35</td>
<td>110.99</td>
<td>104.55</td>
</tr>
<tr>
<td>Initial 4</td>
<td>105.83</td>
<td>103.54</td>
<td>103.64</td>
<td>53.978</td>
<td>105.23</td>
<td>105.3</td>
<td>110.87</td>
<td>104.54</td>
</tr>
<tr>
<td>Initial 5</td>
<td>105.82</td>
<td>103.55</td>
<td>103.64</td>
<td>104</td>
<td>56.161</td>
<td>105.34</td>
<td>110.88</td>
<td>104.55</td>
</tr>
<tr>
<td>Initial 6</td>
<td>105.91</td>
<td>103.6</td>
<td>103.66</td>
<td>104.1</td>
<td>106.28</td>
<td>55.316</td>
<td>110.94</td>
<td>104.55</td>
</tr>
<tr>
<td>Initial 7</td>
<td>105.87</td>
<td>103.51</td>
<td>103.67</td>
<td>103.98</td>
<td>106.26</td>
<td>105.25</td>
<td>61.045</td>
<td>104.5</td>
</tr>
<tr>
<td>Initial 8</td>
<td>105.77</td>
<td>103.54</td>
<td>103.59</td>
<td>104.04</td>
<td>106.25</td>
<td>105.28</td>
<td>110.88</td>
<td>54.509</td>
</tr>
</tbody>
</table>
B.2.3 Closest Minimum for Different Initializations and Different Mirrors

Now we assess the pairwise distances between different initial points and final points (global minima) obtained by all different initializations and all different mirrors (Table 8). The smallest element in each row is exactly the final point obtained by that mirror from that initialization, among all the mirrors and all the initial points.

<table>
<thead>
<tr>
<th>Initial</th>
<th>Final 1</th>
<th>Final 2</th>
<th>Final 3</th>
<th>Final 4</th>
<th>Final 5</th>
<th>Final 6</th>
<th>Final 7</th>
<th>Final 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial 1</td>
<td>$2.64 \times 10^{-8}$</td>
<td>$2.89 \times 10^{-8}$</td>
<td>$2.99 \times 10^{-8}$</td>
<td>$2.81 \times 10^{-8}$</td>
<td>$2.85 \times 10^{-8}$</td>
<td>$2.82 \times 10^{-8}$</td>
<td>$2.66 \times 10^{-8}$</td>
<td>$2.82 \times 10^{-8}$</td>
</tr>
<tr>
<td>Initial 2</td>
<td>$2.79 \times 10^{-8}$</td>
<td>$2.65 \times 10^{-8}$</td>
<td>$2.83 \times 10^{-8}$</td>
<td>$2.83 \times 10^{-8}$</td>
<td>$2.71 \times 10^{-8}$</td>
<td>$2.74 \times 10^{-8}$</td>
<td>$2.69 \times 10^{-8}$</td>
<td>$2.88 \times 10^{-8}$</td>
</tr>
<tr>
<td>Initial 3</td>
<td>$2.89 \times 10^{-8}$</td>
<td>$2.87 \times 10^{-8}$</td>
<td>$2.72 \times 10^{-8}$</td>
<td>$2.94 \times 10^{-8}$</td>
<td>$2.84 \times 10^{-8}$</td>
<td>$2.89 \times 10^{-8}$</td>
<td>$2.78 \times 10^{-8}$</td>
<td>$2.94 \times 10^{-8}$</td>
</tr>
<tr>
<td>Initial 4</td>
<td>$2.79 \times 10^{-8}$</td>
<td>$2.86 \times 10^{-8}$</td>
<td>$2.92 \times 10^{-8}$</td>
<td>$2.67 \times 10^{-8}$</td>
<td>$2.84 \times 10^{-8}$</td>
<td>$2.81 \times 10^{-8}$</td>
<td>$2.69 \times 10^{-8}$</td>
<td>$2.85 \times 10^{-8}$</td>
</tr>
<tr>
<td>Initial 5</td>
<td>$2.76 \times 10^{-8}$</td>
<td>$2.88 \times 10^{-8}$</td>
<td>$2.95 \times 10^{-8}$</td>
<td>$2.93 \times 10^{-8}$</td>
<td>$2.61 \times 10^{-8}$</td>
<td>$2.73 \times 10^{-8}$</td>
<td>$2.66 \times 10^{-8}$</td>
<td>$2.83 \times 10^{-8}$</td>
</tr>
<tr>
<td>Initial 6</td>
<td>$2.80 \times 10^{-8}$</td>
<td>$2.76 \times 10^{-8}$</td>
<td>$2.93 \times 10^{-8}$</td>
<td>$2.79 \times 10^{-8}$</td>
<td>$2.76 \times 10^{-8}$</td>
<td>$2.62 \times 10^{-8}$</td>
<td>$2.71 \times 10^{-8}$</td>
<td>$2.85 \times 10^{-8}$</td>
</tr>
<tr>
<td>Initial 7</td>
<td>$2.73 \times 10^{-8}$</td>
<td>$2.76 \times 10^{-8}$</td>
<td>$2.82 \times 10^{-8}$</td>
<td>$2.79 \times 10^{-8}$</td>
<td>$2.71 \times 10^{-8}$</td>
<td>$2.77 \times 10^{-8}$</td>
<td>$2.55 \times 10^{-8}$</td>
<td>$2.83 \times 10^{-8}$</td>
</tr>
<tr>
<td>Initial 8</td>
<td>$2.73 \times 10^{-8}$</td>
<td>$2.79 \times 10^{-8}$</td>
<td>$2.85 \times 10^{-8}$</td>
<td>$2.78 \times 10^{-8}$</td>
<td>$2.75 \times 10^{-8}$</td>
<td>$2.74 \times 10^{-8}$</td>
<td>$2.73 \times 10^{-8}$</td>
<td>$2.84 \times 10^{-8}$</td>
</tr>
</tbody>
</table>
Figure 9: Different Bregman divergences between all the final points and all the initial points for different mirrors in CIFAR-10 dataset using ResNet-18. Note that the smallest element in every single row is on the diagonal, which confirms the theoretical results.
Figure 10: An illustration of the experimental results. For each initialization $w_0$, we ran different SMD algorithms until convergence to a point on the set $\mathcal{W}$ (zero training error). We then measured all the pairwise distances from different $w_{\infty}$ to different $w_0$, in different Bregman divergences. The closest point (among all initializations and all mirrors) to any particular initialization $w_0$ in Bregman divergence with potential $\psi(\cdot) = \|\cdot\|_q^q$ is exactly the point obtained by running SMD with potential $\|\cdot\|_q^q$ from $w_0$.

B.3 DISTRIBUTION OF THE FINAL WEIGHTS OF THE NETWORK

One may be curious to see how the final weights obtained by these different mirrors look like, and whether, for example, mirror descent corresponding to the $\ell_1$-norm potential induces sparsity. We examine the distribution of the weights in the network for these algorithms starting from the same initialization. Fig. 11 shows the histogram of the initial weights, which follows a half-normal distribution. Figs. 12 (a), (b), (c), (d) show the histogram of the weights for $\ell_1$-SMD, $\ell_2$-SMD (SGD), $\ell_3$-SMD, and $\ell_{10}$-SMD, respectively. Note that each of the four histograms corresponds to an $11 \times 10^6$-dimensional weight vector that perfectly interpolates the data. Even though, perhaps as expected, the weights remain quite small, the histograms are drastically different. The histogram of the $\ell_1$-SMD has more weights at and close to zero, which again confirms that it induces sparsity. However, as will be shown in the next section, this is not necessarily good for generalization (in fact, it turns out that $\ell_{10}$-SMD has a much better generalization). The histogram of the $\ell_2$-SMD (SGD) looks almost identical to the histogram of the initialization, whereas the $\ell_3$ and $\ell_{10}$ histograms are shifted to the right, so much so that almost all weights in the $\ell_{10}$ solution are non-zero and in the range of 0.005 to 0.04. For comparison, all the distributions are shown together in Fig. 12(e).

Figure 11: Histogram of the absolute value of the initial weights in the network (half-normal distribution)
Figure 12: Histogram of the absolute value of the final weights in the network for different SMD algorithms: (a) $\ell_1$-SMD, (b) $\ell_2$-SMD (SGD), (c) $\ell_3$-SMD, (d) $\ell_{10}$-SMD. Note that each of the four histograms corresponds to an $11 \times 10^6$-dimensional weight vector that perfectly interpolates the data. Even though the weights remain quite small, the histograms are drastically different. $\ell_1$-SMD induces sparsity on the weights, as expected. SGD does not seem to change the distribution of the weights significantly. $\ell_3$-SMD starts to reduce the sparsity, and $\ell_{10}$ shifts the distribution of the weights significantly, so much so that almost all the weights are non-zero.
B.4 Generalization Errors of Different Mirrors/Regularizers

In this section, we compare the performance of the SMD algorithms discussed before on the test set. This is important for understanding the effect of different regularizers on the generalization of deep networks.

For MNIST, perhaps not surprisingly, all the four SMD algorithms achieve around 99% or higher accuracy. For CIFAR-10, however, there is a significant difference between the test errors of different mirrors/regularizers on the same architecture. Fig. 13 shows the test accuracies of different algorithms with eight random initializations around zero, as discussed before. Counter-intuitively, $\ell_{10}$ performs consistently best, while $\ell_1$ performs consistently worse. We should reiterate that the loss function is exactly the same in all these experiments, and all of them have been trained to fit the training set perfectly (zero training error). Therefore, the difference in generalization errors is purely the effect of implicit regularization by different algorithms. This result suggests the importance of a comprehensive study of the role of regularization, and the choice of the best regularizer, to improve the generalization performance of deep neural networks.

Figure 13: Generalization performance of different SMD algorithms on the CIFAR-10 dataset using ResNet-18. $\ell_{10}$ performs consistently better, while $\ell_1$ performs consistently worse.